Optimality Analysis of a Two-Server Queuing System with the RID Discipline

XIAN LIU

Department of Systems Engineering

University of Arkansas at Little Rock

Little Rock, USA

CHANGCHENG HUANG

Department of Systems and Computer Engineering

Carleton University

Ottawa, Canada

Latency, jitter, and reliability are three KPIs in QoS of modern edge and cloud computing systems. Their analytical insights can be well gained through queuing theory. In this paper, we develop a queuing model with two parallel heterogeneous servers with a randomly-initial-dispatch discipline. This model generalizes several existing models, yet simple enough as it can be configured by a couple of parameters. With this simplicity, we analytically prove the optimality of several standard measures of effectiveness for the latency and jitter.

CCS CONCEPTS

Networks → Computer-Communication Network

KEYWORDS

Latency and jitter control, optimization, quality of service, queuing systems.

ACM Reference format:

Author 1 and Author 2. 2024. Optimality analysis of a two-server queuing system with the RID discipline. In Proceedings of the 8th International Conference on Cloud and Big Data Computing, Oxford, United Kingdom, 12 pages.

1 INTRODUCTION

Latency, jitter, and reliability are three *key performance indicators* (KPIs) in *quality of service* (QoS) of modern edge and cloud computing systems. Toward the 6G era, these KPIs need to be further improved ([1], [2]). Through the 3GPP initiative of *enhanced ultra-reliable and low latency communication* (eURLLC), the considered benchmarks are *sub-ms* and *7-nine* for the latency/jitter and reliability, respectively. If implementable, these benchmark KPIs will significantly promote several active paradigms in adv-5G and 6G, such as *vehicle-to-everything* (V2X), *augmented/virtual reality* (AR/VR), *mobile edge computing* (MEC), and *non-terrestrial networking* (NTN). The progress of these paradigms, together with other state-of-the-art visions, would bring

significant advancements of information science and technology, including *industrial Internet of things* (IIoT) and *military metaverse* (MM), which will interweave with the evolution of modern cloud architectures.

The infrastructural concepts of eURLLC, and its predecessor URLLC, certainly pays the main attention to latency and reliability. Reliability is a highly system-wide notion, involving many cross-layer issues in modern communication systems and networks. However, if touching its cores, the parallel structure of services is fundamental. On the other hand, in most communication systems, the *end-to-end* (E2E) latency basically consists of four types of delays: processing, propagation, transmission, and queuing. The processing delay is primarily due to the CPU rating, while the propagation delay is directly proportionally to distance. The transmission delay in modern systems usually occupies a small portion of E2E latency. It is not difficult to understand that the first three types are mainly related to the physical layer of wireless communication systems. However, queuing delays may incur in any layer. Driven by the diversity of the coming 6G wireless paradigms, sufficient efforts in regaining queuing intrinsic must be made. The present work develops a basic queuing model that takes both reliability and latency into account. The former is reflected by a parallel-server structure, and the latter is marked by a detailed analysis for the *measures of effectiveness* (MoE). In particular, a main attention will be paid to the optimality of second order MoEs, which generally characterizes the jitter of information transfer. Minimizing jitters is one of the crucial tasks in several emerging architectures like IIoT and MM.

Several works on elementary queuing analysis, well customized for conventional communication systems, were already widely known (e.g., [3], [4]). As the technologies evolved from 4G to 5G, there was a recurring interest in queuing models [5-8]. In particular, the authors of [8] developed a generic queuing model to describe the behavior of information flow in V2X. Their work showed that the results of queuing analysis were consistent with simulation. Nevertheless, the work of [8] was based on a single-server Markovian model. In principle, several QoS metrics can be improved by introducing multiple parallel servers, such as *M/M/c* $(c \geq 2)$. Besides latency and jitter, the reliability is also promoted. The Markovian property of M/M/c is equivalent to the Poisson process and usually provides the analytical tractability. In many disciplines of *operations research* (OR), the M/M/c model is often used as a benchmark for more sophisticated queuing systems ([3, Ch.3], [4, Ch.12]). In some applications such as V2X, it is reasonable to adopt small values of c. Accordingly, considering both applicability and trackability, in this paper we investigate a queuing model with two parallel heterogeneous servers (Fig. 1).

