On the Large Deviations of Resequencing Queue Size: 2-M/M/1 Case

Ye Xia and David Tse

***Abstract*—In data communication networks, packets that arrive at the receiving host may be disordered for reasons such as retransmission of dropped packets or multi-path routing. Reliable protocols such as TCP require packets to be accepted, i.e., delivered to the receiving application, in the order they are transmitted at the sender. In order to do so, the receiver's transport layer is responsible to temporarily buffer out-of-order packets and to resequence them as more packets arrive. In this paper, we analyze a model where the disordering is caused by multi-path routing. Packets are generated according to a Poisson process. Then, they arrive at a disordering network modelled by two parallel M/M/1 queues, and are routed to each of the queues according to an independent Bernoulli process. A resequencing buffer follows the disordering network. In such a model, the packet resequencing delay is known. However, the size of the resequencing queue is unknown. We derive the probability for the large deviations of the queue size.**

***Index Terms*—Resequencing queue, large deviations, transport protocol**

I. Introduction

Data packets can be disordered by the communication networks for various reasons [1]. For instance, with the help of the destination address contained in every packet, the network can deliberately route packets via different paths to the destination, possibly for load balancing or for reducing transfer delay. Some packets may be dropped when the network is congested or when the packet is corrupted. For reliable communication, the sender must retransmit the dropped packet, possibly causing it to arrive out-of-order at the receiver.

Most applications can only accept packets (which contain application-level data) in the same order they are transmitted at the sender. They typically rely on reliable transport protocols, such as the Transmission Control Protocol (TCP), to temporarily buffer out-of-order packets and to resequence them as new packets arrive. The study of packet disordering and resequencing is important because of the following performance implications.

- Insufficient buffer size causes packet losses and reduced throughput.
- Even when the application can consume the packets infinitely fast, the packets may still suffer resequencing delay, which increases the response time of the application.
- The large number of queued packets create bursty load to the processor. Long queue length is typically the result of one or a few very late packets. During the time of queue build up, the processor stays idle most of the time. When the late packets finally arrive, all queued packets are suddenly eligible for processing.
- The out-of-order packets that have arrived at the receiver must wait at the transport layer, consuming precious system resources such as memory and computation cycles. Since they are shared resources, an unusually large amount of out-of-order packets can negatively affect all applications in the same system.

In our earlier paper [2], we model packet disordering by adding an IID random propagation delay to each packet and derive simple expressions for the required buffer size and the resequencing delay. We demonstrate that these two quantities can be significant and show that the resequencing problem becomes worse as the link speed increases. In this paper, we analyze a model with correlated delays where the disordering is caused by multi-path routing. Packets are generated according to a Poisson process. Then, they arrive at a disordering network modelled by two parallel M/M/1 queues, and are routed to each of the queues according to an independent Bernoulli process. A resequencing buffer follows the disordering network. In such a model, the packet resequencing delay is known. However, the size of the resequencing queue is unknown. We derive the probability for the large deviations of the queue size.

This paper is organized as follows. In Section II, we describe the resequencing model and give the main theorem of the paper. We also discuss the relation of this study with previous studies. Sections III, IV and V constitute the bulk of the paper, which is a proof for the main theorem. We show some implications of the theorem in the concluding section, VI.

II. The Model and the Main Result

The detailed network and resequencing model is shown in Figure 1. Sequentially-numbered customers (or packets) arrive at the disordering network (DN) according to a Poisson process with rate λ. Each customer either enters queue 1 with probability p, or enters queue 2 with probability $1-p$, independent of other customers. Then, the arrival processes to the queues in the DN are independent Poisson processes with rate λ_i, $i \in \{1, 2\}$, where

$$\lambda_1 = p\lambda, \qquad \lambda_2 = (1-p)\lambda.$$

The service times for the customers at queue i are IID exponentially distributed with mean $1/\mu_i$, $i = 1, 2$. Hence, we have two M/M/1 queues in the DN. Due to the multi-path routing, customers may be disordered after the DN. They are resequenced at the resequencing queue (RSQ) that follows the DN. Customers immediately leave the RSQ after they are

Ye Xia is with Computer and Information Science and Engineering Department, University of Florida, Gainesville, FL 30611-6120.

David Tse is with Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720–1770.

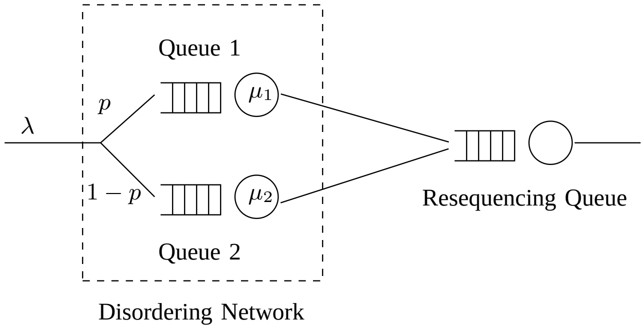


Fig. 1. Network and resequencing model

properly resequenced. That is, customer j leaves the RSQ as soon as all customers $i < j$ have arrived at the RSQ. Note that the server of the RSQ is assumed to have infinite processing capacity. We are interested in computing the stationary queue size of the RSQ. Let q^r be the stationary size of the RSQ. The main result of this paper is the following theorem. Without the loss of generality, let us assume $\mu_1 - \lambda_1 \leq \mu_2 - \lambda_2$. Then,

Theorem 1:

$$\lim_{n\to\infty} \frac{1}{n} \log P\{q^r(t) \geq n\}$$
$$= \max\{\log \frac{\lambda_2}{\lambda_2 + \mu_1 - \lambda_1}, \log \frac{4\lambda_1\mu_1}{(\lambda_1 + \mu_1 + \mu_2 - \lambda_2)^2}\}. \quad (1)$$

The studies that deal with packet disordering due to multipath routing (also including parallel processing or load balancing, etc.) typically analyze an open queueing network, of which the model in Figure 1 is a special case. In some models, a FIFO queue follows the resequencing buffer. The DN is also modelled as a queueing system, whose type typically distinguishes different studies. For instance, the DN is an M/M/∞ queue in [3], an M/GI/∞ queue in [4], a GI/GI/∞ queue in [5], an M/M/2 queue in [6], an M/M/K queue in [7], an M/H_2/K queue in [8], an M/M/2 queue with a threshold-type server assignment policy in [9], two parallel M/M/1 queues with additional fixed propagation delays in [10], and K parallel M/GI/1 queues in [11]. A survey is given in [12]. Most of these studies are concerned mostly with finding the distribution and/or mean of the resequencing delay or end-to-end delay. Several also give results about the number of packets in the resequencing queue. Among the previous studies reviewed here, the most relevant one is [11], where the DN consists of K parallel M/GI/1 queues. In [11], Jean-Marie and Gun derive the distribution of the resequencing delay. In contrast, our results are (i) for the resequencing queue size, (ii) of the large-deviations type, and (iii) for the 2-M/M/1-queue case.

Packet disordering caused by the retransmissions of dropped packets is studied within the context of automatic repeat request (ARQ) protocols [13] [14] [15] [16] [17] [18] [19]. In these studies, ARQ is typically considered as a link-layer protocol running between a sender-receiver pair over a noisy link with constant propagation delay. The sender must retransmit corrupted or dropped packets based on the feedback information it gets from the receiver. Models in this family can not be easily combined into a generic model. Their details and analytical techniques involved differ greatly. Their strength lies in that they typically can model the feedback from the receiver to the sender.

Many previous studies on ARQ models focused on the throughput of the ARQ protocol, or the delay and queue size at the sender side. For instance, Miller and Lin [15] analyzed the throughput for certain Selective-Repeat ARQ schemes. Towsley and Wolf analyzed the queue size and delay at the sender side for the Stop-and-Wait ARQ and the Go-Back-N ARQ in [13], and mean queue length for the Stutter-Go-Back-N ARQ in [20]. Konheim [14] analyzed a Go-Back-N ARQ and a Selective-Repeat ARQ. Anagnostou and Protonotarios [17] analyzed the queue size and delay at the sender side in a Selective-Repeat ARQ model. There are also several studies on the resequencing delay and queue size at the receiver in the ARQ literature. Rosberg and Shacham [18] analyzed a specific Selective-Repeat ARQ protocol over a noisy forward channel from the sender to the receiver and a perfect feedback channel. The distributions of the buffer occupancy and the resequencing delay at the receiver were derived. Rosberg and Sidi [19] extended the above model to allow non-greedy source. In several other studies, Shacham and Towsley [21] considered the resequencing problem for a multicast Selective-Repeat ARQ. Shacham and Shin [22] analyzed the resequencing problem of a Selective-Repeat-ARQ with parallel channels, using a discrete-time model. Varma [23], Ayoun and Rosberg [24] considered optimal control problems in a queue with two servers of different service rates. The question is how to assign the customers to the servers so as to minimize the end-to-end delay [23] or the long-run average holding costs of the customers [24]. Packets get disordered at the server-assignment stage and are required to be resequenced after leaving the two-server queue.

In the remaining part of the paper, we will prove Theorem 1. The basic argument of the proof is as follows. Suppose the oldest customer in the DN is C_* and is being serviced at queue 1 in the DN. We wish to find out the probability that the RSQ has at least n customers. The customers in the RSQ must have all arrived at the DN after C_*, and all gone through queue 2 in the DN during the time C_* spent in queue 1, which is (roughly) an exponential random variable, independent of the queue 2 process. Therefore, the probability that the RSQ has at least n customers is the same as the probability that at least n customers arrive at queue 2, an M/M/1 queue, and at least n of those customers depart the queue during an exponential random time T that is independent of the queue 2 process. There is also the symmetric case where the oldest customer is in queue 2 and all customers in the RSQ come from queue 1. In Section III, we set up the two different cases and write the quantities to be computed. In Section IV, we compute the key quantity, $P\{M(T) \geq n\}$, where the function $M(t)$ is the number of those customers who arrived at the M/M/1 queue on the interval $[0, t]$ and who departed by time t, and T is an exponential random variable independent of the M/M/1 queue. In Section V, we combine results of the previous two sections and give the proof for Theroem 1.

