Front. Eng. Manag. 2017, 4(3): 295–303
DOI 10.15302/J-FEM-2017047

REVIEW ARTICLE

Zhe George ZHANG, Xiaoling YIN

Hierarchical modeling of stochastic manufacturing and service systems

Abstract This paper presents a review of methodologies for analyzing stochastic manufacturing and service systems. On the basis of the scale and level of details of operations, we can study stochastic systems using micro-, meso-, and macro-scopic models. Such a classification unifies stochastic modeling theory. For each model type, we highlight the advantages and disadvantages and the applicable situations. Micro-scopic models are based on quasi-birth-and-death process because of the phase-type distributed service times and/or Markov arrival processes. Such models are appropriate for modeling the detailed operations of a manufacturing system with relatively small number of servers (production facilities). By contrast, meso-scopic and macro-scopic models are based on the functional central limit theorem (FCLT) and functional strong law of large numbers (FSLLN), respectively, under heavy-traffic regimes. These high-level models are appropriate for modeling large-scale service systems with many servers, such as call centers or large service networks. This review will help practitioners select the appropriate level of modeling to enhance their understanding of the dynamic behavior of manufacturing or service systems. Enhanced understanding will ensure that optimal policies can be designed to improve system performance. Researchers in operation analytics and optimization of manufacturing and logistics also benefit from such a review.

Keywords stochastic modeling, QBD process, PH distribution, heavy traffic limits, diffusion process

Received June 6, 2017; accepted July 31, 2017

Zhe George ZHANG (✉)
Department of Decision Sciences, Western Washington University Bellingham, Bellingham, WA 98225, USA; Beedie School of Business, Simon Fraser University Burnaby, Burnaby, BC V5A 1S6, Canada
E-mail: gzhang@sfu.ca

Xiaoling YIN
School of Management, Lanzhou University, Lanzhou 730000, China

1 Introduction

A manufacturing or service system can be studied as a Markovian model because of random factors such as customer order time or amount, job processing time, and/or the reliability of machines or servers. Buzacott and Shanthikumar (1993) proposed a systematic treatment of modeling stochastic manufacturing systems. For stochastic service systems, a typical example is a large call center. Koole and Mandelbaum (2002) surveyed several stochastic models on telephone call centers. Both types of stochastic systems can be modeled as a waiting line or a queueing model of Markovian type. The main distinction between a manufacturing and a service system is usually the scale (or the size) of the system, which is measured by the number of servers. Manufacturing systems are typically involved with a small number of servers, whereas service systems, such as call centers, have a large number of servers. Such a difference entails different modeling approaches. In this paper, we briefly review the modeling hierarchy for analyzing stochastic manufacturing or service systems. We consider three levels of modeling — micro-, meso-, and macro-scopic models — and identify the advantages and disadvantages of each model type.

This paper is structured as follows: Section 2 defines the three levels of modeling and presents the related literature. Section 3 provides typical example of each modeling type in detail (for full details, we refer to the related literature). Section 4 concludes the paper.

2 Stochastic modeling hierarchy

The main approach to modeling a stochastic system is to formulate a Markovian process for the state variable of interest. For example, a production system in a make-to-order mode is a typical setting. Assuming that the production process for a product is triggered by a randomly arriving customer order, such a random arrival process is

modeled as a Poisson process. If the production time of completing the ordered product is also random, the model generally has independent and identically distributed (i.i.d.) service random variables. Considering a production system with c facilities (machines), we can model the process of fulfilling these orders as a basic M/G/c queue. We generally call each service request as a "customer" and the service provider as a "server". Numerous factors, such as customers with different priorities, servers subject to failures or absence, customer reneging or balking, and customer retrial or feedback, can definitely be considered. We intend to model the details of the system dynamics of this system type and treat the queue length or the waiting time as the main performance measure. Therefore, as long as the i.i.d. service times and/or Poisson process arrivals can be justified, we can obtain the stationary distribution of the queue length or waiting time by developing a Markovian process.

Considering that the system's scale is small (only a few servers) and the details of the system dynamics are desired, we call this type of queueing system as a micro-scopic model. This type of models is mainly continuous-time Markov chains (CTMCs). The main approach is to generalize the exponential distribution to phase-type (PH) distribution to model the time interval and to generalize the Poisson process to Markov arrival process (MAP) to model the counting (or arrival) process. With these two generalizations, we can theoretically model the G/G/c queue with any desired accuracy if the number of c servers is not large (small-scale system) because a PH-distributed random variable can estimate any arbitrarily distributed non-negative variable (like service time) and MAP can approximate any arbitrary counting process (like customer arrivals) (Neuts, 1981; He, 2014). The cost of using this approximation approach is the increase of the state space of the CTMC. For example, the basic birth-and-death (BD) process for M/M/c queue will be extended to quasi-birth-and-death (QBD) process for MAP/PH/c queue. Although the dense property of PH distribution and MAP enables us to develop the CTMC for a G/G/c queue with any desired accuracy, the "curse of dimensionality" or the explosion of the state space will make the model an NP-hard problem for a system with a large number of servers. However, given that most manufacturing systems are of small scales, the micro-scopic model still works well and produces detailed performance measure, such as the stationary distribution of the queue length for these stable systems.

