# Optimal Policy for Multi-Class Scheduling in a Single Server Queue

Natalia Osipova
INRIA Sophia Antipolis, France,
E-mail:
Natalia.Osipova@ymail.com

Urtzi Ayesta
LAAS-CNRS, France,
BCAM, Zamudio, Spain,
Email: urtzi@laas.fr

Konstantin Avrachenkov
INRIA Sophia Antipolis, France,
E-mail:
K.Avrachenkov@sophia.inria.fr

***Abstract*—In this paper we apply the Gittins optimality result to characterize the optimal scheduling discipline in a multi-class $M/G/1$ queue. We apply the general result to several cases of practical interest where the service time distributions belong to the set of decreasing hazard rate distributions, like Pareto or hyper-exponential. When there is only one class it is known that in this case the Least Attained Service policy is optimal. We show that in the multi-class case the optimal policy is a priority discipline, where jobs of the various classes depending on their attained service are classified into several priority levels. Using a tagged-job approach we obtain, for every class, the mean conditional sojourn time. This allows us to compare numerically the mean sojourn time in the system between the Gittins optimal and popular policies like Processor Sharing, First Come First Serve and Least Attained Service (LAS). We implement the Gittins' optimal algorithm in NS-2 and we perform numerical experiments to evaluate the achievable performance gain. We find that the Gittins policy can outperform by nearly $10\%$ the LAS policy.**

## I. Introduction

We are interested to schedule the jobs in the $M/G/1$ queue with the aim to minimize the mean sojourn time in the system as well as the mean number of jobs in the system. In our study we restrict ourselves to the non-anticipating scheduling policies. Let us recall that the policy is non-anticipating if it does not use information about the size of the arriving jobs. In [1] Gittins considered an $M/G/1$ queue and proved that the so-called Gittins index rule minimizes the mean delay. At every moment of time the Gittins rule calculates, depending on the service times of jobs, which job should be served. Gittins derived this result as a byproduct of his groundbreaking results on the multi-armed bandit problem. The literature on multi-armed bandit related papers that build on Gittins' result is huge (see for example [2], [3], [4], [5], [6], [7], [8]). However, the optimality result of the Gittins index in the context of the $M/G/1$ queue has not been fully exploited, and it has not received the attention it deserves.

In the present work we generalize the Gittins index approach to the scheduling of the multi-class $M/G/1$ queue. We emphasize that Gittins' optimality in a multi-class queue holds under much more general conditions than the condition required for the optimality of the well-known $c\mu$-rule. We recall that the $c\mu$-rule is the discipline that gives strict priority in descending order of $c_k\mu_k$, where $c_k$ and $\mu_k$ refer to a cost and the inverse of the mean service requirement, respectively, of class $k$. Indeed it is known (see for example [9], [10], [11]) that the $c\mu$-rule minimizes the weighted mean number of customers in the queue in two main settings: (i) generally distributed service requirements among all non-preemptive disciplines and (ii) exponentially distributed service requirements among all preemptive non-anticipating disciplines. In the preemptive case the $c\mu$-rule is only optimal if the service times are exponentially distributed. On the other hand, by applying Gittins' framework to the multi-class queue one can characterize the optimal policy for arbitrary service time distributions. We believe that our results open an interesting avenue for further research. For instance well-known optimality results in a single-class queue like the optimality of the Least Attained Service (LAS) discipline when the service times are of type decreasing hazard rate or the optimality of FCFS when the service time distribution is of type New-Better-than-Used-in-Expectation can all be derived as corollaries of Gittins' result. The optimality of the $c\mu$-rule can also easily be derived from the Gittins' result.

