Population effects in Multiclass Processor Sharing Queues

Abdelghani Ben Tahar
Laboratoire de Mathématiques Raphaël Salem
UMR 6085 CNRS-Université de Rouen*
Avenue de l'Université, BP.12,
F-76801 Saint-Étienne-du-Rouvray
abdelghani.bentahar@univ-rouen.fr

Alain Jean-Marie
INRIA / LIRMM
UMR 5506 CNRS-Université Montpellier 2
161 Rue Ada
F-34392 Montpellier
ajm@lirmm.fr

ABSTRACT

Consider a single server queueing system with several classes of customers, each having its own renewal input process and its own general service times distribution. Upon completing service, customers may leave, or reenter the queue, possibly as customers of a different class. The server is operating under the egalitarian or the discriminatory processor sharing discipline.

In this paper, we consider the fluid approximation of this multiclass processor sharing queue. We first provide the results allowing to compute the trajectories for this model, under the egalitarian PS discipline. Asymptotic results for overloaded queues are also stated. Next, we show that a simple transformation allows to compute the solution for the discriminatory PS queue as well. Finally, we illustrate the different results through numerical experiments. We compare transient trajectories with simulations, and we discuss the fairness issue that may arise in overloaded PS queues.

Keywords

Fluid Limit, Fluid model, measure valued process, multiclass networks, Processor sharing

1. INTRODUCTION

The processor sharing queue is a central model for evaluating the performances of various computer and telecommunication systems. Relatively few analytical results seem to be available for the calculation of performance metrics in multiclass PS queues: see for instance [12]. On the other hand, when the "load" of the queue is important (be it measured in terms of arrival intensity or just the amount of work present), fluid approximations become an interesting complement to a purely stochastic analysis. In the last ten years, many detailed results have been obtained on the fluid limits of the PS queue in [6, 10], refining the initial findings of [4]. We have recently extended this analysis to the multiclass case, where customers may re-enter the queue, possibly changing classes. Fluid approximations for PS queues are an active topic of the current literature, see *e.g.* [7, 13, 8]. Some of these variants consider features like impatiences, limited service or network of queues, but no work has so far addressed the multiclass, single server queue.

The purpose of this paper is twofold. First, we present in a synthetic and self-contained way several results recently obtained on the fluid approximation for the Multiclass PS queue, and the correspondence that can be established between the egalitarian PS queue and the discriminatory PS (DPS) queue. The complete analysis and the proofs are too long to be reported here, and can be found in [3].

The second purpose of this paper is to illustrate the usefulness of these results through several numerical experiments. First, we discuss the computation of trajectories for the population within classes, and compare with simulation, for standard PS and DPS examples. We also illustrate the asymptotic results when time goes to infinity. Next, we exploit a correspondence between the single-class queue and the multiclass queue to revisit the dependence of the response time on the service time duration in an overloaded queue. Finally, still for an overloaded queue, we discuss the effect of the service time distribution in classes, on the growth rate of the different populations.

The paper is organized as follows. In Section 2, we introduce the queuing model. We describe the fluid approximation in Section 3, together with the theoretical results obtained for this model: existence, calculation of the solution, asymptotic behavior. Section 4 is devoted to the illustration of the different results.

2. THE MODEL

We consider first a single-server queue operating under the egalitarian Processor Sharing discipline. Then we shall extend the analysis to the discriminatory PS discipline.

Customers belong to a set $\mathcal{K}$ of K classes. For each $k \in \mathcal{K}$, customers arrive to the queue from the "exterior" according to a renewal process of intensity α_k. Customers of class k have a service duration generically denoted as v_k and distributed according to some probability distribution B_k. Upon service completion, a customer of class k may re-enter the queue as a customer of class l with a fixed probability

*Part of this work was performed while the author was a CNRS Postdoctoral fellow at LIRMM.

VALUETOOLS 2009, October 20-22, Pisa, Italy

DOI 10.4108/ICST.VALUETOOLS2009.7622

p_{kl}, or leave the system with probability $p_{k0} := 1 - \sum_{l \in \mathcal{K}} p_{kl}$.

The service requirements of different customers are assumed to be independent. It is assumed that arrival processes, service durations and routing decision processes are mutually independent as well.

As it will be clear shortly, the analysis of the processor sharing queue is centrally based on a state-space representation using a *measure* of residual service times. For this reason, many results are conveniently expressed using this formalism, which we briefly describe now. For any Borel measure μ, and some function f, we shall use the scalar product $\langle f, \mu \rangle = \int f(x)\mu(\mathrm{d}x)$. In particular, $\langle 1, \mu \rangle$ is the total mass of the measure. If μ is a probability measure associated with some cumulative distribution function $B(\cdot)$, and some random variable X, then $\langle 1_{[0,x]}, \mu \rangle = \mathbb{P}(X \leq x) = B(x)$. In particular, $\langle 1, \mu \rangle = 1$.

Introduce ν_k, the Borel probability measure associated with B_k. It is assumed that for each $k \in \mathcal{K}$, the distribution ν_k does not charge the origin, $\nu_k(\{0\}) = 0$, and we define $\beta_k = \mathbb{E}v_k$ and $\beta_k^{(2)} = \mathbb{E}(v_k)^2$. It is assumed that $\beta_k < +\infty$. The Laplace transform of the distribution will be denoted with $\widehat{B}_k(\cdot)$.

Finally, define the non-negative matrix $P = ((p_{kl}))_{(k,l) \in \mathcal{K} \times \mathcal{K}}$ and its transpose P'. The system is assumed to be *open* in the following sense. We allow that $\alpha_k = 0$ for some classes k, but at least one of them must be strictly positive. Moreover, the matrix $Q = I + P' + (P')^2 + ...$ is assumed to be finite, which is equivalent to requiring that $(I - P')$ be invertible, or that P has a spectral radius less than 1. In that case, $Q = (I - P')^{-1}$.

Initial conditions. For each $k \in \mathcal{K}$, we assume that the system initially contains an integer random number of customers of class k, denoted with $Z_k(0)$. Those customers have a (residual) service time distributed according to a distribution B_k^0. These service times are assumed to be mutually independent and independent from the other variables already introduced. Any customer belonging to class k at time zero in the system is referred to as an "initial customer of class k". After service, an initial customer of class k is routed exactly as an external customer.

Mapping to the single-class case. When a customer leaves the queue to re-enter it immediately, the total number of customers in the system is not changed. Therefore, this event has no effect on the other customers: everything is "as if" the considered customer has simply completed a phase of its service and begun the next phase. Consider the variable V_k representing the total service requirement of some customer, entering initially the queue as a customer of class k, up to the moment when it leaves the queue forever. From the point of view of the total number of customers, the multiclass queue with routing is equivalent to a multiclass queue without routing and a service time V_k for customers of class k. Next, consider the variable V constructed as a mixture of the V_k, with probabilities proportional to the external arrival rates to each class. This variable represents the service requirement of a "typical" customer taken at random in the global input flow. In the case where the arrival process of each class is constructed as a Bernoulli sampling of some renewal process (this includes Poisson-distributed arrival processes), the multiclass queue is equivalent to a single-class one with services distributed as V. This equivalence turns out to happen in the fluid limit, even if the coupling of input processes and service durations is not exact in the stochastic and discrete queue.