As mentioned earlier, service systems such as call centers are usually of large scale with hundreds or thousands of servers. Clearly, the micro-scopic model does not work. The main issue in such a system is on how the system can be staffed to ensure that sufficient level of customer service can be achieved. One feature of this type of systems is that the traffic is heavy and most of the time servers are busy or even overloaded with customer abandonments (this condition is in sharp contrast to a manufacturing system receiving customer orders). Therefore, the adoption of heavy-traffic methods for the approximation of performance measures is an appropriate approach. Another characteristic of the service system is the time-varying arrival rate. If the system manager is only concerned with the average (first moment) performance measure such as mean queue length or mean waiting time, then the macro-scopic model with appropriate time and space scaling, also called fluid model, is appropriate. However, for a one-stage queueing system (like M/G/c type), the fluid model does not capture the random variability of the system dynamics. For a stable system (the arrival rate is less than the service rate), the fluid limit becomes zero; for an unstable system (the arrival rate is greater than the service rate), the fluid limit becomes a linear function of time with slope of $\lambda-\mu$, where λ and μ are the arrival rate and service rate, respectively. To capture the randomness of the queue length process or to stochastically refine the fluid model, the meso-scopic model needs to be considered by using the diffusion space scaling. This two-level modeling hierarchy was first proposed and discussed by Chen and Mandelbaum (1994). Heavy-traffic conditions can be identified by taking the limit via space and time scaling. Whitt (2002) presented different modes of determining the heavy-traffic limit.

The advantage of using macro- or meso-scopic models is that the approximation is increasingly improving when the system scale is becoming large and the traffic load is becoming heavy. These models will thus complement the micro-scopic models and are appropriate in analyzing large-scale service systems such as call centers or large queue networks.

3 Examples of micro-, meso-, and marco-scopic models

3.1 A micro-scopic model of a production system

As mentioned earlier, the micro-scopic model is appropriate for analyzing manufacturing systems with a small number of servers. We present a practical example in hot-rolling process for slabs in the steel industry. When a customer order arrives, it will trigger several slabs (order size) to proceed to the hot-rolling process. Given that order arrivals are random, the requests for the sets of slabs for the hot-rolling process (called jobs or customers) are also random and can be modeled as a Poisson process with arrival rate λ. We only consider the first stage of this hot-rolling process which is "pre-heating". These jobs triggered by customer orders will thus be pre-heated by a few pre-heating furnaces. Real data analysis (Jia et al., 2017) reveals that pre-heating times follow a PH distribution (best-fit distribution). Owing to the limited nature of

this operation, the process can be modeled as an M/PH/c-type multi-server queue. The pre-heating system usually consists of only two to four furnaces (e.g., expensive and large facilities in Shanghai Baogang Steel Company). We can use the micro-scopic model to analyze this pre-heating operation and estimate the expected delay time of customer order and other related performance measures. The system can be formulated as QBD process if we define the state space as follows:

Define:

$N(t)$: The number of customers in the system at time t;
$J_i(t)$: The phase of the PH service time at time t for server i.

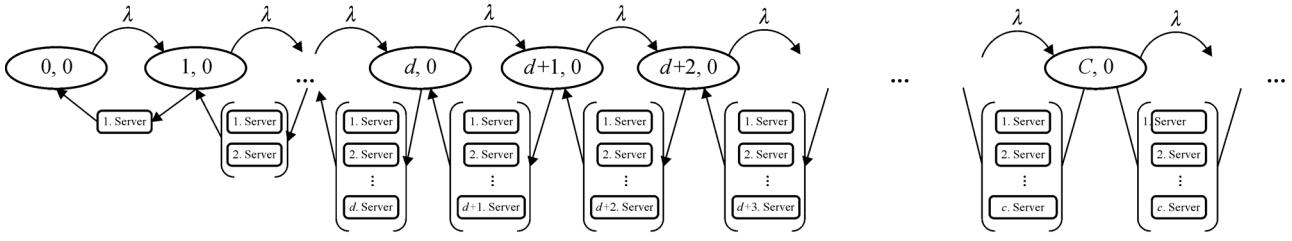
Assuming that the service time's PH distribution is represented as $(\beta,\boldsymbol{S})$ with m phases, the system state space is $\{N(t);J_1(t);J_2(t);...;J_c(t)\}$ where $N(t) = 0,1,...;\ J_i(t) = 1,2,...,m;\ i = 1,2,...,c$. We also assume that the PH representation of the service time is $(\beta,\boldsymbol{S})$ and with $\boldsymbol{t} = -\boldsymbol{S}\cdot\boldsymbol{e}$ where $\boldsymbol{S}$ is a transition rate matrix among the m phases (for example, from State i to State j, the transition rate should be s_{ij}) and $\boldsymbol{e}$ is the column vector of **1**'s (all elements are 1) with appropriate dimension.

Figure 1 shows the state transition diagram of the QBD process for this operation, and Fig. 2 presents a specific example with three servers (furnaces). In these diagrams (without showing the service time phase), we use the first number in the state node to represent the number of customers in the system and "0" to represent all servers that are on (it can be set to 1 to represent a model with changeable staffing levels as in Jia et al. (2017)). Each server node below each state node represents the service phase of each busy server.

We present more details about the QBD model for a three-server case.

