Population Effects in Multiclass Processor Sharing Queues

Abdelghani Ben Tahar\textsuperscript{1} Alain Jean-Marie\textsuperscript{2}

\textsuperscript{1}University Hassan I, Settat, Marocco
previously at the University of Rouen
Research funded by a CNRS Post-doctoral grant

\textsuperscript{2}INRIA
LIRMM CNRS/Univ. Montpellier 2

Valuetools 2009, 20 October 2009
Outline

1. Introduction
   - The model
   - The questions
   - The literature
   - This talk

2. The fluid model
   - Fluid equations
   - Convergence results
   - Existence results
   - Asymptotic results

3. Results for the DPS

4. Illustrations
   - Trajectories
   - Proportions of populations
1 Introduction
   - The model
   - The questions
   - The literature
   - This talk

2 The fluid model
   - Fluid equations
   - Convergence results
   - Existence results
   - Asymptotic results

3 Results for the DPS

4 Illustrations
   - Trajectories
   - Proportions of populations
The Model

Customers arrive to a single-server, infinite buffer, Processor Sharing station.

They belong to classes which govern their service requirement and routing behavior.

When they complete some service, they may re-enter the queue, possibly with a different class, according to specified probabilities (Jackson-like routing).
The Questions

Questions we are interested in are related with *fluid limits*: When either:

- the number of customers initially present
- or time

goes to infinity, *normalized* quantities

- of customers
- of work in progress

evolve according to deterministic differential equations.

As an application, we are interested in *fairness* issues between classes.
Related Literature

The Processor Sharing queue has a long history: see e.g. Yashkov & Yashkova (2007). However, multiclass results are rare.
The Discriminatory Processor Sharing has been actively studied lately, due to its applications in networking: see e.g. Avratchenko & et al., Altman et al. (2005).
Fluid limits of the Processor Sharing queue have been studied before. Partial bibliography:

In single-class

- Robert and Jean-Marie (1994)
- ...

With variants

- customers are impatient: Gromoll et al. (2006)
- server accepts a limited number of customers: Zhang et al. (2008)
- ...

Our analysis is a generalization of Puha, Stolyar and Williams (2006).
The talk is aimed at:

- provide an overview of fluid results for the multiclass PS queue, including DPS
- provide some illustrations
Introduction
- The model
- The questions
- The literature
- This talk

2 The fluid model
- Fluid equations
- Convergence results
- Existence results
- Asymptotic results

3 Results for the DPS

4 Illustrations
- Trajectories
- Proportions of populations
Model Parameters

The principal parameters of the model are:

- the vector of *external* arrival rates $\alpha$
- the routing matrix $P$
- the service time distribution in class $k$: $\sigma_k$ with
  - measure $\nu_k$
  - average $\beta_k$
  - Laplace transform $\hat{B}_k(\cdot)$

Additional notation in the text

- the “vector of ones”, $e$
- $Q = (I - P')^{-1}$
- vectors and diagonal matrices of per-class quantities
State representation

The state of the queue at time $t$ is represented by:

- $A_k(t)$, amount of (fluid) arrivals of class $k$ up to $t$
- $D_k(t)$, amount of (fluid) departures of class $k$ up to $t$
- $\mu_k(t)$, distribution of residual workload in class $k$

In particular:

- $Z_k = \langle 1, \mu_k \rangle$ is the quantity of customers of class $k$ in queue
- $\langle 1_{[x,\infty]}, \mu_k \rangle$ is the quantity of customers in queue with remaining service $\geq x$
- $\langle \text{id}, \mu_k \rangle$ is the total workload of customers of class $k$
Fluid equations

The input/output traffics satisfy:

\[ A(t) = \alpha t + P'D(t) . \]

For every \( k \in K, x \in \mathbb{R}_+, t \geq 0, \)

\[ \langle 1, \mu_k(t) \rangle = \langle 1, \mu_k(0) \rangle + A_k(t) - D_k(t) \]

\[ \langle 1_{[x, \infty)}, \mu_k(t) \rangle = \langle 1_{[x, \infty)}(. - S(0, t)), \mu_k(0) \rangle \]

\[ + \int_0^t \langle 1_{[x, \infty]}(. - (S(s, t))), \nu_k \rangle dA_k(s) , \]

where the cumulative service amount is:

\[ S(s, t) = \int_s^t \frac{1}{\langle 1, e.\mu(u) \rangle} du . \]
The queueing model

Consider a sequence of (discrete stochastic) Processor sharing queues indexed by $r$. They are described by:

- i.i.d. inter-arrival sequences for each class $\{u^r_k(n); n = 1, 2, \ldots\}$; the arrival rate is $\alpha^r_k$
- i.i.d. routing sequences $\varphi^r_{k\ell}$ with probabilities $p^r_{k\ell}$
- i.i.d. service time sequences,
- initial workload measures $\nu^0_r$.

