

BOTTLENECKS IN PRODUCTION NETWORKS: AN OVERVIEW*

Yongcai WANG Qianchuan ZHAO Dazhong ZHENG

*Center for Intelligent and Networked Systems, Department of Automation,
Tsinghua University, Beijing, P.R.China, 100084
wangyongcai@mails.tsinghua.edu.cn*

Abstract

Bottlenecks, the key ingredients for improving the performances of the production networks, have been profoundly studied during the last decade. Yet, because of the complexity of the research results, there is still a significant gap between theory and practice. In this paper, we review various bottleneck definitions, detection methods and the asymptotic results and provide a practical guidance for recognizing and utilizing the bottlenecks in production networks. Queueing theory works as the mathematical foundation in our study. Various definitions of the bottlenecks are classified as either Performance in Processing (PIP) based or sensitivity based definitions, which reflect the preferences of the managers. Detection methods are surveyed closely based on the definitions. These methods are used to recognize the bottlenecks and to provide diagnosis results to managers. Comparisons show that different detection methods may lead to vastly different conclusions. The recognition of the bottlenecks has another advantage: the ultimate phenomena of the bottlenecks can greatly reduce the computation complexity in calculating the system performances. Bottlenecks based approximation and asymptotic results are studied to exhibit the contribution of bottlenecks in performance estimation and theoretical analysis.

Keywords: Bottleneck definition, bottleneck detection method, asymptotic result, queueing theory, production network

1. Introduction

Performances of a production system, such as the throughput, the circle time and the average delay, etc., are affected by the capacities of machines and resources available in the system. Some of them may affect the system

performances more than others. Usually, the limitation of a system can be traced to the limitation of one or two machines or one or two kinds of resources, commonly called bottlenecks. From system point of view, bottlenecks are the congestion points of the system, which slow

* This work was supported by NSFC Grant. No. (60074012,60274011) and NCET-04-0094 program.

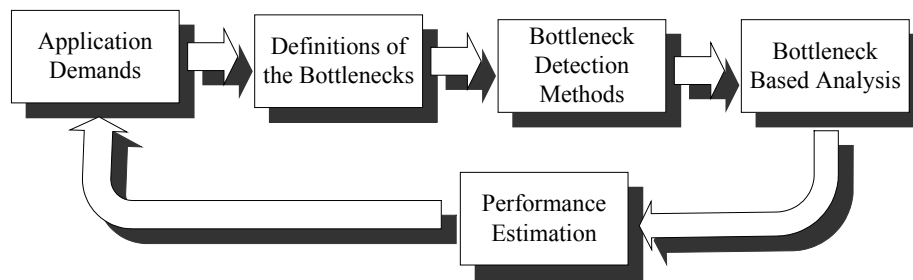


Figure 1 Bottlenecks improvements iteration

down the whole operation chain. In order to improve the performances of the system, it is necessary to improve the bottlenecks. Yet, recognition and improvement for bottlenecks is not a trivial task. In design and research, we cannot directly ask for the knowledgeable employees, as suggested by Cox and Spencer (Cox and Spencer 1997). We need to work on numerous floor data or log data to recognize the bottlenecks and to improve the bottlenecks. The iteration of improving bottlenecks is shown in Figure 1. Bottlenecks are defined based on the application demands. They are recognized by different detection methods. Then, the bottleneck-based approximation and asymptotic methods are applied to estimate the performances of the system. Bottlenecks are improved by adjusting the system parameters. This iteration will be repeated until the application demands are satisfied.

Many factors of a system contribute to the bottlenecks, such as the machine capacity and the number of operators. The bottleneck of a system may be different from different perspective of view and may be different for different class of customers. It becomes much complex for large systems. Numerous efforts have been made in the last decade and various

definitions, detection methods, approximation and asymptotic results have been presented. But there is still no commonly accepted definition or detection technique. This is mainly due to the diversity of the bottlenecks in different application scenarios. It brings difficulties in applying theoretical results to real applications. It is necessary to make clear which definition, detection method and asymptotic result are suitable for an specific application scenario.

In this paper, the main causes of what contribute to a bottleneck and how to define, recognize and utilize the bottlenecks in production networks are reviewed from an application point of view. Queueing theory works as the mathematical foundation in our study. Various definitions are classified into two primary categories: performance in processing (PIP) based and sensitivity based definitions. The former class emphasizes the real-time performance of the system, and the latter focuses more on the potential improvements. Detection methods are reviewed based on the definitions. They are classified into measurement based methods and sensitivity based indicators. Comparisons of these detection methods are presented, which show that different detection methods may lead to vastly different conclusions.

We use an example production system to demonstrate the difference of the various detection methods. Bottleneck based approximation and asymptotical algorithms are summarized. They give fast estimations of system performances. Base on the overview of these results, we propose guidelines for proper selection of the practical definition, recognition method, and approximation method of bottlenecks for industry applications.

The organization of this paper is as following: Section 2 describes the various definitions of the bottlenecks. Section 3 discusses the bottleneck detection methods. Section 4 reviews the bottleneck based approximation and asymptotical results. Conclusion and possible future work are presented in Section 5.

2. Definitions of Bottlenecks

Bottlenecks are generally recognized as some resources or utilities, which heavily limit the performances of a production system. For different application demands and different operation manners, numerous definitions of what contributes to a bottleneck can be found in the literature. But, there is still not a consensus definition of bottlenecks. Several basic definitions are summarized by (Lawrence and Buss 1995):

1. Congestion points occur in product flowing.
2. The resource whose capacity is less than the demands placed upon it.
3. A process that limits throughput.
4. Temporary blockades to increased output.
5. A facility, operator etc., that impedes production...
6. Any operation that limits output...

