

Simulation-Based Optimization Using Greedy Techniques and Simulated Annealing for Optimal Equipment Selection Within Print Production Environments

Sudhendu Rai, Eric Gross and Ranjit Kumar Ettam

Abstract Xerox has invented, tested, and implemented a novel class of operations-research-based productivity improvement offerings, marketed as Lean Document Production (LDP), for the \$100 billion printing industry in the United States. The software toolkit that enables the optimization of print shops is data-driven and simulation-based. It enables quick modeling of complex print production environments under the cellular production framework. The software toolkit automates several steps of the modeling process by taking declarative inputs from the end user and then automatically generating complex simulation models that are used to determine improved design and operating policies. This chapter describes the addition of another layer of automation consisting of simulation-based optimization using simulated annealing and greedy search techniques that enable the search of a large number of design alternatives in the presence of operational and cost constraints. The greedy search procedure quickly determines an acceptable solution in a web-based online application environment. The simulated annealing technique is more time consuming and is performed offline. The results of the application of this approach to real-world problems are described.

1 Introduction

Xerox is the world's leading enterprise for business process and document management solutions. Xerox produces and sells a range of color and black-and-white printers, multifunction systems, photocopiers, digital production printing presses, and related consulting services and supplies. Xerox participates in the printing industry by providing services, via Xerox Managed Services (XMS), to manage print

S. Rai (✉) · E. Gross · R.K. Ettam
Xerox Corporation, 800 Phillips Road, Webster, NY 14450, USA
e-mail: Sudhendu.Rai@xerox.com

E. Gross
e-mail: Eric.Gross@xerox.com

R.K. Ettam
e-mail: Ranjit.Kumar2@xerox.com



Fig. 1 A print production workflow showing the various production operations

operations for clients who choose to outsource their in-plant print operations. Xerox has invented, tested, and implemented a novel class of operations-research-based productivity improvement offerings for the printing industry that has been extensively described in [17]. This work was a finalist in the 2008 Franz Edelman competition.

Print service centers are document manufacturing systems which take raw material and information as input and through a series of processing steps create final finished document products such as books, brochures, checks, invoices, and the like. They are designed to manufacture highly customized documents that are often embedded in their workflows. The document production steps associated with print jobs are indicated in Fig. 1. Typically print service centers have departments that support individual steps in this workflow. Each department supports many different types of internal workflows resulting from the use of different types of software tools, printing machines (e.g., offset, digital), and finishing equipments such as cutting, binding, laminating, and shrink wrapping. For further description of the steps we refer the reader to [17].

The LDP software toolkit automates several steps of the print production modeling process by taking declarative inputs from the end user and then automatically generating complex simulation models that are used to determine improved design and operating points for print shops. In this chapter, we describe the addition of another layer of automation to the LDP toolkit consisting of simulation-based optimization using greedy search techniques and simulated annealing that enables the automated search of a large number of design alternatives in the presence of operational constraints to determine a cost-effective solution for the print production environment.

The printing industry is highly fragmented with thousands of print shops that are geographically distributed. This approach lends itself to being utilized for optimizing print shops across multiple geographies by users less skilled in the art of modeling, simulation and optimization, thereby allowing unprecedented scalability of a simulation-based optimization approach to a wide user base. This is important since users are able to utilize the simulation-based optimization toolkit to make complex design and operational decisions and develop optimized designs without the arduous task of building the simulation models and the associated optimization framework.

This chapter is organized as follows. Section 2 provides a literature review on simulation-based optimization approaches. Section 3 describes the specifics of the problem being addressed in this chapter. Section 4 provides an overview of the Lean Document Production toolkit. Section 5 describes the existing procedure of selecting the optimal printing equipment. Section 6 describes the simulation-based optimization techniques using the LDP toolkit. Section 7 describes some applications and case studies using real-world examples. Lastly in Sect. 8 we present our conclusions and future scope of work.

2 Literature Review

The problem of constrained simulation optimization over a finite discrete set of decision variables is common and has received significant attention. A two-phase statistically valid procedure that detects feasibility of systems in the presence of one constraint with a prespecified probability of correctness was presented in [3]. This procedure was extended to the case of multiple constraints in [4]. An algorithm for optimal sampling allocations using large deviation theory was provided under stochastic settings [19]. Iterative heuristic algorithms [11], optimal computing budget allocation framework [15] was proposed for selecting the best design from a discrete number of alternatives in the presence of a stochastic constraint via simulation experiments with limitations on simulation budget or probability of correctness. A novel method [13] that converts constrained optimization into unconstrained optimization by using the Lagrangian function was proposed for the problem over discrete sets with noisy constraints.

The approaches discussed above either visit all the designs or convert the problem into a single objective function to find the best system. Suppose we conduct n simulation replications for each of θ designs, we need $n\theta$ total simulation replications. If the precision requirement is high, and if the total number of designs in a problem is large, then $n\theta$ can be very large, making the system evaluation computationally expensive using the existing methods. In such cases, stochastic search algorithms such as simulated annealing, tabu search, and genetic algorithms prove to be the best choice. Simulated annealing [12] has shown successful applications in a wide range of combinatorial optimization problems, and this fact has motivated researchers to use simulated annealing in many simulation optimization problems. But these search techniques need to be adapted for the stochastic environment associated with discrete-event systems optimization.

Haddock and Mittenthal have investigated the feasibility of using a simulated annealing algorithm in conjunction with a simulation model [7]. A variant of the simulated annealing algorithm was developed for solving discrete unconstrained stochastic optimization problems by using a constant temperature and convergence criteria as the number of visits made by the different states in the first m -iterations to estimate the optimal solution [2]. Two variants of the simulated annealing algorithm with a decreasing cooling schedule that are designed for solving unconstrained discrete simulation optimization problems was presented in [14]. For solving stochastically constrained simulation optimization systems, an integrated approach using the simulated annealing algorithm for parameter selection followed by Monte Carlo simulation for performance evaluation was presented in [1].

Unlike ranking and selection procedures, the application of metaheuristics techniques to simulation optimization problems in stochastic settings may not guarantee that an acceptable solution, if one exists, will be found. But in most cases we observe that they converge to acceptable solutions in a reasonable amount of time which is most desirable in many real-world applications. In this paper we have present the modified simulated annealing approach that can handle uncertainty in simulation

output and stochastic constraint(s). The algorithm starts with an initial feasible solution and utilizes a decreasing cooling schedule. We perform the *student's t* hypothesis test for determining the feasibility of a solution at the current iteration [1]. Our algorithm is distinct to the procedure in [1] by not restricting the neighborhood search to feasible moves only.

