Lecture Notes in Electrical Engineering 170

Haeng Kon Kim
Sio-Iong Ao
Burghard B. Rieger *Editors*

IAENG Transactions on Engineering Technologies

Special Edition of the World Congress on Engineering and Computer Science 2011

Springer

Lecture Notes in Electrical Engineering

Volume 170

For further volumes:
http://www.springer.com/series/7818

Haeng Kon Kim · Sio-Iong Ao
Burghard B. Rieger
Editors

IAENG Transactions on Engineering Technologies

Special Edition of the World Congress on Engineering and Computer Science 2011

Springer

Editors
Haeng Kon Kim
Engineering College,
School of IT Engineering
Catholic University of DaeGu
DaeGu
Republic of South Korea

Burghard B. Rieger
FB II Linguistische Datenverarbeitung,
Computerlinguistik
Universität Trier
Trier
Germany

Sio-Iong Ao
Unit 1, 1/F
International Association of Engineers
Hong Kong
Hong Kong, SAR

ISSN 1876-1100 ISSN 1876-1119 (electronic)
ISBN 978-94-007-4785-2 ISBN 978-94-007-4786-9 (eBook)
DOI 10.1007/978-94-007-4786-9
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012946626

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

A large international conference on Advances in Engineering Technologies and Physical Science was held in San Francisco, CA, USA, October 19−21, 2011, under the World Congress on Engineering and Computer Science (WCES 2011). The WCECS 2011 is organized by the International Association of Engineers (IAENG). IAENG is a non-profit international association for the engineers and the computer scientists, which was founded originally in 1968 and has been undergoing rapid expansions in recent few years. The WCECS congress serves as good platforms for the engineering community to meet with each other and to exchange ideas. The congress has also struck a balance between theoretical and application development. The conference committees have been formed with over two hundred committee members who are mainly research center heads, faculty deans, department heads, professors, and research scientists from over 30 countries with the full committee list available at our congress web site (http://www.iaeng.org/WCECS2011/committee.html). The congress is truly international meeting with a high level of participation from many countries. The response that we have received for the congress is excellent. There have been more than six hundred manuscript submissions for the WCECS 2011. All submitted papers have gone through the peer review process and the overall acceptance rate is 52.68 %.

This volume contains 30 revised and extended research articles written by prominent researchers participating in the conference. Topics covered include chemical engineering, circuits, communications systems, control theory, engineering mathematics, systems engineering, manufacture engineering, and industrial applications. The book offers the state of art of tremendous advances in engineering technologies and physical science and applications, and also serves as an excellent reference work for researchers and graduate students working with/on engineering technologies and physical science and applications.

Sio-Iong Ao
Haeng Kon Kim
Burghard B. Rieger

Contents

Chapter 1
Adaptive Three-Bit LDPC Decoder Quantization

Raymond Moberly, Michael E. O'Sullivan and Khurram Waheed

Abstract This chapter presents two related 3-bit quantizations for the sum-product algorithm that are suitable for an adaptive decoder implementation using programmable logic. Our decoder design combines the parity-check and variable-node-update steps into a single computation. The hardware requirements are considered and compared to the published work of Planjery et al. Decoder performance in the waterfall region is obtained by simulation.

Keywords Belief propagation · Finite precision · Iterative decoding · Low density parity check codes · Nonlinear quantization · Sum-product algorithm

1.1 Introduction

Low density parity check (LDPC) codes are well suited for error-correction applications. However, the challenge is to find strategies that will enable efficient implementations while ensuring good performance. Iterative decoder designs with limited-precision quantization, suitable for digital logic implementation, appear in the works of T. Zhang and Parhi [1], and Planjery et al. [2], and Z. Zhang et al. [3].

R. Moberly (✉) · M. E. O'Sullivan · K. Waheed
San Diego State University, 5500 Campanile Drive, San Diego, CA, USA
e-mail: raymond.moberly@ieee.org

M. E. O'Sullivan
e-mail: mosulliv@math.sdsu.edu

K. Waheed
e-mail: khurram.waheed@ieee.org

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_1,

In [4] we presented two 3-bit quantizations for a sum-product algorithm LDPC decoder on a Gaussian channel. This chapter expands upon the decoder design and refines the synthesis results. In our examinations of many 3-bit quantizations, our best choice of quantization changes as the channel conditions change. We propose an adaptive design that changes between our two selected quantizations based upon the channel condition.

Our experiments are with a rate-(1/2) length 1162 binary LDPC code; it is from a family of codes that our research group has generated using permutation matrices [5, 6]. The cyclic permutation structure is known to have efficient hardware implementations [7–9].

1.2 Scope

The sum product algorithm (SPA) was simulated on a computer cluster, using look-up tables based upon 3-bit quantization, for 10 iterations. We determine the per-iteration computational latency and evaluate trade-offs, between iterations and computation per-iteration, which contribute to total latency and total decoding gain. Our quantization, with 10 iterations, surpasses the performance of the decoder by Planjery et al. with 100 iterations. Gain versus latency is our comparison criteria, although we discuss other potential criteria. In an engineering application, the designer could attempt to maximize throughput or minimize power consumption.

We are particularly motivated to achieve low-latency decoding in the waterfall region. For voice communication and video streaming, a partial packet (with one or more uncorrected errors) is preferable to an entirely lost packet; and a correctly re-transmitted or excessively delayed out-of-order packet is useless. Moderate error rates (10^{-3}–10^{-5} BER) in the content coming out from the decoder are acceptable for these applications.

1.2.1 Circulant Permutation Matrix

The experiments are performed with a quasi-cyclic LDPC code constructed using a method that yields permutation-based parity-check matrices with large girth [5]; this particular graph is constructed with a girth of 10. Here σ is an 83-by-83 matrix of the form

$$\sigma = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

We can create circular shifts: σ^{κ} is the original submatrix circularly shifted by $\kappa-1$ positions, therefore σ^0 has the form of an identity matrix. H is a block parity-check matrix comprised of circulant permutations. The code, C, is the set of vectors in the null space of H. The 1162-bit length of C is in proximity to the lengths of LDPC codes in other research.

$$H = \begin{bmatrix} \sigma^0 & \sigma^0 & \sigma^0 & 0 & 0 & 0 & 0 & \sigma^0 & \sigma^0 & \sigma^0 & 0 & 0 & 0 & 0 \\ \sigma^0 & 0 & 0 & \sigma^0 & \sigma^0 & 0 & 0 & \sigma^1 & 0 & 0 & \sigma^0 & \sigma^0 & 0 & 0 \\ \sigma^0 & 0 & 0 & 0 & 0 & \sigma^0 & \sigma^0 & 0 & 0 & \sigma^2 & \sigma^3 & 0 & \sigma^0 & 0 \\ 0 & \sigma^0 & 0 & \sigma^2 & 0 & \sigma^4 & 0 & 0 & \sigma^9 & 0 & 0 & \sigma^{13} & \sigma^{16} & 0 \\ 0 & \sigma^0 & 0 & 0 & \sigma^5 & 0 & \sigma^1 & \sigma^{19} & 0 & 0 & 0 & 0 & \sigma^{11} & \sigma^0 \\ 0 & 0 & \sigma^0 & 0 & \sigma^6 & \sigma^{13} & 0 & 0 & \sigma^{32} & 0 & \sigma^{40} & 0 & 0 & \sigma^{30} \\ 0 & 0 & \sigma^0 & \sigma^7 & 0 & 0 & \sigma^{17} & 0 & 0 & \sigma^{26} & 0 & \sigma^{49} & 0 & \sigma^{53} \end{bmatrix}$$

1.2.2 FPGA Implementation

Because the field programmable gate array (FPGA) offers a very rapid pathway to concept verification, it fostered our exploration of the trade-offs between precision and computational speed. The application specific integrated circuit (ASIC) also offers customized precision and designer-defined variable data types that are not available in microprocessors, but at a high development cost. The FPGA synthesis results in this chapter, serve as an indicator of hardware complexity, size, and speed trade-offs; we anticipate that the comparisons of our designs in the FPGA domain would translate to proportional advantages in the ASIC domain. An FPGA solution [1] in the literature achieved LDPC decoding using operands with just 5-bits. Our own prior research [10] explored tradeoffs between the number of bits of precision and the number of decoding iterations. The regular LDPC decoder has a very repetitive structure; for our (6, 3)-regular code, each variable node outputs three update messages. We implemented the logic of one output message, determined the latency, and then implemented all three outputs in order to observe the consequent speed and size.

The Altera DE2 development board was selected for this work and requested from and provided to us by the Altera Corporation as a university research grant. The FPGA on the DE2 board is the Cyclone II EP2C35F672C6 N with 33,216 programmable logic elements.

1.2.3 Formulation of the Iterative Algorithm

Our quantization is applied to the computationally-efficient SPA formulation of [11]. Figure 1.1 shows the iterative algorithm formulations of three formulations of the SPA [12–15] that we analyzed in the ISIT 2006 paper. Each is illustrated

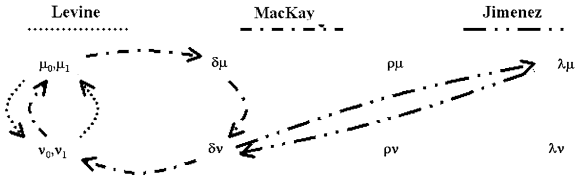


Fig. 1.1 Iterative formulations in the literature

cycling through probability representations, where the variable bit-to-check (μ) messages and the check-to-bit (v) messages can be expressed in terms of probabilities, differences $\delta p = P(0) - P(1)$, ratios $\rho p = P(0)/P(1)$, or log-likelihood ratios $\lambda p = log\ \rho p$. The δp representation transforms [0, 1] probability values to the range of [–1, +1].

Our two formulations, shown in Fig. 1.2, which represent probabilities as differences (δp) or as log-likelihood ratios (LLR) offered significant computational advantages by requiring fewer processor instructions [11]. Transforming multiplication operations into addition operations in the log domain also increases performance on computer processors with arithmetic logic units that can perform addition more rapidly than multiplication [16–20]. The differences diminish when only a few bits of precision are in use.

As Han and Sunwoo showed [21], the LLR calculations involve one particularly obstructive computation, an inverse hyperbolic tangent function; their limited-precision computation involves a lookup table for this calculation. Z. Zhang et al. have also looked at fixed-point LLR quantizations using 5, 6 and 7 bits [3]. In these implementations, the hyperbolic tangent function is a substantial part of the design effort and computational work. The algorithm formulations that we devised do not contain a hyperbolic tangent calculation.

In this paper, instead of looking at the parity check and variable-node update as two separate actions, we will present the cycle as a single computation with one quantization applied per iteration.

1.2.4 Comparing BSC and AWGN

This chapter compares decoding results on an additive white gaussian noise (AWGN) channel with competing published results that use the Binary Symmetric Channel (BSC). The BSC bit-crossover probability, α, can be determined from the Gaussian signal to noise ratio (SNR), E_b/N_0, by

$$\alpha = {}^1/_2 \operatorname{erfc}(\sqrt{2E_b/N_0}).$$

For decoders with floating-point belief propagation, there is an almost 2 dB difference in performance. As Fig. 1.3 shows, the difference is about the same for bit error rate (BER) and frame error rate (FER). These serve as reference curves; we have 2-bit limited-precision sigmoid-based quantizations that approach the

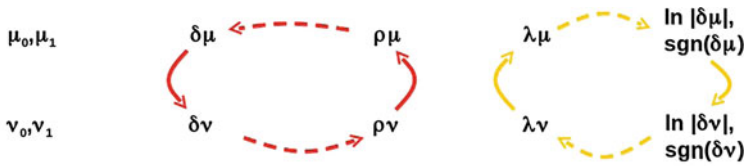


Fig. 1.2 Our proposed formulations of the sum-product algorithm

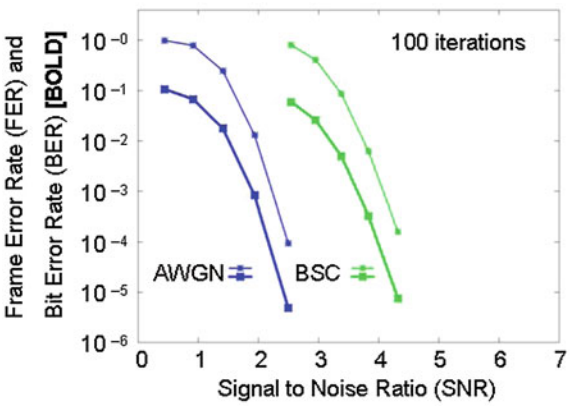


Fig. 1.3 Decoding difference between AWGN and BSC channels, using C

BSC curve, and the 3-bit quantizations presented in this chapter surpass it. Considering the 2 dB loss, it seems appropriate to receive soft decisions when there is soft-information in the decoder. Our decoder design assumes a soft-decision receiver.

1.3 Planjery's Beyond Belief Propagation

Planjery et al. devised 2-bit and 3-bit quantized decoding designs for the decoding of LPDC codes on the BSC. These algorithms begin with a single bit quantization (a hard decision) at the receiver. Another quantization occurs at each parity check, and messages are quantized at each variable-node update. Other algorithms in the literature quantize in a similar two-quantizations-per-iteration fashion; as illustrated in Fig. 1.4.

We replicated the quantized 3-bit algorithm specified in Planjery's paper [2]. To verify our implementation, we replicated their decoding results (100 iterations) using the published codes (benchmarks) that they used for testing. We then ran simulations upon C, with both 10 and 100 iterations. The results are the two upper curves of Figs. 1.5 (BER) and 1.6 (FER).

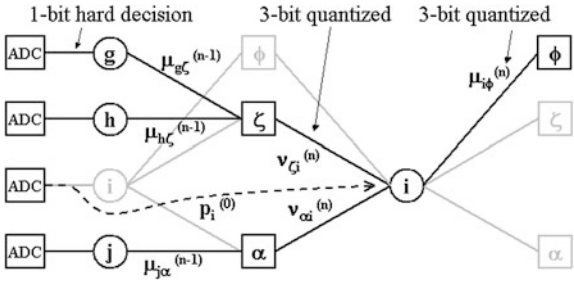


Fig. 1.4 Quantization of the variable nodes and the parity computation

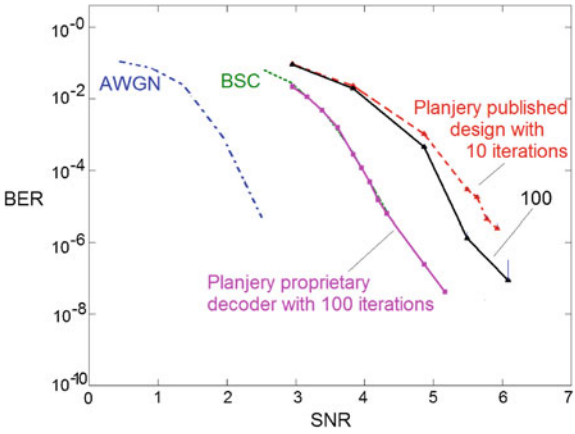


Fig. 1.5 BER for the published and proprietary decoders of Planjery et al., using *C*

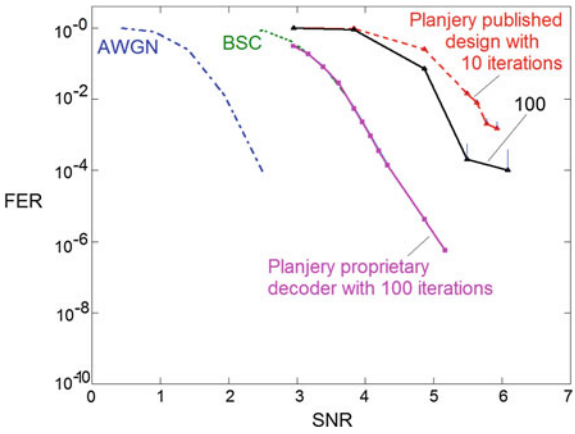


Fig. 1.6 FER for the published and proprietary decoders of Planjery et al., using *C*

Planjery also produced, using a specialized 3-bit proprietary quantization and algorithm, improved results through an approach designed to overcome the influence of trapping sets. With Shiva Planjery's gracious cooperation we were able to obtain the resulting performance curve of their proprietary decoder applied to the LDPC

code that came from our own permutation construction. Transformed from their crossover probability to our SNR axis, this curve is shown in Figs. 1.5 (BER) and 1.6 (FER) and repeated in Figs. 1.10 and 1.11 as a comparison for our quantizations.

1.3.1 Synthesis of the Planjery-Vasic 3-bit Decoder

We implemented the published 3-bit logic in Verilog HDL. The synthesis results, targeting our Cyclone II FPGA, were reported by the Altera Quartus II software, giving a baseline for the cost of their published algorithm. The single bit computation used 138 logic elements and had a longest path delay of 20.489 ns. If we were to compute 1162 bits (the length of our LDPC code) simultaneously, the footprint would expand to 160356 logic elements. If we were to compute, sequentially, the 100 iterations used in Planjery and Vasic's simulations, the decoding latency would be multiplied to 2.0489 microseconds. We programmed this design into the DE2 board for verification and demonstration.

Their second stage proprietary rule accounts for about 1.5 dB additional decoding gain and it increases the implementation logic and latency by an amount unknown to us. The quantizations that we propose in the following sections require more logic elements than their baseline, but our performance results show a great return from the additional logic.

1.4 Our Quantization Work

This section explains how and where quantization is applied within the algorithm, what quantizations we chose to use, and the results that we obtained. We start from the $\delta\mu$ SPA formulation we proposed in [11].

1.4.1 One Computation per Iteration

The SPA is typically described as two computational steps. We treat the iteration as a combined-step instead of the two separate steps; quantization is applied once rather than twice in an iteration. The intermediate parity-check values are indirectly quantized, but not specifically by the design. Figure 1.7 illustrates the whole-iteration computation.

1.4.2 Quantization Scales

Our formulation and quantization values are expressed in $\delta\mu$ representation. Several 5-bit quantizations proved to be very effective in LDPC decoding in our previous

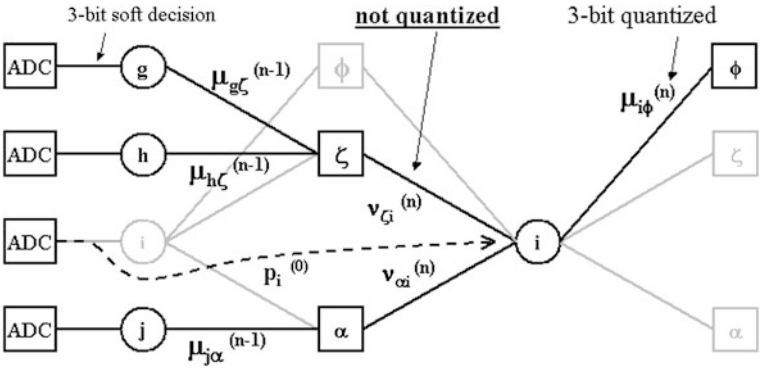


Fig. 1.7 Quantization of the variable nodes. One quantization per iteration

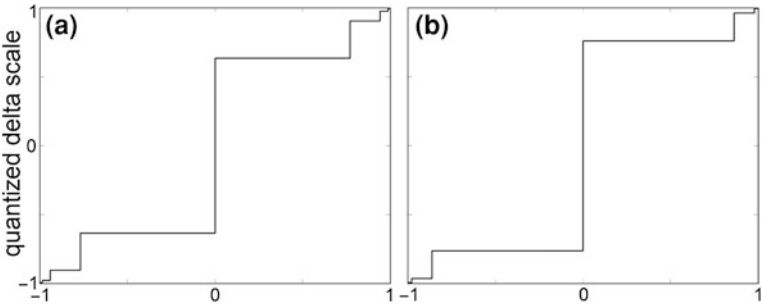


Fig. 1.8 Sigmoid-based quantizations (a) "635" and (b) "762"

effort [10]. Among the quantization schemes tested was one using the sigmoid function, $S(x) = 1/(1 + e^x)$. In [4] we presented two related 3-bit sigmoid-based quantizations, using sigmoid function evaluations at certain intervals to determine the discrete scale values $S(x)$: $x = \pm 1.5$ $n = \pm \{1.5,3.0,4.5,6.0\}$ and $x = \pm 2.0$ $n = \pm \{2.0,4.0,6.0,8.0\}$. These show particular promise for decoder quantization over a tested range of Gaussian channel SNR values.

The chosen step thresholds are the means between the step heights. The step-function mapping of δp assigns the quantized value s_i, choosing i such that $t_{i-1} \leq \delta p \leq t_i$. The two tested quantization scales are titled the "635" sigmoid-based quantization, illustrated in Fig. 1.8a, and the "762" sigmoid-based quantization, illustrated in Fig. 1.8b. Step and threshold values for both quantizations are given in Tables 1.1 and 1.2.

Notice how, for both scales, the precision is concentrated in the regions of greatest certainty; the step functions have finely-spaced steps at the two extremes. This family of quantizations suggest an implementation strategy for varying the decoder precision; such a strategy could compete with other adaptive error correction technologies that have been developed (rate compatible codes, etc.).

Table 1.1 Step values "S" ($\delta\mu$) for the "635" and "762" quantizations

	$-S_4$	$-S_3$	$-S_2$	$-S_1$	S_1	S_2	S_3	S_4
"635"	−0.995	−0.98	−0.90	−0.64	0.64	0.90	0.98	0.995
"762"	−0.999	−0.995	−0.96	−0.76	0.76	0.96	0.995	0.999

Table 1.2 Threshold values "T" for "635" and "762" quantizations

	$-T_3$	$-T_2$	$-T_1$	T_0	T_1	T_2	T_3
"635"	−0.99	−0.95	−0.77	0.0	0.77	0.95	0.99
"762"	−0.99	−0.98	−0.86	0.0	0.86	0.98	0.99

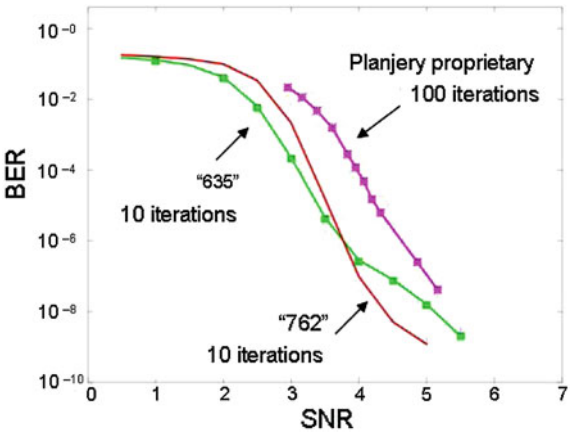


Fig. 1.9 BER for our sigmoid-based "635" and "762" quantizations, using *C*

The two quantizations tested differ only in how the x values of the sigmoid $S(x)$ are selected, their similarities might simplify the implementation of an adaptive design offering both quantizations.

1.4.3 Decoder Performance

One of our quantization scales was better for low SNR conditions and the other was better for high SNR conditions, as SPA simulation results show in Figs. 1.9 (BER) and 1.10 (FER). Our quantized designs were tested with 10 iterations; increasing to 100 iterations resulted in only minor additional gains (1/4 dB BER, 1/3 dB FER). The graphs show comparable results from a simulation by Planjery, using their proprietary decoder upon our code C.

The small thin vertical bars on the graphs show the upper end of a 95 % confidence interval for each of our simulation result values. These confidence intervals can be reduced with longer simulations (more samples). The confidence

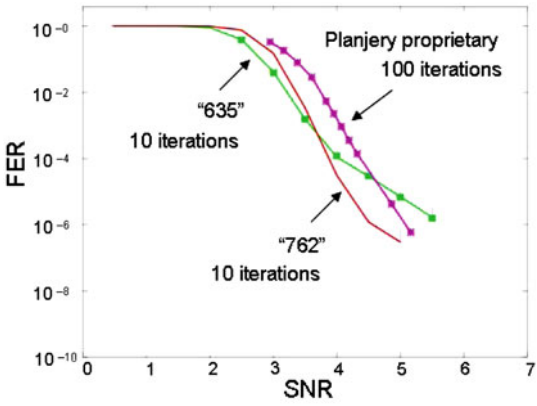


Fig. 1.10 FER for our sigmoid-based "635" and "762" quantizations, using C

intervals that we present are small enough to firmly assert the claims that: *"762" outperforms "635" over the [4.0, 5.0] SNR range* and *"635" outperforms "762" over the [1.0, 3.5] SNR range.*

The BER gain is about 0.9 dB better than the Planjery and Vasic proprietary algorithm over a substantial range. Somewhat less substantial FER gains, around 0.5 dB, are also seen over most of the tested SNR region. A design adapting between our two quantizations outperforms their approach over the entire tested range.

1.4.4 Synthesis Results

In our quantization approach, as described above, limited precision is applied to the receiver sampling and to the variable-node updates. Using this, we implemented a combined parity-check and variable-node-update calculation using a mixture of calculations, logic, and a lookup table. The 3-bit inputs into each (6,3) parity check yield one of 112 possible values (that is far less than the $2^{(3\times5)}$ apparent input combinations). Another way to express this is as an imputed quantization—the parity check output requires no more than seven bits, since $112 < 2^7$. Two parity checks and the original sample factor into the update calculation, specified as a $112 \times 112 \times 8$ lookup table. Additional symmetries make it unnecessary to implement this complete table. Our technique for finding the simplifications was to allow the Altera Quartus II synthesis tool to do the simplifying for us. For our tested quantizations, the tool consistently digested the lookup table (specified in Verilog HDL) and produced a result with a complexity reduced by a factor of about 1000. The cost for each effort was an overnight (8 1/2 h) run of the Quartus II synthesis, place, and routing tool.

The tool returns the number of logic elements (LE), which are required for the design and it computes, after placing and routing in an optimal manner, the longest path delay (LPD) between any pair among the inputs and outputs. The inverse of

Table 1.3 Synthesis results for each quantization

	msgs	LEs	LPD (ns)	Adjusted LPD (ns)
Planjery's algorithm	3	138	20.489	8.928
Sigmoid "635" Scale	1	4,743	36.255	24.694
	3	11,111	43.099	31.538
Sigmoid "762" Scale	1	4,471	37.518	25.957
	3	10,070	42.485	30.924

the LPD is the highest appropriate clock frequency for the logic when used in a clock-synchronous design. The synthesis results for our two quantizations are reported in Table 1.3.

Calculating one variable update using two associated parity checks, synthesized to less than 5,000 logic elements. When the expressed design was expanded to include all three associated parity checks and compute all three of the resulting variable node updates, the design footprint more than doubled, but it did not triple and the delay increased by less than 20 %. We can deduce that the three-message logic synthesized to a blend of shared computation and parallelism.

The chosen Cyclone II FPGA is too small for the 1162 replications of this design needed to handle all of the bits of a code word simultaneously. With limited parallelization [7] or serial implementation [9] a complete FPGA-based decoder is still entirely feasible.

The LPD figures include some amount of input/output (I/O) delay that is characteristic of the FPGA. Since a multiple iteration decoding operation might be able to omit I/O between iterations, we sought to isolate this contribution. Building one simple model with a single exclusive-or (XOR) gate and another design with a cascade of two XOR gates, we determined from an extrapolation of the two design's LDP values the contribution of the I/O to be 11.561 ns. Adjusted LPD figures are shown in the rightmost column of Table 1.3.

1.5 Comparing Decoders

Our synthesized designs have three to four times the adjusted per-iteration latency of Planjery's published design (per our implementation of their design and our synthesis results). Since our decoder exceeds, in 10 iterations, the decoding gain of Planjery's proprietary decoder with 100 iterations, we compute the total decoding time for one bit to be $10 \times 31.538 = 315.38$ ns for our design and $100 \times 8.928 = 892.8$ ns for Planjery's published design. The timing advantage of our decoder, having accounted for a worst-case FPGA I/O contribution, is at least 65 %.

The logic circuitry of our decoder, with its quantizations, was larger than the logic to implement their decoder, but our decoding operation was faster and obtains better decoding results for the tested ranges of SNR, BER and FER. Our

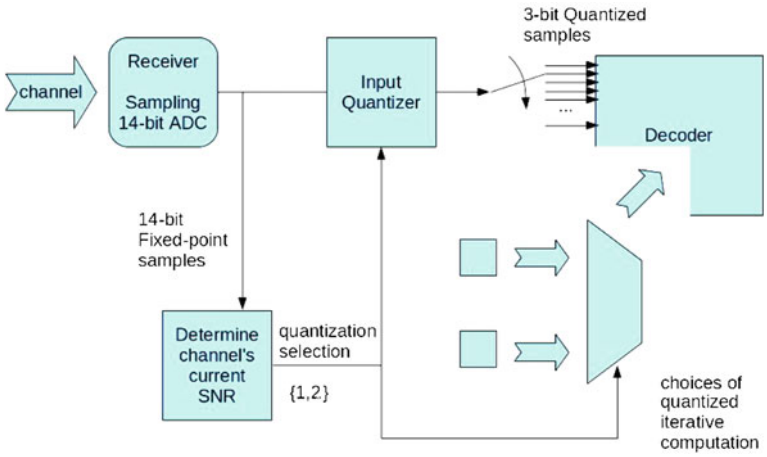


Fig. 1.11 Adaptive decoder that uses the "635" and "762" sigmoid quantizations

Table 1.4 Design comparisons

	Our design	Planjery's design	
		Published	Proprietary
Decode 1 Bit (ns)	315.3	892.8	–
dB gain @ 10^{-4} BER	+8.5	+6.5	+7.6
Chip Area (LEs)	21,181	138	–

computation for one code symbol fits within the selected FPGA; and we could readily use this chip to decode a full codeword in a serial fashion. Alternatively, we could increase throughput by using a larger chip or by adapting this design for an ASIC. Using a larger chip would give us greater throughput and parallelization opportunities; these can be explored more thoroughly under the engineering constraints of a specific application.

Our results, using 3-bit samples from a Gaussian channel, have 0.5–0.9 dB better gain than the hard-decision receiver approach used by Planjery et al. [2]. A conclusion from this is that a receiver that can sample incoming symbols with three bits is better than one that makes a hard-decision. The fidelity available at the receiver sampling point should not be discarded. The quantization selected for 3-bits of precision does make a difference; considering the channel conditions is necessary when trying to choose the best possible quantization. Because we found that one of our quantizations was better in the lower SNR range and the other was better in the higher SNR range, we proposed a decoder that adapts between our two quantizations according to variations of the channel conditions. As channel conditions change, the current noise level could be estimated from the sample variance. The 33,216 LE capacity of our FPGA could accommodate the logic of both of our quantizations with enough additional room for the logic to measure the

channel SNR and select the quantization adaptively. The adaptive decoder, illustrated in Fig. 1.11 surpasses Planjery's decoder on the AWGN channel by approximately 0.9 dB over a substantial waterfall range (BER 10^{-2}–10^{-7}).

Although the single iteration latency is greater than that of the Planjery et al. design, our decoding success with 10 iterations means that a decoder solution that is better for a range of SNR conditions can be reached in less time. We believe there is a potential for parallelization and pipelining, but even working through the bits one at a time in a serial fashion, the 430 ns-per-bit processing would support over 2 Mbps decoding throughput. FPGA-based signal-processing solutions are of interest for applications in software-defined radio (SDR) which require reconfigurability [22]. The FPGA-based decoding capability we propose is adequate to fulfill the diverse narrowband requirements of one particular contemporary system and achieves the lower throughput threshold for wideband operations [23].

Our synthesis is of Planjery's published design. Two assumptions allow us to compare our decoder to their proprietary design: (1) that the proprietary enhancements increase latency and (2) that the proprietary design requires only a modest increase in their resulting logic. With these assumptions, the comparison, summarized in Table 1.4, favors our decoder on two of three evaluation criteria.

Acknowledgments This research was supported in part by NSF grants CCF 0635382 and CHE 0216563. FPGA hardware and development tools were provided by the Altera Corporation. The authors acknowledge Shiva Planjery's graciousness in providing simulation results of their proprietary decoder using the LDPC code that we constructed.

References

1. Zhang T, Parhi KK (2002) A 54 Mbps (3,6)-regular FPGA LDPC decoder. In: IEEE workshop on signal processing systems, pp 127–132
2. Planjery SK, Chilappagari SK, Vasic B, Declercq D, Danjean L (2002) Iterative decoding beyond belief propagation. In: Information theory and applications workshop, pp 1–10
3. Zhang Z, Dolecek L, Nikolic L, Anantharam V, Wainwright M (2009) Design of LDPC decoders for improved low error rate performance: quantization and algorithm choices. IEEE Trans Commun 57(11):3258–3268
4. Moberly R, O'Sullivan ME (2011) Quantization of three-bit logic for LDPC decoding, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2011, WCECS 2011, 19–21 Oct 2011, San Francisco, USA, pp 860–865
5. O'Sullivan ME, Smarandache R (2003) High-rate, short length, (3,3 s)-regular LDPC of girth 6 and 8. In: Proceedings, IEEE International Symposium on Information Theory, p 59
6. Greferath M, O'Sullivan ME, Smarandache R (2004) Construction of good LDPC codes using dilation matrices. In: Proceeding IEEE International Symposium on Information Theory, p 235
7. Chen Y, Parhi KK (2004) Overlapped message passing for quasi-cyclic low-density parity check codes. IEEE Trans Circuits Syst 51(6):1106–1113
8. Mansour MM, Shanbhag NR (2002) Low-power VLSI decoder architectures for LDPC codes. In: Proceeding international symposium on low power electronics and design, pp 284–289
9. Byrne A, Popovici E, O'Sullivan ME (2005) Versatile architectures for decoding a class of LDPC Codes. In: IEEE European conference on circuit theory and design, (3):269–272

10. Moberly R, O'Sullivan ME (2006) Representing probabilities with limited precision for iterative soft-decision LDPC decoding, In: Proceedings 9th international symposium on wireless personal multimedia conference, San Diego, California, pp 292–296
11. Moberly R, O'Sullivan ME (2006) Computational performance of various formulations of the iterative soft-decision decoder algorithm. IEEE International Symposium on Information Theory, pp 1703–1707
12. Pearl J (1988) Probabilistic reasoning in intelligent systems—networks of plausible inference. Morgan Kaufmann, San Francisco
13. Levine B, Taylor RR, Schmit H (2000) Implementation of near Shannon limit error-correcting codes using reconfigurable hardware. In: IEEE symposium on field-programmable custom computing machines, pp 217–226
14. Davey D, MacKay MC (1998) Low-density parity check codes over GF(q). IEEE Commun Lett 2:165–167
15. Jimenez A, Zigangirov KSh (1998) Periodic time-varying convolutional codes with low-density parity-check matrices. In: Proceeding IEEE International Symposium on Information Theory, p 305
16. Gokhale M, Graham P (2005) Reconfigurable computing: accelerating computation with field-programmable gate arrays. Springer, Dordrecht
17. Flynn M, Oberman SF (2001) Avanced computer arithmetic design. Wiley, New York
18. Parhami B (2000) Computer arithmetic—algorithms and hardware designs. Oxford University Press, New York
19. Koren I (2002) Computer arithmetic algorithms. A.K. Peters Ltd., Natick
20. Ercegovac M, Lang T (2004) Digital arithmetic. Kaufmann, San Francisco
21. Han JH, Sunwoo MH (2009) Simplified sum-product algorithm using piecewise linear function approximation for low complexity LDPC decoding. In: Proceedings of the 3rd international conference on ubiquitous information management and communication, pp 302–309
22. Skey K, Bradley J, Wagner K (2006) A reuse approach for FPGA-based SDR waveforms, Military Communications Conference (MILCOM), pp 1–7
23. U.S. Department of Defense Joint Requirements Oversight Council (2003) Joint tactical radio system (JTRS) Operational requirements document (ORD), this is made readily Available at http://www.fas.org/man/dod-101/sys/land/docs/jtr23_mar.htm

Chapter 2
Modeling, Simulation and Analysis of Video Streaming Errors in Wireless Wideband Access Networks

Aderemi A. Atayero, Oleg I. Sheluhin and Yury A. Ivanov

Abstract Analysis of simulated models has become a veritable tool for investigating network behavioral patterns vis-à-vis transmitted content. The streaming video research domain employs modeling extensively due to availability of relevant tools. A vast majority of which are presented on the FOSS platform. The transmission of audio and video streaming services over different media is becoming ever more popular. This widespread increase is accompanied by the difficult task of maintaining the QoS of streaming video. The use of very accurate coding techniques for transmissions over wireless networks alone cannot guarantee a complete eradication of distortions characteristic of the video signal. A software-hardware composite system has been developed for investigating the effect of single bit error and bit packet errors in wideband wireless access systems on the quality of H.264/AVC standard video streams. Numerical results of the modeling and analysis of the effect of interference robustness on quality of video streaming are presented and discussed. Analytic results also suggest that the Markov model of packetization of error obtained from a real network for streaming video can be used in the simulations of transmission of video across networks in the hardware-software complex developed by the authors in a previous work.

A. A. Atayero (✉)
Department of Electrical and Information Engineering,
Covenant University, PMB 1023, Ota, Nigeria
e-mail: atayero@ieee.org

O. I. Sheluhin · Y. A. Ivanov
Department of Information Security, Moscow Technical University
of Communication and Informatics, Moscow, Russia
e-mail: sheluhin@mail.ru

Y. A. Ivanov
e-mail: yurasic@bk.ru

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_2,

Keywords Codec · H.264/AVC · Modeling · Simulation · SNR · Video streaming · Wireless network

2.1 Introduction

One of the most important Quality of Service (QoS) parameters for wireless networks is the probability of occurrence of bit and packet errors measured by the Bit Error Rate (BER) and Packet Error Rates (PER) respectively. Neither single packet losses nor single bit errors can provide a comprehensive imitation modeling of fading channels. In digital systems, errors often occur in packets as a result of transmission conditions. Specifically, signals are attenuated during transmission and this consequently leads to packetization of errors. A group of erroneous packets is essentially a sequence of packets that are either lost in transit or received with error after transmission over a communication channel within a given period of time. Burst Error Length (BEL) is defined as the number of erroneous packets included in a given group of errors [1].

2.2 Method

The H.264/AVC standard is a compendium of innovations and improvements on prior video coding technologies vis-a-vis enhancement of coding efficiency and effective usage over a wide gamut of networks and applications [2]. For a complete analysis of the impact of errors on resultant signal quality, we investigate the influence of the conduit's (i.e., wireless transmission medium's) robustness on the perceivable quality of streaming video standard H.264/AVC using the developed Hardware and Software Complex (HSC) [3, 4]. Objective and subjective indicators of video quality were obtained using methods described in [4]. For qualitative assessment, it is imperative to have the video file data before transmission over the network (on the transmitting end), and after reception from the network (at the receiving end). Data required for the qualitative assessment at the transmitting end are: (a) the original unencoded video in YUV format, (b) the encoded video in MPEG-4, (c) transmission start time and (d) type of each packet sent to the network.

The following data are required for qualitative assessment at the receive end: (a) time of reception of each packet from the network, (b) type of each packet received from the network, (c) the encoded video (possibly distorted) in MPEG-4 format, and (d) the decoded video in YUV format for display.

We performed data evaluation by comparing the transmitted and received files.

A. Structure of the Hardware-Software Complex

Data processing is carried out in the three phases described below:

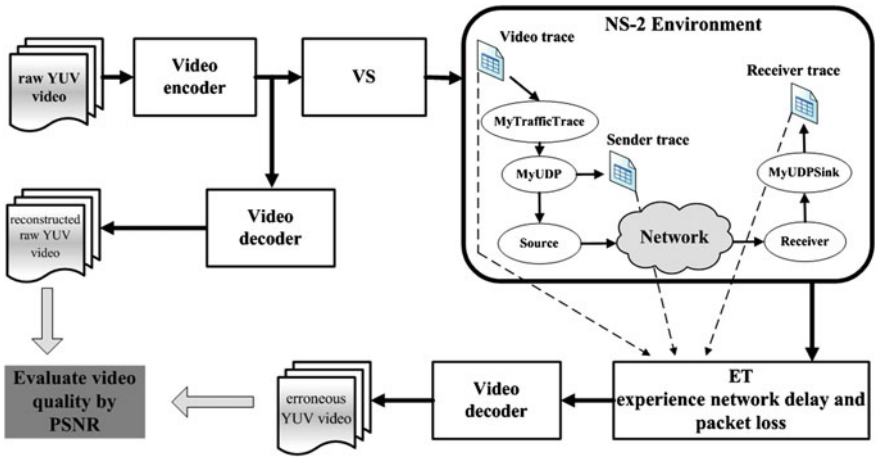


Fig. 2.1 Block diagram of HSC

First phase: the time taken in sending and receiving each packet on both sides as well as the packet type are analyzed. This results in a record of the type of frame and the time elapsed between transmitted and received packets. The distorted video file at the receive end is restored using the originally encoded video file and information about lost packets. Subsequently, the video is decoded for playback to the viewer. It is at this stage that the general task of assessing video quality is considered. Video quality indicators always require a comparison of the received (possibly distorted) video frame and the corresponding source frame. In the case of a total loss of frame in transit, the necessary frame synchronization before and after transmission over the network becomes impossible.

Second phase: In this phase of data processing, the problem of quality assessment is resolved on the basis of the analysis of information about frame loses. Substituting the last relayed frame for the lost frame restores frame synchronization. This methodology allows for subsequent frame-wise assessment of video quality.

Third phase: At this stage, the assessment of the quality of decoded video is achieved by means of both the restored and source video files.

Figure 2.1 shows a block diagram of the HSC for assessing the quality of streaming video. The schematic diagram reflects the interaction between modules in the transmission of the digital video from a source through the network connection to the viewer.

The HSC modules interact with the network by using traces containing all the necessary data listed above. Thus, for proper functioning, the HSC requires two traces, the source video and the decoder. The data network can be considered simply as a two-port *black-box* that introduces delay, packet loss, and possibly packet rearrangement. The network was simulated based on the aforementioned assumptions [5] in the NS2 environment. A detailed description of the functional modules of the HSC is given in [4].

2.3 Simulation Preparation

Video encoding begins with color space conversion from RGB to YUV also known as Y, Cr, Cb i.e., one luma and two chroma components [1]. It is common knowledge that there is a significant correlation of color components in any typical image of the RGB format, which makes it an obviously redundant format in terms of compression. The standard television uses a different representation of images, which also employs three components of the signal, but these components are uncorrelated (i.e., void of inter- componential redundancy). R, G and B components are converted to luminance Y component and two color difference components U and V of the YUV format. Since most information is stored in the luminance component, little information is lost if a thinning of the U and V components is done.

Standard test videos in the YUV format may be used as initial test video sequences. However, these videos have limited playback time and hence do not allow for the assessment of change in quality under prolonged video broadcast. Similarly, a vast amount of experimental data cannot be obtained from them. It is for these reasons that we recorded our own 30-min video in YUV format (send.YUV) with a resolution of 640 × 480 pixels and frame rate of 25 fps using a special software.

The first step is to encode the source video to H.264 format (video stream file). This is done by the video codec (a device used for encoding and decoding video signals).

Video codecs are usually characterized by (a) channel throughput, (b) decoded video distortion rate, (c) startup latency, (d) end-to-end delay, computational complexity, and (e) memory capacity. A good codec is one capable of providing the necessary trade-off vis-a-vis these characteristics [1]. In the next step, the encoded video stream is packaged in an MP4-container for onward transport over the network using the User Datagram Protocol (UDP). The result of encoding the original video is an MP4-file. Since it is necessary to evaluate the quality of video transmitted over the network, the need arises to create a spare decoded YUV file from the newly created MPEG layer-4 file, which serves as the control in evaluating the quality of video transmitted over the network, excluding the impact of the codec. It is thus possible to estimate the influence of a wireless network on the received visual video quality, while excluding encoding and decoding losses. For simulation purposes, it is necessary to create a video trace file that contains the following information: frame number, frame type, frame size, and the number of segments in which the frame is divided into packets. This video trace serves as the input to the simulator network, where the sending and reception of video data occurs. As a result of video transmission over the network, it is necessary to obtain transmission trace files and reception trace files, which contain the following packet data: the transmission/reception time, a unique identifier and trace file size. These two traces are used to determine lost packets in the network. In the end, we obtain files of the sent and received packets containing detailed information about the time of sending from the transmitter and the time of reception by the receiver.

2.4 Modeling and Simulation of Transmission Over Wireless Network

The HSC allows for the simulation of the main types of errors encountered when transmitting video data over wireless networks. The two types of simulation required are as listed below:

A. Bit error simulation

Simulated transmission over a wireless channel model with Additive White Gaussian Noise (AWGN) is conducted. In the process of simulation, certain bits in the sequence are distorted (i.e., inverted) with a given probability. The probability value used is defined by the Bit Error Rate (BER).

B. Packet error simulation

The UDP packets can be manually deleted from the received trace file. This allows for the observation of codec functionality and analysis of change in visual quality in cases of packet loss. At the same time, both the received and undistorted files can be obtained during transmission over an "ideal" channel with unlimited bandwidth and no delay, with subsequent removal of some packets.

2.5 Calculation of Losses and Estimation of Obtained Video Quality

Calculation of losses given the availability of unique id of the package is quite easily achieved. With the aid of the video trace, each packet is assigned a type. Each package of the assigned type that is not included in the received trace is deemed lost. Loss of frame is calculated for any (and all) frame(s) with a lost packet. If the first packet in a frame is lost, then the whole frame is considered lost since the video decoder cannot decode a frame, which is missing the first part. The module for trace assessment estimates received traces. The recovered file must be decoded in YUV format. There are two major methods of estimating the quality of digital video, namely, the subjective and objective methods:

Subjective quality assessment is always based on viewer impression. It is extremely costly, very time consuming and requires specialized equipment. Traditionally, subjective video quality is determined by expert assessment and calculation of the average Mean Opinion Score (MOS), which is assigned a value from 1 to 5 (ITU scale) [6, 7], where 1 and 5 represent worst and best received video quality respectively.

Objective video quality assessment is usually done by measuring the average luminance peak Signal-to-Noise Ratio (PSNR). The PSNR is a traditional metric, which allows for the comparison of any two images [8]. The PSNR module evaluates the objective quality of received video stream in Polynomial Approximation Coding (PAC). The end result is the values of PSNR calculated for the original and distorted image (as shown in Fig. 2.2). MOS values are calculated from the PSNR indicator.

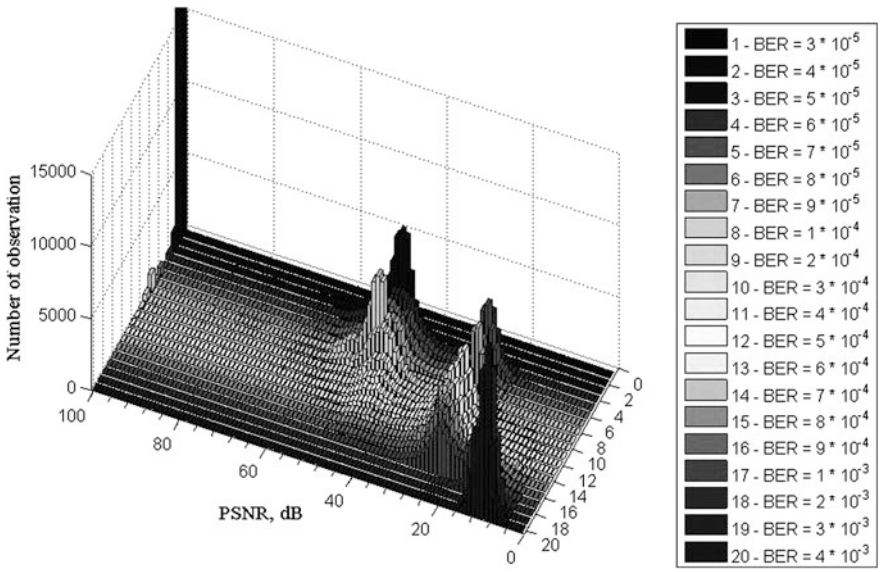


Fig. 2.2 PSNR value distribution histogram of video sequence for different values of wireless channel BER

2.6 Results, Analysis and Discussion

In order to study the effect of transmission errors on the resulting video quality, the transmission of a 30-min video over a wireless network with random packet errors in the channel was simulated. Characteristics of sequences used are listed in Table 2.1.

For modeling purposes, the encoded video stream was split into RTP/UDP-packets using the hardware-software tool reported in [4]. Bit error simulation for transmission over a wireless channel was done using an AWGN error generator contained within the PAC structure. Simulation of packet errors during transmission over a wireless channel was done by deleting packets from the received trace file [4]. This allowed us to explore and analyze the change in visual quality during loss of packets. The received and undistorted trace files were obtained for transmission over an "ideal" channel with unlimited bandwidth and no delay in using the NS2 software environment [5], followed by a random removal of packets, according to PER and BER parameters. Quality assessment was carried out using PSNR and MOS indicators, calculated by using hardware and software tools [5]. The standard deviation of the quality of the average PSNR values was calculated using Eq. (2.1) [9].

$$S'_{PSNR} = \sqrt{\frac{1}{N-1}\sum_{n=0}^{N-1}(PSNR_n - \overline{PSNR'})^2}. \tag{2.1}$$

Table 2.1 Characteristics of encoded video

Format	MPEG-4 Part 14 (MP4)
Codec	H.264
Bit rate	Constant @ 1150 kbps
Frame frequency	25 fps
Resolution	640 × 480 pixels
GOP type	IBBPBBPBB

A. Effect of bit error

Figure 2.2 shows the effect of BER on the quality of video streaming.

Analysis of the results of streaming video over the simulated wireless network with different values of BER revealed the following:

(i) Simulating a wireless channel using AWGN model, and additive, bit errors with a value of BER $\leq 3 \times 10^{-5}$ does not affect the quality of the video. However, when BER $\geq 4 \times 10^{-3}$ packet loss in the network reaches its maximum value of $\geq$99.9 %.

(ii) Objectively, excellent quality of video transmission over a channel can be guaranteed for all bit error probabilities less than 1×10^{-4}, good quality is in the range of 1×10^{-4} to 4×10^{-4}, satisfactory quality is in the range of 4×10^{-4} to 8×10^{-4}, poor quality is in the range of 8×10^{-4} to 1×10^{-3}, while very bad quality is for any BER $> 1 \times 10^{-3}$.

(iii) The histograms of the distribution of PSNR values during simulation and broadcast over a real network in general are of a twin-peak form. One of the peaks characterizes the PSNR value of error-free video stream (the decoder is able to correct bit errors when they are relatively few in the frame). The second peak characterizes PSNR degradation due to the large number of corrupted video frames in fading moments (the decoder is unable to fix large numbers of bit errors). As the number of errors increases, this maximum increases commensurately with a decrease in the second. During transmission, depending on error level, values of either of the maxima increase. In cases when errors in the communication channel are negligible, the PSNR distribution has only one maximum.

Empirical values of BER transitions from an acceptable quality to the poor, according to the relationship between PSNR and MOS [6], are presented in Table 2.2. However, the AWGN model does not allow for adequately simulation of a fading channel. Typically, errors are often long term, since high probability of bit loss occurs in specific periods of transmission, e.g., during poor propagation. Attenuation of the transmitted signal results in packetizing (grouping) of errors. Another cause of error grouping can be physical defects of, and failures inherent in the information storage system. When using VLC, bit error occurrence results in group errors or packetization of errors.

Table 2.2 Relationship between quality indicators and BER

PSNR (dB)	MOS (%)	BER	ITU Quality scale	Picture degradation
>37	81–100	$<1 \times 10^{-4}$	5 Excellent	Noticeable
31–37	61–80	1×10^{-4}–4×10^{-4}	4 Good	Noticeable, but not irritating
25–31	41–60	4×10^{-4}–8×10^{-4}	3 Satisfactory	Slightly irritating
20–35	21–40	8×10^{-4}–1×10^{-3}	2 Poor	Irritating
<20	0–20	$>1 \times 10^{-3}$	1 Very poor	Very irritating

B. Effect of packet error

Figure 2.3 shows the effect of PER indicator on streaming video quality. The range of values of PER, within which the resulting quality is maximal (i.e., almost equal to the original) and minimal are indicated. It is shown that with PER $\leq 1 \times 10^{-4}$ error does not affect the resultant video quality and can be easily eliminated with decoders and existing methods of error correction. A further change in the quality has a stepwise nature and decreases with increasing PER.

Analyzing the results of streaming video over a simulated wireless network with a given probability of packet loss, we safely conclude that:

(i) A PER value of $\leq 1 \times 10^{-4}$ in simulation of a wireless network does not affect the video quality. When PER $\leq 1 \times 10^{-3}$ impact of errors on video quality is not noticeable and does not irritate during viewing experience. When PER ≥ 0.1, packet loss in the network has the worst effect on visual quality.
(ii) Objectively, excellent quality of video transmission over a channel can be guaranteed for all packet error probabilities less than 1×10^{-3}, good quality is in the range of 1×10^{-3}–3×10^{-3}, satisfactory quality is in the range of 3×10^{-3}–1×10^{-2}, poor quality is in the range of 1×10^{-2}–5×10^{-2}, while very bad quality is for any PER $> 5 \times 10^{-2}$.

Histograms of the distribution of values of PSNR when PER $\leq 6 \times 10^{-4}$, in general, have a bimodal shape. One of the peaks characterizes the value of PSNR of video stream distorted due to packet loss. The second maximum characterizes deterioration in the PSNR of dependent frames. As the number of errors increases, one of the peaks increases due to a decrease in the other.

C. Effect of length of error groups

To study the effect of the length of error groups on resultant quality, the simulation of a 30-min video transfer over a wireless network for the values of PER of 1×10^{-5}–5×10^{-2} is repeated, since a visual change in video quality is observed at this range. The simulation of groups of error packets during transmission over a wireless channel was done by means of random deletion of packet groups from the receive trace file with a given BEL. For this particular example BEL = 100 implies that the total random number of consecutively deleted packets does not exceed 100. The total sum of erroneous (deleted) packets in the video sequence for the whole experiment given PER = const. remained the same, irrespective of the value of BEL. Figure 2.4 shows the effect of BEL on the quality of streaming video for PER $= 1 \times 10^{-3}$.

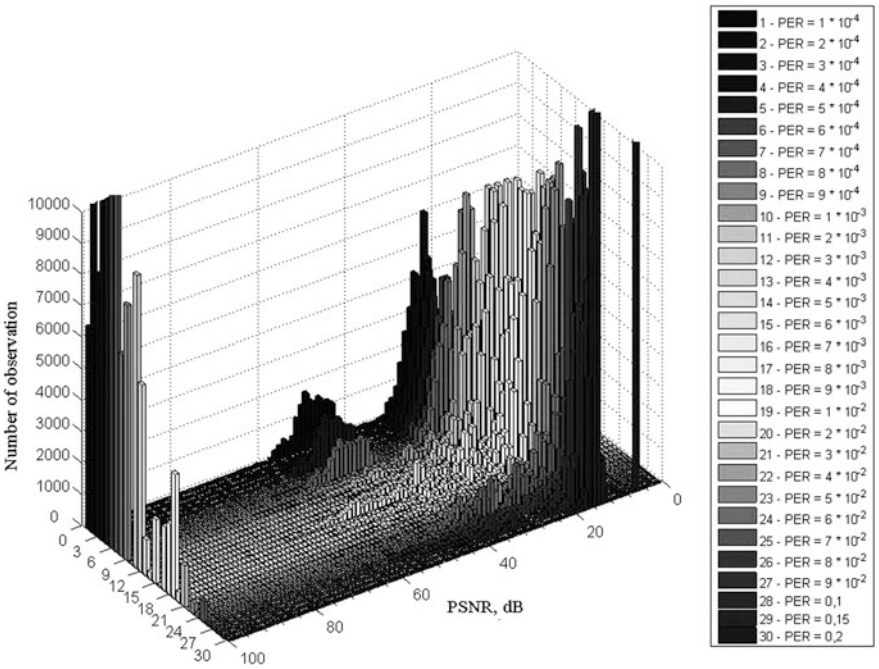


Fig. 2.3 PSNR value distribution histogram of video sequence for different values of wireless channel PER

Analyzing the results of streaming video over the simulated wireless network with a given grouping of erroneous packets, we can draw the following conclusions:

(iv) For PER $\leq 1 \times 10^{-3}$ the effect of single packet errors on quality is insignificant and does not irritate the viewing experience.
(v) Histograms of the distribution of values of PSNR have two maxima. One of the peaks characterizes the value of PSNR of video frames distorted due to the loss of packets. The second maximum characterizes the deterioration of PSNR of dependent frames. With increasing quantities BEL is one of the peaks decreases as the number of dependent frames are also reduced, whereas the second peak remains unchanged. This is explained by the fact that the single scattered throughout the video sequence error number of distorted frames is large due to error propagation to dependent frames. An increase in the BEL value leads to a decrease in one of the maxima, since the number of dependent frames also decreases, while the second maximum remains the same. This is due to the fact that under singular errors spread across the whole video sequence, the number of distorted frames is large because of the distribution of errors on dependent frames.
(vi) Increasing the length of the error groups leads to an increase in the average quality of the video sequence.

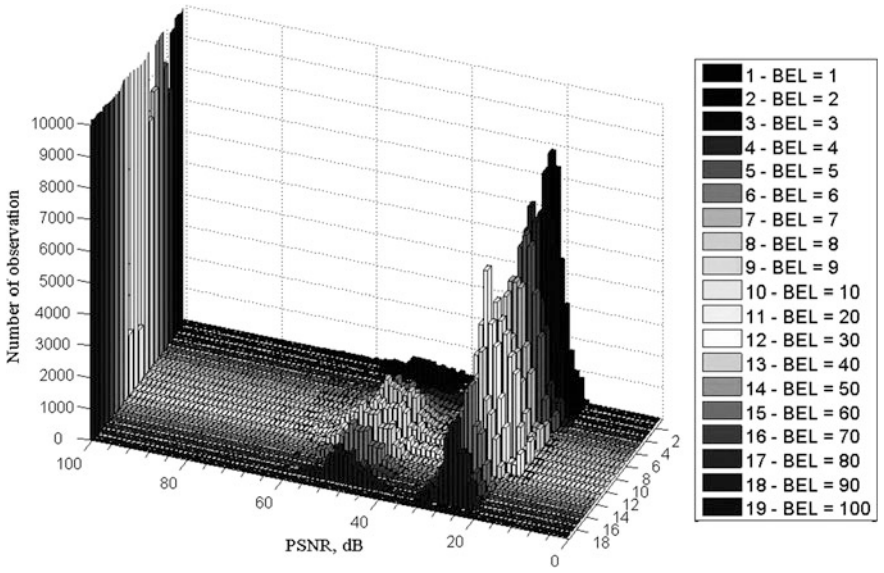


Fig. 2.4 PSNR value distribution histogram of video sequence for PER $= 1 \times 10^{-3}$ and varying values of wireless channel BEL

(vii) Effect of error groups on the quality is more powerful because of the local concentration of errors. However, the average quality of the video sequence increases with increase in the length of the grouping for a given value of probability of occurrence of packet errors.

For BEL $\geq$ 60 the average quality is almost identical to the original video.

D. Relationship between PER and BEL

The average quality of the experimental video sequence for different values of PER and BEL is shown in Fig. 2.5.

In assessing the impact of erroneous packets received on quality, it is necessary to analyze not only the likelihood of occurrence of errors, but also their structure and length of their grouping. Additionally, the following conclusions can be drawn:

(i) Increasing the length of error groupings leads to an increase in the average quality of the video sequence. This is due to the deterioration of a small section of video, where error groups are concentrated, whereas in the case of single bit errors deterioration in the quality of video may be observed across the whole sequence;

(ii) When the length of erroneous packets is BEL $\leq$ 6 the change in quality is minor and identical to the influence of single packet errors (BEL $=$ 1);

(iii) When BEL $\geq$ 60 the average quality is almost identical to the original (PSNR $<$ 90 dB). It is logical to assume that the value of BEL in the longer video sequences, with the same average quality may have a higher value;

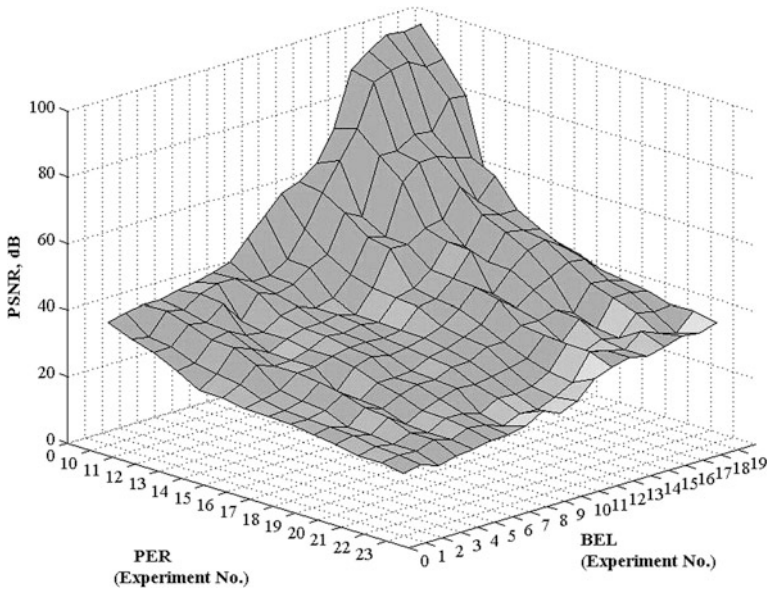


Fig. 2.5 Estimate of video sequence quality for different values of PER and BEL

(iv) The highest dynamics of change in PSNR = 60 dB is observed in two cases: (a) for a fixed PER = 1×10^{-3} and the variable values of BEL; and (b) at BEL $\geq$ 80 and the varying values of the PER. In other cases, the dynamics is not essential and minimal in the absence of clustering of errors (BEL = 1);

(v) With increasing PER, the effect of BEL on quality decreases due to increase in denseness of single errors;

(vi) Analysis of the results of PER and BEL shows that for effective assessment of the impact of transmission errors on resultant quality it is necessary to analyze not only the likelihood of errors, but also their structure and length of their grouping. The most realistic and accurate method of modeling statistical errors in communication channels is the use of probability data obtained from real networks.

At BER values $\leq 3 \times 10^{-5}$ bit errors do not affect the quality of the received video and are easily eliminated by well-known methods of error correction implemented in WiMAX. When BER $\geq 4 \times 10^{-3}$ packet loss in the network reaches its maximum value and leads to an unacceptable quality of the received video. Ensuring objectively *excellent* quality of video sequence over a channel can be done for probabilities of bit error rate less than 1×10^{-4}; *good* quality in the range of 1×10^{-4}–4×10^{-4}; *satisfactory* quality in the range of 4×10^{-4}–8×10^{-4}; *poor* quality in the range of 8×10^{-4}–1×10^{-3} and *very bad* at BER $\geq 1 \times 10^{-3}$. The use of H.264/AVC video in wireless access systems with VLC codec of variable length leads to a disruption of the synchronization of decoded video sequences and

the occurrence of additional grouping of errors, whose impact on the quality for video decoding is much stronger than that of the bit error, since it leads to loss of large segments of the information. It is shown that the quality of the video affects not only the probability of error, but also the structure and length of errors. Analysis of individual errors showed that at PER $\leq 1 \times 10^{-4}$ packet errors do not affect the quality of the received video and are easily eliminated by well-known methods of error correction deployed in wireless networks. When PER $\leq 1 \times 10^{-3}$, the effect of errors on quality is not noticeable and does not irritate the viewing experience. For values of PER ≥ 0.1 packet loss in the network leads to an unacceptable quality of the received video. Ensuring objectively *excellent* quality of video sequence over a channel can be done for probabilities of bit error rate less than 1×10^{-3}; *good* quality in the range of 1×10^{-3}–3×10^{-3}; *satisfactory* quality in the range of 3×10^{-3}–1×10^{-2}; *poor* quality in the range of 1×10^{-2}–5×10^{-2} and *very bad* at BER $\geq 5 \times 10^{-2}$.

To assess the impact on quality of video playback under error grouping conditions of error groups BEL, the use of a regular (deterministic) model is proposed. It is shown that the effect of errors on the average quality is stronger due to local concentration of errors. The average quality of the video sequence at the same time increases with increase in the length of the grouping for a given value of probability of occurrence of packet errors. For groupings of length BEL ≥ 60, average quality is almost identical to that of the source video. With increasing PER, the effect of BEL on quality decreases due to increase in the denseness of single errors. Increase in the BEL leads to an increase in the average quality of the video sequence irrespective of the PER value. The highest dynamics of change in PSNR is observed for fixed PER $= 1 \times 10^{-3}$ and the variable values of BEL; at BEL ≥ 80 as well as for the changing values of the PER. In other cases, the dynamics is not significant and is minimal in the absence of clustering of errors. To assess the quality of video under packetization of errors under real conditions, it is necessary to investigate the actual distribution of packetization of errors in the communication channel.

E. Analysis of error packetization phenomenon

The need to create realistic simulation and mathematical models of behavior of losses in the communication channels based on the apparatus of Markov chains for wireless access systems is a scientific problem of important consequence. Markov processes with the necessary number of states sufficiently describe the mechanism of transmission of information [10], the knowledge of which is necessary to analyze network problems during packet video transmission. The parameters of the model make it possible to determine the quality of transmitted video as well as the statistical parameters of the network. In an experiment carried out by the authors, the matrix of values in Fig. 2.6 was obtained for the Markov model developed for investigating the error packetization phenomenon.

A model describing the length of error intervals and error-free reception for streaming video transmission was developed based on the experimental data obtained as a result of streaming video from a moving source on WiMAX network [11]. Based on the graph of packet loss distribution, an array was formed in which the lost

$$\Gamma = \begin{vmatrix} 0.999 & 0 & 0 & 2.4*10^{-5} & 8.69*10^{-4} & 1.1*10^{-4} \\ 0 & 0.9944 & 0 & 1.344*10^{-4} & 0.0049 & 6.16*10^{-4} \\ 0 & 0 & 0.965 & 8.4*10^{-4} & 0.0304 & 0.0039 \\ 1.8*10^{-5} & 3.6*10^{-4} & 3.6*10^{-4} & 0.9991 & 0 & 0 \\ 4.2*10^{-5} & 8.4*10^{-4} & 8.4*10^{-4} & 0 & 0.9979 & 0 \\ 2.58*10^{-4} & 0.0052 & 0.0052 & 0 & 0 & 0.9871 \end{vmatrix}$$

Fig. 2.6 The matrix of values

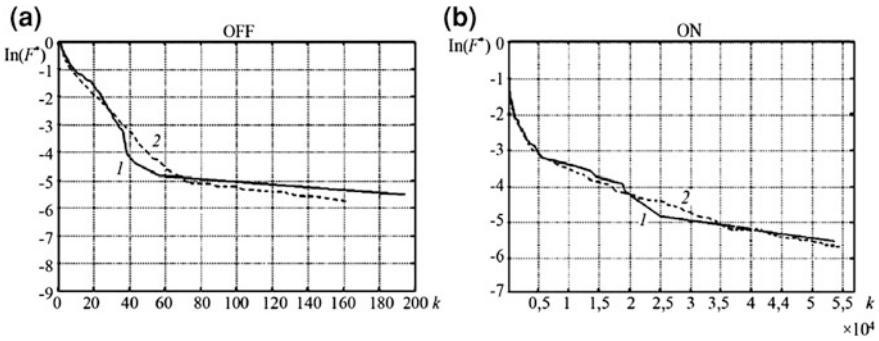


Fig. 2.7 DF of simulated samples of the length of OFF-(**a**) and ON (**b**). Periods: *curve* 1 experiment, *curve* 2 simulation

packet corresponds to a logic zero (0) and received packet corresponds to a logic unit (1) (Fig. 2.7).

Analysis of the quality of received video sequence when simulating Markov model of error packetization shows that the average quality of video sequences is slightly worse than during transmission over a real network. The subjective MOS quality indicator also shows a difference in values: a real WiMAX network returned a mean value of 3.59 (corresponding to *satisfactory*), while the experiments returned values of 2.72 (corresponding to *poor*) and 3.01 (corresponding to *satisfactory*), respectively. The average quality of video sequences when simulating Markov model packetization of errors are similar to those obtained when simulating single packet errors with PER index in the range of 3×10^{-3} to 1×10^{-2}. While the length of error group depending on the PER of the specified range attain values of $BEL \leq 10$.

2.7 Conclusions

We have presented in this paper the modeling, simulation and analysis of errors inherent in video streams over wireless broadband access networks, by presenting results of investigating the effect of single bit error and bit packet errors on the

quality of H.264/AVC standard bursty video streams. A software-hardware composite system that was developed specifically for this purpose was employed in the investigation. Analysis of simulation results led to conclusions and postulations discussed in detail in sections VI A through E for BER, PER, BEL, relationship between PER and BEL, and the effect of error packetization phenomenon respectively.

References

1. Atayero AA, Sheluhin OI, Ivanov YA, Iruemi JO (2011) Effect of wideband wireless access systems interference robustness on the quality of video streaming. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2011, WCECS 2011. San Francisco, USA, pp 848–854, 19–21 October 2011
2. Sheluhin OI, Atayero AA, Ivanov YA, Iruemi JO (2011) Effect of video streaming space–time characteristics on quality of transmission over wireless telecommunication networks. In: Proceedings of the world congress on engineering and computer science 2011, WCECS 2011, vol I, San Francisco, USA, pp 572–577, October 19–21 2011
3. Sheluhin OI, Ivanov YA (2009) Assessment of the quality of streaming video in telecommunication networks using software-hardware methods. Electrotech Inf Complexes Syst 5(4):48–56
4. Ivanov YA, Pryanikov VS (2010) Imitation modeling of wireless networks using hardware-software complex for the assessment of streaming video quality. Chuvash Univ Dig 1(1):35–48
5. NS-2 documentation [online] (2011) http://bit.ly/ysayeJ. Accessed 11 Nov 2011
6. ITU P.800 (1996) Methods for subjective determination of transmission quality [Online]. http://www.itu.int/rec/T-REC-P.800-199608-I/en
7. Atayero AA (2000) Estimation of the quality of digitally transmitted analogue signals over corporate VSAT networks. PhD thesis, Moscow, Jan 2000
8. Ostermann J et al (2004) Video coding with H.264/AVC: tools, performance, and complexity. IEEE Circuits Syst Mag 4(1):7–28 (First Quarter (Q1) 2004)
9. Lemmon JJ (2002) Wireless link statistical bit error model, NTIA Report 02-394, U.S. Department of Commerce, June 2002
10. Wang H, Moayeri N (1995) Finite state Markov channel—a useful model for radio communication channels. IEEE Trans Veh technol 44(2):163–171
11. Hohlfeld O (2008) Markovian packet loss generators and video QoE, T Systems, Feb 2008

Chapter 3
Bayesian Based Intrusion Detection System

Hesham Altwaijry

Abstract In this paper intrusion detection using Bayesian probability is discussed. The systems designed are trained a priori using a subset of the KDD dataset. The trained classifier is then tested using a larger subset of KDD dataset. Initially, a system was developed using a naive Bayesian classifier that is used to identify possible intrusions. This classifier was able to detect intrusion with an acceptable detection rate. The classier was then extended to a multi-layer Bayesian based intrusion detection. Finally, we introduce the concept that the best possible intrusion detection system is a layered approach using different techniques in each layer.

Keywords Bayesian filter · Intrusion detection · KDD dataset · Multi-layer filters · Training engine · U2R and R2L attacks

3.1 Introduction

Intrusion detection systems have become an integral part of the security infrastructure of most organizations due to the increased number and severity of network attacks. These instruction detection systems can be software, hardware, or a combination of both. These systems automate the process of monitoring and analysis of network traffic with the goal of capturing and detecting security problems [1, 2].

H. Altwaijry (✉)
Computer Engineering Department, King Saud University, 51178
Riyadh 11543, Saudi Arabia
e-mail: twaijry@ksu.edu.sa

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_3,

The Deployment of highly effective IDS systems is extremely challenging. For example until an IDS is properly tuned to a specific environment, there will be thousands of alerts generated daily, with most of these alerts being incorrect and thus are false alerts. However, it is not obvious whether the alert is positive or negative until after they have been investigated thereby creating a large burden on the IT department. There have been many techniques proposed to lessen these false alerts and improve the performance of the system. Agarwal and Joshi [3] used a two-stage general-to specific framework for learning a rule-based model (PNrule). This model can classify models of a data set that has widely different class distributions in the training data set. Levin [4] used a data-mining tool for classification of data and prediction of new cases using automatically generated decision trees. In this paper will show that the use of Bayesian probability is very promising in reducing the false positive alert rate and the use of multi-stage Bayesian probability is extremely effective in reducing the false positive alert rate.

Bayesian probability is an interpretation of the probability calculus which holds that the concept of a probability can be defined as the degree to which a person (or community) believes that proposition is true. Currently Bayesian theory is used in email spam-filters [5–7], Speech recognition [8], Pattern Recognition [9], and Intrusion Detection [10–12].

3.2 Bayesian Theory

Bayesian theory is named after Thomas Bayes (1702–1761), his theory can be explained as follows:

If the events A_1, A_2,……and A_n constitute a partition of the sample space S such that $P(A_k) \neq 0$ for $k = 1, 2, \ldots, n$, then for any event B such that $P(B) \neq 0$:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{K=1}^{n} P(A_k)P(B|A_k)} = \frac{P(A_i)P(B|A_i)}{P(B)}$$

In recent years Bayesian networks have been used across a wide range of fields in computer science [13] because of their ability to obtain a coherent result from probabilistic information about a situation. Additionally there are many efficient algorithms that can be used to derive the results from the information. It is believed that this ability and readily available algorithms would allow one to construct an efficient IDS system.

3.3 KDD-99 Dataset

To test our IDS system we used the DARPA KDD99 Intrusion Detection Evaluation dataset [14]. This dataset was created by Lincoln Laboratory at MIT and was used in The Third International Knowledge Discovery and Data Mining Tools

Table 3.1 Basic characteristics of the KDD 99 intrusion detection datasets in terms of number of samples

Dataset	Normal	DoS	Probing	R2L	U2R	Total
Whl KDD	972,780	3,883,370	41,102	1,126	52	4,898,430
10 % KDD	97,278	391,458	4,107	1,126	52	494,020
KDD corr	60,593	229,853	4,166	16,189	228	311,029

Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining [14]. This dataset is one of the most realistic publicly available sets that include actual attacks [15]. Therefore, researchers have been using this dataset to design and evaluate their intrusion detection systems. An added benefit is that a common dataset allows researchers to compare experimental results. The data set was acquired from nine weeks tcp dump data. It is made up of a large number of network traffic activities including both normal and malicious connections. The KDD99 data set includes three independent sets; "whole KDD", "10 % KDD", and "corrected KDD". In our experiments we have used the "10 % KDD" and the "corrected KDD" as our training and testing set, respectively. Table 3.1 summarizes the number of samples in each dataset:

The training set contains a total of 22 training attack types. Additionally the "corrected KDD" testing set includes an additional 17 attack types. Therefore there are 39 attack types that are included in the testing set and these attacks can be classified into one of the four main classes;

- DOS: Denial of Service attacks.
- Probe: another attack type sometimes called Probing.
- U2R: User to Root attacks.
- R2L: Remote to Local attacks.

DoS and Probe attacks are different from the normal traffic data and can be easily separated from normal activities. They come in a greater frequency in a short period of time. On the other hand, U2R and R2L attacks are embedded in the data portions of the packets and normally involve only a single connection. Therefore these types of attacks are harder to identify and it is difficult to achieve satisfactory detection accuracy for these two attacks [16].

The KDD-99 network TCP connections have 41-features per connection (record). These features can be divided into four categories [17]:

Basis features: Features 1–9 are the basic features that are derived from packet header without inspecting the payload.

Content features: Domain knowledge is used to assess the payload of the original TCP packets. This includes features such as the number of failed login attempts.

Time-based traffic features: These are features that capture properties that mature over a 2-s temporal window. An example is the number of connections to the same host over the 2-s interval.

Host-based traffic features: These features utilize a historical window estimated over the number of connection instead of time. They are designed to assess attacks, which span intervals longer than 2 s.

3.4 Bayesian Filter

The Bayesian IDS is built out of a Naive Bayesian classifier. This classifier is anomaly based. It works by recognizing that feature values have different probabilities of occurring in attacks and in normal TCP traffic. The filter is trained by giving it pre-classified traffic. It will then adjust the probabilities for each feature. After training, the filter will calculate the probabilities for each TCP connection and classify it as either normal TCP traffic or an attack. Therefore our Bayesian filter consists of the following two components:

A. Training Engine:

Figure 3.1 shows the block diagram for the Bayesian filter that is constructed for the IDS system. For each input record there is a label describing the type of connection. We use this label to train the engine as follows:

- First the numbers of good records and bad records in the training dataset are calculated.
- Then two hash tables are created; the first one includes the frequency of each attribute for normal records, and the second one includes the frequency of each attribute for the not normal records.
- Finally, a third hash table is created. This table contains each attribute from the normal and not normal records and it is scored using the following formula

$$score(attribute) = \frac{\frac{B}{Num\_Bad\_Rec}}{\frac{B}{Num\_Bad\_Rec} + \frac{G}{Num\_Good\_Rec}}$$

Where

- B is the frequency of that attribute in the hash table related to not-normal file.
- G the frequency of that attribute in the hash table related to normal file

B. Testing Engine:

After training the engine is tested by loading the KDD corrected dataset. The following formula is applied to obtain a probability of whether the record is normal or not

$$P(record) = \frac{\Pi_{i=1}^{n} score(i)}{\Pi_{i=1}^{n}\ score(i) + \Pi_{i=1}^{n}(1 - score(i))}$$

where

- n: number of attributes that we need to use to test the required record
- Score(i): the score of the attribute

The record is considered to be an attack if the P(record) is greater than a specified threshold.

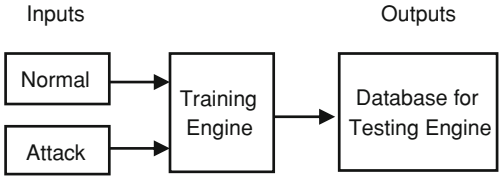


Fig. 3.1 The training engine

3.5 Simple Bayesian Experimental Setup

Many experiments were conducted until we finally achieved results that are comparable to what has been published in the literature. In this section, the most important experiments will be explained. Experiments differ basically in the training data used to build the database, which accordingly affects the accuracy of the test. Additionally, the number of features and level of threshold used in the testing engine makes a big difference in the results. Therefore, all experiments presented differ due to manipulation of the following inputs: training data, features and threshold.

A. Training data

Using the 10 % KDD data set we have 494,020 records that can be used to train the training engine. These training records consist of normal (non-attack) records and known attack records distributed among the four attacks types: DoS, Probe, U2R and R2L. In all the experiments that we will present we will use the normal records (non-attack), adding to them the appropriate not-normal (attack) records. In each experiment, we select one type of not-normal record, except for the first experiment, in which all types of records were used. The objective of varying between not-normal records in each experiment is to see the effect of each different attack on the results. This method of altering the attack type in each experiment shows some interesting results (see experiment 5), as the detection rate of U2R attacks was higher when using R2L records to train the engine than when using U2R (see experiment 4).

B. Features

Since the data record consists of 41 features, we can select between them and perform a very large number of combinations. We have selected the features as follows:

(1) Using specific features like basic (features 1–9), content (features 10–22) and traffic features (23–31).
(2) Using all the 41 features.
(3) Using selected features by inspecting the score map.

The experiments performed show that the first method does not yield good results compared to the second method. However, the results from using all features are worse than results previously published. Consequently, it was necessary to find key features that can lead to better results. To do so, we analyzed the score map that was built by the training engine to see the highest value that can result in a score that is above the threshold so that the detection rate may be increased.

Among the many features and after many experiments, we ended up with three features that raised the detection rate of R2L attack to 85 % as will be explained in experiment 5.

C. Threshold

This is the level that we used to distinguish between normal records and attacks. The threshold value was adjusted between the experiments to increase the detection rate.

After performing each experiment, we analyzed the results based upon the number of normal and not-normal records that the testing engine succeeded or failed in classifying. We use the following expressions to analyze the data:

True Negative (TN): The percentage of valid records that are correctly classified.
True Positive (TP): The percentage of attack records that are correctly classified.
False Positive (FP): The percentage of records that were incorrectly classified as attacks whereas in fact they are valid activities.
False Negative (FN): The percentage of records that were incorrectly classified as valid activities whereas in fact they are attacks.

$$Detection\ Rate(DR) = \frac{TP}{TP + FN}$$

$$Overall\ Classification\ Rate(CR) = \frac{TP + TN}{TP + TN + FP + FN}$$

3.6 Naïve Bayesian Filter

The results obtained using a single Bayesian filter were presented in [18]. These results are reproduced here. Each of the following five experiments consists of five sub-experiments. However, we categorize them into five main experiments since we use the same training engine to perform the five sub-experiments. For example, using the normal and not-normal records in the training engine for the first group of experiments, we test not-normal records, DOS, Probing, U2R and R2L. Each test has its own setup in terms of features and threshold. Table 3.2 summarizes the records used in training the Bayesian filter for all five experiments.

Table 3.3 shows what was used for each test in this experiment, after the Bayesian filter had been trained using 65,593 normal records and 401,195 attack records. We then performed five tests on the trained network to see the effects on each attack type. Normal records were used in all tests and each test used one attack at a time. For example, the first test in the experiment used 60,593 normal records and 250,436 not-normal records.

From Table 3.4 we can see that normal records were detected with high accuracy = 99.03 %, while the best detection rate was for the DOS attack type

Table 3.2 Number of records used in experiments

Experiment	Type	Number	Type	Number
Experiment 1	Normal	65,593	All Attacks	401,195
Experiment 2	Normal	92,827	DOS	280,504
Experiment 3	Normal	92,827	Probing	4,107
Experiment 4	Normal	92,827	U2R	188
Experiment 5	Normal	92,827	R2L	1,126

Exp1 using all attack records and normal(non-attack) records

Table 3.3 Experiment 1 testing environment

Testing data	No of Records	Features	Threshold
Normal	60,593	All	0.9
Not Normal	250,436	All	0.9
DOS	165,299	All	0.9
Probing	4,166	All	0.9
U2R	188	All	0.9
R2L	16,180	All	0.9

Table 3.4 Test results percentages

Test	TN	TP	FN	FP	DR	CR
All Attacks	99.03	89.70	0.97	10.30	89.70	94.37
DOS	99.03	99.36	0.97	0.64	99.36	99.20
Probing	99.03	57.15	0.97	42.85	57.15	78.09
U2R	99.03	0.00	0.97	100.00	0.00	49.52
R2L	99.03	0.00	0.97	100.00	0.00	49.52

which had 99.36 % accuracy while the worst results were for U2R, and R2L attacks with 0 %. The overall detection rate did not reflect the actual accuracy of the filter if the TP rate was low, as in U2R and R2L.

D. Experiment 2: using DOS records and normal(non-attack) records

Table 3.5 shows what was used to test the Bayesian filter for the second experiment. While in Table 3.6 we can see that normal records were detected with high accuracy: TN = 99.6 %. The best DR was for the DOS attack type, which was 99.24 %, as expected since it was trained using only DOS attacks, and the worst result was for U2R and R2L attacks with 0 %. The overall DR did not reflect the actual accuracy of the filter if the TP rate was low as in U2R and R2L.

E. Experiment 3: using Probing records and normal records

Table 3.7 shows what was used to test the Bayesian filter for the third experiment. While Table 3.8 shows that normal records are detected with high accuracy: TN = 99.4 %. The best accuracy was for the U2R attack type, which was 91.5 % although the training data was using PROBING attacks.

Table 3.5 Experiment 2 testing environment

Testing data	No of Records	Features	Threshold
Normal	60,593	All	0.9
All Attacks	250,436	All	0.9
DOS	165,299	All	0.9
Probing	4,166	23,24,31	0.6
U2R	188	23,24,31	0.6
R2L	16,180	23,24,31	0.6

Table 3.6 Experiment 2 test result percentages

Test	TN	TP	FN	FP	DR	CR
All Attacks	99.60	65.50	0.40	34.50	65.50	82.55
DOS	99.60	99.24	0.40	0.76	99.24	99.42
Probing	99.60	17.73	0.40	82.27	17.73	58.67
U2R	99.60	0.00	0.40	100.00	0.00	49.80
R2L	99.60	0.00	0.40	100.00	0.00	49.80

Table 3.7 Experiment 3 testing environment

Testing data	No of Records	Features	Threshold
Normal	60,593	All	0.9
All Attacks	250,436	All	0.9
DOS	165,299	All	0.9
Probing	4,166	All	0.9
U2R	188	23,24,31	0.6
R2L	16,180	23,24,31	0.6

Table 3.8 Experiment 3 test result percentages

Test	TN	TP	FN	FP	DR	CR
Not Normal	99.40	71.00	0.60	29.00	71.00	85.20
DOS	99.40	70.30	0.60	29.70	70.30	84.85
Probing	99.40	81.90	0.60	18.10	81.90	90.65
U2R	99.40	91.50	0.60	8.50	91.50	95.45
R2L	99.40	61.50	0.60	38.50	61.50	80.45

Table 3.9 Experiment 4 testing environment

Testing data	No of Records	Features	Threshold
Normal	60,593	All	0.9
Not Normal	250,436	23,24,31	0.6
DOS	165,299	All	0.6
Probing	4,166	23,24,31	0.6
U2R	188	23,24,31	0.6
R2L	16,180	23,24,31	0.6

Table 3.10 Experiment 4 test result percentages

Test	TN	TP	FN	FP	DR	CR
Not Normal	99.70	76.50	0.30	23.50	76.50	88.10
DOS	99.70	0.02	0.30	99.98	0.02	49.86
Probing	99.70	71.60	0.30	28.40	71.60	85.65
U2R	99.70	93.00	0.30	7.00	93.00	96.35
R2L	99.70	61.50	0.30	38.50	61.50	80.60

Table 3.11 Experiment 5 testing environment

Testing data	No of Records	Features	Threshold
Normal	60,593	All	0.9
Not Normal	250,436	All	0.9
DOS	165,299	All	0.9
Probing	4,166	23,24,31	0.6
U2R	188	23,24,31	0.6
R2L	16,180	23,24,31	0.6

Table 3.12 Experiment 5 test result percentages

Test	TN	TP	FN	FP	DR	CR
Not Normal	68.03	10.00	31.97	90.00	10.00	39.02
DOS	68.03	2.00	31.97	98.00	2.00	35.02
Probing	68.03	63.60	31.97	36.40	63.60	65.82
U2R	68.03	96.30	31.97	3.70	96.30	82.17
R2L	68.03	85.35	31.97	14.65	85.35	76.69

By analyzing the features, it is found that U2R and R2L attacks can be detected with a better rate if selected features are chosen. These features are 23(count), 24(serror_rate) and 31(Srv_diff_host_rate)

F. Experiment 4: using U2R records and normal records

Table 3.9 shows what was used to test the Bayesian filter for the fourth experiment. While Table 3.10 shows that normal records were detected with high accuracy: TN = 99.7 %. The best DR was for the U2R attack type which was 93.0 % and the worst result was for DOS attacks with 0.02 %. Also in this experiment, like the previous one, attacks show sensitivity if using features 23, 24 and 31.

G. Experiment 5: using L2R records and normal records

Table 3.11 shows what was used to test the Bayesian filter for the fifth experiment. While Table 3.12 shows that normal records accuracy decreased to 68.03 % since the threshold value was decreased to improve the TP for R2L. The best DR was for the U2R attack type which was 96.3 % although the training data used the R2L type. The worst result was for DOS attacks with 2 %. Also in this experiment, like the previous two experiments, showed sensitivity to using features 23, 24 and 31.

Table 3.13 Detection rate for various algorithms [17]

Algorithm	DOS	Probe	U2R	R2l
KDD cup winner	97.10	83.30	12.30	8.40
SOM map	95.10	64.30	22.90	11.30
Gaussian classifier	82.40	90.20	22.80	9.60
K-means clustering	97.30	87.60	29.80	6.40
Nearest clustering	97.10	88.80	2.20	3.40
Radial basis	73.00	93.20	6.10	5.90
C4.5 decision tree	97.00	80.80	1.8	4.6
Linear GP	96.70	85.70	1.30	9.30
SVM	99.90	67.31	0.00	29.09
KMO + SVM	75.76	99.61	49.45	22.24
Backpropagation	97.23	96.63	87.71	30.97

The previous five experiments showed that by tweaking selected features it is possible to achieve high accuracy for all four types of attacks (Table 3.13).

The results obtained by using Bayesian filters are comparable to what has been presented in Chou's PhD thesis [17] where he reported the results of most algorithms. However, Bayesian filters were able to achieve superior results for in detecting U2R and R2L attacks.

3.7 Multilayer Filters

3.7.1 Improved Bayesian Filter

To improve the performance of the IDS system for the U2R and R2L attacks we implemented multiple Bayesian filter layer. This section will describe the promising results that we achieved [19].

3.7.1.1 Improved Bayesian Filter 1 (IBF1):

Although the testing engine classifies its inputs records to normal and attacks, its accuracy varies according to the records that are incorrectly classified (FN and FP). Since the FP percentages were very low for Bayesian Filters, with values less than 1 %, we suggest an improved Bayesian Filter. The Improved Bayesian Filter will trust the testing engine for its classification of attacks records, however normal records will enter a different engine to be filtered. The process can be repeated as many times as needed to seek the required accuracy. The improved Bayesian Filter is illustrated in Fig. 3.2 where we have a nested loop of filters where each filter uses its preceding's normal output as an input. The attack records are collected from each engine.

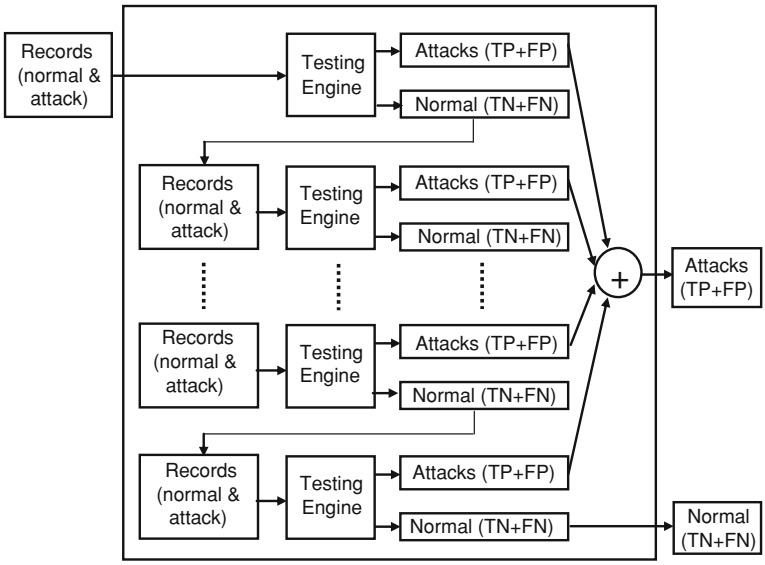


Fig. 3.2 Improved Bayesian filter 1

However, each engine should use a different setup to be able to catch more attacks otherwise nothing will be changed. The settings that can be changed are the threshold and most important is the features' selections. By in depth study of the records' behavior, appropriate features can be selected to increase the accuracy.

3.7.1.2 Improved Bayesian Filter 2 (IBF2):

We have noticed that the accuracy of each filter and the DR varies based on the database used each time. Moreover, mostly the attack type that is used to train the filter could score the best accuracy We, therefore, we suggest using multilayer engines with different databases for each layer as illustrated in Fig. 3.3.

This filter is optimized as follows:

Testing Engine 1: this engine will use the database optimized to detect DOS attacks. Therefore, it will have the best results in detecting DOS attacks type. The output of this engine that is classified as normal will be sent as inputs to testing engine 2 for more filtration.
Testing Engine 2: this engine will use the database optimized for PROBING attack type. The output of this engine that classified as normal will be sent as inputs to testing engine 3 for more filtration.
Testing Engine 3: this final engine will use the database and setup that is optimized to detects U2R and R2L attacks. We assumed that before reaching this stage we would have removed most of DOS and PROBING attacks and this engine will score the best DR for U2R and R2 l attacks.

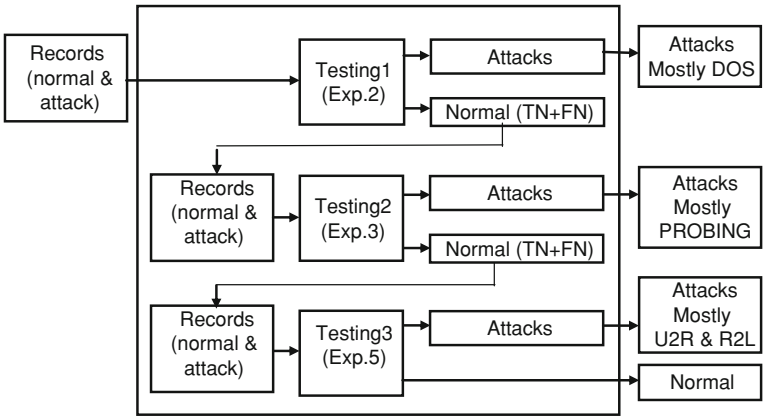


Fig. 3.3 Improved Bayesian filter 2

This 3-layer improved Bayesian filter can detect the four attacks types with high DR. However, the FP percentage reaches 31.97 % considered as a weakness in this filter.

3.7.1.3 Improved Bayesian Filter 3(IBF3):

Using the ideas that were explained in IBF1 and IBF2, we conducted many experiments. We got the best results so far by using two layers filter (Fig. 3.4). The first layer used all attack records and the second layer used the R2L attacks records in addition to the normal traffic.

The Training Engine (layer 1): the training data was all the records available for training which classified as normal (92,827 records) and attacks or not normal (401,195 records).

The Testing Engine (layer 1): the setting used for the training engine was: threshold = 0.9 and all features used to build the score map for the training engine. The data that was tested are normal (60,593 records) and attacks or not normal (250,436 records).

The results (layer 1) : they are classified into four categories as follows:

TN: 60,163 normal records classified successfully with 99.3 % as normal records.

TP: 224,893 attack records classified successfully with 90 % as attacks records.

FP: 430 normal records classified by mistake with 0.7 % as attacks records.

FN: 25,543 attack records classified by mistake with 10 % as normal attacks.

Therefore the DR = 90 % and the CR = 94.65 %.

Although we achieved good results in general and especially to detect attacks with very low of FP, we wanted better results to improve DR and to reduce the FN. Therefore, we will trust the first filter (layer 1) when it classified records as attacks since it gave excellent results with just 0.7 % FP. However, the normal records

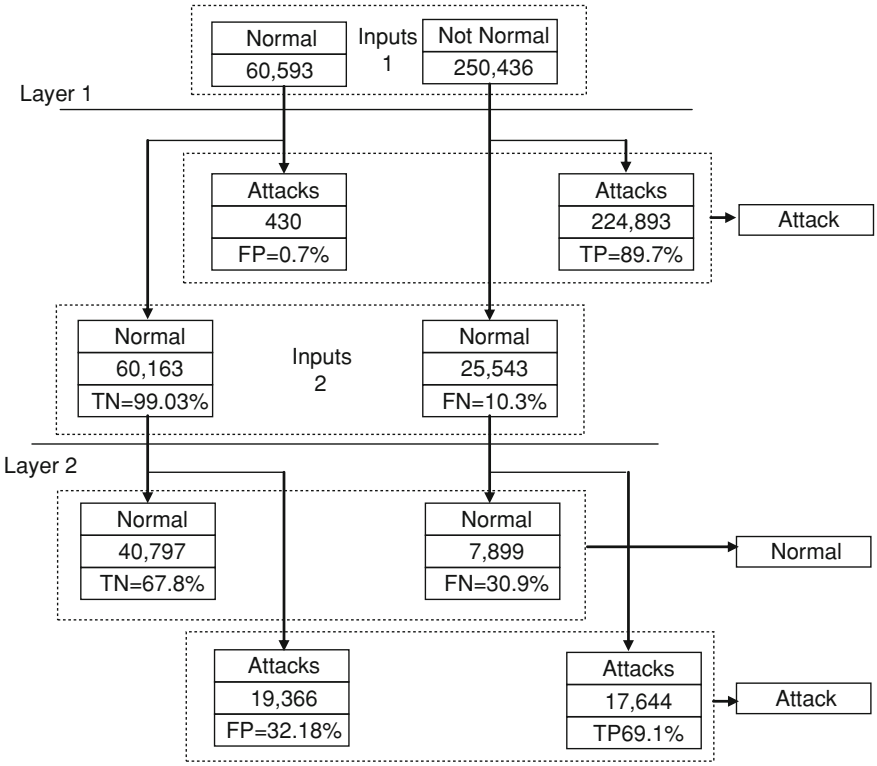


Fig. 3.4 Improved Bayesian filter 3

that classified by layer 1′s filter, need more filtration. Thus, we will add another filter as follows:

The Training Engine (layer 2): the training data was normal (92,827 records) and R2L attacks (1,126 records).

The Testing Engine (layer 2): the setting used for the training engine was: threshold = 0.6 and features 23,24,31 to build the score map for the training engine. The data that was tested are the data that was classified by filter 1 as normal data. However, these data contains (60,163 normal records) and (25,543 attacks records). The question might be raised: why did we select R2L attacks to train the engine? And the answer is simply because by analysing the attacks' types, we found that R2L attacks were 63 % of the attacks found while DOS, Probing and U2R were 29, 7 and 0.9 % respectively. Thus, we chose the attack type that has majority among the attacks.

The results (layer 2) : they are classified to four categories as follows:

TN: 40,797 normal records classified successfully with 67.8 % as normal records.

TP: 17,644 attack records classified successfully with 69 % as attacks records.

FP: 19,366 normal records classified by mistake with 32.18 % as attacks records.

FN: 7,899 attack records classified by mistake with 30.9 % as normal attacks.

The DR = 69 % and the CR = 68.4 %.

However, since the IBF3 contains both layers, we will have an overall results as follows:

TN: 40,797 normal records classified successfully with 67.8 % as normal records.

TP: 17,644 + 224,893 = 242,537 attack records classified successfully with 96.85 % as attacks records calculated as 242,537/250,436.

FP: 19,366 + 430 = 19,796 normal records classified by mistake with 32.67 % as attacks records calculated as 19,796/60,593.

FN: 7,899 attack records classified by mistake with 3.15 % as normal attacks calculated as 7,899/250,436.

The overall DR = 96.85 % and the overall CR = 82.1 %.

These results are the best results comparing to all the experiments that we conducted for all data. Moreover, this experiment is considered much realistic and practical since the data came to the filter not classified as normal nor attacks but as raw data and the filter should classify them according to its setup and database.

H. Improved Multi-method Filter

Finally, we decided to build a multistage filter, that is optimized by using different techniques in each stage. The initial stage would use a rule based filter, e.g. Snort [20] that has the a very low FP rate but is not able to detect new attacks that do not have rules. This filter will be used to capture all known attacks. Then this filter is followed by a Bayesian based filter that has been designed to catch the new unknown types of attacks.

3.8 Conclusions

Since the goal of this research was to improve the accuracy of the R2L attack using Bayesian methods, we have succeeded in achieving our target by using the Bayesian method as an engine to classify the data accordingly. We achieve results superior to Chou in his PhD dissertation [17], where he achieved a DR of 69.82 % for the R2L. Our research results show that we could have better results for R2L attack with a DR of 85.35 % by using the three features: 23, 24 and 31 and a threshold value of 0.6. However, the CR which equals 76.69 % is considered low comparing to Chou result because we used a low threshold value which reduces the accuracy of detection of normal records (TN) but increases the DR for R2L attack. To improve the accuracy of an IDS system we propose that we should use several Bayesian filters in parallel with each filter optimized to detect one type of record; this can be a good subject for further research in this field.

Furthermore we have shown that using multiple Bayesian filters in series with each filter optimized for a specific attack type achieves results that are better than what can be

achieved by a single filter. Moreover, using a two optimized Bayesian filters we were able to achieve an overall DR = 96.85. We believe that having multiple Bayesian filters in series will allow us to detect attacks with a high degree of confidence.

Finally, we believe that the best possible intrusion detection systems are multilayer system, with each layer built using a different technique. Where we have the initial layers built using rule based filters that are able to capture all known attacks. These layers are followed by Bayesian filters that can capture new types of attacks with a high degree of confidence.

References

1. Crothers T (2003) Implementing intrusion detection systems: a hands-on guide for securing the network. Wiley, Indianapolis
2. Bace R, Mell P (2001) NIST special publication on intrusion detection systems, National Institute of Standards and Technology
3. Agarwal R, Joshi M (2000) PNrule: a new framework for learning classifier models in data mining (a case-study in network intrusion detection)
4. Levin I (2000) KDD-99 classifier learning contest LLSoft's results overview. ACM SIGKDD Explorations I(2):67–75
5. Grapham P (2004) Hackers and painters: big ideas from the computer age, O'Reilly
6. Issac B, Jap W, Sutanto J (2009) Improved bayesian anti-spam filter Iimplementation and analysis on independent spam corpuses. In: international conference on computer engineering and technology, ICCET, Singapore, 2009
7. Alkabani Y, El-Kharashi M, Bedor H (2006) Hardware/software partitioning of a bayesian spam filter via hardware profiling. In: IEEE international symposium on industrial electronics, Canada, 2006
8. Chien J-T, Huang C-H, Shinoda K, Furui S (2006) Towards optimal bayes decision for speech recognition. In: IEEE international conference on acoustics, Speech and Signal Processing, ICASSP, Toulouse, 2006
9. Shi X, Manduchi R (2003) A study on bayes feature fusion for image classification. In: conference on computer vision and pattern recognition workshop, CVPRW, Madison, 2003
10. Kruegel C, Mutz D, Robertson W, Valeur F (2003) Bayesian event classification for intrusion detection. In: 19th annual computer security applications conference (ACSAC), IEEE Computer Society, Las Vegas
11. Cemerlic A, Yang L, Kizza J (2008) Network intrusion detection based on bayesian networks. In: Proceedings of the twentieth international conference on software engineering and knowledge engineering, SEKE, CA, 2008
12. Mehdi M, Zair A, Anou A, Bensebti M (2007) A bayesian networks in intrusion detection systems. J Comput Sci 3(5):259–265
13. Darwiche A (2010) Bayesian networks. Commun ACM 53(12):80–90
14. KDD Cup (1999) Data, 1999. [Online]. Available. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
15. Aickelin U, Twycross J, Hesketh-Roberts T (2007) Rule generalization in intrusion detection systems using SNORT. Int J Electron Secur Digit Forensics 1(1):101–116
16. Lee W, SSJ, Mok K (1999) A data mining framework for building intrusion detection models. In: Proceedings of the 1999 IEEE symposium on security and privacy, Oakland
17. Chou TS (2007) Ensemble fuzzy belief intrusion detection design, Florida International University, Paper AAI3299199
18. Altwaijry H, Algarni S (2012) Bayesian based intrusion detection system. CCIS J, 1:1–6

19. Altwaijry H, Algarny S (2011) Multi-layer bayesian based intrusion detection system. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science, WCECS 2011, San Francisco, 19-21 October, pp 918–922
20. Snort—Homepage, [Online]. Available. http://www.snort.org/

Chapter 4
The MAC Poisson Channel: Capacity and Optimal Power Allocation

Samah A. M. Ghanem and Munnujahan Ara

Abstract The majority of worldwide data and voice traffic is transported using optical communication channels. As the demand for bandwidth continues to increase, it is of great importance to find closed form expressions of the information capacity for the optical communications applications at the backbone as well as the access networks. In particular, we introduce an information-theoretic derivation of the capacity expressions of Poisson channels that model the application. The closed form expression for the capacity of the single input single output (SISO) Poisson channel-derived by Kabanov in 1978, and Davis in 1980 will be revisited. Similarly, we will derive closed form expressions for the capacity of the multiple accesses Poisson channel (MAC) under the assumption of constant shot noise. This provides a framework for an empirical form of the *k*-users MAC Poisson channel capacity with average powers that are not necessarily equal. Moreover, we interestingly observed that the capacity of the MAC Poisson channel is a function of the SISO Poisson channel and upper bounded by this capacity plus some quadratic non-linear terms. We have also observed that the optimum power allocation in the case of Poisson channels follows a waterfilling alike interpretation to the one in Gaussian channels, where power is allotted to less noisy channels. Therefore, we establish a comparison between Gaussian channels and Poisson optical channels in the context of information theory and optical communications.

S. A. M. Ghanem (✉)
Institute of Telecommunications, University of Porto, Porto, Portugal
e-mail: samah.ghanem@fe.up.pt

M. Ara
Institute of Telecommunications, University of Porto, Porto, Portugal
e-mail: munnujahan.ara@dcc.fc.up.pt

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_4,

Keywords Gaussian channels · MAC · Parallel channels · Poisson channels · Power allocation · SISO

4.1 Introduction

Information Theory provides one of its strongest developments via the notion of maximum bit rate or channel capacity. Determining an ultimate limit to the rate at which we can reliably transmit information over a physical medium in a given environment is an earnest attempt of fundamental and practical consideration. Such a limit is referred to as the channel capacity and the process of evaluating this limit leads to an understanding of the technical solutions required to approach it. Therefore, if the capacity can be found, then the goal of the engineer is to design an architecture which achieves that capacity. Capacity evaluations require information theory which must be adapted to the specific characteristics of the channel under study. The seminal work of Shannon published in 1948 [1] gave birth to information theory. Shannon determined the capacity of memoryless channels, including channels impaired by additive white Gaussian noise (AWGN) for a given signal-to-noise ratio (SNR). However, applying concepts of information theory to the optical communications channels encounters major challenges. The most important difficulty is dealing with the simultaneous interaction of specifically: The noise, filtering, and Kerr nonlinearity phenomena in the optical channel. These phenomena are distributed along the propagation path, and influence each other leading to deterministic as well as stochastic impairments [2].

Therefore, in this chapter, we accomplish an information-theoretic approach to derive the closed form expressions for the capacity of the SISO Poisson channel already found by Kabanov [3] and Davis [4], as well as for the k-user MAC Poisson channel using a direct detection or photon counting receiver and under constant noise; therefore, we simplify the framework of derivation. Several contributions have been done using information theoretic approaches to derive the capacity of Poisson channels under constant and time varying noise via martingale processes [3–7], or via approximations using Bernoulli processes [8], to define upper and lower bounds for the capacity and the rate regions of different models [9, 10], to define relations between information measures and estimation measures [11], in addition to deriving optimum power allocation for such channels [6, 7, 12]. However, in this contribution, we introduce a simple framework for deriving the capacity of Poisson channels for the model of consideration—The MAC Poisson channel—with the assumption of constant stochastic martingale noise, i.e. for the sake of simplicity, we didn't model the noise as Gaussian within the stochastic intensity rate process. In addition, we build upon derivations for the optimal power allocation.

In Poisson channels, the shot noise is the dominant noise whenever the power received at the photodetector is high; such noise is modeled as a Poisson random process. In fact, such framework has been investigated in many researches; see [2–7, 9–13]. Capitalizing on the expressions derived on [3, 4, 6, 7] and on the

results by [6, 7, 9], we investigate the derivation process of the channel capacity in a straightforward way; we then determine the optimal power allocation that maximizes the information rates. To derive the optimal power allocation for different channel frameworks, it's worth to notice that different optimization criteria could be relevant. In particular, the optimization criteria could be the peak power, the average optical power, or the average electrical power. The average electrical power is the standard power measure in digital and wireless communications and it helps in assessing the power consumption in optical communications, while the average optical power is an important measure for safety considerations and helps in quantifying the impact of shot noise in wireless optical channels. In addition, the peak power, whether electrical or optical, gives a measure of tolerance against the nonlinearities in the system, for example the Kerr nonlinearity which is identified by a nonlinear phase delay in the optical intensity or in other words as the change in the refractive index of the medium as a function of the electric field intensity.

4.2 The Communication Framework

In a communication framework, the information source inputs a message to a transmitter. The transmitter couples the message onto a transmission channel in the form of a signal which matches the transfer properties of the channel. The channel is the medium that bridges the distance between the transmitter and the receiver. This can be either a guided transmission such as a wire or a wave guide, or it can be an unguided free space channel. A signal traverses the channel will suffer from attenuation and distortion. For example, electric power can be lost due to heat generation along a wire, and optical power can be attenuated due to scattering and absorption by air molecules in a free space. Therefore, channels are characterized by a transfer function which models the input–output process. The input–output process statistics is dominated by the noise characteristics the modulated input experiences during its propagation along the communication medium, in addition to the detection procedure experienced at the channel output. In particular, when the noise $n_G(t)$ is a zero-mean Gaussian process with double-sided power spectral density *N0/2*, the channel is called an additive white gaussian channel (AWGN). However, when the electrical input is modulated by a light source, like a laser diode, the channel will be an optical channel with the dominant shot noise $n_d(t)$ arising from the statistical nature of the production and collection of photoelectrons when an optical signal is incident on a photodetector, such statistics characterized by a Poisson random process.

Figure 4.1 illustrates both the AWGN and the Poisson optical channels. In this chapter, we focus on the Poisson optical communication channel shown in Fig. 4.1b and we derive capacity closed form expression for the MAC Poisson channel capitalizing on the framework of derivation of the SISO Poisson channel capacity under a constant shot noise.

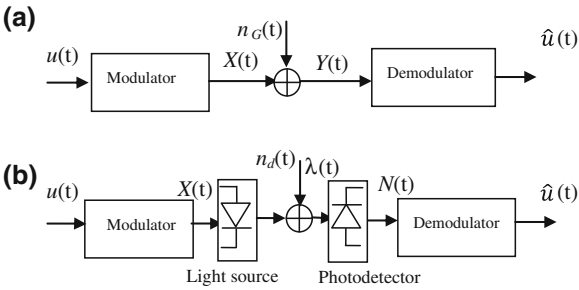


Fig. 4.1 **a** The AWGN channel. **b** The Poisson optical channel

4.3 The SISO Poisson Channel

Consider the SISO Poisson channel ***P*** shown in Fig. 4.2. Let $N(t)$ represent the channel output, which is the number of photoelectrons counted by a direct detection device (photodetector) in the time interval [0, T]. $N(t)$ has been shown to be a doubly stochastic Poisson process with instantaneous average rate $\lambda(t) + n$. The input $\lambda(t)$ is the rate at which photoelectrons are generated at time t in units of photons per second. And n is a constant representing the photodetector dark current and background noise.