III. The Setup

At time t, let $V(t)$ be the event {the DN is empty at time t}. If $\bar{V}(t)$, let $C_*(t)$ be the oldest customer in the DN, let $W_*(t)$ be the time $C_*(t)$ has spent in the DN, and let $I_*(t)$ be the queue in the DN which $C_*(t)$ goes through. For $n \geq 0$, let

$$\begin{aligned} E(t,s,n) = \{&\text{at least } n \text{ customers arrived at the DN} \\ &\text{on the interval } (t-s,t]\text{, out of which} \\ &\text{at least } n \text{ have left the DN by } t\}. \end{aligned}$$

Let the size of the resequencing queue (RSQ) at time t be $q^r(t)$, and let $q_i(t)$ be the size of queue i at time t, where $i = 1$ or 2. Then, for $n > 0$,

$$P\{q^r(t) \geq n\} = P\{\bar{V}(t) \text{ and } E(t, W_*(t), n)\}. \tag{2}$$

Next, we will explain equality (2). When the RSQ size is greater than or equal to n, where $n > 0$, it must be waiting for some customer still in the DN. In particular, the next packet gap the RSQ is trying to fill is $C_*(t)$. The customers in the RSQ are exactly those who arrived at the DN later than $C_*(t)$, but who have left the DN by time t. We are interested in computing $\lim_{t\to\infty} P\{q^r(t) \geq n\}$. Alternatively, let us assume all relevant processes are stationary.

Let us extend the definition of $W_*(t)$, $W_*(t) = 0$ if $V(t)$. Then, when $n = 0$,

$$P\{q^r(t) \geq n\} = 1.$$

$$\begin{aligned} &P\{\bar{V}(t) \text{ and } E(t, W_*(t), n)\} \\ &= P\{E(t, W_*(t), n)|\bar{V}(t)\}P\{\bar{V}(t)\} = P\{\bar{V}(t)\}. \end{aligned}$$

$$\begin{aligned} &P\{V(t) \text{ and } E(t, W_*(t), n)\} \\ &= P\{E(t, W_*(t), n)|V(t)\}P\{V(t)\} = P\{V(t)\}. \end{aligned}$$

Hence, for $n = 0$,

$$P\{q^r(t) \geq n\} = P\{E(t, W_*(t), n)\}. \tag{3}$$

For $n > 0$, (3) is still true because

$$\begin{aligned} &P\{V(t) \text{ and } E(t, W_*(t), n)\} \\ &= P\{E(t, 0, n)|V(t)\}P\{V(t)\} = 0. \end{aligned}$$

Note that, because customers are served on first-come-first-serve basis in each of the queues, the oldest customers in the non-empty DN must be in service at one of the queues. If queue i is not empty, $i \in \{0,1\}$, let $W_i(t)$ be the duration for which the customer in service at queue i has stayed in the queue. If queue i is empty, let $W_i(t) = 0$. By using a simple reversibility argument, $W_i(t)$ has the same distribution as the waiting time in queue i (not including the service time) by an arbitrary customer. This distribution and the density are (page 213 in [25]), for $x \geq 0$,

$$F_{W_i}(x) = P\{W_i(t) \leq x\} = 1 - \rho_i e^{-(\mu_i - \lambda_i)x}, \tag{4}$$

$$f_{W_i}(x) = (1-\rho_i)\delta(x) + \lambda_i(1-\rho_i)e^{-(\mu_i-\lambda_i)x}, \tag{5}$$

where $\rho_i = \lambda_i/\mu_i$, and $\delta(x)$ is the Dirac delta function, representing the point probability mass at $x = 0$. We will occasionally omit the dependency on t for brevity.

Let $\hat{M}_i(t,s)$ be the number of those customers who arrived at queue i on the interval $(t-s,t]$ and who departed by time t. Note that for $n > 0$,

$$\begin{aligned} &P\{\hat{M}_1(t, W_*(t)) \geq n \mid W_1(t) = W_2(t) = 0\} \\ &= P\{\hat{M}_1(t, 0) \geq n \mid W_1(t) = W_2(t) = 0\} = 0. \end{aligned}$$

Also,

$$P\{W_1(t) = W_2(t) \neq 0\} = 0.$$

Therefore,

$$\begin{aligned} &P\{\hat{M}_1(t, W_*(t)) \geq n \mid W_1(t) = W_2(t)\} \\ &\cdot P\{W_1(t) = W_2(t)\} = 0. \end{aligned}$$

Then, for $n > 0$,

$$\begin{aligned} &P\{q^r(t) \geq n\} \\ &= P\{E(t, W_*(t), n)\} \\ &= P\{\hat{M}_2(t, W_*(t)) \geq n \mid W_1(t) > W_2(t)\} \\ &\quad \cdot P\{W_1(t) > W_2(t)\} \\ &\quad + P\{\hat{M}_1(t, W_*(t)) \geq n \mid W_2(t) > W_1(t)\} \\ &\quad \cdot P\{W_2(t) > W_1(t)\}. \end{aligned} \tag{6}$$

This can be explained as follows. If $W_1(t) > W_2(t)$, then the oldest customer, $C_*(t)$, in the DN must be in service at queue 1. Hence, $W_1(t) = W_*(t)$. All customers who came to the DN after $C_*(t)$ and who have left the DN by time t must have been routed to the RSQ via queue 2.

For $n > 0$,

$$\begin{aligned} &P\{\hat{M}_2(t, W_*(t)) \geq n \mid W_1(t) > W_2(t)\} \\ &= \int_{0^+}^{\infty} P\{\hat{M}_2(t, s) \geq n \mid W_1(t) = s, W_1(t) > W_2(t)\} \\ &\quad \cdot f_{W_1|W_1>W_2}(s)ds \\ &= \int_{0^+}^{\infty} P\{\hat{M}_2(t, s) \geq n \mid W_1(t) = s, W_2(t) < s\} \\ &\quad \cdot f_{W_1|W_1>W_2}(s)ds \\ &= \int_{0^+}^{\infty} P\{\hat{M}_2(t, s) \geq n \mid W_2(t) < s\} \\ &\quad \cdot f_{W_1|W_1>W_2}(s)ds. \end{aligned} \tag{7}$$

In the above, $f_{W_1|W_1>W_2}(s)$ denotes the conditional density of $W_1(t)$ given $\{W_1(t) > W_2(t)\}$. In the last step, we used the fact that the two queue processes are independent. Note that, in the integral, the (conditional) probability mass at $s = 0$ does not contribute to the probability on the left hand side.

We will compute the conditional density by starting with the joint probability. For $x \geq 0$,

$$\begin{aligned} &P\{W_1 > x, W_1 > W_2\} \\ &= \rho_1 e^{-(\mu_1-\lambda_1)x} - \rho_1\rho_2 \frac{\mu_1 - \lambda_1}{\mu_1 - \lambda_1 + \mu_2 - \lambda_2} \\ &\quad \cdot e^{-(\mu_1-\lambda_1+\mu_2-\lambda_2)x}. \end{aligned} \tag{8}$$

From (8), we have

$$\begin{aligned} P\{W_1 > W_2\} &= P\{W_1 > 0, W_1 > W_2\} \\ &= \rho_1 - \rho_1\rho_2 \frac{\mu_1 - \lambda_1}{\mu_1 - \lambda_1 + \mu_2 - \lambda_2}. \end{aligned} \tag{9}$$

From (8) and (9), we get the conditional density for $x \geq 0$,

$$f_{W_1|W_1>W_2}(x) = K_1 e^{-(\mu_1-\lambda_1)x} - K_2 e^{-(\mu_1-\lambda_1+\mu_2-\lambda_2)x}, \tag{10}$$

where K_1 and K_2 are constants, given by,

$$K_1 = \frac{\mu_1-\lambda_1}{1-\rho_2\frac{\mu_1-\lambda_1}{\mu_1-\lambda_1+\mu_2-\lambda_2}}, \tag{11}$$

$$K_2 = \frac{\rho_2(\mu_1-\lambda_1)}{1-\rho_2\frac{\mu_1-\lambda_1}{\mu_1-\lambda_1+\mu_2-\lambda_2}}. \tag{12}$$

Note that the second term in (10) decays much faster than the first term. If we ignore it, the conditional probability density decays exponentially.

Next, we will bound (7) from above and below.

$$\begin{aligned}
&\int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n \mid W_2(t) < s\} f_{W_1|W_1>W_2}(s)ds \\
&= \int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n, W_2(t) < s\} \frac{f_{W_1|W_1>W_2}(s)}{P\{W_2(t) < s\}} ds \\
&\leq \int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n\} \frac{f_{W_1|W_1>W_2}(s)}{P\{W_2(t) = 0\}} ds \\
&\leq \frac{1}{1-\rho_2} \int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n\} f_{W_1|W_1>W_2}(s)ds. \qquad (13)
\end{aligned}$$

For a lower bound,

$$\begin{aligned}
&\int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n \mid W_2(t) < s\} f_{W_1|W_1>W_2}(s)ds \\
&= \int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n, W_2(t) < s\} \frac{f_{W_1|W_1>W_2}(s)}{P\{W_2(t) < s\}} ds \\
&\geq \int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n, W_2(t) = 0\} f_{W_1|W_1>W_2}(s)ds \\
&= \int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n, q_2(t) = 0\} f_{W_1|W_1>W_2}(s)ds. \qquad (14)
\end{aligned}$$

In the next section, we will prepare to compute the upper and lower bound.

IV. Computation of $P\{M(T) \geq n\}$

In this section, we consider a stationary M/M/1 queue whose arrival rate is λ_1 and whose departure rate is μ_1. We assume $\lambda_1 < \mu_1$ so that the queue is stable. Let T be an exponential random variable independent of the queue process with mean $1/(\mu_2-\lambda_2)$, where $\lambda_2 < \mu_2$. Let $M(t)$ be the number of those customers who arrived on the interval $[0,t]$ and who departed by time t. We wish to compute $P\{M(T) \geq n\}$ for large n. The main result of this section is Theorem 2. A similar result is Lemma 4.