In order to get insights into the structure of the optimal policy in the multi-class case we consider several relevant cases where the service time distributions are Pareto or hyper-exponential. We have used these distributions due to the evidence that the file size distributions in the Internet are well presented by the heavy-tailed distributions such as Pareto distributions with the infinite second moment. Also it was shown that the job sizes in the Internet are well modelled with the distributions with the decreasing hazard rate (DHR). We refer to [12], [13], [14] for more details on this area. In particular, we study the optimal multi-class scheduling in the following cases of the service time distributions: two Pareto distributions, several Pareto distributions, one hyper-exponential and one exponential distributions. Using a tagged-job approach and the collective marks method we obtain, for every class, the mean conditional sojourn time. This allows us to compare numerically the mean sojourn time in the system between the Gittins optimal and popular policies like Processor Sharing (PS), First Come First Serve (FCFS) and LAS. We find that in a particular example with two classes and Pareto-type service time distribution the Gittins' policy outperforms LAS by nearly 25% under moderate load.

From an application point of view, our findings could be applied in Internet routers. Imagine that incoming packets

are classified based on the application or the source that generated them. Then it is reasonable to expect that the service time distributions of the various classes may differ from each other. A router in the Internet does not typically have access to the exact required service time (in packets) of the TCP connections, but it may have access to the attained service of each connection. Thus we can apply our theoretical findings in order to obtain the optimal (from the connection-level performance point of view) scheduler at the packet level. We implement the Gittins' scheduling in the NS-2 simulator and perform experiments to evaluate the achievable performance gain. We found that in particular examples the Gittins policy can outperform the LAS policy by nearly $10\%$.

The structure of the paper is as follows: In Section 2 we review the Gittins index policy for the single-class $M/G/1$ queue and then provide a general framework of the Gittins index policy for the multi-class $M/G/1$ queue. In Section 3, we study the Gittins index policy for the case of two Pareto distributed classes and we generalized the results to multiple Pareto classes. In Section 4 we study the case of exponential and hyper-exponential distributions, we obtain analytical results and provide numerical examples. Section 5 concludes the paper.

## II. Gittins policy in multi-class $M/G/1$ queue

Let us first recall the basic results related to the Gittins index policy in the context of a single-class $M/G/1$ queue.

Let $\Pi$ denote the set of non-anticipating scheduling policies. Popular disciplines such as PS, FCFS and LAS, also called FB, belong to $\Pi$. Important disciplines that do not belong to $\Pi$ are SRPT and Shortest Processing Time (SPT).

We consider a single-class $M/G/1$ queue. Let $X$ denote the service time with distribution $P(X \leq x) = F(x)$. The density is denoted by $f(x)$, the complementary distribution by $\overline{F}(x) = 1 - F(x)$ and the hazard rate function by $h(x) = f(x)/\overline{F}(x)$. Let $\overline{T}^{\pi}(x)$, $\pi \in \Pi$ denote the mean conditional sojourn time for the job of size $x$ in the system under the scheduling policy $\pi$, and $\overline{T}^{\pi}$, $\pi \in \Pi$ denote the mean sojourn time in the system under the scheduling policy $\pi$.

Let us give some definitions.

*Definition 1:* For any $a, \Delta \geq 0$, let

$$J(a,\Delta) = \frac{\int_0^{\Delta} f(a+t)dt}{\int_0^{\Delta} \overline{F}(a+t)dt} = \frac{\overline{F}(a) - \overline{F}(a+\Delta)}{\int_0^{\Delta} \overline{F}(a+t)dt}. \quad (1)$$

For a job that has attained service $a$ and is assigned $\Delta$ units of service, equation (1) can be interpreted as the ratio between (i) the probability that the job will complete with a quota of $\Delta$ (interpreted as payoff) and (ii) the expected processor time that a job with attained service $a$ and service quota $\Delta$ will require from the server (interpreted as investment). Note that for every $a > 0$

$$J(a,0) = \frac{f(a)}{\overline{F}(a)} = h(a),$$

$$J(a,\infty) = \frac{\overline{F}(a)}{\int_0^{\infty} \overline{F}(a+t)\,dt} = 1/E[X-a|X>a].$$

Note further that $J(a,\Delta)$ is continuous with respect to $\Delta$.