In a web-based simulation optimization applications, approaches that result in optimal/near optimal solutions in a reasonable time are desirable. A greedy approach is frequently a good alternative which makes locally optimal choices at each stage with the hope of finding a global optimum. An efficient greedy approach to allocate ambulance fleet in emergency medical services system was presented [20]. To determine the optimal configuration of a conveyor-based automatic material handling systems in wafer fabs, a greedy heuristic was proposed [10]. Discussion on greedy approximation for dock allocation in a food distribution center can be found in [6]. In this chapter we present the greedy approach for optimal allocation of equipment in print production environment in a web-based online application. The greedy algorithm initially starts with sufficient number of production equipment and systematically reduces the number. The algorithm removes one or more devices such that the customer's performance criteria are not violated. This process is repeated until no more cost reduction is possible subject to the constraint. Alternatively, the algorithm can begin with no or a minimal set of equipment and systematically increases that number.

3 Problem Description

Print service centers experience many sources of variability that make them hard to analyze and optimize. They exhibit high levels of job size variations, routing complexity, and demand fluctuations as shown in Fig. 2. These service centers are primarily make-to-order service systems that cater to specific requests of each incoming customer. The incoming service requests have random arrival and due date requirements that vary from job to job and often exhibit variability within the same job type. The job sizes are often characterized by highly non-normal distributions and sometimes heavy-tailed [16]. In addition to the above challenges, print shops also experience long bid times, variability in labor and equipment characteristics, etc.

The LDP toolkit automates the workflow modeling and analysis of the print service center. In order to optimize the cost and performance of a print service center, the user manually evaluates a limited number of designs and selects the best design among them. For example, to select the optimal equipment that minimizes the cost of equipment, while simultaneously meeting the performance of a print service center such as turnaround time, number of late jobs, operator or equipment utilization, process cycle efficiency, etc., the user have to simulate multiple equipment configuration scenarios manually and select the cost-effective solution among them. This process is labor intensive, time consuming, and often ad hoc. In this chapter we have described an automated method to assist in selecting cost-effective solutions for a print service center by integrating the optimization algorithm with the LDP solution.

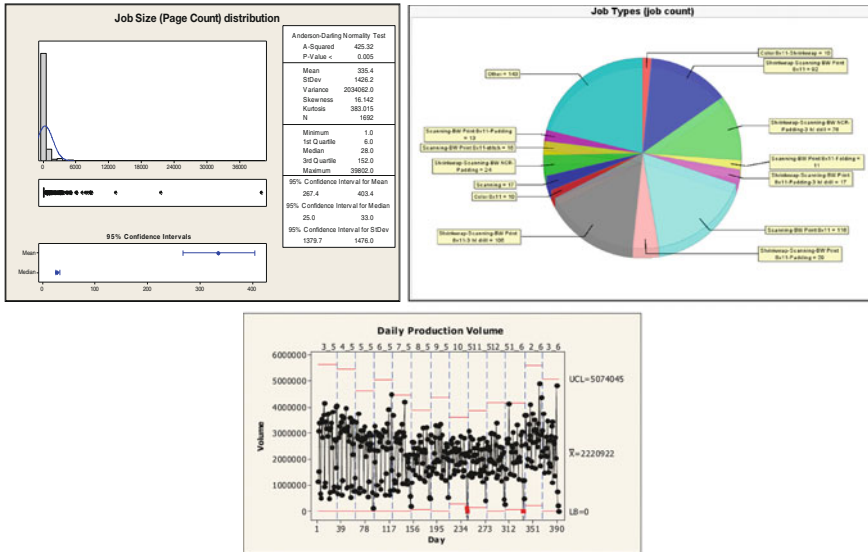


Fig. 2 Multiple sources of variability in a print production environment such as power law job size distributions, multiple coexistent job types, and high demand fluctuation

4 The LDP Solution for Print Service Center Environment

To address the complexity of operations associated with the print production processes, the service center resources are organized into autonomous cells [17]. As a result, the most common jobs can be finished autonomously inside (at least) one of these cells. Figure 3 shows how traditional print service centers are organized based on a departmental structure operated by specialized workers and compare it to the redesigned operational framework based on autonomous cells where diverse pieces of equipment are collocated and operated by cross-trained workers.

To orchestrate the flow and control of jobs through the parallel hierarchical cell structure, the Lean Document Production Controller (LDPC) uses 2-level architecture for production management. The LDPC has:

- A service center controller module (Service center CM)—high-level controller, in charge of global service center management.
- Several cell controller modules (Cell CMs)—low-level controllers, in charge of local management inside cells.

4.1 Simulation

Simulation is performed to assess the results of improvements resulting from changes in workflow grouping, operator cross-training, grouping diverse equipment into

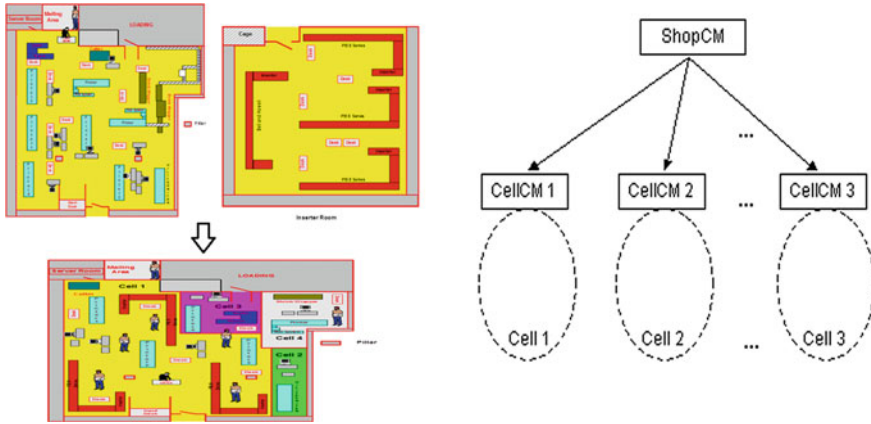


Fig. 3 Figure showing how a departmental configuration of a print service center is transformed into a cellular structure utilizing autonomous cells and the corresponding two-level architecture for the Lean Document Production Controller

autonomous cells and scheduling policies. Building discrete-event simulation models is often a time-intensive effort especially when various scenarios have to be investigated to determine improved solutions. To facilitate the model building process, the LDP tool is employed to build the simulation models from a declarative user interface (Fig. 4). This allows for fast and efficient evaluation of a large number of what-if scenarios and greatly aids in determining an improved solution out of a large search space.