As a consequence, it is possible to consider the fluid multiclass processor sharing queue as a single-class one and use results from the literature for this system, *e.g.* [9, 10, 11, 6]. Note however that this reduction concerns uniquely global quantities such as the total number of customers. It is not possible to infer *a priori* per-class quantities from the results about the single-class queue.

The following lemma provides the formulas for passing from the multiclass description to the single-class one. Those formulas involve matrix manipulations and we introduce here the notation which will be used throughout the rest of this paper. Let $B(x) = diag\{B_k(x); k \in \mathcal{K}\}$, $\widehat{B}(x) = diag\{\widehat{B}_k(x); k \in \mathcal{K}\}$ and $\beta = diag\{\beta_k; k \in \mathcal{K}\}$. Let e be the (row) "vector of ones". For two matrices of measurable functions $F(.)$ and $G(.)$ defined on $\mathbb{R}_+$, we denote by the matrix-valued functions $(F * G)(x)$ for $x \in \mathbb{R}_+$, the matrix convolution formed of the elements: $(F * G)_{ij}(x) = \sum_k (F_{ik} * G_{kj})(x)$. This operation is associative and distributive over matrix addition. The multiplication by a constant matrix C can be seen as a convolution, where each element C_{ij} is interpreted as the function $C_{ij} 1_{x \geq 0}$. Associativity therefore holds for mixed scalar products and convolutions. The n-th convolution power of a matrix $F(x)$ is denoted with $F^{*n}(x)$.

The following renewal-like matrix function is central in the calculations.

$$\mathcal{B}(x) = \sum_{n=0}^{\infty} (BP')^{*n}(x) \ . \tag{1}$$

Finally, let $\alpha_e := \sum_{k \in \mathcal{K}} \alpha_k = e.\alpha$ be the "equivalent" arrival rate of single-class customers. The "." denotes here the inner product for vectors.

LEMMA 1. *We have the following properties.*

i) The distribution function of the variable V_k is given by:

$$V_k(x) \ = \ \big(e(I - P')(\mathcal{B} * B)(x))\big)_k \ , \tag{2}$$

with first moment $\mathbb{E}(V_k) = (e\beta Q)_k$ and Laplace transform given by:

$$\widehat{V}_k(s) \ = \ \Big(e(I - P')(I - \widehat{B}(s)P')^{-1}\widehat{B}(s))\Big)_k \ . \tag{3}$$

ii) The variable V, formed as a mixture of the V_k, proportionally to the arrival rates α_k, has a distribution function $V(x) = \sum_{k \in \mathcal{K}} \alpha_k V_k(x)/\alpha_e$ with Laplace transform given by:

$$\widehat{V}(\theta) \ = \ \frac{1}{\alpha_e} \ e(I - P')(I - P'\widehat{B}(\theta))^{-1}\widehat{B}(\theta)\alpha \ .$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622

These results can be proved as follows. The total service time of a customer of class k is the sum of one service time v_k and of the service it requires after the end of this service. The latter duration is distributed according to V_j with probability p_{kj} and is zero with probability p_{k0}. The service time after re-entering the queue is independent from the first service. Accordingly, we have the identity for distribution functions:

$$V_k(x) \;=\; \sum_j p_{kj}(B_k * V_j)(x) \;+\; p_{k0}B_k(x)\ .$$

Expressed in vector-matrix form, with $V(x) = (V_k(x); k \in \mathcal{K})$ (a row vector), we have:

$$V(x) \;=\; \big(V * (P'B)\big)(x) \;+\; e(I - P')B(x)\ ,$$

and this is a multidimensional renewal equation, of the sort studied in [1] for instance. By application of Lemma 2.1 in this reference, we obtain the existence and uniqueness of the solution, and the value $V(x) = e(I - P')(\mathcal{B} * B)(x)$, whence (2).

3. THE FLUID APPROXIMATION AND ITS SOLUTION

3.1 Definition of the Fluid Model

The fluid limit arises from a normalization of the stochastic, discrete queueing process. Quantities of interest are: the number of arrivals in class k to date t, $A_k(t)$, the number of departures $D_k(t)$, the current population in class k, $Z_k(t)$ and the measure $\mu_k(t)$ which counts the residual service times of all customers of class k in the queue.

Given a sequence of discrete Multiclass Processor Sharing systems with parameters and descriptors indexed by an integer number r, we obtain a sequence of arrival, departure and residual service time measures: $A^r_k(t)$, $D^r_k(t)$, $\mu^r_k(t)$ for each class k. The superscript r denotes the dependence of the process on this parameter r, through initial populations, arrivals, service time and routing distributions. The scaled processes that will give rise to a fluid limit are defined as:

$$\bar{A}^r_k(t) \;=\; \frac{A^r_k(rt)}{r}, \quad \bar{D}^r_k(t) \;=\; \frac{D^r_k(rt)}{r},$$
$$\bar{Z}^r_k(t) \;=\; \frac{Z^r_k(rt)}{r}, \quad \bar{\mu}^r_k(t) \;=\; \frac{\mu^r_k(rt)}{r}\ .$$

The fluid model shares the following parameters with the discrete model: the non-negative vector $\alpha = (\alpha_1, ..., \alpha_K)$, the vector of Borel probability measures $\nu = (\nu_1, ..., \nu_K)$ and the non-negative routing matrix P.

Define the vector $\lambda = Q\alpha$. The global arrival rate to the class k is then λ_k, and the load factor of the queue is $\rho = \sum_{k\in\mathcal{K}} \lambda_k \mathbb{E}v_k$. The adjectives *subcritical*, *critical* and *supercritical* will be used to refer to data (α, ν, P) that satisfy $\rho < 1$, $\rho = 1$, $\rho > 1$ respectively.

Let $\mathcal{M}^{c,K} = \{\xi \in \mathcal{M}^K : \xi_k(\{x\}) = 0 \text{ for all } x \in \mathbb{R}_+ \text{ and } k \in \mathcal{K}\}$ be a set of finite, non-negative Borel measures on $\mathbb{R}_+$ that have no atoms, and let $\mathcal{M}^{c,p,K} = \{\xi \in \mathcal{M}^{c,K} : \xi \neq 0\}$ be the set of positive measures of $\mathcal{M}^{c,K}$.

DEFINITION 1 (FLUID SOLUTION MODEL). *Let (α, ν, P) be some data and $\xi \in \mathcal{M}^{c,K}$ be an initial state. A fluid solution is a triple $(A(t), D(t), \mu(t))$ of two real-, and one measure-valued vectors of continuous functions: A, $D : \mathbb{R}_+ \to \mathbb{R}^K_+$, and $\mu = (\mu_1, ..., \mu_K) : \mathbb{R}_+ \to \mathcal{M}^K$ such that $\mu(0) = \xi$, and for all $t < t_\rho(\xi)$ defined below,*

i) A and D are increasing componentwise,

ii) The triple satisfies the relations

$$A(t) \;=\; \alpha t + P'D(t) \tag{4}$$
$$\langle 1, \mu_k(t)\rangle \;=\; \langle 1, \xi_k\rangle + A_k(t) - D_k(t) \tag{5}$$
$$\langle 1_{[x,\infty[}, \mu_k(t)\rangle \;=\; \langle 1_{[x,\infty[}(. - S(0,t)), \xi_k\rangle + \int_0^t \langle 1_{[x,\infty[}(. - (S(s,t)), \nu_k\rangle dA_k(s) \tag{6}$$

for every $k \in \mathcal{K}$, $x \in \mathbb{R}_+$ and:

$$S(s,t) = \int_s^t \varphi(\langle 1, e.\mu(u)\rangle) du\ , \tag{7}$$

where $\varphi(x) = 1/x$ for $x \in (0, \infty)$, and $\varphi(0) = 0$.

