Large Deviation Bounds for a Polling System with Two Queues and Multiple Servers

Wei FENG*
Nagare College

(Received August 19, 2003)

In this paper, we present large deviation bounds for a discrete-time polling system consisting of two-parallel queues and m servers. The arrival process in each queue is an arbitrary, and possibly correlated, stochastic process. Each server (serves) independently serves the two queues according to a Bernoulli service schedule. Using large deviation techniques, we analyze the tail behavior of the stationary distribution of the queue length processes, and derive upper and lower bounds of the buffer overflow probability for each queue. These results have important implications for traffic management of high-speed communication networks such as call admission, bandwidth allocation, and server control.

Keywords: polling system, multiple servers, Bernoulli service schedule, effective bandwidth, large deviation, queue length tail distribution

1. Introduction

Polling systems have found many applications in the areas of high-speed ATM (Asynchronous Transfer Mode) networks, computer-communication networks, and multiprocessor systems, where service capacity (CPU, bandwidth) has to be shared among different users. In particular, polling systems consisting of two-parallel queues and multiple servers have been used to model communication systems with two different types of traffic: the real-time traffic (e.g., voice and video) and the non-real-time traffic (e.g., data), and wireless mobile communication systems with originating calls and handoff calls (e.g., Bluetooth system) (see [20] and [21] for detailed analyses and surveys on this subject). In order to meet diverse *Quality of Service* (QoS) requirements, various service schedules such as the exhaustive, k-limited ([10], [18]), Bernoulli ([11], [16]) and Markovian ([4], [6], [17], [19], [22]) have also been proposed. In this paper, we consider a discrete-time fluid polling system consisting of two-parallel queues (Q_1 and Q_2) and $m(\geq 1)$ servers. By *discrete-time fluid*, we mean that all arrival and service happen at discrete-time slots indexed by integers, and are in the form of fluid. The arrival process in Q_i is an arbitrary, and possibly correlated stochastic process. Each server visits Q_1 and Q_2 independently each other according to a Bernoulli service schedule: at the beginning of each discrete-time, if both queues are not empty, then the kth server just completing the service in Q_i makes a random decision: with probability p_i^k, $0 < p_i^k < 1$, it continues to deal with cells of Q_i in the next slot, and with probability $q_i^k = 1 - p_i^k$, it switches to Q_j ($j \neq i$) and deals with cells of Q_j in the next slot. The service rate of each server is c, and the service policy is assumed to be work-conserving. That is, each server is permitted to devote its residual service capacity to another queue whenever the present queue becomes empty. Furthermore, each server is assumed not to experience switching times in its transition from one queue to the other. All arrival processes and service processes are mutually independent.

The motivation for this work is three-fold. First, as development of high-speed packet-switched communication networks employing ATM technology, discrete-time fluid models become more and more important. Up to now, most of work for discrete-time fluid models is mainly devoted to performance analysis of single queueing systems, the problem of analyzing discrete-time fluid polling systems have received remarkable little attention in the literature. Secondly, the polling system considered here is actually deduced from discretization of a continuous-time Markovian fluid polling system, with continuous-time arrival processes

---

* Department of Engineering physics, Electronics and Mechanics, Graduate School of Engineering, Nagare College
Supported in part by Grant-in-Aid for Scientific Research (No.15510124), Ministry of Education, Culture, Sports, Science and Technology

and Markovian service processes. This model is important in describing the dynamics of high-speed ATM networks by using fluid mechanics. Finally, for the polling systems with general, possibly autocorrelate arrival processes and multiple servers, getting the exact stationary distribution of queue-length processes is extremely difficult, because of the autocorrelate structure of the arrival processes, and the complexity of the service processes. To the best of our knowledge, no any analytic results on the discrete-time fluid polling system considered here have been obtained. As a mathematical problem, therefore, it is also a challenging work to study the behavior of the polling system.

In this paper, we utilize large deviation techniques to analyze the tail behavior of the stationary distribution of the queue length processes, and derive upper and lower bounds of the buffer overflow probability for each queue. In recent decade, large deviation techniques have been extensively applied to problems of estimating tail probability of rare events in single queueing systems ([3], [5], [9], [13]) and queueing networks ([1], [2], [7], [12], [18], [23], [24]). In [1], [2] and [24], large deviation results for a network system consisting of two-parallel queues and a single sever have been derived under *Generalized Process Sharing* (GPS) service discipline. In [12], we obtained large deviation bounds for a polling system consisting of two-parallel queues and a single server. For a Weighted Round Robin (WRR) polling system, Massoulie[18] derived logarithmic equivalents of the stationary tail distribution of the queue length for each queue, by using sample path large deviation techniques. For an Markovian polling system with a single server, Poisson arrival processes and exponentially distributed service times, Delcoigne and Fortelle [7] presented the local rate function governing the sample path large deviation principle.

The paper is organized as follows. In Section 2, we briefly review some conceptions and results from large deviation theory on the real-line $\mathbf{R}$. Then, we define exactly the potential service process of each server by using Markov chains, and give some large deviation results for these service processes. In Section 3, we show our main theorem —the upper and lower bounds of the overflow probabilities, and in Section 4, we prove the theorem. Finally, some conclusions are included in Section 5.

2. Preliminaries

Throughout the paper, all time indices t, τ, etc., are always integers. $\mathbf{N}=\{0,1,2,\cdots\}$. We denote by $S^X_{\tau,t}=\sum_{k=\tau}^{t-1}X_k$, $\tau<t$ and $S^X_t=\sum_{k=0}^{t-1}X_k$ the partial sums of the random sequence $X=\{X_t;\, t\in\mathbf{N}\}$, and by $S^X_t(s)=\sum_{k=0}^{\lceil ts\rceil}X_k/t$, $0\le s\le 1$ the scaled partial sum of X, respectively. Furthermore, we denote by $\Lambda_X(\theta)$ the limit logarithmic moment generating function of the partial sum process of X, and by $\Lambda^*_X(\alpha)$ the Legendre-Fenchel transform of $\Lambda_X(\theta)$. Namely,

$$\Lambda_X(\theta)=\lim_{t\to\infty}\frac{1}{t}\log E[e^{\theta S^X_t}],\quad \theta\in\mathbf{R};\quad \Lambda^*_X(\alpha)=\sup_{\theta\in\mathbf{R}}\{\theta\alpha-\Lambda_X(\theta)\},\ \alpha\in\mathbf{R}. \tag{1}$$

A. Basic assumptions and definitions

We first give some definitions and results from large deviation theory on the real-line $\mathbf{R}$. A *rate function* I from $\mathbf{R}$ to $[0,\infty]$ is *good* if all level sets $\{y\in\mathbf{R};\, I(y)\le x\}$, $x\in[0,\infty)$ are compact. A sequence of probability measures $\{\mu_t;\, t\in\mathbf{N}\}$ on $\mathbf{R}$ satisfies *Large Deviation Principle* (LDP) with a good rate function I if

1. Upper Bound: For every closed set F, $$\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(F)\le-\inf_{x\in F}I(x). \tag{2}$$

2. Lower Bound: For every open set G, $$\liminf_{n\to\infty}\frac{1}{n}\log\mu_n(G)\ge-\inf_{x\in G}I(x). \tag{3}$$

Let $\{X_t;\, t\in\mathbf{N}\}$ be a sequence of random variables on $\mathbf{R}$ and $\{\mu_t;\, t\in\mathbf{N}\}$ the corresponding sequence of probability measures. If $\{\mu_t;\, t\in\mathbf{N}\}$ satisfies the large deviation principle with a good rate function I, we say $\{X_t;\, t\in\mathbf{N}\}$ also satisfies the large deviation principle with a good rate function I.

Assumption A1:

(1) The limit logarithmic moment generating function $\Lambda_X(\theta)$ exists for all θ as an extended ratl function, i.e., $\pm\infty$ are permitted as limit points.

(2) The origin is in the interior of the domain $D_{\Lambda_X}\equiv\{\theta;\ \Lambda_X(\theta)<\infty\}$ of $\Lambda_X(\theta)$.

(3) $\Lambda_X(\theta)$ is differentiable in the interior of D_{Λ_X}, and derivative tends to infinity as θ approaches the boundary of D_{Λ_X}

(4) $\Lambda_X(\theta)$ is lower semicontinuous, i.e., $\liminf_{\theta_n\to\theta}\Lambda_X(\theta_n)\ge\Lambda_X(\theta)$.

Theorem 2.1 (Gärtner-Ellis): Under Assumption A1, the large deviation upper bound (2) and lower bound (3) hold with the good rate function $I=\Lambda_X^*$.