The user specifies the equipment characteristics, elements of the cell, scheduling policies, number of operators and their skill level, and workflow/job characteristics

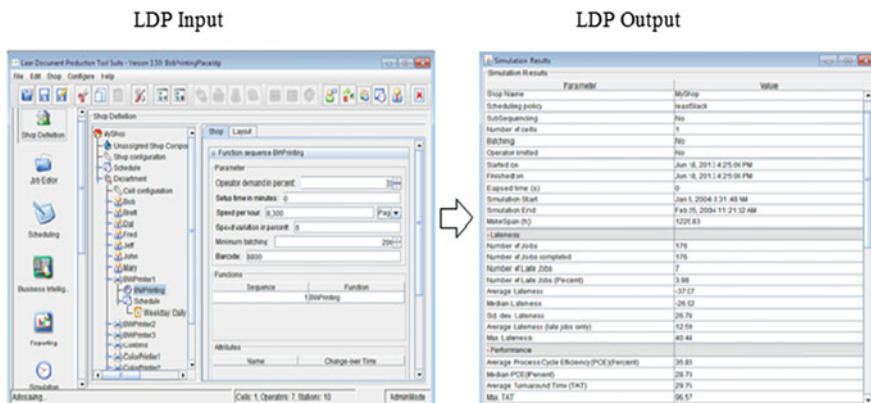


Fig. 4 Illustrates the user interface for defining the printing equipment, operators and shop policies, and simulation results for a sample print service center

as inputs to the simulation model using the LDP user interface (Fig. 4). Before the shop is simulated, the user schedules the jobs automatically using the scheduling architecture as described above. Next the tool simulates the operation of the print service center and outputs various performance metrics such as average turnaround time, number of late jobs, operator and equipment utilization, maximum turnaround time, and process cycle efficiency, etc., as shown in Fig. 4.

5 Existing Procedure for Selecting Optimal Equipment Design in a Print Service Center

The selection of optimal printing equipment in the print service center is currently carried out manually. The user first defines the necessary equipment type, cost, and other characteristics (speed, setup time, failure, and repair rates, etc.) in each cell. The job workflow characteristics and other shop operating policies (job sequencing policy, batching, and work in progress, etc.) are collected from the shop and uploaded to the LDP tool. An equipment design is defined as a combination of different numbers of equipment types in each cell. The user has to create different equipment designs that he is interested in by varying the quantity of each equipment type in each cell. Each of these equipment designs is simulated N times in order to create performance metric distributions (in the case where the simulation is subject to random events such as machine failures and job variation). Then, the mean performance measure of interest and total cost of the equipment is computed. Finally, the user selects the equipment design that has the least cost and meets the desired print service center performance goal as specified by the user. This process of evaluating multiple design configurations is labor intensive, time consuming, and can lead to solutions far from optimal. Figure 5 illustrates the detailed process flow diagram of the existing procedure.

6 Simulation-Based Optimization Using the LDP Toolkit

The main idea presented here is the integration of the optimization routine and simulation module within the LDP toolkit that embodies many elements of shop specification and modeling automation. This enables the automatic search of an optimal solution for the print production service center. For more detailed discussion, applications and benefits of integrating optimization with simulation can be found in [5, 8, 18, 21].

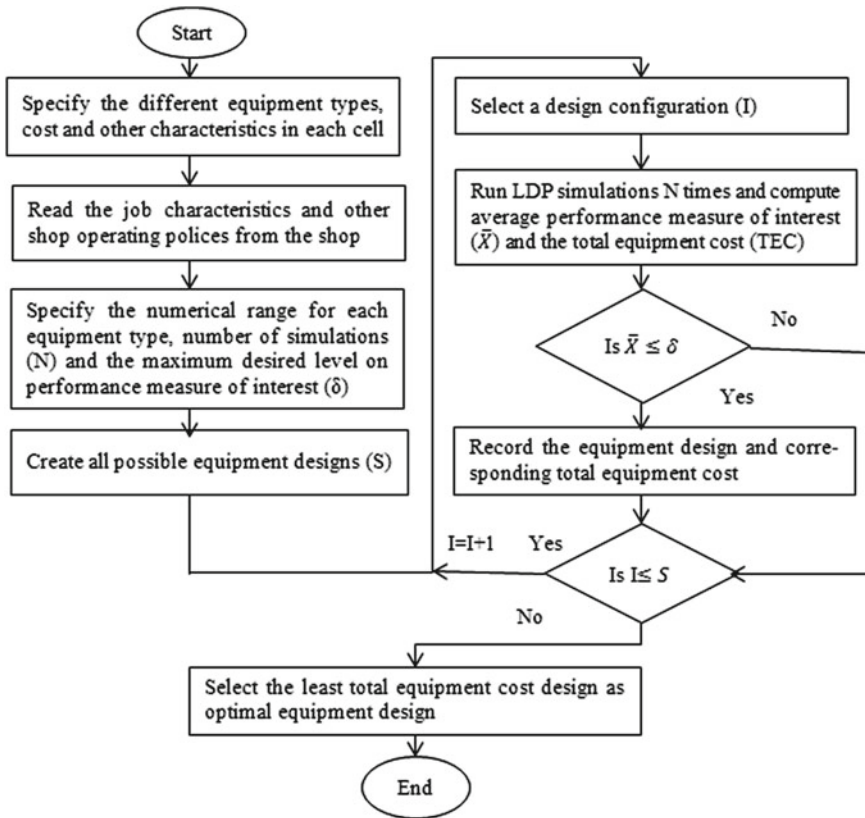


Fig. 5 The existing procedure for selecting optimal equipment configuration using LDP tool in print service center

6.1 Problem Formulation

The problem of selecting the cost-optimal equipment solution for the print production environment in the presence of stochastic operational constraints such as average turnaround time, number of late jobs, maximum turnaround time, etc., over a large number of design alternatives can be formulated as below.

$$\begin{aligned}
 \text{Objective : } & \min_{X_k \in S} f_0(X_k) & (1) \\
 \text{Subject to : } & f_1(X_k) \leq \delta \\
 & lb_{ij} \leq x_{ij} \leq ub_{ij}, i = 1..n_j, j = 1..m \\
 & X_k = [x_{ij}]
 \end{aligned}$$

where S , the search space, is a finite and discrete set of equipment design configurations; X_k is the k th equipment design configuration, which is the vector combination in the number of each type of equipment in each cell; k is the index of equipment design configuration; x_{ij} is the number of the i th type of equipment in the j th cell; n_j is the number of unique equipment types in cell j ; m represents the total number of cells in the print service center; lb_{ij} and ub_{ij} are the lower and upper bounds on the number of the i th type of equipment in the j th cell; $f_0(X_k)$ is the total equipment cost defined as $C_{ij} \times x_{ij}$, where C_{ij} is the cost of i th equipment in j th cell; $f_1(X_k)$ is the print service center performance measure such as average turnaround time, number of late jobs, maximum turnaround time, etc., which cannot be evaluated exactly, but needs to be estimated via the LDP simulation. Let A_{kl} be the print service center performance observation observed from simulation replication l of system k , then $f_1(X_k) = E[A_{kl}]$; and δ is the maximum desirable level of the print service center performance measure.

6.2 Modified Simulated Annealing Algorithm

Here, we present the modified simulated annealing algorithm used for solving Eq. 1. The algorithm consists of two phases: initial feasible solution phase and optimal solution phase. In the initial feasible solution phase, the algorithm starts with searching for an initial feasible solution by randomly selecting a solution from the design search space until the stopping criteria is met. If an initial feasible solution was found, the algorithm starts with this solution and identifies the optimal solution by utilizing a decreasing cooling schedule in the optimal solution phase. In the case of initial unfeasible solution, the algorithm is terminated.