4.3.1 Derivation of the Capacity of SISO Poisson Channels

Let $p(N_T)$ be the sample function density of the compound regular point process $N(t)$ and $p(N_T|S_T)$ be the conditional sample function of $N(t)$ given the message signal process $S(t)$ in the time interval [0, T]. Then we have,

$$p(N_T|S_T) = e^{-\int_0^T (\lambda(t)+n)d(t)+\int_0^T \log(\lambda(t)+n)dN(t)} \tag{4.1}$$

$$p(N_T) = e^{-\int_0^T \left(\widehat{\lambda(t)}+n\right)d(t)+\int_0^T \log\left(\widehat{\lambda(t)}+n\right)dN(t)} \tag{4.2}$$

We use the following consistent notation in the paper, $\widehat{\lambda(t)}$ is the estimate of the input $\lambda(t)$. $\mathbb{E}[\cdot]$ is the expectation operation over time. Therefore, the mutual information is defined as follows,

$$I(S_T; N_T) = \mathbb{E}\left[\log\left(\frac{p(N_T|S_T)}{p(N_T)}\right)\right] \tag{4.3}$$

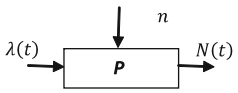


Fig. 4.2 The SISO Poisson channel model

Theorem 1 (Kabanov'78 [1]-Davis'80 [4]):
The capacity of the SISO Poisson channel is given by:

$$C = \frac{K}{P}(P+n)\log(P+n) + \left(1 - \frac{K}{P}\right) n \log(n) - (K+n)\log(K+n) \tag{4.4}$$

Proof:
Substitute (4.1, 4.2) in (4.3), we have,

$$I(S_T; N_T) = \mathbb{E}\left[-\int_0^T \lambda(t) - \widehat{\lambda(t)} dt + \int_0^T \log\left(\frac{\lambda(t)+n}{\widehat{\lambda(t)}+n}\right) dN(t) \right]$$

Since $\mathbb{E}\left[\widehat{\lambda(\mathrm{t})}\right] = \mathbb{E}[\mathbb{E}[\lambda(\mathrm{t})|N(t)]] = \mathbb{E}[\lambda(\mathrm{t})]$, it follows that,

$$I(S_T; N_T) = \mathbb{E}\left[\int_0^T \log\left(\frac{\lambda(t)+n}{\widehat{\lambda(t)}+n}\right) dN(t) \right]$$

And $N(t) - \int_0^{\mathrm{T}} \log(\lambda(t)+n)$ is a martingale from theorems of stochastic integrals, see [6, 11] therefore,

$$I(S_T; N_T) = \mathbb{E}\left[\int_0^T (\lambda(t)+n)\log\left(\frac{\lambda(t)+n}{\widehat{\lambda(t)}+n}\right) dt \right]$$

$$= \int_0^{\mathrm{T}} \mathbb{E}\lceil(\lambda(t)+n)\log(\lambda(t)+n)\rceil - \mathbb{E}\lceil(\lambda(t)+n)\log\left(\widehat{\lambda(t)}+n\right)dt\rceil$$

$$= \int_0^{\mathrm{T}} \mathbb{E}\lceil(\lambda(t)+n)\log(\lambda(t)+n)\rceil - \mathbb{E}\lceil\mathbb{E}[(\lambda(t)+n)]\log\left(\widehat{\lambda(t)}+n\right)|N_T dt\rceil$$

$$= \int_0^{\mathrm{T}} \mathbb{E}\lceil(\lambda(t)+n)\log(\lambda(t)+n)\rceil - \mathbb{E}\lceil\mathbb{E}([\lambda(t)+n])|N_T]\rceil\log\left(\widehat{\lambda(t)}+n\right)dt$$

$$= \int_0^T \mathbb{E}\lceil(\lambda(t)+n)\log(\lambda(t)+n)\rceil - \mathbb{E}\lceil\left(\widehat{\lambda(t)}+n\right)\log\left(\widehat{\lambda(t)}+n\right)dt\rceil \tag{4.5}$$

See [6, 7] for similar steps. In [11], it has been shown that the derivative of the input–output mutual information of a Poisson channel with respect to the intensity of the dark current is equal to the expected error between the logarithm of the actual input and the logarithm of its conditional mean estimate, it follows that,

$$\frac{dI(S_T;N_T)}{d\lambda(t)} = \mathbb{E}\left[\log\left(\frac{\lambda(t)+n}{\widehat{\lambda(t)}+n}\right)\right] \tag{4.6}$$

The right hand side term of (4.6) is the derivative of the mutual information corresponding to the integration of the estimation errors. This plays as a counter part to the well known relation between the mutual information and the minimum mean square error (MMSE) in Gaussian channels in [14].

The capacity of the SISO Poisson channel given in Theorem 1 (4.4) is defined as the maximum of (4.5) solving the following optimization problem,

$$\max I(S_T;N_T) \tag{4.7}$$

Subject to average and peak power constraints,

$$\frac{1}{T}\mathbb{E}\left[\int_0^T \lambda(t)dt\right] \le \sigma P$$

$$0 \le \lambda(t) \le P \tag{4.8}$$

With P is the maximum power and the ratio of average to peak power σ is used with $0 \le \sigma \le 1$. We can easily check that the mutual information is strictly convex via it $\lambda(t)$s second derivative with respect to as follows,

$$\frac{d^2I(S_T;N_T)}{d\lambda^2(t)} = \log\left(\frac{\lambda(t)+n}{\widehat{\lambda(t)}+n}\right) > 0.$$

Therefore, the mutual information is convex with respect to $\lambda(t)$.

Now solving:

$$\max\left(\int_0^T \mathbb{E}\lceil(\lambda(t)+n)\log(\lambda(t)+n)\rceil - \mathbb{E}\lceil\left(\widehat{\lambda(t)}+n\right)\log\left(\widehat{\lambda(t)}+n\right)\rceil - \frac{\xi}{T}\mathbb{E}\lceil\lambda(t)\rceil\right)$$

With ξ as the Lagrangian multiplier.

The possible values of $\mathbb{E}\lceil(\lambda(t)+n)\log(\lambda(t)+n)\rceil$ must lie in the set of all y-coordinates of the closed convex hull of the graph $\mathrm{y} = (\mathrm{x}+\mathrm{n})\log(\mathrm{x}+\mathrm{n})$. Hence, the maximum mutual information achieved using the distribution $p(\lambda =$

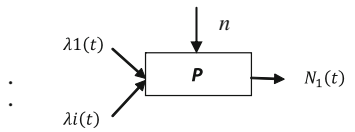


Fig. 4.3 The MAC Poisson channel model

$P) = 1 - p(\lambda = 0) = \alpha$. Where $0 \leq \alpha \leq 1$, so that $\mathbb{E}[\lambda(t)] = K$. So, we must have $\mathbb{E}[\lambda(t)] = \sum \lambda(t)p(\lambda)$. It follows that, $K = Pp(\lambda = P) = P\alpha$. Then, $\alpha = \frac{k}{P}$ and the capacity in Theorem 1 (4.4) is proved.

4.3.2 Optimum Power Allocation for SISO Poisson Channels

We need to solve the following optimization problem,

$$\max\left(\frac{K}{P}(P+n)\log(P+n) + (1-\frac{K}{P})n\log(n) - (K+n)\log(K+n) - \frac{\xi}{T}K\right) \tag{4.9}$$

Since (4.9) is concave with respect to K, i.e. the second derivative of (4.9) with respect to K is negative. Using the Lagrangian corresponding to the derivative of the objective with respect to K, and the Karush–Kuhn–Tucker (KKT) conditions, the optimal power allocation is the following,

$$K^* = (P+n)e^{-\left(1+\frac{\xi}{T}\right)+\frac{n}{P}\log\left(1+\frac{P}{n}\right)} - n \tag{4.10}$$

4.4 The MAC Poisson Channel

Consider the MAC Poisson channel shown in Fig. 4.3. Let us consider a 2-input MAC Poisson channel, then, $N_1(t)$ is a doubly stochastic Poisson process with instantaneous average rate $\lambda_1(t) + \lambda_2(t) + n$.

4.4.1 Derivation of the Capacity of MAC Poisson Channels

Let $p(N_1)$ and $p(N_1|S_1, S_2)$ be the joint density and conditional sample function of the compound regular point process $N_1(t)$ given the message signal processes $S_1(t)$ in the time interval [0, T]. Then we have,

$$p(N_1|S_1,S_2) = e^{-\int_0^T(\lambda_1(t)+\lambda_2(t)+n)dt+\int_0^T \log(\lambda_1(t)+\lambda_2(t)+n)dN(t)} \tag{4.11}$$

$$p(N_1) = e^{-\int_0^T(\widehat{\lambda_1(t)}+\widehat{\lambda_2(t)}+n)dt+\int_0^T \log\left(\widehat{\lambda_1(t)}+\widehat{\lambda_2(t)}+n\right)dN(t)} \tag{4.12}$$

Therefore, the mutual information is defined as follows,

$$I(S_T;N_T) = \mathbb{E}\left[\log\left(\frac{p(N_1|S_1,S_2)}{p(N_1)}\right)\right] \tag{4.13}$$

Theorem 2:
The capacity of the 2-input MAC Poisson channel is given by:

$$C = \left(\frac{k1}{P}+\frac{k2}{P}\right)(P+n)\log(P+n) + \left(1-\left(\frac{K1}{P}+\frac{K2}{P}\right)\right)n\log(n) - (K1+K2+n)\log(K1+K2+n) \tag{4.14}$$

Proof:
Substitute (4.11, 4.12) in (4.13), we have,

$$I(S_T;N_T) = \mathbb{E}\left[-\int_0^T\left(\lambda 1(t)-\widehat{\lambda 1(t)}\right)dt - \int_0^T\left(\lambda 2(t)-\widehat{\lambda 2(t)}\right)dt + \int_0^T \log\left(\frac{\lambda 1(t)+\lambda 2(t)+n}{\widehat{\lambda 1(t)}+\widehat{\lambda 2(t)}+n}\right)dN(t)\right]$$

Since $\mathbb{E}\left[\widehat{\lambda 1(t)}+\widehat{\lambda 2(t)}\right] = \mathbb{E}[\mathbb{E}[\lambda 1(t)+\lambda 2(t)|N_T]] = \mathbb{E}[\lambda 1(t)+\lambda 2(t)]$, it follows that,

$$I(S_T;N_T) = \mathbb{E}\left[\int_0^T \log\left(\frac{\lambda 1(t)+\lambda 2(t)+n}{\widehat{\lambda 1(t)}+\widehat{\lambda 2(t)}+n}\right)dN(t)\right]$$

And $N(t) - \int_0^T \log(\lambda 1(t)+\lambda 2(t)+n)$ is a martingale from theorems of stochastic integrals, see [6, 11] therefore,

$$I(S_T;N_T) = \mathbb{E}\left[\int_0^T (\lambda 1(t)+\lambda 2(t)+n)\log\left(\frac{\lambda 1(t)+\lambda 2(t)+n}{\widehat{\lambda 1(t)}+\widehat{\lambda 2(t)}+n}\right)dt\right]$$

$$
\begin{aligned}
&= \int_0^T \mathbb{E}\lceil(\lambda 1(t) + \lambda 2(t) + n)\log(\lambda 1(t) + \lambda 2(t) + n)\rceil \\
&\quad -\mathbb{E}\lceil(\lambda 1(t) + \lambda 2(t) + n)\log(\widehat{\lambda 1(t)} + \widehat{\lambda 2(t)} + n)\rceil dt \\
&= \int_0^T \mathbb{E}\lceil(\lambda 1(t) + \lambda 2(t) + n)\log(\lambda 1(t) + \lambda 2(t) + n)\rceil \\
&\quad -\mathbb{E}\lceil\mathbb{E}[(\lambda 1(t) + \lambda 2(t) + n)]\log(\widehat{\lambda 1(t)} + \widehat{\lambda 2(t)} + n)|N_T\rceil dt \\
&= \int_0^T \mathbb{E}\lceil(\lambda 1(t) + \lambda 2(t) + n)\log(\lambda 1(t) + \lambda 2(t) + n)\rceil \\
&\quad -\mathbb{E}\lceil\mathbb{E}\lceil(\lambda 1(t) + \lambda 2(t) + n)|N_T\log(\widehat{\lambda 1(t)} + \widehat{\lambda 2(t)} + n)\rceil dt \\
&= \int_0^T \mathbb{E}\lceil(\lambda 1(t) + \lambda 2(t) + n)\log(\lambda 1(t) + \lambda 2(t) + n)\rceil \\
&\quad -\mathbb{E}\lceil\mathbb{E}\lceil(\lambda 1(t) + \lambda 2(t) + n)|N_T\log(\widehat{\lambda 1(t)} + \widehat{\lambda 2(t)} + n)\rceil dt
\end{aligned} \tag{4.15}
$$

The capacity of the MAC Poisson channel given in Theorem 2 (4.14) is defined as the maximum of (4.15) solving the following optimization,

$$\max I(S_T; N_T) \tag{4.16}$$

Subject to average and peak power constraints,

$$
\begin{aligned}
&\frac{1}{T}\mathbb{E}\left[\int_0^T (\lambda 1(t) + \lambda 2(t))dt\right] \le \sigma P \\
&0 \le \lambda 1(t) \le P1 \\
&0 \le \lambda 2(t) \le P2
\end{aligned} \tag{4.17}
$$

With $P1$ and $P2$ are the maximum power and the ratio of average to peak power σ is used with $0 \le \sigma \le 1$. Now, solving:

$$
\max\left(\int_0^{\mathrm{T}} \mathbb{E}\lceil(\lambda 1(t) + \lambda 2(t) + n)\log(\lambda 1(t) + \lambda 2(t) + n)\rceil - \mathbb{E}\lceil(\widehat{\lambda 1(t)} + \widehat{\lambda 2(t)} + n)\right.
$$

$$
\left.\log\left(\widehat{\lambda 1(t)} + \widehat{\lambda 2(t)} + n\right)\rceil - \frac{\xi}{\mathrm{T}}\mathbb{E}[\lambda 1(t) + \lambda 2(t)]\right),
$$

with ξ as the Lagrangian multiplier. The possible values of $\mathbb{E}\lceil(\lambda 1(t)+\lambda 2(t)+n)\log(\lambda 1(t)+\lambda 2(t)+n)\rceil$ must lie in the set of all y-coordinates of the closed convex hull of the graph $y=(x1+x2+n)\log(x1+x2+n)$. Suppose that the maximum power for both inputs is $P1+P2=\sigma P$. Hence, the maximum mutual information achieved using the distribution $p(\lambda=P)=1-p(\lambda=0)=\alpha$. Where $0\leq\alpha\leq 1$ so that $\mathbb{E}[\lambda 1(t)]=K1$, $\mathbb{E}[\lambda 2(t)]=K$. So, we have $\mathbb{E}[\lambda 1(t)+\lambda 2(t)]=\sum(\lambda 1(t)p(\lambda 1)+(\lambda 2(t)p(\lambda 2)$. It follows that, $K1=Pp(\lambda 1=P)=P\alpha$. $K2=Pp(\lambda 2=P)=P(1-\alpha)$. Then, $\alpha=\frac{k1}{P}$ and $1-\alpha=\frac{k2}{P}$ and then the capacity in Theorem 2 (4.14) is proved and can be maximized when $\frac{k1}{P}=\frac{k2}{P}$.

It's worth to note that we also have $K3=P1p(0\leq\lambda 1(t)\leq\sigma P)+P2p(0\leq\lambda 2(t)\leq\sigma P)=P1\alpha+P2(1-\alpha)$, however, K3 is not considered in the capacity equations since we only need the maximum and the minimum powers for both $\lambda 1(t)$ and $\lambda 2(t)$ to get the maximum expected value. Therefore, our framework of derivation differs from [9] by solving the problem geometrically.

4.4.2 Optimum Power Allocation of MAC Poisson Channels

We need to solve the following optimization problem,

$$\max\left(\left(\frac{K1}{P}+\frac{K2}{P}\right)(P+n)\log(P+n)+\left(1-\left(\frac{K1}{P}+\frac{K2}{P}\right)\right)n\log(n)\right.$$
$$\left.-(K1+K2+n)\log(K1+K2+n)-\frac{\xi}{T}(K1+K2)\right) \tag{4.18}$$

Using the Lagrangian corresponding to the derivative of the objective with respect to K, and the Karush–Kuhn–Tucker (KKT) conditions, the optimal power allocation is the solution of the following equation,

$$K1^*+K2^*=(P+n)e^{-\left(1+\frac{\xi}{T}\right)+\frac{n}{P}\log\left(1+\frac{P}{n}\right)}-n \tag{4.19}$$

The optimum power allocation solution introduces the fact that orthogonalizing the inputs via time or frequency sharing will achieve the capacity; therefore, it follows the importance for interface solutions to aggregate different inputs to the Poisson channel.

4.5 MAC Poisson Channel Capacity and Rate Regions

We dedicate this section to analyze the result of Theorem 2. We will introduce the two-user MAC Poisson channel rate regions and we will then define the MAC capacity with respect to the SISO capacity and to bounds found mainly in [9]. The rate regions for the two-user MAC Poisson channel is given by,

$$R1 \leq I(S_1; N_1|S_2) \tag{4.20}$$

$$R2 \leq I(S_2; N_1|S_1) \tag{4.21}$$

$$R1 + R2 \leq I(S_1, S_2; N_1) \tag{4.22}$$

The mutual information that defines the sum of the rates $I(S_1, S_2; N_1)$ is defined in [Eq. 3.21, 9] under the condition that the average inputs for the two users are equal; in particular when both inputs are equiprobable. Here, we can manipulate this result into a sum rate upper bound with the two users having different average input powers as follows,

$$\begin{aligned} I(S_1, S_2; N_1) = &\left(\frac{K1}{P} + \frac{K2}{P}\right)(P+n)\log(P+n) + \left(1 - \left(\frac{K1}{P} + \frac{K2}{P}\right)\right) n\log(n) \\ &- (K1 + K2 + n)\log(K1 + K2 + n) \\ &- 2\left(\frac{K1^2}{P^2} + \frac{K2^2}{P^2}\right)(P+n)\log(P+n) \\ &+ \left(\frac{K1K2}{P^2}\right)(2P+n)\log(2P+n) + \left(\frac{K1K2}{P^2}\right) n\log(n) \end{aligned} \tag{4.23}$$

Where, $I(S_1, S_2; N_1)$ is maximized when $\frac{K1}{P} = \frac{K2}{P}$. It is important to notice that the first non-quadratic terms of $I(S_1, S_2; N_1)$ is the capacity of the SISO Poisson channel with the input as $\lambda 1(t) + \lambda 2(t)$. Therefore, we can see through Theorem2 that the capacity is approximately defined by the first term of $I(S_1, S_2; N_1)$,

$$I(S_1, S_2; N_1) = C_{SISO}(\lambda_1 + \lambda_2) + \beta \tag{4.24}$$

Where,

$$\begin{aligned} \beta = &-2\left(\frac{K1^2}{P^2} + \frac{K2^2}{P^2}\right)(P+n)\log(P+n) + \left(\frac{K1K2}{P^2}\right)(2P+n)\log(2P+n) \\ &+ \left(\frac{K1K2}{P^2}\right) n\log(n) \end{aligned} \tag{4.25}$$

Therefore, we can deduce that the rate region as defined in [9] is an upper bound for the capacity, and thus we can write an empirical form for the k-user MAC Poisson capacity, using the first non-quadratic terms of the above equation as follows,

$$C_{k-\text{user MAC}} = C_{\text{SISO}}(\lambda 1 + \ldots + \lambda k) \tag{4.26}$$

We can also verify Theorem 2 comparing it to the results in [9] for different setups, for example, consider the case when $K1 = K2 = K$, the capacity will be, $C = 2\frac{K}{P}(P+n)\log(P+n) + \left(1 - \frac{2K}{P}\right) n\log(n) - (2K+n)\log(2K+n)$.

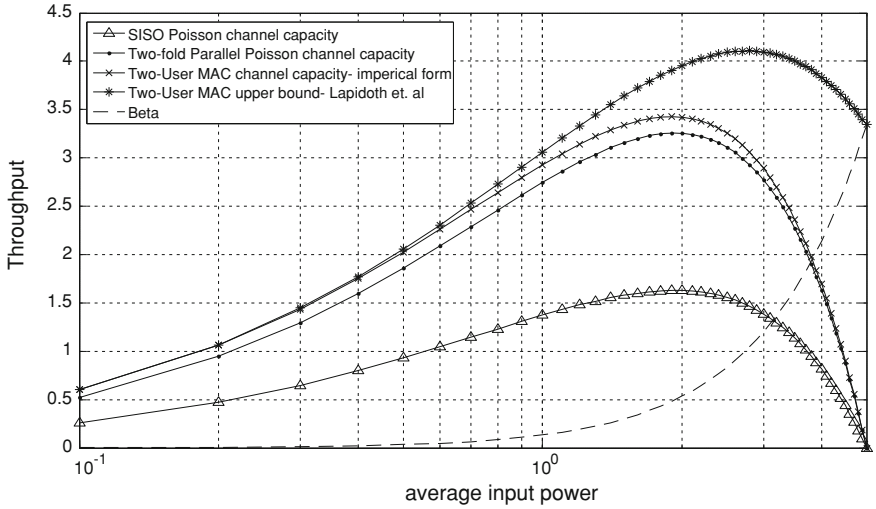


Fig. 4.4 Capacity of the Poisson channels versus the average power

When $K1 = K2 = K = P$, the negative terms indicates a zero capacity, $C = 0$, and when $K1 = K2 = K \neq P$, and $n = 0$ the capacity will be, $C = 2K \log \frac{P}{2K}$, and the rate sum will be, $I(S_1, S_2; N_1) = 2K \log \frac{P}{2K} + 2\frac{K^2}{P} \log 2$.

Therefore, $I(S_1, S_2; N_1)$ given in [Eq. 3.21, 9] upper bounds the capacity by the term $2\frac{K^2}{P} \log 2$, and via the constraints over the average power, $2\frac{K^2}{P} \log 2 \leq 2 \log 2$ it follows that this upper bounds the capacity with a value always less than or equal to 1.4 nats/sec for the two-user MAC. In a more generalized way, the empirical form differs from the upper bound by less than or equal to $k \log k$, where k corresponds to the number of inputs/users to the MAC Poisson channel. We can also verify that the maximum capacity achieved by orthognalizing the inputs such that the capacity approaches P/e nats/sec for each user. Therefore, non-orthognalizing the inputs incurs a maximum of around 0.5256 P power loss in the two-user MAC case. This well explains the limitation in the number of users for the MAC Poisson channel.

Figure 4.4 shows the capacity of different Poisson channels under a total power constraint of $P = 5$ on the SISO channel and each user's input of the parallel channel and the MAC channel, an equal average input power $K1 = K2 = K$, and shot noise $n = 0.1$. When the average input power is around one quarter the total power $K = P/4$, the rate is the maximum achievable rate, this explains the power loss in the two-user MAC case explained before. We can notice that the maximum mutual information presented by Lapidoth et al. in [Eq. 3.21, 9] upper bounds the rate region of all given channels, however, we can see that the maximum achievable rate is always $C \leq P/e$ nats/sec. In particular, for the MAC channel the maximum achievable rate with total power $P = 10$ is 3.425 nats/sec which is $C \leq 10/e \leq 3.7037$ nats/sec, i.e. the capacity for the k-user MAC is always $C \leq kP/e$.

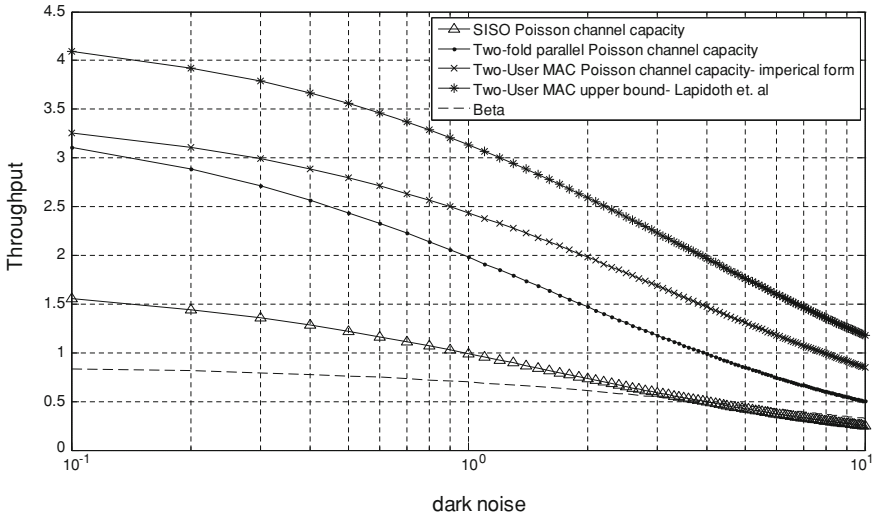


Fig. 4.5 Capacity of the Poisson channels versus the shot noise

We can further see that in the low average power regime, both the upper bound and the empirical capacity of the MAC matches, while it logarithmically differs at the high average power regime, this is due to the quadratic part that is missed in the empirical capacity formula denoted by β. Notice also that the MAC channel under the given conditions upper bounds the parallel channel, or in other words the parallel channel defines a lower bound over the MAC when both inputs are active.

Figure 4.5 shows the capacity of different Poisson channels with respect to the noise where naturally the capacity decreases with respect to the increase in the shot noise. However, it is of particular relevance to notice that in the low noise power regime, that Lapidoth upper bound for the MAC maximum achievable rate [9] indeed cannot be achieved due to existence of the quadratic terms, this gives rise of the achieved capacity over the right one, $C \leq kP/e$ however, our empirical form of the MAC capacity shows consistency regarding this relation and can be generalized to k-users.

4.6 Discussion

The solutions provided in this chapter show that the capacity of Poisson channels is a function of the average and peak power of the input. As a normal consequence to the expressions of the SISO Poisson channel, the Poisson parallel channels throughput is the sum of their independent SISO channels, proof is provided in [7]. For the MAC Poisson channel, the capacity expression derived here gives a generalization of a

closed form expression for the k-user MAC Poisson channel. The authors in [9] studied the capacity regions of the two-user MAC Poisson channels.

They also pointed out an interesting observation that we can also emphasize and verify via Theorem 2; that is; in contrary to the Gaussian MAC, in the Poisson MAC the maximum throughput is bounded in the number of inputs, and similar to the Gaussian MAC in terms of achieving the capacity via orthogonalizing the inputs or via the usage of a limited average input power for each user that is equal to one quarter the total power in the two-user MAC case. In fact, for the Poisson MAC, when equal input powers up to half the total power for each are used, the capacity faces a decay to zero, while when they differ i.e. inputs are orthogonal, the capacity is again maximized. In addition, we can also verify that the two main factors in the MAC capacity is the orthogonalization and the maximum power, while increasing the average power for one or the two inputs above a certain limit will not add positively to the capacity, see [7]. We can also see that the maximum power is a function of the average power through which both can be optimized to maximize the capacity.

Moreover, it can be deduced via the mathematical formulas that the power allocation is a decreasing value with respect to the dark current for all Poisson channels. It means that the power allocation for the Poisson channels in some way or another follows a waterfilling alike interpretation to the one for the Gaussian setup where less power is allotted to the more noisy channels [7, 15]. However, it's well known that the optimum power allocation is an increasing function in terms of the maximum power.

4.6.1 Gaussian Channels Versus Poisson Channels

Here, we summarize some important points about the capacity of Poisson channels in comparison to Gaussian channels within the context of this work. Firstly, in comparison to the Gaussian capacity, the channel capacity of the Poisson channel is maximized with binary inputs, i.e. [0, 1], while the distribution that achieves the Gaussian capacity is a Gaussian input distribution. Secondly, the maximum achievable rates for the Poisson channel is a function of its maximum and average powers due to the nature of the Poisson process which follows a stochastic random process with martingale characteristics, while in Gaussian channels, the processes are random and modeled by the normal distribution. Thirdly, the optimum power allocation for the Poisson channels is very similar for different models depending on the defined power constraints, and in comparison to the Gaussian optimum power allocation; it follows a similar interpretation to the waterfilling, at which more power is allocated to stronger channels, i.e. power allocation is inversely proportional to the more noisy channel. However, although the optimal inputs distribution for the Poisson channel is a binary input distribution, the optimal power allocation is a waterfilling alike, i.e. unlike the Gaussian channels with arbitrary inputs where it follows a mercury-waterfilling interpretation to compensate for the non-Gaussianess

in the binary input [16]. Finally, it is worth to emphasize two more important differences that were already shown in [10], which can be straight forward to proof here: Unlike the Gaussian channels, in Poisson channels, due to the characteristics of the Poisson distribution, we cannot implement interference cancellation techniques, since it is not possible to construct the probability of $p(N_1 = \lambda 1 + n)$ from the probability $p(N_1 = \lambda 1 + \lambda 2 + n)$ if $\lambda 2$ is considered as an interferer to $\lambda 1$. Besides, unlike Gaussian channels, Poisson channels are scale-invariant, since $p(N_1 = \lambda 1 + n/a) \neq p(N_1 = a\lambda 1 + n)$, if a scaling factor $a \neq 1$ is multiplied to the inputs, the mutual information $I(S_1, S_2; N_1 = a\lambda 1 + a\lambda 2 + n) \neq I(S_1, S_2; N_1 = \lambda 1 + \lambda 2 + n/a)$.

4.7 Conclusions

In this chapter, we show via an information theoretic approach that the capacity of optical Poisson channels is a function of the average and maximum power of the inputs, the capacity expressions have been derived as well as the optimal power allocation for the SISO and the MAC channel models. We provide a closed form expression for the *k*-user MAC Poisson channel with any average input powers. It is shown-through the limitation on users within the capacity of the Poisson MAC- that the interface solutions for the aggregation of multiple users/channels over a single Poisson channel are of great importance. However, a technology like orthogonal frequency division multiplexing (OFDM) for optical communications stands as one interface solution. While it introduces attenuation via narrow filtering, etc. it therefore follows the importance of optimum power allocation which can mitigate such effects, hence, we build upon optimum power allocation derivations.

Acknowledgments The authors would like to thank Prof. Dr. Izzat Darwazeh for his insightful comments that help improve the presentation of this work.

References

1. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423 and 623–656
2. Essiambre R-J, Kramer G, Winzer PJ, Foschini GJ, Goebel B (2010) Capacity limits of optical fiber networks. J Lightwave Technol 28(4):662–701
3. Kabanov YM (1978) The capacity of a channel of the Poisson type. Theory Prob Appl 23(1):143–147
4. Davis MHA (1980) Capacity and cutoff rate for Poisson-type channels. IEEE Trans Inf Theory 26(6):710–715
5. Fray M (1991) Information capacity of the Poisson channel. IEEE Trans Inf Theory 37(2):244–256
6. Han H (1981) Capacity of multimode direct detection optical communication channel. TDA Progress Report 42–63

7. Ghanem SAM, Ara M (2011) Capacity and optimal power allocation of Poisson optical communication channels, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2011, WCECS 2011, 19–21 October 2011, San Francisco, pp 817–821
8. Johnson DH, Goodman IN (2008) Inferring the capacity of the vector Poisson channel with a Bernoulli model. Netw Comput Neural Syst 19(1):13–33
9. Lapidoth A, Shamai S (1998) The Poisson multiple access channel. IEEE Trans Info Theory 44(1):488–501
10. Lifeng L, Yingbin L, Shamai S (2010) On the capacity region of the Poisson interference channels. IEEE International Symposium on Information Theory, Austin
11. Ongning G, Shamai S, Verdu S (2008) Mutual information and conditional mean estimation in Poisson channels. IEEE Trans Inf Theory 54(5):1837–1849
12. Alem-Karladani MM, Sepahi L, Jazayerifar M, Kalbasi K (2009) Optimum power allocation in parallel Poisson optical channel. 10th international conference on telecommunications, Zagreb
13. Segall A, Kailath T (1975) The modeling of randomly modulated jump processes. IEEE Trans Inf Theory IT-21(1):135–143
14. Dongning G, Shamai S, Verdu S (2005) Mutual information and minimum mean-square error in Gaussian channels. IEEE Trans Inf Theory 51
15. Cover TM, Thomas JA (2006) Elements of information theory, Vol 1, 2nd edn. Wiley, New York
16. Lozano A, Tulino A, Verdú S (2005) Mercury/waterfilling: optimum power allocation with arbitrary input constellations. IEEE International Symposium on Information Theory, Australia

Chapter 5
An Efficient Dispersion Control Chart

Saddam Akber Abbasi and Arden Miller

Abstract Control chart is the most important Statistical Process Control tool used to monitor reliability and performance of industrial processes. For monitoring process dispersion, *R* and *S* charts are widely used. These control charts perform better under the ideal assumption of normality but are well known to be very inefficient in presence of outliers or departures from normality. In this study we propose a new control chart for monitoring process dispersion, namely the *D* chart, and compared its performance with *R* and *S* charts using probability to signal as a performance measure. It has been observed that the newly proposed chart is superior to *R* chart and is a close competitor to S chart under normality of quality characteristic. When the assumption of normality is violated, *D* chart is more powerful than both *R* and *S* charts. This study will help quality practitioners to choose an efficient and robust alternative to *R* and *S* charts for monitoring dispersion of industrial processes.

Keywords Control Chart · Monte Carlo Simulations · Non-Normality · Probability to Signal · Process Dispersion · Process Monitoring

S. A. Abbasi (✉) · A. Miller
Department of Statistics, The University of Auckland,
Pri-vate Bag 92019, Auckland 1142, New Zealand
e-mail: sabb025@aucklanduni.ac.nz

A. Miller
e-mail: miller@stat.auckland.ac.nz

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_5,

5.1 Introduction

Control chart introduced by Walter A. Shewhart in 1920's, is the most important Statistical Process Control (SPC) tool used to monitor reliability and performance of industrial processes. The basic purpose of implementing control chart procedures is to detect abnormal variations in the process (location & scale) parameters. Although first proposed for manufacturing industry, control charts have recently been applied in a wide variety of disciplines, such as in nuclear engineering [7], health care [16], education [15], analytical laboratories [10] etc.

Monitoring process dispersion is an important component of SPC. Dispersion control charts are a well known tool used for improving process capability and productivity by reducing variability in the process. *R* and *S* charts are the two most widely used control charts for monitoring changes in process dispersion [11]. The design of these charts is based on estimating the process standard deviation σ using sample range and sample standard deviation respectively. These charts perform better under the ideal assumptions but are well known to be very inefficient when the assumption of normality is violated. In this study we propose a new dispersion control chart, namely the *D* chart, based on Downton's based estimate of process standard deviation. The design of *D* chart is established and is shown to be more efficient as compared to both *R* and *S* charts particularly for non-normal processes.

Assume X be a normally distributed quality characteristic with in-control mean μ and standard deviation σ (i.e. $X \sim N(\mu, \sigma^2)$). Let $X_1, X_2, \ldots, X_n$ represents a random sample of size n and the corresponding order statistics are represented by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The Downton's estimator is defined as (see [2, 5] and [1]):

$$D = \frac{2\sqrt{\pi}}{n(n-1)} \sum_{i=1}^{n} \left[i - \frac{1}{2}(n+1) \right] X_{(i)} \tag{5.1}$$

For normally distributed quality characteristic, *D* is an unbiased estimator of σ [3] and it has been shown in the past that *D* is not much affected by non-normality. The purpose of this study is to develop a variability chart based on *D* that performs better than existing variability charts, such as *R* or *S* charts, under the existence and violation of normality assumption. The rest of study is organized as follows: In the next section the widely used 3-sigma and probability limit structure of *D* chart is established following [14] and [6]. The following section compares the performance of *D*, *R* and *S* charts assuming normality of quality characteristics. The comparison is made using probability to signal as a performance measure. Fourth section presents comparison of these charts when the assumption of normality is violated and quality characteristic is assumed to follow non-normal (heavy tailed symmetric and skewed) distributions following [14] and [12]. Finally conclusions have been made in the last section.

5.2 Design of *D* Control Chart

Suppose the relationship between D and σ be defined by a random variable Z as $Z = D/\sigma$ (similar to $W = R/\sigma$ for R chart; [11]). For setting up control limits of the proposed D chart, estimates of σ and σ_D are required. By taking expectations on both sides of Z, we obtain:

$$E(Z) = E(D/\sigma) = E(D)/\sigma \tag{5.2}$$

$E(D)$ can be replaced with average of sample $D's$ ($\overline{D}$), computed from an appropriate number of random samples obtained from a process during normal operating conditions (similar to $\overline{R}$ and $\overline{S}$ used in the construction of R and S charts). Let $E(Z) = z_2$, as D is an unbiased estimator of σ hence we have $z_2 = 1$ (for every value of n). Thus under normality, an unbiased estimator of σ based on Downton's estimator is given as $\widehat{\sigma} = \overline{D}$.

Similarly for an estimate of σ_D we have $\sigma_Z = \sigma_D/\sigma$. Let $\sigma_Z = z_3$, hence we have

$$\sigma_D = z_3\sigma \tag{5.3}$$

Barnett et al. [3] showed that

$$var(D) = \frac{\sigma^2}{n(n-1)}\left\{n\left(\frac{1}{3}\pi + 2\sqrt{3} - 4\right) + \left(6 - 4\sqrt{3} + \frac{1}{3}\pi\right)\right\} \tag{5.4}$$

From Eqs. (5.3) and (5.4) we have

$$z_3 = \frac{1}{\sqrt{n(n-1)}}\sqrt{n\left(\frac{1}{3}\pi + 2\sqrt{3} - 4\right) + \left(6 - 4\sqrt{3} + \frac{1}{3}\pi\right)} \tag{5.5}$$

Replacing an estimate of σ (i.e. $\widehat{\sigma} = \overline{D}$) in Eq. (5.3), we obtain $\widehat{\sigma}_D = z_3\overline{D}$

Hence the widely used 3-sigma control limits for the proposed D chart are defined as

$$LCL = max(0, \overline{D} - 3z_3\overline{D}), \quad CL = \overline{D} \quad \text{and} \quad UCL = \overline{D} + 3z_3\overline{D} \tag{5.6}$$

$$LCL = Z_3\overline{D}, \quad CL = \overline{D} \quad \text{and} \quad UCL = Z_4\overline{D} \tag{5.7}$$

where $Z_3 = max(0, 1 - 3z_3)$ and $Z_4 = 1 + 3z_3$. The coefficients z_3, Z_3 and Z_4 entirely depends on sample size n and are given in Table 5.1 for some representative values of sample size n. After setting up control limits, sample statistic D is plotted against time or sample number. If all the plotted points lie inside the control limits we can say that the process variability is in statistical control otherwise if one or more points lie outside the control limits, the process variability is said to be out-of-control.

The use of 3-sigma limits is based on the symmetric assumption of the plotted statistic, we will see that the distribution of D is not symmetric atleast for small to moderate values of n. Hence there is a need to develop the probability limit structure for the proposed D chart. Probability limits for D chart can be computed by using the quantile points of the distribution of Z. Let α be the specified probability of making Type-I error, denoting α-quantile of the distribution of Z by Z_α, the probability limits based on D are given as:

$$\begin{aligned} LCL &= Z_{(\alpha/2)}\overline{D} \quad \text{with} \quad \Pr(Z \le Z_{(\alpha/2)}) \quad = \alpha/2 \\ UCL &= Z_{(1-\alpha/2)}\overline{D} \quad \text{with} \quad \Pr(Z \ge Z_{(1-\alpha/2)}) = 1 - \alpha/2 \end{aligned} \tag{5.8}$$

These quantile points have been computed through extensive Monte Carlo simulation routines. The distribution of Z is obtained by generating 10,000 samples of size $n = 2, 3, \cdots, 15, 20, 25, 35, 50, 75$ and 100 from standard normal distribution. For a specified Type-I error probability α, $(\alpha/2)^{th}$ and $((1-\alpha)/2)^{th}$ quantile points have been computed from the distribution of Z for every combination of α and n. The same procedure is repeated 1000 times and the mean values of the quantile points together with their standard errors are reported in Table 5.1. The 3-sigma and probability limit structure of R and S charts with their respective control chart constants and quantile points can be seen in [13].

5.3 Comparison of D, R and S Charts for Normal Processes

In this section we provide comparison of the D, R and S Charts for normally distributed quality characteristic using probability to signal as the performance measure. For a fair comparison, different competing procedures needs to be adjusted to have the same false alarm probability and then comparison is made with respect to out-of-control detection probabilities. In our case, the process is said to be out-of-control whenever process standard deviation σ shifts from an in-control value, say σ_0 to another value say σ_1, where σ_1 is defined as $\sigma_1 = \sigma_0 + \delta\sigma_0$. For a fixed false alarm rate, control chart structure which gives highest probability to signal for out-of-control situations will indicate best performance as compared to other charts.

By setting up probability limits for $\alpha = 0.002$, probability to signal have been computed for both in-control and out-of-control situations for D, R and S charts using their respective control chart coefficients and quantile points. To save space and to aid in visual clarity, power curves have been constructed instead of presenting results in tabular form. The power curves of the three charts for normally distributed quality characteristics for $n = 5, 10$ and 15 are shown in Fig. 5.1.

From power curves in Fig. 5.1 we can observe that for zero sigma shift in process standard deviation, the probability of signaling is very close to 0.002 for all the charts and for every sample size, representing the case for an in-control process.

Table 5.1 Control chart constants and quantile points of the distribution of Z (standard errors)

n	$Z_{0.001}$	$Z_{0.01}$	$Z_{0.05}$	$Z_{0.95}$	$Z_{0.99}$	$Z_{0.999}$
2	0.00166	0.01565	0.0785	2.45975	3.23694	4.15268
	(0.00056)	(0.00015)	(0.001)	(0.00027)	(0.00111)	(0.00033)
3	0.03551	0.11492	0.25502	1.96152	2.439	3.00633
	(0.00084)	(0.00057)	(0.00086)	(0.00052)	(8e−04)	(0.00086)
4	0.09823	0.21368	0.37075	1.76101	2.12345	2.53082
	(0.00032)	(0.00111)	(0.00031)	(0.00109)	(0.00096)	(0.00069)
5	0.15217	0.28511	0.44493	1.64577	1.9552	2.30548
	(0.00025)	(0.00044)	(0.00027)	(0.00047)	(0.00076)	(0.00089)
6	0.20846	0.34765	0.50023	1.56683	1.8354	2.15353
	(0.00025)	(0.00048)	(0.00032)	(0.00057)	(0.00093)	(0.00036)
7	0.25049	0.39398	0.54303	1.51488	1.75368	2.02625
	(0.00013)	(2e−04)	(0.00061)	(0.00034)	(0.00076)	(0.00035)
8	0.30186	0.43455	0.57671	1.47484	1.68721	1.94266
	(0.00033)	(0.00087)	(0.00023)	(0.00025)	(0.00086)	(0.00028)
9	0.33462	0.46845	0.60199	1.443	1.64455	1.87114
	(0.00016)	(0.00025)	(0.00079)	(0.00011)	(0.00035)	(0.00081)
10	0.37155	0.49531	0.62393	1.41335	1.60123	1.81688
	(0.00028)	(0.00024)	(0.00046)	(0.00078)	(0.00063)	(0.00035)
11	0.38631	0.51896	0.63972	1.39357	1.56754	1.7785
	(9e−04)	(0.00036)	(0.00091)	(0.00024)	(0.00068)	(0.00073)
12	0.41242	0.53575	0.6575	1.37317	1.54059	1.7294
	(0.00109)	(0.00087)	(0.00039)	(0.00056)	(0.00076)	(0.00085)
13	0.43215	0.55201	0.6712	1.35789	1.52125	1.69148
	(0.00067)	(0.00034)	(0.00079)	(0.00021)	(0.00025)	(0.00063)
14	0.45467	0.57071	0.68466	1.34069	1.49412	1.68263
	(0.00031)	(0.00056)	(0.00077)	(0.00041)	(0.00033)	(0.00039)
15	0.46773	0.58356	0.6935	1.32709	1.47725	1.63709
	(0.00066)	(0.0011)	(0.00098)	(0.00022)	(0.00051)	(0.00081)
20	0.54297	0.63953	0.73694	1.27848	1.4035	1.54339
	(0.00042)	(0.00033)	(0.00092)	(0.00059)	(0.00017)	(0.00011)
25	0.57977	0.67448	0.76569	1.24488	1.35738	1.47917
	(0.00074)	(0.00028)	(0.00107)	(0.00053)	(0.00088)	(0.00072)
35	0.64444	0.72547	0.8022	1.20638	1.29519	1.4015
	(0.00022)	(0.00091)	(0.00012)	(0.00031)	(0.00067)	(0.00079)
50	0.70201	0.77009	0.83552	1.17129	1.24661	1.33503
	(7e−04)	(0.00084)	(0.00019)	(5e−04)	(0.00052)	(0.00094)
75	0.75708	0.8122	0.86561	1.13826	1.1971	1.26395
	(0.00031)	(0.00099)	(0.00053)	(0.00036)	(0.00097)	(0.00025)
100	0.78607	0.83591	0.88294	1.12025	1.17153	1.2276
	(0.00094)	(0.00079)	(0.00055)	(0.00105)	(0.00077)	(0.0011)

When the process is out-of-control, D chart is equally efficient to S chart for detecting shifts in process variability and have significantly higher probability to signal as compared to R chart, as the power curves of D chart coincides with that of S chart and always higher than the power curves of R chart for every choice of n. Hence

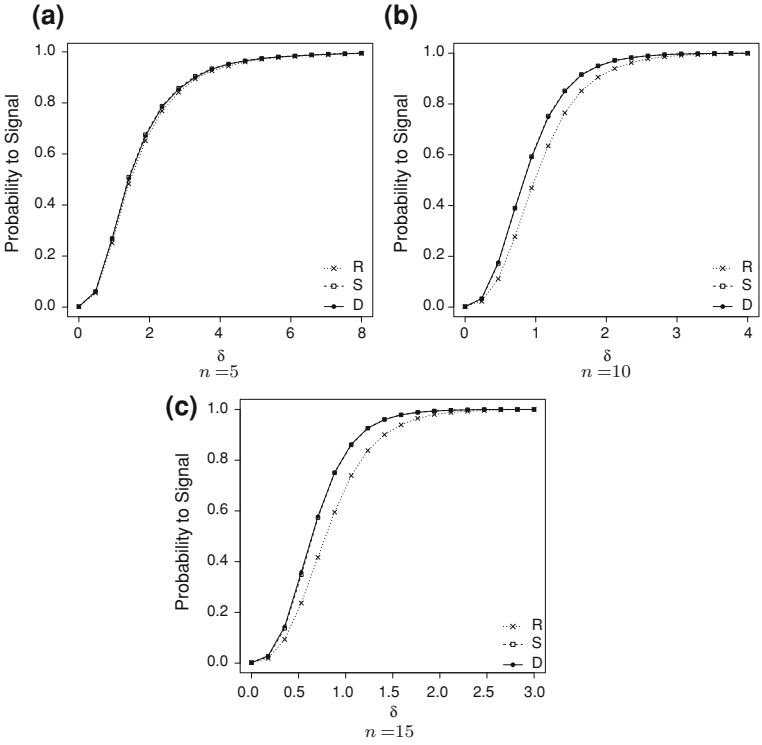


Fig. 5.1 Power curves of D, R and S charts for $n = 5, 10$ and 15 under Normal distribution when $\alpha = 0.002$

we can say that under the ideal assumption of normality D chart is more efficient than R chart and is a close competitor to S chart.

5.4 Comparison of D, R and S Charts for Non-Normal Processes

Normal distribution have wide applications in statistics and almost all SPC charts are based on this assumption. But in practice data from many real world processes follow non-normal distributions. To mention a few of such cases: [4] and [8] pointed out that quality characteristics such as capacitance, insulation resistance, surface finish, roundness, mold dimensions follow non-normal distributions. Levinson and Polny [9] indicates that impurity levels in semiconductor process chemicals follow Gamma distribution. Many other characteristics such as straightness, flatness, cycle time are not distributed normally. Hence there is a need to study the performance of these variability charts for different parent non normal

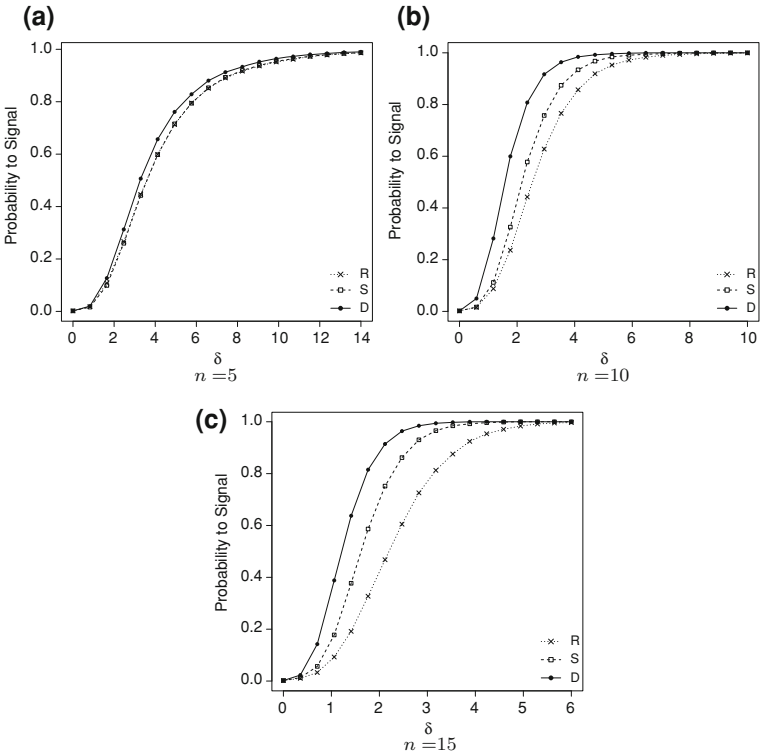


Fig. 5.2 Power curves of D, R and S charts for $n = 5, 10$ and 15 under Student's t distribution when $\alpha = 0.002$

distributions. To represent the case of non-normal processes, the performance of D, R and S charts is investigated by assuming that the quality characteristic follows heavy tailed symmetric Student's t and skewed Gamma and Weibull distributions. The density function of these non-normal distributions are given below:

$$\textbf{Student's t } (t_k): \quad f(x|k) = \frac{\Gamma[(k+1)/2]}{\sqrt{k\pi}\Gamma(k/2)}\left(1+\frac{x^2}{k}\right)^{-(k+1)/2}, \quad -\infty < x < \infty, k > 0$$

$$\textbf{Gamma}(\alpha,\beta): \quad f(x|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, \quad x > 0, \alpha > 0, \beta > 0$$

$$\textbf{Weibull}(\alpha,\beta): \quad f(x|\alpha,\beta) = \frac{\alpha}{\beta}x^{\alpha-1}e^{-x^{\alpha}/\beta}, \quad x \geq 0, \alpha > 0, \beta > 0$$

Probability to signal of D, R and S charts have been computed for these non-normal distributions using similar simulation routines as were used earlier for the case of normal distribution. In our simulation study we used Student's t distribution with $k = 5$, Gamma distribution with $\alpha = 2$ and $\beta = 1$, and finally Weibull distribution with $\alpha = 1.5$ and $\beta = 1$. The power curves of the three charts when quality characteristic is assumed to follow Student's t, Gamma and Weibull distributions are presented in Figs. 5.2, 5.3, 5.4 respectively. From Figs. 5.2, 5.3, 5.4 we can clearly see that the power curves of D chart are always higher than the power

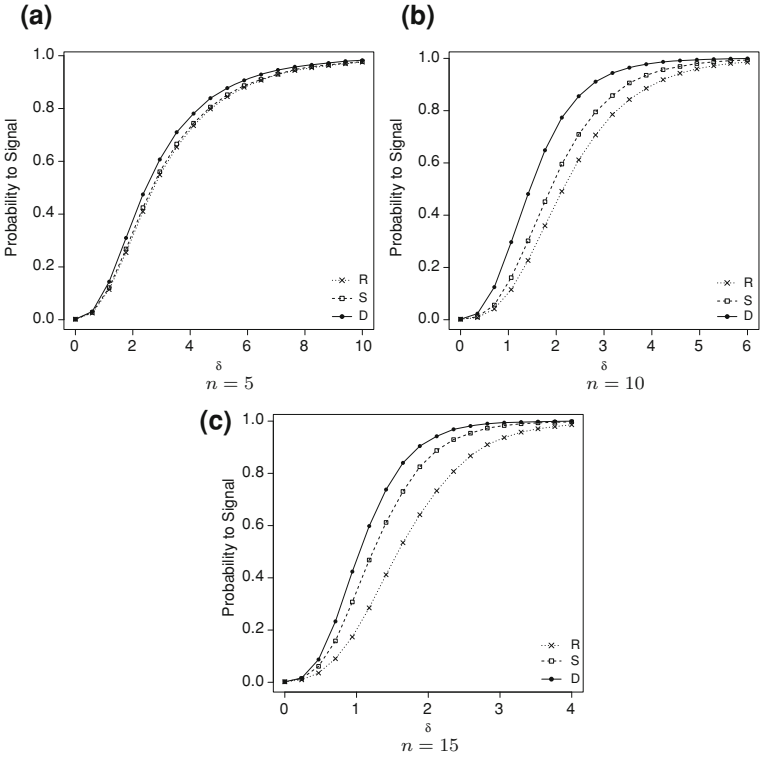


Fig. 5.3 Power curves of D, R and S charts for $n = 5, 10$ and 15 under Gamma distribution when $\alpha = 0.002$

curves of both R and S charts for all non-normal cases and for every choice of sample size n. This indicates that D chart has higher probability to signal shifts in process variability as compared to both R and S charts when the assumption of normality is violated. We can also observe that the difference in the detection ability of these charts increases with an increase in n. Relatively R chart is extremely affected while D chart is least affected by non-normality. Hence for non-normal processes, we can easily say that D chart is always superior than both R and S charts.

5.5 Conclusions

This study proposes an efficient control chart, namely the D chart, to monitor changes in process dispersion. The performance of the D chart is compared to the widely used R and S charts. It has been shown that for normally distributed quality characteristic, D chart is equally efficient to the S chart in terms of detecting shifts in process variability and has significantly better detection ability as compared to

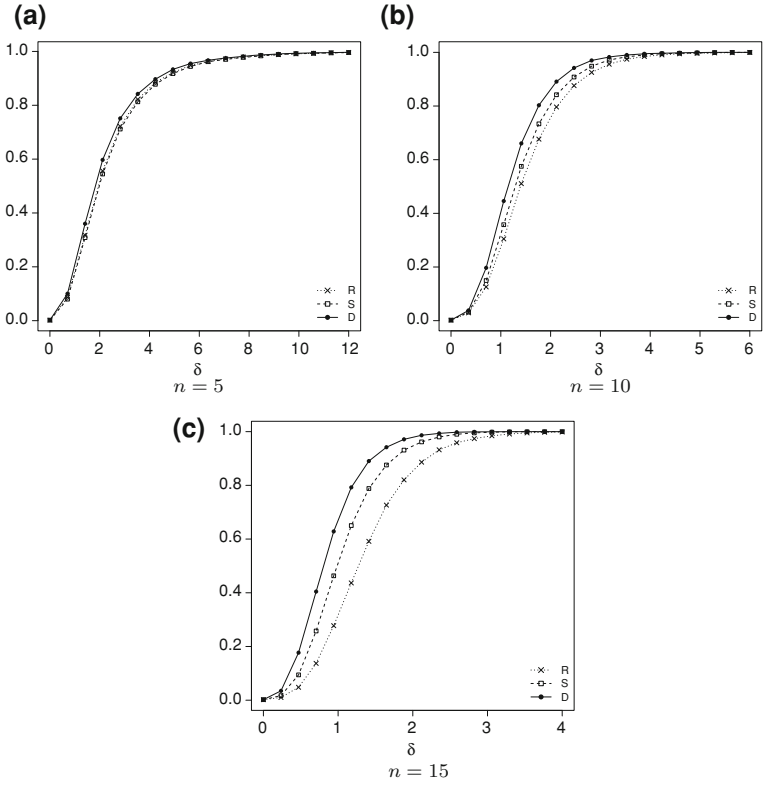


Fig. 5.4 Power curves of D, R and S charts for $n = 5, 10$ and 15 under Weibull distribution when $\alpha = 0.002$

the R chart. For non-normal processes, D chart clearly showed superiority over both R and S charts. Quality control practitioners can now easily choose D chart as a superior alternative to both R and S charts due to its efficient detection ability.

References

1. Abbasi SA, Miller A (2011) D chart: an efficient alternative to monitor process dispersion. Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science (2011) Vol II, WCECS 2011, 19–21 October, 2011. San Francisco, USA, pp 933–938
2. Abu-Shawiesh MO, Abdullah MB (2000) Estimating the process standard deviation based on downton's estimator. Qual Eng 12(3):357–363
3. Barnett FC, Mullen K, Saw JG (1967) Linear estimates of a population scale parameter. Biometrika 54:551–554
4. Bissell D (1994) Statistical methods for SPC and TQM. Chapman& Hall, New York
5. Downton F (1966) Linear estimates with polynomial coefficients. Biometrika 53(1):129–141

6. Gonzalez IM, Viles E (2001) Design of R control chart assuming a gamma distribution. Econ Qual Control 16:199–204
7. Hwang SL, Lin JT, Liang GF, Yau YJ, Yenn TC, Hsu CC (2008) Application control chart concepts of designing a pre-alarm system in the nuclear power plant control room. Nucl Eng Des 238(12):3522–3527
8. James PC (1989) C_pk equivalencies. Quality 28(9):75
9. Levinson WA, Polny A (1999) Spc for tool particle counts. Semicond Int 22(6):117–121
10. Masson P (2007) Quality control techniques for routine analysis with liquid chromatography in laboratories. J Chromatogr A 1158(1–2):168–173
11. Montgomery DC (2001) Introduction to statistical quality control. Wiley, New York
12. Ramalhoto MF, Morais M (1999) Shewhart control charts for the scale parameter of a weibull control variable with fixed and variable sampling intervals. Journal of Applied Statistics 26(1):1129–160
13. Ryan PR (2000) Statistical methods for quality improvement. Wiley, New York
14. Shewhart WA (1931) Economic control of quality manufactured product. In: Van Nostrand D (ed) Reprinted by the american society for quality control in 1980. Milwauker, New York
15. Wang Z, Liang R (2008) Discuss on applying spc to quality management in university education. In: Proceedings of the 9th international conference for young computer scientists. ICYCS 2008, pp 2372–2375
16. Woodall WH (2006) The use of control charts in health-care and public-health surveillance. J Qual Technol 38(2):89–104

Chapter 6
Operational Cost Reduction of an Activated Sludge System: Correlation Between Setpoint and Growth Substrate

George Simion Ostace, Anca Gal, Vasile Mircea Cristea and Paul Şerban Agachi

Abstract This work presents the optimization of two control strategies of the wastewater treatment plant. The control architectures are assessed from an operational costs point of view, and improved by adding an upper, supervisory level of control. The upper control level dictates the optimal set-point for the two control structures by taking into consideration the quantity of ammonia nitrogen that enters the wastewater treatment plant. The relationship between the quantity of ammonia nitrogen that enters the WWTP and the optimal setpoint was established by means of linear and polynomial interpolations. The study is based on the modified Benchmark Simulation Model No. 1 (BSM1). Two modifications were made to the BSM1. The first one is the implementation of an enhanced Activated Sludge Model No. 3. This model includes two additional processes that describe the direct growth of the heterotrophic biomass on readily biodegradable substrate, in both anoxic and oxic conditions. The 2nd modification of the BSM1 regards the secondary settler, which is considered to be reactive. The simulation results show that by using the supervisory level of control the total operational cost can be reduced with almost 15.5 %, while effluent standards are maintained.

G. S. Ostace (✉) · A. Gal · V. M. Cristea · P. Ş. Agachi
Faculty of Chemistry and Chemical Engineering, Babes-Bolyai University,
11 Arany Janos, 400028 Cluj-Napoca, Romania
e-mail: george.ostace@ubbcluj.ro

A. Gal
e-mail: anca.gal27@gmail.com

V. M. Cristea
e-mail: mcristea@chem.ubbcluj.ro

P. Ş. Agachi
e-mail: serban.agachi@ubbcluj.ro

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_6,

Keywords BSM1 · MPC · Operational costs · PI · Reactive settler · Setpoint optimization

6.1 Introduction

Considered to be one of the most economical efficient and technological sustainable processes for the wastewater treatment, the activated sludge process (ASP), is widely used for treating both household and industrial wastewaters. The process relies on different genders of bacteria that, under diverse environmental conditions, use pollutants from the wastewater as substrate for growth. Depending on the targeted effluent quality and wastewater characteristics, activated sludge systems (ASS) have different configurations that vary by alternating the environmental conditions (e.g. anaerobic, aerobic, and anoxic).

Due to the intricate behavior of the microorganisms and the constantly changing influent characteristics, the ASP is characterized by a strong nonlinearity, being hard to control and predict. In the last decades, mathematical models have become important tools for process prediction and development of new control strategies meant to reduce effluent pollutants and operational costs.

The current state of biological wastewater treatment modeling consists in the Activated Sludge Model suite, which includes the Activated Sludge Model No. 1, 2, 2d and 3 (ASM1, ASM2, ASM2d, ASM3) [1]. The ASM1 was first presented by Henze et al. [2]. It is mainly used for municipal activated sludge wastewater treatment plants and it describes the removal of organic carbon and ammonium nitrogen. ASM1 is based on eight biological processes describing the growth and decay of heterotrophic and autotrophic bacteria involved in the activated sludge process.

Henze et al. [3] presented the ASM2, a mathematical model for the ASP that also includes phosphorus removal. Later, this model was further improved by Henze et al. [4], and named ASM2d, a model similar to the ASM2, but incorporating the storage of poly-phosphate under anoxic conditions.

The ASM3 was introduced by Gujer [5] and it is the first activated sludge model that considers the internal storage of rapidly biodegradable substrate by the heterotrophic biomass. The model is developed to describe the removal of organic carbon and ammonium nitrogen and it is more complex than the ASM1. All these models have suffered extensions over the past years [6–15].

Wastewater treatment plants (WWTPs) should be controlled in a way that minimizes plant operational costs (OC), while effluent standards are carefully maintained [16]. The performance of WWTPs has been improved over the years by using automatic control [17, 18]. Attention has been also paid to set-point optimization [16, 19], in order to improve control performance.

The objective of this research is to investigate and propose a setpoint optimization scheme for two control strategies of the WWTP. The improved architecture

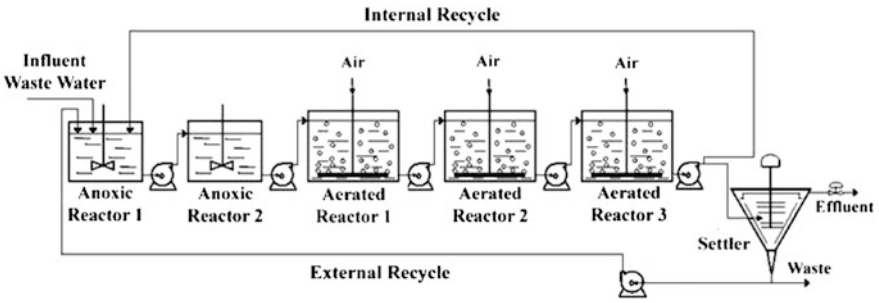


Fig. 6.1 Layout of the BSM1 benchmark simulation plant

is intended to reduce the operational costs by manipulating the setpoint of these control structures.

6.2 Materials and Methods

6.2.1 Simulated Plant Description

The Benchmark Simulation Model No. 1 wastewater treatment plant was developed by The International Association of Water Quality (IAWQ) and European Cooperation, in the field of Scientific and Technical Research (COST) 624 group in 2002 [20]. The purpose of the BSM1 is to study different control strategies for biological wastewater treatment plants. The wastewater treatment plant considered in the BSM1 is based on is as a Modified Ludzack-Ettinger (MLE) process. The MLE is one of the most common architectures used for biological nitrogen removal in municipal wastewater treatment.

The BSM1 plant has five biological reactors arranged in series (Fig. 6.1). The first two reactors are anoxic and each has a volume of 1,000 m^3. The last three reactors are aerated, each of them with a volume of 1,333 m^3. The total biological volume is 6,000 m^3.

The reactors are followed by a secondary settler that has a depth of 4 m and a cross-section of 1,500 m^2.

The plant has two recycle flows:

- The sludge recycle flow, from the bottom of the settler to the first anoxic tank Qrs = 18,446 m^3/day.
- The nitrates recycle flow, from the last aerated reactor to the first anoxic tank Qa = 55,338 m^3/day.

The aeration is indirectly manipulated with the oxygen transfer coefficient (KLa), which is constrained to a maximum of 240 day^{-1}. The waste flow rate

(Qw) is set to 358 $m^3 day^{-1}$, to ensure a sludge age close to 8 days just like for the original BSM1 simulator.

For this study, the biological kinetic model used to describe the carbon and nitrogen removal, is a modified ASM3 that considers the direct growth of the heterotrophic biomass on the primary substrate and on the internal storage products. The modified ASM3 has two extra biological processes which describe the direct growth on readily biodegradable substrate of the heterotrophic biomass in both anoxic and oxic conditions. Also, the growth on internal storage polymers is considered to take place only after the depletion of the primary substrate. Further details on the model may be found in [7].

The BSM1 considers that no biological activity occurs in the secondary settler of the simulated WWTP. The present study strives to achieve a better agreement between real WWTP behavior and the simulated mathematical model, and therefore assumes that the same reactions take place in the secondary settler as in the WWTP reactors. The reactive settler model is the combination of the model described by Takács in 1991 [21] and the activated sludge model [6, 7, 22, 23].

The influent composition was generated by adapting the original BSM1 influent files to the modified ASM3 [7]. These three influent files provide input data for a period of 14 days of operation, at an interval of 15 min and mimic dry, rain and storm weather conditions.

In this work, the same two-step simulation procedure was used as the one proposed in the BSM1. The first step is defined by the simulations up to steady state, followed by dynamic simulations using the three influent data files [20]. The steady state is reached by simulating the plant for 100 days with constant influent, defined by flow-weighted dry weather data file. The dynamic regime was simulated for 28 days with each influent file using as starting point the steady state solution.

6.2.2 Operational Costs Function Development

The operational costs were calculated with the following formula:

$$OC = \gamma(AE + ME) + EF \tag{6.1}$$

where: AE is the aeration energy [kWh/day]; ME—mixing energy [kWh/day]; EF—effluent fines; γ—electricity price 0.1 [€/kWh];

Because the external recycle flow rate (Qr), waste flow rate (Qw) and internal flow rate (Qa) were set to constant values during all the simulations, the pumping energy costs are excluded from the formula.

The average aeration energy costs were calculated with the equation proposed in [20]:

Table 6.1 Parameters used for the effluent fines

Effluent variable	$\Delta\alpha_j$(€·kg^{-1})	$\Delta\beta_j$ (€·kg^{-1})	$\Delta\beta_0$(€·m^{-3})	$C_{L,j}$ (mg·L^{-1})
Ammonia	4.00	12.00	2.70×10^{-3}	4.00
Total nitrogen	2.70	8.10	1.40×10^{-3}	18.00

$$AE = \frac{24}{T} \int_{t=22d}^{t=28d} \sum_{i=1}^{i=5} \left[0.4032 \cdot K_{Lai}(t)^2 + 7.8408 \cdot K_{Lai}(t) \right] dt \tag{6.2}$$

where: $K_L ai(t)$ is the mass transfer coefficient in the ith aerated reactor at time t [h^{-1}] and T = 7 days.

The mixing energy is a function of the compartment volume and it was calculated with the equation suggested in [24]:

$$ME = \frac{24}{T} \int_{t=22d}^{t=28d} \sum_{i=1}^{i=5} \begin{bmatrix} 0.005 \cdot V_i & if & K_{Lai}(t) < 20d^{-1} \\ 0 & otherwise & \end{bmatrix} dt \tag{6.3}$$

where: V_i is the reactor volume [m^3].

The effluent fines [25] were calculated by comparing the total nitrogen and ammonia in the effluent with their maximum allowed discharge limits. A mathematical description of the cost function used for the effluent fines is presented in Eq. (6.4):

$$Cost(t) = \begin{cases} \Delta\alpha_j \cdot C_{ef.j} \cdot Q_{ef} & if \quad C_{ef.j} \leq C_{Lj} \\ \Delta\alpha_j \cdot C_{Lj} \cdot Q_{ef} + \beta_{0j} \cdot Q_{ef} + \Delta\beta_j \cdot \left(C_{ef.j} - C_{Lj} \right) \cdot Q_{ef} & if \quad C_{ef.j} > C_{Lj} \end{cases} \tag{6.4}$$

The total nitrogen concentration was calculated with Eq. (6.5) and as a result, it can be noted that ammonia is penalized twice.

$$\begin{aligned} N_{tot.ASM3} = {} & S_{NH.ASM3} + S_{NO} + i_{N.SI} S_I + i_{N.SS} S_S + \\ & + i_{N.XI} X_I + i_{N.XS} X_S + i_{N.BM} (X_H + X_A) \end{aligned} \tag{6.5}$$

Where the variables denote: $i_{N.SI}$ nitrogen content of SI; $i_{N.SS}$ nitrogen content of S_S; $i_{N.XI}$ nitrogen content of X_I; $i_{N.XS}$ nitrogen content of X_S; $i_{N.BM}$ nitrogen content of X_H and X_A.

The ammonia and total nitrogen parameter values used in this research were obtained from [16]. The parameters used to compute the EF are presented in Table 6.1.

6.3 Control Approach

6.3.1 Description of the Applied Control Strategies

Two control approaches were investigated and optimized in this research study. The simulation results of the control strategies were compared to the open loop operation of the WWTP and between each other. The open loop approach is the simplest because it has no feedback loops. The exploitation of the WWTP plant is done by keeping the system variables (manipulated variables) constant at predefined values. Although rudimentary, this approach is still widely used in practice today. For the open loop control, constant K_La values of 240 day^{-1} were assumed for each aerated reactor, i.e. maximum aeration.

The first control architecture evaluated in this work has three control loops. These control loops have to keep the Dissolved Oxygen (DO) concentration in the aerated reactors at the predefined setpoints. The control is achieved by manipulating the air flow rate (indirectly, by the oxygen transfer coefficient K_{La}). This flow rate is constrained to a maximum of 240 day^{-1}. The control scheme is built of three PI controllers, one for each control loop. The PI controllers are tuned as suggested in [20] with a proportional gain of $K = 500$, integral time constant of $T_i = 0.001$ and anti-windup time constant set to $T_t = 0.0002$. This control architecture will be further referred to as 3DO.

The second control architecture is a cascade control scheme. On the outer level of the cascade control architecture, a multi input multi output (MIMO) Model Predictive Controller (MPC) adjusts the DO setpoint values for the aerated reactors. The outer control level is based on the MPC algorithm because it has the ability to provide different and predictive DO setpoints for each reactor. The controlled variable of this architecture is the nitrate (S_{NO3}) in the third aerated reactor. The inner control level consists of PI controllers that keep the DO concentration in the aerated reactors at the set-points imposed by the MPC. To prevent excessive aeration, the set-points provided by the MPC are constrained to a maximum of 2.5 mg/L. The sampling time of the MPC controller was set to $\Delta t = 1$ min. The prediction horizon and the control horizon have the values of $H_p = 200$ and $H_c = 3$. For the tuning of the PI controllers the same parameters were used, as the ones used for the first control strategy. This control scheme will be further referred to as NO5.

6.3.2 Optimization of the Applied Control Strategies

6.3.2.1 Setpoint Optimization

The first approach to optimize the control strategies was the setpoint optimization. The setpoints were optimized using a pattern search (PS) algorithm, so that the total operational costs of the WWTP were minimized as much as possible,

similarly to [19]. Only the last 7 days of the simulation were taken into consideration for performance assessment and optimization. The optimized setpoints were kept constant during the simulation.

6.3.2.2 Correlation Between Setpoint and Biodegradable Substrate

Because both control strategies control the kinetics of the nitrification process, the second approach to optimize the control strategies was the correlation of the influent ammonia nitrogen to the setpoints of each control strategy.

Nitrification is a two step biological process that converts ammonia nitrogen to nitrate nitrogen, in the presence of oxygen. The process is carried out by the nitrifying bacteria and is described by Eq. (6.6):

$$NH_4^+ + 2O_2 \overset{bacteria}{\longrightarrow} NO_3^- + 2H^+ + H_2O \tag{6.6}$$

From Eq. (6.6) it can be deduced that for high levels of ammonia in the system, a large amount of oxygen is required and vice versa. Also, the nitrate levels are directly influenced by the available ammonia in the system.

Figure 6.2 presents the variation of the inlet ammonia, DO in the aerated reactors and nitrate nitrogen in reactor 5, under open loop configuration of the WWTP. It can be observed that when the inlet ammonia has a low value (around 500 kg/day), the DO concentration in the aerated reactor drastically increases. This increase is due to the low activity of the nitrifying microorganisms which, in shortage of growth substrate, do not consume the oxygen in the system. This shows that when S_{NH} concentrations are low in the aerobic compartment, the oxygen uptake rate is smaller, and during these periods the aeration is excessive and energy is wasted.

The end product of the nitrification process, nitrate nitrogen, presents low values when ammonia nitrogen is low in the system. The NO5 control approach will try to correct this fact by imposing high setpoint values for the DO in the aerated reactor. This will lead to an excessive aeration, while the nitrate nitrogen values will still be low since it can not be produced because of the low ammonia concentration.

All these facts lead to the conclusion that when the S_{NH} is low in the aerated part of the plant, the DO requirements are low, and the control schemes carry out excessive aeration. Therefore the setpoint of the control strategies should be correlated with the available ammonia in the system.

Linear Interpolation Method

The first approach to correlate the substrate to the setpoint was the Linear Interpolation Method (LIM). For this approach the relationship between the quantity of ammonia nitrogen that enters the WWTP and the optimal setpoint for

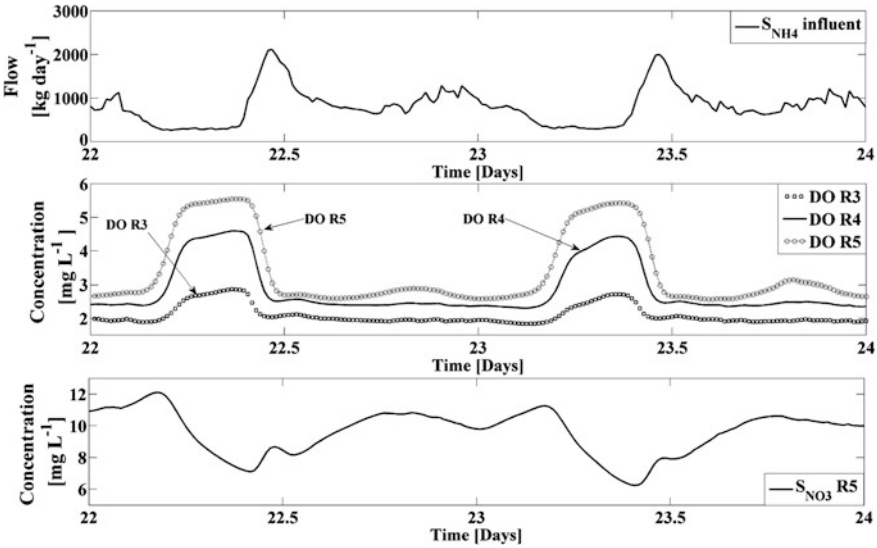


Fig. 6.2 Open loop dynamic variation of: **a** ammonia nitrogen quantity; **b** DO in the aerated reactors; **c** NO in reactor 5, between day 22 and 24 of simulation

any given control scheme was considered to be represent by a linear function. For both studied control strategies a maximum and a minimum setpoint value was attributed, therefore the setpoint changing strategy is built on Eq. (6.7):

$$OSP = SP_{\min} + \frac{(SP_{\max} - SP_{\min})(S_{NH.in} - S_{NH.\min})}{(S_{NH.\max} - S_{NH.\min})} \tag{6.7}$$

where SP_{min} and SP_{max} are the minimum and maximum setpoints [mg/L]; $S_{NH.min}$ and $S_{NH.max}$ are the minimum and maximum values of the mass flow of ammonia nitrogen [kg/day] that enters the WWTP; $S_{NH.in}$ the mass flow of ammonia that enters the WWTP.

In order to have the best results by using the LIM, proper values for the minimum and maximum quantities of influent ammonia nitrogen must be defined. If the difference between the minimum and maximum is large then the setpoint change will be slow and vice versa. The best values for the minimum and maximum inlet S_{NH} were determined by model based optimization for each control strategy.

For the 3DO control strategy the minimum setpoint was considered to be equal to 0.5 mg/L, because below this value the growth rate is minimal. The maximum allowable setpoint was 2 mg/L in order to avoid excessive aeration. For each reactor a different setpoint optimization function was determined. The optimization algorithm returned the following functions:

$$SP\_DOR3 = -0.7008 + 0.0026 SNH_{in} \tag{6.8}$$

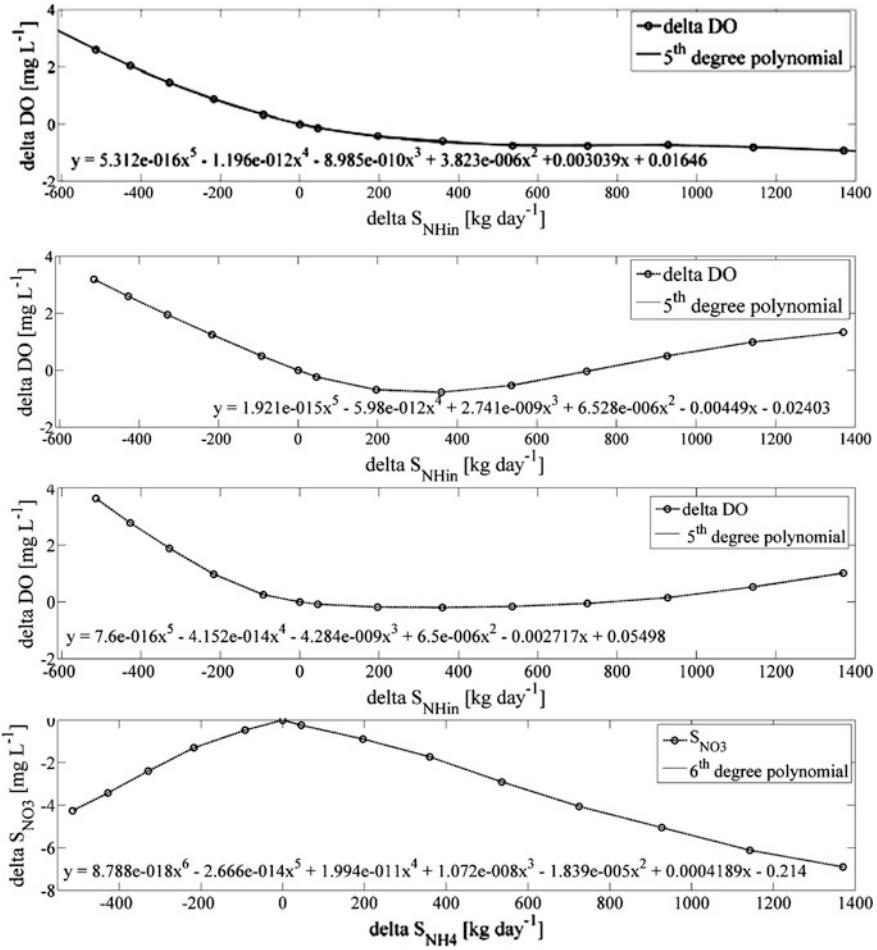


Fig. 6.3 The fitted polynomial functions from domain of the ammonia gradient to DO co-domain gradient, respectively to S_{NO3} gradient in the aerated reactors: **a** DO reactor 3; **b** DO reactor 4; **c** DO reactor 5; **d** S_{NO3} reactor 5

$$SP\_DOR4 = 0.0608 + 0.0022 SNH_{in} \tag{6.9}$$

$$SP\_DOR5 = -6.27 \cdot 10^{-4} + 0.0024 \cdot 10^{-3} SNH_{in} \tag{6.10}$$

For the NO5 control strategy the setpoint was considered to span between 5 and 12 mg/L. After the optimization procedure, the function which varies the setpoint became:

$$SP\_NOR5 = 5.1429 + 0.0043 SNH_{in} \tag{6.11}$$

Table 6.2 Results of the control approaches and open loop simulations for all influent conditions

Influent	Control strategy	AE €/day	ME €/day	SNH €/day	NO_{tot} €/day	EF €/day	OC €/day
Dry	OL	854.84	0	283.94	686.78	970.72	1825.56
	3DO	633.65	0	363.40	632.83	996.23	1629.88
	3DO LIM	610.98	0.16	331.79	600.18	931.97	1543.12
	3DO PIM	613.24	0	348.94	605.72	954.66	1567.91
	NO5	708.61	0.92	313.11	637.65	950.76	1660.31
	NO5 LIM	705.95	3.18	328.30	611.89	940.20	1649.34
	NO5 PIM	720.12	1.09	311.24	642.57	953.81	1675.03
Rain	OL	854.84	0.00	456.66	796.63	1253.30	2108.14
	3DO	592.11	0.00	632.81	761.74	1394.55	1986.67
	3DO LIM	568.46	0.14	622.11	739.37	1361.48	1930.08
	3DO PIM	570.36	0.00	636.37	741.70	1378.07	1948.43
	NO5	690.97	0.48	514.06	759.60	1273.66	1965.11
	NO5 LIM	682.93	2.52	526.76	740.16	1266.92	1952.37
	NO5 PIM	702.05	0.57	512.91	762.97	1275.88	1978.50
Storm	OL	854.84	0.00	433.80	750.78	1184.58	2039.42
	3DO	626.42	0.00	558.68	709.63	1268.31	1894.73
	3DO LIM	604.32	0.14	534.83	684.45	1219.28	1823.74
	3DO PIM	603.18	0.00	565.87	689.28	1255.15	1858.33
	NO5	724.83	0.29	467.63	712.68	1180.31	1905.42
	NO5 LIM	710.69	2.64	486.08	691.56	1177.65	1890.98
	NO5 PIM	721.99	1.00	477.21	714.88	1192.09	1915.08

Polynomial Interpolation Method

The second approach to linking the setpoints to the biodegradable substrate was the Polynomial Interpolation Method (PIM).

The polynomial relationship between the influent ammonia and the control variables was established using data generated by simulation of the plant with constant influent for 100 days. Altogether, fourteen simulations with different values for the influent ammonia, varying from 200 to 1,600 kg/day, were made. Because the flow rate has a great impact on the nitrifying microorganisms, the ammonia quantity was increased by raising the flow and the ammonia concentration in the same time. Only the ammonia concentration was modified from one simulation to the other, while the rest of the components remained constant during all the simulations.

The steady state solution was considered to be a threshold for the generated data. The threshold for the inlet ammonia was equal to the steady state inlet, and had a value of 755.36 kg/day. The threshold value for each variable was subtracted from the generated data and was plotted against the inlet ammonia variation (Fig. 6.3). The relationship between the setpoints and influent ammonia was established by fitting a polynomial function to each of these plots. The final equation for the setpoint determination had the form:

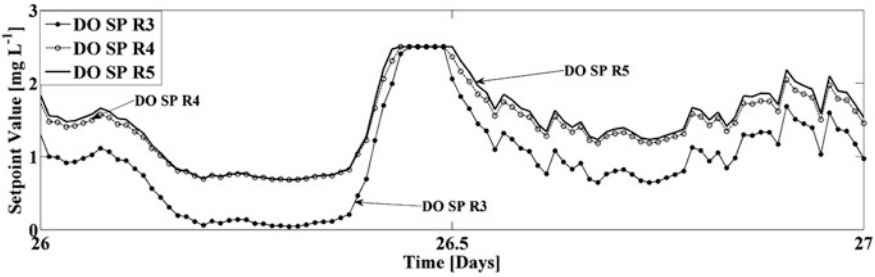


Fig. 6.4 Variation of the DO setpoint values of the 3DO LIM control scheme, between day 26 and 27 of dynamic simulation

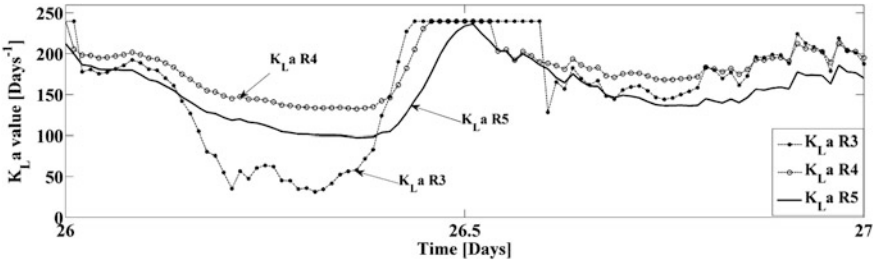


Fig. 6.5 Variation of K_L a values of the 3DO LIM control scheme in the aerated reactors, between day 26 and 27 of dynamic simulation

$$SP = SP_{base} - SP_{poly} \tag{6.12}$$

where SP_{base} is the root setpoint and SP_{poly} is generated by the polynomial function.

In order to achieve the optimal performance, SP_{base} was optimized for each controller. The results returned by the optimization algorithm was 2 mg/L for the DO controllers and 7.5 mg/L for the NO5 control strategy.

6.4 Results and Discussions

Table 6.2 presents the results for the 7 days of simulation for all control strategies, including the open loop simulations.

The setpoint optimization algorithm of the 3DO control strategy returned the following optimal setpoint values: 1.67 mg/L for the first aerated reactor and 1.86 mg/L for the second and third aerated reactor. With these setpoint values, in case of the dry weather influent file, the 3DO control returned 1,629.88 €/d as total operational costs, 195.68 €/d (10.72 %) lower than the open loop approach.

The reduction of the operational costs is due to the decreased aeration, the effluent fines being only 1 % higher than in case of the open loop simulation.

Further improvement in the operational costs is achieved with the 3DO PIM control scheme. The operational costs are lower with 257.65 €/d (14.11 %) compared to the open loop simulation, and with 86.76 €/d compared to the simple 3DO approach.

The best results for the 3DO control was returned by the LIM approach with a mean value for the total operational costs of 1,543.12 €/d, i.e. a cost cutback of 15.47 % compared to the open loop results.

For the simulations with the rain and storm influent files the 3DO control scheme presents significant improvement in the operational costs compared with the open loop. The LIM optimization approach proved to be the best way to optimize the operational costs while maintaining a good effluent quality.

Figures 6.4 and 6.5 present the variation of the manipulated variable for the 3DO LIM control scheme. It can be observed that there are no rapid changes in the manipulated variables and therefore the exploitation of the aeration unit will be minimal.

The NO5 control strategy presents reductions of the operational costs as well. The simulations based on the optimal setpoint (9.88 mg/L) resulted a mean value of 1,660.31 €/d, 9 % lower than the open loop simulation.

The best results in case of the NO5 control strategy were obtained with the LIM optimization approach. For this simulation the total operational costs were reduced with almost 10 % compared with the open loop.

6.5 Conclusions

This work proposed two methods for optimizing the setpoints of two control strategies of the WWTP. The first method was the model based setpoint optimization which managed to improve the operational costs with 60,000–71,000 €/year, depending on the control strategy.

The second approach to improve the operational costs was the association of the setpoint with the growth substrate. This was accomplished with two methods: Linear and Polynomial Interpolation. The LIM proved to achieve a better performance returning cost cutbacks of 103,000 €/year in case of the 3DO control scheme and showing promising incentives for its implementation.

Acknowledgments The authors wish to thank for the financial support provided from programs co-financed by the Sectoral Operational Program for Human Resources Development 2007–2013, Contract no.: POSDRU/88/1.5/S/60185–"Innovative doctoral studies in a Knowledge Based Society.

References

1. Henze M, Gujer W, Mino T, van Loosdrecht MCM (2000) Activated sludge models ASM1, ASM2, ASM2d, and ASM3, IWA scientific and technical report no. 9. IWA Publishing, London, UK
2. Henze M, Grady CPL Jr, Gujer W, Marais GVR, Matsuo T (1987) Activated sludge model No. 1, IAWQ scientific and technical report no. 1, London, UK
3. Henze M, Gujer W, Mino T, Matsuo T, Wentzel MC, Marais GvR (1995) Activated sludge model No. 2. IAWQ scientific and technical report no. 3. London, IAWQ
4. Henze M, Gujer W, Mino T, Matsuo T, Wentzel MC, Marais GvR, van Loosdrecht MCM (1999) Activated sludge model no. 2d, ASM2d. Water Sci Technol 39(1):165–182
5. Gujer W, Henze M, Mino T, van Loosdrecht MCM (1999) Activated sludge model no. 3. Water Sci Technol 39(1):183–193
6. Ostace GS, Cristea VM, Agachi PS (2011a) Cost reduction of the wastewater treatment plant operation by MPC based on modified ASM1 with two-step nitrification/denitrification model. Comput Chem Eng 35:2469–2479
7. Ostace GS, Gal A, Cristea VM, Agachi PS (2011) Operational costs reduction for the WWTP by means of substrate to dissolved oxygen correlation—a simulation study. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2011, WCECS 2011, 19–21 Oct 2011, San Francisco, USA, pp 945–950
8. Giusti E, Marsili-Libelli S, Spagni A (2011) Modelling microbial population dynamics in nitritation processes. Environ Model Softw 26:938–949
9. Iacopozzi I, Innocenti V, Marsili-Libelli S, Giusti E (2007) A modified activated sludge model no. 3 (ASM3) with two-step nitrification/denitrification. Environ Model Softw 22:847–861
10. Marsili-Libelli S, Ratini P, Spagni A, Bortone G (2001) Implementation, study and calibration of a modified ASM2d for the simulation of SBR processes. Water Sci Technol 43:69–76
11. Sin G, Vanrolleghem PA (2006) Evolution of an ASM2d-like model structure due to operational changes of an SBR process. Water Sci Technol 53:237–245
12. Kaelin D, Manser R, Rieger L, Eugster J, Rottermann K, Siegrist H (2009) Extension of ASM3 for two-step nitrification and denitrification and its calibration and validation with batch tests and pilot scale data. Water Res 43:1680–1692
13. Ni BJ, Yu HQ (2008) An approach for modeling two-step denitrification in activated sludge systems. Chem Eng Sci 63:1449–1459
14. Ossenbruggen PJ, Spanjers H, Klapwik A (1996) Assessment of a two-step nitrification model for activated sludge. Water Res 30:939–953
15. Karahan-Gül Ö, van Loosdrecht MCM, Orhon D (2003) Modification of activated sludge model no. 3 considering direct growth on primary substrate. Water Sci Technol 47(11):219–225
16. Stare A, Vrecko D, Hvala N, Strmcnik S (2007) Comparison of control strategies for nitrogen removal in an activated sludge process in terms of operating costs. Water Res 41:2004–2014
17. Benedetti L, de Baets B, Nopens I, Vanrolleghem PA (2010) Multi-criteria analysis of wastewater treatment plant design and control scenarios under uncertainty. Environ Model Softw 25:616–621
18. Cecil D, Kozlowska M (2010) Software sensors are a real alternative to true sensors. Environ Model Softw 25:622–625
19. Guerrero J, Guisasola A, Vilanova R, Baeza AJ (2011) Improving the performance of a WWTP control system by model-based setpoint Optimisation. Environ Model Softw 26:492–497
20. Copp JB (2002) The COST simulation benchmark: description and simulator manual. Office for Official Publications of the European Communities, Luxembourg

21. Takács I, Patry GC, Nolasco D (1991) A dynamic model of the clarification-thickening process. Water Res 25:1263–1271
22. Ostace GS, Cristea VM, Agachi PS (2010) Investigation of different control strategies for the BSM1 waste water treatment plant with reactive secondary settler model. In: 20th European symposium on computer aided process engineering, Ischia, pp 1841–1846
23. Gernaey KV, Jeppsson U, Batstone DJ, Ingildsen P (2006) Impact of reactive settler models on simulated WWTP performance. Water Sci Technol 53:159–167
24. Alex JL, Benedetti L, Copp JB, Gernaey KV, Jeppsson U, Nopens I, Pons MN, Rosen C, Steyer JP, Vanrolleghem P, Winkler S (2008) Benchmark simulation model no. 1 (BSM1), Technical report no. LUTEDX/(TEIE- 7229)/1-62/2008
25. Carstensen J (1994) Identification of wastewater processes. Ph.D. Thesis, Institute of Mathematical Modelling, Technical University of Denmark

Chapter 7
Periodic Oscillations on Angular Velocity with Maximum Brake Torque ABS Operation

Ivan Vazquez, Juan Jesus Ocampo and Andres Ferreyra

Abstract The appearance of oscillatory processes is inherent to the antilock braking system (ABS) operation, that can represent a problem on performance and comfort, that's why the oscillatory behavior represents an important study area, since in can lead to significant advances in ABS performance. In this paper we show that the ABS operation while the longitudinal contact force applied in a pneumatic system is near to the maximum value produces an oscillatory effect on the angular velocity of the vehicle′s wheel, and that for the time intervals that the system operates the oscillation can be considered periodic.

Keywords Antilock brake system · Contact force · Mathematical model · Periodic oscillation · Pneumatic brake system · Slip rate

7.1 Introduction

Security in modern automotive systems represents important criteria for design, for that reason, research in security systems has been increased in the last years, one of the concerns is brake systems, and more ABS, one of the problems to solve with ABS is the appearance of high frequency vibrations in the angular velocity of the

I. Vazquez (✉) · J. J. Ocampo · A. Ferreyra
Universidad Autonoma Metropolitana, Av San Pablo180, 02200 Mexico, D.F., Mexico
e-mail: iva@correo.azc.uam.mx

J. J. Ocampo
e-mail: jjoh@correo.azc.uam.mx

A. Ferreyra
e-mail: fra@correo.azc.uam.mx

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_7,

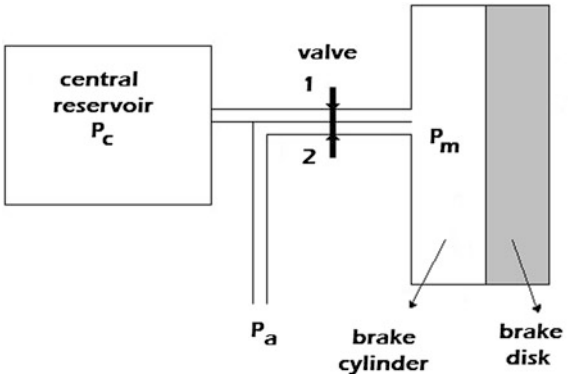


Fig. 7.1 Pneumatic brake model

wheel's rotation, which has been studied by Clover [1], Jansen [2], Kruchinin [3, 4], and Gozdek [5] among others. Modeling and research of forced oscillations in deformable wheel as a result of ABS activity has been discussed by Clover [1], and Jansen [2], while Kruchinin [3, 4] has analyzed the processes of appearance of vibrations during the pressure's relief phase in the brake cylinder of the ABS are analyzed, as well as the algorithms to suppress such vibrations. Gozdek [5] studied the possibility of longitudinal vibrations in the chassis of an airplane during the active phase of ABS is discussed.

The modern ABS systems very often use sliding modes control [6–11] with switching of ABS valves. Simultaneously the nonlinear character of ABS dynamics can lead to specific periodic regimes of angular velocity change for this manner of control algorithms that make programmed switch of the valve with a given period and duty cycle. The condition of existence of periodic changes in the angular velocity of the wheel's rotation due to the presence of specific ABS regimes is discussed in this paper.

The model of a pneumatic brake system is under consideration. The specific configuration of this system includes the next: brake disks, which hold the wheels, as a result of the increment of the air pressure in the brake cylinder (Fig. 7.1). The entrance of the air trough the pipes from the central reservoir and the expulsion from the brake cylinder to the atmosphere is regulated by a common valve. This valve allows only one pipe to be open, when 1 is open 2 is closed and vice versa. The time response of the valve is considered small, compared with the time constant of the pneumatic systems.

We study the case of wheel's rotation control, such that the longitudinal force, due to the contact of the wheel with the road, is near from the maximum value in the period of time valid for the model. This effect is reached as a result of the ABS valve's throttling.

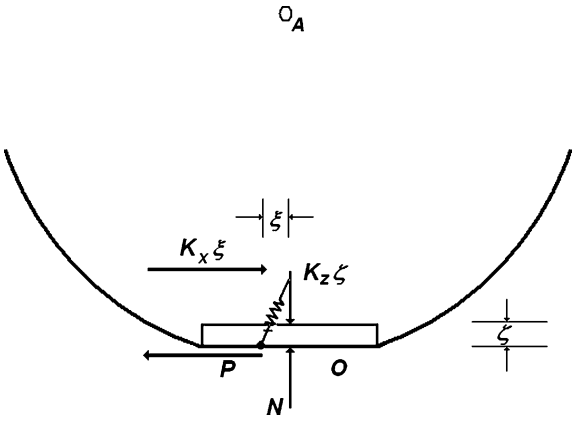


Fig. 7.2 Model for the contact element of the wheel

7.2 Mathematical Model

7.2.1 Wheel Motion Equations

To describe the wheel's motion we use a partial mathematical model of the dynamic system [3, 12]. Let's write the equation of the angular momentum change relative to the rotation axis (Fig. 7.2).

$$I_y \frac{d\Omega_y}{dT} = FR + L \tag{7.1}$$

where I_y—wheel's inertia moment, Ω_y—wheel's angular velocity, R—wheel's radius, F—contact force, L—brake torque.

The expression for longitudinal component of the contact force in the motion's plane according to experimental results [13] is equals

$$F = -vN\varphi(s) \tag{7.2}$$

v is the friction coefficient between the wheel and the road, N—normal reaction.

$$s = \frac{V_x + \Omega_y R + \frac{d}{dT}\breve{\xi}}{V_x} \tag{7.3}$$

s—slip rate, V_x—longitudinal velocity of the wheel mass center, $\breve{\xi}$—longitudinal deformation of the tire's contact area element. The function $\varphi(s)$ is defined experimentally, and it looks like Fig. 7.3.

The motion equation of the contact element with mass M_c is described by the tire longitudinal deformation. The interaction between this element and the rigid part of the wheel can be described with a viscoelastic forces model. The movement equation for the contact element is the next

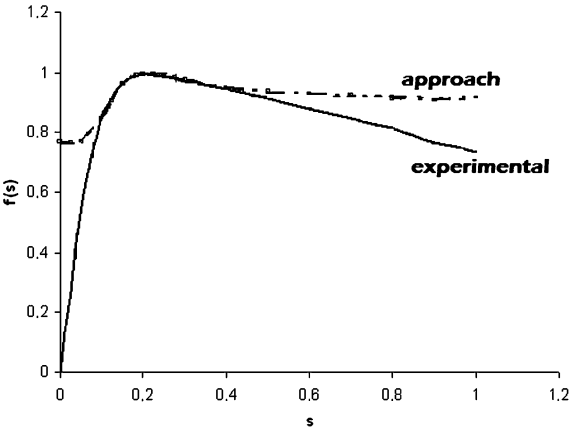


Fig. 7.3 Characteristic function for slip rate

$$M_c \frac{d}{dT}\left(V_x + \Omega_y R + \frac{d}{dT}\breve{\xi}\right) = F - C_x \frac{d}{dT}\breve{\xi} - K_x \breve{\xi} \tag{7.4}$$

Here C_x and K_x are longitudinal constants of viscous and elastic behavior of tire's model. The model to be used is the similar to description of first waveform in model [2].

The Eqs. (7.1)–(7.4) characterize wheel motion. This system is closed if we assume longitudinal velocity V_x and normal reaction N as constants. This approximation is correct for time lag about seconds if longitudinal velocity and normal reaction changes slowly and their variations are small [14].

Model proposed was previously used to describe the wheel's vibration for small values of slip ratio $s < 0.1$ when dependence $\varphi(s)$ is approximately linear $\varphi(s) = K_0 s$ [3]. Under these conditions, it is possible to consider that natural period of contact element vibrations in (7.4) is much smaller than the characteristic time of change of angular velocity and break torque. The fractional analysis method [14] can be used to reduce Eq. (7.4) to terminal form and write approximated relation $F = K_x \xi$. The wheel motion equations in this case is equivalent to pendulum equation [2, 3] with viscous friction. Natural frequency of this pendulum is

$$\omega_n = \sqrt{\frac{K_x R^2}{I_y} - \frac{K_x^2 V_x^2}{4v^2 N^2 K_0^2}}$$

Such as been shown [3], this result is consistent with experimental effects detected in the process of ABS control algorithm tests.

Further we consider the behavior of the system around the maximum value of the brake torque, it means in the region of $\varphi(s)$ maximum. The Tikhonov's theorem [14] condition used for reduction in previous paragraph is correct too, but reduced equations has singularities for $\varphi'(s) = 0$. The analytic and numerically solution of this equations is difficult. Therefore it is necessary to study full system

(7.1)–(7.4) properties in order to analyze periodic oscillation of the angular velocity.

We use the next approximation for $\varphi(s)$

$$\varphi(s) = \frac{a_1 s^2 + a_2 s + a_3}{s^2 + a_4 s + a_5} \tag{7.5}$$

The parameters $a_1 \ldots a_5$ were calculated with the least squares method [15]. We use for calculation the values:

$a_1 = 0.8886$
$a_2 = -0.1776$
$a_3 = 0.0155$
$a_4 = -0.2226$
$a_5 = 0.0201$

These values approximates top neighborhood of tire characteristics, used by Mogamedov [10].

7.2.2 Pneumatic Brake System Equations

We suppose that the brake torque L is proportional to the pressure P_m in the brake cylinder.

$$L = K_L P_m \tag{7.6}$$

For the brake system we use an approximated model of pressure changes in the brake cylinder due to the opening of the valve with a first order relation [1, 16].

$$T_e \frac{dP_m}{dT} + P_m = P_* \tag{7.7}$$

Let's suppose opening and closing of valve is momentary and the parameters of the Eq. (7.7) are given by the next rules:

(a) $P_* = P_c = \text{const}$ $T_e = T_{\text{in}}$ when 1 is opened and 2 is closed
(b) $P_* = P_a = 0$ $T_e = T_{\text{out}}$ when 2 is opened and 1 is closed

Here P_c—pressure inside the central reservoir, P_a—atmospheric pressure, that we'll consider 0. T_{in} and T_{out}—time constants of internal and external pipelines.

7.3 Dimensionless Equations

We desire to rewrite Eqs. (7.1)–(7.3), (7.5) in a more useful form, by ignoring changes in V_x. Taking $\frac{d\Omega_y}{dT}$ from (7.1), and writing in (7.5) we have:

$$\begin{cases} I_y \frac{d\Omega_y}{dT} = vNR\varphi(s) - L \\ M_c \frac{d^2\breve{\xi}}{dT^2} + C_x \frac{d\breve{\xi}}{dT} + K_x \breve{\xi} = -\frac{M_c R}{I_y} L + \left(\frac{M_c R^2}{I_y} - 1\right) vN\varphi(s) \\ s = 1 + \Omega \frac{R}{V_x} + \frac{1}{V_x} \frac{d\breve{\xi}}{dT} \end{cases} \tag{7.8}$$

Equation (7.7) can be modified to following form:

$$T_e \frac{dL}{dT} - K_L P_* + L = 0 \tag{7.9}$$

To reduce the number of parameters we take the variables to a dimensionless form

$$l = \frac{L}{NR}, \ \omega = \frac{\Omega_y R}{V_x}, \ \xi = \frac{\breve{\xi}}{V_x T_1}, \ t = \frac{T}{T_1}$$

where

$$T_1 = \frac{I_y V_x}{NR^2}$$

is the characteristic time of the angular velocity changes, according to (7.1).

The system (7.1), (7.8), (7.9) has the next dimensionless form

$$\begin{cases} \frac{d\omega}{dt} = l - v\varphi(s) & s = 1 + \omega + \frac{d\xi}{dt} \\ \frac{d^2\xi}{dt^2} + q\frac{d\xi}{dt} + p\xi = -l - vk\varphi(s) & \\ \frac{T_e}{T_1}\frac{dl}{dt} = l_s - l & (a) l_s = l_c = \text{const} \quad T_e = T_{\text{in}} \\ & (b) l_s = 0 \qquad\quad T_e = T_{\text{out}} \end{cases} \tag{7.10}$$

where

$$q = \frac{C_x T_1}{M_c}, \quad p = \frac{K_x T_1^2}{M_c}, \quad k = \frac{I_y}{M_c R^2} - 1 \quad l_c = \frac{K_L P_c}{L_*}.$$

7.4 Periodic Solutions Finding

The main goal of this work is the study of periodic regimes produced by programmed switching of the valve with a given period and duty cycle [16].

To search for periodic regimes we analyze an auxiliary task: control with a relay feedback built such that the system switches the valve when the slip ratio s reaches the arbitrary limit values s_1 and s_2. We analyze the values s_1, s_2 for which the function $\varphi(s)$ changes around the maximum value (Fig. 7.3). In this region the contact force has a value less or equal than 10 % down the maximum value, for a constant normal reaction between the wheel and the road.

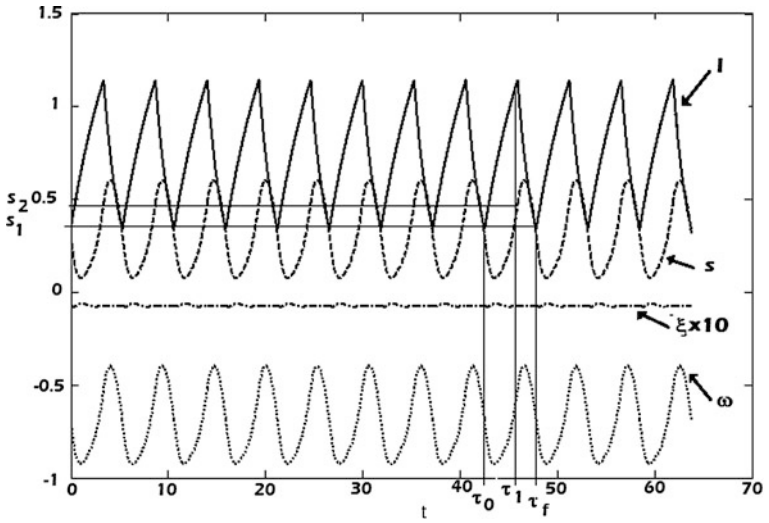


Fig. 7.4 Periodic solution

To find periodic solutions $[l_P, \xi_P, \omega_P]$ we integrate numerically the equation system (7.10) for initial conditions that can be present in real systems [3]. As a result of this integration we have solutions for which the values (a) work in the interval $\Delta_1 = \tau_1 - \tau_0$, and the values (b) in the interval $\Delta_{fr} = \tau_f - \tau_1$ (Fig. 7.4).

We consider that a periodic regime was found if the integration if the next criteria is true

$$\max\left(l_f - l_{P_0}, \xi_f - \xi_{P_0}, \frac{d\xi_f}{dt} - \frac{d\xi_{P_0}}{dt}, \omega_f - \omega_{P_0}\right) \leq 0.01$$

Here $\left(l_{P_0}, \xi_{P_0}, \omega_{P_0}\right)$ and $\left(l_f, \xi_f, \omega_f\right)$ are the variables in two successive periods at the moment of valve's opening. $\left(l_{P_0}, \xi_{P_0}, \omega_{P_0}\right)$— are the initial conditions of computed periodic solution.

All the possible values Δ_1, Δ_f and the corresponding initial conditions of the periodic solutions at the opening moment were obtained by solving the system for different pairs (s_1, s_2) inside the interval $(s_{1\,\min}, s_{2\,\max})$. The region of founded values Δ_1, Δ_f for different friction coefficient value v can be seen in Fig. 7.5.

The parameters for calculations are

$T_{in} = 0.0043$ s
$T_{out} = 0.0085$ s
$p = 1000$
$q = 100$

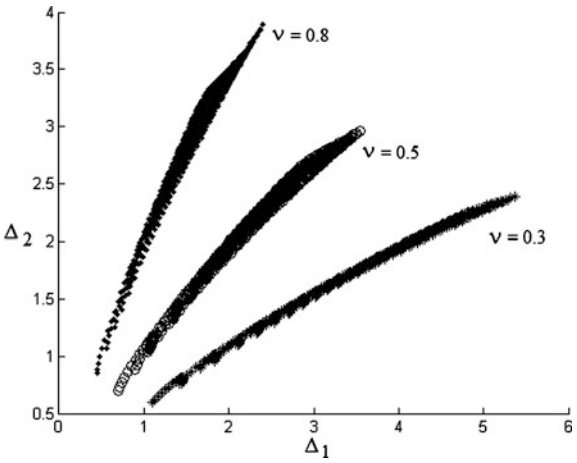


Fig. 7.5 Regions founded for different friction coefficients

$k = 10$
$l_s = 0.4755$
$T_1 = 0.0848$ s

7.5 Conclusion

ABS has become standard equipment in most of the modern vehicles since they can provide a good control response in direction during extreme braking situations. ABS operation is based on a switching process, oscillatory affects are produced, and the results can have consequences on performance, security and comfort of the vehicle, for that reason it is important to analyze the properties of such oscillations. The case of maximum longitudinal force before the wheel locks was considered because operation of ABS starts when this condition occurs. The simulation showed that the oscillations on the angular velocity of the wheel have a periodic behavior for some regions of the analysis, that information can be helpful to design control algorithms, either, to suppress vibrations or to take advantage of the periodic oscillation on the switching process, to increase performance.