The infinitesimal generator can be written as follows:

$$\boldsymbol{Q} = \begin{bmatrix} -\lambda\boldsymbol{I} & \lambda\beta & & & & & \\ -\boldsymbol{S}\cdot\boldsymbol{e} & -\lambda\boldsymbol{I}+\boldsymbol{S} & \lambda\boldsymbol{I}\otimes\beta & & & & \\ & \boldsymbol{S}^{01} & -\lambda\boldsymbol{I}+\boldsymbol{S}_2 & \lambda\boldsymbol{I}\otimes\beta & & & \\ & & \boldsymbol{S}^{02} & -\lambda\boldsymbol{I}+\boldsymbol{S}_3 & \lambda\boldsymbol{I} & & \\ & & & \boldsymbol{S}^{03} & -\lambda\boldsymbol{I}+\boldsymbol{S}_3 & \lambda\boldsymbol{I} & \\ & & & & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 \\ & & & & & \cdots & \cdots \end{bmatrix}$$



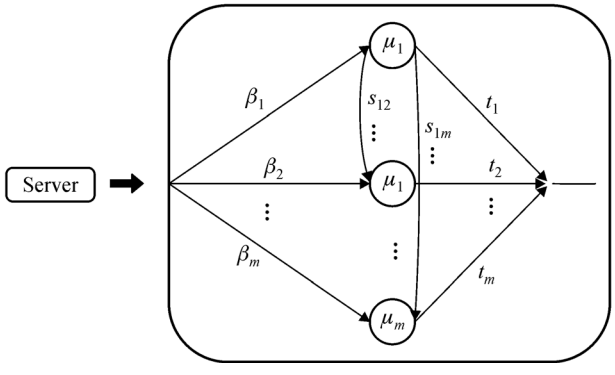


Fig. 1 State transition diagram of an M/PH/c queue

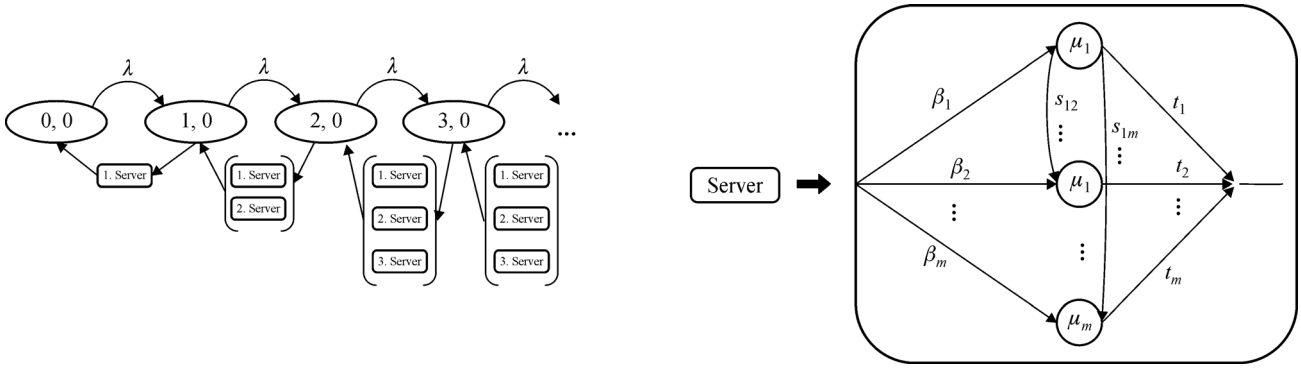


Fig. 2 A manufacturing system with three production facilities modeled as an M/PH/3 queue

$$\boldsymbol{S}^{01} = (-\boldsymbol{S}\cdot\boldsymbol{e}) \otimes \boldsymbol{I} + \boldsymbol{I} \otimes (-\boldsymbol{S}\cdot\boldsymbol{e}),$$

$$\boldsymbol{S}^{02} = (-\boldsymbol{S}\cdot\boldsymbol{e}) \otimes \boldsymbol{I} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes (-\boldsymbol{S}\cdot\boldsymbol{e}) \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{I} \otimes (-\boldsymbol{S}\cdot\boldsymbol{e}),$$

$$\boldsymbol{S}^{03} = (-\boldsymbol{S}\cdot\boldsymbol{e}\cdot\beta) \otimes \boldsymbol{I} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes (-\boldsymbol{S}\cdot\boldsymbol{e}\cdot\beta) \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{I} \otimes (-\boldsymbol{S}\cdot\boldsymbol{e}\cdot\beta),$$

$$\boldsymbol{S}_2 = \boldsymbol{S} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{S},$$

$$\boldsymbol{S}_3 = \boldsymbol{S} \otimes \boldsymbol{I} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{S} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{I} \otimes \boldsymbol{S},$$

$$\boldsymbol{A}_0 = \lambda\boldsymbol{I},\ \boldsymbol{A}_1 = -\lambda\boldsymbol{I} + \boldsymbol{S}_3,\ \boldsymbol{A}_2 = \boldsymbol{S}^{03},$$

where $\otimes$ is the Kronecker multiplication operation. With this model, we can obtain the stationary queue length distribution as a matrix geometric solution (Neuts, 1981). In this model, $\boldsymbol{\pi Q} = \boldsymbol{0}$, where $\boldsymbol{\pi} = [\boldsymbol{\pi}_0,\boldsymbol{\pi}_1,...]$ and $\boldsymbol{\pi}_i$ is the stationary probability vector for level i. If we have D levels of boundary states, then we can find stationary distribution as $\boldsymbol{\pi}_i = \boldsymbol{\pi}_D\boldsymbol{R}^{i-D}$, $i = D+1, D+2, ...$, where $\boldsymbol{R}$ is the rate matrix which is the minimal nonnegative solution to the equation quadratic matrix equation $\boldsymbol{R}^2\boldsymbol{A}_2 + \boldsymbol{R}\boldsymbol{A}_1 + \boldsymbol{A}_0 = 0$, which can be numerically solved by using iteration algorithm (Neuts, 1981).

