Probability in the Engineering and Informational Sciences, **23**, 2009, 61–74. Printed in the U.S.A.
doi:10.1017/S0269964809000060

PERFORMANCE ESTIMATION OF $M/D_K/1$ QUEUE UNDER FAIR SOJOURN PROTOCOL IN HEAVY TRAFFIC

YINGDONG LU
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
E-mail: yingdong@us.ibm.com

We study the performance of a $M/D_K/1$ queue under Fair Sojourn Protocol (FSP). We use a Markov process with mixed real- and measure-valued states to characterize the queuing process of system and its related processor sharing queue. The infinitesimal generator of the Markov process is derived. Classifying customers according to their service time, using techniques in multiclass queuing system, and borrowing recently developed heavy traffic results for processor-sharing queues, we are able to derive approximations for average waiting time for the jobs.

1. INTRODUCTION

Fair Sojourn Protocol (FSP) was proposed by Friedman and Henderson [4] for the purpose of achieving both fairness and efficiency in a single-server queuing system. The main idea of the policy is to take advantage of both processor sharing (PS) for fairness and first come first server (FCFS) for efficiency. More specifically, the policy enforces the departures to follow the order of the departures of the same queue (a queue with almost surely the same sample path of arrivals) under the PS policy, whereas the server serves one job at a time. The performance advantage of the FSP policy was established in [4], more specifically, it was shown that FSP *dominates* PS policy; that is, the jobs in a FSP system always depart earlier than those in a PS system with the same arrival sample path, and no other policy *dominates* FSP. It is understood that quantitative analysis of queuing performance under FSP, such as the calculation of

© 2009 Cambridge University Press 0269-9648/09 $25.00
https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

average waiting time, is very hard to obtain. The difficulty lies at the fact that the performance depends on the status of a PS queue with the same arrival sample path, which, from now on, will be called the associated PS queue. The purpose of this study is to use some recently developed heavy traffic approximation results in a PS queue to approximate the average waiting time for jobs in a FSP system in heavy traffic (i.e., when the arrival rate is close to the service rate).

Another main idea of our study is to treat a single queue under FSP as a multiclass queuing system. We observed that when the service time is constant for all jobs, the FSP is equivalent to a FCFS queue. In a queue with general service time, for any two jobs with the same service time, the order of departures under PS, hence FSP, is the same as the order of the arrivals. Based on these observations, we put jobs that have the same processing time in the same class, and within each class, the jobs will follow FCFS. For the convenience of our analysis, we also assume that our service distribution has finite sample space. For most applications, this should not be a problem because we can have the sample space large enough to cover all of the different choices of service time. From a theoretical standpoint, it is known that any distribution can be approximated within any given degree of accuracy by a distribution with finite sample space; since our study also falls into the category of approximation, this assumption will not become a major restriction.

Multiclass queues and queuing networks, as a family of important mathematical models for manufacturing and communication networks, have been intensively studied in the past several decades. Much progress have been made in the area of performance analysis and optimal control. Among them, approximation schemes, mainly including fluid and diffusion approximations, have been considerably enriched and expanded. Efficient approximations are established under different scheduling and routing policies and applied successfully to the study of stability and control. The works mostly related to ours are that of the fluid and diffusion approximations of general PS queues. In a series of articles, [**?**,6,7] fluid and diffusion approximation for PS queues were obtained through the approximation of the residual service time process, a measure-valued process. To be more specific, a measure-valued process is derived as the fluid limit of the residual service time process under proper scaling; furthermore, this fluid limit has a stationary point as time goes to infinity. It is also shown that, asymptotically, the residual service process can be approximated by applying a lifted map upon the workload process, whose limit is the same as any work-conserving queue and has been known previously. The last two points will play crucial roles in our analysis. Related results can also be found in the work of Grishechkin, such as [5].

To facilitate our analysis, we assume that the arrival process is a Poisson process. We demonstrate that the mixed process of the queue length, residual service time for the FSP system, and residual service time for the PS system is a mixed-valued Markov process. We are able to obtain the probabilistic characterization of this process (i.e., its infinitesimal generator). The expression of the average sojourn time in the form of functionals of the mixed-value Markov process will be derived with the aid of the tagged job arguments used in Kleinrock [9] The property of PASTA (Poisson arrival see time average) allows us to replace the observations of each arrival with the

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

stationary quantities, which are then approximated by the fluid and diffusion limits we discussed earlier.

The rest of the article is organized as the follows, in Section 2, we give detailed description of FSP and some basic facts, then we will derive probabilistic characterization of the state process under Poisson arrival; in Section 3, we develop approximations for the average waiting time for FSP.

