Annales Mathematicae et Informaticae
***39** (2012) pp. 309–323*

Proceedings of the Conference on Stochastic Models and their Applications
Faculty of Informatics, University of Debrecen, Debrecen, Hungary, August 22–24, 2011

Discriminatory processor sharing with access rate limitations

B. Székely, A. Kőrösi, P. Vámos, J. Bíró

Budapest University of Technology and Economics
biro@tmit.bme.hu

Dedicated to Mátyás Arató on his eightieth birthday

Abstract

In the access part of communication networks user access rates are usually limited by technology and are much lower than the bottleneck link transmission capacity carrying the traffic flows aggregated. A possible model for bandwidth sharing of the bottleneck link is the Discriminatory Processor Sharing (DPS) models, in which the server capacity (link bandwidth) is distributed among different classes of users in an unequal manner. Recently, some DPS variants incorporating the access rate limits of users have been analyzed. These models are not bandwidth sparing in a sense, that the capacity share of a class may simply be cut at its access rate limit, and the incidentally residual bandwidth is not reused in other classes. In this paper we introduce and analyze a novel variant of DPS in which the original processor sharing effect and the access rate limit constraints are combined in a bandwidth economical way resulting a truly capacity-conserving operation. Besides the state space characterization of this model, two asymptotic behaviors are also presented. We also argue in the favor of practical significance of these asymptotics, that is it could greatly help in finding high quality approximate solutions of this DPS system, i.e.q in terms of the average waiting times of flows.

1. Introduction

The original discriminatory processor sharing (DPS) model has been presented and analyzed first in [7] and [11] for modeling purposes of time-sharing computer operation. In this model there are K number of classes of users, and the state of the system can be attributed by n_i denoting the number of class-i ($i = 1, \ldots, K$) users in the C capacity processor sharing system. There is also a set of weights ϕ_i, $i = 1, \ldots, K$ which can be used to control the sharing of the processor capacity

among the classes of customers. More formally the (instantaneous) service rate of a class$-i$ customer is

$$c_i = \frac{\phi_i}{\sum_{j=1}^{K} \phi_j n_j} C. \tag{1.1}$$

In [6] Fayolle et al. proved the results for DPS with respect to the steady-state average response times. In [12] Rege and Sengupta showed how to obtain the moments of the queue length distributions as the solutions to linear equations in case of exponential service time requirements, and they also presented a heavy-traffic limit theorem for the joint queue length distribution. These results were extended to phase type distributions by van Kessel et al. [14]. A further remarkable milestone in DPS analysis is [1] in which the authors showed that the mean queue lengths of all classes are finite under reasonable stability conditions, regardless of the higher moments of the service requirements.

Introducing capacity limits for the customers is mainly motivated by involving access rate limitations of users (e.g. in DSL-type access systems) into the modeling framework. In [10] Lindberger analyzed the M/G/R-PS system, which is a single-class processor sharing model with access rate limit b on the users ($R := C/p$ is the "number of servers" in this system). Several improvements of this model were studied for dimensioning purposes of IP access networks, e.g. in [13] and [5] still remaining at the single-class models.

In case of multi-class discriminative processor sharing with limited access rates the question of bandwidth re-distribution is an important issue, which was not addressed in the literature. This means that if users in a class can not fully utilize their service capacity share (bandwidth share) due to their access rate limit, the problem is how this unused bandwidth is re-distributed among the other classes. In one of the extreme cases, there is no re-distribution at all meaning that the possible remaining unused bandwidth due to rate limits is wasted. One can also interpret this as the server capacity may not be fully utilized, even in those cases when there is "enough" customers in the system. This approach is followed for example in the papers [8], [2].

In this paper we present and analyze the capacity conserving case of access rate limited discriminatory processor sharing, in which all the unused bandwidth left by rate limited customers are fully utilized by the other (non-limited) customers. This is referred to as bandwidth economical discriminatory processor sharing with access rate limitations. We characterize the state space of this model, with identifying those traffic classes which are compressed (whose users are not able to utilize its access rates) and those which are not compressed (which can receive service with their access rates) and with feasible computations for their respective service rates. Two asymptotic regimes of this bandwidth economical DPS are shown and their equivalence is proven. We present that the asymptotic equilibrium point of the bandwidth efficient system is always in the non-compressed region and can simply be formulated (for every class of users), as opposed to the more complicated asymptotic equilibrium of the previously analyzed model [2].