The boundary state probability vectors can be obtained by solving a finite number of equations and the normalization condition $\sum_{i=0}^{\infty} \boldsymbol{\pi}_i\boldsymbol{e} = 1$. In this case, the rate matrix can be computed by using several alternative algorithms available (Neuts, 1981; Latouche and Ramaswami, 1999). Many important performance measures, such as expected queue length and the probability that the queue length exceeds a critical level, can be obtained with the queue length distribution. We can also handle both low-traffic and heavy-traffic situations. By contrast, the meso- or marco-scopic models may fail to produce useful performance measures for low-traffic conditions. However, the disadvantage is that the size of the state space is exponentially increasing with the number of servers. That is, for a c-server system with PH-distributed service time of m phases, the QBD process will have mc states. Dealing with a large number of servers for micro-scopic models is thus like an NP-hard problem.

For a queueing model developed for the steel industry, we aim to determine the optimal number of servers. If we turn on an excessive number of reheating furnaces, then the productivity will be increased, but the company will also have the idle cost, which will increase the total cost of the hot rolling process. By contrast, when the number of reheating furnaces is excessively small, the production cost will be reduced, but the order waiting time will increase or the waiting cost will be high. Therefore, a tradeoff between the furnace idle cost and the order waiting cost must be considered to determine the optimal number of furnaces. The micro-scopic model can help decision makers by providing the performance information of the system. Jia et al. (2017) presented a sophisticated model based on the M/PH/c in the steel industry.

3.2 Meso-scopic and macro-scopic models for a service system

Although the PH distribution and MAP turn several general stochastic systems into Markovian ones, the micro-scopic model can only treat small-scale systems such as a single-stage queue with only a few number of servers. For large-scale service systems, the micro-scopic model is impractical to use because of the “curse of dimensionality” mentioned in the previous section. A large-scale system usually implies a single-stage queue with a large number of servers such as call centers or a queueing network with many service stations of either a single server or many servers. Furthermore, customers can be either of single or multiple classes, and servers can be either homogeneous or heterogeneous. For these large-scale systems, we have to

adopt the meso- or marco-scopic models. In this subsection, we provide two simple but typical examples.

The first example is to develop the meso-scopic model for a single-stage service process via reflected Brownian motion (RBM)-based approximations. In this model, the building block is the G/G/1 queue. After approximating the performance measure of G/G/1, we extend it to G/G/s system. Let $A(t)$ be the number of customers that arrived from time 0 to t, and $S(t)$ the number of customers the server would process from time 0 to t if the server was busy during (0, t). Given that the interarrival times and service times are i.i.d., the processes $\{A(t),t\geqslant 0\}$ and $\{S(t),t\geqslant 0\}$ are renewal processes. Therefore, for a large t, $A(t)$ is estimated to be normally distributed with mean λt and variance $\lambda C_a^2 t$, where C_a^2 is the squared coefficient of variation for the interarrival time. In addition, for a large t, $S(t)$ is normally distributed with mean μt and variance $\mu C_s^2 t$, where C_s^2 is the squared coefficient of variation for the service time. Approximation (for some large constant T) of $\{A(t),t\geqslant T\}$ and $\{S(t),t\geqslant T\}$ by Gaussian processes, in particular, Brownian motions (BMs), is practical. This approach is a reasonable approximation and can be justified rigorously by appropriately adopting a scaling argument as done by Chen and Yao (2001). A stochastic model based on BM or generally a diffusion process is called a meso-scopic model as it is an approximation under heavy-traffic condition but retains the random variation component. We first present the relevant performance measures for the queue in terms of $A(t)$ and $S(t)$.

Let $X(t)$ denote the number of customers in the G/G/1 queue at time t with $X(0)=x_0$, a given finite constant number of customers initially. To present $X(t)$ in terms of $A(t)$ and $S(t)$, the duration in which the server was busy and idle during (0, t) needs to be determined. Thus, let $B(t)$ and $I(t)$, respectively, denote the total time the server has been busy and idle from time 0 to t. We emphasize that the server is work conserving, which means that the server would be idle if and only if the system has no customers. This condition is equivalent to $B(t)+I(t)=t$. Thus, we can present an expression for $X(t)$ as $X(t)=x_0+A(t)-S(B(t))$. By using the centering, we can re-write the expression as $X(t)=U(t)+V(t)$ where $U(t)=x_0+(\lambda-\mu)t+(A(t)-\lambda t)-(S(B(t))-\mu B(t))$ and $V(t)=\mu I(t)$. Computing the expected value and variance of $U(t)$ for large t yields $E[U(t)]=x_0+(\lambda-\mu)t$, $Var[U(t)]\approx \lambda C_a^2 t+\lambda C_s^2 t$ because $U(t)$ is approximated by a drifted BM via Donsker's theorem. Although $E[U(t)]$ is exact for any t, $Var[U(t)]$ is reasonable only for a large t, that is, in the asymptotic case. $X(t)$ and $V(t)$ can be specifically determined by satisfying the following conditions:

$$X(t)\geqslant 0,\ \frac{\mathrm{d}V(t)}{\mathrm{d}t}\geqslant 0 \text{ with } V(0)=0, \text{ and } X(t)\frac{\mathrm{d}V(t)}{\mathrm{d}t}=0.$$