2. PROBABILISTIC CHARACTERIZATIONS OF THE SYSTEM

In this section we will first present detailed descriptions of our mathematical model—in particular, a mixed (measure and real)-valued process that characterizes the dynamics of the queuing process under FSP and its associated PS queue. Then we will derive basic probabilistic characterization of this mixed (measure and real)-valued process. Since the arrivals follow a Poisson process, it is evident that the joined process of the queue length and residual service time of the FSP queue and the residual of the associated PS queue is Markovian, so the probabilistic characterization will be realized through the computation of the infinitesimal generator.

The arrival process follows a Poisson process with rate λ. The required service times are mutually independent, independent of the arrival process, and identically distributed random variables in the form of D_K, (i.e., a discrete random variable with a finite sample space of size K). Let us denote the sample space by $\mathcal{S} = \{\ell_1, \ell_2, \ldots, \ell_K\}$. Without loss of generality, we assume that $\ell_1 < \ell_2 < \cdots < \ell_K$. Let us also denote $p_1, p_2, \ldots, p_K$ as the probability associated with each value (i.e., $\mathsf{P}[D_K = \ell_k] = p_k$, $k = 1, 2, \ldots, K$, so the overall service rate $\mu = 1/\sum p_i \ell_i$). To ensure that we have a stable queue with a heavy traffic condition, assume that $\lambda/\mu < 1$ and $\lambda/\mu \approx 1$.

Under the FSP rule, whenever a job is completed or departs, the system needs to rank all of the remaining jobs in the queue according to their departure time in the associated PS queue; then the one with the earliest departure time will be served. Since the event of two jobs having the same remaining service time in the PS queue has negligible probability, it will be ignored in the analysis. In practice, arbitrary tier break rules can be applied, and they will not affect any of the results in this article. From the viewpoint of multiclass queuing system, we can treat this queuing system as a queue with K classes of arrivals, each class of arrivals follow an independent Poisson with rate $p_k\lambda$ for $k = 1, 2, \ldots, K$, and jobs in each class k require deterministic service time ℓ_k.

We define a mixed-valued Markov process to characterize the joined process of the FSP queue and its associated PS queue. By a mixed-valued Markov process, we refer to a Markov process that takes a value in a product of $\mathbb{R}$ and $\mathcal{M}_F$, where $\mathcal{M}_F$ denotes the space of finite, nonnegative Borel measures equipped with the topology of weak convergence of measures (see, e.g., [10]). Let $q_k(t), k = 1, 2, \ldots, K$ be the number of jobs of class k that are waiting in the FSP queue and let $r(t)$ denote the residual service time process of the job that is being processed in the FSP queue at time t and $\mu_k(t)$ denote the measure-valued process of the residual service time for the class k jobs in the associated PS queue. More specifically, at any time t, for any

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

Borel set B on $\mathbb{R}_+$, $\mu_k(t)(B)$ equals to the number of class k jobs in the system that has the residual processing times that are in B. From the definition of FSP and its associated PS queue, we can conclude the following.

THEOREM 1: *The process* $(q_1(t), q_2(t), \ldots, q_K(t), r(t), \mu_1(t), \mu_2(t), \ldots, \mu_K(t))$ *in* $\mathbb{Z}_+^K \times \mathbb{R}_+ \times \mathcal{M}_F^K$ *is a Markov process.*

Next, we derive the infinitesimal generator of this Markov process. To do this, we need to introduce some fundamental functional on the space $\mathcal{M}_F$. For any bounded continuous function $f(\cdot)$ on $\mathbb{R}_+$ and any measure $\sigma \in \mathcal{M}_F$, let us denote $<f, \sigma>= \int_{\mathbb{R}_+} f(x)\, d\sigma$. It is easy to see that $Q^{\mathrm{PS}}(t), t \geq 0$, the queue length process for the associated PS queue, can be expressed $Q^{PS}(t) = <\mathbf{1}, \mu_t>$, where $\mathbf{1}$ denotes the function of $f(x) = 1$, and the work load process $W(t) = <\chi, \mu_t>$, where $\chi(x) = x$. For any bounded continuous function $f(\mathbf{x}, y, \mathbf{z})$ defined on $\mathbb{Z}_+^K \times \mathbb{R}_+ \times \mathcal{M}_F^K$, define

$$<f, (\mathbf{q}, r, \mu)> = \int_{\mathbb{R}_+} f(\mathbf{q}, r, \mathbf{z})\, d\mu,$$

For our mixed-valued process, similar to the measure-valued Markov process (see, e.g., [2,8]), the transition probability operator T_t is defined by

$$T_t <f, (\mathbf{q}, r, \mu)> \triangleq \mathsf{E}[<f, X(t)>|X(0) = (\mathbf{q}, r, \mu)];$$

then the infinitesimal generator for the semigroup of T_t, hence the Markov process, is defined by

$$\mathcal{G} <f, (\mathbf{q}, r, \mu)> \triangleq \frac{d}{dt} T_t <f, (\mathbf{q}, r, \mu)>|_{t=0}.$$