References

1. Clover CL, Bernard JE (1998) Longitudinal tire dynamics. Veh Syst Dyn 29(1):231–259
2. Jansen ST, Zegelaar PW, Pacejka HB (1999) The influence of in-plane tyre dynamics on ABS braking of a quarter vehicle model. Veh Syst Dyn 32(2):249–261
3. Kruchinin PA, Magomedov M, Novozhilov IV (2001) Mathematical model of an automobile wheel for antilock modes of motion. Mech Solids 36(6):52–57

4. Kruchinin PA, Magomedov M, Makarov LM (2003) Active suppression of parasitic oscillations while the operation of a pneumatic antiblock system. Mosc Univ Mech Bull 5:25–29
5. Gozdek VS, Goncharenko VI (2004) About verification of ACS stability in the brake wheel process. Promislova gidravlika I Pnevmatika 3(6):73–76
6. Drakunov SV, Ozguner U, Dix P, Ashrafi B (1995) ABS Control Using Optimum Search via Sliding Modes. IEEE Trans Control Syst Technol 3(1):79–85
7. Ming-Chin Wu, Shih Ming-Chang (2003) Simulated and experimental study of hydraulic anti-lock braking system using sliding-mode PWM control. Mechatronics 13:331–351
8. Unsal C, Pushkin K (1999) Sliding Mode Measurement Feedback Control for Antilock Braking Systems. IEEE Trans Control Syst Technol 7(2):271–281
9. Utkin V, Guldner J, Shi J (2009) Sliding mode control in electromechanical systems. CRC Press, Boca Raton
10. Magomedov M, Alexandrov VV, Pupkov KA (2001) Robust adaptive stabilization of moving a car under braking with ABS in control circuit, SAE Technical Paper (01)
11. El Hadri A, Cadiou JC, M'Sirdi NK (2002) Adaptive sliding mode control for vehicle traction. In: Proceedings of the IFAC World Congress, Barcelona, Spain
12. Novozhilov IV, Kruchinin PA, Magomedov M (2000) Contact force relation between the wheel and the contact surface, Collection of scientific and methodic papers, vol 23(1). Teoreticheskaya mekhanika, MSU, pp 86–95
13. Pacejka HB (1981) In-plane and out-of-plane dynamics of pneumatic tyres. Veh Syst Dyn 8(4–5):221–251
14. Novozhilov IV (1997) Fractional analysis. Methods of motion decomposition. Birkhauser, Boston
15. Lawson C, Henson R (1974) Solving least squares problems. Prentice Hall Inc., Englewood Cliffs
16. Vazquez I, Ocampo JJ, Ferreyra A (2011) Periodic oscillations on angular velocity due to the ABS operation under specific work regimes. Lecture notes in engineering and computer sciences. In: Proceedings of the world congress on engineering and Computer Science 2011, WCECS 2011, San Francisco, USA, pp 997–1000, 19–21 October 2011

Chapter 8
The Computer Simulation of Electrochemical Shaping Processes

Jerzy Kozak

Abstract Electrochemical machining (ECM) is an important manufacture technology in machining difficult-to-cut materials and to shape complicated contours and profiles with high material removal rate without tool wear and without inducing residual stress. This paper presents the physical and mathematical models on the basis of which of the simulation process module in the computer-aided engineering system (CAE–ECM) for ECM has been developed. The results of computer simulation of electrochemical sinking and examples of CAE–ECM system application are discussed.

Keywords Dissolution · Electrochemical · ECM · Machining · Modeling · Simulation · Sinking

8.1 Introduction

Electrochemical machining (ECM) is an important manufacture technology in machining difficult-to-cut materials and to shape complicated contours and profiles with high material removal rate without tool wear and without inducing residual stress.

As shown in Fig. 8.1 machining based on controlled anodic electrochemical dissolution process in which the workpiece is the anode and the tool is the cathode of an electrolytic cell. In the ECM process, a low voltage (8–30 V) is normally applied between electrodes with a small gap size (usually 0.2–0.8 mm) producing

J. Kozak (✉)
Institute of Aviation, Al.Krakowska 110/114, 02-256 Warsaw, Poland
e-mail: jkozak64@wp.pl

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_8,

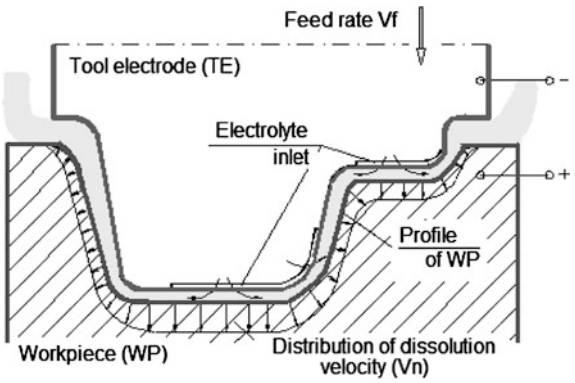


Fig. 8.1 Schematic diagram of ECM sinking

a high current density of the order of (10–100 A/cm^2), and a metal removing rate ranging from an order 0.1 mm/min, to 10 mm/min. Electrolyte (typically NaCl or $NaNO_3$ aqueous solutions) is supplied to flow through the gap with a velocity of 10–50 m/s to maintain the electrochemical dissolution with high rate and to flush away the reactions products (usually gases and hydroxides) and heat generated caused by the passage of current and electrochemical reactions.

As electrochemical dissolution proceeds, the tool electrode–cathode can be fed mechanically towards the workpiece−anode in order to maintain the machining action. Under these conditions, the inter-electrode gap width gradually tends to a steady-state value, and a shape, complementary to that of the cathode-tool, is reproduced approximately on the anode-workpiece. Being a non-mechanical metal removal process, ECM is capable of machining any electrically-conductive material with high stock removal rates regardless of their mechanical properties, such as hardness, elasticity and brittleness. It has been applied in diverse industries such as aerospace, automotive and electronics to manufacture airfoils and turbine blades, die and mold, artillery projectiles, surgical implants and prostheses, etc. [1–5].

The main objective of ECM is to achieve the required shape of workpiece within a given tolerance on the shape and dimensions. The tasks relating to this purpose can be reduced directly (or indirectly) to a problem of searching for a boundary of the area within which the machining, i.e., to a value boundary problems (moving boundary problem, free boundary problem or inverse boundary problem).

Depending on which the electrode surface is to be determined all tasks can be divided into two groups [1–9]:

(1) tasks in which for a known shape of the tool electrode and known condition of machining, the evolution of a shape of workpiece surface has to be determined,
(2) tasks in which the tool electrode shape is searched for, which ensures obtaining the required shape of the workpiece.

The first category of the problems is encountered in the analysis of ECM accuracy. The tasks from the second category mainly deal with the tool electrode design and most frequently are encountered in practice.

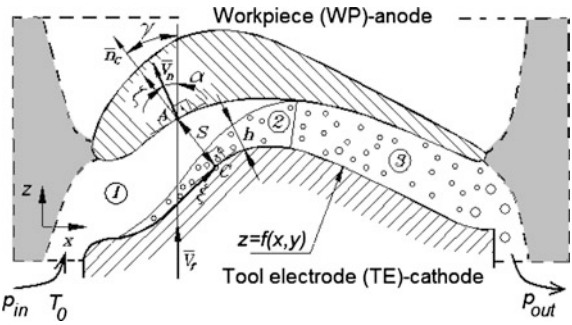


Fig. 8.2 Schematic diagram for mathematical model of ECM process

Industrial practices in ECM have revealed some problems impeding its further development and wider acceptance by industrial users. Among them, prediction and control of the local gap width distribution (and hence, the control of dimensional accuracy), along with the design of tool electrodes for complex workpiece shapes and optimization of process, are the major problems encountered by ECM users.

The optimization and tool-electrode design is carried out using ECM process models. Software for the computer-aided ECM (CAE–ECM) has been developed in Warsaw University of Technology covers basic manufacturing problems ECM [3, 4, 8–10].

The paper presents the physical and mathematical models, basis of which a process simulation module has been developed the ECM sinking. The example of results of simulation by CAE–ECM system are discussed.

8.2 Mathematical Modeling of ECM Shaping Process

The main task of the electrochemical shaping, regardless of variant, is to calculate distribution of the material removed thickness on the anode-workpiece surface after a time step used in numerical calculations, which is determined by current density distribution in the gap, in particular in a medium of varying electrical conductivity, with complex processes occurring on the surfaces of the electrodes and with shape change of machined surface during course of machining. Since properties of electrolyte depend on temperature and gas phase concentration (mainly on concentration of generated during machining hydrogen), which distributions depend on velocity and pressure fields as well as on current density, ECM processes have to be described by set of mass, heat and electric charge transfer equations.

The problem of determining the changes shape of machining surface and physical conditions in the inter-electrode gap for transit and steady state of the ECM with using contoured cylindrical tool-electrode is consider in this paper (Fig. 8.2).

The electrode flow is from left to right in a thin gap of local size S and of length L. The down surface is the tool-electrode which moves upward with feed rate V_f. At a

point opposite the tool, the workpiece surface is moving upward with local velocity V_n. This distribution of the velocity of dissolution and the change of physical conditions along the flow path evoke non-uniform distribution of the gap size S and a shape error of the workpiece-anode.

The mathematical model of ECM process, referring to the formulated problems consists of sequence of mutual conjugated partial models which describe in the gap:

- distribution of the local gap size, $\boldsymbol{S}$,
- distribution of the flow parameters such as the static pressure p(x) and the velocity, $\boldsymbol{w}$,
- distribution of the temperature, $\boldsymbol{T}$,
- distribution of the void fraction, $\boldsymbol{\beta}$ or the thickness layer $\boldsymbol{h}$ with two phase flow (electrolyte and gas),
- distribution of the electrical conductivity, $\boldsymbol{\kappa}$,

The physical model with the following assumptions serves as the basis for mathematical modeling [8]:

1. The three regions can be identified in the gap: ***1***- the region with pure electrolyte, ***2***- the bubble region near cathode with thickness $\boldsymbol{h < S}$ and which consist of mixture of electrolyte and gas (hydrogen), where the void fraction of hydrogen in the bubble region layer is constant and equal $\boldsymbol{\beta^*}$, ***3***- the region of two phase flow witch change of $\boldsymbol{\beta}$,
2. The current density $\boldsymbol{i}$ is depends on medium conductivity in the gap and on the voltage $\boldsymbol{U}$ according to Ohm's law, which is extrapolated to the whole gap size. The electrochemical reaction will be accounted for by introducing the total overpotential $\boldsymbol{E = E_a - E_c}$, where $\boldsymbol{E_a}$ and $\boldsymbol{E_c}$ are the overpotential potential of anode and cathode, respectively,
3. The surface tension effect on the gas bubbles is neglected, and so is the bubble formation time.
4. Since analytical models of $\boldsymbol{K_v = K_v(i)}$ and $\boldsymbol{E = E(i)}$ are not available, experimental results are used with theoretical model,
5. The effect of bubble layers on pressure and the flow velocity of the electrolyte is accounted by consideration of a homogeneous two-phase model flow, where the electrolyte in the gap is treated as a uniformly-mixed pseudo-continuous medium of gas and liquid with local average void fraction:

$$\beta = \beta_* \frac{h}{S} \quad 0 < h < S \tag{8.1}$$

6. The electrical conduction of the two-phase medium can be determined by the Bruggeman equation:

$$\kappa = \kappa_o [1 + \alpha_T \theta](1 - \beta)^{3/2} \tag{8.2}$$

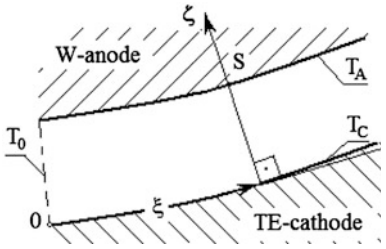


Fig. 8.3 Scheme of the curvilinear coordinate system

where: $\theta = T - T_o$, T_o = inlet electrolyte temperature, α_T = the temperature coefficient of the electrolyte conductivity at T_o, and κ_o = electrolyte conductivity at T_o and $\beta = 0$.

To formulate the mathematical model, a general case describing change in shape of the surface of the workpiece can be examined using coordinate system attached to the workpiece, which is immovable during machining (Fig. 8.3).

To simplify the calculations let us introduce a curvilinear coordinate system (ξ, ζ), connected with the tool-electrode in which a coordinate ξ lies on the given electrode and is measured from the inlet of the electrolyte and let axis ζ overlap its normal n_C (Figs. 8.2 and 8.3).

The surface of the workpiece at a given moment in time can be described by: z = Z(x, y,). According to electrochemical shaping theory, the evolution of the shape of the workpiece F(x, y, t), can be described as follows [3–5]:

$$\frac{\partial Z}{\partial t} = K_v(i_A) i_A \sqrt{1 + \left(\frac{\partial Z}{\partial x}\right)^2 + \left(\frac{\partial Z}{\partial y}\right)^2} \tag{8.3}$$

where: K_V is the coefficient electrochemical machinability, which is defined as the volume of material dissolved per unit electrical charge.

At the beginning of machining: t = 0, z = Z_0(x, y), where: Z_0(x, y) describes an initial shape of the workpiece surface.

To find the current density i_A on the surface of the workpiece, approximation using linearization of electric potential distribution along the segments of distance S between a given point of anode and given point on the TE has been applied [3, 4]. The current density can be obtained by Ohm's law with respect of change of conductivity across the gap:

$$i = \kappa_o \phi_{TG} \frac{U - E}{S} \tag{8.4}$$

where:

$$\phi_{TG} = \left[\frac{1}{S}\int_o^S \frac{d\zeta}{(1 + \alpha_T T)(1 - \beta)^{3/2}}\right]^{-1} \tag{8.5}$$

The equation of mass conservation for the hydrogen generation in the case of cylindrical electrodes and curvilinear coordinates can be obtained from mass balance as:

$$\rho_g \beta_* \int_o^h w(\zeta) d\zeta = \eta_H K_H \int_o^\xi i d\xi \tag{8.6}$$

where: $\rho_g = \frac{p}{RT}$ —specific gas density of hydrogen, R = gas constant for 1 kg of hydrogen, ηH = current efficiency of the hydrogen generation, KH = electrochemical equivalent of hydrogen, $\bar{w}(\xi)$—average velocity in the given cross section of the gap.

The heat transfer in the gap with respect to Joule's heat is described by:

$$w(\xi, \zeta) \cdot \frac{\partial T}{\partial \xi} = \frac{\partial}{\partial \zeta} \left[(a + a_T) \cdot \frac{\partial T}{\partial \zeta} \right] + \frac{i^2}{\rho \cdot C_p \cdot \kappa} \tag{8.7}$$

where: a—thermal diffusivity, a_T —turbulent thermal diffusivity by turbulence pulses (for laminar flow $a_T = 0$), ρ—specific medium density (in the region 1 $\rho = \rho_e$, and in the region 2 and 3 $\rho = \rho_e(1 - \beta)$where ρ_e—density of electrolyte and C_p—heat capacity of electrolyte.

The boundary conditions are as follows: $T(\xi = 0) = To$, $T(\xi, 0) = T_A$, $T(\xi, S) = T_C$, where: T_A and T_C are temperatures of the anode and cathode, respectively.

To complete of the systems of Eqs. (8.2)–(8.7), the formulation the pressure and the flow rate must be included. The continuity equation can be expressed by:

$$\bar{w}_o S = \int_o^S (1 - \beta) w d\zeta \tag{8.8}$$

where: $\bar{w}_o, S_o$—average velocity and gap size in inlet, respectively.

The momentum balance equation for the moving control is:

$$w \cdot \frac{d[\rho_e \cdot (1 - \beta) \cdot w \cdot S]}{d\xi} = -S \cdot \frac{dp}{d\xi} - \tau_a - \tau_c \tag{8.9}$$

where: τ_a and τ_c are the shear stresses on the surfaces of anode and cathode, which are assumed to be equal $(\tau_a = \tau_c)$. In general, the shear stress is expressed as follows:

$$\tau = \lambda \frac{\rho w^2}{8} \tag{8.10}$$

where $\lambda = C/\mathrm{Re}^m$, $\mathrm{Re} = \frac{2wS}{v}$ is Reynolds number (for laminar flow: $C = 96$ and $m = 1$; for turbulent: $C = 0.316$ and $m = 0.25$).

The boundary conditions for Eq. (8.9) is described by: $p(\xi = 0) = p_{in} - \varsigma_1 \frac{\rho \bar{w}_o^2}{2}$; $p(\xi = L) = p_{out} + \varsigma_2 \frac{\rho w^2}{2}$, where: ς_1, ς_2 is the hydraulic loss pressure in the inlet and the outlet, respectively.

The system of Eqs. (8.2)–(8.9) has been solved numerically using the finite difference method and iterative procedure.

In the first iteration one-dimensional approximation has been used for distribution of temperature and for determination of change of thickness the bubble layer $h^{(1)}$, next this distribution of $h^{(1)}$ is used in the second iteration and in the first approximation of two-dimensional (2-D) calculation of temperature distribution. Calculation of a given iterative cycle are finished, when the criteria of accuracy of calculation is satisfied and the simulation ECM process culminates in printout and plots of: the gap distribution and distributions of p, w, T, β.

After analysis of results of computer simulation of ECM at different operating parameters, the critical conditions can be determined from the point of view of the process limitations such as: boiling, choking flow and cavitation.

8.3 Computer Simulation System for Electrochemical Shaping

8.3.1 Structure of the CAE–ECM System

The developed CAE–ECM system is based on the conventional concepts of the expert system, in which the user interface, knowledge base, working data base and inference engine. Generally expert systems are classified as being of either of two types: the analytical type or the synthetic type. A diagnosis expert system is a typical example of analytical type, while a design system is an example of the synthetic type.

First type can be used for:

- selection of electrochemical methods of manufacturing leading to obtain required results (for example: deburring, sinking, EC-drilling etc.),
- simulation of different ECM processes for analyzing machining conditions and the results of machining,
- recognition and diagnosis the causes of the trouble in EC-manufacturing and suggest some remedies for them.
- Second type can be used for:
- planning and optimization of conditions of selected method of EC-manufacturing,
- tool electrode design for ECM,
- tooling design.

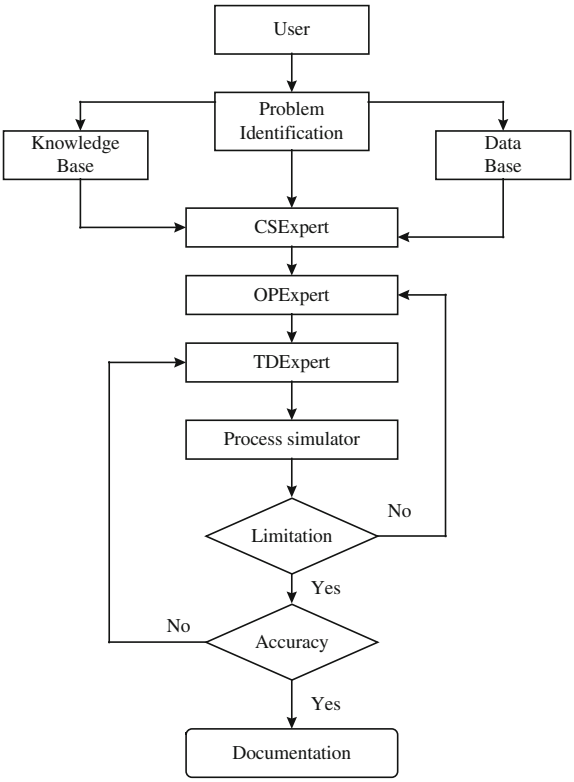


Fig. 8.4 Architecture of the CAE–ECM software

The presented CAE–ECM system can also serve as training and learning tool. Using the system the inexperienced engineers in EC-manufacturing or operators may get to know about the applicability of electrochemical technology for certain job, the selection of methods, the preparation of tooling and planning of operations, the analysis of some defects during machining and trouble-shooting etc.

The general structure of the CAE–ECM software is shown in Fig. 8.4.

The ***knowledge*** base is comprised of two forms of information:

(1) files which hold principal mathematical relations and characteristics of different methods of EC-manufacturing, (2) other data held as facts or in rule-based forms.

The working ***data base*** also holds two forms information. The first type of information relates to data about the electrolytes, materials of workpiece process, limitations of machine tool etc.

This type data is different for each methods of EC-manufacturing (for example: at ECM-shaping and finishing it is relation between the electrochemical machinability and current density; at electropolishing it is volt–ampere curves, etc.).

The second is the information which is inferred and computed by the system. For example, computed function for EC-machinibility after filling experimental data for "new" material of workpiece, which must be included to CAE–ECM

system. The machining condition and parameters for the simulation of process and tool design is provided by the ***conditions selection module CSExpert*** (Fig. 8.4).

The operating parameters are selected by ***OPExpert*** module, uses transferred information from ***Process Simulator*** module and tool design module ***TDExpert***.

The software of CAE–ECM System is based on multithread structure, which enables parallel simulation of many processes for different input data.

During simulation of the process, each of windows, that present graphically the change of the shape, is attended by separate process responsible for calculation and displaying data on the screen. In order to start the simulation, user must open windows with processes (maximum 10). Then program requires definitions of physical parameters of the process and the beginning shape of tool-electrode and workpiece-electrode or demanded shape of surface in case of designing workpiece-electrode. All parameters must be defined separately for each window. Program ECM enables designing of the beginning shape of both electrodes as a sequence of functions or by giving the points. In case the shape of electrodes is given by points—interpolation of splain functions is applied. The window for given electrodes contains options of saving designed shape of electrode or getting the shape from database.

Beside input data connected with physical parameters of modeling, program gives the possibility of establishing (also during simulation process) of the speed and preciseness of presented graphics connected with time step. In case of omitting such parameters, program uses default data. Buttons responsible for graphics presentation (start, pause, next step, previous step, that enable more precise analysis of shape evolution), are situated on the tool strip in the main window. After completed modeling, user can get the display of output data from hand menu of the process window.

Program simulating electrochemical sinking process is integrated with database, which gives possibility to save full set of physical parameters, shapes of electrodes and settings of presentation.

Results of simulation such as the distributions of gap size, current density, static pressure, void fraction, average temperature and temperature across and along the gap can be saved and displayed.

8.3.2 An Example of Results of Simulation of ECM Sinking

To illustrate application of CAE–ECM system simulation of ECM sinking operation for shaping, of airfoils are considered (Fig. 8.5).

For given anode (workpiece) profile, the tool electrode shape was designed using the CAE–ECM system

Examples of simulation of ECM shaping are shown in Fig. 8.6, where subsequent graphs illustrate anode-workpiece shape evolution in time.

The gap size distribution with neglecting changes of properties electrolyte i.e., under “ideal” conditions of ECM, is shown in Fig. 8.7.

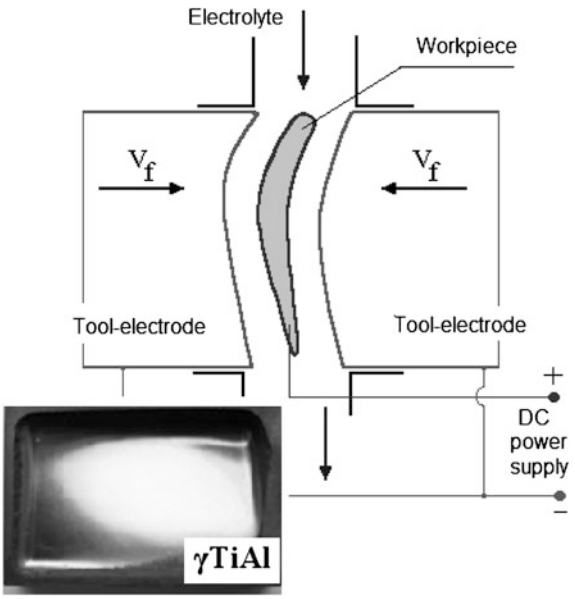


Fig. 8.5 Electrochemical shaping of the airfoil

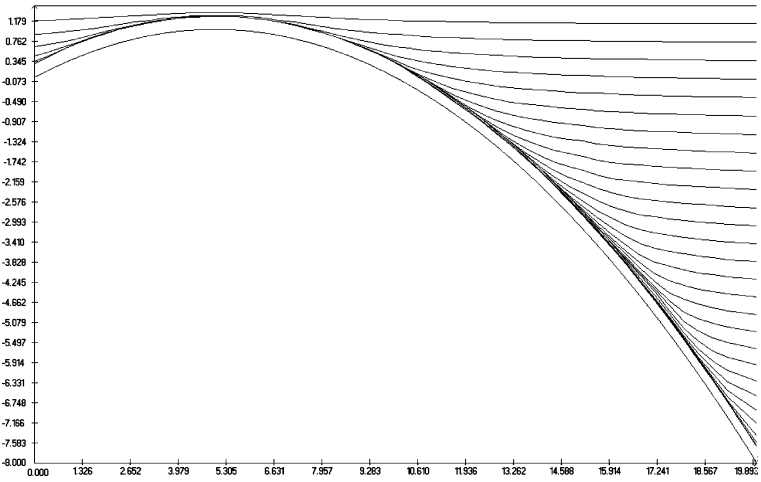


Fig. 8.6 Screen print-out containing the evolution of the shape of the workpiece during machining

In practice, ECM conditions are far to the ideal process of electrochemical shaping. In consequence, the material removal rate is distributed over the anode-workpiece surface in a different way in compare with the distribution in "ideal" process, and the machined part takes on a profile from that found from mathematical modeling and theoretical design. Calculation that is more exact needed simulation of all shaping process with regarding heat and mass transfer, as well as anodic dissolution characteristics.

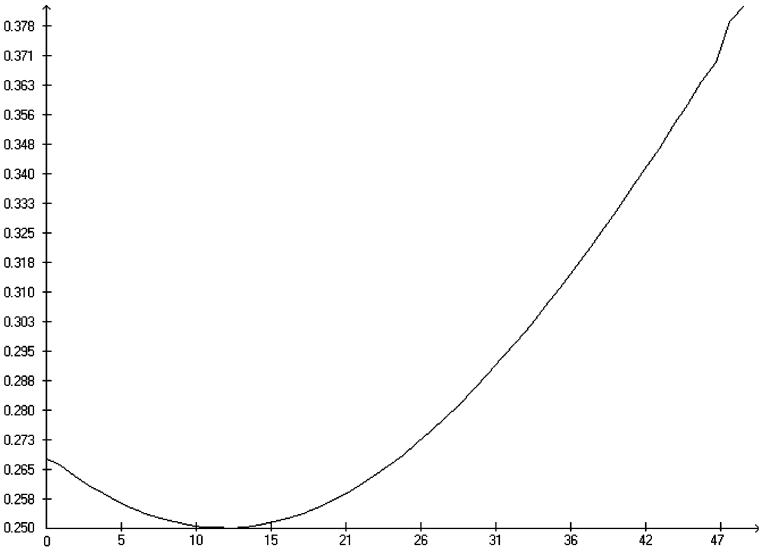


Fig. 8.7 Distribution the gap size in steady state ECM with neglecting influence of the heat and hydrogen generation (i.e., from the model of ideal ECM process)

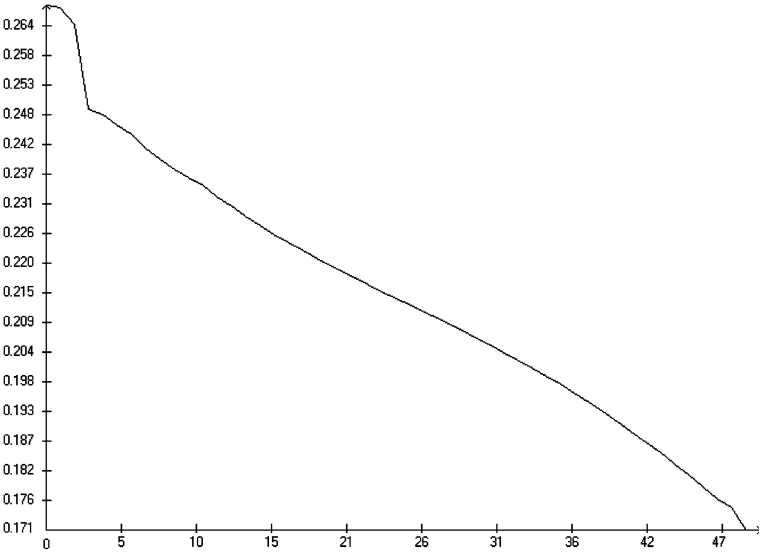


Fig. 8.8 Plot of distribution of the gap size in steady state of ECM with regarding changes of properties of the electrolyte due to heating and gas generation

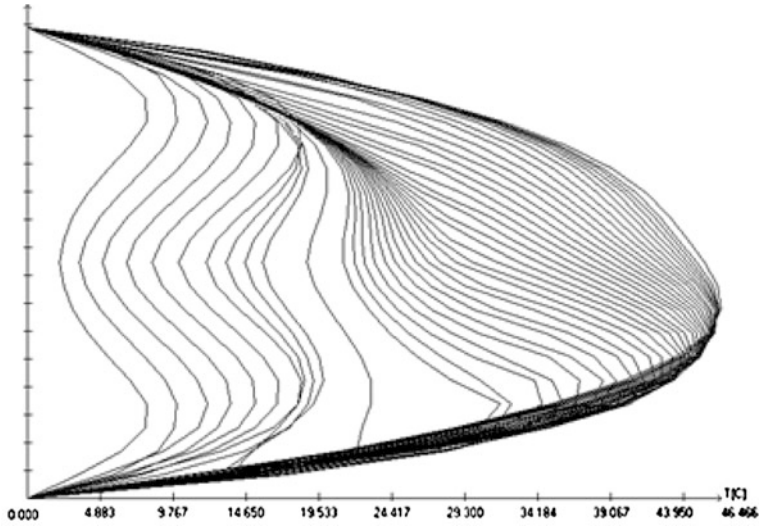


Fig. 8.9 The steady distribution temperature in the gap with size S = 0.1 [mm], (U = 9 [V], E = 3 [V], flow rate 6 [m/s], electrolyte: 12 % $NaNO_3$)

Figure 8.8 illustrates the effect of hydrogen and heat generation on gap size distribution. In described conditions, the heat and gas evolution diametrically changed course of the gap distribution.

The minimum practical tool gap size, which may be employed, however is constrained by the onset of unwanted electrical discharges. These short electrical circuits reduce the surface quality of the workpiece, and led to electro-erosive wear of the tool-electrode, and usually machining cannot progress because of them. Investigations of electrical discharges in an electrolyte reveal that the probability of electrical breakdown the gap is a function of the evolution of gaseous–vapor layers and passivation of the work surface. Intense heating, hydrogen generation sometimes choking phenomena and cavitation within the gap can lead to evaporation and subsequent gas evolution, and it is this gas which is believed to cause the onset of electrical discharge.

The issue of heating of electrolyte is primary importance for the determination of limit condition of ECM process. The distribution of mean temperature in the inter-electrode gap along the flow was determined using one-dimensional mathematical model of ECM process [1]. Further specification of temperature distribution was revealed in [3, 7–9]. Due to the heat exchange through electrodes as well as distribution of electrolyte velocity, the temperature changes along the flow path as well across the gap size.

The electrolyte temperature distributions across the gap width, at the different distances from the inlet electrolyte in ECM process with inter-electrode gap size S = 0.1 [mm] is shown in Fig. 8.9.

The two maxims that can be observed in (Fig. 8.9) in proximity of electrodes are very important in ECM input parameters selection. The input parameters should always be chosen such that the maximum temperature of electrolyte never reaches its boiling point. One-dimensional model, in which only average values of (T, and β) across the gap can be calculated, may not be accurate enough to properly estimate the maximum temperature.

For example, in the Fig. 8.9, the maximum average increment of temperature is $\Delta T_{av} = 32$ [°K], in this time, the maximum temperature, as shown in Fig. 8.9 is $\Delta T_{max} = 46.47$ [°K]. Use of input parameters from simulation that underestimated electrolyte temperature for actual machining may lead to short-circuit between electrodes and, what follows, to damage of tool and workpiece.

8.4 Conclusion

The presented CAE–ECM system can be useful for process planning for different variants of ECM. It can be used for process analysis, tool design and parameters selection. Presented software has significant potential to be used in industry applications.

References

1. McGeough JA (1974) Principles of electrochemical machining. Chapman & Hall, London
2. Rajurkar KP, McGeough JA, Kozak J, De Silva A (1999) New developments in electrochemical machining, Annals of the CIRP, vol 48/2, 567–579
3. Kozak J (1976) Surface shaping by electrochemical machining. Transaction of Warsaw University of Technology (WUT), Warszawa, (in Polish)
4. Davydov D, Kozak J (1990) High rate electrochemical shaping. Ed Nauka, Moscow (in Russian)
5. Davydov AD, Volgin VM (2007) Electrochemical Machining. In: Bard AJ (ed.) Encyclopedia of electrochemistry, vol 5. Chapter 12. Electrochemical Engineering. Willey-VCH, New York
6. Chang CS, Hornung LW (2001) Two-dimensional two-phase numerical model for tool design in electrochemical machining. Appl Electrochem 31:145–154
7. Klokov VV, Filatov EE, Firsov AG, Tikhonov A (1999) The complex computer simulation of the ECM blades shaping. The proceedings of the 15th international conference on computer-aided production engineering CAPE'99, Durham, pp 451–456
8. Kozak J (1998) Mathematical models for computer simulation of electrochemical machining processes. J Mater Process Technol 76(1-3):170–175
9. Kozak J, Dabrowski L, Lubkowski K, Rozenek M, Slawinski R (1999) CAE–ECM system for electrochemical technology of parts and tools. Proceedings of the 15th international conference on computer-aided of production engineering CAPE'99, Durham, pp 431–436
10. Kozak J (2011) Computer simulation of electrochemical machining. Lecture notes in engineering and computer science: proceedings of The World Congress on Engineering and Computer Science 2011, WCECS 2011, San Francisco, 19–21 Oct 2011

Chapter 9
An Analysis of the Genetic Evolution of a Ball-Beam Robotic Controller Based on a Three Dimensional Look up Table Chromosome

Mark Beckerleg and John Collins

Abstract This chapter describes how a robotic controller based on a 3-dimensional lookup table was used to control a ball balancing beam system. The evolved motion of the beam and the corresponding chromosome is analysed. The 3 system states of the ball and beam were translated by the lookup table into a motor speed and direction which maintained the ball in balance. The ball-beam states included the ball position, ball speed, and beam position. The reproduction method used 2-point crossover with a mutation rate of 2 percent. The selection method was tournament, and the population size was 100 individuals. Successful evolution was achieved on 4 lookup tables, each containing different maximum motor speeds. Each evolved lookup table was able to maintain the ball in balance for more than 5 minutes.

Keywords Artificial intelligence · Ball balancing beam · Ball-beam · Evolvable robotics · Evolution · Evolving lookup tables · Genetic algorithms · Lookup table-based controllers · Metaheuristic optimization · Robotics

M. Beckerleg (✉) · J. Collins
School of Engineering, AUT University, New Zealand. AUT City Campus,
Private Bag 92006, Auckland 1142, NZ
e-mail: mark.beckerleg@aut.ac.nz

J. Collins
e-mail: john.collins@aut.ac.nz

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_9,

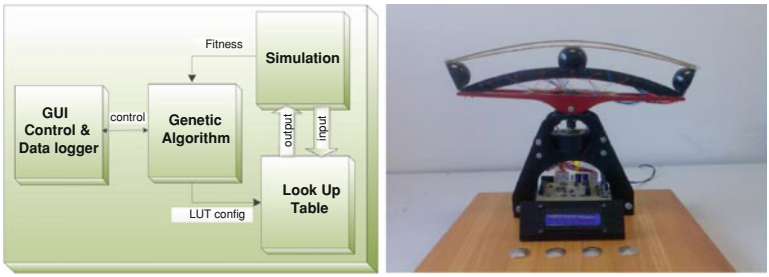


Fig. 9.1 Block representation and connections between the 4 units that were implemented on a computer and the physical beam that the simulation was modeled on

9.1 Introduction

An investigation was undertaken to determine if a genetic algorithm (GA) could be used to generate a ball-beam controller by evolving a population of lookup tables (LUT) employed to control the motion of the beam. The system that was developed (Fig. 9.1) consisted of 4 parts: (i) the graphical user interface (GUI), which displayed the motion of the ball and beam with control and data logging capabilities, (ii) the GA, which generated the final controller by evolving a population of LUTs, (iii) the simulation, which modeled the characteristics of the ball-beam system and (iv) the LUT, which provided the new beam motor speed and direction depending on the current ball-beam state [1].

In this experiment a curved beam was chosen rather than a standard straight beam, as the curved beam provided a more complex control system, and made the ball control more challenging.

The mathematical model and corresponding simulation was based on a physical ball-beam system that was developed at AUT University as part of a student project (Fig. 9.1). The position of the ball was determined using 19 infrared detectors while the motion of the beam was controlled by a stepper. The pulse rate of the stepper motor controlled the angular velocity of the beam. The maximum pulse rate was 125 pulses per second and the angular movement of the beam per pulse was 0.22 degrees. This resulted in a maximum angular velocity of the beam of 27.5 degrees per second.

9.2 Background

The ball-beam has historically been used to demonstrate control systems, owing to its non-linear dynamics and behaviour. It has now become a benchmark for research in this field. Studies using the ball-beam system have been implemented using a range of control systems such as proportional integral differential (PID)

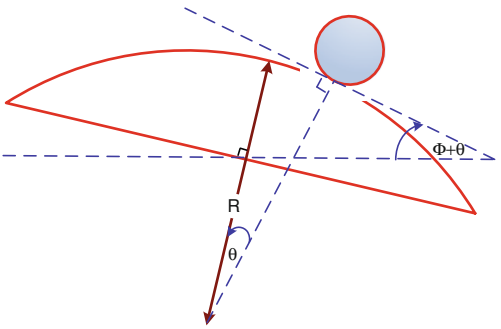


Fig. 9.2 The ball and beam showing the relationships between the angles and motion

control [2, 3], fuzzy logic [4, 5], and neural networks [6, 7]. Ball-beam controllers have been investigated using GAs. Some examples of evolved robotic controllers are the evolution of rules and classes of a fuzzy logic controller [8, 9], the weightings and connectivity of artificial neural networks [10, 11], and the coefficients of a PID controller [12, 13]. However the authors were unable to find any research into the use of a LUT for a ball-beam controller evolved by a GA.

LUTs have been used in evolution in a variety of ways. Robotic simulation has been replaced by lookup tables thus reducing real-time computation requirements [14, 15]. Cellular automata rules have been configured within a LUT and then evolved to create 2- and 3-dimensional shapes [16]. LUT's have been encoded with simulated DNA sequences and evolved to create robotic motion [17]. LUTs contained within a FPGA functional element have been evolved to create a robotic controller [18]. Finally, the authors have used GAs to evolve LUT based robotic controllers for 2 systems including a mobile inverted pendulum [19] and the gait of a hexapod robot [20].

9.3 Mathematical Model

The beam position (Fig. 9.2), is the angle φ from horizontal, while the ball position is the angle θ from the centre of the beam. Eqs. (9.1) and (9.2) outline the final equations for the ball acceleration. The full derivation for this mathematical model has been described by the authors in other literature [21].

$$\ddot{\theta} = A(\theta + \phi) \tag{9.1}$$

$$A = \frac{g}{R\left(1 + \frac{1}{mr^2}\right)} \tag{9.2}$$

Where

g — gravitational acceleration

I — moment of inertia of the ball
R — radius of curvature of the beam
m — mass of the ball
r — radius of the ball
θ — ball position (angle from the centre)
Ø — beam position (angle from horizontal)
x — ball position
v — ball velocity
b — beam position
a — acceleration of the ball

The value for acceleration (a) of the ball on the beam was determined by physical experimentation, as a factor of the ball position (x) and beam position (b) in Eq. (9.3). Using this acceleration the new ball position can be found dependant on it current velocity (v), acceleration and position as shown in Eq. (9.4), and the new speed of the ball dependant on its current speed and acceleration in Eq. (9.5). The simulation was adjusted to a time period of every millisecond in Eqs. (9.6) and (9.7).

$$a = 12x + 2.8b \tag{9.3}$$

$$x_{new} = x + vt + \frac{at^2}{2} \tag{9.4}$$

$$v_{new} = v + at \tag{9.5}$$

$$x_{new} = x + \frac{v}{10^3} + \frac{12x + 2.8b}{2 \times 10^6} \tag{9.6}$$

$$v_{new} = v + \frac{12x + 2.8b}{10^3} \tag{9.7}$$

9.4 Genetic Algorithm

A genetic algorithm is a metaheuristic optimization method that uses natural selection as a search engine. It acts on a population of individuals or chromosomes which are potential candidate solutions to the problem needing to be solved. Chromosomes are comprised of various forms such as bits, numbers or parameter sequences, depending on the problem. The genetic algorithm is iterative and is comprised of 3 main processes including reproduction, fitness evaluation, and selection. Reproduction is the generation of offspring from the surviving population of chromosomes. It uses 2 genetic operators: (a) crossover, where chromosomes are exchanged between parents, and (b) mutation, where parts of the parents' chromosomes are randomly altered. Fitness evaluation determines how

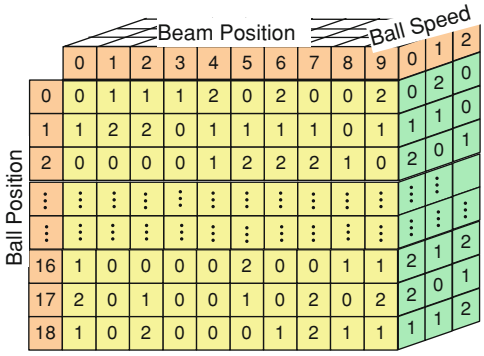


Fig. 9.3 Three dimensional LUT showing 19 ball positions, 10 beam positions, 3 ball speeds

well each chromosome in the population performs as a potential solution to the problem. Selection determines which chromosomes within the population will survive to the next generation based on their fitness.

The steps in a genetic algorithm are: (a) generation of an initial random population of chromosomes, (b) reproduction of offspring, (c) fitness evaluation of each new chromosome, and (d) selection, where the chromosomes with the best fitness are kept. The processes of reproduction, fitness evaluation and selection are repeated until the required fitness is reached or a set number of generations have been completed.

9.4.1 Chromosome

The heart of the controller was a 3-dimensional lookup table (Fig. 9.3). The lookup table contained the motor speed and direction required to drive the motor in such a way as to balance the ball. The 3-dimensions of the lookup table were linked to the ball and beam states. These were ball position (19 inputs), beam position (10 inputs), and ball speed (3 inputs). Several lookup tables were evaluated with a range of motor speeds varying from 2 to 11. The elements of the array were defined as char variables initialized with a randomly generated number quantised into 11 discrete steps ranging from 0 to 250. This enabled each location in the array to describe a motor speed with 5 left speeds, 5 right speeds and 1 stopped. The speed range was reduced when evaluating different speeds by adjusting the threshold so that the motor had limited speeds. For example with a 2 speed range, values below 125 would drive the motor hard left, while values above 125 would drive the motor hard right.

The LUT was used to control the beam's motor depending on the beam states. This was achieved by connecting the simulation's current ball-beam states to the axis of the LUT. The parameter at that location was sent back to the simulation to control the simulation's motor speed and direction.

Table 9.1 Search space within the LUT dependent on the number of motor speeds

Speeds	Search space
2	3.9×10^{171}
3	9.1×10^{271}
5	2.6×10^{398}
11	3.9×10^{593}

The search space that the GA explores was dependent on the total number of combinations that the chromosome can have. This was dependent on the number of locations within the LUT, and the number of speeds that were employed at each location. The experiments were repeated with 4 ranges of motor speeds: 2 (left and right), 3 (left, stopped and right), 5 (2 left, stopped and 2 right) and 11 speeds (5 left, stopped and 5 right). The total search space for each LUT was calculated using Eq. (9.8) and illustrated in Table 9.1. As evident in this table, the search space rapidly increased as the number of speeds increased. The exponent 570 was derived from the size of the LUT ($19 \times 10 \times 3$).

$$\text{Search space} = \text{speeds}^{\text{size of LUT}} = \text{speeds}^{570} \tag{9.8}$$

9.4.2 Reproduction and Selection

The objective of reproduction is to generate new offspring from a population of chromosomes which will have a higher fitness than their parents. The purpose of selection is to decide which offspring and parents to keep, with the goal that the population will move rapidly up the fitness landscape while maintaining enough diversity to bypass local maxima. The 2 parts of reproduction are crossover and mutation. Crossover is a method that is used to split and recombine the chromosomes from 2 or more parents into 1 offspring. Two-point crossover was used in this experiment using the x-axis (ball position) and y-axis (beam position) of the array as the positions within the array to be cut. The first cut points of the crossover were determined by randomly choosing points between 0–18 and 0–9. The end cut points of the crossover were determined by randomly choosing points between the first cut points and the end of the array, 18 or 9. Mutation will randomly alter the parameters within a single chromosome with the purpose to maintain the diversity of the population. It has a low probability of occurring which is typically between 0.1 to 2 percent. In this experiment a mutation rate of 2 percent was chosen, with every individual in the population being mutated after crossover occurred.

The selection pressure is an important parameter in genetic selection. The selection operator has a high selective pressure if it severely reduces the difference between individuals, or a low selective pressure if it allows many different individuals to survive. A low selection pressure will have a slow rate of convergence

to the optimum solution and can possibly stagnate, whereas a selection pressure that is too high may get stuck on local maxima due to loss of diversity. The choice of which selection method to use is dependent on what type of problem is to be solved, with each method having its advantages and disadvantages. Tournament selection with a group size of 2 was employed for this experiment. The choice of this group size was primarily to maintain a large diversity within the population with a moderate selection pressure.

9.4.3 Fitness Criteria

Each chromosome was evaluated by the simulation to see how well it functioned. This fitness evaluation was then used by the genetic selection to determine if the chromosome would be kept. The maximum fitness that a chromosome could have was 420 seconds. This was made up from 7 simulation runs with the beam positioned horizontally and the ball placed at rest on various positions on the beam. Each simulation run would last either until 60 seconds had passed, or until the ball reached a beam end stop.

9.5 Simulation

Using a genetic algorithm to evolve a physical robot controller is difficult due to the potential damage to the robot and its environment. As well, the complexity of the robot's actions has a large search space requiring a large number of generations before a suitable controller can be evolved. This is time consuming if performed in real time on an actual robot. To overcome these problems, evolutionary robotic genetic algorithms are normally performed using a software simulation of the robot. Once a suitable solution is found it can be transferred to the actual robot. However creating a simulation for a robot means modelling the real world which can never be entirely accurate, thus the final solution will carry with it the flaws in the simulation.

In this experiment the new ball position and speed were determined by the simulation every millisecond, based on the new beam position and the previously described Eqs. (9.6) and (9.7). Two maximum beam velocities of 22.7 and 45.4 degrees per second were evaluated. The motor speed and direction was converted by the simulation into a new beam position, and from this, the new ball position and speed were calculated. The motor speed and direction produced by the lookup table, was converted by the simulation into a new beam position. From this, a new ball position and speed were determined by the simulation based on the previously described Eqs. (9.6) and (9.7). The new ball-beam states were then feedback to the lookup table to determine the next motor speed and direction. The simulation recalculated the ball position and speed every millisecond. Two maximum beam velocities of 22.7 and 45.4 degrees per second were evaluated.

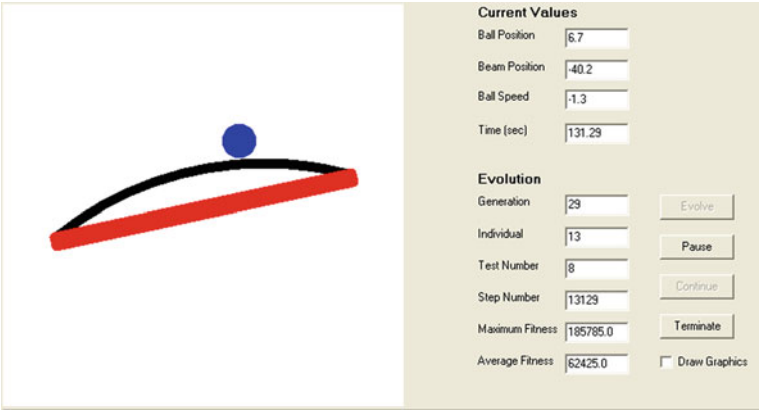


Fig. 9.4 The graphical user interface

9.6 Graphical User Interface

The GUI (Fig. 9.4) displayed the current ball position, ball speed, beam position, and the current time that the ball had remained balanced. A dynamic visual representation of the ball and beam in motion could be turned on or off, allowing the user to see how the ball and beam were responding at various stages of the evolutionary process. The visual representation was normally turned off, as when it was on, the evolutionary process was slowed to real time. Evolutionary control buttons were used to start, pause, and terminate the evolutionary process. The evolutionary parameters of generation number, current individual under test, average fitness of the population, and maximum fitness that had been reached was provided. A text display showed the number of speed settings, the maximum beam speed, the maximum fitness, average population fitness, generation number and time taken for the evolutionary process. These values were stored for later analysis.

9.7 Results

Two ranges of experiments were performed with 2 maximum motor speeds equating to a maximum beam angular velocity of 22.7 and 45.4 degrees per second. For each experiment, 4 ranges of motor speeds (2, 3, 5 and 11 speeds) were evaluated. Each simulation was repeated 7 times with a different ball start positions lying between ± 12 degrees from the top of the beam. The simulation was run until either the ball reached a beam end stop or 60 seconds had passed. This gave a maximum fitness of 420 seconds.

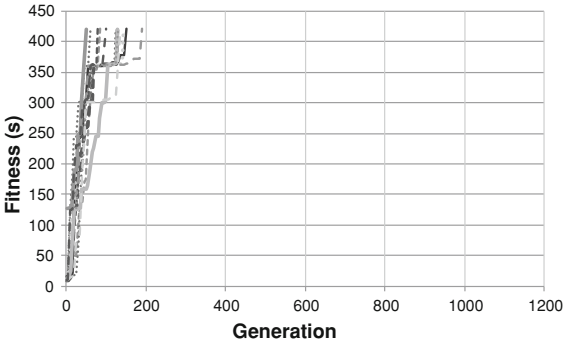


Fig. 9.5 Two motor speeds, maximum angular velocity of 22.7 ° per second

9.7.1 Evolved Motion of the Ball

The relationship between the fitness of number of generations and corresponding fitness of the best individual for 4 different maximum motor speeds is shown (Figs. 9.5, 9.6, 9.7, and 9.8).

The first observation of these results is that the LUT using only 2 motor speeds evolved in a shorter number of generations and time. This was due to 2 factors: (1) the smaller search space of the chromosome, and (2) no requirement for multiple speeds with an associated smoother response of the beam was built into the fitness.

The second observation was that the fitness increased in discrete steps, improving rapidly until it reached a plateau at a fitness of approximately 320 and 360 seconds. On investigation of the ball motion, it was found that it was difficult to capture the ball when it was started at the furthest position from the centre of the beam. To avoid failure at these start positions, the beam was required to move at maximum angular velocity in the opposite direction. This problem was more apparent in a LUT with 5 or 11 motor speeds as there was a greater possibility that the maximum beam angular velocity would not be used.

The motion of the ball-beam during the evolution process could be monitored on the GUI. It could be seen that the ball beam motion evolved through 4 stages. These were: (1) the beam remained stationary and the ball simply fell to an end stop; (2) the beam would react once, reversing the motion of the ball. However the ball would then fall to the opposite end stop; (3) the ball would be captured in 1 position, where the beam would perform an oscillation motion causing the ball to stay within 2 points. This pattern would last for between 5 and 10 seconds, but eventually the ball would break free and reach an end stop; (4) finally the beam was able to trap the ball between 2 points for the full 60 seconds test. This method of balancing by capturing the ball between 2 points by oscillating the beam was observed in all motor speed ranges.

It was observed that the ball tended to be captured near either end of the beam. This seemed unusual as it is a more risky position than a ball captured near the center of the beam. This was due to the design of the beam's sensor location, with

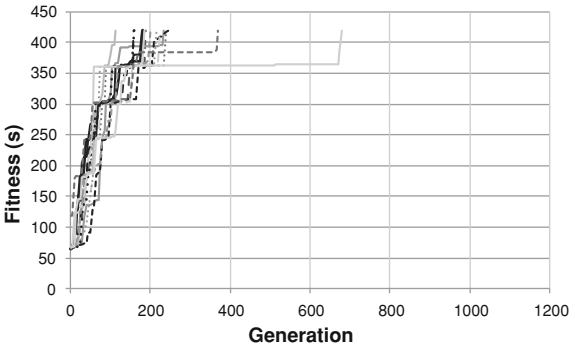


Fig. 9.6 Three motor speeds, maximum angular velocity of 22.7 ° per second

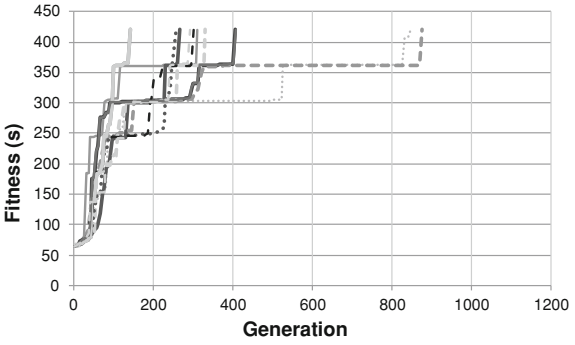


Fig. 9.7 Five motor speeds, maximum angular velocity of 22.7 ° per second

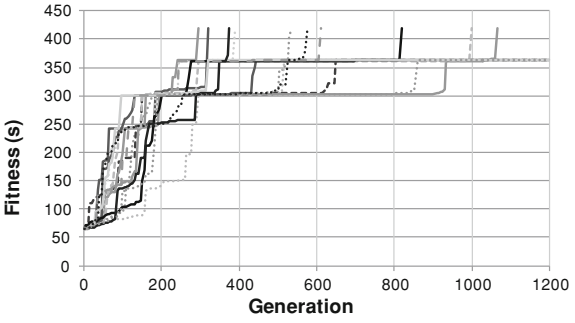


Fig. 9.8 Eleven motor speeds, maximum angular velocity of 22.7 ° per second

more sensors placed near ends of the beam as it was thought that this was a more important position. Unintentionally this gave the simulation a more accurate position of the ball and its speed at this point. Subsequently the evolved controller used the end locations to balance the ball.

A peculiarity of using a robotic simulation was observed when the ball was started at rest in the center of the beam. The evolved chromosome learnt to keep the motor off, thus achieving a perfect score for that run. This behaviour of course would not work well with a real ball-beam system.

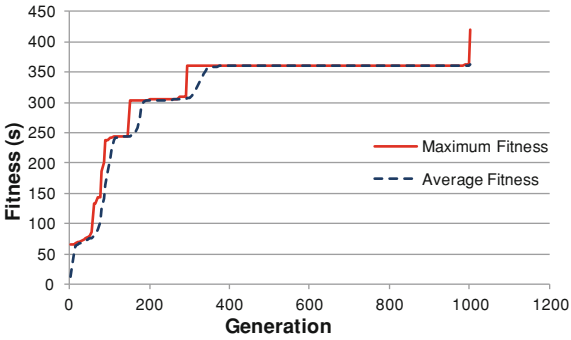


Fig. 9.9 Comparison between the maximum and average fitness, with 11 motor speeds and a maximum angular velocity of 22.7 ° per second

9.7.2 Evolved Chromosome

When the best chromosomes from several successful evolutionary runs were compared, it was found that each chromosome was different, producing a varied pattern for the beam and ball motions. This variation in successful chromosomes was due to the initial random population and the multiple pathways that the ball and beam could interact with the LUT. This variation in chromosomes was compounded by the fact that the evolution stopped once a successful pattern had been found.

Most successful simulation runs did not use a large part of the LUT because the ball would simply be moved to a position on the beam, and be kept in place by beam oscillations.

The maximum and average fitness of a typical run shows that the maximum fitness increased in abrupt steps, and then reached a plateau (Fig. 9.9). The average fitness would then converge with the maximum fitness at each plateau. It was thought that the population had converged at these points allowing only mutation to modify the chromosome and its associated fitness. However an investigation of the individual chromosomes and associated ball-beam motion showed that there was still population diversity, and that the reason for the convergence of fitness was due to the difficulty of evolving a chromosome that could start the ball at the edge of the beam.

9.7.3 Comparison of Two Maximum Motor Speeds

Multiple experiments were run using the 2 maximum beam angular velocities of 22.7 and 45.4 degrees/second with speeds ranging from 2 to 11 settings. A comparison of these results is shown in Table 9.2, which details the average fitness, number of generations and time the evolution was in progress at the end of the evolution. From this table it can be seen that the faster motor and minimum number of motor speeds had the best results in terms of the number of generations

Table 9.2 Comparison of the average fitness, average number of generations and the average time taken to evolve

22.7 ° per second			45.4 ° per second		
Generation	Av fitness	Time(s)	Generation	Av fitness	Time(s)
118	347726	197	42	268456	35
268	364240	592	56	327891	76
398	357240	3624	98	351811	297
861	359427	25794	103	349563	467

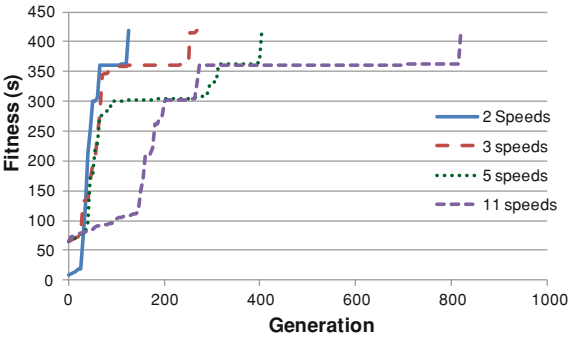


Fig. 9.10 Four motor speeds with maximum beam angular velocity of 22.7 ° per second

and the time taken to come to a successful evolution. It was noted that the time taken for the 5 and 11 motor speeds to successfully evolve was also acceptable despite the much larger search space. This was due to the constrained motion of the beam and the path that the ball took, with only a limited part of the chromosome being used for the beam control.

A comparison of the 4 motor speeds within each maximum motor pulse rate is shown (Figs. 9.10 and 9.11). From these graphs it can be seen that doubling the beam angular velocity had a significant improvement on the ability of the system to evolve, especially at the 5 and 11 speed range. The fitness plateau at 320 and 360 seconds can clearly be seen. All the solutions had difficulty with either one or both of the extreme starting points.

9.8 Conclusion

These experiments have shown that a 3-dimensional lookup table used as a controller for a ball-beam system can be successfully evolved to allow the ball to remain in balance for a total of 5 minutes. The motion of the ball and beam are unique for each successful evolved chromosome. The experiments were performed on 2 maximum beam angular velocities, and a range of 2, 3, 5 and 11 motor speeds. It was found that the higher beam velocity and lower number of motor speeds had the best evolutionary performance.

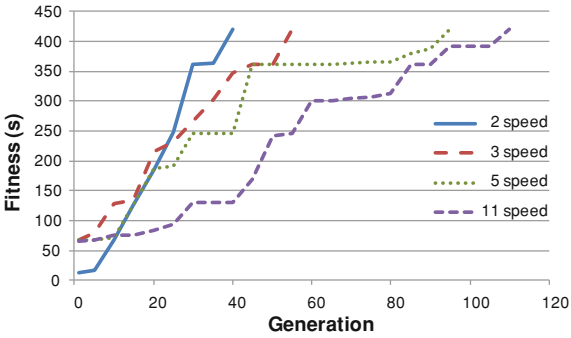


Fig. 9.11 Four motor speeds with maximum beam angular velocity of 45.4 ° per second

Future research will involve modifying the fitness evaluation such that the behaviour of the ball is more accurately controlled, for example to bring the ball to rest at a set point in the shortest amount of time. This could then be compared with other controllers such as proportional, differential and integral control. In addition further research will explore how the evolved chromosome can be moved from simulation to a physical beam.

References

1. Beckerleg, M, Collins J (2011) Evolving a three dimensional lookup table controller for a curved ball and beam system, lecture notes in engineering and computer science: proceedings of The World Congress on Engineering and Computer Science 2011, WCECS 2011, San Francisco, 19–21 Oct 2011, pp 360–366
2. Ka C, Nan L (1987) A ball balancing demonstration of optimal and disturbance-accomodating control. Control Syst Mag IEEE 7(1):54–57
3. Gordillo, F et al (2002) On the ball and beam problem: regulation with guaranteed transient performance and tracking periodic orbits, in proceedings of the international symposium on mathematical theory of networks and systems University of Notre Dame
4. Dadios EP et al (2000) Vision guided ball-beam balancing system using fuzzy logic. In: Industrial Electronics Society, IECON. 26th Annual Confjerence of the IEEE
5. Iqbal J et al (2005) Implementing ball balancing beam using digital image processing and fuzzy logic. In: Electrical and computer engineering, Canadian conference
6. Ng, KC, Trivedi MM (1996) Neural integrated fuzzy controller (NiF-T) and real-time implementation of a ball balancing beam (BBB). In: Proceedings IEEE international conference on Robotics and automation.
7. Eaton PH, Prokhorov DV, Wunsch DC II (2000) Neurocontroller alternatives for fuzzy ball-and-beam systems with nonuniform nonlinear friction. Neural Netw IEEE Trans 11(2):423–435
8. Tettamanzi, AGB (1995) An evolutionary algorithm for fuzzy controller synthesis and optimization. In: IEEE international conference on systems, man and cybernetics, intelligent systems for the 21st century.
9. Hoe Sung (2001) A self-organized fuzzy controller for wheeled mobile robot using an evolutionary algorithm. Ind Electron IEEE Trans 48(2):467–474
10. Bianco R, Nolfi S (2003) Evolving the neural controller for a robotic arm able to grasp objects on the basis of tactile sensors, 2829 edn. Lecture notes in computer science 375–384

11. Kim K-J, Cho S-B (2001) Dynamic selection of evolved neural controllers for higher behaviors of mobile robot. In: Proceedings 2001 IEEE international symposium on computational intelligence in robotics and automation, 2001.
12. Yi Z, Xiuxia Y (2004) Design for beam-balanced system controller based on chaos genetic algorithm. In: Information acquisition: proceedings, international conference
13. Pedersen, GKM, Butz MV (2010) Evolving robust controller parameters using covariance matrix adaptation: proceedings of the 12th annual conference on genetic and evolutionary computation, ACM, Portland, pp 1251–1258
14. Lund HH, Hallam J (1997) Evolving sufficient robot controllers. In: Evolutionary computation, IEEE international conference
15. Lund HH (2001) Co-evolving control and morphology with LEGO robots. In: Proceedings of workshop on morpho-functional machines
16. Chavoya A, Duthen Y (2006) Using a genetic algorithm to evolve cellular automata for 2D/3D computational development. In: Proceedings of the 8th annual conference on genetic and evolutionary computation, ACM, Seattle, pp 231–232
17. Greenfield G (2008) Evolved look-up tables for simulated DNA controlled robots. In: Proceedings of the 7th international conference on simulated evolution and learning, Springer-Verlag, Melbourne, pp 51–60
18. Z, RY et al (2002) Evolving FPGA-based robot controllers using an evolutionary algorithm, in first international conference on artificial immune systems
19. Beckerleg M, Collins J (2007) An analysis of the chromosome generated by a genetic algorithm used to create a controller for a mobile inverted pendulum. Studies in computational intelligence. 76
20. Currie J, Beckerleg M, Collins J (2008) Software evolution of a hexapod robot walking gait. In: 15th international conference on Mechatronics and machine vision in practice, M2VIP 2008.
21. Beckerleg M, Collins J (2008) Evolving electronic circuits for robotic control. In: 15th international conference on mechatronics and machine vision in practice. Auckland, New Zealand

Chapter 10
Development of a Bottom-up Compact Model for Intel®'s High-K 45 nm MOSFET

David E. Espejo Rodriguez and Alba G. Ávila Bernal

Abstract MOSFETs models have been critical components for evaluation of devices design and technology. These models face the challenge of being scalable to match the available semiconductor technologies. For the 45 nm MOSFET production, dielectric and metal gates were integrated. With new high dielectric materials and thinner oxide layers, new physics effects emerged that were not considered or integrated into the early models used in circuit simulators. Here an analytical model for 45 nm MOSFETs is presented. The model includes Short Channel Effects (Channel Length Modulation, the threshold voltage variation and carriers velocity saturation). The Drain-Source current and voltage equations derived from the model are implemented as a circuit device in SPICE 3F5. A comparison between the experimental data provided by the manufacturer and the simulation results obtained with the developed model integrating the technological and electrical parameters published by Intel®, demonstrates good agreement between both sets of data.

Keywords Compact · High-K · Model · MOSFET · Short channel effects · SPICE

D. E. E. Rodriguez (✉) · A. G. Á. Bernal
Microelectronics Center, Universidad de los Andes, Carrera 1 Este # 19A-40, Mario Laserna Building.ML-320 Office, Bogotá, Colombia
e-mail: de.espejo71@uniandes.edu.co

A. G. Á. Bernal
e-mail: a-avila@uniandes.edu.co

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_10,

10.1 Introduction

The semiconductor industry has continuously increased the density of transistors on a chip, successfully reducing their physical dimensions. This trend, predicted by Moore's Law [1], has imposed challenges on device modeling, which is an essential tool to simulate the operation of Integrated Circuit (IC) prior to the fabrication process. One of the main challenges of device modeling is to describe the behavior of nanometer scaled Metal Oxide Semiconductor Field Effect Transistors (MOSFETs).

Table 10.1 summarizes the state of the art of compact models developed for SPICE:

The behaviour of a MOSFET in an electrical circuit has been studied using circuit simulators. The *Simulation Program with Integrated Circuits Emphasis* SPICE, is an example of them.

From the table it can be observed that:

- There is a high complexity involved on recently developed compact models
- Models have been strongly focused on Silicon dioxide

The complexity arises from the top-down approach focusing on the application rather than on the electrical transport properties on the device. The models above studied, do have a challenge to include others dielectric materials beyond SiO_2. An example of high dielectric materials is the hafnium dioxide HfO_2. This chapter describes a methodology proposed to build a compact model based on the transport properties at nanoscale, aiming to integrate the outputs of analytical equations into ABM blocks to simulate the output of a basic circuit topology (half-wave rectifier).

10.2 Field Effect Transistors State of the Art

Emergent novel structures of Field Effect Transistors (FET) were investigated prior developing the model, both in research (R) and production (P) stages. Table 10.1 summarize the structures reported by the ITRS in 2009. The Table 10.2 intends to review three key parameters: gate length and description and the structures power dissipation.

Intel®'s 45 nm High-k MOSFET is selected as the focused of the model given the barriers to keep the high number of transistors on a single chip: dissipated power and sacrificing device performance (thin SiO_2 layers affect the transistors leakage current and speed) [10]. The target model aims to integrate and understand the short channel effects and to keep the compatibility with one of the most widely used IC simulation language: SPICE.

It can also be observed that the devices in production involve new materials different form the conventional CMOS devices materials. Among the available High-κ dielectric materials (such as ZnO_2) Hafnium dioxide is considered to be a promise for replacing the silicon dioxide due to the high dielectric constant, good thermal stability in direct contact with silicon substrates and low leakage current [6].

Table 10.1 State of the art of compact models for MOSFETs. Information gathered from [2–5]

Model	Year of release	Modeled dielectric material	Gate length lower limit	Complexity level
Level-1	1972	SiO_2	50 μm	Low
Level-2	1975	SiO_2	10 μm	Low
Level-3	1978	SiO_2	2 μm	Medium
BSIM1	1985	SiO_2	1 μm	Low
BSIM2	1990	SiO_2	200 nm	Medium
BSIM3	1995	SiO_2	50 nm	High
BSIM4.4	2009	SiO_2, with EOT derived from Silicon dioxide	10 nm	High

Table 10.2 Novel FET structures and some of their main parameters

Device	FinFET	GAA-FET	High-K	VRG-FET	TriGate
Stage	R	R	P	R	P
L_G (nm)	20	30	45	50	60
Gate description	Si segment (*Fin*) divides the polysilicon gate	Metal gate that surrounds the cannel	High-κ metal gate	dielectric	HfO_2 gate dielectric grown by Atomic Layer
Deposition	Three effective gates, sources and drains				
Dissipated power (W)	44 n	30 n	61 n	39 μ	72 μ
References	[7]	[8]	[9, 10]	[11]	[12, 13]

GAA-FET Gate All Around FET, *VRG-FET* Vertical Replacement Gate FET, *TriGate* Triple Gate FET

10.3 Intel®'s High-k MOSFET Electrical and Physical Parameters

Table 10.3 describes in detail the geometrical and electrical properties of the device, as reported by the manufacturer [11, 14].

Where L_G is the Gate length, *Tox* is the oxide layer thickness I_{TH} is the average leakage current, x_{dD} and x_{ds} are the depth of the depletion region associated with the Drain and Source respectively, ρ is the electrical conductivity of the HfO_2 at 1100 °C, V_{GS} is the gate-source voltage, I_D is the Drain current, V_{TH} is the threshold voltage, ε_r is the relative permittivity of HfO_2 and V_{DS} is the Drain-Source voltage.

From the data at Table 10.2, it is possible to recognize that for the 45 nm gate length High-κ MOSFET, the oxide thickness T_{ox} has reach the size of a few atomic layers. This device also is considered to be a *short channel device*, because its gate

Table 10.3 Intel®'s 45 nm MOSFET electrical and physical parameters

Physical dimensions	Material parameters	Electrical parameters
L_G	ε_r	V_{gs}
45 nm	20	0–1.7 V
T_{ox}	K	I_D
1 nm	14–22	0–104 μA
x_{dD}, x_{ds}	P	V_{TH}
20–90 nm	4.5×10^3 @1100 °C	0.32 V @ Vds = 1 V
Gate pitch	Dielectric band gap	
160 nm	6 eV	

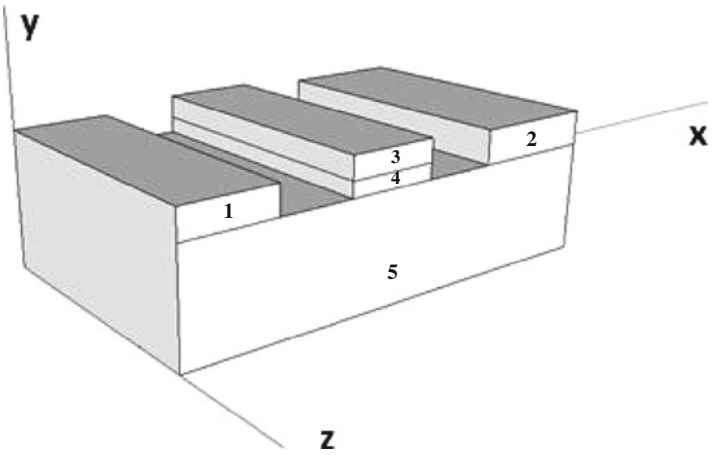


Fig. 10.1 Schematic view of a high-κ MOSFET. The main components are identified by number where *1* Source, *2* Drain, *3* Gate, *4* High-k dielectric layer, *5* Substrate

length (L_g) is in the same order of magnitude as x_{dD} and x_{ds}, which sets up a scenario where the electrical field along the y axis (see Fig. 10.1) is about 10^5 V/cm for N-channel High-κ MOSFET at $V_{DS} = 5V$ with $L_g = 100$nm [14], and where phenomena like carrier tunneling, Drain-Induced Barrier Lowering (DIBL), Channel Length Modulation (CLM), among others known as *Short Channel Effects* (SCE), are most likely to occur. SCE have not been fully implemented on modern MOSFET compact models mainly due to top-down approaches that make difficult to reach a particle-size level of detail in device analysis and authors considered the development of a model that accurately could describe the carrier transport in the presence of such phenomena.

10.4 A Model for Carrier Quantization

The model is based on the coupled Poisson-Schrodinger equations in order to describe the electron transport along the channel. The proposed procedure starts from Poisson equation describing 2-D potential distribution ϕ(x,y) along the channel.

$$\frac{\partial^2\phi(\mathrm{x,y})}{\partial \mathrm{x}^2} + \frac{\partial^2\phi(\mathrm{x,y})}{\partial \mathrm{y}^2} = \frac{\mathrm{q}}{\varepsilon_{\mathrm{HFO_2}}}(\mathrm{N_A(X)} + \mathrm{n(x,y)}) \tag{10.1}$$

Where q is the electron charge, ε_{HFO_2} is the electrical permittivity of Hafnium dioxide, $\mathrm{N_A}(x)$ is the density of acceptors along the channel and n(x, y) is the electron concentration along the channel.

In the scenario in which the charge is quantized, the continuous conduction band is now divided into two sub-bands. The wave function of each sub-band is given by the Schrödinger equation:

$$\frac{\hbar^2}{2m}\nabla^2\psi - q\phi(x,y)\psi_i(x,y) = E_i\psi_i(x,y) \tag{10.2}$$

Where $\hbar$ is the reduced Planck constant, $\psi_i(x,y)$ is the carrier wavefunction on the ith subband, m is the electron mass and the eigenvalue E_i is the energy level associated with wavefunction ψ_i [15]. Carrier's confinement is higher in the direction perpendicular to the Gate [16], so the one-dimensional Schrödinger equation can be used to assess the problem. Assuming a MOS structure with uniform potential distribution along the direction perpendicular to the Gate, the component in y axis can be removed from Poisson's equation, and the problem is reduced to coupled one-dimensional equations of Schrödinger-Poisson:

$$\frac{d^2\phi(x)}{dx^2} = \frac{qN_A(x)}{\varepsilon_{Hf}} - \sum_i \frac{Q_{inv,i}}{q\varepsilon_{Hfo_2}}|\psi_i(x)|^2 \tag{10.3}$$

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} - q\phi(x)\psi(x) = E\psi(x) \tag{10.4}$$

Where $Q_{inv,i}$ denotes the inversion charge on the ith subband. Reaching an analytical solution for this pair of equations is not an easy task; however, numerical simulations provide some insights into reaching this solution.

Figure 10.1 shows the results of simulating the carriers population per sub-band, using the parameters of Intel®'s High-κ MOSFET. $N_d = 1^{18}\mathrm{cm}^3$, $T_{ox} =$ 1 nm and $\kappa = 22$.

From Fig. 10.2, as we move on the voltage range between $0\mathrm{V}<\mathrm{V_{gs}}<3\mathrm{V}$, it is possible to see that between 45 and 85 % of the carriers are located in the lower sub-band $E_{1,1}$, so only the lowest level of energy can be considered to approach to the analytical solution. With that approximation, the pair of Eqs. (10.3) and (10.4)

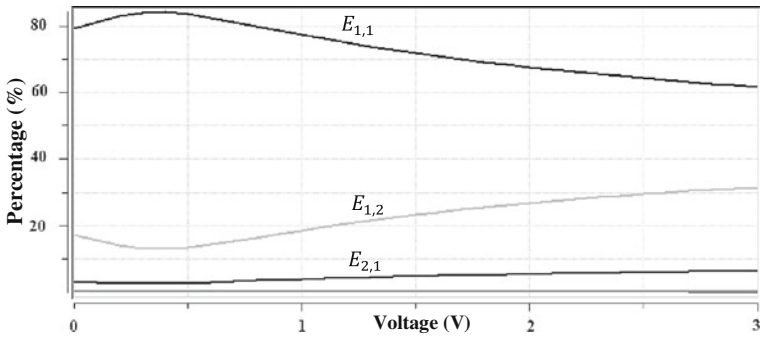


Fig. 10.2 Numeric simulation of carrier populations per sub-band, simulation conducted using SCHRED [18] with parameters $T_{ox} = 1$ nm, T = 300°K and $N_d = 10^{-18}\text{cm}^{-3}$

for each valley or group of sub-bands associated with the crystal structure of the interface *Si-Hfo*$_2$ are reduced to:

$$\frac{d^2\phi(x)}{dx^2} = \frac{qN_A(x)}{\varepsilon_{Hfo_2}} + \frac{Q_{inv,1}}{q\varepsilon_{Hfo_2}}\left|\psi_{1,1}(x)\right|^2 + \frac{Q_{inv,2}}{q\varepsilon_{Hfo_2}}\left|\psi_{1,2}(x)\right|^2 \tag{10.5}$$

$$-\frac{\hbar^2}{2m_i}\frac{d^2\psi_{1,1}(x)}{dx^2} - q\phi(x)\psi_{1,i}(x) = E_{i,1}\psi_{i,1}(x) \tag{10.6}$$

Where *i* denotes each one of the two valleys where the lowest energy level will be calculated.

Using calculus of variations, wave functions are assumed to have a shape similar to $\psi_{1,1}$ and $\psi_{1,2}$, thereby ensuring a good level of accuracy for the calculated energy levels [19]. The next step is to integrate the simplified Poisson equation from bulk to surface and also, in order to find the lowest expression of energy, expected value of the Hamiltonian of the wave function $\psi_{1,1}$ is calculated obtaining the following expression:

$$E_{1,1} = \int_0^\infty \psi_{1,1}(x)\frac{\hbar^2}{2m_1}\frac{d^2}{dx^2}\psi_{1,1}(x)dx + \int_0^\infty q\phi(x)\psi_{1,1}(x)^2 dx \tag{10.7}$$

A simulation of the sub-bands wave functions (shown in Fig. 10.3), provides elements to reach an analytical solution for Eq. (10.7)

Simulation of Fig. 10.2 shows that the peak of carrier density is a few nanometers (no more than 5 nm) below the channel surface, which leads to approximate:

$$d \gg \frac{1}{\alpha_1} \quad \text{and} \quad d \gg \frac{1}{\alpha_2}, \tag{10.8}$$

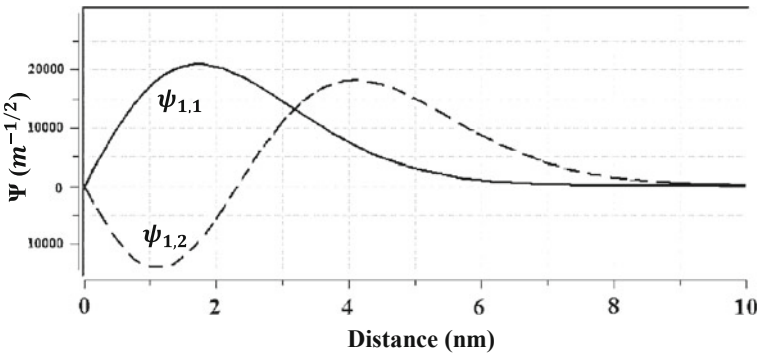


Fig. 10.3 Sub-bands wavefunctions, simulation conduced on SCHRED [20]

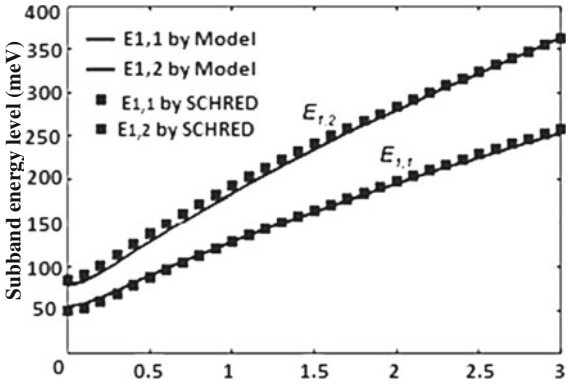


Fig. 10.4 Energy levels at each valley, by model expressions and by SCHRED simulation

Then, this approximation is applied and the result is factorized, based on the parameter of proportionality γ. According to the variations method, α_1 and α_2 should minimize the energy level, i.e:,

$$\frac{dE_{1,1}}{d\alpha_1} \quad \text{and} \quad \frac{dE_{1,1}}{d\alpha_1} = 0 \tag{10.9}$$

Calculating the derivatives, we obtain the expressions for α_1 and α_2 then the mean value of γ is obtained. Finally, replacing the values obtained and the approximations described above, leads to the expressions for $E_{1,1}$ and $E_{1,2}$:

$$E_{1,1} = \frac{3\hbar^2\alpha_1^2}{2m_1} \quad \text{and} \quad E_{1,2} = \frac{3\hbar^2\alpha_2^2}{2m_2} \approx 1,432E_{1,1} \tag{10.10}$$

Verification of model accuracy is performed by comparing its results with SCHRED simulation of energy levels for sub-bands $E_{1,1}$ and $E_{1,2}$ as shown in Fig. (10.4).

Accuracy of the quantization model obtained, allows us to approach to a short channel I-V model that takes into account the quantization of charge that occurs at this scale, and enables some approximations for that model.

10.5 I-V Short Channel Model

The fact that, in short channel devices, there is no complete control of the channel charge is an indicator that E_y is not negligible compared with E_x and the effects associated with this behavior should be incorporated in a compact model. In terms of I_d, the most significant effect to be included is velocity saturation, that may result in reduced effective drain saturation current. One of the empirical relations in use to model the dependence of carrier velocity V_d with respect to E_x was adopted for this step [19]:

$$|V_d| = \frac{|V_d| \max \frac{|E_x|}{|E_c|}}{1 + \frac{|E_x|}{|E_c|}} \tag{10.11}$$

Where E_c is defined as the intersection of the $v_d = \mu V_{GS} E_x$ line and an imaginary horizontal asymptote, as can be seen on Fig. 10.5 (Figs. 10.5, 10.6, 10.7).

The first step is to find an expression for I_d in non-saturation regime I_{dsn} E_x is expressed as the differentials of the potential between the polarization of the inversion layer and the end of the piece. The result of integrating these differences along the channel is presented in Eq. 10.12.

$$I_{DS} = \frac{W}{L_G} \mu c_{ox} \left[\frac{(V_{GS} - V_{TH}) V_{DS} - 0.5 \alpha v_{DS}^2}{1 + \frac{V_{DS}}{L_G . E_c}} \right] V_{DS} \leq V'_{DS} \tag{10.12}$$

Where W is the channel width and C_{ox} is the oxide capacitance. For the expression in the saturation region, it is necessary to include the effect of *Channel Length Modulation* (CLM) [18] finding the value of V_{ds} at which saturation occurs, so the expression of I_{ds} in presence of velocity saturation is given by:

$$\mathrm{I_{DS}} = \mathrm{W} \mu \mathrm{C_{ox}} \left[\frac{(\mathrm{V_{GS}} - \mathrm{V_{TH}}) \mathrm{V'_{DS}} - 0.5 \alpha \mathrm{v_{DS}^2}}{\mathrm{L_g} \left(1 - \frac{1_\mathrm{p}}{L_G} + \frac{\mathrm{V'_{DS}}}{L_G . E_c} \right)} \right] \tag{10.13}$$

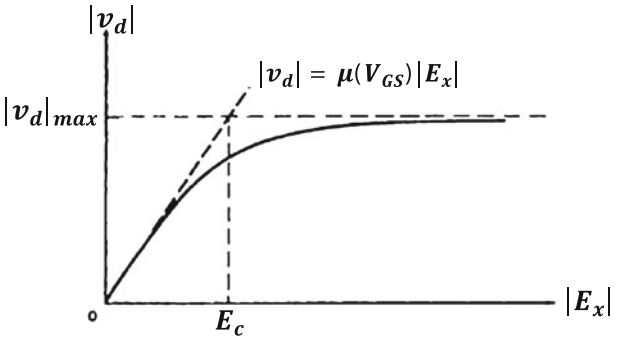


Fig. 10.5 Magnitude of carrier velocity in the inversion layer versus magnitude of the longitudinal component of the electric field [17]

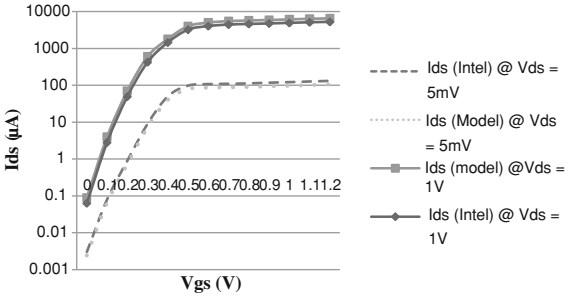


Fig. 10.6 Comparison of I-V characteristics provided by the manufacturer and calculated by the model equations

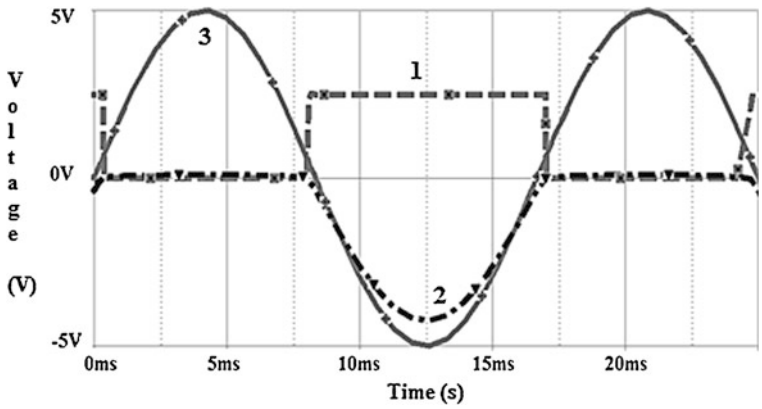


Fig. 10.7 Transient analysis using Orcad PSICE^{TM} 9.1 for three signals labeled. *1* Output voltage of native N-MOS MOSFET with High-k transistor physical and electrical parameters running under BSIM 4.4 model, *2* output of ABM blocks running model expressions obtained with this methodology, *3* input voltage signal with amplitude of 5 V and 60 Hz frequency. This results shows higher accuracy of simulations of a common circuit configuration compared with performance of state-of-the-art compact models, like BSIM 4.4

10.6 Compact I-V Short Channel Model

Finally, to obtain a unified expression of I_d in presence of velocity saturation, it becomes necessary to incorporate the smoothing equation of V_{ds}:

$$V_{deff} = V_{DSAT} - \frac{1}{2}\left[V_{DSAT} - V_{DS} - \delta_s + \sqrt{(V_{DSAT} - V_{DS} - \delta_s)^2 + 4\delta_s V_{DSAT}}\right] \tag{10.14}$$

Thus, V_{deff} is used to replace V_{ds} in the I_{ds} expression, as well as the early effective voltage V_{Aeff} to incorporate CLM in the unified expression, that is then obtained:

$$I_{ds} = \frac{I_{deff}}{1 + \frac{(R_{sd} I_{deff})}{V_{deff}}} \tag{10.15}$$

With

$I_{deff} = \left(1 + \frac{V_{DS} - V_{deff}}{V_{Aeff}}\right) I_{ds0}$ and $V_{Aeff} = \frac{E_{SAT} L_{eff} (E_{SAT} L_{eff} + V_{DS})}{\zeta v_{deff}}$

10.7 Comparison of Model Equations with Intel® Data

With the physical and electrical parameters given in Table 10.2 and those obtained by the *Arizona State University Predictive Technology Model* for 45 nm technology node [20], we compare the results using the IV model expressions obtained from the curve provided by the manufacturer (see Fig. (10.6)) [21].

The results show a correlation coefficient R^2 equal to 98 % and an average error of 0.33 %, indicating a relatively strong relationship between the data obtained by the model and those reported by the manufacturer. This level of accuracy, allows us to get to the next stage, which is the model equations testing in SPICE.

10.8 Testing the Model Expressions in a SPICE Circuit Simulation

To ensure the portability of the model for any SPICE-based simulator, obtained model expressions were represented by using Analog Behavioral Modeling blocks, included in Orcad© PSPICE 9.1. Those blocks have a maximum of three inputs and one output of voltage or current. They use mathematical relationships to model a circuit segment. When connected in cascade, and at netlist generation stage, the simulator concatenates the blocks to make the entire expression. Equations are entered in the model and tested in a configuration of half-wave rectifier-inverter,

compared with an N-channel MOSFET in the same configuration, whose electrical and physical parameters have been replaced by the ones of Intel's High-k MOSFET to verify the performance of BSIM4 model at sub-50 nm scale and the operation of the model obtained in a circuital implementation. Simulation (Fig. (10.7)) shows the degradation in the description of the MOSFET behavior in sub-threshold regime by BSIM4 model and a good performance of the equations obtained for the model.

10.9 Conclusions and Recommendations for Future Work

A compact model for High-κ 45 nm MOSFET was described including the short channel effects present at nanoscales. The methodology has consolidated in SCORM-compatible learning material (Sharable Content Object Reference Model), which is available at https://nanohub.org/resources/10024.

Future work could include additional effects such as temperature dependence and body effects. The model could be simplified to have as few parameters as possible, avoiding a great number of ABM blocks.

References

1. Moore G (1965) Cramming more components onto integrated circuits. Electronics 38(8):114–117
2. University of California, Irvine, Department of Electrical Engineering and Computer Science, HSPICE User's Manual, 2007
3. Vladimirescu A, Lius S (1980) Simulation of MOS using SPICE2. Electronics Research Laboratory, University of California, Berkeley
4. Mississippi State University (2005) MOSFET devices and their SPICE models. Department of Electrical and Computer Engineering
5. Ytterdal T, Cheng Y, Fjeldly T (2003) Device modeling for analog and RF CMOS circuit design, Wiley online library
6. Ávila A, Espejo D (2011) A SPICE-compatible model for intel®'s high-k 45 nm MOSFET. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2011, WCECS 2011, San Francisco, USA, pp 762–765, 19–21 October
7. Inaba S, Okano K, Izumida T, Kaneko A, Kawasaki A, Yagishita A, Kanemura T, Ishida T, Aoki N, Ishimaru K, Suguro K, Eguchi K, Tsunashima Y, Toyoshima Y, Ishiuchi H (2006) FinFET: the prospective multi-gate device for future SoC applications. In: Proceedings of the 36th European solid-state device research conference
8. Song J, Choi W, Park J, Lee J, Park B (2006) Design optimization of gate-all-around (GAA) MOSFETs. IEEE Trans Nanotechnol 5(3):186–191
9. Auth C, Capellani J, Dalis A, Ghani T, Mistry K (2009) 45 nm high-k + metal gate strain-enhanced transistors. Intel Press
10. Intel™ Press (2008) A 45 nm logic technology with high-k + metal gate transistors, strained silicon, 9 cu interconnect layers, 193 nm dry patterning, and 100 % Pb-free packaging

11. Hergenrother JM et al (2000) The vertical replacement-gate (VRG) MOSFET: a high performance vertical MOSFET with lithography-independent critical dimensions IEEE Electron Devices Meeting
12. Kavalieros J, Doyle B, Datta S, Dewey G, Doczy M, Jin B, Lionberger D, Metz M, Rachmady W, Radosavljevic M, Shah U, Zelick N, Chau R (2006) Tri-gate transistor architecture with high-k gate dielectrics, metal gates and strain engineering. Symposium on VLSI technology digest of technical papers
13. Doyle BS, Datta S, Doczy M (2003) High performance fully-depleted tri-gate CMOS transistors. IEEE Electron Device Lett 24(4):263–265
14. Bohr M, Chau R, Ghani T, Mistry K (2007) The high-k solution. IEEE Spectr 44(10):23–29
15. Ming L (2007) Semi-empirical device model for nanoscale MOSFET. PhD thesis, Universiti Teknologi Malaysia
16. Wang L (2006) Quantum mechanical effects on MOSFET scaling limit. PhD thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology
17. Hareland SA, Manassian M, Shih WK, Jallepalli S, Wang H, Chindalore GL, Tasch A, Maziar CM (1998) Computationally efficient models for quantization effects in MOS electron and hole accumulation layers. IEEE Trans Electron Devices 45(7):1487–1493
18. Vasileska D, Ahmed S, Mannino M, Matsudaira A, Klimeck G, Lundstrom M *SCHRED*, available on http://nanohub.org/resources/schred
19. Ren Z (2001) Nanoscale MOSFETs: physics, simulation and design. PhD thesis, Purdue University
20. University of California, Berkeley, Arizona State University, Berkeley Predictive Technology Model, available on http://ptm.asu.edu/modelcard/LP/45nm_LP.pm, retrieved on 26/04/10
21. Mistry K, Allen C, Auth C (2007) A 45 nm logic technology with High-k metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning and 100 % free Pb packaging, IEDM

Chapter 11
Power-Aware Topology Generation Based on Clustering for Application-Specific Network on Chip

Fen Ge and Ning Wu

Abstract A clustering-based topology generation approach is proposed to construct Network on Chip (NoC) topologies for given applications. The approach consists of four phases and constructs irregular NoC topology with design constraints, according to the communication requirements of the given application and characteristics of the router architectures. Specially, a recursion based link construction algorithm embedded in the topology generation is proposed to construct links between routers. The evaluation performed on various multimedia benchmark applications confirms the efficiency of the proposed approach. Experimental results show that the approach saves 61.5 % of power consumption on average in comparison with using regular Mesh topology. Significant network resource improvement is also achieved. Moreover, the approach performs well for two multimedia applications compared to existing algorithms.

Keywords 3D · Application-specific · Cluster · Network on chip · Power consumption · Topology generation

F. Ge (✉) · N. Wu
College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, 210016 Nanjng, P.R.China
e-mail: gefen@nuaa.edu.cn

N. Wu
e-mail: wunee@nuaa.edu.cn

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_11,

11.1 Introduction

The rapid advancement of semiconductor technologies makes it possible to integrate dozens of cores on a single chip. With more and more cores, the on-chip communication architecture design encounters more challenges in various aspects including the throughput, latency, power consumption, signal integrity, and clock synchronization. Traditional bus-based interconnect architectures are inherently non-scalable, which constitutes a bottleneck for the on-chip communication. The emerging Network on Chip (NoC) provides an effective, reliable and flexible infrastructure for system modules based on data packet transmission scheme. It has become an effective solution to overcome difficulties associated with global interconnections and communications in complex System on Chip (SoC) designs [1].

NoC architectures are constructed using topologies. A topology describes the overall connection forms between routers and resource nodes. The floorplan of a topology determines the length and complexity of the on-chip connections, and as a result, significantly affects the network latency, throughput, area cost and power consumption. Network topologies of NoC can be classified into two categories, regular and irregular architectures. Regular topologies, as used in most NoC designs (e.g., mesh and torus), have the advantage of reusability and low design complexity. However, with regular topologies, applications cannot be well optimized. This may lead to large-scale redundant routers, low link utilization rate, and local congestion. For example, the number of routers on a mesh architecture is fixed irrespective of how many of them are actually used. The same happens to the links between routers. Even if unused routers and links can be shut down, they still occupy area on the chip. Irregular topologies, on the other hand, are designed to be application specific and therefore, are tailorable for each design. Compared to regular topologies, they usually use fewer routers and links, while offering better system performance and lower cost [2].

In this chapter, we focus on network topology generation for the custom irregular architecture. Specifically, we propose a clustering-based topology generation approach for application-specific NoC. Parts of our work have been presented in [3] to minimize the network communication power consumption. This chapter expands the previous work with a further analysis of the feasibility to address the problem of application-specific 3D NoC topology generation using the proposed approach.

The rest of the chapter is organized as follows: Section 11.2 summarizes related work; Sect. 11.3 describes the problem definitions; Sect. 11.4 presents our topology generation approach with an example; Sect. 11.5 discusses the possible extension of the current approach; experimental results are discussed in Sect. 11.6, and finally the conclusion is made in Sect. 11.7.

11.2 Related Work

There are many advantages of using irregular topologies over regular topologies for application-specific NoC [4]. However, generation of irregular topologies calls for scalable topology generation algorithms [5–11]. In [5], the authors present a technique for constraint driven communication architecture synthesis of point to point links. The technique results in network topologies that have only two routers between each source and sink, and does not address routing for each communication trace. The work in [6] presents the mixed integer linear programming (MILP) based topology generation. However, this method is constrained by the exponentially increasing solution times for large communication trace graphs. Different optimization techniques have been proposed to address the problem of topology generation within reasonable time [7–10]. In [7] and [8], genetic algorithm based topology generation approaches are proposed, which obtain better results and less runtimes compared to the MILP technique. The author of [10] proposes a combination of the depth first search and the AO* algorithm to generated a near-optimal topology. However, these techniques have greater computational complexity due to a sufficient number of iterations.

In [11], a three-step topology generation algorithm called PATC is presented, which includes core cluster, core cluster optimizing and physical router mapping. The author of [12] proposes another simpler method called TopGen to cluster the given application based on the communication characteristic, and thereafter, construct the topology by connecting the clusters to each other one by one.

In this chapter, we propose a four-phase approach of topology generation analogous to those used in [11] and [12], but completely different in the algorithm design. The proposed approach is verified and compared to those using regular NoC topology and existing algorithms on multimedia benchmarks, which shows that our approach achieves better results.

11.3 Problem Formulation and Definitions

An NoC architecture consists of interconnected routers that are responsible for routing data packets on the communication architecture. As shown in Fig. 11.1a, a router is composed of switch fabrics, a routing and arbiter unit, an input port and output port module. Every resource node (IP core) should be connected to a router through input and output port channels, which consist of two unidirectional links. Each link can connect to a core by a network interface (NI) implemented with open core protocol (OCP), or connect to other routers directly to expand the architecture [11], as shown in Fig. 11.2b. In this case, designers can construct different regular or irregular NoC topologies based on the requirements and design constraints.

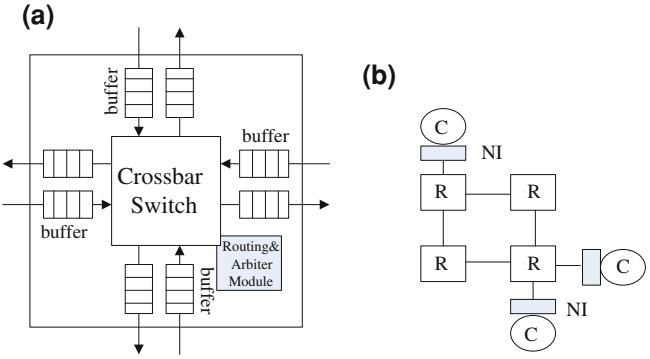


Fig. 11.1 The router structure and NoC architecture, **a** The router structure of NoC **b** NoC architecture

The topology generation problem can be formulated as follows.

Given a core communication graph denoted by $CCG(C, A)$, where each vertex $c_i \in C$ represents an IP core, and each directed edge $a_{i,j} \in A$ represents the communication trace from IP c_i to IP c_j. Every edge has two attributes, denoted by $b(a_{i,j})$ and $l(a_{i,j})$, which represent the bandwidth requirement in bits per second (Mbps) and the latency constraint in hops respectively.

Given a characterized library £ of the router architectures, with η denoting the number of input and output ports of the router, and Ω denoting the peak bandwidth that can be supported by the router on any one port.

Find a NoC topology T(R, E), where R ∈ £ represents the set of routers chosen to use from library £ in the topology generation, and E represents the set of links between the routers.

Such that:

(1) Each IP core c can be mapped onto a port of a router r, and the maximum number of cores mapped on a router should less than η.
(2) For each $a_{i,j} \in A$, there exists a unique path $p_{i,j} = \{(r_i, r_k), (r_k, r_m), \ldots (r_n, r_j)\} \in P$ in T that satisfies communication latency and bandwidth constraints.
(3) The total communication power consumption is minimized:

$$\min E(A) = \sum_{\forall a_{i,j} \in A} b(a_{i,j}) \times E_{bit}^{c_i,c_j} \tag{11.1}$$

where

$$\begin{aligned} E_{bit}^{c_i,c_j} &= \sum_{r \in p_{i,j}} E_{Rbit} + \sum_{e \in p_{i,j}} E_{Lbit} \\ &= (d(p_{i,j}) + 1) \times E_{Rbit} + d(p_{i,j}) \times E_{Lbit} \end{aligned}$$

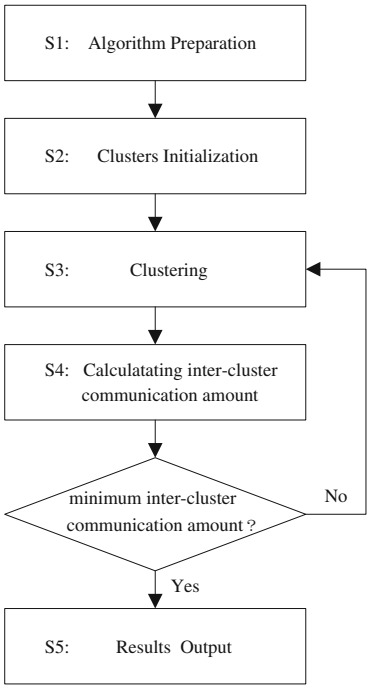


Fig. 11.2 The flowchart of the clustering algorithm

$E_{bit}^{c_i,c_j}$ represents the energy consumed when one bit of data is transported through the routing path $p_{i,j}$; E_{Rbit} and E_{Lbit} are the energy consumed on the router and the link respectively [11].

Since E_{Rbit} and E_{Lbit} are constants, the NoC power consumption varies linearly with the communication amount and routing distance, which can be represented by:

$$\min E(A) = \sum_{\forall a_{i,j} \in A} b(a_{i,j}) \times d(p_{i,j}) \tag{11.2}$$

Therefore, we try to cluster high communicative cores into the same router so that data exchanges among these cores consume minimized communication power consumption as calculated by (11.2).

11.4 Topology Generation Approach

The main idea of our proposed approach is to assign high communicative cores to the same routers or nearby routers, and subsequently, determine the optimal connection between routers. The goal is to minimize the total number of communication hops for communication IP core pairs, as well as to reduce the

number of used routers and links in the NoC topology. The approach consists of four phases: (1) core clustering, (2) cluster and router mapping, (3) router connection construction, and (4) topology optimization. Each phase of the approach is described in detail as follows.

11.4.1 Core Clustering

In the first phase, we partition the IP core set for a given application into several clusters under the design constraints. The flowchart of the clustering algorithm is shown in Fig. 11.2.

Step 1: Algorithm Preparation. We define a variable N_{max}, which denotes the maximum number of cores in each cluster. Since IP cores in the same cluster will be mapped to different ports of the same router in a topology, and each router must be connected to the topology on at least one port, $N_{max} = \eta - 1$. Then, we sort each communication trace $a_{i,j}$ in descending order according to the communication weight $b(a_{i,j})$.

Step 2: Clusters Initialization. Clustering is to partition vertices of $CCG(C, A)$ into k non-empty sets $C_1, C_2, \ldots, C_k$. Each cluster C_i ($i = 1, 2, \ldots, k$) contains N_{max} cores at most. In the initialization, each vertex of $CCG(C, A)$ forms a cluster partition, that is $CP = \{C_1, C_2, \ldots, C_n\}$, where $C_i = \{c_i\}$, $i = 1, 2, \ldots, N$, N is the number of vertices of CCG.

Step 3 and 4: Clusters Merging. According to the order of communication traces in step 1, we first process the edge $a_{i,j}$ with highest communication weight. Let $a_{i,j} = (c_i, c_j)$, if c_i and c_j belong to different clusters, and if the core number in the new cluster is not greater than N_{max} after merging, calculate the inter-cluster communication amount among clusters after merging. If the calculated amount is less than the previous one, merge the clusters, otherwise not.

Step 5: Results Output. When all the edges have been processed in sequence, we obtain the best number of clusters with minimum inter-cluster communication amount.

For example, we give CCG in Fig. 11.3a, in which the labels of the edges in CCG denote the bandwidth requirement. Assuming the number of router ports η is 4, each partitioned cluster contains $N_{max} = 4 - 1 = 3$ cores at most. According to the above clustering algorithm, the CCG can be divided into four clusters C_1, C_2, C_3, C_4, as shown in Fig. 11.3b.

11.4.2 Cluster and Router Mapping

In the second phase, we map each cluster to a router. The router number used in the generated topology is equal to the number of clusters. Every IP core in the cluster is mapped to a port of a router randomly.

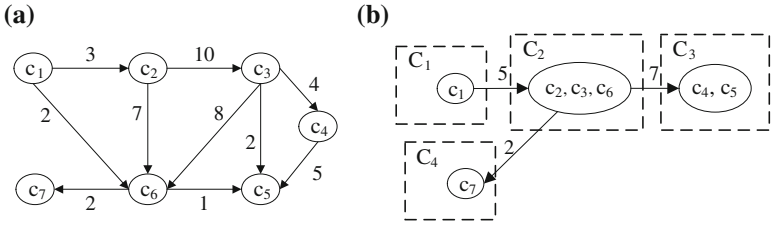
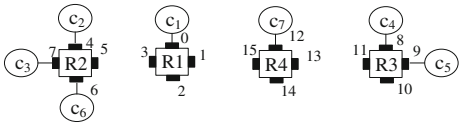


Fig. 11.3 Core clustering example, **a** Core communication graph **b** Clustering result

Fig.11.4 Cluster and router mapping



For the core clustering results shown in Fig. 11.3b, the clusters need to be mapped to four routers, denoted by r_1, r_2, r_3, r_4 respectively. As shown in Fig. 11.4, the core c_1 in the cluster C_1 is mapped to port 0 in the router r_1, and the cores in the cluster C_2 are mapped to three ports in the router r_2.

11.4.3 Router Connection Construction

In the third phase, the routers mapped with IP cores are connected to form the initial topology. We sort the clusters in ascending order according to their number of cores. For clusters with the same number of cores, we sort them in descending order according to their communication amount. Then, we use a recursion based link construction algorithm to generate router connections.

Before describing the recursion based link construction algorithm, it is worth pointing out that, the communication amount of a certain cluster is calculated as the sum of the inter-cluster communication amounts between this cluster and all others. Such sort will make the communication trace with high communication weight get shortest communication path in advance, and as a result, minimize the communication power consumption.

The idea of our proposed recursion based link construction algorithm is as follows. First, the source and destination routers for each communication trace are obtained according to current router selection and port mapping results; then, under the bandwidth and latency constraints, the following three ways are attempted to recursively search the path from the source router to the destination router:

(1) Use the existing links between source and destination routers;
(2) Use the empty port of routers without placing IP core between the source and destination router to build new links;
(3) Use the links built by previous communication trace from the source or destination router to other routers.

Through the above recursively search process, we can construct router connections by allocating a routing path for each communication trace.

The pseudo code of the recursion based link construction algorithm is shown in Fig. 11.5. The return value of the routine *get_next_rtr*(r_i) is r_{next} which is connected to the router r_i. The constructed link between router r_i and r_{next} should satisfy the bandwidth and latency constraints. The adjacency matrix $RAdj[M_R][M_R]$ represents the interconnection relation among routers, where M_R is the number of used routers in the topology generation. The initial value of the matrix elements is 0, and the value is between 0 and ∞ if there exists a link among routers. After allocating paths for all the communication traces, each element in $RAdj[M_R][M_R]$ is checked to ensure that its value does not exceed the supported bandwidth Ω. The port information list *PortList* is used to record the status of each router port. The status indicates whether the port is empty or connected with IP cores or other routers.

As an example, the number of cores in cluster C_1 and C_4 is identical as shown in Fig. 11.3b, and the communication amount of cluster C_1 is 5 which is larger than that of cluster C_4. As a result, the routing path for communication trace between cluster C_1 and C_2 is allocated first, and port 3 is connected to port 5 to construct a routing path. Then, the routing paths for other two communication traces between C_4 and C_2, C_3 and C_2 can be allocated. Eventually, after completing path allocations for all the communication traces, connection among routers can be constructed. The initial topology of the mapping results in Fig. 11.4 is shown in Fig. 11.6.

11.4.4 Topology Optimization

The last phase is to merge adjacent routers with empty ports until no adjacent routers can be merged. This further reduces communication power consumption and resources costs. As an example shown in Fig. 11.6, there exist empty ports in router r_1 and r_4, thus router r_1 can be merged with router r_4, leading to the final NoC topology as is shown in Fig. 11.7.

In order to evaluate the time complexity of our proposed approach, let n be the number of vertices in the core communication graph, and a be the number of edges in the core communication graph *CCG*. Since each cluster contains at most n elements and there exists a maximum of n clusters, the complexity of inter-cluster communication amount calculation is $O(n^2)$. All the edges should be

Algorithm Input：the corresponding source router r_{src} and destination router r_{dest} for each communication trace $a_{i,j}=(c_i, c_j)$.

Algorithm Output：the routing path $p_{rsrc,\ rdest}=\{(r_{src}, r_{next}), \ldots (r_n, r_{dest})\}$.

Recursive terminative condition：if $r_{src}==r_{dest}$ or find no path for the communication trace.