Moreover, the constraints in Eq. 1 are stochastic and the general-purpose simulated annealing approach has to be adapted to consider the feasibility of a solution when it moves from one solution to another. A solution is feasible if it meets the print service center performance goal as specified by the user. To test the feasibility of a solution, we use the following procedure [1].

Let us consider, an arbitrary stochastic constraint $g(\theta) \leq \delta$, where $g(\theta)$ is the stochastic simulation output for design θ and δ being the maximum desirable level specified by the user. Letting $g_i(\theta)$ denote the i th simulation replication and running simulation n times, the mean and variance estimate for $g(\theta)$ could be determined over n replications as:

$$\hat{g}(\theta) = \sum_{i=1}^n g_i(\theta) / n$$

$$\hat{\sigma}_{\hat{g}(\theta)} = \sum_{i=1}^n (g_i(\theta) - \hat{g}(\theta))^2 / (n - 1)$$

The hypothesis statements for feasibility conditions are as follows:

$$\begin{aligned} \text{Null Hypothesis } H_0 &: \hat{g}(\theta) \leq \delta \\ \text{Alternate Hypothesis } H_1 &: \hat{g}(\theta) > \delta \end{aligned}$$

Accept the null hypothesis H_0 , if $\omega = \hat{g}(\theta) + t_{n-1,1-\alpha} \times \hat{\sigma}_{\hat{g}(\theta)} / \sqrt{n} \leq \delta$ where,

- $n - 1$ is the degrees of freedom
- $1 - \alpha$ is the upper critical point for the t distribution
- $\hat{g}(\theta)$ is the mean value of n simulation observations
- $\hat{\sigma}_{\hat{g}(\theta)}$ is the standard deviation of $\hat{g}(\theta)$.

Unlike to the approach [1], our algorithm does not restrict the neighborhood search to feasible moves only. In their approach the temperature length (M) parameter is not incremented until a neighboring feasible solution is found, resulting in unknown/more number of evaluations. When the probability of finding a feasible neighborhood solution is very low, this may result in indefinite looping. In the modified simulated annealing algorithm, a move to a neighborhood solution is irrespective of the feasibility of the solution, providing more control on the total number of evaluations by the algorithm. Let T_0 be the initial temperature, T_f be the final temperature (T_0/T_{depth}) and r the temperature decay rate. This results in the following series of annealing temperatures:

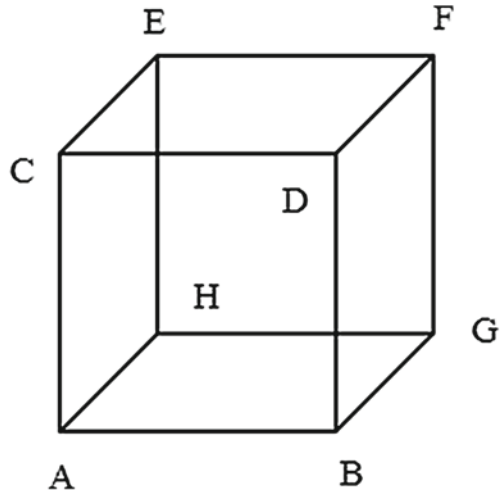
$$\begin{aligned} T_0, T_0 \times r, T_0 \times r^2, T_0 \times r^3 \dots \dots + T_0 \times r^n, \\ T_f = \frac{T_0}{T_{depth}} = T_0 \times r^n, \\ n = \frac{\log^1/T_{depth}}{r}. \end{aligned}$$

If the number of times to search a neighborhood solution at a given temperature is L , then the number of evaluations is $n \times L$. The value of L is fixed throughout the algorithm and is determined using trial-and-error approach. To estimate the performance measure the algorithm makes use of all the historical observations obtained at that solution. Next, we define the following:

Definition 1 The search space S is a set of equipment design configurations whose cardinality or $|S|$ is $\prod_{j=1}^m \prod_{i=1}^{n_j} (ub_{ij} - lb_{ij} + 1)$.

Definition 2 For each $X_k \in S$, there exists a subset $N(\theta)$ of $S - \{X_k\}$ which is called the set of neighbors of X_k , such that each point in $N(\theta)$ can be reached from X_k in a single transition. For example in Fig. 6, the search space $S = \{A, B, C, D, E, F, G\}$ and the set of neighbors of $X_k = B$ is $N(\theta) = \{A, D, G\}$.

Fig. 6 Illustrates the discrete search space of a cube



6.2.1 Algorithm

Parameters

Number of times to run the simulations at design (\mathbf{X}_k) : n

Temperature depth: T_{depth}

Temperature decay rate: r

Maximum desirable level of secondary performance measure: δ

No. of times to search a neighborhood solution at a given temperature: L

Fraction of the total search space S for obtaining initial feasible solution in %: β

Significance value for t -test: α

Phase I: Finding initial feasible solution

1. $feasibility = false$
2. $max = |S| * \beta$
3. $i = 0$
4. Repeat:
 - 4.1. Randomly select the design configuration: $X_i \in S$
 - 4.2. Generate n simulation observations for performance measures: $\{f_0(X_i)\}_{j=1}^n, \{f_1(X_i)\}_{j=1}^n$
 - 4.3. Evaluate: $\hat{f}_0(X_i), \hat{f}_1(X_i), \hat{\sigma}_{\hat{f}_0(X_i)}, \hat{\sigma}_{\hat{f}_1(X_i)}$ and $t_{n-1, 1-\alpha}$
 - 4.4. If $\hat{f}_1(X_i) + t_{n-1, 1-\alpha} \times \hat{\sigma}_{\hat{f}_1(X_i)} / \sqrt{n} \leq \delta$ then
 - 4.5. $feasibility = true$
 - 4.6. End if
 - 4.7. $i = i + 1$
5. Until $feasibility = true$ or $i > max$

6. If *feasibility* = *true* then
7. Return X_i as initial feasible design configuration
8. Else
9. Return initial design configuration cannot be found in *max* iterations
10. End If

Phase II: Finding optimal design solution

1. $value = \hat{f}_0(X_i)$, $T_{initial} = value/2$, $T_{final} = T_{initial}/T_{depth}$
2. Repeat:
 - 2.1. For $j = 1 \dots L$
 - 2.2. Randomly select the neighborhood design X_j , where $X_j \in N(X_i)$ and $N(X_i)$ is the set of neighborhood of X_i
 - 2.3. Generate n simulation observations for performance measures:
 $\{f_0(X_j)\}_{p=1}^n, \{f_1(X_j)\}_{p=1}^n$
 - 2.4. Evaluate: $\hat{f}_0(X_j), \hat{f}_1(X_j), \hat{\sigma}_{\hat{f}_0(X_j)}, \hat{\sigma}_{\hat{f}_1(X_j)}$ and $t_{n-1, 1-\alpha}$
 - 2.5. If $\hat{f}_1(X_j) + t_{n-1, 1-\alpha} \times \hat{\sigma}_{\hat{f}_1(X_j)} / \sqrt{n} \leq \delta$
 - 2.5.1. $newvalue = \hat{f}_0(X_j)$
 - 2.5.2. $delta = newvalue - value$
 - 2.5.3. Generate uniform random number $U_k \sim U[0, 1]$
 - 2.5.4. If $delta < 0$ or $e^{-delta/T} \geq U_k$ then
 - 2.5.5. $value = newvalue$
 - 2.5.6. $X_i = X_j$
 - 2.5.7. End If
 - 2.6. End If
 - 2.7. Next j
3. Reduce the temperature: $T = r \times T$
4. Until $T \geq T_{final}$
5. Return X_i as the optimum equipment design configuration and $value$ as optimum total equipment cost value.