We know that for Δt sufficient small, the probability that there is an arrival of type k is $\lambda_k \Delta t + o(\Delta t)$. In that event, the queue length will become $\mathbf{q} + \mathbf{e}^k$ and the residual equals to $r - \Delta$. For $\mu_k(0) = \sum_{i=1}^{Q_k(0)} \delta_{x_i}$, where δ_x denotes the Dirac measure at x,

$$\mu_k(\delta) = \sum_{i=1}^{Q_k(0)} \delta_{x_i - \Delta t / \sum_{k=1}^K <\mathbf{1}, \mu_k>};$$

hence,

$$\mathsf{E}[<f, X(\Delta t)>|X(0) = (\mathbf{q}, r, \mu)] = \sum_{k=1}^K \lambda_k \Delta t \int f(\mathbf{q} + \mathbf{e}^k, r - \Delta t, z)\, d(\tilde{\mu} + \delta_{\ell_k})$$

$$+ \left(1 - \sum_{k=1}^K \lambda_k \Delta t\right) \int f(\mathbf{q} + \mathbf{e}^k, r - \Delta t, z)\, d\tilde{\mu} + o(\Delta t),$$

where

$$\tilde{\mu} = \sum \delta_{x - \Delta t / \sum_{k=1}^K <\mathbf{1}, \mu_k>},$$

when $\mu = \sum \delta_x$.

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

Take the limit as $\Delta t \to 0$, we have, when $r > 0$,

$$\mathcal{G}<f,(\mathbf{q},r,\mu)> = \sum_{k=1}^{K} \lambda_k(<f,(\mathbf{q}+\mathbf{e}^k,r,\mu+\delta_{\ell_k})> - <f,(\mathbf{q},r,\mu)>) + <\frac{\partial f}{\partial y},(\mathbf{q},r,\mu)> + \sum_{k=1}^{K} \frac{<\partial f/\partial z_k,(\mathbf{q},r,\mu)>}{\sum_{k=1}^{K}<\mathbf{1},\mu_k>};$$

similarly, when $r = 0$,

$$\mathcal{G}<f,(\mathbf{q},r,\mu)> = \sum_{k=1}^{K} \lambda_k[<f,(\mathbf{q}+\mathbf{e}^k,r,\mu+\delta_{\ell_k})> - <f,(\mathbf{q},r,\mu)>] + (<f,(\mathbf{q}-\mathbf{e}^{k^*},r,\mu)> - <f,(\mathbf{q},r,\mu)>) + <\frac{\partial f}{\partial y},(\mathbf{q},\delta_{\ell_{k^*}},\mu)> + \sum_{k=1}^{K} \frac{<\partial f/\partial z_k,(\mathbf{q},r,\mu)>}{\sum_{k=1}^{K}<\mathbf{1},\mu_k>},$$

where

$$a = \sup\left\{s : \sum_{k=1}^{K} <\mathbf{1}\{s,\infty\},\mu_k> = \sum_{k=1}^{K} q_k\right\},$$

$$k^* = \{k : (<\mathbf{1}\{a+0,\infty\},\mu_k> - <\mathbf{1}\{a,\infty\},\mu_k>) > 0,\}.$$

Following the same arguments, we have the following.

COROLLARY 2: *The process* $(\mu_1(t),\mu_2(t),\ldots,\mu_K(t))$ *in* $\times\mathcal{M}_F^K$ *is a measure-valued Markov process. Its infinitesimal generator is given by,*

$$\mathcal{G}<f,(\mu_1,\mu_2,\ldots,\mu_K)> = \sum_{k=1}^{K} \lambda_k(<f,\mu+\delta_{\ell_k})> - <f,(\mu_1,\mu_2,\ldots,\mu_K)>) + \sum_{k=1}^{K} \frac{<\partial f/\partial z_k,(\mu_1,\mu_2,\ldots,\mu_K)>}{\sum_{k=1}^{K}<\mathbf{1},\mu_k>}.$$

3. AVERAGE WAITING TIME APPROXIMATIONS

In this section we will derive our approximation scheme for the average waiting time for a FSP queue. It requires that we have knowledge of the associated PS queue at the arrival epoches (in other words, the state of the PS queue at any arrival epoch). It is unrealistic to obtain such a characterization exactly; therefore, we will use recently developed fluid and diffusion approximations for the PS queue instead. Hence, this

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

section will be divided into two parts. In the first part we present the basic results of fluid and diffusion approximation of a multiclass queue under the PS discipline; then, in the second part, we will make use of these approximations and the "tagged job" argument of Kleinrock to derive an approximation scheme for average waiting time under FSP.