Furthermore, they can be expressed in terms of $U(t)$ as follows:

$$V(t)=\sup_{0\leqslant s\leqslant t}\max\{-U(s),0\},$$

$$\text{and } X(t)=U(t)+\sup_{0\leqslant s\leqslant t}\max\ \{-U(s),0\}.$$

The proof of these relations can be found in the work of Chen and Yao (2001). On the basis of the characteristics of this result, $X(t)$ is called the reflected process of $U(t)$ and $V(t)$ is the regulator of $U(t)$. The $\{X(t),t\geqslant 0\}$ process is then a corresponding RBM. Using heavy-traffic approximations for queueing process, we focus on the workload process (which is inherently continuous) as an RBM. Considering that the RBM is a continuous function of the BM, according to continuous mapping theorem, we can conclude that $X(t)$ will approach the queue length process limit. By denoting $W(t)$ as the workload at time t, Chen and Yao (2001) used the approximation to relate the workload in the system to the number in the system. Evidently, if $\{X(t),t\geqslant 0\}$ is an RBM with initial state x_0, drift $(\lambda-\mu)$, and variance $\lambda(C_a^2+C_s^2)$, then $\{W(t),t\geqslant 0\}$ is also an RBM with initial state x_0/μ, drift $(\lambda-\mu)/\mu$, and variance $\lambda(C_a^2+C_s^2)/\mu^2$ when $W(t)=X(t)/\mu$. For the G/G/1 queue described earlier, considering that we approximated the workload process $\{W(t),t\geqslant 0\}$ as an RBM, we have the expected workload in steady state as $\lambda(C_a^2+C_s^2)/[2(1-\rho)\mu^2]$, which is the approximation to the expected waiting time of a customer. Other performance measures such as L_q, W, and L, can then be obtained by using $L_q=\lambda W_q$, $W=W_q+1/\mu$ and $L=\lambda W$. Such an approximation only needs the mean and variance of the interarrival time and service time instead of the entire distribution. This G/G/1 queue approximation can be extended to the multi-server case or G/G/s queue. Under the first come first serve (FCFS) discipline in a stable system, the average total system time can be approximated by

$$w\approx\frac{1}{\mu}+\frac{\rho^2C_s^2+C_a^2}{2\lambda(1-\rho)},$$

where $\rho=\lambda/(s\mu)$. This approximation is appropriate for G/M/s systems and can be proved using fluid and diffusion scaling (Halfin and Whitt, 1981). In addition, an empirical approximation for G/G/s queues (originally developed for M/G/s queues) is

$$w\approx\frac{1}{\mu}+\frac{\alpha_s}{\mu}\left(\frac{1}{1-\rho}\right)\left(\frac{C_a^2+C_s^2}{2s}\right),$$

where a_s should be chosen such that

$$\alpha_s=\begin{cases}\dfrac{\rho^s+\rho}{2} & \text{if}\quad \rho>0.7\\ \rho^{\frac{t+1}{2}} & \text{if}\quad \rho<0.7\end{cases}.$$

Hence, we briefly describe the diffusion approximation

to a multi-server queue under heavy-traffic condition (extensive literature is available in this area; Whitt, 2002). The main idea is to obtain the stochastic process limits as approximations of the actual stochastic processes. These limits are either deterministic processes called fluid limits or stochastic processes called diffusion limits that use appropriate scaling methods for time and space. Clearly, fluid limits correspond to macro-scopic models and diffusion limits correspond to meso-scopic models. We start with the fluid limit scaling. Let $A(t)$ be the number of arrivals into a system during $(0, t)$. The average arrival rate is λ, which is expressed as

$$\lambda = \lim_{t\to\infty} \frac{A(t)}{t},$$

based on the strong law of large numbers (SLLN).

To obtain the fluid limit of the discrete arrival process $\{A(t),\ t\geqslant 0\}$, we define

$$\overline{A_n}(t) = \frac{A(nt)}{n},$$

for any $n>0$ and $t\geqslant 0$. We show that as $n\to\infty$, $\overline{A_n}(t)\to\lambda t$, which we refer to as the fluid limit. This result can be seen as the functional strong law of large numbers (FSLLN), a generalization of the SLLN. Similarly, we define $\hat{A}_n(t)$ as

$$\hat{A}_n(t) = \sqrt{n}\left[\overline{A_n}(t)-\lambda t\right] = \frac{A(nt)-n\lambda t}{\sqrt{n}},$$

for any $n>0$ and $t\geqslant 0$. We would like to study $\hat{A}_n(t)$ as $n\to\infty$, which we will call diffusion scaling. Such a scaling is similar to magnifying the fluid space scaling by a factor $\sqrt{n}$ after centering. In this way, some random fluctuations around the center can be captured. $\hat{A}_n(t)$ converges to a BM with drift 0 and variance term λC_a^2 as $n\to\infty$. Such a result is called the FCLT, also called Donsker's theorem, a generalization of central limit theorem. Similar to the arrival process, the service time process when scaled in a similar manner also converges to a BM. Thus, approximate expressions for the distribution of waiting times or system times can be obtained using the diffusion approximation when the traffic intensity is close to one (heavy-traffic approximations).