The significance of the fluid limits lies in the following. There is still no solution

in the literature for the multi-class access rate limited DPS system (in case Poisson arrivals and exponential service time requirements the equilibrium of the underlying Markov chain, consequently, the expected response times are not known). Therefore, achieving high quality approximations of system parameters have an utmost importance, e.g. from viewpoint of dimensioning tasks of communication channels for elastic flows in aggregation part of access networks or of processing capacity in highly loaded computer systems like data centers [3]. One "extreme" type of access rate limited multi-class DPS is the limitless case (no compression imposed on the classes), for which Fayolle et al. have already given the solution [6] in terms of the steady-state average response times (by integro-differential equations), and also showed that in the special case of exponential distribution of the service time requirements, the steady-state average response times of classes can be obtained by solving a system of linear equations. The fluid limit is the other extreme case of this DPS system in the sense that some of (or all) classes are "infinitely" compressed (due to infinitely speed up the system), whilst the scaled down performance parameters remain (tend to) finite values. Operational systems to be modeled or dimensioned based on DPS models stand between these two extremes, surprisingly sometimes very close to the fluid limit.

2. DPS extended by the limits of service rates

In DPS for every pair of classes i, j the ratio of the service rates allocated to class-i and class-j users is equal to the ratio of the class weights (see formula (1.1)), that is

$$\frac{c_i}{c_j} = \frac{\phi_i}{\phi_j}, \ \forall i, j \in 1, \ldots, K. \tag{2.1}$$

The total amount of capacity (in a non-empty system) used by the users of classes is evidently C, i.e.

$$\sum_{i=1}^{K} n_i c_i = C. \tag{2.2}$$

Regarding the incorporation of access rate (customer service capacity share) limits into the DPS model, in [8] and [2] a very simple approach is followed. Namely, first computing the bandwidth shares of class-i users according to (1.1) and then cutting at the access rate limits p_i, i.e.

$$c_i = \min\left(\frac{\phi_i}{\sum_{j=1}^{K} \phi_j n_j} C, p_i\right). \tag{2.3}$$

The benefit of this bandwidth share calculation is its simplicity. Nevertheless the price for simplicity is that this approach is not a bandwidth saving one, because it may happen that the total amount of capacity used by the customers is smaller

then the server capacity (the server capacity is not completely shared among the users), i.e.

$$\sum_{i=1}^{K} n_i c_i < C \tag{2.4}$$

even in those cases when there are "enough" users in the system, that is

$$\sum_{i=1}^{K} n_i p_i > C. \tag{2.5}$$

In this paper we follow the other "extreme" approach, in which all the unused parts of capacity shares due to access rate limits are redistributed among users which are not imposed by these limits on. Because redistribution and sharing the whole capacity C is possible when $\sum_{i=1}^{K} n_i p_i > C$, hereafter we assume the system is in this regime. Otherwise, when $\sum_{i=1}^{K} n_i p_i \leq C$, the bandwidth shares are trivially $c_i = p_i$. In what follows we define our bandwidth economical DPS.

Definition 2.1. The bandwidth economical DPS is such a discriminatory processor sharing system in which the bandwidth shares c_i of the users of K classes at a given state $\mathbf{n} = \{n_1, \ldots, n_K\}$ are determined by the following equations:

$$c_i = \min\left\{p_i, \frac{\phi_i}{\phi_j} c_j\right\} \ \forall i, j \in \{1, \ldots, K\}, \ c_j < p_j \tag{2.6}$$

and

$$\sum_{i=1}^{K} n_i c_i = C \tag{2.7}$$

where p_i is the service rate limit of class$-i$ users, $0 < p_i \leq C$.

For the next lemma without loss of generality let us assume that

$$\frac{\phi_K}{p_K} \leq \frac{\phi_i}{p_i}, \forall i = 1, \ldots, K. \tag{2.8}$$

Lemma 2.2. *For class$-K$ users $c_K < p_K$ always holds.*

Proof. The proof is based on contradiction. Assume that $c_K = p_K$. Due to (2.6) and the assumption (2.8) above it follows that

$$c_i = \min\left\{p_i, \frac{\phi_i}{\phi_K} c_K\right\} = p_i, \ \forall i = 1, \ldots, K. \tag{2.9}$$

But in this case $\sum_{i=1}^{K} n_i c_i = \sum_{i=1}^{K} n_i p_i > C$ which contradicts to equation (2.7). □

In the next corollary we show the following statement:

Corollary 2.3. *There is a unique solution of equations (2.6) and (2.7) with respect to* c_i, $i = 1, \ldots, K$.