From these definitions, we can see the

diversity of the bottlenecks. They are not only caused by the physical constraints, such as resource, process, facility etc, but also influenced by the function, operator, etc. Some bottlenecks may appear temporarily and some may remain static. A common sense of bottleneck is “something” that limits system’s production rate. But the bottlenecks are not identical from different point of views. In this section, various bottleneck definitions will be introduced. We classify these definitions into two categories: PIP based and sensitivity based, aiming at providing a practical guidance for application perspective.

2.1 PIP Based Definitions

PIP(Performance in Processing) based definitions define system bottlenecks according to the system performance measurement. In PIP definitions, measuring of average waiting time and capacity workload (utilization) are important results.

2.1.1 Measuring the Average Waiting Time

When measuring the average waiting time, the machine with the longest average waiting time is considered to be the bottleneck (Pollett 2000).

$$B = \{i \mid W_i = \max(W_1, W_2, \dots, W_n)\} \quad (1)$$

In Equation (1), W_i is the mean waiting time of products in the i th machine. For the Little’s law, measurement of average queue length is also within this category. This method is suitable for analyzing networks with unlimited intermedial buffers. For systems containing only limited buffers and systems without buffers, it is not a suitable choice. If several machines have the same largest waiting time, this method can not

determine the unique bottleneck.

2.1.2 Measuring the Average Utilization

The machine with the largest busy/idle ratio is considered as the bottleneck (Knessl and Tier 1998, Schweitzer and Serazzi 1993), with average utilization measuring method.

$$B = \{i \mid \rho_i = \max(\rho_1, \rho_2, \dots, \rho_n)\} \quad (2)$$

In Equation (2), ρ_i is the utilization of the i th machine. $\rho_i = \lambda_i / \mu_i$ in which λ_i μ_i are the product arriving rate and service rate of the i th machine respectively. As more than one machine may have a similar workload, the difference between the utilizations of the machines may be very small. Although this method is easy to automate, it may result in multiple bottlenecks. The bottleneck detection method of Berger (Berger and Bregman 1999) investigated all possible combinations of bottlenecks, which rapidly became more complicated for larger systems.

2.2 Sensitivity Based Definitions

Another way to define the bottleneck is to find the machine whose throughput mostly affects the overall system throughput. The sensitivity of the system performance to the perturbation of machine parameters is used as the measurement.

2.2.1 Production Bottleneck

(Chiang and Kuo et. al. 1998, 1999, 2000, Kuo and Lim 1996) used a system theoretic approach to determine the sensitivity of the machine throughput to the system throughput. They studied this problem in a Markovian production line. The production rate is the average number of parts produced by the last

machine, and it is a function of all machine and buffer parameters:

$$\overline{PR} = \overline{PR}(p_1, r_1, \dots, P_m, r_m, N_1, \dots, N_{m-1}, c_1, \dots, c_m) \quad (3)$$

in which, N_i is the buffer size before the i th machine, c_i is the circle time. The uptime and the downtime of each machine m_i are random variables distributed exponentially with parameters p_i , r_i respectively. Three kinds of bottlenecks are presented. The definition of *up-time bottleneck(UT-BN)* was given in (Chiang and Kuo 1998). If

$$\frac{\partial PR}{\partial T_{up_i}} > \frac{\partial PR}{\partial T_{up_j}}, j \neq i \quad (4)$$

then m_i is the up-time bottleneck(*UT-BN*). They also gave the definitions of *down-time bottleneck(DT-BN)*: If

$$\left| \frac{\partial PR}{\partial T_{down_i}} \right| > \left| \frac{\partial PR}{\partial T_{down_j}} \right|, j \neq i \quad (5)$$

In (5) m_i is the down-time bottleneck(*DT-BN*). Absolute values are used here because $\frac{\partial PR}{\partial T_{down}}$ is negative: increase in T_{down} leads to a decrease of PR . Machine m_i is bottleneck(*BN*) if it is both *UT-BN* and *DT-BN*. Bottleneck definition based on the sensitivity to the machines' circle time was given in (Chiang and Kuo 1999). A machine is *c-bottleneck* if

$$\frac{\partial \overline{PR}}{\partial c_i} > \frac{\partial \overline{PR}}{\partial c_j}, j \neq i \quad (6)$$

Then m_i is defined as *c-bottlenecks(c-BN)*. Besides *UT-BN*, *DT-BN* and *c-BN*, another definition of bottleneck based on the sensitivity to the production rate was given by (Kuo and Lim 1996). A machine is the bottleneck if the sensitivity of the system performance index to its production rate in isolation is the largest, as

compared to all other machines. m_i is the bottleneck if:

$$\frac{\partial \overline{PR}(p_1, \dots, p_m, N_1, \dots, N_m)}{\partial p_i} > \frac{\partial \overline{PR}(p_1, \dots, p_m, N_1, \dots, N_m)}{\partial p_j}, \forall i \neq j \quad (7)$$

In (7), p_i is the production rate of the i th machine in isolation. Note that these definitions are not mutually exclusive and that a particular work center may satisfy one or more of them at any given time. Since none of these definitions considers costs, revenues, or profitability of the firm, but focuses solely on the output of the process, they are referred to *production bottlenecks*.