Different than most studies in public literature, however, our model is a generalized M/M/2 system with two controllable parameters (a, β), where $0 \leq a \leq 1, 0 < \beta \leq 1,$ and $\beta = \mu_2 / \mu_1$. In this system, the item arriving at the queue-head is controlled with the following discipline: (1) if both servers are idle, then it will go to server 1 with probability a or server 2 with probability $(1-a)$; (2) if one of the two servers is idle, then it goes to the idle server; (3) if both servers are busy, then it will be queued. Note that the usual FCFS (first-come-first-served) discipline is still adopted in queuing. The discipline defined above will be referred to as *randomly-initial-dispatch* (RID) in this paper. In theory, this queuing model is not elementary Markovian since its overall service time follows a hybrid distribution. However, as described in Sections III and V, this hybridity only rises in a few states. Therefore, for simplicity, throughout the present work, the model is named the *M/M/2H RID* system, where "H" stands for heterogeneous servers.

The rest of this paper is organized as follows. The main differences of our study than some early works are summarized in Section II. Then, in Section III, the system essentials are detailed. Next, in Section IV, several MoEs are derived, and the optimality is proven. Section V is devoted to the distribution of queuing time. Finally, the paper is concluded in Section VI. Note that the terms "delay" and "latency" are interchangeably used in this paper. The notation "$E(\bullet)$" is the expectation of its argument.

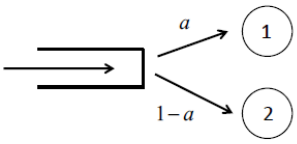

Figure 1: M/M/2H with RID discipline.

2 RELATED WORKS

The discussions in most queuing textbooks are restricted to the homogeneous-server systems (e.g., [3], [4]). It is well known that the M/M/c system with homogeneous servers is superior to the system built on *c* identical M/M/1 facilities with regard to the sojourn time of packets [3, pp. 213-214]. Relaxing the homogeneous condition, several researchers have also investigated the heterogeneous-server systems. Most reported studies reasonably selected two heterogeneous servers to develop the canonical model. In the context of electrical engineering, one of the early noteworthy papers was [9]. The authors investigated an M/M/2H system with a special dispatch discipline: the new arrival does not have to go to the available slow server. In the literature of OR, this treatment is sometimes referred to as a queuing system with *vacations* (or other similar terms). The authors of [9] proved that there exists an optimal queuing threshold that minimizes the mean system size. Their accomplishment has inspired a series of studies along this avenue (see [10], [11], and the references therein).

Our M/M/2H RID model does not consider the vacation-type discipline. However, the M/M/2H RID model can be refined into several special models: (1) $a = 0.5$ leads to the standard model in which the arrivals are dispatched to either server with the same chance; (2) $a = 1$ leads to the "fast-server-first" model; and (3) $a = 1/(1+\beta) = \mu_1/(\mu_1+\mu_2)$.

Moreover, we note that most existing studies in M/M/2H did not include the analysis for the second order *measures of effectiveness* (MoEs) such as the variance. A reason was perhaps due to the analytical complexity since that the handy Little's Theorem in the elementary queuing theory ([3, Ch.3], [4, Ch.12]) is only applicable to the first order MoEs. However, the second order MoEs have a leading importance in a variety of multimedia applications as they characterize the jitter behavior of concerned traffic flows. Therefore, the optimality details of the second order MoEs will provide a clear analytical guideline for these applications. In the present work, significant attention is paid to the optimality of several second order MoEs.

3 SYSTEM ESSENTIALS

Consistent with the conventions in queuing theory, the following notations are adopted throughout this paper: N_q = queuing size; T_q = queuing time; N = system size (the packet number in the queue and servers); T = system sojourn time (the sum of queuing time and server time). Note that N_q and N are discrete *random variables* (RVs), while T_q and T are continuous RVs. The theoretical studies would involve their probability functions and the induced moments (typically the 1st and 2nd orders). The present work conducts a theoretical analysis the *probability mass functions* (PMFs) of (N_q, N) and the *probability density function* (PDF) of T_q. The *cumulative distribution function* (CDF) can be derived from the PDF. The distribution of T is more involved and is worthwhile developing a companion paper.

The analysis complexity mainly comes from the selected queuing discipline. The state transition diagram is used to conduct the analysis (Fig. 2), where two types of notations are adopted: In the left part corresponding to the initial states, doublet (j_1, j_2) stands for the number of customers in the server 1 and 2, respectively, while in the right part the single number, representing the total number of customers in the system, is sufficient to describe the states. In addition, the following notations will be used to represent the state probabilities:

$P_0 = P(0,0)$, $P_{1a} = P(1,0)$, $P_{1b} = P(0,1)$, $P_2 = P(1,1)$, and $P_k = P(k)$ $(k > 2)$. Aware of that the state transition diagram consists of the initial state and the equilibrium state, we derive the system size PMF in two steps.

