Analysis of Multiserver Queues with Batch Arrivals and Setup Times

Tuan Phung-Duc[1,2]

[1]Institute of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
[2]Center for Artificial Intelligence Research (C-AIR), Tsukuba Institute for Advanced Research, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
tuan@sk.tsukuba.ac.jp

Abstract This paper considers a batch Poisson-arrival multiserver queue with setup time, motivated by power-saving data centers. A server is immediately turned off once it has no job to process. An off-server (if any) is turned on again upon the arrival of a job. The server needs some setup time during which it cannot process the job. In this model, the number of setup servers depends on the number of jobs in the system. Although some methods have been developed to solve special cases with Poisson arrivals or batch Poisson arrivals with a single setup server, this paper is the first to consider the model with batch Poisson arrivals without the limitation on the number of setup servers. Using a generating function approach, we derive the joint stationary distribution of the number of active servers and the number of jobs in the system. Furthermore, we obtain a conditional decomposition formula with clear physical meaning that shows the effect of the setup time on the number of jobs in the system.

1. Introduction

Data centers are the core infrastructure for our information society and AI era. These data centers host a large number of servers that consume significant amounts of electricity. These servers are not always busy processing jobs, but are idle for a significant portion of the time [2]. Thus, it is natural to turn off idle servers to save energy and turn them on when jobs arrive. One of the simplest policies in this line is the ON-OFF policy, which turns off an idle server immediately and turns it back on once a job is waiting to be processed. From a queueing model perspective, data centers with the ON-OFF policy can be modeled using a queueing system with multiple servers and setup time [1, 3, 4, 8, 10].

Models with Poisson arrivals have been extensively studied, and efficient algorithms for the stationary distribution of the underlying Markov chain have been proposed and evaluated [1, 3, 4, 8, 10]. In this paper, we relax the arrival process to a Batch Poisson process, which captures the case where jobs arrive in batches. This allows modeling more realistic situations with bursty traffic. In our previous work, we considered the case where only one server can be set up at a time [7, 9]. This paper removes this assumption, allowing an arbitrary number of servers in the setup process.

The rest of the paper is organized as follows. Section 2 presents the queueing model and its underlying Markov chain. Section 3 shows the analysis of the stationary distribution of the Markov chain via the generating function method and the conditional decomposition for the queue length.

2. Model and Markov Chain

2.1. Model

We consider $M^X/M/c$/Setup queueing systems with the ON-OFF policy. We assume that the distribution of the batch size X is given by $\beta_i = P(X = i)$ $(i = 1, 2, \dots)$ and the mean batch size is given by $E[X] < \infty$. Batches of jobs arrive at the system according to a Poisson process with rate λ. We assume that the service time of jobs follows an exponential distribution with mean

$1/\mu$. In this system, upon service completion, a server is turned off immediately if there are no waiting jobs. Upon the arrival of a job, an OFF server is turned on, and the job is placed in the buffer. The off-server needs some setup time to be ready to serve waiting jobs. We assume that the setup time follows an exponential distribution with mean $1/\alpha$. Suppose that there are two jobs in the system. One job is receiving service, and the other job in the buffer is waiting for a server in the setup process. In this situation, if the service completes before the setup, the waiting job is served immediately, and the server in the setup process is turned off. Under these settings, the number of active servers is smaller than or equal to the number of jobs in the system.

Let j denote the number of customers in the system and i denote the number of active servers. The number of servers in the setup process is given by $\min(j-i, c-i)$, where $j-i$ is the number of waiting jobs and $c-i$ is the number of servers that are not active (serving a job). In this model, a server is in one of the following states: ACTIVE, OFF, or SETUP. We assume that waiting jobs are served according to a first-come, first-served (FCFS) manner. The exponential assumptions for the inter-arrival, setup time, and service time allow us to construct a Markov chain for which an efficient solution for the stationary distribution is presented.

2.2. Markov chain and notations

It is easy to see that the stability condition for the system is $\lambda E[X] < c\mu$ because all the servers are eventually active if the number of jobs in the system is large enough. Let $C(t)$ and $N(t)$ denote the number of active servers and the total number of jobs in the system at time t, respectively. Under the assumptions made in Section 2.1, it is easy to see that $\{X(t) = (C(t), N(t)); t \geq 0\}$ forms a Markov chain in the state space

$$\mathcal{S} = \{(i,j); i = 0, 1, \ldots, c, j = i, i+1, \ldots\}.$$

See Figure 1 for the transitions among states for a special case with two servers and the maximum batch size of 2.