2.2.2 Economic Bottleneck

In contrast, when cost and revenues are taken into account, bottlenecks are identified as those resources which limit profitability. Such resources are called economic bottlenecks (Lawrence 1995). Below is a formal definition of an *economic bottleneck*. Consider a production facility with existing capacity levels μ and fixed production demands λ . The *congestion costs* function $F(\mu)$ was assumed to be proportional to the queue length of each workcenter. It was defined as the long-run flow cost in a unit time:

$$F(\mu) = \sum_{k=1}^m \frac{F_k \gamma_k}{\mu_k - \gamma_k} \quad (8)$$

In the equation, F_k is the unit cost for a job staying in the k th work-center for a unit time. γ_k and μ_k are the arrival rate and service rate of the k th machine respectively. Since

$\frac{F_k \gamma_k}{\mu_k - \gamma_k}$ is the average queue length of machine

k , $F(\mu)$ is proportional to the queue length of the work-centers. If any machine is congested, the congestion cost will increase immediately with the increasing of the queue length. Based on the definition of the congestion cost, the *economic bottleneck* is defined as the station for which marginal increases in capacity provides the largest decrease in congestion costs; that is, the work center k for which

$$\left| \frac{\partial F(\mu)}{\partial \mu_k} \right| > \left| \frac{\partial F(\mu)}{\partial \mu_j} \right|, \forall j \neq k \quad (9)$$

Since $F(\mu)/\mu_k$ is negative, from Equation (8), the definition of *economic bottleneck* can be further formulated as, the j th station which satisfies:

$$-\frac{F_k \gamma_k}{(\mu_k - \gamma_k)^2} < -\frac{F_j \gamma_j}{(\mu_j - \gamma_j)^2}, \forall j \neq k \quad (10)$$

So far, we have discussed various definitions of bottlenecks. In the next section, we will introduce the bottleneck detection methods based on these definitions.

3. Bottlenecks Detection Methods

Definitions of bottlenecks integrates the application demands and the system's real-time behaviors to define conditions for the bottlenecks. Detection method processes the observed factory data or simulation data to locate who satisfies these conditions. Systems' diversity and the state explosion of the production network make accurately recognition of bottlenecks difficult in large systems. Various

PIP Based Detection Methods	Measuring Average Waiting Time	Measuring Average Workload	Measuring The Average Active Duration
	Law and Kelton, 1991 George, et.al , 1999 Pollett, et.al , 2000	Law and Kelton, 1991 Luthi and Haring, 1997 Berger, et.al, 1999 Casale and Serazzi, 2003	Roser, et.al, 2001 Roser, et.al, 2003
Shift Bottlenecks Detection	Roser, et.al, 2002a Roser, et.al, 2003	Roser, et.al, 2002b	Roser, et.al, 2002
Sensitivity Based Detection	Kuo, et.al , 1995 Chiang, Kuo, Meekov, 2000	Chiang, et.al , 1998	Chiang, et. al, 1999
Other Methods	Cox and Spencer 1997 Jibiki, et.al, 1999	Dina, et.al, 1997 Pollett, 2000	Wang, et.al, 1997 Luthi, et.al, 1998 Delp, et.al, 2003

Figure 2 Summarizing of bottleneck detection methods

efforts have been devoted in this area in the last decade. From an implementation point of view, we summarized these bottleneck detection methods in Figure 2, providing a clear classification to aid selection in practice.

3.1 PIP (Performance In Processing) Bottleneck Detection

Corresponding to PIP based definitions of bottlenecks, evaluating PIP using simulation is an important bottleneck detection method. Within this category, there are different branches.

3.1.1 Measuring the Average Waiting Time

Measuring the average waiting time and recognizing the machine with the longest waiting time to be the bottleneck is described by (Law and Kelton 1991). Measuring of the queue length (Pollett 2000) is within the same category according to the Little’s formula. Response time is the sum of waiting time and processing time.

Checking for the maximum average per-hop delay (Elmasry and John 2000) is based on the same idea. Measuring the average waiting time, average queue length, and average per-hop delay are intuitive and easy in implementation, but they have the same drawbacks. The accuracy of this approach is compromised if the system contains buffers of limited size. Furthermore, this approach analyzes only the processing machines of the manufacturing system. Other elements, for example the supply and demand, or human workers, do not have a buffer in the classical sense and require additional consideration or may not be considered at all.

3.1.2 Measuring the Average Workload

When measuring the workload, the machine with the largest workload (utilization) is considered as the bottleneck (Law and Kelton 1991). Yet, as more than one machine may have a similar percentage of being active, the difference between the workloads of the

machines may be very small. Since workload measurements may have errors due to the random variation of the data, it is often hard to decide which entity is the bottleneck. While this method is easy to automate, the results are not always accurate. An approach was described in (Luthi 1998) to determine the likelihood of multiple bottlenecks based on the percentage of the time the machines were active using a bottleneck probability matrix. The bottleneck detection method from (Berger and Bregman 1999) also investigated all possible combinations of bottlenecks, which rapidly became more complicated for larger systems. Recently, convex polyhedra based bottleneck detection method was proposed in (Casale and Serazzi 2003) according to the workload matrix $L_{m,n}$ of the customers, where m was the number of customer classes, and n was the number of workstations. A variant of convex hull algorithm was proposed, which was polynomial time complexities of m and n .

3.2 Measuring the Average Active Duration

When measuring the active duration, the machine with the longest average active period is recognized as the bottleneck (Roser and Nakano et.al. 2001). The *active* state of machine is different from traditional *busy* concept. All activities towards improving the system throughput, including repair and service states are active states. For example, work, repair, tool change, etc are all active states. In (Roser and Nakano et.al. 2001), a bottleneck detection method was proposed to determine bottlenecks by measuring the longest average consecutive active duration of machines. With simulation results in a serial production line, they showed

that the proposed method can more accurately detect the bottleneck based on the sensitivity definition. So, by PIP measurements, not only PIP based bottlenecks can be detected, sensitivity based bottlenecks can also be detected by the appropriate use of the simulation log data. Simulation results indicated that active duration based recognition can be used in AGV systems (Roser and Nakano et.al. 2003).