Proof. Because of Lemma 2.2 and (2.7) and the monotone increasing property of $\min\{p_i, \frac{\phi_i}{\phi_K} x\}$ w.r.t. x, a class$-K$ user bandwidth share is a unique solution of the equation

$$\sum_{i=1}^{K} n_i \min\left\{p_i, \frac{\phi_i}{\phi_K} x\right\} = C \tag{2.10}$$

with respect to x. Therefore, every other bandwidth share is also unique and can be calculated by using c_K and the equality

$$c_i = \min\{p_i, \frac{\phi_i}{\phi_K} c_K\}. \qquad \square$$

Let a numerical example be presented for this calculation. Let $C = 100$ [Mbit/s] and five classes (with index 1 to 5 in sequence) are set up with the following paramaters: $\mathbf{n} = (8, 15, 20, 10, 30)$, $\mathbf{p} = (2, 2, 1.5, 2, 10)$ [Mbit/s], $\phi = (10, 9, 5, 4, 1)$. The following table shows the ϕ_i/p_i ratios, the access rate limits p_i, the bandwidth shares in case of original DPS (without access rate limit), of DPS with access rate limit with simple cutting at the limits using formula (2.3), and the new bandwidth economical DPS according to equations (2.6) and (2.7).

class index	1	2	3	4	5
ϕ_i/p_i	5	4.5	3.33	2	1
p_i	2	2	1.5	2	10
orig. DPS	2.5974	2.3377	1.2987	1.0389	0.2597
equ (2.3)	2	2	1.2987	1.0389	0.2597
bw eco. DPS	2	2	1.5	1.3714	0.343

Table 1: Example of bandwidth shares of different DPS systems

The fifth line of the table clearly shows that in case of simple cutting DPS (using equation (2.3), or simple comparing the third and fourth lines of the table), the class-1 and class-2 users can utilize their access rates (they are uncompressed), while classes 3, 4 and 5 are compressed (they can not reach their access rates). It can also be observed that $\sum_{i=1}^{5} n_i c_i = 90.16$ Mbit/s, that is from the total capacity 100 Mbit/s almost ten percent is wasted.

On the contrary, the last row presenting the bandwidth share of the new DPS system shows, that not only class-1 and class-2 can achieve their access rate limits, but also class-3 became uncompressed, thanks to the redistribution[1] of the unused

[1]The term 'redistribution' is used because it can be shown that the following process results exactly the same solution: start with the original DPS bandwidth share, cut at the access rate limits, and redistribute the residual banwidths among the still compressed classes, which may result some classes become uncompressed. Repeat this until the bandwidth shares no longer change.

bandwidth left by class-1 and class-2 customers. Furthermore, class-4 and class-5 bandwidth shares are also higher than in the previous case, because they can also gain from bandwidth reuse. In this case, of course $\sum_{i=1}^{5} n_i c_i = 100$ Mbit/s, hence this is attributed as bandwidth economical.

Although the computational approach above is straightforward, it is worth exploring further the structure of the system. For this, let us assume again without restriction that

$$\frac{\phi_1}{p_1} \geq \frac{\phi_2}{p_2} \geq \ldots \geq \frac{\phi_K}{p_K} \ . \tag{2.11}$$

Lemma 2.4. *If $\sum_{i=1}^{K} n_i p_i > C$ there exists an i^*, $1 \leq i^* \leq K-1$ such that*

$$\sum_{k=1}^{i^*-1} n_k p_k + \sum_{k=i^*}^{K} n_k \phi_k \frac{p_{i^*}}{\phi_{i^*}} \leq C \text{ and} \tag{2.12}$$

$$\sum_{k=1}^{i^*} n_k p_k + \sum_{k=i^*+1}^{K} n_k \phi_k \frac{p_{i^*+1}}{\phi_{i^*+1}} > C \ . \tag{2.13}$$

Proof. Note that the function

$$f(i) = \sum_{k=1}^{i-1} n_k p_k + \sum_{k=i}^{K} n_k \phi_k \frac{p_i}{\phi_i}$$

is increasing w.r.t. i due to (2.11) and exceeds C for some $i^* + 1 \leq K$, otherwise $f(K) = \sum_{i=1}^{K} n_i p_i \leq C$ would hold which is not true. □

As an important consequence of this lemma it is also worth noting that

$$\{1, \ldots, i\} \subset \mathcal{U}(\underline{n}) \text{ iff } \sum_{k=1}^{i-1} n_k p_k + \sum_{k=i}^{K} n_k \phi_k \frac{p_i}{\phi_i} \leq C \tag{2.14}$$

where $\mathcal{U}(\underline{n}) := \{1, \ldots, i^*\}$ is the set of uncompressed classes in the state $\underline{n}$.

Now the main theorem of this section is the following:

Theorem 2.5. *The unique solution of (2.6) and (2.7) can be expressed through i^* in the following way:*

$$c_k = p_k, \text{ if } k \leq i^* \text{ and} \tag{2.15}$$

$$c_k = \frac{\phi_k}{\sum_{i=i^*+1}^{K} \phi_i n_i} \left(C - \sum_{j=1}^{i^*} n_j p_j \right), \text{ if } i^* < k. \tag{2.16}$$

Proof. The validity of (2.7) can easily be checked. Next we show that (2.6) is fulfilled by c_k, c_l for which $k, l \in \mathcal{Z}(\underline{n}) := \{1, \ldots, K\} \setminus \mathcal{U}(\underline{n})$. In this case due to (2.13) and (2.11) $c_k < p_k$ and $c_l < p_l$. Moreover $c_k/p_k = c_l/p_l$ holds, therefore (2.6) is satisfied, that is $c_k = \min\{p_k, \frac{\phi_k}{\phi_l} c_l\}$.