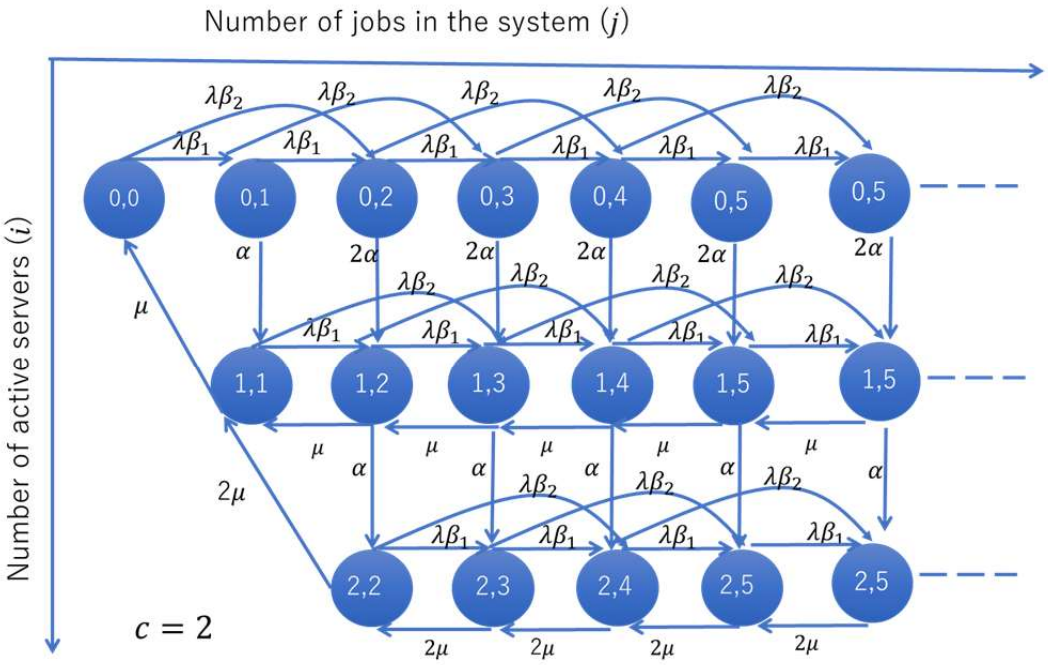

Figure 1: Transition among states ($c = 2, \beta_1 + \beta_2 = 1$).

We define

$$\pi_{i,j} = \lim_{t\to\infty} \mathbb{P}(C(t) = i, N(t) = j), \qquad (i,j) \in \mathcal{S}.$$

It should be noted that at the state (i,j) the number of waiting jobs is $j-i$. Furthermore, we define the following partial generating functions.

$$\Pi_i(z) = \sum_{j=c}^{\infty} \pi_{i,j} z^{j-c}, \qquad i = 0, 1, \ldots, c.$$

3. Generating Function Approach

In this section, we derive expressions for the partial generating functions.

3.1. Recursive expressions

The balance equations for the case $i = 0$ read as follows.

$$\lambda\pi_{0,0} = \mu\pi_{1,1}, \qquad j = 0,$$

$$(\lambda + j\alpha)\pi_{0,j} = \lambda \sum_{i=1}^{j} \beta_i \pi_{0,j-i}, \qquad j = 1, 2, \ldots, c-1, \tag{1}$$

$$(\lambda + c\alpha)\pi_{0,j} = \lambda \sum_{i=1}^{j} \beta_i \pi_{0,j-i}, \qquad j \geq c. \tag{2}$$

Multiplying (2) by z^{j-c} and summing over $j \geq c$, we obtain

$$\begin{aligned}
(\lambda + c\alpha)\Pi_0(z) &= \lambda \sum_{j=c}^{\infty} \left(\sum_{i=1}^{j} \beta_i \pi_{0,j-i} \right) z^{j-c} \\
&= \lambda \sum_{j=c}^{\infty} \left(\sum_{i=1}^{j-c} \beta_i \pi_{0,j-i} + \sum_{i=j-c+1}^{j} \beta_i \pi_{0,j-i} \right) z^{j-c} \\
&= \lambda\beta(z)\Pi_0(z) + \lambda \sum_{k=0}^{c-1} \beta_k(z)\pi_{0,k},
\end{aligned}$$

where

$$\beta_k(z) = z^{k-c} \left(\beta(z) - \sum_{i=1}^{c-k-1} \beta_i z^i \right). \tag{3}$$