The main goal in a diffusion approximation is to obtain the transient or steady-state distribution of a stochastic process $\{Z(t),t\geqslant 0\}$ (this value is usually the number in the system process $\{X(t),t\geqslant 0\}$ but we will keep it more generic in this case). For a detailed description, see Gautam (2012). The process is scaled by a factor "n" across time and $\sqrt{n}$ across "space". Thus, we define

$$\hat{Z}_n(t) = \frac{Z(nt)-\overline{Z}(nt)}{\sqrt{n}}.$$

The term $\overline{Z}(nt)$ is the deterministic fluid model of the stochastic process $\{Z(t),t\geqslant 0\}$ by fluid scaling. Usually, $\overline{Z}(nt) = E[Z(nt)]$ or is a heuristic approximation for it, indicating that the fluid model is focused on the first moment value. Therefore, the fluid model is called a macro-scopic model. Assuming that the deterministic fluid limit $\overline{Z}(nt)$ exists and can be computed, the meso-scopic model is mainly intended to study $\left\{\hat{Z}_n(t),t\geqslant 0\right\}$, which captures the randomness feature of the process. The application of appropriate space scaling reveals that as $n\to\infty$, the stochastic process $\left\{\hat{Z}_n(t),t\geqslant 0\right\}$ converges to a diffusion process. A diffusion process is a continuous-time stochastic process with almost certainly continuous sample paths and satisfies the Markov property. Examples of diffusion processes are BM, Ornstein-Uhlenbeck process, Brownian bridge process, and branching process. The diffusion process can be expressed as $\left\{\hat{Z}_\infty(t),t\geqslant 0\right\}$. For detailed information, see Whitt (2004). The key idea of diffusion approximation is to start by using the properties of $\left\{\hat{Z}_\infty(t),t\geqslant 0\right\}$ such as the stationary distribution of $\hat{Z}_\infty(\infty)$. For large n, $\hat{Z}_n(\infty)$ is approximately equal in distribution to $\hat{Z}_\infty(\infty)$. We can thus estimate the distribution for $Z(\infty)$ using $Z(\infty) = \overline{Z}(\infty)+\sqrt{n}\hat{Z}_n(\infty)$ such that $\hat{Z}_n(\infty)\approx\hat{Z}_\infty(\infty)$.

We use an M/M/s queue to illustrate the development of meso-scopic model first. Whitt (2004) has presented general cases such as the G/G/s or M/G/s queue. For M/M/s queues, the Markov property leads to diffusion processes. However, in the G/G/s queue, although the marginal distribution at any time in steady state converges to Gaussian, the process itself may not be a diffusion process (because the Markov property would not be satisfied). For such an M/M/s queue, let $X(t)$ be the number of customers in the system at time t. We are interested in applying diffusion scaling to the stochastic process $\{X(t),t\geqslant 0\}$. Further, we define

$$\hat{X}_n(t) = \frac{X(nt)-\overline{X}(nt)}{\sqrt{n}},$$

for any $n>0$ and $t\geqslant 0$. As a heuristic approximation for the fluid model, we use the well-known steady-state system size L for the M/M/s queue

$$L = \frac{\lambda}{\mu} + \frac{p_0(\lambda/\mu)^s\lambda}{s!s\mu[1-\lambda/(s\mu)]^2},$$

where

$$p_0 = \left[\sum_{n=0}^{s-1}\left(\frac{1}{n!}(\lambda/\mu)^n\right) + \frac{(\lambda/\mu)^s}{s!}\frac{1}{1-\lambda/(s\mu)}\right]^{-1}.$$

Given that L=$E[X(\infty)]$, for large nt, we use the approximation $\overline{X}(nt) = L$. We can thus study the diffusion scaled process as

$$\hat{X}_n(t) = \frac{X(nt) - L}{\sqrt{n}},$$

by increasing n and show that this process converges to a diffusion process under three different scalings for n.

In the first scaling, we consider a sequence of M/M/s queues where μ and s are held constant and only λ is increased to ensure that ρ approaches 1. We can use the scale $n = 1/(1-\rho)^2$ to ensure that n increases as ρ increases. This condition is a special case of general scaling called by Halfin and Whitt (1981) as the regime.

It can be shown that $\hat{X}_n(t/n) = (X(t) - L)/\sqrt{n}$ process converges to a diffusion process as n is scaled. This scaling can be used when the system has high traffic intensity but not a large number of servers.

The next scaling is to fix ρ but increase λ and s. In this scaling, we consider a sequence of M/M/s queues where μ and ρ are held constant but λ and s are increased to ensure that s approaches ∞. We use the scale $n = s$ to ensure that n increases as s increases. Again, it can be shown that

$$\hat{X}_n(t/n) = \frac{X(t) - L}{\sqrt{n}},$$

$\{\hat{X}_n(t/n), t \geqslant 0\}$ process converges to a diffusion process as n is scaled. This scaling would be more powerful if the scaled time is also available, that is, plotted $\hat{X}_n(t)$ instead of $\hat{X}_n(t/n)$. This scaling could be used when the system has a large number of servers but not a high traffic intensity.