Now assume that $l \in \mathcal{U}(\underline{n})$ and $k \in \mathcal{Z}(\underline{n})$. In this case $c_k < p_k$, therefore

$$\frac{\phi_l}{\phi_k} c_k = \frac{\phi_l}{\sum_{i=i^*+1}^{K} \phi_i n_i} \left(C - \sum_{j=1}^{i^*} n_j p_j \right) \tag{2.17}$$

which is not less than p_l due to (2.12) and (2.11). Hence,

$$c_l = \min\{p_l, \frac{\phi_l}{\phi_k} c_k\} = p_l$$

that is (2.6) is again fulfilled. □

3. Asymptotic behaviors of the bandwidth economical DPS

In this section we first show that the so-called fluid limit of the processor sharing model investigated in this paper exists. Then we find the equilibrium of the fluid limit. The stability of this equilibrium has been also proved, however, not presented in this paper. Assume that the service times are exponentially distributed and the arrival processes follow Poisson processes. Then in this case the number of jobs (of customers) in the system can be modeled by a Markov chain. The equilibrium of the Markov chain, consequently, the expected response times are not known. Fluid scaling is a possible asymptotic regime in which one may expect computing the equilibrium at least for the limiting structure. In fluid limit the arrival processes are accelerated by a common factor and the capacity of the server is speed up by the same factor. If the accelerating factor goes to infinity then in limit one gets the fluid limit of the number of waiting jobs. The limiting process of the number of waiting jobs is deterministic, it is a solution of a differential equation. The equilibrium of this differential equation can be found using analytical considerations. We remark that the fluid limit of many processor sharing model, as well as the one investigated in this paper, can be determined by using classical results presented in e.g. [4, Chapter 11].

For finding the fluid limit of our model first the transition rates are to be determined $q(\underline{n}, \underline{n} + \underline{l})$ from state $\underline{n}$ to $\underline{n} + \underline{l}$. Let $\underline{e}_k$ be a vector such that in $\underline{e}_j$ 1 stands at coordinate j and except this coordinate each coordinate is 0. For any $j = 1, \ldots, K$

$$\begin{aligned}
&q(\underline{n}, \underline{n} + \underline{e}_j) = \lambda_j \\
&q(\underline{n}, \underline{n} - \underline{e}_j) = \mu_j n_j p_j && \text{if } j \in \mathcal{U}(\underline{n}) \\
&q(\underline{n}, \underline{n} - \underline{e}_j) = \mu_j n_j \phi_j \frac{C - \sum_{i \in \mathcal{U}(\underline{n})} p_i n_i}{\sum_{i \in \mathcal{Z}(\underline{n})} \phi_i n_i} && \text{if } j \in \mathcal{Z}(\underline{n}) \\
&q(\underline{n}, \underline{n} + \underline{l}) = 0 && \text{if } \underline{l} \neq \pm \underline{e}_k \\
& && \text{for some } k = 1, \ldots, K.
\end{aligned} \tag{3.1}$$

Let $c_j(\underline{n})$ denote the bandwidth that a stream of class j obtains. We have

$$c_j(\underline{n}) = p_j \mathbf{I}\{j \in \mathcal{U}(\underline{n})\} + \phi_j \frac{C - \sum_{i\in\mathcal{U}(\underline{n})} p_i n_i}{\sum_{i\in\mathcal{Z}(\underline{n})} \phi_i n_i} \mathbf{I}\{j \in \mathcal{Z}(\underline{n})\}. \tag{3.2}$$

We remark that using (2.14), $c_j(\underline{n})$ can be given as an explicit function of $\underline{n}$ as follows:

$$c_j(\underline{n}) = p_j \mathbf{I}\left\{\sum_{k=1}^{j-1} n_k p_k + \sum_{k=j}^{K} n_k \phi_k \frac{p_j}{\phi_j} \le C\right\} + \phi_j \frac{C - \sum_{i\in\mathcal{U}(\underline{n})} p_i n_i}{\sum_{i\in\mathcal{Z}(\underline{n})} \phi_i n_i} \mathbf{I}\left\{\sum_{k=1}^{j-1} n_k p_k + \sum_{k=j}^{K} n_k \phi_k \frac{p_j}{\phi_j} > C\right\}. \tag{3.3}$$

Of course, this definition makes sense for $\underline{n} \in \mathbb{R}_+^K$.