From this result, we have

$$\Pi_0(z) = \frac{\lambda \sum_{k=0}^{c-1} \pi_{0,k}\beta_k(z)}{\lambda + c\alpha - \lambda\beta(z)}. \tag{4}$$

In case $c = 1$, (4) is further simplified to

$$\Pi_0(z) = \frac{\lambda\pi_{0,0}\beta_0(z)}{\lambda + \alpha - \lambda\beta(z)} = \frac{\lambda\pi_{0,0}\beta(z)/z}{\lambda + \alpha - \lambda\beta(z)}. \tag{5}$$

The partial generating function for the number of waiting jobs for the case $C(t) = 0$ is given by

$$\widehat{\Pi}_0(z) = \pi_{0,0} + z\Pi_0(z) = \frac{(\lambda + \alpha)\pi_{0,0}}{\lambda + \alpha - \lambda\beta(z)}.$$

It should be noted that (1) implies that $\pi_{0,j}$ $(j = 0, 1, \ldots, c-1)$ is recursively expressed in terms of $\pi_{0,0}$. Furthermore, (4) implies that all the probabilities $\pi_{0,j}$, $j \geq c$ is also expressed in terms of $\pi_{0,0}$.

Remark 1. *Furthermore, $\pi_{1,1}$ is calculated in terms of $\pi_{0,0}$ as follows.*

$$\mu\pi_{1,1} = \sum_{j=1}^{c-1} j\alpha\pi_{0,j} + c\alpha\Pi_0(1).$$

We shift to the case $i = 1$. The balance equations are given as follows.

$$\begin{aligned}
(\lambda+\mu)\pi_{1,1} &= \alpha\pi_{0,1} + \mu\pi_{1,2} + 2\mu\pi_{2,2}, \\
(\lambda+\mu+(j-1)\alpha)\pi_{1,j} &= j\alpha\pi_{0,j} + \lambda\sum_{i=1}^{j-1}\beta_i\pi_{1,j-i} + \mu\pi_{1,j+1}, \qquad 2 \le j \le c-1, \qquad (6) \\
(\lambda+\mu+(c-1)\alpha)\pi_{1,j} &= c\alpha\pi_{0,j} + \lambda\sum_{i=1}^{j-1}\beta_i\pi_{1,j-i} + \mu\pi_{1,j+1}, \qquad j \ge c. \qquad (7)
\end{aligned}$$

Multiplying (7) by z^{j-c} and summing up over $j \ge c$ yields,

$$(\lambda+\mu+(c-1)\alpha)\Pi_1(z) = c\alpha\Pi_0(z) + \lambda\sum_{j=c}^{\infty}\left(\sum_{i=1}^{j-1}\beta_i\pi_{1,j-i}\right)z^{j-c} + \frac{\mu}{z}(\Pi_1(z) - \pi_{1,c}). \qquad (8)$$

We transform the second term on the right-hand side of (8) as follows.

$$\begin{aligned}
\lambda\sum_{j=c}^{\infty}\left(\sum_{i=1}^{j-1}\beta_i\pi_{1,j-i}\right)z^{j-c} &= \lambda\sum_{j=c}^{\infty}\left(\sum_{i=1}^{j-c}\beta_i\pi_{1,j-i} + \sum_{i=j-c+1}^{j-1}\beta_i\pi_{1,j-i}\right)z^{j-c} \\
&= \lambda\sum_{j=c}^{\infty}\sum_{i=1}^{j-c}\beta_i\pi_{1,j-i}z^{j-c} + \sum_{j=c}^{\infty}\sum_{i=j-c+1}^{j-1}\beta_i\pi_{1,j-i}z^{j-c} \\
&= \lambda\beta(z)\Pi_1(z) + \lambda\sum_{k=1}^{c-1}\pi_{1,k}\beta_k(z),
\end{aligned}$$

where $\beta_k(z)$ is defined in (3). Substituting this into (8) and rearranging the results yields

$$\Pi_1(z) = \frac{c\alpha z\Pi_0(z) + \lambda\sum_{k=1}^{c-1}\pi_{1,k}z\beta_k(z) - \mu\pi_{1,c}}{(\lambda+\mu+(c-1)\alpha)z - \lambda z\beta(z) - \mu}. \qquad (9)$$