6.3 A Greedy Algorithmic Approach for Equipment Allocation

In this section we consider another approach in a different class from that of simulated annealing. That is an approach formulated from a greedy perspective. The greedy methodology to optimization applies a heuristic that makes the locally optimal choice at each step. Often the globally optimal solution will not be found but the greedy heuristic may yield an adequate solution in reasonable time.

Consider the case of assigning a subset of N devices to a production shop. For N devices, there are 2^N possible assignments. Even for moderate values of N this can be prohibitively large. Also each assignment may require the completion of a time-consuming simulation run since there is no analytic model capable of expressing

the production process characteristics of interest except in the simplest cases. Each run itself may need to be repeated to obtain distributions on performance metrics. Also if it is desirable to provide timely production design services via, for example, a web-based tool, then there may be additional constraints on the timely completion of a solution. The greedy algorithm initially starts with a sufficient number of devices. Next, the algorithm removes one or more devices such that the customer's performance criteria are not violated. This process is repeated until no more cost reduction is possible. Alternatively, the algorithm can start with none, or a minimal set of devices (such that each required function can be performed) and from this configuration devices can be added one or more at a time until the constrained performance criteria is achieved. The device chosen to be added or removed at each iteration is the device with the best (as detailed below) cost to benefit trade-off. The method is analogous to forward selection, backward selection, and mixed selection methods applied in the area of parsimonious model selection discussed in [9].

To illustrate the approach, we consider a shop model in the form of a discrete-event simulation that must be exercised over some duration and some job list condition. We define best performance as that with least cost with a job turnaround time metric below a specified upper constraint. The discrete-event model provides as output the turnaround metric and the costs incurred in processing the set of jobs. We will assume for the example below that the performance metric is the average turnaround time (TAT). Other performance metrics can be selected. The approach proceeds in the following steps:

1. Complete a simulation with all N machines, $\{M_1, M_2 \dots M_N\}$, assigned to the shop. This will produce as output a turnaround time metric (TAT) and cost (or if runs are stochastic then the output will be in the form of distributions). Check that the TAT metric is below the performance constraint(s). If not then stop since there is no solution possible that would satisfy the constraint. If no solution exists then one must start with more than N machines. If a solution exists then proceed to step 2.
2. Run N more simulations. Each of the N simulations will consist of $N-1$ machines. For the first simulation remove M_1 and retain $\{M_2 \dots M_N\}$, for the second simulation replace M_1 and remove M_2 , so that we now retain $\{M_1, M_3, M_4 \dots M_N\}$. Repeat until each machine has been removed in turn. From these N simulations determine the set of TAT metrics and cost that is output from the simulation.
3. Consider the average TAT metric and cost output from step 1 above, and the N from step 2. This is shown in Fig. 7 (For clarity only 4 of the N results, labeled points A, B, C, and D, are shown from step 2). Here a decision is made in which one of the machines is removed so as to reduce the set from N machines to $N-1$. Any point, such as point D that results in the constraint being violated is not a candidate and is to be removed. If all points lie above the constraint then no reduction in machines is permissible, and so the machine removal portion of this method is stopped. A number of sensible rules can be applied to define which machine to remove. Example performance heuristics are:

- Remove the machine which resulted in the largest cost reduction without violating the *TAT* constraint. This would correspond to point *A* in Fig. 7.
- Remove the machine which resulted in the smallest increase in *TAT* metric, this would be point *B*. Or,
- Remove the machine that resulted in the greatest cost reduction per unit *TAT* increase which would be point *C*. This would also capture the case in which a *TAT* “increase” is actually negative—a very favorable condition.

All three of the above rules can be applied in three completely independent applications of the method and the best result chosen. This will require approximately 3 times the number of runs.

4. Steps 2 and 3 above are repeated until the costs can no longer be achieved subject to the constraint. This would result in an absolute maximum $(N + 1) * N / 2$ simulations. So for a pool of 30 machines that would be 465 simulations maximum. The maximum is unlikely to be run and certain policies can be adopted to reduce further the number of simulations. For example,
 - Step 2 is likely to stop because of *TAT* constraint violations with N substantially larger than 1 in all except the most trivial of problems.
 - Also the assumption that the machines are unique is highly unlikely. For example, if machines M_1 , M_2 , and M_3 are identical then fewer simulations would be required since removing M_1 is equivalent to removing M_2 or M_3 . In this way the existence of equivalent machines may reduce the required number of simulations.
 - A final way in which the number of simulations can be reduced is to use as a guide the utilization results. So, for example, after the completion of step 1, we have as output the utilization rates of the set of N machines. These should be ordered from low utilization to high utilization. Step 2 can begin by removing the machines in the order of lowest to highest utilization. One can then choose not to run simulations for the machines that are highly utilized.
 - Lastly, a modification that may be particularly effective would be to look at larger groups of machines to be removed and replaced. For example, with the case of 10 unique machines, instead of removing 1 machine at a time we may consider removing groups of 2 (or 3, etc...) machines at a time. Suppose we consider the removal of 2 machines at a time. Then there are $10 * 9 / 2 = 45$ unique 2-tuples ($10 * 9 * 8 / 6 = 120$ unique 3-tuples). And so 45 evaluations would be run after which 2 machines are removed that yielded the best trade-off in cost and performance. There are now 8 machines left in the pool. The number of simulations increased from 19 ($10 + 9$) to 45, but the method explores a larger set of possibilities and is therefore more likely to produce a solution closer to optimal. If the group size is increased from 1 to 2, then from 2 to 3, and then from 3 to 4, etc... a more extensive area of the search space is evaluated. As the group size approaches N we approach the state of exhaustive search. If a time constraint is set on the simulation time then one can proceed logically through configurations as outlined in this chapter until the time limit is reached.