The third scaling is based on Halfin and Whitt's (1981) regime. We use the scale $n = s$ to ensure that n increases as s increases and let $\rho \to 1$ in such a way that β is held constant where $\beta = (1-\rho)\sqrt{s}$. Under this scaling, $\{\hat{X}_n(t/n), t \geqslant 0\}$ process converges to a diffusion process. This scaling would be more powerful if we were to have scaled time as well, that is, plotted $\hat{X}_n(t)$ instead of $\hat{X}_n(t/n)$. This scaling is practical to use when the system has both high traffic intensity and a large number of servers (which is typical in inbound call centers). Such a system is called quality and efficiency driven. Evidently, the first two scalings are special cases of this scaling. Clearly, $E[\hat{X}_n(t)]$ converges to zero because $E[X(nt)]$ would be L. Halfin and Whitt (1981) presented the probability behavior for $\hat{X}_n(t)$ as $t \to \infty$ and $n \to \infty$. If we define

$$\hat{X}_n(t) = \frac{X(nt) - s}{\sqrt{n}},$$

and use the scale $n = s$, then $\lim_{n\to\infty} P(\hat{X}_n(\infty) \geqslant 0) = \theta$, where θ is a constant satisfying $0 < \theta < 1$. We provide an intuition for the first one and subsequently describe θ. For $\hat{X}_n(t) \geqslant 0$, the process converges to an RBM with negative drift (because for $X(nt) \geqslant s$, the CTMC is a BD process with constant parameters λ and $s\mu$, which is a random walk that converges to a BM upon scaling). However, for $\hat{X}_n(t) < 0$, the process converges to an Ornstein-Uhlenbeck process (because for $X(nt) < s$, the CTMC is a BD process with parameters λ and $X(nt)\mu$, which is a random walk that converges to an Ornstein-Uhlenbeck process upon scaling). Such a property also holds for G/M/s queues. Therefore, the probability in which $\hat{X}_n(t) \geqslant 0$ can be obtained as $t \to \infty$ and $n \to \infty$. Thus, under the scaling $n = s$,

$$P\{X(\infty) \geqslant s\} \to \theta,$$

where $\theta = [1 + \sqrt{2\pi}\beta\varphi(\beta)\exp(\beta^2/2)]^{-1}$, and $\varphi(x)$ is the probability that a standard normal random variable is less than x. Thus, with probability θ, an arriving customer in steady state will experience a delay. Notably, θ can be obtained with the other two scalings considered earlier. In those cases, θ would only be 0 or 1. In particular, if we fix s and increase λ, then θ approaches 1 (that is, an arriving request with probability 1 will be delayed for service to begin). Such a case is said to be in efficiency driven (ED) regime. By contrast, if we fix ρ but increase λ and s, then θ approaches zero (that is, an arriving customer with probability 1 will find a free server). This case is said to be in quality driven regime. Aside from these probability measures, we can also obtain the expected queue length for either the stationary or transient system (Halfin and Whitt, 1981). A major advantage of the meso-scopic model is that we can obtain major performance measures for large-scale service systems consisting hundreds of servers under heavy-traffic conditions. These systems are practical in the service sector.

Next, we discuss the macro-scopic models or fluid models. The fluid model itself is useful in studying the stability of networks and analyzing queues with time-varying arrivals. For a traditional single-stage multi-server queue, because of the first moment nature, the stable, critically loaded, and overloaded cases correspond to simple queue length results of zero, the same as the initial content, and linearly increase with time, respectively (Whitt, 2002). Thus, to provide an interesting case, we consider a multi-server queue with a special feature, namely, overloaded with customer abandonment behavior. We use a Markovian M/M/s/$r+M$ queue as an example to demonstrate the benefit of using the macro-scopic model or fluid deterministic model, in which $r+M$ represents a finite waiting room r-s and the time to abandon is exponentially distributed with rate α. The analysis is based on Whitt (2004), and we only present a few simple fluid model results. Overloaded queue is also called a queue under the ED regime, as mentioned earlier. Consider a sequence of M/M/s/$r+M$ queues indexed by s, which is the number of servers and we would use to scale. In particular, λ_s and r_s are the scaled arrival rate and system capacity, respectively. However, the service rate μ and abandonment rate α are not scaled. In addition, the traffic

intensity ρ is not scaled and remains fixed for the entire sequence of queues with $\rho > 1$.

Define

$$q = \frac{\mu(\rho-1)}{\alpha},\ \lambda_s = \rho s\mu,\ r_s = s(\eta+1),$$

where $\eta > q$ to ensure that asymptotically no arriving customers are rejected due to a full system (Whitt, 2004). With this scaling, we can obtain the meso-scopic model by using the diffusion scaling as presented above. Let $X_s(t)$ be the number of customers in the system at time t when s servers are used. We define the diffusion term as follows:

$$\hat{X}_s(t) = \frac{X_s(t) - \overline{X}_s(t)}{\sqrt{s}},$$

where $\overline{X_s}(t)$ is a deterministic marco-scopic model of $X_s(t)$. We use a heuristic approximation for the deterministic quantity $\overline{X_s}(t)$, which is where $X_s(t)$ tends to linger around in the ED regime. In particular, we select $\overline{X_s}(t)$ as an "equilibrium" point where the system growth rate equals the shrinkage rate. Thus, we have $\overline{X_s}(t)$ as the solution to the flow balance equation

$$\lambda_s = s\mu + \left[\overline{X_s}(t) - s\right]\alpha,$$

by estimating that $\overline{X_s}(t)$ must be greater than s (as $\rho > 1$ results in $\lambda_s > \min\{i,s\}\mu$ for any $i \geqslant 0$). We have the following:

$$\overline{X_s}(t) = \frac{\lambda_s - s\mu}{\alpha} + s = (1+q)s,$$

where the last equality is achieved by using the expressions for λ_s and q. Thus, we represent the diffusion term as

$$\hat{X}_s(t) = \frac{X_s(t) - (1+q)s}{\sqrt{s}} \text{ for all } t \geqslant 0.$$