Let $A^r_k(t)$ and $D^r_k(t)$ be the arrival and departure count processes, and $\mu^r_k(t)$ the residual workload measures. Define

$$\bar{A}^r(t) = \frac{A^r(rt)}{r}, \quad \bar{D}^r(t) = \frac{D^r(rt)}{r}, \quad \bar{\mu}^r(t) = \frac{\mu^r(rt)}{r}.$$
Convergence result

Theorem

Assume that:

- arrival rates $\alpha'_k$ and routing frequencies $\varphi'_{k\ell}$ converge
- initial measures converge in the appropriate sense.

Then the normalized quantities $(\bar{A}'_r, \bar{D}'_r, \bar{\mu}'_r)$ converges in distribution to the solutions of the fluid model with initial measure $\bar{\mu}_k(0)$ when $r \to \infty$.

Complete proof in Ben Tahar and Jean-Marie (2009).
There exists a unique solution to the system of fluid equations. It is given by

\[
A(t) = \lambda t + QP'(Z(0) - Z(t)) \\
D(t) = \lambda t + Q(Z(0) - Z(t)) \\
Z(t) = \tilde{Z}(T^{-1}(t)) \\
\tilde{Z}(s) = Q^{-1}(U \ast C)(s)Z(0) + Q^{-1}(U \ast (I - B) \ast (TI))(s)\lambda \\
T(s) = (H \ast U_e)(s)
\]

for some functions $C(\cdot), H(\cdot), U_e(\cdot), U(\cdot)$ directly constructed from the data.
According to the results of Gromoll, Puha, Williams:

**Lemma**

Let \((A(t), D(t), \mu(t))\) be a solution such that \(\mu(0) = \xi \neq 0\).

- Let \(t^* = \inf\{t : e.\mu(t) = 0\}\). Then,

\[
\begin{align*}
  t^* &= \begin{cases} 
    +\infty & \text{if } \rho \geq 1 \\
    \frac{e(\beta^0 + \beta QP')Z(0)}{1 - \rho} & \text{if } \rho < 1
  \end{cases}
\end{align*}
\]

- The function \(S : [0, t^*) \to [0, \infty)\) is well defined and strictly increasing. So is \(T \equiv S^{-1} : [0, \infty) \to [0, t^*)\).
Existence of solutions, supercritical case

Lemma

Assume that the queue is supercritical (\( \rho > 1 \)). Then there exists a unique positive real number \( \theta_0 \) solution to the equation:

\[
\theta_0 = e (I - \hat{B}(\theta_0))(I - P'\hat{B}(\theta_0))^{-1} \alpha .
\]

The Laplace transform in the RHS is that of the service of a “typical customer”.

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

\[
m = (I - \hat{B}(\theta_0))(I - P'\hat{B}(\theta_0))^{-1} \alpha .
\]

This \( \theta_0 \) is the global growth rate of the population. The vector \( m \) is its repartition among classes.
Existence of solutions, supercritical case (ctd.)

Define \( p_k : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \), for each \( k \in \mathcal{K} \) by

\[
p_k(x) = \frac{m_k}{1 - \hat{B}_k(\theta_0)} \int_x^\infty \theta_0 e^{-\theta_0(y-x)} dB_k(y),
\]

and let \( s_k \in \mathcal{M}_+ \) be the measure:

\[
s_k(x) = p_k(x) \, dx.
\]

**Theorem**

Assume that the system is supercritical, and let \( \theta_0 \) be as above. Then the triple

\[
(A, D, \mu)(t) = t \times \left( (I - \hat{B}(\theta_0))^{-1} m, (I - \hat{B}(\theta_0))^{-1} \hat{B}(\theta_0)m, s \right)
\]

is the unique fluid solution of the model starting from the origin, that is, with \( \mu(0) \equiv 0 \). As a consequence, \( Z(t) = mt \).
Assume that the data is supercritical ($\rho > 1$).

**Theorem**

*Given a supercritical data $(\alpha, P, \nu)$ and $\xi \in \mathcal{M}^{c,K}$, there holds:*

$$\lim_{t \to \infty} \frac{\mu_k(t)}{t}(.) \implies s_k(.) .$$

*As a consequence,*

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

These properties follow from the result of Athreya and Rama Murthy (1976) on systems of renewal equations.
Progress

1. Introduction
   - The model
   - The questions
   - The literature
   - This talk

2. The fluid model
   - Fluid equations
   - Convergence results
   - Existence results
   - Asymptotic results

3. Results for the DPS

4. Illustrations
   - Trajectories
   - Proportions of populations
Extension to the DPS

Under the Discriminatory Processor Sharing discipline with weights \((g_k)_k\), the service delivered to some customer of class \(k\) grows therefore as

\[
\frac{g_k}{\sum_k g_k Z_k(t)} = \frac{g_k}{g \cdot Z(t)}.
\]

Then the cumulative of service per customer of class \(k\) can be expressed as:

\[
S_k(s, t) = \int_s^t \frac{g_k}{\langle 1, g, \mu(u) \rangle} \, du.
\]

The dynamics of the measure \(\mu_k\) becomes:

\[
\langle 1_{[x, \infty)}, \mu_k(t) \rangle = \langle 1_{[x, \infty)}(\cdot - S_k(t)), \mu_k(0) \rangle
\]

\[
+ \int_0^t \langle 1_{[x, \infty)}(\cdot - (S_k(s, t)), \nu_k \rangle dA_k(s).
\]
Let $G$ be the diagonal matrix obtained from $g$.