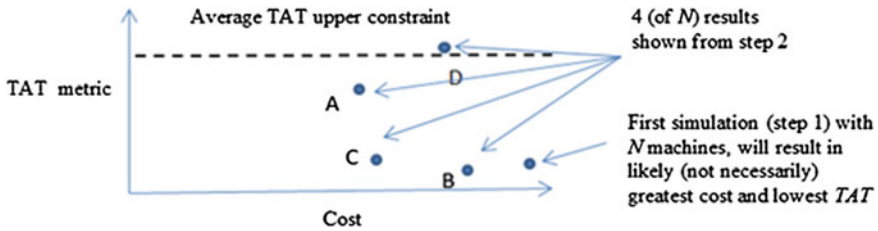


Fig. 7 Plot of *TAT* metric versus *cost*

- All the above stated approaches aim at reducing the number of required simulations. Since the operations in step 2 and step 3 can be performed independently, the method easily lends itself to parallel processing which may further reduce the computation time.

The final phase of this method can be to introduce machine failures. This final step is a more conventional approach and so it is only outlined here for completion (unlike steps 1–4 above, the final step does not constitute a core idea of this chapter). Directionally to cope with machine failures more machines may need to be added (not subtracted as in the steps above). A number of simulations are to be run and the *TAT* metric distribution estimated. Machine(s) that are bottleneck devices and/or highly utilized and therefore vulnerable failure points are identified and if their reliability levels are sufficiently low, backup machines are added until the distribution of the *TAT* metric is adequate.

7 Application and Case Study

Print service centers can be classified into three categories based on the activity that they perform: transaction printing, on-demand publishing, or a combination of both. A transaction printing environment produces documents such as checks and invoices. Each document set is different. Mail metering and delivery are part of the workflow. On-demand publishing environments focus on producing several copies of identical documents with more finishing options such as cutting, punching, and binding. Examples of such products include books, sales brochures, and manuals. Other environments perform both types of document production simultaneously with varying emphasis on each one.

In this section we illustrate the selection of the equipment configuration in three print service centers using existing, simulated annealing, and greedy algorithmic approaches for different performance criteria's. The total equipment cost is deterministic and defined $C_{ij} \times x_{ij}$ as where, C_{ij} is the fixed cost of i th equipment in the j th cell and x_{ij} is the number of i th type of equipment in the j th cell. The print service center performance measure $f_1(X_k)$ is problem specific and can only be estimated by running simulations using the LDP toolkit.

7.1 Print Service Center 1

This print service center has two cells and six stations and can process printing and inserting job workflows. Table 1 shows the equipment in each cell and their fixed cost.

Job data over a period of 10 days is collected from the print service center with a total of 2692 jobs during that period. The number of equipment of each function/station type in a cell is varied between 1 and 3 and so the total number of possible equipment configuration is 729. Table 2 illustrates a sample of all the possible equipment configurations.

Next, we illustrate the selection of optimal or near-optimal equipment configurations for the print service center using the existing approach with N equal to 30 (N is the number of simulations replications for each design configuration), simulated annealing approach with the parameters $n = 5$, $L = 5$, $T_{depth} = 100$, $r = 0.9$, $\beta = 5\%$ and $\alpha = 0.01$, and greedy algorithm starting initially with a solution having 3 number of equipment of each type in each cell for two test cases.

7.1.1 Test Case 1

In this problem, we have consider the print service center performance measure $f_1(X_k)$ as the average turnaround time less than or equal to 5h. The average turnaround time is defined as the arithmetic average of turnaround times (difference between the completion time and arrival time of job) of all the jobs. Table 3 illustrates the results summary.

Table 1 The printing equipment in each cell and their fixed cost

Cell	Station	Fixed cost (\$)
Cell one	Printer A	2,448,874
Cell one	Inserter A	423,366
Cell one	Inserter B	1,443,304
Cell one	Printer B	2,448,874
Cell two	Printer C	3,000,000
Cell two	Inserter B	1,443,304

Table 2 A sample of equipment configuration in print service center 1

Design no	No. of printer A's in cell one	No. of inserter A's in cell one	No. of inserter B's in cell one	No. of printer B's in cell one	No. of printer C's in cell two	No. of inserter B's in cell two
1	1	1	1	1	1	1
2	1	1	1	1	1	2
.
729	3	3	3	3	3	3

Table 3 Test case 1 results summary

	Existing approach	Simulated annealing		Greedy algorithm	
		Run1	Run2	Run1	Run2
Printer A's in cell one	1	1	1	1	1
Inserter A's in cell one	1	1	1	1	1
Inserter B's in cell one	3	3	3	3	3
Printer B's in cell one	1	1	1	1	1
Printer C's in cell two	1	1	1	1	1
Inserter B's in cell two	2	2	2	2	2
Optimal total station cost (\$)	15,537,634	15,537,634	15,537,634	15,537,634	15,537,634
Average turnaround time (h)	4.8	4.75	4.82	4.78	4.78
Number of simulations	21870	1120	1115	72	72
Time in hours	29.94	1.44	1.66	0.106	0.11

7.1.2 Test Case 2

In this problem, we have considered the print service center performance measure $f_1(X_k)$ as number of late jobs less than or equal to 0. A print job is late if the completion date exceeds the due date. Table 4 illustrates the results summary.

7.2 Print Service Center 2

This print service center has 4 cells and 70 stations and can process job workflows having printing, cutting, binding, punching, and other finishing and mailing services. The search for the optimal equipment configuration is performed only for the printing equipment in the print service center. Only two cells in the print service have printing equipment. Table 5 shows the printing equipment in each cell and their monthly fixed costs.

Job data for a period of 20 days is collected from the print service center with 2593 jobs in the period. The number of equipment of each type in a cell is varied between 1 and 3 and the total number of possible equipment configuration is 2187.

Table 4 Test case 2 results summary

	Existing approach		Simulated annealing		Greedy algorithm	
	Run1	Run2	Run1	Run2	Run1	Run2
Printer A's in cell one	1	1	1	1	1	1
Insertor A's in cell one	2	3	1	3	3	3
Insertor B's in cell one	2	2	2	2	2	2
Printer B's in cell one	1	1	1	1	1	1
Printer C's in cell two	2	2	2	2	2	2
Insertor B's in cell two	2	2	3	2	2	2
Optimal total	17,517,696	17,941,062	18,537,634	17,941,062	17,941,062	17,941,062
Station cost (\$)						
No of late jobs	0	0	0	0	0	0
No of simulations		21,870	1105	1120	64	64
Time in hours		29.94	1.77	1.50	0.094	0.095

Table 5 The printing equipment in each cell and their fixed cost

Cell	Station	Monthly fixed cost (\$)
Cell one	Printer A	1601
Cell one	Printer B	6771
Cell one	Printer C	3907
Cell two	Printer D	6771
Cell two	Printer E	1544
Cell two	Printer F	2472
Cell two	Printer G	2120

Next, we illustrate the selection of optimal or near-optimal equipment configurations for the print service center using the existing approach with N equal to 5, the simulated annealing approach with parameters $n = 5$, $L = 5$, $T_{depth} = 100$, $r = 0.9$, $\beta = 5\%$ and $\alpha = 0.01$, and the greedy algorithm starting initially with a solution having 3 number of equipment of each types in each cell for two test cases.