3.1. Approximations for the Multiclass PS Queue

In this subsection we will present fluid and diffusion approximations to the associate PS queue as a multiclass queuing system. Most of the results are direct adaptions of those in [6,7,11], the only exception is the fluid limit, which is weaker in the limit sense, but more proper to our application, meanwhile, its proof come directly from applying Dynkin's formula using the infinitesimal generator we derived in the last section.

As we have defined, for each $k = 1, 2, \ldots, K$, $\mu_k(t)$ denotes the measure-valued process of the residual service time for the class k jobs in a PS queue. $\mu(t) \in \mathcal{M}_F$ is the measure-valued process of residue service time for the queue at time t; that is, $\mu(t) = \sum_{k=1}^{m} \mu^k(t)$, equivalently, $\mu(t) = \sum_{k=1}^{Q^{PS}(t)} \delta_{r_k}$, where $Q^{PS}(t)$ denotes the number of jobs in the PS queue at time t and R_k denotes their remaining service time. In the following, we will discuss its fluid limit, the asymptotic behavior of the fluid limit, and the diffusion approximations.

Fluid and diffusion limits are obtained through proper scaling of time and space. For this purpose, define

$$\bar{\mu}_k^r(t) = \frac{1}{r}\mu_k(rt), \bar{\mu}(t) = \frac{1}{r}\mu(rt), \hat{\mu}_k^r(t) = \frac{1}{r}\mu_k(r^2t), \hat{\mu}(t) = \frac{1}{r}\mu(r^2t).$$

Let us investigate their behavior as $r \to \infty$.

3.1.1. Tightness and fluid limit. Our goal is to show that the measure-valued processes $\{\bar{\mu}_k^r(t)\}$ is tight. By Jakubowski's criterion (see, e.g., [3]), it suffices to show the following:

1. For each $T > 0$ and $0 < \eta < 1$, there exists a compact set $K_{T,\eta} \in \mathcal{M}_F$ such that
$$\liminf_{r\to\infty} \mathsf{P}[\bar{\mu}_k^r(t) \in K_{T,\eta} \quad \forall t \in [0, T]] \geq 1 - \eta.$$
2. For each $g \in C_b^1(\mathbb{R}_+)$, the space of a bounded and continuous function on $\mathbb{R}_+$, the real-valued processes $\{<g, \bar{\mu}_k^r(t)>\}$, is tight.

The second statement is relatively conventional to verify; in fact, the proof in [7] using fundamental results in [3] in the case of a single-class PS queue can be directly applied here. For the first one, we will prove by contradiction. Suppose that it is not

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

true, then there exist a sequence of T_n such that for each compact set $K \in \mathcal{M}_F$, we have

$$\lim_{n\to\infty} \lim \inf_{r\to\infty} \mathsf{P}[\bar{\mu}_k^r(t) \in K \quad \forall t \in [0, T_n]] = 0 \tag{1}$$

On the other hand, for each compact set $K' \subset \mathbb{R} \times \mathbb{R}_+$, let us denote $K \in \mathcal{M}_F$ to be its preimage of the mapping $\mu \to <\chi, \mu>$. Then we know that K is compact since the mapping is closed, and

$$\mathsf{P}[\bar{\mu}_k^r(t) \in K, \forall t \in [0, T_n]] = \mathsf{P}[<\chi, \bar{\mu}_k^r(t)> \in K' \quad \forall t \in [0, T_n]].$$

Then (1) contradicts with the fact that the workload processes $<\chi, \mu^k(t)>$ is tight. Therefore, we can conclude the following.

THEOREM 3: *For each k, the process $\bar{\mu}_k^r(t)$ is tight.*

Next, we demonstrate that $\bar{\mu}_k^r(t)$ converge, to a fluid solution, which will be called the fluid limit. To define a fluid solution, let us first define the following function space $\mathcal{S}_0 = \{f \in \mathcal{S}, f(0) = 0, f'(0) = 0\}$, where $\mathcal{S}$ denotes the space of Schwartz functions. Then a fluid solution is a measure-valued process satisfies the following conditions:

- $\bar{\mu}(t)$ is continuous;
- For each $t \geq 0$, $< \mathbf{1}\{0\}, \bar{\mu}(t) >= 0$,
- For each continuous function $f \in \mathcal{S}_0$,

$$<f, \bar{\mu}^k(t)> = <f(\cdot), \bar{\mu}^k(0)> - \int_0^t \frac{<f', \bar{\mu}^k(s)>}{<1, \bar{\mu}(s)>}\, ds + \lambda t <f, \nu^k>, \quad \forall k \tag{2}$$

for all $t < \infty$.