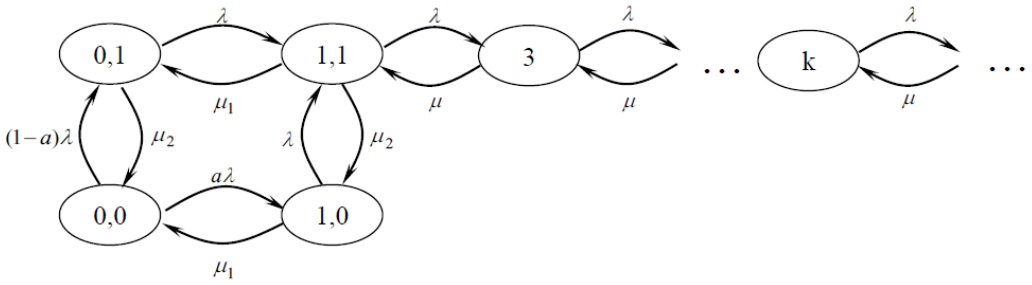

Figure 2: The state transition diagram of the M/M/2H system with the RID discipline.

3.1 Step 1

Let λ be the mean arrival rate and μ_j $(j = 1, 2)$ the mean service rate. According to the state transition diagram shown in Figure 2, the balance equations for the initial state can be formulated and solved as below:

$$(\lambda + \mu_1)P_{1a} = a\lambda P_0 + \mu_2 P_2, \quad (1)$$

$$(\lambda + \mu_2)P_{1b} = (1-a)\lambda P_0 + \mu_1 P_2, \quad (2)$$

$$\lambda P_0 = \mu_1 P_{1a} + \mu_2 P_{1b}. \quad (3)$$

Since this linear equation system involves four unknowns $(P_0, P_{1a}, P_{1b}, P_2)$, it can be solved by fixing one unknown as the parameter, called the *pivot*. In the following analysis we use P_2 as the pivot. With a little more effort, the above system can be converted into:

$$P_{1a} + P_{1b} = P_2 / \rho, \quad (4)$$

$$[2\mu_1(a-1) - \lambda]P_{1a} + (2a\mu_2 + \lambda)P_{1b} = (\mu_1 - \mu_2)P_2, \quad (5)$$

where $\rho = \lambda / \mu$ and $\mu = \mu_1 + \mu_2$. Without of loss generality, we assume $\mu_2 \le \mu_1$ throughout this paper.

[Definition 1] The parameter $\beta = \mu_2 / \mu_1$ is called the *service ratio*.

Solving the system consisting of eqs. (4) and (5), we obtain:

$$P_{1a} = \frac{(1 + a/\rho)\mu_2 P_2}{\lambda + a\mu_2 + (1-a)\mu_1}, \quad (6)$$

$$P_{1b} = \frac{[1 + (1-a)/\rho]\mu_1 P_2}{\lambda + a\mu_2 + (1-a)\mu_1}. \quad (7)$$

Substituting eqs. (6) and (7) into eq. (3), we have:

$$P_0 = \frac{(2+1/\rho)\mu_1\mu_2 P_2}{\lambda[\lambda+(1-a)\mu_1+a\mu_2]}. \qquad (8)$$

As a result, up to this point, the linear system of equations has been defined by eqs. (1) to (3) has been solved in terms of P_2.

3.2 Step 2

The balance equations for the equilibrium states can be formulated and solved as below. First, in state (1, 1) we have:

$$(\lambda+\mu)P_2 = \lambda P_{1a} + \lambda P_{1b} + \mu P_3. \qquad (9)$$

Substituting eqs. (6) and (7) into eq. (9) and reducing the intermediary terms, we obtain:

$$P_3 = \rho P_2. \qquad (10)$$

Next, it follows from the flow balance at state 3 that:

$$(\lambda+\mu)P_3 = \lambda P_2 + \mu P_4. \qquad (11)$$

Putting the result of eq. (10) into eq. (11) leads to:

$$P_4 = \rho^2 P_2. \qquad (12)$$

Recursively this way, the general solution can be derived by induction:

$$P_k = \rho^{k-2} P_2. \qquad (k = 2,3,\cdots) \qquad (13)$$

As a result, the equilibrium state part has been solved in terms of P_2. Consequently,

$$\sum_{k=2}^{\infty} P_k = P_2 \sum_{k=2}^{\infty} \rho^{k-2} = \frac{P_2}{1-\rho} \overset{def.}{=} r. \qquad (14)$$