For $t \geq t_\rho(\xi)$, $A(t) = D(t) = \lambda t$, $\mu(t) = 0$. The number $t_\rho(\xi)$ is the time range of the solution, and is defined as:

$$\begin{cases} t_\rho(\xi) = \inf\{t : e.\mu(t) = 0\} & \text{if } \xi \neq 0 \\ t_\rho(0) = 0 & \text{if } \rho \leq 1 \\ t_\rho(0) = \infty & \text{if } \rho > 1\ . \end{cases} \tag{8}$$

It turns out that under quite general assumptions, the sequence of normalized processes $\bar{A}^r$, $\bar{D}^r$ and $\bar{\mu}^r$ converge when $r \to \infty$ to a fluid solution model according to Definition 1. The precise assumptions, the sense in which convergence of these stochastic processes occurs, and the proofs are given in [3]. Basically, arrival, service and routing distributions should converge simply, whereas initial populations should be asymptotically proportional to r.

The proof techniques are adapted from [6, 10] with two main differences. First, the per-class processes are coupled through Eq.(6), which prevents to reduce the problem to independent one-dimensional problems. The analysis is truly multidimensional, with in particular the use of multidimensional renewal equations. The second and main difference is that arrivals to class k are the superposition of exogenous arrival, which are easy to characterize, and endogenous ones, the nature of which is more difficult to assess. Most of the difficulty lies in the proof that the customer feedback processes are "as tame" as external renewal processes.

Total populations and the initial measure. The definition above involves the detailed residual service time measure $\mu_k(t)$, but a less detailed description is also possible. The total population of class k, $Z_k(t)$, is related to μ_k by:

$$Z_k(t) \;=\; \langle 1, \mu_k(t)\rangle\ .$$

On the other hand, the measure ξ_k appearing in (6) represents the distribution of the workload among the customers of class k initially present in the queue. Since ξ_k is a finite measure for each $k \in \mathcal{K}$, there exists a probability measure ν^0_k such that $\xi_k = Z_k(0)\nu^0_k$. We shall denote with v^0_k the

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622

generic random variable with distribution ν_k^0, and we shall assume that $\beta_k^0 := \mathbb{E}v_k^0 < +\infty$.

As a particular case of (6), we have the law of evolution for Z_k:

$$Z_k(t) = Z_k(0)\mathbb{P}(v_k^0 > S(0,t)) + \int_0^t \mathbb{P}(v_k > S(s,t))\mathrm{d}A_k(s) \tag{9}$$

since $\langle 1_{[x,\infty[}, \mu_k(0)\rangle = Z_k(0)\mathbb{P}(v_k^0 > x)$ for all $x \geq 0$.

Attained service and response time. The function $S(s,t)$, defined in Equation (7) and appearing in (6) or (9), represents the service accumulated by *any* customer in the queue which would be present during the time interval $[s,t]$. Indeed, the derivative of this function is $1/\sum_{k\in\mathcal{K}} Z_k(t)$ (when the queue is not empty), in accordance with the rules of the egalitarian Processor Sharing discipline. Consider a customer arriving at time s and requiring a service of σ units. The time t at which the customer leaves the queue is such that the attained service at time t is σ. Therefore, $\sigma = S(s,t)$. Introduce the abbreviated form $S(t) = S(0,t)$ and the function $T(u) = S^{-1}(u)$. Since $S(s,t) = S(t) - S(s)$, we can express the departure time as $t = T(S(s) + \sigma)$ and the response time as $R = T(S(s) + \sigma) - s$. In particular, for customers present initially, the response time is simply $R = T(\sigma)$.

3.2 Results

3.2.1 Time range of the solution

The first result concerns the interval over which the fluid queue is not empty. The time range of the solution, $t_\rho(\xi)$, has been defined above in (8).

LEMMA 2. *Assume that $\xi \neq 0$. The the value of $t_\rho(\xi)$ is*

$$\begin{cases} t_\rho(\xi) &= +\infty & \text{if } \rho \geq 1 \\ t_\rho(\xi) &= \dfrac{e(\beta^0 + \beta QP')Z(0)}{1-\rho} & \text{if } \rho < 1\ . \end{cases}$$

In this result, the term $e(\beta^0 + \beta QP')Z(0)$ has the following interpretation. It is the "virtual workload" of customers which are initially present. Indeed, each customer of class k brings its initial workload β_k^0, but will also bring work to the queue when it re-enters as a customer of class l (which happens with probability, or fluid proportion, p_{kl}). According to Lemma 1, this quantity is given by $(e\beta Q)_l$.

3.2.2 Existence and Construction

The next result provides existence, uniqueness of the solution, and provides the explicit formulas for constructing it. The result holds whether the queue is subcritical, critical or supercritical.

THEOREM 3. *Given data (α, ν, P) and $\xi \in \mathcal{M}^{c,K}$, there exists a unique fluid solution $(A(t), D(t), \mu(t))$ of the model such that $\mu(0) = \xi$. For every $t \leq t_\rho(\xi)$, the elements $(A(t), D(t), Z(t))$ of this solution are given by*

$$\begin{aligned} A(t) &= \lambda t + QP'(Z(0) - Z(t)) \\ D(t) &= \lambda t + Q(Z(0) - Z(t)) \\ Z(t) &= \widetilde{Z}(T^{-1}(t)) \\ \widetilde{Z}(s) &= Q^{-1}(\mathcal{B} * C)(s)Z(0) + Q^{-1}(\mathcal{B} * (I-B) * (TI))(s)\lambda \\ T(s) &= (H * U_e)(s) \\ U_e(u) &= \sum_{n\geq 0} \rho^n (V_e)^{*n}(u) \\ H(x) &= \int_0^x e.Q^{-1}(\mathcal{B} * C)(y)Z(0)\,dy \\ C(t) &= (I - B^0(t)) + (I - B(t))QP' \end{aligned}$$

where $\mathcal{B}$ is defined in (1)*, and V_e is the excess lifetime distribution associated to the random variable V.*

The measure $\mu(t)$ can also be computed in closed form. We omit this expression which we shall not use in this paper. For computational purposes, most of these functions are better described by their Laplace-Stieltjes transforms. Using the generic notation $\widehat{F}(\theta) = \int e^{-x\theta}\mathrm{d}F(x)$, we have:

$$\begin{aligned} \widehat{\mathcal{B}} &= (I - \widehat{B}P')^{-1} \\ \widehat{\widetilde{Z}} &= Q^{-1}\widehat{\mathcal{B}}\left(\widehat{C}Z(0) + (I - \widehat{B})\widehat{T}\lambda\right) \\ \widehat{T}(\theta) &= \frac{\widehat{H}(\theta)}{1 - \psi(\theta)} \\ \widehat{H}(\theta) &= \theta^{-1}\ e\left(I - Q^{-1}(I - \widehat{B}(\theta)P')^{-1}\widehat{B}^0(\theta)\right)Z(0) \\ \psi(\theta) &= \theta^{-1}eQ^{-1}(I - \widehat{B}(\theta)P')^{-1}(I - \widehat{B}(\theta))\lambda\ . \end{aligned}$$

The use of these equations is illustrated in Section 4.1.