Recursive function：**route_construction**(r_{src}, r_{dest})

{ **if** ($r_{src}==r_{dest}$) exit;
if (no link between r_{src} and r_{dest})
{ **if** (existing empty port $port_i$ and $port_j$ in r_{src} and r_{dest})
{construct link between $port_i$ and $port_j$, let $r_{src}=r_{dest}$, add $b(a_{i,j})$to $RAdj[r_{src}][\ r_{dest}]$, and update *PortList*, exit;}
else if (no empty port in r_{src})
{ r_{next} = **get_next_rtr**(r_{src});
if (r_{next} != NULL)
{ let $r_{src}=r_{next}$, add $b(a_{i,j})$to $RAdj[r_{src}][\ r_{next}]$;
route_construction(r_{src}, r_{dest});}
else add $a_{i,j}$ to *PathUnAssignedSet*, exit;}
else if (no empty port in r_{dest})
{ r_{next} = **get_next_rtr**(r_{dest});
if (r_{next} != NULL)
{ let $r_{dest}=r_{next}$, add $b(a_{i,j})$to $RAdj[r_{next}][\ r_{dest}]$;
route_construction(r_{src}, r_{dest});}
else add $a_{i,j}$ to *PathUnAssignedSet*, exit;};}
else if (existing link between r_{src} and r_{dest})
{ **if** ($RAdj[r_{src}][\ r_{dest}]+ b(a_{i,j})\leq$ && $d(p_{rsrc,\ rdest}) \leq l(a_{i,j})$)
{ let $r_{src}=r_{dest}$, add $b(a_{i,j})$to $RAdj[r_{src}][\ r_{dest}]$, exit;}
else {
r_{next} = **get_next_rtr**(r_{src});
if (r_{next} != NULL && r_{next} != r_{dest})
{ let $r_{src}=r_{next}$, add $b(a_{i,j})$to $RAdj[r_{src}][\ r_{next}]$;
route_construction(r_{src}, r_{dest});}
else if (existing empty port $port_i$ and $port_{next}$ in r_{src} and r_{next})
{ construct link between $port_i$ and $port_{next}$;
let $r_{src}= r_{next}$, add $b(a_{i,j})$to $RAdj[r_{src}][\ r_{next}]$;
update *PortList*;
route_construction(r_{src}, r_{dest});}
else add $a_{i,j}$ to *PathUnAssignedSet*, exit;};}

Fig. 11.5 The pseudo code of the link construction algorithm

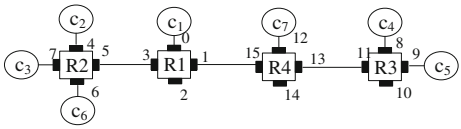


Fig. 11.6 Initial topology

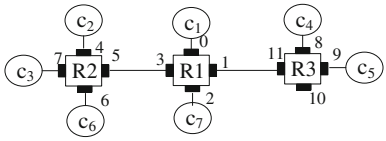


Fig. 11.7 Final topology

Table 11.1 Graph Characteristics

Graph	Graph ID	Nodes	Edges
MP3 decoder	G1	6	6
H.263 decoder	G2	7	8
MP3 encoder	G3	7	8
H.263 encoder	G4	8	11
MWD	G5	12	13
VOPD	G6	12	15
MPEG4 decoder	G7	12	26
H.263 enc MP3 dec	G8	12	17
H.263 enc MP3 enc	G9	14	19
H.263 enc H.263 dec	G10	15	19

traversed, so the time complexity of cluster partitioning is $O(a \times n^2)$. Consequently, the overall time complexity of the algorithm is estimated to be $O(a \times n^2)$.

11.5 Experimental Results

In this section, we present the experimental results obtained by executing the proposed approach on various multimedia benchmark applications. We generated custom irregular NoC topologies for seven combinations of four multimedia benchmarks: MP3 audio encoder, MP3 audio decoder, H.263 video encoder, and H.263 video decoder [5]. In addition, we obtained results for three other benchmarks: MPEG4 decoder, video object plane decoder (VOPD), and multi-window display (MWD) [2]. Table 11.1 lists the graph IDs and sizes of the *CCG* of the various benchmarks.

In order to evaluate the efficiency of the proposed approach, we compared the results produced by our clustering-based topology generation approach (Cluster-TG) against the solution of mapping benchmark applications onto regular Mesh topology. The selection of Mesh topology for comparison is due to the fact that,

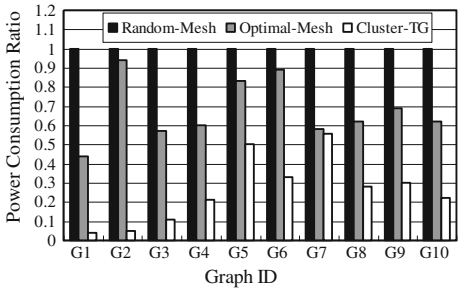
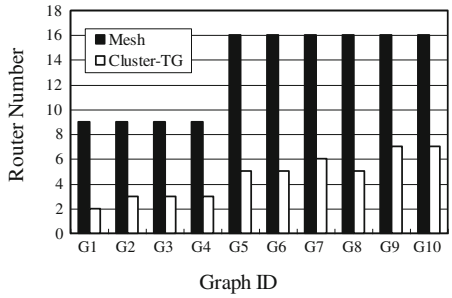
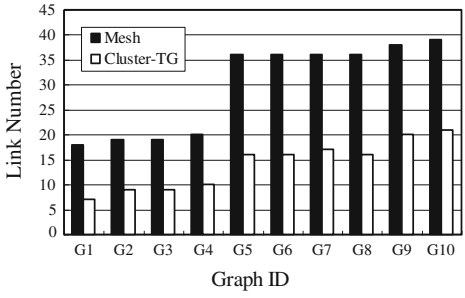


Fig.11.8 Communication power consumption comparison



(a) The number of routers



(b) The number of links

Fig.11.9 Resource costs comparison **a** The number of routers, **b** The number of links

Mesh topology is proved to outperform other regular NoC topologies with respect to power consumption and area costs, and it can be easily implemented on chips. The number of router ports η is set to be 4, and the supported bandwidth Ω is set to be 1 GB/s.

Figure 11.8 presents the results of the comparison in communication power consumption of NoC topology generated by Random-Mesh, Optimal-Mesh and Cluster-TG. 'Random-Mesh' represents the solution of mapping IP cores in benchmark applications onto regular Mesh topology randomly. 'Optimal-Mesh' represents the solution of mapping IP cores onto optimized regular Mesh topology by the genetic algorithm based approach in [13]. Figure 11.9 shows the comparison of router and link utilities. As seen from the figures, a much better

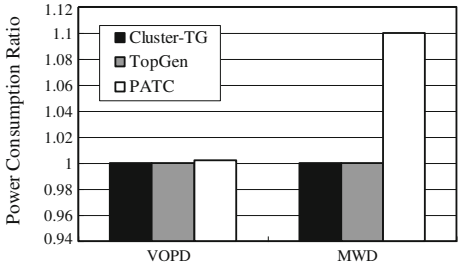
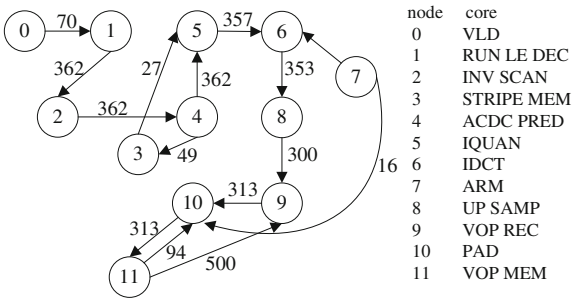
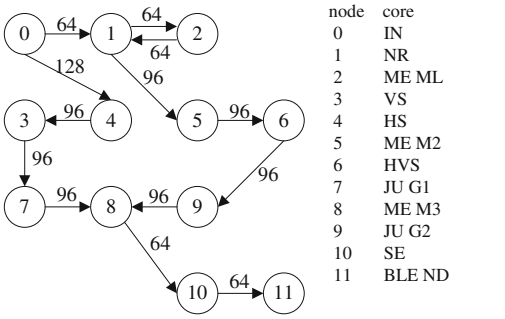


Fig. 11.10 Power consumption comparison for different approach



(a) The CCG of VOPD

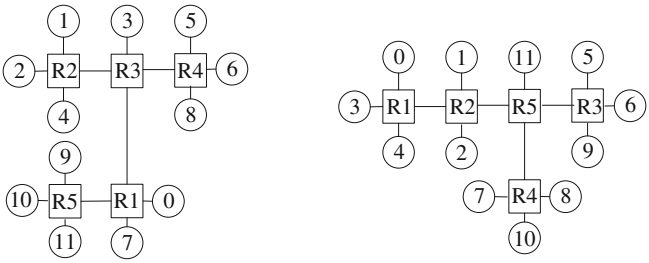


(b) The CCG of MWD

Fig. 11.11 The CCG of application VOPD and MWD, **a** The CCG of VOPD, **b** The CCG of MWD

performance in communication power consumption and resource costs has been achieved using our approach compared to that of the regular Mesh topology. On average, our approach saves about 61.5 % of communication power consumption compared to Optimal-Mesh.

Another experiment is conducted to compare the results of two multimedia applications, VOPD and MWD, generated by Cluster-TG, TopGen [12] and PATC [11] respectively. The resource costs of the applications using different approaches turn out to be about the same, and the power consumptions are compared in Fig. 11.10. It can be seen that our proposed approach achieves results that are



(a) The topology of VOPD **(b)** The topology of MWD

Fig. 11.12 The final irregular topology of application VOPD and MWD

better than PATC, and commensurate with TopGen. As an example, the CCGs and the generated irregular topologies of the VOPD and MWD benchmarks are illustrated in Figs. 11.11 and 11.12 respectively.

11.6 Possible Extension

The advent and increasing viability of 3D silicon integration technology make it possible to scale NoC over the third dimension [14]. As a result, 3D NoC is arousing more and more research interest. My proposed approach can be extended to application-specific 3D topology generation with metrics of 3D NoC taken into consideration.

In 3D NoC, IP cores are distributed on different 2D layers, and multiple device layers are stacked on top of each other with direct vertical interconnects tunneling through them using through-silicon vias (TSVs). Every IP core also should be connected to a router in 2D layers. The router connects to other routers in the same layer using horizontal links, and connects to other routers in the adjacent layers using up/down port and vertical links.

The approach for application-specific 3D NoC topology generation also should consist of four phases: core clustering, cluster and router mapping, router connection construction, and topology optimization. However, the problem introduces new issues, such as the technology constraint on the number of TSVs that can be supported, accurate power models for 3D interconnects.

In the phase of core clustering we first partition the IP core set for a given application (the example CCG is shown in Fig. 11.13a) into several clusters under the constraint on the number of TSVs, and make the IP cores in different clusters distribute on different 2D layers, as shown in Fig. 11.13b. Then IP cores in the same layer are further partitioned into clusters according to the algorithm in Sect. 11.4.1. In the phase of router connection construction, the routing path allocations for communication traces maybe use the vertical links among routers in

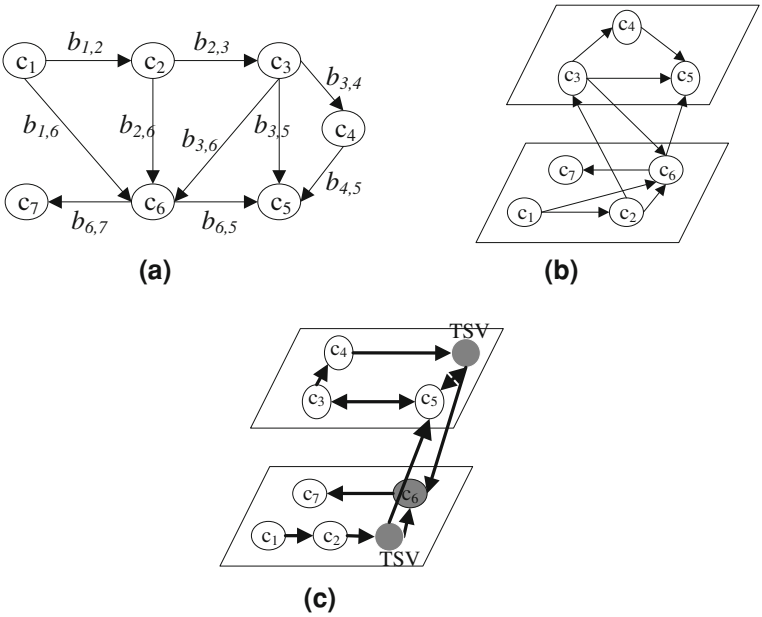


Fig. 11.13 Illustration of the 3D NoC topology generation problem, **a** Core communication graph **b** Clustering result, **c** The construction of vertical links

adjacent layers, as shown in Fig. 11.13c. The construction of vertical links should meet the constraint on the number of TSVs

Additionally, the power model in 2D NoC should be extended to 3D NoC by including the power consumed on vertical links.

11.7 Conclusion and Future Work

This chapter presents a four-phase clustering-based topology generation approach for application-specific NoC. The aim is to reduce the network communication power consumption. Under the constraints of the bandwidth and latency, the approach designs custom irregular NoC topologies according to the communication requirements of the given application and characteristics of router architectures. Specially, a recursion based link constructing algorithm embedded in the topology generation is proposed to construct links between routers. Applying our approach on various multimedia benchmark applications gives experimental results showing significantly improved performance as compared to those using regular Mesh topology and existing algorithms. The detail analysis of 3D NoC topology generation using our approach will be done as future work.

Acknowledgments This work was supported by the Natural Science Foundation of China under Grant 61076019 and 61106018, the Aeronautical Science Foundation of China under Grant 20115552031, the China Postdoctoral Science Foundation under Grant 20100481134, and the NUAA Scientific Research Foundation for talent introduction.

References

1. Benini L, De Micheli G (2002) Networks on chips: a new SoC paradigm. IEEE Comp 35(1):70–78
2. Bertozzi D, Jalabert A, Murali S et al (2005) NoC synthesis flow for customized domain specific multiprocessor systems-on-chip. IEEE Trans Parallel Distrib Sys 16(2):113–129
3. Ge F, Wu N, Qin X, Zhang Y (2011) Clustering-based topology generation approach for application-specific network on chip. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2011, WCECS 2011, Oct 19–21, 2011, San Francisco, USA, pp 753–757
4. Ogras U, Marculescu R (2006) It's a small word after all: NoC performance optimization via long-rang link insertion. IEEE Trans Very Larg Scale Integr Sys 14(7):693–706
5. Pinto A, Carloni LP, Sangiovanni-Vincentelli AL (2003) Efficient synthesis of networks on chip. In: Proceedings of the international conference on computer design, 2003, pp 146–150
6. Srinivasan K, Chatha KS, Konjevod G (2006) Linear-programming-based techniques for synthesis of network-on-chip architectures. IEEE Trans Very Larg Scale Integr Sys 14(4):407–420
7. Srinivasan K, Chatha KS (2005) ISIS: a genetic algorithm based technique for custom on-chip interconnection network synthesis. In: Proceedings of the international conference on VLSI Design, 2005, pp 623–628
8. Leary G, Srinivasan K, Mehta K, Chatha KS (2009) Design of network on chip architectures with a genetic algorithm-based technique. IEEE Trans Very Larg Scale Integr Sys 17(5):674–687
9. Choudhary N, Gaur MS, Laxmi V, Singh V (2010) Genetic algorithm based topology generation for application specific network-on-chip. In: Proceedings of the IEEE international symposium on circuits and systems (ISCAS), 2010, pp 3156–3159
10. Liu Z, Cai J, Yao L, Du M (2009) Application-aware generation and optimization for NoC topology. In: Proceedings of the IEEE youth conference on information, computing and telecommunication, 2009, pp 259–262
11. Chang KC, Chen TF(2008) Low-power algorithm for automatic topology generation for application-specific networks on chips. IET Comp Digit Tech 2(3):239–249
12. Ar Y, Tosun S, Kaplan H (2009) TopGen: a new algorithm for automatic topology generation for network on chip architectures to reduce power consumption. In: Proceedings of the AICT, 2009, pp 1–5
13. Ge F, Wu N (2010) Genetic algorithm based mapping and routing approach for network on chip architectures. Chin J Electron 19(1):91–96
14. Yan S, Lin B(2008) Design of application-specific 3D networks-on-chip architectures. In: Proceedings of the IEEE international conference on computer design, 2008, pp 142–149

Chapter 12
Remote Hand Motion Detection and Monitoring with Noise Reduction

Jing Pang

Abstract The digital triaxial accelerometer has become more and more popular in the research and industrial world due to its small size, low power consumption, low cost and its sensing capability in the x, y and z-axis directions. This paper presents several noise reduction schemes for hand motion detection with the triaxial digital accelerometer ADXL345 by using the RCM3365 board as the web server for control and also for data monitoring.

Keywords Digital accelerometer · Hand motion · Kalman filter · Median filter · Moving average filter · Noise reduction · Output data rate · Remote data monitoring · Web server

12.1 Introduction

The triaxial accelerometer based hand motion detection has a wide range of potential applications, including handsets, mobile device, gaming, personal navigation devices, learning, sports medicine, health-care and other gesture-based customer devices.

The small size integrated microelectromechanical system (iMEMS) accelerometer such as ADXL345 [1] has an integrated on-chip analog-to-digital converter (ADC) and it provides digital outputs with SPI and I^2C digital interface signals. Such accelerometers can be connected with the microcontroller through the SPI or I^2C

J. Pang (✉)
Department of Electrical and Electronic Engineering, Computer Engineering Program, California State University, 6000 J Street, Sacramento, CA 95819, USA
e-mail: pangj@gaia.ecs.csus.edu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_12,

interface, and can be used for acquiring x-axis, y-axis and z-axis direction sensing data. Usually, one limitation of microcontrollers is that they have limited memory for data storage. As a result, this design work used the RCM 3365 module as a web server to control the ADXL345 device and monitor hand motion. At the same time, the remote PC could obtain the acquired accelerometer data through Ethernet and then store them into the Excel database in real time for remote monitoring.

One 8-bit Rabbit 3000 microprocessor is mounted on the RCM3365 module [2] and it runs with a maximum clock frequency of 44.2 MHz. RCM3365 has 512 K Flash, 512 K program SRAM, and 512 K data SRAM. It works with a DC power supply that ranges from 3.15 to 3.45 V. With a 3.3 V power supply and a 44.2 MHz clock, the current consumption is about 250 mA. RCM3365 supports Dynamic C with the royalty-free TCP/IP stack to enable rapid and secure web interface development.

A C# program running on a remote PC was developed to acquire real-time accelerometer data on the internet and also to store data in Microsoft Excel sheets on the remote PC. Noise reduction is important for getting reliable and high resolution triaxial accelerometer measurements. This is especially true for some sensitive applications that require a high level of precision. The major noise sources for the surface technology silicon digital accelerometers are from Brownian mechanical noise, electronic thermal noise, and internal analog to digital conversion quantization noise.

The noise reduction methods considered in this work include techniques to be implemented on the hardware in real time and also techniques for post processing using software.

The rest of the paper consists of multiple sections. Section 12.2 describes the hardware system. Section 12.3 explores multiple real-time noise reduction techniques, including using different output data rates, applying a median filter [3] and applying a moving average filter [4]. Section 12.4 discusses the Kalman filter [5] noise reduction technique for post processing. Finally, Sect. 12.5 gives conclusion.

12.2 Hardware System

Several hardware devices are used in the remote hand motion detection system shown in Fig. 12.1. They are the accelerometer ADXL345, the RCM3365 module, the NETGEAR wireless router, and the PC. The ADXL345 is used to detect hand motion. The RCM3365 communicates with the ADXL345 through an I^2C interface to acquire hand motion data. The NETGEAR wireless router connects the RCM3365 with the internet so that the server data stored inside the RCM3365 module can be posted on the website in real time. Moreover, the remote PC gets access to the website accelerometer data and stores them into the Excel database in real time.

In Fig. 12.1, the ADXL345 chip works with a 3.3 V power supply. The clock signal SCL and the data signal SDA are connected with 2 K ohm pull-up resistors

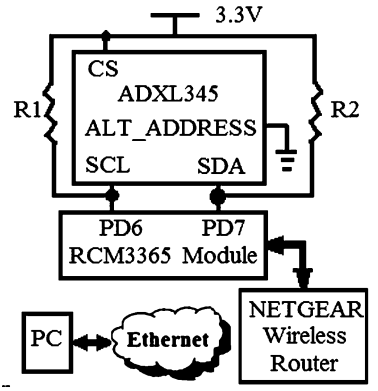


Fig. 12.1 Remote hand motion detection and monitoring hardware system

R1 and R2. They are controlled by the RCM3365 module through the I^2C interface. The I^2C mode is enabled when the CS pin is tied high to 3.3 V. The ALT_ADDRESS pin is the alternate I^2C address selection signal. The alternate I^2C address of 0x53 is chosen when the ALT_ADDRESS pin is grounded and this translates to 0xA6 for writing and 0xA7 for reading. By default, the Rabbit 3000 microprocessor on the RCM3365 module uses the PD6 and PD7 pins for I^2C communication. They are connected with the SCL and SDA pins of ADXL345 respectively.

The ADXL345 can be attached to the top of the user's hand. Six different hand positions have been tested successfully based on my previously conducted work: straight up, left-tilted facing up, left-tilted facing down, upside down, right-tilted facing down, and right tilted facing up [6]. In my previous work, the proposed algorithm covers a relatively wide angle range for each hand motion position and it is not sensitive to small sensor errors. The wireless router NETGEAR is used to interface the RCM3365 with the Internet. The RCM3365 serves as the web server to control and monitor hand motion data collected by the accelerometer. The remote PC can access the hand motion accelerometer data through the Ethernet using the C# program, and then store them into the Excel database in real time.

In order to detect a more accurate accelerometer angle for hand motion, noise reduction techniques need to be studied. At the same time, the noise reduction techniques need to consider the tradeoff between the algorithm complexity and the real-time angle detection.

12.3 Real Time Noise Reduction Schemes

Three schemes were implemented in this design work to study the effect of real-time noise reduction including: changing the RCM3365 output data rates, applying a moving average filter and applying a median filter.

Table 12.1 The BW_RATE 0x2C register map [2]

D7	D6	D5	D4	D3	D2	D1	D0
0	0	0	LOW_POWER	Rate			

12.3.1 BW_RATE Register and Rate Code

The BW_RATE 0x2C register map of the ADXL345 chip is shown in Table 12.1. The rate bits set the output data rate and the LOW_POWER bit sets the power mode.

In this design work, the LOW_POWER bit in the BW_RATE register is set to 0 to allow the ADXL to work in the normal mode to reduce noise. According to the ADXL345 datasheet, when the LOW_POWER bit is set to 1, the ADXL operates with reduced power consumption. However, the noise is somewhat higher.

In order to check the relationship between the output data rate and the noise level, several output data rates were tested: 400, 200 and 100 Hz. The rate codes are set according to the Table 12.2.

12.3.2 Performance Comparison

Figures 12.2, 12.3 and 12.4 illustrate the hand motion detection in the x-axis, y-axis, and z-axis directions with the RCM3365 output data rate set to 400 Hz.

For Fig. 12.2 the mean and standard deviation of accelerometer x-axis angle are 1.123325 and 0.244906, respectively.

For Fig. 12.3 the mean and standard deviation of accelerometer y-axis angle are 5.260033 and 0.372795, respectively.

For Fig. 12.4 the mean and standard deviation of accelerometer z-axis angle are 5.385482 and 0.363293, respectively.

In order to reduce noise, both the moving average filter and the median filter applied a window size of four to the accelerometer data samples. According to the results shown in Table 12.3, both methods get similar performance results and they achieve much smaller standard deviation values compared with the original x-axis, y-axis, and z-axis data.

Since the ADXL345 hardware itself presumably supports mechanisms internally to improve the accuracy of the measurements when using a lower sampling rate, different output data rates were tested, along with a moving average filter and a median filter for a window size equal to 4. The results are shown in Figs. 12.5, 12.6, and 12.7. These figures show that reducing the output rate also reduces the standard deviation of the noise in the measurements. With the window size equal to four, either the moving average filter or the median filter can achieve the similar performance to the method of changing the output data rate to four times smaller.

Table 12.2 Output data rate and rate code [2]

Output data rate (Hz)	Rate code
400	1100
200	1011
100	1010

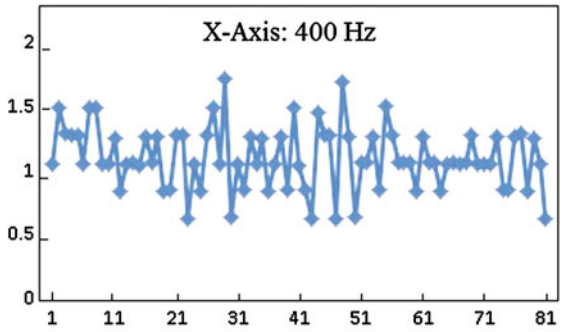


Fig. 12.2 Hand motion detection in the x-axis with the output data rate of 400 Hz

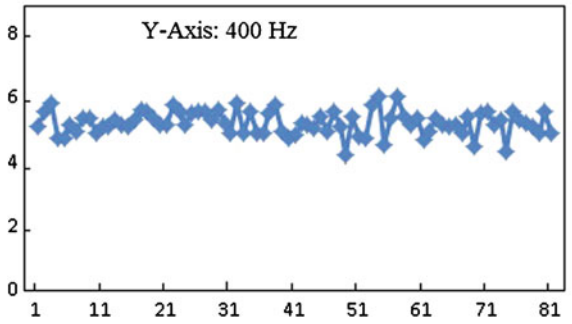


Fig. 12.3 Hand motion detection in the y-axis with the output data rate of 400 Hz

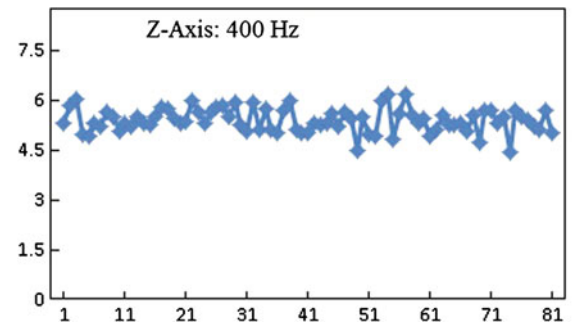


Fig. 12.4 Hand motion detection in the z-axis with the output data rate of 400 Hz

Figure 12.5 shows that if the moving average filter or the median filter is applied to the accelerometer data, the additional ADXL345 output data rate reduction will not lower the measurement noise standard deviation a lot.

Table 12.3 Mean and standard deviation for accelerometer data with output data rate 400 Hz

	Original data	Moving average	Median
X-axis mean	1.123325	1.127214	1.130681
X-axis standard deviation	0.244906	0.09544	0.111451
Y-axis mean	5.260033	5.259894	5.267801
Y-axis standard deviation	0.372795	0.167936	0.187202
Z-axis Mean	5.385482	5.386154	5.392974
Z-axis standard deviation	0.363293	0.166385	0.186787

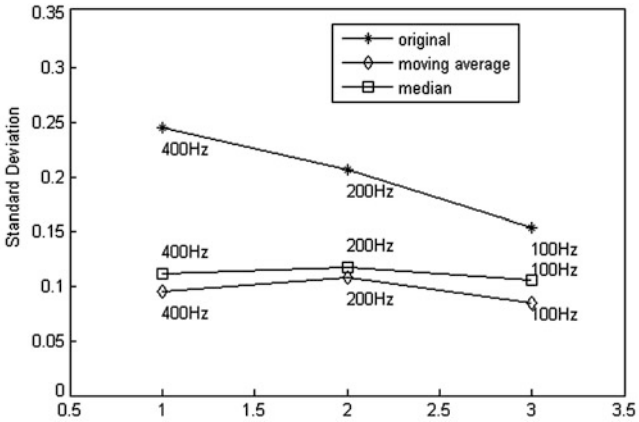


Fig. 12.5 Standard deviation for hand motion detection schemes in the x-axis direction

Figure 12.6 looks similar to Fig. 12.7 because y-axis and z-axis direction angles are similar in the testing case. Since the hand is approximately still, the mean and standard deviation of both angles show similar values.

Both the moving average and median filters smooth the input data high frequency noise. The median filter is especially good for smoothing the impulse noise. For a small window size of 4, only a small amount of storage is required and it allows for the fast computation speed for the moving average and median filter implementations. However, if the window size becomes bigger, both the moving average and median filters will create data lags.

12.4 Kalman Filter

In comparison to the moving average filter and the median filter methods which require past measurements, the Kalman filter uses only the present measurements and the previously calculated state [7].

The measurements from accelerometers are perturbed by noise. In this work, all process and measurement errors are assumed to have Gaussian distributions. The

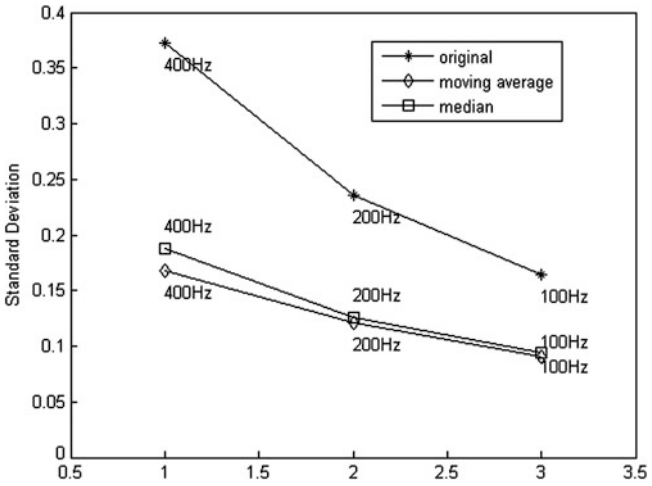


Fig. 12.6 Standard deviation for hand motion detection schemes in the y-axis direction

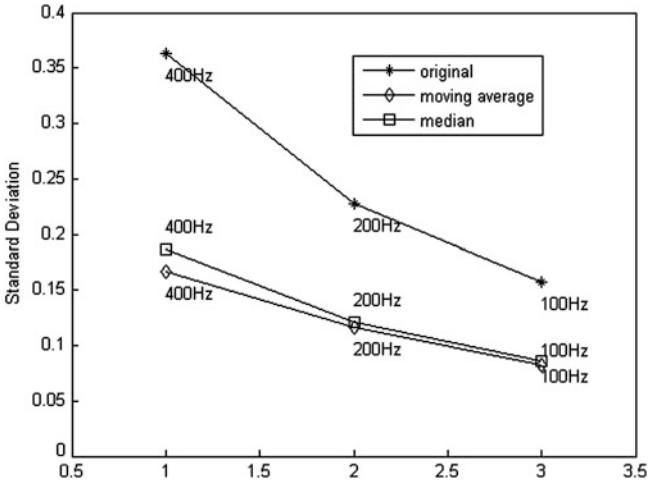


Fig. 12.7 Standard deviation for hand motion detection schemes in the z-axis direction

accelerometer measurement is modeled as a linear dynamical system according to the Kalman filter framework in the following equations. X(t) is the vector that includes the accelerometer x-axis, y-axis, z-axis angle measurements as well as the angular speed signals: mx(t), my(t), mz(t), vx(t), vy(t), and vz(t). X(t) and X(t − 1) are the internal states of the process at time t, and time t − 1 respectively. X(t) is represented as:

$$X(t) = F*X(t - 1) + W(t) \tag{12.1}$$

$$F = \begin{bmatrix} 1 & 0 & 0 & dt & 0 & 0 \\ 0 & 1 & 0 & 0 & dt & 0 \\ 0 & 0 & 1 & 0 & 0 & dt \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{12.2}$$

$$X(t) = [mx(t) my(t)\ mz(t)\ vx(t)\ vy(t)\ vz(t)]^T \tag{12.3}$$

$$dt = 1 \tag{12.4}$$

Where W(t) is the process noise which has zero mean and covariance Q, All speed signals are normalized based on the time interval of dt equal to 1.

Moreover, Z(t) is the vector representing the noisy accelerometer x-axis, y-axis and z-axis angle measurements according to the following equations.

$$Z(t) = H * X(t) + V(t) \tag{12.5}$$

$$H = \begin{bmatrix} 1 & 0 & 0 & dt & 0 & 0 \\ 0 & 1 & 0 & 0 & dt & 0 \\ 0 & 0 & 1 & 0 & 0 & dt \end{bmatrix}; \tag{12.6}$$

Where V(t) is the measurement noise which has zero mean and covariance M.

To solve the above equations, the Kalman filter provides computational means to obtain the state of the process using the prediction Eqs. (12.7) and (12.8), the Kalman gain Eq. (12.9) and the update Eqs. (12.10), (12.11), and (12.12) recursively. The P matrix is the error covariance matrix and the final filtered output matrix is OUTX. The Kalman gain is optimized to minimize the error covariance matrix. One advantage of the Kalman filter is that the new state estimate only requires the estimated state from the previous time step and the current measurement.

$$X_{prediction} = F * X_{estimation} \tag{12.7}$$

$$P_{prediction} = F * P_{estimation} * F^T + Q \tag{12.8}$$

$$K_{gain} = \left(H * P_{prediction}{}^T\right) * \left(H * P_{prediction}{}^T * H^T + R\right)^{-1} \tag{12.9}$$

$$X_{estimation} = X_{prediction} + K_{gain} * \left(Z - H * X_{prediction}\right) \tag{12.10}$$

$$P_{estimation} = P_{prediction} - K_{gain} * H * P_{prediction} \tag{12.11}$$

$$OUTX = H * X_{estimation} \tag{12.12}$$

In Figs. 12.8, 12.9 and 12.10, the Kalman filter is applied to the same data used in Figs. 12.2, 12.3 and 12.4 with the output data rate of 400 Hz.

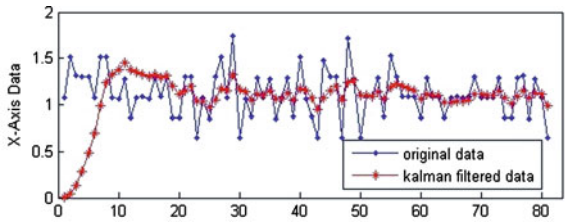


Fig. 12.8 Kalman filtered hand motion detection in x-axis direction

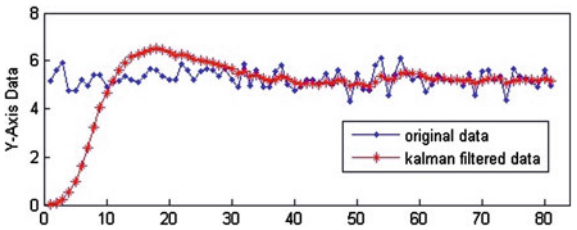


Fig. 12.9 Kalman filtered hand motion detection in y-axis direction

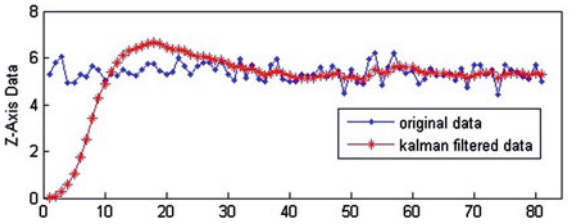


Fig. 12.10 Kalman filtered hand motion detection in z-axis direction

In Fig. 12.8, the mean is 1.1074 and the standard deviation is 0.0688 for the x-axis accelerometer data with the output data rate of 400 Hz.

In Fig. 12.9, the mean is 5.1836 and the standard deviation is 0.1374 for the y-axis accelerometer data with the output data rate of 400 Hz.

In Fig. 12.10, the mean is 5.3082 and the standard deviation is 0.1362 for the z-axis accelerometer data with the output data rate of 400 Hz.

Compared with the other schemes used in Table 12.3, the Kalman filter method gets the best standard deviation for the filtered result based only on the present measurements and the previously calculated state.

12.5 Conclusion and Future Work

Different noise reduction schemes are studied in this paper for hand motion detection using the accelerometer ADXL3365 and the rabbit RCM3365 module. In general, reducing the output data rate lowers the noise level. For real-time operation, both the moving average filter and the median filter techniques require the small window size

in order not to get data lag. For the small window size such as four samples, the moving average filter and the median filter techniques have similar performance when they are applied to the accelerometer data. The Kalman filter scheme has the best standard deviation compared with the other methods. Furthermore, it is based only on the present measurements and the previously calculated state, which will allow for the real-time implementation of the noise reduction.

In this work, the different noise reduction schemes for hand motion detection are studied. However, the advantage of the Kalman filter scheme is only studied for the data post processing using software. In the future, further study will be explored to implement the Kalman filter scheme on hardware in real-time to allow noise reduction and tracking of hand motion for more complicated gesture recognition.

References

1. Analog Devices Inc., 3-Axis, ±2 g/±4 g/±8 g/±16 g Digital Accelerometer ADXL345, Product Datasheet, Norwood, MA (2011)
2. Digi International Inc., RabbitCore RCM3365/RCM3375 C-Programmable Core Module with NAND Flash Mass Storage and Ethernet, Product Manual, Minnetonka, MN (2008)
3. Arce GR (2005) Nonlinear signal processing: a statistical approach. Wiley, New Jersey
4. Demosthenous P, Nicolaou N, Georgiou J (2010) A hardware-efficient lowpass filter design for biomedical applications. 2010 IEEE biomedical circuits and systems conference, pp 130–133
5. Zhao Y, Yang Y, Kyas M (2011) Comparing centralized Kalman Filter schemes for indoor positioning in wireless sensor network. International conference on indoor positioning and indoor navigation, 2011, pp 1–10
6. Pang J, Singh I (2011) Accelerometer based real-time remote detection and monitoring of hand motion. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2011, WCECS 2011, vol II, 19–21 Oct 2011, San Francisco, USA, pp 744–747. http://www.iaeng.org/publication/WCECS2011/WCECS2011_pp744-747.pdf
7. Kalman RE (1960) A new approach to linear filtering and prediction problems 1. Trans ASME J Basic Eng 82(Series D):35–45

Chapter 13
Development of Precision High Speed AC Power Monitoring Device for Power Regulation and Control

Aditya P. Kulkarni and N. K. S. Rajan

Abstract Precision measurement of AC power at conversion rates comparable to the driving frequency is a challenging task but is a desirable feature in applications involving regulation of power, loaded on a device. Practically, the application assumes significance, typically in developing a test-bench for characterization of a power generation system such as engine (DG set), or to measure and manage the power load in a specified pattern. A good analysis of demands and deliveries of power loads require capturing of transient responses, leading to a need for faster conversion rates in the measurement. Since the AC power measurement involves RMS integration, known to be inherently a slower process, the transient phenomena like a step loading tends to get smudged during measurement and often could lose its identity. It is known that typical conversion rates of commercially available power meters are about 0.5–1 Hz. In the approach presented here, a specifically developed analog circuitry, that is built using industrially common analog devices is shown to provide a high definition RMS integration of the AC power (with a typical conversion rate better than 35 Hz for a 50 Hz signal source) that is developed, tested and qualified for its functional integrity.

Keywords Analog circuitry · Engine testing · Higher conversion rates · Power regulation · Precision AC transient capture · RMS measurement

A. P. Kulkarni (✉)
Indian Institute of Science, E-10 New Housing Colony, 560012 Bangalore, India
e-mail: adityakulkarni990@gmail.com

N. K. S. Rajan
Indian Institute of Science, E-5 New Housing Colony, 560012 Bangalore, India
e-mail: nksr@cgpl.iisc.ernet.in

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_13,

13.1 Introduction

Quality of electrical power is an important factor for the life today, with the AC power being made as the widest from of energy distribution and consumption. To maintain a good quality of service, having backup power generation units and in the other upcoming trend of having distributed power generation, use of engines of appropriate size is becoming more practical. This leads to an important need for assessment and qualifying of the engines on their performance in meeting the power demands for varied applications in tandem with fuel efficiency. With medium to small engines, this performance testing is a lucid indicator of its economic viability and evaluating this critically for varying loads is not a commonly addressed problem. In addressing this aspect, precision of measurement of power under transient states of operation is necessary for evaluating the compliance of the system under test. Building a test rig that has ability to load the system in a designated pattern is another important factor to render such tests carried out. In designing such a facility, a common and most important link is the high quality measurement system. The commercially available true RMS meters generally have a limited functional scope as regards to their response times and are not quite adoptable for using them for generating a control signal and in analyzing the transients cost effectively. The approach presented imbibes a cost effective electronic circuitry that is designed, built and tested in an attempt to address the issues mentioned.

The device presented has been developed and tested with an intention for producing an affordable industry standard electronic unit (EU) and is expected to enhance the scope of measurement of AC power to applications that are currently limited as mentioned before. The design imbibes multi-functionality to the module allowing it to be used also as precision voltmeter and ammeter, each of which can be independently used.

In realizing the test facility discussed, that calls for a regulated or controlled power levels that effectively corrects for fluctuations in power inputs and on fluctuations in load the design needs a control with an error correcting system having feedback loop of power output. The EU designed provides options for using precision feedback control of voltage and thereby controlling the RMS power of the system. This serves to either to regulate the power to constant value or to control it precisely in a predefined pattern. This aspect of control could also be found useful in certain other applications such as establishing of constant fluid flow rate or in constant torque devices.

13.2 Approach

The design concept involves rendering a high speed and a strongly linearized conversion of analog AC signals that correspond to the applied current and voltage on the device drawing the power, into the true RMS of power with all the signals

processed in analog condition. The analog design adopted in the present approach is an important aspect that eliminates the 'conversion time' involved in a digital processing. The design is focused on a reliable measurement of power and uses lower number of components. The unit is designed to respond faster, to be better than half of the cycle period of the source signal. Active filters are integrated to selectively and dynamically filter out the ripples resulting at the integrators. The ripple free output could be reliably used in the high quality measurements or in controlling of external devices for power regulation or for other sub-systems with pre-designated actions such as power loading panels.

The operational schematics of the design features are shown Fig. 13.1. The input stages draw signals from a current transformer (CT) or from an alternate current sensor and another input from voltage transformer (PT) or an alternate sensor for it, to acquire separately the 'current' and 'voltage' of the monitored power device concurrently. The inputs are individually conditioned to suit subsequent processing.

The conditioned AC signals are rendered non–negative with a precision rectifier built out of active components and Schottky diodes [1, 2]. The output of this stage is fed to a two stage integrator taking care to see the integration is manage to be linear enough by selecting suitable range of charge pump. A bank of two analog low pass active filters of Butterworth design is introduced in cascade with a cut off frequency of 25 Hz (with the background that the source under measurement is at 50 Hz) that determines the adequateness of the response of the RMS conversions of the parameters—voltage and current. Two independent and identical processing streams consisting of these functional blocks of precision rectification, RMS Integration and low pass active ripple filtering constitute the analog linearized AC to DC converter for the current and voltages being measured. These parameters are multiplied in the analog state using AD632—a precision device from analog devices [3] that provides the quantity of power derived from the product of current and voltage signals. Provisions for adjustable gain and offset corrections are embedded at suitable points for end-use customization.

The design uses standard analog devices such as op-amps and other passive devices allowing for an option for standardization of the EU built for this at the end-user level. The required design information on the components is available in corresponding data sheets of the devices concerned. The design of active filter used, is based on the earlier well-established and long standing research activities as reported in [4–7].

13.3 The EU Built and Tested

The EU was built to test the designed circuitry.

Figure 13.2 gives a schematic capture of the circuit designed and Fig. 13.3 give the view of the test bench model built. The test board is powered by an AC power pack and commercially available CT and PTs are used as the sourcing sensors as can be seen in the Fig. 13.3.

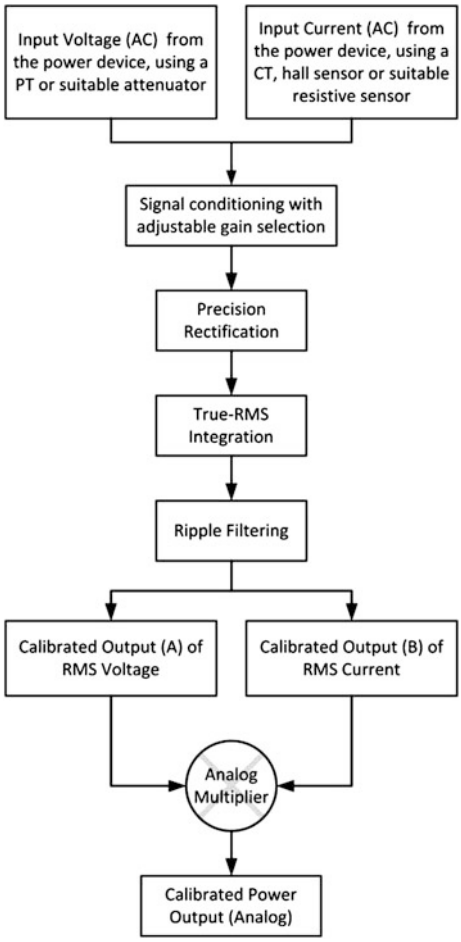


Fig. 13.1 Schematic block diagram of the design indicating the features involved

13.4 Salient Operating Features

The EU has functionally three distinguishable operations. Sequence of their occurence in the measurement process is as listed.

1. Capturing and conditioning of inputs from voltage and current sensors.
2. Integration of the conditioned signals with true RMS feature.
3. Analog processing for multiplication of the filtered RMS outputs from the above stage rendering it to be the measure of the power.

CTs and VTs (used in this testing) but could be changed to more The measurement of voltage and current, the components for the power measurement is drawn from signals from sensors like precision sensing devices like Hall sensors or resistive sensors for a more reliable and linear measurements.

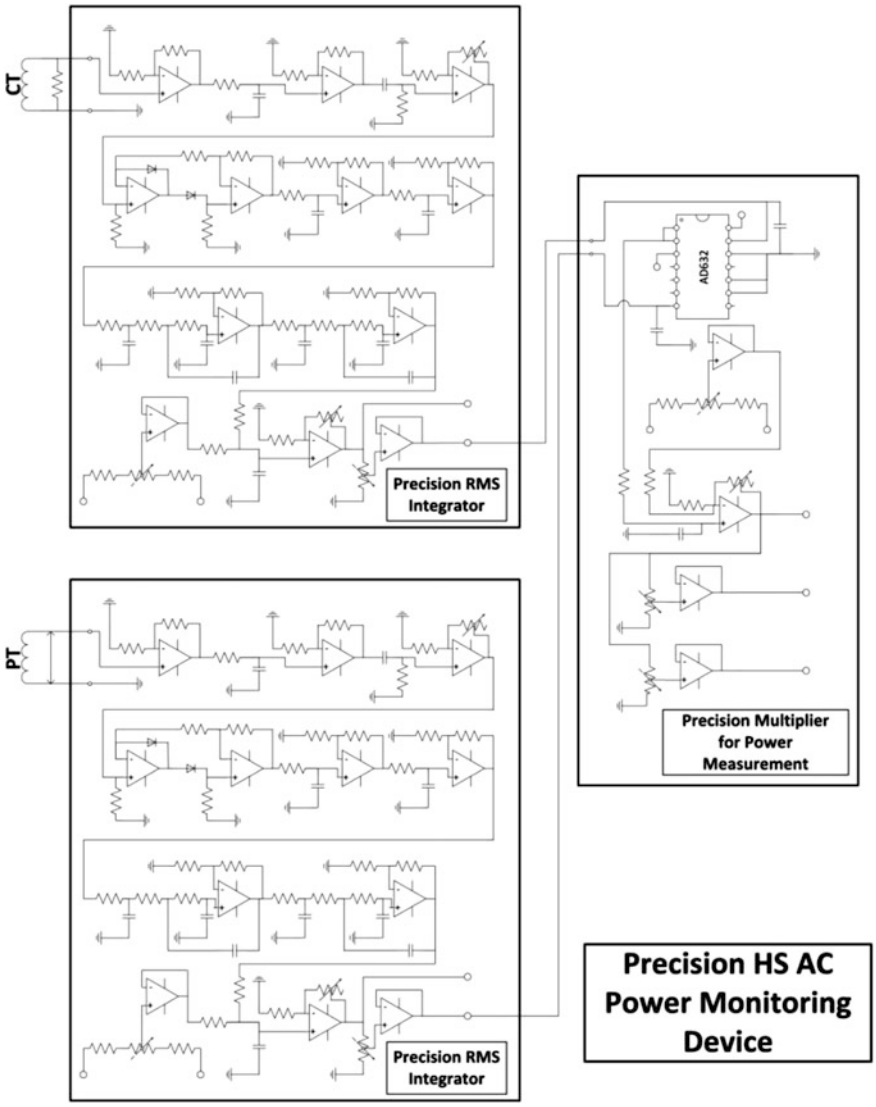


Fig. 13.2 Schematic capture of the circuit designed

The signals from the transformers are conditioned (gained and trimmed off with offsets, if any) in two independent channels allowing for finer corrections on the inputs. Care is taken to include a rectifier with low small signal error [8, 9] before passing the signal on to an integrator. The integrator section has two stage integration and the RC combinations used are to reduce the non-linearity during the integration. The integration stages are allowed to have significant ripple factor, inherently to retain a better dynamic response.

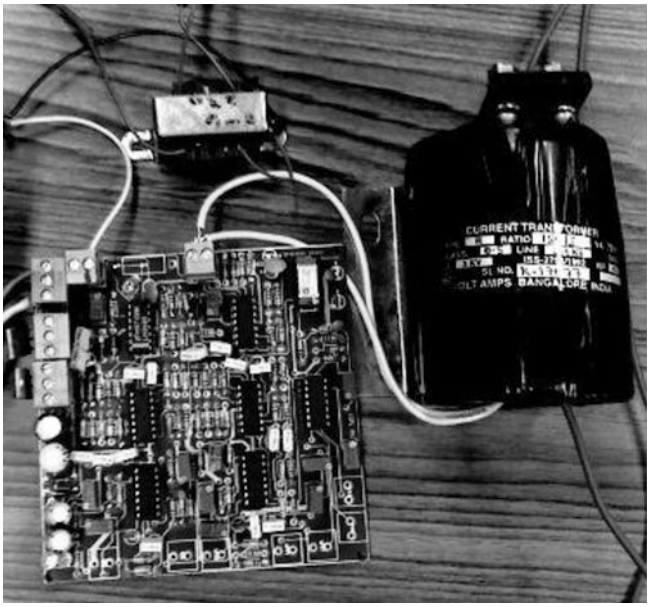


Fig. 13.3 Assembled unit on laboratory test

A cascaded two stage active low pass filters of third order Butterworth design [10, 11] is used in sequence to these integrators so as to have the signal passing through them get off with the ripples and still retain the dynamic response at their best. While it is known that integration of the ACs will impose a loss of dynamic response the aspects discussed above take adequate measures in limiting this loss to a practical lower limit, thereby enabling an achievable best dynamic response for the unit.

Low pass active filters are designed for a cut off center frequency of 25 Hz. This is center frequency is selected to see that the major frequency component of the ripples being at around 100 Hz (twice the line frequency due to the rectification) would be well be eliminated and still will be responsive enough to the input variations. In this consideration, the EU selectively responds to variation of power (the parameter measured) and not to the native AC components of the signal. It can be viewed that the design stands at an optimum point having a trade–off between the fastest transients captured over the allowable ripple factors in the output. The active filter provides a good enhancement in the achievable frequency with no loss of measurement accuracy. Fine tuning options are provided to render the usage of the EU offering wider range in measurement and in meeting larger requirements. This is to make the unit more field–worthy. Distinct sample points for a diagnostic data acquisition were used for characterizing and studying the performance of the EU.

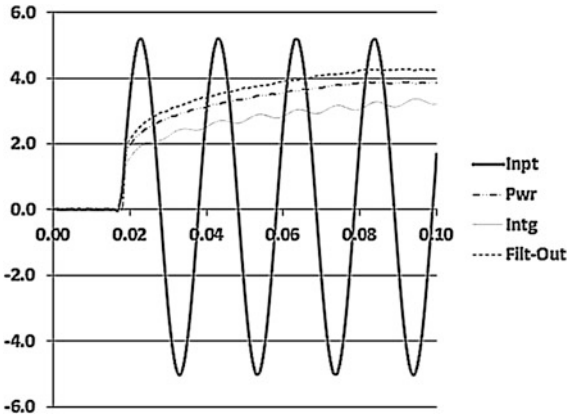


Fig. 13.4 Transient feature monitored as captured

13.5 Results and Discussions

The EU was tested for its compliance to the design and its performance has been validated with a set of field as well as simulated test conditions [12]. The diagnostic data at different stages of the circuit useful in evaluation and characterizations was captured. With the AC power signal running 50 Hz, a sampling interval of 1 ms was chosen to capture the transient responses at different test points with adequate time resolution. During the tests, a PC based data acquisition system with eight channels of acquisition at 12 bit resolution for the digital conversion was used.

The tests were conducted in two categories, one with simulated signals for the inputs so as to capture the response characteristics of the unit and to evaluate it for its compliance to the design features and the other set of tests in field condition (with the inputs received from CT and PT, as in Fig. 13.3) to validate its usability and reliability.

Simulated signals from a controllable analog function generator were used for the two AC input signals representing current and voltage components of the power being measured. This arrangement allowed for creating different test conditions and to get comprehensive performance characterization of the EU.

Figure 13.4 represents the capture of transient response of the integrator and active filters. It can be seen clearly that the ripple factors that have the dominating harmonics of the rectification peaks that runs at the double the native frequency of the signal are considerably high at a magnitude of about 5 % of the peak. This is filter off in the subsequent stage of active filters with virtually no loss of additional response, as one can infer from the figure. The response reaches to its asymptotic steady value for the step input in about 80 ms indicating that the large signal response to be easily crossing 10 Hz while the dynamic response for small signal variations would be about much higher than that.

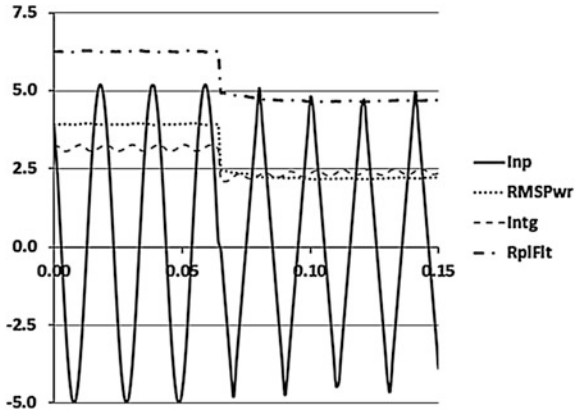


Fig. 13.5 Illustration of true RMS measurement

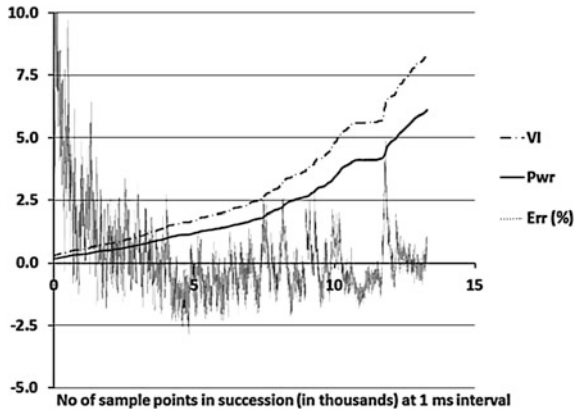


Fig. 13.6 History of estimated error in measurement

The response seen in Fig. 13.5 shows the behavior of the true RMS capture during the AC integration. In the test carried out a simulated train of signal that switches between sine to triangular waves are given at the inputs (both current and voltage) and the transients are captured as seen in the figure. One can clearly observe that for the amplitudes of the shifted waves are set nearly being same the integrator output and the final dc output of the unit have shifted swiftly on the changeover indicating its good characteristics of having true RMS capture. It can also be seen that the change in the wave forms has resulted in a transient settlement in less than 1 ms indicating that small signal dynamic response is easily better than 50 Hz, the native frequency of the AC signals. This shows one of the important design features established.

The Fig. 13.6 shows a history of estimated errors. Expectedly the errors near zero inputs for a rectifier is higher since zero crossing ideally cannot be captured. However, with the use of improved precision rectifier [13] this error is reduced to manageable low and the range of measurement is enhanced. The estimate of error is made by providing same simulated signals for both voltage and current and the error is

estimated as the difference between the calculated RMS–square of the input signals and the output. One can see that the error band comes down to less than 5 % when the inputs cross 10 % of the peak measurable range and much less than less than 1 % in the other zones. Taking off the stray noises in the plot, one can see that the unit provides a good response that goes with a nominal error band of 0.1–0.5 % and is expected to be good in the industrial standards for such measurements.

The analysis of the behavior shown up in the test indicates clearly that all the design features are realized and system is functioning satisfactorily. Subsequent these simulated tests, field tests made with an electrical loading panel have shown a good consistency and performance.

13.6 Concluding Remarks

Though AC power measurement is not a new concept and many alternatives for this measurement are available, fast responding devices with good precision for such measurements are not seen commonly. More distinctly analog based approaches for such measurements are not in practice. However, for a cost effective delivery of the criteria considered, analog designs can provide a good alternative as compared to the digital approaches. In the present effort, this line of thinking is pursued and a distinctly convincible alternative is arrived at.

It may be useful to note that this development was a result of a need felt for providing a power regulation of the test load for an engine at the laboratory for a critical transient measurement and the device built on this concept was successfully used for in controlling with the analog feedback derived from this device. Another quick reference for a good usage comes out in its high usability in managing excellently the load sharing of an engine in a set of synchronized engines. In this context, the concept presented is considered to have addressed the technology gap wherein normally available power measurement devices do not capture the transients well and are seldom useable for good control logic.

The design concept presented is built, tested and found to have met the design criteria and the analysis and characterization presents show the benefits of the design concept clearly that stand out as an alternative to relatively highly expensive digital devices that match to this class of performance. In the background of the benefits the design concept presented could offer, it is considered to be of a good contribution in the area of power monitoring and regulation.

References

1. Data sheet for small signal Schottky diode (1N 5711), STMicroelectronics August 1999 Ed1A. http://www.st.com/
2. Data sheet for LM136-2.5/LM236-2.5/LM336-2.5 V, National Semiconductor Corporation DS005715, 2005. http://www.national.com/

3. Data sheet for AD 632 (internally trimmed precision IC multiplier)—Analog Devices Inc., 1997. http://www.analog.com
4. Bin Md Idros MF, Hassan SFbA (2009) A design of Butterworth low pass filter's layout basideal filter approximation on the ideal filter approximation (published conference proceedings style). In: IEEE symposium on industrial electronics & applications (ISIEA) 2009, Kuala Lumpur, Malaysia
5. Vural RA, Yildirim T (2010) Component value selection for analog active filter using particle swarm optimization (published conference proceedings style). In: The 2nd international conference on computer and automation engineering (ICCAE) 2010, Singapore
6. Miao H, Liu X, Tan B, Shen J, Wang Y (2010) A universal active filter design method (published conference proceedings style). In: 2010 international conference on information networking and automation (ICINA), Kunming, China
7. Tow J (1969) A step-by-step active-filter design (periodical style—accepted for publication). Spectrum, IEEE, December 1969, pp 64–68
8. Djukic' S, Veskovic' M, Vulovic (2010) An improved precision full—wave rectifier for low-level signal (published conference proceedings style). 2010 9th international symposium on electronics and telecommunications (ISETC), Timisoara, Romania
9. Antoniou A (1979) Design of precision rectifiers with operational amplifiers. In: Proceedings of the Institution of Electrical Engineers IEEE, vol 121, No. 10. October 1979 (published conference proceedings style), pp 1041–1044
10. John Bishop, Bruce Trump and R. Mark Stitt (2012) FilterProTM MFB and Sallen-key low-pass filter design program (application report). SBFA001A-November 2001, Texas Instruments. http://www.ti.com/
11. Bruce C, Hruelsman LP (2012) Handbook of operational amplifier active RC network (application report). SB0A093A—October 2001, Texas Instruments. http://www.ti.com/
12. Kulkarni AP, Rajan NKS (2012) A precision high speed AC power monitoring device for power regulation and control, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2011, WCECS 2011, 19–21 October 2011, San Francisco, pp 748–752
13. Rod Elliot (ESP) (2010) Precision rectifiers (application report). AN-001, 18/12/2010. (http://sound.westhost.com/

Chapter 14
The Co-Design of Test Structure and Test Data Transfer Mode for 2D-Mesh NoC

Ying Zhang, Ning Wu and Fen Ge

Abstract NoC(Network-on-Chip) has been proposed as a new solution to deal with the global communication problem of complex SoC(System-on-Chip). However, there are many difficulties in testing and verification for NoC. We propose novel co-design of test architecture and data transfer schemes for 2D-Mesh topology NoC to improve the parallelism of test packets transmission. The testing efficiencies of different structures or transfer modes are evaluated under a coverage-driven and hierarchical NoC testbench, which is based on the VMM verification methodology and SystemVerilog language. The evaluation results of testing cost, testing time and hardware overhead show that the shortening of transmission path and parallel testing effectively decreases the power consumption and testing time. Furthermore, one of these test structures can be proved to the optimal scheme.

Keywords Data transfer · NoC · Parallel testing · Testbench · Test structure · VMM

Y. Zhang (✉) · N. Wu · F. Ge
The College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, 29# Yudao Street, College No.4, NUAA, 210016 Nanjing, Jiangsu Province, China
e-mail: tracy403@nuaa.edu.cn

N. Wu
e-mail: wunee@nuaa.edu.cn

F. Ge
e-mail: gefen@nuaa.edu.cn

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_14,

14.1 Introduction

With the development of the semiconductor industry and growing scale of the chip design, NoC (Network-on-Chip) will become the solution to the design of complicated system chip instead of traditional bus system chip SoC (System-on-Chip), which will also pose the greatest challenge to associated validation and testing. Furthermore, validation and testing are expected to account for more than 70 % of the whole design work for complex system chip [1]. There is urgent demand for novel testing methods or schemes aiming at testing complicated system chips, especially NoCs.

However, there are many difficulties in NoC testing. Such as the controllability and observability are poor for the network structure and communication scheme, GALS (Global Asynchronous Local Synchronous) method make it necessary for adopting various testing clocks and there is no standard benchmark for evaluation NoC testing, etc. As far as the first difficulty is concerned, the test structure of NoC should be distributed to cooperate with the communication architecture and decrease the hardware overhead. The test control components are inserted to the network and the basic strategy is to transmit test vectors to resources under test through tested components [2]. There are two steps of NoC testing, the first one is to test communication fabric when the router (switch) can be regarded as a IP core and we can apply the same test vectors to all the routers, the second one is to reuse the communication fabric as the test access to resources and transfer test data packets through particular testing input. The verification of communication fabric can be accomplished by the associated testbench [3–5], while the key problem of the second step is how to choose or configure the testing input.

Authors of [6] choose one switch to be connected to the source of test data (called Test Access Switch, TAS) and all the test data are broadcasted from TAS. Though the test scheme is prior to traditional bus-based SoCs in test time and power consumption, it needs complicated routing algorithm and may cause the increase of test time especially for large NoCs. The same situation may be worse when the TAS is the center of the network [7]. Because the center switch will be the hot-spot when all its neighbors communicate with it, the congestion will happen and greatly increase the test time. The solution to this problem is to increase the input/output port number so as to improve the efficiency of parallel testing. Sedghi et al. [8] apply two TASs, one is located on the lower left of the network and the other is at the upper right. The advantage in test time was showed, but the location of TASs can not change and specific routing algorithm is needed.

On the other side, the parallelism of test data transfer is needed to consider [9, 10]. Network communication technology of test data has been widely applied for NoC testing process [11–13]. The communication means can be divided into unicast (point-to-point packet transfer) and multicast (one-to-many packet transfer). Among them, multicast can efficiently improve the parallelism of data transmission, so how to apply multicast transfer mode to testing is worth studying. Fang Fang et al. [9] proposed a multicast paths testing method which

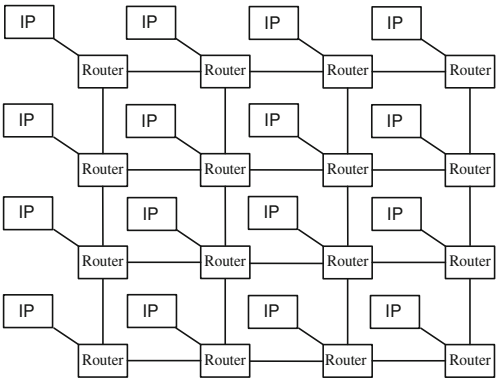


Fig. 14.1 4 × 4 Mesh network

modify multicast communication protocol for test and improve the testing parallelism based on Virtual Channels. However, the testing scheme is based on ordinary test structure.

We propose a novel test structure of 2D-Mesh NoC which has configurable TASs and applied with multicast transfer mode. It is greatly adaptable for parallel testing of routers and resources. The evaluation on the VMM-based testbench shows the reasonable test structure configuration is effective in improving testing efficiency and reducing testing cost.

Hereafter, Sect. 14.2 give a brief introduction of the NUT (NoC Under Test) and the associated testbench. Section 14.3 presents the co-design of NoC test structures and multicast test data transfer. Section 14.4 explains the results of performance evaluation for the test schemes. Section 14.5 draws the conclusions.

14.2 NOC Under Test and the Testbench

14.2.1 NoC Under Test

NoCs can be defined as a set of structured routers and point-to-point channels interconnecting IP cores (Resources). The topology of NoC can be represented as an undirected connected graph *G(N,L)*, where $N = \{n_1,n_2,\ldots n_i\}$ is the set of nodes and $L = \{l_1,l_2,\ldots l_j\}$ is the set of links in the corresponding network. The widely applied topologies are Mesh, Torus, Ring, Hypercube and Fat-tree [14]. Figure 14.1 shows a 4 × 4 2D-Mesh network, which is the structure of our NoC.

For regularity of its structure, Mesh network are easy to implement and have good scalability. Each node in Mesh network is connected to neighbors through regular grid point-to-point links.

The key component of the network fabric is communication node (router). Our router under test is realized in RTL description. The architecture of router is shown in Fig. 14.2. It includes SRAM, Cross Switches, Read/Write control, Routing

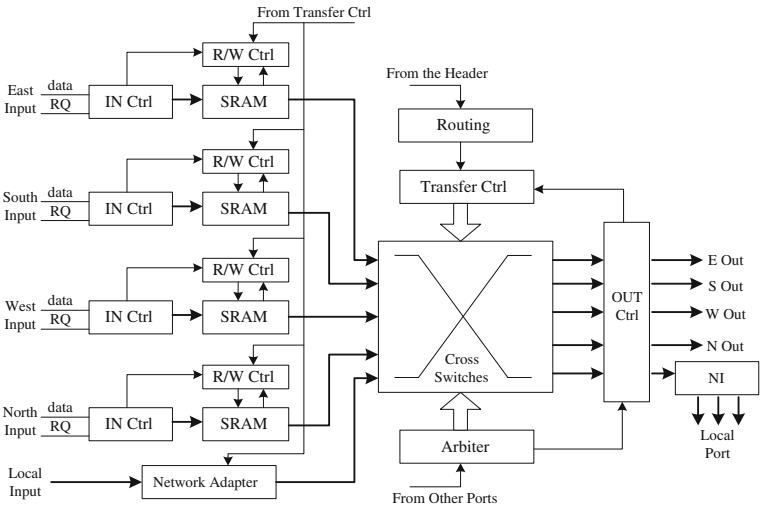


Fig. 14.2 Architecture of router under test

module, Arbiter, Transfer control, Input/Output control, and Network Interface (NI).

Each router has four input channels and four output channels and one network interface for providing communication service to resource nodes. The input control port will test the first packet, checking 32 bits data is in accordance with the parity, if parity checking is correct, data will be written into SRAM, otherwise it will be rejected and retransmission request will be sent to the upper router. Routing module obtains the destination address from the packet header and transfer data with the confirmation of Arbiter. Routing and arbitration module can be adjusted according to certain routing algorithms.

Our NoC employs XY routing algorithm, which is the most commonly used deterministic routing algorithm. XY routing algorithm is related to source node and destination node addresses while irrelevant to network status. It is suitable for 2D-Mesh topology and other similar topology NoCs. It is known to be deadlock-free in meshes (unicast mode) and easy to implement in hardware.

NoCs typically use the message communication model and messages are formed in packets. Each packet is composed of a header and some payloads [15]. The original data packets format of our NoC is shown in Fig. 14.3.

Packet_Type demonstrates the type of data packets, *Destination_ID* and *Source_ID* are destination and source address, *Packet_length* gives the total packet number of the message, *Reserve* field is for extension or user-defined function. *HEC* is used to validate the packet. *Payload* is the actual communication data.

The header of Multicast data packet has many flits, in addition to original unicast packet header information, it will also define the sorting of each destination node of the data in the whole multicast data packets. If the number of the

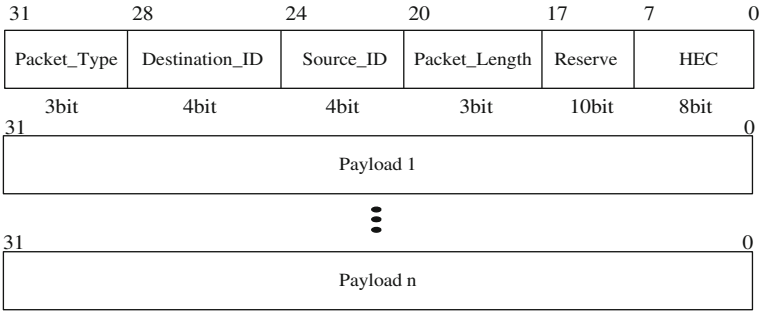


Fig. 14.3 Format of transmission data packets

destination nodes is more than two, the destination nodes information will be added to the header flits, at most 32 destination nodes address. The last flit of the header is the length of sub-packet information, which is the flits number of each sub-packet.

14.2.2 Coverage-Driven and VMM-Based Testbench

The hierarchical NoC testbench is based on the VMM methodology, which integrates assertions, abstraction, automation, and reuse mechanism to improve validation efficiency and productivity [16]. The testbench consists of Test Layer, Scenario Layer, Function Layer, Instruction Layer and NoC Under Test. The transmission between different layers is achieved through channel mechanism [17].

The test process contains four steps and they are generating random stimuli, establishing the testbench environment, executing the test and producing the verification report.

Each module of the testbench is realized in the form of certain class, the UML class diagram to explain their relationship is shown as Fig. 14.4.

The *Environment* class is the core of the NoC testbench. It is almost related to all the modules and controls the building, operating and ending of test process.

The *Config* class determines whether send packets to the specific node or not and the quantity of packets, etc. According to the configuration information, *RU_generator* is responsible for generating random packets with constraints and sending them to next layer through the channel.

The *Driver* receives packets from *RU_generator*, then sends them to NoC and transmits to the *Scoreboard* through callback at the same time. *Monitor* concentrates on output packets from NoC and transmit data to the *Scoreboard* and *Coverage* class. *Scoreboard* achieves the expected packets from *Driver*, gets actual packets from *Monitor*, then compares them and induces the result.

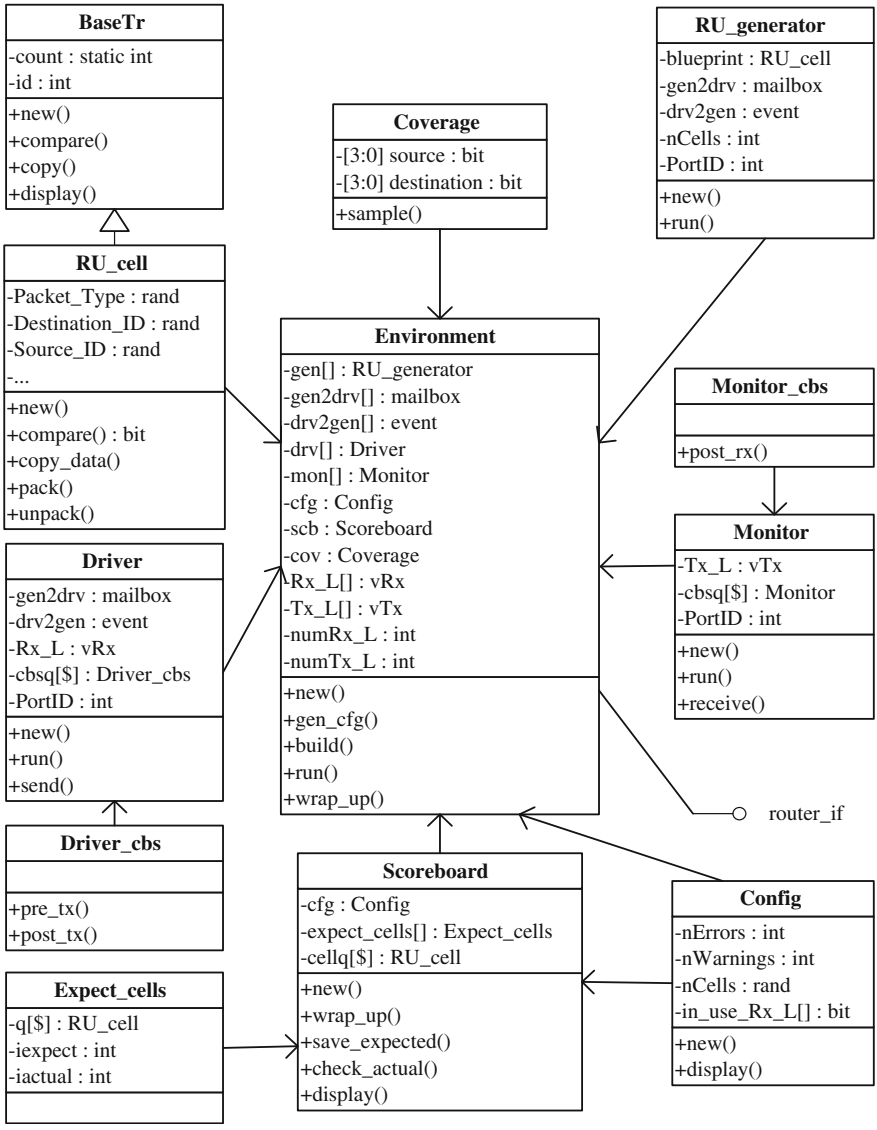


Fig. 14.4 UML class diagram of NoC testbench

The *Coverage* class defines the related cover point and can modify restraint based on the result of simulation. Function coverage is the important parameter to evaluate the efficiency of the testbench, which refers to the validated function of all functions percentage.

Adopting different random seeds and modifying design repeatedly, our testbench finally achieves 100 % function coverage. The testbench can accomplish

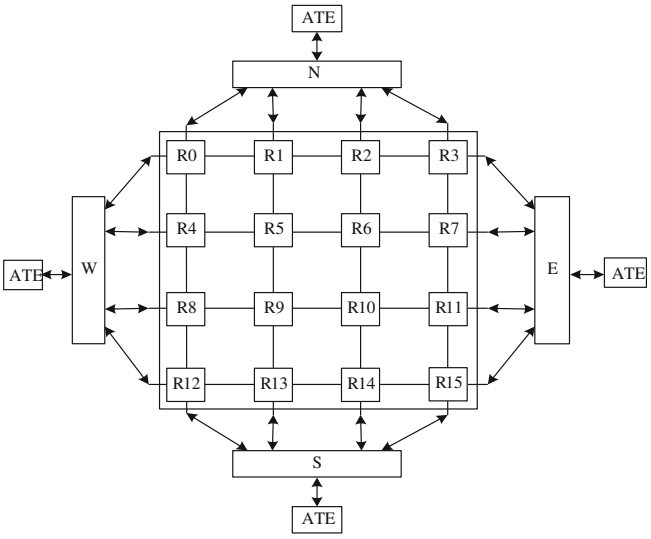


Fig. 14.5 Test structure of NoC with 16 TASs

function verification effectively with good adaptability and expansibility. Only provided the packets format and the transmission timing sequence are known, the testbench can be easily applied to different NoCs without the consideration of topology structure or routing algorithm.

14.3 Design of NoC Test Structures and Test Data Transfer Mode

We propose the test structure which adopts edge of Mesh as TASs, as shown in Fig. 14.5. ATE and peripheral circuits are connected in four directions, while TAS selection can be flexible with the configuration of related circuit. The structure has less test pins and can greatly shorten the transmission path of test data packets, which will be helpful to the parallel testing and decrease testing time as well.

The structure of periphery circuits is shown in Fig. 14.6. The input of periphery circuits comes from output ports of routers and input ports of ATE, while the output of periphery is connected to input ports of routers and output ports of the ATE.

Test data packets are transferred from the ATE and sent to certain router according to the input configuration. Test response packets are sent into the peripheral circuits and passed through selectors to output port of the ATE. Moreover, the arbitration mechanism is added to prevent the competition between routers.

In order to reduce the area overhead, we can configure partial peripheral circuits as shown in Fig. 14.7. Only two ports are selected in each direction. The number

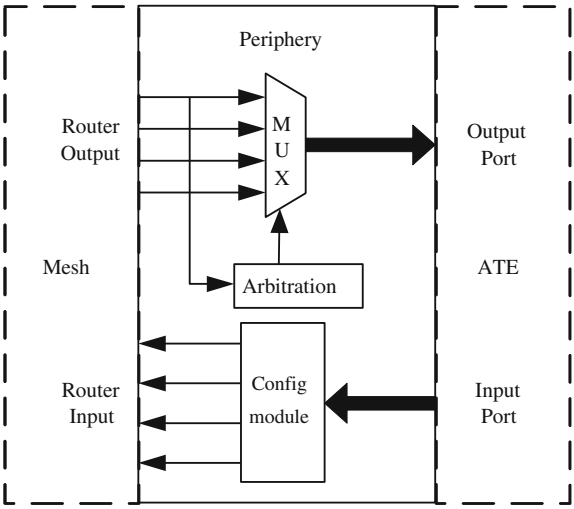


Fig. 14.6 Architecture of periphery circuits

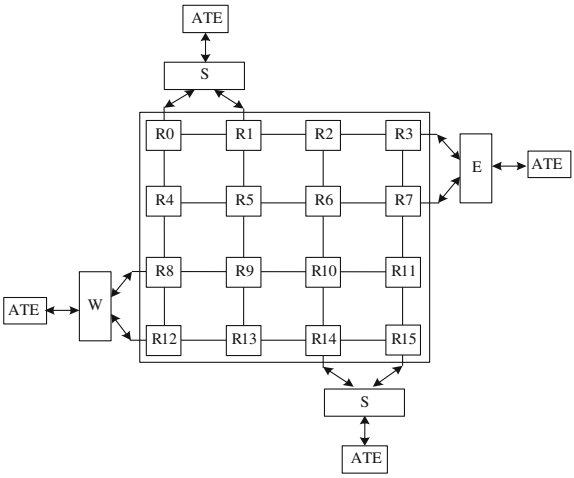


Fig. 14.7 Test structure of NoC with 8 TASs

and position of TASs can be adjusted according to the network size and the application characteristic.

The NoC test data transfer modes can be divided into unicast transmission mode and multicast mode. Unicast mode is sending packets from one port to other single port, while multicast is from one port to more than one ports. Compared with unicast mode, multicast has average shorter delay and smaller network bandwidth, which will improve the efficiency of the NoC testing. Moreover, its advantage will be more apparent when the number of transmission nodes increases.

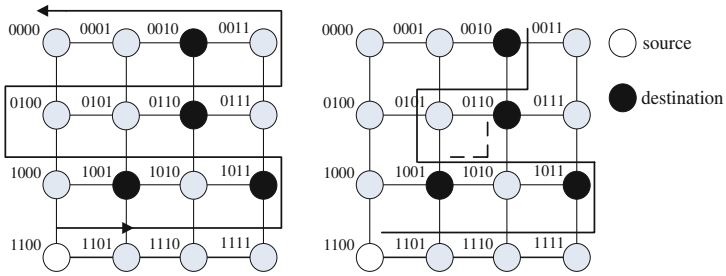


Fig. 14.8 Sorting method and transmission path when source node in four corners

However, there is a key problem need to be solved for multicast mode, that is deadlock avoiding. Provided that the transfer information is test data, we proposed the deadlock-free routing schemes as Figs. 14.8 and 14.9 show.

As the design of test structure is described, source node is located in the external 12 router of NoC. When the source node belongs to one of the four corners, shown in Fig. 14.8, starting from the source node, draw a line cross all the routers, the first destination node is the last multicast node while the last destination node is the first multicast one. The destination node sequence for Fig. 14.8 is 0010, 0110, 1001, and 1011.

When the source node is the other eight edge routers, the line starts from source node, connected to the four corners of the router firstly and then line to the destination node in the same way as the source nodes is the four corners, the sorting mode is shown in Fig. 14.9.

Furthermore, XY routing algorithm also needs to be improved when multicast transfer mode is applied. The rules are as the follows.

(1) The line turning to the 180° angel is not allowed;
(2) When data packets are sent to the last destination node in one row, if the next destination node of the X direction underside the line in the current node (as shown in Fig. 14.8, destination node 1001 is on the next column of 0110), the packet will send along the Y direction to next node;
(3) For source nodes not located in four corners of the router, packet is transmitted to nearest source node in the four corner routers.

After the improvement of destination node sorting and routing algorithm, multicast transmission path become a forward line, which will never repeat the path has taken, so as to efficiently avoid the deadlock.

14.4 Performance Evaluation

For 4 × 4 Mesh network, Fig. 14.10 give four testing schemes with different TAS configuration and each resource chooses the nearest TAS for test vectors transmission.

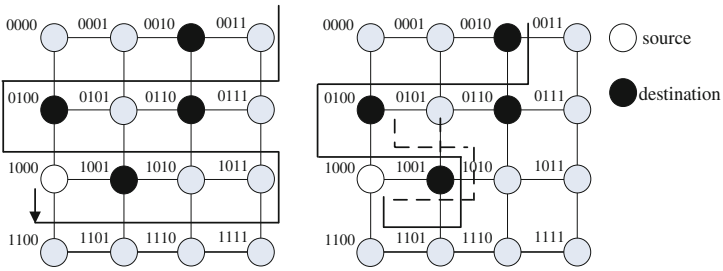


Fig. 14.9 Sorting method and transmission path when source node not in four corners

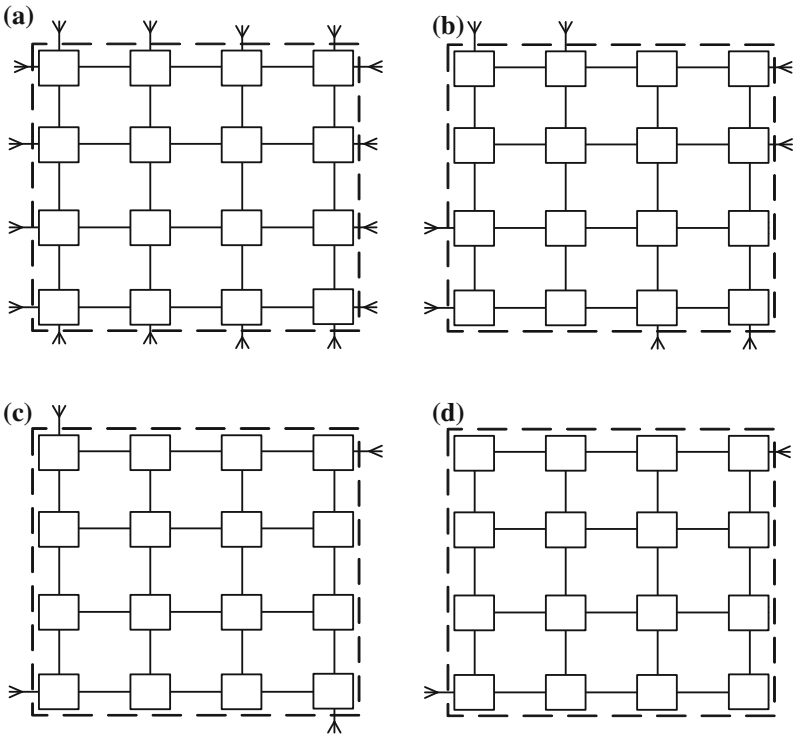


Fig. 14.10 Schemes of different test ports configuration. **a** 16 test ports. **b** 8 test ports. **c** 4 test ports. **d** 2 test ports

How to choose TASs greatly affects test cost, test time and corresponding synthesis area. The test cost mainly refers to test power, which is relevant to the number of test packets and the length of transmission path [18, 19]. The test time can be evaluated by the modified VMM-based testbench and the area overhead of Scheme (a) and Scheme (b) are calculated and compared with the original Mesh NoC on Synopsys EDA platform.

Table 14.1 Comparison of test costs

Scheme	$N_{Ri} = 1$	$N_{Ri} = 2$	$N_{Ri} = 3$	$N_{Ri} = 4$	Sum
(a)	12	4	0	0	20
(b)	8	8	0	0	24
(c)	4	8	4	0	32
(d)	2	4	6	4	44

14.4.1 Estimation of Test Cost

For simplifying and standardizing the estimation of test cost, we assume that the number of test data packets for each resource to be constant C. N_{Ri} and N_{Ci} are respectively the number of routers and channels in the test path for $Resource_i$ and $N_{Ci} = N_{Ri} - 1$. T_R and T_C are respectively the average cost of test packets passing through one router and through one channel. n is the total number of resources, so the total test costs T_{all} can be calculated as follows.

$$\begin{aligned} T_{all} &= \sum_{i=1}^{n} (N_{Ri} * T_R + N_{Ci} * T_C) * C \\ &= \sum_{i=1}^{n} [(N_{Ri} * T_R + (N_{Ri} - 1) * T_C] * C \\ &= \sum_{i=1}^{n} [(T_R + T_C) * N_{Ri} - T_C] * C \end{aligned} \tag{14.1}$$

Most parameters in equation (14.1) are constants except N_{Ri}, so the total test cost can be measured by $\sum_{i=1}^{n} N_{Ri}$, that is the number of routers in test paths for all resources. Test costs of four schemes in Fig. 14.10 are shown as Table 14.1.

It is evident that the test cost increases with the decrease of TASs and obviously Scheme (c) and scheme (d) have too high test costs to be good testing schemes. Scheme (a) has lowest test cost but 16 TASs will lead to unacceptable size overhead. Compared with Scheme (a), Scheme (b) has a little increase on test cost and much decrease on test ports, so it should be optimal scheme. Reasonable choice of the position and number of TASs is extremely important especially when the scale of network increases.

14.4.2 Evaluation of Test Time

For the function verification, the driver and monitor module of the testbench are connected to each router to verify the communication between resources. However, for the testing, the driver and monitor module should be connected to periphery circuits to make the original test platform equivalent to the ATE.

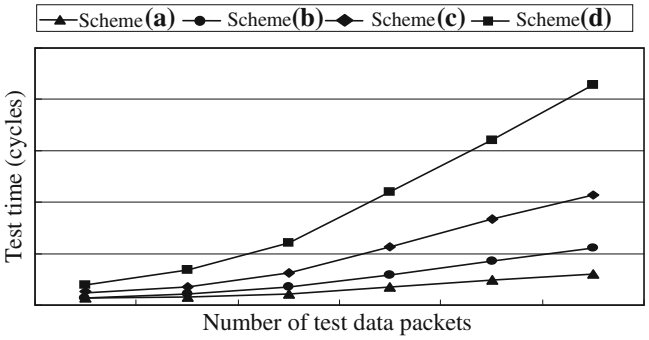


Fig. 14.11 Test time comparison

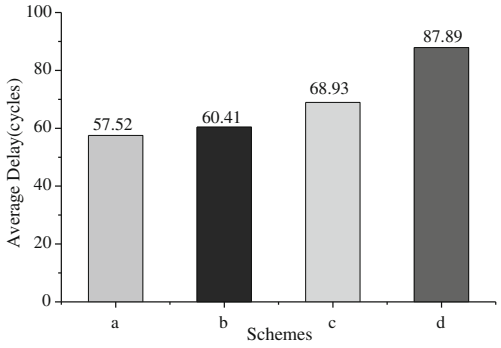


Fig. 14.12 Average delay time comparison

The test time of four testing schemes is shown in Fig. 14.11. X axis is the number of test data packets received by each resource node and Y axis is the time when the last packet is received. When packets number is small, there is no obvious different in test time for all the schemes, however, with increase in packets number, the test time of four schemes gradually increase and the difference between them also becomes increasing. The increase of TASs greatly enhances testing parallelism and effectively reduces the test time.

The average delay time between sending test data packets and receiving response packets is shown as Fig. 14.12. With the decrease of TASs, the average delay also increases gradually, because average path from TAS to resource gets longer. However the difference between Scheme (a) and Scheme (b) is not obvious.

Based on the analysis of the test time with variable packets number and the statics of test packet transmission time, we can conclude that test time is mainly affected by the number of test ports, especially when the amount of communication information increases greatly.

Table 14.2 Area overhead comparison

	Original Mesh (μm^2)	Scheme(a) (μm^2)	Ratio of increase	Scheme(b) (μm^2)	Ratio of increase
Combination logic	959583.24	972296.74	1.325 %	964227.32	0.484 %
Sequence logic	2134660.71	2156475.23	1.022 %	2145984.61	0.530 %
Interconnect	30881190.62	31257792.38	1.220 %	31008210.76	0.411 %
Total size	33975434.56	34386564.36	1.210 %	34118422.69	0.421 %

14.4.3 Evaluation of Area Overhead

For the advantage in test cost and test time, the Scheme (a) and Scheme (b) were synthesized and optimized based on Synopsys DC Compiler and SMIC 0.18 CMOS technology library. Table 14.2 gives the area overhead of original Mesh network and the Scheme (a) and (b). The ratio of increase is relative to the synthesized area of original Mesh network.

The majority of synthesized areas are interconnection lines, so the increase of TASs will surely increase the area overhead. The results of Table 14.2 show the Mesh network with additional test structure has bigger size than original Mesh network and Scheme (b) is superior to Scheme (a) on size overhead.

The testability design can reduce the testing time, but it would increase the cost of the area. We manage to find the strategy to make balance between two aspects. This paper puts forward four testing schemes, Scheme (b) is the best one which has relatively shorter test time and delay time, small synthesized area and its test costs is modest.

14.5 Conclusion and Future Work

We proposed a configurable co-design of test structure and test data transfer mode for 2D-Mesh NoC and given four practical schemes with different TASs configuration. The evaluation platform is the coverage-driven and VMM-based testbench which is originally designed for communication verification on NoC. By adjusting the packet format and routing mechanism, the testbench is applied to evaluating the performance of different testing schemes. Combined with the circuit synthesis results, the optimal scheme can be verified. The experimental results showed that the shortening of transmission path and parallel testing effectively decreases the test cost and test time.

We will research on the optimization of test schedule algorithm based on the designed test structure later.

Acknowledgments This work was supported by the Natural Science Foundation of China under Grant 61076019 and 61106018, the Aeronautical Science Foundation of China under Grant 20115552031, the China Postdoctoral Science Foundation under Grant 20100481134.

References

1. Chris Spear (2006) SystemVerilog for Verification, Synopsys, Inc
2. Grecu C, Pande P, Wang B et al. (2005) Methodologies and algorithms for testing switch-based NoC interconnects. In: Proceedings of the 20th. DFT conf. defect and fault tolerance in VLSI systems, Montery(CA), pp 238–246
3. Li Y, Liang L (2010) An NoC modeling and simulating method with systemC, microelectronic & computer, China, 27(3):78-82
4. Banerjee N, Vellanki P, Chaha KS (2004) "A power and performance model for Network-on-Chip architectures." In: Proceedings of the DATE Conf. France, pp1250–1255
5. Kreku J, Hoppari M, Kestilä T (2008) Application-platform performance modeling and evaluation. In Proceedings of the forum on FDL conf. specification, verification and design languages, Sttargart, pp 43–48
6. Hosseinabady M, Banaiyan A, Bojnordi MN et al. (2006) A concurrent testing method for NoC switches. In Proceedings of the design automation & test in Europe conf. vol.1, Munich, p 244
7. Zhang F (2007) Design-for-testability optimal strategy and design-for-testability research based on network-on-chip, M.S. thesis, Dept. Electron. Eng., Tsinghua Univ., China
8. Sedghi M, Koopahi E, Alaghi A et al. (2008) An NoC test strategy based on flooding with power, test time and coverage considerations. In: Proceedings of the VLSI design conf. Hyderabad, pp 409–414
9. Fang F, Dong J, Han Y et al. (2010) Parallel testing method for NoC using multicast path. J Electron Meas Instrum. 4(10):911–917
10. Giles G, Wang J, Sehgal A et al. (2008) Test access mechanism for multiple identical cores. In: Proceedings of the international test conference (ITC), pp 1–10
11. OuYang YM, Feng W, Liang HG (2008) An optimized test ports selecting method under power constraint in NoC. J Comput Appl. 28(4):1026–1028
12. Li J, Xu Q, HU Y, et al. (2008) Channel width utilization improvement in testing NoC-based systems for test time reduction. In: Proceedings of electronic design, test and applications, 4th IEEE International Symposium 2008, pp 26–31
13. Zhao JW, Shi YB, Wang Zh G (2009) Research on test strategy for hierarchical network-on-chips interconnection infrastructure. J Electron Meas Instrum. 23(5):34–39
14. Tayan O (2009) Networks-on-chip: challenges, trends and mechanisms for enhancements. In Proceedings of the ICICT'09 conf. information and communication technologies, Karachi, pp 57–62
15. Zhang Y, Wu N, Ge F (2011) Novel test structures for 2D-Mesh NoC with evaluation on the coverage-driven&VMM-based testbench, Lecture Notes in Engineering and Computer Science. In: Proceedings of the world congress on engineering, WCECS 2011, 19–21 October, 2011, San Francisco, USA, pp 797–801
16. Bergeron J, Cerny E, Hunter A, Nightingale A (2006) Verification methodology manual for system verilog, Synopsys & ARM Inc
17. Mao Ye (2008) Research and implementation of VMM-based verification platform, M.S. thesis, Dept. Electron. Eng., Huazhong University of Science and Technology
18. Cota E, Carro L, Wagner F et al. (2002) Power-aware noc reuse on the testing of core-basedsystems. In Proceedings of the ITC'02 Conf. Baltimore, pp 612–621
19. Zeferino CA, Susin AA (2003) SoCIN: a parametric and scalable network-on-chip. In Proceedings of the intergrated circuits and systems design (SBCCI)'03 conf. Sao Paulo, pp 169–174

Chapter 15
Overhead- and Performance-Aware Fault-Tolerant Architecture for Application-Specific Network-on-Chip

Fathollah Karimi Koupaei, Ahmad Khademzadeh and Majid Janidarmian

Abstract Defect in manufacturing of integrated circuits is almost inevitable, and fast scaling in technology has caused the components of a Network-on-Chip (NoC) to be more susceptible to faults. Therefore, it is crucial to sustain chip production yield and reliable operation in the presence of defects. The permanent faults are a consequence of manufacturing defects that occur during fabrication or aging defects that occur during system lifetime. A fault-tolerant application-specific NoC should be able to detect a fault and recover the system to correctly operate the mapped application. In this paper, a fault-tolerant NoC architecture designed in VHDL and synthesized using Xilinx ISE is presented which not only is able to recover from single permanent router failure, but also improves the average response time of the system in the different traffic loads. As hardware overhead is a major issue while considering fault tolerance, a new component, called Link Interface (LI) is also developed to reduce cost overhead. The Video Object Plan Decoder (VOPD) and MPEG-4 core graphs are used as two real applications in this study.

F. Karimi Koupaei (✉)
CE Department, Arak Branch of Azad University, Arak, Iran
e-mail: karimi.fathollah@gmail.com

A. Khademzadeh
Iran Telecommunication Research Center, Tehran, Iran
e-mail: zadeh@itrc.ac.ir

M. Janidarmian
CE Department, Science and Research Branch of Azad University, Tehran, Iran
e-mail: jani@ieee.org

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_15,

Keywords Application-specific network-on-chip · Deadlock-free routing algorithm · Fault-tolerant design · Link interface · Mapping algorithm · Permanent failure

15.1 Introduction

The number of processor, memory and accelerator cores on systems-on-chip is rapidly increasing to support evolving standards and new applications. Computation and communication complexity is skyrocketing, and scalability-centric design paradigms are critically needed [1]. Networks-on-Chip (NoCs) have emerged as the best alternative to provide high performance in communication for futures Systems-on-Chip (SoCs) with dozens of cores integrated on a single silicon die. Mapping an application to on-chip network is the first and the most important step in the design flow as it will dominate the overall performance and cost [2]. Several approaches have been proposed in literature in the context of topological mapping in NoCs [3]. Mapping algorithms are mostly focused on 2D mesh topology which is the most popular topology in NoC design due to its layout efficiency, good electrical properties and simplicity in addressing on-chip resources. Another concern in NoC implementation is selecting an efficient routing strategy while providing freedom from deadlock. The routing algorithm determines the path that each packet follows between a source–destination pair. In the future chip generations, faults will appear with increasing probability due to the susceptibility of shrinking feature sizes to process variability, age-related degradation, crosstalk, and single-event upsets. To sustain chip production yield and reliable operation, very large numbers of faults will have to be tolerated [4, 5]. This argument strengthens the notion that chips need to be designed with some level of built-in fault tolerance. Furthermore, relaxing the requirement of 100 % correctness in the operation of various components and on-chip channels profoundly reduces the manufacturing cost as well as cost incurred by test and verification [6].

This paper is an extension of [7], and the remainder of this paper is organized as follows: in Sect. 15.2, an overview of some fault-tolerant research efforts in NoC is given. Section 15.3 illustrates the basic concepts of application-specific NoC design, and a new fault-tolerant architecture is introduced in Sect. 15.4. Simulation results i.e. average response time and hardware overhead will be presented in Sect. 15.5 followed by the concluding remarks in Sect. 15.6.

15.2 Related Work

Scaling of interconnects exacerbates the already challenging reliability of on-chip networks. As process technology scales, the integration of billions of transistors comes with an increased likelihood of failures. Smaller dimension circuits are

getting more sensitive to a particle strike, increasing the probability that the charge due to a high-energy particle flips the value in a cell [8]. With technology trends in device scaling, high clock frequencies, and supply voltage decreases, fault rates are increasing, which makes a reliable design a real challenge. Transient and Permanent errors are two main kinds of errors which generally occur in an NoC. The most common transient error recovery technique is the retransmission mechanism following error detection techniques like coding [9]. Many recently developed solutions focus on methods keeping the system working in spite the fact that some parts of the system are shut down [10]. Permanent faults in NoCs due to fabrication challenges in sub-65 nm CMOS technologies and due to wear-out underscore the need for fault-tolerant design [11]. Fault-tolerant routing algorithms are recently investigated to bypass the failed hardware. The inherent redundancy in NoCs due to multiple paths between packet sources and sinks can greatly improve communication fault resiliency [11]. Fault-tolerant routing algorithms should be able to find a path from source to destination in presence of the faults in NoC with a certain degree of tolerance [12]. Many algorithms in this area have been proposed which follow their own optimization aims. To name just a few, in [11], the authors propose a novel low-overhead neighbor aware, turn model based fault-tolerant routing scheme (NARCO) for NoCs which combines threshold-based replication in network interfaces, a parameterizable region-based neighbor awareness in routers, and the odd–even and inverted odd–even turn models. Shi et al. [13] presents the scalable and fault-tolerant distributed routing (SFDR) mechanism. It supports three routing modes including corner-chains routing, boundary-chains routing and fault-ring routing. Inspired by divide-and-conquer concept, system is partitioned into nine regions. Each region promises fault-tolerance of one's own when packets bounded into its area to guarantee total fault tolerance. The main problem with the fault-tolerant routings is that if a router fails, considering mesh architecture, recovery cannot be accomplished only by rerouting. In addition, hardware redundancy is inevitable in order to repair the lost connection to the network of the core directly connected to the failed router. A fault-tolerant mesh-based NoC architecture with the ability of recovering from single permanent failure is presented in [14]. This method adds a redundant link between each core and one of its neighboring routers, resulting in significant improvement in reliability while has little impact on performance. In this architecture, only one spare router should be selected among all possible alternative ones. This has an influential effect on overall performance in terms of the average response time and reliability of the system. Regarding to this work, in [15] a new fast and optimum algorithm based on performance measurement and extra communication cost is proposed to find the best configuration that also results in a more reliable system. It also shows that mapping algorithm has a great impact on mentioned parameters. Following this concept, in this paper, a hardware and performance-aware design for the fault-tolerant NoC architecture is presented which takes into account the specific application mapped onto mesh topology.

15.3 Prerequisites of Application-Specific NOC Design

15.3.1 Mapping Problem

To formulate mapping problem in a more formal way, we need to first introduce the following two concepts borrowed from [16]:

Definition 1 The core graph is a directional graph $G(V, E)$ whose each vertex $v_i \in V$ shows a core, and a directional edge $e_{i,j} \in E$ illustrates connection between v_i and v_j. The weight of $e_{i,j}$ that is shown as $comm_{i,j}$, represents the communication volume from v_i to v_j . The IP core along with a router connected to it by Network Interface (NI) is displayed as a tile.

Definition 2 The NoC architecture graph is a directional graph $A(T, L)$, whose each vertex $t_i \in T$ represents a tile in the NoC architecture, and its directional edge that is shown by $l_{i,j} \in L$ shows a physical link from t_i to t_j.

The core graph mapping $G(V, E)$ on NoC architecture graph $A(T, L)$ is defined by a one to one mapping function.

$$map : V \rightarrow T, s.t.map(v_i) = t_j, \forall vi \in V, \exists t_j \in T, |V| \leq |T|$$

Mapping algorithms are mostly focused on 2D mesh topology (Architecture Graph) which is the most popular topology in NoC design due to its suitability for on-chip implementation and low cost. The definitions are presented in Fig. 15.1.

15.3.2 Routing Algorithm

The routing algorithm determines the path that each packet follows between a source-destination pair. Routing algorithms noticeably affect the cost and performance of NoC parameters i.e. area, power consumption and average message latency [17]. Due to determined sources and destinations in application-specific NoC, minimum-distance routing algorithms are mostly considered in this area which the routing is computed off-line and admissible paths stored into the routing tables.

In general, every routing algorithm should include deadlock freedom feature [18]. So channel dependency graphs (CDG) is used to avoid any possible deadlocks. The CDG is a directed graph with the network channels as the vertices and the direct dependencies between the two channels as the edges. A dependency exists between the links $l_{i,j}$ and $l_{j,k}$ whenever there is a path to route packets from v_i to v_k through v_j which uses those links. An extension to CDG as a sub graph is the concept of application specific channel dependency graph (ASCDG) introduced in [19]. The ASCDG is a sub graph of the CDG and an edge in CDG between channels, $l_{i,j}$ and $l_{j,k}$ is removed if there was no application-specific dependency

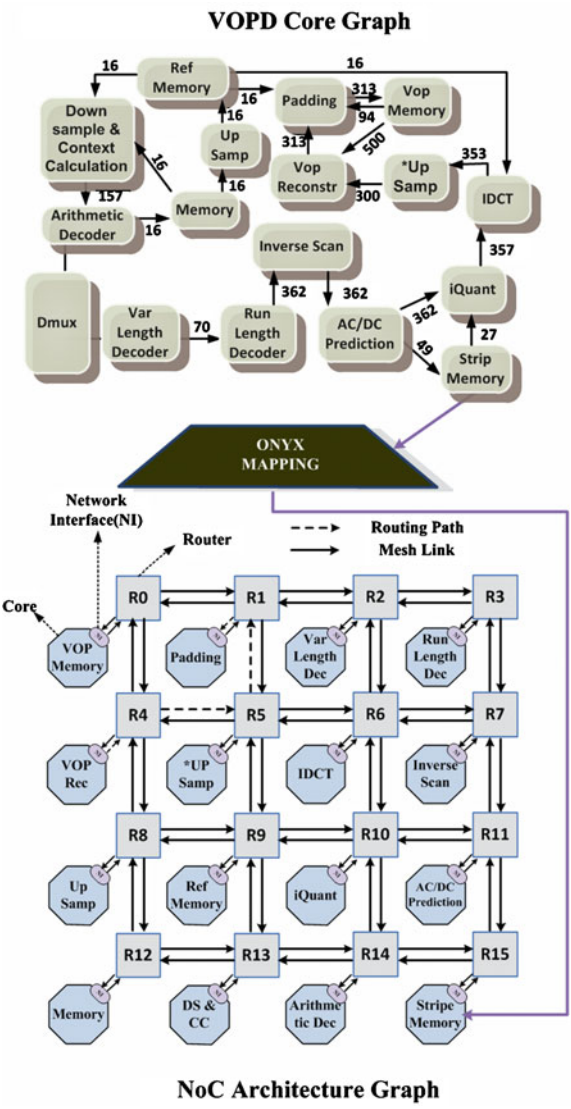


Fig. 15.1 Mapping problem concepts

between $l_{i,j}$ and $l_{j,k}$. Deadlock is inevitable when there are any cycles through ASCDG graph. A cycle in the ASCDG is a succession of application specific direct dependencies, $D = \{d_1, d_2, \cdots, d_n\}$, where $d \in D$ is a pair $(l_{i,j}, l_{j,k})$ with $l_{i,j}, l_{j,k} \in L$. If there is any cycle, we need to break it. By removing a dependency, all of the corresponding paths to that dependency will be removed. Using this method guarantees that routing algorithm is still deadlock-free. It is worth noting that using deterministic routing algorithm and efficient mapping algorithm, a few existing cycles can be easily broken.

Although the proposed methodology is topology and application agnostic, the state-of-art mapping algorithm proposed in [18] is used to map Video Object Plan Decoder (VOPD) cores onto mesh topology. We have also used two minimum-distance routing algorithms in mesh i.e. XY and YX. In XY (YX) routing algorithm, a packet is routed first in the X(Y) direction and then along the perpendicular Y(X) dimension. Because of using both algorithms together to route and reroute the packets, deadlock problem which is easily solved by described ASCDG graph should be taken into consideration.

15.4 The Proposed Fault-Tolerant Application-Specific Architecture

In the mesh-based architecture which is the simplest and most dominant topology for today's regular network-on-chips, each core is connected to a single router as depicted in Fig. 15.1. If a failure occurs in a router in this topology, the failed router cannot be used any more for routing packets and the directly connected core obviously loses its communication with the network, so the expected requirements of the mapped application are not satisfied and the whole system breaks down. As shown in Fig. 15.1, each router consists of five identical input/output ports and each port is a bidirectional link with a circular FIFO on its input side. In order to recover the inaccessibility of the core, a fault-tolerant architecture which each core is connected to two routers i.e. main and spare have been proposed [14] shown in Fig. 15.2. The spared router is used while a permanent fault is detected in the main one and in this situation; average response time of the system degrades due to rerouting phase. All essential modifications in Network Interface (NI), routers and packets were explained in [14]. The authors also supposed that fault detection is done by self-testing facilities embedded in each router which is responsible for continuously testing its operation. A router can inform the neighbors about its faulty status by setting a fault_status flag. This flag is checked by the neighboring routers and cores before starting any communication with the router.

Adding one port to all non-edge routers in this architecture leads to a great waste of hardware. Although it is able to tolerate one permanent router failure, it does not exploit the full potential of the architecture. However, in this paper the provided path diversities between adjacent routers are used to improve performance of the system and a new component called Link Interface (LI) instead of router port is developed to reduce hardware overhead.

According to the mentioned architecture shown in Fig. 15.2, after mapping an application onto mesh topology, an efficient spare router selection should be considered to find the best spare router for any core. To minimize the hardware cost, only one of the possible spare routers for a core is chosen [14]. On the other hand, with respect to the possible places for each core as shown in Fig. 15.2, there are two constraints to select a spare router.

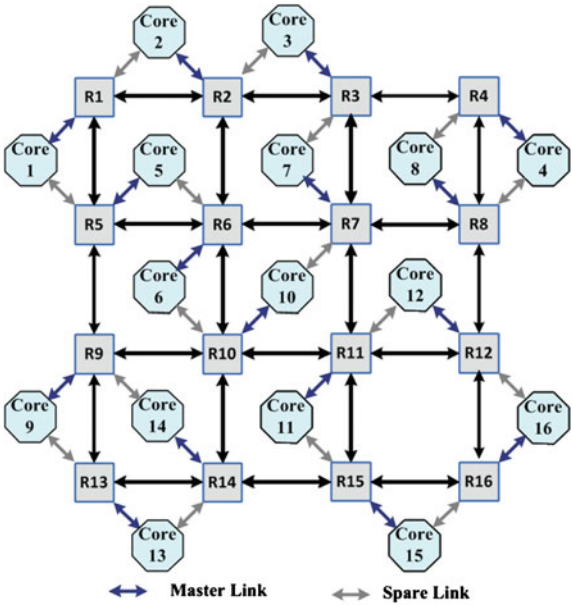


Fig. 15.2 The proposed architecture in [14]

1. Each router is limited to be linked as a spare by only one core.
2. Each core is located in the neighborhood of its local and adjacent routers and all cores should be placed in different locations.

The spare router selection algorithm in [15] called FERNA results in better average response time, extra communication cost and system reliability than greedy algorithm and has a polynomial time complexity. Since both main and spare routers can be used at the same time in the developed solution, we need to modify FERNA algorithm with regards to more routing path opportunities (explained in details in Fig. 15.3).

As you can see in Fig. 15.4, in the proposed architecture, each core is connected to its router via main local port and to the links using Link Interface via backup local port. The architecture of LI will be discussed in the following subsection.

In this architecture, if both main and spare routers are working properly, the best (minimum-distance) paths to send and receive packets are derived from path diversities and if main (spare) is faulty, spare (main) will be responsible to transfer packets through rerouting paths. Because this architecture is supposed to recover from only single permanent failure, all the best and rerouting paths are easily found while are deadlock-free by applying the ASCGD concept. For example, the ASCDG graph has been drawn in Fig. 15.5 while all routers are working properly. All admissible paths are offline stored into the routing tables and used with regards to routers conditions.

```
Initialize (G (E, V));
For (All Routers)
   Router_Unused[i] =1;
Do
{
   Selected_Core=Find_Max_Comm (G (E, V));
   Available_Palces=Find_All_Available_Places (Selected_Core);
   K=1;
   For (All Neighbor Routers)
      If (Router_Unused[i] =1)
      {
         Attach (Router[i], Selected_Core (Backup_Port));
         Response_Time[k] =Calculate_Response_Time (Architecture);
         Detach (Router[i], Selected_Core (Backup_Port));
         K=K=1;
      }
   Selected_Router=Find_Best_Neighbor_Router (Response_Time []);
   If (Selected_Router_Unused_Port=True)  // Edge Routers
         {
      Attach (Router[i], Selected_core (Backup_Port));
      Router_Unused [Selected_Router] =0;
         }
   Else //Non-Edge Router
            Attach (Link_Interface, Selected_Core (Backup_Port));
   Update_Available_Places;
} While (Find Spare Router For All Cores);
```

Fig. 15.3 The pseudo code of spare router selection algorithm

15.4.1 Link Interface

In the previous design [14], it is necessary to add one port to all non-edge routers resulting in much hardware redundancy. In order to reduce the overhead, in this section a Link Interface is suggested. After entrance of header flit into this module, destination address is decoded. As an example, if the address shows that connected core is the destination, header and its following flits will be sent to backup network interface of the core, otherwise they are routed to another output towards next router.