Note that r is the *Erlang C* probability. Finally, the value of P_2 can be determined by the probability axiom $\sum_{k=0}^{\infty} P_k = 1$:

$$P_0 + P_{1a} + P_{1b} + \sum_{k=2}^{\infty} P_k$$
$$= P_2\left[\frac{(2+1/\rho)\mu_1\mu_2}{\lambda[\lambda+(1-a)\mu_1+a\mu_2]} + \frac{1}{\rho} + \frac{1}{1-\rho}\right] = 1. \qquad (15)$$

Therefore,

$$P_2 = \left[\frac{(2+1/\rho)\mu_1\mu_2}{\lambda[\lambda+(1-a)\mu_1+a\mu_2]} + \frac{1}{\rho} + \frac{1}{1-\rho}\right]^{-1}$$
$$= \frac{\rho^2(1-\rho)[(a+\rho)\beta^2+(2\rho+1)\beta+(1-a+\rho)]}{(a+\rho)\rho\beta^2+(2\rho+1)\beta+(1-a+\rho)\rho}. \qquad (16)$$

4 THE MEASURES OF EFFECTIVENESS

One of the fundamental MoEs, the first order moment of the system size, can be readily determined by the state probabilities derived in Section IV:

$$E(N) = P_{1a} + P_{1b} + \sum_{k=2}^{\infty} kP_k = \frac{P_2}{\rho(1-\rho)^2}. \qquad (17)$$

In order to derive another MoE, the first order moment of the queuing size, we notice that the queue is empty when the system size is 0, 1, or 2. For this reason, we have:

$$E(N_q) = \sum_{k=3}^{\infty} (k-2)P_k$$
$$= P_2 \sum_{k=3}^{\infty} (k-2)\rho^{k-2} = \frac{\rho P_2}{(1-\rho)^2}. \quad (18)$$

In practice, many applications (e.g., the real-time packet switching communication paradigm) are also characterized by the variances of the system size or the queuing size. To determine these variances, we firstly derive the second order moments:

$$E(N^2) = \sum_{k=1}^{\infty} k^2 P_k = P_{1a} + P_{1b} + \sum_{k=2}^{\infty} k^2 P_k$$
$$= \frac{P_2}{\rho} + P_2 \sum_{k=2}^{\infty} k^2 \rho^{k-2} = \frac{(1+\rho)P_2}{\rho(1-\rho)^3}. \quad (19)$$

$$E(N_q^2) = \sum_{k=3}^{\infty} (k-2)^2 P_k$$
$$= P_2 \sum_{k=3}^{\infty} (k-2)^2 \rho^{k-2} = \frac{\rho(1+\rho)P_2}{(1-\rho)^3}. \quad (20)$$

The variances, therefore, can be determined as follows:

$$\text{var}(N) = E(N^2) - E^2(N)$$
$$= \frac{(1+\rho)P_2}{\rho(1-\rho)^3} - \frac{P_2^2}{\rho^2(1-\rho)^4}, \quad (21)$$

$$\text{var}(N_q) = E(N_q^2) - E^2(N_q)$$
$$= \frac{\rho(1+\rho)P_2}{(1-\rho)^3} - \frac{\rho^2 P_2^2}{(1-\rho)^4}. \quad (22)$$

Finally, the first moments of the system time and the queuing time can be obtained by using Little's Theorem ([3, Ch.3], [4, Ch.12]):

$$E(T) = \frac{E(N)}{\lambda} = \frac{P_2}{\rho\lambda(1-\rho)^2}, \quad (23)$$

$$E(T_q) = \frac{E(N_q)}{\lambda} = \frac{\rho P_2}{\lambda(1-\rho)^2} = \frac{P_2}{\mu(1-\rho)^2}. \quad (24)$$

In the following, we conduct the analysis of optimality.

4.1 First Oeder MoE

[Proposition 1] Given λ and ρ, if $1/2 \le a \le 1$, there exists an optimal service ratio β^* that minimizes $E(N_q), E(T_q), E(N)$, and $E(T)$.