For supercritical systems which are initially empty, we have a more direct construction. This particular solution appears also in the asymptotics. The construction is based on the *global growth rate* of the population, given in the following Lemma.

LEMMA 4. *Let (α, ν, P) be a supercritical data. Then there exists a unique positive real number θ_0 solution to the equation:*

$$\theta_0 \ = \ e\ (I - \widehat{B}(\theta_0))(I - P'\widehat{B}(\theta_0))^{-1}\ \alpha\ . \tag{10}$$

Define the vector $m = (m_1, \ldots, m_K)'$ as:

$$m \ = \ (I - \widehat{B}(\theta_0))(I - P'\widehat{B}(\theta_0))^{-1}\ \alpha\ . \tag{11}$$

Given a supercritical data (α, ν, P), define $p_k : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, for each $k \in \mathcal{K}$ by

$$p_k(x) \ = \ \frac{m_k}{1 - \widehat{B}_k(\theta_0)} \int_x^\infty \theta_0 e^{-\theta_0(y-x)}\mathrm{d}B_k(y)\ , \tag{12}$$

and let $s_k \in \mathcal{M}$ denote the measure that is absolutely continuous with respect to Lebesgue measure and which Radon-Nikodym derivative is $p_k(.)$:

$$s_k(x) \ = \ p_k(x)\ \mathrm{d}x \qquad \text{for all } x \in \mathbb{R}_+. \tag{13}$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622

Note that $\int_0^\infty p_k(x)\mathrm{d}x = m_k$ and $\langle 1, s_k\rangle = m_k$. Finally, let $s := (s_1, \ldots, s_K)'$.

THEOREM 5. *Assume that (α, ν, P) is a supercritical data, and let θ_0 and s be given respectively by (10) and (11)–(13). Then the triple*

$$(A, D, \mu)(t) \quad (14)$$
$$= \; t \times \left((I - \widehat{B}(\theta_0))^{-1} m, (I - \widehat{B}(\theta_0))^{-1} \widehat{B}(\theta_0) m, s\right)$$

is the unique fluid solution of the model starting from the origin, that is, with $\mu(0) \equiv 0$. As a consequence, $Z(t) = mt$.

3.2.3 Asymptotic results

Asymptotics for the response time. The following asymptotic results concern the scalar function $T(\cdot)$. Since we know that this function is the same as in the single-class model, the results of [10] or [11] apply. Translated to the multiclass notation, we have:

THEOREM 6. *Given a data (α, ν, P) and $\xi \in \mathcal{M}^{c,p,K}$, we have:*

(i) If $\rho < 1$ then $\lim_{t\to+\infty} T(t) = \dfrac{e(\beta^0 + \beta QP')Z(0)}{1-\rho}$;

(ii) If $\rho = 1$ and $\beta_k^{(2)} < +\infty$ for all $k \in \mathcal{K}$, then $\dot{T}(t) \sim c_1$ and $T(t) \sim c_1 t$ as $t \to +\infty$, where:

$$c_1 \;=\; \frac{e(\beta^0 + \beta QP')Z(0)}{e(\frac{1}{2}\beta^{(2)} + \beta P'Q\beta)\lambda} \; . \quad (15)$$

(iii) If $\rho > 1$ then $T(t) \sim c_2 \exp(\theta_0 t)$ as $t \to +\infty$, where

$$c_2 \;=\; -\; \frac{\widehat{H}(\theta_0)}{\rho\theta_0 \dot{\widehat{V}}_e(\theta_0)} \; . \quad (16)$$

These results predict that the response time of a customer requesting σ units of service grows: linearly with σ in case (ii), and exponentially with σ in case (iii).

Asymptotics for the trajectories. We now state the asymptotic results concerning the trajectories of the fluid limit when $t \to \infty$.

In the subcritical case, the trajectories have a non-trivial evolution until $t = t_\rho(\xi)$ (the value of which is given in Lemma 2); for $t > t_\rho(\xi)$, the queue is empty: $\mu(t) = 0$, $Z(t) = 0$. Consequently, $A(t) = D(t) = \lambda t + QP'Z(0)$, according to Equation (4).

For the critical and supercritical cases, we have the following results.

THEOREM 7. *Given a critical data (α, ν, P) and $\xi \in \mathcal{M}^{c,p,K}$ and assuming that $\beta_k^0 < \infty$ and $\beta_k < \infty$ for all $k \in \mathcal{K}$, then, as $t \to \infty$,*

$$\mu_k(t)(.) \implies \frac{e(\beta^0 + \beta QP')Z(0)}{e(\frac{1}{2}\beta^{(2)} + \beta P'Q\beta)\lambda}\; \beta_k \lambda_k \nu_k^e(\cdot) \; .$$

If $\beta_j^{(2)} = +\infty$ for some $j \in \mathcal{K}$, the limit is 0.

THEOREM 8. *Given a supercritical data (α, ν, P) and $\xi \in \mathcal{M}^{c,K}$, there holds:*

$$\frac{\mu_k(t)}{t}(.) \implies s_k(.) \; .$$

As a consequence,

$$\lim_{t\to\infty} \frac{A(t)}{t} \;=\; \lambda - QP'm \qquad \lim_{t\to\infty} \frac{D(t)}{t} \;=\; \lambda - Qm$$

and

$$\lim_{t\to\infty} \frac{Z(t)}{t} \;=\; m \; .$$

This theorem predicts that populations in the queue grow asymptotically linearly. In the case where the queue is initially empty, the growth is actually exactly linear, according to Theorem 5. The growth rate of the population of class k is m_k, given in (11). From this equation, it is readily seen that $\alpha_k > 0$ if and only if $m_k > 0$. Therefore, *all* classes which have some external arrival will see their population grow to infinity as $t \to \infty$. Contrary to some service disciplines, for instance priority-based ones, it is not possible that some subset of classes behave in a stable manner whereas other classes are unstable.

3.2.4 The Discriminatory Processor Sharing queue

A natural generalization of multiclass processor sharing (egalitarian), commonly encountered in the literature, is the "discriminatory" processor sharing (DPS), where all customers present in the system are served simultaneously with rates controlled by a vector of weights $\{g_k > 0, k \in \mathcal{K}\}$. Under the DPS discipline, any individual customer of class k is served at a speed which is proportional to g_k. The service delivered to some customer of class k grows therefore as

$$\frac{g_k}{\sum_{k\in\mathcal{K}} g_k Z_k(t)} \;=\; \frac{g_k}{g.Z(t)} \; .$$

Since $Z_k(t) = \langle 1, \mu_k(t)\rangle$, it is clear that the cumulative of service per customer of class k can be expressed as:

$$S_k(s,t) = \int_s^t g_k \varphi(\langle 1, g.\mu(u)\rangle)\mathrm{d}u \; , \quad (17)$$

and that the dynamics of the measure μ_k is:

$$\langle 1_{[x,\infty)}, \mu_k(t)\rangle \;=\; \langle 1_{[x,\infty)}(. - S_k(t)), \mu_k(0)\rangle \quad (18)$$
$$+ \int_0^t \langle 1_{[x,\infty)}(. - (S_k(s,t)), \nu_k\rangle \mathrm{d}A_k(s) \; .$$

When all the weights are multiplied by some scalar γ, the dynamics are not changed since $\langle 1, \gamma g.\mu(s)\rangle = \gamma\langle 1, g.\mu(s)\rangle$ which implies $\gamma g_k \varphi(\langle 1, \gamma g.\mu(s)\rangle) = g_k\varphi(\langle 1, g.\mu(s)\rangle)$. In particular, when all weights are equal, this coincides with the equations of the multiclass egalitarian PS system.