Theorem 2:

$$\lim_{n\to\infty} \frac{1}{n} \log P\{M(T) \geq n\} = \begin{cases} \log \frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2} & \text{if } \mu_1-\lambda_1 \geq \mu_2-\lambda_2 \\ \log \frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2} & \text{if } \mu_1-\lambda_1 < \mu_2-\lambda_2 \end{cases}. \tag{15}$$

In the next two subsections, we will prove Theorem 2. We will frequently use the following fact. For $a > 0$ and integer $k \geq 0$,

Fact 3:

$$\int_0^{\infty} \frac{e^{-at}t^k}{k!} dt = (\frac{1}{a})^{k+1}. \tag{16}$$

A. Case of $\mu_1-\lambda_1 \geq \mu_2-\lambda_2$

1) The Upper Bound:

$$\begin{aligned}
&P\{M(T) \geq n\} \\
&\leq P\{\text{the number of customer arrivals on the interval } [0,T] \text{ is at least } n\} \qquad (17) \\
&= \sum_{k=n}^{\infty} \int_0^{\infty} \frac{e^{-\lambda_1 t}(\lambda_1 t)^k}{k!} (\mu_2-\lambda_2) e^{-(\mu_2-\lambda_2)t} dt \\
&= \sum_{k=n}^{\infty} (\mu_2-\lambda_2) \int_0^{\infty} \frac{e^{-(\lambda_1+\mu_2-\lambda_2)t}(\lambda_1 t)^k}{k!} dt \\
&= \sum_{k=n}^{\infty} \frac{\mu_2-\lambda_2}{\lambda_1+\mu_2-\lambda_2} (\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2})^k \quad \text{(by (16))} \\
&= \frac{\mu_2-\lambda_2}{\lambda_1+\mu_2-\lambda_2} \frac{(\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2})^n}{1-\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2}} \\
&= (\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2})^n.
\end{aligned}$$

2) The Lower Bound: For a function that grows as $\exp(\alpha n + o(n))$ when n increases, α is the rate of growth. The method that estimates the rate of growth of an integral by that of the maximum of the integrand is known as the Laplace principle (See page 12 of [26].). In our case, we will consider the following integral, as n gets large,

$$\int_0^{\infty} \frac{e^{-\lambda_1 t}(\lambda_1 t)^n}{n!} (\mu_2-\lambda_2) e^{-(\mu_2-\lambda_2)t} dt.$$

It can be shown easily that the integrand is maximized at,

$$t_o = n/(\lambda_1+\mu_2-\lambda_2). \tag{18}$$

This information will be useful in the proof for the lower bound.

Let $q(t)$ be the queue size at time t. Let $D(t)$ be the number of departures on the interval $[0,t]$.

$$\begin{aligned}
&P\{M(t) = k\} \\
&= \sum_{m=0}^{\infty} P\{M(t) = k | q(0) = m\} P\{q(0) = m\} \\
&\geq P\{M(t) = k | q(0) = 0\} P\{q(0) = 0\} \\
&= (1-\rho_1) P\{D(t) = k | q(0) = 0\} \\
&\geq (1-\rho_1) P\{D(t) = k, q(t) = 0 | q(0) = 0\}. \qquad (19)
\end{aligned}$$

From [27] (page 199),

$$
\begin{aligned}
&P\{D(t)=k, q(t)=0|q(0)=0\}\\
&=\sum_{i=0}^{\infty}\frac{(1+i)\rho_1^k}{k!(k+i+1)!}(\mu_1 t)^{2k+i}e^{-(\lambda_1+\mu_1)t}\\
&=\frac{(\lambda_1 t)^k e^{-\lambda_1 t}}{k!}\sum_{i=0}^{\infty}\frac{1+i}{(k+i+1)!}(\mu_1 t)^{k+i}e^{-\mu_1 t}\\
&\geq \frac{1}{k+1}\frac{(\lambda_1 t)^k e^{-\lambda_1 t}}{k!}\sum_{i=0}^{\infty}\frac{1}{(k+i)!}(\mu_1 t)^{k+i}e^{-\mu_1 t}\\
&=\frac{1}{k+1}\frac{(\lambda_1 t)^k e^{-\lambda_1 t}}{k!}P\{Y_{(\mu_1 t)}\geq k\}, \qquad (20)
\end{aligned}
$$

where $Y_{(\mu_1 t)}$ is a Poisson random variable with mean $\mu_1 t$. Now, with the definition of t_o as in (18),

$$
\begin{aligned}
&P\{M(T)\geq n\}\\
&\geq P\{M(T)\geq n, T\geq t_o\}\\
&\geq P\{M(t_o)\geq n, T\geq t_o\}\\
&= P\{M(t_o)\geq n\}P\{T\geq t_o\} \qquad (21)\\
&= \sum_{k=n}^{\infty}P\{M(t_o)=k\}P\{T\geq t_o\}. \qquad (22)
\end{aligned}
$$

The equality in (21) is because of independence between the queue process and the random variable T. Then, by (22), (19) and (20),

$$
\begin{aligned}
&P\{M(T)\geq n\}\\
&\geq (1-\rho_1)\sum_{k=n}^{\infty}\frac{1}{k+1}\frac{(\lambda_1 t_o)^k e^{-\lambda_1 t_o}}{k!}\\
&\quad\cdot P\{Y_{(\mu_1 t_o)}\geq k\}e^{-(\mu_2-\lambda_2)t_o}\\
&\geq (1-\rho_1)\frac{1}{n+1}\frac{(\lambda_1 t_o)^n e^{-\lambda_1 t_o}}{n!}\\
&\quad\cdot P\{Y_{(\mu_1 t_o)}\geq n\}e^{-(\mu_2-\lambda_2)t_o}. \qquad (23)
\end{aligned}
$$

We will show $P\{Y_{(\mu_1 t_o)}\geq n\}$ is greater than a constant as n tends to infinity. By the definition of t_o and by the assumption $\mu_1-\lambda_1\geq\mu_2-\lambda_2$,

$$
\mu_1 t_o=\frac{\mu_1}{\lambda_1+\mu_2-\lambda_2}n\geq n.
$$

Let $n_o=\lfloor\mu_1 t_o\rfloor$. Then, $n_o\geq n$. Let $X_1, X_2, ..., X_{n_o}$ be IID. Poisson random variables with mean 1. Then,

$$
\begin{aligned}
P\{Y_{(\mu_1 t_o)}\geq n\}&\geq P\{\frac{X_1+X_2+...+X_{n_o}}{n_o}\geq\frac{n}{n_o}\}\\
&\geq P\{\frac{X_1+X_2+...+X_{n_o}}{n_o}\geq 1\}\\
&= P\{\frac{X_1+X_2+...+X_{n_o}-n_o}{\sqrt{n_o}\sqrt{n_o}}\geq 0\}\\
&= P\{\frac{X_1+X_2+...+X_{n_o}-n_o}{\sqrt{n_o}}\geq 0\}.
\end{aligned}
$$

By the central limit theorem,

$$
\begin{aligned}
&\lim_{n_o\to\infty}P\{\frac{X_1+X_2+...+X_{n_o}-n_o}{\sqrt{n_o}}\geq 0\}\\
&=\int_0^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\frac{1}{2}.
\end{aligned}
$$

Therefore, for any $\epsilon>0$, there exists some integer $N>0$ such that for all $n>N$,

$$
P\{Y_{(\mu_1 t_o)}\geq n\}\geq\frac{1}{2}-\epsilon. \qquad (24)
$$

Continuing from (23), for all $n>N$,

$$
\begin{aligned}
&P\{M(T)\geq n\}\\
&\geq(1-\rho_1)(\frac{1}{2}-\epsilon)\frac{1}{n+1}\frac{(\lambda_1 t_o)^n e^{-\lambda_1 t_o}}{n!}e^{-(\mu_2-\lambda_2)t_o}. \qquad (25)
\end{aligned}
$$

By Stirling's approximation,

$$
n!=\sqrt{2\pi n}n^n e^{-n}(1+O(1/n)).
$$

For sufficiently large n,

$$
n!\leq 2\sqrt{2\pi n}n^n e^{-n}.
$$

Therefore, for large enough n, using the definition for t_o in (18), we have

$$
\begin{aligned}
&P\{M(T)\geq n\}\\
&\geq\frac{1}{4}(1-\rho_1)(1-2\epsilon)\frac{1}{n+1}(\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2})^n\\
&\quad\cdot\frac{n^n\exp(-\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2}n)}{\sqrt{2\pi n}n^n e^{-n}}\exp(-\frac{\mu_2-\lambda_2}{\lambda_1+\mu_2-\lambda_2}n)\\
&=\frac{(1-\rho_1)(1-2\epsilon)}{4\sqrt{2\pi n}(n+1)}(\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2})^n. \qquad (26)
\end{aligned}
$$

B. Case of $\mu_1-\lambda_1<\mu_2-\lambda_2$

1) The Lower Bound: By (19) and (20),

$$
\begin{aligned}
&P\{M(T)\geq n\}\\
&\geq\int_0^{\infty}\sum_{k=n}^{\infty}(1-\rho_1)P\{D(t)=k, q(t)=0|q(0)=0\}\\
&\quad\cdot(\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt \qquad (27)\\
&\geq(1-\rho_1)(\mu_2-\lambda_2)\frac{1}{n+1}\\
&\quad\cdot\int_0^{\infty}\frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!}P\{Y_{(\mu_1 t)}=n\}e^{-(\mu_2-\lambda_2)t}dt\\
&=(1-\rho_1)(\mu_2-\lambda_2)\frac{1}{n+1}\\
&\quad\cdot\int_0^{\infty}\frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!}\frac{(\mu_1 t)^n e^{-\mu_1 t}}{n!}e^{-(\mu_2-\lambda_2)t}dt\\
&=(1-\rho_1)(\mu_2-\lambda_2)\frac{1}{n+1}\frac{(2n)!}{n!n!}(\lambda_1\mu_1)^n\\
&\quad\cdot\int_0^{\infty}\frac{t^{2n}e^{-(\lambda_1+\mu_1+\mu_2-\lambda_2)t}}{(2n)!}dt\\
&=(1-\rho_1)\frac{\mu_2-\lambda_2}{\lambda_1+\mu_1+\mu_2-\lambda_2}\frac{1}{n+1}\frac{(2n)!}{n!n!}\\
&\quad\cdot(\lambda_1\mu_1)^n\frac{1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^{2n}}.
\end{aligned}
$$