Now, let us construct k measure-valued processes $\mu_k(t)$ that can be verified to be a fluid solution. We will also show that the fluid solution is uniquely determined by the initial system data, hence implying its uniqueness. For any measure $\xi \in \mathcal{M}_F$, define

$$H_\xi(x) := \int_0^x <\mathbf{1}(y, \infty), \xi > dy, \qquad x \in \mathbb{R}_+;$$

also, denote $U_e(u)$ as the renewal measure for the service time:

$$U_e(u) = \sum_{n=0}^{\infty} (F_e^{*n})(u).$$

Then, given the fluid limit $\bar{\mu}_k(t)$ and $\bar{\mu}(t)$, define the system size as

$$\bar{Q}(t) = <1, \bar{\mu}(t)>,$$

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

and the cumulative service per job can be expressed as

$$\bar{S}(t) = \int_0^t \frac{ds}{\bar{Q}(s)}.$$

In [7], it is shown that

$$\bar{S}'(t) = \frac{1}{\bar{T}'(S(t))}, \bar{T} = H_\xi \star U_e.$$

In other words, they are uniquely determined by the initial condition of the fluid solution. From the definition of the fluid solution, we can see that they in turn uniquely determine the fluid solution. On the other hand, construct the following $\bar{\mu}_k \in \mathcal{M}_F$ for $k = 1, 2, \ldots, K$:

$$<\mathbf{1}_{(0,w)}, \bar{\mu}_k(t)> = <\mathbf{1}_{(0,w)}(\cdot - \bar{S}(t)), \xi_k(t)> + \alpha \int_0^t <\mathbf{1}_{(0,w)}(\cdot - \bar{S}(t) - \bar{S}(s)), \nu_k(s)> ds.$$

It is easy to verify that $\bar{\mu}_k$ satisfy all of the conditions in the definition of the fluid solution. Therefore, we have Theorem 4.

THEOREM 4: *The fluid solution as we defined uniquely exists.*

In turn, we will have Theorem 5.

THEOREM 5: *$\bar{\mu}_k^r(t)$ converges to the fluid solution weakly, as $r \to \infty$ in the following sense:*

$$\mathsf{E}[<f, \bar{\mu}_k^r(t)>] \to <f, \bar{r}_k(t)>$$

for any continue function function f.

PROOF: This weak form of fluid approximation can be obtained through Dynkin's formula (see, e.g. [12, Sect. III.10]):

$$\mathsf{E}[<f, \bar{\mu}_k^r(t)>] = \int_0^{nt} \mathcal{G}<f, \mu> ds.$$

Then the fluid limit can be obtained by the usual functional law of large number and the following:

$$\frac{1}{n} \int_0^{nt} \frac{<f, \mu_k(s)>}{\sum_{k=1}^K <\mathbf{1}, \mu_k(s)>} ds = \int_0^t \frac{<f, \bar{\mu}_k(u)>}{\sum_{k=1}^K <\mathbf{1}, \bar{\mu}_k(u)>} du,$$

which is the result of change of variable $nu = s$. ■

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

Remark:

- The fluid limit in the strong sense of those in [7] can be obtained following the arguments there; the reason we elect to present a weak form of result is twofold. First, for our application, the weak form suffices; second, the derivation from the Dynkin's formula seems to reveal more connections between the original process and the fluid limit.
- We use a different function space in characterizing the action on $\mathcal{M}_F$ than the one used in [6,7,11] because for the problems we deal with in this article, both are sufficient. In general, the Schwartz function space, of course, is much more widely used.
- The fluid limit can also be characterized by a differential equation. From the definition, we know that the fluid limit as a measure is absolute continuous with respect to the Lebesgue measure. Let $h_k(t,\cdot)$ be its Radon–Nykodim derivative with respect to the Lebesgue measure at any time t, then $h_k(t,x)$ should satisfy

$$\frac{\partial}{\partial t} f_k(t,x) = \frac{\partial/\partial x f_k(t,x)}{\sum_k \int_0^\infty f(t,x)\,dx} + \lambda f_{\nu_k},$$

where f_{ν_k} is the density of ν. The existence and uniqueness can also be derived from the theory of nonlinear differential equation (see, e.g., [1]).

3.1.2. Stationary behaviors of the fluid limits. Under the heavy traffic condition, the fluid limit has a stationary limit as $t \to \infty$. This limit is, of course, a measure in $\mathcal{M}_F$ and will serve as an approximation for the remaining service status in the later part of our analysis. Here, we will first identify this limit and then show that the fluid limit indeed converges.

We start with a definition.