This module has been implemented with three processes which run concurrently; therefore it is able to transmit three dataflow as shown in Fig. 15.6. This module has been also designed without using clock pulse that leads us to achieve better response time and power consumption. To this end, as soon as data_ready line is activated; the input port will run its process of management and data control flow to guide flits towards output port. It is worth noting that if two input ports

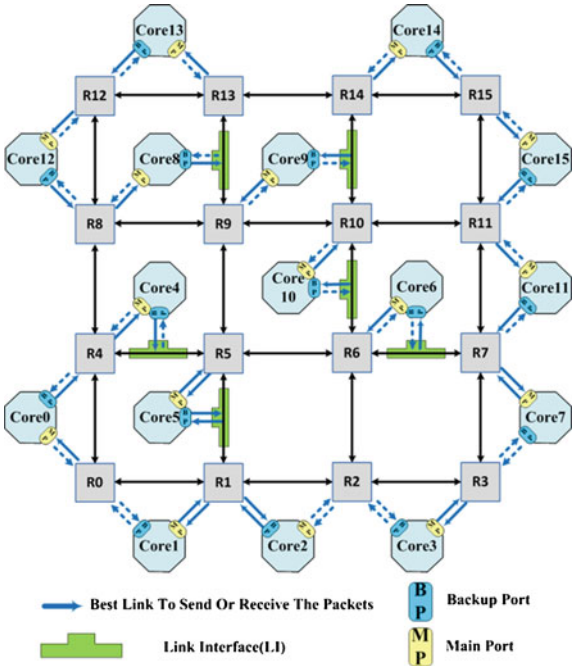


Fig. 15.4 The proposed fault-tolerant architecture in this paper

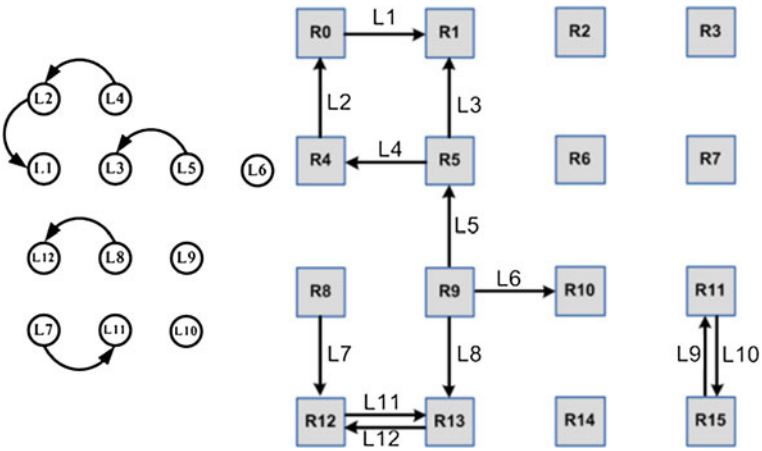


Fig. 15.5 The ASCDG when all routers are working properly

simultaneously request one output port, priority mechanism is used to tackle with this problem (Fig. 15.7). In order to prevent overwriting, a flag has been also considered for each output port to inform its free or busy status.

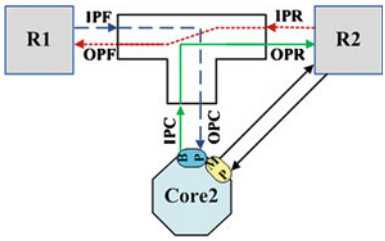


Fig. 15.6 The proposed link interface

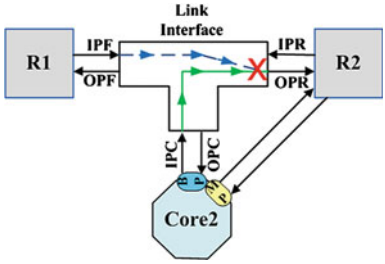


Fig. 15.7 The possible conflict in the link interface

15.5 Experimental Results

In order to compare the average response time and hardware cost of the reliable architecture and traditional mesh, they have been designed in VHDL and synthesized using Xilinx ISE (on FPGA—Xilinx VitexE). Janidarmian et al. [15] shows the effect of mapping algorithm on system reliability, so Video Object Plan Decoder (VOPD) and MPEG-4 as two case studies have been mapped on a 4 × 4 mesh topology using the best mapping technique proposed in [18]. For our experiments, packets are generated according to a uniform distribution with the rates extracting from communication volumes in the VOPD or MPEG-4 core graphs. It means that the number of packets generated in the source cores are derived from the core graph edges and in order to increase the traffic load, communication volumes are multiplied by a traffic factor *i*.

As can be seen in Fig. 15.8, it takes 137,265 cycles to reach all packets to the destination routers in the traditional mesh when the traffic factor *i* is 1 and VOPD application is considered. As mentioned before, the proposed architecture is able to tolerate one router failure and guarantees the 100 % packet delivery.

Because it potentially is possible both main and spare routers are used to send or receive packets by each core, the proposed architecture also significantly decrease the average response time on the faultless and 16 possible faulty routers compared with the mesh architecture as illustrated in Fig. 15.8. It should be pointed that it actually is a great achievement to develop a fault-tolerant NoC design which also has better performance. To explain in details, when all routers are correctly operating, new architecture improves the average response time by 41 % comparing to mesh. The worst observed average response time (123,377 cycles) occurs when the eleventh router fails, and in this case it also improves the

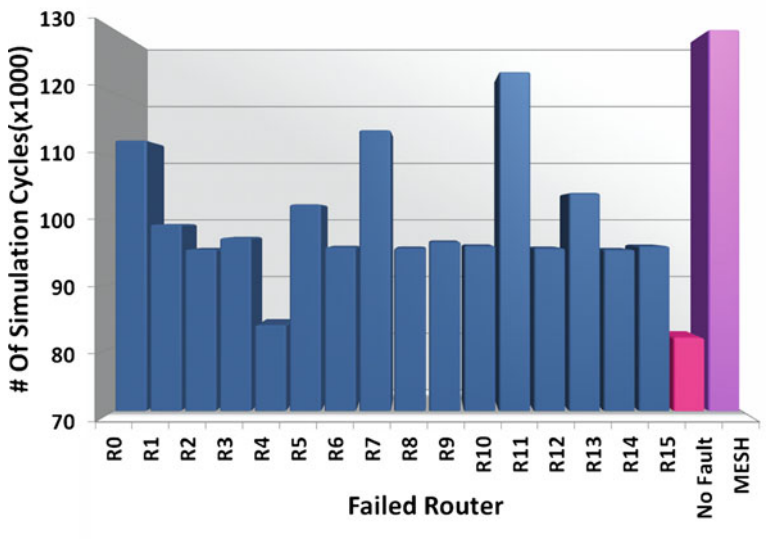


Fig. 15.8 Number of simulation cycles when all packets are received by the destinations (traffic factor = 1)

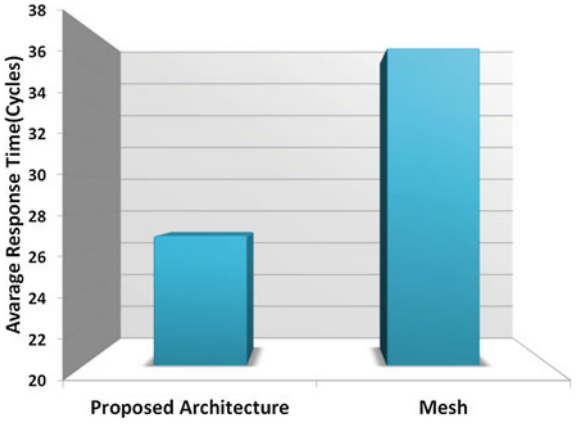


Fig. 15.9 Comparing proposed architecture with mesh in terms of average response time (all situations are considered)

response time by 10 %. Conclusively, we observe that the proposed approach allows to decrease the response time of system by 27 % and tolerate permanent failure of each single router (Fig. 15.9).

To better estimation, the proposed architecture has been simulated in different traffic loads form traffic factor 1–10 for VOPD core graph. As shown in Figs. 15.10 and 15.11, much more improvements are achieved for higher traffic loads even when the worst scenario happens. The worst case (failure of the router 7) signifies that we have the minimum improvements in this situation while comparing performance in all traffic loads.

Traffic Load	Mesh - XY Routing	Best Case (No Fault)		Worst Case (R7 Failed)	
		Proposed Arch	Improvement	Proposed Arch	Improvement
Traffic 1X	**36.79**	**21.83**	**40.66%**	**33.07**	**10.11%**
Traffic 2X	38.03	22.16	41.73%	33.90	10.86%
Traffic 3X	**40.47**	**23.40**	**42.18%**	**35.43**	**12.45%**
Traffic 4X	49.65	27.16	45.30%	42.38	14.64%
Traffic 5X	**58.36**	**29.30**	**49.79%**	**47.79**	**18.11%**
Traffic 6X	72.50	34.23	52.79%	57.58	20.58%
Traffic 7X	**86.92**	**34.82**	**59.94%**	**66.96**	**22.96%**
Traffic 8X	115.92	37.26	67.86%	78.56	32.23%
Traffic 9X	**154.83**	**45.24**	**70.78%**	**102.85**	**33.57%**
Traffic 10X	285.31	53.07	81.40%	173.33	39.25%

Fig. 15.10 Comparing proposed architecture with mesh in terms of average response time in the different traffic loads

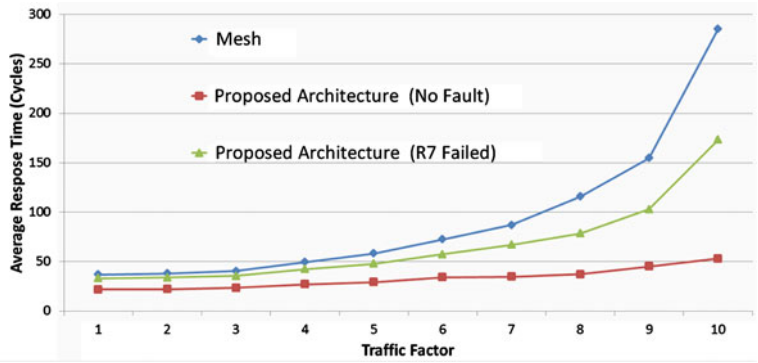


Fig. 15.11 Comparing proposed architecture with mesh in terms of average response time in the different traffic loads

This procedure has been also applied on MPEG-4 core graph shown in Fig. 15.12 and obtained results while traffic factor i is equal to 1 are depicted in Fig. 15.13. Without loss of generality, the edges of the graph are not considered bidirectional and in this case, the improvement of average response time is about 23 % compared to mesh.

To recover from a permanent fault, hardware redundancy is mandatory and reduction of this overhead has always been an important issue in this area. In our design, we do not add any router port contrary to what [14] does and instead a link interface was developed which helps to achieve less hardware overhead.

The router port and LI have been designed and implemented in the VertexE FPGA (v50ecs144-6). Synthesized results (Fig. 15.14) indicate that LI overhead translates to approximately 32 % of the router port area. Therefore, the proposed fault-tolerant architecture (with only 6 LI) introduces better overhead compared to previous work in the literature.

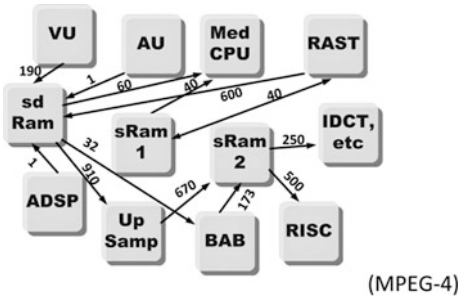


Fig. 15.12 MPEG-4 core graph

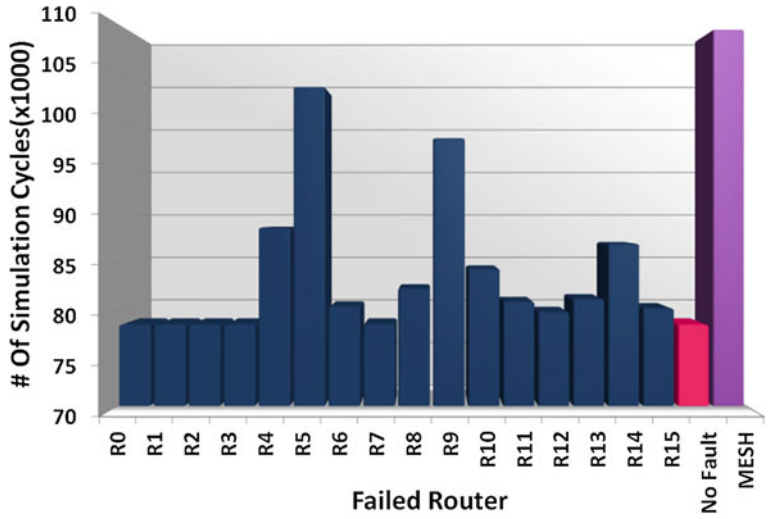


Fig. 15.13 Number of simulation cycles when all packets are received by the destinations (traffic factor = 1)

15.6 Conclusion

In this paper, a new fault-tolerant application-specific network-on-chip was proposed which is able to tolerate one router failure and guarantees the 100 % packet delivery. Considering fault tolerance in designing forces us to accept performance degradation. However, this architecture also improves the average response time of system by 27 and 23 % comparing to traditional mesh for VOPD and MPEG-4 applications, respectively. It was also shown that our architecture obtains more improvements in terms of performance in the higher traffic loads.

Link Interface as a solution for reducing hardware redundancy was suggested and synthesized results demonstrated that each new router port is almost equal to

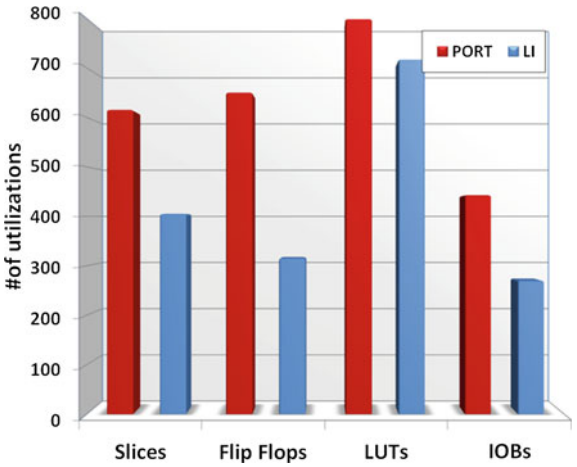


Fig. 15.14 Hardware overhead of the router port and link interface (LI)

three Link Interfaces in terms of hardware overhead. Although the proposed methodology is topology and application agnostic, best mapping algorithm to map Video Object Plan Decoder (VOPD) and MPEG-4 cores onto 2D mesh topology was simulated and investigated.

References

1. Benini L (2006) Application specific NoC design, date. In: Proceedings of the design automation & test in Europe conference, vol 1. p 105
2. Shen W, Chao C, Lien Y, Wu A (2007) A new binomial mapping and optimization algorithm for reduced-complexity mesh-based on-chip network, NOCS. In: First international symposium on networks-on-chip (NOCS'07), pp 317–322
3. RoshanFekr A, Janidarmian M, Samadi Bokharaei V, Khademzadeh A (2011) Yield enhancement with a novel method in design of application-specific networks on chips. Electr Eng Appl Comput 90:247–257
4. Furber S (2008) The future of computer technology and its implications for the computer industry. Comput J 51(6):735–740
5. Borkar S (2005) Designing reliable systems from unreliable components: the challenges of transistor variability and degradation. IEEE Micro 25(6):10–16
6. Dumitraş T, Mărculescu R (2003) On-chip stochastic communication, date. In: Design, automation and test in Europe conference and exhibition (DATE'03), vol 1. p 10790
7. Karimi Koupaei F, Khademzadeh A, Janidarmian M (2011) Fault-tolerant application-specific network-on-chip. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2011, WCECS 2011, San Francisco, USA, pp 734–738, 19–21 Oct 2011
8. Shivakumar P, Kistler M, Keckler SW, Burger D, Alvisi L (2002) Modeling the effect of technology trends on the soft error rate of combinational logic. dsn. In: International conference on dependable systems and networks (DSN'02), p 389
9. Ali M, Welzl M, Hessler S, Hellebrand S (2007) An Efficient fault tolerant mechanism to deal with permanent and transient failures in a network on chip. Int J High Perform Sys Archit 1(2):113–123

10. Rantala V, Lehtonen T, Liljeberg P, Plosila J (2009) Multi network interface architectures for fault tolerant network-on-chip, isscs. International symposium on signals, circuits and systems. pp 1–4
11. Zou Y, Pasricha S (2010) NARCO: neighbor aware turn model-based fault tolerant routing for NoCs. Embed Sys Lett IEEE 2:85–89
12. Dumitras T, Kerner S, Marculescu R (2003) Towards on-chip fault-tolerant communication. In: Proceedings of the Asia and South Pacific design automation conference
13. Shi Z, You K, Ying Y, Huang B, Zeng X, Yu Z (2010) A scalable and fault-tolerant routing algorithm for NoCs, iscas. International symposium on circuits and systems (ISCAS), 30 May 2010–2 June 2010, pp 165–168
14. Refan F, Alemzadeh H, Safari S, Prinetto P, Navabi Z (2008) Reliability in application specific mesh-based NoC architectures. On-line testing symposium, 2008. IOLTS apos; 08. 14th IEEE international, pp 207–212
15. Janidarmian M, Tinati M, Khademzadeh A, Ghavibazou M, RoshanFekr A (2010) Special issue on a fault tolerant network on chip architecture. AIP Conf Proc 1247:191–204
16. Janidarmian M, Khademzadeh A, Tavanpour M (2009) Onyx: a new heuristic bandwidth-constrained mapping of cores onto tile-based network on chip. IEICE Electron Express 6(1):1–7
17. Janidarmian M, Samadi Bokharaie V, Khademzadeh A, Tavanpour M (2010) Sorena: new on chip network topology featuring efficient mapping and simple deadlock free routing algorithm. 2010 10th IEEE international conference on computer and information technology, pp 2290–2299
18. Janidarmian M, RoshanFekr A, Samadi Bokharaei V (2011) Application-specific networks-on-chips design. IAENG Int J Comp Sci 38(1):16–25
19. Palesi M, Holsmark R, Kumar S (2006) A methodology for design of application specific deadlock-free routing algorithms for NoC systems. Hardware/software codesign and system synthesis, CODES + ISSS '06. In: Proceedings of the 4th international conference, pp 142–147, Oct 2006

Chapter 16
Co-Existance of High Assurance and Cloud Based Computing

William R. Simpson and Coimbatore Chandersekaran

Abstract Cloud computing is emerging as an attractive, cost effective computing paradigm. However, many of the applications require high assurance, attribution and formal access control processes including defense, banking, credit, content distribution, etc. Current implementations of cloud services do not meet high assurance requirements. The high assurance requirement presents many challenges to normal computing and some rather precise requirements that have developed from high assurance issues for web service applications. The challenges of high assurance associated with cloud computing are primarily in five areas. The first is virtualization and the loss of attribution that accompanies a highly virtualized environment. The second is the loss of ability to perform end-to-end communications. The third is the extent to which encryption is needed and the need for a comprehensive key management process for public key infrastructure, as well as session and other cryptologic keys. The fourth is in monitoring and logging for attribution, compliance and data forensics. The fifth is in cloud content storage. We explore each of these challenges and discuss how they may be able to be overcome. Our view of high assurance and the issues associated with web services is shaped by our work with DoD and the Air Force, but applies to a broader range of applications, including content delivery and rights management.

Keywords Attribution · Authentication · Cloud computing · Content management · High assurance · Hypervisor · IT security · Virtualization

W. R. Simpson (✉) · C. Chandersekaran
Institute for Defense Analyses, 4850 Mark Center Drive,
Alexandria, VA 22311, USA
e-mail: rsimpson@ida.org

C. Chandersekaran
e-mail: cchander@ida.org

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_16,

16.1 Introduction

This paper is based in part on a paper published in WCECS [1]. Cloud computing has come to mean many different things. To some, it is simply putting one's data on a remote server. However, in this paper, we utilize the definition provided by NIST [2]. They define five essential characteristics of any cloud computing environment:

1. On demand self-service,
2. Broad network access,
3. Resource pooling,
4. Rapid elasticity, and
5. Measured service.

It is important to note that multi-tenancy and virtualization are not essential characteristics for cloud computing. For our discussion we will assume no multi-tenancy, which adds the largest element of security complication.

Arguments below do not require either. Cloud computing is, at its core, a *service*. There are three primary models of this service. In the lowest level Infrastructure as a Service (IaaS), storage, computation, and networking are provided by the cloud provider to the cloud consumer. In the next level up of Platform as a Service (PaaS), all of the trappings of IaaS plus an operating system and perhaps some application programming interfaces (APIs) are provided and managed by the cloud provider. The highest service model is Software as a Service (SaaS), in which the cloud provider provides an end-user service such as webmail. The higher the service model, the more control the cloud provider has as compared to the cloud consumer.

There are four different models for deploying cloud services. Primarily, they are public or private clouds. In a public cloud, the infrastructure—although generally not the data on it—may be used by anyone willing to agree to its terms of use. Public clouds exist off the premises of the cloud consumer. Private cloud infrastructure is used only by one organization. It may exist either on or off the organization's premises. There are two twists to these infrastructures. In a community cloud, a group of organizations with similar interests or needs share a cloud infrastructure. That infrastructure is not open to the general public. The community may adopt a single security approach and the same security mechanisms or it may not. Community clouds are best formed in this manner. In this form the shared cloud is similar to an enterprise with restricted sharing. In the latter there is a restricted form of multi-tenancy which may lead to security issues unless low assurance satisfies the basic requirement. In a hybrid cloud, two or more cloud deployment models are connected in a way that allows data or services to move between them. An example of this would be an organization's private cloud that makes use of a community cloud during loads of high utilization.

16.2 Cloud Computing

Cloud computing benefits emerge from economies of scale [2]. Large cloud environments with multiple users are better able to balance heavy loads, since it is unlikely that a large proportion of cloud consumers will have simultaneously high utilization needs. The cloud environment can therefore run at a higher overall utilization, resulting in better cost effectiveness. In a large cloud computing environment, rather than having a number of information technology generalists, the staff has the ability to specialize and become the masters of their own domains. In many cloud environments this balancing is done by virtualization and the use of a hypervisor. With regard to information security, the staff can become even more specialized and spend more time hardening platforms to secure them from attacks. In the homogeneous cloud environment, patches can be rolled out quickly to the nearly identical hosts.

16.2.1 Drawbacks of the Cloud

Cloud computing is not without its drawbacks. In cases where services are outsourced, there is a degree of loss of control. This can affect compliance with laws, regulations, and organizational policies. Cloud systems have additional levels of complexity to handle intra-cloud communications, scalability, data abstraction, and more. To be available to cloud consumers, cloud providers may need to make their services available via the Internet. And critically, many clouds use multi-tenancy, in which multiple organizations simultaneously utilize a single host and virtualization. If one tenant organization is compromised or malicious, it may be able to compromise the data or applications of the other organizations on the same host. The load balancing may use a single identity for all instances of a service whether it is virtual or real.

16.2.2 Differences from Traditional Data Centers

Cloud computing relies on much of the same technical infrastructure (e.g., routers, switches, operating systems, databases, web servers) as traditional data centers and as a result, many of the security issues are similar in the two environments. The notable exception in some cases is the addition of a hypervisor for managing virtual machines. The Cloud Security Alliance's security guidance states "Security controls in cloud computing are, for the most part, no different than security controls in any IT environment. Cloud computing is about gracefully losing control while maintaining accountability even if the operational responsibility falls upon one or more third parties." While many of the controls are similar, there are

two factors at work that make cloud computing different: perimeter removal and trust. With cloud computing, the concept of a network or information perimeter changes radically. Data and applications flow from cloud to cloud via gateways along the cloud perimeters. However, since the data may be stored in clouds outside the organization's premises or control, perimeter controls become less useful. In exchange for the lack of a single perimeter around one's data and applications, cloud consumers must be able to trust their cloud providers. A lack of trust in a cloud provider does not necessarily imply a lack of security in the provider's service. A cloud provider may be acceptably secure, but the novelty of cloud computing means that many providers have not had the opportunity to satisfactorily demonstrate their security in a way that earns the trust of cloud consumers. Trust must be managed through detailed Service Level Agreements (SLAs), with clear metrics and monitoring mechanisms, and clear delineation of security mechanisms [3].

Cloud computing benefits emerge from economies of scale [4]. Large cloud environments with multiple users are better able to balance heavy loads, since it is unlikely that a large proportion of cloud consumers will have simultaneously high utilization needs. The cloud environment can therefore run at a higher overall utilization, resulting in better cost effectiveness. In a large cloud computing environment, rather than having a number of information technology generalists, the staff has the ability to specialize and become the masters of their own domains. In many cloud environments this balancing is done by virtualization and the use of a hypervisor. With regard to information security, the staff can become even more specialized and spend more time hardening platforms to secure them from attacks. In the homogeneous cloud environment, patches can be rolled out quickly to the nearly identical hosts.

16.2.3 Some Changes in the Threat Scenario

There are clear differences in many of the threat scenarios as detailed below [2]:

1. Loss of governance (or visibility and/or control of the governance process),
2. Lock-in (threats may be present and locked into the cloud environment),
3. Isolation failure (e.g., hypervisor attack, lack of accountability),
4. Compliance risks (if provider cannot provide compliance evidence or will not permit audit by customer, lack of accountability),
5. Management interface compromise (and or inheritance of threats and/or malicious code from other users of the cloud),
6. Data protection (how does customer verify protection, lack of accountability),
7. Insecure or incomplete data deletion,
8. Malicious insider (often the cloud insider is not vetted as well as the organizational insider, and insiders from other customers could bring in contagious viruses—see 5 above).

16.3 High Assurance Computing

While the current implementations of Cloud Computing provide efficient and operationally friendly solutions to data computing and content distribution, they are not up to the challenge of high assurance. In certain enterprises, the network is continually under attack. Examples might be; banking industry enterprise such as a clearing house for electronic transactions; defense industry applications; credit card consolidation processes that handle sensitive data; both fiscal and personal, medical with concerns for privacy and statutory requirements; content distributor's worried about rights in data, or theft of content.

The attacks have been pervasive and continue to the point that nefarious code may be present, even when regular monitoring and system sweeps clean up readily apparent malware. Despite this attack environment, the web interface is the best way to provide access to many of its users. One way to continue operating in this environment is to not only know and vet your users, but also your software and devices. Today we regularly construct seamless encrypted communications between machines through SSL or other TLS. These do not cover the "last mile" between the machine and the user (or service) on one end, and the machine and the service on the other end. This last mile is particularly important when we assume that malware may exist on either machine, opening the transactions to exploits for eaves dropping, ex-filtration, session high-jacking, data corruption, man-in-the-middle, masquerade, blocking or termination of service, and other nefarious behavior. Before we examine the challenges of Cloud Computing systems, let us first examine what high assurance architecture might look like.

16.3.1 Architectural Features

In order to build an architecture that conforms to these tenets, there must be elements that insure that they are built into the systems. In the architecture we espouse, the basic formulation is based on web services and uses Organization for the Advancement of Structured Information Standards (OASIS) standards of security [5].

16.3.1.1 Naming and Identity

Identity will be established by the requesting agency. To avoid collision with the names, the identity used by all federated exchanges shall be the name as it appears on the primary credential provided by the certificate authority. The name must be unique over time and space which means that retired names are not reused and ambiguities are eliminated. Naming must be applied to all active entities (persons, machines, and software).

16.3.1.2 Credentials

Credentials are an integral part of the federation schema. Each identity (all active entities) requiring access shall be credentialed by a trusted credentialing authority.

16.3.1.3 Bi-Lateral End-to-End Authentication

The requestor will not only authenticate to the service (not the server), but the service will authenticate to the requestor. This two way authentication avoids a number of threat vulnerabilities.

16.4 Challenges in Bringing the Cloud and High Assurance Together

Despite the obvious advantages of cloud computing, the large amount of virtualization and redirection poses a number of problems for high assurance. In order to understand this, let's examine a security flow in a high assurance system (Fig. 16.1).

The application system consists of a web application (for communication with the user), one or more aggregation services that invoke one or more exposure services and combines their information for return to the web application and the user, The exposure services retrieve information from one or more Authoritative Data Sources (ADSs).

Once the authentication is completed, an SSL connection is established between the requestor and the service provider, within which a WS-Security package will be sent to the service. The WS-Security [5] package contains a SAML token generated by the Security Token Server (STS) in the requestor domain. The primary method of authentication will be through the use of public keys in the X.509 certificate, which can then be used to set up encrypted communications (either by X.509 keys or a generated session key). Session keys and certificate keys need to be robust and sufficiently protected to prevent malware exploitation. The preferred method of communication is secure messaging using WS Security, contained in SOAP envelopes. The encryption key used is the public key of the target (or a mutually derived session key), ensuring only the target can interpret the communication.

The problem of scale-up and performance is the issue that makes cloud environments and virtualization so attractive. The cloud will bring on assets as needed and retire them as needed. Let us first examine scale-up in the unclouded secure environment. We will show only the web application, although the same rules apply to all of the communication links between any active elements shown in the Fig. 16.2. The simplest form of dividing the load is to stand up multiple

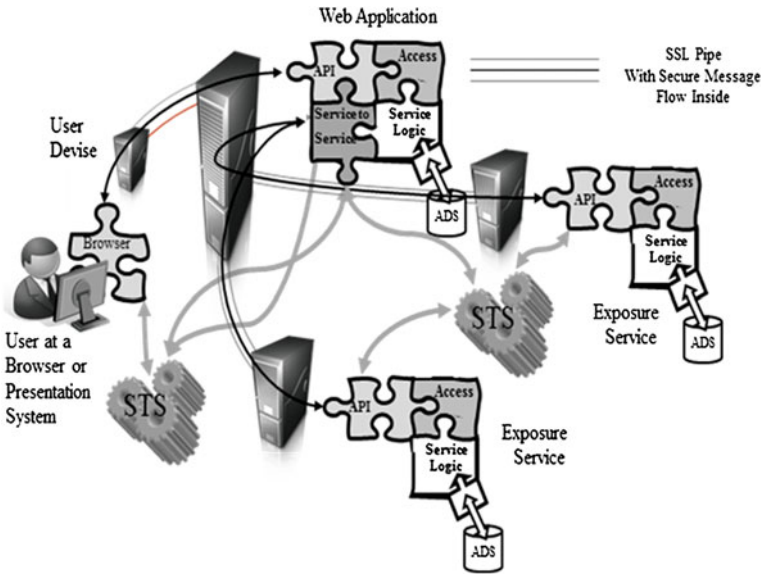


Fig. 16.1 High assurance security flows

independent instances and divide users into groups who will use the various instances. Dependent instances that extend the thread capabilities of the server are considered single independent instances. Remember, all independent instances are uniquely named and credentialed and provisioned in the attribute stores. A representation that is closer to the cloud environment is shown in Fig. 16.3.

A traffic cop (load balancer) monitors activity and posts a connection to an available instance. In this case all works out since the new instance has a unique name, end-point, and credentials with which to proceed. All of this, of course needs to be logged in a standard form and parameters passed to make it easy to reconstruct for forensics. We have shown a couple of threats that need mitigation where one eavesdrops on the communication and may actually try to insert himself into the conversation (man-in-the-middle). This highlights the importance of bi-lateral authentication and encrypted communications. The second is present on instance 4 and highlights the need to protect caches and memory spaces.

When a cloud environment runs out of resources for computing, it builds additional instances, some of these may be thread extension schemas, and some may be independent instances. The traffic cop here is often called a hypervisor and it keeps track of the instances and connections. Figure 16.3 shows notionally how this operation works. When thread capacity is saturated at the server, the hypervisor would nominally redirect the request to an independent virtual or real instance of the web application. If none exists, it will build one from elements in the resource pool as depicted in instance 4 on the chart. If the last user signs off of an independent virtual or real instance (instance 3 in the Fig. 16.3), the hypervisor

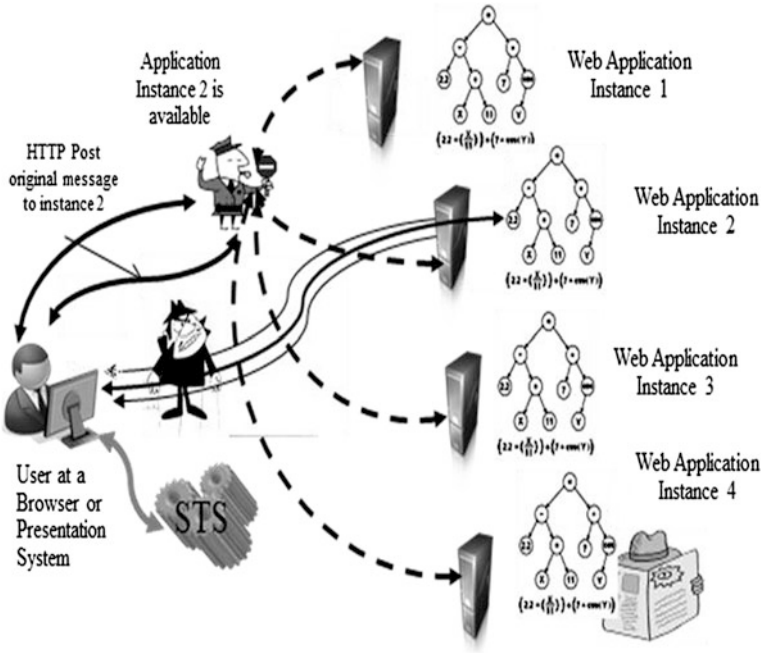


Fig. 16.2 High assurance load balancing

tears down the instance and places the resources back into the resource pool. This provides an efficient re-allocation of resources.

There are several steps that must be taken to preserve the security, if we are interested in a high assurance computing environment. The number of independent instances must be anticipated. Names, credentials and end points must be assigned for their use. The attribute stores and HSMs must be provisioned with properties and key to be used. The simple re-direct must be changed to a re-post loop as in Fig. 16.3. The user will then have a credentialed application to authenticate with bi-laterally and an end point for end-to-end message encryption. Key management is complex and essential. When a new independent instance is required it must be built, and activated (credentials and properties in the attribute store, as well as end point assignment). All of these activities must be logged in a standard format with reference values that make it easy to reassemble the chain of events for forensics. When a current independent instance is retired, it must be disassembled, and de-activated (credentials and properties in the attribute store, as well as end point assignment).

All of these activities must be logged in a standard format with reference values that make it easy to reassemble the chain of events for forensics. The same threats exist, and the same safeguards must be taken. In fact, in Fig. 16.3 nefarious code is built right into the virtual or real instance 4, which underscores the need for trusted and verified software to do the virtualization, and protection of the resources while they are in the resource pool.

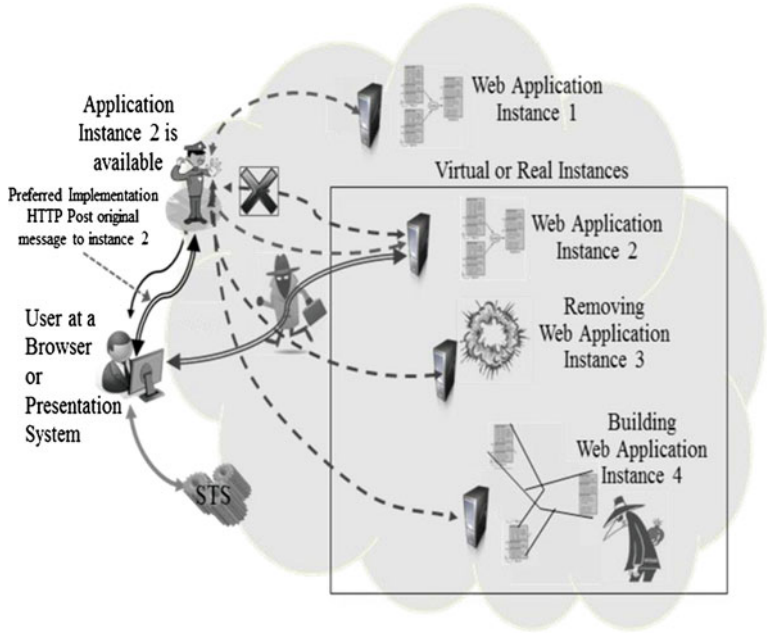


Fig. 16.3 High assurance virtualized hypervisor activity

A recap of these challenges is listed below:

1. Shared Identities and credentials break the accountability paradigm.

- Each independent instance of a virtual or real machine or virtual or real service must be uniquely named [6, 7] and provided a PKI Certificate for authentication. The Certificate must be activated while the virtual machine is in being, and de-activated when it is not, preventing hijacking of the certificate by nefarious activities. Each instance of an independent virtual or real machine or virtual or real service must have a unique end point. This means that simple re-direct will not work. Extensions of the thread mechanism by assigning resources to the operating system may preserve this functionality.

2. Multi-tenancy (multiple tenants using a single host) must be prohibited.
3. No virtualization across machines (each virtual machine must reside in a single real machine).
4. Each potential independent instance of a service must have an account provisioned with appropriate elements in an attribute store. This is required for SAML token issuance.
5. A cloud based Security Token Service (STS) needs to be installed and implemented and it must meet all of the requirements listed her for uniqueness of names and end points as well as instantiated certificates and cryptographic capability.

6. The importance of cryptography cannot be overstated, and all internal communications as well as external communications should be encrypted to the end point of the communication. Memory and storage should also be encrypted to prevent theft of cached data and security parameters.
7. Private keys must reside in Hardware Storage Modules (HSMs).
    a. Stand-up of an independent virtual or real machine or virtual or real service must link keys in HSM, and activate credentials pre-assigned to the virtual service.
    b. Stand-down of an independent virtual or real machine or virtual or real service must de-link keys in HSM, and de-activate credentials pre-assigned to the virtual service.
    c. Key Management in the virtual environment is a particular concern and a complete management schema including destruction of session keys must be developed.
8. Proxies and re-directs break the end-to-end paradigm. When end points must change, a re-posting of communication is the preferred method.
9. Resource pools must be protected from persistent malicious code.
10. All activities must be logged in a standard format with reference values that make it easy to reassemble the chain of events for forensics.

The aforementioned challenges are daunting, but provisions must be made if high assurance computing environments are take advantage of the cloud computing environment.

16.5 Content in the Cloud

In the high assurance enterprise, content stored in general cloud areas can only be protected in one of two ways. The first is total isolation and restrictive gateways for access, which is contrary to the cloud computing paradigm. The second is in encrypted form when unauthorized access is an issue. The latter implies a rights management system of some type.

16.5.1 The Rights Management Function

The Rights management is a collective concept that includes the automated and manual processes to accomplish the steps below:

a. The information asset must be labeled for access and distribution; this is done by the author. Defaults may be assigned absent author input.
b. The information asset must be signed by the author for content integrity (additional signatures may be affixed for authority).

c. Generation of associated metadata for search and discovery.
d. Assignment of an identity (name)—defaulted by the system but can be changed by the user.
e. Author assignment of the actual storage location on the network and filing of the cross-reference between the location and the identity of the asset. The location may be physical or logical.
f. Presentation of a rights information request page (defaulted to read/write/delete rights to the creator and read/delete rights to all others and signature. If additional rights are required (e.g., interest group, special-access group), these are specified at this time.
g. Examination of the access control labels and where an information asset is restricted and not available to all (internal/external), encryption of the information asset and the attachment of an appliqué to the information asset which is used to communicate to the Rights Manager for access control. If the information asset is not access control labeled and is available to all internal and external, the information asset is not encrypted. As a consequence, both encrypted and unencrypted assets may be further distributed without consequence.

To access an information asset, the appliqué attached to the content program for the information asset examines the information asset and if it is encrypted communicates to the Rights Manager via a secure web session to verify claims.

16.5.2 The Components of a Stored Information Asset

The components of a stored information asset are provided in Fig. 16.4 and must be created in steps as described below:

Formatted Document Section a. Information Labeled

Provided by the rights management software with defaults based upon user interest group memberships or by user from approved list, also includes "draft", "final", or Approved as previously described.

Formatted Document Section b. Information Asset Signature(s)

The author's signature (and others) are added and further changes to the information asset at this point are prohibited.

External Information c. MDE Metacard

The Meta Data Environment (MDE) Metacard is prepared. This involves a number of items described below: It should be noted that most information assets are not directly retrievable and it must be retrieved by the content retrieval service for checking of ACLs, MAC issues and restricted authorities. The exception is unclassified, unlimited distribution.

- ***Access Control Labels***

  These are taken directly from the trusted labeling of the information asset.

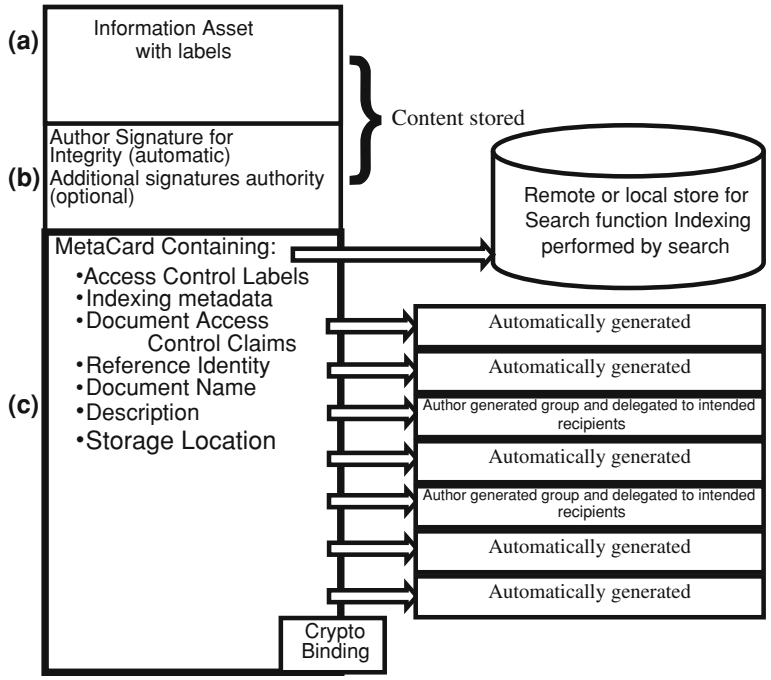


Fig. 16.4 Authoritative content information asset format

- ***Key Word MetaData***

The key words are developed from full information asset text scan and/or can be manually entered.

- ***ACL Lists and Associated Data***

The primary ACL is provided by the author (example, "MyGroup"). Once the label is chosen the rights manager inserts this as a delegatable claim of the author.

- ***Reference Identity and Information asset Description***

This is the mechanism for retrieval. The rights manager software defines the identity to prevent duplication, ambiguity or confusion in the information asset file keeping system.

- ***Information asset Name***

The rights management software will provide a default name. It may be modified by the author.

- ***Information asset Description***

The rights management software will suggest a description based upon a title or lead heading. It may be modified by the author.

- ***Storage Location(s)***

This is the actual storage location of the information asset in network asset store. Each time an unmodified copy of the information asset is stored in a different location, the appliqué provides that location and the unique id of the information asset to the rights manager for updating the metacard. The metacard may contain any number of storage locations. This latter allows cleanup when archiving old content.

16.6 Summary

We have reviewed the basic approaches to clouds and their potentials for savings in computing environments. We have also discussed at least one high assurance architecture and its' requirements which provide direct challenges to the way cloud computing environments are organized. Notably the extensive use of virtualization and re-direction is severe enough that many customers who need high assurance have moved away from the concept of cloud computing [8, 9]. Content storage in high assurance cloud environments is also a concern requiring some tools and processes. We believe, however, that a precise statement of the high assurance requirements will lend themselves to solutions in the cloud computing environment, and expand the potentials use of this technology. These concepts are part of a more comprehensive enterprise architecture for high assurance that is web-service based and driven by commercial standards. Portions of this architecture are described in references [10–14].

References

1. Simpson WR, Chandersekaran C (2011) High assurance challenges for cloud computing. In: Proceedings of the world congress on engineering and computer science 2011, Lecture notes in engineering and computer science, vol I. San Francisco, Oct 2011, pp 61–66
2. Jansen W, Grance T (2011) NIST SP 800-144 Draft: guidelines on security and privacy in public cloud computing, security division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899-8930, Jan 2011. http://csrc.nist.gov/publications/drafts/800-144/Draft-SP-800-144_cloud-computing.pdf
3. Mell P, Grance T (2011) NIST SP 800-145 Draft: cloud computing, computer security division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899-8930, Jan 2011. http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf
4. Cloud Security Alliance (2009) Security guidance for critical areas of focus in cloud computing V2.1, Dec 2009, https://cloudsecurityalliance.org/csaguide.pdf
5. OASIS Identity Federation (2011) Liberty alliance project, Available at http://projectliberty.org/resources/specifications.php. Accessed 19 Feb 2011

6. OASIS profiles for the OASIS security assertion markup language (SAML) V2.0. Available at http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security. Accessed 19 Feb 2011
7. Standard for Naming Active Entities on DoD IT Networks, Version 3.5, Sept 23, 2010
8. Remarks-Debra Chrapaty, Corporate Vice President, Global Foundation Services, Microsoft Mgt Summit, Las Vegas, May 2008. http://www.microsoft.com/Presspass/exec/debrac/mms2008.mspx. Accessed 19 Feb 2011
9. Plesser A (2008) Executive producer, Beet.tv, cloud computing is hyped and overblown, Forrester's Frank Gillett.Big Tech Companies have "Cloud Envy". http://www.beet.tv/2008/09/cloud-computing.html, Sept 26, 2008. Accessed 19 Feb 2011
10. Catteddu D, Hogben G, European Network Information Security Agency (ENISA) (2009) Cloud computing risk assessment, Nov 2009. http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment
11. Simpson WR, Chandersekaran C, Trice A (2008) A persona-based framework for flexible delegation and least privilege. In: Electronic digest of the 2008 system and software technology conference, Las Vegas, Nevada, May 2008
12. Simpson WR, Chandersekaran C, Trice A (2008) Cross-domain solutions in an era of information sharing. In: The 1st international multi-conference on engineering and technological innovation (IMET 2008), vol I. Orlando, FL, pp 313–318
13. Simpson WR, Chandersekaran C (2009) Information sharing and federation. In: The 2nd international multi-Conference on engineering and technological innovation (IMETI 2009), vol I. Orlando, FL, pp 300–305
14. Chandersekaran C, Simpson WR (2010) A SAML framework for delegation, attribution and least privilege. In: The 3rd international multi-Conference on engineering and technological innovation (IMETI 2010), vol 2. Orlando, FL, pp 303–308

Chapter 17
Estimation of Susceptibility to Hot Tearing in Solidifying Casting

Norbert Sczygiol and Zbigniew Domański

Abstract Hot tearing, also called hot cracking, is a serious defect that appears during the solidification of an alloy. Due to the low recurrence of the phenomena occurring during alloy solidification, such as the evolution of grained structure or stress redistributions, the casting's susceptibility to hot tearing can be estimated only in an approximate way. Predicting the appearance of hot tears in alloys is thus an important issue in industrial practice. This work concerns with a new criterion for hot tearing evaluation in castings. An algorithm for the computer simulations of the phenomena accompanying the casting formation is introduced and discussed.

Keywords Casting · Computer simulation · Hot tears · Finite element method · Solidification processing · Susceptibility

17.1 Introduction

The production of castings is an important technology that involves many factors of significant impact on the quality of the finished product. In shape casting, an equiaxial structure is formed. During solidification processes in the solid–liquid

N. Sczygiol (✉)
Institute of Computer and Information Sciences, Czestochowa University of Technology, Dabrowskiego 69, 42-201 Czestochowa, Poland
e-mail: norbert.sczygiol@icis.pcz.pl

Z. Domański
Institute of Mathematics, Czestochowa University of Technology, Dabrowskiego 69, 42-201 Czestochowa, Poland
e-mail: zbigniew.domanski@im.pcz.pl

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_17,

areas various types of defects may appear. Among them the shrinkage leads to macro-porosity or/and micro-porosity effects, while the stress reveals the so-called hot tearing in the casting. Hot tearing of solid–liquid areas occurs when the stresses acting on them are able to break the backbone of solid phase, filled with the liquid phase.

Hot tearing of castings was and still is of founders and scientists interest [1, 2]. Initially, the problems of hot tearing formation were solved by experimental estimation of the hot tearing susceptibility of foundry alloys. Then, the mathematical models have been developed. Unfortunately, the studies, predominantly focused on the formation of a single crack, were not relevant to industrial practice. The next step in the development of testing methods for hot tearing occurrence was the use of advanced numerical methods, through the computer simulations [3].

Approaches based on computer simulations can be divided into two groups. The first group concerns the analysis of development of a single crack, whereas the second group involves a comprehensive analysis of thermo-mechanical phenomena, accompanying the production process of castings. The analysis of thermo-mechanical phenomena attempts to draw conclusions for assessing the degree of risk of the appearance of defects in the continuity in the entire casting or in its selected parts [4]. Such an approach would allow to built up commercial engineering programs. However, such programs do not contain any criteria for hot tearing evaluation in castings. Users have to decide which of the calculated values characterizing the state of stress and/or deformation are appropriate to the rupture-susceptibility evaluation. It should be noted, however, that such an analysis is very time consuming, requires good knowledge of the phenomena in casting formation, skills in simulation of these phenomena and possessing specialized engineering software, usually based on a finite element method.

In this paper, we are focused on the analysis of the susceptibility to hot tearing during an equiaxial structure casting. We propose a new stress criterion to assess the level of risk of rupture in selected fragments of the casting. The evaluation of castings hot tearing with the use of this criterion is possible only after conducting a series of calculations, according to the algorithm proposed here. The result is the information about the degree of rupture risk in selected areas of the casting. Studying the susceptibility to hot tearing by means of the method proposed here is time-consuming, but the already achieved calculations' speed will result in applications of the proposed solutions in foundry practice.

17.2 The Criterion for Hot Tearing Evaluation

Metal alloys often solidify by increasing equiaxial dendrites. It can be assumed that initially each dendrite grows individually. As the dendrites are in contact with each other they form the backbone of the solid phase. Dendrite arms are intertwined with the arms of their neighbors. From this point, there appears tension in the solidifying solid–liquid area, carried by each entangled dendrite arm. The dendrites are

separated by layers largely filled with the liquid phase. Such two-phase areas (consisting of the growing equiaxial grains and separating them layers of the liquid phase) in the numerical modeling are represented by hexagonal solid phase grains and the surrounding layers of the liquid phase. The solid phase is presented using regular hexagons, while the liquid phase by means of flattened hexagons. The relevant parts of these two types of hexagons are on the border area, see Fig. 17.1. The size of both areas (solid and liquid phases) is characterized by the participation of the solid phase, calculated at the stage of solidification simulation.

Modeling solidifying casting area enables operating at the micro level of analysis separately for growing grains and for narrowing layers of the liquid phase. Solidifying metal grains almost always are much smaller than finite elements used in calculations and in the macroscopic stress analysis. In the macroscopic analysis, a two-phase area is treated as isotropic, ignoring the grain nature of the casting construction. However, since the solidification simulation is carried out based on the coupled model, i.e. macro-microscopic, so after the solidification simulation there can be easily reconstructed the accumulation of grains in two-phase areas, which combined with the analysis of stress at a microscopic level enables to analyze the phenomena that could lead to hot tearing.

The most important causes of stress in the casting are uneven temperature gradients and the resistance posed by the wall of the mold to the shrinking casting. Conditions of heat evacuation from the casting to the mold and to the environment determine the speed of the alloy solidification, i.e. the equiaxial grains growth speed, but also the speed of the stress generated in the casting.

To evaluate hot tearing of the solidifying casting there is a new stress criterion proposed that takes into account the stress-speed ratio of effective stress in the layers separating the congealed particles to the speed of effective strain in these grains. The proposed criterion is expressed by the so-called local coefficient of susceptibility to hot tearing, marked as Θ. This is the criterion, which deals with stress states in micro scale, but these conditions are obtained under the stress states in a macro scale. During the solidification, the changes in geometry (size) of grains and separating layers are obtained from microscopic analysis conducted on the basis of macroscopic modeling. This is possible because in the macroscopic modeling the growth of equiaxial grains is represented by the connection of diffusion phenomena (micro scale) with thermal phenomena (macro scale). The calculation of the local coefficient of susceptibility to hot tearing proceeds in the following time steps, beginning with the participation of the solid phase, in which the backbone of solid phase is formed, until complete solidification. Effective strain rate can be written as:

$$\dot{\bar{\sigma}} = \frac{|\Delta\bar{\sigma}|}{\Delta t}, \tag{17.1}$$

where: $\Delta\bar{\sigma}$ is the effective stress increment in the time step Δt. The study shows, however, that much better results are obtained by an introduction of relative effective stresses to the criterion. Therefore it can be written that:

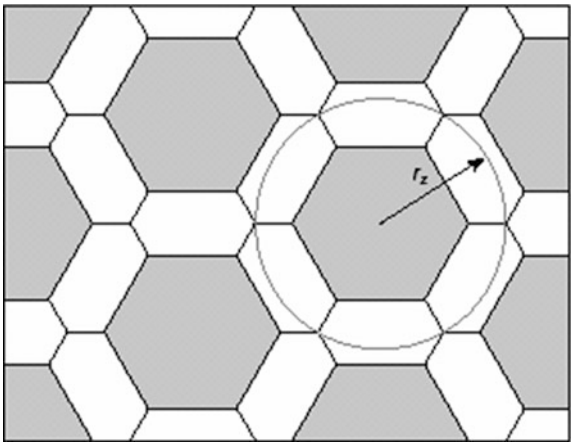


Fig. 17.1 Model of a two-phase area for the alloy solidifying in the form of equiaxial grains

$$\Theta = \frac{|\Delta\bar{\sigma}_l|}{|\Delta\bar{\sigma}_g|}\frac{\bar{\sigma}_g}{\bar{\sigma}_l}, \tag{17.2}$$

where: l is a sub-layer separation, while g denotes the sub-grain-solidified parts. The quotient of relative increment of effective stress in the layers and the grains tends to zero with increasing fraction of the solid phase. Thus, for the sake of clarity, the criterion for hot tearing may be transformed to the form:

$$\Theta = -\ln\left(\frac{|\Delta\bar{\sigma}_l|}{|\Delta\bar{\sigma}_g|}\frac{\bar{\sigma}_g}{\bar{\sigma}_l}\right). \tag{17.3}$$

The application of the criterion Eq. (17.3) requires a computer simulation in the macro scale, and then in the micro scale. At the macro level macroscopic (standard) finite elements are used, while at the micro level we use microscopic elements covering the macroscopic element area. Also, formulating the finite element method is different for both types of simulation. At the macro level it is a traditional formulation, e.g. based on the method of weighted residuals, while at the micro level a hybrid formulation was used [5–7].

The local coefficient of susceptibility Θ given by Eq. (17.3) describes the local susceptibility to hot tearing of a small macroscopic area, corresponding to one finite macroscopic element, subdivided into two areas, i.e. grains and layers separating them. Used in Eq. (17.3) stress values and their increments are determined for the subdivisions of layers and grains, receiving two tensors which describe the resultant state of stress in all the grains and the resultant state of stress in all the layers of separation, which belong to the analyzed area. Tensors are obtained as a result of the so-called homogenization, based on the integration of the stress function in the above-mentioned subdivision, and then dividing the resulting value by the area of integrated subdivision.

Large values of factor Θ indicate high susceptibility to hot tearing. However, the value of Θ increases with increasing equiaxial grain, as a result of stress growing with an increasing solidification area. Therefore, the criterion Θ does not indicate a specific limit value, above which the casting will crack. Θ values are used to indicate the areas of analyzed casting, where a damage, i.e. the rupture, is most likely to appear. The criterion Θ can also be used to determine the conditions most conducive to the production of a given type casting.

17.2.1 Algorithm for Preparatory Calculations of Factor Θ

Computation of factor Θ is possible after a complex computer simulations that provide data on which you can only determine (estimate) the susceptibility to the casting hot-tearing. At this stage, a number of preparatory tasks are performed. The steps leading to denoting the casting susceptibility to hot tearing cover the following.

- Simulation of solidification. For succeeding time steps the temperature field, the distribution of the solid phase participation and the mean radii of equiaxial grains, are determined.
- Calculating distributions of stress in consecutive time steps.
- Identification (selection) of subdivided areas for the hot tearing analysis.
- Division of the macroscopic finite elements into the microscopic-hexagon-hybrid-finite elements in order to obtain the solid–liquid areas.
- Calculation of the stress in all solid–liquid areas corresponding to the macroscopic finite elements.
- Calculation of the value of Θ for each macroscopic finite element in the selected areas.
- Preparation of the scale of susceptibility to hot-tearing based on the simulations and calculations carried out for all the analyzed variants of the task.
- Execution of the local distribution coefficient diagrams for susceptibility to hot tearing for different variants of the task.
- Drawing conclusions.

The first two steps are described in literature [8]. Therefore, only the remaining steps will be described below.

17.2.2 Identification of Subdivisions

There is no point in conducting the analysis of the susceptibility to hot tearing for all the macroscopic finite elements of the casting. As the practice indicates, the cracks appear only in selected, easy identified, fragments of the casting. To proceed to the identification of such fragments, a group of finite elements and the area around, must

be selected beforehand. This selection can be done with the help of the probability of the hot tearing localization. The group of selected macroscopic elements should be slightly greater than the area of the analysis. There should be also a group of macroscopic elements (of a similar number of the elements) selected for the analysis in the area least subject to hot tearing. If there is a suspicion about the possibility of hot tearing appearance in other parts of the casting, then another group of elements should be created and analyzed. For the casting shown in Fig. 17.2, three groups of elements were selected.

The major group is located in the central part of the casting and includes elements collected under infusion and forming a notch around the bottom of the casting. This is the group at high risk of hot tearing. In the group of the elements located on the left, in the casting arm, there is no danger of hot tearing. Rupture should not occur either, by design, in the third group, comprising the area around the 'notch' connecting the right shoulder with the casting 'head' located at its end.

Comparing the areas selected for the analysis with the size of the entire area of the casting (Fig. 17.2) shows that the number of elements selected for analysis is relatively small in comparison to the number of finite elements in the whole casting.

17.2.3 Figures Division of the Macroscopic Finite Elements into the Microscopic Finite Elements

Macroscopic finite elements belonging to the group of the elements analyzed from the point of view of susceptibility to hot tearing are divided into hybrid microscopic finite elements [8]. The hybrid finite element mesh is generated on the basis of the characteristic dimension of the grains (grain radius), determined in the solidification simulation. The formed mesh can also be taken into account in further analysis of two areas of material properties: densely tangled dendrites (solid phase) and the layers separating them in a solid–liquid state [9].

The number of microscopic finite elements is determined by the surface area of a macroscopic finite element. Whatever its original shape, the hybrid elements mesh is always built on a rectangular plan (similar to a square) with an area equal to or close to the macro element area. Such an approximation of projecting a macroscopic element to microscopic elements is dictated by the polygonal shape of hybrid elements.

In the areas of separating layers there is not only the liquid phase, but also the solid phase in the form of dendrite arms. Therefore, the participation of the grains area in the region of the whole solid–liquid area can be written as:

$$q = \frac{A_g}{A} = \frac{f_s(1-u)}{1-u \cdot f_s}, \quad u \in \langle 0, 1 \rangle , \tag{17.4}$$

where: u is the part of the solid phase in the separating layers, while f_s is a volumetric solid state fraction. The course of q, depending on the fraction of the

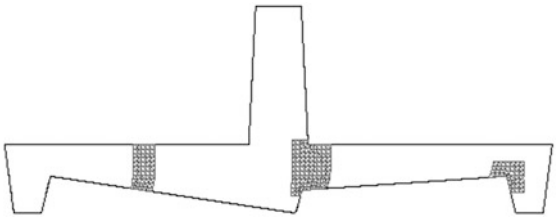


Fig. 17.2 Location of selected groups of macroscopic elements for the analysis of hot tearing

solid phase, including a displacement of half share of the solid phase to the area of layers ($u = 0.5$) and assuming that all of the solid phase is in the grain area ($u = 0$), is shown in Fig. 17.3.

Grain growth due to the increase in solid phase participation in the analyzed solid–liquid area was implemented through the appropriate displacement of the finite element mesh nodes, according to the relation:

$$x = x_c - \sqrt{\frac{q}{q'}}(x_c - x'), \tag{17.5}$$

where: x is the coordinate of the node, x_c is the coordinate of the so-called measure of the solid phase increase, while the primes denote the current location of the node and the output quotient of the grain area.

17.2.4 Calculation of the Stress in Microscopic Areas

The macroscopic calculation yields a number of instantaneous fields. Among them, the temperature profile, the characteristic grain's size and the stress field are relevant for further simulations. In the selected area the macroscopic finite elements are isolated from the rest of the elements mesh of the casting. The parameters describing the state of the macro elements are used as input for further calculations leading to the determination of the susceptibility to hot tearing. The equiaxial grain radius assigned to the macro element is used to determine the dimension of the hybrid finite elements. The solid phase participation function $f_s(t)$ and the temperature profile $T(t)$ are used to control the growth of the grains area and the change in material properties in successive time steps. The appropriate boundary conditions are formulated with the help of the stress tensor $\sigma(t)$ (see Fig. 17.4). Due to the lack of symmetry for loading the system, the stress tensor is converted into an equivalent tensor of main stresses. As the result of this approach it is possible to analyze only a quarter of the system, suitably mounted on symmetry axes and charged by the main stress. For the purpose of numerical modeling of the solid–liquid center cracking it was necessary to separate the macroscopic properties as the properties of the two subdivisions appointed in an experimental way before. The "amount" of the subdivision is determined by the share of its surface area q in the whole solid–liquid area. Thus, the value of material property W is:

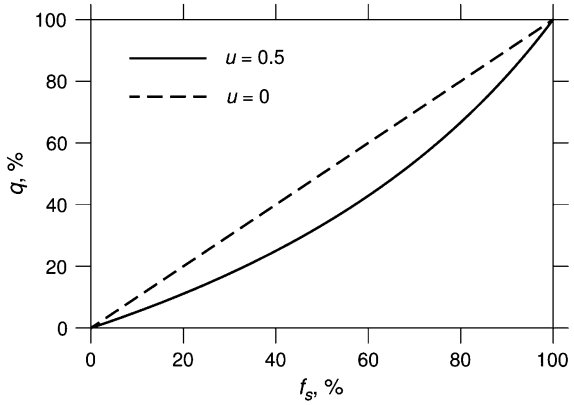


Fig. 17.3 The course of the function q for different u values

$$W = qW_g + (1 - q)W_l, \tag{17.6}$$

where: W_g is the value of material properties for the solid phase area (grains), while W_l is the value of material properties for the area of separating layers. Furthermore, it has been assumed that the material properties in the subdivisions are in a relationship expressed as:

$$\frac{W_l}{W_g} = p, \quad p \in (0, 1]\,. \tag{17.7}$$

Unlike pure metals, alloys solidify over a range of temperatures. Thus, Eq. (17.7) involves the distribution $p = p(T)$ of material properties that can be defined as follows

$$p(T) = \frac{T_L - T}{T_L - T_S}, \tag{17.8}$$

where: T_L and the T_S are the liquidus and the solidus temperatures, respectively. These temperatures determine the range of the solidification temperatures. Substituting Eq. (17.7) to (17.6) we obtain the relationship describing material properties for the solid phase:

$$W_g = \frac{W}{p + (1 - p)q}. \tag{17.9}$$

The microscopic finite element mesh, covering the analyzed solid–liquid area, is charged by the macroscopic state of stress. The boundary conditions are updated on grains arising from the simulation at the macro level. Material properties of sub-grains and separating layers are also determined with the use of the current temperature values. The calculations are carried out from the 'appearance' of stress, i.e., when the solid phase fraction exceeds a critical value (e.g. 25 %) until the complete solidification.

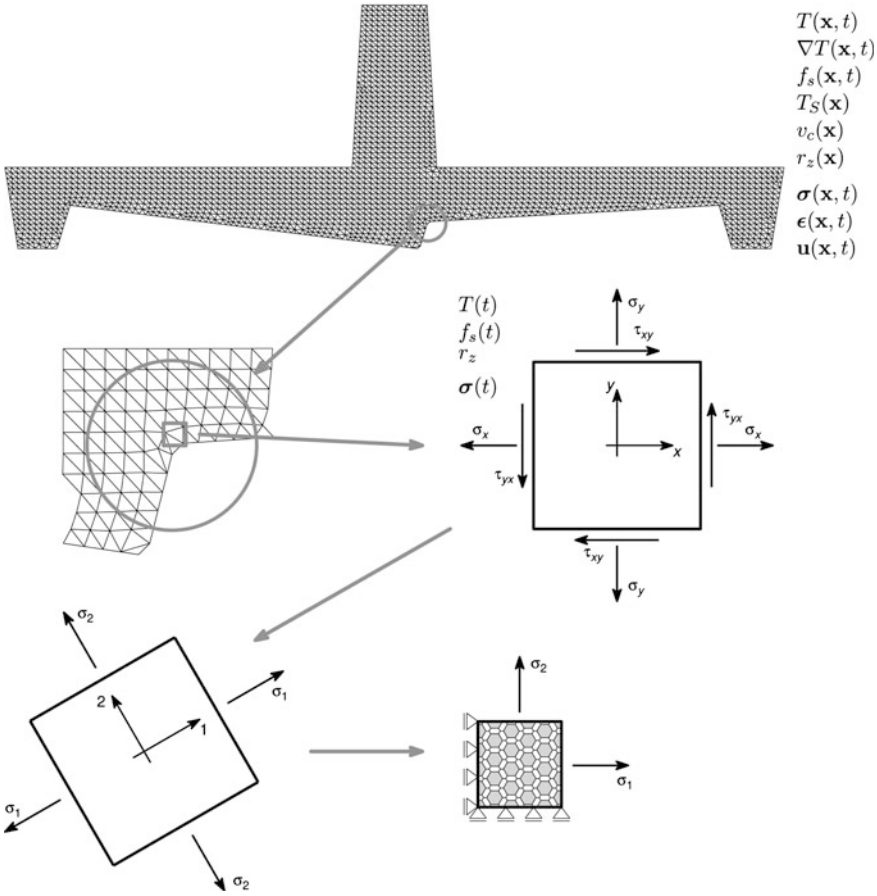


Fig. 17.4 The division of macroscopic (hybrid) finite elements and the way of supporting and loading of the analyzed area

17.2.5 Calculation of the Value of Θ for the Macroscopic Finite Element

The values of the local coefficient of susceptibility to hot-tearing Θ are calculated according to Eq. (17.3). Since different areas of the casting solidify at different time intervals it is convenient, for the sake of further analysis, to present the course of Θ in the function of the solid phase fraction. A sample graph presenting such a course is shown in Fig. 17.5. The presentation of the results in the function of the solid phase fraction enables a direct comparison of the coefficient value Θ of all the solved task variants.

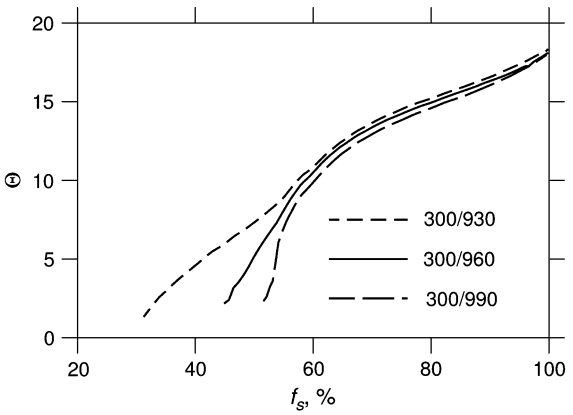


Fig. 17.5 Sample courses of T for given casting conditions (temperature of the mold/ casting temperature)

17.2.6 Drawing Up the Scale of Θ

In order to compare the values Θ for different tasks, different conditions for pouring and solidification have been drawn up on the scale of susceptibility to hot tearing, based on the critical value Θ_{cr}. It was assumed that the scale is dependent on the participation of the solid phase. The critical value Θ_{cr} is determined from the maximum values Θ for all the variants of the simulation for the solid phase participation, ranging from 50 to 95 %, in steps of 5 %. On the basis of the received values the function determining critical values of Θ in the function of the solid phase may be constructed. This function is the basis for determining the degrees of the susceptibility to hot tearing. Thus one should decide whether further analysis of susceptibility to hot tearing will run for four degrees. Values Θ larger and equal to Θ_{cr} have been adopted as a high (the highest) degree, as the average-values from 0, $9\Theta_{cr}$ to Θ_{cr}, as low degree-values from 0, $8\Theta_{cr}$ to 0, $9\Theta_{cr}$. For values Θ below 0, $8\Theta_{cr}$ the lack of susceptibility to hot tearing is accepted.

17.2.7 Execution of Diagram of Θ Distribution

Proposed in the previous section, the scale is the basis for drawing up diagrams (maps) of the coefficient Θ distribution for the main group of elements and for the control groups (see Fig. 17.2). The maps are drawn up for certain selected values of the solid phase participation (see Fig. 17.6).

Since the factor Θ distribution maps are only comparative, there are compared elements with the same fraction of the solid phase in a single casting. So they do not represent any real situations, i.e. those which may occur in the solidifying casting. Such maps are made to indicate that while the main group values Θ indicates the possibility of hot tearing, in the control groups the coefficient values Θ are so small, that they are not at risk from cracking.

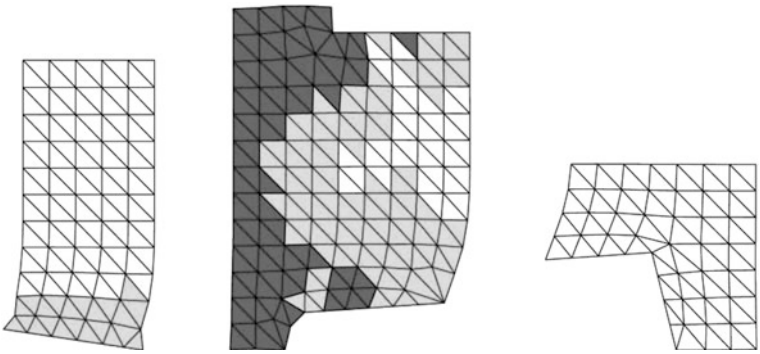


Fig. 17.6 Sample map of the coefficient Θ distribution (a *darker color* means a grater susceptibility to hot tearing)

17.2.8 Conclusions From the Simulations

After preparing maps of the coefficient Θ distribution some relevant, for the casting practice, conclusions arise. These conclusions may involve the casting hot tearing at different stages of solidification. What is also important, is the evaluation of infusion conditions, which determine the temperature of the mold or flooding temperature, to ensure obtaining sound castings.

17.3 Example of Application of the New Criterion

Application of the proposed criterion for the hot tearing evaluation has been illustrated by the simulations and analysis of the casting made of Al-2 % Cu alloy, solidifying in a metal form. For all the simulations, the initial mold temperature was set to 300 K. The variable parameter was the pouring temperature, that was equal to: 930, 960 and 990 K, respectively. Distributions of the local coefficient of susceptibility to hot tearing for the major group and control groups are presented in Fig. 17.7. The upper distributions were made for the pouring temperature 930 K, the middle—for 960 K and the lower distributions for 990 K.

The analysis shows that in all the cases, there is a high risk of the rupture of hot casting. It is therefore concluded that the initial mold temperature is too low. The obtained results were confirmed by experimental research. The hot tearing occurred for an initial mold temperature of 300 K, while raising the temperature to 600 K guaranteed to receive a sound casting.

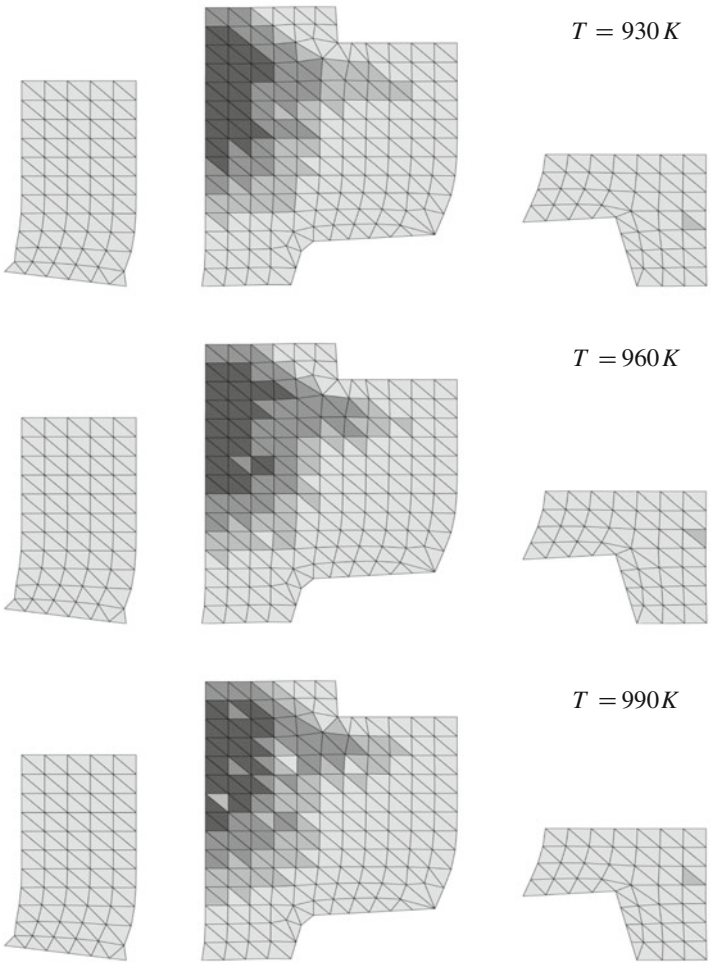


Fig. 17.7 The distribution of Θ for the solid phase fraction of 60 % in each macroscopic finite element

17.4 Summary

The proposed new stress criterion for hot tearing estimation in alloy castings is a local criterion, covering the area of a single macroscopic finite element. The application of this criterion for compact groups of finite elements, in selected casting areas, allows for a global evaluation of the casting susceptibility to hot tearing. The analysis of the susceptibility to hot tearing can be carried out jointly for several ranges of the initial and the boundary conditions. From the point of view of the rupture risk this analysis yields the most advantageous variant of the casting.

Prediction of hot tears appearance using simulation software is a complicated task because it requires detailed information about the evolution of stresses and strains in solidifying casting. Nevertheless some physically based indicators, as e.g. our Θ factor [9], RDG [3], PSD [10] or other criteria [11], can be derived using the data from the numerical simulation of solidification and then help engineers to manufacture defect-controlled alloys. Application of the proposed Θ criterion is time-consuming because it requires a lot of preparatory work and computer simulations. However, modern computing systems are powerful enough, so that our criterion can be taken into account and yield valuable results.

In contemporary manufacturing of metallic composites the designer has to relay on robust failure criteria that can be applied to optimize alloy performance and to control production processes. In the near future we plan to merge the presented estimation of susceptibility to hot tearing with some optimization procedure [12, 13] and apply them to modeling the mechanical properties of steel castings.

References

1. Lin Z, Monroe ChA, Huff RK, Beckermann Ch (2009) Prediction of hot tear defects in steel castings using a damage based model. In: Cockcroft SL, Maijer DM (eds) Modeling of casting, welding, and advanced solidification processes—XII. The Minerals, Metals & Materials Society, Warrendale, pp 329–336
2. Bellet M, Cerri O, Bobadilla M, Chastel Y (2009) Modeling hot tearing during solidification of steels: assessment and improvement of macroscopic criteria through the analysis of two experimental tests. Metall Mater Trans A 40(11):2705–2717
3. Rappaz M, Drezet J-M, Gremaud M (1999) A new hot tearing criterion. Metall Mater Trans A 30(2):449–455
4. Szwarc G (2003) PhD Thesis, Czestochowa University o f Technology. (unpublished)
5. Gosh S, Moorthy S (1995) Elastic-plastic analysis of arbitrary heterogeneous materials with the Voronoi Cell finite element method. Comput Methods Appl Mech Eng 121(1):373–409
6. Moorthy S, Gosh S (1998) A Voronoi cell finite element model for particle cracking in elastic-plastic composite materials. Comput Methods Appl Mech Eng 151(3–4):377–400
7. Parkitny R, Sczygiol N, Szwarc G (2002) Application of the hybrid finite element formulation to numerical modeling of hot tearing of castings. Proceedings of international symposium ABDM, 9–11 Sept 2002, Cracow, Poland
8. Desbiolles J–J, Droux J–J, Rapapaz J, Rappaz M (1987) Simulation of solidification of alloys by the finite element method, Computer. Phys Rep 6:371–383
9. Sczygiol N, Domański Z (2011) Numerical evaluation of hot tearing in the solidifying casting. Lecture notes in engineering and computer science 2011. Proceedings of the world congress on engineering and computer science 2011, WCECS 2011, 19–21 Oct 2011, San Francisco, USA, pp 7–12
10. Monroe C, Beckermann C (2005) Development of a hot tear indicator for steel castings. Mater Sci Eng A 413–414:30–36
11. Suyitno WHK, Katgerman L (2005) Hot tearing criteria evaluation for direct-chill casting of an Al-4.5 Pct Cu alloy. Metall Mater Trans A 36(6):1537–1546
12. Grzybowski AZ (2009) An optimal stopping rule for approaching a border that should not be crossed. Lecture notes in engineering 2009. Proceedings of the world congress on engineering 2009, WCE 2009, 1–3 July 2009, London, U.K., pp 1000–1002
13. Grzybowski AZ (2011) Monte Carlo analysis of risk measures for blackjack type optimal stopping problems. Eng Lett 19(3):147–154

Chapter 18
On Mathematics Software Equipped with Adaptive Tutor System

Hisashi Yokota

Abstract In this article, we describe how an educators' knowledge structure map is utilized to assess a knowledge state of a learner in college mathematics courses such as calculus and linear algebra. We also describe how an adaptive tutoring system is implemented into our mathematics learning software JCALC using the relative distance and the knowledge score.

Keywords Adaptive tutoring system · Concept map · Knowledge score · Knowledge state · Knowledge structure map · Relative distance

18.1 Introduction

Well known effective educational model for less prepared learners is one-on-one tutoring [1]. But, one-on-one tutoring is not a realistic solution for many learners because of cost. This motivated us to develop Intelligent Tutoring Systems (ITSs) with one-on-one tutoring capability for calculus and linear algebra for college level learners. Even though ITSs are becoming popular among learners at pre-college level mathematics courses [4], designing ITS which accurately diagnose learners' knowledge structure, skills, and styles is not easy. According to [7], to diagnose learners' knowledge structure, the generated question should be short answer question but not multiple choice questions.

H. Yokota (✉)
College of Engineering, Shibaura Institute of Technology,
307 Fukasaku Minuma-ku, Saitama 337-8570, Japan
e-mail: hyokota@shibaura-it.ac.jp

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_18,

In this article, how educators' knowledge structure map can be utilized to diagnose a learner's knowledge structure is shown. Then how the knowledge score can be used to assess a learner's understanding of the material is shown. Furthermore, how the relative distance is utilized for implementing an adaptive feedback system into JCALC is shown.

18.2 Assessing Learner's Knowledge

18.2.1 Experienced Mathematics Educator's Knowledge Structure

It is often said that experienced mathematics educators can often tell what causes him/her to make a mistake in the exam or what types of problems learners might fall into by grading exams or looking at what learners are writing. This forces us to study that how experienced mathematics educators can tell the cause of problems by reading a learner's solution written on the paper. Here, ten experienced mathematics educators are chosen from the mathematics department of our school. Then they are given the following learner's responses, and asked why these learners made mistakes.

(1) A learner writes $2(x^2 + 3x)^3(2x + 3)$ as to the question of "Find the derivative of $(x^2 + 3x)^4$".
(2) A learner writes $-\sin(3x + 1)$ as to the question of "Find the derivative of $\cos(3x + 1)$".
(3) A learner writes $-2/(x + 1)^2$as the answer to the question of "Find the derivative of $(x - 1)/(x + 1)$".
(4) A learner writes $xe^x + x + c$ as the answer to the question of "Evaluate $\int xe^x dx$".
(5) A learner writes $\det\begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix}$ as the answer to the question of "Find the (1, 2) minor of $\begin{pmatrix} 1 & 3 & 2 \\ 2 & -1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$".
(6) A learner writes $\det\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ as the answer to the question of "Find the (1, 2) cofactor of $\begin{pmatrix} 1 & 3 & 2 \\ 2 & -1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$".
(7) A learner writes $\begin{pmatrix} 1 & 3 & 2 \\ 2 & -1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$ as the answer to the question of "Find the cofactor expansion of $\det\begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix}$ along the 1st row".

Table 18.1 Typical responses by experienced mathematics educators

(1)	Every experienced mathematics educator has responded by saying that this learner knows how to differentiate the composite function and how to apply the chain rule. But the learner somehow made a mistake multiplying by 2 instead of multiplying by 4
(2)	Most of the experienced mathematics educators responded by saying that this learner knows how to differentiate the cosine function. But the learner probably does not know how to apply the chain rule
(3)	Most of the experienced mathematics educators agreed that this learner knows what to do. But this learner somehow memorized the quotient rule in the wrong way
(4)	Most of the experienced mathematics educators agreed that this learner knows about the integration rule called by parts. But the learner did not apply the integration rule correctly. So, the learner's knowledge about integration is not enough
(5)	Most of the experienced mathematics educators responded by saying that this learner knows that the minor of a matrix is given by determinant. But the learner does not know how to find it
(6)	Most of the experienced mathematics educators responded by saying that this learner may know a little bit about cofactor. But forgetting a sign means that his/her knowledge about cofactor is not enough
(7)	Most of the experienced mathematics educators responded by saying that this learner has no idea about cofactor expansion

The learners' responses are collected and sorted and shown in Table 18.1.

The experienced mathematics educators' responses can be explained by using a concept map. For example, consider the response (1). The experienced mathematics educators read the learner's solution $2(x^2+3x)^3(2x+3)$. Then they compared the learner's solution to the right solution. To do so, they have differentiated the given function by themselves. In other words, they have to recall the chain rule and apply it correctly. Furthermore, they have to recall the differentiation of the power function and apply it correctly within a short period of time. With all these process, they have noticed that the derivative of power function is essentially in the right form. Furthermore, the chain rule is applied correctly. Therefore, the experienced mathematics educators' response for the question (1) becomes like the one in Table 18.1.

Now notice that every experienced mathematics educator used the chain rule and the derivative of the power function. Thus, every experienced mathematics educators' knowledge structure is very similar. Even though the knowledge of an individual expert consists of both a cognitive element—the individual's viewpoints and beliefs, and a technical element—the individual's context specific skills and abilities [2, 6], experienced mathematics educators' knowledge structure can be used as the basic knowledge structure about how to solve problems in calculus and linear algebra.

18.2.2 Assessing Learner's Knowledge by the Relative Distance

By defining the ratio or the difference of the evaluated values of a learner's input and a generated correct answer, it is possible to assess a knowledge state of a learner. As in [10, 11], we first define the distance d: Let *vin* and *vca* be defined as follows:

$$vin = \text{the value of the input evaluated at certain point.}$$
$$vca = \text{the value of the correct answer evaluated at certain point.}$$

Then define the distance d as follows:

$$d = \left\{ \begin{array}{c} \text{the difference of } vin \text{ and } vca \\ \text{the ration of } vin \text{ and } vca \end{array} \right\}$$

For if a learner's input value is far from the correct value, the distance d becomes large. This type of phenomena can occur if a learner does not know a material at all or some. In this case, even an experience educator cannot conclude whether the learner knows a material a little or none. Thus, it is necessary for alternative way to assess learner's understandings.

Define rd by the following equation:

$$rd = \frac{d}{\text{evaluated value of correct answer}} \tag{18.1}$$

If the value rd is large, then d must be very large compared with the evaluated value of the correct answer. Then it is quite natural to assume that the learner does not know much about the material. On the other hand, if the value rd is small, it is natural to assume that the learner knows the material a little. With this reason the value of rd is called the relative distance, now to assess a learner's knowledge structure, the following example explains the usage of the relative distance. Suppose that the system generated question is given by "Differentiate $2(x^2 + 3x)^4$" and a learner's input is $2(2x + 3)(x^2 + 3x)^2$. Furthermore, the system generated solution is $8(x^2 + 3x)^2(2x + 3)$. Then the evaluated value of the learner's solution at $x = 1.315$ is equal to 2057.11, and the evaluated value of the system generated solution at $x = 1.315$ is 8228.45. Since the evaluated values of these two expressions are not equal to each other. Thus, it is possible to tell the learner's solution is wrong. Now, calculate the relative distance defined above. Then

$$rd = \frac{\left[8(x^2 + 3x)^3(2x + 3)|_{1.315} - 2(2x + 3)(x^2 + 3x)^3|_{1.315}\right]}{8(x^2 + 3x)^3(2x + 3)|_{x=1.315}} = \frac{3}{4} \tag{18.2}$$

Here rd is given by the simple fraction 3/4. Note that by the Sect. 18.2.1, the experienced mathematics educators assess the learner's knowledge structure by

Table 18.2 Performance criteria

Differentiation	Polynomials	Sum of derivatives
		Difference of derivatives
		Constant multiples
	Rational functions	Sum of derivatives
		Difference of derivatives
		Constant multiples
		Quotient rule
	Trig functions	Differentiation formula of sin x, cos x, tan x, sec x
		Sum, difference, product, quotient rule
		Differentiation formula of inverse trig functions
	Composite functions	Composition of polynomials, rational functions, trig functions, exponential functions, logarithmic functions, inverse trig functions, hyperbolic functions
		Differentiation formula of composite functions
		Derivatives of composite functions
	Higher order derivatives	Property of the second derivatives
		nth derivatives
		Leibnitz formula
	Applications of derivatives	Tangent line
		Normal line
		Taylor, MacLaurin expansion
		Estimating remainder term

checking each step necessary to obtain the right solution. This tells us that the learner's knowledge structure can be assessed by checking the relative distance. We also note that the experienced mathematics educators concluded that the learner probably made the simple mistake. Therefore, the relative distance is given by the simple fraction implies that the learner's knows the material, but made a careless mistake.

18.2.3 Assessing Learner's Knowledge Structure by the Experienced Mathematics Educator's Knowledge Map

To assess a learner's knowledge structure, one well known method is concept mapping. According to [7], to construct a good concept map, it is important to begin with a domain of knowledge structure that is very familiar to the person constructing the map. Following this suggestion, we first made sure that the learning outcomes of the subject such as calculus and linear algebra usually taught in college mathematics. Then the performance criteria for each concept for which learners are expected to learn is created. An example of the differentiation is shown in the following Table 18.2.

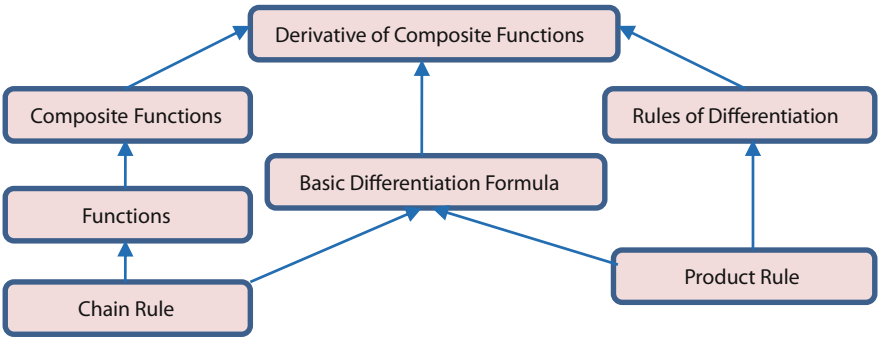


Fig. 18.1 The concept maps for derivative of composite functions

In the process of deciding what to expect for students, we learned that a relational is helpful. Note also that superior performance is dependent not only on domain knowledge but also on intimate familiarity with the relational structure of domain objects in a problem situation [5, 8]. This suggested that it is necessary for our mathematics software to be equipped with not only concept maps but also relational structures.

Now to check to see whether the learners understand the concept, questions containing necessary knowledge must be created and tested. This suggests that questions generated by our system must be split into finer questions which are more familiar to the learner. Then by examining how well learners answer to the finer short-answer questions, a concept map for each learner can be created. Note that in the analysis for the use of abstract concepts, a larger number of links were expected to be attached to abstract concepts in the high performer network than in the low network. Now using the knowledge structure of experienced mathematics educators, it is possible to identify the key concepts that apply to this domain. Thus, the concept map of experienced mathematics educators' knowledge structures is used to refine a short-answer question.

For example, the concept map of the differentiation of composite function is shown in the following Fig. 18.1.

Experts characteristically use more abstract concepts to solve a problem than novices [9]. Since experts chunk or group their knowledge differently, their mental models should be characterized by groupings around abstract concepts.

Now to implement the concept map into JCALC, we introduce the "knowledge score". The marks such as 0.2 and 0.3 on arrows are called "knowledge score" which indicate the basic knowledge needed to obtain a correct concept. Note that the knowledge scores on top row adds up to 1, and the knowledge score added vertically adds up to the one of top scores. In other words, to be able to differentiate a composite function, the knowledge about composite function consists of 20 %, the knowledge about the chain rule consists of 50 %, and the knowledge about the basic rules of differentiation consists of 30 %. These percentages are

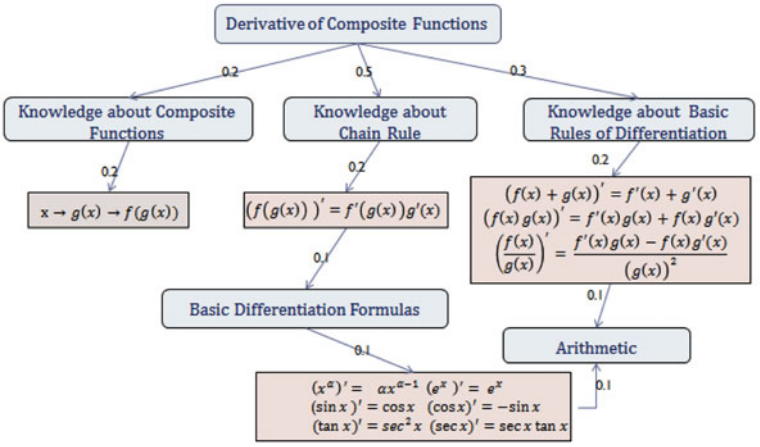


Fig. 18.2 Educators' knowledge structure map

derived by adding the necessary knowledge needed to acquire before completing the top row knowledge.

The explained knowledge structure map is shown in the following Fig. 18.2.

18.3 Adaptive Tutoring System JCALC

18.3.1 How to Tell a Right Answer from a Wrong Answer

Look at the question of finding the integral of the function "$1/1 + \sin x$". Then a learner's input such as "$\tan x - \sec x + c$" is a right answer. But another learner's input "$2\tan\frac{x}{2}\big/\left(1+\tan\frac{x}{2}\right)+c$" is also a right answer. Thus to judge a learner's input is a right answer or not, it is not possible to list the right answers and compare words by words. For this reason, it should be noted that any system which produces multiple choice questions and answers is not suited for college level mathematics. This suggests that to develop an adaptive system for college level mathematics, any system should be able to handle short-answer questions. Furthermore, according to [4], short-answer questions have the advantage of avoiding cueing rather than selecting or guessing from options supplied. Thus, in our system, a learner must enter his/her answer in the form of mathematical expressions. This also suggests that our system must be able to read a learner's input and be able to tell whether it is a right answer or not.

The expressions for right answers are not unique, rather unlimited. Thus, preparing all expressions for right answers in database, and checking whether learners' answer is in database to decide the learners answer is correct is not

plausible. So, developing some other way to tell right answers form wrong answers is expected.

Look at the values of the power of x, say at $x = 1.3$. Then the values of the power of x are given as follows: $x^2 = 1.69$, $x^3 = 2.197$, $x^4 = 2.8561$,…. Now note that the last digit's decimal place increases one place each time. This means that the counting the decimal places appear in expression, it is possible to tell which power of x the expression contains. Which in turn implies that two polynomial expressions are evaluated to be equal imply that they are exactly the same expressions. So to decide the learner input is a right answer or not, it is only necessary to evaluate the learner's input and the system generated solution at some point. Noting that every elementary function can be approximated by the polynomial, we can determine that a learner input is right answer or not by checking the value of a learner input and a correct answer at some point. More detailed discussion is given in [10].

18.3.2 Inferring Learner Knowledge Structure

It was shown in the Sect. 18.3.1 that it is possible to determine the learner input is right or not by evaluating the correct answer generated by JCALC and a learner input at certain value. We studied this method carefully to notice that when the value of the correct answer and the learner input are different, their difference or ratio has some tendency among group of learners. Suppose that a displayed question is "find a derivative of $(x^2 + 2x + 3)^4$" and a learner's input is "$(x^2 + 2x + 3)^3(2x + 3)$". Furthermore, the correct answer generated by JCALC is "$4(x^2 + 2x + 3)^3(2x + 3)$". Looking at the learner's input, anyone with calculus teaching experience judges that the learner has the knowledge of derivative of composite function because he/she has took care of derivative of power function then worked inside function.

Suppose this time that the learner input is "$4(x^2 + 2)^2$". Then again anyone with calculus teaching experience would say that this learner did not master the rule of derivative of composite functions. It is because the derivative of the inside function is taken before the derivative of the power function. This time it is not easy to design our system to judge the same way as the experienced mathematics educator. For learners inputs vary many ways and it is impossible to cover all.

Now as explained in Sect. 18.2.3, the knowledge score for each performance criterion is calculated. Then by adding the knowledge scores to the learner's knowledge structure, the complete knowledge structure of the learner can be obtained. Thus to infer a learner's knowledge structure, adding the knowledge scores for each question is only thing to do.

18.3.3 Feedback

It is noted in [3] that any educational software needs to give a quick feedback to encourage a learner to study more. This suggested that our system ought to have a few types of feedbacks. For example, using the relative distance explained in Sect. 18.2.2, the simple fraction *rd* can give the feedback like "Careless mistakes. Try again". The small value of the relative distance can give the feedback like "Close to the right answer. Try again". One more feedback needed to our system is the statement such as "Expression has not been simplified". One way to accomplish this is to compare the number of terms in the generated solution and the learner's solution. After a learner's input are read, interpreted, and the relative distance and the knowledge score are calculated, three different types of hints will be displayed. Since the knowledge score is supposed to assess a learner's knowledge structure, the sum of knowledge scores gives more valuable information about how much learner knows.

Now to show how adaptive hints are generated and displayed, all subjects and performance criteria of calculus and linear algebra are checked. Then short-answer questions are divided into 44 groups depending on the number of steps used to solve these questions. For example, when to generate a question of differentiating a product of functions, a collection of techniques which gives a product rule is searched. Then using the function generated, a hint explaining what to do to solve this question is displayed. If a knowledge score is less than 1, then using the displayed question, hint for which experienced mathematics educator might give will be displayed. If learner cannot get a right answer, another hint will be displayed.

18.4 Conclusion and Future Work

We implemented the assessing method explained above into JCALC and tested the assessing method is valid or not. To verify whether the hypothesis is true or not, we use the hypothesis test for slope of regression line. The null hypothesis is the slope $= 0$. Then we obtain the following results: the standard error SE is given by $SE = 0.253$, T-score is given by $T = 1.926$. From these, we obtain that $P(t > 0.1926) = 0.0415$. This shows that the p-value is less than the significant level (0.05), and we cannot accept the null hypothesis.

We are currently running our system JCALC on web and collecting learners' data as much as possible. The data collected contains the following information: learner's ID, subject selected, section selected, and number of questions tried, number of right answer, time spent for solving each question, expression inputted by learner, which type of feedback is shown, system generated solution. From the information above, each generated question is rated for difficulties. Then using this rating, the future system should be able to provide questions which emphasize to fill the learner's week point.

Acknowledgments This work was supported in part by Shibaura Institute of Technology, Grant-in-Aid for Scientific Research in 2011–2012.

We thank all colleagues and learners participated in this project and suggested useful ideas to refine our adaptive learning system.

References

1. Ainsworth RG (1995) Turning potential school dropouts into graduates: the case for school-based one-to-one tutoring. Research report 95–07. National Commission for Employment Policy, 35
2. Alavi M, Leidner DE (2001) Knowledge management and knowledge management systems: conceptual foundations and research issues. MIS Q Rev 25(1):107–136
3. Bana P (1999) Artificial intelligence in educational software: has its time come. Br J Educ Technol 30(1):79–81
4. Bloom BS (1984) Taxonomy of educational objectives. Pearson Education, Boston
5. Brehmer B (1980) In one word: not from experience. Acta Psychol 45:223–241
6. Keyes J (1990) Where's the ''expert'' in expert systems. AI Expert 5(3):61–64
7. Newble D, Cannon R (2000) A handbook for teachers in university and colleges: a guide to improving teaching methods. Kogan page Ltd, London
8. Rentsch J, Heffner T (1994) Group Organ Manag 19(4):450–474
9. Sembugamoorthy V, Chandresekaren B (1986) Functional representation of devices and compilation of diagnostic problem-solving systems. In: Kolodner JL, Reisbeck CK (eds) Experience, memory, and reasoning. Erlbaum, Hillsdale
10. Yokota H (2006) On development of e-math-learning system for short-answer type questions. Res Bull HIT, 40:319–325
11. Yokota H (2011) On a development an adaptive tutoring system utilizing educator's knowledge structure. Lecture notes in engineering and computer science. In: proceedings of the world congress on engineering and computer science 2011, WCECS 2011, 19–21 October 2011, San Francisco, pp 260–264

Chapter 19
Effect of pH on the Floatability of Base Metal Sulphides PGMs

Ayo Samuel Afolabi, Edison Muzenda and Saka Ambali Abdulkareem

Abstract This study investigated the effect of pH on the recovery and grade of the Platinum Group Metals (PGMs) and base metal sulphides from the UG2 ore of the bushveld complex. This was achieved through running a series of test work in a Denver flotation cell at varying pH 6–11 at constant reagent dosage. The UG-2 reef is characterized by two predominant gangue phases i.e., chromite and silicate, that have significantly different physical and chemical properties. The test work was aimed at evaluating which pH produces the best recoveries, and finding the effect of the chrome content in these recoveries. A pH of 9 produced the highest recovery compared to other pH values. However, the highest PGM grade was attained at a pH of 6 which is slightly acidic. Ideally this trend could be expected since the collectors (xanthates) are more stable in alkaline medium. The higher PGM recovery was also accompanied by higher chrome content as a result of their similar chemical properties.

Keywords Base metals · Bushveld · Collectors · Gangue phases · Platinum group metals · Recoveries · Sulphides

A. S. Afolabi (✉) · S. A. Abdulkareem
Department of Chemical Engineering, University of South Africa, Bag X6, Florida, Johannesburg, 1710, South Africa
e-mail: afolaas@unisa.ac.za

S. A. Abdulkareem
e-mail: kasaka2003@yahoo.com

E. Muzenda
Department of Chemical Engineering, University of Johannesburg, 17011 Johannesburg, 2028, South Africa
e-mail: emuzenda@uj.ac.za

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_19,

19.1 Introduction

This is a continuation and extension of our previous work on effect of pH on the recovery and grade of base metal sulphides by flotation [1]. Flotation is described as a physio-chemical process that exploits the differences in the electrochemical properties of mineral surfaces. It depends directly on the nature and properties of mineral-water interface. Flotation therefore, depends directly on the nature and properties of mineral-water interface. Other two important factors that influence the flotation process are the interaction of water molecules with the mineral surface, both in liquid and gaseous environments, and the electrical double layer at solid-water interface [2]. Flotation depends on a number of factors that affect the degree of recovery of the mineral to be extracted. For instance, the platinum group metals are extracted by selective process that can be used to achieve specific separation from the ore and this is flotation. One needs to grind the pulp so as to liberate the minerals and separate them using flotation. The froth flotation process is used to extract the desired mineral from the unwanted gangue; this is done by treating a finely ground ore with chemical reagents. The efficiency of flotation is dependent on a number of factors which among others are: particle size, pulp density, reagent dosage and pH. The principle of unlocking the minerals is induced by comminution process. Most minerals are finely disseminated and intimately associated with the gangue, they must be initially "unlocked" or liberated before separation can take place or can be undertaken. This can be achieved by comminution, in which particles size of the ore is progressively reduced until the clean particles of minerals can be separated by such methods as are available as flotation [3]. The effectiveness of froth flotation is limited to a narrow particle size range of about 10–100 μm [4]. For the sizes above or below the range the effectiveness is reduced/decreased. Over grinding leads to particles being too fine to float and the loss of minerals to slimes [5]. Under grinding on the other hand hinders the valuable metals to float due to the unlocked metal bearing mineral, therefore this will increase the tailings grade thereby reducing the overall recovery. Pulp density is another parameter that plays an important role in the addition of reagents as the dosages are calculated in grams per ton, therefore it is of utmost importance to maintain a constant pulp density throughout for the correct dosages to be maintained. Also affecting the efficiency of froth flotation is the reagent dosage. The addition of reagents to the pulp is for the manipulation of the pulp chemistry and to enhance differences in the surface hydrophobicity to facilitate separation [6]. In optimising a flotation plant it is important to consider the effect of the various parameters on the recovery and grade of precious metals. The process is sometimes complicated due to the presence of gangue phase that has a severe consequence in the downstream processing of flotation concentrate. The reagents normally used are in the froth flotation process are collectors, frothers, depressants and activators.

The addition of flotation reagents in the flotation process is to selectively render the surfaces of mineral particles either hydrophobic or hydrophilic as a result of

the ionic interactions and exchanges that can occur at the double layer. During which the hydrophobic particles become attached to air bubbles and are carried upwards through the slurry to a froth layer that forms at the top of the flotation cell. This froth layer is removed and usually becomes the concentrate. Hydrophilic particles remain in the slurry in the flotation cell. It has been reported that most minerals are naturally hydrophilic i.e., polar in molecular structure and they need collectors to render them hydrophobic (non-polar in molecular structure). Collectors are heterogeneous compounds with a functional inorganic group and a hydrocarbon chain. There is a difference in electric charge on the ends of the inorganic group caused by an uneven distribution of polar bonds on its ends. The hydrocarbon chain is non-polar and has no difference in charge between its ends. The inorganic group is the portion which adsorbs on the mineral surface (provides hydrophobicity), causing the mineral to attach itself to the bubble. The most common mineral types that are recovered this way are sulphides and the widely used sulphide collectors are sodium and potassium salts of certain acids containing a hydrocarbon group. These are anionic collectors and include xanthates and dithiophosphates [7–9]. The use of co–collectors or promoters in flotation is well established [10]. For instance, it has been reported that the most important aspects of froth flotation is the formation of a froth in which the valuable minerals are retained for further Upgrading. Thus the presence of a surface-active or frothing agent, either as a neutral frother or in the dual function as collector and frother, is vital to the process [11]. When mineral surfaces have been made hydrophobic by the use of a collector, they have to attach to a stable air bubble for them to be recovered. The stability of the bubble depends on the type of frother used. A good frother should have no collecting power but should be stable enough to ensure that the floated minerals are transferred from the float cell to a collecting launder [12]. Frothers are therefore, heteropolar organic reagents which are capable of being adsorbed on the air–water interface. The heteropolar structure of the frother molecules makes non-polar group to orientate towards air and the polar groups towards water. Frothers must be to some extent soluble in water, otherwise they would be distributed very unevenly in an aqueous solution and their surface-active properties would not be fully effective. The alcohols (OH) are the most widely used frothers, since they have practically no collector properties, and in this respect are preferable to other frothers, such as the carboxyl, which are also powerful collectors [12]. During froth flotation, the ore is crushed and ground to a specific size sufficient for mineral liberation. The ore is then suspended in slurry and mixed with reagent or collectors. The collectors react with sulphide mineral particles to make them hydrophobic. The treated ore is introduced to a water-filled aeration tank and a frother (usually alcohol based) is added. Air is then induced and the air bubbles attach to the hydrophobic minerals are skimmed off. These skimming are generally subjected to a cleaner-scavenger cell to remove excess silicates and to improve the grade of the final product which is sent for down-stream processing [12]. It is therefore important to state that in practice, selectivity in complex separation is dependent on a delicate balance between reagent con-centration and pH [3]. This is an indication that the reagents together with pH

should have an optimum balance so as to get optimum results. This study reports the effect of pH on the floatability of base metal sulphides on PGMs recovery of UG2 ore. This is aimed at evaluating the effect of varying pH on the recovery of PGMs. This project is carried out using the UG2 ore of the Rustenburg Bushveld complex.

The Bushveld complex is the world's largest known source of PGMs with very complex sulfides and PGMs mineralogy. Depending on the type of reef and geography the Bushveld complex is the predominant PGM mineral types and their association and can vary to a great degree [13]. There are about fourteen type locality platinum minerals that have been discovered in the Bushveld deposit. There are also 150 unnamed and, the most part, inadequately characterized platinum group phases are recorded from the Bushveld complex. Over 50 % of these unnamed minerals are documented from the UG2 chromitite, majority of which are platinum phases. The UG2 layer is the host unit of most Bushveld minerals such as Iridium, Osmium, Rhodium, Ruthenium, Platinum and Palladium [14–17].

19.1.1 Importance of pH

It is evident that from the fore going that pulp alkinity plays a very important and complex role in flotation. In practice, selectively the separation is dependent on a delicate balance between reagent concentration and pH. Flotation where possible is carried out in a alkaline medium, as most collectors including xanthates are stable under these conditions [6]. The alkalinity of the pulp is maintained by addition of lime (sodium carbonate) to a lesser extent sodium hydroxide or ammonia. The lowering of the pH can be attained by the addition of sulphuric acid or sulphurous acid. These chemicals are often used in very significant amounts in almost all flotation operations [6].

The pH regulators are cheaper than the frothers and collectors, however the overall costs is generally higher with the pH regulators per ton of ore treated. For example: the cost of lime in sulphide minerals flotation is roughly double the amount of the collector used, so the significant operation cost saving can be attained by proper selection of the pH regulators [6]. Lime being the cheapest, is widely used to regulate pulp alkalinity and used in the form of milk of lime, a suspension of calcium hydroxide particle in a saturated aqueous solution. Lime or soda ash is often added to the slurry prior to flotation to precipitate heavy metal ions from solution. In this sense the alkali is acting as "deactivator" as these heavy metal ions can activate sphalerite and pyrite and prevent their selective flotation from lead or copper minerals. Since the heavy metal precipitated by the alkali can dissociate to a limited extent and thus allows ions into solution, cyanide is often used with the alkali to complex them. The hydroxyl and hydrogen ions modify the electrical double layer and the zeta potential surrounding the mineral; hence the hydration of the surface and their floatability is affected.

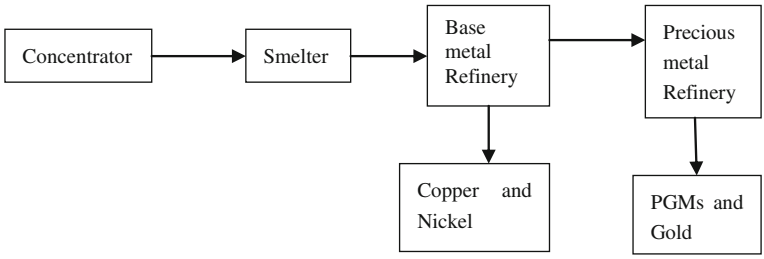


Fig. 19.1 The platinum group metal supply chain

19.2 Mineralogy of the UG2 Ore Body

The UG2 ore is a platinum- bearing chromitite rich ore that contributes a growing proportion of platinum group metals production from the Bushveld Ingenious complex in the northern region of South Africa. This ore is ideal for study of entrainment as it contains two major gangue phases which are chromite and pyroxene that are significantly different in chemical and physical properties however, poses certain problems and challenges [18]. High recovery of valuable minerals from UG2 is accompanied by high chrome recoveries to the concentrate and this is detrimental to the downstream smelting process. This is due to spinal formation, which is an insoluble species in respect to the conditions that prevail in the smelters Furthermore; these species are of intermediate density (between matte and slag) and form a mushy layer at the matte-slag interface [18].

The UG2 consists mainly of Chomitite of about (60–90 %) and Cr_2O_3 content averages of 43.5 %. The dominant base metal sulphides are mainly pemtalandite and chalcopyrite, with lasser pyrrhotite, pyrite, arsenopyrite, bornite chalcocite, covellite galena and millerite. Approximately 85 % of base metal sulpides are present at grain boundaries, 4 % occur within chromite and the remainder occurs within silicates grains [19]. The platinum group metals (PGMs) assemblage of UG2 chromitite is dominated by Pt–Pd sulphides (35 %) and laurite (30 %) with Pt–Fe alloy and intergrowths (21 %), Rh sulphide (11 %) and Pd alloys (3 %) [19]. Figure 19.1 illustrates the first metallurgical stage in the production of Platinum group metals and gold which is the production of a concentrate as a feedstock to the smelter. The smelter—converter stage is followed by a refinery of the base metal oxides and lastly the precious metal refinery which process the final products which are the PGMs and gold [20].