In deriving the last step, (16) has been used. By Stirling's approximation, for large enough n,

$$
\frac{(2n)!}{n!n!}=\frac{\sqrt{4\pi n}(2n)^{2n}e^{-2n}(1+O(1/n))}{(\sqrt{2\pi n}(n)^n e^{-n}(1+O(1/n)))^2}\geq\frac{C_1}{\sqrt{n}}4^n,
$$

for some constant $C_1 > 0$. Therefore,

$$\begin{aligned} & P\{M(T) \ge n\} \\ & \ge C_1(1-\rho_1)\frac{\mu_2-\lambda_2}{\lambda_1+\mu_1+\mu_2-\lambda_2} \\ & \cdot \frac{1}{\sqrt{n}(n+1)}(\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2})^n. \end{aligned} \tag{28}$$

2) The Upper Bound: The computation for the upper bound in the previous case does not apply here. To see the reason, consider the integral in the lower bound calculation. Suppose, as n becomes large,

$$\begin{aligned} & \int_0^\infty \frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!}\frac{(\mu_1 t)^n e^{-\mu_1 t}}{n!}e^{-(\mu_2-\lambda_2)t}dt \\ \approx & \max_{t\ge 0}\frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!}\frac{(\mu_1 t)^n e^{-\mu_1 t}}{n!}e^{-(\mu_2-\lambda_2)t}. \end{aligned}$$

It can be shown easily the above maximum is achieved at

$$t_o = \frac{2n}{\lambda_1+\mu_1+\mu_2-\lambda_2}. \tag{29}$$

Note that when $\mu_1-\lambda_1 < \mu_2-\lambda_2$,

$$\mu_1 t_0 = \frac{2\mu_1 n}{\lambda_1+\mu_1+\mu_2-\lambda_2} < n.$$

Therefore, $\{Y_{(\mu_1 t_o)} \ge n\}$ is a large deviations type of event instead of an event with constant probability, as n becomes large. It is not tight enough to bound $P\{M(t) \ge n\}$ from above by only looking at the arrival processes, as was done in (17).

We will next carry out the analysis on the upper bound.

$$\begin{aligned} & P\{M(t) \ge n\} \\ & \le P\{\text{at least } n \text{ customers arrived on the interval } [0,t], \text{ and} \\ & \quad \text{at least } n \text{ customers are served on the same interval}\} \\ & \le \sum_{k=n}^\infty \frac{e^{-\lambda_1 t}(\lambda_1 t)^k}{k!}P\{\sum_{i=1}^n X_i \le t\}, \end{aligned} \tag{30}$$

where $\{X_1, X_2, ..., X_n\}$ are IID service times. The sum $\sum_{i=1}^n X_i$ has the Gamma distribution with density,

$$f(t) = \frac{\mu_1 e^{-\mu_1 t}(\mu_1 t)^{n-1}}{(n-1)!}.$$

Hence,

$$\begin{aligned} & P\{M(T) \ge n\} \\ & \le \sum_{k=n}^\infty \int_0^\infty \frac{e^{-\lambda_1 t}(\lambda_1 t)^k}{k!} \\ & \int_0^t \frac{\mu_1 e^{-\mu_1\tau}(\mu_1\tau)^{n-1}}{(n-1)!}d\tau(\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt \\ & = \mu_1(\mu_2-\lambda_2)\sum_{k=n}^\infty\int_0^\infty\int_\tau^\infty \frac{e^{-(\lambda_1+\mu_2-\lambda_2)t}(\lambda_1 t)^k}{k!}dt \\ & \cdot \frac{e^{-\mu_1\tau}(\mu_1\tau)^{n-1}}{(n-1)!}d\tau. \end{aligned}$$

Let $t = \tau + u$. The above becomes,

$$\begin{aligned} & P\{M(T) \ge n\} \\ & \le \mu_1(\mu_2-\lambda_2)\sum_{k=n}^\infty\int_0^\infty\int_0^\infty \\ & \frac{e^{-(\lambda_1+\mu_2-\lambda_2)(\tau+u)}(\lambda_1(\tau+u))^k}{k!}du\frac{e^{\mu_1\tau}(\mu_1\tau)^{n-1}}{(n-1)!}d\tau \\ & = \mu_1(\mu_2-\lambda_2)\sum_{k=n}^\infty\lambda_1^k\int_0^\infty\int_0^\infty \\ & \frac{e^{-(\lambda_1+\mu_2-\lambda_2)u}\sum_{i=0}^k\frac{k!}{i!(k-i)!}u^i\tau^{k-i}}{k!}du \\ & \frac{e^{-(\lambda_1+\mu_1+\mu_2-\lambda_2)\tau}(\mu_1\tau)^{n-1}}{(n-1)!}d\tau \\ & = \mu_1(\mu_2-\lambda_2)\sum_{k=n}^\infty\lambda_1^k\sum_{i=0}^k\frac{1}{(k-i)!} \\ & \int_0^\infty\int_0^\infty\frac{e^{-(\lambda_1+\mu_2-\lambda_2)u}u^i}{i!}du \\ & \frac{e^{-(\lambda_1+\mu_1+\mu_2-\lambda_2)\tau}\tau^{k-i}(\mu_1\tau)^{n-1}}{(n-1)!}d\tau \\ & = (\mu_2-\lambda_2)\mu_1^n\sum_{k=n}^\infty\lambda_1^k \\ & \sum_{i=0}^k\frac{1}{(k-i)!}\frac{1}{(\lambda_1+\mu_2-\lambda_2)^{i+1}} \\ & \int_0^\infty\frac{e^{-(\lambda_1+\mu_1+\mu_2-\lambda_2)\tau}\tau^{n-1+k-i}}{(n-1)!}d\tau \\ & = (\mu_2-\lambda_2)\mu_1^n\sum_{k=n}^\infty\lambda_1^k\sum_{i=0}^k\frac{(n-1+k-i)!}{(k-i)!(n-1)!} \\ & \frac{1}{(\lambda_1+\mu_2-\lambda_2)^{i+1}}\frac{1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^{n+k-i}} \\ & = \frac{\mu_2-\lambda_2}{\lambda_1+\mu_2-\lambda_2}(\frac{\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^n \\ & \sum_{k=n}^\infty(\frac{\lambda_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^k \\ & \cdot\sum_{i=0}^k\frac{(n-1+k-i)!}{(k-i)!(n-1)!}(\frac{\lambda_1+\mu_1+\mu_2-\lambda_2}{\lambda_1+\mu_2-\lambda_2})^i. \end{aligned} \tag{31}$$

For $i = 0, 1, ..., k$, define

$$a(k,i) = \frac{(n-1+k-i)!}{2^k(k-i)!(n-1)!}.$$

Let

$$\beta = \frac{\lambda_1+\mu_1+\mu_2-\lambda_2}{\lambda_1+\mu_2-\lambda_2}.$$

Note that for $\mu_1-\lambda_1 < \mu_2-\lambda_2$, $\beta < 2$.

$$\frac{a(k+1,i)}{a(k,i)} = \frac{n+k-i}{2(k+1-i)} = \frac{1+\frac{n-1}{k+1-i}}{2}.$$

Then, for each fixed $i \in \{0, 1, ..., k\}$,

$$\frac{a(k+1,i)}{a(k,i)}\begin{cases} \ge 1 & \text{if } k \le n+i-2 \\ < 1 & \text{if } k > n+i-2 \end{cases}.$$

Therefore, $a(k,i)$ is maximized at $k = n+i-1$ for each i [1]. Then,

$$a(n+i-1,i) = \frac{(2n-2)!}{(n-1)!(n-1)!2^{n+i-1}}.$$

Then, the sum in (31) index by i becomes,

$$\begin{aligned}
&\sum_{i=0}^{k} \frac{(n-1+k-i)!}{(k-i)!(n-1)!}(\frac{\lambda_1+\mu_1+\mu_2-\lambda_2}{\lambda_1+\mu_2-\lambda_2})^i \\
&= 2^k \sum_{i=0}^{k} a(k,i)\beta^i \\
&\le 2^{k-n} \sum_{i=0}^{k} \frac{2(2n-2)!}{(n-1)!(n-1)!}(\frac{\beta}{2})^i \\
&\le 2^{k-n} \frac{2(2n-2)!}{(n-1)!(n-1)!} \sum_{i=0}^{\infty}(\frac{\beta}{2})^i \\
&= 2^{k-n} \frac{2(2n-2)!}{(n-1)!(n-1)!} \frac{2}{2-\beta}.
\end{aligned}$$

The infinite sum above is finite because $\beta/2 < 1$. Going back to (31), we get,

$$\begin{aligned}
&P\{M(T) \ge n\} \\
&\le \frac{4(\mu_2-\lambda_2)}{(2-\beta)(\lambda_1+\mu_2-\lambda_2)} \frac{(2n-2)!}{2^n(n-1)!(n-1)!} \\
&(\frac{\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^n \sum_{k=n}^{\infty}(\frac{2\lambda_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^k \qquad (32) \\
&= \frac{4(\mu_2-\lambda_2)}{\mu_2-\lambda_2-(\mu_1-\lambda_1)} \frac{(2n-2)!}{2^n(n-1)!(n-1)!} \\
&(\frac{\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^n \\
&\cdot (\frac{2\lambda_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^n/(1-\frac{2\lambda_1}{\lambda_1+\mu_1+\mu_2-\lambda_2}). \qquad (33)
\end{aligned}$$

The sum in (32) is finite because, for $\mu_1-\lambda_1 < \mu_2-\lambda_2$ and $\lambda_1 < \mu_1$,

$$\frac{2\lambda_1}{\lambda_1+\mu_1+\mu_2-\lambda_2} < 1. \qquad (34)$$

By Stirling's approximation,

$$\frac{(2n)!}{n!n!} = \frac{\sqrt{4\pi n}(2n)^{2n}e^{-2n}(1+O(1/n))}{(\sqrt{2\pi n}(n)^n e^{-n}(1+O(1/n)))^2} \le \frac{C_2}{\sqrt{n}}4^n, \qquad (35)$$

Combining (33) and (35), we have, for some constant $C_4 > 0$,

$$P\{M(T) \ge n\} \le \frac{C_4}{\sqrt{n}}(\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2})^n.$$

This completes the analysis on the upper bound.

C. A Related Lemma to Theorem 2

The following lemma is also used in the proof of Theorem 1. Its proof is directly related to that of Theorem 2. The notations will be the same as those used in the proof of the lower bound in Theorem 1.