This fluid queueing model is described by the data (α, P, ν, g). Definition 1 naturally extends to this system: we shall call a DPS Fluid Solution a triple of vector functions and measures $(A(t), D(t), \mu(t))$ that satisfy Equations (4)–(5) and (17)–(18). Such a DPS Fluid solution can be constructed

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622

from an equivalent (egalitarian) PS Fluid solution with the following transformations.

Let G be the diagonal matrix obtained from g. Define (α^g, P^g, ν^g) by

$$\left\{ \begin{array}{ll} \alpha^g & = G\alpha \\ P^g & = GPG^{-1} \\ \nu_k^g(\cdot) & = \nu_k(g_k \times \cdot) \ . \end{array} \right. \tag{19}$$

This transformation actually consists in multiplying, for each class, the external arrival rate by the weight, while dividing service times by the same factor. It would be interesting to see how the same transformation for the discrete, stochastic DPS queue could be exploited.

Now consider a triple $(A^g(t), D^g(t), \mu^g(t))$ of vector functions and measures as usual. Define then the transformed functions A, D and measures μ by:

$$\left\{ \begin{array}{ll} A(t) & = G^{-1}A^g(t) \\ D(t) & = G^{-1}D^g(t) \\ \mu_k(\cdot)(t) & = \frac{1}{g_k}\ \mu_k^g(\frac{1}{g_k} \times \cdot)(t) \ . \end{array} \right. \tag{20}$$

REMARK 1. *Observe that the triple (α^g, P^g, ν^g) may not be a valid data for an Egalitarian PS queue, since it may be that some entries in P are larger than 1. However, it is always true that P^g is a positive matrix with the same spectral radius as P. We conjecture that the results for the fluid process do hold even if P is not sub-stochastic, under the condition $\rho(P) < 1$, and although the interpretation of the entries p_{ij} as routing probabilities does not necessarily hold.*

The characteristics of the transformed (Egalitarian) PS queue are easily derived from that of the original queue. In particular, we have:

$$Q^g \ = \ GQG^{-1}, \qquad \lambda^g \ = \ G\lambda$$
$$\beta^g \ = \ \beta G^{-1} \ = \ G^{-1}\beta \qquad \beta^{(2),g} \ = \ \beta^{(2)}G^{-2} \ = \ G^{-2}\beta^{(2)} \ .$$

In particular, the load factor of the queue is $\rho^g = e.\beta^g.\lambda^g = e.\beta.\lambda = \rho$. The transformation (19) has the property of multiplying by the weight both the external and the internal arrival rates to class k. Combined with the division of service times in class k by the same factor, the load due to class k remains the same.

PROPOSITION 9. *Assume that the triple (α^g, P^g, ν^g) is a valid data. The triple $(A^g(t), D^g(t), \mu^g(t))$ is a solution of the egalitarian Fluid model with data (α^g, P^g, ν^g) defined in* (19)*, and an initial state described by the measures $\mu_k^g(\cdot)(0) = (g_k)^{-1}\mu_k((g_k)^{-1} \times \cdot)(0)$, if and only if the triple $(A(t), D(t), \mu(t))$ defined by* (20) *is a DPS Fluid solution for the DPS model, and initial measure $\mu(0)$.*

This result allows to compute the fluid trajectories for the DPS queue, by applying the formulas of Theorem 3 with the matrices obtained with (19). This provides the functions A^g, D^g and μ^g. The trajectories for the DPS are then obtained with (20). In particular, the total population of class k is obtained as $Z(t) = \langle 1, \mu_k(t) \rangle = g_k^{-1} Z_k^g(t)$. These calculations are illustrated in Section 4.1.

We conclude this section on the DPS with a qualitative observation. Let $T_k(s) = S_k^{-1}(s)$. We have seen that the function T_k is related with the computation of response times; here, there is one specific function for each class. The transformation from the egalitarian to the discriminatory queue and the identity (17) provide the following comparison results for response time mappings.

PROPOSITION 10. *Let (α, ν, P) be a given data, let $g = \{g_k > 0, k \in \mathcal{K}\}$ be a vector of weights and ξ be a non-zero initial state. If $g_k \geq g_l$, then for all $t \geq 0$: $T_k(t) \leq T_l(t)$.*

This result is to be expected, since it has been proved in [2] that in the (discrete) $M/GI/1/DPS$ queue (without reentries), steady state response times, conditioned on the service duration t, are stochastically ordered in the opposite directions of weights.

4. ILLUSTRATIONS

4.1 Trajectories

We illustrate in this section the effective construction of trajectories, in a case where most computations can be performed in closed form.

Consider a queue with two classes. Customers of class 1 have no external arrivals ($\alpha_1 = 0$), their service is distributed as $\text{Exp}(\mu_1)$ and when they complete service, they turn into customers of class 2: $P_{12} = 1$. Customers of class 2 arrive from the exterior with rate α. Their service is distributed as $\text{Exp}(\mu_2)$ and when they complete service, they exit the system: $P_{21} = P_{22} = 0$. The initial situation is that there is one unit of fluid of class 1, and no fluid of class 2: $Z(0) = (1, 0)'$. Service times of customers present initially have the same distribution as that of regular customers. We shall consider the Discriminatory PS queue with weights $g_1 = 1$ and $g_2 = G$. The Egalitarian PS queue is obtained with $G = 1$.

The first step is to compute the function $T(\cdot)$. The value of its Laplace Transform is provided below Theorem 3. We apply these formulas with the matrices resulting from the transformation (19). The resulting expression is:

$$\widehat{T}(\theta) \ = \ \frac{\theta + G(\mu_1 + \mu_2)}{(\theta + \mu_1)(\theta + G(\mu_2 - \alpha))} \ , \tag{21}$$

and the inversion of the Laplace transform gives:

$$T(t) \ = \ \frac{\mu_1 + \mu_2}{\mu_1(\mu_2 - \alpha)} \ + \ \frac{(\mu_1(G-1) + G\mu_2)e^{-\mu_1 t}}{\mu_1(\mu_1 + G(\alpha - \mu_2))} + \ \frac{(\alpha + \mu_1)e^{G(\alpha - \mu_2)t}}{(\alpha - \mu_2)(\mu_1 + G(\alpha - \mu_2))} \ .$$

Observe that $\widehat{T}(0) = T(+\infty) = t_\rho$ does not depend on G. This is natural, since the time range of the solution depends only on the workload of the system, and not on the service weights.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622

The next step is to compute the function $S(s)$ by solving $s = T(t)$, which cannot always be done in closed form, even in this simple case. We proceed with specific values.

A stable queue. We consider first the case where $\mu_1 = 1/4$ and $\mu_2 = 1$, and the arrival rate is $\alpha = 1/2$. In that case, the load factor is $\rho = 1/2$. Referring to Section 3.2.1, the "virtual workload" due to initial customers is

$$e(\beta^0 + \beta QP')Z(0) = \begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}4 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}\begin{pmatrix}1\\ 0\end{pmatrix} = 5.$$

According to Lemma 2, the time range of the solution is therefore $t_\rho = 5/(1 - 1/2) = 10$.