Let $0 < z_1 < 1$ denote the unique solution of the equation $(\lambda+\mu+(c-1)\alpha)z - \lambda z\beta(z) - \mu = 0$ in $(0,1)$. Because $\Pi_1(z)$ is analytic at $z = z_1$, the numerator of (9) must be 0 at $z = z_1$, leading to

$$\pi_{1,c} = \frac{c\alpha z_1\Pi_0(z_1) + \lambda\sum_{k=1}^{c-1}\pi_{1,k}z_1\beta_k(z_1)}{\mu}.$$

In case $\beta(z) = z$, z_1 is explicitly given by

$$z_1 = \frac{\lambda+\mu+(c-1)\alpha - \sqrt{(\lambda+\mu+(c-1)\alpha)^2 - 4\lambda\mu}}{2\lambda}.$$

We can rewrite the equation for $\pi_{1,c}$ as follows.

$$\pi_{1,c} = a_c^{(1)} + \sum_{k=1}^{c-1} b_{c,k}^{(1)}\pi_{1,k},$$

where

$$a_c^{(1)} = \frac{c\alpha z_1 \Pi_0(z_1)}{\mu}, \qquad b_{c,k}^{(1)} = \frac{\lambda z_1 \beta_k(z_1)}{\mu}.$$

Assuming that

$$\pi_{1,j+1} = a_{j+1}^{(1)} + \sum_{k=1}^{j} b_{j+1,k}^{(1)} \pi_{1,k}, \qquad j \le c-1.$$

Substituting this formula into (6), we obtain

$$(\lambda + \mu + (j-1)\alpha)\pi_{1,j} = j\alpha\pi_{0,j} + \lambda \sum_{k=1}^{j-1} \beta_{j-k}\pi_{1,k} + \mu \left(a_{j+1}^{(1)} + \sum_{k=1}^{j} b_{j+1,k}^{(1)} \pi_{1,k} \right).$$

Thus, we obtain the following recursion.

$$\pi_{1,j} = a_j^{(1)} + \sum_{k=1}^{j-1} b_{j,k}^{(1)} \pi_{1,k}, \qquad j = 1, 2, \ldots, c-1,$$

where

$$a_j^{(1)} = \frac{j\alpha\pi_{0,j} + \mu a_{j+1}^{(1)}}{\lambda + \mu + (j-1)\alpha - \mu b_{j+1,j}^{(1)}}$$

$$b_{j,k}^{(1)} = \frac{\lambda\beta_{j-k} + \mu b_{j+1,k}^{(1)}}{\lambda + \mu + (j-1)\alpha - \mu b_{j+1,j}^{(1)}}, \qquad k = 1, 2, \ldots, j-1.$$

Remark 2. *At this moment, we can express $\pi_{1,j}$ $(2 \le j \le c)$ in terms of $\pi_{1,1}$ which is further expressed in terms of $\pi_{0,0}$. Furthermore, the probability $\pi_{1,j}$ $(j > c)$ and $\Pi_1(1)$ are calculated from the generating function (9).*

Next, we shift to the general case with h active servers. As in the case of one active server, we write the balance equations.

$$(\lambda + h\mu)\pi_{h,h} = \alpha\pi_{h-1,h} + h\mu\pi_{h,h+1} + (h+1)\mu\pi_{h+1,h+1}, \qquad j = h,$$

$$(\lambda + h\mu + (j-h)\alpha)\pi_{h,j} = (j-h+1)\alpha\pi_{h-1,j} + \sum_{i=1}^{j-h} \lambda\beta_i\pi_{h,j-i} + h\mu\pi_{h,j+1}, \tag{10}$$

$$h+1 \le j \le c-1,$$

$$(\lambda + h\mu + (c-h)\alpha)\pi_{h,j} = (c-h+1)\alpha\pi_{h-1,j} + \sum_{i=1}^{j-h} \lambda\beta_i\pi_{h,j-i} + h\mu\pi_{h,j+1}, \qquad j \ge c, \tag{11}$$