19.3 Materials and Methods

A UG2 ore sample was crushed in stages using a laboratory jaw and cone crusher to 45 % passing 75 μm. The crushed ore was then blended and rotary split to 1 kg representative sub-samples for the test work. The test work was carried out in a

Table 19.1 Reagent dosages and floatation times

Milling (45–75 %)	Time (mins)	SIBX (g/t)	Sascol 105	M47	DF 200 (g/t)
Condition 1	2	40	40		
Condition 2	2			40	10
Float 1	2	2			
Float 2	4	4			
Condition 3	2	2	10	10	
Condition 4	2			10	10
Float 8	8				
Float 8	8				
Float 8	8				

Denver flotation cell of 1 kg capacity. All tests were conducted using the standard flotation procedure. The milled ore was then agitated in a Denver machine to ensure homogenous suspension of solids. Reagents specific to each test were added and conditioned as illustrated in Table 19.1, thereafter air was manually induced using an air rotameter. The concentrates were collected by manually scraping the froth using scraper blades. Either sulphuric acid or lime was used to achieve the desired pH. The flotation studies were performed at pHs of 6, 8, 9, 10 and 11. To ensure reproducibility the runs were carried out in quadruplets. The runs were carried out in 4 batches. Sulphuric acid and lime were added in a proportion so as to achieve the desired pH. The following pHs were studied 6.0, 8.0, 9.0, 10.0, 11.0. The pH prior flotation was measured; sulphuric acid or lime was used as a pH regulator added to get the desired pH. After flotation, wet samples were dried in an oven, then weighed and lump broken. Concentrates and tailings were put in separate packages to avoid contamination and these were analyzed for PGMs.

19.4 Results and Discussion

Figure 19.2 shows PGMs recovery versus cumulative recovery. The highest grade was achieved at pH of 6 while the highest recovery was obtained at a pH of 9. According to Wills [6], the pulp alkalinity plays a very important role in flotation. Xanthate which is the collector used in this work is stable in alkaline medium. It selects mineral water repellent by adsorption of molecule or ions to the mineral surface [21]. The grade and recovery of PGMs are at the lowest values at a higher pH of 11. This is because the collection capacity of xanathates is reduced when they become more stable at higher pH values. Wills [6] emphasized the balancing of reagents suites with pH for maximum activation for all reagents to achieve both high grades and recoveries.

UG2 ore has a lot of chrome content with similar chemical properties to those of the PGMs and base metals. During flotation the amount of chrome recovered with PGMs need to be minimized to avoid smelter feed containing high chrome content. The relationship between chrome grade and recovery is shown in Fig. 19.3.

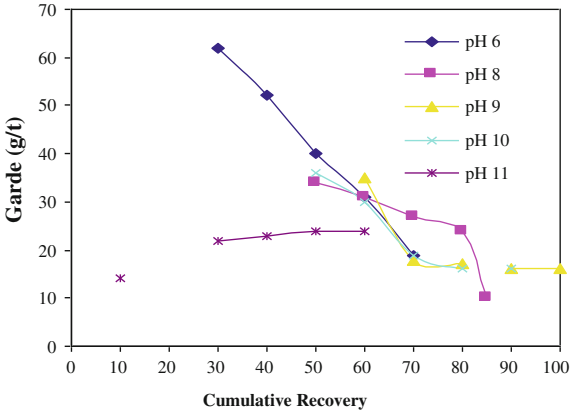


Fig. 19.2 PGMs grade versus cumulative recovery

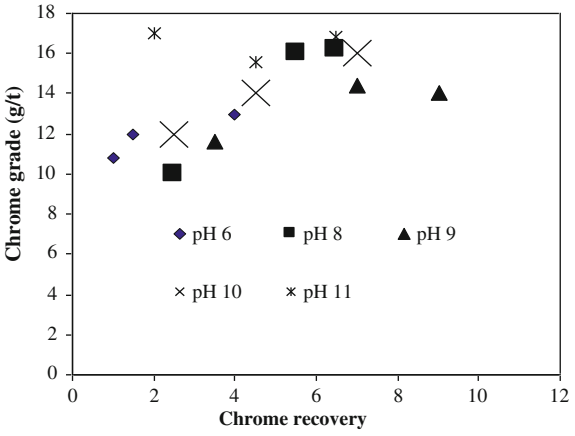


Fig. 19.3 PGMs grade versus cumulative recovery

The highest chrome grades are achieved at pH of 11 whilst the highest recovery is at pH of 9. High recoveries for both PGMS and chrome were achieved at pH of 9 supporting the theory that higher PGMs recovery is coupled with high chrome recoveries. The high recovery of valuable minerals from UG2 ore accompanied by high chrome recoveries to the concentrate is detrimental to the downstream smelting process. This is due to spinal formation, which is an insoluble species in respect to the conditions that prevail in the smelters. In addition, these species are of intermediate density (between matte and slag) and form a mushy layer at the matte-slag interface [17]. For effective flotation of valuable minerals, gangue materials in the pulp should be depressed. This is complicated as certain silicate gangue phases are activated in alkali solutions [21]. Figure 19.4 shows recovery versus time relationship. The highest and lowest recoveries with time were achieved at pH values of 9 and 11 respectively.

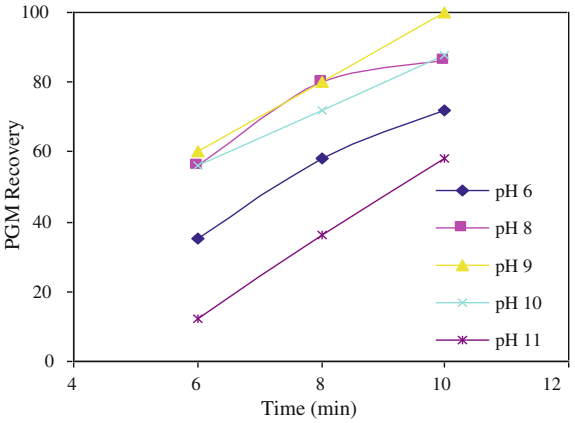


Fig. 19.4 PGMs grade versus cumulative recovery

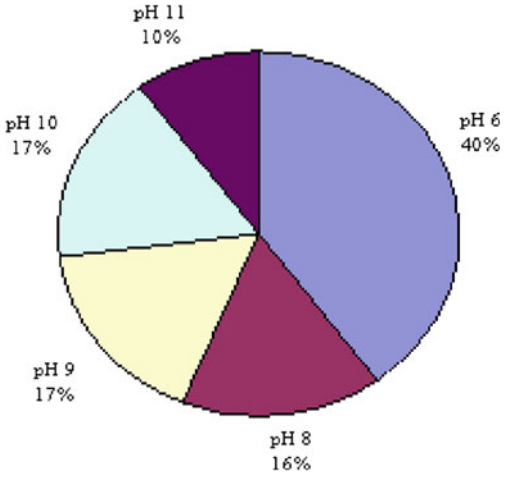


Fig. 19.5 Recovery of Copper and Nickel at various pHs

The recovery of copper and nickel at various pHs is represented in Fig. 19.5. Highest recovery was achieved at pH of 6 indicating that the recovery of base metals is favoured by acidic conditions. About 40 % of the recovery was achieved at pH of 6 compared to about 10 % at pH of 11. As discussed above the high grade of PGMs attained at low pH is accompanied by high recovery of base metals.

19.5 Conclusion

This work demonstrated the importance of pH and its balancing with reagent dosages to improve both recovery and grade. In future studies it is recommended to widen the pH range and to find ways of reducing high chrome recoveries

associated with high PGM recoveries. The recovery behaviour of chrome with time and its dependence on pH was found to be similar to that of PGMs. The results of this work can be used to optimize industrial flotation plants.

Acknowledgments The authors gratefully acknowledge the financial supports of the National Research Foundation (NRF) and Universities of South Africa and Johannesburg.

References

1. Muzenda E, Afolabi AS, Ntuli F, Abdulkareem AS (2011) Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2011, WCECS, San Francisco, USA, pp 609–612, 19–21 October 2011
2. Fuerstenau DW (1982) Mineral-water interfaces and the electrical double layer, in: principles of flotation. In: King RP (ed) (SAIMM, Monograph, Johannesburg Series, 1982) pp 17–30
3. Wills BA, Napier-Munn T (2006) Wills' mineral processing technology: an introduction to the practical aspects of ore treatment and mineral recovery (Butterworth-Heinemann 2006). Elsevier Publisher, Great Britain
4. Tasdemir A, Tasdemir T, Oteyaka B (2007) The effect of particle size and some operating parameters in the separation tank and the downcomer on the Jameson cell recovery. Miner Eng 20:1331–1336
5. Fredriksson A, Holmgren A, Forsling W (2006) Kinetics of collector adsorption on mineral surface. Miner Eng 19:6–8
6. Wills BA, Napier-Munn T (1997) Wills' mineral processing technology: An introduction to the practical aspects of ore treatment and mineral recovery 7th edn. Elsevier, Great Britain
7. Wiese J, Harris P, Bradshaw D (2005) The influence of reagent suite in the flotation of ores from the Merensky reef. Miner Eng 18:189–198
8. Wiese J, Harris P, Bradshaw D (2005) Investigation of the role and interactions of dithiophosphate collector in the flotation of sulphides from the Merensky reef. Miner Eng 18:792–800
9. Wiese J, Harris P, Bradshaw D (2006) The role of reagent suite in optimizing pentlandite recoveries from the Merensky reef. Miner Eng 19:1290–1300
10. Kelebek S, Demir U, Sahbaz O, Ucar A, Cinar M, Karaguzel O, Oteyaka B (2008) The effect of dodecylamine, kerosene and pH on batch flotation of Turkey's Tuncbilek coal. Int J Miner Process 88:3–4
11. Harris PJ (1982) Principles of flotation: mineral-water interfaces and the electrical double layer. S Afr Inst Min Metall 3:237
12. Hughes TC (2005) AM-2 a hydroxamate flotation collector reagent for oxides and oxide mineral systems. vol 3. Aust J Min, 58–59
13. Bruckard WJ, Kyriakidis I, Woodcock JT (2007) The flotation of metallic arsenic as a function of pH and pulp potential—a single mineral study. Int J Miner Process 84:1–4
14. Peyerl W (1983) The metallurgical implications of the mode of occurrence of platinum group metals in Merensky reef and UG2 chromitite of the Bushveld igneous complex, vol 7. Special Publication of Geology Society of South Africa, South Africa pp 295–300
15. Schouwstra P, Kinloch ED (2000) A short geological review of the Bushveld complex. Amplats Research Centre, South Africa
16. Ballhaus C, Sylvester P (2000) PGE enrichment processes in the Merensky reef. J Petroleum 41:454–561
17. Cawthorn RG, Merkle RKW, Viljoen MV (2002) Platinum—group elements deposits in the Bushveld complex, South Africa. In: Cabri LJ (ed) The geology, geochemistry, mineralogy, mineral benefiation of the platinum group elements, vol 54. Canadian Institute of Mining, Metallurgy and Petroleum, Canada pp 389–430

18. Cilek EC (2009) The effect of Hydrodynamic conditions on true flotation and intrainment flotation of complex sulphide ore. Int J Miner Process 90(1–4):34–44
19. Viljo AM, Viljoen B, Van Wyk E, Van Heerden FR (1998) Distribution and chemotaxonomic significance of flavonoids in Aloe (Asphodelaceae). Plant Syst Evol 211:31–42
20. Valenta MM (2007) Balancing the reagent suite to optimize grade and recovery. Miner Eng 20(10):1–6
21. Wiese J, Harris P, Bradshaw D (2007) The response of sulphide and gangue minerals in selected Merensky ores to increased depressant dosages. Miner Eng 20:986–995

Chapter 20
Investigation of Cu (II) Removal from Synthetic Solution by Ion Exchange Using South African Clinoptilolite

John Kabuba, Edison Muzenda, Freeman Ntuli and Antoine Mulaba-Bafubiandi

Abstract The objective of this study was to investigate the effect of NaCl, KCl and acid (HCl), on South Africa clinoptilolite used as an adsorbent in the ion-exchange process for the removal of cations (Cu II) from wastewater. The kinetic parameters such as ΔH, ΔS and ΔG affecting the adsorption of Cu (II) ions were studied. The adsorption of Cu (II) from synthetic waste water was found to be dependent on pH, temperature, contact time and initial adsorbate concentration. The pH was varied from 2.5–6 and the optimum pH for Cu (II) removal was found to be 4.0. The removal of Cu (II) ions increased with time and attained saturation in about 60–70 min. The equilibrium data showed that the adsorption was endothermic in nature. Kinetics data showed that at higher temperatures, the rate of adsorption is higher for the clinoptilolite in natural zeolite and that Langmuir equation successfully described the adsorption process.

Keywords Adsorption · Clinoptilolite · Copper removal · Langmuir equation · Ion exchange · Kinetics · Saturation

J. Kabuba · E. Muzenda (✉) · F. Ntuli
Department of Chemical Engineering, University of Johannesburg, Doornfontein, PO Box 17011 Johannesburg, 2028, South Africa
e-mail: emuzenda@uj.ac.za

J. Kabuba
e-mail: johnk@uj.ac.za

F. Ntuli
e-mail: fntuli@uj.ac.za

A. Mulaba-Bafubiandi
Faculty of Engineering and the Built Environment, School of Mining, Metallurgy and Chemical Engineering, Minerals Processing and Technology Research Center, University of Johannesburg, PO Box 17011 Johannesburg, South Africa
e-mail: amulaba@uj.ac.za

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_20,

20.1 Introduction

This work is an extension of our previous work [1]. Heavy metals are common pollutants found in various mining and industrial discharges. As environmental regulations on heavy metals discharge are getting stricter and tighter, more efficient remediation methods for waste water are required [2]. Waste streams containing low to medium levels of heavy metals are usually found in metal plating facilities, mining operations, fertilizers, chemical, pharmaceutical, electronic device manufacturing and many others [3]. The United States Environmental Protection Agency in 1978 published a list of organic and inorganic pollutants including copper found in wastewater, and which constitute serious health hazards. Copper is one of the most important metals often found in industrial effluents such as acid mine drainage, galvanizing plants, natural ores and municipal wastewater treatment plants. It is not biodegradable and it travels through the food chain via bioaccumulation [4].

Removal of copper from waste water is crucial and its toxicity for human beings is at levels of 100–500 mg/day [5, 6]. The World Health Organization in 2006 recommended 2.0 mg/l as the maximum acceptable concentration of copper in drinking water. Recently, various studies have focussed on the removal of toxic heavy metal ions from sewage, industrial and mining waste effluents. The presence of heavy metals in streams and lakes has been responsible for several health problems in animals, plants and human beings [7]. Available methods for heavy metal remediation include chemical precipitation, the most economic but inefficient for dilute solutions; adsorption and reverse osmosis, generally effective but have the drawback of fouling as well as high maintenance and operation costs. Ion exchange is among the few promising alternatives for this purpose especially when low cost natural adsorbets such zeolites, clay material and agricultural wastes are used [8].

Natural zeolites are alumino silicate minerals with high cation exchange capacities and heavy metals selective properties [9]. The zeolites structure consist of a three dimensional arrangement of SiO_4 and AlO_4 tetrahedral. The aluminum ion is small enough to occupy the position in the center of the tetrahedron of four oxygen atoms, and isomorphous replacement of Al^{3+} for Si^{4+} results a negative charge in the lattice. The net negative charge is balanced by the exchangeable cation (sodium, potassium and calcium) and these are exchangeable with certain heavy metal cations in the solution. The relative innocuous nature of zeolite exchangeable ions makes zeolites suitable in the remediation of heavy metals from contaminated streams [3].

Society is eager for new and innovative ideas particularly in health and environmental matters. Zeolites offers a low cost and environmentally safe method of treating municipal water supplies, domestic, industrial and mining waste-water discharge.

The aim of this work was to remove copper from synthetic water using South African clinoptilolite zeolite. In this study the temperature, pH, time and as well as zeolites and synthetic Cu^{2+} solution concentrations were varied.

20.2 Adsorption Isotherm Studies

Ion-exchange isotherms studies are of fundamental importance in the design of ion-exchange systems because they indicate the nature of partitioning between the adsorbent and liquid phases at equilibrium as a function of ion concentrations [10]. When adsorbent and ion solutions are in contact, the concentration of ions on the adsorbent will increase until a dynamic equilibrium is reached.

At this point, a defined distribution of ions between the solid and liquid phases exists [10, 11]. Adsorption isotherms are widely used and are of importance in the optimum use of adsorbents.

Freundlich and Langmuir isotherms were used in this study to describe the adsorption process.

20.2.1 Langmuir Isotherm

The Langmuir isotherm is applied to the ion-exchange data using the linear expression of [12] in Eq. (20.1).

$$\frac{Ce}{qe} = \frac{1}{K_l b} + \frac{Ce}{K_l} \tag{20.1}$$

where q_e is the amount of cation-exchanged per unit weight of clinoptilolite at the equilibrium (mg/g) and is expressed as in Eq. (20.2).

$$q_e = [(C_o - C_e)V]/M \tag{20.2}$$

V is the volume of solution, M the amount of clinoptilolite added to the solution, Ce is the metal concentration in the aqueous phase, b is the maximum adsorption capacity and K_1 is the Langmuir constant related to the ion-exchange capacity and energy of ion-exchange respectively.

20.2.2 Freundlich Isotherm

It is one of the widely used mathematical descriptions which fit experimental data over a wide range of concentrations.

The Freundlich Isotherm [13] expressed as in Eq. (20.3) relates the ion exchange capacity to surface heterogeneity as well as the exponential distribution of active sites and their energies.

$$\log q_e = \log K + \frac{1}{n} \log C_e \tag{20.3}$$

Where K is the Freundlich constant related to the ion exchange capacity of the sorbent, and $1/n$ is the Freundlich constant related to the energy heterogeneity of the system and the size of the exchanged molecule.

20.3 Materials and Methods

20.3.1 Materials

The Natural zeolite used in this study was supplied by Prattely South Africa and was sourced from the Vulture Creek in KwaZulu Natal province of South Africa. The sample was crashed and pulverized to 80 % passing -75 μm then analyzed using a scanning electron microscope (SEM). Synthetic wastewater was prepared at five different Cu^{2+} ion concentrations of 0.361, 1.099, 1.969 and 2.748 g/l [14].

20.3.2 Experimental Procedure

Fifty-two clinoptilolite zeolite samples were prepared and these were exposed to different conditions to optimize the removal of Cu^{2+} ions. For pH variation, 16 zeolite samples of 10 g each were prepared. These were activated at 30 °C for 24 h at varying HCl concentration of 5.726, 3.181, 1.725 and 0.255 M. The adsorption studies were performed at varying pHs of 2.5–6. The rest of the samples were prepared for temperature, zeolite concentration and time variation investigations. These were activated at a pH of 4 corresponding to a concentration of 0.255 M at 30 °C for 24 h. The procedure involved the continuous mixing of zeolite samples in aqueous solutions in rolling bottles for 24 h. The solids were separated from solution by filtration. The filtrate was titrated with 0.1 M NaOH solution to determine HCl concentration after adsorption. The zeolite solids, separated from the solution, were dried at 90 °C and then analyzed by SEM. For pH variation, 10 g/l of zeolite was mixed with 1L synthetic water solution at concentration of 0.361, 1.099, 1.969 and 2.748 g/l and held in a closed polyethylene flask at 90 °C for 24 h. Zeolite concentrations were varied from 4 to 10 g for the same pH variation conditions. Temperature was varied from 30 to 90 °C at pH of 4 and the same conditions used to study the influence of pH and zeolite concentration were employed. To study the effect of contact time, samples were analyzed at 10 min intervals at pH of 4 and temperature of 90 °C. Zeolite and synthetic waste water concentrations were the same as for pH and temperature variation [1].

Table 20.1 Composition of the natural clinoptilolite as determined by XRD

Ion in clinoptilolite	Raw zeolite Abundance (%)	Activated zeolite Abundance (%)
SiO_2	74	70
Al_2O	12.4	11.6
K_2 O	3.8	2.6
Fe_2 O_3	1.5	1.0
Na_2O	1.3	0.9
CaO	1.5	1.3
MgO	1.1	0.8
Cu^{2+}	3.5	3.5
Cl^-	–	3.4
TiO_2	0.2	0.2

20.4 Results and Discussion

20.4.1 Clinoptilolite Characterization

The mineralogical composition for natural and activated clinoptilolite zeolite is shown in Table 20.1. After the activation with HCl, clinoptilolite was incorporated with significant amount of Cl^-.

The SEM micrographs of natural and activated zeolite are shown in Figs. 20.1 and 20.2 respectively. The SEM micrographs show a more open structure of the activated clinoptilolite compared to natural zeolite due morphological changes brought by acid activation [1].

20.4.2 The FTIR Analyses

20.4.2.1 The FTIR for Original and HCl Activated Clinoptilolite Forms

The FTIR-spectra in Fig. 20.3 gives a clear picture about the effect of chemical conditioning on the clinoptilolite. Acid treatment is said to remove the non-zeolitic components thus increasing the concentration of zeolite minerals.

In Fig. 20.3 at the range of 4000 and 3000 cm^{-1} the original and 0.04 and the 0.02 M HCl-activated forms of clinoptilolite showed distinct stretching at this range which are typical of water absorption. This shows that water adsorption and retention by clinoptilolite is increased by HCl activation at 0.02 M concentration. At the range of 2000 and 1500 cm^{-1}, the 0.02 M HCl-activated clinoptilolite showed two intensive peaks and yet again the original and the 0.04 M activated forms showed none. This could be attributed to the washing out the non zeolitic impurities present in the original clinoptilolite as confirmed by XRD, XRF and SEM–EDS when activated with 0.02 M HCl. The disappearance of these peaks

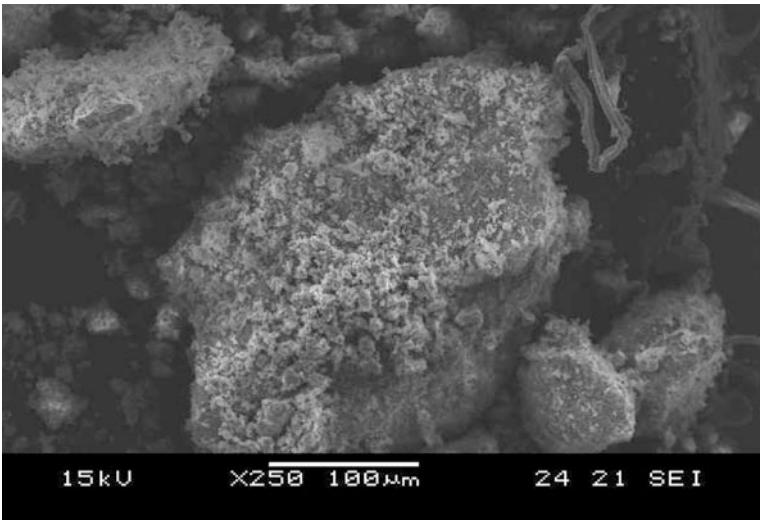


Fig. 20.1 SEM micrograph of original clinoptilolite at X250 magnification

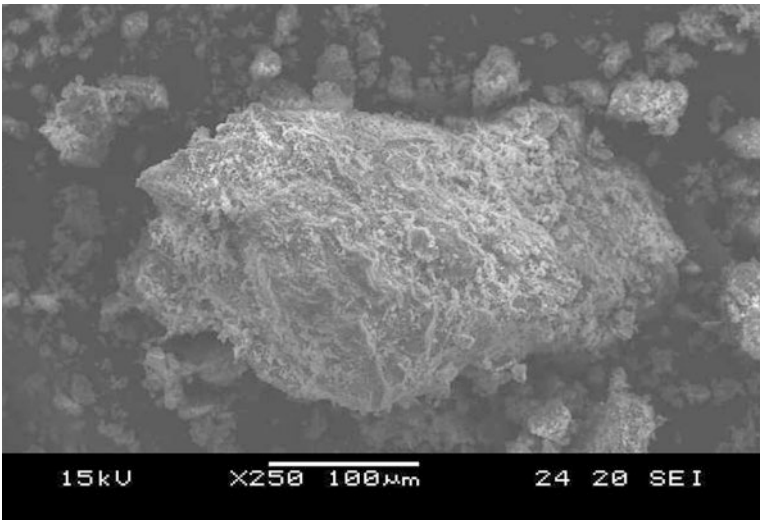


Fig. 20.2 SEM micrograph of HCl activated clinoptilolite at X250 magnification

with the 0.04 M HCl-activated form could be due to the higher acid strength resulting in the destruction of the active sites observed with the 0.02 M HCl activated clinoptilolite. There were peaks observed for all the clinoptilolite forms at 1558 cm^{-1} which may be due to the bending vibrations of adsorbed water. This was expected given the porous clinoptilolite structure thus, desiccation of the zeolite at temperatures above 50 °C will increase its hydrophilic (water

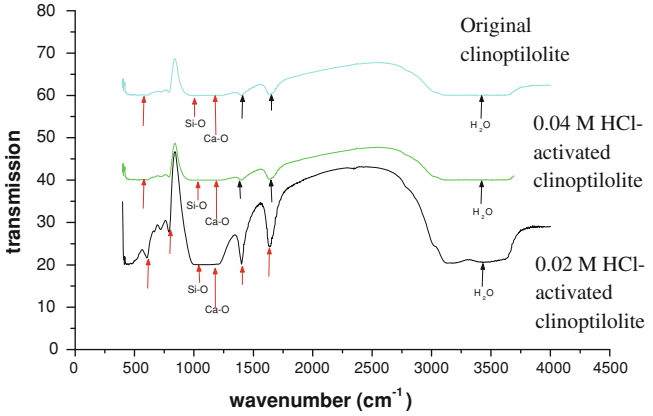


Fig. 20.3 The FTIR spectra for original and HCl-activated clinoptilolite forms at concentrations of 0.02 and 0.04 M

absorption) properties. Zeolites that have K^+ in high amounts have low water absorption capacity. This could probably explain why the original clinoptilolite does not show such peaks. It is also possible that the K^+ in the acid activated zeolite was leached out by the acid. The intensity of the peak at this range is more pronounced with the 0.02 M HCl-activation than it is with the 0.04 M HCl-activated and the original form. Since this is a water sorption peak, it could be possible that water sorption capacity is low with original clinoptilolite, high with 0.02 M HCl-activated clinoptilolite and an increase in HCl-activation to 0.04 M concentration diminishes the water sorption capacity of the zeolite [14].

The stretching between 1500 and 1000 cm^{-1} observed in Fig. 20.3 indicates the presence of a high content of calcite in the sample as confirmed by XRD results. The strong band at 1341 cm^{-1} (due to Si–O stretching) is the main characteristic band for quartz. The peaks observed between 1000 and 600 cm^{-1} are present in all the forms of clinoptilolite, one characteristic band appears at 836 cm^{-1} for all the forms. This is the quartz band. Quartz is common with zeolites, especially those of the Heulandite family. The peak that appears at 753 cm^{-1} for the original clinoptilolite form appears at 759 cm^{-1} for the activated zeolite. There is a peak that appears at 686 cm^{-1} for the original form while for the acid-activated forms it appears at 635 cm^{-1}. This shift could be attributed to the action of the acid [14].

It has been documented that acid treatment of natural clinoptilolite improves its sorption properties in ion exchange applications which is due to decatination, dealumination and the dissolution of amorphous silica fragments blocking the channels. A study by Mamba et al. (2009) also revealed that there is a change in the clinoptilolite structure after acid treatment with dilute acid activations accounting for improved heavy metal removal capacity of the clinoptilolites [14].

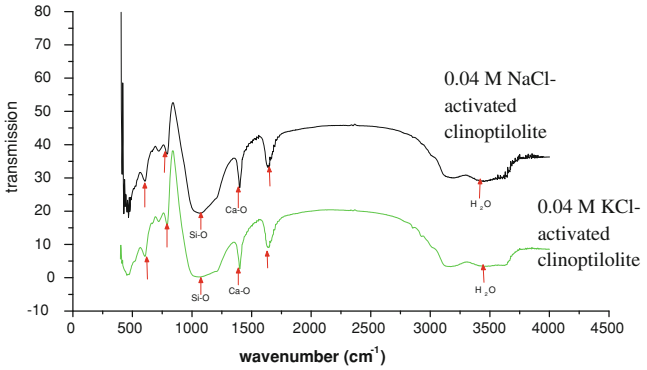


Fig. 20.4 FTIR spectra for KCl- and NaCl-activated clinoptilolite both activated at 0.04 M

20.4.2.2 FTIR for the KCl- and NaCl-Activated Clinoptilolite Forms

In essence, the KCl- and NaCl-activated clinoptilolite samples were comparable to the HCl-activated samples. The difference was in the intensities and a slight shift in the positions of the peaks i.e. they were less pronounced in the KCl- and NaCl-activated clinoptilolite forms than in the HCl-activated forms as observed in Fig. 20.4.

There were peaks between 2000 and 1250 cm^{-1}. These indicate the presence of a high content of calcite in the sample. The strong band at 850 cm^{-1} is the main characteristic band for quartz. Other characteristic quartz bands appear between 850 and 500 cm^{-1}. The positions of the peaks appear to have shifted when compared to the HCl-activated forms. This could be attributed to the mild action of the KCl and NaCl. The 0.04 M K-activated and Na-activated clinoptilolite showed an FTIR spectra that was super imposable over their 0.02 M activated counterparts, as a result only the 0.04 M activated clinoptilolite' spectra is shown. It can therefore be deduced that the two concentrations used did not markedly change the zeolite's structure. One may therefore expect these activations to differ only slightly in performance from the acid activated clinoptilolite [14].

20.4.2.3 Effect of pH

An optimum pH of 4 was obtained for all initial Cu^{2+} ion concentrations. The mechanism of adsorption at the clinoptilolite zeolite surface reflects the nature of physicochemical interaction of the metal ions in solution and the zeolite active sites [15]. Acid treatment of natural clinoptilolite improves its adsorbent properties [16, 17].

This is due to de-cantination, de-alumination and the dissolution of amorphous silica fragments blocking the channels. The availability of sites relates to the equilibrium behaviour whereas accessibility relates to the kinetic behaviour of the ion exchange system (Fig. 20.5).

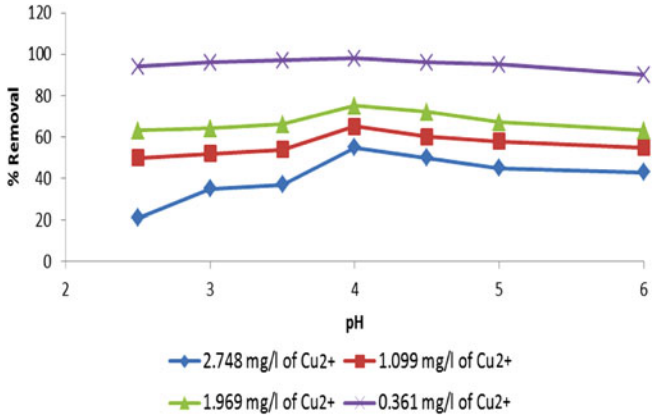


Fig. 20.5 Effect of pH on adsorption of Cu (II) ions

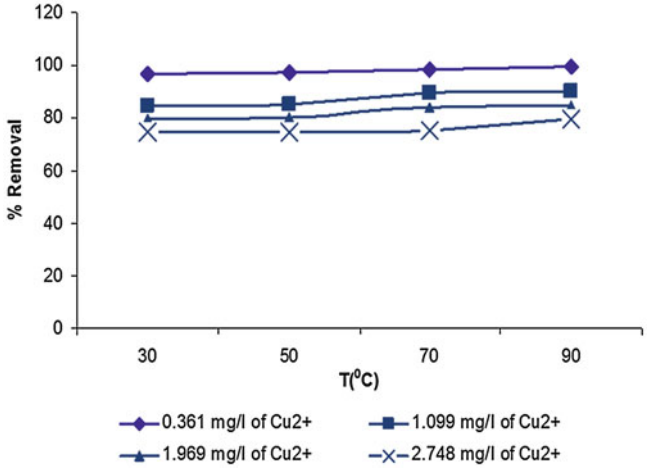


Fig. 20.6 Effect of temperature on adsorption of Cu (II) ions

20.4.2.4 Effect of Temperature

Effect of temperature on adsorption of Cu (II) ions was investigated by varying temperature from 30 to 90 °C at various initial concentrations at a pH of 4 for 24 h and adsorption was found to be temperature dependent. It is proposed that temperatures above 50 °C increase the pore size of the zeolites, enhancing the rate of intraparticle diffusion of ions (Fig. 20.6).

Table 20.2 Estimated Thermodynamic Parameters for Cu^{2+} at 90 °C and pH of 4

Isotherm models	Estimated isotherm parameters		
Langmuir equation	R^2	K(mg/g)	b (l/g)
$\frac{Ce}{qe} = \frac{1}{K_l b} + \frac{Ce}{K_l}$	0.9345	0.24	88.9
Freudlich Equation	R^2	K(mg/g)	n
$\log q_e = \log K + \frac{1}{n} \log c_e$	0.9768	4.23	3.21
Thermodynamic Parameters	ΔG	ΔH	ΔS
	−15.81	8.31	0.0442

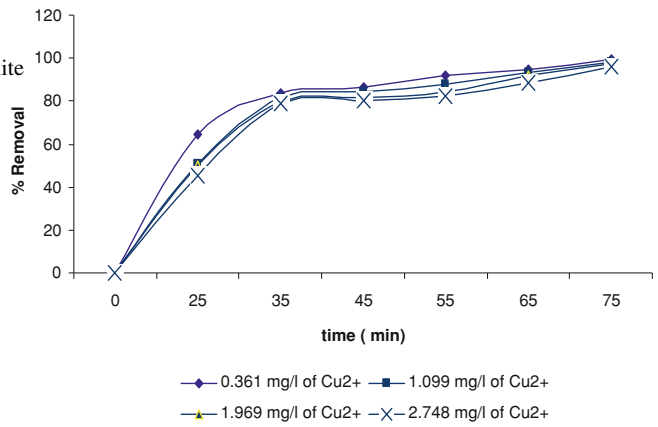


Fig. 20.7 Effect of time on adsorption of Cu (II) onto 10 g/l of Clinoptilolite Zeolite at pH 4.0 and 90 °C

20.4.2.5 Effect of Contact Time

The removal of Cu (II) ions increased with time and attained saturation in about 60–70 min (Fig. 20.7) as was previously observed [16].

The synthetic wastewater was prepared at five different Cu^{2+} ion concentrations, at 0.361, 1.099, 1.969 and 2.748 g/l.

20.4.2.6 Effect of Zeolite Concentration

Adsorption increased with increase in zeolite amount for fixed initial concentration (Fig 20.8). This is because the increase in zeolite amount leads to increase adsorption surface area and available active sites [17].

20.4.2.7 Adsorption Isotherm Studies

The correlation coefficient (R^2) of the adsorption isotherm data shown in Table 20.2, and Figs. 20.9, 20.10 confirms that adsorption of Cu (II) ions on clinoptilolite zeolite gave the best fit using the Freundlich isotherm model.

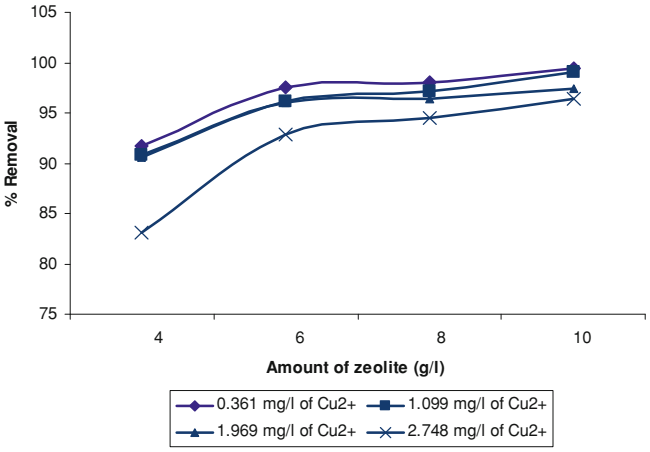


Fig. 20.8 Effect of amount of zeolite on adsorption of Cu (II) at pH 4.0 and 90 °C

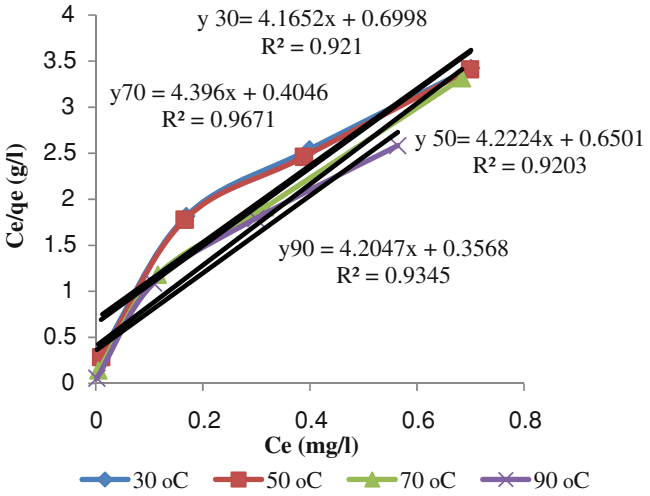


Fig. 20.9 Langmuir isotherm plots for Cu (II) adsorption by Zeolite at various temperatures

The Langmuir constant b, increases with the increase in temperature as sorption capacities and intensities are enhanced. The nature of adsorption is indicated by the separation factor (R_L) [18] expressed in Eq. (20.4).

$$R_L = \frac{1}{1 + K_1 C_0} \tag{20.4}$$

C_o is the initial cations concentration and K_1 the Langmuir constant. For $R_L = 0$ (irreversible), $0 < R_L < 1$ (favourable) and $R_L = 1$ (unfavourable) [18] (Fig. 20.9).

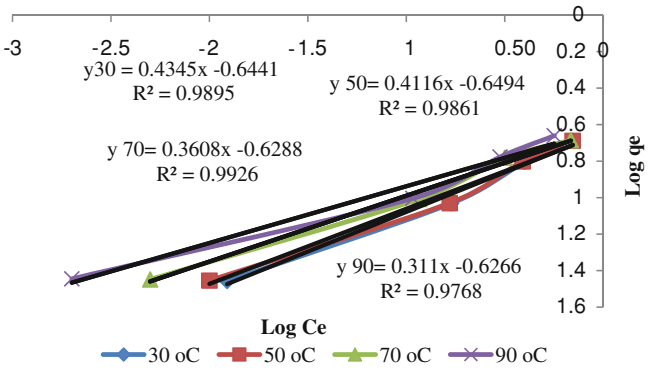


Fig. 20.10 Freudlich isotherm in Cu^{2+} adsorption by zeolite at various temperatures

The R_L value obtained for the data in Table 20.2 was 0.202, showing that adsorption was favourable.

20.5 Conclusion

The removal of Cu (II) ions from synthetic waste water was dependent on pH, amount of adsorbent, initial concentration of waste water solution and contact time. The optimum pH was 4 and equilibrium saturation was reached after 75 min. Thermodynamic parameters such as ΔG, ΔH and ΔS were highly dependent on temperature. The adsorption process was endothermic in nature and the rise in temperature increased the randomness of the solid–solution interface. The Freundlich isotherm model fitted the adsorption of Cu (II) ions into clinoptilolite better compared to the Langmuir model. The calculated values of the dimensionless separation factor R_L from the Langmuir isotherm constants confirm favourable sorption of Cu (II) onto clinoptilolite zeolite.

Acknowledgments The authors acknowledge financial support from the University of Johannesburg.

References

1. Muzenda E, Kabuba J, Ntuli F., Mollagee M., Mulaba Bafubiandi AF (2011) Cu(II) removal from synthetic waste water by ion exchange process. In: Proceedings of the world congress on engineering and computer science 2011, WCECS 2011, 19–21 October 2011 (Lecture notes in engineering and computer science). San Francisco, pp 685–689
2. Clement RE, Eiceman GA, Koester CJ (1995) Environmental analysis. Anal Chem 67:221–255
3. Massadeh AM, Baker HM (2008) Natural Jordanian zeolite: removal of heavy metal ions from water samples using column and batch methods. J Environ Monit Assess 157:319–330

4. Akar T, Tunali S (2005) Biosorption performance of Botrytis cinerea fungal by-products for removal of Cd (II) and Cu (II) ions from aqueous solutions. Mineral Eng 18:1099–1109
5. Norton L, Baskaran K, McKenzie T (2004) Biosorption of zinc from aqueous solutions using bio solids. Adv Environ Res 8:629–635
6. Chong KH, Volesky B (1995) Description of 2–metal bio sorption equilibrium by Langmuir-type models. Biotechnol Bioeng 47:451–460
7. Hui KS, Chao CYH, Kot CS (2005) Removal of mixed heavy metal ions in wastewater by zeolite 4A and residual products from recycled coal fly ash. J Hazard Mater 127:89–101
8. Matis KA, Lazaridis NK, Zouboulis AI, Gallios GP, Mavrov V (2005) A hybrid flotation-microfiltration process for metal ions recovery. J Memb Sci 247:29–35
9. Tewari DK, Behari J, Sen P (2008) Application of nanoparticles in wastewater treatment. World Appl Sci J 3:417–433
10. Argun ME (2008) Use of clinoptilolite for the removal of nickel ions from water. Kinetics and thermodynamics. J Hazard Mater 150:585–595
11. Ozay O, Ekici S, Baran Y, Aktas N, Sahiner N (2009) Removal of toxic metal ions with magnetic hydrogels. Water Res 43:4403–4441
12. Altin O, Ozbelge HO, Dogu T (1998) Use of general purpose adsorption isotherms for heavy metal-clay mineral interactions. J Colloid Interf Sci 198:130–140
13. Barci S (2004) Nature of ammonium ion adsorption by sepiolite: analysis of equilibrium data with several isotherms. Water Res 38:1129–1138
14. Mamba BB, Nyembe DW, Mulaba-Bafubiandi, AF (2009) Removal of copper and cobalt from aqueous solutions using natural clinoptilolite. Water SA 35(3):307–314
15. Korkuna O, Leboda R, Skubiszewska J, Vrublevs'ka T, Gun'ko VM, Ryczkowski J (2006) Structural and physicochemical properties of natural zeolites: clinoptilolite and mordenite. Microporous Mesoporous Mater 87:243–254
16. Vasylechko VO, Gryshchouk GV, Lebedynets LO, Leboda R, Skubiszewska-Zieba J (1999) Investigation of usefulness of Transcarpathian zeolites in trace analysis of waters. Application of mordenite for the pre concentration of trace amounts of copper and cadmium. Chem Anal (Warsaw) 44:1013–1024
17. Hernandez MA (2000) Nitrogen-sorption characteristics of the microporous structure of clinoptilolite-type zeolites. J Porous Master 7:443–454
18. Dyer H (1981) The plotting and interpretation of ion-exchange isotherms in charcoal systems. Sep Sci Technol 16:173–183

Chapter 21
Performance of RANS, URANS and LES in the Prediction of Airflow and Pollutant Dispersion

Salim Mohamed Salim and Kian Chuan Ong

Abstract The performance of 3 different CFD numerical approaches, namely RANS, URANS and LES are evaluated to determine their suitability in the prediction of airflow and pollutant dispersion in urban street canyons. Numerical results are evaluated against wind tunnel experimental data available from an online database (www.codasc.de

Keywords Air pollution · CFD · LES · RANS · URANS · Urban street canyon

21.1 Introduction

Air quality in urban and industrial complexes has garnered great interest because of numerous implications on human health, pedestrian comfort and environmental concerns. There is need to understand the mechanism of pollutant dispersion in urban street canyon and its ramifications to environmental and structural engineering practices. This has resulted in continuous development of new simulation tools and improvement of existing numerical modelling techniques in order to assist regulators, policy makers, architects and urban planners to mitigate air pollution problems in their cities. This is in addition to enabling emergency authorities develop evacuation plans following natural disasters, accidents or deliberate release of harmful airborne matter.

S. M. Salim (✉)
School of Engineering, Taylor's University, 47500, Subang Jaya, Selangor, Malaysia
e-mail: salim@alumni.nottingham.ac.uk

K. C. Ong
University of Nottingham, Malaysia Campus, Semenyih, Malaysia

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_21,

This has motivated a number of field, experimental and numerical investigations to assess the interaction of buildings, trees, moving vehicles and other obstacles with the atmospheric boundary layer flow and resulting pollutant accumulation and/or dissipation patterns within urban and industrial complexes, including the effects of thermal stratification. Computational fluid dynamics (CFD) is the preferred tool of investigation at the micro scale [1, 2] especially with the ever increasing computer resources available to civilian researchers.

However, many previous CFD studies have employed conventional Reynolds-averaged Navier–Stokes (RANS) turbulence closure schemes, which although performed qualitatively well, poorly predicted the pollutant levels and distribution in comparison to wind tunnel (WT) experiments. The assumption of steady-state solution in the analyses has been identified as one of the major shortcomings and possible cause of the discrepancies [3–5].

Recently, Salim et al. [6, 7] compared RANS against large eddy simulation (LES) to evaluate their respective performances and corresponding computational cost. LES, although more resource demanding, was observed to predict better than RANS because it resolved the unsteady fluctuations in the flow field, thus accounting for the turbulent mixing process that the dispersion of pollutants rely on. Similar inferences were reported in a study by Tominaga and Stathopoulos [8].

A question than arises as to whether unsteady RANS (URANS) would perform equally well, as it also solves for the transient solution similar to LES but at a fraction of the computational cost. This is addressed by Salim et al. [9] and presented extensively in this chapter.

The simulation of wind and pollutant dispersion within an urban street canyon of width to height ratio, $W/H = 1$ is examined using 2 steady-state and unsteady RANS models i.e., the standard k - ε and RSM; and LES to evaluate their relative performance. The numerical results are validated against WT measurement data available from the online database CODASC www.codasc.de [10].

The findings of the study are not only limited to environmental issues in urban areas, but can be applied to any flow simulation where large scale eddies dominate and resolving transience is paramount to achieving accurate and reliable results.

21.2 Methodology

21.2.1 Computational Domain and Boundary Conditions

In the present study, CFD is employed to evaluate the prediction performance of RANS, URANS and LES using the commercial software FLUENT® carried out with the aim of reproducing the WT experiment by Gromke and Ruck [11, 12]. The computational domain and boundary conditions are summarized in Fig. 21.1 above.

An inlet boundary condition is defined at the entrance and no-slip conditions are applied for the building walls and floor. Symmetry conditions are specified for the

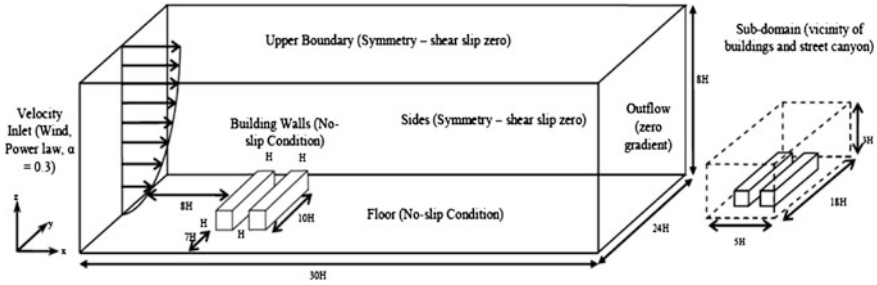


Fig. 21.1 Computational domain and boundary conditions for the CFD simulation setup

top and lateral sides of the domain to enforce parallel flow. At the face downwind of the obstacles an outflow boundary condition is imposed to ensure all the derivatives of the flow variables to vanish.

A high resolution computational grid is generated using hexahedral cells with a total mesh count of 1 million, incorporating recommendations based on the wall y^+ approach by Salim et al. [13]. Cells are arranged in a horizontally-structured and vertically-structured grid with fine resolution close to the wall and in regions with large flow gradients (with half the total cells placed in the sub-domain defining the vicinity of the street canyon: see Fig. 21.1). A minimum spacing of 0.05 H in the x, y, and z direction is set with an expansion ratio of below 1.2 between consecutive cells.

Velocity profile and turbulence quantity profiles are modeled numerically using user defined functions (UDFs) and based on wind tunnel measurements, were assumed to follow the power law profile:

$$u(z) = 4.7\left(\frac{z}{0.12}\right)^{0.3}, \tag{21.1}$$

while turbulent kinetic energy and dissipation rate profiles are specified as

$$k = \frac{u_*^2}{\sqrt{c_\mu}}\left(1 - \frac{z}{\delta}\right) \tag{21.2}$$

and

$$\varepsilon = \frac{u_*^3}{kz}\left(1 - \frac{z}{\delta}\right). \tag{21.3}$$

where u is the vertical velocity profile, z the vertical distance, k the kinetic energy profile, ε the dissipation rate profile, δ is the boundary layer depth ($\approx$0.5 m), $u_* = 0.54$ ms^{-1} the friction velocity, κ the von Kàrmàn constant (= 0.4) and $C_\mu = 0.09$.

21.2.2 Numerical Scheme

Steady-state RANS mean solutions are computed using standard k-ε and RSM turbulence closure schemes. Second order upwind scheme is selected for the transport equations to improve accuracy and reduce numerical diffusion except for pressure, where STANDARD interpolation is employed instead. The scaled residual for all flow properties are set at 10^{-6}. The RANS equations are

$$\frac{\partial \overline{u_i}}{\partial x_i} = 0, \tag{21.4}$$

and

$$\overline{u_j}\frac{\partial \overline{u_i}}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x_i} + \upsilon\frac{\partial^2 \overline{u_i}}{\partial x_j^2} - \overline{u_j'\frac{\partial u_i'}{\partial x_j}}. \tag{21.5}$$

A non-dimensional time step of 4×10^{-2} is implemented for the unsteady (i.e., time-advancement) solution in the URANS simulations. All other settings are maintained as above. The equations are

$$\frac{\partial \overline{u_i}}{\partial xi} = 0, \tag{21.6}$$

and

$$\overline{\frac{\partial u_i}{\partial t}} + \overline{u_j}\frac{\partial \overline{u_i}}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x_i} + \upsilon\frac{\partial^2 \overline{u_i}}{\partial x_j^2} - \overline{u_J\frac{\partial u_i}{\partial x_j}}. \tag{21.7}$$

The flow properties are disintegrated into their mean and fluctuating components and integration over time (i.e., time-averaging) is performed for the RANS approach. The difference between steady-state and unsteady RANS is that an addition unsteady term is present in the later.

For LES simulation, the dynamic Smagorinsky-Lily sub-grid scale (SGS) model is selected. Bounded central differencing scheme for momentum, 2nd order time-advancement and 2nd order upwind for energy and species transport equations are chosen. PRESTO and SIMPLEC are employed for pressure and pressure–velocity coupling, respectively. Convergences at 1×10^{-3} for the scaled residual are set. A dimensionless time-step of 2.5×10^{-3} is chosen. The LES equations are

$$\frac{\partial \overline{u_i}}{\partial x_i} = 0, \tag{21.8}$$

and

$$\overline{\frac{\partial u_i}{\partial t}} + \overline{u_j}\frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x_i} + \upsilon\frac{\partial^2 \overline{u_i}}{\partial x_j^2} - \frac{\partial \tau_{ij}}{\partial x_j}. \tag{21.9}$$

The over-bar signifies spatial filtering, and not time-averaging as is the case of RANS. It is worth identifying that the filtered (i.e., LES) momentum equation is similar to the RANS equation. The spatial-filtering is an integration just like time-averaging, the difference being that the integration is in space and not over time as in the case of RANS.

All simulations were performed in parallel on an Intel® Xeon® workstation (4 CPU processors). Further details regarding the different numerical schemes can be obtained from FLUENT® user manual [14].

21.2.3 Pollutant Dispersion

The advection–diffusion (AD) method is employed for modeling the dispersion of pollutants species and FLUENT® computes it as follows for turbulent flows:

$$J = -\left(\rho D + \frac{\mu_t}{Sc}\right)\nabla Y. \tag{21.10}$$

Where D is the molecular diffusion coefficient for the pollutant in the mixture, μ_t is the turbulent eddy viscosity, Y is the mass fraction of the pollutant, ρ is the mixture density. Line sources are used to model the release of traffic exhaust and are simulated by ear-marking sections of the volume in the geometry and demarcating them as source volumes, with Sulfur hexafluoride (SF_6) discharged at a rate of $Q = 10\ \mathrm{gs}^{-1}$, acting as the pollutant species.

The position of the line sources in the WT setup and computational domain are presented in Fig. 21.2.

21.3 Results and Discussion

21.3.1 Steady-State Versus Transient Solution (RANS versus LES)

First, mean solution obtained by the conventional steady-state RANS are compared to LES in order to demonstrate the necessity for resolving the transience for the flow field within the street canyon.

The normalized concentration of pollutant concentration at the leeward wall (Wall A) and windward wall (Wall B) are compared to WT measurement as illustrated in Figs. 21.3 and 21.4.

In Fig. 21.3, it can be seen that LES reproduces the pollutant concentration distribution much better than the 2 steady-state RANS turbulence models. This is particularly evident in the vicinity of the centerline $(\mathrm{y/H} = 0)$ at both walls, where the maximum concentration occurs and is determined to be the most critical zone especially when considering pedestrian exposure.

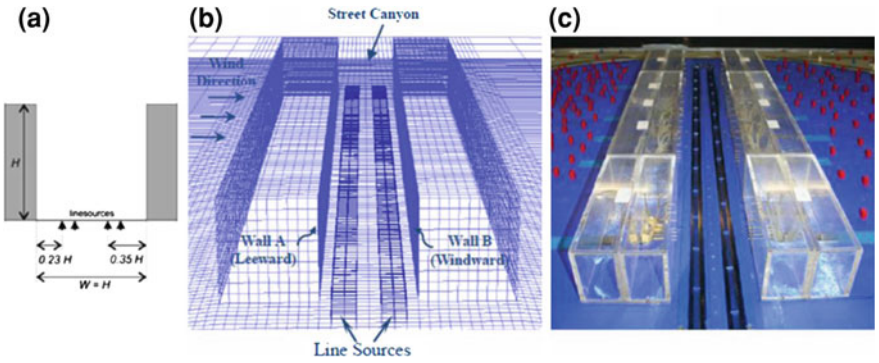


Fig. 21.2 Position of line sources **a** sketch, **b** computational domain (FLUENT) and **c** wind tunnel (CODASC database)

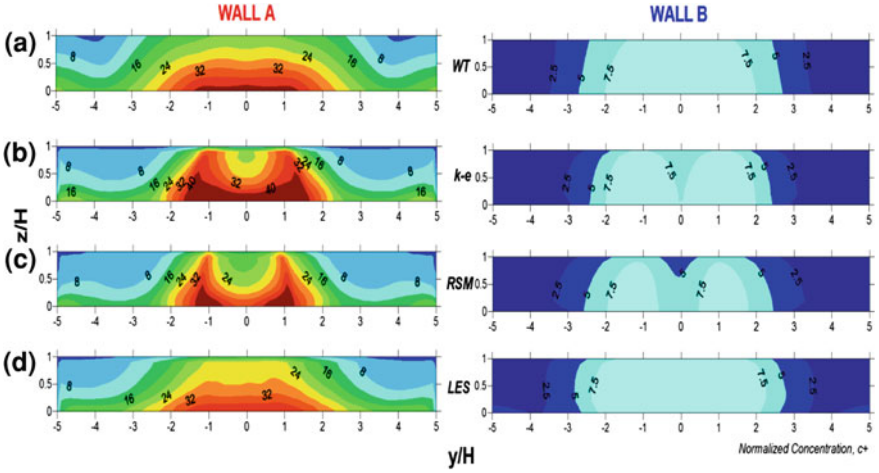


Fig. 21.3 Mean concentration profiles along the leeward wall (Wall A) and windward wall (Wall B) obtained by RANS, LES and WT

The quantitative comparison of the vertical concentration profiles along different locations on the leeward and windward walls are illustrated for the 2 CFD techniques in Fig. 21.4. LES not only predicts much better than RANS, but provides more consistent results in relation to WT data. It predicts well for all locations along the leeward wall and only slightly overpredicts along the windward wall. The standard k-ε and RSM on the other hand have varying degree of accuracies at different locations, overpredicting at some locations and underpredicting at others.

Figure 21.5 demonstrates the mean normalized velocities and pollutant concentration contours along the mid-plane $(y/H = 0)$ within the street canyon.

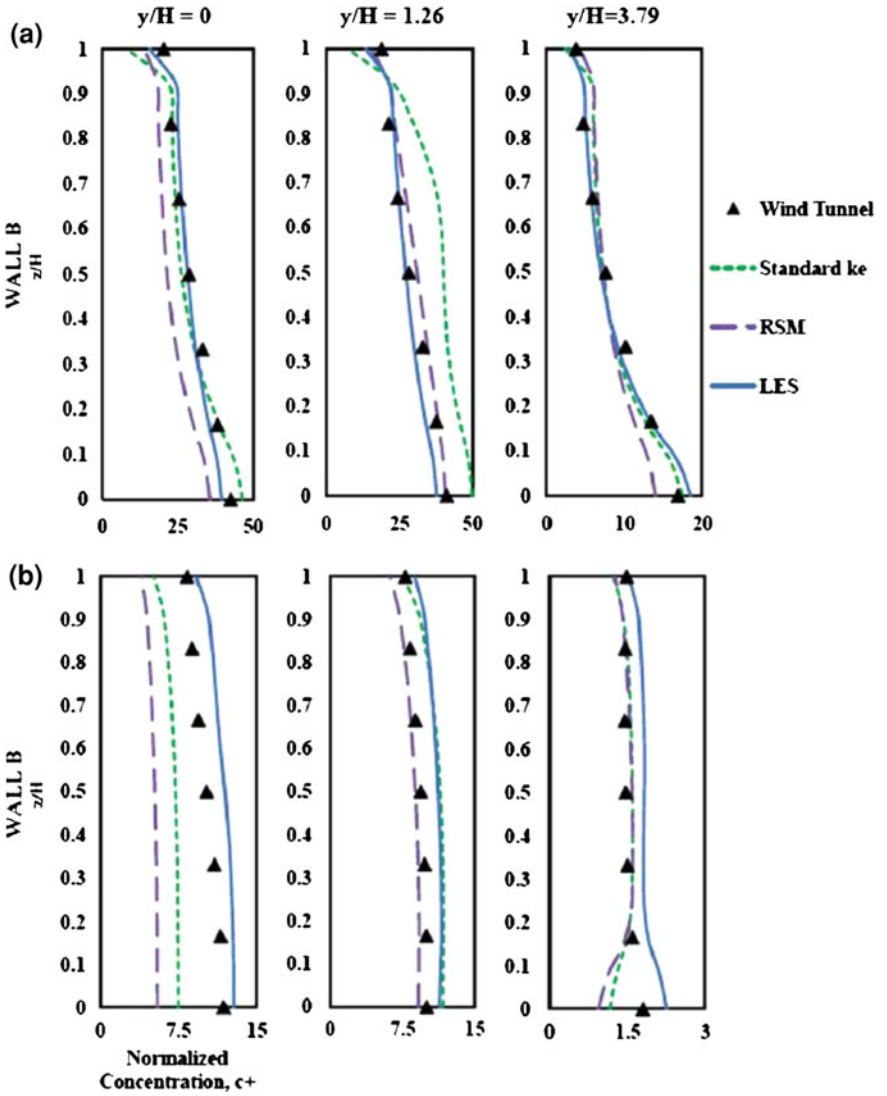


Fig. 21.4 Mean concentration profiles at different locations along the **a** Leeward wall and **b** windward wall comparing RANS and LES against WT experiment

RANS turbulence models predict an accumulation of pollutant towards the leeward walls; whereas LES reproduces a better spread and this can be explain by the fact that it resolves the turbulent mixing within the canyon by accounting for the inherent fluctuations in the flow field as further illustrated in Fig. 21.6 for the wall concentrations.

Flow variables are shown to vary significantly over time and LES is able to capture pockets of intertwining bubbles of opposing velocities.

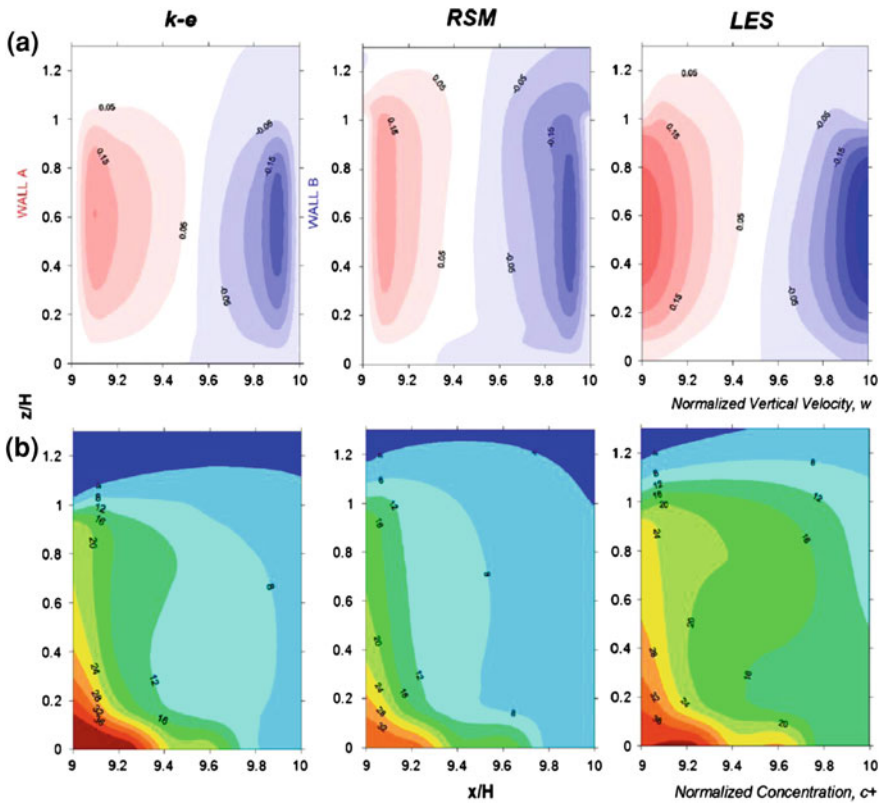


Fig. 21.5 **a** Normalized vertical velocity contours and **b** normalized concentration contours comparing k - ε, RSM and LES

21.3.2 Transient Solution (URANS Versus LES)

It has been determined in the previous section that there is a need to resolve the unsteadiness of the flow field within the street canyon in order to obtain accurate predictions. Therefore, 2 transient CFD techniques, namely URANS and LES, are evaluated in this section to determine their relative performance.

Comparison of numerical results between URANS and LES at the mid-plane of the canyon (i.e., $y/H = 0$), and at the leeward (Wall A) and windward (Wall B) walls are presented in Figs. 21.7 and 21.8. LES resolves the fluctuations of the flow variables, which are shown to vary significantly over time thus capturing transient mixing which is important to accurately predict pollutant dispersion.

URANS are only suitable for non-stationary flows such as periodic or quasi-periodic flows involving deterministic structures (for example, they can occasionally predict vortex shedding i.e., largest unsteady scales) and falls most often short of capturing the remaining eddy scales [14]. This is because they still solve

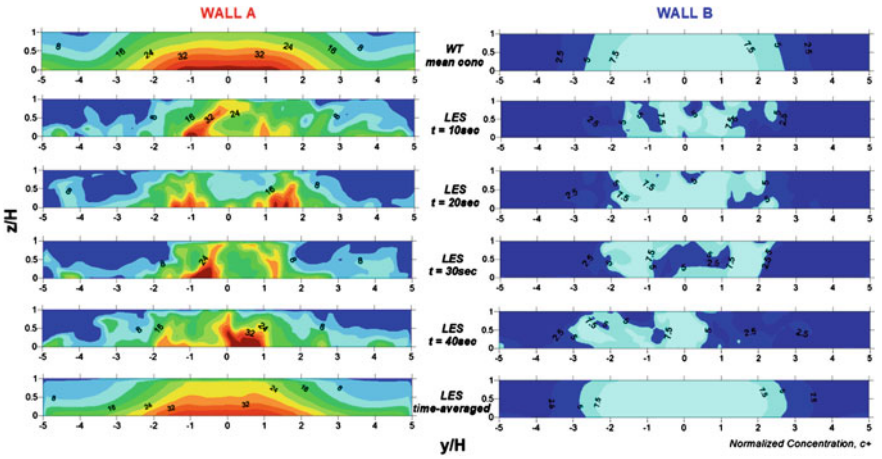


Fig. 21.6 Time-evolution of the normalized concentration along Wall A and Wall B at different time, obtained by LES

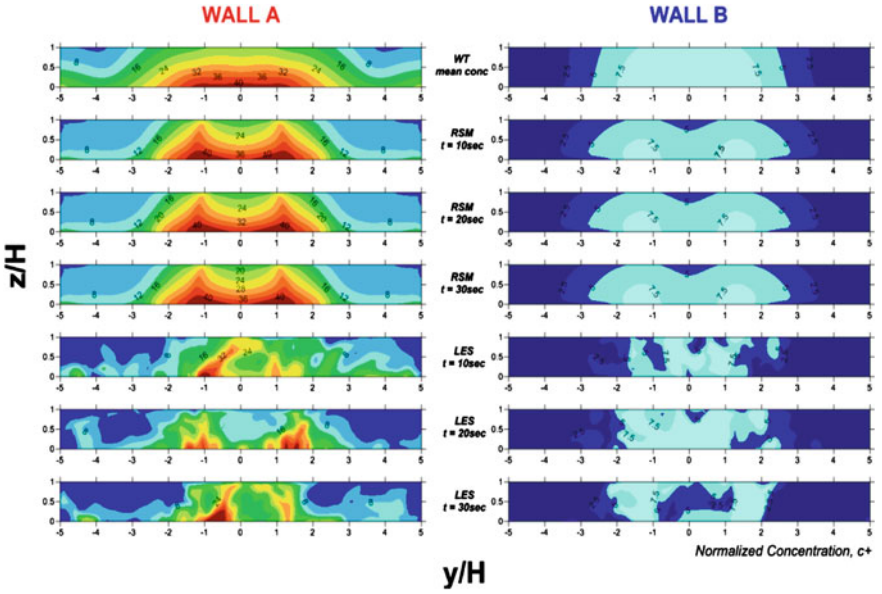


Fig. 21.7 URANS against LES for unsteady simulations at different time instances showing the mid-plane normalized vertical velocities and corresponding normalized concentrations

for the mean flow equations but perform ensemble averaging instead (i.e., realizations of the mean flow over many instances).

The ratio of the approaching boundary layer thickness to the obstacle height plays an important role on the resulting flow structure, especially in the separation

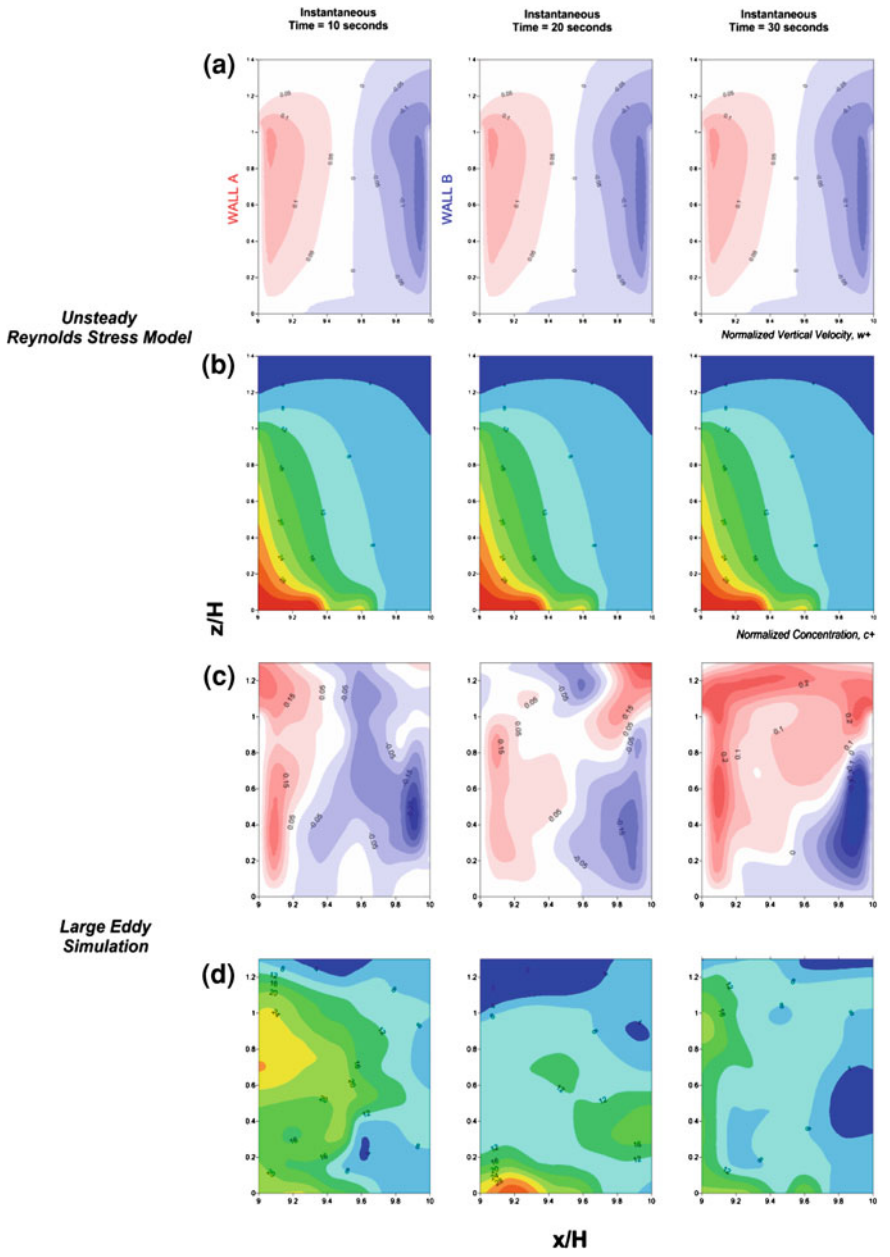


Fig. 21.8 URANS against LES for unsteady simulations at different time instances showing the wall concentration levels

regions upstream and downstream of the obstacle. In the present study, the ratio is $\delta/\mathrm{H} = 5$ explaining why the URANS calculations failed to detect periodic motion within the street canyon and in the wake region behind the leeward building, thus reproducing similar poor predictions to RANS.

There is evidence in the literature that a thick and turbulent boundary layer tends to suppress periodic motion. For example, Castro [15] did not observe periodicity in the wake region of a cube in a thick boundary layer, $\delta/\mathrm{H} = 6.6$ during his experimental investigations. On the other hand, strong periodicity for a vertical square-cross section cylinder with $\mathrm{W}/\mathrm{H} = 1/3$ placed in a relatively smaller boundary layer of thickness of 0.8 H was observed by Sakamoto and Haniu [16].

The present study suggests that the fluctuations dominate the flow field over a possibly weak mean field. Unlike LES, both RANS and URANS are unable to account for fluctuating velocity and turbulent quantities, i.e., $\overline{u_i'} = 0$ which are clearly seen in Fig. 21.6.

21.4 Conclusion

Three different CFD techniques, namely steady-state RANS, unsteady RANS (URANS) and LES were employed for the simulation of airflow and pollutant dispersion within an urban street canyon and validated against wind tunnel experimental data, to evaluate their relative prediction performance.

It is shown that in order to accurately reproduce the flow and concentration fields within urban street canyons, it is imperative to account for the transient solution by resolving the internally and externally induced fluctuations on which the dispersion of pollutants depends on.

Steady-state RANS poorly predicted the pollutant concentrations and did not give consistent results. URANS was unable to account for the fluctuations of the flow field, although solving for the transient solution, as it is limited to externally induced fluctuations (i.e., periodic motion) which is absent due to the relatively large turbulent boundary layer of $\delta/\mathrm{H} = 5$. Therefore, although URANS is comparatively cheaper than LES in terms of computational cost, it is not a suitable replacement for air pollution problems, or any other flow situation where small scale eddies are prominent in the flow field development.

Pollutant concentrations at the leeward wall are underestimated in the RANS/URANS simulations, while both slightly over- and under-estimate at the windward wall.

The study also suggests that typical urban street canyon in atmospheric boundary layer flows could be treated as statistically stationary for high δ/H parameter values since periodic flow motions are negligible. Fluctuations dominate the flow field over a possibly weak mean field, supporting why LES was able to reproduce the solutions most accurately and consistently, as it resolved the small scale unsteady fluctuations.

The combination of wind tunnel experimental and LES can provide viable approach to investigate pollutant dispersion in street canyons, to obtain the necessary data for assessment, planning and implementation of exposure mitigation in urban areas.

References

1. Britter RE, Hanna SR (2003) Flow and dispersion in urban areas. Annu Rev Fluid Mech 35:1817–1831
2. Britter RE, Schatzmann M (2007) Background and justification document to support the model evaluation guidance and protocol, COST Action 732
3. Gromke C, Buccolieri R, Di Sabatino S, Ruck B (2008) Dispersion study in a street canyon with tree planting by means of wind tunnel and numerical investigations—evaluation of CFD data with experimental data. Atmospheric Environ 42:8640–8650
4. Di Sabatino S, Buccolieri R, Pulvirenti B, Britter RE (2008) Flow and pollutant dispersion in street canyons using FLUENT and ADMS-urban. Environ Model Assess 13:369–381
5. Buccolieri R, Salim SM, Leo LS, Di Sabatino S, Chan A, Ielpo P, de Gennaro G, Gromke C (2011) Analysis of local scale tree-atmosphere interaction on pollutant concentration in idealized street canyons and application to a real urban junction. Atmospheric Environ 45:1702–1713
6. Salim SM, Buccolieri R, Chan A, Di Sabatino S (2011) Numerical simulation of atmospheric pollutant dispersion in an urban street canyon: comparison between RANS and LES. J Wind Eng Ind Aerodyn 99:103–113
7. Salim SM, Chan A, Cheah SC (2011) Numerical simulation of atmospheric pollutant dispersion in tree-lined street canyons: comparison between RANS and LES. Build Environ 46:1735–1746
8. Tominaga Y, Stathopoulos T (2010) Numerical simulation of dispersion around an isolated cubic building: model evaluation of RANS and LES. Build Environ 45:2231–2239
9. Salim SM, Ong KC, Cheah SC (2011) Comparison of RANS, URANS and LES in the prediction of airflow and pollutant dispersion. Lecture notes in engineering and computer science: proceedings of the World Congress on Engineering and Computer Science 2011, WCECS 2011, San Francisco, 19–21 Oct 2011, pp 673–678
10. CODASC (2008) Concentration data of steet canyons, laboratory of building and environmental aerodynamics, IfH, Karlsruhe Institue of Technology
11. Gromke C, Ruck B (2007) Influence of trees on the dispersion of pollutants in an urban street canyon—experimental investigations of the flow and concentration field. Atmospheric Environ 41:3287–3302
12. Gromke C, Ruck B (2009) On the impact of trees on dispersion processes of traffic emissions in street canyons. Boundary-Layer Meteorol 131:19–34
13. Salim SM, Ariff M, Cheah SC (2010) Wall y+ approach for dealing with turbulent flows over a wall mounted cube. Prog Comput Fluid Dyn 10:1206–1211
14. FLUENT 6.3 documentation, Lebanon, U.S.A., 2005
15. Castro IP (1981) Measurements in shear layers separating from surface-mounted bluff bodies. J Wind Eng Ind Aerodyn 7:253–272
16. Sakamoto H, Haniu H (1988) Aerodynamic forces acting on two square prisms placed vertically in a turbulent boundary layer. J Wind Eng Ind Aerodyn 31:41–66

Chapter 22
Methodology for Extraction of Soluble Non-Starch Polysaccharides and Viscosity Determination of Aqueous Extracts from Wheat and Barley

Rodica Caprita and Adrian Caprita

Abstract Cereal grains contain various amounts of non-starch polysaccharides (NSP), which are composed predominantly of arabinoxylans, β-glucans and cellulose. The detrimental effect of soluble NSP is mainly associated with the viscous nature of these polysaccharides and their physiological effects on the digestive medium. Our study had in view to investigate the influence of some extraction conditions on the viscosity of wheat and barley aqueous extracts. Water extract viscosities (WEV) appeared to be related to the particle size, the extraction time and temperature, and to the time elapsed after isolation of the extract. The experiments revealed as optimum conditions for obtaining the soluble NSP extract from wheat and barley and for WEV determination: granulation of 0.5 mm size, extraction temperature 40 °C, extraction time 60 min, and viscosity measurements immediately after extract isolation.

Keywords Arabinoxylans · Barley · Dynamic viscosity · β-glucans · Non-starch polysaccharides · Wheat

R. Caprita (✉)
Department of Exact Sciences, Banat University of Agricultural Sciences and Veterinary Medicine Timisoara, Calea Aradului 119, 300645 Timisoara, Romania
e-mail: rodi.caprita@gmail.com

A. Caprita
Department of Chemistry, Banat University of Agricultural Sciences and Veterinary Medicine Timisoara, Calea Aradului 119, 300645 Timisoara, Romania

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_22,

22.1 Introduction

Polysaccharides are widespread biopolymers, which quantitatively represent the most important group of nutrients in botanical feed. Carbohydrates constitute a diverse nutrient category ranging from sugars easily digested by the monogastric animals in the small intestine to dietary fibre fermented by microbes in the large intestine [1].

Dietary fibre (DF) is now defined as food material, particularly plant material, that is not hydrolysed by enzymes secreted by the human digestive tract but that may be digested by microflora in the gut.

The types of plant material that are included within the definitions of DF may be divided into two forms, based on their water solubility.

- Insoluble dietary fibre (IDF) which includes celluloses, some hemicelluloses and lignin;
- Soluble dietary fibre (SDF) which includes β-glucans, pectins, gums, mucilages and some hemicelluloses.

The IDF and SDF compounds, apart from lignin, are known collectively as non-starch polysaccharides (NSP), which was one of the earlier definitions of DF.

In animal nutrition, as "non-starch-polysaccharides" are summarized polysaccharides, which cannot be degraded by endogenous enzymes and therefore reach the colon almost indigested. Individual NSP groups have different chemical and physical characteristics that result in various effects on physiology of intestine and on organism in general.

Non-starch polysaccharides are principally non-α-glucan polysaccharides of the plant cell wall. They are a heterogeneous group of polysaccharides with varying degrees of water solubility, size, and structure.

Non-starch polysaccharides, according to Ebihara and Kiriyama [2], and Englyst and Hudson [3], refer to all carbohydrate fractions and types of dietary fibre, with the exception of lignin, either soluble or insoluble. Included are pectic substances, hemicelluloses, celluloses and gums (guar) and mucilages [4, 5].

Cellulose, hemicellulose and pectic substances are known as plant cell wall NSP since they comprised 80–90 % of the plant cell wall [5]. Resistant starch, theoretically, falls outside the NSP concept, but practically it depends on the method used to eliminate starch. Southgate [4] divided NSP in plant foods into structural and non-structural polysaccharides. Table 22.1 illustrates major types of NSP in plant foods.

Sasaki et al. [6, 7] classified NSPs into water-soluble and water-insoluble fractions which delineate their functions and chemical structure [8–10]. The solubility of NSP is determined not only by their primary structure, but also by how they are bound to other cell wall components (protein and lignin). Water-soluble NSP have opposite effects on water binding capacity and viscosity than the insoluble fiber fraction [7].

Table 22.1 Major types of NSP in plant foods

Primary source	Major groups	Components present	Summary of structures	Distribution in foods
Structural materials of the plant cell wall	Cellulose		Long chain β-glucans	All cell walls
	Non-cellulosic polysaccharides	Pectic substances	Galacturonans	Mainly in fruits and vegetables
		Hemicelluloses	Arabinogalactans	Cereals
			Arabinoxylans	Cereals
			Glucurono-arabinoxylans	Cereals
			Glucuronoxylans	Fruits and vegetables
			Xyloglucans	Fruits and vegetables
			β-Glucans	Cereals
Non-structural polysaccharides	Gums, mucilages		Wide range of heteropoly-saccharides	Seeds and fruits

The water insoluble fraction include cellulose, galactomannans, xylans, xyloglucans, and lignin, while the water-soluble fibers are the pectins, arabinogalactans, arabinoxylans, and β-(1,3)(1,4)-D-glucan (β-glucan) [11].

Cereal grains contain various amounts of non-starch polysaccharides, which are composed predominantly of arabinoxylans (pentosans), ß-glucans and cellulose [12]. The detrimental effect of soluble NSP is mainly associated with the viscous nature of these polysaccharides and their physiological effects on the digestive medium.

The NSP content and type differ among grains. The NSP content relative to dry matter is lower in wheat kernel (11.4 %) than in rye (13.2 %) and barley (16.7 %). Arabinoxylans are the predominant NSP in wheat (6–8 %) and rye (8.9 %), while β-glucans are the predominant NSP in barley (7.6 %) [13].

β-Glucans are glucose polymers containing a mixture of β1-3 and β1-4 linkages that make their physicochemical properties totally different from cellulose that is a straight-chain glucose polymer with only β1-4 linkages. Barley contains a high level of mixed-linked β-glucans (3–4 %) [14]. Barley also contains an appreciable amount of soluble NSP other than β-glucans [15].

The structures of cereal pentosans (arabinoxylans) are composed predominantly of two pentoses, arabinose and xylose, and their molecular structure consists of a linear β1-4-xylan backbone to which substituents are attached through O2 and O3 atoms of the xylosyl residues [16]. Most of the arabinoxylans in cereal grains are insoluble in water, but the arabinoxylans not bound to the cell walls can form highly viscous solutions and can absorb about ten times their weight of water.

Soluble NSP increase the viscosity of the small intestinal chime, generally hampering the digestion process, whereas insoluble NSP impede the access of endogenous enzymes to their substrates by physical entrapping [17, 18].

The anti-nutritive effect of soluble NSP is manifested through inhibition of digestion of starch, lipid and protein in the foregut [19]. The mechanism of action of soluble NSP is thought to involve increased viscosity of digesta which limits contact between digestive enzymes and substrates, and contact between nutrients and absorption sites on the intestinal mucosa [20–22].

The retention time of digesta during the passage through the gastrointestinal tract under the presence of NSP is an important factor for digestion and absorption of nutrients. Soluble NSP lead to a prolonged passage rate in the preceacal digestive tract, while insoluble NSP mostly showed no or only slight influence [23]. Soluble NSP in the form of β-glucans showed viscosity-elevating effects due to an increased stimulation of digestive juices secretion, which was also observed for insoluble NSP, and to their enormous water-binding capacity [24, 25]. While most experiments examine the influence of NSP on gastric emptying, there are only a few which consider physiological effects in the small intestine.

In addition to changed retention times in the digestive tract, the intestine motility inhibiting effects of soluble NSP play an important role [26–28]. The formation of a water layer between digesta and mucosa, and decreased motility lead to a reduced contact between digestible substrate and the specific enzymes and might cause problems involving nutrient absorption due to a reduced contact between resorbable substrates and mucosa.

The degree of thickening when exposed to fluids depends on the chemical composition and concentration of the polysaccharide [13]. Concentration, conformation and molecular weight distribution of soluble NSP is important for rheological properties such as viscosity and gel formation.

Measurement of WEV in cereals is an indirect means of estimation of their soluble non-starch polysaccharide content [29, 30].

Experiments were conducted to investigate the influence of some extraction conditions: granulation, extraction temperature and time, on the viscosity of wheat and barley aqueous extracts, and to optimize the method for obtaining the soluble NSP extract [31].

2.2 Experimental

The water-soluble fraction was obtained without endogenous enzyme inactivation, using a simple water extraction, with constant shaking, at three different temperatures: 20, 40 and 60 °C. Incubation was performed for 15, 30 and 60 min for each temperature.

The grains were milled by a laboratory grinder to pass through a 500 μm sieve. The extracts were obtained at a ratio of flour to deionised water of 1:2, by shaking the mixtures at 150 rpm in a LabTech LSB-015S water bath.

The extracts were centrifuged with a Hettich 320R centrifuge for 10 min at 5,000 rpm and 25 °C.

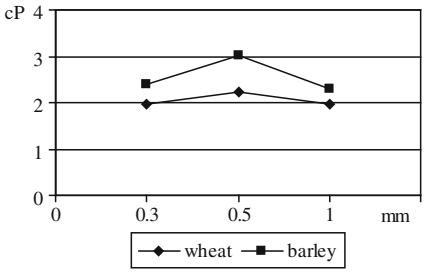


Fig. 22.1 The effect of particle size on the dynamic viscosity of the soluble fraction in wheat and barley

Following the centrifugation, an aliquot of 0.5 mL supernatant was removed and assayed for dynamic viscosity at different times after the extract separation. Viscosity measurements were carried out at 25 °C using a Wells Brookfield Cone/Plate Digital Viscometer Model DVIII Cone CP-40. All results were expressed in cP and calculated also as values relative to that of water.

22.3 Results and Discussion

First experiment studied the effect of granulation on WEV. Grinding grain samples breaks cell walls that contain NSP (cellulose, arabinoxylans, β-glucans) that are resistant to digestive enzymes [32, 33] and facilitates exposure of digestible components (starch, protein) to the action of digestive enzymes. Grinding increases the total surface relative to volume and thus exposure of nutrients to the action of digestive enzymes is improved [34]. Grinding contributes to digestive fluid, better mixing nutrients with digestive enzymes, and implicitly to more efficient use of cereals [35].

WEV appeared to be related to the particle size of the meals obtained after grinding. We studied the relationship between WEV and grain size for three granulations: 0.3, 0.5, and 1 mm. The results of experiments show that viscosities of aqueous extracts of wheat and barley samples have maximum values at 0.5 mm size. At lower or higher granulation the observed viscosities are lower and approximately equal (Fig. 22.1).

Second experiment studied the effect of extraction temperature and time on the dynamic and relative viscosity of aqueous extracts. The influence of the time elapsed from the extract isolation until the viscosity measurement (time after centrifugation) on WEV was also studied. In order to observe better the variations in viscosity values, all experiments were carried out with wheat and barley grains milled by a laboratory grinder to pass through a 500 μm sieve.

The experimental results are presented in Tables 22.2 and 22.3. WEV values for barley were higher than those for wheat, because of the presence of very high molecular weight β-glucans in barley.

Table 22.2 Dynamic and relative viscosities of wheat extracts

Incubation temperature (°C)	Incubation time (min)	Time after centrifugation (min)	Dynamic viscosity (cP)	Relative viscosity (cP)
20	15	0	1.94	2.20
20	30	0	2.20	2.50
20	60	0	2.30	2.61
20	60	60	2.08	2.36
40	15	0	2.34	2.65
40	15	60	2.04	2.31
40	30	0	2.36	2.68
40	30	60	1.93	2.19
40	60	0	2.46	2.79
40	60	30	2.30	2.61
40	60	60	2.04	2.31
60	15	0	7.20	8.18

Graphic representations of WEV obtained from wheat and barley at 20 and 40 °C (Figs. 22.2 and 22.3) reveal higher dynamic viscosities for extraction at 40 °C compared with extraction at 20 °C, NSP solubility increasing with temperature in this range.

The viscosity of the extracts prepared from wheat and barley raised considerable when the extraction temperature was increased from 40 to 60 °C. The significant increase of WEV at 60 °C was probably caused by the gelatinization of the starch. The increase in WEV might be also explained by an aggregation of the polymers. Therefore, we consider that the optimum temperature for extraction of soluble NSP is 40 °C.

The results demonstrated that the viscosities of the water extracts increased with the extraction time up to 60 min. The increase in viscosity with extraction time is the result of the NSP slowly solubilizing, so we recommend an incubation period of 60 min for a good soluble NSP extraction.

WEV values of aqueous extracts obtained at 40 °C and 60 min extraction time show a decrease as the time elapsed from the extract isolation increased up to 60 min (Fig. 22.4). Soluble NSP were hydrolyzed by endogenous enzymes and consequently their molecular mass was reduced. The dynamic viscosity decreased with 22 % for barley and 17 % for wheat when measured after 60 min from the extract isolation.

All cereal grains possess enzymes able to degrade the endosperm cell walls and make accessible starch granules to amylolytic attack during seed germination. Low levels of endogenous activity detectable in cereal flour [36] appear constantly present during grain storage and may slowly degrade soluble NSP.

Careful management of endogenous activity during feed processing could reduce or remove the need for exogenous addition. More usually, however, thermal methods of feed processing rapidly destroy endogenous activities. In addition, thermal processing encourages the release of soluble NSP.

Table 22.3 Dynamic and relative viscosities of barley extracts

Incubation temperature (°C)	Incubation time (min)	Time after centrifugation (min)	Dynamic viscosity (cP)	Relative viscosity (cP)
20	15	0	2.44	2.77
20	30	0	2.64	3
20	60	0	2.74	3.11
20	60	60	2.60	2.95
40	15	0	2.80	3.18
40	15	60	2.40	2.73
40	30	0	3.11	3.53
40	30	60	2.70	3.07
40	60	0	3.32	3.77
40	60	30	2.98	3.38
40	60	60	2.60	2.95
60	15	0	6.9	7.84

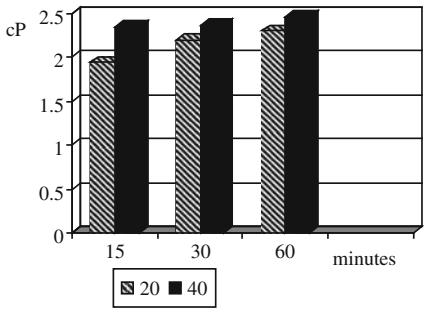


Fig. 22.2 Effect of temperature and extraction time on the dynamic viscosity in wheat

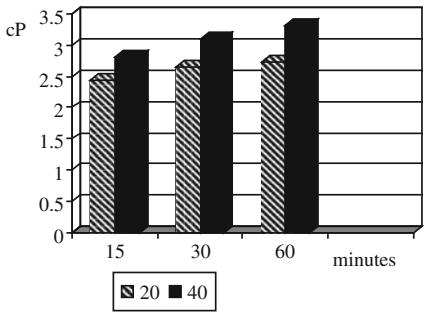


Fig. 22.3 Effect of temperature and extraction time on the dynamic viscosity in barley

Germination of the grains either before or after harvesting decreased β-glucan and pentosan content and viscosity, due to the activation of endogenous enzymes. Newly harvested wheat is often reported to cause nutritional problems in broilers. Prolonged storage of the grains tends to reduce the β-glucan content of barley [37].

The endogenous enzyme content of barley grain was shown by Burnett [38] to be the main reason for the differences found in the nutritive value of U.S. barleys.

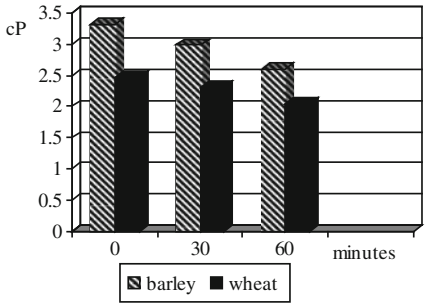


Fig. 22.4 Effect of the time elapsed after centrifugation on the dynamic viscosity in wheat and barley

Preece and McKenzie [39] had previously identified endogenous enzyme systems in barley grain responsible for the decrease in viscosity without an increment in reduced sugars content; that is, these enzymes had the capacity to modify the structure, but not the content of β-glucans. As long as barley grain has endogenous enzyme activity, there could be a transformation from soluble to insoluble β-glucans (with a concomitant decrease in viscosity). Thus, the existence of endo 1-3 β-glucanase activity in barley is possible.

22.4 Conclusion

The experiments revealed as optimum conditions for obtaining the soluble NSP extract from wheat and barley and for WEV determination:

- granulation of 0.5 mm size
- extraction temperature 40 °C
- extraction time 60 min
- viscosity measurements immediately after extract isolation.

Acknowledgments This work was supported by CNCSIS–UEFISCSU, project number 1054/2009 PNII—IDEI code 894/2008, and project number 1055/2009 PNII—IDEI code 898/2008

References

1. Bach Knudsen KE, Jørgensen H (2001) Intestinal degradation of dietary carbohydrates—from birth to maturity. In: Lindberg JE, Ogle B (eds) Digestive physiology of pigs. CABI publishing, Wallingford, pp 109–120
2. Ebihara K, Kiriyama S (1982) Comparative effects of water-soluble and water-insoluble dietary fibers on various parameters relating to glucose tolerance in rats. Nutr Rep Int 26(2):193–201
3. Englyst HN, Hudson GJ (1987) Colorimetric method for routine measurement of dietary fibre as non-starch polysaccharides. A comparison with gas-liquid chromatography. Food Chem 24:63–76

4. Southgate DAT (1995) The diet as a source of dietary fiber. Eur J Clin Nutr 49(Suppl 3): S22–S26
5. Cummings JH (1997) Bowel habit and constipation. In: Bruxelles (ed) The large intestine in nutrition and disease. Institut Danone pp 87–101
6. Sasaki T (2004) Effect of water-soluble and insoluble non-starch polysaccharides isolated from wheat flour on the rheological properties of wheat starch gel. Carbohydr Polym 57(4):451–458
7. Sasaki T, Yasui T, Matsuki J (2000) Effects of amylose content on gelatinization, retrogradation, and pasting properties of starches from waxy and nonwaxy wheat and their F1 seeds. Cereal Chem 77:58–63
8. Izydorczyk MS, Biliaderis CG, Bushuk W (1991) Physical properties of water-soluble pentosans from different wheat varieties. Cereal Chem 68:139–144
9. Izydorczyk MS, Biliaderis CG, Bushuk W (1991) Comparison of the structure and composition of water-soluble pentosans from different wheat varieties. Cereal Chem 68:145–150
10. Izydorczyk MS, Macri LJ, MacGregor AW (1998) Structure and physicochemical properties of barley non-starch polysaccharides. I. Water-extractable β-glucans and arabinoxylans. Carbohydr Polym 35(3):249–258
11. Johnson IT (2001), New food components and gastrointestinal health". Proc Nutr Soc 60:481–488
12. Lineback DR, Rasper VF (1988) Wheat carbohydrates. In: Pomeranz Y (ed) Wheat: chemistry and technology, vol 1. American Association of Cereal Chemistry, Saint Paul, pp 277–371
13. Schweizer TF, Würsch P (1981) Analysis of dietary fiber. In: James WPT, Theander O (eds) The analysis of dietary fiber in food. Marcel Dekker, New York, pp 203–216
14. Burnett GS (1966) Studies of viscosity as the probable factor involved in the improvement of certain barleys for chickens by enzyme supplementation. Br Poult Sci 7:55–75
15. Choct M, Kocher A (2000) Non-starch carbohydrates: digestion and its secondary effects in monogastrics. Proc Nutr Soc 24:31–38
16. Sinha AK, Kumar V, Makkar HPS, De Boeck G, Becker K (2011) Non-starch polysaccharides and their role in fish nutrition—a review. Food Chem 127:1409–1426
17. Bedford MR (1995) Mechanism of action and potential environmental benefits from the use of feed enzymes. Animal Feed Sci Technol 53:145–155
18. Steenfeldt S, Knuden KE, Boring CF, Eggum BO (1995) The nutritive value of decorticated mill fractions of wheat. 2. Evaluation with raw and enzyme treated fractions using adult cockerels. Animal Feed Sci Technol 54:249–265
19. Choct M, Annison G (1992) The inhibition of nutrient digestion by wheat pentosans. Br J Nutr 67:123–132
20. Annison G (1993) The role of wheat non-starch polysaccharides in broiler nutrition. Aust J Agric Res 44:405–422
21. Bedford MR, Morgan AJ (1996) The use of enzymes in poultry diets. World's Poult Sci J 52:61–68
22. Smits HM, Veldman A, Verstegen MWA, Beynen AC (1997) Dietary carboxymethyl cellulose with high instead of low viscosity reduces macronutrient digestion in broiler chickens. J Nutr 127:483–487
23. Low AG (1990) Nutritional regulation of gastric secretion, digestion and emptying. Nutr Res Rev 3:229–252
24. Johansen HN, Bach Knudsen KE, Sandström B, Skjøth F (1996) Effects of varying content of soluble dietary fibre from wheat flour and oat milling fractions on gastric emptying in pigs. Br J Nutr 75:339–351
25. Miquel N, Bach Knudsen KE, Jørgensen H (2001) Impacts of diets varying in dietary fibre characteristics on gastric emptying in pregnant sows. Archiv Tierernährung 55:121–145
26. Edwards C (1990) Mechanisms of action of dietary fibre on small intestine adsorption and motility. Adv Exp Med Biol 270:95–104

27. Flourié B (1992) Influence of dietary fibre on carbohydrate digestion and adsorption. In: Schweizer TF, Edwards CA (eds) Dietary fibre-A component of food: nutritional function in health and disease. Springer-Verlag, London, pp 181–196
28. Furda I (1990) Interaction of dietary fibre with lipids-mechanistic theories and their limitations. In: Furda I, Brine CJ (eds) New developments in dietary fibre: physiological, physiochemical and analytical aspects. Plenum Press, New York, pp 67–82
29. Boros D, Marquardt RR, Slominski BA, Guenter W (1993) Extract viscosity as an indirect assay for water-soluble pentosan content in rye. Cereal Chem 70:575–580
30. Saulnier L, Peneau N, Thibault JF (1995) Variability in grain extract viscosity and water-soluble arabinoxylan content in wheat. J Cereal Sci 22:259–264
31. Caprita R, Caprita A, Cretescu I (2011) Effective extraction of soluble non-starch polysaccharides and viscosity determination of aqueous extracts from wheat and barley. Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2011, WCECS 2011, San Francisco, USA, pp 613–616, 19–21 October 2011
32. Jørgensen H, Zhao X-Q, Eggum BO (1996) The influence of dietary fibre and environmental temperature on the development of the gastrointestinal tract, digestibility, degree of fermentation in the hind gut and energy metabolism in pigs. Br J Nutr 75:365–378
33. De Lange CFM (2000) Characterisation of nonstarch polysaccharides in animal feeds. In: New developments in feed evaluation, International Postgraduate Seminar, Wageningen University, The Netherlands, pp 77–92
34. Wondra KJ, Hancock JD, Kennedy GA, Hines RH, Behnke KC (1995) Reducing particle size of corn in lactation diets from 1,200 to 400 micrometers improves sow and litter performance. J Animal Sci 73:421–426
35. Ohh SJ, Allee GL, Behnke KC, Deyoe CW (1983) Effect of particle size of corn and sorghum grain on performance and digestibility of nutrients for weaned pigs. J Animal Sci 57(Suppl. 1):260
36. Cleemput G, Hessing M, van Oort M, Deconynck M, Delcour JA (1997) Purification and characterisation of a β-D-xylosidase and an endo-xylanase from wheat flour. Plant Physiol 113:377–386
37. Brufau J, Francesch M, Perez-Vendrell AM, Esteve-García E (1993) Effect of post-harvest storage on nutritional value of barley in broilers. 1st European Symposium on Feed Enzymes, Ittingen, pp 13–16
38. Burnett GS (1962) The effect of damaged starch, amylolytic enzymes, and proteolytic enzymes on the utilization of cereals by chickens. Br Poult Sci 3:89–103
39. Preece IA, McKenzie KG (1952) Non starch polysaccharides of cereal grains. Fraction of the barley gums. J Inst Brewing 58:353–362

Chapter 23
Selective Ti-Based Homogeneous Catalyst for Ethylene Dimerization Using New Haloethane Promoters and Electron Donor Ligands

Seyed Hamed Mahdaviani, Davood Soudbar and Matin Parvari

Abstract This research deals with investigation of the effects of two haloethanes as new promoters and two electron donor compounds as suitable ligands in the presence of triethylaluminum (TEAl) as activator for ethylene dimerization to 1-butene. We first compared effects of tetrahydropyran (THP) and 2,5-dimethoxytetrahydrofuran (2,5-DMTHF) as donor ligands on the catalyst performance in the absence of promoters. The experimental results showed that in general the performance of the catalyst system containing 2,5-DMTHF was better than that containing THP in the selected reaction conditions. Then, the effects of chloroethane (CE) and bromoethane (BE) as new promoters were examined on the $Ti(OC_4H_9)_4$/TEAl/2,5-DMTHF catalyst system over a very wide promoter/Ti molar ratio. The data obtained from the relevant experiments showed that ethylene conversion, overall selectivity to 1-butene, and yield of reaction were all higher for the CE compared to BE in the aforementioned catalyst system. The further postulations were drawn for elucidation of the suitable effect of these promoters.

Keywords Electron donor ligand · Ethylene dimerization · Haloethane · PE · Promoter · TEAl · Yield

S. H. Mahdaviani (✉) · M. Parvari
Chemical Engineering Department, Iran University of Science and Technology (IUST), Tehran, Iran
e-mail: hamed_mahdaviani@chemeng.iust.ac.ir

M. Parvari
e-mail: parvari@iust.ac.ir

S. H. Mahdaviani · D. Soudbar
Research and Development Center (R&D Division), Arak Petrochemical Company (ARPC), Arak, Iran

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_23,

23.1 Introduction

Linear α-olefins (LAOs) such as 1-butene, 1-hexene, and 1-octene are important co-monomers for the production of linear low-density polyethylene (LLDPE) or high-density polyethylene (HDPE), PVC plasticizers, alcohols, aldehydes, carboxylic acids, and detergents. LAOs are predominantly produced by metal-catalyzed oligomerization of ethylene, which usually leads to a mathematical distribution of α-olefins like the Schulz-Flory or Poisson with different chain length that frequently does not match the market demand and provides a serious challenge to subsequent separation processes to economically obtain the specific desirable LAOs [1, 2]. Increasing the selectivity of these processes by developing better catalysts is therefore one of the major challenges. Accordingly, there is a driving force behind the multiple endeavors to perform the scientific investigations and the industrial developments for finding new catalytic systems in the selective di-, tri-, and tetramerization of ethylene that leads to the production of the certain LAOs (1-butene, 1-hexene, and 1-octene) [3–5].

1-butene, a basic petrochemical building block of captive requirements, is the first member of LAOs. Although the properties of LLDPE (for example tear resistance) prepared by using 1-hexene and 1-octene as co-monomers is superior to the properties of LLDPE prepared by using 1-butene as co-monomer [6], but conventional full range producers of LAOs have to meet a formidable challenge to match the market demand. On the basis of a scientific report, the average annual increase in demand for 1-butene is 5.3 % from 2006 to 2020 even more than estimated average annual growing rate for 1-hexene production (i.e. 4.7 %) [7]. Since 1-butene is a versatile chemical intermediate to a wide variety of the industrial products with a desirable price, a considerable amount of research effort has been dedicated both recently and in the past to the development of the improved ethylene dimerization processes [7–9]. On the other hand, the dimerization of ethylene has been of great interest from both academic and technological viewpoints, since there are a number of different processes that can happen during the reaction, some of which lead to isomers of 1-butene, linear oligomer formation, and polymeric products [10, 11]. By investigation of previously reported studies, two mechanisms for the ethylene dimerization are more valuable: (1) The bimetallic Titanium-Aluminum complex that was proposed by Angelescu et al. [8, 12]. It has been depicted in Fig. 23.1. (2) Cyclic intermediate mechanism which proceeds by four steps involving the complexation of two molecules of ethylene on a titanium atom affording a Ti(IV) cyclopentane intermediate species. Subsequently, it is converted to the π-bonded 1-butene complex via β-H transfer and indeed reductive elimination is occurred. Finally, Ti species are regenerated [12]. This mechanism has been shown in Fig. 23.2.

Generally, the process of ethylene dimerization is conducted in liquid phase with the homogeneous catalytic system e.g. $Ti(OC_4H_9)_4$–$Al(C_2H_5)_3$-modifier (an electron donor compound) [7]. $Ti(OC_4H_9)_4$ is the main catalyst [12]. TEAl is an activator which can release free coordination sites in titanate complex and generate

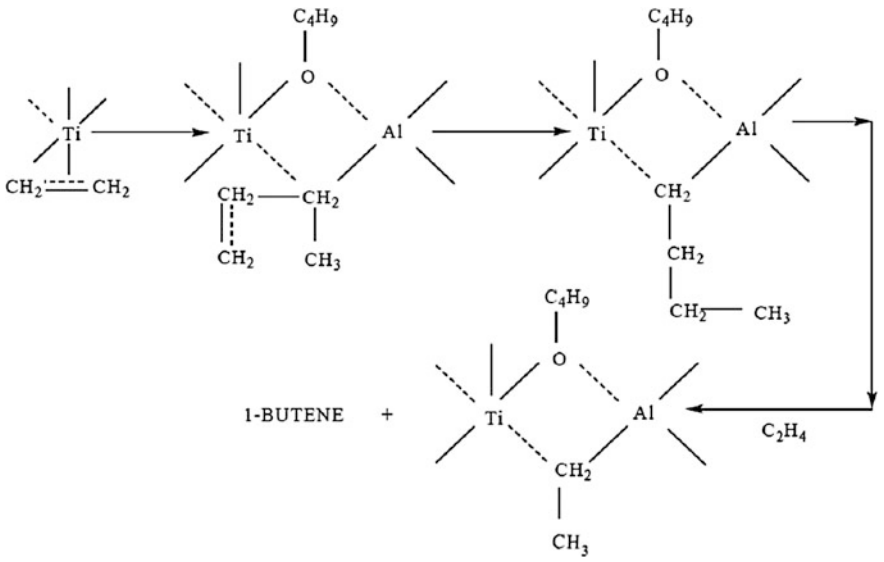


Fig. 23.1 Dimerization of ethylene to bimetallic titanium–aluminium complex

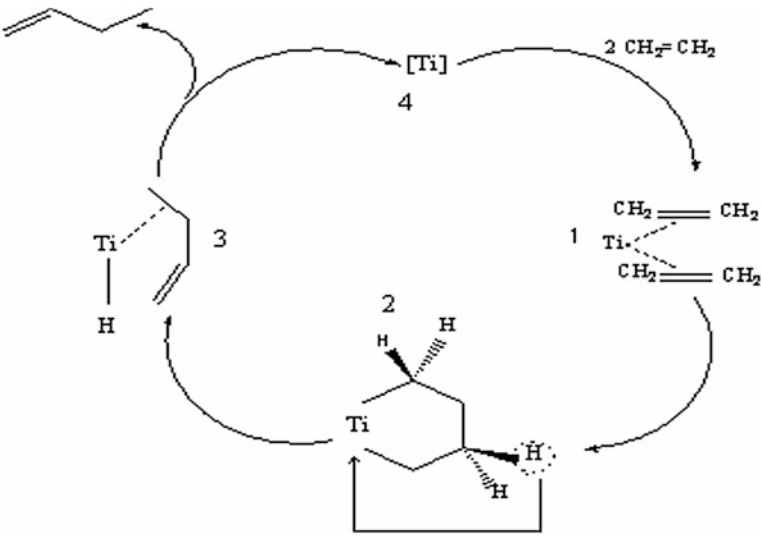


Fig. 23.2 Schematic representation of the cyclic intermediate mechanism

one or more Ti–C bonds by exchanging its ethyl groups with the butoxide groups of the titanate complex [12]. It also enhances the dimerization rate and has a profound effect on the course of the reaction. Usually, the catalyst modifiers are polar additives (such as tert-phosphine, phosphate, amine, and cyclic ether) which, when added to the catalyst system, provide better selectivity for the reaction.

The modifiers influence the mode of linkage of ethylene molecules and inhibit the formation of heavy compounds [9]. Several studies were conducted to elucidate the necessary characteristics of these ligands to catalyze the selective dimerization [9, 11, 12].

Promoters, which almost always are halide compounds, have been extensively studied in tri- and tetramerization of ethylene [13–17]. The promoters play an important role in assisting the central metal of the relevant catalyst system to achieve high selectivity in favor of formation of the desired product and high catalytic activity [16]. This account is based on the findings of research studies of Yang et al. [13] and Chen et al. [15]. They reported that the addition of geminal chloro compounds and halides to the catalyst systems at ethylene trimerization and tetramerization reactions resulted in significant improvement of selectivity to 1-hexene and 1-octene, respectively.

To our knowledge, there is not any report in literature on the use of halocompounds as promoters for titanium-based catalysts in homogeneous ethylene dimerization reactions. In our previous research work, we introduced 1,2-dichloroethane (DCE) to be a suitable vicinal chloro compound for use in the ethylene dimerization reaction [18]. This result indicated that promoters suitable for the titanium-catalyzed ethylene dimerization are different from those suitable for chromium-based catalysts in the tri- and tetramerization of ethylene. Also in another study, we could find another new promoter i.e. ethylene chlorobromide which was effective in combination with 2,5-dimethoxytetrahydrofuran (2,5-DMTHF) as electron donor compound for production of 1-butene in the dimerization of ethylene [19].

In the present study, we first compared effects of two electron donor ligands i.e. tetrahydropyran (THP) and 2,5-DMTHF as modifiers on the homogeneous $Ti(OC_4H_9)_4$/TEAl catalyst system. The experimental results showed that the performance of the catalyst system containing 2,5-DMTHF was generally better than that containing THP. Then we proceeded to investigate the influences of the addition of two halohydrocarbons to the $Ti(OC_4H_9)_4$/TEAl/2,5-DMTHF catalyst system in the various molar ratios of promoter/Ti. Indeed, this paper is the extended of our recent study [20] with more details of our results.

23.2 Experimental

23.2.1 Materials and Instruments

2,5-DMTHF, THP, CE (2 M solution in diethyl ether), and BE were purchased from Merck. TEAl was obtained from Crompton Chemicals and was diluted to a 0.5 M solution in heptane before use. n-Heptane was distilled from anhydrous sodium carbonate under dry nitrogen and stored over pre-activated molecular sieves (4 Å) until its water content was below 1 ppm. Other chemicals were obtained commercially and used as received.