3.3 Shift Bottleneck Detection

The active duration based recognition method was further developed and a shift bottleneck detection method was proposed. The method has been proven to work reliably for non-AGV systems (Roser and Nakano et.al. 2002) and is further developed to work in AGV systems (Roser and Nakano et.al. 2002). The method recognizes the machine or AGV with the longest active period as the bottlenecks, and further distinguishes them as shifting bottlenecks and sole bottlenecks. Active period of shifting bottlenecks overlaps with the next bottleneck. Sole bottlenecks do not overlap with previous or subsequent bottlenecks. Figure 3.3 shows an example of a two-machine system, where at the beginning Machine M1 has the longest active period, and therefore is the bottleneck. Later, the bottleneck shifts from Machine M1 to M2, and then M2 is the sole bottleneck. The likelihood of a machine being the bottleneck can be measured easily by determining the percentage of the time a machine is a sole or shifting bottleneck. In (Roser and Nakano et.al. 2001), the author showed that the shifting bottleneck method could accurately detect the sensitivity based bottlenecks, verified by simulation results.

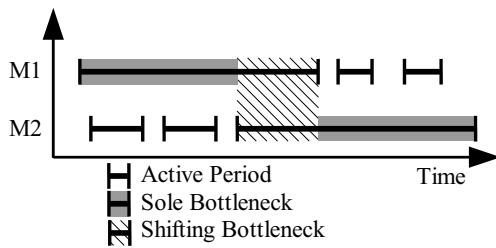


Figure 3 Shifting bottlenecks

3.4 Sensitivity Bottleneck's Indicator

Although the sensitivity based definitions for bottlenecks seem appealing from the systems point of view, these definitions suffer a deficiency due to the fact that the sensitivities involved cannot be either measured on-line or efficiently calculated off-line. Several methods are proposed to tackle this problem. In (Kuo and Lim 1996), the bottleneck machine in a serial production line is identified by analyzing relationships between the so-called manufacturing blockage and manufacturing starvation of each machine. The DT-BN Indicator was proposed by (Chiang and Kuo 1998) based on the data available on the factory floor through real-time measurement (such as average uptime and downtime, starvation and blockage time, etc.) or on the data that can be constructively calculated using the machine and buffer parameters. An inequality function was

proposed as DT-BN indicator for two machines production system. The indicator was further developed in (Chiang and Kuo et. al. 2000). Directed arrows were assigned to the machines, according to their flowing conditions. And Bottleneck severity was proposed to measure the level of congestion. Based on the arrows assignment rules, they proposed the DT-BN Indicator, which could be used to detect the bottleneck in serial production systems. They also developed the c-bottleneck indicator (Chiang and Kuo et. al. 1999) corresponding to the c-bottleneck definition with similar flowing condition function. But currently, the bottleneck indicators can only be applied in serial Markovian production lines. Every machine must have exponentially distributed service time and the system should be a serial line without operators and AGVs.

3.5 Comparison of the Bottleneck Detection

Methods

After introducing various definitions and detection methods of bottlenecks, we will use an example to compare the different detection methods. A production system with automatically guided vehicles (AGV) is used in the comparison. The presented system consists of two machines and three AGVs as shown in

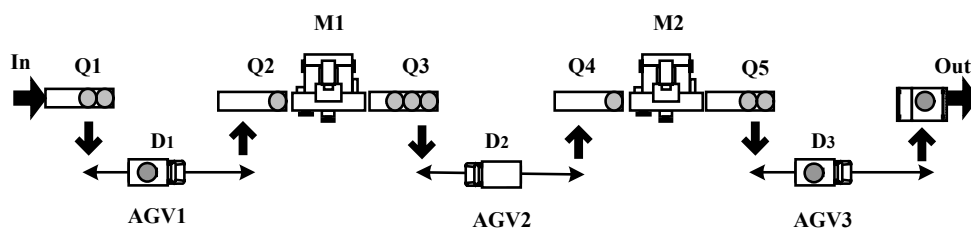


Figure 4 Example system with two machines and three AGV's

Table 1 Machine parameters

Machine	Cycle	Time	MTBF	MTTR
In	180	0	0	
M1	120	1000	50	
M2	153	1000	50	

Table 2 AGV parameters

From	To	Distance(m)	Time
In	M1	11500	76.7
M1	M2	10000	66.7
M2	Out	12000	80

Table 3 Measuring average waiting time

Mean	Waiting Time	Queue length
AGV1	13.52	4.80
AGV2	0.09	0.03
AGV3	1.17	0.40
M1	0.18	0.06
M2	7.72	2.72

Table 4 Measuring the average workload

Mean	Working	Repair	Utilization
AGV1	88.9%	-	88.9%
AGV2	77.3%	-	77.3%
AGV3	91.5%	-	91.5%
M1	69.6%	4.68%	74.3%
M2	83.2%	6.95%	90.1%

Figure 4. Products arrive from the “In” station and are carried by AGV1 to the first machine. After being processed by M1, products are carried by AGV2 to the second machine. After processed by M2, they are carried by AGV3 to the “Out” station. The supply and demand of products at the “In” and “Out” stations are infinite.

Table 1 shows the arrival rate of the products

and the machine parameters. The time between arrivals is exponentially distributed with mean 180 minutes. Each machine has a deterministic cycle time and randomly occurring failures. The time between failures and the time to repair are exponential distributed. The mean time between failures (MTBF) is 1000 minutes and the mean time to repair (MTTR) is 50 minutes for M1 and M2. Table 2 shows the distances the AGV has to travel between the stations and the travel time with a speed of 150m/minute.

The simulation was implemented using the Arena simulation software and run for 2000 hour simulation time and with 500 hour warm-up time. To detect the bottleneck, the average waiting time method measures the waiting time of products to determine the bottleneck. An alternative method may look for the longest queue instead of the longest waiting time. In our example, the waiting time of product and queue length were measured and shown in Table 3. Averagely, Q1 has the maximum queue length and products will wait the longest time in Q1. AGV1 is detected as the bottleneck with this method.