The solution of the fluid model in the Egalitarian PS case ($G = 1$) is given by:

$$\begin{aligned} T(t) &= 10 - 16e^{-t/4} + 6e^{-t/2} \\ S(s) &= -4\log(\frac{4}{3} - \frac{\sqrt{4+6s}}{6}) \\ Z_1(s) &= \frac{4}{3} - \frac{\sqrt{4+6s}}{6} \\ Z_2(s) &= 3\ Z_1(s)\ (1 - Z_1(s))\ . \end{aligned}$$

One checks that $t_\rho = \lim_{t\to\infty} T(t)$ and $\lim_{s\to t_\rho} S(s) = +\infty$.

Figure 1 displays one trajectory obtained by simulation with an initial population of $r = 1000$ customers and periodic arrivals (all simulations presented in the paper have these parameters), together with the properly scaled fluid trajectories. Simulations with a smaller initial population exhibit significant differences with the fluid trajectory, certainly due to the fact that the approximation of the initial random workload by its fluid counterpart has a bad precision then.

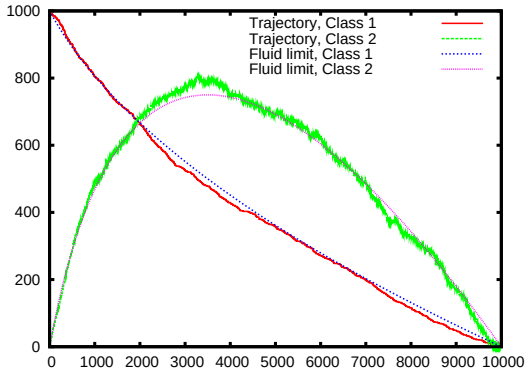

Figure 1: Trajectories in a stable case

In order to assess the precision of the fluid approximation, we have collected statistics on the maximal distance between real and fluid trajectories. With the notation of Section 3.1, we define the random variable:

$$\begin{aligned} M_k^{r,T} &= \max_{0\le t\le T} |rZ_k(t/r) - Z_k^r(t)| \\ &= r \max_{0\le u\le T/r} |Z_k(u) - \bar{Z}_k^r(u)|\ . \end{aligned}$$

We know that $M_k^{r,rT_0}/r \to 0$ for every T_0. The question is: how fast? Figure 2 displays the empirical distribution of $M_k^{r,T}$ for class $k = 2$ in the same situation as in Figure 1. We have selected $T = 1.1 \times r \times t_\rho = 11\ r$ (to account for some "overshooting" of simulated trajectories with respect to the theoretical t_ρ) and geometrically increasing values of r. The data was obtained using 1000 independent replications of the simulation for each value of r. Figure 3 displays the mean and standard deviation for the same experiment. It is clear from that figure that both quantities are proportional to $r^{1/2}$: a scaling compatible with some "central limit"-like result. However, the empirical distribution of $M^{r,T}$ does not match a Normal distribution with the same two first moments (at least for the values of r we have considered) as can be seen from Figure 2 (the curve labeled "Normal").

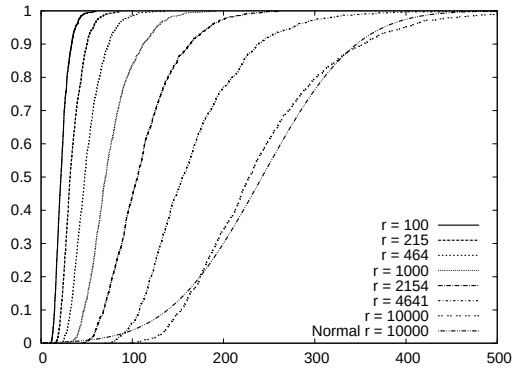

Figure 2: Distribution of the metric $M_1^{r,T}$ for increasing r (curves from left to right)

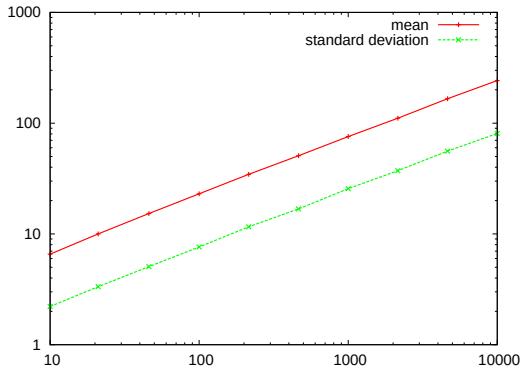

Figure 3: Mean and standard deviation of the metric $M_1^{r,T}$ for increasing r

A stable DPS queue. Let us assume now that $G = 2$. Through numerical inversion, we obtain the values of $S(\cdot)$ and $Z_i(t)$, which are represented in Figure 4, together with simulated trajectories.

The effect of the increased weight on the population of customers of class 2 is clear. Of course, increasing further the weight will have the effect of protecting external arrivals from the slowness due to the initial workload.

It is interesting to observe that the fluid model accurately predicts the actual trajectory, although the data of the problem is not valid in the sense of Section 3.2.4. Indeed, the

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622

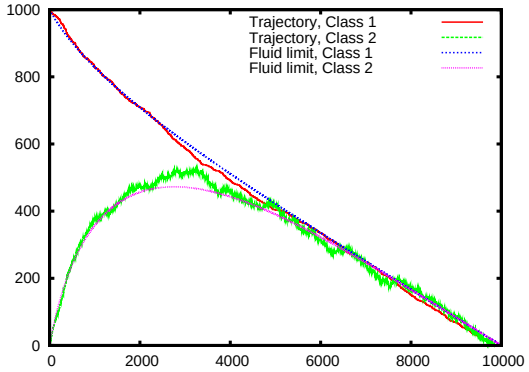

Figure 4: DPS Trajectories in a stable case

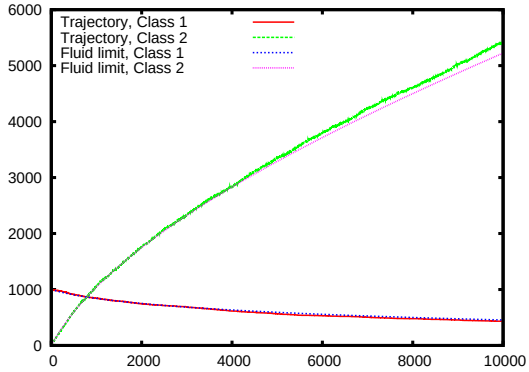

Figure 5: Trajectories in an unstable case

matrix $P^g = GPG^{-1}$ constructed using (19) is here:

$$P^g = \begin{pmatrix} 0 & 0 \\ G & 0 \end{pmatrix} .$$

It is therefore not a sub-stochastic matrix, but it has a spectral radius less than 1. This numerical observation and other "common sense" arguments have led us to the conjecture expressed in Remark 1. The investigation of this issue is however beyond the scope of the present paper.