Note that $\Lambda_X(\cdot)$ and $\Lambda_X^*(\cdot)$ are convex dual, namely, $\Lambda_X(\theta)$ is also the Legendre-Frenchel transform of $\Lambda_X^*(\alpha)$.

$$\Lambda_X(\theta)=\sup_{\alpha\in\mathbf{R}}\{\theta\alpha-\Lambda_X^*(\alpha)\},\quad \theta\in\mathbf{R}. \tag{4}$$

The following properties of $\Lambda_X(\theta)$ and $\Lambda_X^*(\alpha)$ hold (cf. [24]).

Proposition 2.2: (i) $\Lambda_X(\theta)$ and $\Lambda_X^*(\alpha)$ are all strictly convex and essentially smooth.

(ii) $\Lambda_X(\theta)$ and $\Lambda_X'(\theta)$ are all strictly increasing.

(iii) *dom* $\Lambda_X^*(\alpha)=$*int*(*ran* Λ_X') and $\Lambda_X^*(\alpha)$ is continuous in *int*(*dom* Λ_X^*). In particular, $\inf_{\alpha>a_l}\Lambda_X^*(\alpha)=\inf_{\alpha\geq a_l}\Lambda_X^*(\alpha)$ and $\inf_{\alpha<a_r}\Lambda_X^*(\alpha)=\inf_{\alpha\leq a_r}\Lambda_X^*(\alpha)$, where a_l and a_r the left and right end points of *dom* Λ_X^*, respectively.

(iv) Let $\bar{x}=E[X_1]$, then $\Lambda_X^*(\bar{x})=0$ and $\Lambda_X'(0)=\bar{x}$.

(v) $\Lambda_X(\theta)+\Lambda_X^*(\alpha)=\theta\alpha$ if and only if $\alpha=\Lambda_X'(\theta)$.

A stronger concept than LDP for the partial sum process S_t^X is the *Sample Path Large Deviation Principle* (LDPps) for the partial process $S_t^X(s)$. Let $D([0,1],\mathbf{R})$ denote the space of right continuous function with left limits from [0, 1] to $\mathbf{R}$ equipped with the supremum norm topology. A sequence of probability measures $\{\mu_t, t\in\mathbf{N}\}$ on $D([0,1],\mathbf{R})$ satisfies the LDPsp with a good rate function $I(\phi)$ if (i) $I(\phi)$ is a function from $D([0,1],\mathbf{R})$ to $[0,\infty]$ with compact level sets, and (ii) the upper bound (2) and the lower bound (3) hold for any closed and open sets in $D([0,1],\mathbf{R})$.

Assumption A2: $\{X_t; t\in\mathbf{N}\}$ is adapted to a filtration $\{\mathcal{F}_t^X; t\in\mathbf{N}\}$ with the following property: for any $\theta\in\mathbf{R}$, there exists a function $\Gamma_X(\theta)$, $0\leq\Gamma_X(\theta)$, such that for any $\tau,t\geq 0$

$$\Lambda_X(\theta)t-\Gamma_X(\theta)\leq\log E[e^{\theta S_{\tau,\tau+t}^X}\mid\mathcal{F}_\tau^X]\leq\Lambda_X(\theta)t+\Gamma_X(\theta)\quad a.s.. \tag{5}$$

Let μ_t be the probability measure of $S_t^X(\cdot)$. Then, under Assumption A2, $\{\mu_t, t\in\mathbf{N}\}$ satisfies the LDPsp with the good rate function $I(\cdot)$ defined as follows: for any $\phi\in D([0,1],\mathbf{R})$,

$$I_X(\phi)=\begin{cases}\int_0^t\Lambda_X^*(\phi'(t))dt, & \text{if}\ \phi\in AC_0([0,1],\mathbf{R}),\\ \infty, & \text{otherwise,}\end{cases} \tag{6}$$

where $AC_0([0,1],\mathbf{R})$ is the space of absolutely continuous function from [0, 1] to $\mathbf{R}$ with $\phi(0)=0$, and $\phi'(t)$ is the derivative of $\phi(t)$.

As we allow dependence between the number of customers at different slots in each arrival process, this will lead to the fact that at steady state, the arrival magnitude at t may be dependent on the stationary queue length at $\tau(<t)$. The following assumption permits us to deal with such dependence in deriving the large deviation results for each queue.

Assumption A3: Let $\mathcal{F}_{(-\infty,k)}^X=\sigma\{X_t; -\infty<t\leq k\}$ and

$$\upsilon^X(n)=\sup_{U\in\mathcal{F}_{(-\infty,k)}^X,U'\in\mathcal{F}_{(k+n,\infty)}^X,P\{U\}>0}|P(U'|U)-P(U')|, \tag{7}$$

then, $\lim_{n\to\infty}\upsilon^X(n)=0$.

The assumptions (A1), (A2) and (A3) are satisfied by processes that are commonly used to model burst traffic in communication networks, e.g., renewal processes, Markov-modulated processes and more generally stationary processes with mild mixing conditions.

B. Arrival processes and potential service processes

The arrival process in Q_i is denoted by $\{A_t^i; t\in\mathbf{N}\}$. We assume that $\{A_t^i; t\in\mathbf{N}\}$ satisfies the following conditions:

(1) The arrival process $\{A_t^i; t\in\mathbf{N}\}$ is ergodic and strictly stationary.

(2) The arrival process $\{A_t^i; t\in\mathbf{N}\}$ satisfies Assumptions A1, A2 and A3.

Now we define the service process in Q_i. Since each server serves the two queues according to Bernoulli service schedule, the lengths of the *service period*(duration that one sever continues to serve Q_i) and the *non-service period* are two independent sets of i.i.d. random variables with geometric distributions as follows: for $t \geq 1$

$$P\{\text{ service period of } Q_i \text{ contains t slots }\} = (1-p_i)p_i^{t-1},$$
$$P\{\text{ non-service period of } Q_i \text{ contains t slots }\} = (1-p_j)p_j^{t-1},\ j \neq i.$$

Let b_t^i be the number of servers during the slot t. Then from the geometric nature of the service and non-service periods, it follows that b_{t+1}^i for $i=1,2$ can be derived from b_t^i as follows:

$$b_{t+1}^1 = \sum_{k=1}^{b_t^1} \sigma_k + \sum_{k=1}^{m-b_t^1} (1-\eta_k), \qquad b_{t+1}^2 = \sum_{k=1}^{m-b_t^2} (1-\sigma_k) + \sum_{k=1}^{b_t^2} \eta_k, \tag{8}$$

where $\{\sigma_k, k=1,2,\cdots,m\}$ and $\{\eta_k, k=1,2,\cdots,m\}$ are two independent sets of i.i.d. Bernoulli random variables with probability distributions

$$P\{\sigma_k=1\}=p_1, \quad P\{\sigma_k=0\}=q_1 \quad \text{and} \quad P\{\eta_k=1\}=p_2, \quad P\{\eta_k=0\}=q_2.$$

Note that $b_t^1 + b_t^2 = m$ for any t, and m servers are independent. We have the following proposition.

Proposition 2.3: For $i=1,2$, the process $\{b_t^i;\ t \in \mathbf{N}\}$ is an irreducible Markov chain with state space $\{0,1,2,\cdots,m\}$ and transition probability matrix $\mathrm{P}_{b^i} = (p_{lk}^i)$, where for $l,k \in \{0,1,2,\cdots,m\}$

$$p_{lk}^1 \equiv P\{b_{t+1}^1 = k | b_t^1 = l\} = \sum_{n=0}^{\min\{l,k\}} \binom{l}{n} p_1^n q_1^{l-n} \binom{m-l}{k-n} q_2^{k-n} p_2^{(m-l)-(k-n)},$$

$$p_{lk}^2 \equiv P\{b_{t+1}^2 = k | b_t^2 = l\} = P\{b_{t+1}^1 = m-k | b_t^1 = m-l\}$$
$$= \sum_{n=0}^{\min\{m-l,\,m-k\}} \binom{m-l}{n} p_1^n q_1^{m-l-n} \binom{l}{m-k-n} q_2^{m-k-n} p_2^{l-(m-k-n)}.$$

As the service rate of each server is c, $B_t^i = b_t^i c$ denotes the total service rate devoted to Q_i in the slot t. Obviously, $\{B_t^i;\ t \in \mathbf{N}\}$ is also a Markov chain with the state space $\{0,c,2c,\cdots,mc\}$ and the transition probability matrix $\mathrm{P}_{B^i} = (p_{lc,kc}^i)$, where $p_{lc,kc}^i = p_{lk}^i$. We call $\{B_t^i;\ t \in \mathbf{N}\}$ *the potential service process (MSP)*, and denote its equilibrium distribution by $\pi_B^i = (\pi_0^i, \pi_1^i, \cdots, \pi_m^i)$ and its mean by $\mathcal{B}^i = E[B_t^i] = \Sigma_{k=0}^m kc\pi_k^i$.