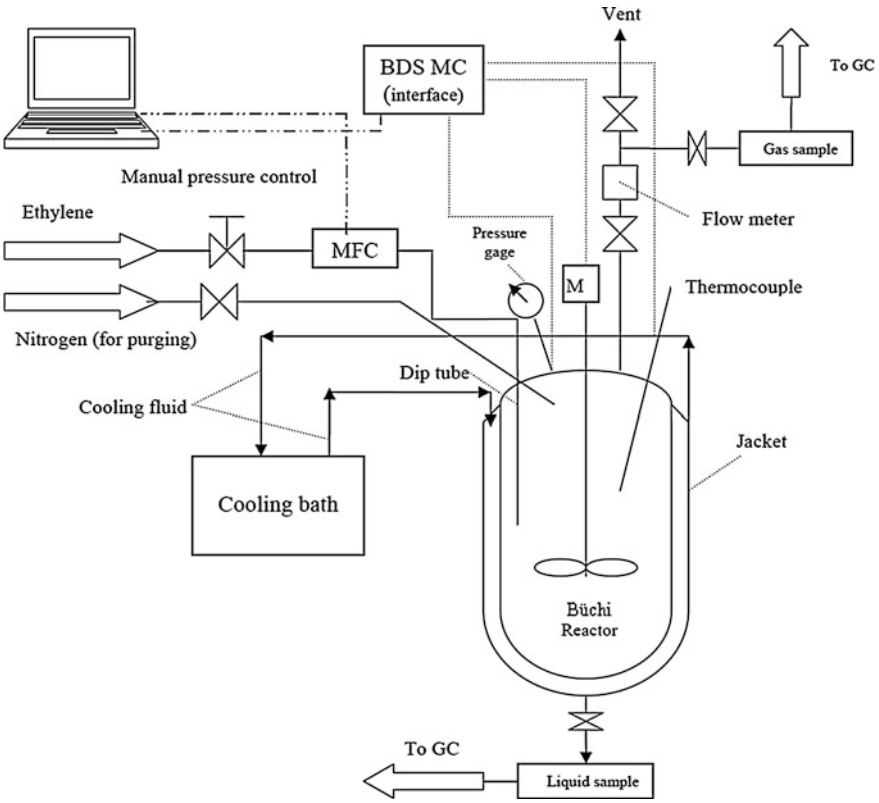


Fig. 23.3 Schematic set-up of the büchi reactor used in this study

A 1-L double-walled stainless steel büchi pressure reactor equipped with an external circulation bath for temperature control, a magnetically driven mechanical stirrer for saturation of inlet gas in the liquid phase, an internal thermocouple, gas inlet and outlet ports, and a liquid sampling port was used. The reactor was set up with a büchi multi channel data system (BDS MC) to display and record the temperature, pressure, and stirrer speed profiles with reaction time. A schematic representation experimental set-up used in this research is depicted in Fig. 23.3.

Gas chromatography with flame-ionization detector (GC/FID) analyses of reaction products were carried out on a Varian 3800 chromatograph using a CP Sil 8 capillary column (25 m × 0.53 mm). The column oven temperature of the GC was programmed to run from 40 to 280 °C at a rate of 10 °C/min.

A TA4000 system differential scanning calorimeter (DSC) was used for exact identification and characterization of the polymers formed in certain runs based on the thermal properties (i.e. melting and degradation behavior and degree of crystallinity). The DSC measurements were recorded during second heating/cooling cycle with the heating rate 10 °C/min.

23.2.2 Reaction Procedure and Product Analysis

All manipulations involving air- and moisture-sensitive materials have been performed under atmosphere of dry nitrogen using standard Schlenk techniques.

Before conducting a catalytic batch experiment, the reactor was heated to 100 °C for an hour to eliminate traces of water, air, and impurities. Then, it was cooled to ambient temperature and was swept with dry nitrogen for 30 min. Thereafter, the reactor was charged with 400 ml of n-heptane as solvent to provide sufficient liquid height to ensure successful operation of the gas entrainment stirrer for both mass and heat transfer. Subsequently, the reactor was heated to a temperature lower than the desired temperature. It is noteworthy that the internal temperature of the reactor was adjusted to a level 4–6 °C lower than the desired value because of the highly exothermic character of ethylene dimerization. Ethylene was then introduced into the reactor to the desired pressure. The temperature inside of the reactor was controlled using cooling fluid, if required. Then, the calculated quantities of $Ti(OC_4H_9)_4$, TEAl, ligand, and promoter were immediately injected into the reactor. At this moment, agitation was started. The speed of the stirrer was initially set to 900 rpm. As the reaction progressed, a drop in the ethylene pressure was observed. The speed of the stirrer was also reduced with increasing reaction time.

The volume of ethylene introduced through the inlet was measured using a Brooks mass-flow controller (MFC). Also, the total volume of gaseous components was measured by means of a gas flowmeter. After the dimerization was allowed to proceed for 0.5 h, the reaction was terminated by switching off the gas entrainment stirrer and the products were withdrawn.

A gas sample was collected in a 150 ml stainless steel bomb and then was analyzed by GC-FID. A liquid sample was distilled to remove catalyst and TEAl. Finally, a sample was analyzed with GC-FID. The polymers formed in certain runs were removed, washed with hexane, dried in a vacuum oven at 100 °C, weighed, and ultimately characterized by differential scanning calorimetry (DSC). The melting and degradation points were measured as 129 and 221 °C, respectively, and the degree of crystallinity was 57 %. It was found that the polymers formed were polyethylene (PE).

The conversions and product selectivities were evaluated from the mass balance for ethylene consumption based on measured values from the MFC, flowmeter and GC analyses of the gaseous and liquid products, the liquid product weight and the gas volume.

The yield of the reaction was calculated as:

$$Yield(\%) = \frac{(ethylene\ conversion(\%)) \times (overall\ selectivity\ to\ 1-butene(\%))}{100} \tag{23.1}$$

Table 23.1 The effects of electron donor ligands on the ethylene dimerization

No.	L* symbol	L/Ti (mol/mol)	Con. (%)	Overall selectivity to product (%wt)		PE (mg)	Yield (%)
				1-C_4	Oligomers		
1		2	72.24	74.77	25.23	45	54.00
2		3	74.82	75.32	24.68	40	56.35
3	(a)	4	77.45	75.90	24.10	32	58.78
4		5	75.38	74.96	25.04	45	56.50
5		6	70.66	73.62	26.38	50	52.00
6		2	71.88	74.86	25.14	40	53.81
7		3	73.55	75.41	24.59	35	55.46
8	(b)	4	76.12	76.10	23.90	30	57.92
9		5	74.90	75.20	24.80	35	56.32
10		6	70.22	73.97	26.03	40	51.94

Reaction conditions: reaction temperature: 55 °C, initial pressure of ethylene: 25 bar, reaction time: 45 min, solvent: n-heptane, Al/Ti = 4 molar ratio

* L is ligand (modifier)

23.3 Results and Discussion

23.3.1 Comparison of Two Electron Donor Ligands

In present work, we first examined the influence of 2,5-DMTHF and THP as modifiers under selected operating conditions. Table 23.1 lists the effects of these electron donor ligands on the properties of the $Ti(OC_4H_9)_4$/TEAl/ligands catalyst systems. Fig. 23.4 is a schematic chemical structure of the electron donor compounds as modifiers.

As shown in Table 23.1, the performance of the $Ti(OC_4H_9)_4$/TEAl/2,5-DMTHF catalyst system is better than $Ti(OC_4H_9)_4$/TEAl/THP catalyst system in terms of conversion and yield in the selected operating conditions. A slight decrease of overall selectivity to 1-butene in the catalyst system containing 2,5-DMTHF compared with that containing THP may be probably due to inductive electron-withdrawing effect of two methoxy groups on the oxygen atom of THF ring which leads to the decrease of oxygen atom basicity of THF ring and consequently the coordination power of 2,5-DMTHF ligand over titanium active center is decreased and Ti(IV) complex is not well stabilized. It is also necessary to mention that a considerable feature obtained from our results for two aforementioned catalyst systems was the enhancement of PE produced in the catalyst system containing 2,5-DMTHF about 5-10 mg. This disparity in the results obtained may conceivably be due to reaction of the methoxy moiety of the ligand with the TEAl co-catalyst which is a Lewis acid [21].

From examination of the results in Table 23.1, it was apparent that the increase of ligand/Ti molar ratio from 2 to 4 led to increases in ethylene conversion, 1-butene selectivity and yield to a maximum value. However, when the ligand/Ti

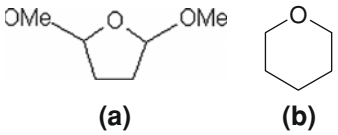


Fig. 23.4 Schematic chemical structure of electron donor ligands. **a** 2,5-DMTHF, **b** THP

molar ratio was further increased, the values fell. This is expected since any excess ligand can interfere with the formation of the active Ti species or may prevent coordination of ethylene at the active dimerization sites [19].

23.3.2 Effect of CE and BE as New Efficient Promoters in the Ethylene Dimerization

Based on the four-membered homogeneous titanium catalyst system [$Ti(OC_4H_9)_4$)/2,5-DMTHF/TEAl/halide compound], effects of CE and BE as haloethanes with various molar ratios of haloethane/Ti on ethylene conversion, overall selectivity to 1-butene, and yield were investigated. The results are summarized in Table 23.2.

According to the Table 23.2, it is apparent that an increase of halide/Ti molar ratio until 6 resulted in a continuous enhancement of ethylene conversion, overall selectivity for 1-butene, and yield (runs 12–17 and 23–28). Further augmentation of halide/Ti molar ratio had a negative effect on the catalytic performance (runs 18–21 and 29–32). The detrimental effect of halide/Ti at high molar ratios is due to a coordinative loading of the active sites by more than one halide per metal atom and thus it leads to the preventing ethylene coordination [13, 15]. On the other hand, this trend can be rationalized that any excess promoter may interfere with the formation of the active Ti species or by over-reduction of active Ti species [18, 20]. From Table 23.2, it is seen that conversion, overall selectivity for 1-butene, and yield (respectively 86.10, 83.12, and 71.56 %) in the optimum molar ratio of halide/Ti equal to 6 of the catalyst system containing CE is better than of a corresponding BE. This result can be likely attributed to the steric environment of the Br atoms around the catalyst center [15]. It is proposed that since the spatial encumbrance of BE is greater than CE, it can suppress the process of facile coordination and repetitive insertion of ethylene molecules at the metal center. Also, the steric hindrance of the Br atoms is responsible for a decrease in the amount of β-hydrogen transfer process and hence the liberation of the dimer from the active sites. Indeed, considering that the coordination ability of the bromo group towards the central metal of the catalyst may be stronger than the coordination ability of the chloro group towards the central metal of the catalyst [16, 22], it may be suggested that the better improvement effect of chloride on both the ethylene conversion and the overall selectivity to 1-butene is due to more suitable coordination ability with titanium center.

Table 23.2 The effects of promoters on the ethylene dimerization

No.	P*	P/Ti (mol/mol)	Con. (%)	Overall selectivity to product (%wt)		PE (mg)	Yield (%)
				1-C_4	Oligomers		
11		0	77.45	75.90	24.10	30	58.78
12		1	79.41	77.47	22.53	27	61.52
13		2	80.66	78.66	21.34	25	63.45
14		3	81.93	79.52	20.48	22	65.15
15		4	83.14	80.72	19.28	20	67.11
16	CE	5	84.88	81.85	18.15	15	69.47
17		6	86.10	83.12	16.88	10	71.56
18		7	85.15	82.34	17.66	15	70.11
19		8	82.56	80.80	19.20	20	66.71
20		9	78.22	77.83	22.17	25	60.88
21		10	73.60	74.10	25.90	30	54.54
22		0	77.45	75.90	24.10	30	58.78
23		1	79.00	76.33	23.67	27	60.30
24		2	80.50	77.10	22.90	25	62.06
25		3	81.77	78.15	21.85	22	63.90
26		4	82.83	79.24	20.76	20	65.63
27	BE	5	84.43	80.17	19.83	17	67.69
28		6	85.90	81.16	18.84	14	69.71
29		7	84.14	80.64	19.36	18	67.85
30		8	81.91	78.12	21.88	25	63.99
31		9	78.11	75.82	24.18	30	59.22
32		10	73.40	72.45	27.55	35	53.18

Reaction conditions: reaction temperature: 55 °C, initial pressure of ethylene: 25 bar, reaction time: 45 min, solvent: n-heptane, Ti(IV)/2,5-DMTHF/TEAl molar ratios = 1:4:4

* P is promoter

With more depth investigation on the remarkable effect of CE in the aforementioned catalytic system, the following concluding remarks can be drawn:

Generally, because of compatibility of aluminum alkyls and chlorinated hydrocarbons, an electrophilic process generating what may be looked on as a carbonium counter ion pair, a free carbonium ion, or a well-defined polar complex $Et_3Al + RX \rightleftharpoons [R]^+ [Et_3AlX]^-$ (1) (at this complex X represents Cl group) which then may react further depending on the reaction conditions and the nature of reactants [23]. Indeed, owning the acidity of TEAl and nature of CE and with polarization of the C–Cl bond, the compatible polar complex (1) is generated at reaction media which its chemical property and consecutive specific interaction in a head-on orientation toward metal center of the catalyst system may cause monomeric TEAl is released [13]. Considering the fact that the conventional homogeneous catalyst system of ethylene dimerization is a dual functional catalyst system, monomeric TEAl can modify spatial and electronic properties of Ti active sites giving concomitant generation of the increased number of these active

species whereby the dimerization component of dual functional catalyst is provoked and finally amounts of polymeric compounds are suppressed. As another proposal, the suitable coordination ability may exist between chloro group of CE and central titanium of five-membered Ti metallacyclic transition state which may generate a specific structural arrangement of catalyst system at reactional mixture [20]. Therefore, the chemical micro-surrondings of titanium center in the catalyst system is changed. This geometric regulation stabilizes titanium cationic complex in a higher oxidation state which is responsible for the ethylene dimerization and production of 1-butene.

The exact geometric structures of possible molecular complexes between the halogen groups of the promoters used and the titanium center of the catalyst in the reaction media should be detected by strong experimental evidence such as X-ray crystallographic analysis and variable-temperature NMR spectroscopy. Also, spectroscopic in situ measurements would certainly be preferable, but are not straightforward due to the paramagnetic nature of titanium complexes and the fast reaction. However, in order to elucidate the detailed mechanism of ethylene dimerization in the catalytic cycle at presence of promoters, more extensive research and studies by quantum chemistry calculations and density functional theory (DFT) modeling are still needed.

23.4 Conclusion

In this research, we first investigated the effects of two electron donor compounds as modifiers on the catalytic performance of the ethylene dimerization reaction. In general with comparison of yields calculated, 2,5-DMTHF gave a better performance than THP in the catalyst system. Then, we found the addition of a small amount of halohydrocarbon as promoter had a significant effect on the relevant catalytic system. Comparative studies showed that the CE-promoted [$Ti(OC_4H_9)_4$/TEAl/2,5-DMTHF] catalyst system was very active and selective for the dimerization of ethylene.

Acknowledgments We appreciate the R&D department of Arak Petrochemical Company (ARPC) for their fruitful collaboration. Especially, S.H. Mahdaviani is very grateful to Mr. A. Saadatmand, head of the research laboratory of the R&D department of ARPC, for assistance and advice during working with set up of the 1 L büchi reactor and also would like to express his indebtedness towards Mr. A. Nassiri who is one of the staff of the R&D department of ARPC for help in performing the dimerization experiments and the characterization of polymers produced in reactions by DSC analysis as part of a fellowship awarded to him.

References

1. Vogt D (2002) In: Cornils B, Herrmann W (eds) Applied homogeneous catalysis with organometallic compounds, vol 1. Wiley-VCH, Weinheim, p 240
2. Peitz S, Aluri BR, Peulecke N, Müller BH, Wöhl A, Müller W, Al-Hazmi MH, Mosa FM, Rosenthal U (2011) An alternative mechanistic concept for homogeneous selective ethylene oligomerization of chromium-based catalysts: binuclear metallacycles as a reason for 1-octene selectivity. Chem Eur J 16(26):7670–7676
3. McGuinness DS (2011) Olefin oligomerization via metallacycles: dimerization, trimerization, tetramerization, and beyond. Chem Rev 111(3):2321–2341
4. Speiser F, Braunstein P, Saussine L (2005) Catalytic ethylene dimerization and oligomerization: recent developments with nickel complexes containing P,N-chelating ligands. Acc Chem Res 38(10):784–793
5. Fritz PM, Bölt H, Wöhl A, Müller W, Winkler F, Wellenhofer A, Rosenthal U, Hapke M, Peulecke N, Müller BH, Al-Hazmi MH, Aliyev VO, Mosa FM (2009) Catalyst composition and process for di-,tri- and/or tetramerization of ethylene. WO 2009/006979A2
6. Kissin YV (2005) Polymers of Higher Olefins. Kirk-Othmer encyclopedia of chemical technology. Wiley, New York
7. Belov GP, Matkovsky PE (2010) Processes for the production of higher α-olefins. Petrol Chem 50(4):283–289
8. Pillai SM, Ravindranathan M, Sivaram S (1986) Dimerization of ethylene and propylene catalyzed by transition-metal complexes. Chem Rev 86(2):353–399
9. Al-Jaralleh AM, Anabtawi JA, Siddiqui MAB, Aitani AM, Al-Sadoun AW (1992) Ethylene dimerization and oligomerization to 1-butene and linear alpha-olefins: a review of catalytic system and processes. Catal Today 14:1–121
10. Belov GP (2008) Selective dimerization, oligomerization, homopolymerization and copolymerization of olefins with complex organometallic catalysts. Russ J Appl Chem 81(9):1655–1666
11. Pillai SM, Tembe GL, Ravindranathan M, Sivaram S (1988) Dimerization of ethylene to 1-butene catalyzed by the titanium alkoxide-trialkylaluminum system. Ind Eng Chem Res 27(11):1971–1977
12. Al-Sadoun AW (1993) Dimerization of ethylene to butene-1 catalyzed by $Ti(OR')_4$-AlR_3. Appl Catal A 105(1):1–40
13. Yang Y, Kim H, Lee J, Paik H, Jang HG (2000) Roles of chloro compound in homogeneous [$Cr(2\text{-ethyhexanoate})_3$/2,5-dimethylpyrrole/triethylaluminum/chloro compound] catalyst system for ethylene trimerization. Appl Catal A 193(1–2):29–38
14. Mahomed H, Bollmann A, Dixon JT, Gokuln V, Griesel L, Grove C, Hess F, Maumela H, Pepler L (2003) Ethylene trimerisation catalyst based on substituted cyclopentadienes. Appl Catal A 255(2):355–359
15. Chen H, Liu X, Hu W, Ning Y, Jiang T (2007) Effects of halide in homogeneous Cr(III)/PNP/MAO catalytic systems for ethylene tetramerization toward 1-octene. Mol Catal A 270(1–2): 273–277
16. Leo H-K, Li DG, Li S (2004) The effect of halide and the coordination geometry of chromium center in homogeneous catalyst system for ethylene trimerization. Mol Catal A 221(1–2):9–17
17. Jiang T, Ning Y, Zhang B, Li J, Wang G, Yi J, Huang Q (2006) Preparation of 1-octene by the selective tetramerization of ethylene. Mol Catal A 259(1–2):161–165
18. Mahdaviani SH, Soudbar D, Parvari M (2010) Selective ethylene dimerization toward 1-butene by a new highly efficient catalyst system and determination of its optimum operating conditions in a Büchi reactor. Int J Chem Eng Applic 1(3):276–281
19. Mahdaviani SH, Soudbar D, Parvari M (2011) [$Ti(OC_4H_9)_4$/2,5-dimethoxytetrahydrofuran/TEA/ethylene chlorobromide] as a novel homogeneous catalyst system effective for ethylene dimerization reaction. World Acad Sci Eng Tech 77:300–305

20. Mahdaviani SH, Parvari M, Soudbar D (2011) Effect of two halohydrocarbons as new promoters suitable for titanium-catalyzed ethylene dimerization toward 1-butene. In: Proceedings of the world congress on engineering and computer science (WCECS 2011). Oct 19–21, 2011, San Francisco, USA, pp 646–650
21. Killian E, Blann K, Bollmann A, Dixon JT, Kuhlmann S, Maumela MC, Maumela H, Morgan DH, Nongodlwana P, Overett MJ, Pretorius M, Höfener K, Wasserscheid P (2007) The use of bis(diphenylphosphino) amines with *N*-aryl functionalities in selective ethylene tri- and tetramerisation. Mol Catal A 270(1–2):214–218
22. Carrasquillo A, Jeng J-J, Barriga RJ, Temesghen WF, Soriaga MP (1997) Electrode-surface coordination chemistry: ligand substitution and competitive of halides at well-defined Pd(100) and Pd(111) single crystals. Inorg Chim Acta 255(2):249–254
23. Miller DB (1966) Reaction of trialkylaluminums with halohydrocarbons. Org Chem 31(3): 908–912

Chapter 24
Distribution of the Distance Between Receptors of Ordered Micropatterned Substrates

Zbigniew Domański and Norbert Sczygiol

Abstract We study the statistics of equally spaced pairs of receptors on a family of ordered flat microsubstrates whose adhesive centers form regular tessellations. We establish relationship between the symmetry of substrates and the probability density of the end-to-end polymer separation in terms of the so-called Manhattan distance.

Keywords Distance distribution · Micropatterned substrates · Polymer adhesion · Surface chemistry · Tessellations · Zigzag path statistics

24.1 Introduction

Adhesion of polymers to metal and ceramics surface is the subject of extensive theoretical and experimental studies. It is mainly due to such industrial applications as composite manufacturing, electronic packaging or production of demanding anti corrosive coatings. In this context the adherence using micropatterned nanosubstrates is an important engineering problem, with very diverse applications ranging from chemical processing to biological applications. Biopolymers play an important role in the exploration of complex-polymer-adhesion processes. Detailed knowledge of the attachment of biopolymers to different substrates is desirable to

Z. Domański (✉)
Institute of Mathematics, Czestochowa University of Technology,
Dabrowskiego 69, 42-201 Czestochowa, Poland
e-mail: zbigniew.domanski@im.pcz.pl

N. Sczygiol
Institute of Computer and Information Sciences, Czestochowa University of Technology,
Dabrowskiego 69, 42-201 Czestochowa, Poland
e-mail: norbert.sczygiol@icis.pcz.pl

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_24,

numerous biotechnologies, such as the fabrication of nanostructures for biomedical applications [1].

The polymer adhesion depends not only on chemical and physical factors. The structure of polymer-solid interfaces is modulated by the relative positions of substrate uptake centers and polymer sticker groups. Therefore, the geometric characteristics like surface topography and topology also come into the picture [2]. From the mathematical point of view a substrate receptor group can be represented by the nodes of an appropriate flat lattice. The geometrical properties of lattices still attract much attention, mainly due to progress in the fields of applied information theory, computer science, statistical physics, biology, and nanotechnology. It is interesting to note that the micropatterned substrates are built with the use of methods borrowed from the semiconductor industry [3, 4]. Such methods generally employ the fabrication of highly ordered microscale structures [5, 6]. A very recent and challenging advancement in this field deals with protein-based programming of quantum nanostructures, as e.g. the self-assembly of quantum dot complexes using nanocrystals capped with specific sequences of DNA [7].

The objective of this work is the theoretical analysis of the polymer adhesion in terms of polymer sticker groups and substrate receptor groups [8]. The problem we consider is the polymer chain trapped by the receptors placed within the nodes of an ordered lattice. In order to enhance the adhesion to the substrate our polymer is functionalized by addition of specified stickers to its ends [9]. In the proximity of the substrate's surface the functionalized polymer feels an attractive, non-homogenous electrochemical potential generated by the receptors. In such circumstances the end points of the polymer do not move freely. They are forced to slide mainly between the nodes, and thus their trajectories resemble zigzag lines. We focus our analysis on the distribution of distances between the end points of the polymer trapped on the ordered surface depending on the surface's symmetry. Because of the already mentioned zigzag-like-shaped trajectories we think that the Euclidean norm is not appropriate to measure the distance traveled by the sticker. Instead of the Euclidean norm the end-point-path lengths are measured in terms of the Manhattan distances.

24.2 Theory and Methodology

In this section, we present an approach we use to calculate, for a given finite substrate, the number of pairs of receptors that are separated by a prescribed distance.

24.2.1 Technological Aspects

In the field of biotechnology micropatterned substrates serve as tools for the creation of novel biologically-inspired materials and for studying mechanisms of cell function [6, 10–12]. The mechanical properties of the substrate to which

polymers adhere mediate many aspects of polymer physicochemical function, as e.g. the DNA ability to take up different signaling molecules. Current efforts to understand the efficiency of adhesion are focused on the surface engineering aspects, whereas the influence of the substrate-receptor-group symmetry is less pronounced and sometimes ignored.

24.2.2 Substrate Space Geometry

Artificial substrate surface architectures employ the lattices which have edges and vertices formed by a regular tiling of a plane, so that all corners are equivalent. For instance in biomedicine, in the context of micropatterned substrates the so-called Archimedean lattices [13] are of special interest. This is because the patterns of extracellular matrix proteins of varying geometries modulate the organization of cells grown on the patterns. Three of the Archimedean lattices: triangular, square and hexagonal are drawn in a plane such a way that all faces are the same whereas the remaining eight lattices need more than one type of a face. The former lattices belong to the regular tessellations of the plane and the latter ones are called semiregular lattices. Another important group of lattices contains dual lattices of the Archimedean ones. The given lattice G can be mapped onto its dual lattice DG in such a way that the center of every face of G is a vertex in DG, and two vertices in DG are adjacent only if the corresponding faces in G share an edge. The square lattice is self-dual, and the triangular and hexagonal lattices are mutual duals. The lattices are labeled according to the way they are drawn [13]. Starting from a given vertex, the consecutive faces are listed by the number of edges in the face, e.g. a square lattice is labeled as (4, 4, 4, 4) or equivalently as (4^4). Consequently, the triangular and hexagonal lattices are (3^6) and (6^3), respectively. Other, frequently encountered lattice is (3, 6, 3, 6)—called Kagomé lattice. In some ways these four lattices are representative to study polymer adhesion problems in two dimension. The regular lattices form pairs of mutually dual lattices and also share some local properties as e.g. the coordination number z being the number of edges with common vertex. One of the most interesting lattices in two dimension is the Kagomé lattice. Each its vertex touches a triangle, hexagon, triangle, and a hexagon. Moreover the vertices of the lattice correspond to the edges of the hexagonal lattice, which in turn is the dual of a triangular lattice. The Kagomé lattice is also related to the square lattice, they have the same value, $z = 4$, of the coordination number. The regular lattices and the Kagomé lattice are presented in Fig. 24.1.

24.2.3 The Manhattan Distance

Many questions lead to a problem of the analysis of properties of random walk path and end-to-end distances distributions on regular networks [14, 15]. Examples

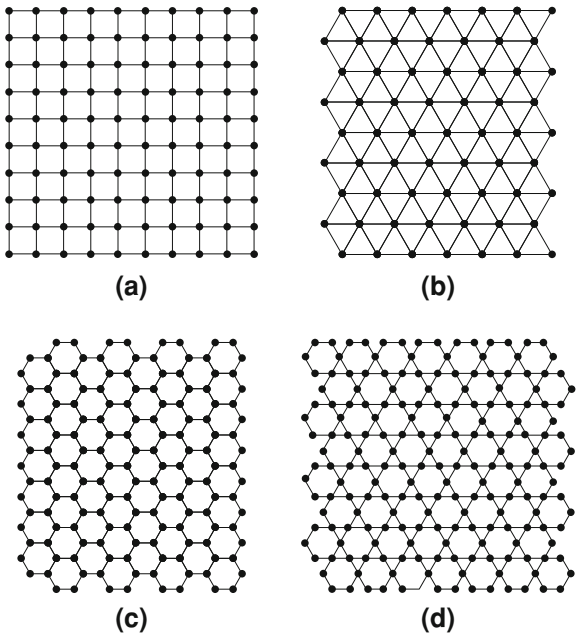


Fig. 24.1 Two dimensional lattices used in this work. They are represented by tessellations with: **a** square, **b** triangular, **c** hexagonal, and **d** Kagomé symmetries. For all tessellations the receptors are identified by the nodes. The edges form the shortest allowed paths between the pairs of receptors

include, but are not limited to, material science or biology. For instance, in the field of computer science an important problem concerns the allocation of processors to parallel tasks in a grid of a large number of processors. This problem relays on the nontrivial correlation between the sum of the pair-wise distances between the processors allocated to a given task and the time required to complete the task [16].

The common question of the above mentioned problems is how many pairs of points separated by a given number q of steps can be found in a bounded region of a two-dimensional lattice. Such number q is referred to as the so-called Manhattan distance.

More specifically, because the distance should be measured in terms of a process and its activities, therefore functional distance should take into account the symmetry of the underlying lattice. A distance measure that accounts for the symmetry can be constructed around the p—norm

$$\|\mathbf{x}\|_p = \left(\sum\nolimits_{i=1,\ldots,n} |x_i|^p\right)^{1/p}. \tag{24.1}$$

For $p = 2$ we have the familiar Euclidean norm and for $p = 1$ we get the Manhattan norm also called the taxicab norm.

This definition is equivalent to the definition of the distance between nodes in the graph that represents the lattice, i.e. the distance between two nodes u and v in a graph is the length of the shortest path from u to v.

24.2.4 Graph Theory Approach

Graphs are useful for representing networks. In this subsection, we briefly present some definitions and background on the graph theory and the method that we use to count the pairs of substrate's nodes separated by a given value of the Manhattan distance. This question belongs to an ample class of the combinatorial properties of Archimedean lattices. It is efficient to study such properties by using tools provided by the graph theory. To do this we map an ordered micropatterned substrate onto a finite, connected graph $G(V, E)$ whose vertices (nodes) $v_{i=1,\ldots,n} \in V$ represent receptors, n is the number of vertices in G. Two vertices are adjacent if they are connected by an edge in E. A walk is a sequence of vertices each of which is adjacent to the previous one. If all vertices are distinct the walk is called a path. The length of a path is the sum of the lengths of all component edges. Since our graph represents the Archimedean lattice then all its edges have the same length and, consequently, the path length is given by the number of visited edges.

An useful concept in the graph theory is the correspondence between graphs and so-called adjacency matrices, sometimes called connectivity matrices. An adjacency matrix $\mathbf{A}_G$ of G is the $n \times n$ matrix whose entries $(\mathbf{A}_G)_{ij} = 1$ if v_i and v_j are adjacent and zero otherwise. An adjacency matrix is very convenient to work with. For instance, let $\mathbf{A}_G^k = \mathbf{A}_G \ldots \mathbf{A}_G$ be the k-times matrix product of $\mathbf{A}_G$, then $(\mathbf{A}_G^k)_{ij}$ is the number of walks of the length k from v_i to v_j in G.

Our approach consists of two steps:

- with the help of the family of matrices $\mathbf{A}_G^k$, each pair of nodes is assigned the smallest value of k so that the corresponding entry of $\mathbf{A}_G^k$ is nonzero;
- for each value of k we count the number of pairs of nodes related to this value.

Since the graph is finite, this approach yields a partition of Manhattan distances.

24.3 Results

We present the detailed calculations of distance distributions for three regular tessellations and for the Kagomé lattice.

24.3.1 Square Lattice

First we analyze the square lattice of receptors with the lattice constant $a \equiv 1$. Without loss of generality, let us assume that the substrate has the shape of a square whose side contains L nodes. Thus, the maximum value of the end-to-end distance $q_{\max} = 2(L - 1)$ corresponds to two pairs of receptors located in the opposite corners of the substrate. On the other hand $q_{\min} = 1$ is related to the number of pairs

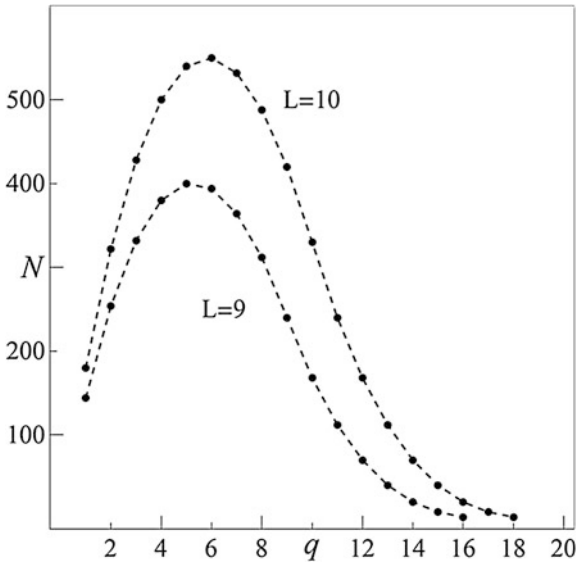


Fig. 24.2 Distribution of the receptor–receptor distance for the square lattice related to two values of the substrate's length L. The lines are drawn using Eq. (24.2) and they are only visual guides

of nodes connected by edges of the substrate. Each of the L rows and columns contain L–1 edges and this means that there are $2L(L–1)$ such distances. Following the approach described in Sect. 24.2.4 we obtain an expression that describes the number $N(L, q)$ of distances q within the square-shaped substrate

$$N(L,q) = \begin{cases} 2Lq(L-q) + \frac{1}{3}(q^2-1)q, & q \le L, \\ \frac{1}{3}\left[(2L-q)^2 - 1\right](2L-q), & L < q \le 2(L-1). \end{cases} \tag{24.2}$$

In Fig. 24.2 we show $N(L, q)$ for different values of L. Equation (24.2) can be written in the form of the probability distribution function of distance with the help of the normalization condition, namely

$$P(L,q) = \frac{2}{L^2(L^2-1)} N(L,q). \tag{24.3}$$

Probability distribution function Eq. (24.3) is useful for small substrates. When L grows significantly an appropriate description is based on the concept of probability density function. The probability density function can be introduced as follows. Let $a \equiv 1 \to \tilde{a} = 1/L$, thus all distances are scaled by the factor $1/L$. Due to this scaling, $P(L, q)$ Eq. (24.3) is replaced by the probability distribution function $\tilde{P}_{1/L}(x_q)$ for the discrete set of distances

$x_q = q/L$ in the unit square with the step $1/L$, i.e.

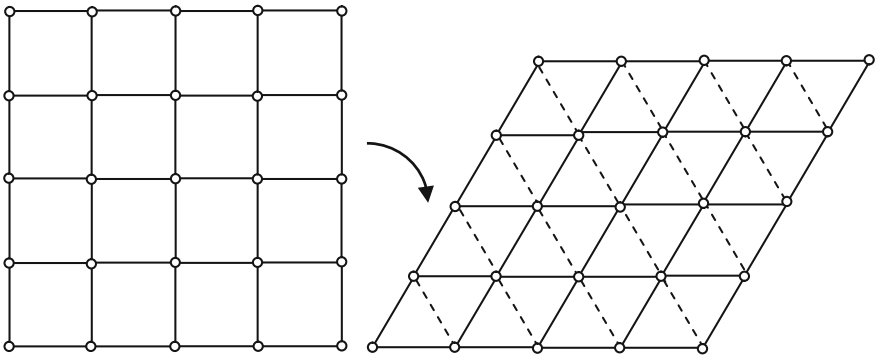


Fig. 24.3 Deformation of the square reduces the diagonals marked by *dashed lines*

$$\tilde{P}_{1/L}(x_q) = \frac{2}{3(L-1/L)} \times \begin{cases} 6x_q(1-x_q) + x_q^3 - x_q/L^2, & x_q \le 1, \\ (2-x_q)^3 - (2-x_q)/L^2, & 1 < x_q \le 2 - 2/L. \end{cases} \tag{24.4}$$

Keeping only terms of the order $1/L$ in Eq. (24.4) we approximate $\tilde{P}_{1/L}(x_q)$ by $g(x_q)dx_q$ with $dx_q = 1/L$. In the limit of $L \to \infty$, g becomes the probability density function of Manhattan distance inside the unit square

$$g(x) = \begin{cases} 4x(1-x) + \frac{2}{3}x^3, & x \le 1, \\ \frac{2}{3}(2-x)^3, & 1 < x \le 2. \end{cases} \tag{24.5}$$

Function $g(x)$ Eq. (24.5) is presented in Fig. 24.4.

24.3.2 Deformed Square Lattice

Here we analyze the case of the substrate with square symmetry of the underlying receptor's group, whose shape has been changed, for example, under the influence of shear stress as illustrated in Fig. 24.3.

Assume that we start from the unperturbed square-shape substrate and then we smoothly increase the stress. Under a sufficiently strong deformation the path between at least one pair of nodes appears. The distance along this path will be shorter than that in the original substrate. Thus, shear deformations shift the mean value of the distribution of distances toward the smaller value in comparison with nonstressed substrates, as it is seen in Fig. 24.4. Below we present formulas obtained for heavily distorted substrate, i.e. under shear stress generating an effective triangular symmetry among the node's positions. The number of distances Eq. (24.2) takes the following form

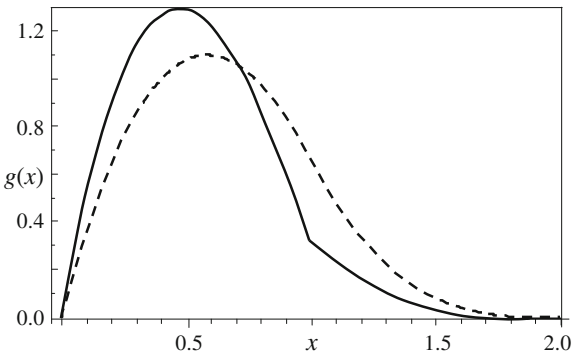


Fig. 24.4 Probability density of distances Eq. (24.5) within the square-shaped substrate (*dashed curve*) compared to that one Eq. (24.7) for the same set of receptors but within the deformed substrate (*solid curve*)

$$N(L,q)=\begin{cases} q(L-q)(3L-q)+\dfrac{1}{6}\left(q^2-1\right)q, & q\le L,\\ \dfrac{1}{6}\left[(2L-q)^2-1\right](2L-q), & L<q\le 2(L-1). \end{cases} \tag{24.6}$$

The corresponding probability density function of distance Eq. (24.5) now is given by

$$g(x)=\begin{cases} 2x(1-x)(3-x)+\dfrac{1}{6}x^3, & x\le 1,\\ \dfrac{1}{6}(2-x)^3, & 1<x\le 2. \end{cases} \tag{24.7}$$

The probability density functions Eqs. (24.5) and (24.7) are shown in Fig. 24.4.

24.3.3 Rectangular Lattice

The statistical characteristics (Eqs. 24.2–24.5) of the square-shaped substrates arise from the more general expressions corresponding to the rectangular-shaped substrates. These expressions are reported here to exhaust all possible symmetric arrangements of receptors with the underlying square symmetry. Consider a rectangular array of receptors with L_X and $L_Y \neq L_X$ receptors in rows and columns, respectively. Let $l = \mathrm{Min}\ (L_X, L_Y)$ and $L = \mathrm{Max}(L_X, L_Y)$. Then, the number $N(L_X, L_Y, q)$ of distances q is equal to

$$N(L_X,L_Y,q)=$$
$$\begin{cases} 2L_X\cdot L_Y q-(L_X+L_Y)q^2+\frac{1}{3}(q-1)q(q+1), & q\le l,\\ \left[l\cdot L+\frac{1}{3}\left(l^2-1\right)\right]l-l^2q, & l<q<L,\\ \frac{1}{3}\left[(L_X+L_Y-q)^2-1\right](L_X+L_Y-q), & L\le q\le L_X+L_Y-2. \end{cases} \tag{24.8}$$

With the help of the normalization condition the probability distribution function reads

$$P(L_X, L_Y, q) = \frac{2}{L_X \cdot L_Y (L_X \cdot L_Y - 1)} N(L_X, L_Y, q). \tag{24.9}$$

The Eqs. (24.8, 24.9) reduce to the forms given by Eqs. (24.2, 24.3) only for $L_X \equiv L_Y$, i.e. for the square. It is interesting to note that for $l < q < L$ the function $N(L_X, L_Y, q)$ (Eq. 24.8) is linear with respect to q. Such linear dependence of the distance distribution was previously observed in granular compaction experiments carried out on single-layers of compounds created by welding ball bearings together [17]. The angular shape of the compounds prevents them from the rotation and thus, near the jamming, they change positions in such a way that their trajectories resemble zigzag lines.

24.3.4 Triangular Lattice

Discussed in Sect. 24.3.2, a highly deformed square-shaped substrate possesses triangular lattice of receptor locations. Here and in the following subsections we analyze the triangle-shaped substrates. Such shape is rather artificial. Nevertheless, it is worth analyzing it because the triangular, the hexagonal and the Kagomé lattices are built around the same set of nodes, see Fig. 24.5, and this will enable us to directly compare the results. Our approach applied to the graph represented in Fig. 24.5a yields the distance distribution in the form

$$N(L, q) = \frac{3}{2} q(L - q)(L - q + 1), \quad q = 1, 2 \ldots, L - 1\,. \tag{24.10}$$

Thus, the corresponding probability distribution of distances $P(L, q)$ and probability density $g(x)$ are as follows

$$P(L, q) = \frac{12}{(L^2 - 1)L(L + 2)} N(L, q), \tag{24.11}$$

$$g(x) = 18x(1 - x)^2. \tag{24.12}$$

The lattice size $L \geq 2$ is shown in Fig. 24.5c.

24.3.5 Hexagonal Lattice

The hexagonal lattice, see Fig. 24.5b, viewed as a graph, possess fewer nodes and edges than the graph of the underlying triangular lattice presented in Fig. 24.5a. Thus, within the same support, all functions representing the hexagonal symmetry

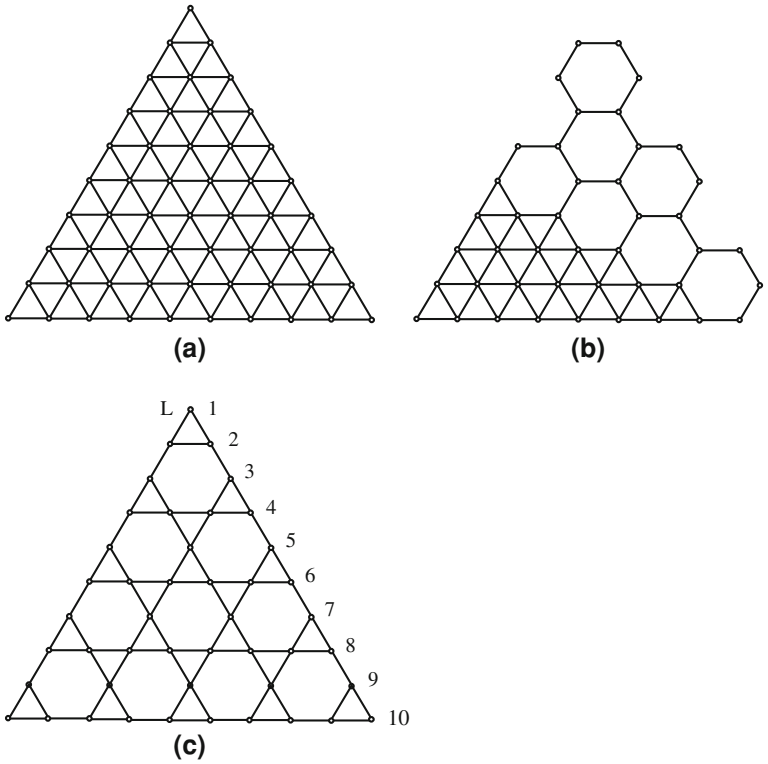


Fig. 24.5 Finite triangular lattice (**a**) viewed as an undirected graph. Subgraphs of this graph represent the hexagonal lattice (**b**) and the Kagomé lattice (**c**). In part (**b**) it is seen how the hexagonal lattice emerges from the triangular one. The lattice size L is shown in (**c**)

take smaller values than these related to the triangular symmetry. The functions appropriate to the hexagonal symmetry read

$$N(L,q) = \frac{3}{8}Lq(L-2q+4) - 3 + \frac{3}{8}q\left[q + \frac{1-(-1)^q}{2}\right](q-4), \quad q = 1,\ldots,L-1, \tag{24.13}$$

$$P(L,q) = \frac{32}{(L-2)(L+6)(L^2+4L-8)}N(L,q), \tag{24.14}$$

The lattice size $L \geq 4$ is shown in Fig. 24.5c.

24.3.6 Kagomé Lattice

Arguments, similar to these stated in the case of the hexagonal symmetry, make the relevant functions defined as follows

$$N(L,q) = \begin{cases} \frac{3}{4}L^2, & q = 1, \\ \frac{3}{16}(5q-2)(L-q)(L-q+2), & q = 2,\ldots,L-2, \\ \frac{3}{8}(L+1-q)[(2q+1)L+3-q(2q-1)], & q = 3,\ldots,L-1. \end{cases} \tag{24.16}$$

and

$$P(L,q) = \frac{128}{3L(L+2)(3L^2+6L-8)} N(L,q). \tag{24.17}$$

The lattice size $L \geq 2$ is shown in Fig. 24.5c.

24.4 Conclusion

As micropatterned substrates play an increasing role in the understanding of basic cell biology, there is an increasing need to understand the interplay between substrate geometry and surface chemistry. In this paper we have studied a minimalist model of an earlier stage of a polymer adhesion to flat ordered substrate. Our results could serve as an initial class of probability density functions of the end-to-end distance.

They can also be applied to other problems concerning the movement of the system components restricted to zigzags. Such movements occur, e.g. in granulates near the jamming transition. Grains close to jamming become stuck and only few of them change their positions in such a way that their trajectories resemble zigzag lines. Thus, we believe that the Manhattan-like distance measure is the metric appropriate to highly compacted granulates. Some of the above mentioned results, concerning the square and the triangular symmetries, have been obtained elsewhere [18, 19]. It is worth to mention that apart from the four tessellations discussed here other two dimensional tessellations have already been analyzed in the context of statistical properties of micro-flows through flat, microfluidic devices [20].

References

1. See, e.g. Intelligent Substrate Inc., Micropatterned substrates: Highlights from the literature. Available: www.intelligentsubstrates.com/Applications/Applications.html
2. Crosby AJ, Hageman M, Duncan A (2005) Controlling polymer adhesion with 'Pancakes'. Langmuir 21:11738–11743
3. Lee I, Zheng H, Rubner MF, Hammond PT (2002) Controlled cluster size in patterned particle arrays via directed adsorption on confined surfaces. Adv Mater 14(8):572–577
4. Zheng H, Lee I, Rubner MF, Hammond PT (2002) Two component particle arrays on patterned polyelectrolyte multilayer template. Adv Mater 14(8):569–572
5. Azioune A, Capri N, Tseng Q, Théry M, Piel M (2010) Protein micropatterns: a direct protocol using deep UVs. Methods Cell Biol 97:133–146
6. Otsuka H (2010) Nanofabrication of nonfouling surfaces for micropatterning of cell and microtissue. Molecules 15:5525–5546. Available: www.mdpi.com/1420-3049/15/8/5525/pdf
7. Tikhomirov G, Hoogland S, Lee PE, Fisher A, Sargent EH, Kelly SO (2011) DNA-based programming of quantum dot valency, self-assembly and luminescence. Nat Nanotechnol 6:485–490
8. Lee I, Wool RP (2000) Polymer adhesion vs. substrate receptor group. Macromolecules 33:2680–2687
9. Rana N, Kossow C, Eisenbraun ET, Greer RE, Koloyeros AE (2011) Controlling interfacial adhesion of self-assembled polypeptide fibrils for novel nanoelectromechanical systems (NEMS) applications. Micromachines 2:1–16. Available: www.mdpi.com/2072-666X/2/1/1/pdf
10. Raghavan S, Chen ChS (2004) Micropatterned Environments in Cell Biology. Adv Mater 16(15):1303–1313
11. Amin R, Hwang S, Park SH (2011) Nanobiotechnology: an interaction between nanontechnology and biotechnology. NANO 6(2):101–111
12. Pannier AK, Anderson BC, Shea LD (2005) Substrate-mediated delivery from self-assembled monolayers: effect of surface ionization, hydrophilicity, and patterning. Acta Biomater 1(5):511–522
13. Grünbaum B, Shepard G (1986) Tilings and patterns. W. H. Freeman, New York
14. Janse van Rensburg EJ (2003) Statistical mechanics of directed models of polymer in the square lattice. J Phys A: Math Gen 36(15):R11–R61
15. Bender CM, Bender MA, Demaine ED, Fekete SP (2004) What is the optimal shape of a city? J Phys A: Math Gen 37(1):147–159
16. Leung VJ et al (2002) Processor allocation on Cplant: achieving general processor locality using one-dimensional allocation strategies. In: Proceedings of the 4th IEEE international conference on cluster computing. Willey-Computer Society Press, Chicago, pp 296–304
17. Rakenburg IC, Zieve RJ (2001) Influence of shape on ordering of granular systems in two dimensions. Phys Rev E 63(6):61303
18. Domański Z (2011) Geometry-induced transport properties of two dimensional networks. In: Schmidt M (ed) Advances in Computer Science and Engineering. InTech, Rijeka, 2011, pp 337–352. Available: www.intechweb.org/books
19. Domański Z, Sczygiol N (2011) Distribution of the End-to-End Distance of Polymers Trapped onto Ordered Substrates. Lecture Notes in Engineering and Computer Science 2011, WCECS 2011, 19–21 October, 2011, San Francisco, pp 604–608
20. Domański Z (2011) Efficiency of two-dimensional microfilters. AIP Conf Proc 1373:211–220

Chapter 25
Quantification of Athlete's Heart Condition: A Detrended Fluctuation Analysis

Toru Yazawa, Yukio Shimoda and Albert M. Hutapea

Abstract Detrended fluctuation analysis (DFA) technology was used to check the heartbeats of athletes in the Indonesian olympic training center. We report results of time series analysis of heartbeat on subjects who underwent ergometer exercise. The objective of this research was to determine whether DFA could function as a useful computation method for the evaluation of the subject's quality of cardiovascular system. Since there are no 2 individuals that are identical physiologically, we present case studies but novel findings regarding how wellness of subjects can be evaluated by the electrocardiography. Even from the case studies, we can propose a general conclusion that DFA is a new, useful numerical method for quantifying the degree of wellness through the heartbeat recording.

Keywords Athlete · Ergometer exercise · Detrended fluctuation analysis · Heart · Heartbeat interval · Scaling exponent

T. Yazawa (✉)
Biophysical Cardiology Research, Neurobiology, Department of Biological Science, Tokyo Metropolitan University, Hachioji, 192-0397 Tokyo, Japan
e-mail: yazawa-tohru@tmu.ac.jp

Y. Shimoda
Medical Research Institute, Tokyo Women's Medical University, Shinjuku, 162-0054 Tokyo, Japan
e-mail: yshimoda@lab.twmu.ac.jp

A. M. Hutapea
Faculty of Science, Universitas Advent Indonesia, Jl. Kolonel Masturi 288, Parongpong, Bandung 40559, Indonesia
e-mail: amhutapea@unai.edu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_25,

25.1 Introduction

Cardiovascular disease is one of major social health problems. While the default setting is in general good health, there is always an onset of a process of alteration to an anatomical remodeling or a disease that never returns to the default health. Particularly, in the heart, this onset may result in "silent" angina, and a "silent" heart attack as the worst-case scenario. Early detection of the onset is a crucial technological solution to combat this pathogenesis, and making prediction of the "onset" is one of the main goals of science.

Cardiac remodeling is typically associated with disease but also occurs in the athlete's heart as an adaptive physiological response. The aim of our work was to detect the onset of a process of a cardiac remodeling, i.e., compensatory gradual changes in structure and function which will be leading the heart to pathogenic condition or adaptation [1]. Bio-medical computation could be very suitable method in investigating how to detect the onset of a cardiac remodeling. Therefore, our ultimate aim is to quantitatively analyze the heart condition, thus making numerical prediction.

Traditionally, cardiac studies employ heart rate variability (HRV) to detect the onset of cardiac problems, including the disorders of the autonomic nervous system. Problems arise, however, when patients are usually assumed to be healthy before the appearance of symptoms associated with HRV. The detrended fluctuation analysis (DFA) [2] was proposed as a potentially useful method in determining a sign of cardiovascular disease [3], however, it has not yet developed to be a practical medical tool such as the electrocardiogram (EKG). (We prefer the word "EKG," instead of "ECG," with due respect to the inventor, Dutch physiologist, Nobel laureate, Willem Einthoven.)

We recently tested practical usefulness of DFA with using the heart of crustacean-animal models. In the test, we succeeded to show that DFA can distinguish the beating of intact hearts from isolated hearts [4]. In that study, we found that the scaling exponent of the isolated hearts shifted and approached to 0.5. In turn, the scaling exponent of the intact hearts showed a value about 1.0. As the results, we realized that DFA was reliable and useful, because DFA probably accurately reflects cardiac and systemic physiology. Unlike HRV, the excelling point for DFA is that it has a baseline value of one (1), like a standard body temperature (37 °C), a standard blood pH (7.4), and so on. Therefore, DFA was simple as a tool, we hypothesized.

One (1) is nonlinearly determined as "healthy" outcome resulted from complex interactions between structure and function of molecules, cells and organs. Thereby we have a hope that DFA can "numerically" determine the state of health through the quality of the functioning of the heart. DFA seemed to be reflecting the state of not only the heart itself but also its physiological interaction with the nervous system. We considered that DFA might be a tool to detect the onset of cardiac remodeling, including functional changes of the autonomic nervous system.

In this chapter, we show case studies of heartbeat analysis of subjects who performed exercise on ergometer stationary bicycle to provide empirical evidence for the practical usefulness of DFA and a new EKG amplification device that facilitates automatic DFA computation in practical use. We will show that DFA is a potentially helpful engineering tool for the early identification of physiological disorders, as it reveals information that is not provided by an EKG.

25.2 Materials and Method

25.2.1 Peak Detection

Interval analysis requires detection of the precise timing of the heartbeat. A consecutive and perfect detection without missing any beat is necessary. According to our preliminary tests, about 2,000 consecutive heartbeats were required for obtaining a reliable computation of scaling exponent. We hypothesized that longer recording of the heartbeats would results in a better diagnosis. However, we found that long recording was not justifiably useful and a recording of about 2,000 consecutive heartbeats is preferable.

To detect the timing of the heartbeats, one may assume that common EKG recording is sufficiently useful. However, the problem with conventional EKG was the drifting of the baseline of the recording. Due to the drift and the contamination of unexpected electric power-line noise, recording failures may happen.

Another obstacle arose from the premature ventricular contraction (PVC). Among the "normal" subjects (age over 40 years old), about 60 % of subjects have PVC arrhythmic heartbeats. Normally, this PVC is believed to be benign arrhythmia, and in fact during our recording, we found many healthy-looking individuals exhibited this arrhythmia. However, PVC is an obstacle to a perfect detection of the timing of the heartbeat, because the height of its signal could sometimes vary much. If the baseline of EKG recording could be extremely stable, the heartbeats would automatically be detectable even when irregular beats appeared sporadically. Unfortunately, in commercial EKG recording devices, baseline of the record is not stable.

25.2.2 Stable Baseline

To capture heart beat peaks without missing any detection, we made an EKG amplifier that stabilizes baseline of the recording. Important issue was: we discovered that time-constant for input-stage of recording must be adjusted to an appropriate level.

Having stable baseline recording was an advantage to our DFA research. However, in some cases, inevitable noises would contaminate the recording. In such case, we removed the noises by identifying them visually on PC screen thus making a perfect (without miscounting) heartbeat interval time series. We have already identified how this inconvenience came about. About one-half of these cases were due to the sweat on the skin under the electrodes. We were able to overcome this problem by cleaning the skin with an appropriate solution.

25.2.3 DFA: Background

DFA is based on the concepts of "scaling" and "self-similarity" [5]. It can identify "critical" phenomena because systems near critical points exhibit self-similar fluctuations [2, 5, 6], which means that recorded signals and their magnified/contracted copies are statistically similar. Self-similarity is defined as follows: In general, statistical quantities, such as "average" and "variance," of fluctuating signals can be calculated by taking the average of the signal through a certain section; however, the average is not necessarily a simple average. In this study, we took a squared average of the data. The statistical quantity calculation depended on the section size. The signal was self-similar when the statistical quantity was λ^{α} times for a section size magnified by λ. Here, α is the "scaling exponent" and characterizes self-similarity.

Stanley and colleagues considered that scaling property can be detected in biological systems because most of these systems are strongly nonlinear and resemble the systems in nature that exhibit critical phenomena. They applied DFA to DNA arrangement and EKG data in the late 1980s and early 1990s and discovered the usefulness of the scaling property [2, 6], and emphasized the potential utility of DFA in the life sciences [6]. Although DFA technology has not progressed to a great extent, nonlinear technology is now widely accepted, and rapid advances are being made in this technology.

25.2.4 DFA: Technique

We made our own computation program based on the previous publication [2], which is described in one of the references [7].

25.2.5 EKG Recording

For EKG recordings, we used a power lab system (AD Instruments, Australia). For EKG electrodes, a set of ready-made 3 AgAgCl electrodes (+, −, and ground;

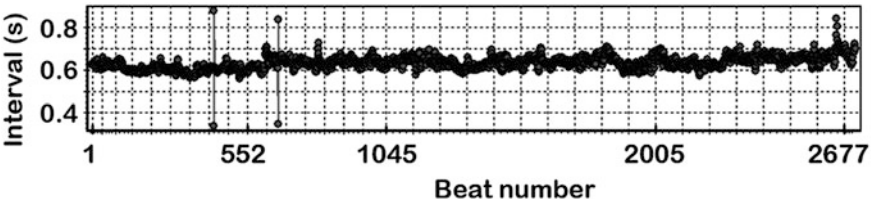


Fig. 25.1 Heartbeat recording from *subject one* at rest. Figure shows 2719 beats in total during 28 min, which were 100 % accurately detected, as shown in next Fig. 25.2

Nihonkoden Co. Ltd. disposable model Vitrode V) were used. Wires from EKG electrodes were connected to our newly made amplifier. These EKG signals were then connected to a power lab system.

25.2.6 Volunteers and Ethics

EKG recordings were performed at Indonesian national olympic training center in Bandung. Subjects were selected at random at the Indonesian olympic training center. All subjects were treated as per the ethical control regulations of our universities, Tokyo Metropolitan University, Tokyo Women's Medical University and Universitas Advent Indonesia.

25.3 Results and Discussion

25.3.1 Case Study 1 (Subject One)

Subject one is a 49-year old healthy looking Indonesian. His resting state heartbeats were recorded for 28 min while the subject was sitting on ergometer, relaxing and answering to several questions made by researchers. Two pre-matured ventricular contractions (so called PVC) can be seen (Fig. 25.1). These PVCs (see also PVC in Fig. 25.3) are a benign type of heartbeat, as classified by guidelines used by medical doctors. It is described that PVCs are observable among 60 % of healthy persons over the age of 40. However, we should acknowledge that a hypothesis associates the occurrence of PVC with sudden death [8].

At the beginning of EKG recording, his heart rate was about 90 beat per min (BPM) (Fig. 25.1, interval 0.6 s, see also Fig. 25.4). In 3 min, the heart rate was approaching over 100 BPM after the start of the recording (Fig. 25.4). He seemed to be nervous because the measurement equipment as well as the foreigner researcher doing the recording were new to him. During the period for 28 min of

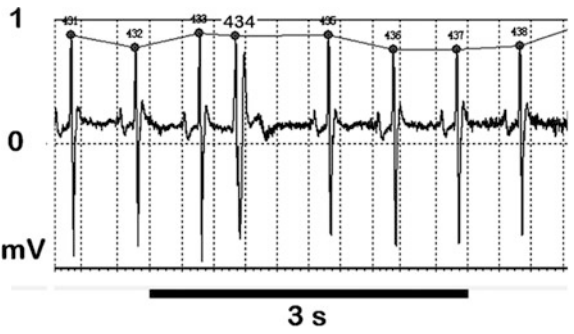


Fig. 25.2 Example recording of heartbeats and accurate peak detections. The heartbeat number 434 is PVC. The figure shows that our amplifier produced a steady baseline EKG even though subject was freely moving. (Here, subject was answering to question)

Table 25.1 Comparison between at rest and during exercise

	DFA (α)	
Box size	Rest	Exercise
51–100	1.01	1.32
30–140	0.99	1.25

resting stage measurement, he got relaxed by conversations with us, and his heart rate gradually decreased (Fig. 25.1, interval about 0.7 s, see also Fig. 25.4).

At the start of the exercise stage his heart rate immediately increased to over 100 BPM, and kept slow increment until reaching a plateau before the end of the exercise (Figs. 25.3 and 25.4).

After 20 min of the exercise he mentioned to us that he was not tired yet. He also mentioned at the end of exercise session that he felt he could continue with this low load (50 watt) aerobic exercise for another 30 min. One PVC appeared at near the end of exercise (Figs. 25.3 and 25.4) as was seen at rest condition (Fig. 25.1).

After obtaining heartbeat-interval time series, we proceeded to next step, calculating the scaling exponents (α) by our computing method the DFA (Table 25.1). One can see that exercise increased α, which can be seen in overall slops (see Fig. 25.5). It means that the scaling exponents were pushed up by exercise.

At rest, this 49 years old man exhibited healthy scaling exponents nearly 1.0 (see the windows, both 51–100 and 30–140 in Table 25.1 and corresponding graphs of Fig. 25.5). From our previous studies we determined a guideline in interpreting the scaling exponents to define whether an individual is healthy or unhealthy (Table 25.2). We were surprised to discover that an apparently healthy person (athlete's heart) did get risky high value of the scaling exponent "during exercise" (Table 25.1). Interestingly, we also discovered that the same exercise induced an increase in the scaling exponents of the other subject, a swimmer girl (see below).

From the results from *subject one*, we started to consider that "athlete's heart," which is a remodeling heart, may not be so healthy than we first believed. In the class room, general physiology teaches that athlete's hearts undergoes

Fig. 25.3 The same subject shown in Figs. 25.1 and 25.2. Continuous recording from Fig. 25.1. He started ergometer exercise, at the heartbeat number 1 (one) with a 50 watt load strength and a 96 rpm speed, lasting for 32 min. 1–3746 beats shown

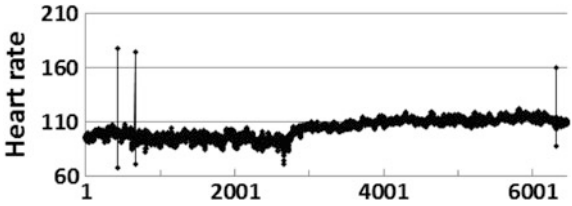


Fig. 25.4 Entire time series of *subject one*. Connected data of Fig. 25.1 and Fig. 25.3. Y axis, heart rate (BPM). X axis, beat number

Table 25.2 Our temporary guideline (see [7])

DFA (α)	Physiological interpretation
0.5 ~ 0.899	Stressfulness, PVC, Alternans, Naturally dying
0.9 ~ 1.199	Healthy, 1/f fluctuation
1.2 ~ 1.5	This heart is at risk of catastrophic circulation stoppage

physiological cardiac hypertrophy instead of pathological hypertrophy, e.g., caused by hypertension. However, a literature pointed out difficulties associated with distinguishing the athlete's heart from hypertrophic cardiomyopathy (HCM). HCM is the leading cause of sudden cardiac death in young athletes [1].

The EKGs of the other 3 subjects shown in this study were recorded at rest and while engaging in exercise in the same room at the same environment, temperature of 25 °C, together with the *subject one* aforementioned (Figs. 25.1, 25.2, 25.3, 25.4, 25.5).

25.3.2 Case Study 2 (Subject two)

Subject two, age 28, an Indonesian female swimmer. Her average-heart-rate at resting state is rather high with unknown reason (Figs. 25.6 and 25.7). According to the doctor's guideline: the heart rate above 100 bpm is referred to as tachycardia. However, we did not investigate the reason for this seemingly abnormal condition. (We assume Indonesian tropical room-temperature, 25 °C, is the reason.)

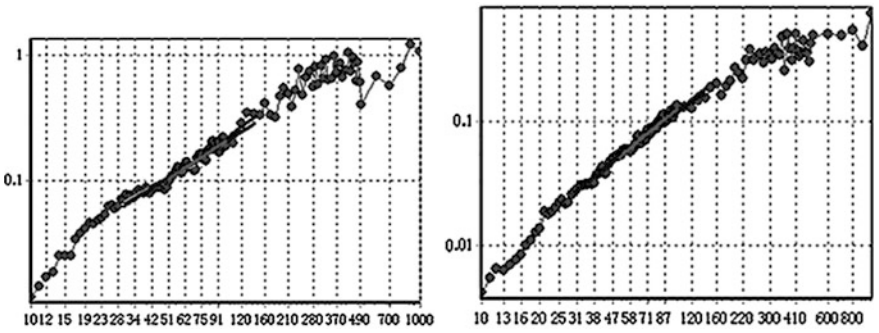


Fig. 25.5 DFA computation. *Left* graph, at rest. *Right* graph, at exercise. For the same subject shown in Figs. 25.1, 25.2, 25.3, and 25.4. Y axis, variance. X axis, box size, from 10 to 1000. Log scale in both axis (see original article [2] for the basics of DFA). The slope of the graph gives the scaling exponent, which is calculated at various "windows," i.e., "box size," as shown in Table 25.1

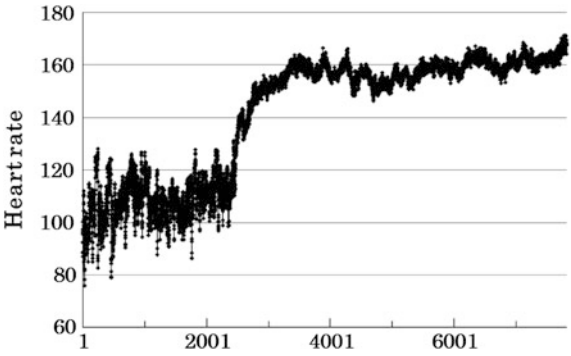


Fig. 25.6 Time series, 2471 beats for 28 min resting state, and 5341 beats for 32 min exercise state. Y axis, heart rate. X axis, beat number. Ergometer exercise session started at the heartbeat number 2472 with a 50 watt load strength and a 96 rpm speed, lasting for 32 min, identical strength in the case study 1

After starting ergometer exercise, heart rate quickly increased (see *A in Fig. 25.7). Heart rate soon attained a steady state, about 160 BPM. Apparently a high rate was maintained over the period of exercise. This ability indicates that she has athlete's heart, i.e., her heart seems to have adapted to the physiological demand of a long distance swimming.

Her heart rate quickly decreased at the end of exercise (see *B in Fig. 25.7). A significant characteristic, "maintained" rate during exercise, were seen in both *subject one* (Fig. 25.1) and *two* (Fig. 25.6).Therefore, we can conclude that both *subjects one* and *two*, Case studies 1 and 2, show characteristics of "athlete's heart."

From Fig. 25.7, one can clearly see an exponential rise (*A) and decay (*B). This exponential behavior is due to the changes of the cardio-inhibitory nerve activity, i.e., parasympathetic nerve activity, was switched-off and switched-on (respectively, at *A and *B in Fig. 25.7). We have already described a mathematical model for this exponential function of neurotransmitter release regarding to cardio-inhibitory nerve control [9].

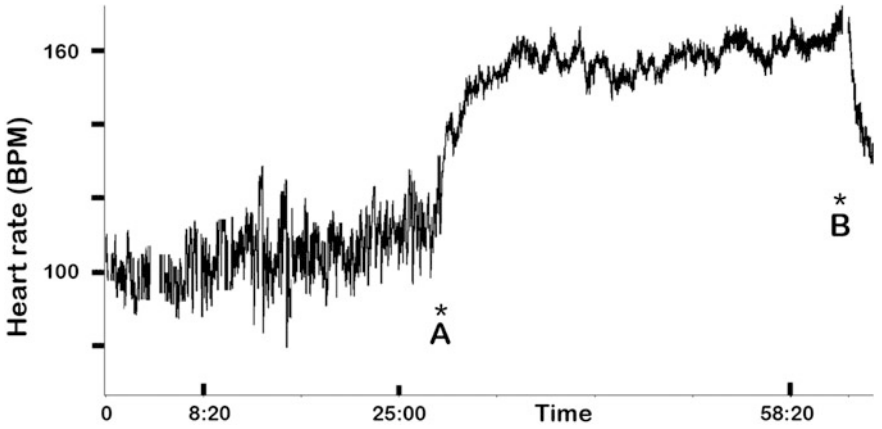


Fig. 25.7 Exponential rise and decay in heart rate. The same record as in Fig. 25.6 but X axis is shown in time (min) instead of beat number (see Fig. 25.6). *A, exercise started. *B, exercise stopped. Y axis, heart rate (beat per min)

Table 25.3 Comparison between at rest and during exercise (A female swimmer)

	DFA (α)	
Box size	Rest	Exercise
51–100	0.62	1.36
30–140	0.8	1.42

Figures 25.6 and 25.7 indicate that fluctuation of the heartbeat interval, i.e., variability in rate, is greater at rest than during exercise. This must be due to a change of vagal tone governing the heart: We can interpret that during exercise the heart received a decreased discharge frequency in the inhibitory autonomic nerve fibers innervating the heart. In other words, acceleration-dominant-state was induced by exercise. From neurophysiological consideration, this acceleration-dominant-state can be explained by reduction of inhibitory influence that caused the acceleration (*A in Fig. 25.7), and thereafter getting dis-inhibition (*B in Fig. 25.7). This consideration was experimentally proven: we have already demonstrated real EKG data and mathematical model in crustacean heart [9].

The scaling exponents of the subject two are very low at rest (see Table 25.3) as can be seen in the left graph where slope is less steep (Fig. 25.8) and the scaling exponents ranges around 0.6–0.8 at rest (Table 25.3). We already know that if one has perfect health condition, the slope must exhibit 45 degree, i.e., the scaling exponent is one (1) at relaxed condition. This means that her general health condition and especially heart condition is not perfect, although DFA cannot tell the physiological reason(s) that contributes to this (see our temporary guideline Table 25.2).

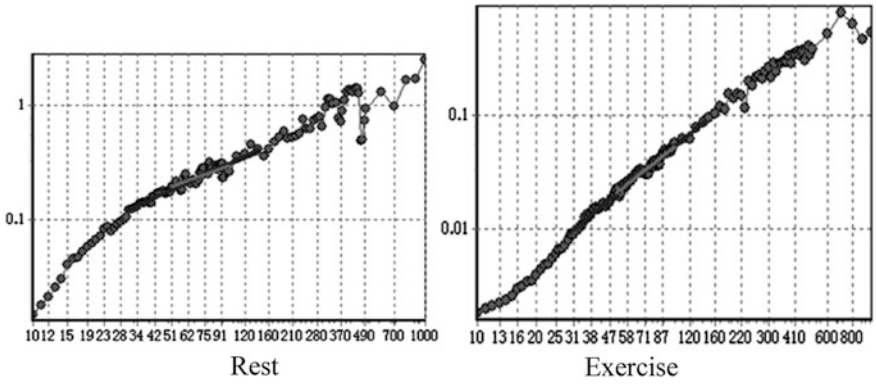


Fig. 25.8 DFA computation. *Left* graph shows results at rest and *right* graph at exercise. The same subject shown in Figs. 25.6 and 25.7. Y axis, variance. X axis, the box size from 10 to 1000. Log scale in both axis

When she was engaged in exercise, the DFA-slope became steeper in almost entire ranges of window size (Fig. 25.8) and thus computed scaling exponents were significantly increased. The values are astonishingly high, ranging from 1.3 to 1.4 (see Table 25.3). This significant increase of the scaling exponents during exercise was also found in other athlete's heart, *subject one* (see Case study 1). According to our guideline (Table 25.2), we may interpret that this high values during exercise may indicate that their hearts are at a risky state. If this consideration would be proven in the future investigations, we must conclude that athlete's heart that is remodeled heart is not normal. It is intriguing that we have already found that a high scaling exponent is associated with a subject who has ischemic heart, such as those received stent placement and/or bypass implantation [10].

There are a lot of elements in the body such as molecules, cells, tissues and organs. Nonlinear interactions between the elements of athlete's body must be contributing to this scaling behavior of the heartbeat.

25.3.3 Case Study 3 (Subject Three)

Subject three, age 29, an Indonesian male basketball player (Fig. 25.9 and Table 25.4). It should be noted that here the load was 75 watt instead of 50 watt in the case of other 3 subjects, which are case studies 1, 2, and 4.

It took long time before reaching a plateau phase due to higher load (Fig. 25.9). Figure 25.9 shows that plateau started at about 5,500 in heartbeat number and lasted until the end of exercise. He (*subject three*) mentioned to us that, at the end of exercise, he was tired and he wished very much to stop the exercise session.

Here, one can see again that fluctuation during exercise is smaller than that at rest. As is in the *subject two* (Table 25.3), again the scaling exponents during

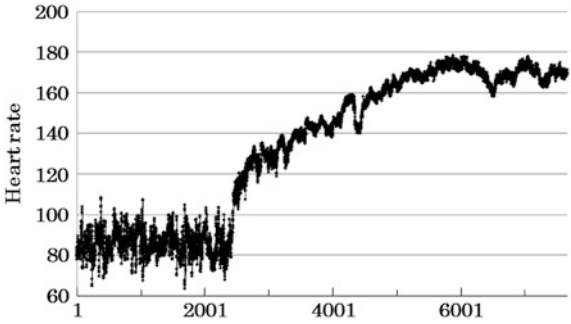


Fig. 25.9 Time series, of *subject three*, 2438 beats for 28 min resting state, and 5224 beats for 32 min exercise state. Y axis, heart rate. X axis, beat number. Ergometer exercise was started at the heartbeat number 2439 with a 75 watt load and a 96 rpm speed, lasting for 32 min

Table 25.4 Comparison between at rest and during exercise (a basketball player)

	DFA (α)	
Box size	Rest	Exercise
51–100	0.76	1.62
30–140	0.81	1.52

exercise were very high (Table 25.4). His scaling exponent at rest is low (Table 25.4) which characteristics are similar to the recording from the *subject two* (Table 25.3). We conclude that his (*subject three*) heart system at rest is not perfect in terms of DFA analysis.

25.3.4 Case Study 4 (Subject Four)

Subject four, age 27, male, an Indonesian futsal player (Spanish futbol de salon). In this subject, the scaling exponents were increased by exercise as those shown by the rest of the 3 cases in this study. However, his heart stayed in a healthy range of the scaling exponent during exercise (see Table 25.5). He mentioned that, at the end of exercise, he was not tired at all with 50 watt and 96 rpm. We may conclude that he is the most appropriate sports-aspiring person among 4 subjects in this case study, because he has no risky value, even in terms of DFA. During exercise, his autonomic nervous system was still capable to send inhibitory command to the heart, which is observable as sporadically occurring "slowing down" in heart rate (see enlargement of time series, Fig. 25.10). This "slowing down" in heart rate is derived from inhibitory discharge in the autonomic nerves, i.e., the parasympathetic nerve or the vagus nerve. Figure 25.10 shows that his vagus nerve still regulates the heart properly. He might be able to have much high load exercise though we have not yet tested. His exercise period was not "up to the chin" condition. That is why his scaling exponent shows nearly one (1), which means his heartbeat can behave dynamically, that is, responding dynamically to internal demands. The ability to meet demands is a good condition of the heart that can respond properly and

Table 25.5 Comparison between at rest and during exercise

	DFA (α)	
Box size	Rest	Exercise
51–100	0.85	1.15
30–140	0.89	1.12

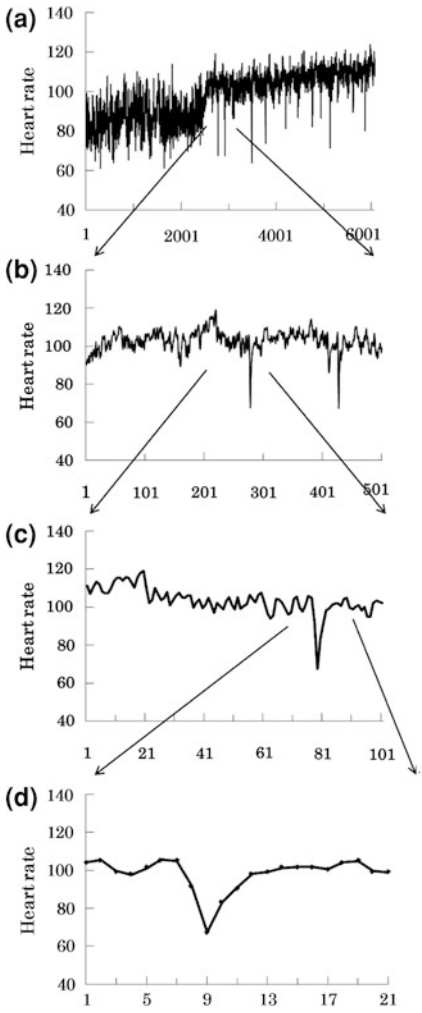


Fig. 25.10 Heartbeat-interval time series of *subject four*. A, 6072 beats in total. B, C, D, partially enlarged to show detail of the time series. B, Beat number 2500–3000. C, 2700–2800. D, 2770–279

dynamically to the internal and external environment. However, his heart condition at rest shows that he might have a stress in terms of DFA analysis. This fact is similar to that of both *subjects two* and *three* in this study.

Athlete who has a healthy scaling exponent at rest in the present report, is only *subject one* (Case study 1). The subject who has a healthy scaling exponent at exercise, is only *subject four* (Case study 4).

25.4 Concluding Remarks

The techniques and experimental results in the present study are new as far as we know. In the present observations with DFA computation we tried to find any apparent correlations between the scaling exponent and the state of heart during exercise. While data of the heartbeat time series were obtained from subjects who were healthy looking individuals, 3 of 4 subjects (Case studies 1, 2, and 3) exhibited surprising results: exercise brings them to a risky state in terms of the scaling exponents. We would like to suggest that their state of heartbeats during exercise is the state that the heart is ready to stop any time, as demonstrated before in animal model experiments and human ischemic cardiac conditions [10].

Athlete's heart is believed to be a benign adaptation [1]. Four subjects in this study have obviously different genomic structure from each other. However, outcome of control system commanding the heart performance was all identical: exercise increases the scaling exponents, that is, to a normal level in one subject and to an alarming level in the rest of 3 subjects in this study. This is probably normal function of healthy subjects who have complex internal nonlinear physiological interactive pathways. However, it is important to know that there has been some debate over whether the athlete's heart is a truly physiological phenomenon or whether long-term, chronic exercise training is maladaptive and leads to heart disease or sudden cardiac death [11]. This investigation may cause a stir in the debate.

Preliminary work has been appeared in a proceeding [12].

Acknowledgments This work was supported by an international collaboration agreement (code number: I-240211) made in 2011 between the Adventist University of Indonesia (AMH), Tokyo Women's. Medical University (YS), and Tokyo Metropolitan University (TY). This work was also supported in part by Grant-in-Aid for scientific research (C) No. 23500524 (TY).

References

1. Weeks KL, McMullen JR (2010) The athlete's heart vs. the failing heart: can signaling explain the two distinct outcome? Physiology 26:97–105
2. Peng C-K, Havlin S, Stanley HE, Goldberger AL (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5:82–87
3. Stanley HE, Amaral LAN, Goldberger AL, Havlin S, Ivanov PCh, Peng C-K (1999) Statistical physics and physiology: monofractal and multifractal approaches. Phys A 270:309–324

4. Yazawa T, Kiyono K, Tanaka K, Katsuyama T (2004) Neurodynamical control systems of the heart of Japanese spiny lobster, Panulirusjaponicus. IzvestiyaVUZ. Appl Nonlinear Dyn 12(1–2):114–121
5. Stanley HE (1995) Phase transitions. Power laws and universality. Nature 37(8):554
6. Goldberger AL, Amaral LAN, Hausdorff JM, Ivanov PC, Peng C-K, Stanley HE (2002) Fractal dynamics in physiology: alterations with disease and aging. PNAS 99(suppl 1):2466–2472
7. Yazawa T, Shimoda Y, Hutapea AM (2010) Evaluation of sleep by detrended fluctuation analysis of the heartbeat, IAENG Transactions on Engineering Technologies, 6, in AIP conference proceedings series, the Congress WCECS AIP press, USA, pp 199–210
8. Lown B, Wolf M (1971) Approaches to sudden death from coronary heart disease. Circulation 44:130–142
9. Yazawa T, Katsuyama T (2001) Spontaneous and repetitive cardiac slowdown in the freely moving spiny lobster, Panulirusjapnicus. J. Comp Physiol A 187:817–824
10. Yazawa T, Tanaka K, Kato A, Katsuyama T (2008) The scaling exponent calculated by the detrended fluctuation analysis, distinguishes the injured sick hearts against normal healthy hearts. IAING (WCECS 2008) vol II.pp 7–12. International conference on computational biology, San Francisco, 22–24 October
11. Drews O, Tsukamoto O, Liem D, Streicher J, Wang Y, Ping P (2010) Differential regulation of proteasome function in isoproterenol-induced cardiac hypertrophy. Circ Res 107:1094–1101
12. Yazawa T, Hutapea AM, Shimoda Y (2011) Quantification of athlete's heartbeats engaged in ergometer exercise: a detrended fluctuation analysis study checking the heart condition. Lecture notes in engineering and computer science: proceedings of The World Congress on Engineering and Computer Science 2011, WCECS 2011, 19–21 Oct 2011, San Francisco, pp 585–590

Chapter 26
Black Globe Temperature Estimate for the WBGT Index

Vincent E. Dimiceli, Steven F. Piltz and Steve A. Amburn

Abstract The wet bulb globe temperature (WBGT) index is used in industry, sports and other areas to indicate the heat stress level for humans and animals. One of the values needed to calculate the WBGT Index is the black globe temperature. The black globe temperature is measured using a Black Globe Temperature Sensor which includes a black globe with a thermometer inserted in the center. However, the Black Globe Temperature Sensor can be costly and many of these instruments may be needed to get measurements in many locations. The lead author has derived a formula to estimate the black globe temperature using readily available data collected by the National Weather Service (NWS). The formula was derived from a formula suggested by Kuehn, which was based on heat transfer theory. The resulting equation was a fourth degree polynomial in terms of the black globe temperature. It was determined that the black globe temperature can be very accurately approximated by taking a fourth degree polynomial in terms of the black globe temperature to create a linear approximation for black globe temperature. Some preliminary tests indicate accuracy within 0.5 °F.

Keywords Black globe · Heat index · Heat safety · Heat stress · Temperature measurement · Wet bulb globe temperature

V. E. Dimiceli (✉)
Oral Roberts University, 7777 S Lewis Ave, Tulsa, OK 74171, USA
e-mail: vdimiceli@oru.edu

S. F. Piltz · S. A. Amburn
National Oceanic and atmospheric Administration, National Weather Service Forecast Office, 10159 E 11th St. Ste 300, Tulsa, OK 74128, USA
e-mail: steven.piltz@noaa.gov

S. A. Amburn
e-mail: steve.amburn@noaa.gov

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_26,

26.1 Introduction

One of the government regulations instituted by the Occupational Safety and Health Administration (OSHA) is heat stress management [4]. The manual states in Section III: Chapter 4 the second paragraph of the introduction:

> " …Outdoor operations conducted in hot weather, such as construction, refining, asbestos removal, and hazardous waste site activities, especially those that require workers to wear semipermeable or impermeable protective clothing, are also likely to cause heat stress among exposed workers…".

A rating is calculated which indicates the safe amount of time a person can work outside on a hot day. This quantity is called the Wet Bulb Globe Temperature Index (WBGT). In the past, WBGT data has been collected manually using a portable instrument. The OSHA manual includes the following formulas for the WBGT [11]:

1. For indoor and outdoor conditions with no solar load, WBGT is calculated as:

$$\mathrm{WBGT} = 0.7\mathrm{NWB} + 0.3\mathrm{GT}$$

2. For outdoors with a solar load, WBGT is calculated as

$$\mathrm{WBGT} = 0.7\mathrm{NWB} + 0.2\mathrm{GT} + 0.1\mathrm{DB}$$

Where: GT = globe temperature
NWB = natural wet-bulb temperature
DB = Dry-bulb temperature

This index is important to the military, sports teams, construction workers, and anyone who will be exerting effort in hot weather. The American Academy of Pediatrics references this index [2] for child safety in hot temperatures. Also, athletes should cancel any outdoor training activity when the WBGT Index is above about 82 degrees Fahrenheit (about 28 degrees Celsius) [3]. The value at which marines will cease all outdoor training activity is about 90 degrees Fahrenheit (about 32 degrees Celsius).

The International Standards (ISO) number 7243 is also based on the WBGT. Parsons [12] gives a good description of how the WBGT Index can be applied globally by considering ISO 7243. One drawback can be the cost of the WBGT device, as well as having personnel trained to use the device.

Recently, the crews that were working on the Gulf oil spill in the summer of 2010 needed this index all along the Gulf Coast. However, in order to collect the data necessary for calculating the WBGT, a relatively expensive device was needed for many different locations along the Gulf coast making this task cost prohibitive. Consequently, the National Weather Service (NWS) was asked to provide the WBGT using only data that is routinely collected by the NWS. The main problem is that one of the variables in the equation to calculate the WBGT

index is the "globe temperature". This temperature is found by using a copper globe painted in black matte paint with a thermometer inserted so that the bulb is in the center of the globe. This temperature is not routinely collected by the NWS.

Turco et al. [13] derived equations to estimate the black globe temperature based on meteorological data. However, their model was a statistical model, not a physical model. The equations derived were regression equations computed from meteorological data. Although the equations were extremely accurate, a more accurate model may be derived from the heat equations for the black globe. According to the authors:

"The models developed resulted in great performance to predict the black globe temperature, allowing the estimation of bioclimatic indices to assess the conditions of the environment, to accomplish regional studies, and to indicate best house designs for animals" [13].

"This paper shows how the globe temperature can be approximated using only data routinely collected by the NWS. A fourth degree polynomial equation is derived for globe temperature with the coefficients dependent on readily available data. Then, it is shown that the fourth degree polynomial can be reasonably approximated by a linear equation, thus making computation less costly and time-consuming. Finally, some experiments were done to verify the accuracy of the estimate using the linear expression in terms of temperature" [6].

26.2 Derivation

The theory to support this derivation may be found in Kwuhn [1]. The following heat equation was taken from a paper by Hunter and Minyard [6], with the exception of the constant (h) in the second term on the right:

$$\left(1-\alpha_{sps}\right)S(f_{db}S_{sp}+(1+\alpha_{es})f_{dif}+\left(1-\alpha_{spl}\right)\sigma\varepsilon_a T_a^4=\varepsilon\sigma T_g^4+hu^{0.58}\left(T_g-T_a\right) \tag{26.1}$$

The coefficient in the second term on the right side of equation (h) is from the convective heat transfer coefficient. It was determined during testing that this coefficient varied according to the Solar Irradiance and the cosine of the zenith angle. A multiple power regression was performed to determine an equation for h in terms of S and cos(z), where z is the solar angle to zenith. The equation that approximates h is:

$$h=\mathrm{a}\left(\mathrm{S}^{\mathrm{b}}\right)\left([\cos(\mathrm{z})]^{\mathrm{c}}\right) \tag{26.2}$$

where a, b and c are determined experimentally from data using multiple power regression. Now, putting all T_g terms on the left of Eq. (26.1), and dividing by $\varepsilon\sigma$ we get:

$$T_g^4 + \frac{hu^{0.58}}{\varepsilon\sigma} T_g = \frac{(1-\alpha_{sps})S(f_{db}S_{sp} + (1+\alpha_{es})f_{dif}) + (1-\alpha_{spl})\sigma\varepsilon_a T_a^4}{\varepsilon\sigma} + \frac{hu^{0.58}}{\varepsilon\sigma} T_a \tag{26.3}$$

The values of all variables except T_g are either given or can be calculated from available data from the NWS. The following values are provided below.

Globe albedo for short and long wave radiation:

$\alpha_{sps} = \alpha_{spl} = 0.05$ so $1 - \alpha_{sps} = 1 - \alpha_{spl} = 0.95$.

Black globe emissivity: $\varepsilon = 0.95$.

Stephan–Boltzman constant:$\sigma = 5.67 \times 10^{-8}$ is used.

Albedo for grassy surfaces: $\alpha_{es} = 0.2$.

When these values are entered into Eq. (26.3) we get:

$$T_g^4 + \frac{hu^{0.58}}{0.95(5.67 \times 10^{-8})} T_g = \frac{0.95S(f_{db}S_{sp} + (1.2)f_{dif}) + 0.95(\varepsilon_a)\sigma T_a^4}{0.95(5.67 \times 10^{-8})} + \frac{hu^{0.58}}{0.95(5.67 \times 10^{-8})} T_a \tag{26.4}$$

Hunter and Minyard, in their paper [9], show that $S_{sp} = \frac{1}{4\cos(Z)}$, where z is the solar angle to zenith. Putting this into (26.4), we get

$$T_g^4 + \frac{hu^{0.58}}{(5.3865 \times 10^{-8})} T_g = S\left(\frac{f_{db}}{4\sigma\cos(z)} + \frac{(1.2)}{\sigma} f_{dif}\right) + \varepsilon_a T_a^4 + \frac{hu^{0.58}}{0.95\sigma} T_a \tag{26.5}$$

where S is solar irradiance, f_{db} is the direct beam radiation from the sun and f_{dif} is the diffuse radiation from the sun. The Stefan–Boltzman Constant is σ and h is the convective heat transfer coefficient. The convective heat transfer coefficient is calculated experimentally as indicated above. Finally, the ambient temperature is represented by T_a and the wind speed by u in meters per hour. All of these are given or may be calculated directly from data given by the NWS.

The last parameter on which the globe temperature depends is the thermal emissivity, ε_a. According to Hunter and Minyard [6], thermal emissivity can be calculated using

$$\varepsilon_a = 0.575 e_a^{(1/7)} \tag{26.6}$$

where e_a is atmospheric vapor pressure, which may be calculated by

$$e_a = \exp\left(\frac{17.67(T_d - T_a)}{T_d + 243.5}\right) \times (1.0007 + 0.00000346P) \times 6.112 \exp\left(\frac{17.502 T_a}{240.97 + T_a}\right) \tag{26.7}$$

where P is the barometric pressure and T_d is the dew point temperature.

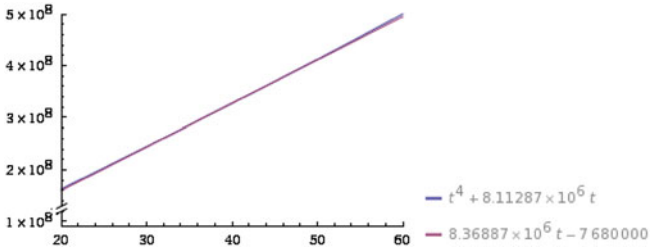


Fig. 26.1 The graph of $y = t^4 + ct$and y1 = Ct−7,680,000 for C ≈ 8,389,000 and t between 20 and 60. As can be seen in the graph, the two graphs are so close on the interval it is hard to distinguish between the two

When we take into consideration the fact that all parameter values in Eq. (26.5) are constants that can be entered at constant time intervals, we can reduce the equation to

$$T_g^4 + CT_g = B + CT_a, \tag{26.8}$$

where

$$C = \frac{hu^{0.58}}{(5.3865 \times 10^{-8})} \quad \text{and} \quad B = S\left(\frac{f_{db}}{4\sigma\cos(z)} + \left(\frac{1.2}{\sigma}\right)f_{dif}\right) + (\varepsilon_a)T_a^4.$$

By doing this, we can treat (26.8) as a fourth degree polynomial in terms of T_g. The values of T_g in which we are interested are in the interval [20, 60], since values below 20 °C are too cold to cause heat stress and values above 60 °C, in general, do not occur. Figure 26.1 shows a graph of $y = t^4 + ct$ and y1 = Ct – 7,680,000 (the tangent line approximation for the function y at t = 40) on the interval [20, 60] (C was calculated for a wind speed of about 15 mph).

Notice that the curve appears to be very close to the linear graph. We compute the curvature for y to see how close to a linear function y is. The curvature of $y = t^4 + ct$ is given by

$$k = \frac{12t^2}{(3t^3 + c)^{(3/2)}} \tag{26.9}$$

In order to get an understanding of the magnitude of the curvature, consider the graph (Fig. 26.2) of the function $k(t,u) = \frac{12t^2}{(3t^3+c(u))^{(3/2)}}$for u between 1 mph and 40 mph (1,690 meters per hour to about 65,000 meters per hour) and t between 20 °C and 60 °C.

Notice that the curvature is on the order of 4×10^{-10} or less on the domain of interest. This confirms the assumption that y is nearly linear for values of t and u that make sense for this context. It is therefore reasonable to use a linear

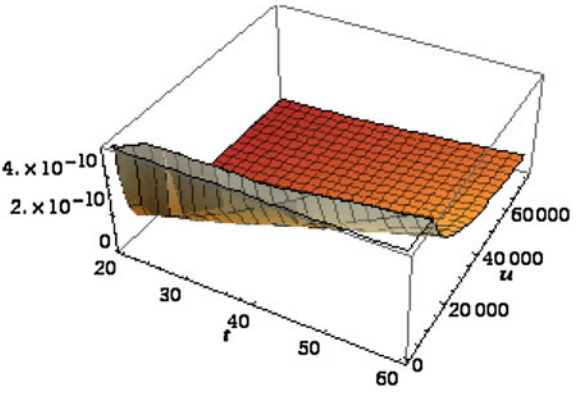


Fig. 26.2 Pictured above is the curvature of y for $20 < t \leq 60$ and $1 \leq u \leq 76{,}000$

approximation for y to solve for t (=T_g). In other words we may use a linear approximation on the left side of Eq. (26.8) to estimate the value of the globe temperature.

Using differential calculus to find the equation of the tangent to the curve at $t = 40$ (the midpoint of the interval [20, 60]), we find that the left side of Eq. (26.8) may be substituted by.

$$y_{est} = CT_g + 256000T_g - 7680000. \tag{26.10}$$

Putting this in place of the left side of Eq. (26.8) and solving for T_g, we get

$$T_g = \frac{B + CT_a + 7680000}{C + 256000} \tag{26.11}$$

with B, C and T_a as defined previously. Now we have an estimate of Tg dependent only on values which are either readily available from the NWS or may easily be calculated from data available from the NWS. Also, the equation is linear making for easier computation than what was necessary to solve the original fourth degree polynomial.

26.3 Preliminary Tests

In September 2010, three preliminary tests were conducted to test the accuracy of the globe temperature estimate. A Black Globe Temperature Sensor for heat stress was created by an employee of the NWS using specifications from an article by Purswell and Davis [8]. A picture of the unit is included in Fig. 26.3. This unit was used to get some preliminary readings to check for accuracy of the equation.

The first preliminary test was done on September 9, 2010 in front of the NOAA offices in Tulsa, Oklahoma. The weather conditions were hazy that day with air temperature 86 °F, and dew point temperature 69 °F. The barometric pressure was

Fig. 26.3 Black Globe Temperature Sensor used in preliminary tests (*left*). Black Globe Temperature Sensor for Heat Stress used to verify equations at the Oklahoma Mesonet test facility in Norman Oklahoma (*right*)

30.08 in. of Hg (about 993 mb for pressure not adjusted for sea level) and the solar irradiance was 336 W/m^2. The wind speed averaged around 5–6 mph during the measurement. The globe temperature was measured to be 91 °F using a black globe as described earlier in this paper. Using Excel, a spread sheet was created to use the derived equation to estimate globe temperature. The equation estimated the globe temperature to be about 91.434 °F.

Another preliminary test was performed on September 10, 2010. The conditions were sunny with air temperature 93 °F and dew point temperature 76 °F. The barometric pressure was 29.75 in. of Hg (about 982 mb for pressure not adjusted for sea level) and the solar irradiance was 754 W/m^2. The wind speed was measured at about 7 mph during the measurement. The globe temperature was measured to be 103 °F using a black globe. The equation estimated the globe temperature to be about 102.757 °F.

The third preliminary test was performed on September 17, 2010. The conditions were similar to the conditions on September 10. The air temperature was 94 °F and dew point temperature 76 °F. The barometric pressure was 30.05 in. of Hg (about 992 mb for pressure not adjusted for sea level) and the solar irradiance was 579 W/m^2. The wind speed was measured at about 3.7 mph during the measurement. The globe temperature was measured to be 105 °F using a black globe. The equation estimated the globe temperature to be about 105.175 °F.

The three preliminary tests indicate that the formula used to estimate the globe temperature is very accurate. If the estimate is within 1 °F, it is sufficient to estimate the WBGT index. As can be seen from the preliminary tests, the estimates are within about 0.5 °F. The main problem with our tests was that we had to estimate the wind speed and the formula is very sensitive to the value of the wind speed. However, the estimates for wind speed should be within about 0.5–1 mph. Also, these tests were all done at about the same time of day. Therefore, the solar irradiance and the cosine of the zenith angle were about the same for all preliminary tests.

26.4 Tests with Official WBGT Sensor

An official black globe was set up at the Oklahoma Climatological Survey site in Norman Oklahoma for data collection in June, 2011. The data from this site was then analyzed to determine accuracy of the black globe temperature equation. After extensive analysis and the derivation of the Multiple Power Regression Equation for the heat transfer coefficient, h, Eq. (26.11) was successful in calculating the black globe temperature to within one degree Celsius. An Excel document has been included in the appendix indicating the accuracy of the equations for different times of the day on two days in July.

The instrument was set to collect data every minute. The data collected included the black globe temperature, natural wet bulb temperature, ambient temperature, and relative humidity. Wind speed at two meters above ground, solar irradiance, and barometric pressure were collected, as well. The data was then retrieved via the internet.

As can be seen from the Excel document in the appendix, the black globe temperature was estimated to within 0.666 °C in every case. Since the factor for the black globe temperature in the WBGT Index equation is 0.2, this error will contribute less than 0.15 °C to the WBGT Index error.

26.5 An Algorithm

In this section, an algorithm is created for the calculation of globe temperature estimates. First, we will consider the values readily available from the NWS. These will be input values to be entered at the beginning of the program.

1. The values to be entered are wind speed (u in meters per hour), ambient temperature (T_a in degrees Celsius), dew point temperature (T_d in degrees Celsius), solar irradiance (S in Watts per meter squared), direct beam radiation from the sun (f_{db}) and diffuse radiation from the sun (f_{dif})
2. The zenith (z) angle may be entered or calculated. (The angle, z, must be in radians for Excel.)
3. The thermal emissivity must be calculated next. Using the following two equations:

   a.

$$e_a = \exp\left(\frac{17.67(T_d - T_a)}{T_d + 243.5}\right) \times (1.0007 + 0.00000346P) \times 6.112 \exp\left(\frac{17.502T_a}{240.97 + T_a}\right)$$

   b.

$$\varepsilon_a = 0.575 e_a^{(1/7)}$$

4. Now B and C can be calculated using the following equations

a.

$$B = S\left(\frac{f_{db}}{4\sigma\cos(Z)} + \frac{(1.2)}{\sigma}f_{dif}\right) + (\varepsilon_a)T_a^4$$

b.

$$C = \frac{hu^{0.58}}{5.3865 \times 10^{-8}}$$

where h is computed using (26.2).

5. Finally the estimate for globe temperature is calculated using Eq. (26.10).

$$T_g = \frac{B + CT_a + 7680000}{C + 256000}$$

26.6 National Weather Service Applications

The National Weather Service produces forecasts and observations of numerous meteorological parameters. One of these is the apparent temperature or Heat Index, based on work by R.G. Steadman [5]. In this work, Steadman constructed a table which uses relative humidity and dry bulb temperature to produce the "apparent temperature" or the temperature the body "feels." The OSHA uses a more detailed approach based on the WBGT Index [4] which provides guidelines to protect workers. In an effort to provide decision makers with forecast information on the WBGT, the National Weather Service in Tulsa has implemented and is testing the algorithm described earlier in this paper.

The algorithm is being tested for feasibility and accuracy in the forecast and also the hourly analysis of WBGT. In the current test phase, the first goal is to provide real-time, hourly estimates of WBGT across the WFO Tulsa forecast area of eastern Oklahoma and northwest Arkansas. The second goal is to make forecasts of WBGT out seven days to provide decision makers with important planning information. Ultimately, the algorithm will be offered to all National Weather Service Offices across the United States and elsewhere.

The National Weather Service uses a forecast system called the Graphical Forecast Editor, or simply GFE [7]. Virtually all weather variables are included in the system in a gridded format at either 2.5 km resolution or 5 km resolution. Grid fields of forecast parameters such as air temperature, dew point temperature, wind, pressure, cloud amount, the probability of rain, rainfall amounts, and many others are produced through a forecast period of seven days. From these parameters, an assortment of secondary forecast variables can be created through various algorithms. Some of these currently include relative humidity, wind chill index, heat

Table 26.1 Preliminary tests

	9/9/2010	9/10/2010	9/17/2010
Fdb	0.67	0.75	0.75
Fdif	0.33	0.25	0.25
Ta °F °C	86, 30	93, 33.89	94, 34.44
Td °F °C	69, 20.56	76, 24.44	70, 21.11
P in. of merc	30.08	29.75	30.05
S W/m^2	336	754	579
z degrees	38.44	36.65	41.41
u mph	6	7	3.7
Rh	67.5	54.27	52
Tw °C	26.77550	26.575561	26.34527
Ea	22.64868	28.487086	22.48619
Epsa	0.897936	0.9278434	0.897013
B	3614652151	7098450550	56176782
C	1197110170	1309071170	90440592
Tg °C °F	33.02, 91.43	39.31, 102.76	40.65, 105.18
E	28.54383	28.584421	28.24707
actual Tg °F °C	91, 32.78	103, 39.44	105, 40.56
WBGT			
Dimiceli	28.35, 83.02	29.85, 85.74	30.02, 86.03
Actual	28.30, 82.94	29.88, 85.79	30.00, 86.00
Australia	32.17, 89.90	34.39, 93.9	34.57, 94.23

index, icing index, wild fire spread index, and others. The WBGT algorithm described here will be tested and eventually added to the suite of products issued by WFO Tulsa.

An example of the WBGT Index graphics page can be found at http://www.srh.noaa.gov/tsa/dsp/element.php?element=WBGT. The example indicates the variability of the index across the forecast area due to differences in the contributing parameters. The darker areas indicate higher index values. Graphical forecasts of the WBGT may eventually be available to the public and decision makers, along with point forecasts of WBGT for any location on the map.

26.7 Conclusion And Future Work

The black globe temperature, until now, has either been measured using a Black Globe Temperature Sensor on site, or by using a regression equation to estimate the WBGT. One example of this equation is used by the Australian Bureau of Meteorology, but "does not take into account variations in the intensity of solar radiation or of wind speed" [10]. Dimiceli has derived a linear expression to estimate the black globe temperature to within about 0.666 °C, thus estimating the WBGT Index to within about 0.333 °C or .733 °F. More tests have been

Table 26.2 Some results from official Black Globe Sensor

	7/3/2011; 13:00	7/10/2011; 13:00	7/10/2011; 13:00
Fdb	0.75	0.75	0.75
Fdif	0.25	0.25	0.25
Ta (°C)	28.8	29.5	31.06
Td (°C)	17.57	16.2	16.59
P mb	973.71	971.03	971.37
S W/m^2	448.7	173	387.98
cos(z)	0.9244	0.9689	0.9226
u (mph)	2.164	2.734	3.86
Rh	50.7	44.65	41.86
H	0.1325695	0.098524678	0.127660528
Tw (°C)	21.454124	20.46198409	20.86084154
Ea	18.591840	16.74872176	16.93891306
Epsa	0.872970592	0.860047363	0.86143581
B	3979819932	1506448296	3444243993
C	444878843.2	378649115.5	599274858.9
Tg (°C °F)	37.74, 99.93	33.5, 109.1	36.8, 92.3
E	20.00916235	18.471	18.8050
actual Tg °C	37.16	33.58	37.47
Error	−0.58139651	0.103869211	0.665554019

performed using an official Black Globe Temperature Sensor. These tests provided data to get a multiple power regression equation to estimate the convective heat transfer coefficient (h) used in the estimation of the black globe temperature in the formula (26.11) by Dimiceli.

The formula has been added to an internal experimental web page in order to allow computation of the WBGT Index for a limited area. This web page allows interested parties to calculate the WBGT Index for their local area by entering the ambient temperature (T_a) and approximate average wind speed at their location. A tool which will be available to anyone interested in computing the danger of heat stress at their particular location is being created at the writing of this article. By doing this, the NWS will give sports teams, construction companies, military personnel, and other affected people, the ability to determine the danger of working in the heat at different times of the day. Eventually an application for cell phones will be designed to calculate the WBGT Index. It is the investigators' hope that this index will eventually be used nationally (or even internationally as Parsons [12] proposed in his article) by the NWS to indicate the dangers of heat stress in any location in the country.

After further research it has been determined that the direct beam irradiation and the diffuse beam irradiation values have more of an effect on the accuracy of the black globe temperature estimate than previously thought. In order to improve the estimate for the black globe temperature, equations will be derived to calculate these values. The direct beam irradiation and the diffuse beam irradiation equations will be a function of time of day, sun angle, and cloud conditions. The NWS has cloud data and the other two variables are readily available. Table 26.1, 26.2.

Acknowledgments The authors would like to thank Bruce Sherbon of NOAA for his construction of an inexpensive Black Globe Temperature Sensor. The authors would also like to thank Jamie Frederick of NOAA for programming the web page to calculate the WBGT Index. Finally, V. E. Dimiceli would like to thank Glenn Wiley of NOAA for helping V. E. Dimiceli to get a research sabbatical at NOAA/NWS in the fall of 2010.

References

1. Keuhn LA et al (1970) Theory of the globe thermometer. J Appl Physiol 25(5):750–757
2. Committee on Sports Medicine and Fitness (1991) Sports medicine: health care for young athletes, 2nd edn. Elk Grove Village IL, American Academy of Pediatrics, 1991:98
3. Helen MB, Joseph B, Douglas JC, Douglas MK, Paul EP (2002) National athletic trainers' association positional statement: exertional heat illnesses. J Athl Train 37(3):329–343
4. OSHA Technical Manual, TED 01-00-015 [TED 1-0.15A], updated June 2008. Available http://www.osha.gov/dts/osta/otm/otm_iii/otm_iii_4.html
5. Steadman RG (1979) Part I: A Temperature-humidity index based on human physiology and clothing science. J Appl Meteorol 18(7):861–873
6. Dimiceli VE, Amburn SA, Piltz SF (2011) Estimation of black globe temperature for calculation of the wet bulb globe temperature index. Lecture Notes in Engineering and Computer Science: Proceedings of the world congress on engineering and computer science 2011, WCECS 2011, 19–21 Oct 2011, San Francisco, pp 591–599
7. J Wakefield (2007) Information Generation Section, National Oceanic and Atmospheric Administration Earth System Research Laboratory. Available http://gfesuite.noaa.gov/EFTHome.html
8. Purswell JL, Davis JD (2008) Construction of a low cost black globe thermometer. Applied Engineering in Agriculture. ASABE 24(3):379–381
9. Hunter CH, Minyard CO (1999) Estimating wet bulb globe temperature using standard meteorological measurements. Department of Energy, Office of Scientific and Technical Information, U S A, WSRC-MS-99-00757, 2.7
10. Commonwealth of Australia (2010). Thermal comfort observations. Bureau of Meteorology, ABN 92 637 533.532
11. Campbell Scientific Corporation (2010) Black globe instruction manual. Campbell Scientific Corporation, Canada. Available http://www.osha.gov/dts/osta/otm/otm_iii/otm_iii_4.html
12. Parsons K (2006) Heat stress standard 7243 and its global application. Ind Health 44:368–379
13. Turco SHN, da Silva TGF, de Oliveira GM, Leitão MMVBR, de Moura MSB, Pinheiro C, da Silva Padilha CV (2008). Livestock Environment VIII, 31 Aug–4 Sept 2008, Iguassa Falls, Brazil 701PO408

Chapter 27
Using MOEAs to Outperform Stock Benchmarks in the Presence of Typical Investment Constraints

Andrew Clark and Jeff Kenyon

Abstract Portfolio managers are typically constrained by turnover limits, minimum and maximum stock positions, cardinality, a target market capitalization and sometimes the need to hew to a style (such as growth or value). In addition, portfolio managers often use multifactor stock models to choose stocks based upon their respective fundamental data. We use multi-objective evolutionary algorithms (MOEAs) to satisfy the above real-world constraints. The portfolios generated consistently outperform typical performance benchmarks and have statistically significant asset selection.

Keywords Asset selection · Financial constraints · Multi-objective evolutionary algorithms (MOEA) · Multi-period MOEAs · Mean-variance optimization (MVO) · Portfolio construction

27.1 Introduction

In finance, a portfolio is a collection of assets held by an institution or a private individual. The portfolio selection problem seeks the optimal way to distribute a given monetary budget on a set of available assets. The problem usually has two criteria: expected return to be maximized and risk to be minimized. Classical mean-variance portfolio selection aims at simultaneously maximizing the expected

A. Clark (✉) · J. Kenyon
Thomson Reuters, 707 17th Street, 22nd Floor, Denver, CO 80202, USA
e-mail: andrew.clark@thomsonreuters.com

J. Kenyon
e-mail: jeff.kenyon@thomsonreuters.com

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_27,

return of the portfolio and minimizing portfolio risk. In the case of linear equality and inequality constraints, the problem can be solved efficiently by quadratic programming, i.e. variants of Markowitz's critical line algorithm. What complicates this simple statement of portfolio construction are the typical real-world constraints that are by definition non-convex, e.g., cardinality constraints which limits the number of assets in a portfolio and minimum and maximum buy-in thresholds. In what follows, we use multi-objective evolutionary algorithms (MOEAs) as an active set algorithm optimized for portfolio selection. The MOEAs generate the set of all feasible portfolios (those portfolios meeting the constraints), calculates the efficient frontier for each and also their respective Sharpe ratio. The portfolio with the best Sharpe ratio becomes the portfolio used for the next time period. We chose MOEAs to solve a non convex optimization problem because there are certain outstanding problems in terms of their use:

1 In the literature MOEAs have not been used to solve multi-period financial problems (or multi-period problems in general),
2 The number and types of constraints in a real world financial portfolio problem exceeds what has been done with MOEAs so far, and
3 It is not known if MOEA stock selection is statistically significant.[1]

We answer all of these questions with a yes thereby advancing the understanding and use of MOEAs.