An unstable queue. For the same values of μ_1 and μ_2, when the arrival rate is $\alpha = 5/4$, the load factor is $\rho = 5/4$. When $G = 1$, the solution is given by:

$$\begin{aligned} T(t) &= 12e^{t/4} + 8e^{-t/4} - 20 \\ Z_1(s) &= \frac{5}{4} + \frac{s}{16} - \frac{\sqrt{16+40s+s^2}}{16} \\ Z_2(s) &= 3\left(\frac{1}{Z_1(s)} - Z_1(s)\right) . \end{aligned}$$

The trajectories are displayed in Figure 5. The asymptotic growth rate of the population of class 2 is predicted by Theorem 8 and Lemma 4. The non-negative value which solves Equation (10) is $\theta_0 = \alpha - \mu_2 = 1/4$. This value can also be obtained from the explicit expression for $Z_2(s)$: an asymptotic expansion gives

$$Z_2(s) = \frac{s}{4} + 5 + O(s^{-1}) .$$

The accuracy of this asymptotic approximation is rather poor in this case. This means in practice that using the asymptotic formula $Z_k(t) \sim m_k t$ as an approximation may not be sufficient for a good prediction, even in relative terms.

The critical case. The critical case $\rho = 1$ can be solved for arbitrary μ_1 and G. Assume that $\alpha = \mu_2 = 1$. In that case, the inversion of the Laplace transform (21) gives:

$$T(t) = \frac{1+\mu_1}{\mu_1} Gt - \frac{G+\mu_1(G-1)}{\mu_1^2}(1 - e^{-\mu_1 t}) .$$

The function T is asymptotically linear, as predicted by Theorem 6 in this case. Observe the dependence of the asymptotic slope on the factor G. Continuing with $G = 1$ in order to simplify a bit the formulas, the solution of $T(t) = s$ gives:

$$S(s) = \frac{1}{\mu_1}\left(W\left(-\frac{e^{-Y}}{1+\mu_1}\right) + Y\right)$$

where $Y := (1+s\mu_1^2)/(1+\mu_1)$ and $W(\cdot)$ is Lambert's function defined by: $W(z)e^{W(z)} = z$. Finally, the populations in each class are given by:

$$\begin{aligned} Z_1(s) &= -(1+\mu_1)W\left(-\frac{e^{-Y}}{1+\mu_1}\right) \\ Z_2(s) &= \frac{1+\mu_1}{\mu_1}(1 - Z_1(s)) . \end{aligned}$$

In particular, we have as $s \to \infty$, $Z_1(s) \to 0$ and $Z_2(s) \to 1 + 1/\mu_1$. It can be verified that this is the value predicted by Theorem 7 for this situation. Simulated trajectories and the fluid limit are compared in Figure 6. The empirical and theoretical curves for class 1 are almost superimposed. For class 2, the empirical trajectory exhibits random oscillations which do not seem to vanish, for the time scale used in the diagram. Gromoll [5] has investigated this sort of phenomenon for the single-class case. The generalization to the multiclass case is currently under way.

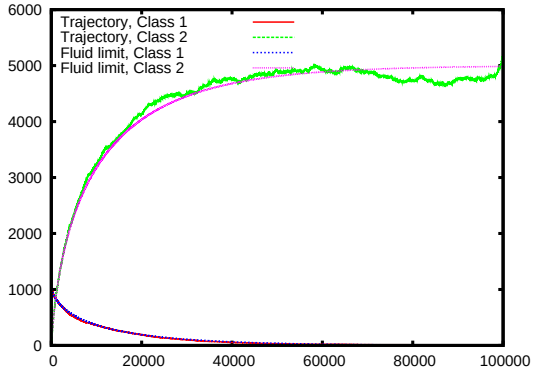

Figure 6: Trajectories in a critical case

4.2 Slowness as a function of the service time

Consider a single-class, processor sharing with service time distribution v. Assume that this distribution is discrete,

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622

with $\mathbb{P}(v = k\delta) = \pi_k$, $\sum_{k=1}^{\infty} \pi_k = 1$, for some parameter $\delta > 0$. Denote also $f_k = \sum_{j=k}^{\infty} \pi_k = \mathbb{P}(v \geq k\delta)$.

According to the discussion in Section 2, this single-class queue can be seen as a multiclass queue. Let k denote the class of customers having their service time equal to $k\delta$ in the single-class queue. In the multiclass queue, such customers are considered to have a service time equal to δ, and a routing probability $P_{k,k+1} = \mathbb{P}(v \geq (k+1)\delta | v \geq k\delta) = f_{k+1}/f_k$. If the support of the distribution is not bounded, there is an infinite number of classes. For the discussion to follow, we shall informally consider that the results of Section 3 apply with infinitely many classes. For a rigorous discussion, we might as well truncate the distribution, then let the truncation threshold go to infinity.

In the multiclass view of the system, the matrix $Q = (I - P')^{-1}$ is given by:

$$Q = \begin{pmatrix} 1 & 0 & 0 & \dots \\ f_2/f_1 & 1 & 0 & \dots \\ f_3/f_1 & f_3/f_2 & 1 & \ddots \\ \vdots & & & \ddots \end{pmatrix},$$

in other words, $Q_{ij} = f_i/f_j$ for $i \geq j$, 0 otherwise. Observe that $f_1 = 1$. The vector of external arrival rates is $(\alpha, 0, 0, \dots)'$, and the vector of (theoretical) global arrival rates is, as expected by construction, $\lambda = Q\alpha = \alpha(1, f_2, f_3, \dots)$.

Assuming that the system is supercritical, and following Lemma 4, Equation (10), we define θ_0 as the solution to the equation:

$$\theta_0 = \alpha \left(1 - \sum_{k \geq 0} \pi_k e^{-k\delta\theta_0}\right),$$

and apply (11) to obtain the value of the vector m. We have simply $\widehat{B}(\theta_0) = e^{-\delta\theta_0} I$, and it is easily seen that:

$$\begin{aligned} (I - P'\widehat{B}(\theta_0))^{-1}\alpha &= \alpha(1, f_2\widehat{B}(\theta_0), f_3\widehat{B}(\theta_0)^2, \dots) \\ &= \alpha(1, f_2 e^{-\delta\theta_0}, f_3 e^{-2\delta\theta_0}, \dots). \end{aligned}$$

Accordingly, applying Theorem 8, we obtain for the asymptotic growth rate, arrival rate and departure rate of customers of class k, respectively:

$$\begin{aligned} m_k &= \alpha f_k \left(1 - e^{-\delta\theta_0}\right) e^{-(k-1)\delta\theta_0}, \\ a_k &= \alpha f_k \, e^{-(k-1)\delta\theta_0}, \quad d_k = \alpha f_k \, e^{-k\delta\theta_0}. \end{aligned}$$

This shows that the departure rate of customers is decreased *exponentially* as a function of their service length. The factor of this exponential decay is the factor θ_0. This is in accordance with prior findings that the response time of customers grows exponentially with their service time. This is also in accordance with the fact that response times grow exponentially fast with the service requirement.

4.3 Competition between classes

We illustrate here how the processor sharing discipline "distorts" the throughputs of classes, in the case of overload. Consider a multiclass queue in which customers of class k arrive with a rate α_k, receive service, then leave the system. The routing matrix is $P' = 0$.

The reference situation is that the available service capacity is "fairly" shared among classes, proportionally to their load factor $\rho_k = \alpha_k \beta_k$. This situation is that of a stable server whatever its conservative service discipline, and that of an overloaded FIFO queue.

In the supercritical case, the "fair" situation is therefore that the throughput of class k is α_k/ρ, and the accumulation rate of customers in the queue is $\alpha_k(1 - 1/\rho)$. The workload accumulation rate is $\rho_k(1 - 1/\rho)$. In particular, if arrival rates are equal, customer throughputs and accumulation rates are equal. If, in the PS queue, one class accumulates faster (or has a smaller throughput) than some other one, arrival rates being equal, we say it is unfairly treated.