The second bottleneck detection method, workload method measures the utilization of the machines and AGVs and defines the machine/AGV with the largest utilization as the bottleneck. In our example the utilization is measured and is shown in Table 4. AGV3 has the largest utilization and is detected as the bottleneck. An interested phenomenon is that although products wait the longest time in Q1, AGV1 is not the busiest. This can be easily understood, since the time interval between arrivals is exponential distributed, while the

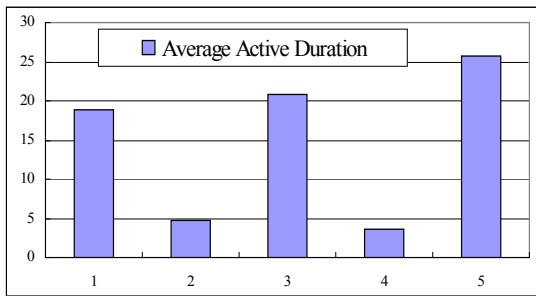


Figure 5 Measuring the average active duration

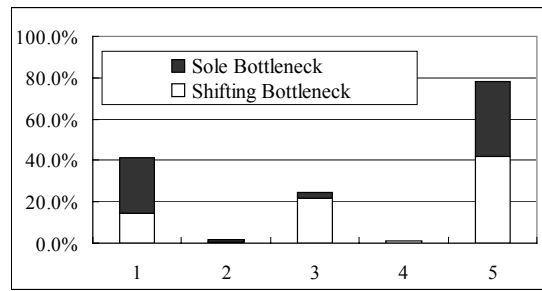


Figure 6 Shifting bottleneck detection technique

Table 5 Comparison of bottleneck detection methods

Detection Methods	Average Time	Waiting	Average Workload	Average Duration	Active	Shifting Bottleneck
Detected Bottleneck	AGV1		AGV3	M2		M2

cycle time of AGV1 is deterministic. When the products arrive in batch, the queue length of Q1 will increase immediately. When products arrive less frequently, AGV1 trends to be idle.

The third bottleneck detection method defines the bottleneck as the machine or transporter with the longest average active duration. An active duration is the time interval between two idle states. This method is different from the utilization method, since it measures the length of the active duration instead of the total proportion of the active time among the simulation time. The measured average active duration is shown in Figure 5 and M2 is detected as the bottleneck.

The shifting bottleneck detection method also measures the active duration of machines. Instead of simply calculating the average length of the active durations, it further distinguishes at any time which active duration limits the system

performance furthest. The definitions of the shifting bottleneck and sole bottleneck are the same as in Section 3.4. The percentage of the time a machine/transporter is a sole or shifting bottleneck is measured and shown in Figure 6. M2 has the maximum likelihood to be a bottleneck.

Table 5 summarizes the bottleneck detection results. We can see vastly different conclusions are drawn by different bottleneck detection methods.

So, which machine or AGV is really the bottleneck? In experiments, we experimentally check which machine or AGV has the maximum sensitivity. We reduce the cycle time of machines and the transfer distances of AGVs independently to 95% of their origin values and run the simulation to calculate the sensitivity. The mean response time of products is used as performance metric and the sensitivity of

Table 6 Measured machine/AGV sensitivities

Mean	Transport time of AGV1 (minutes)	Transport time of AGV2 (minutes)	Transport time of AGV3 (minutes)	Cycle time of M1 (minutes)	Cycle time of M2 (minutes)
Original value	76.7	66.7	80	120	153
Mean Response Time	1852	1852	1852	1852	1852
Improved value	72.8	63.3	76	114	145.4
Mean Response Time	1157	1777	1246	1736	1143
Sensitivity	231.8	25.0	202.0	38.6	236.2
Mean	Transport time of AGV1 (minutes)	Transport time of AGV2 (minutes)	Transport time of AGV3 (minutes)	Cycle time of M1 (minutes)	Cycle time of M2 (minutes)
Sensitivity	231.8	25.0	202.0	38.6	236.2

machines and AGVs are shown in Table 6. The results show that M2 has the maximum sensitivity, coinciding with the results of the average activation duration method and shifting bottleneck method.

The presented example shows that different bottleneck detection methods may draw vastly different conclusions. The average workload method and the average waiting time method can accurately detect the PIP based bottleneck, while the average active duration method and the shifting bottleneck method can more accurately detect the sensitivity based bottleneck.

3.6 Other Methods

An analytical method was described in (Pollett 2000) allowing to recognize regions of congestion in closed Markovian queueing networks. Given flow chart and layout of system, to detect the bottleneck by analyzing the structure of system is another proposed method (Cox and Spencer 1997). However, this is a complex manual task, difficult to automate, and applicable only to simple systems. Dynamic

system regulation method was proposed by (Delp and Hwang et. al. 2003), using the ratio of the cycle time divided by the processing time to determine the bottleneck (X-factor). They studied the relationship between the X-factor and bottlenecks, and systematically studied the relationships among processing time, utilization, X-factor and bottlenecks. Matrix based approach was used in (Luthi 1998) to determine the overall system constraint. They use VU-list, which was an interval to model the workload associated with uncertainty and/or variability and propose interval-based bottleneck identification matrices. Artificial Neural Network (ANN) was proposed in (Jibiki 1999) to detect the bottlenecks of large scale networks, such as production systems and internet. Using the traffic information collected by appropriate observation points, the ANN was trained and non-linear optimization is applied to recognize the congestion point of system. The bottlenecks in oil refinery industry were studied by (Wang and Chen et.al 1997).. Maximum production network was defined and a bottleneck detection algorithm for maximum production network was

presented. A passive approach for detecting shared bottlenecks in the internet was presented by (Katabi and Bazzi et.al. 2001). By measuring the time difference between consecutive packets, a passive observer can cluster the flows into groups, such that by analyzing the properties of groups, the bottlenecks are identified.