C. The stability condition

Let L_t^i be the queue length (backlog) of Q_i at time t. $L_t = L_t^1 + L_t^2$. Note that $B_t^1 + B_t^2 = mc$ and no switching times are experienced during each server transitions. It follows from Loynes's Stability Theorem [15] that the stable condition of the polling system is

$$\mathcal{A}^1 + \mathcal{A}^2 < mc \tag{9}$$

where $\mathcal{A}^i = E[A_1^i]$. Throughout the paper, we assume that the condition (9) hold. Thus, the aggregate queue length process L_t converges in distribution to a finite random variable. As $L_t^i \leq L_t$, L_t^i converges also in distribution to a finite random variable.

D. Large deviation results on the potential service processes

For any $\theta \in \mathbf{R}$ and $i=1,2$, define $(m+1)\times(m+1)$ matrices $\Psi_{B^i}(\theta) = (p_{lc,kc}^i e^{\theta kc})$. Let $\rho_{B^i}(\theta) = sp(\Psi_{B^i}(\theta))$ be the spectral radii of $\Psi_{B^i}(\theta)$ and $\boldsymbol{x}^{B^i}(\theta) = (x_0^i(\theta), x_1^i(\theta), \cdots, x_m^i(\theta))^T$ the positive right eigenvector corresponding to $\rho_{B^i}(\theta)$. Furthermore, define $\Gamma_{B^i}(\theta) = \max_{0 \leq k,j \leq m} x_k^i(\theta)/x_j^i(\theta)$. When $m=1$, $\rho_{B^i}(\theta)$, $\boldsymbol{x}^{B^i}(\theta)$ and $\Gamma_{B^i}(\theta)$ can be directly calculated. We have for $i, j=1,2$; $i \neq j$, $\rho_{B^i}(\theta) = (p_j + p_i e^{\theta c} + \sqrt{(p_j - p_i e^{\theta c})^2 + 4\, q_1 q_2 e^{\theta c}}\,)/2$, $\boldsymbol{x}^{B^i}(\theta) = ((\rho_{B^i}(\theta) - p_i e^{\theta c}) / (\rho_{B^i}(\theta) + q_i - p_i e^{\theta c}),\ q_i / (\rho_{B^i}(\theta) + q_i - p_i e^{\theta c_i}))^T$, and $\Gamma_{B^i}(\theta) = \max\{q_i/(\rho_{B^i}(\theta) - p_i e^{\theta c}),\ (\rho_{B^i}(\theta) - p_i e^{\theta c})/q_i\}$. Utilizing the general discussion on the large deviation of Markov chains (see [6],[7],[8] and [11] for details), we obtained the following large deviation

results on the Markov chain $\{B_t^i, t\in\mathbf{N}\}$.

Theorem 2.4: (i) $\Lambda_{B^i}(\theta)=\log(\rho_{B^i}(\theta))$, and

$$\Lambda^*_{B^i}(\alpha)=\sup_{\theta\in\mathbf{R}}=\{\theta\alpha-\Lambda_{B^i}(\theta)\}=\begin{cases}\sup_{\theta\geq 0}\{\theta\alpha-\Lambda_{B^i}(\theta)\} & \text{if } \mathcal{B}_i<\alpha\leq mc\\ \sup_{\theta\geq 0}\{\theta\alpha-\Lambda_{B^i}(\theta)\} & \text{if } 0<\alpha\leq\mathcal{B}_i\\ \infty & \text{otherwise.}\end{cases}\tag{10}$$

(ii) The processes $\{S_t^{B^i}/t\,; t\in\mathbf{N}\}$ satisfies the large deviation principle with the convex, good rate function $\Lambda^*_{B^i}(\alpha)$.

(iii) Let $\mathcal{F}_t^{B^i}=\sigma\{B_\tau^i;\ \tau\leq t\}$, then for all $\theta\in\mathbf{R}$ and $\tau, t\leq 0$,

$$\Lambda_{B^i}(\theta)t-\Gamma_{B^i}(\theta)\leq\log E[e^{\theta S^{B^i}_{\tau,\tau+t}}|\mathcal{F}_\tau^{B^i}]=\log E[e^{\theta S^{B^i}_{\tau,\tau+t}}|B_\tau^i]\leq\Lambda_{B^i}(\theta)t+\Gamma_{B^i}(\theta),\ a.s..$$

(v) The process $\{B_t^i, t\in\mathbf{N}\}$ satisfies Assumption A3, i.e., $\lim_{n\to\infty}\upsilon^{B^i}(n)=0$.

3. Large deviation bounds of the overflow probability

In this section, we present our main theorem-the large deviation upper and lower bounds of the overflow probability for the polling system. The proof of the theorem is relegated to the next section. By $\alpha_{A^i}(\theta)$ and $\alpha_{B^i}(\theta)$ we express the effective bandwidths of $\{A_t^i, t\in\mathbf{N}\}$ and $\{B_t^i, t\in\mathbf{N}\}$, respectively, i.e., $\alpha_{A^i}(\theta)=\Lambda_{A^i}(\theta)/\theta$ and $\alpha_{B^i}(\theta)=\Lambda_{B^i}(\theta)/\theta$. Furthermore, we define $\alpha_{D^i}(\theta)=\Lambda_{D^i}(\theta)/\theta$ and $\alpha_{E^i}(\theta)=\Lambda_{E^i}(\theta)/\theta$ for $i=1,2$, where $\Lambda_{D^i}(\theta)$ and $\Lambda_{E^i}(\theta)$ are given as follows.

For any $\theta\geq 0$,

$$\Lambda_{D^i}(\theta)=\begin{cases}
\textit{CASE1. } \mathcal{A}^i<\Lambda'_{A^i}(\delta_i^*)\leq\mathcal{B}^i<\min\{a_r^i, mc\}\\
\quad\Lambda_{A^i}(\theta) & \text{if } \theta\leq\delta_i^*\\
\quad\Lambda_{A^i}(\delta_i^*)+\Lambda_{B^i}(\theta-\delta_i^*) & \text{if } \delta_i^*<\theta \text{ and } \mathcal{B}^i\leq\Lambda'_{B^i}(\theta-\delta_i^*)\leq\min\{a_r^i, mc\}\\
\quad\Lambda_{A^i}(\delta_i^*)+(\theta-\delta_i^*)\min\{a_r^i, mc\}-\Lambda^*_{B^i}(\min\{a_r^i, mc\}) & \text{if } \delta_i^*<\theta \text{ and } \min\{a^r, mc\}<\Lambda'_{B^i}(\theta-\delta_i^*)\\
\textit{CASE2. } \mathcal{A}^i<\mathcal{B}^i<\Lambda'_{A^i}(\delta_i^*)\leq\min\{a_r^i, mc\}\\
\quad\Lambda_{A^i}(\theta) & \text{if } \theta:\Lambda'_{A^i}(\theta)\leq\mathcal{B}^i\\
\quad J_i(\theta) & \text{if } \theta:\Lambda'_{A^i}(\theta)>\mathcal{B}^i, \theta\leq\delta_i^* \text{ or } \Lambda'_{A^i}(\theta)>\mathcal{B}^i, \theta>\delta_i^*;\ \Lambda'_{B^i}(\theta-\delta_i^*)\leq\Lambda'_{A^i}(\delta_i^*)\\
\quad\max\{J_i(\theta), \Lambda_{A^i}(\delta_i^*)+\Lambda_{B^i}(\theta-\delta_i^*)\} & \text{if } \theta:\Lambda'_{A^i}(\theta)>\mathcal{B}^i, \theta>\delta_i^* \text{ and } \Lambda'_{A^i}(\delta_i^*)<\Lambda'_{B^i}(\theta-\delta_i^*)\leq\min\{a_r^i, mc\}\\
\quad\max\{J_i(\theta), \Lambda_{A^i}(\delta_i^*)+(\theta-\delta_i^*)\min\{a_r^i, mc\}-\Lambda^*_{B^i}(\min\{a_r^i, mc\})\} & \text{if } \theta:\Lambda'_{A^i}(\theta)>\mathcal{B}^i, \theta>\delta_i^* \text{ and } \Lambda'_{B^i}(\theta-\delta_i^*)>\min\{a_r^i, mc\}\\
\textit{CASE3. } \mathcal{A}^i<\mathcal{B}^i<\min\{a_r^i, mc\}<\Lambda'_{A^i}(\delta^*)\\
\quad\Lambda_{A^i}(\theta) & \text{if } \theta:\Lambda'_{A^i}(\theta)\leq\mathcal{B}^i\\
\quad K_i(\theta) & \text{if } \theta:\Lambda'_{A^i}(\theta)>\mathcal{B}^i\\
\textit{CASE4. } \mathcal{A}^i\geq\mathcal{B}^i\\
\quad\Lambda_{B^i}(\theta)
\end{cases}\tag{11}$$