7.2.1 Test Case 3

In this problem, we have considered the print service center performance measure $f_1(X_k)$ as the average turnaround time less than or equal to 2 h. Table 6 illustrates the results summary.

7.2.2 Test Case 4

In this case, we have considered the print service center performance measure $f_1(X_k)$ as the maximum turnaround time less than or equal to 48 h. The maximum turnaround time is defined as the maximum value of turnaround times over all the jobs. Table 7 illustrates the results summary.

7.3 Print Service Center 3

This print service center has two cells and four stations and can process job workflows having printing, and inserting. Table 8 shows the printing equipment in each cell and their monthly fixed costs.

Job data over a period of 30 days is collected from the print service center with a total of 2833 jobs in the period. The number of equipment of each type in a cell is varied between 1 and 8 and the total number of possible equipment configuration is 4096.

Table 6 Test case 3 results summary

	Existing approach		Simulated annealing		Greedy algorithm	
			Run1	Run2	Run1	Run2
Printer A's in cell one	1	2	1	1	2	2
Printer B's in cell one	1	1	1	1	1	1
Printer C's in cell one	1	1	1	1	1	1
Printer D's in cell two	1	1	1	1	1	1
Printer E's in cell two	2	1	2	2	1	1
Printer F's in cell two	1	1	1	1	1	1
Printer G's in cell two	1	1	1	1	1	1
Optimal total station cost (20 days)	\$17,822	\$17,860	\$17,822	\$17,822	\$17,860	\$17,860
Avg turnaround time (h)	2.0	1.94	1.98	2.0	1.91	1.93
Number of simulations	10,935		1105	1110	61	61
Time in hours	58.76		5.55	5.53	0.316	0.305

Next, we illustrate the selection of optimal or near-optimal equipment configurations for the print service center using the existing approach with N equal to 5, simulated annealing approach with the parameters $n = 5$, $L = 5$, $T_{depth} = 100$, $r = 0.9$, $\beta = 5\%$ and $\alpha = 0.01$, and greedy algorithm starting initially with a solution having 8 number of each equipment type in each cell for two test cases.

7.3.1 Test Case 5

In this problem, we have consider the print service center performance measure $f_1(\mathbf{X}_k)$ as the average turnaround time less than or equal to 5 h. Table 9 illustrates the results summary.

Table 7 Test Case 4 results summary

	Existing approach		Simulated annealing		Greedy algorithm	
			Run1	Run2	Run1	Run2
Printer A's in cell one	1	1	1	1	1	1
Printer B's in cell one	1	1	1	1	1	1
Printer C's in cell one	1	1	1	1	1	1
Printer D's in cell two	1	1	1	1	1	1
Printer E's in cell two	2	1	2	2	2	2
Printer F's in cell two	1	1	1	1	1	1
Printer G's in cell two	1	2	1	1	1	1
Optimal total station cost (20 days)	\$17,822	\$18,206	\$17,822	\$17,822	\$17,822	\$17,822
Max turnaround time (h)	41.52	44.30	40.80	41.22	40.97	41.75
Number of simulations	10,935		1110	1110	67	67
Time in hours	58.76		5.58	5.67	0.33	0.35

Table 8 The printing equipment in each cell and their fixed cost

Cell	Station	Monthly fixed cost (\$)
Cell one	Printer A	19,156
Cell one	Printer B	3907
Cell two	Insenter A	21,267
Cell two	Insenter B	11,485

7.3.2 Test Case 6

In this case, we have considered the print service center performance measure $f_1(X_k)$ as the maximum turnaround time less than or equal to 48 h. Table 10 illustrates the results summary.

Table 9 Test case 5 results summary

	Existing approach		Simulated annealing		Greedy algorithm	
			Run1	Run2	Run1	Run2
Printer A's in cell one	8	8	8	8	8	8
Printer B's in cell one	1	2	2	2	1	1
Inserter A's in cell two	1	1	1	1	1	1
Inserter B's in cell two	5	5	5	5	5	5
Optimal total station cost (\$)	235,847	239,754	239,754	239,754	235,847	235,847
Average turnaround time (h)	4.95	4.88	4.9	4.92	4.97	5
No of simulations	20,480		1155	1165	67	67
Time in hours	57.41		1.58	1.83	0.103	0.092

Table 10 Test Case 6 results summary

	Existing approach		Simulated annealing		Greedy algorithm	
			Run1	Run2	Run1	Run2
Printer A's in cell one	5	5	5	5	5	5
Printer B's in cell one	4	2	4	2	4	4
Inserter A's in cell two	1	2	1	2	1	1
Inserter B's in cell two	3	2	3	2	3	3
Optimal total station cost (\$)	167,130	169,098	167,130	169,098	167,130	167,130
Max turnaround time (h)	47.82	47.38	47.79	47.29	47.8	47.76
No of simulations	20,480		1145	1110	99	99
Time in hours	57.41		1.44	1.30	0.135	0.133

7.4 Results and Discussion

We have demonstrated the selection of optimal equipment in print service environments using modified simulated annealing and greedy algorithm techniques for different test cases. These test cases differ either in the performance measures or the problem size. In test cases 1 and 4, the simulated annealing and greedy algorithm finds the optimal solutions for both the experimental runs. Whereas, in test case 3 the simulated annealing outperforms the greedy algorithm solutions and in test case 5 the greedy algorithm outperforms simulated annealing solutions. In test case 2 and 6, the simulated annealing and greedy algorithm performs equally in one experimental run, but the greedy technique outperforms annealing algorithm in the second experimental run.

The results show that the greedy algorithm and simulated annealing perform adequately for a set of tasks typical in the improvement of print operations irrespective of the size of the problem. The simulated annealing technique is more time consuming and is performed offline and used during preliminary print service center cost evaluations. The simulated annealing algorithm is wrapped around the stand-alone LDP modeling framework, enabling the users to determine the optimal equipment configuration by evaluating a very large number of possible configurations automatically. In addition, by enabling automated simulation-based optimization, we can enable less skilled users to utilize the power of the LDP toolkit in making informed and optimized decisions offline. The business value of this automated simulation optimization solution can be enhanced further by incorporating this into an online web-based LDP optimization framework. As the greedy algorithm is much faster than simulated annealing, it is used in a web-based online application as shown in Fig. 8.