**Theorem**

The DPS Fluid solution can be constructed from an equivalent (egalitarian) PS Fluid solution with the following data $(\alpha^g, P^g, \nu^g)$:

\[
\begin{align*}
\alpha^g &= G\alpha \\
P^g &= GPG^{-1} \\
\nu^g_k(\cdot) &= \nu_k(g_k \times \cdot).
\end{align*}
\]

Observe that $GPG^{-1}$ is not necessarily stochastic, but has spectral radius less than 1. In examples, everything works fine with this matrix!
Introduction
- The model
- The questions
- The literature
- This talk

The fluid model
- Fluid equations
- Convergence results
- Existence results
- Asymptotic results

Results for the DPS

Illustrations
- Trajectories
- Proportions of populations
Trajectories, stable case

Two-class experiment:
- Class 1 customers have 0 arrival rate, and route to Class 2 Services Expo mean 4.0
- Class 2 customers have 1/2 arrival rate, route to the outside Services Expo mean 1.0
- Load is 1/2, initial workload normalized to 1

The solution is given by:

\[
T(t) = 10 - 16e^{-t/4} + 6e^{-t/2}
\]

\[
S(s) = -4 \log\left(\frac{4}{3} - \frac{\sqrt{4 + 6s}}{6}\right)
\]

\[
Z_1(s) = \frac{4}{3} - \frac{\sqrt{4 + 6s}}{6}
\]

\[
Z_2(s) = 3 Z_1(s) (1 - Z_1(s))
\]
Population Effects in the PS queue

Abdelghani Ben Tahar, Alain Jean-Marie

Introduction

The model

The questions

The literature

This talk

Model

Fluid equations

Convergence results

Existence results

Asymptotic results

Results for the DPS

Illustrations

Trajectories

Proportions of populations

Trajectories, stable case (ctd.)
Trajectories, stable case (ctd.)

Not always so lucky...
Trajectories with DPS, stable case

Same situation with $G = 2$:
Closeness of the approximation

How to quantify the closeness of the approximation?

$$\implies \text{maximum deviation between the real (random) trajectory } Z(t) \text{ and the fluid trajectory at the same scale.}$$

Define the random variable:

$$M_{k,T}^{r} = \max_{0 \leq t \leq T} |rZ_k(t/r)-Z_k^r(t)| = r \max_{0 \leq u \leq T/r} |Z_k(u)-\bar{Z}_k^r(u)|.$$
Closeness of the approximation

Empirical distribution of the metric $M_{1}^{r,T}$ for increasing $r$ in the previous example, class 2.
Mean and std. dev. of the metric $M_{1,T}^r$ for increasing $r$: compatible with a $\sqrt{r}$ scaling.
Trajectories, unstable case

Two-class experiment:

- Class 1 customers have 0 arrival rate, and route to Class 2
  Services Expo mean 4.0
- Class 2 customers have $\frac{5}{4}$ arrival rate, route to the outside
  Services Expo mean 1.0
- Load is $\frac{5}{4} > 1$, initial workload normalized to 1

\[
Z_1(s) = \frac{5}{4} + \frac{s}{16} - \frac{\sqrt{16 + 40s + s^2}}{16}
\]

\[
Z_2(s) = 3 \left( \frac{1}{Z_1(s)} - Z_1(s) \right)
\]
Trajectories, unstable case (ctd.)

- Trajectory, Class 1
- Trajectory, Class 2
- Fluid limit, Class 1
- Fluid limit, Class 2
Proportions of populations

Principle of the experiment:

- In a “fair” unstable queue, proportions of customers in queue should be proportional to the arrival rate. This happens for FIFO.
- In a PS queue, there is a bias:
  - influence of the mean service, even within the same family of one-parameter distributions
  - influence of the distribution, within distributions with identical means.

Illustration: two classes of customers with identical arrival rates.
Example 2: exponential and deterministic distributions with same mean.

Proportion of customers of class 1 & 2 in queue, as a function of $\rho$. 

Proportions of populations
Example 3: exponential and Pareto distributions with same mean.

Proportion of customers of class 1 (Exponential distribution, upper curves) & 2 (Pareto distribution, lower curves), as a function of $\rho$. Different Pareto shape parameters.
Questions?
Bibliography

References and surveys on the PS and DPS queues, and for the proofs


Fluid limits and overloaded Processor Sharing


Fluid limits, recent developments


Renewal theory


Proportions of populations

Example 1: exponential distributions with different means

Proportion of customers of class 1 in queue, as a function of $\rho$ and $\xi = \mu_2/\mu_1$. 