We proceed in the same manner as in the case of one active server. Equation (11) is transformed as follows.

$$(\lambda + h\mu + (c-h)\alpha)\Pi_h(z) = (c-h+1)\alpha\Pi_{h-1}(z) + \lambda\beta(z)\Pi_h(z) + \lambda \sum_{k=h}^{c-1} \pi_{h,k}\beta_k(z)$$

$$+ \frac{h\mu}{z}(\Pi_h(z) - \pi_{h,c}).$$

Transforming this equation, we obtain

$$[(\lambda + h\mu + (c-h)\alpha)z - \lambda z\beta(z) - h\mu]\Pi_h(z)$$
$$= (c-h+1)\alpha z\Pi_{h-1}(z) + \lambda\sum_{k=h}^{c-1}\pi_{h,k}z\beta_k(z) - h\mu\pi_{h,c}. \tag{12}$$

Let $0 < z_h < 1$ denote the unique solution of $(\lambda + h\mu + (c-h)\alpha)z - \lambda z\beta(z) - h\mu = 0$ in $(0,1)$.

In case $\beta(z) = z$, z_h is explicitly given by

$$z_h = \frac{\lambda + h\mu + (c-h)\alpha - \sqrt{[\lambda + h\mu + (c-h)\alpha]^2 - 4h\lambda\mu}}{2\lambda}.$$

Substituting $z = z_h$ into (12) yields,

$$h\mu\pi_{h,c} = (c-h+1)\alpha z_h\Pi_{h-1}(z_h) + \lambda\sum_{k=h}^{c-1}\pi_{h,k}z_h\beta_k(z_h).$$

Thus, we have

$$\pi_{h,c} = \frac{(c-h+1)\alpha z_h\Pi_{h-1}(z_h)}{h\mu} + \sum_{k=h}^{c-1}\frac{\lambda z_h\beta_k(z_h)}{h\mu}\pi_{h,k}.$$

So, we have

$$\pi_{h,c} = a_c^{(h)} + \sum_{k=h}^{c-1} b_{c,k}^{(h)}\pi_{h,k},$$

where

$$a_c^{(h)} = \frac{(c-h+1)\alpha z_h\Pi_{h-1}(z_h)}{h\mu}, \qquad b_{c,k}^{(h)} = \frac{\lambda z_h\beta_k(z_h)}{h\mu}.$$

By mathematical induction using (10), as in the case of one active server, we can obtain the following recursion.

$$\pi_{h,j} = a_j^{(h)} + \sum_{k=h}^{j-1} b_{j,k}^{(h)}\pi_{h,k}, \qquad j = h+1,\ldots,c, \tag{13}$$

where

$$a_j^{(h)} = \frac{(j-h+1)\alpha\pi_{h-1,j} + h\mu a_{j+1}^{(h)}}{\lambda + h\mu + (j-h)\alpha - h\mu b_{j+1,j}^{(h)}}, \tag{14}$$

$$b_{j,k}^{(h)} = \frac{\lambda\beta_{j-k} + h\mu b_{j+1,k}^{(h)}}{\lambda + h\mu + (j-h)\alpha - h\mu b_{j+1,j}^{(h)}}. \tag{15}$$

Thus, at this moment, $\pi_{h,h}$ can be computed in terms of the known quantities. Indeed, by the horizontal cut between states with less than h active servers and those with at least h active servers, we have

$$h\mu\pi_{h,h} = \sum_{j=h-1}^{c-1}\min(j-h+1, c-h+1)\alpha\pi_{h-1,j} + (c-h+1)\alpha\Pi_{h-1}(1).$$

Once the probability $\pi_{h,h}$ is known, using the recursions (13), (14) and (15), we compute all the probabilities $\pi_{h,j}$ $(j = h+1, h+2, \ldots, c)$.