Let $\Pi_j^a(t), t \ge 0$ and $\Pi_j^d(t), t \ge 0$ for $j = 1, \dots, K$ be $2K$ independent Poisson processes with rate 1. Let $N_j(t)$ be the number of flows from class j in the system at time t. Then by the rates in (3.1) we have

$$N_j(t) = N_j(0) + \Pi_j^a\big(\lambda_j t\big) - \Pi_j^d\left(\int_0^t \mu_j N_j(s) c_j\left(\underline{N}(s)\right) \mathrm{d}s\right). \tag{3.4}$$

Let $\lambda_j^L = \lambda_j L$, $j = 1, \dots, K$, $C^L = CL$. Let $N_j^L(t)$ be the number of flows from class j in the system at time t if the arrival intensities to the classes are $\lambda_1^L, \dots, \lambda_K^L$ respectively and the capacity is C^L. Simply rewriting the equation (3.4) for $N^L(t), t \ge 0$ and dividing by L we get

$$\frac{N_j^L(t)}{L} = \frac{N_j^L(0)}{L} + \frac{1}{L}\Pi_j^a\big(L\lambda_j t\big) - \frac{1}{L}\Pi_j^d\left(L\int_0^t \mu_j \frac{N_j^L(s)}{L} c_j\left(\frac{\underline{N}^L(s)}{L}\right) \mathrm{d}s\right) \quad j = 1, \dots, K.$$

For the ease of notations we rewrite this equation. Introducing $n_j^L(t) = \frac{N_j^L(t)}{L}$ $j = 1, \dots, K$ we have

$$n_j^L(t) = n_j^L(0) + \frac{1}{L}\Pi_j^a\big(\lambda_j L t\big) - \frac{1}{L}\Pi_j^d\left(L\int_0^t \mu_j n_j^L(s) c_j\left(\underline{n}^L(s)\right) \mathrm{d}s\right) \quad j = 1, \dots, K \tag{3.5}$$

The theory presented in [4, Ch 6.4 and Ch 11.2] can be applied to the process $\underline{n}^L(t), t \geq 0$ for obtaining convergence to $\underline{n}(t), t \geq 0$ the solution of the system of equations

$$n_j(t) = n_j(0) + \lambda_j t - \int_0^t \mu_j n_j(s) c_j\left(\underline{n}(s)\right) \mathrm{d}s, \quad j = 1, \ldots, K \tag{3.6}$$

as it is stated in the following theorem.

Theorem 3.1. *Assume that* $\lim_{L\to\infty} n_j^L(0) = n(0) \in [0, \infty)$ *for any* $j = 1, \ldots, K$. *Then for every* $t \geq 0$,

$$\lim_{L\to\infty} \sup_{s\leq t} \left|\underline{n}^L(s) - \underline{n}(s)\right| = 0 \quad a.s. \tag{3.7}$$

Proof. We will apply Theorem 2.1 of [4, p 456]. We have to check three conditions. First, for any compact set $B \subset [0, \infty)^K$ the following bound holds

$$\sup_{\underline{n}\in B} n_j c_j(\underline{n}) < \infty \quad j = 1, \ldots, K, \tag{3.8}$$

second, there exist M_B such that for any $j = 1, \ldots, K$

$$|n_j c_j(\underline{n}) - m_j c_j(\underline{m})| \leq M_B \, |\underline{n} - \underline{m}| \quad \underline{n}, \underline{m} \in B. \tag{3.9}$$

Third,

$$\lim_{L\to\infty} n_j^L(0) = n(0) \in [0, \infty) \quad j = 1, \ldots, K. \tag{3.10}$$

Using (3.3) Simple calculations show that (3.8) and (3.9) hold. The condition (3.10) is the same as the assumption of Theorem 3.1. Therefore, the convergence (3.7) holds. □

The main results of this section is the following.

Theorem 3.2. *If the function* $\underline{n}(t), t \geq 0$ *satisfies the equations (3.6) then in the stationary state* n_j^*, $j = 1, \ldots, K$ *each class is uncompressed and the the following holds:*

$$n_j^* = \frac{\lambda_j}{\mu_j p_j} \quad j = 1, \ldots, K.$$

Proof. For finding the stationary state $\underline{n}^*$ of the fluid limit differentiate $n_j(t)$, $j = 1, \ldots, K$ with respect to t and find the solution of the system $n_j'(t) = 0$, $j = 1, \ldots, K$. Using (3.6) and (3.2) one gets

$$0 = n_j'(t) = \lambda_j - \mu_j n_j(t) \cdot \left(p_j \mathbf{I}\{j \in \mathcal{U}(\underline{n}(t))\} + \phi_j \frac{C - \sum_{i\in\mathcal{U}(\underline{n})} p_i n_i(t)}{\sum_{i\in\mathcal{Z}(\underline{n})} \phi_i n_i(t)} \mathbf{I}\{j \in \mathcal{Z}(\underline{n}(t))\} \right)$$