*Definition 2:* The Gittins index function is defined by

$$G(a) = \sup_{\Delta \geq 0} J(a,\Delta), \quad (2)$$

for any $a \geq 0$.

We call $G(a)$ the *Gittins index* after the author of book [1], which handles various static and dynamic scheduling problems. Independently, Sevcik defined a corresponding index when considering scheduling problems without arrivals in [15]. In addition, this index has been dealt with by Yashkov, see [16] and references therein, in particular the works by Klimov [17], [18].

*Definition 3:* For any $a \geq 0$, let

$$\Delta^*(a) = \sup\{\Delta \geq 0 \mid J(a,\Delta) = G(a)\}. \quad (3)$$

By definition, $G(a) = J(a, \Delta^*(a))$ for all $a$.

*Definition 4:* The Gittins index policy $\pi_g$ is the scheduling discipline that at every instant of time gives service to the job in the system with highest $G(a)$, where $a$ is the job's attained service.

*Theorem 1:* The Gittins index policy minimizes the mean sojourn time in the system between all non-anticipating scheduling policies. Otherwise, in the $M/G/1$ queue for any $\pi \in \Pi$,

$$\overline{T}^{\pi_g} \leq \overline{T}^{\pi}.$$

*Proof:* See [1]. ■

Note that by Little's law the Gittins index policy also minimizes the mean number of jobs in the system.

We generalize the result of Theorem 1 to the case of the multi-class single server queue. Let us consider a multi-class $M/G/1$ queue. Let $X_i$ denote the service time with distribution $P(X_i \leq x) = F_i(x)$ for every class $i = 1, \ldots, N$. The density is denoted by $f_i(x)$ and the complementary distribution by $\overline{F}_i(x) = 1 - F_i(x)$. Jobs of every class-$i$ arrive with the Poisson process with rate $\lambda_i$, the total arrival rate is $\lambda = \sum_{i=1}^{N} \lambda_i$. For every class $i = 1, \ldots, N$ we define $J_i(a,\Delta) = \frac{\int_0^{\Delta} f_i(a+t)dt}{\int_0^{\Delta} \overline{F}_i(a+t)dt}$ and then the Gittins index of a class-$i$ job is defined as $G_i(a) = \sup_{\Delta \geq 0} J_i(a,\Delta)$.

The mean conditional sojourn time $\overline{T}_i^{\pi}(x)$ for the class-$i$ job of size $x$, $i = 1, \ldots, N$, and the mean sojourn time $\overline{T}^{\pi}$ in the system under the scheduling policy $\pi \in \Pi$ are defined as in the previous section.

*Proposition 1:* In a multi-class $M/G/1$ queue the policy that schedules the job with highest Gittins index $G_i(a)$, $i = 1, \ldots, N$ in the system, where $a$ is the job's attained service, is the optimal policy that minimizes the mean sojourn time.

*Proof:* The result follows directly from the application Gittins Index Definition 2 and Theorem 1 to a multi-class $M/G/1$ queue. ■

Let $h_i(x) = f_i(x)/\overline{F}_i(x)$ denote the hazard rate function of class $i = 1, \ldots, N$. Let the service time distribution of class-$i$ have a decreasing hazard rate. It is possible to show,

see [19], that if $h_i(x)$ is non-increasing, the function $J_i(a,\Delta)$ is non-increasing in $\Delta$. Thus

$$G_i(a) = J_i(a,0) = h_i(a). \tag{4}$$

As a consequence we obtain the following proposition.

*Proposition 2:* In a multi-class $M/G/1$ queue with non-increasing hazard rates functions $h_i(x)$ for every class $i = 1,\dots,N$, the policy that schedules the job with highest $h_i(a)$, $i = 1,\dots,N$ in the system, where $a$ is the job's attained service, is the optimal policy that minimizes the mean sojourn time.