Lemma 4:

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log P\{M(T) \ge n, q(T) = 0\} \\
&= \begin{cases} \log \frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2} & \text{if } \mu_1-\lambda_1 \ge \mu_2-\lambda_2 \\ \log \frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2} & \text{if } \mu_1-\lambda_1 < \mu_2-\lambda_2 \end{cases}. \qquad (36)
\end{aligned}$$

Proof: Since

$$P\{M(T) \ge n, q(T) = 0\} \le P\{M(T) \ge n\},$$

The upper bound is immediate from Theorem 2. We only need to show the left hand side of (36) is no less than the right hand side.

Let us first consider the case where $\mu_1-\lambda_1 \ge \mu_2-\lambda_2$. Conditional on T, which is independent of the queue process, we have

$$\begin{aligned}
&P\{M(T) \ge n, q(T) = 0\} \\
&= \int_0^{\infty} P\{M(t) \ge n, q(t) = 0\}(\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt \\
&= \sum_{k=n}^{\infty} \int_0^{\infty} P\{M(t) = k, q(t) = 0\}(\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt \\
&\ge \sum_{k=n}^{\infty} \int_0^{\infty} P\{M(t) = k, q(t) = 0 | q(0) = 0\}P\{q(0) = 0\} \\
&\quad \cdot (\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt \\
&= \sum_{k=n}^{\infty} \int_0^{\infty} (1-\rho_1)P\{D(t) = k, q(t) = 0 | q(0) = 0\} \\
&\quad \cdot (\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt \qquad (37) \\
&\ge \sum_{k=n}^{\infty} \int_0^{\infty} (1-\rho_1)\frac{1}{k+1}\frac{(\lambda_1 t)^k e^{-\lambda_1 t}}{k!}P\{Y_{(\mu_1 t)} \ge k\} \\
&\quad \cdot (\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt \qquad (38) \\
&\ge \sum_{k=n}^{\infty} \int_{t_o}^{\infty} (1-\rho_1)\frac{1}{k+1}\frac{(\lambda_1 t)^k e^{-\lambda_1 t}}{k!}P\{Y_{(\mu_1 t)} \ge k\} \\
&\quad \cdot (\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt \\
&\ge \int_{t_o}^{\infty} (1-\rho_1)\frac{1}{n+1}\frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!}P\{Y_{(\mu_1 t)} \ge n\} \\
&\quad \cdot (\mu_2-\lambda_2)e^{-(\mu_2-\lambda_2)t}dt. \qquad (39)
\end{aligned}$$

To obtain (38), we have used (20). In the last two steps above, t_o is as given in (18). Note that, for all $t \ge t_o$,

$$P\{Y_{(\mu_1 t)} \ge n\} \ge P\{Y_{(\mu_1 t_o)} \ge n\}. \qquad (40)$$

By (40) and (24), we obtain that, for any $\epsilon > 0$, there exists some integer $N > 0$ such that for all $n \ge N$ and for all $t \ge t_o$,

$$P\{Y_{(\mu_1 t)} \ge n\} \ge \frac{1}{2} - \epsilon.$$

Continuing from (39), we get

$$\begin{aligned}
&P\{M(T) \ge n, q(T) = 0\} \\
&\ge (1-\rho_1)(\frac{1}{2}-\epsilon)\frac{(\mu_2-\lambda_2)}{n+1}\int_{t_o}^{\infty}\frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!}e^{-(\mu_2-\lambda_2)t}dt. \\
&\qquad (41)
\end{aligned}$$

[1] We assume that n is large enough when necessary. In this case, $n \ge 1$.

As noted in (18), the above integrand achieves the maximum value at t_o. Furthermore, it is easy to show that, for $t \geq t_o$, the integrand is a decreasing function of t. We must have,

$$
\begin{aligned}
&\int_{t_o}^{\infty} \frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!} e^{-(\mu_2-\lambda_2)t} dt \\
\geq &\int_{t_o}^{t_o+1} \frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!} e^{-(\mu_2-\lambda_2)t} dt \\
\geq &\frac{(\lambda_1 (t_o+1))^n e^{-\lambda_1 (t_o+1)}}{n!} e^{-(\mu_2-\lambda_2)(t_o+1)} \\
\geq &\frac{(\lambda_1 t_o)^n e^{-\lambda_1 t_o}}{n!} e^{-(\mu_2-\lambda_2)t_o} e^{-(\lambda_1+\mu_2-\lambda_2)}. \qquad (42)
\end{aligned}
$$

Combining (41) and (42), for sufficiently large n, we get,

$$
\begin{aligned}
&P\{M(T) \geq n, q(T) = 0\} \\
&\geq (1-\rho_1)(\frac{1}{2}-\epsilon)\frac{(\mu_2-\lambda_2)}{n+1} e^{-(\lambda_1+\mu_2-\lambda_2)} \\
&\cdot \frac{(\lambda_1 t_o)^n e^{-\lambda_1 t_o}}{n!} e^{-(\mu_2-\lambda_2)t_o}. \qquad (43)
\end{aligned}
$$

Following the steps in (25) to (26), for sufficiently large n, we get the lower bound

$$
\begin{aligned}
&P\{M(T) \geq n\} \\
\geq &\frac{(1-\rho_1)(1-2\epsilon)(\mu_2-\lambda_2)}{4\sqrt{2\pi n}(n+1)} e^{-(\lambda_1+\mu_2-\lambda_2)} (\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2})^n. \\
&\qquad (44)
\end{aligned}
$$

Next, let us consider the case where $\mu_1 - \lambda_1 < \mu_2 - \lambda_2$. From (37), we have

$$
\begin{aligned}
&P\{M(T) \geq n, q(T) = 0\} \\
\geq &\sum_{k=n}^{\infty} \int_0^{\infty} (1-\rho_1) P\{D(t) = k, q(t) = 0 | q(0) = 0\} \\
&\cdot (\mu_2-\lambda_2) e^{-(\mu_2-\lambda_2)t} dt.
\end{aligned}
$$

The rest of the steps are identical to the lower bound proof in Theorem 2 from (27) to (28). ■

V. Proof of Theorem 1

We will combine the results of the previous two sections and prove the main theorem. We wish to show that, without the loss of generality, when $\mu_1 - \lambda_1 \leq \mu_2 - \lambda_2$,

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log P\{q^r(t) \geq n\} \\
= \ &\max\{\log \frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1}, \log \frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2}\}. \\
&\qquad (45)
\end{aligned}
$$

Proof: We need to find the asymptotic exponent for the two terms in (6), as n approaches infinity. The factors $P\{W_1(t) > W_2(t)\}$ and $P\{W_2(t) > W_1(t)\}$ are constants on $(0,1)$, not dependent on n. For instance, $P\{W_1(t) > W_2(t)\}$ is given by (9). We will not carry these factors around in the subsequent analysis.

We will first consider the first term in (6). We wish to show

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log P\{\hat{M}_2(t, W_*(t)) \geq n \mid W_1(t) > W_2(t)\} \\
&= \begin{cases} \log \frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1} & \text{if } \mu_2-\lambda_2 \geq \mu_1-\lambda_1 \\ \log \frac{4\lambda_2\mu_2}{(\lambda_2+\mu_2+\mu_1-\lambda_1)^2} & \text{if } \mu_2-\lambda_2 < \mu_1-\lambda_1 \end{cases}. \qquad (46)
\end{aligned}
$$

For the lower bound, we combine (7), (14) and (10), and get

$$
\begin{aligned}
&P\{\hat{M}_2(t, W_*(t)) \geq n \mid W_1(t) > W_2(t)\} \\
= &\int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n \mid W_2(t) < s\} f_{W_1|W_1>W_2}(s) ds \\
\geq &\int_{0^+}^{\infty} P\{\hat{M}_2(t,s) \geq n, q_2(t) = 0\} f_{W_1|W_1>W_2}(s) ds \\
= &\int_{0^+}^{\infty} P\{M_2(s) \geq n, q_2(s) = 0\} f_{W_1|W_1>W_2}(s) ds \\
= &K_1 \int_0^{\infty} P\{M_2(s) \geq n, q_2(s) = 0\} e^{-(\mu_1-\lambda_1)s} ds - \\
&K_2 \int_0^{\infty} P\{M_2(s) \geq n, q_2(s) = 0\} e^{-(\mu_1-\lambda_1+\mu_2-\lambda_2)s} ds, \\
&\qquad (47)
\end{aligned}
$$

where $K_1 > 0$ and $K_2 > 0$ are constants given in (11) and (12). In the above, we have used the conditional density of W_1 given $\{W_1 > W_2\}$ from (10). By Lemma 4 with suitable substitution of variables, and since

$$
\mu_2 - \lambda_2 < \mu_1 - \lambda_1 + \mu_2 - \lambda_2,
$$

the second term in (47) has the following asymptotic exponent,

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log \int_0^{\infty} P\{M_2(s) \geq n, q_2(s) = 0\} \\
&\cdot e^{-(\mu_1-\lambda_1+\mu_2-\lambda_2)s} ds \\
&= \log \frac{4\lambda_2\mu_2}{(\lambda_2+\mu_2+\mu_1-\lambda_1+\mu_2-\lambda_2)^2} \\
&= \log \frac{4\lambda_2\mu_2}{(2\mu_2+\mu_1-\lambda_1)^2}. \qquad (48)
\end{aligned}
$$

By Lemma 4 with suitable substitution of variables, the first term in (47) has the following asymptotic exponent,

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log \int_0^{\infty} P\{M_2(s) \geq n, q_2(s) = 0\} e^{-(\mu_1-\lambda_1)s} ds \\
&= \begin{cases} \log \frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1} & \text{if } \mu_2-\lambda_2 \geq \mu_1-\lambda_1 \\ \log \frac{4\lambda_2\mu_2}{(\lambda_2+\mu_2+\mu_1-\lambda_1)^2} & \text{if } \mu_2-\lambda_2 < \mu_1-\lambda_1 \end{cases}. \qquad (49)
\end{aligned}
$$

Now,

$$
\frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1} = \frac{4\lambda_2\mu_2}{4\lambda_2\mu_2+4\mu_1\mu_2-4\lambda_1\mu_2},
$$

$$
\frac{4\lambda_2\mu_2}{(2\mu_2+\mu_1-\lambda_1)^2} = \frac{4\lambda_2\mu_2}{4\mu_2^2+4\mu_1\mu_2-4\lambda_1\mu_2+(\mu_1-\lambda_1)^2}.
$$

Hence, we have

$$
\frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1} > \frac{4\lambda_2\mu_2}{(2\mu_2+\mu_1-\lambda_1)^2}.
$$

Also, because $\lambda_2 < \mu_2$, we have

$$
\frac{4\lambda_2\mu_2}{(\lambda_2+\mu_2+\mu_1-\lambda_1)^2} > \frac{4\lambda_2\mu_2}{(2\mu_2+\mu_1-\lambda_1)^2}.
$$

Therefore, we can ignore the contribution from (48) when considering the lower bound of the left hand side in (46). Then, (49) gives the lower bound.