Whitt (2004) showed that the stochastic process $\left\{\hat{X}_s(t), t \geqslant 0\right\}$ as $s \to \infty$ converges to an Ornstein-Uhlenbeck diffusion process. In this case, we focus on the fluid model only. Let $E[Q_s(\infty)]$ be the expected steady-state number of customers waiting in queue. Under heavy traffic conditions, for large t, we have $\overline{X_s}(t) \approx s + E[Q_s(\infty)]$. Thus, $E[Q_s(\infty)] \approx qs \equiv (\rho-1)s/\alpha$. We can also obtain the abandonment probability, which is denoted by $P_s(ab)$. The relation $\lambda_s P_s(ab) = \alpha E[Q_s(\infty)]$ shows that $P_s(ab) = (\rho-1)/\rho$. Such simple performance measures indicate the advantage of using the macro-scopic model. We only present a small sample of the fluid model results in this case. Detailed studies on fluid models can be found in Whitt (2004; 2005; 2006). In theory, when the scaling conditions (i.e., heavy-traffic condition) are satisfied, the meso- and micro-scopic models must generate acceptable approximations. In practice, we can use simulations to verify the accuracy of using these approximations.

In contrast to these meso-scopic and micro-scopic models, the micro-scopic model cannot handle a system with a large number of servers and its performance measures can be obtained only by numerical approach.

The examples in this section demonstrate the strengths and weaknesses of each type of modeling in this stochastic modeling hierarchy. Different levels of models complement one another to better represent the stochastic manufacturing and service systems in practice.

4 Conclusions

We present different levels of modeling stochastic systems subject to congestions (queues). Micro-scopic models are practical in analyzing manufacturing systems with a small number of production facilities (servers). Using PH distribution and MAP to model the random service times and the customer arrivals makes continuous-time Markov chain flexible in modeling a real-world system. In theory, the PH distribution can approximate any non-negative continuous distribution and the MAP can approximate any non-renewal process. In practice, this condition implies that real data can be used to estimate the parameters of PH distribution and MAP to make the model fit the real system well. Another benefit of using the micro-scopic model is that we can obtain the stationary distribution of the system size. However, this level of modeling is not appropriate for analyzing large-scale systems, such as a call center with hundreds of servers or a queue network with many nodes or service stations. For these large-scale service systems, we can rely on meso-scopic and macro-scopic models. For these two levels, we consider the scaled processes of system size or customer waiting time and obtain the performance measures based on FLLNs and FCLTs. The macro-scopic model is the deterministic fluid model. The meso-scopic model becomes the diffusion model by adding the stochastic refinements to the fluid model. With the meso-scopic model, we can obtain probability-based performance measures as the macro-scopic model only provides first moment-based performance measures. Given that these two higher levels of modeling are based on stochastic process limits, they can be used as powerful tools to study large-scale service systems. In addition, several closed-form formulas for major performance measures may be obtained. Finally, the meso- or macro-scopic models can be used to analyze not only stationary systems but also transient systems. In practice, the appropriate level of modeling should be selected depending on the situation.

Acknowledgements The authors thank the NSERC of Canada for the partial support of this research.

References

Buzacott J A, Shanthikumar J G (1993). Stochastic Models of

Manufacturing Systems. New York: Prentice Hall

Chen H, Mandelbaum A (1994). Stochastic modeling and analysis of manufacturing systems. In: Yao D D, ed. Operations Research. Berlin: Springer

Chen H, Yao D (2001). Fundamentals of Queueing Networks, Performance, Asymptotics, and Optimization. New York: Springer

Gautam N (2012). Analysis of Queues. New York: CRC Press

Halfin S, Whitt W (1981). Heavy-traffic limits for queues with many exponential servers. Operations Research, 29(3): 567–588

He Q (2014). Fundamentals of Matrix-Analytic Methods. New York: Springer

Jia Y, Zhang Z G, Tang L (2017). Modeling hot rolling process in steel industry by M/PH/*c* queues. Working paper. Simon Fraser University, WP0170056

Koole G, Mandelbaum A (2002). Queueing models of call centers: An introduction. Annals of Operations Research, 113(1–4): 41–59

Latouche G, Ramaswami V (1999). Introduction to Matrix Geometric Methods in Stochastic Modeling. Philadelphia: SIAM

Neuts M F (1981). Matrix-Geometric Solutions in Stochastic. New York: Dover Publications

Whitt W (2002). Stochastic Process Limits. New York: Springer

Whitt W (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Management Science, 50 (10): 1449–1461

Whitt W (2005). Two fluid approximations for multi-server queues with abandonments. Operations Research Letters, 33(4): 363–372

Whitt W (2006). Fluid models for multiserver queues with abandonments. Operations Research, 54(1): 37–54