Finally, the case $i = c$ requires special treatment. Balance equations read as follows.

$$(\lambda + c\mu)\pi_{c,c} = \alpha\pi_{c-1,c} + c\mu\pi_{c,c+1}, \qquad j = c, \tag{16}$$

$$(\lambda + c\mu)\pi_{c,j} = \alpha\pi_{c-1,j} + \lambda\pi_{c,j-1} + c\mu\pi_{c,j+1}, \qquad j \geq c+1. \tag{17}$$

Multiplying (16) by z^0 and (17) by z^{j-c} and summing up over $j \geq c$ yields

$$\begin{aligned}(\lambda + c\mu)\Pi_c(z) &= \alpha\Pi_{c-1}(z) + \lambda\sum_{j=c}^{\infty}\left(\sum_{i=1}^{j-c}\beta_i\pi_{c,j-i}\right)z^{j-c} + \frac{c\mu}{z}(\Pi_c(z) - \pi_{c,c}) \\ &= \alpha\Pi_{c-1}(z) + \lambda\beta(z)\Pi_c(z) + \frac{c\mu}{z}(\Pi_c(z) - \pi_{c,c}).\end{aligned}$$

Arranging this equation, we obtain

$$f_c(z)\Pi_c(z) = \alpha z\Pi_{c-1}(z) - c\mu\pi_{c,c}, \tag{18}$$

or equivalently,

$$\Pi_c(z) = \frac{\alpha z\Pi_{c-1}(z) - c\mu\pi_{c,c}}{f_c(z)}, \tag{19}$$

where

$$f_c(z) = (\lambda + c\mu)z - \lambda z\beta(z) - c\mu.$$

In case $c = 1$, plugin $\Pi_0(z)$ into (19), we obtain

$$\Pi_1(z) = \frac{\lambda(\lambda+\alpha)\pi_{0,0}(\beta(z)-1)}{(\lambda + \alpha - \lambda\beta(z))f_1(z)}. \tag{20}$$

It should be noted that the numerator and denominator of the right-hand side of (19) vanish at $z = 1$. The former vanishes due to the horizontal cut between the states with c active servers and those with $c-1$ active servers.

Thus, applying L'Hopital's rule, we obtain

$$\Pi_c(1) = \frac{\alpha\Pi'_{c-1}(1) + \alpha\Pi_{c-1}(1)}{c\mu - \lambda\beta'(1)}.$$

At this moment, all the probabilities $\pi_{i,j}$ $(j \leq c)$ and the generating functions $\Pi_i(z)$ $(i = 0, 1, \ldots, c)$ are expressed in terms of $\pi_{0,0}$, which is uniquely determined using the following the normalization condition.

$$\sum_{i=0}^{c-1}\sum_{j=i}^{c-1}\pi_{i,j} + \sum_{i=0}^{c}\Pi_i(1) = 1.$$

Especially, for the case $c = 1$, from (5), (20) and

$$\pi_{0,0} + \Pi_0(1) + \Pi_1(1) = 1,$$

we obtain

$$\pi_{0,0} = \frac{\alpha}{\lambda+\alpha}(1-\rho),$$

where $\rho = \lambda\beta'(1)/\mu$. Thus, in case $c = 1$, the generating function of the number of jobs in the system is given as follows.

$$\Pi(z) = \widehat{\Pi}_0(z) + z\Pi_1(z) = \frac{(\mu - \lambda\beta'(1))(z-1)}{(\lambda+\mu)z - \mu - \lambda z\beta(z)} \times \frac{\alpha}{\lambda + \alpha - \lambda\beta(z)}.$$

The first term on the right-hand side represents the number of jobs in an $M^X/M/1$ queue with batch arrivals, whose batch sizes have generating function $\beta(z)$. Furthermore, the second term represents the generating function for the number of jobs arrived during an exponentially distributed time with mean $1/\alpha$. This shows the decomposition formula for the number of jobs in the system for the single-server queue with batch arrival and setup time.