where, δ_i^* is the largest solution of the equation $\Lambda_{A^i}(\theta)+\Lambda_{B^i}(-\theta)=0$, and $J_i(\theta)=(\theta-\theta^*_{A^i}(\theta)-\tilde{\theta}^*_{B^i}(\theta)\eta^{A^iB^i}(\theta)+\Lambda_{A^i}(\tilde{\theta}^*_{A^i}(\theta))+\Lambda_{B^i}(\tilde{\theta}^*_{B^i}(\theta))$, here $\eta^{A^iB^i}(\theta)$ is the maximum point of function $\theta\alpha-\Lambda^*_{A^i}(\alpha)-\Lambda^*_{B^i}(\alpha)$ in the interval $[\mathcal{B}^i, \Lambda'_{A^i}(\delta^*)]$. For θ fixed, $\tilde{\theta}^*_{A^i}(\theta)$ and $\tilde{\theta}^*_{B^i}(\theta)$ are the unique solutions of the equations $\Lambda'_{A^i}(\tilde{\theta})=\eta^{A^iB^i}(\theta)$ and $\Lambda'_{B^i}(\tilde{\theta})=\eta^{A^iB^i}(\theta)$, respectively. $K_i(\theta)=(\theta-\hat{\theta}^*_{A^i}(\theta)-\hat{\theta}^*_{B^i}(\theta))\xi^{A^iB^i}(\theta)+\Lambda_{A^i}(\hat{\theta}^*_{A^i}(\theta))+\Lambda_{B^i}(\hat{\theta}^*_{B^i}(\theta))$. $\xi^{A^iB^i}(\theta)$ is the maximum point of the

function $\theta\alpha - \Lambda^*_{A^i}(\alpha) - \Lambda^*_{B^i}(\alpha)$ in the interval $[\mathcal{B}^i, \min\{a_r^i, mc\}]$, here a_r^i is the right end point of $dom\,\Lambda^*_{A^i}$. For θ fixed, $\hat{\theta}^*_{A^i}(\theta)$ and $\hat{\theta}^*_{B^i}(\theta)$ are the unique solutions of the equations $\Lambda'_{A^i}(\hat{\theta}) = \xi^{A^iB^i}(\theta)$ and $\Lambda'_{B^i}(\hat{\theta}) = \xi^{A^iB^i}(\theta)$ respectively.

$$\Lambda_{E^i}(\theta) = \begin{cases} CASE1.\ \mathcal{A}^i < \mathcal{B}^i & \\ \quad \Lambda_{A^i}(\theta) & \text{if } \theta : \Lambda'_{A^i}(\theta) \le \mathcal{B}^i \\ \quad K_i(\theta) & \text{if } \theta : \Lambda'_{A^i}(\theta) > \mathcal{B}^i \\ CASE2.\ \mathcal{A}^i \ge \mathcal{B}^i & \\ \quad \mathcal{B}^i\theta & \end{cases} \tag{12}$$

where $K_i(\theta)$ is given in the definition (11).

Remark 1. As will be seen, $\Lambda_{E^i}(\theta)$ and $\Lambda_{D^i}(\theta)$ are respectively the effective bandwidths of the transient departure process (e.g., the departure process from an empty queue at time 0) and the stationary departure process from an *G/MSP/m* queueing system with the arrival process $\{A_t^i, t \in \mathbf{N}\}$ and the service process $\{B_t^i, t \in \mathbf{N}\}$.

Theorem 3.1: Under the stability condition (9), the queue length L_0^i of Q_i at steady state satisfies the following bounds.

(i) *upper bound*:

$$\limsup_{x\to\infty} \frac{1}{x} \log P\{L_0^i > x\} \le -\Theta^*_{ij}, \tag{13}$$

where, Θ^*_{ij} is the unique solution of the equation:

$$\alpha_{A^i}(\theta) + \alpha_{D^j}(\theta) = mc, \quad i \ne j. \tag{14}$$

(ii) *lower bound*:

$$\liminf_{x\to\infty} \frac{1}{x} \log P\{L_0^i > x\} \ge -\theta^*_{ij}, \tag{15}$$

where, θ^*_{ij} is the unique solution of the equation:

$$\alpha_{A^i}(\theta) + \alpha_{E^j}(\theta) = mc, \quad i \ne j. \tag{16}$$

4. The proof of Theorem 3.1

In this section we prove Theorem 3.1. First we give some large deviations of the stationary queue length and the effective bandwidths of the stationary and transient departure processes from a *G/MSP/m* queueing system.

4.1. Large deviations of a *G/MSP/m* queueing system

Consider a *G/MSP/m* queueing system with the arrival process $\{A_t = A_t^2; t \in \mathbf{N}\}$ and the service process $\{B_t = B_t^2; t \in \mathbf{N}\}$. According to Loynes's Stability Theorem, this system is stable if $\mathcal{A} < \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are the mean arrival rate and the mean service rate, respectively. Let L_t be the queue length of the queueing system at time t. Then under the stable condition, L_t converges in distribution to a finite random variable L_∞. Here we assume that the queue process have reached its steady state. Thus L_0 has the same distribution as L_∞. Applying the large deviations of a single *G/G/* 1 queueing system (cf. [3], [5], [9], [13]) to the *G/MSP/m* queueing system considered here, we obtain the following large deviations.

Theorem 4.1: Under $\mathcal{A} < \mathcal{B}$, the tail of the distribution of the steady state queue length L_0 is characterized by

$$\lim_{x\to\infty} \frac{1}{x} \log P\{L_0 > x\} = -\delta^*, \tag{17}$$

where $\delta^* > 0$ is the largest solution of the equation:

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0. \tag{18}$$

In the case $\delta^* = \infty$, the equality (17) holds trivially. To avoid such a case, we assume that δ^* is finite in the sequel. Let E_t be the number of departures from the transient *G/MSP/m* queueing system at time t, and D_t the number of departures from the stationary *G/MSP/m* queueing system at time t. Under steady state, these departure processes are also stationary. E_t, D_t and their

partial sum process S_t^E, S_t^D are governed by the following recursive equations:

$$E_t = \min\{L_{t-1}+A_{t-1}, B_{t-1}\}, \quad S_t^E = \inf_{0\le\tau\le t}\{S_\tau^A + S_{\tau,t}^B\}, \tag{19}$$

$$D_t = \min\{L_{t-1}+A_{t-1}, B_{t-1}\}, \quad S_t^D = \min\{L_0 + \inf_{0<\tau\le t}\{S_\tau^A + S_{\tau,t}^B\},\ S_t^B\}. \tag{20}$$

Furthermore, we define the processes $\tilde{S}_t^E$ and $\tilde{S}_t^D$ as followe:

$$\tilde{S}_t^E = \min\{S_t^A, S_t^B\}, \quad \tilde{S}_t^D = \min\{L_0 + S_t^A, S_t^B\}, \quad t\in\mathbf{N}. \tag{21}$$

Then under the assumption that $\{A_t; t\in\mathbf{N}\}$ satisfy A1, A2 and A3, we obtain the following result according to Theorem 2 in Chang and Zajic [5].