This means that in the stable state we have

$$\lambda_j = \mu_j p_j n_j^* \text{ if } j \in \mathcal{U}(\underline{n}^*),$$
$$\lambda_j = \mu_j n_j^* \frac{\phi_j}{\sum_{i\in\mathcal{Z}} \phi_i n_i^*} \left(C - \sum_{i\in\mathcal{U}} p_i n_i^* \right) \text{ if } j \in \mathcal{Z}(\underline{n}^*).$$

If there is at least one compressed class, that is, $\mathcal{Z}(\underline{n}^*) \neq \emptyset$ then we have for $j \in \mathcal{Z}(\underline{n}^*)$

$$\begin{aligned}\lambda_j &= \mu_j n_j^* \frac{\phi_j}{\sum_{i\in\mathcal{Z}(\underline{n}^*)} \phi_i n_i^*} \left(C - \sum_{i\in\mathcal{U}(\underline{n}^*)} p_i n_i^* \right) \\ &= \mu_j n_j^* \frac{\phi_j}{\sum_{i\in\mathcal{Z}(\underline{n}^*)} \phi_i n_i^*} \left(C - \sum_{i\in\mathcal{U}(\underline{n}^*)} \frac{\lambda_i}{\mu_i} \right) \\ &= \mu_j n_j^* \frac{\phi_j}{\sum_{i\in\mathcal{Z}(\underline{n}^*)} \phi_i n_i^*} \left(C - \sum_{i\in\mathcal{U}(\underline{n}^*)} C\varrho_i \right)\end{aligned}$$

since the definition $\varrho_i = \frac{\lambda_i}{\mu_i C}$. Dividing by $\mu_j C$ and using $\varrho_j = \frac{\lambda_j}{\mu_j C}$ one gets

$$\varrho_j = \frac{\phi_j n_j^*}{\sum_{i\in\mathcal{Z}(\underline{n}^*)} \phi_i n_i^*} \left(1 - \sum_{i\in\mathcal{U}(\underline{n}^*)} \varrho_i \right),$$

rearranging the terms on the right we have

$$\frac{\varrho_j}{1 - \sum_{i\in\mathcal{U}(\underline{n}^*)} \varrho_i} = \frac{\phi_j n_j^*}{\sum_{i\in\mathcal{Z}(\underline{n}^*)} \phi_i n_i^*},$$

then summing both sides over $j \in \mathcal{Z}(\underline{n}^*)$ one has

$$\frac{\sum_{j\in\mathcal{Z}(\underline{n}^*)} \varrho_j}{1 - \sum_{i\in\mathcal{U}(\underline{n}^*)} \varrho_i} = 1,$$

this is equivalent to $\sum_{j=1}^{K} \varrho_j = 1$ which is contradiction. Consequently, $\mathcal{Z}(\underline{n}^*) = \emptyset$ and for any $j = 1, \ldots, K$ $n_j^* = \frac{\lambda_j}{\mu_j p_j}$. □

It can also be shown that the equilibrium $\underline{n}^*$ is stable, nevertheless, due to the lack of space it is not performed here. It has been elaborated following the argumentation in [2, pp 48–49] and also using [9, Lemma 3].

Here we note that in this bandwidth economical DPS the fluid limit lies completely in the uncompressed region (every classes in the limit are uncompressed), and the closed form expression of the fluid limit of a class depends only on the class parameters (λ_j, μ_j, p_j), and is quite simple.

On the contrary, in case of the previously analyzed DPS [2] (based on equation (2.3)) the fluid limit has no closed form solution, an algorithm is needed to determine the compressed and uncompressed classes and the corresponding limits in the asymptotics. Furthermore, the limit of a class may depend on the parameters of other classes (see Proposition 1.3. in [2]).

4. Fluid limit as the number of servers goes to infinity

In the concept of fluid limit the intensity of the arrival processes and the capacity of the server increase in the same pace by a multiplier L. Consequently, the number of packets under service increases and the number of served packets in unit time increases as well. The first property can be rephrased as the number of servers $(\frac{C}{p_j})$ increases. It is natural to ask whether one can take an asymptotic regime in which the number of servers increases but the intensities of the arrivals and the capacity are fixed. If so, then what can be said about the limit process. A possible way of considering such an asymptotic is that we decrease the access rates by L and take $p_j^L = p_j/L$. This is not enough to obtain fluid scaling like set up because the number of served packets per unit time does not increase. One can get over this problem and obtain limit of similar kind as the fluid limit if the time of the system is accelerated too. This regime will be described in this section.

Let us fix C and λ_j and decrease the access rate limits p_j, such that $p_j^L = \frac{p_j}{L}$, $j = 1, \ldots, K$ for $L > 0$. Let $M_j^L(t)$ be the number of flows from class j in the system at time t if the access rate limits are p_j^L. It can be proved that the rescaled and time accelerated process has fluid limit.