DEFINITION 6: *A nontrivial measure $\nu \in \mathcal{M}_F$ (i.e., $\nu \neq 0$) is called an invariant state if $\mu_t(\nu) = \nu, \forall t$. and the set of all invariant states,*

$$I = \{(\nu_k)_k \in \mathcal{M}_F : \mu_k(t)(\nu_k) = \nu_k \quad \forall t\},$$

is called the invariant manifold of the fluid limit.

A class of measures can be easily shown to be part of the invariant manifold; that is, $\xi_k = c\nu_e^k$, where ν_e^k is the lift with respect to ν^k (i.e., ν_e^k is defined as $\nu_e^k([0,x]) = <\chi, \nu^k>^{-1} \int_0^x \mu^k(y,\infty)\,dy$ for any $x \in \mathbb{R}$). To see that, we need the following fundamental relationship;

LEMMA 7: *For each $g \in \mathcal{S}_0$, we have*

$$\alpha <g, \nu> = <g', \nu_e>. \tag{3}$$

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

This is the Proposition 3.1 in [11], a direct application of integration by parts. An argument similar to that in [11], we can show that the class in fact contains all of the measures that are invariant. Hence, we have Theorem 8.

THEOREM 8: *The invariant manifold is* $\{cv_e^k, c \in \mathbb{R}_+\}$.

PROOF: We can adapt the proof in [11]. It is easy to verify that each member in the defined set is an invariant measure. Now, let us show the other direction, given any invariant measure vector (ξ_k), then we know that its corresponding fluid limit $\bar{\mu}_{\xi_k}(t) = \xi_k$ for any $t \geq 0$. Now, the fluid limit tells us that for any $g \in \mathcal{S}_0$, we have $<g_k', \xi_k> = <1, \xi>\alpha_k <g_k, \mu_k>$. Hence, according to Lemma 7, we have $<g_k', \xi_k> = <1, \xi><g_k', \nu_k^e>$. Since $g \in \mathcal{S}_0$, we know that this implies $\xi_k = cv_k^e$ for $c = <1, \xi>$. ■

It is also evident that if the fluid limit has stationary limit, then it must be a point of the invariant manifold. In order to show the convergence, let us revisit the fluid limit. For simplicity, let us assume that we start from a empty system, then

$$<g, \mu_k(t)> = \int_0^t \frac{<g', \mu_k(s)>}{<1, \mu(s)>} ds + \alpha t <g, \nu_k> \quad \forall k;$$

therefore,

$$<g, \mu_k(t)> = \alpha \left(\sum_{n=1}^{\infty} <g^{(n)}, \nu_k> \int_{\Delta_n(t)} \frac{s_n}{\prod_1^n <1, \mu(s_i)>} d\sigma_n \right) \quad \forall k,$$

where we denote

$$\int_{\Delta_n(t)} f(s_1, s_2, \ldots, s_n)\, d\sigma_n = \int_0^t \int_0^{s_1} \cdots \int_0^{s_{n-1}} f(s_1, s_2, \ldots, s_n)\, ds_1\, ds_2 \cdots ds_n.$$

From the results in [11], we know that there exists an $\delta > 0$ such that

$$\prod_1^n (< 1, \mu(\cdot) >) \geq \delta;$$

then we have

$$<g, \mu(t)> \leq \alpha \sum_{n=1}^{\infty} \frac{(t/\delta)^n}{n!} <g^n, \mu(0)>.$$

Since g is a Schwartz function, $<g, \mu(t)>$ is bounded for any $t \geq 0$. From Theorem 7, we know that the invariance measure is uniquely determined by the initial data and we can conclude that $<g, \mu(t)>$ converges as $t \to \infty$.

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

3.1.3. Diffusion approximations. In this subsection we introduce a result in [6] needed in the computation later in the article.

To derive diffusion approximations of the PS queue, Gromoll [6] proposed a "bootstrap"-type approach. The key idea is to replace the diffusion-scaled process $\hat{\mu}^r(t)$ by the so-called *shifted* fluid-scaled process, $\bar{\mu}^{r,m}(t) := \bar{\mu}^r(mt)$, with $m \in [0, rT]$. Thus, as $r \to \infty$, it can be shown that $\bar{\mu}^{r,m}(t)$ behaves asymptotically like its stationary limit, to be more precise, we have

$$\bar{\mu}^{r,m}(t) \approx \Delta_v < \chi, \bar{\mu}^{r,m}(t) >$$

and this leads to

$$\hat{\mu}^r(t) \approx \Delta_v < \chi, \hat{\mu}^r(t) >.$$

$< \chi, \hat{\mu}^r(t) >$ is the workload process of the queue, which is known to converge to a reflected Brownian motion.

THEOREM 9 (Gromoll): *Under the heavy traffic condition, the residue service process of the process sharing queue can be approximated by*

$$\mu^* = \Delta_v W^*,$$

where W^* denotes the measure of the reflected Brownian motion that approximates the workload process.