For the upper bound of the left hand side in (46), we combine (7) and (13), and get

$$\begin{aligned}&P\{\hat{M}_2(t, W_*(t)) \geq n \mid W_1(t) > W_2(t)\}\\ &\leq \frac{1}{1-\rho_2}\int_{0+}^{\infty} P\{\hat{M}_2(t,s) \geq n\} f_{W_1|W_1>W_2}(s)ds. \qquad (50)\end{aligned}$$

Using a similar argument as in the derivation of the lower bound, but with Theorem 2 substituting the role of Lemma 4, we get

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}\log\int_{0+}^{\infty} P\{\hat{M}_2(t,s) \geq n\} f_{W_1|W_1>W_2}(s)ds\\ &=\begin{cases}\log\frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1} & \text{if } \mu_2-\lambda_2 \geq \mu_1-\lambda_1\\ \log\frac{4\lambda_2\mu_2}{(\lambda_2+\mu_2+\mu_1-\lambda_1)^2} & \text{if } \mu_2-\lambda_2 < \mu_1-\lambda_1\end{cases}. \qquad (51)\end{aligned}$$

Since the upper and lower bound agree with each other, we get (46).

To determine $P\{q^r(t) \geq n\}$ for large n, we also need to consider the second term in (6), $P\{\hat{M}_1(t, W_*(t)) \geq n \mid W_2(t) > W_1(t)\}$. By symmetry,

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{M}_1(t, W_*(t)) \geq n \mid W_2(t) > W_1(t)\}\\ &=\begin{cases}\log\frac{\lambda_1}{\lambda_1+\mu_2-\lambda_2} & \text{if } \mu_1-\lambda_1 \geq \mu_2-\lambda_2\\ \log\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2} & \text{if } \mu_1-\lambda_1 < \mu_2-\lambda_2\end{cases}. \qquad (52)\end{aligned}$$

When $\mu_1-\lambda_1 \leq \mu_2-\lambda_2$, combining (6), (46) and (52), we get (45). ■

VI. Conclusion

To conclude, we discuss the implications of Theorem 1. When, $\lambda_1 = \lambda_2$ and $\mu_1 = \mu_2$,

$$\lim_{n\to\infty}\frac{1}{n}\log P\{q^r(t) \geq n\} = \log\rho_1.$$

When $\mu_1 - \lambda_1 = \mu_2 - \lambda_2$,

$$\lim_{n\to\infty}\frac{1}{n}\log P\{q^r(t) \geq n\} = \max\{\log\rho_1, \log\rho_2\}.$$

Like all GI/GI/1 queues, the resequencing queue size depends on the arrival and departure rates through a dimensionless parameter. This implies that the resequencing queue size does not change with the link speed of the network, if all links involved scale their bandwidth by the same factor and if the traffic characteristics are not altered by the technology change. This is in contrast with the models from our previous paper [2], where the improvement of network speed worsens the packet resequencing problem in terms of both the queue size and the delay. In the current model, there can be many ways to produce the large resequencing queue size, which, in general, depends on parameters for both queues in the DN. According to Theorem 1, for $\mu_1 - \lambda_1 \leq \mu_2 - \lambda_2$ and for large n,

$$\begin{aligned}&P\{q^r(t) \geq n\}\\ &\approx \max\{(\frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1})^n, (\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2})^n\}.\end{aligned}$$

In the first term above, $\frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1} \approx 1$ if $\mu_1 - \lambda_1 \ll \lambda_2$. In the second term above, $\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2} \approx 1$ if $\mu_2 - \lambda_2 \ll \lambda_1 + \mu_1$ and $\lambda_1 \approx \mu_1$. This implies that, for the second term to decay slowly, we need to have

$$\mu_1 - \lambda_1 \leq \mu_2 - \lambda_2 \ll 2\mu_1 \approx 2\lambda_1.$$

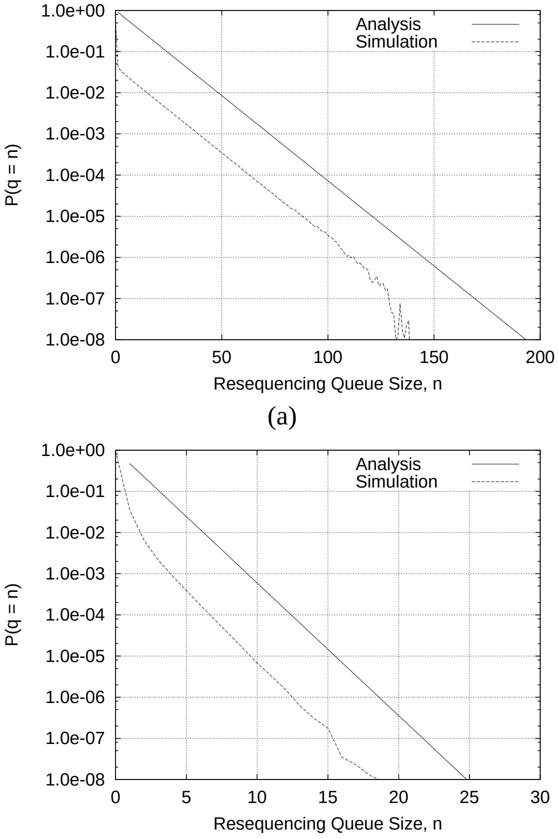
An interesting observation is that it can be large even when the queue sizes in the DN are both small. This occurs when the two disordering queues are "mismatched", that is, when one of the disordering queues is much faster than the other in terms of both the arrival rate and the service rate. For example, suppose $\mu_i = 2\lambda_i$ for $i = 1$ and 2. Hence, $\rho_1 = \rho_2 = 1/2$, and for $i = 1$ and 2,

$$\lim_{n\to\infty}\frac{1}{n}\log P\{q_i(t) \geq n\} = \log\frac{1}{2}.$$

Suppose, in the DN, queue 2 is ten times faster than queue 1, i.e., $\lambda_2 = 10\lambda_1$ and $\mu_2 = 10\mu_1$. Then,

$$\lim_{n\to\infty}\frac{1}{n}\log P\{q^r(t) \geq n\} = \log\frac{10}{11}.$$

Intuitively, the queue size of the RSQ can be large because many later packets can go through queue 2 in the DN and end up waiting in the RSQ, while some earlier packets are waiting in queue 1.



(a)

(b)

Fig. 2. $P\{q^r = n\}$: Simulation results. (a) $\lambda_1 = 1$, $\mu_1 = 2$, $\lambda_2 = 10$, $\mu_2 = 20$; (b) $\lambda_1 = 10$, $\mu_1 = 20$, $\lambda_2 = 1$, $\mu_2 = 12$

In Figure 2, we show the simulation results for $P\{q^r = n\}$ and compare them with the analytical results in Theorem 1.

In Figure 2 (a), the parameters are chosen so that

$$\frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1}=\frac{10}{11}=0.9091,$$

$$\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2}=\frac{8}{169}=0.0473.$$

In Figure 2 (b), the parameters are chosen so that

$$\frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1}=\frac{1}{21}=0.0476,$$

$$\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2}=\frac{800}{1681}=0.4759.$$

Loosely speaking, Theorem 1 says, for $\mu_1-\lambda_1\leq\mu_2-\lambda_2$ and for large n,

$$P\{q^r(t)\geq n\}=e^{-\delta n+o(n)}, \tag{53}$$

where $o(n)$ is a function that grows more slowly than n, i.e., $o(n)/n\to 0$ as n tends to infinity. The large deviations analysis of this paper is able to give an expression for the parameter δ,

$$\delta=-\max\{\log\frac{\lambda_2}{\lambda_2+\mu_1-\lambda_1},\log\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2}\},$$

but cannot capture the nature of $o(n)$. In each plot of Figure 2, the gap between the two curves shows the “imprecision” of the large deviations result. That is, it shows how much the large deviations result misses the actual tail probability of the queue size.

From the modelling point of view, compared with those in [2], the model in this paper allows non-IID packet delays in the DN and it specifically models packet disordering caused by routing on different paths. As for generalization, our preliminary work shows that there are similar large deviations results for more complex arrival and service processes for the queues in the DN, even for the case of non-IID arrival processes. However, in order to generalize, we must rely on more generalizable arguments than many probabilistic arguments used in this paper, which specifically depend on the underlying probability distributions. In another direction of generalization, we can consider a disordering network with k parallel M/M/1 queues, where $k\geq 3$. Preliminary investigation seems to show that there are no conceptual hurdle in that direction but careful book-keeping is required. Finally, one weakness of these models is that they do not allow situations that yield heavy-tailed distributions for the RSQ.