Theorem 4.1. *Assume that* $\lim_{L\to\infty} \frac{M_j^L(Lt)}{L} = m(0) \in [0, \infty)$ *for any* $j = 1, \ldots, K$. *For the processes* $M_j^L(t)$, $j = 1, \ldots, K$ *defined above we have the following fluid limit*

$$\lim_{L\to\infty} \sup_{s\le t} \left| \frac{M_j^L(Lt)}{L} - \underline{n}(s) \right| = 0 \quad a.s. \tag{4.1}$$

where $\underline{n}(s)$ *is the solution of the differential equation (3.6). Consequently,*

$$\lim_{t\to\infty} \lim_{L\to\infty} \frac{M_j^L(Lt)}{L} = n_j^* \quad j = 1, \ldots, K, \tag{4.2}$$

where n_j^* *is defined in Theorem 3.2.*

Proof. We will prove that the process $\frac{\underline{M}^L(Lt)}{L}, t \ge 0$ satisfies equation (3.5). Consequently, Theorem 3.1 can be applied for $\frac{\underline{M}^L(Lt)}{L}, t \ge 0$ yielding the same convergence (4.1). Then one can conclude that Theorem 3.2 holds for the limit process (4.2) without any further modification.

Proving the process $\frac{\underline{M}^L(Lt)}{L}, t \geq 0$ satisfies equation (3.5), we first rewrite the equation (3.4) for $M^L(t), t \geq 0$. We have

$$M_j^L(t) = M_j^L(0) + \Pi_j^a\big(\lambda_j t\big) - \Pi_j^d\left(\int_0^t \mu_j M_j^L(s) c_j^{L*}\left(\underline{M}^L(s)\right)\,\mathrm{d}s\right),$$

where for any $\underline{m} \in [0,\infty)^K$ we define

$$c_j^{L*}(\underline{m}) = p_j \mathbf{I}\left\{\sum_{k=1}^{j} m_k \frac{p_k}{L} + \sum_{k=j+1}^{K} m_k \phi_k \frac{p_j}{\phi_j L} \leq C\right\} + \mu_j \phi_j \frac{C - \sum_{i\in\mathcal{U}(\underline{m})} p_i m_i}{\sum_{i\in\mathcal{Z}(\underline{m})} \phi_i n_i} \mathbf{I}\left\{\sum_{k=1}^{j} m_k \frac{p_k}{L} + \sum_{k=j+1}^{K} m_k \phi_k \frac{p_j L}{\phi_j} > C\right\}.$$

As previously we devide by L and for having fluid limit we speed up the time by L:

$$\frac{M_j^L(Lt)}{L} = \frac{M_j^L(0)}{L} + \frac{1}{L}\Pi_j^a\big(\lambda_j Lt\big) - \frac{1}{L}\Pi_j^d\left(\int_0^{Lt} \mu_j M_j^L(s) c_j^{L*}\left(\underline{M}^L(s)\right)\,\mathrm{d}s\right).$$

Using the fact that $\int_0^{Lt} f(s)\,\mathrm{d}s = \int_0^t Lf(Ls)\,\mathrm{d}s$ we have

$$\frac{M_j^L(Lt)}{L} = \frac{M_j^L(0)}{L} + \frac{1}{L}\Pi_j^a\big(\lambda_j Lt\big) - \frac{1}{L}\Pi_j^d\left(L\int_0^t \mu_j M_j^L(Ls) c_j^{L*}\left(\underline{M}^L(Ls)\right)\frac{1}{L}\,\mathrm{d}s\right). \tag{4.3}$$

From the definition of c_j^{L*} and c_j it follows that

$$c_j^{L*}\left(M_j^L(Lt)\right) = \frac{1}{L} c_j\left(\frac{M_j^L(Lt)}{L}\right).$$

This equation and (4.3) implies that

$$\frac{M_j^L(Lt)}{L} = \frac{M_j^L(0)}{L} + \frac{1}{L}\Pi_j^a\big(\lambda_j Lt\big) - \frac{1}{L}\Pi_j^d\left(\int_0^t \mu_j M_j^L(Ls) c_j\left(\frac{\underline{M}^L(Ls)}{M}\right)\,\mathrm{d}s\right). \tag{4.4}$$

Introducing $m_j^L(t) = \frac{M_j^L(Lt)}{L}$, (4.4) can be written as

$$m_j^L(t) = m_j^L(0) + \frac{1}{L}\Pi_j^a\big(\lambda_j Lt\big) - \frac{1}{L}\Pi_j^d\left(L\int_0^t \mu_j m_j^L(s) c_j\left(\underline{m}^L(s)\right)\,\mathrm{d}s\right).$$

which is the same as equation (3.5) and for the processes $m_j^L(t), t \geq 0$ we have fluid limit. □