27.2 Financial Theory

We will briefly define the modern portfolio theory terms used in the problem and its solution [4].

The first term is efficient frontier. This frontier is calculated by trading off mean stock returns and their related variances. In essence, a combination of stocks, often referred to as the portfolio, is called efficient if it has the best possible expected level of return for its level of risk (where risk is usually proxied by the standard deviation of the portfolio's return). Every possible combination of stocks can be plotted in risk-expected return space and the collection of all such possible portfolios defines a region in this space. The upward-sloped part of the boundary of this region, a hyperbola, is called the efficient frontier.

The Sharpe ratio is a measure of the excess return (or risk premium) per unit of risk in an investment or a trading strategy. The Sharpe ratio is the standard measure of the risk premium when an efficient frontier is calculated. So the Sharpe ratio is used to characterize how well the return of a portfolio compensates the investor for the risk taken. The higher the Sharpe ratio the better the portfolio trades off risk and return.

[1] Asset selection is a test performed to determine if the portfolio outperformance is due to stock selection skills.

When comparing portfolios with differing expected returns against the risk-free rate (in our case 3-month U.S. Treasury bills), the portfolio with the higher Sharpe ratio gives more return for the same risk. Investors are often advised to pick investments with high Sharpe ratios. The best Sharpe ratio on the efficient frontier is by definition the best portfolio to invest in.

The Information ratio is another measure of the risk-adjusted return of a financial portfolio. The Information ratio is often used to gauge the skill of managers of mutual funds, hedge funds, etc. In these cases, it measures the expected active return of the manager's portfolio divided by the amount of risk that the manager takes relative to her benchmark. The higher the Information ratio, the higher the active return[2] of the portfolio given the amount of risk the manager has taken. Top-quartile investment managers typically achieve information ratios of about one-half.

Generally, the information ratio compares the returns of the portfolio with those of a benchmark such as the yield on three-month Treasury bills or an equity index such as the S&P 500.

27.3 Multi-Objective Optimization

A multi-objective optimization problem (MOP) differs from a single objective optimization problem because it contains several objectives that require optimization. When optimizing a single objective problem, the best single design solution is the goal. But for multi-objective problem with several (possibly conflicting) objectives, there is usually no single optimal solution. Because of this the decision maker is required to select a solution from a finite set of possible solutions by making compromises. A suitable solution should provide acceptable performance over all objectives. Given a set of multi-objective solutions, some of the set will be dominated by others in this set. Those that are not dominated by any others form what is called the Pareto set. In objective space, the set of objective vectors corresponding to the Pareto set is called the Pareto front.

Any single point on the Pareto front is called Pareto optimal. It is usually not optimal in the single objective sense since it usually does not minimize each of the objectives. However it represents a compromise such that if any solution exists that improves upon one objective then that solution will be worse on at least one other objective.

The Pareto set for any problem contains for each objective a point that truly minimizes that objective. For example, if you are trying to find a bridge design that has minimal mass, minimal cost and whose construction has a minimal carbon footprint, we can expect three solutions on the Pareto front to be the best possible solutions amongst those members of the vector X that are feasible bridge designs [5].

[2] Active return is defined as the return in excess of the compensation for the risk borne.

The main motivation for using evolutionary algorithms (EAs) to solve multi-objective optimization problems is that EAs can deal simultaneously with a set of possible solutions which allows us to find several members of what is called the Pareto optimal set in a single run of the algorithm. This differs from deterministic mathematical programming techniques where a series of separate runs is required. Additionally EAs are less susceptible to the shape or continuity of the Pareto front, e.g., they can easily deal with discontinuous and concave Pareto fronts. Discontinuity and concavity problems are known obstacles for deterministic mathematical programming.

Any solution on the Pareto front can be identified formally by the fact that it is not *dominated* by any other possible solution. A solution X is said to be dominated by solution Y if Y is at least as good on all counts (constraints) and better on at least one constraint. Stated mathematically (and assuming all the constraints are to be minimized): $f_i(Y) \leq f_i(X) \, \forall i = 1, M$ and $f_i(Y) < f_i(X)$ for some i.

Adapting any stochastic optimization algorithm (such as an EA) so it can perform a multi-objective optimization requires a change to the method of archiving possible solutions. As several possible solutions can be generated, an archive of the non-dominated (Pareto optimal) solutions needs to be maintained. A possible archiving scheme is:

- All feasible solutions (Pareto optimal vectors) generated are candidates for archiving.
- If a candidate solution dominates any existing members of the archive, the dominated solutions are removed.
- If the new solution is dominated by any existing member of the archive, the new solution is not archived.
- If the new solution neither dominates nor is dominated by any members of the archive, the new solution is added to the archive.

Using such a scheme as the search progresses, the archive will converge to the true trade-off surface between constraints.

27.4 Evolutionary Algorithms

A generic EA assumes a discrete search space H and a function $f : H \rightarrow \Re$.

The general problem is to minimize f given $X \in H$ where X is a vector of the decision variables and f is the objective function.

With EAs it is customary to distinguish *genotype*—the encoded representation of the variables from *phenotype*—the set of variables themselves. The vector X is represented by a string (or chromosome) s of length l made up of symbols drawn from an alphabet A using the mapping $c : A^l \rightarrow H$.

If the domain of c is total, i.e. the domain of c is all of A^l, c is called a decoding function. The mapping c is not necessarily surjective. The range of c determines the subset of A^l available for exploration by an evolutionary algorithm.

The range of c, Ξ where $\Xi \subseteq A^l$ is needed in order to account for the fact that some strings in the image A^l under c may represent invalid solutions to the original problem.

The search space Ξ can be determined by either Shannon or 2nd order Renyi entropy. If the decision variables X are independent, Shannon entropy applies. If the decision variables are correlated then 2nd order Renyi entropy applies. A minimization of either entropy will help define the feasible search space Ξ.

The string length l depends on the dimensions of both H and A with the elements of the string corresponding to *genes* and values to *alleles*. This statement of genes and alleles is often referred to as *genotype-phenotype mapping*.

Given the statements above, the optimization becomes one of minimize g given $S \in L$ given the function $g(s) = f(c(s))$.

With EAs it is helpful if c is a bijection. The important property of a bijection as it applies to EAs is that bijections have an inverse, i.e. there is unique vector X for every string and a unique string for each X.

In the implementation of EAs mechanisms inspired by evolution such as reproduction, mutation, recombination, selection and survival of the fittest are used. Candidate solutions to the optimization problem play the role of individuals in a population and the objective function determines the environment within which the solutions "live." Evolution of the population then takes place after the repeated application of the above operators.

In this process, there are two main forces that form the basis of EAs, recombination and mutation, which create the necessary diversity and thereby facilitate novelty. Selection acts as a force increasing quality.

Many aspects of EAs are stochastic. Changed pieces of information due to recombination and mutation are randomly chosen. On the other hand, selection operators can be either deterministic or stochastic. In the latter case, individuals with a higher fitness have a higher chance to be selected than individuals with a lower fitness but typically even weak individuals have a chance to become a parent or to survive.

To mathematically define the EA operators and functions we will define the EA fitness function first. As H is a nonempty set, $c : A^l \rightarrow H$ and $f : H \rightarrow \mathbb{R}$, we can define the fitness scaling function $T_s : \mathbb{R} \rightarrow \mathbb{R}$ and a related fitness function $\Phi \triangleq T_s \circ f \circ c$.

In this definition it is understood that the objective function f is determined by the application while the specification of the decoding function c and the fitness scaling function T_s are design issues.[3]

Execution of an EA typically begins by randomly sampling with replacement from A^l. The resulting collection is the initial population denoted P.

[3] If the domain of c is total, i.e. the domain of c is all of A^l, c is called a decoding function. The mapping c is not necessarily surjective. The range of c determines the subset of A^l available for exploration by the evolutionary algorithm.

More generally a population is a collection $P = \{a_1, \ldots, a_\mu\}$ of individuals $a_i \in A^l$. The number of individuals μ is referred to as the population size.

Following initialization, execution proceeds iteratively. Each iteration consists of an application of one or more evolutionary operators. The combined effect of the evolutionary operators applied in a particular generation $t \in N$ is to transform the current population $P(t)$ into a new population $P(t+1)$.

In the population transformation, $\mu, \mu' \in \mathbf{Z}^+$ (the parent and offspring population sizes respectively). A mapping $T : H^\mu \to H^{\mu'}$ is called a population transformation. If $T(P) = P'$ then P is a parent population and P' is the offspring population. If $\mu = \mu'$ then they are called simply the population size.

The population transform (PT) resulting from an evolutionary operatory (EO) often depends on the outcome of a random experiment. In Merkle and Lamont [6], this result is referred to as a random population transform (RPT) or random PT.

To define RPT, let $\mu \in \mathbf{Z}^+$ and Ω be a set (the sample space). A random function $R : \Omega \to T(H^\mu, \cup_{\mu' \in \mathbf{Z}^+} H^{\mu'})$ is called a random population transformation. The distribution of PTs resulting from the application of an EO depends on the operator parameters, in other words an EO maps its parameters to a RPT.

Now that we have defined both the fitness function and RPT, we can define in general an evolutionary operator: let $\mu \in \mathbf{Z}^+$, X be a set (the parameter space) and Ω a set. The mapping

$$Z : X \to T(\Omega, T[H^\mu, \cup_{\mu' \in \mathbf{Z}^+} H^{\mu'}]) \tag{27.1}$$

is an evolutionary operator. The set of evolutionary operators is denoted as $EVOP(H, \mu, X, \Omega)$.

There are three common evolutionary operators: recombination, mutation and selection. These three operators are roughly analogous to their similarly named count

In Merkle and Lamont's definition of the recombination operator [6], $r \in EVOP(H, \mu, X, \Omega)$. If there exists $P \in H^\mu, \Theta \in X$ and $\omega \in \Omega$ such that one individual in the offspring population $r_\Theta(P)$ depends on more than individual of P, then r is referred to as a recombination operator.

A mutation is defined in the following manner. Let $m \in EVOP(H, \mu, X, \Omega)$. If for every $P \in H^\mu$, for every $\Theta \in X$ and for every $\omega \in \Omega$ and if each individual in the offspring population $m_\Theta(P)$ depends on at most one individual of P then m is called a mutation operator.

Finally, for selection let $s \in EVOP(H, \mu, X \times T(H, \mathbb{R}), \Omega)$. If $P \in H^\mu$, $\Theta \in X$, $\Phi : H \to \mathbb{R}$, in all cases and if s satisfies $a \in s_{(\Theta,\Phi)}(P) \Rightarrow a \in P$ then s is a selection operator.

Choosing the numerical values or techniques that will compute/simulate mutation, recombination and so forth tends to be heuristic. In the MOEA descriptions below, we use standard procedures and values to set the evolutionary parameters. For those interested in reading more about the various values and techniques that are used to select evolutionary parameters, the authors suggest Deb [3].

27.5 Problem Statement

The two optimization problems we face are: generate a series of monthly portfolios that outperform the S&P 500 over the last 30 years and generate a set of monthly portfolios that outperform the Russell 3000 Growth index over the last 15 years.

The constraints we will operate under are: turnover is not to exceed 8 % per month, the minimum stock position is set at 0.35 % of the net asset value of the portfolio, the maximum stock position is set at 4 % of the net asset value of the portfolio and a target market capitalization constraint where the average market capitalization of the portfolio must be greater than the average market capitalization of all stocks available to purchase in the current month (this last constraint will mean both portfolio sets will be what is called "large-cap" The S&P 500 and the Russell 3000 Growth are large-cap benchmarks). Another constraint common to both problems is we must choose stocks that maximize the scores generated by a multi-factor stock model . This constraint typifies the use of what is called fundamental financial data to select stocks that are potential candidates for the final monthly portfolios.

An additional constraint was added for the Russell 3000 Growth problem: we cannot exceed the average book-to-price value of all stocks available for purchase in the current month. Meeting this constraint will mean we will generate the required growth portfolios for the Russell 3000 Growth pool.

We solve the issues of the constraints by breaking them into two sets and use two MOEAs. The first MOEA generates potential portfolios that lie within the bounds of all the constraints except turnover and position. In the second MOEA, we trade off the turnover and position constraints as well as mean return and variance (the last being the typical factors used in mean-variance optimization). We set the rebalance period to quarterly versus monthly but stay within the stated turnover constraint (not to exceed 8 % per month). The reader can view the pseudo code which outlines the steps each MOEA takes to solve the problem, as well as the values used for the EA operators, at [2].

27.6 Results

In Table 27.1 are the 1, 3, 5 and 10 year annualized (transaction cost adjusted) returns for the large-cap MOEA portfolios and the S&P 500. The period covered is from December 1979 through December 2009 (121 months).

Table 27.2 has the annualized risk and cumulative return on 10,000 USD for the S&P 500 and the MOEA portfolios for the same time period.

In Table 27.3 are the 1, 3, 5 and 10 year annualized (transaction cost adjusted) returns for the large-cap growth MOEA portfolios and the Russell 3000 Growth. The period covered is from December 1996 through December 2009 (53 months).

Table 27.1 Annualized Returns: S&P 500 and MOEA

	1 Year (%)	3 Year (%)	5 Year (%)	10 Year (%)
S&P 500	9	30	54	137
MOEA	13	44	84	239

Table 27.2 Cumulative and Risk Adjusted Returns: S&P 500 and MOEA

	Sharpe ratio	Information ratio	Cumulative return on 10,000 USD
S&P 500	1.1	N/A	37,070
MOEA	1.8	0.14	49,500

Table 27.3 Annualized Returns: Russell 3000 Growth and MOEA

	1 Year (%)	3 Year (%)	5 Year (%)	10 Year (%)
R3000 Growth	10	33	61	159
MOEA	14	48	93	271

Table 27.4 Cumulative and Risk Adjusted Returns: Russell 3000 Growth and MOEA

	Sharpe ratio	Information ratio	Cumulative return on 10,000 USD
R3000 Growth	0.14	N/A	25,688
MOEA	0.24	0.32	31,298

Table 27.4 has the annualized risk and cumulative return on 10,000 USD for the Russell 3000 Growth and the MOEA portfolios for the same time period.

As seen in Tables 27.1, 27.2, 27.3, 27.4 the percent return and USD return of the MOEA portfolios is approximately double the value of their benchmarks.

On a risk-adjusted basis, the results are somewhat mixed. The information ratio for the large-cap growth MOEA is very significant while its Sharpe ratio is only a little larger than the Russell 3000 Growth Sharpe ratio and both are little different from 0 (zero). For the large-cap MOEA, its Sharpe ratio is significantly larger than its benchmark, but its information ratio is very small. So it is not clear that the MOEA portfolios are the better risk-adjusted portfolios in all cases. By not underperforming their benchmarks on a risk-adjusted basis, the implication is that at a minimum the MOEAs reside on a higher curve in risk-return space and could be the more attractive portfolios to investors.

As to the other constraints:

1 In 100 % of all cases, the MOEA portfolios on a weighted market capitalization basis met or exceeded the market capitalization constraint,
2 For the smallest and largest positions constraints based on net asset value, none of the MOEA portfolios broke this constraint on either the minimum or maximum side, and
3 Turnover did occasionally exceed the 8 % limit per month. These occurrences tended to happen early in the 1980's portfolios just as the MOEA was getting on

its feet. And the turnover limit was broken in a small number of later portfolios as well. The authors conjecture that in the latter cases the non-dominated feasible solutions handed off to the second MOEA were composed of individual stocks different enough from the prior quarter's portfolio that all the resulting portfolios prevented the turnover constraint from being met. This is an open question however and needs further investigation.

As to one of the reasons why we chose to use MOEAs—is the stock selection of the quarterly portfolios statistically significant—we find the answer to be yes. As measured by John Guerard the MOEA asset selection was very significant. This is a very pleasant surprise to the authors, especially as the portfolios typically contained 150–200 stocks. Our results demonstrate that MOEAs can generate statistically significant asset selection while operating under real world constraints and using fundamentally driven stock scores.

27.7 Conclusions

In this paper we demonstrate that MOEAs in the presence of real world constraints can generate portfolios that have higher returns than their benchmarks, comparable (if not better) risk adjusted returns versus their benchmarks and statistically significant asset selection.

We arrive at these portfolios by dividing the MOEA in two: the first MOEA generates all the non-denominated feasible sets that meet all the constraints except turnover and minimum and maximum position. The second MOEA trades off the last two constraints along with mean and variance to come up with the final portfolio that has the best Sharpe ratio (or meets the maximum turnover if the other constraints are not met).

Finally, as best as we know, this is the first multi-period use of MOEAs in stock portfolio construction [1]. We are encouraged by the results and hope others will extend and improve upon our work.[4]

References

1. Clark A, Kenyon J (2011) Using MOEAs to outperform stock benchmarks in the presence of typical investment constraints, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science (2011) WCECS 2011, 19–21 October, 2011. USA, San Francisco, pp 1050–1053

[4] The authors would like to thank John Guerard of McKinley Capital for the challenge he set us. We would not have tested MOEAs in the way described above without John's challenge.

2. Clark A, Kenyon J (2011) Using MOEAs to outperform stock benchmarks in the presence of typical investment constraints. SSRN (http://ssrn.com/abstract=1893644 or 10.2139/ssrn.1893644 and arXiv (http://ssrn.com/abstract=1893644)
3. Deb K (2000) Multi-objective optimization using evolutionary algorithms. Wiley, New York
4. Goetzmann W (1996) An introduction to investment theory. Available at http://viking.som.yale.edu/will/finman540/classnotes/notes.html
5. Knowles J, Corne D, Deb K (eds) (2008) Multiobjective problem solving from nature: from concepts to applications. Springer, Heidelberg
6. Merkle LD, Lamont GB (1997) A random function based framework for evolutionary algorithms. In: Proceedings of the 7th international conference on genetic algorithms, East Lansing, MI, USA, July 19–23, 1997 Morgan Kaufman, San Francisco

Chapter 28
Continuous Integration and Automation for DevOps

Andreas Schaefer, Marc Reichenbach and Dietmar Fey

Abstract The task of managing large installations of computer systems presents a number of unique challenges related to heterogeneity, consistency, information flow and documentation. The emerging field of DevOps borrows practices from software engineering to tackle complexity. In this paper we provide an insight in how automation can to improve scalability and testability while simultaneously reducing the operators' work.

Keywords Administration · Automation · DevOps · eLearning · Heterogeneous Systems · System Management

28.1 Introduction

A common sight at universities is a highly complex and heterogeneous computer system infrastructure with low turnaround times in terms of both, hardware and personnel. Ensuring a high quality of service is mandatory for both, teaching and research, but many institutes find themselves hard-pressed for manpower. Thus, the solution to the plethora of challenges presented by this setting is not to work harder, but smarter. Working smarter almost always means that work is automated.

A. Schaefer (✉) · M. Reichenbach · D. Fey
Chair for Computer Architecture, Friedrich-Alexander-University, Erlangen, Germany
e-mail: andreas.schaefer@informatik.uni-erlangen.de

M. Reichenbach
e-mail: marc.reichenbach@informatik.uni-erlangen.de

D. Fey
e-mail: dietmar.fey@informatik.uni-erlangen.de

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_28,

The key to good automation is to make it flexible, so that it can be adapted to various use cases. For instance instead of requiring the admins to manually search for errors in the network, a set of automated tests should verify the availability of all systems. Of course this set of tests needs to be tailored for each machine's role in the network. Another important example is the management of knowledge in an eLearning system. A good practice for collaboration is to store all data in a revision control system (RCS) such as Git. An RCS however needs to be interfaced with the eLearning system (e.g. Moodle). Instead of manually synchronizing both data stores, which is tedious and error prone, the eLearning system should automatically fetch updates from the RCS repository.

In this context, continuous integration means that there are no monolithic changes to the system configuration (e.g. major updates), but all updates ripple through the systems in a series of small changes. This may initially sound like hell to a sysadmin as each update may introduce new errors. But our approach makes use of configuration management and systems monitoring, which together enable the operators to test changes in an isolated environment before rolling them out and to track failures in near real-time.

The network found at our chair [8] is of medium size, but high complexity. On the one hand computer architects need fat nodes with lots of memory to carry out synthesis and system simulation. On the other hand they require servers to house FPGA boards with direct access to the PCIe bus. The high performance computing folks need servers with lots of PCIe slots suitable for GPUs (Graphics Processing Units), and also medium sized MPI [3] clusters. Those MPI clusters are then again relevant to grid computing research, as they are well suited for high throughput computing jobs. The constellation of our systems is illustrated in Fig. 28.1.

Previously, our approach to system administration was to have one or two experts who would take care of all installed systems. Homogeneity was ensured by running the same Linux distribution (Debian stable) on all nodes. This made it easy to automate basic tasks such as backup and updates via homebrew scripts. But as our chair grew and research interests became more diverse, this approach did not scale: the GPU machines required frequent updates to the Nvidia drivers and CUDA libraries while the systems sporting IBM Cell BE did not work well with Debian, but did call for a Redhead based distribution. Additionally, the closed source software for the hardware engineering tasks put the admins under unexpected load, since they were not familiar with the pitfalls of its installation. In other words: users could not work efficiently because they had to wait for the admins, who were feeling their powers spread thin between an increasingly complex range of specialized servers.

The alternative was to move administrative powers to the actual users of the system. Each admin could then use the best operating system and configuration for his use case, and would only have to deal with software and hardware he is used to. The challenge with this mode of operation is to prevent the individual admins from being swamped by having to replicate basic functionality such as login services, home directories and backup. We saw a need to reform our way of system administration. Our goals for this were:

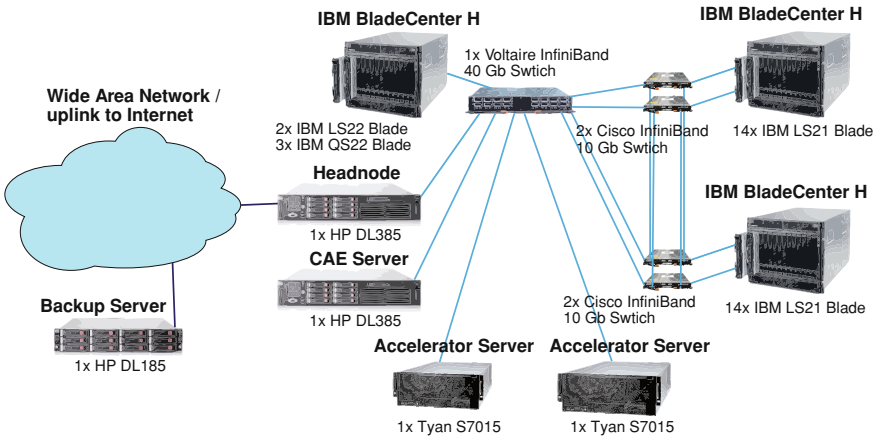


Fig. 28.1 Map of the systems in our chair's HPC laboratory and the network spanning across them. What is most striking is the heterogeneity of the systems involved: while the LS21 blades in the *whistler* cluster on the right are with just 8 GB RAM and 4 cores rather lightweight, the CAE server has what constitutes a fat node: 96 GB of RAM and 48 cores. The accelerator servers special pieces of hardware whose cases, board layouts and power supplies have been optimized to provide a maximum number of PCIe 2.0 $\times$ 16 double width slots (8 in each node). They are used for testing GPU codes and FPGA designs. The backup server is located off-site to increase disaster safety

1. high degree of automation, to keep the total workload low, despite maintaining a steadily increasing number of servers,
2. flexibility, in order to be able to accommodate the heterogeneity of the servers in use,
3. scalability, to share the load among all admins,
4. traceability, which allows admins to track down who made which changes when and why,
5. *don't repeat yourself*, an advice from Hunt et al. [5], teaches the avoidance of redundancy. In our case this means to prevent systems from entering an inconsistent state when two databases store different versions of the same data (e.g. */etc/hosts* storing IPs which differ from the actual host addresses),
6. testability, to quickly diagnose and remedy failures,
7. repeatability, for automatically applying the configuration to new machines (not just already running ones), too, thereby greatly reducing the deployment time—and cost.

Being software developers, we turned to DevOps practices. This allowed us to use tools from software engineering (e.g. a revision control system) together with dedicated administration tools (e.g. Puppet, Nagios). With respect to network architecture, our basic approach was to outsource basic functionality to a new, central head node, while leaving the details of specific expert hosts to their corresponding admins. The next section describes the basic infrastructure we

build, while Sects. 28.3 and 28.4 describe the DevOps inspired parts. These address the heterogeneity of the systems and the team oriented aspect of our approach.

28.2 The Network

First, we identified a number of common services which each system would require and which could be incorporated by the headnode.

1. shared user database
2. common home directories
3. secure backup of confidential user files
4. resource arbitration
5. monitoring of system health and performance
6. documentation of changes to the configuration

Afterwards we tried to identify the most suitable tools for each task. Perhaps the crucial point was the network file system for the home directories. Since both, hardware and high performance computing groups work with large datasets easily in the Terabyte range, it needs to be fast. But, because for some of portions of the data our users had to sign NDAs (for industry projects), it also needs to be secure. Also, we wanted POSIX semantics to ensure compatibility with existing software. We chose NFSv4 over InfiniBand, as this is fast enough for our uses, and also allows for encryption. The user database is implemented using Kerberos and LDAP. Thereby we can implement strict access control for NFS shares, while simultaneously being open to other user databases, e.g. in order to import the department's list of student accounts. PAM plugins (e.g. *pam_listfile*) allow us to limit the access of users to certain systems, thereby preventing e.g. students from swamping the staff's CAE (computer aided engineering) server.

In our case resource arbitration refers to the task of automatically allocating pieces of hardware to a given user for a certain period of time. Typically this is done by a batch queuing system. Originally we planned to use this only to manage the flow of jobs on our compute cluster, but soon we realized that we were facing a similar problem on our PCIe servers: a varying number of users were competing for a significantly smaller number of GPUs and FPGA boards. With a small user base it was sufficient to use IRC to let the colleagues know who was using a certain PCIe device, but as more and more students started using the devices for their projects, we had to come up with an automated method for resource arbitration.

We chose the Sun Grid Engine [6] (SGE, now renamed to Oracle Grid Engine) as a batch scheduler as it is one of the most mature systems freely available and comes with all the features we need. The SGE consists of three types of nodes: the execution hosts are those who run the actual jobs. Submission hosts are used to send jobs to the system. The planning of when to run which job on what machines is done by the host running the central scheduler. For this the scheduler maintains a

number of queues. The queues basically function as FIFOs, but with a twist: not just waiting time, but also job priorities and user/project fairness are taken into consideration. This prevents individual users from clogging the queues with a high number of jobs. Administrators can configure the queues to give certain user groups (e.g. staff) prioritized access, while limiting the resource usage of other (e.g. students).

However, setting it up for the PCIe servers was a bit of a challenge: in our installation larger number (up to eight) PCIe devices may be located in a single server, so the batch system should be able to schedule multiple jobs on a single server, as long as their resource allocations permit this. For measurements however the system should also allow jobs to use a node exclusively, e.g. for precise performance measurements. Finally, a job might require multiple accelerators, possibly on multiple nodes, in parallel.

Our initial approach was to create a single queue for the PCIe servers and let them process one job after another. While this would allow the jobs to access all PCIe devices exclusively, the resulting resource utilization was poor. Another attempt was to create a queue for each requestable PCIe device. This would allow us run multiple jobs on each server while still allowing exclusive scheduling where necessary. However,jobs couldn't reserve multiple accelerators simultaneously. Our final setup uses a single queue *gpu.q* and each node has a number of *complex values*. In the SGE these complex values can be used to model hardware resources which a job can request and (temporarily) block. This is most commonly used to model the available RAM or CPU cores, but can equally be used to arbitrate a guaranteed IO bandwidth or, as in our case, available GPUs and other PCIe devices.

When a job is started by the SGE and has reserved a number of devices, it needs to know e.g. the corresponding CUDA device numbers. For this we wrote a custom script *alloc* which maintains a database of all present and reservable resources. It returns all required IDs for a given device and allocates the device to the current job. Using this mapping, the tool can also free resources if a job has ended but failed to deallocate its devices, so crashed jobs do not place a problem.

Monitoring needs to satisfy two demands: first of all we need an automated way to check the functionality of our installation, similar to what unit tests are to software. Examples include a working SSH daemon on all nodes or a running SGE execution daemon. For this we chose Nagios, as its architecture allows for custom tests and it is well suited for sending alarm messages on multiple channels.

Second, a metering tool is required to identify possible performance bottlenecks, which may affect system availability in the future, e.g. temporary high load situations or an exhausted network bandwidth on certain servers. Nagios is good at telling if a certain measured value has crossed a certain threshold, but it is bad at reporting how this value has developed across time. Therefore we use Ganglia, which can provide plots of basic performance metrics for all nodes. It may be extended with custom metrics, e.g. to plot the temperature measured by an external probe.

For disaster security our backup server is located in a different server room. We need to ensure security in this context in multiple ways:

- The backups should not be lost if one or two disks fail. Therefore the backup server features a RAID6 device which will only fail if three drives fail simultaneously.
- The WAN connection between both server rooms is not to be trusted, so all access has to be protected. We export storage using CHAP protected ISCSI volumes.
- Multiple systems will store their backups on this system. Some may carry data for which NDAs have been signed. So admins of the different systems should not be able to access the backups of other systems. Thus we encrypt all backups using LUKS. Only the servers mounting the shares can decrypt them. To be able to restore data, the corresponding admins keep an offline copy of those keys.

28.3 Configuration Management

This section outlines how we use Puppet to automate the installation of packages and changes to the systems' configurations. The beauty of this approach is that new nodes only need a basic operating system installed, along with the simple Puppet client. All other configuration and installation work is then taken over by Puppet. This greatly reduces our deployment time for new systems. Also, nodes may be migrated from other forms of administration to this one step by step, as Puppet's configuration catalog may include setups tailored for each node individually (see Figs. 28.2 and 28.3).

The infrastructure illustrated above requires an extensive configuration of each node. If the configuration was static and all nodes would use the same Linux distribution, we could use a system image to fill the node with a suitable initial configuration. Our experience however is that the configuration needs to change frequently (e.g. because new packages are installed) and also different Linux distributions are most suitable for the different machines.

The standardized formulation of system configurations and their deployment has gained a lot of attention in the recent years [2, 9]. We chose Puppet, as it is more feature rich than the aging Cfengine, but is simultaneously more mature than Chef. Puppet consists of a central configuration server which is being polled by clients for changes to the configuration. The configuration itself is described via a set of scripts written in a domain specific language. It offers a unified interface to tasks like starting system services or handling packages, which may require different actions on each operating system.

The different roles each of our systems need to play are reflected by a custom class hierarchy as shown in Fig. 28.2. Basic services are defined in modules, which are then included in the classes. For instance the root class *UnixNode* includes the basic package and config file modules as well as—among others—the module

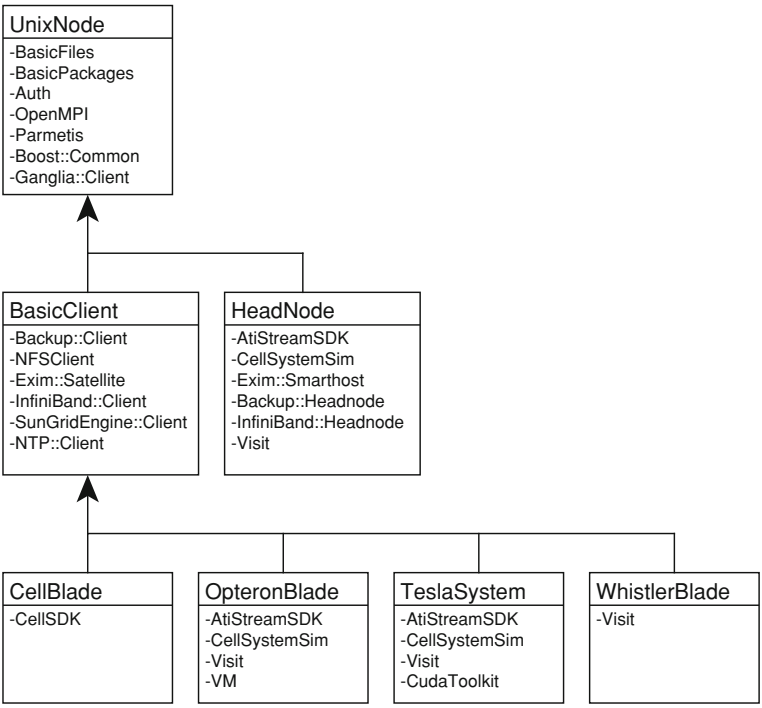


Fig. 28.2 Class diagram of our Puppet setup. Each individual class stores the setup of one category of nodes. For this it may inherit settings from another class and include a number of modules

Fig. 28.3 Excerpt from our Puppet installation's manifest file. Two nodes are configured: *faui36i* is an accelerator server which mainly houses GPUs, while *whistler*01 is an LS21 blade and part of our medium sized cluster computer

```
node "faui36i.informatik.uni-erlangen.de" {
  $remergepackages = "nvidia-drivers"
  $openglinterface = "nvidia"
  include teslasystem
}
node "whistler01.informatik.uni-erlangen.de" {
  include whistlerblade
  debian_net_config { private:
    ip_eth0 => "192.168.1.20",
    ip_ib0  => "192.168.0.30",
    ip_ib1  => "192.168.0.31"
  }
}
```

Auth, which sets up LDAP and Kerberos clients and configures PAM to use those. A short example of a custom Puppet module can be seen in Fig. 28.4.

The class *CellBlade* is specific to our IBM QS22 blades (each with two PowerXCell 8i processors), which run Fedora. *OpteronBlade* refers to those blades featuring two AMD Opteron six-cores, which are a lot beefier than the smaller

```
class ganglia {
  class client {
    package { $operatingsystem ? {
      Gentoo => "ganglia",
      Fedora => "ganglia-gmond",
      default => "ganglia-monitor" } :
    }
    file { "/etc/ganglia/gmond.conf":
      owner => root,
      group => root,
      mode => 644,
      source => "puppet:///ganglia/gmond.conf"
    }
    service { $operatingsystem ? {
      Debian => "ganglia-monitor",
      default => "gmond" } :
      subscribe => File["/etc/ganglia/gmond.conf"],
      ensure => running,
      enable => true
    }
  }
}
```

Fig. 28.4 Example for a Puppet module. In this case we see the client part of our Ganglia installation. It first installs the monitor package, which unfortunately has different names on each Linux distribution, and then ensures that the service is up and running. It gets restarted each time the configuration file (also managed by Puppet) is updated

*WhistlerBlade*s, and are thus suited for a larger range of applications. For instance running VisIt [1]—our tool of choice for visualizing the 3D results of our simulation codes. The classification *TeslaSystem* is used by the PCIe servers, which were originally only used to house Nvidia Tesla GPUs, but changed over the time to accommodate all sorts of accelerators, including AMD GPUs and FPGA boards. The *HeadNode* has a dedicated class, which shares some modules with the client classes, but generally needs tweaks to its configuration as it mostly runs the server parts of the services. We found it useful to keep the headnode's configuration in the Puppet repository, too, even though no other node needs to duplicate it, because this makes it much easier to trace changes.

For some packages, especially those that communicate via the network, we have to have the same version installed on each node. That feature may not be achievable with the stock packages available in the different Linux distributions, as their versions may differ. Therefore custom puppet modules handle the installation of the OFED InfiniBand drivers, our MPI environment Open MPI[4] and the SGE. The class hierarchy allows us to examine new packages and modules selectively on single nodes and only enable them on all nodes after they have been thoroughly tested.

Puppet itself handles the distribution of configuration changes to the nodes, but it does not facilitate the communication and documentation of these changes among a group of administrators. We therefore decided to place our puppet configuration in a Mercurial (HG) repository and let a Trac (by Edgewall) installation interface with it. Trac's integrated Wiki allows the users to maintain system documentation in a single place. We use the ticket system to assign tasks to the different admins. One benefit of a single, integrated system is that the wiki, tickets, commit messages and files in the repository may reference and interact with each other (e.g. tickets may be automatically closed via certain commit messages and the wiki may link to parts of the source code).

28.4 Tracking Changes

This section is meant to give an overview of how we use our setup to manage the systems' configurations.

The core is the Mercurial repository, which stores the configuration data. An admin would first update his local copy of the repository by pulling changes from the server. After making modifications and committing these locally, he would push them back to the server. The client machines regularly poll the Puppet server and apply the retrieved configuration catalog. Admins can view changes to the repository in Trac's timeline. Also, after making changes to the configuration, it is often sensible to update the user documentation, which is then accessible to users, too.

Figure 28.3 shows an excerpt of our *manifests/site.pp* file which defines the setup for all client nodes. *faui36i* is one of the PCIe servers, which we mainly use to house GPUs. Its type is set to *teslasystem*, a historical name which stems from the node's first use. It will—among a variety of other packages—install Nvidia's GPU drivers (and update them after each kernel update) and additionally the CUDA toolkit and SDK. *whistler*01 is one of the smaller blade servers. The *debian_net_config* passage defines *whistler*01's network setup. While this first appears to be inferior to running an DHCP server, we actually found this approach to better fit our needs: by placing the address information within Puppet, we avoid repeating the same data across different databases (e.g. */etc/hosts* and the DHCP server) and can furthermore automatically extract the data and reformat it for other uses (e.g. the aforementioned host file).

Because of the high number of nodes with identical configurations, we wrote a short Ruby script which can generate the manifest from a short template. This means that we have to maintain significantly less code (the manifest is slightly larger than 13 kB while the generator script weights only less than 3 kB) and also adding a new node (with an configuration identical to some previous node) now boils down to adding a single line of code to the generator script.

28.5 Application Specific Infrastructure

In the previous section, we presented a hardware and software environment with an automated system maintenance infrastructure. The usage was limited to standard architectures e.g. servers, PCs, GPUs. Also, most of the software was developed for these standard architectures. Hence, there was no need to implement a custom tool set. In a more specialized environment we have to extend our infrastructure by developing secure and fast tools for automated setup and easy maintenance. Therefore, we will present in this section methods for the handling of heterogeneous, application specific infrastructures. As an example we use a remote hardware laboratory as part of an online lecture.

In our hardware development lectures for students, we gained the experience that most of the practical time was wasted to understand the development tools and to set up the working environment, e.g. window manager and *rc scripts*. Therefore, we decided to create a new virtual lecture, where the students can do their exercises at their PCs at home. But it is not possible that every student can get an experimental hardware board from the university for the practical tests. On the other hand, practical experience is important for the appreciation of the topic. Therefore, we created a remote hardware laboratory, based on FPGAs, were the students can test their own hardware projects on real hardware via a remote connection over the internet. Hence, our online hardware lecture contains two parts, which were firstly described in [7]. First the content management system which provides the lectures content and second the remote hardware laboratory, for practical hardware tests. In the following, we describe how this application specific architecture is integrated in our chair computer infrastructure system.

eLearning Content. To create eLearning content some challenges have to be meet. First the content should be provided in HTML for online access, but also as PDF for printing and archiving the lectures. Moreover a version control system is required. Because such lessons are created from different employees of the chair with special knowledge at their topic, an easy way to share the source is necessary. Moreover, it is preferable to have a version history, where you can view, merge and revert changes. Therefore, we chose an XML based flexible and automated content creation flow, based on *eLML*, *git* and *moodle*.

For an easy creation of different output formats (PDF and HTML), an XML based flow has been chosen. The eLML is a XML based markup language which provides special tags for eLearning content such as lecture, exercise and so on. With the help of XSL transformation scenarios, the source files given in eLML can easily be transformed to the output formats like PDF or HTML. For an automation of this process, *saxon9* and *fop* is used. Because of using XML and, hence, the text based development flow, the development process can gain efficiency by using a version control system. Because of the local development strategy and the well working *index concept*, we have chosen git. Git was developed for the Linux kernel development and is now used for several large open source projects where efficient version control is necessary. The last important point is, how the created

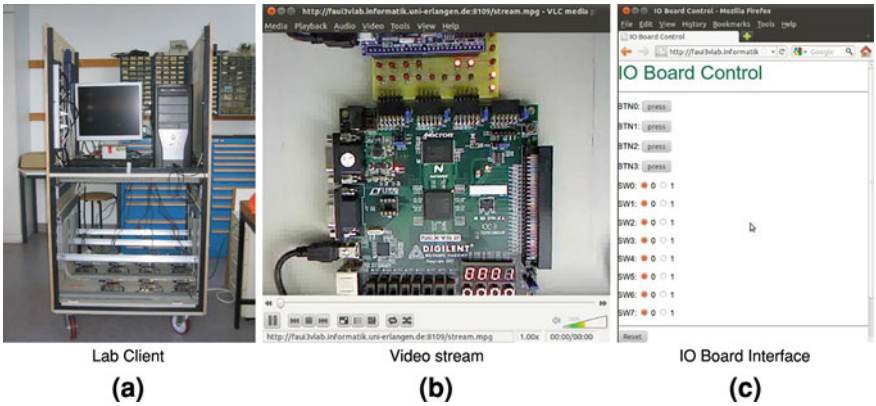


Fig. 28.5 A lab client system an their web user interface after programming a board

content is published for the students. For this we are using a *Content Management System (CMS)* called *moodle* which was immediately created for hosting eLearning content. This CMS was written in PHP and allows a sufficient integration in our apache web servers. Unfortunately, no accessible interface was created to allow an automated integration of content from an version control system. Therefore, we implemented *git hooks* and a moodle interface for an automated lesson import from git. Now, if a teacher commits and pushes changes in their XML lessons to the git repository, the git hook is executed which updates the content, executes the XSL transformation and publishes the results on the moodle server. Moreover, it is possible to control the update process by writing a commit message in a special syntax. This allows a flexible but easy to use interface for the complex automated publishing process. Because the transformations of the sources can take a view moments, a feedback in terms of an email response will be created. Also the complete build process log is attached to this email to find possible errors in the highly automated system.

Remote Hardware Laboratory. With the development of *Field Programmable Gate Arrays (FPGA)* a powerful device for hardware developers was created for testing of user-specific circuits, from simple boolean functions up to complex parallel multiprocessors. The SRAM based device allows a reconfigurability, wherewith it is possible to load any hardware architectures as often as required. Such programmable chips are an efficient trade of between performance and flexibility compared to an *Application-Specific Integrated Circuit (ASIC)*.

The usage of such FPGA devices at a remote hardware laboratory leads to a very heterogeneous hardware setup, which has to be controlled by a PC system.Because many small and different devices, e.g. the FPGA boards are connected to this server, special requirements for controlling these devices have to be meet. In contrast to this, a pure software systems, where every server is running an operating system and is connected to the internal network, controlling and communication can be done at a higher abstraction level. Moreover, no software is

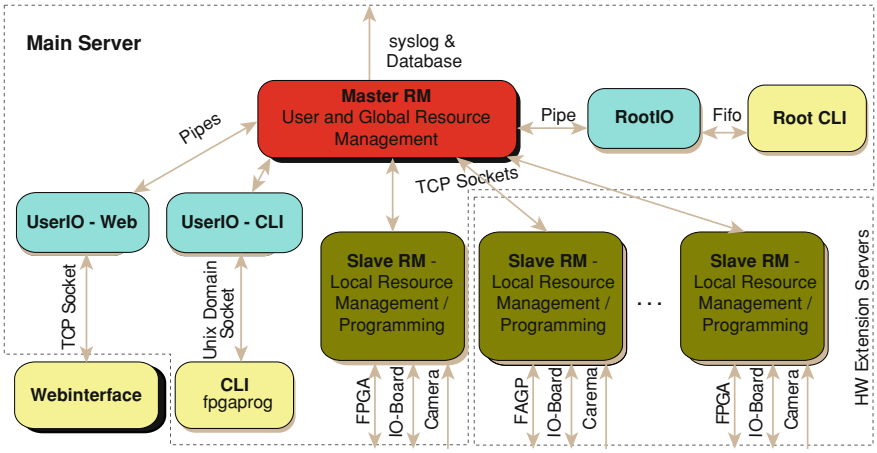


Fig. 28.6 Schematic overview about the FPGA lab system

available for an automated setup, stable and secure operation and easy to use maintenance of such a system. Therefore, we had to develop our own tool set to handle these difficulties.

The FPGA is integrated in a board which contains several IO components, such as LEDs, 7-segment display, switches and buttons. These boards are connected to a PC, which is again connected to the internet, which is called *lab client*. Finally, some of such lab clients can be connected together via the internet to offer more FPGA resource, however, with the advantages of a single sign on and a centralized user database. Hence, the students can connect to this system and program a board, with their own hardware. Figure 28.5a shows a lab client, containing the PC at the top, and the FPGA boards with webcams and IO boards at the bottom. The problem is, that via remote access the visual outputs can not be seen and no physically movements of switches and buttons are possible. To solve these two problems, every FPGA board is captured by a webcam and connected to a so called IO Board. The captured video data is send via the internet to the students at home. To stimulate the FPGAs inputs, the FPGA board is connected via electrical wires to an IO Board, which runs a web server. By accessing the web server the wires are stimulated by the IO board and the FPGA can get the inputs. In Fig. 28.5 the user interface with video stream (28.5b) and IO control (28.5c) is shown.

To get the described system working, we developed a software system which controls the programming of the FPGAs, the capturing of the boards via webcams and the access to the IO boards for FPGA control. Also the system should fulfill the requirements of high security, high stability, good performance, fairness of usage, ability to system monitoring and easy scalability. All these requirements could be met by implementing a software architecture containing daemons and end user programs shown in Fig. 28.6. Every green blob figures out a daemon (Slave Resource Manager) at the PC of a lab client. It handles and controls the resources

which are attached locally. The arrows at the bottom show the control signals of the components, where the PC is connected to. All of these *Slave Resource Managers* are connected via TCP sockets to a *Master Resource Manager*, also a daemon, which is located at an arbitrary server and is called *lab server*. The Master Resource Manager handles all available resources of all Slave Resource Managers. Thereby, the Slave Resource Managers can physically placed around the globe. For example, we have tested an inter-continental connection from Argentina to Germany with moderate latencies. There are some ways, you can interact with the Master Resource Manager. The simplest way is a ssh connection to the server where the Master Resource Manager is located. With a special program (fpgaprog) and an additional security layer you can request FPGA resources to test your hardware, analyze the video stream and make inputs utilizing the IO board. For root access an other program (Root CLI) was developed to monitor and maintain the system. A third interface, a web GUI is still in development and not finished yet. This own development was mainly written in C for the daemons and CLIs. For Database access SQL was used. Installing new and maintaining existing systems is mainly done by using bash scripts. Every connection via the TCP sockets uses our own protocol for remote hardware experiments and is encrypted by using SSL.

Coming back to the aforementioned requirements in the previous paragraph, we can say that our system is highly secure and stable. Where it is possible, the rights management system of the underlying Linux kernel for our security concept is used. Moreover, the system is running for one year now, without any system crash during this time. Also we could determine that our system is fast, because of the application-specific C implementation. The fairness of usage is given by a scheduler, which allows every user access to a board by assigning time slots, if necessary. Because of storing every event of the daemons in a database, and every important event to *syslog*, its very easy to monitor the system. Finally, the usage of TCP sockets with SSL between the daemons, allows a distributed and secure system around the globe. By using the install scripts, a new lab client can be set up easily and integrated in the existing flexible system infrastructure.

28.6 Summary

We have presented a tool centric approach to collaborative system management. It draws ideas from the DevOps movement to transform administration for the most part into writing source code, which can be shared, reviewed and developed in a team. Its heart is a set of scripts written for the Puppet configuration management system. The Puppet server and clients facilitate the communication of the configuration catalog among the nodes, while Mercurial as a revision control system and Trac are used to share and document changes among the developers/operators and users. This approach is flexible enough not only for managing hardware, but also information in an eLearning environment. The beauty of this approach is that it allows us to achieve a high degree of automation, thereby

removing the need for dedicated admins. Simultaneously it makes the process of administration scalable and repeatable.

References

1. Childs H, Brugger ES, Bonnell KS, Meredith JS, Miller M, Whitlock BJ, Max N (2005) A contract-based system for large data visualization. In: Proceedings of IEEE Visualization, pp 190–198
2. Delaet T, Joosen W (2007) Podim: a language for high-level configuration management. In: Proceedings of the 21st conference on Large Installation System Administration Conference, Berkeley, CA, USA, USENIX Association, pp 21:1–21:13
3. Forum, Message Passing Interface, editor (2009) MPI: A Message-Passing Interface Standard—Version 2.2. High-Performance Computing Center Stuttgart, Stuttgart, Germany
4. Gabriel E et al (2004) Open MPI: goals, concept, and design of a next generation MPIimplementation. In: Proceedings of the11th European PVM/MPI Users' Group Meeting. Budapest, Hungary, pp 97–104
5. Hunt A, Thomas D (1999) The pragmatic programmer: from journeyman to master. Addison-Wesley, Boston
6. Oracle (2011). Oracle Grid Engine. http://www.oracle.com/us/products/tools/oracle-grid-engine-075549.html
7. Reichenbach M, Schmidt M, Pfundt B, Fey D (2011) A new virtual hardware laboratory for remote fpga experiments on real hardware. In: Proceedings of the 2011 international conference on e-Learning, e-Business, enterprise information systems, e-Goverment, EEE '11
8. Schaefer A, Reichenbach M, Fey D (2011) In: Proceedings of the world congress on engineering and computer science 2011, vol I of 1, chapter collaborative administration in the context of research computing systems, pp 1092–1097, Newswood Limited
9. Vanbrabant B, Delaet T (2010) Authorizing and directing configuration updates in contemporary it infrastructures. In: Proceedings of the 3rd ACM workshop on assurable and usable security configuration, SafeConfig '10, New York, NY, USA, ACM, pp 79–82

Chapter 29
Verification of Virtual Prototypes of Mining Machines for Technical Criterion

Jarosław Tokarczyk

Abstract Methods for verification of virtual prototypes of powered roof support and falling object protective structure (FOPS) for operator for technical criterion were presented in the paper. In the case of powered roof support resistance strain gauge measurements and geometrical measurements were used. Method of building of computational models and their further modifications for the purpose of comparison of results was presented. Potential reasons of differences between results of stand tests and results of virtual tests were showed. Virtual prototype of FOPS protective structure was verified with use of reverse engineering method (RE).

Keywords Computational methods · Experimental test · Finite element method · Reverse · Engineering · Stand test · Virtual prototyping

29.1 Introduction

Each machine or equipment in the Polish hard coal mining industry before its actual use should be subjected to a series of obligatory stand tests, on the basis of which Notified Body makes assessment of product conformity.

Depending on a type and range of use of a given machine or equipment, there is a different range of required tests. Each machine designed for operation in underground mining industry, e.g. machine that belongs to so-called longwall system, is subjected to strict strength tests at the stage of product certification.

J. Tokarczyk (✉)
Institute of Mining Technology KOMAG, ul. Pszczyńska 37, 44-101 Gliwice, Poland
e-mail: jtokarczyk@komag.eu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_29,

Most of the tests belong to non-destructive tests, although after which a given machine can not be used. In the case of not meeting the assumed requirements it is necessary to manufacture the next copy of the material prototype. Due to this, at present before final manufacturing of machine or equipment, which is designed for experimental tests, virtual prototyping is applied. Virtual tests, which are conducted in this way at the KOMAG Institute of Mining Technology (Poland), minimize a risk of not meeting the assumed requirements during conducting of experimental tests. On the other hand, results of experimental tests are the basis for verification of virtual prototypes. Due to higher and higher possibility of complication of virtual prototype, which results from increasing computational power of present computers and development of the next versions of software of CAE (Computer Aided Engineering) class, it is possible to create complex finite element meshes of computational models and to simulate wide range of physical phenomena [1]. However, it requires continuous verification of virtual prototypes with the results of stand tests. Verification of virtual prototype for strength criterion [2] which belongs to the technical assessment criteria [3], and it needs using the following measuring methods: resistance strain gauge measurements, geometrical measurements and reverse engineering method (RE), was presented in the paper on the basis of powered roof support and FOPS.

29.2 Powered Roof Support

Powered roof supports have to protect workers against roof fall during underground mining of hard coal by a longwall system, which is especially popular in Poland and in Europe [4]. Powered roof supports are a part of longwall system, including also longwall shearer with armored face conveyor (AFC).

Main components of powered roof support are as follows, Fig. 29.1:

- Base (1).
- Canopy (2).
- Gob shield (3).
- Hydraulic legs (4).
- Lemniscate links (5).

Links with gob shield make the lemniscate system, which ensures rectilinear, vertical movement of canopy in a required range of height of support.

29.2.1 Virtual Prototyping

On the basis of 3D geometrical model a solid computational model of powered roof support was created. Boundary conditions of computational model were in accordance with support scheme A.1.1a—A.2a, according to [5], Table 29.1. It is

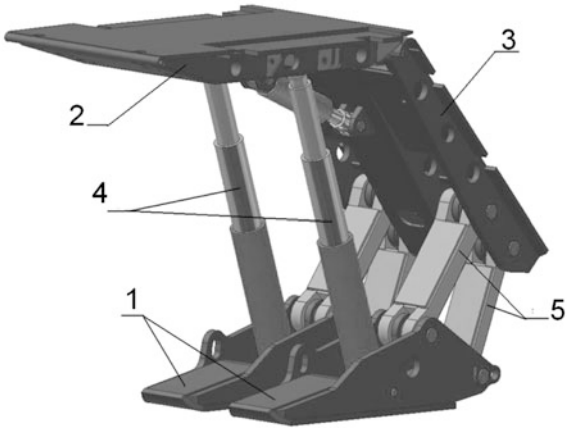


Fig. 29.1 3D geometrical model of powered roof support

Table 29.1 Components of scenario of virtual prototyping

Type of supporting	Scheme of supporting	
	Canopy	Base
A.1.1a–A.2a		

required to create a computational model consisting of solid elements, with use of a tool for automatic creation of meshes of spatial elements, to include construction details in a computational model at so complex geometrical model.

Due to this, computational model consisted of the following:

- 716081 nodes.
- 396443 TET10 solid elements.
- 48 MPC elements (articulated joints).
- 4 MPC elements (supports).
- 24 BEAM 1D elements (bolts).
- 38 BEAM 1D elements (virtual strain gauges).

Linear-and-elastic model of material of the following properties was assumed:

- Young's modulus—205 [GPa],
- Poisson's ratio—0.3.
- Density—7850 [kg/m^3].

According to passive load, forces coming from front hydraulic legs F_1, and rear hydraulic legs F_2, Fig. 29.2.

Values of forces were calculated on the basis of pressure in under-piston areas of hydraulic supports during stand tests. Forces coming from masses of front legs

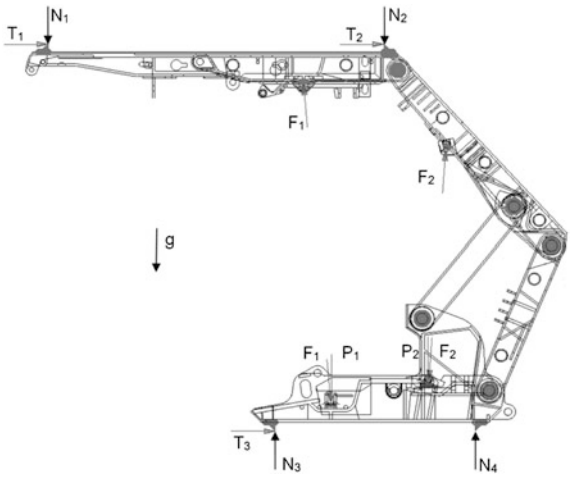


Fig. 29.2 Boundary conditions of computational model of powered roof support

m_1, rear legs m_2, canopy m_3 and gob shield m_4 acted on the base. Frictional forces between supporting beams and roof T_1, T_2 and supporting beam and floor T_3 were determined in supports, in which movement along OX axis is possible. Vectors of frictional forces were directed in opposite direction to the movement of a given support. Calculation of frictional forces (base—floor, canopy—roof) required making of initial calculations for each load variant.

Values of active forces, which are in computational model of exemplary support were as follows:

- $F_1 = 1.72$ [MN].
- $F_2 = 1.22$ [MN].
- $T_{1,2,3} = N_{1,2,3} * \mu$.
- $P_1 = m_1*g + m_3*g$.
- $P_2 = m_2*g + m_4*g$.

Where:

- $m_1 = 615$ [kg].
- $m_2 = 368$ [kg].
- $m_3 = 5270$ [kg].
- $m_4 = 4030$ [kg].
- $m_4 = 4030$ [kg].
- $\mu = 0.1$—steel—steel friction coefficient.
- N—calculated values of reactions in points of support. These values depend on a variant of support of powered roof support model.
- g—direction of gravity.

Additionally, virtual strain gauges (measuring points of strains) were created in the computational model. These were beam elements of BEAM2 type of radius

equal to 10^{-6} [m].They had common nodes with TET10 solid elements to obtain identical values of deformations.

Beam elements were placed in accordance with the placement of measuring strain gauges on a real object.

29.2.2 Experimental Test

According to the European standard requirements [5] within strength criterion, the tests of powered roof supports include a cycle of static and fatigue loads at the test stand prepared specially for that purpose. Such a stand is at the KOMAG Institute of Mining Technology.

Stand tests, by suitable methods of loading of powered roof support, i.e. methods of supporting of canopy and bases, recreate real conditions of mining in laboratory conditions in a simplified way. According to standard assumptions, different methods of supporting of powered roof support, which are applied during static tests, can be distinguished. Selection of proper methods of supporting of the system depends on a structure of tested powered roof support. Passive load of support is realized by supplying the hydraulic components, which decide about load bearing capacity of support, i.e. hydraulic legs, canopy cylinder. Active load of the support system is realized by acting of stand roof on the support with simultaneous control of increase of pressure in hydraulic components of support. Both methods enable gradual increase of support load.

Applied schedule of loading of powered roof support include symmetrical and asymmetrical methods of roof supporting.

During static stand tests of powered roof support the measurements of the following amounts, which characterize load condition of support, are made:

- Values of strains in selected points of support sub-systems.
- Deflections (deformations) of basic support sub-systems (bases, gob shield, canopy).
- Pressure in operational spaces of hydraulic components (legs, canopy cylinder).

29.2.3 Comparative Analysis of Results of Calculations and Measurements

Observed differences between results of stand tests and FEM calculations are caused mainly by arrangement of measuring and virtual strain gauges close to the places where there is a high gradient of material deformation. Moreover, discretization errors may additionally occur in those places. Such errors appear in

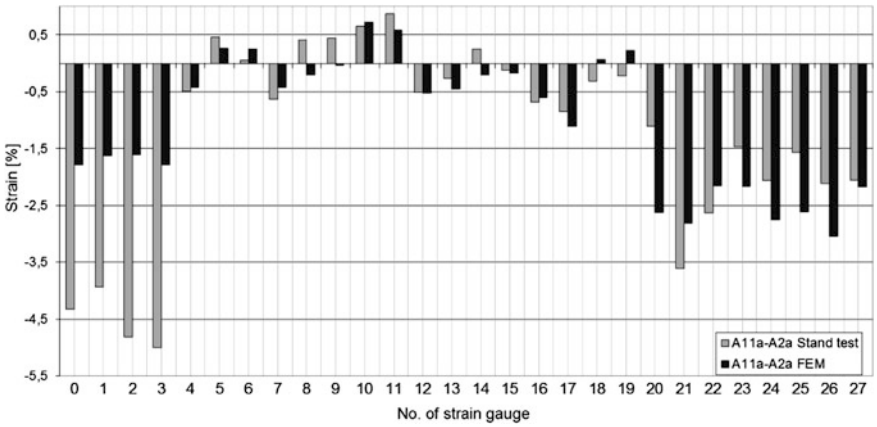


Fig. 29.3 Comparison of values of strains in measuring points for A.1.1a–A.2a support type

finite elements, geometrical form of which is degenerated, i.e. in flattened or shapeless elements. In the case of geometrically complex computational models degenerated elements are in roundings of small radius, bevelling of edges, precise mapping of screws in holes, notches, etc. Discretization error appears by high gradient of resulted values between each node of a single degenerated finite element.

Elimination of discretization errors requires re-making of FEM calculations within global-and-local task, which is solved in this model area where such error occurred. Solving of the task begins with densification of FEM mesh in a given part of tested computational model. Moreover, geometrical model can not have the features, which make creation of consistent finite elements mesh impossible. It is difficult to identify them, especially in expanded geometrical models, such as: installation clearances, edge adhesion of metal sheets and tangential connection of cylindrical surface and flat surface.

29.2.3.1 Global Task

Global task includes computational model, which comprises all sub-systems of analyzed machine or equipment. Tensors of axial deformation of measuring and virtual strain gauges were compared. Obtained results are presented in a form of diagram, Fig. 29.3.

Value of deformation of measuring strain gauge and beam component is presented on a diagram on OY axle. Numbers of measuring strain gauges are on OX axle.

Differences between values of strains of virtual and real strain gauges result from the following reasons:

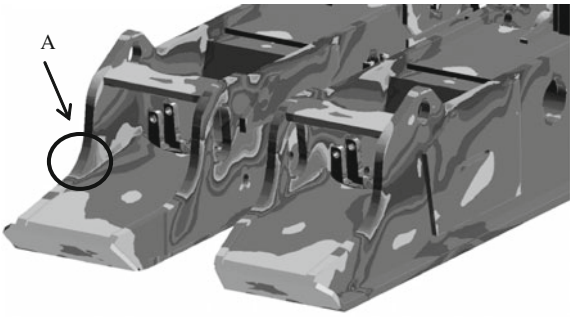


Fig. 29.4 Map of reduced stresses on a base, isometric view

- Strain gauges are placed in areas of high gradient of reduced stresses, where even a small change of placement of measuring point causes a big difference in the results (Fig. 29.4 area A).
- There are differences between placement of real and virtual strain gauges. These differences result form uncertainty of measurement of placement of strain gauges on a powered roof support.
- Computational model is made strictly according to the documentation, while in a real object there are inaccuracies associated with tolerance of manufacturing and clearances in bolt connections. These factors can disturb symmetrical operation of powered roof support.
- There are differences in homogeneity of real material.
- Non-linear behavior of material after exceeding of yield point was not included.
- At test stand the roof is supported by 12 cylinders—there is a possibility of temporary non-parallelism of roof and base.

Comparison of deformations (deflections) of the whole systems of powered roof support is the other method for verification of virtual prototype for strength criterion. Values of deflections of base and canopy are measured on their side edge. Deflection is a relative value between vertical displacement of canopy ends or base ends and vertical displacement of the middle part of the edge. During stand tests location of measuring point at the edge of base or canopy is initially estimated and corrected during strength tests. To obtain maximal value of deflection for a given component it is required to determine deflection diagram.

Comparison of differences between maximal deflections of canopy and base obtained from calculations and stand tests is included in Table 29.2.

Results obtained from FEM calculations were modified to read the value of gob shield deflection. It was necessary because of movement and rotation of gob shield, in the case of each method of supporting of powered roof support, what made direct reading of deflection impossible.

Due to that, application started in the environment of AutoCad software was created for that purpose. There were the following input data:

- Deformed, surface FEM mesh of gob shield upper surface.

Table 29.2 Components of scenario of virtual prototyping

Supporting variant	Component	Deflection value [mm]		Difference [%]
		Stand tests	FEM calculations	
A.1.1a–A.2a	Base	5.4	5.11	5.37
	Canopy	17.6	20.81	18.18

Table 29.3 Comparison of values of gob shield deflections for exemplary powered roof support

Type of supporting	Value of gob shield deflection [mm]		Difference [%]
	Stand tests	FEM calculations	
A.1.1a–A.2a	2.7	2.13	21.11

- Coordinates of reference node—the node was created in the middle of diagonal of gob shield upper surface and its location was in accordance with measuring point at test stand.

Deformed FEM mesh and coordinates of node were exported in a file of Patran Neutral File format. The results are shown in the Table 29.3.

Difference between deflection value that was calculated and deflection value obtained on the basis of stand tests was observed. Clearances present in bolt connections and tolerance of manufacturing of each part were not included in computational model. Average value of deflection obtained from stand tests is equal to 2.05 [mm], while value of clearance in bolt connection is equal to 1–2 [mm] between diameter of hole and diameter of shaft, due to what they have an impact on obtained results [6].

Insensitivity to discretization errors is an advantage of the method of verification by comparison of deflections of each sub-system. Additionally, value of deflection of a given system is a determinant of its global exerting and this result is not disturbed by local differences in values of stresses, what happens in the case of strain gauge measurements.

29.2.3.2 Local Task

A part of computational model, which requires local densification of the mesh and more accurate representation of design form, is extracted from the whole, Fig. 29.5 and treated as a separate task, including the boundary constrains that connect it with the surrounding. These constrains are represented by displacements. In the latest versions of computational software local task is a part of global task and it is solved simultaneously.

For the task needs elastic-and-plastic model of material with linear strengthening of the following properties corresponding to material, which was used in powered roof support, was defined:

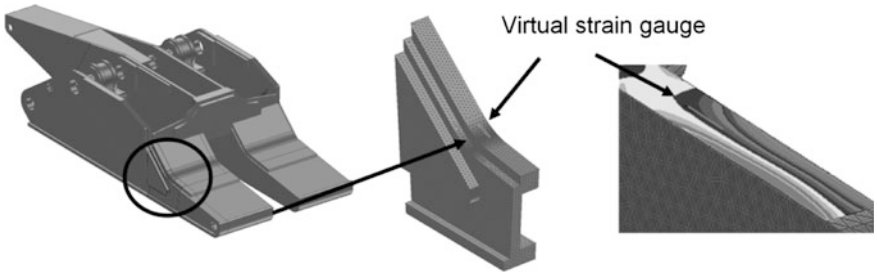


Fig. 29.5 Point of measurement of deflection at test stand

Table 29.4 Ranges of fields of deformations around each elongation measuring point

No. of measuring strain gauge	Elongation of measuring strain gauge (stand tests) [%]	Range of field of deformation around measuring strain gauge [%]
4	−0.49	−0.1 ÷ 1.75
5	0.46	−5.9 ÷ 7.45
8	0.41	−1.96 ÷ 0.488
14	0.26	−0.426 ÷ 0.2
16	−0.68	−1.14 ÷ −0.36
17	−0.85	−2.95 ÷ −0.494
18	−0.31	−1.66 ÷ 2.6
19	−0.22	−2.00 ÷ 1.52
20	−1.11	−4.77 ÷ 0.733

- Young's modulus—205 [GPa].
- Poisson's ratio—0.3.
- Density—7850 [kg/m^3].
- Yield point—690 [MPa].
- Strength—930 [MPa].
- Elongation—10 %.

Due to the time of calculations, material of these properties was assigned only to the small part of local model, in which exceeding of yield point should have been expected. Linear-and-elastic properties were assigned to the rest part of the model.

Selected parts of the model were covered with the mesh of TRIA6 surface elements of the thickness equal to 10^{-6} [m]. This mesh is stretched in a local coordinate system, in which one of the axles is in accordance with a direction of virtual strain gauge. Elements of the mesh had joint nodes with TET10 elements of solid model, what caused identical deformations.

Ranges of fields of deformations close to strain gauges were obtained in a result of solving of global-and-local tasks with use of surface mesh. Comparison of values of deformations of measuring strain gauges with the ranges of fields of deformations obtained for surface elements is given in Table 29.4. "−"sign means

compression. The ranges in perpendicular direction to longitudinal axis of assessed strain gauge were compared.

Obtained results confirm sensitivity of result of deformation to change of strain gauge location. Due to the above, use of several independent methods of verification is required to verify virtual prototype with the results of stand tests and to obtain objective comparison. In some cases it is indispensable to use RE method.

29.3 Falling Object Protective Structure

Falling Object Protective Structure (FOPS) and roll-over protective structures (ROPS) types, which protect operators of self-propelled vehicles are used in the mining industry. FOPS protects the operator against falling rock slides, while ROPS protects against crushing during vehicle overturn. These structures are subjected to destructive tests, according to standards, which are in force [7, 8]. ROPS are also used in general automotive industry [9] and in special vehicles. The results of virtual prototyping of protective structure for the operator of side-discharge loader are verification subject. Verification for strength criterion was conducted, i.e. protective structure was loaded by falling weight.

29.3.1 Virtual Prototyping

FOPS virtual prototyping for safety criterion was conducted by FEM method in MSC.Dytran software environment [10]. Non-linear, time-dependent calculations were carried out. Nonlinearities resulted from taking into account contact phenomena and elastic-and-plastic phenomena of material model. Computational model for the weight and FOPS protective structure was developed. A model of weight of 520 [kg] consisted of 6074 TETRA 4 solid elements. Initial distance between the weight and upper surface of FOPS was equal to 1 [mm]. A model of weight had initial speed equal to 6.67 [m/s], what corresponds to free fall from the height of 2.3 [m] from upper surface of of FOPS, to obtain impact energy of 11600 [J]. The following simplifications were assumed, basing on experience from previously realized work [11], [Praca badawcza U/BDC-8643/OR: Obliczenia wytrzymałościowe konstrukcji chroniącej operatora przed spadającymi przedmiotami. CMG KOMAG, Gliwice 2003 (in Polish)(unpublished)],[Praca badawcza U/BDM—8752/OR: Badanie daszka ochronnego. CMG KOMAG, Gliwice 2004 (in Polish) (unpublished)]:

- Computational model consisted of QUAD 4 and TRIA 3 shell elements.
- Grooved joints of sheets were replaced by finite elements of thickness equal to total thickness of sheets in a given joint.

The following material properties were defined:

- Model of material: elastic-and-plastic,
- Young's modulus $E = 2.068*10^{11}$ [Pa].
- Poisson's ratio $\eta = 0.29$.
- Yield point $Re_{min} = 330$ [MPa].
- Tensile strength $Rm = 490 \div 630$ [MPa].

The same supporting method as for the object at test stand: rigid anchoring at the lower edges of vertical supports, was assumed for the computational model.

Calculation process was stopped, when the model of weight reached the speed equal to 0 [m/s], i.e. when the verified construction took over the whole impact energy (maximal values of displacements and stresses were reached). Values of displacements and stresses present in FOPS structure were obtained in a result of calculations. Displacements of node on the surface of lower sheathing of FOPS, over operator's head, were assumed as criterial ones.

Maximal deflection (elastic and plastic strain) of lower surface was equal to 10 [mm] and it did not exceed the value of 50 [mm] determined in a Standard as permissible value. Maximal deflection was equal to 41 [mm] in the place of weight drop after 0.011 [s] from the time of drop.

29.3.2 Experimental Test

Conducted laboratory tests aimed at checking of protective structure for operator of loader against local puncture caused by impact load and indirectly by a method for verification of ability of transferring of impact load [Praca badawcza niepublikowana SP/BDM—9877: Badania weryfikacyjne rzeczywistej konstrukcji chroniącej operatora. CMG KOMAG, Gliwice 2005 (in Polish) (unpublished)]. The tests were conducted according to the Standard requirements [7]. Under lower sheathing of structure, at the place of operator's seat a DLV (Deflection—Limiting Volume) model of space of dimensions determined in the Standard [12], which can not be affected by deformable parts of the structure and by the weight itself, is placed. View of test stand and DLV dummy are presented in Fig. 29.6.

The weight was hanged on a crane hook. The outline of the weight was marked with a line to improve clarity of the picture. According to the Standard the energy of weight impact should be equal to 11.600 [J]. It requires lowering of the weight of 520 [kg] from the height $H = 2.3$ [m]. Microexplosive, which rapidly breaks hoisting rope of the weight, is used for that purpose.

29.3.3 Comparison of Results

A reconstruction of upper sheathing after stand tests was conducted by RE method [13] to verify virtual prototype of FOPS structure, Fig. 29.7.

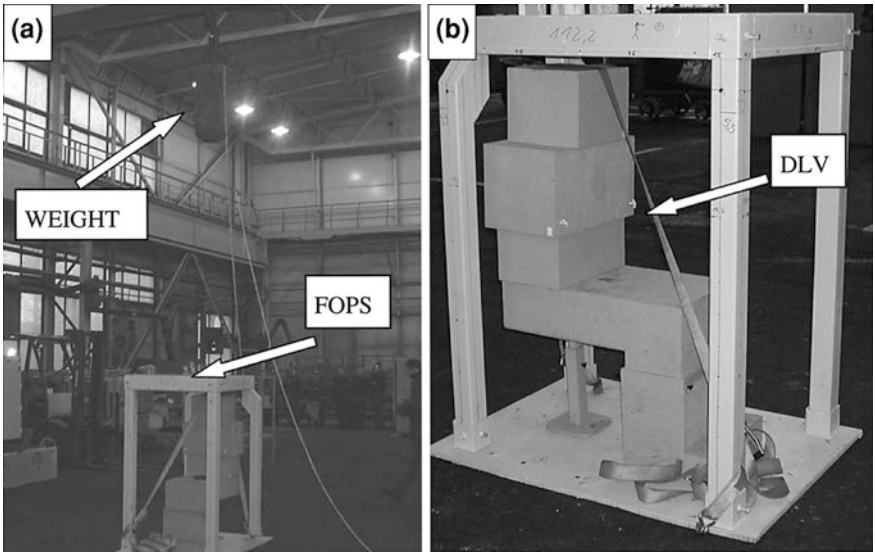


Fig. 29.6 View of test stand (**a**) and DLV dummy (**b**)

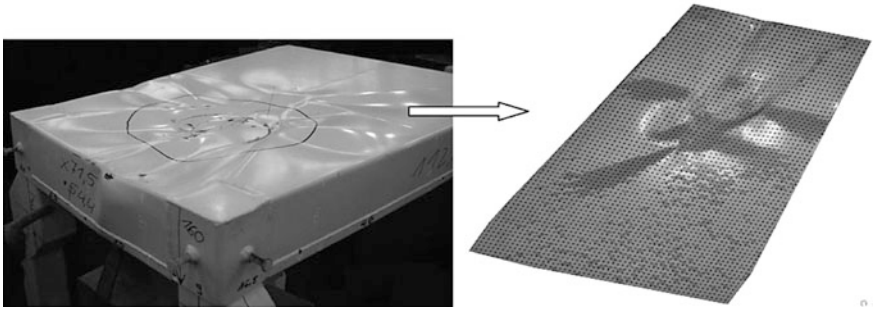


Fig. 29.7 Surface model of damaged protective structure—RE method

This method is used for reconstruction of real objects in computer environment. Coordinate measuring machines, optical methods or laser scanning are used in RE method.

At the same time initial positions of nodes of finite elements mesh was modified in FEM post-processor on the basis of magnitudes and directions of displacement vectors, obtained from deformed upper sheathing of the same FOPS structure, Fig. 29.8. It is one of functions of present post-processors [14].

Shell models obtained form different sources were set together to compare displacements of 25 points located in the same places. The were some differences in displacement (point 4, Fig. 29.9) values as a result from repeated bounces of the weight during stand tests. These phenomena were not included in computational

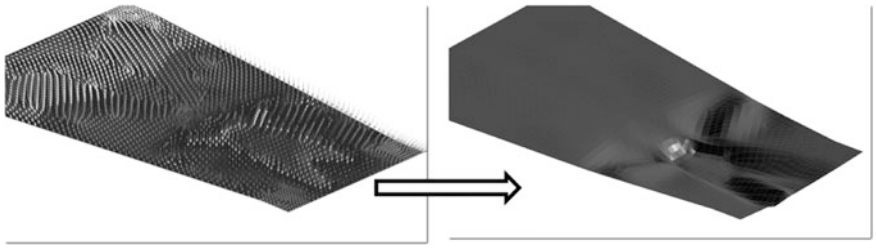


Fig. 29.8 Modification of positions of nodes of computational model mesh—result of simulation

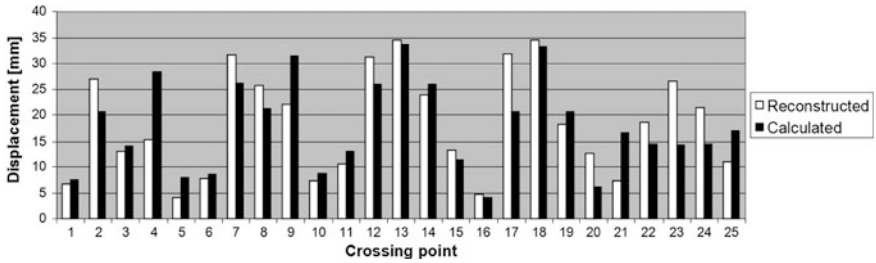


Fig. 29.9 Presentation of values of displacements of crossing points 1–25

model, due to the time of calculations. Use of RE method enabled verification of calculations results by comparison of two geometrical models obtained from two different sources.

This method is especially useful in the case of damaged objects of irregular shape, where use of traditional measuring methods is difficult.

29.4 Summary

Verification of the virtual prototype enables to identify the reasons of differences between results obtained during simulation and tests on a test stand. Verified virtual prototype enables to assess future machine or equipment in conditions, which can not be obtained at test stand or which are difficult to be obtained. High cost of stand tests, especially in the case of destructive tests in a unit production increases the final price of product that is sold. Additionally, it is required to repeat the test in the case of its negative result, what needs manufacturing of the next material prototype.

Work as regards virtual prototyping for strength criterion has been conducted at the KOMAG Institute of Mining Technology for more than 10 years. In this time a lot of research work on the basis of which the most significant reasons of differences between results obtained from simulation and stand tests was carried out.

The following factors have the biggest impact on differences between results obtained from virtual prototyping and stand tests:

- Differences between dimensions of geometrical model and real object.
- Boundary conditions.
- Material properties.
- Simplifications of computational model.
- Discretization errors.

Constant and methodical process of verification of virtual prototype enables to develop a method for creation of computational model, on the basis of which results of acceptable error are obtained for identified loading conditions. In this way in the next step it will be possible to use computational methods for certification of such products, which at present require obtaining of positive results of strength tests at test stand. Taking into account increasing possibilities of CAE software, newly designed machines and equipment will be multi-criterial assessed and optimized.

References

1. Quey R, Dawson PR, Barbe F (2011) Large-scale 3D random polycrystals for the finite element method:Generation, meshing and remeshing. Comput Methods Appl Mech Eng 200:1729–1745
2. Tokarczyk J (2011) methods for verification of virtual prototypes of mining machines for strength criterion, Lecture Notes in Engineering and Computer Science. In: Proceedings of The World Congress on Engineering and Computer Science WCECS 2011, 19–21 October, 2011, San Francisco, USA, pp 1106–1112
3. Winkler T, Tokarczyk J (2010) Multi-criteria assessment of virtual prototypes of mining machines. In: Proceedings: WCECS 2010, World Congress on Engineering and Computer Science, vol 2, San Francisco, USA, 20–22 October, 2010 pp. 1149–1153
4. Scientific and business news and events from the Polish and the world of mining www.teberia.pl
5. PN-EN 1804-1 + A1 (2011) Machines for underground mines. Safety requirements for hydraulic powered roof supports. Support units and general requirements. Polish Standard. (in Polish)
6. Ober G (2001) Wpływ luzów w prototypowych sekcjach obudów zmechanizowanych na tor końca stropnicy. Maszyny Górnicze nr 85/2001, pp. 50–54. (in Polish)
7. PN-EN 3449 (2009) Earth-moving machinery—Falling-object protective structures—Laboratory tests and performance requirements. Polish Standard. (in Polish)
8. PN-92/G-59001 Samojezdne maszyny górnicze. Konstrukcje chroniące operatora przed obwałami skał. Wymagania i badania. Polish Standard. (in Polish)
9. Barszcz Z (2006) Ocena bezpieczeństwa dużych pojazdów do przewozu osób w zakresie wytrzymałości ich konstrukcji nośnej. Konferencja: Spotkanie użytkowników oprogramowania MSC. Mszczonów (in Polish)
10. Dytran MSC 2008r1: Theory manual. MSC.Software Corporation
11. Bojara S, Całus Ł, Tokarczyk J (2003) Modelowanie wybranych zjawisk dynamicznych na przykładzie struktur ochronnych maszyn samojezdnych. Materiały na konferencję

KOMTECH: Nowoczesne, niezawodne i bezpieczne systemy mechaniczne w świetle wymagań Unii Europejskiej. Szczyrk (in Polish)
12. PN-EN ISO 3164 (2009) Earth-moving machinery—Laboratory evaluations of protective structures—Specifications for deflection-limiting volume. Polish Standard. (in Polish)
13. Winkler T, Tokarczyk J, Bojara S (2008) Use of reverse engineering method in verification of virtual prototypes. Computer assisted mechanics and engineering sciences 1:53–65
14. Patran® (2010) Finite element modeling. MSC.Software Corporation

Chapter 30
Project Scheduling with Fuzzy Cost and Schedule Buffers

Pawel Blaszczyk, Tomasz Blaszczyk and Maria B. Kania

Abstract The aim of this research was the trial of modelling and optimizing the time-cost trade-offs in project planning problem with taking into account the behavioral impact of performers' (or subcontractors') estimations of basic activity parameters. However, such a model must include quantitative measurements of budget and duration, so we proposed to quantify and minimize the apprehension of their underestimations. The base of the problem description contains both safe and reasonable amounts of work estimations and the influence factors matrix. We assumed also the pricing opportunity of performance improving. Finally we introduce fuzzy measurements for work amount. This paper is a revised, extended version of Blaszczyk et al. 2011, presented on the World Congress on Engineering and Computer Science 2011.

Keywords Buffer management · Project planning · Time-cost trade-off · Fuzzy numbers · Scheduling · Fuzzy linear programming

P. Blaszczyk (✉) · M. B. Kania
Institute of Mathematics, University of Silesia, Katowice, Poland
e-mail: pblaszcz@math.us.edu.pl

M. B. Kania
e-mail: mkania@math.us.edu.pl

T. Blaszczyk
Department of Operations Research, University of Economics in Katowice, Katowice, Poland
e-mail: blaszczyk@ue.katowice.pl

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 170, DOI: 10.1007/978-94-007-4786-9_30,

30.1 Introduction

The time-cost trade-off analysis, allowing for establishment of such a project plan which satisfies the decision-maker's expectations for the soonest completion date with as low budget as possible, is one of the basic multicriterial problems in project planning. The first researches in this subject, conducted by [8, 13], have been publicated in 1960's. Precise reviews of temporary results were widely described by several authors, for instance by [5]. The aim of the follow-ing research is to consider the critical chain approach described by [9] in multiple-criteria environment. The primal description of the method was based on verbal language, rather than formal. The chain and time buffers quantification methods were the results of successive authors. One of the detailed approaches was formally described by [19]. The issues of buffering some project characteristics, other than duration, were considered by [1, 10, 14]. The general critical chain approach, widely discussed by various authors (compare [11, 17, 20]), is not drawback-free. So that, the range of its practical implementation is not as wide as the regular CPM and PERT methods. However, the critical chain has an important advantage because of the behavioral aspects inclusion, what can make it more useful in the real-life planning problem descriptions. Thus, by including the impact of the human factor on measurable project features, we are capable of using it to improve these features in return for financial equivalent. An example of such a solution, with using the extraordinary premium fund, was described by [1]. The following part of the paper is the consequence of continuing this research on buffering different project features. Here we took into consideration the duration and budget expectations by modelling the project with time and cost buffers. Apart from temporary results in the described procedure, we introduced buffers on overestimated amounts of labor which are given by employees or subcontractors. In this paper we introduce fuzzy measure of amount of labor to represent the uncertainty of afford estimations. Fuzzy approach to critical chain modelling has been considered by [7, 15, 18]. The model we are proposing assumes the opportunity to motivate them to participate in the risk of delays and budget overrunning in return for probable profits, in case of faster and cheaper realization.

30.2 The First Mathematical Model: Cost and Time Buffers

We consider project which consist $x_1, \ldots, x_n$ activities characterized by cost and time criteria. We assume that only q factors has any influence on the cost and the time of the project. Let us consider the following matrix X:

$$X = \begin{bmatrix} x_{11} & \ldots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nq} \end{bmatrix} \tag{30.1}$$

Elements of the matrix X equals 0 or 1. If x_{ij} equals 1 it means that factor j has influence on the completion of activity x_i. In the other case there is no influence of factor j on activity x_i. The matrix X we will call the *factor's matrix*. Let

$$K = [k_{ij}]_{i=1,\ldots,n;j=1,\ldots,q} \tag{30.2}$$

to be the matrix of cost's ratios of all q factors for all activities and

$$W^m = [w_1^m, \ldots, w_n^m] \tag{30.3}$$

to be the vector of minimal amounts of work for the activities $x_1, \ldots, x_n$. On the basis of matrix X and vector W^m for activity x_i we can calculate the total amount of work w_i by:

$$w_i = f_{w_i}(x_{i1}, \ldots, x_{iq}, w_i^m) \tag{30.4}$$

where f_{w_i} is a work assigning function. Moreover we assume that there is vector

$$R = [r_1, \ldots, r_q] \tag{30.5}$$

describing the restrictions of accessibility of factors for whole project. Let

$$T = [t_{ij}]_{i=1,\ldots,n;j=1,\ldots,q} \tag{30.6}$$

be the matrix of amounts of work for each factor in each activity. On the basis of the matrix X, T and K we calculate the cost and the duration of each activity by:

$$k_i = f_{i_k}(x_{i1}, \ldots, x_{iq}, t_{i1}, \ldots, t_{iq}, k_{i1}, \ldots, k_{iq}) \tag{30.7}$$

and

$$t_i = f_{i_t}(x_{i1}, \ldots, x_{iq}, t_{i1}, \ldots, t_{iq}) \tag{30.8}$$

where f_{i_k} and f_{i_t} are some functions. We called this functions the *cost* and the *time* functions, respectively. Thus the total cost and the total duration of the project are given by

$$K_c = \sum_{i=1}^{n} k_i \tag{30.9}$$

and

$$T_c = \max_{i=1,\ldots,n} (ES_i + t_i) \tag{30.10}$$

where ES_i is the earliest start of activity x_i. Under the following assumptions we minimize total cost of the project. If the functions f_{i_k} and f_{i_t} are linear functions than this optimization problem can be solved by Linear Programming (LP). In typical case the linear programming model is given by

$$c \cdot x \longrightarrow \min \tag{30.11}$$

$$A \cdot x \leq b \tag{30.12}$$

$$x \geq 0 \tag{30.13}$$

where c, x, A, b are coefficient vector of object function, coefficient vector of decision variables, matrix of coefficient of restriction and vetor of absolute terms respectively. In our case we have the following linear programming model

$$\sum_{i=1}^{n} f_{i_k}(x_{i1}, \ldots, x_{iq}, t_{i1}, \ldots, t_{iq}, k_{i1}, \ldots, k_{iq}) = \sum_{i=1}^{n} k_i \longrightarrow \min \tag{30.14}$$

$$X' \cdot T \leq R \tag{30.15}$$

$$X \cdot T' = W \tag{30.16}$$

$$t_i \geq 0. \tag{30.17}$$

It leads to find the optimal work assignments for every factor in each activities. From the set of alternate optimal solutions we choose this one, for which the total duration of project is minimal. In this way we obtain the optimal solution in safe case. According to the contractors' safe estimations the amount of work could be overestimated. It leads up to overestimations of the activities' cost and duration expected values and afterwards the total cost and the total duration of the whole project. That means

$$k_i = k_i^e + k_i^B \tag{30.18}$$

and

$$t_i = t_i^e + t_i^B \tag{30.19}$$

where k_i^e, t_i^e are the reasonable cost and reasonable duration for activity x_i and k_i^B, t_i^B are the buffers of budget and time for activity x_i, respectively. Therefore we can write the total cost and total duration of project by

$$K_c = K^e + K^B \tag{30.20}$$

and

$$T_c = T^e + T^B \tag{30.21}$$

where K^e, T^e are the reasonable cost and reasonable duration of the project and K^B, T^B are the buffers of budget and time, respectively. To set the buffers K^B, T^B

up we must estimate the most probable amounts of work. We do that by changing appropriate elements x_{ij} in matrix X from 1 to 0 or vice versa. It means that some factors which had influence on activity x_i in safe estimation case does not have it in real estimation case and vice versa. Than we using the function w_i for each activity x_i. In this way we get the new factor's matrix $X^{\star}$ and the new vector of amounts of work $W^{\star}$. Then we execute the same procedure for the most probable amount of work but under additional condition $t_{ij} \geq t_{ij}^{\star}$ for $i = 1, \ldots, n; j = 1, \ldots, q$, where $T^{\star} = \left[t_{ij}^{\star}\right]$ is the matrix of amounts of work for each factor in each activity calculated for the new data. Since that is unlikelihood that all factors will occur, we can reduce the buffers for project by:

$$K_r^B = \alpha K^B \tag{30.22}$$

and

$$T_r^B = \beta T^B \tag{30.23}$$

where $\alpha, \beta \in [0, 1]$ are the ratios revising amount of buffers.

$$K^P = K^e + K_r^B \tag{30.24}$$

and

$$T^P = T^e + T_r^B \tag{30.25}$$

Part of saved money can generate bonus pool B and be divided between the factors. Let us introduce the weight of importance of activities

$$S = [s_i]_{i=1,\ldots,n}, \tag{30.26}$$

where $s_i \in [0, 1]$. To share the bonus pool we define function which depends on saved amount of work, importance of activity x_i and if the activity is critical or not and on the reduced buffers of cost and time. In the general case that factor i can receive the amount of money b_i

$$b_i = f_{b_i}(s_i, D_i^W, c, D_B^K, D_B^T) \tag{30.27}$$

where s_i is the importance of activity x_i, D_i^W is the saved amount of work for activity x_i, $c = 1$ if the activity is on critical path or $c = 0$ if it is not on the critical path, D_B^K is the amount of saved cost, D_B^T is the amount of saved time and f_{b_i} is some function. For example we can used the following function

$$b_i = \begin{cases} \frac{s_j}{s^1}\frac{D_j^W}{D^1}\gamma_1 B & \text{if} \quad x_i \quad \text{is on critical path} \\ \\ \frac{s_j}{s^2}\frac{D_j^W}{D^2}\gamma_2 B & \text{else} \end{cases} \tag{30.28}$$

where B is the bonus pool $\gamma_2 < \gamma_1$, $\gamma_1 + \gamma_2 = 1$, s^1 is the sum of importances of activities which is on critical path, s^2 is the sum of importances of activities which is not on critical path, D_j^W is the sum of saved amounts of work for activity x_i, D_i^1 is the saved amounts of work for activities which are critical and D_i^2 is the sum saved amounts of work for activities which are beside any critical path.

30.3 The Second Mathematical Model: Work Amount Buffer

In this section we discuss another mathematical model for the project considered above, which was introduced in [2]. Like in the first model we introduce factor's matrix X, matrix of cost's ratios K, vector of minimal amounts of work W^m, vector R describing the restrictions of accessibility of factors and the matrix of amounts of work for each factor in each task T (see (30.1)–(30.3), (30.5), (30.6), respectively). On the basis of the matrix X, T, K we calculate the cost and the duration of each activity using formulas (30.7) and (30.8) and then using formulas (30.9) and (30.10) the total cost and total duration of the project. Like in previous model we minimize the total cost of the project. If the functions f_{i_k} and f_{i_t} are linear than this optimization problem can be solved by LP. From the set of alternate optimal solution we choose this one, for which the total duration of project is minimal.

For task x_i the amount of work could be written as

$$w_i = f_{w_i}(x_{i1}, \ldots, x_{iq}, w_i^m) = w_i^e + w_i^B. \tag{30.29}$$

Therefore we can write the total amount of work of project by

$$W_c = W^e + W^B, \tag{30.30}$$

where W^e is the reasonable amount of work of the project and W^B is the buffer of amount of work. To set the buffer W^B up we must estimate the most probable amounts of work. We do that by changing appropriate elements in matrix X and using the function w_i for each task x_i. In this way we get the new factor's matrix X^* and the new vector of amounts of work W^*. Since that is unlikelihood that all factors will occur, we can reduce the buffer for project by:

$$W_r^B = [\alpha_1, \ldots, \alpha_n]W^B, \tag{30.31}$$

where $\alpha \in [0, 1]$ for $i \in \{1, \ldots, n\}$ are the ratios revising amount of work for tasks $x_1, \ldots, x_n$. So the total project amount of work is given by

$$W^P = W^e + W_r^B. \tag{30.32}$$

The overestimation of amount of work leads up to overestimations of the tasks' cost and duration expected values and afterwards the total cost and the total duration of the whole project. Because the amount of work changed the duration and cost of project also changed. Therefore we can write the total cost and total duration of project as in (30.20) and (30.21), respectively. Like in previous model part of saved money can generate bonus pool B and be divided between the factors. The weight of importance of tasks S is given by formula (30.26). The bonus pool for the factor i can be shared by using function (30.27). Like above we can you the function (30.28).

30.4 Fuzzy Approach

Deterministic values in Classic Linear Programming model usually does not correspond with real and uncertain conditions expected during project execution. To deal with this problem we propose extention of the models above using fuzzy approach. The proposed method use trapezoidal fuzzy numbers (TrFN). First, let us introduce some basic facts, which we use in our fuzzy model extension.

Definition 1 Let A be a subset in some space X. A fuzzy set A in X is a set of ordered pair

$$(x; \mu_A(x)) : x \in X \tag{30.33}$$

where

$$\mu_A : X \to \mathbb{R} \tag{30.34}$$

is membership function of set A.

For each $x \in A$, $\mu_A(x)$ is called the grade of membership of x in A in the literature it is often used $A(x)$ instead $\mu_A(x)$ to describe the membership function of set A. To define fuzzy number, first we must introduce some basic facts.

Definition 2 The set A is called normal if

$$h(A) = \sup_{x \in X} \mu_A(x) = 1. \tag{30.35}$$

Definition 3 The set

$$supp(A) = \{x \in A : \mu_A(x) > 0\} \tag{30.36}$$

is called the support of A.

Definition 4 Let $\alpha \in [0, 1]$. The set

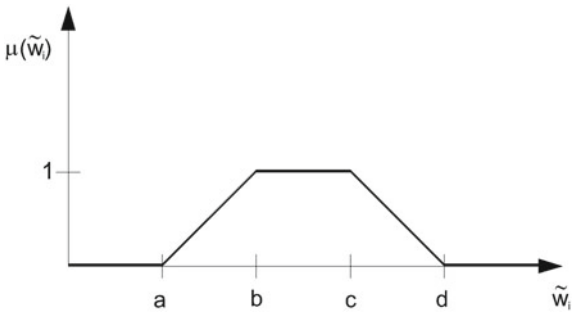


Fig. 30.1 An example of trapezoid fuzzy number (TrFN)

$$A_\alpha = \{x \in X : A(x) \geq \alpha\} \tag{30.37}$$

is called alpha cut.

Definition 5 Let $X = \mathbb{R}$. A fuzzy number is such fuzzy set $A \in F(\mathbb{R})$ which satisfy following conditions:

1. A is normal set,
2. A_α is closed for each $\alpha \in [0, 1]$,
3. $supp(A)$ is bounded,

Definition 6 The trapezoidal fuzzy numbers $TrFN(a, b, c, d)$ (see Fig. 30.1) is a fuzzy number for which the membership function is given by the following formula

$$\mu(x) = \begin{cases} (x-a)/(b-a) & \text{for} \quad x \in [a, b] \\ 1 & \text{for} \quad x \in [b, c] \\ (d-x)/(d-c) & \text{for} \quad x \in [c, d] \\ 0 & \text{for} \quad x \notin [a, d] \end{cases} \tag{30.38}$$

The membership function μ depends on expert's judgment about availability of factors, workers, materials etc.

Definition 7 Let $x \in \mathbb{R}$ and $\epsilon \in [0, 1]$ be sufficient small. The trapezoid fuzzy number $\widetilde{x}$ is called the fuzzy number close to real number x when is given by:

$$\widetilde{x} = (x - \epsilon, x, x, x + \epsilon) \tag{30.39}$$

In the hereinafter of this article we denote the fuzzy number close to real number x by $\widehat{x}$ We write that trapezoid fuzzy number $A(a, b, c, d) \geq \delta$, where δ is some real number, if $a \geq \delta, A > \delta, A \leq \delta$ for $d \leq \delta$ and $A(a, b, c, d) < \delta$ id $d < \delta$. If A, B are two fuzzy subset of set a space X, than $A \leq B$ mean that $A(x) \leq B(x)$ for all $x \in X$, or A is a subset of B, $A < B$ holds when $A(x) < B(x)$ for all x. There is a potential problem with the symbol $\leq$. In this article $A \leq B$ for fuzzy numbers A, B means that A is less than or equal to B.

Definition 8 For two fuzzy numbers the basic four arithmetic operation are given by the following formulas

$$\mu_{B=A_1\oplus A_2}(y) = \sup_{x_1,x_2\in X, y=x_1+x_2} \min\{\mu_{A_1}(x_1), \mu_{A_2}(x_2)\} \tag{30.40}$$

$$\mu_{B=A_1\ominus A_2}((y) = \sup_{x_1,x_2\in X, y=x_1-x_2} \min\{\mu_{A_1}(x_1), \mu_{A_2}(x_2)\} \tag{30.41}$$

$$\mu_{B=A_1\odot A_2}((y) = \sup_{x_1,x_2\in X, y=x_1\cdot x_2} \min\{\mu_{A_1}(x_1), \mu_{A_2}(x_2)\} \tag{30.42}$$

$$\mu_{B=A_1\oslash A_2}((y) = \sup_{x_1,x_2\in X, y=x_1/x_2} \min\{\mu_{A_1}(x_1), \mu_{A_2}(x_2)\} \tag{30.43}$$

In all above cases the result is also a fuzzy number, but not necessary trapezoid fuzzy number. In the case when objective functions and and restrictions are given by fuzzy numbers the Fuzzy Linear Programming (FLP) model is given by the following formula

$$\widetilde{c}\cdot x \longrightarrow \min \tag{30.44}$$

$$\widetilde{A}\cdot x \leq \widetilde{b} \tag{30.45}$$

$$x \geq 0 \tag{30.46}$$

where $\widetilde{c}, \widetilde{A}, \widetilde{b}$ are fuzzy coefficient vector of object function, matrix of fuzzy coefficient of restriction and vector of fuzzy numbers respectively.

Theorem 1 Let $\tilde{c_j}, \tilde{a_{ij}}$ be a fuzzy quantities. Than the fuzzy set $\tilde{c_1}x_1 + \ldots + \tilde{c_n}x_n$ and $\tilde{a_1}x_1 + \ldots + \tilde{a_n}x_n$ defined by the extension principle is again fuzzy quantity.

Detailed information about solving fuzzy linear programming can be found in [6, 12, 16].

30.5 The Third Mathematical Model: Fuzzy Work Amount and Fuzzy Buffers

In this section we discuss third mathematical model for the project considered above. This model has been primarily introduced in [4]. Like in the first two models we introduce factor's matrix X, matrix of cost's ratios K, vector of minimal amounts of work W^m, vector R describing the restrictions of accessibility of factors and the matrix of

amounts of work for each factor in each task T (see (30.1)–(30.3), (30.5), (30.6), respectively). On the basis of the matrix X, T, K we calculate the cost and the duration of each activity using formulas (30.7) and (30.8) and then using formulas (30.9) and (30.10) the total cost and total duration of the project. Like in previous model we minimize the total cost of the project. If the functions f_{i_k} and f_{i_t} are linear than this optimization problem can be solved by LP. From the set of alternate optimal solution we choose this one, for which the total duration of project is minimal. Like in the second model, for task x_i, the amount of work could be written using the formula (30.29). Therefore the total amount of work of project is given by formula (30.30). To set the buffer W^B up we must estimate the most probable amounts of work. In some cases it could be hard to estimate the amount of work for task x_i and therefore, for some tasks, it could be impossible to set up deterministic value of amount of work. To solve this problem we can use the trapezoid fuzzy numbers described in (30.38). For the safe estimation the amount of work for task x_i is given by real number. Before we estimate the real amount of work we must rewrite this real numbers as a fuzzy number close to real number using definition 7 and the formula (30.39). Now the amount of work can be written using the following formula

$$\hat{w_i} = \widetilde{w_i}^e + \widetilde{w_i}^B. \tag{30.47}$$

where $\widehat{w_i}$ is fuzzy number close to real number w_i, $\widetilde{w_i}^e$ fuzzy number describing to real estimation for work amount of the task x_i and $\widetilde{w_i}^B$ is the buffer for the work amount for the task x_i. Therefore we can write the total amount of work of project by

$$\widetilde{W_c} = \widetilde{W^e} + \widetilde{W^B}, \tag{30.48}$$

where $\widetilde{W^e}$ is the reasonable amount of work of the project and $\widetilde{W^B}$ is the buffer of amount of work. The buffer $\widetilde{W^B}$ is setting by expert's judgment about availability of factors in the matrix X. Under the following assumptions we minimize total cost of the project. If the functions f_{i_k} and f_{i_t} are linear than this optimization problem can be solved by FLP. From the set of alternate optimal solution we choose this one, for which the total duration of project is minimal. Since that is unlikelihood that all factors will occur, we can reduce the buffer for project by:

$$\widetilde{W_r^B} = [\alpha_1, \ldots, \alpha_n]\widetilde{W^B}, \tag{30.49}$$

where $\alpha \in [0, 1]$ for $i \in \{1, \ldots, n\}$ are the ratios revising amount of work for tasks $x_1, \ldots, x_n$. So the total project amount of work is given by

$$\widetilde{W^P} = \widetilde{W^e} + \widetilde{W_r^B}. \tag{30.50}$$

The overestimation of amount of work leads up to overestimations of the tasks' cost and duration expected values and afterwards the total cost and the total duration of the whole project. Because the amount of work changed the duration and cost of project also changed. Therefore we can write the total cost and total duration of project as

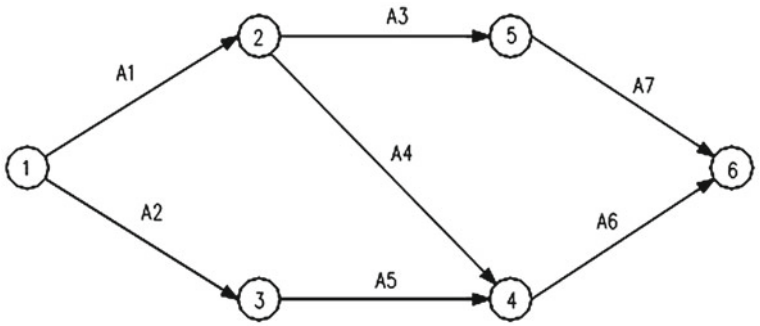


Fig. 30.2 Project diagram (Activity on Arc)

$$K_c = K^e + K^B \tag{30.51}$$

$$\widetilde{T_c} = \widetilde{T^e} + \widetilde{T^B} \tag{30.52}$$

where $K^e, \widetilde{T^e}$ are the reasonable cost and reasonable duration of the project and $K^B, \widetilde{T^B}$ are the buffers of budget and time, respectively. The $\widetilde{T^e}$ and $\widetilde{T^B}$ are trapezoid fuzzy number.

Like in previous models part of saved money can generate bonus pool B and be divided between the factors. The weight of importance of tasks S is given by formula (30.26). The bonus pool for the factor i can be shared by using function (30.27). Like above we can use the function (30.28).

30.6 Example

To illustrate the applicability of the fuzzy mathematical model we propose to consider the model of project composed of seven activities: $A_1, A_2, \ldots, A_7$ by the diagram shown on a Fig. 30.2.

Each of these activities could be realized by resources represented by number 1 in the binary influence matrix (Table 30.1).

The required work amounts to complete consecutive activities are uncertain, depends on several conditions and are estimated by fuzzy triangular numbers described in Table 30.2.

Costs of single unit of work are fixed for all resources in every activity and shown in Table 30.3. The accessibility of resources is also known and fixed (E1-600, E2-450, E3-650, E4-200, E5-350 man-hours). While the problem is defined as a need of optimal resources distribution with minimal project cost and duration we introduce planned individual employees workloads as decision variables in the model minimizing the total project cost (including indirect work costs and bonus amounts according to (30.27)) with respect to project limited resources availability and the

Table 30.1 The matrix of influence of resources on activities

Activity	Employee 1	Employee 2	Employee 3	Employee 4	Employee 5
A_1	0	0	1	1	0
A_2	1	1	0	0	0
A_3	0	0	0	0	1
A_4	0	0	1	1	0
A_5	1	1	0	0	0
A_6	0	1	0	1	0
A_7	1	0	1	0	1

Table 30.2 The triangular fuzzy estimations of work amounts

Activity	Fuzzy workload		
A_1	180	200	225
A_2	210	240	270
A_3	140	150	165
A_4	275	300	340
A_5	300	340	380
A_6	170	200	240
A_7	600	690	750

Table 30.3 The unitary cost matrix

Activity	Employee 1	Employee 2	Employee 3	Employee 4	Employee 5
A_1	50	60	33	35	28
A_2	45	50	52	48	35
A_3	32	40	30	68	43
A_4	67	55	63	57	47
A_5	55	60	45	54	51
A_6	45	42	50	50	38
A_7	50	50	45	57	49

Table 30.4 The optimal workload distribution

Activity	Employee 1	Employee 2	Employee 3	Employee 4	Employee 5
A_1	0	0	30	175	0
A_2	120	120	0	0	0
A_3	0	0	0	0	150
A_4	0	0	255	25	0
A_5	268	60	0	0	0
A_6	0	170	0	0	0
A_7	5	0	395	0	200

structure of precedence. To solve this class of problems we use implement Fuzzy Linear Programming algorith inthe Matlab enviroment. As an optimal solution of this case we achieved the values of workloads shown in Table 30.4.

Optimal values of project duration and cost (including the bonus success fees) should be obtained by scheduling the work of employee 1 in activities A_2, A_5, A_7, employee 2 in activityies A_2, A_5, A_6, employee 3 in activities A_1, A_4, A_7, employee 4 in activities A_1, A_4 and employee 5 in activities A_3, A_7.

30.7 Conclusion

According to the authors of the paper, in the project planning issues it is possible to extract the safety buffers hidden in schedule estimations. The results of prior researches indicate that this mechanism is useful in project budgeting processes. The main thesis of our study, stating the existence of required labor overestimations, seems to be justified as well. The theoretical consideration is compliant with the project cost buffering approach, described in [1] in terms of the procedure of buffers sizing and profits distributing. Another extension of the prior approach was done with including the influence matrix describing the hypothetical dependence of resources on several time and cost drivers. The introduction of fuzzy measures allowed us to improve the representation of estimations on the required amounts of labor. It has to be highlighted, though, that proving its efficiency requires further empirical study in real-life conditions. It will be a subject of the next stage of this research.

References

1. Blaszczyk T, Nowak B (2008) Project costs estimation on the basis of critical chain approach (in Polish), Trzaskalik T (ed.) Modelowanie preferencji a Ryzyko'08, Akademia Ekonomiczna w Katowicach, Katowice
2. Blaszczyk P, Blaszczyk T, Kania MB (2009) Task duration buffers or work amount buffers? In: Proceedings of the first earned value analysis conference for the continental Europe, Geneva, 2009. vol 1, pp 345–375
3. Blaszczyk P, Blaszczyk T, Kania MB (2011) The bi-criterial approach to project cost and schedule buffers sizing. Lecture notes in mathematics and economy, Springer, Berlin 2011. pp 105–114
4. Blaszczyk P, Blaszczyk T, Kania MB (2011) Theoretical foundations of fuzzy bi-criterial approach to project cost and schedule buffers sizing. In: Proceedings of the world congress on engineering and computer science WCECS 2011, San Francisco, 19–21 October 2011. Lecture notes in engineering and computer science, pp 1121–1125
5. Brucker P, Drexl A, Mohring R, Neumann K, Pesch E (1999) Resource-constrained project scheduling: notation, classification, models and methods. Eur J Oper Res 112:3–41
6. Buckey JJ, Eslami E, Feuring E (2002) Fuzzy mathematics in economy and engineering. Springer, Heidelberg
7. Chen L, Liang F, Xiaoran S, Deng Y, Wang H (2010) Fuzzy-safety-buffer approach for project buffer sizing considering the requirements from project managers and customers. The 2nd IEEE international conference on information management and engineering (ICIME), 2010. pp 482–486
8. Fulkerson DR (1961) A network flow computation for project cost curves. Manag Sci 7:167–178

9. Goldratt E (1997) Critical chain. North River Press, Great Barrington
10. Gonzalez V, Alarcon LF, Molenaar K (2009) Multiobjective design of Work-In-Process buffer for scheduling repetitive projects. Autom Constr 18:95–108
11. Herroelen W, Leus R (2009) On the merits and pitfalls of critical chain scheduling. J Oper Manag 19:559–577
12. Jamison KD, Lodwick WA (2001) Fuzzy linear programming using a penalty method. Fuzzy Sets Syst 119:97–110
13. Kelley JE (1961) Critical-path planning and scheduling: mathematical basis. Oper Res 9:296–320
14. Leach L (2003) Schedule and cost buffer sizing: how account for the bias between project performance and your model. Proj Manag J 34:34–47
15. Long LD, Ohsato A (2008) Fuzzy critical method for project schedulling under resource constraints and uncertainty. Int J Proj Manag 26:688–698
16. Ramik J (2006) Duality in fuzzy linear programming with possibility and necessity relations. Fuzzy Sets Syst 157:1283–1302
17. Rogalska M, Bozejko W, Hejducki Z (2008) Time/cost optimization using hybrid evolutionary algorithm in construction project scheduling. Autom Constr 18:24–31
18. Shi Q, Gong T (2010) An improved project buffer sizing approach to critical chain management under resources constraints and fuzzy uncertainty. International conference on artificial intelligence and computational intelligence AICI '09, November 2010. pp. 486–490
19. Tukel OI, Rom WO, Eksioglu SD (2006) An investigation of buffer sizing techniques in critical chain scheduling. Eur J Opera Res 172:401–416
20. Van de Vonder S, Demeulemeester E, Herroelen W, Leus R (2005) The use of buffers in project management: The trade-off between stability and makespan. Int J Prod Econ 97:227–240