Theorem 4.2: Under $\mathcal{A}<\mathcal{B}$, for any $\alpha\in\mathbf{R}$,

(i)

$$\lim_{t\to\infty}\frac{1}{t}\log P\{S_t^E>\alpha t\} = \lim_{t\to\infty}\frac{1}{t}\log P\{\tilde{S}_t^E>\alpha t\} = -\inf_{x\ge\alpha}\Lambda_E^*(x), \tag{22}$$

$$\lim_{t\to\infty}\frac{1}{t}\log E[e^{\theta S_t^E}] = \lim_{t\to\infty}\frac{1}{t}\log E[e^{\theta\tilde{S}_t^E}] = \Lambda_E(\theta), \qquad \theta\ge 0, \tag{23}$$

where,

$$\Lambda_E^*(\alpha) = \inf_{x\ge\alpha}\Lambda_A^*(x) + \inf_{x\ge\alpha}\Lambda_B^*(x) = \begin{cases} 0 & \text{if } \alpha<\mathcal{A} \\ \Lambda_A^*(\alpha) & \text{if } \alpha\le\mathcal{B} \\ \Lambda_A^*(\alpha)+\Lambda_B^*(\alpha) & \text{if } \mathcal{B}<\alpha\le\min\{a_r, mc\} \\ \infty & \text{if } \alpha>\min\{a_r, mc\}, \end{cases}$$

here, a_r is the right end point of $dom\,\Lambda_A^*$, and

$$\Lambda_E(\theta) = \sup_{\mathcal{A}\le\alpha}\{\theta\alpha-\Lambda_E^*(\alpha)\} = \sup_{\mathcal{A}\le\alpha\le\min\{a_r, mc\}}\{\theta\alpha-\Lambda_E^*(\alpha)\}. \tag{24}$$

(ii)

$$\lim_{t\to\infty}\frac{1}{t}\log P\{S_t^D>\alpha t\} = \lim_{t\to\infty}\frac{1}{t}\log P\{\tilde{S}_t^D>\alpha t\} = -\inf_{x\ge\alpha}\Lambda_D^*(x), \tag{25}$$

$$\lim_{t\to\infty}\frac{1}{t}\log E[e^{\theta S_t^D}] = \lim_{t\to\infty}\frac{1}{t}\log E[e^{\theta\tilde{S}_t^D}] = \Lambda_D(\theta), \qquad \theta\ge 0, \tag{26}$$

where,

$$\Lambda_D^*(\alpha) = \delta^*\alpha - \sup_{x\le\alpha}\{\delta^* x - \Lambda_A^*(x)\} = \inf_{x\ge\alpha}\Lambda_B^*(x)$$

$$= \begin{cases} 0 & \text{if } \alpha<\mathcal{A} \\ \Lambda_A^*(\alpha) & \text{if } \alpha\le\Lambda_{A,i}'(\delta^*) \text{ and } \mathcal{A}<\alpha\le\mathcal{B} \\ \Lambda_A^*(\alpha)+\Lambda_B^*(\alpha) & \text{if } \alpha\le\Lambda_A'(\delta^*) \text{ and } \mathcal{B}<\alpha\le\min\{a_r, mc\} \\ \delta^*\alpha-\Lambda_A(\delta^*) & \text{if } \alpha>\Lambda_A'(\delta^*) \text{ and } \mathcal{A}<\alpha\le\mathcal{B} \\ \delta^*\alpha-\Lambda_A(\delta^*)+\Lambda_B^*(\alpha) & \text{if } \alpha>\Lambda_A'(\delta^*) \text{ and } \mathcal{B}<\alpha\le\min\{a_r, mc\} \\ \infty & \text{if } \alpha>\min\{a_r, mc\}, \end{cases}$$

here, δ^* is the largest solution of the equation (18), and

$$\Lambda_D(\theta) = \sup_{\mathcal{A}\le\alpha}\{\theta\alpha-\Lambda_D^*(\alpha)\} = \sup_{\mathcal{A}\le\alpha\le\min\{a_r, mc\}}\{\theta\alpha-\Lambda_D^*(\alpha)\}. \tag{27}$$

From (24) and (27) we can directly calculate $\Lambda_D(\theta)$ and $\Lambda_E(\theta)$. Here we omit their long derivation because of the space limitation.

Theorem 4.3: Under $\mathcal{A}<\mathcal{B}$, the limit logarithmic moment generating functions $\Lambda_E(\theta)$ and $\Lambda_D(\theta)$ of the departure processes $\{E_t, t\in\mathbf{N}\}$ and $\{D_t, t\in\mathbf{N}\}$ are given by (11) and (12) for $i=2$.

The following conclusions are obtained from (24), (27) and the convexity of $\Lambda_D^*(\cdot)$ and $\Lambda_E^*(\cdot)$.

Corollary 4.4:

(i) For any $\alpha \in \mathbf{R}$, $\Lambda_D^*(\alpha) \le \Lambda_E^*(\alpha)$. In particular, $\Lambda_D^*(\alpha) = \Lambda_E^*(\alpha)$ if $\alpha \le \Lambda_A'(\delta^*)$.

(ii) For any $\theta \ge 0$, $\Lambda_D(\theta) \ge \Lambda_E(\theta)$. In particular, $\Lambda_D(\theta) = \Lambda_E(\theta)$ if $\theta \le \delta^*$.

4.2 Proof of the large deviation bounds

Having the previous preparation, we prove the large deviation bounds given in Theorem 3.1.The methodology is similar to one used in [27] and [17]. Without loss of generality, we establish the upper (13) and the lower (15) for Q_1, i.e,, for the case $i=1$, $j=2$. For convenience, we look backwards in time from $t=0$, and assume that the polling system have reached its steady state. Thus, L_0^i has the same distribution as L_∞^i.

1. Upper bound: Since the Bernoulli service schedule is work-conserving, we have that for any $t \ge 0$, $\min\{1, (L_{-t-1}^i + A_{-t-1}^i)/B_{-t-1}^i\}$ denotes the time length spent in Q_i by b_{-t-1}^i servers during the slot $(-t-1, -t]$. When $L_{-t-1}^i + A_{-t-1}^i \ge B_{-t-1}^i$, the time length is 1, which means that all b_{-t-1}^i servers have been serving Q_i during the slot $(-t-1, -t]$, and Q_j $(j \ne i)$ receives B_{-t-1}^j service amounts at most. When $L_{-t-1}^i + A_{-t-1}^i < B_{-t-1}^i$, the time length becomes $(L_{-t-1}^i + A_{-t-1}^i)/B_{-t-1}^i < 1$, and during the remained time $1-(L_{-t-1}^i + A_{-t-1}^i)/B_{-t-1}^i$, all m servers serve $Q^j (j \ne i)$. Hence, Q^j receives actually $B_{-t-1}^j (L_{-t-1}^i + A_{-t-1}^i)/B_{-t-1}^i + mc(1-(L_{-t-1}^i + A_{-t-1}^i)/B_{-t-1}^i) = B_{-t-1}^j (L_{-t-1}^i + A_{-t-1}^i)/B_{-t-1}^i + (B_{-t-1}^j + B_{-t-1}^i)(1-(L_{-t-1}^i + A_{-t-1}^i)/B_{-t-1}^i) = mc - (L_{-t-1}^i + A_{-t-1}^i)$ amount of service. Basing on the above observation, the dynamics of the polling system can be expressed by the following recursions:

$$L_{-t}^1 = \max\{L_{-t-1}^1 + A_{-t-1}^1 - \max\{B_{-t-1}^1,\ mc-(L_{-t-1}^2 + A_{-t-1}^2)\},\ 0\}, \tag{28}$$

$$L_{-t}^2 = \max\{L_{-t-1}^2 + A_{-t-1}^2 - \max\{B_{-t-1}^2,\ mc-(L_{-t-1}^1 + A_{-t-1}^1)\},\ 0\}. \tag{29}$$

Define $R_{-t}^i = \max\{B_{-t}^i,\ mc-(L_{-t}^j + A_{-t}^j)\}$, $i, j=1,2$; $i \ne j$. Then, R_{-t}^i denotes the amount of service actually received by Q_i during the slot $(-t-1, -t]$. Expanding (28) and (29) recursively we have

$$L_0^i = \max_{t \in \mathbf{N}} \{S_{-t}^{A^i} - S_{-t}^{R^i}\}, \quad i=1, 2, \tag{30}$$

where, $S_{-t}^{R^i} = \sum_{\tau=-t}^{-1} R_\tau^i$ is the total amount of service actually received by Q_i in $[-t, 0)$. Observing that

$$S_{-t}^{R^i} = L_{-t}^i + S_{-t}^{A^i} - L_0^i. \quad i=1, 2, \tag{31}$$

we have $S_{-t}^{R^i} \ge S_{-t}^{A^i} - L_0^i$. Therefore, the maximum in (30) for $i=1$ must be achieved at the time when $L_t^1 = 0$. Let $-t \le 0$ be the first time such that $L_{-t}^1 = 0$ and $L_{-\tau}^1 > 0$ for $\tau \in (0, t)$. Since the queue Q_1 is busy during the interval $(-t, 0]$ and the Bernoulli service schedule is a work-conserving policy, the queue Q_1 gets at least $S_{-t}^{B^1}$ amount of service (by considering the situation that the queue Q_2 may become empty during $[-t, 0)$). Thus, $S_{-t}^{R^1} \ge S_{-t}^{B^1}$. On the other hand, since mct is the total amount of service devoted by m servers during $[-t, 0)$, it always holds that

$$S_{-t}^{R^1} + S_{-t}^{R^2} = mct. \tag{32}$$

Also from the definition of $\{B_t^i, t \in \mathbf{N}\}$, we have that for any $t \ge 0$,

$$S_{-t}^{B^i} = mct - S_{-t}^{B^j}, \quad i, j = 1, 2; \quad i \ne j \tag{33}$$

Hence,

$$S_{-t}^{R^1} = mct - S_{-t}^{R^2} = \max\{mct - S_{-t}^{R^2},\ S_{-t}^{B^1}\} = \max\{mct - S_{-t}^{R^2},\ S_{-t}^{B^1}\} = mct - \min\{S_{-t}^{R^2},\ S_{-t}^{B^2}\}. \tag{34}$$