Appendix
Alternative Proofs for the Upper Bound in the Proof of Theorem 2: Case of $\mu_1-\lambda_1<\mu_2-\lambda_2$

We will work directly with the departure process. Following the approach in [27] (chapter 2, section 4), let

$$H_{ij}(t)=P\{q(t)=i,D(t)=j|q(0)=0\}, \tag{54}$$

where $D(t)$ is the number of departures on the interval $[0,t]$. We will show

Lemma 5:

$$H_{ij}(t)=\frac{\rho_1^{i+j}e^{-(\lambda_1+\mu_1)t}(\mu_1 t)^{2j+i}}{j!}\sum_{l=0}^{\infty}\frac{(i+l+1)(\mu_1 t)^l}{(j+i+l+1)!}. \tag{55}$$

Therefore,

$$\begin{aligned}&P\{M(t)=j|q(0)=0\}\\ =&\sum_{i=0}^{\infty}H_{ij}(t)\\ =&\frac{\rho_1^{j}e^{-(\lambda_1+\mu_1)t}(\mu_1 t)^{2j}}{j!}\sum_{i=0}^{\infty}\rho_1^i(\mu_1 t)^i\sum_{l=0}^{\infty}\frac{(i+l+1)(\mu_1 t)^l}{(j+i+l+1)!}.\end{aligned} \tag{56}$$

Proof: We will start with the integral transform of $H_{ij}(t)$. Define

$$H^*(p,q,s):=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}p^iq^j\int_0^{\infty}e^{-s\tau}H_{ij}(\tau/\mu_1)d\tau, \tag{57}$$

where $|p|<1$, $|q|<1$ and Re $s>0$. It is shown in [27] (page 198),

$$H^*(p,q,s)=\frac{(q-p)x_2(q)-(q-x_2(q))p}{(q-x_2(q))(\rho_1p^2-(1+\rho_1+s)p+q)}, \tag{58}$$

where $x_2(q)$ is one solution to the equation

$$\rho_1p^2-(1+\rho_1+s)p+q=0.$$

The two solutions for p are

$$\begin{aligned}x_1(q)=&\frac{1+\rho_1+s+\sqrt{(1+\rho_1+s)^2-4\rho_1q}}{2\rho_1},\\ x_2(q)=&\frac{1+\rho_1+s-\sqrt{(1+\rho_1+s)^2-4\rho_1q}}{2\rho_1}.\end{aligned}$$

It can be shown that for Re $s>0$ and $|q|\leq 1$,

$$|x_1(q)|>1,\qquad |x_2(q)|<1.$$

We will use the fact that, for $n=1,2,...$,

$$\begin{aligned}&x_2^n(q)\\ =&\int_0^{\infty}e^{-(1+\rho_1+s)\tau}\frac{nq^n}{\tau(\rho_1q)^{\frac{n}{2}}}I_n(2\tau\sqrt{\rho_1q})d\tau\\ =&\sum_{m=0}^{\infty}q^{n+m}\int_0^{\infty}e^{-(1+\rho_1+s)\tau}\frac{n\rho_1^m}{m!(m+n)!}\tau^{2m+n-1}d\tau.\end{aligned} \tag{59}$$

where $I_n(x)$ is the modified Bessel function of the first kind with series expansion

$$I_n(x)=\sum_{m=0}^{\infty}\frac{(\frac{1}{2}x)^{n+2m}}{m!(m+n)!},\qquad n=0,1,...$$

To find an expression for $H_{ij}(t)$, we will expand $H^*(p,q,s)$ into power series of p. Starting with (58),

$$\begin{aligned}H^*(p,q,s)=&\frac{q(x_2(q)-p)}{\rho_1(q-x_2(q))(p-x_1(q))(p-x_2(q))}\\ =&\frac{-q}{\rho_1(q-x_2(q))(p-x_1(q))}\\ =&\frac{q}{\rho_1(q-x_2(q))}\frac{1}{x_1(q)}\sum_{i=0}^{\infty}(\frac{p}{x_1(q)})^i.\end{aligned} \tag{60}$$

In the above, the series expansion is valid when $|p| < |x_1(q)|$, which is satisfied when $|p|$ is small enough. By (60) and by the definition of $H^*(p,q,s)$ in (57), the coefficient for p^i in the power series expansion of $H^*(p,q,s)$ with respect to p is

$$\sum_{j=0}^{\infty} q^j \int_0^{\infty} e^{-s\tau} H_{ij}(\tau/\mu_1) d\tau = \frac{q}{\rho_1(q-x_2(q))} \frac{1}{x_1^{i+1}(q)}. \tag{61}$$

Next, we use the fact

$$x_1(q)x_2(q) = \frac{q}{\rho_1},$$

and we express $\frac{q}{q-x_2(q)}$ in power series of $\frac{x_2(q)}{q}$. We get,

$$\sum_{j=0}^{\infty} q^j \int_0^{\infty} e^{-s\tau} H_{ij}(\tau/\mu_1) d\tau = \rho_1^i \sum_{l=0}^{\infty} \frac{x_2^{i+l+1}(q)}{q^{i+l+1}}. \tag{62}$$

The above series expansion is true when $|x_2(q)| < |q|$, which can be satisfied if $|s|$ is large enough. By (59),

$$\begin{aligned}&\sum_{j=0}^{\infty} q^j \int_0^{\infty} e^{-s\tau} H_{ij}(\tau/\mu_1) d\tau \\ =&\rho_1^i \sum_{l=0}^{\infty} \sum_{j=0}^{\infty} q^j \int_0^{\infty} e^{-(1+\rho_1+s)\tau} \frac{(i+l+1)\rho_1^j}{j!(j+i+l+1)!} \tau^{2j+i+l} d\tau.\end{aligned} \tag{63}$$

Matching both sides term-by-term, we get

$$H_{ij}(\tau/\mu_1) = \rho_1^i \sum_{l=0}^{\infty} \frac{(i+l+1)\rho_1^j}{j!(j+i+l+1)!} e^{-(1+\rho_1)\tau} \tau^{2j+i+l}. \tag{64}$$

Replacing τ/μ_1 by t, we get (55). ■

We will now show the upper bound. First, we notice that for all integer $i \geq 0$,

$$P\{M(t) = j | q(0) = i\} \leq P\{M(t) = j | q(0) = 0\}.$$

This is because

$$\begin{aligned}&P\{M(t) = j | q(0) = i\} \\ =& P\{\text{At least } j \text{ customers arrived on } [0,t], \text{ the } i \text{ customers} \\ &\text{in the queue at time 0 and the first } j \text{ customers who} \\ &\text{arrived on } [0,t] \text{ are served by time } t\},\end{aligned}$$

and this probability should be monotonically decreasing in i. Hence,

$$\begin{aligned}&P\{M(T) = n\} \\ =&\int_0^{\infty} P\{M(t) = n\}(\mu_2 - \lambda_2)e^{-(\mu_2-\lambda_2)t} dt \\ =&\int_0^{\infty} \sum_{i=0}^{\infty} P\{M(t) = n | q(0) = i\} P\{q(0) = i\} \\ &\cdot (\mu_2 - \lambda_2)e^{-(\mu_2-\lambda_2)t} dt \\ \leq&\int_0^{\infty} \sum_{i=0}^{\infty} P\{M(t) = n | q(0) = 0\} P\{q(0) = i\} \\ &\cdot (\mu_2 - \lambda_2)e^{-(\mu_2-\lambda_2)t} dt \\ =&\int_0^{\infty} P\{M(t) = n | q(0) = 0\}(\mu_2 - \lambda_2)e^{-(\mu_2-\lambda_2)t} dt.\end{aligned}$$

By (56),

$$\begin{aligned}&P\{M(T) = n\} \\ \leq&\int_0^{\infty} \frac{\rho_1^n e^{-(\lambda_1+\mu_1)t}(\mu_1 t)^{2n}}{n!} \sum_{i=0}^{\infty} \rho_1^i (\mu_1 t)^i \\ &\cdot \sum_{l=0}^{\infty} \frac{(i+l+1)(\mu_1 t)^l}{(n+i+l+1)!} (\mu_2 - \lambda_2) e^{-(\mu_2-\lambda_2)t} dt.\end{aligned}$$

Because $n+i+l+1 \geq i+l+1$ for all $n \geq 0$,

$$\begin{aligned}&P\{M(T) = n\} \\ \leq&\int_0^{\infty} \frac{\rho_1^n e^{-(\lambda_1+\mu_1)t}(\mu_1 t)^{2n}}{n!} \sum_{i=0}^{\infty} \rho_1^i (\mu_1 t)^i \\ &\cdot \sum_{l=0}^{\infty} \frac{(\mu_1 t)^l}{(n+i+l)!} (\mu_2 - \lambda_2) e^{-(\mu_2-\lambda_2)t} dt \\ =&(\mu_2 - \lambda_2) \sum_{i=0}^{\infty} \sum_{l=0}^{\infty} \frac{\rho_1^{n+i}(2n+i+l)!}{n!(n+i+l)!} \\ &\cdot \int_0^{\infty} \frac{e^{-(\lambda_1+\mu_1+\mu_2-\lambda_2)t}(\mu_1 t)^{2n+i+l}}{(2n+i+l)!} dt \\ =&\frac{\mu_2 - \lambda_2}{\lambda_1+\mu_1+\mu_2-\lambda_2} \\ &\cdot \sum_{i=0}^{\infty} \sum_{l=0}^{\infty} \frac{\rho_1^{n+i}(2n+i+l)!}{n!(n+i+l)!} (\frac{\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^{2n+i+l}.\end{aligned}$$

In deriving the last step, (16) has been used. Note that

$$\begin{aligned}\frac{(2n+i+l)!}{n!(n+i+l)!} &= \frac{(2n)!}{n!n!} \frac{(2n+1)(2n+2)...(2n+i+l)}{(n+1)(n+2)...(n+i+l)} \\ &\leq \frac{(2n)!}{n!n!} 2^{i+l}.\end{aligned}$$

Hence

$$\begin{aligned}&P\{M(T) = n\} \\ \leq&\frac{\mu_2 - \lambda_2}{\lambda_1+\mu_1+\mu_2-\lambda_2} \frac{(2n)!}{n!n!} \rho_1^n (\frac{\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^{2n} \\ &\cdot \sum_{i=0}^{\infty} 2^i \rho_1^i (\frac{\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^i \\ &\cdot \sum_{l=0}^{\infty} 2^l (\frac{\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^l \\ =&\frac{\mu_2 - \lambda_2}{\lambda_1+\mu_1+\mu_2-\lambda_2} \frac{(2n)!}{n!n!} (\frac{\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2})^n \\ &\cdot \sum_{i=0}^{\infty} (\frac{2\lambda_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^i \sum_{l=0}^{\infty} (\frac{2\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2})^l.\end{aligned} \tag{65}$$

Since $\mu_1 - \lambda_1 < \mu_2 - \lambda_2$ and $\lambda_1 < \mu_1$, it follows that

$$\frac{2\lambda_1}{\lambda_1+\mu_1+\mu_2-\lambda_2} < 1, \tag{66}$$

$$\frac{2\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2} < 1. \tag{67}$$

Hence, the two infinite sums in (65) are both finite. Furthermore, by Stirling's approximation,

$$\frac{(2n)!}{n!n!} = \frac{\sqrt{4\pi n}(2n)^{2n}e^{-2n}(1+O(1/n))}{(\sqrt{2\pi n}(n)^n e^{-n}(1+O(1/n)))^2} \leq \frac{C_2}{\sqrt{n}} 4^n, \tag{68}$$

for some constant $C_2 > 0$. Therefore,

$$P\{M(T) = n\} \leq \frac{C_3}{\sqrt{n}}\left(\frac{4\lambda_1\mu_1}{(\lambda_1+\mu_1+\mu_2-\lambda_2)^2}\right)^n, \qquad (69)$$

for some constant $C_3 > 0$. We are done with the proof for the upper bound.