The following experiments will show that unfairness (in this sense) arises naturally when several classes with different service requirements are present. In order to demonstrate that the situation cannot be reduced to the comparison of moments/averages, or the comparison of tails of distributions, we have performed the following experiments. We study first the case of service distributions of the same "family" with short (exponential) tails. Then we study the case of service distributions with the same mean but different tails.

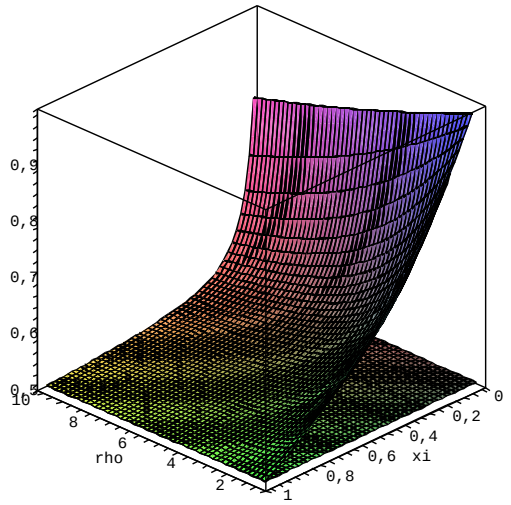

Figure 7: Proportion of customers of class 1 in queue, exponential distributions

The effect of service length. Consider an example with two classes, in which the distributions have the same "shape", the same arrival rate, but not the same average service time. Figure 7 represents the proportion of customers of class 1 in queue, as a function of ρ and $\xi = \mu_2/\mu_1$. The range $[1, 10]$ for ρ has been chosen to illustrate the global behavior of the functions represented; of course, such values are not supposed to occur in practice.

The effect of the distribution. Consider now an example with two classes, in which the distributions have the same arrival and average service time (hence the same individ-

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622

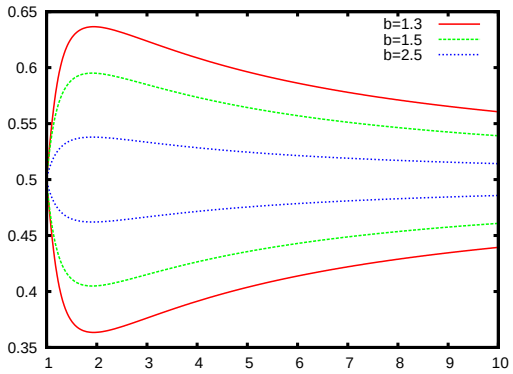

Figure 8: Proportion of customers of class 1 (Exponential distribution, upper curves) & 2 (Pareto distribution, lower curves)

ual load factor), but not the same "shape". The service for the first class has an exponential distribution, whereas the service for the second class has a Pareto distribution $\mathbb{P}(\sigma > t) = (a/(a+t))^b$. Figure 8 represents the proportion of customers of class 1 & 2 in queue, as a function of ρ, for different Pareto shape parameters b.

Finally, when customers of class 2 have a constant service time, the proportions are as shown in Figure 9. The figures show that customers with an exponentially distributed service times accumulate faster when confronted with customers with a Pareto distribution, but win the competition against customers with deterministic service times. Fairness in the sense above would require that proportions be equal. The diagrams also show that the proportions do not behave monotonically as a function of the global load factor ρ, and the interaction of classes and the residual service times corresponding to different distributions remains a complex issue.

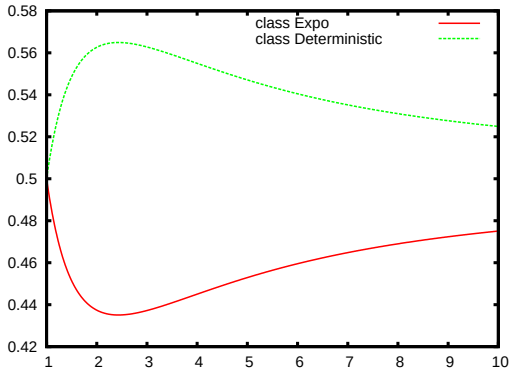

Figure 9: Proportion of customers of class 1 (Exponential distribution) & 2 (Deterministic)

5. CONCLUSION

We have given a tour of results available for the fluid approximation to the multiclass PS queue, both for the egalitarian and the discriminatory versions. The effective construction of trajectories has been demonstrated. Cases where explicit formulas can be obtained are quite rare, but numerical solutions have been obtained as well. It remains to study more complex cases with a larger number of classes and/or more complex service distributions (in particular, with heavy tails): the numerical efficiency of Laplace transform-based computations has to be checked more thoroughly.

Other possibilities are offered by the results contained in this paper. To detail just one: the fact that a single-class queue can be seen as a multiclass one through the decomposition of service durations, coupled with the use of the discriminatory service discipline, may provide a new analysis of scheduling disciplines based on service durations.

6. REFERENCES

[1] K. Athreya and K. Rama Murthy. Feller's renewal theorem for systems of renewal equations. *J. Ind. Inst. Sci.*, 58(10):437–459, 1976.

[2] K. Avrachenkov, U. Ayesta, P. Brown, and R. Nuñez Queija. Discriminatory processor sharing revisited. In *Proc. INFOCOM'2005*, volume 2, pages 784–795, Miami, FL, USA, 2005.

[3] A. Ben Tahar and A. Jean-Marie. The fluid limit of the multiclass processor sharing queue. Research Report RR6867, INRIA, April 2009. http://hal.inria.fr/inria-00368246_v2/.

[4] H. Chen, O. Kella, and G. Weiss. Fluid approximations for a processor sharing queue. *Queueing Systems Theory Appl.*, 27:99–125.

[5] H. C. Gromoll. Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab.*, pages 555–611, 2004.

[6] H. C. Gromoll, A. L. Puha, and R. J. Williams. The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.*, pages 797–859, 2002.

[7] H. C. Gromoll, P. Robert, and B. Zwart. Fluid limits for processor-sharing queues with impatience. *Math. Ops. Res.*, 33(2):375–402, 2008.

[8] H. C. Gromoll and R. Williams. Fluid limits for networks with bandwidth sharing and general document size distributions. *Ann. Applied. Prob.*, 10(1):243–280, 2009.

[9] A. Jean-Marie and P. Robert. On the transient behavior of the processor sharing queue. *Queueing Systems: Theory Appl.*, 17:129–136, 1994.

[10] A. L. Puha, A. L. Stolyar, and R. J. Williams. The fluid limit of an overloaded processor sharing queue. *Math. Ops. Res.*, 31(2):316–350, May 2006.

[11] A. L. Puha and R. J. Williams. Invariant states and rates of convergence for the fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.*, 14:517–554, 2004.

[12] S. F. Yashkov and A. S. Yashkova. Processor sharing: A survey of the mathematical theory. *Automation and Remote Control*, 68(9):1662–1731, 2007.

[13] J. Zhang, J. Dai, and B. Zwart. Law of large number limits of limited processor sharing queues. *Math. Ops. Res. to appear*, June 2008. Tech. Rep. Georgia Institute of Technology. `http://www2.isye.gatech.edu/people/faculty/dai/publications/draft_zhangDaiZwart08.pdf`.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2009.7622
http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7622