[Proof] With the given conditions, it is sufficient to show that P_2 attains the optimality. According to eq. (16),

$$P_2 = \left[\frac{1+2\rho}{\rho^2 y} + \frac{1}{\rho(1-\rho)} \right]^{-1}, \qquad (25)$$

where

$$y = \frac{(1+\beta)[(a+\rho)\beta + 1 + \rho - a]}{\beta}. \qquad (26)$$

Let $\partial y / \partial \beta = 0$. We obtain:

$$\beta^* = \sqrt{\frac{1+\rho-a}{\rho+a}}. \qquad (27)$$

Note that $\beta^* \in (0,1]$ since $1/2 \le a \le 1$. Moreover,

$$\partial^2 y / \partial \beta^2 > 0. \qquad (28)$$

Therefore, β^* is the minimizer of y. It is also the minimizer of P_2. Accordingly, the minimum of y is:

$$y^* = (1+2\rho)\mu + 2\mu\sqrt{\rho(1+\rho) + a(1-a)}. \qquad (29)$$

Finally, substituting eq. (29) into (25) then (23), we obtain the minimum of $E(T)$:

$$E^*(T) = \left[\mu\rho(1-\rho) + \frac{\mu(1+2\rho)(1-\rho)^2}{1+2\rho+2\sqrt{\rho(1+\rho)+a(1-a)}} \right]^{-1}. \qquad (30)$$

Q.E.D.

Two example profiles of $E(T)$ are illustrated in Fig. 3. The optimality of $E(T_q)$ can be similarly proven. Moreover, the optimality of $E(N_q)$ and $E(N)$ is attained according to Little's Theorem ([3, Ch.3], [4, Ch.12]).

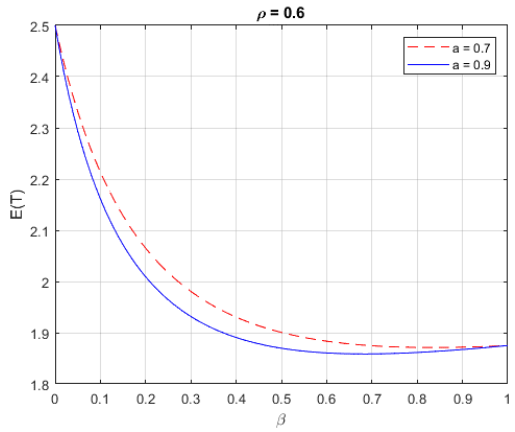

Figure 3: Example profiles of $E(T)$.

It is also interesting to compare with the classic counterpart, shown in the following:

[Corollary 1] The mean sojourn time of the M/M/2H system operating at $\beta = \beta^*$ is not greater than that of the M/M/2 system with homogeneous servers.

[Proof] Denote the mean sojourn time of the homogeneous M/M/2 system by $E_g(T)$. Then, according to elementary queuing theory [3, Ch.3], we have:

$$E_g(T) = \frac{2}{\mu(1-\rho^2)}. \tag{31}$$

On the other hand, it follows from eq. (30) that, when $a = 1/2$,

$$\begin{aligned} & E^*(T)\big|_{a=1/2} \\ &= \left[\mu\rho(1-\rho) + \frac{\mu(1+2\rho)(1-\rho)^2}{1+2\rho+2\sqrt{\rho(1+\rho)+(1/4)}} \right]^{-1} \\ &= \frac{2}{\mu(1-\rho^2)} = E_g(T). \end{aligned} \tag{32}$$

Therefore, $E^*(T) \le E_g(T)$, since the term $a(1-a)$ reaches its maximum at $a = 1/2$.

Q.E.D.

4.2 Second Order MoE

In queuing theory, Little's Theorem ([3, Ch.3], [4, Ch.12]) is only applicable to the first order MoEs. The analysis for the second order MoEs needs the knowledge of the probability distributions. As seen in Section IV, with the aid of PMFs, we have derived the variances of N_q and N. In the following, we conduct the optimality analysis.

For this purpose, it has proven that adopting the Erlang C probability as an intermediate variable is a wise idea. Due to (14) and (16), the Erlang C probability can be elaborated as follows:

$$r = \frac{\rho^2[(a+\rho)\beta^2 + (2\rho+1)\beta + (1-a+\rho)]}{(a+\rho)\rho\beta^2 + (2\rho+1)\beta + (1-a+\rho)\rho}. \tag{33}$$

[Proposition 2] Given λ and ρ, if $1/2 \le a \le 1$, there exists an optimal service ratio β^* that minimizes $\mathrm{var}(N_q)$.