Substituting (34) into (30) for $i=1$ yields

$$L_0^1 = \max_{t \in \mathbf{N}} \{S_{-t}^{A^1} + \min\{S_{-t}^{R^2},\ S_{-t}^{B^2}\} - mct\}. \tag{35}$$

*CASE*1: $\mathcal{A}^2 < \mathcal{B}^2$

In this case we can upper bound L_t^2 by the queue length at a virtual system. Consider a *G/MSP/m* queueing system with the arrival process $\{A_t^2, t \in \mathbf{N}\}$ and the service process $\{B_t^2, t \in \mathbf{N}\}$. Let D_{-t}^2 be the number of departures, and $\tilde{L}_{-t}^2$ the queue length at time $-t$ of the virtual system, respectively. Since this virtual system does not receive extra service except $S_{-t}^{B^2}$, it always holds that $L_{-t}^2 \le \tilde{L}_{-t}^2$. We have by (31)

$$S_{-t}^{R^2} \le L_{-t}^2 + S_{-t}^{A^2} \le \tilde{L}_{-t}^2 + S_{-t}^{A^2}. \tag{36}$$

Combining (35) with (36) yields

$$L_0^1 \le \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + \min \{ \tilde{L}_{-t} + S_{-t}^{A^2},\ S_{-t}^{B^2} \} - mct \} = \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + \tilde{S}_{-t}^{D^2} - mct \}, \tag{37}$$

where $\tilde{S}_{-t}^{D^2} = \min \{ \tilde{L}_{-t}^2 + S_{-t}^{A^2},\ S_{-t}^{B^2} \}$. From the fact that $\{A_t^2, t \in \mathbf{N}\}$ satisfies Assumption A2 and Theorem 4.2, it follows that for any $\theta > 0$,

$$\lim_{t \to \infty} \frac{1}{t} \log E[e^{\theta S_{-t}^{A^1}}] = \Lambda_{A^1}(\theta), \quad \text{and} \quad \lim_{t \to \infty} \frac{1}{t} \log E[e^{\theta \tilde{S}_{-t}^{D^2}}] = \Lambda_{D^2}(\theta). \tag{38}$$

Then, for any $\varepsilon > 0$, there exists a sufficiently large t_ε such that for any $t \ge t_\varepsilon$

$$E[e^{\theta S_{-t}^{A^1}}] \le e^{(\Lambda_{A^1}(\theta) + \varepsilon)t} \quad \text{and} \quad E[e^{\theta \tilde{S}_{-t}^{D^2}}] \le e^{(\Lambda_{D^2}(\theta) + \varepsilon)t}. \tag{39}$$

We have

$$E[\, e^{\theta L_0^1}] \le E[e^{\theta \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + v_1 S_{-t}^{M^2} - mc_1 t \}}] \le \sum_{t \in \mathbf{N}} E[\, e^{\theta (S_{-t}^{A^1} + \tilde{S}_{-t}^{D^2} - mct)}] \tag{40}$$

$$= \sum_{t \in \mathbf{N}} E[e^{\theta S_{-t}^{A^2}}] E[e^{\theta \tilde{S}_{-t}^{D^2}}] e^{-\theta mct} \le C_\varepsilon + \sum_{t \ge t_\varepsilon} e^{(\Lambda_{A^1}(\theta) + \Lambda_{D^2}(v_1 \theta) + 2\varepsilon - mc\theta)t},$$

where the last second equality follows from the independence of $S_{-t}^{A^2}$ and $\tilde{S}_{-t}^{D^2}$, and C_ε is a constant dependent on ε.

Therefore, we have $E[e^{\theta L_0^1}] < \infty$ if $\Lambda_{A^1}(\theta) + \Lambda_{D^2}(\theta) + 2\varepsilon - mc\theta < 0$. By Chebyshev's inequality, $\mathrm{P}\{L_0^1 > x\} \le e^{-\theta x} E[\, e^{\theta L_0^1}]$ for any $x \ge 0$, and the definitions of $\alpha_{A^1}(\theta)$ and $\alpha_{D^2}(\theta)$, we have if $\alpha_{A^1}(\theta) + \alpha_{D^2}(\theta) + 2\varepsilon / \theta - mc < 0$,

$$\limsup_{x \to \infty} \frac{1}{x} P\{ L_0^1 > x \} \le -\theta. \tag{41}$$

Taking $\varepsilon \to 0$ and getting the tightest upper bound, we establish the upper (13).

CASE2: $\mathcal{A}^2 \ge \mathcal{B}^2$

By (35), we have

$$L_0^1 \le \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + S_{-t}^{B^2} - mct \}. \tag{42}$$

Similarly, for any $\theta > 0$, if $\Lambda_{A^1}(\theta) + \Lambda_{B^2}(\theta) + 2\varepsilon - mc\theta < 0$, then,

$$E[\, e^{\theta L_0^1}] \le \sum_{t \in \mathbf{N}} E[\, e^{\theta (S_{-t}^{A^2} + S_{-t}^{B^2} - mct)}] < \infty. \tag{43}$$

Again by Chebyshev's inequality, if $\alpha_{A^1}(\theta) + \alpha_{B^2}(\theta) + 2\varepsilon / \theta - mc < 0$,

$$\limsup_{x \to \infty} \frac{1}{x} P\{ L_0^1 > x \} \le -\theta. \tag{44}$$

Taking $\varepsilon \to 0$ and getting the tightest upper bound (note that $\alpha_{D^2}(\theta) = \Lambda_{B^2}(\theta) / \theta$ in this case), we establish the upper (13).

The next lemma will be used repeatedly in deriving the lower bound.

Lemma 4.5 For $\Lambda_X^*(\cdot)$ and $\Lambda_X(\cdot)$ being convex conjugate, it holds

$$\inf_{\alpha > c} \frac{\Lambda_X^*(\alpha)}{\alpha - c} = \theta^* \tag{45}$$

where θ^* is the unique positive root of the equation $\Lambda_X(\theta) = c\theta$, and $c > E[X]$ is a constant.

2. Lower bound: In section 2, we have defined the aggregate queue length of the two queues as $L_t = L_t^1 + L_t^2$. The aggregate service process for L_t is $R_t = R_t^1 + R_t^2 = mc$. Hence,

$$L_0 = \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + S_{-t}^{A^2} - mct \} \tag{46}$$

Similarly, the maximum must be achieved at the time when $L_{-t} = 0$. Let $-t^* \le 0$ be the first time such that $L_{-t^*} = 0$ (which implies $L_{-t^*}^1 = L_{-t^*}^2 = 0$) and $L_{-t} > 0$ for $t \in (0, t^*)$. Further, we have a similar expression for the queue length L_0^2, namely,

$$L_0^2 = \max_{t \in \mathbf{N}} \{ S_{-t}^{A^2} + \min \{ S_{-t}^{R^1}, S_{-t}^{B^1} \} - mct \}. \tag{47}$$