Another perhaps shorter proof for the upper bound starts with (30). By writing

$$P\{\sum_{i=1}^{n} X_i \leq t\} = \sum_{l=n}^{\infty} \frac{e^{-\mu_1 t}(\mu_1 t)^l}{l!},$$

we have

$$\begin{aligned}
& P\{M(T) \geq n\} \\
\leq & \int_0^\infty \sum_{i=n}^{\infty} \frac{e^{-\lambda_1 t}(\lambda_1 t)^i}{i!} \sum_{l=n}^{\infty} \frac{e^{-\mu_1 t}(\mu_1 t)^l}{l!} \\
& \cdot (\mu_2 - \lambda_2) e^{-(\mu_2-\lambda_2)t} dt \\
= & \int_0^\infty \sum_{i=0}^{\infty} \frac{e^{-\lambda_1 t}(\lambda_1 t)^{n+i}}{(n+i)!} \sum_{l=0}^{\infty} \frac{e^{-\mu_1 t}(\mu_1 t)^{n+l}}{(n+l)!} \\
& \cdot (\mu_2 - \lambda_2) e^{-(\mu_2-\lambda_2)t} dt \\
= & (\mu_2 - \lambda_2) \sum_{i=n}^{\infty}\sum_{l=n}^{\infty} \frac{\rho_1^{n+i}(2n+i+l)!}{(n+i)!(n+l)!} \\
& \cdot \int_0^\infty \frac{e^{-(\lambda_1+\mu_1+\mu_2-\lambda_2)t}(\mu_1 t)^{2n+i+l}}{(2n+i+l)!} dt \\
= & \frac{\mu_2 - \lambda_2}{\lambda_1+\mu_1+\mu_2-\lambda_2} \\
& \cdot \sum_{i=n}^{\infty}\sum_{l=n}^{\infty} \frac{\rho_1^{n+i}(2n+i+l)!}{(n+i)!(n+l)!}\left(\frac{\mu_1}{\lambda_1+\mu_1+\mu_2-\lambda_2}\right)^{2n+i+l}.
\end{aligned} \qquad (70)$$

Next,

$$\begin{aligned}
\frac{(2n+i+l)!}{(n+i)!(n+l)!} &= \frac{(2n)!}{n!n!}\frac{(2n+1)(2n+2)...(2n+i+l)}{(n+1)...(n+i)(n+1)...(n+l)} \\
&\leq \frac{(2n)!}{n!n!}2^{i+l}.
\end{aligned} \qquad (71)$$

The rest steps are exactly the same as those starting at (65).

References

[1] V. Paxson, “End-to-end Internet packet dynamics,” *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 277–292, 1999.

[2] Y. Xia and D. Tse, “Analysis on Packet Resequencing for Reliable Network Protocols,” in *Proceedings of the IEEE Infocom 2003*, San Francisco, CA, April 2003.

[3] F. Kamoun, L. Kleinrock, and R. Muntz, “Queueing analysis of the reordering issue in a distributed database concurrency control mechanism,” in *Proceedings of the 2nd International Conference on Distributed Computing Systems*, Versailles, France, April 1981, pp. 13–23.

[4] G. Harrus and B. Plateau, “Queueing analysis of a reordering issue,” *IEEE Transaction on Software Engineering*, vol. SE-8, no. 2, pp. 113–123, March 1982.

[5] F. Baccelli, E. Gelenbe, and B. Plateau, “An end-to-end approach to the resequencing problem,” *Journal of the Association for Computing Machinery, Vol. 31, No. 3*, pp. 474–485, July 1984.

[6] Y.-C. Lien, “Evaluation of the resequence delay in a Poisson queueing system with two heterogeneous servers,” in *Proceedings of the International Workshop on Computer Performance Evaluation*, Tokyo, Japan, September 1985, pp. 189–197.

[7] T.-S. P. Yum and T.-Y. Ngai, “Resequencing of messages in communication networks,” *IEEE Transactions on Communications*, vol. COM-34, no. 2, pp. 143–149, February 1986.

[8] S. Chowdhury, “An analysis of virtual circuits with parallel links,” *IEEE Transactions on Communications*, vol. 39, no. 8, pp. 1184–1188, August 1991.

[9] I. Iliadis and L. Y.-C. Lien, “Resequencing delay for a queueing system with two heterogeneous servers under a threshold-type scheduling,” *IEEE Transactions on Communications*, vol. 36, no. 6, pp. 692–702, June 1988.

[10] N. Gogate and S. S. Panwar, “On a resequencing model for high speed networks,” in *Proceedings of INFOCOM ’94*, Toronto, Canada, June 1994, pp. 40–47.

[11] A. Jean-Marie and L. Gün, “Parallel queues with resequencing,” *Journal of the Association for Computing Machinery*, vol. 40, no. 5, pp. 1188–1208, November 1993.

[12] F. Baccelli and A. M. Makowski, “Queueing models for systems with synchronization constraints,” *Proceedings of the IEEE, Vol. 77, No. 1*, pp. 138–161, January 1989.

[13] D. Towsley and J. K. Wolf, “On the statistical analysis of queue lengths and waiting times for statistical multiplexers with ARQ retransmission schemes,” *IEEE Transactions on Communications*, vol. COM-27, no. 4, pp. 693–702, April 1979.

[14] A. G. Konheim, “A queueing analysis of two ARQ protocols,” *IEEE Transactions on Communications*, vol. COM-28, no. 7, pp. 1004–1014, July 1980.

[15] M. J. Miller and S.-L. Lin, “The analysis of some selective-repeat ARQ schemes with finite receiver buffer,” *IEEE Transactions on Communications*, vol. COM-29, no. 9, pp. 1307–1315, September 1981.

[16] B. H. Saeki and I. R. Rubin, “An analysis of a TDMA channel using stop-and-wait, block, and selective-and-repeat ARQ error control,” *IEEE Transactions on Communications*, vol. COM-30, no. 5, pp. 1162–1173, May 1982.

[17] M. E. Anagnostou and E. N. Protonotarios, “Performance analysis of the selective repeat ARQ protocol,” *IEEE Transactions on Communications*, vol. COM-34, no. 2, pp. 127–135, February 1986.

[18] Z. Rosberg and N. Shacham, “Resequencing delay and buffer occupancy under the selective-repeat ARQ,” *IEEE Transactions on Information Theory, Vol. 35, No. 1*, pp. 166–172, January 1989.

[19] Z. Rosberg and M. Sidi, “Selective-repeat ARQ: the joint distribution of the transmitter and the receiver resequencing buffer occupancies,” *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1430–1438, September 1990.

[20] D. Towsley, “The Stutter Go Back-N ARQ Protocol,” *IEEE Transactions on Communications*, vol. COM-27, no. 6, pp. 869–875, June 1979.

[21] N. Shacham and D. Towsley, “Resequencing Delay and Buffer Occupancy in Selective Repeat ARQ with Multiple Receivers,” *IEEE Transactions on Communications*, vol. COM-39, no. 6, pp. 928–937, June 1991.

[22] N. Shacham and B. C. Shin, “A Selective-Repeat-ARQ Protocol for Parallel Channels and Its Resequencing Analysis,” *IEEE Transactions on Communications*, vol. 40, no. 4, pp. 773–782, April 1992.

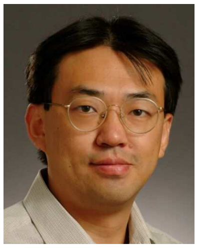
[23] S. Varma, “Optimal Allocation of Customers in a Two Server Queue with Resequencing,” *IEEE Transactions on Automatic Control*, vol. 36, no. 11, pp. 1288–1293, November 1991.

[24] S. Ayoun and Z. Rosberg, “Optimal Routing to Two Parallel Heterogeneous Servers with Resequencing,” *IEEE Transactions on Automatic Control*, vol. 36, no. 12, pp. 1436–1449, December 1991.

[25] L. Kleinrock, *Queue Systems, Volume I: Theory*. Jonh Wiley & Sons, 1975.

[26] A. Schwartz and A. Weiss, *Large Deviation for Performance Analysis - Queues, Communications and Computing*. Chapman & Hall, 1995.

[27] J. W. Cohen, *The Single Server Queue*, 1st ed. North-Holland Publishing Company, 1969.



Ye Xia received the BA degree in 1993 from Harvard University, the MS degree in 1995 from Columbia University and the PhD degree in 2003 from University of California, Berkeley, all in Electrical Engineering. Between June 1994 and August 1996, he was a member of the technical staff at Bell Laboratories, Lucent Technologies in New Jersey, where he worked in performance evaluation of a shared-memory ATM switch and studied traffic control for ATM networks. Since August 2003 to the present, he is an assistant professor in the Computer and Information Science and Engineering department at the University of Florida. His primary research interest is in communication networks.

David Tse received the B.A.Sc. degree in systems design engineering from University of Waterloo, Canada in 1989, and the M.S. and Ph.D. degrees in electrical engineering from Massachusetts Institute of Technology in 1991 and 1994 respectively. From 1994 to 1995, he was a postdoctoral member of technical staff at A.T. & T. Bell Laboratories. Since 1995, he has been at the Department of Electrical Engineering and Computer Sciences in the University of California at Berkeley, where he is currently a Professor. He received a 1967 NSERC 4-year graduate fellowship from the government of Canada in 1989, a NSF CAREER award in 1998, the Best Paper Awards at the Infocom 1998 and Infocom 2001 conferences, the Erlang Prize in 2000 from the INFORMS Applied Probability Society, the IEEE Communications and Information Theory Society Joint Paper Award in 2001, and the Information Theory Society Paper Award in 2003. He was the Technical Program co-chair of the International Symposium on Information Theory in 2004, and was an Associate Editor of the IEEE Transactions on Information Theory from 2001 to 2003. He is a coauthor, with Pramod Viswanath, of the text ”Fundamentals of Wireless Communication”. His research interests are in information theory, wireless communications and networking.