Remark 3. *In order to calculate* $\Pi_c(1)$*, we need the derivative* $\Pi'_{c-1}(1)$ *for which the derivative* $\Pi'_i(1)$ *(*$i = c-2, \ldots, 1, 0$*) is needed. Fortunately, the derivative* $\Pi'_0(1)$ *is explicitly obtained because of the explicit expression of* $\Pi_0(z)$*. We can further derive the derivatives of an arbitrary order* n*,* $\Pi_i^{(n)}(1)$ *(*$i = 0, 1 \ldots, c$*) using equations for the generating functions (4), (12) and (18) as in [8, 10] for the models without batch arrivals, i.e.,* $\beta(z) = z$*.*

Remark 4. *The computational complexity of the generating function approach is* $O(c^2)$*. Indeed,* $\pi_{i,j}$ *(*$i \le j, 0 \le j \le c$*) by the order*

$$(0,0) \to (0,1) \to (0,c) \to (1,1) \to (1,2) \to \cdots \to (1,c) \to \cdots \to (c,c).$$

As a result, the complexity is of order $\sum_{i=1}^{c} i = c(c+1)/2 = O(c^2)$*.*

Remark 5. *In this section, we obtain a numerical procedure to compute all the probabilities* $\pi_{i,j}$ *(*$j \le c$*) and the generating functions. In case* z_h *(*$h = 1, 2 \ldots, c-1$*) is explicitly obtained (*$\beta(z) = z$ *or geometric batch-size distribution etc.), we obtain explicit expressions for all the probabilities* $\pi_{i,j}$ *and the generating functions.*

Remark 6. *Our model is a special case of triangular M/G/1-type Markov chains considered in [6]. Here, our results are derived using the generating function method, while matrix analytic methods are used in [6].*

3.2. Interpretations for the conditional decomposition

Let $\widehat{\Pi}_{c-1}(z)$ denote the generating function of the number of waiting jobs when the number of active servers is $c-1$. By definition, we have

$$\widehat{\Pi}_{c-1}(z) = \sum_{j=c-1}^{\infty} \pi_{c-1,j} z^{j-(c-1)}.$$

Thus, we have

$$\widehat{\Pi}_{c-1}(z) = z\Pi_{c-1}(z) + \pi_{c-1,c-1}.$$

We have derived the following result.

$$\Pi_c(z) = \frac{\alpha\widehat{\Pi}_{c-1}(z) - c\mu\pi_{c,c} - \alpha\pi_{c-1,c-1}}{f_c(z)},$$

$$\Pi_c(1) = \frac{\alpha\widehat{\Pi}'_{c-1}(1)}{c\mu - \lambda\beta'(1)}.$$

Let $Q^{(c)}(t)$ denote the conditional queue length given that all c servers are active in the steady state, i.e.,

$$\mathbb{P}(Q^{(c)} = i) = \mathbb{P}(N(t) = i + c \mid C(t) = c).$$

Let $P_c(z)$ denote the generating function of $Q^{(c)}$. We have

$$\begin{aligned}
P_c(z) &= \frac{\Pi_c(z)}{\Pi_c(1)} \\
&= \frac{\alpha\widehat{\Pi}_{c-1}(z) - c\mu\pi_{c,c} - \alpha\pi_{c-1,c-1}}{\alpha\Pi'_{c-1}(1)} \frac{c\mu - \lambda\beta'(1)}{f_c(z)} \\
&= \frac{\widehat{\Pi}_{c-1}(z) - \widehat{\Pi}_{c-1}(1)}{\widehat{\Pi}'_{c-1}(1)(z-1)} \frac{(c\mu - \lambda\beta'(1))(z-1)}{f_c(z)} \\
&= \frac{\sum_{j=1}^{\infty} \pi_{c-1,c-1+j}(z^j - 1)}{\widehat{\Pi}'_{c-1}(1)(z-1)} \frac{(c\mu - \lambda\beta'(1))(z-1)}{f_c(z)} \\
&= \frac{\sum_{j=1}^{\infty} \pi_{c-1,c-1+j} \sum_{i=0}^{j-1} z^i}{\widehat{\Pi}'_{c-1}(1)} \frac{(c\mu - \lambda\beta'(1))(z-1)}{f_c(z)} \\
&= \frac{\sum_{i=0}^{\infty} \left(\sum_{j=i+1}^{\infty} \pi_{c-1,c-1+j}\right) z^i}{\widehat{\Pi}'_{c-1}(1)} \frac{(c\mu - \lambda\beta'(1))(z-1)}{f_c(z)},
\end{aligned}$$

where we have used $c\mu\pi_{c,c} = \alpha\Pi_{c-1}(1) = \alpha(\widehat{\Pi}_{c-1}(1) - \pi_{c-1,c-1})$ in the second equality.