[Proof] It follows from (33) that:

$$\begin{aligned} r &< \frac{\rho^2[(a+\rho)\beta^2 + (2\rho+1)\beta + (1-a+\rho)]}{(a+\rho)\rho\beta^2 + (2\rho+1)\rho\beta + (1-a+\rho)\rho} \\ &= \rho < \frac{1}{2\rho} + \frac{1}{2}. \end{aligned} \tag{34}$$

Next, substituting (33) into (22), we have:

$$\mathrm{var}(N_q) = \frac{\rho(1+\rho)r}{(1-\rho)^2} - \frac{\rho^2 r^2}{(1-\rho)^2}. \tag{35}$$

Consequently,

$$\frac{\partial\,\mathrm{var}(N_q)}{\partial\beta} = \frac{\partial\,\mathrm{var}(N_q)}{\partial r}\frac{\partial r}{\partial\beta}. \tag{36}$$

Due to (34) and (35),

$$\frac{\partial \operatorname{var}(N_q)}{\partial r} = \frac{\rho(1+\rho-2\rho r)}{(1-\rho)^2} > 0.$$

Therefore, the solution β^* in (27) minimizes $\operatorname{var}(N_q)$. *Q.E.D.*

Two example profiles of $\operatorname{var}(N_q)$ are illustrated in Fig. 4.

Note that the relation (34) is a pivot in the proof for $\operatorname{var}(N_q)$. However, it is not sufficient to prove the optimality of $\operatorname{var}(N)$. Prior to conducting a further analysis, we introduce two Lemmas first.

[Lemma 1] There exists an interval of β such that the following inequality holds:

$$r < \frac{\rho(1+\rho)}{2}. \qquad (37)$$

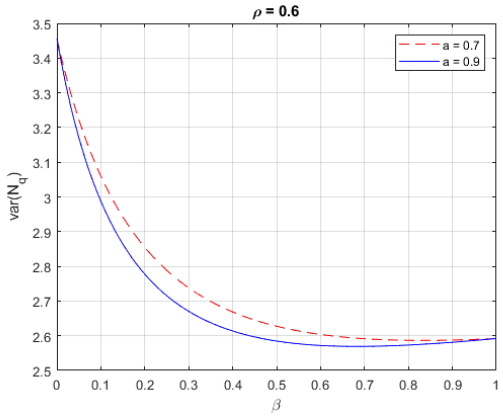

Figure 4: Example profiles of $\operatorname{var}(N_q)$.

[Proof] Using (33), we need to show:

$$\frac{\rho^2[(a+\rho)\beta^2+(2\rho+1)\beta+(1-a+\rho)]}{(a+\rho)\rho\beta^2+(2\rho+1)\beta+(1-a+\rho)\rho} < \frac{\rho(1+\rho)}{2}. \qquad (38)$$

This is equivalent to:

$$\rho(a+\rho)\beta^2-(2\rho+1)\beta+(1-a+\rho)\rho < 0. \qquad (39)$$

To ensure the existence of real roots in the left side, we need to verify the discriminant:

$$\Delta = (2\rho+1)^2 - 4p(a+p)(1-a+\rho)\rho \geq 0. \qquad (40)$$

One key fact to prove (40) is:

$$a(1-a) \leq \frac{1}{4}. \qquad (0 \leq a \leq 1) \qquad (41)$$

Accordingly, for $a \neq 1/2,$ we have two distinct real roots:

$$\begin{cases} \beta_1 = \dfrac{(2\rho+1)-\sqrt{(2\rho+1)^2-4\rho(a+\rho)(1-a+\rho)\rho}}{2\rho(a+\rho)}, \\ \beta_2 = \dfrac{(2\rho+1)+\sqrt{(2\rho+1)^2-4\rho(a+\rho)(1-a+\rho)\rho}}{2\rho(a+\rho)}. \end{cases} \tag{42}$$

Therefore, when $\beta_1 < \beta < \beta_2$, the relation (37) holds. *Q.E.D.*

Next, with the expressions given in (27) and (42), it is easy to show:

[Lemma 2] The minimizer of the first order MoE satisfies the following relation:

$$\beta_1 < \beta^* < \beta_2. \tag{43}$$

[Proposition 3] Given λ and ρ, if $1/2 < a \le 1$, there exists an optimal service ratio β^* that minimizes $\mathrm{var}(N)$.