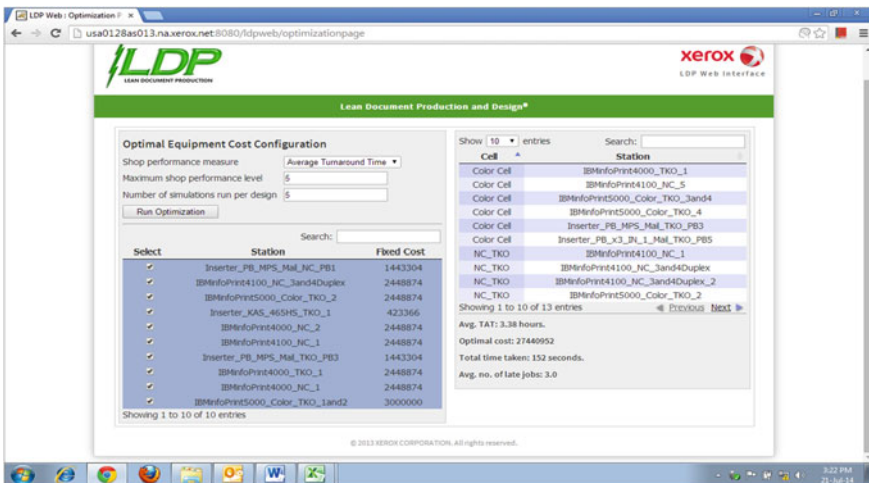


Fig. 8 Online web-based LDP tool kit

8 Conclusions and Future Work

This chapter presents a simulation-based optimization solution using simulated annealing as an offline approach and a greedy methodology as an offline or online approach for optimal print shop equipment selection. It describes how suitable abstractions and automation of the simulation tool can enable deployment of the Lean Document Production solution for cost-optimal equipment selection within a highly fragmented printing industry, while optimizing key performance objectives such as average turnaround time, number of late jobs, operator or equipment utilization, process cycle efficiency, etc. Though the techniques described here are applied within printing industry, they can also be utilized in other service-based operations with similar workflow characteristics.

Here we have used simulated annealing as an optimization approach, other evolutionary approaches such as ant colony and genetic algorithms can also be utilized for this purpose. But, these techniques need to be adapted to suit to the stochastic environments. The computational speed of these algorithms can be improved further by parallelizing, running on cloud-based platforms. We carried the above study by considering a single performance measure; further study can be made to extend the algorithm for multiple shop performance measures such as labor cost and operational cost.

References

1. Ahmed MA, Alkhamis TM, Hasan M (1997) Optimizing discrete stochastic systems using simulated annealing and simulation. *Computers & Industrial Engineering* 32:823–836
2. Alkhamis TM, Ahmed MA (2004) Simulation based optimization using simulated annealing with confidence interval. In: Ingalls RG, Rossetti MD, Smith JS, Peters BA (eds) *Proceedings of the 2004 winter simulation conference*, Washington DC, 2004
3. Andradóttir S, Goldsman D, Kim SH (2005) Finding the best in the presence of a stochastic constraint. In: Kuhl ME, Steiger NM, Armstrong FB, Joines JA (eds) *Proceedings of the 2005 winter simulation conference*, Florida, 2005
4. Batur D, Kim SH (2005) Procedures for feasibility detection in the presence of multiple constraints. In: Kuhl ME, Steiger NM, Armstrong FB, Joines JA (eds) *Proceedings of the 2005 winter simulation conference*, Florida, 2005
5. Fu MC, Andradóttir S, Carson JS, Glover F, Harrell CR, Ho YC, Kelly JP, Robinson SM (2000) Integrating optimization and simulation: research and practice. In: Joines JA, Barton RR, Kang K, Fishwick PA (eds) *Proceedings of the 2000 winter simulation conference*, Florida, 2000
6. Gopakumar B, Sundaram S, Wang S, Koli S, Srihari K (2008) A simulation based approach for dock allocation in a food distribution center. In: Mason SJ, Hill RR, Moench L, Rose O, Jefferson T, Flower JW (eds) *Proceedings of the 2008 winter simulation conference*, Florida, 2008
7. Haddock J, Mittenthal J (1992) Simulation optimization using simulated annealing. *Computers & Industrial Engineering* 22:387–395
8. Harkan IA, Hariga M (2007) A simulation optimization solution to the inventory continuous review problem with lot size dependent lead time. *The Arabian Journal for Science and Engineering* 2:327–338

9. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R. Springer Heidelberg, New York
10. Johnson A, Carlo HJ, Jimenez JA, Nazzal D, Lasrado V (2009) A greedy heuristic for locating crossovers in conveyor based ahms in wafer fabs. In: Rossetti MD, Hill RR, Hohansson B, Dunkin A, Ingalls RG (eds) Proceedings of the 2009 winter simulation conference, Texas, 2009
11. Kabirian A, Olafsson S (2009) Selection of the best with stochastic constraints. In: Rossetti MD, Hill RR, Hohansson B, Dunkin A, Ingalls RG (eds) Proceedings of the 2009 winter simulation conference, Texas, 2009
12. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220: 671–680
13. Luo Y, Lim E (2011) Simulation based optimization over discrete sets with noisy constraints. In: Jain S, Creasey RR, Himmerspach J, White KP, Fu M (eds) Proceedings of the 2011 winter simulation conference, Arizona, 2011
14. Prudius AA, Andradóttir S (2005) Two simulated annealing algorithms for noisy objective functions. In: Kuhl ME, Steiger NM, Armstrong FB, Joines JA (eds) Proceedings of the 2005 winter simulation conference, Florida, 2005
15. Pujowidianto NA, Lee LH, Chen CH, Yap CM (2009) Optimal computing budget allocation for constraint optimization. In: Rossetti MD, Hill RR, Hohansson B, Dunkin A, Ingalls RG (eds) Proceedings of the 2009 winter simulation conference, Texas, 2009
16. Rai S (2008) Fat tail inputs in manufacturing systems. In: Flower J, Mason S (eds) Proceedings 2008 industrial engineering research conference, Norcross, GA
17. Rai S, Duke CB, Lowe V, Trotter CQ, Scheermesser T (2009) LDP lean document production -O. R. - enhanced productivity improvements for the printing industry. *Interfaces* 39: 69–90
18. Sandeman T, Stanford C, Fricke C, Bodon P (2010) Integrating optimization and simulation a comparison of two case studies in mine planning". In: Johansson B, Jain S, Montoya - Torres J, Hagan J, Yücesan E (eds) Proceedings of the 2010 winter simulation conference, Maryland, 2010
19. Szechtman R, Yücesan E (2008) A new perspective on feasibility determination. In: Mason SJ, Hill RR, Moench L, Rose O, Jefferson T, Flower JW (eds) Proceedings of the 2008 winter simulation conference, Florida, 2008
20. Yue Y, Marla L, Krishnan R (2012) An efficient simulation based approach to ambulance fleet allocation and dynamic redeployment. In: proceedings of the 26th AAAI conference on artificial intelligence, Toronto, Ontario, Canada, 2014
21. Zeng Q, Yang Z (2009) Integrating simulation and optimization to schedule loading operations in container terminals. *Computers and Operations Research* 36: 1935–1944