It should be noted that

$$\frac{(c\mu - \lambda\beta'(1))(z-1)}{f_c(z)}$$

is the generating function of the number of jobs in an $M^X/M/1$ queue with the arrival rate, the service rate and the generating function of the batch size are λ, $c\mu$ and $\beta(z)$, respectively. We give a clear interpretation for the generating function

$$\frac{\sum_{i=0}^{\infty} \left(\sum_{j=i+1}^{\infty} \pi_{c-1,c-1+j}\right) z^i}{\widehat{\Pi}'_{c-1}(1)}.$$

For simplicity, we define

$$p_{c-1,i} = \frac{\sum_{j=i+1}^{\infty} \pi_{c-1,c-1+j}}{\widehat{\Pi}'_{c-1}(1)}, \qquad i \in \mathbb{Z}_+,$$

where $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$. We have

$$\sum_{j=i+1}^{\infty} \pi_{c-1,c-1+j} = \mathbb{P}(N(t) - C(t) > i \mid C(t) = c-1)\mathbb{P}(C(t) = c-1),$$

and

$$\widehat{\Pi}'_{c-1}(1) = \mathbb{E}[N(t) - C(t) \mid C(t) = c-1]\mathbb{P}(C(t) = c-1).$$

Thus, we have

$$p_{c-1,i} = \frac{\mathbb{P}(N(t) - C(t) > i \mid C(t) = c-1)}{\mathbb{E}[N(t) - C(t) \mid C(t) = c-1]}.$$

It should be noted that $N(t) - C(t)$ is the number of jobs in the system that are waiting for the last server (in setup mode) to be active.

Thus, $p_{c-1,i}$ $(i = 0, 1, 2, \dots)$ represents distribution of the number of waiting jobs in front of an arbitrary waiting customer under the condition that $c-1$ servers are active and the

last server is in setup mode. Let Q_{Res} denote the random variable with the distribution $p_{c-1,i}$ $(i = 0, 1, 2, \dots)$. Thus our decomposition result is summarized as follows.

$$Q^{(c)} \stackrel{d}{=} Q^{(c)}_{ON-IDLE} + Q_{Res}.$$

We observe that Q_{Res} represents the number of extra jobs due to the setup time. This interpretation is also presented in [5].

References

[1] Artalejo, J. R., Economou, A. and Lopez-Herrero, M. J. (2005). Analysis of a multiserver queue with setup times. Queueing Systems, 51(1-2), 53-76.

[2] Barroso, L. A. and Holzle, U. (2007). The case for energy-proportional computing. Computer, 40 (12), 33-37.

[3] Gandhi, A, Harchol-Balter, M. and Adan, I. (2010). Server farms with setup costs. Performance Evaluation, 67, 1123–1138.

[4] Gandhi, A., Doroudi, S., Harchol-Balter, M. and Scheller-Wolf, A. (2014). Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. Queueing Systems, 77(2), 177-209.

[5] Tian N., Li Q. L. and Gao J. (1999). Conditional stochastic decompositions in the M/M/c queue with server vacations. Stochastic Models, 15, 367-377.

[6] Van Houdt B. and van Leeuwaarden J. S.H. (2011), Triangular M/G/1-type and tree-like QBD Markov chains, INFORMS Journal on Computing, 23(1), 165-171, 2011.

[7] Phung-Duc, T. (2014). Server farms with batch arrival and staggered setup. Proceedings of the 5th Symposium on Information and Communication Technology, 240-247.

[8] Phung-Duc, T. (2017). Exact solutions for M/M/c/setup queues. Telecommunication Systems, 64(2), 309-324.

[9] Phung-Duc, T. (2020). Batch arrival multiserver queue with state-dependent setup for energy-saving data center. In Applied Probability and Stochastic Processes, 421-440. Singapore: Springer Singapore.

[10] Le-Anh, T. and Phung-Duc, T. (2025). A fast algorithm for multiserver queueing systems with setup times and power-saving modes. Proceedings of the 36th International Teletraffic Congress (ITC-36) (pp. 1-9). IEEE.

Acknowledgments

This work was supported by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University. TP was supported in part by JSPS KAKENHI Grant Number JP 21K11765, F-MIRAI: R&D Center for Frontiers of MIRAI in Policy and Technology, the University of Tsukuba and Toyota Motor Corporation Collaborative R&D center, and Kayamori Foundation of Informational Science Advancement.