[Proof] Substituting (33) into (21), we have:

$$\mathrm{var}(N) = \frac{(1+\rho)r}{\rho(1-\rho)^2} - \frac{r^2}{\rho^2(1-\rho)^2}. \tag{44}$$

Consequently,

$$\frac{\partial \mathrm{var}(N)}{\partial \beta} = \frac{\partial \mathrm{var}(N)}{\partial r}\frac{\partial r}{\partial \beta}, \tag{45}$$

where

$$\frac{\partial \mathrm{var}(N)}{\partial r} = \frac{\rho(1+\rho)-2r}{\rho^2(1-\rho)^2}. \tag{46}$$

Due to Lemmas 1 and 2 as well as the function continuity, in the neighborhood of β^*,

$$\frac{\partial \mathrm{var}(N)}{\partial r} > 0. \tag{47}$$

Therefore, the solution β^* in (27) minimizes $\mathrm{var}(N)$. *Q.E.D.*

The profile of $\mathrm{var}(N)$ is similar to $\mathrm{var}(N_q)$.

5 DISTRIBUTION OF QUEUING TIME

As mentioned in Section IV, the variance analysis needs the knowledge of the probability functions. For the discrete RVs such as N_q and N, the analysis was done in the preceding section. In this section, we shift the attention to the queuing time T_q, which is a continuous RV when $t > 0$. However, at $t = 0$, there is a single-point probability.

[Proposition 4] When both servers are busy, the partial PDF of T_q takes the following form:

$$f_{T_q,r}(t) = r(\mu-\lambda)\exp[-(\mu-\lambda)t], \tag{48}$$

where $\mu = \mu_1 + \mu_2$. The term "partial" is used for the simplicity of presentation, since the integration of (48) is less than 1.

[Proof] When both servers are busy, a new arrival will see that there are k $(k \ge 2)$ items in the system (including two items in the servers). For those two items in the servers already, let T_s be the time until the next departure

(so a server will be available). Due to the memoryless Markov property ([3, Ch.3], [4, Ch.12]), the complementary CDF (CCDF) of T_s can be expressed as:

$$\Pr(T_s > t) = \Pr(T_{s1} > t, T_{s2} > t) = \exp[-(\mu_1 + \mu_2)t]. \quad (49)$$

Accordingly, the PDF is:

$$f_{T_s}(t) = (\mu_1 + \mu_2)\exp[-(\mu_1 + \mu_2)t] = \mu \exp(-\mu t), \quad (50)$$

Therefore, for the new arrival, the queuing time will be the sum of $k-1$ such T_s. With the Laplace transform or direct convolution, we obtain the PDF of the queuing time:

$$f_{T_q}(t;k) = \frac{\mu^{k-1}t^{k-2}}{(k-2)!}\exp(-\mu t). \quad (51)$$

With the PMF (13), the average over all realizations $k \geq 2$ is:

$$f_{T_q,r}(t) = \sum_{k=2}^{\infty} P_k f_{T_q}(t;k) = r(\mu - \lambda)\exp[-(\mu - \lambda)t]. \quad (52)$$

Q.E.D.

[Corollary 2] When both servers are busy, the partial CDF of T_q takes the following form:

$$F_{T_q,r}(t) = r - r\exp[-(\mu - \lambda)t]. \quad (53)$$

The overall CDF is the CDF that characterizes the situations with empty and non-empty queues. Therefore, we have

[Corollary 3] The overall CDF of T_q is:

$$\begin{aligned} F_{T_q}(t) &= (1-r)u(t) + F_{T_q,r}(t) \\ &= (1-r)u(t) + r - r\exp[-(\mu - \lambda)t], \end{aligned} \quad (54)$$

where $u(t)$ is the *unit-step function*.

[Corollary 4] The overall PDF of T_q is:

$$f_{T_q}(t) = (1-r)\delta(t) + f_{T_q,r}(t), \quad (55)$$

where $\delta(t)$ is the *impulse function* (i.e., *Dirac delta function*):

$$\int_{-\infty}^{\infty} \delta(t)dt = 1; \qquad \delta(t) = 0. \qquad (\forall t \neq 0) \quad (56)$$

With the PDF available in (55), we have:

$$E(T_q) = \frac{r}{\mu - \lambda}. \quad (57)$$

Note that (57) is the same as (24) due to (14). Furthermore,

$$\mathrm{var}(T_q) = \frac{r(2-r)}{(\mu - \lambda)^2}. \quad (58)$$

Finally, parallel to Propositions 1, 2, and 3, we have:

[Proposition 5] Given λ and ρ, if $1/2 \leq a \leq 1$, there exists an optimal service ratio β^* that minimizes $\mathrm{var}(T_q)$.