*Proof:* Follows immediately from the Gittins policy Definition 4, Proposition 1 and equation (4). ■

The policy presented in Proposition 2 is an optimal policy for the multi-class single server queue between all non-anticipating scheduling policies. Let us notice that for the single class single server queue the Gittins policy becomes a LAS policy, as the hazard rate function is the same for all jobs and so the job with the maximal value of the hazard rate function is the job with the least attained service. When we serve jobs with the Gittins policy in the multi-class queue to find a job which has to be served next, we need to calculate the hazard rate of every job in the system. The job which has the maximal value of the hazard rate function is served the next.

Now let us consider several subcases of the described general approach. Depending on the behavior of the hazard rate functions of the job classes the policy is different. We consider the case with two job classes in the system and two subcases: (a) both job classes are distributed with Pareto and the hazard rate function do not cross and (b) job size distributions are hyper-exponential with one and two phases and they cross at one point. Then we extend the case of two Pareto job classes to the case of $N$ Pareto job classes.

## III. Two Pareto classes

### A. Model description

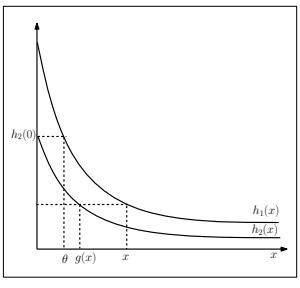


Figure 1. Two Pareto classes, hazard rates

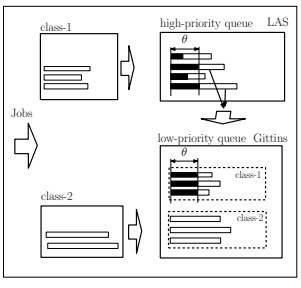


Figure 2. Two Pareto classes, policy scheme

We consider the case when the job size distribution functions are Pareto. We consider the two-class single server $M/G/1$ queue. Jobs of each class arrive to the server with Poisson process with rates $\lambda_1$ and $\lambda_2$. Job sizes are distributed according to the Pareto distributions, namely

$$F_i(x) = 1 - \frac{b_i^{c_i}}{(x+b_i)^{c_i}}, \quad i = 1,2. \tag{5}$$

Here $b_i = m_i(c_i - 1)$, where $m_i$ is the mean of class-$i$, $i = 1,2$. Then $f_i(x) = b_i^{c_i} c_i/(x+b_i)^{c_i+1}$, $i = 1,2$ and the hazard rate functions are

$$h_i(x) = \frac{c_i}{(x+b_i)}, \quad i = 1,2.$$

This functions cross at the point $a^{**} = \frac{c_2 b_1 - c_1 b_2}{c_1 - c_2}$. Without loss of generality suppose that $c_1 > c_2$. Then the behavior of the hazard rate functions depends on the values of $b_1$ and $b_2$. Let us first consider the case when the hazard rate function do not cross, so $a^{**} < 0$. This happens when $b_1/b_2 < c_1/c_2$. Then the hazard-rate functions are decreasing and never cross and $h_1(x) > h_2(x)$, for all $x \geq 0$.

Let us denote $\theta$ and function $g(x)$ in the following way that

$$h_1(x) = h_2(g(x)), \quad h_1(\theta) = h_2(0).$$

We can see that $g(\theta) = 0$. For given expressions of $h_i(x)$, $i = 1,2$ we get $g(x) = \frac{c_2}{c_1}(x + b_1) - b_2$, $\theta = \frac{c_1 b_2 - c_2 b_1}{c_2}$. According to the definition of function $g(x)$, the class-1 job of size $x$ and the class-2 job of size $g(x)$ have the same value of the hazard rate when they are fully served, see Figure 1. Then the optimal policy scheme is given on Figure 2.