4. Bottleneck Based Analysis

A distinguishing advantage of bottleneck analysis over other solution techniques is its limited computational complexity. This is because in limitation condition of heavy traffic, flow of production system has some typical features: the traffic intensities at some queues are sufficiently high, while the traffic intensities of all the other queues are substantially lower. This phenomenon is formulated as heavy-traffic limit theorem, and can greatly reduce the computation complexity of bottleneck based analysis. But due to the explosion of state space for large systems, the dynamic of system's behavior and the difficulty of congestion analysis, theoretical analysis is mainly on the stage of approximation and asymptotic analysis. In this section, we will briefly introduce these results as summarized in figure 4, emphasizing the function and the evolutionary process.

4.1 Approximation Results

Heavy-traffic limit theorem was proposed by (Iglehart and Whitt 1970, Reiman 1983, Reiman 1984, Chen and Mandelbaum 1988), indicating that for a system with bottlenecks, the standard steady-state random variables such as the waiting time at each queue are distributed nearly the same (relatively to the level of congestion at the bottleneck queue) as if all the service times

in the non-bottleneck queues were set equal to 0. In (Suresh and Whitt 1990), it was showed that if the traffic intensity of one queue was allowed to approach 1, then the waiting-time distribution at the bottleneck queue was approximately the same as if the immediate arrival process were replaced by the external arrival process to the first queue with squared coefficient of variation $c^2 = 1$. A switching approximation method was proposed in (Suresh and Whitt 1990), to analyze the performances of serial production system. The variation of arrival for bottleneck machines and non-bottleneck machines were calculated separately using different formulas. After calculating the first two moments of the arrival process to every machine and the first two moments of service time of each machine, the mean waiting time of product at each machine can be calculated. Based on this, The queueing network analyzer (QNA) was proposed in (Whitt 1991), which can approximately analyze the queueing system performances with assigned parameters. Motivated by the heavy traffic theorem, QNET method was proposed in (Harrison and Nguyen 1990, Dai and Harrison 1993). The basic idea of QNET method is to compute the stationary distribution of an approximating Brownian model. Not only the first moment information, but also the second moment information can be approximately calculated. (The sequential bottlenecks decomposition (SBD) method was proposed in (Dai 1994). It is based on the heavy traffic limit theorem and QNET method. k-dimension Brownian motion was formulated based on workload similar sub-networks partition, then the sub-networks were analyzed sequentially with a variant of the QNET method. The

performance of QNA, QNET and SBD were also compared, which showed that SBD was generally better than QNA and QNET in approximating the performances of generalized Jackson networks. Another approximation proposed in (Bolch and Greiner et. al. 1998) is called Bottapprox (BOTT). It is an iterative method, which is an improvement over MVA method. Using BOTT, the initial value of throughput can be selected based on the bottleneck's throughput, and this reasonable guess can effectively reduce the number of iterations of approximation methods.

4.2 Asymptotic Analysis

Asymptotic analysis is important both for the theoretical aspects and the practical aspects in production system analysis. Asymptotic formulas often clearly show how the system behaves in terms of variables or parameters in the model, and can be used to assess and improve system's performance based on the understanding of system's behavior. For early works of asymptotic bottleneck analysis in close queueing networks, a complete survey was given in (Schweitzer and Serazzi 1993). In (Knessl and Tier 1998), the asymptotic approximations in multi-class queueing networks with large populations were studied. The perturbation scheme in (Schweitzer and Serazzi 1993) was re-examined and a new analysis to correct the non-uniformities when there were multi bottleneck nodes was proposed. After that, the asymptotic analysis has been extended to multi-class queueing networks with multi bottleneck nodes. Due to the difficulty in studying unlimited condition, existing results have mainly focused on the asymptotic condition of the network when the population

size trending infinity. In (Balbo and Serazzi 1996), the author presented computationally simple formulas for immediate calculation of the performance indices of per-class behavior of the network based the assumption that in these networks all the classes of customers identify a single station as their bottleneck. The results were further extended in (Balbo and Serazzi 1997) to more general multi-class queueing networks, allowing each class identifies a different station as the potential bottleneck of the network. A new technique for computing the normalization constant of closed models was shown in (Harrison and Coury 2002), which can help in understanding the asymptotic behavior of the network. A number of theoretical estimations of the system performance were compared in (Bukchin 1998). Comparison results showed that an estimator based on the machine bottlenecks works best. Recently, convex polytopes method was proposed (Casale and Serazzi 2004) to identify the bottlenecks in multi-class queueing networks, and also proposed the asymptotic analysis algorithm for computing performance indices for the closed product form queueing networks. The algorithm has polynomial time complexity in the number of classes and in the number of stations, but the construction of convex polytopes needs pre-obtained workload data.

5. Conclusion

In this paper, a large number of approaches studying the bottlenecks in production networks have been discussed and analyzed. We restrict our attention in three aspects: the definitions, detection methods, and approximation and asymptotic results of bottleneck based analysis. We summarize current results from practical

point of view, aiming at providing a useful guidance in the application. Our observations can be summarized as follows:

- ◆ Definitions of the bottlenecks rely on the application preferences, which can be summarized into two categories: PIP based and Sensitivity based. The former emphasizes realtime performances and the latter pays more attention to potential improvements.
- ◆ Detection methods are feature excavating methods to locate what resources or utilities satisfy the bottleneck conditions, which are closely based on the definitions. Neither PIP based detection methods, nor the sensitivity based indicators can provide a universal solution. Different detection methods may have vastly different results. For application, detection method should be selected according to application demands.
- ◆ Benefiting from the heavy traffic limit behavior, bottleneck based analysis give fast system's characters analysis with less

computation effort. Approximation methods can be applied for quick performance estimation, and asymptotic formulas can help the understanding of the systems' behavior in terms of parameters changing. For complex networks, it is more important to understand the parameter effects, which is often computation consuming for simulation, and bottleneck based analysis greatly advances in this aspect.