[Proof] It follows from (58) that

$$\frac{\partial}{\partial \beta}\mathrm{var}(T_q) = \frac{\partial}{\partial r}\mathrm{var}(T_q)\frac{\partial r}{\partial \beta} = \frac{2(1-r)}{(\mu-\lambda)^2}\frac{\partial r}{\partial \beta}. \qquad (59)$$

Therefore, the optimality of $\mathrm{var}(T_q)$ is the same as the Erlang probability r, since in (59) $(1-r)/(\mu-\lambda)^2 > 0$.

Q.E.D.

Two profiles of $\mathrm{var}(T_q)$ are illustrated in Fig. 5.

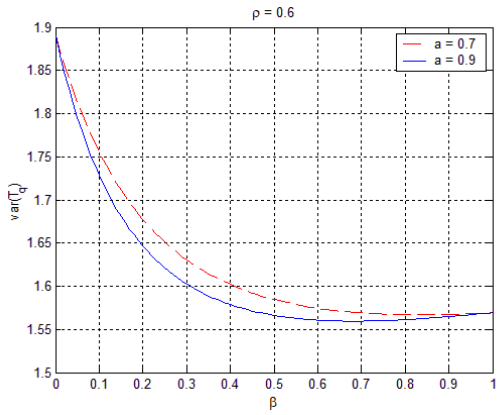

Figure 5: Example profiles of $\mathrm{var}(T_q)$.

6 CONCLUSION

As the URLLC service is to be promoted in several emerging edge and cloud paradigms, there is a regained interest in further investigating some queuing models. In this paper, we describe a generalized queuing model with two configurable parameters. Much attention is paid to the second order MoEs, due to their equal importance in QoS for many edge and cloud paradigms. These MoEs can be readily used as building blocks to design a variety of QoS control models when incorporating the transforms through the notion of utility function in the optimization theory.

REFERENCES

[1] W. Saad, M. Bennis, and M. Chen. 2020. A vision of 6G wireless systems: applications, trends, technologies, and open research problems. IEEE Network 34, 3 (May/Jun. 2020), 134–142.

[2] X. Lin. 2022. An overview of 5G Advanced Evolution in 3GPP Release 18. IEEE Communications Standards Magazine 6, 3 (Sept. 2022), 77-83.

[3] D. Bertsekas and R. Gallager. 1992. Data Networks (2nd ed). Prentice Hall, NJ.

[4] A. L.-Garcia. 2008. Probability, Statistics, and Random Processes for Electrical Engineering (3rd ed). Pearson, Upper Saddle River, NJ.

[5] Y. M. Abdelradi, A. A. El-Sherif, and L. H. Afify. 2019. A queueing theory approach to small-cell assisted IoT traffic offloading. In Proc. of Global Conference on Internet of Things. IEEE. Dubai, United Arab Emirates.

[6] P. Chindanonda, V. Podolskiy, and M. Gerndt. 2019. Metrics for self-adaptive queuing in middleware for Internet of things. In Proc. of the 4th International Workshops on Foundations and Applications of Self Systems. IEEE. Umea, Sweden. 130-133.

[7] Y. Yang, J. Ding, and G. Zhu. 2020. A large-scale access scheme for Internet of Things based on distributed queuing. In Proc. of the 7th International Conference on Information Science and Control Engineering. Changsha, China. 31-35.

[8] X. Wang, Z. Ning, and L. Wang. 2018. Offloading in Internet of vehicles: A fog enabled real-time traffic management system. IEEE Trans. Ind. Informat.14, 10 (Oct. 2018), 4568-4578.

[9] W. Lin and P. Kumar. 1984. Optimal control of a queueing system with two heterogeneous servers. IEEE Transactions on Automatic Control 29, 8 (Aug. 1984), 696-703.

[10] I. Iliadis and Y. Lien. 1993. Resequencing control for a queueing system with two heterogeneous servers. IEEE Transactions on Communications 41, 6 (June 1993), 951-961.

[11] N. Gogate and S. Panwar. 1999. Assigning customers to two parallel servers with resequencing. IEEE Communications Letters 3, 4 (April 1999), 119-121.