Note that the workload process of a PS queue is the same as the workload process of a FCFS queue; therefore, we know that W^* is a reflected Brownian process with mean $\lambda - \mu$ and variance $\lambda + \lambda C_s^2$.

3.2. Calculate the Average Waiting Time for FSP

To estimate the average waiting time for FSP, we follow the "tagged arrival" argument used in Kleinrock [9]. Suppose that an arrival belongs to class $k = 1, 2, \ldots, K$; its waiting time, then, is made of three parts—first, the unfinished service at the server; second, the service time of those jobs that already in the queue and will be processed before the tagged job; third, the service time of those jobs that arrives later but will be served before the tagged arrival. In summary, we have

$$W_k = W_0 + \sum_{i=1}^{K} (N_{ik} + M_{ik}) \ell_i, \qquad k = 1, 2, \ldots, K, \tag{4}$$

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

where

$W_k \triangleq$ waiting time for class k job

$N_{ik} \triangleq$ number of class i customers founded in the queue by a tagged job (from class k) and receive service before the tagged job

$M_{ik} \triangleq$ number of class i customers arriving to the system while the tagged job (from class k) is in the queue and receiving serve before the tagged job

$W_0 \triangleq$ remaining service observed by the tagged job

W_0, since we have Poisson arrivals, can be treated as long-run average remaining service; hence, see, for example, [9] and

$$\mathsf{E}W_0 = \sum_{k=1}^{K} \frac{\lambda_i \ell_i^2}{2}. \tag{5}$$

Now, let us consider other terms in eqn (4). We start with N_{ip}. Suppose that at time t, when the tagged job arrives, one observes the status of the associated PS queue being

$$(\mu_1(t), \mu_2(t), \ldots, \mu_K(t)).$$

We know that for any $z \geq 0$, $<\mathbf{1}\{z, \infty\}, \mu_k(t)>$ denotes the number of class k jobs in the associated PS system that has more than z units of workload; meanwhile, if these jobs are in the queue of the FSP system, the workload is ℓ_k. Hence, there exist $a \geq 0$ such that

$$a = \min\left\{ z : \sum_{k=1}^{K} <\mathbf{1}\{z, \infty\}, \mu_k(t)>\ell_k - < \chi, \mu(t)> \ < 0 \right\}. \tag{6}$$

Then for each $k = 1, \ldots, K$, $<\mathbf{1}\{a, \infty\}, \mu_k(t)>$ represents the number of the class k jobs currently in the queue of the FSP system, and $<\mathbf{1}\{a, \infty\}, \mu(t)>$ represents the total number of jobs in the queue in FSP system. Meanwhile, $<\mathbf{1}\{\ell_p, \infty\}, \mu_i(t)>$ is the number of the class i jobs that will be served after the tagged job. Hence, their difference will be exactly what we desired; that is,

$$N_{ip} = <\mathbf{1}\{a, \infty\}, \mu_i(t)> - <\mathbf{1}\{\ell_p, \infty\}, \mu_i(t)>. \tag{7}$$

Next, let us look at the quantity M_{ip}, which refers to the number of class i jobs that arrive after the tagged class p job but receive service before the tagged job. Of course, when $i \geq p$ and $\ell_i \geq \ell_p$, $M_{ik} = 0$. When $i < p$, M_{ip} equals the number of

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

class i jobs that arrive before the remaining process time for the tagged job reaches $\ell_p - \ell_i$. Therefore, with the Poisson arrival assumption, we have

$$\mathsf{E}[M_{ip}] = \lambda_i \mathsf{E}[\tau_{ip}], \qquad \tau_{ip} := \inf\{t : \int_0^t \frac{ds}{Q^{\mathrm{PS}}(s)} \geq \ell_p - \ell_i\}.$$

Assume that we use $Q^{\mathrm{PS}}(\infty)$, the stationary distribution of the PS queue length, to replace $Q^{\mathrm{PS}}(s)$ by Wald's identity; we have

$$\mathsf{E}\tau_{ip} = \frac{\ell_i - \ell_p}{\mathsf{E}[1/Q^{\mathrm{PS}}(\infty)|Q^{\mathrm{PS}}(\infty) \geq 1]}. \tag{8}$$

We will employ the fluid and diffusion approximations described in the previous section. Recall that in the associated multiclass PS queue, for any initial data ξ, the fluid limit of the residual time descriptor $\bar{\mu}_k(t)$ has limit $<1, \xi > \Delta_{\nu_k}$. Especially, we take ξ to have the first moment as the stationary queue length. Plug this into (6); we have

$$a = \left(\sum_n \lambda_n \ell_n\right) \Big/ \left(2 \sum_n \lambda_n\right).$$