Although bottlenecks in production systems have been studied for more than ten years, future directions are still promising. A common definition is possible by parameterizing the demands, and the detection methods may provide more accurate and reliable detection with the advantage of the data mining technology and computing powers. Bottleneck based analysis will be more important in understanding complex system's behavior, and will probably be extended to the internet traffic scheduling and wireless channel allocation.

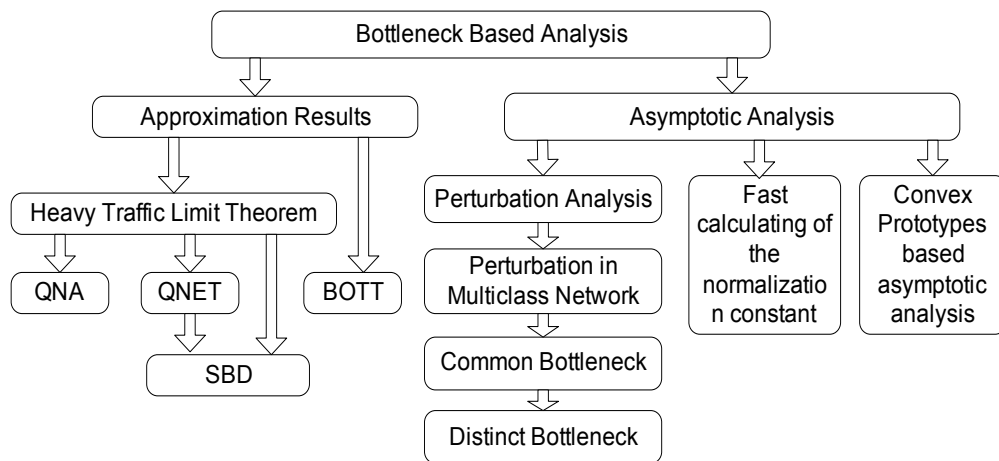


Figure 7 Bottleneck based analytical results

References

- [1] Balbo G., G. Serazzi, "Asymptotic analysis of multiclass closed queueing networks: common bottleneck", *Performance Evaluation*, Vol.26, No.1, pp51-72, 1996.
- [2] Balbo G., G. Serazzi, "Asymptotic analysis of multiclass closed queueing networks: multiple bottlenecks". *Performance Evaluation*, Vol.30, No.3, pp115-152, 1997.
- [3] Berger A., L. Bregman, et. al., "Bottleneck analysis in multiclass closed queueing networks and its application", *Queueing Systems*, 31(3-4), pp. 217-237, 1999.
- [4] Bolch G., S. Greiner, et. al., *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, John Wiley and Sons, 1998.
- [5] Bukchin J., "A comparative study of performance measures for throughput of a mixed model assembly line in a JIT environment", *International Journal of Production Research*, Vol.36, No.10, pp2669-2685, 1998.
- [6] Casale G., G. Serazzi, "Estimating bottlenecks of very large models", *Performance Evaluation Stories and Perspectives-G.Kotsis Editor*, Austrian Computing Society, pp89-104, 2003.
- [7] Casale G., G. Serazzi, "Bottlenecks identification in multiclass queueing networks using convex polytopes", *In Proc. IEEE/ACM MASCOTS 2004, IEEE Comp. Soc.*, pp223-230, 2004.
- [8] Chen H., A. Mandelbaum, "Stochastic discrete flow networks: diffusion approximations and bottlenecks", Graduate School of Business, Stanford University, 1988.
- [9] Chiang S. Y., C. T. Kuo, et. al. "DT-bottlenecks in serial production lines: theory and application", *IEEE Transactions on Robotics and Automation*, Vol.16, Issue 5, pp567-580, 2000.
- [10] Chiang S. Y., C. T. Kuo, et. al., "Bottlenecks in Markovian production lines: a systems approach", *IEEE Transactions on Robotics and Automation*, Vol. 14, Issue 2, pp352-359, 1998.
- [11] Chiang S. Y., C. T. Kuo, et. al., "c-bottlenecks in serial production lines: identification and application", *Proceedings of the 38th IEEE Conference on Decision and Control*, Vol.1, pp456-461, 1999.
- [12] Chiang S. Y., C. T. Kuo, et. al., "Bottlenecks in Markovian production lines: identification and application", *Proceedings of the 37th IEEE Conference on Decision and Control*, Vol.4 pp4348-4349, 1999.
- [13] Cox F. J., M. S. Spencer, *The Constraints Management Handbook*, Boca Raton, Florida: CRC Press - St. Lucie Press, 1997.
- [14] Dai J. G., "Sequential bottleneck decomposition: an approximation method for generalized Jackson networks", *Operations Research*, Vol.42, No.1, pp119-136, 1994.
- [15] Dai J. G., J. M. Harrison, "The QNET method for two moment analysis of closed manufacturing systems", *Annals of Applied Probability*, Vol.3, pp968-1012, 1993.
- [16] Delp D., J. Si, Y. Hwang, et. al., "A dynamic system regulation measure for increasing effective capacity: the X-factor theory", *Advanced Semiconductor Manufacturing Conference and Workshop*, pp81-88, 2003.
- [17] Elmasry G. F., C. John, et. al., "Bottleneck