Also, this maximum must be achieved at the time when $L_{-t}^2 = 0$. Let $-\tau^* \le 0$ be the first time such tnat $L_{-\tau^*} = 0$ and $L_{-t}^2 > 0$ for $t \in (0, \tau^*)$. Then we have $\tau^* \le t^*$. Utilizing this fact and the relations (46) and (47), we have

$$\begin{aligned} L_0^1 &= L_0 - L_0^2 = \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + S_{-t}^{A^2} - mct \} - \max_{t \in \mathbf{N}} \{ S_{-t}^{A^2} + \min \{ S_{-t}^{R^1}, S_{-t}^{B^1} \} - mct \} \\ &= \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + S_{-t}^{A^2} - mct - \max_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + \min \{ S_{-\tau}^{R^1}, S_{-\tau}^{B^1} \} - mc\tau \} \} \\ &\ge \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + S_{-t}^{A^2} - mct - \max_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + S_{-\tau}^{B^1} \} - mc\tau \} \} = \max_{t \in \mathbf{N}} \{ S_{-t}^{A^1} + \min_{0 \le \tau \le t} \{ S_{-t,-\tau}^{A^2} + S_{-\tau}^{B^2} \} - mct \}. \end{aligned} \tag{48}$$

For any $x \ge 0$, let $t = \lceil x/\beta \rceil$, where $\beta > 0$ is an arbitrary constant. We have

$$\begin{aligned} &\liminf_{x \to \infty} \frac{1}{x} \log P\{ L_0^1 > x \} \ge \frac{1}{\beta} \liminf_{t \to \infty} \frac{1}{t} \log P\{ L_0^1 > \beta t \} \\ &\ge \frac{1}{\beta} \liminf_{t \to \infty} \frac{1}{t} \log P \{ S_{-t}^{A^1} + \min_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + S_{-t,-\tau}^{B^2} \} - mct > \beta t \} \\ &= \frac{1}{\beta} \liminf_{t \to \infty} \frac{1}{t} \log P \{ \frac{S_{-t}^{A^1}}{t} + \frac{\min_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + S_{-t,-\tau}^{B^2} \}}{t} > mc + \beta \} \end{aligned} \tag{49}$$

*CASE*1: $\mathcal{A}^2 \ge \mathcal{B}^2$ Since $\mathcal{B}^1 + \mathcal{B}^2 = mc$,

$$\begin{aligned} &P\{ \frac{S_{-t}^{A^1}}{t} + \frac{\min_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + S_{-t,-\tau}^{B^2} \}}{t} > mc + \beta \} \\ &= P\{ \frac{S_{-t}^{A^1}}{t} + \frac{\min_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + S_{-t,-r}^{B^2} \}}{t} > \mathcal{B}^1 + \mathcal{B}^2 + \beta \} \\ &\ge P\{ \frac{S_{-t}^{A^1}}{t} > \mathcal{B}^1 + \beta, \frac{\min_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + S_{-t,-\tau}^{B^2} \}}{t} > \mathcal{B}^2 \} \\ &= P\{ \frac{S_{-t}^{A^1}}{t} > \mathcal{B}^1 + \beta \} P\{ \frac{\min_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + S_{-t,-\tau}^{B^2} \}}{t} > \mathcal{B}^2 \}, \end{aligned}$$

where the last equality follows from the independence of $\{ A_{-t}^1; t \in \mathbf{N} \}$, $\{ A_{-t}^2; t \in \mathbf{N} \}$ and $\{ B_{-t}^2; t \in \mathbf{N} \}$. Then,

$$\begin{aligned} &\liminf_{x \to \infty} \frac{1}{x} \log P\{ L_0^1 > x \} \\ &\ge \frac{1}{\beta} \Big(\liminf_{t \to \infty} \frac{1}{t} \log P\{ \frac{S_{-t}^{A^1}}{t} > \mathcal{B}^1 + \beta \} + \liminf_{t \to \infty} \frac{1}{t} \log P\{ \frac{\min_{0 \le \tau \le t} \{ S_{-\tau}^{A^2} + S_{-t,-\tau}^{B^2} \}}{t} > \mathcal{B}^2 \} \Big) \\ &\ge -\frac{1}{\beta} \inf_{\alpha \ge \mathcal{B}^1 + \beta} \Lambda_{A^1}^*(\alpha) - \frac{1}{\beta} \inf_{\alpha \ge \mathcal{B}^2} \Lambda_{E^2}^*(\alpha) = -\frac{1}{\beta} \inf_{\alpha \ge \mathcal{B}^1 + \beta} \Lambda_{A^1}^*(\alpha), \end{aligned} \tag{50}$$

where, the last equality follows from the fact that $\inf_{\alpha \ge \mathcal{B}^2} \Lambda_{E^2}^*(\alpha) = \Lambda_{E^2}^*(\mathcal{B}^2) = 0$ if $\mathcal{A}^2 \ge \mathcal{B}^2$. As β is arbitrary we have

$$\begin{aligned} &\liminf_{x \to \infty} \frac{1}{x} \log P\{ L_0^1 > x \} \ge -\inf_{\beta \ge 0} \inf_{\alpha \ge \mathcal{B}^1 + \beta} \left\{ \frac{\Lambda_{A^1}^*(\alpha)}{\beta} \right\} = -\inf_{\alpha \ge \mathcal{B}^1} \inf_{\alpha - \mathcal{B}^1 > \beta} \left\{ \frac{\Lambda_{A^1}^*(\alpha)}{\beta} \right\} \\ &= -\inf_{\alpha > \mathcal{B}^1} \left\{ \frac{\Lambda_{A^1}^*(\alpha)}{\alpha - \mathcal{B}^1} \right\} = -\theta_{12}^*, \end{aligned} \tag{51}$$

where, the last second equality follows from that $1/x$ is a continuous decreasing function for $x > 0$, and the last equality follows

from Lemma 4.5, here θ_{12}^* is the unique positive solution of the equation: $\Lambda_{A^1}(\theta)=\mathcal{B}^1\theta$. However, by the definition (11), $\Lambda_{E^2}(\theta)=\mathcal{B}^2\theta$ in the case $\mathcal{A}^2\geq\mathcal{B}^2$. We have $mc\theta-\Lambda_{E^2}(\theta)=mc\theta-\mathcal{B}^2\theta=\mathcal{B}^1\theta$. Hence, θ_{12}^* is actually the unique solution of the equation $\Lambda_{A^1}(\theta)+\Lambda_{E^2}(\theta)=mc\theta$, which is the equation (16).

CASE2: $\mathcal{A}^2<\mathcal{B}^2$　Let $\alpha_i\geq\mathcal{A}^i$, $i=1,2$ and $\alpha_1+\alpha_2>mc+\beta$. Then,

$$P\{\frac{S_{-t}^{A^1}}{t}+\frac{\min_{0\leq\tau\leq t}\{S_{-\tau}^{A^2}+S_{-t,-\tau}^{B^2}\}}{t}>mc+\beta\}\geq P\{\frac{S_{-t}^{A^1}}{t}>\alpha_1,\frac{\min_{0\leq\tau\leq t}\{S_{-\tau}^{A^2}+S_{-t,-\tau}^{B^2}\}}{t}>\alpha_2\}$$

$$=P\{\frac{S_{-t}^{A^1}}{t}>\alpha_1\}P\{\frac{\min_{0\leq\tau\leq t}\{S_{-\tau}^{A^2}+S_{-t,-\tau}^{B^2}\}}{t}>\alpha_2\}.$$

We have

$$\begin{aligned}&\liminf_{x\to\infty}\frac{1}{x}\log P\{L_0^1>x\}\\ &\geq\frac{1}{\beta}\left(\liminf_{t\to\infty}\frac{1}{t}\log P\{\frac{S_{-t}^{A^1}}{t}>\alpha_1\}+\liminf_{t\to\infty}\frac{1}{t}\log P\{\frac{\min_{0\leq\tau\leq t}\{S_{-\tau}^{A^2}+S_{-t,-\tau}^{B^2}\}}{t}>\alpha_2\}\right)\\ &\geq-\frac{1}{\beta}\left(\inf_{x\geq\alpha^1}\Lambda_{A^1}^*(x)+\inf_{x\geq\alpha^2}\Lambda_{E^2}^*(x)\right)=-\frac{1}{\beta}(\Lambda_{A^1}^*(\alpha_1)+\Lambda_{E^2}^*(\alpha_2)),\end{aligned}\tag{52}$$

where, the last equality follows from the increasing properties of $\Lambda_{A^2}^*(\cdot)$. As β is arbitrary we have