4.1. The one-dimensional case

Let us consider the M/G/1-PS system as a special one-dimensional case of the multiclass Processor Sharing. The average number of customers in the sationary state of the system is $EN = \frac{\rho}{1-\rho}$ where $\rho = \frac{\lambda}{\mu C}$. It can easily be shown (also based on the previous discussion) that M/G/1-PS has a stable fluid limit, which is $n^* = \lim_{L\to\infty} \frac{EN^L}{L} = \rho$ where N^L is the average number of customers in the L times speed up M/G/1-PS system ($\lambda^L = L\lambda, C^L = LC$). Similarly to the multiclass case above, here it is also true that the very same fluid limit results if the number of servers goes to infinity (with C and λ fixed), that is $L := \frac{C}{p}$ tends to infinity (with p tending to zero) where p is the *access rate limit*. This observation is very important, because for every finite L the system is equivalent to the M/G/L-PS (in the literature often referred to as M/G/R-PS [10]) system whose solution is known. It means that in this single class case not only the two 'extreme' systems (the access rate limitless M/G/1-PS case when $L = 1$ and the fluid limit when $L = \infty$) can be characterized, but every system between them can be solved, thus the convergence to the limit can fully be described. Based on the formula for the average number of customers presented in [10] for M/G/L-PS, one can obtain

$$\frac{EN^L}{L} = \rho\left(1 + \frac{E_2(L, L\rho)}{L(1-\rho)}\right) \tag{4.5}$$

where E_2 is Erlang's second formula. It can easily be checked that the formula above gives $\frac{\rho}{1-\rho}$ for $L = 1$, and ρ for $L = \infty$. Of course, the M/G/R-PS itself has also fluid limit, which is $R\rho = \frac{\lambda}{\mu C/R} = \frac{\lambda}{\mu p}$ (see the similarities to the multiclass case in Theorem 3.2) and the convergence to the fluid limit can be characterized by using similar formula as in (4.5).

We strongly believe that this well characterizable convergence to the fluid limit of single class DPS can be utilized for solving the bandwidth economical multiclass access rate limited DPS, because the solution of the original model [6], and the fluid limit of the access rate limited multiclass DPS presented in this paper are already in our hands.

5. Conclusion

In this paper we have analyzed a bandwidth economical discriminatory processor sharing system with access rate limitations, as a possible and realistic model for bandwidth sharing of (elastic) network traffic flows subject to flow control and access rate limits. We have characterized the state-space and determined the unique state-dependent bandwidth shares of such a capacity conserving system, in which the unused capacity of users due to the effect of their access rate limits is fully re-distributed among other users. We have also presented two asymptotic regimes of the system which may help in the further research to obtain computationally tractable methods for evaluating the performance.

References

[1] K. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija. Discriminatory processor sharing revisited. In *In: Proc. IEEE Infocom 2005, Miami FL*, pages 784–795, 2005.

[2] U. Ayesta and M. Mandjes. Bandwidth-sharing networks under a diffusion scaling. *Annals Operation Research*, 170(1):41–58, 2009.

[3] Niklas Carlsson and Martin Arlitt. Towards more effective utilization of computer systems. In *ICPE '11 Proceeding of the second joint WOSP/SIPEW International Conference on Performance Engineering*, 2011.

[4] S.N. Ethier and T.G. Kurtz. *Markov processes: Characterization and convergence.* Wiley New York, 1986.

[5] Z. Fan. Dimensioning Bandwidth for Elastic Traffic. *NETWORKING 2002: Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications*, pages 826–837, 2006.

[6] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *J. ACM*, 27(3):519–532, 1980.

[7] L. Kleinrock. Time-shared systems: A theoretical treatment. *J. of ACM*, 14(2):242–261, 1967.

[8] A. Lakshmikantha, R. Srikant, and CL Beck. Differential equation models of flow-size based priorities in internet routers. *International Journal of Systems, Control and Communications*, 2(1):170–196, 2010.

[9] L. Leskela. Stabilization of an overloaded queueing network using measurement-based admission control. *Journal of Applied Probability*, 43(1):231–244, 2006.

[10] K. Lindberger. Balancing quality of service, pricing and utilisation in multiservice networks with stream and elastic traffic. In *ITC-16: International Teletraffic Congress*, pages 1127–1136, 1999.

[11] T. M. O'Donovan. Direct solutions of m/g/1 processor-sharing models. *Operation Research*, 22:1232–1235, 1974.

[12] K.M. Rege and B. Sengupta. Queue length distribution for the discriminatory processor-sharing queue. *Operation Research*, 44:653–657, 1996.

[13] A. Riedl, T. Bauschert, and J. Frings. A framework for multi-service IP network planning. In *International Telecommunication Network Strategy and Planning Symposium (Networks)*, pages 183–190, 2002.

[14] G. van Kessel, R. Núñez-Queija, and S. Borst. Asymptotic regimes and approximations for discriminatory processor sharing. *ACM SIGMETRICS Performance Evaluation Review*, 32(2):44–46, 2004.