Hence, we have

$$\mathsf{E}N_{ik} = \frac{2\left(\ell_i - (\sum_n \lambda_n \ell_n)/(2\sum_n \lambda_n)\right)}{\ell_p \sum_n \lambda_n \ell_n} \mathsf{E}Q^{\mathrm{PS}}.$$

The key to the calculation of (8) is to compute $\mathsf{E}[1/Q^{\mathrm{PS}}(\infty)|Q^{\mathrm{PS}}(\infty) \geq 1]$; thus,

$$\mathsf{E}\left[\frac{1}{Q^{\mathrm{PS}}(\infty)}|Q^{\mathrm{PS}}(\infty) \geq 1\right] = \int_1^\infty \frac{1}{x} \mathsf{P}[Q^{\mathrm{PS}}(\infty) \in dx | Q^{\mathrm{PS}}(\infty) \geq 1].$$

From the diffusion approximation, we know that $Q^{\mathrm{PS}}(\infty)$ can be approximated by an exponential distribution $\mathrm{Exp}(2(1-\rho)/[\rho(1+\sigma_s^2 \wedge 1)])$. Therefore, we have

$$\mathsf{E}\left[\frac{1}{Q^{\mathrm{PS}}(\infty)}\middle| Q^{\mathrm{PS}}(\infty) \geq 1\right] \approx \frac{2(1-\rho)}{\rho(1+\sigma_s^2 \wedge 1)} \exp\left(\frac{2(1-\rho)}{\rho(1+\sigma_s^2 \wedge 1)}\right) I\left(\frac{2(1-\rho)}{\rho(1+\sigma_s^2 \wedge 1)}\right),$$

where $I(y) := \int_y^\infty (e^{-\sigma x}/x) dx$.

Combine what we have derived; we can have the following approximation, for each $k = 1, 2 \ldots, K$:

$$\mathsf{E}W_k = \sum_{k=1}^K \frac{\lambda_i \ell_i^2}{2} + \sum_{i=1}^{k-1} \left[\frac{\left(\ell_i - \frac{\sum_n \lambda_n \ell_n}{2\sum_n \lambda_n}\right)}{(\ell_p \sum_n \lambda_n \ell_n)} \frac{\lambda(1+\sigma_s^2)}{(1-\rho)}\right]$$
$$+ \sum_{i=k+1}^K (\lambda_k(\ell_k - \ell_i)) \Big/ \left[\frac{2(1-\rho)}{\rho(1+\sigma_s^2 \wedge 1)} \exp\left(\frac{2(1-\rho)}{\rho(1+\sigma_s^2 \wedge 1)}\right) I\left(\frac{2(1-\rho)}{\rho(1+\sigma_s^2 \wedge 1)}\right)\right]^{-1}. \tag{9}$$

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press

References

1. Cabré, X. & Caffarelli, L.A. (1995). *Fully nonlinear elliptic equations*. Providence, R: American Mathematical Society. Colloquium Publications, Vol. 43.
2. Dawson, D.A. (1993). Measure-valued markov processes, *Ecole d'Eté de Probabilités de Saint Flour XX1-1991*, Lecture Notes in Mathematics No. 1541. New york; Springer.
3. Ethier, S. & Kurtz, T. (1986). *Markov processes: Characterization and convergence.* New York: Wiley.
4. Friedman, E.J. & Henderson, S.G. (2003). Fairness and efficiency in web server protocols. In *Proceedings ACM/SIGMETRICS'03*.
5. Grishechkin, S. (1994). *GI/GI*/1 processor sharing queue in heavy traffic. *Advances in Applied Probability* 26: 539–555.
6. Gromoll, H.C. (2004). Diffusion approximations for a processor sharing queue in heavy traffic. *Annals of Applied Probability* 14: 555–611.
7. Gromoll, H.C., Puha, A.L., & Williams, R.J. (2002). The fluid limit of a heavily loaded processir sharing queue. *Annals of Applied Probability* 12: 797–859.
8. Kallenberg, O. (1976). *Random measures*. Berlin: Akademie-Verlag.
9. Kleinrock, L. (1976). *Queueing systems*, Vol. II: *Computer applications*, New York: Wiley.
10. Prohorov, Y.V. (1956). Convergence of random processes and limit theorems in propability theory. *Theory of Probability and its Applications* 1: 157–214.
11. Puha, A.L. & Williams, R.J. Invariant states and rates of convergence for a critical fluid model of a processor sharing queue. *Annals of Applied Probability* 14: 517–554.
12. Rogers, L.C.G. & Williams, D. (2000). *Diffusions, Markov processes and martingales*. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S0269964809000060 Published online by Cambridge University Press