- discovery in large scale network based on the expected value of per-hop delay”, Tech. Report, 2000.
- [18] Harrison P., S. Coury, “On the asymptotic behaviour of closed multiclass queueing networks”, *Performance Evaluation*, Vol.47, No.2-3, pp131-138, 2002.
- [19] Harrison J. M., V. Nguyen, “The QNET method for two moment analysis of open queueing networks”, *Queueing Systems*, Vol.6, pp1-32, 1990.
- [20] Iglehart D. L., W. Whitt, “Multiple channel queues in heavy traffic, ii: sequences, networks, and batches”, *Adv. Appl. Prob.* Vol.2, pp355-369, 1970.
- [21] Jibiki M., T. Terano et. al., “Comprehensive bottleneck detection via nonlinear optimization techniques”, *Internet Workshop IWS*, pp286-293, 1999.
- [22] Katabi D., I. Bazzi et. al., “A passive approach for detecting shared bottlenecks”, *In Proc. International Conference on Computer Communications and Networks*, 2001.
- [23] Knessl C., C. Tier, “Asymptotic approximations and bottleneck analysis in product form queueing networks with large populations”, *Performance Evaluation*, Vol. 33, pp. 219-248, 1998.
- [24] Kuo C. T., J. T. Lim, et. al., “Bottlenecks in serial production lines: a system-theoretic approach”, *Mathematical Problems in Engineering*, Vol.2, pp233-276, 1996.
- [25] Law A. M., D. W. Kelton, *Simulation Modeling and Analysis*. McGraw Hill, 1991.
- [26] Lawrence S. R., and A. H. Buss, “Economic analysis of production bottlenecks”, *Mathematical Problems in Engineering*, Vol.1, No.4, pp341-369, 1995.
- [27] Luthi J., “Interval matrices for the bottleneck analysis of queueing network models with histogram- based parameters”, *In IEEE International Computer Performance and Dependability Symposium*, Durham, NC, USA, pp142-151, 1998.
- [28] Luthi J., G. Haring, “Bottleneck analysis for computer and communication systems with workload variabilities and uncertainties”, *In Proc. of 2nd Intl Symposium on Mathematical Modelling*, pp525-534, 1997.
- [29] Pollett P. K., “Modelling congestion in closed queueing networks”, *International Transactions in Operations Research*, Vol.7, pp.319-330, 2000.
- [30] Reiman M. I., ”Some diffusion approximations with state-space collapse,” *Proc. Int. Seminar on Modeling and Perf. Eval. Methodology*, Springer-Verlag, Berlin, pp209-240, 1983.
- [31] Reiman M. I., “Open queueing networks in heavy traffic”, *Math. Oper. Res.*, Vol.9, pp441-458, 1984.
- [32] Roser C., M. Nakano et. al., “A practical bottleneck detection method”, *Proceedings of the 2001 Winter Simulation Conference*, pp. 949-953, 2001.
- [33] Roser C., M. Nakano et. al., “Comparison of bottleneck detection methods for AGV systems”, *Proceedings of the 2003 Winter Simulation Conference*, pp1192-1198, 2003.
- [34] Roser C., M. Nakano et. al., “Shifting bottleneck detection”, *In Winter Simulation Conference*, San Diego, CA, USA. pp1079-1086, 2002.
- [35] Roser C., M. Nakano et. al., “Tracking Shifting Bottlenecks”, *In Japan-USA*

- Symposium on Flexible Automation*, Hiroshima, Japan, pp745-750, 2002.
- [36] Roser C., M. Nakano, et. al., "Throughput sensitivity analysis using a single simulation", *Simulation Conference Proceedings of the Winter*, Vol. 2, pp1087-1094, 2002.
- [37] Roser C., M. Nakano, et. al., "Shifting bottleneck detection", *Simulation Conference Proceedings of the Winter*, Vol. 2, pp1079-1086, 2002.
- [38] Schweitzer P. J., G. Serazzi, et. al., "A survey of bottleneck analysis in closed network of queues", *Perf. Eval. of Computer and Communication Systems, Joint Tutorial Papers of Performance 93 and Sigmetrics*, Springer-Verlag, pp491-508, 1993.
- [39] Suresh S., W. Whitt, "The heavy-traffic bottleneck phenomenon in open queueing networks", *Operations Research Letters*, Vol. 9, No. 6, pp355-362, 1990.
- [40] Wang J., R. Chen et. al., "Bottleneck analysis and maximum yield solution for a class of production process", *IEEE International Conference on Intelligent Processing Systems*, Vol. 2, pp1361-1365, 1997.
- [41] Whitt W., "The queueing networks analyzer", *Bell Sys. Tech. J.*, Vol. 62, pp2779-2815, 1991.

Yongcai Wang received the B.S. degree in automatic control from Tsinghua University, Beijing, China, in 2001. He is currently pursuing the Ph.D. degree in the Department of Automation at Tsinghua University. His current research interests include bottleneck analysis in complex network systems, integrated layer

design in wireless sensor networks and performance evaluation of parallel and distributed systems.

Qianchuan Zhao received the B.E. degree in automatic control in 1992, and the B.S. degree in applied mathematics and the Ph.D. degree in control theory and its applications from Tsinghua University, Beijing, China, in 1992 and 1996, respectively. Currently, he is a Professor in the Department of Automation at Tsinghua University. He was a Visiting Scholar at Carnegie Mellon University, Pittsburgh, PA, in 2000, and at Harvard University, Cambridge, MA, in 2002. His current research interests include DEDS theory, sensor networks and the optimization of complex systems. He is an associate editor of *Journal of Optimization Theory and Applications*.

Dazhong Zheng received the diploma in automatic control from Tsinghua University, Beijing, China, in 1959. Currently, he is a Professor in control theory and engineering with the Department of Automatic Control at Tsinghua University, Beijing, China, where he has been since 1959. He was a Visiting Scholar in the Department of Electrical Engineering at the State University of New York at Stony Brook from 1981 to 1983 and from April to November 1993. His research interests include linear systems, discrete event dynamic systems, and power systems. He has published many journal papers and five books. He is also a Deputy Editor-In-Chief of *Acta Automatica Sinica*, Beijing, China. Currently, he is a Vice-Chairman of control theory technical committee for Chinese Association of Automation (CAA).