$$\begin{aligned}\liminf_{x\to\infty}\frac{1}{x}P\{L_0^1>x\}&\geq-\inf_{\beta>0}\frac{1}{\beta}\inf_{\{\alpha_i\in\mathbf{R},\alpha_i\geq\mathcal{A}^i,i=1,2;\alpha_1+\alpha_2>mc+\beta\}}\{\Lambda_{A^1}^*(\alpha_1)+\Lambda_{E^2}^*(\alpha_2)\}\\ &=-\inf_{\{\alpha_i\in\mathbf{R},\alpha_i\geq\mathcal{A}^i,i=1,2;\alpha_1+\alpha_2>mc\}}\inf_{\alpha_1+\alpha_2-mc>\beta}\left\{\frac{\Lambda_{A^1}^*(\alpha_1)+\Lambda_{E^2}^*(\alpha_2)}{\beta}\right\}\\ &=-\inf_{\{\alpha_i\in\mathbf{R},\alpha_i\geq\mathcal{A}^i,i=1,2;\alpha_1+\alpha_2>mc\}}\left\{\frac{\Lambda_{A^1}^*(\alpha_1)+\Lambda_{E^2}^*(\alpha_2)}{\alpha_1+\alpha_2-mc}\right\}\\ &=-\inf_{\alpha>mc}\left\{\frac{I^*(\alpha)}{\alpha-mc}\right\}=-\theta_{12}^*,\end{aligned}$$

where,

$$I^*(\alpha)\equiv-\inf_{\{\alpha_i\in\mathbf{R},\alpha_i\geq\mathcal{A}^i,i=1,2;\alpha_1+\alpha_2>mc\}}\{\Lambda_{A^1}^*(\alpha_1)+\Lambda_{E^2}^*(\alpha_2)\},\tag{53}$$

and θ_{12}^* is the unique solution of the equation $\Lambda_{A^1}(\theta)+\Lambda_{E^2}(\theta)=mc\theta$. Let $I(\theta)=\sup_{\alpha\in\mathbf{R}}\{\theta\alpha-I^*(\alpha)\}$. By Lemma 4.5, if we can prove that $I(\theta)=\Lambda_{A^2}(\theta)+\Lambda_{E^2}(\theta)$ and $I'(0)<mc$, then the last equality is obtained. First, we have

$$\begin{aligned}I(\theta)&=\sup_{\alpha\in\mathbf{R}}\{\theta\alpha-\inf_{\{\alpha_i\in\mathbf{R},\alpha_i\geq\mathcal{A}^i,i=1,2;\alpha_1+\alpha_2>mc\}}\{\Lambda_{A^1}^*(\alpha_1)+\Lambda_{E^2}^*(\alpha_2)\}\}\\ &=\sup_{\alpha\in\mathbf{R}}\sup_{\{\alpha_i\in\mathbf{R},\alpha_i\geq\mathcal{A}^i,i=1,2;\alpha_1+\alpha_2>mc\}}\{\theta\alpha-\Lambda_{A^1}^*(\alpha_1)-\Lambda_{E^2}^*(\alpha_2)\}\\ &=\sup_{\alpha_1\in\mathbf{R},\alpha_2\in\mathbf{R}}\{\theta\alpha_1+\theta\alpha_2-\Lambda_{A^1}^*(\alpha_1)-\Lambda_{E^2}^*(\alpha_2)\}\\ &=\sup_{\alpha_1\in\mathbf{R},\alpha_2\in\mathbf{R}}\{(\theta\alpha_1-\Lambda_{A^1}^*(\alpha_1))+(\theta\alpha_2-\Lambda_{E^2}^*(\alpha_2))\}\\ &=\sup_{\alpha_1\in\mathbf{R}}\{\theta\alpha_1-\Lambda_{A^1}^*(\alpha_1)\}+\sup_{\alpha_2\in\mathbf{R}}\{\theta\alpha_2-\Lambda_{E^2}^*(\alpha_2)\}=\Lambda_{A^1}(\theta)+\Lambda_{E^2}(\theta).\end{aligned}\tag{54}$$

Secondly, since $\mathcal{A}^2<\mathcal{B}^2$ and the stability condition (9), we have

$$I'(\theta)|_{\theta=0}=(\Lambda_{A^1}'(\theta)+\Lambda_{E^2}'(\upsilon_1\theta))|_{\theta=0}=\mathcal{A}^1+\mathcal{A}^2<\mathcal{B}^1+\mathcal{B}^2=mc.$$

Hence, we obtained the lower bound (15) when $\mathcal{A}^2<\mathcal{B}^2$.

5. Conclusion

In this paper we have analyzed a discrete-time polling system consisting of two-parallel queues and m servers. The arrival process in each queue is an arbitrary, and possibly correlated stochastic process, and each server serves independently the two queues according to the Bernoulli service schedule. For each queue, we derived the upper and lower bounds of the buffer overflow probability by using the large deviation techniques and the effective bandwidth concept. The results can be used in traffic management of high-speed communication networks such as call admission, server allocation, and congestion control. As have been seen, however, the upper and lower bounds obtained here in general do not match exactly. The main reason is that when each server allocates its service capacity to the two queues randomly, a large deviation in the departure process may be encouraged (see [5]). Therefore, for the polling system with general arrival processes, developing an approach to get matched upper and lower bounds of the overflow probability still is an opening problem for future investigation.

References

[1] D. Bertsimas, I.C. Paschalidis and J.N. Tsitsiklis, Asympotic buffer overflow probabilities in Multiclass multiplexers: An optimal control approach, IEEE Transactions on Automatic Control 43(3) (1998) 315-335.

[2] D. Bertsimas, I.C. Paschalidis and J.N. Tsitsiklis, Lerge deviations analysis of the generalized processor sharing policy, Queueing Systems 32(1999) 319-349.

[3] D.D. Botvich and N.G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexers, Queueing Systems 20 (1995) 293-320.

[4] O.J. Boxma and J.A. Weststrate, Waiting times in polling systems with Markovian server routing, preprint 1989.

[5] C.S. Chang and T. Zajic, Effective bandwidths of departure processes from queues with time varying capacities, in: *Proc. IEEE INFOCOM'95* (1995) pp. 1001-1009.

[6] H. Chung, C.K. Un and W.Y. Jung, Performance analysis of Markovian polling system with single buffer, Performance Evaluation 19 (1994), 303-315.

[7] F. Delcoigne and A. de La Fortelle, Large deviation rate function for polling systems, Queueing System 41 (2002) 12-44.

[8] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications.* Jones and Bartlett Publishers, 1993.

[9] N.G. Duffeld, Exponential bounds for queues with Markovian arrivals, Queueing Systems 17 (1994) 413-430.

[10] M. Eisenberg, Two queues with alternating service, SIAM Journal on Applied Mathematics 36 (1979) 287-303.

[11] W. Feng, M. Kowada and K. Adachi, A two-queue model with Bernoulli service schedule and switching times, Queueing Systems 30(1998) 405-434.

[12] W. Feng, K. Adachi and M. Kowada, Large deviation approximation for a polling system with Markovian on/off sources and Bernoulli service schedule, submitted to publication.

[13] P.W. Glynn and W. Whitt, Logarthmic asymptoics for steady-state tail probabilities in a single-server queue, Journal of Applied Probability 31(1994) 131-156.

[14] T. Hirayama, Mean waiting times in Markovian polling systems, Proceedings of Symposium on Queueing Theory and its Application, (2002) 58-65.

[15] R. M. Loynes, The stability of a queue with non-independent inter-arrival and service times, Proc. Cambridge Philos. Soc. 58 (1962) 497-520.

[16] D.S. Lee, Analysis of a cyclic server queue with Bernoulli schedules, Journal of Applied Prob-ability 34 (1997) 176-191.

[17] M.A. Marsan, S. Donatelli and F. Neri, GSPN models of Markovian multiserver multiqueue system, Performance Evaluation 11 (1990) 227-240.

[18] L. Massoulie, Large deviations estimates for polling and weighted fair queueing service systems (1999) Advances in Performance Analysis 2 (1999) 103-128.

[19] M.M. Srinivasan, Nondeterministic polling systeme, Management Science 37 (1991), 667-681.

[20] H. Takagi, *Analysis of Polling Systems,* MIT Press Cambridge, Massachusetts, 1986.

[21] H. Takagi, Queuing analysis of polling models, ACM Comput. Surveys 20 (1988) 5-28.

[22] R.D. van der Mei and S.C. Borst, Analysis of multiple-server polling systems by means of the power-series algorithm, Commun. Statist.-Stochastic Models, 13(2) (1997) 339-369, 1997.

[23] A. Weiss, An introduction to large deviations for communication networks, IEEE Journal on Selected Areas in Communications 13(6) (1993) 938-385.

[24] Z.-L. Zhang, Large deviations and the generalized processor sharing schedule for a two-queue system, Queueing Systems 26 (1997) 229-254.