### B. Optimal policy

Jobs in the system are served in two queues, low and high priority queues. The class-1 jobs which have attained service $a < \theta$ are served in the high priority queue with LAS policy. When the class-1 job achieves $\theta$ amount of service it is moved to the second low priority queue. The class-2 jobs are moved immediately to the low priority queue. The low priority queue is served only when the high priority queue is empty. In the low priority queue jobs are served in the following way: the service is given to the job with the highest $h_i(a)$, where $a$ is the job's attained service. So, for every class-1 job with $a$ attained service the function $h_1(a)$ is calculated, for every class-2 job with $a$ attained service the function $h_2(a)$ is calculated. After all values of $h_i(a)$ are compared and the job which has the highest $h_i(a)$ is served.

Now let us calculate the expressions of the mean conditional sojourn time for the class-1 and class-2 jobs.

### C. Mean conditional sojourn times

Let us denote by indices $[]^{(1)}$ and $[]^{(2)}$ the values for class-1 and class-2 accordingly.

Let us define as $\overline{X_y^n}^{(i)}$ the $n$-th moment and $\rho_y^{(i)}$ be the utilization factor for the distribution $F_i(x)$ truncated at $y$ for $i = 1,2$. The distribution truncated at $y$ equals to $F(x)$ for $x \leq y$ and equals to 1 when $x > y$. Let us denote $W_{x,y}$ the mean workload in the system which consists only of class-1 jobs of size less than $x$ and of class-2 jobs of size less than $y$. According to the Pollaczek-Khinchin formula

$$W_{x,y} = \frac{\lambda_1 \overline{X_x^2}^{(1)} + \lambda_2 \overline{X_y^2}^{(2)}}{2(1 - \rho_x^{(1)} - \rho_y^{(2)})}.$$

Now let us formulate the following Theorem.

*Theorem 2:* In the two-class $M/G/1$ queue where the job size distributions are Pareto, given by (5), and which is scheduled with the Gittins policy described in Subsection III-B, the mean conditional sojourn times for class-1 and class-2 jobs are

$$T_1(x) = \frac{x + W_{x,0}}{1 - \rho_x^{(1)}}, \quad x < \theta, \tag{6}$$

$$T_1(x) = \frac{x + W_{x,g(x)}}{1 - \rho_x^{(1)} - \rho_{g(x)}^{(2)}}, \quad x \geq \theta, \tag{7}$$

$$T_2(g(x)) = \frac{g(x) + W_{x,g(x)}}{1 - \rho_x^{(1)} - \rho_{g(x)}^{(2)}}, \quad x \geq \theta. \tag{8}$$

*Proof:* Let us give a very general idea of the proof. To obtain expressions (7), (8) we use the fact that the second low priority queue is the queue with batch arrivals. To obtain expressions of the mean batch size with and without the tagged job we apply the Generating function analysis using the method of the collective marks. ■

The obtained expressions (6), (7) and (8) can be interpreted using the tagged-job and mean value approach.

Let us consider class-1 jobs. The jobs of size $x \leq \theta$ are served with the LAS policy, so the mean conditional sojourn time is the known, see [20, Sec. 4.6], $T_1(x) = \frac{x+W_{x,0}}{1-\rho_x^{(1)}}$, $x < \theta$, where $W_{x,0}$ is the mean workload and $\rho_x^{(1)}$ is the mean load in the system for class-1 job of size $x$. The mean workload $W_{x,0}$ and mean load $\rho_x^{(1)}$ consider only the jobs of the high priority queue of class-1.

For the jobs of size $x > \theta$ the expression (7) can be presented in the following way, $T_1(x) = x + W_{x,g(x)} + T_1(x)(\rho_x^{(1)} + \rho_{g(x)}^{(2)})$, where

- $x$ is time which is actually spent to serve the job;
- $W_{x,g(x)}$ is the mean workload which the tagged job finds in the system and which has to be processed before it;
- $T_1(x)(\rho_x^{(1)} + \rho_{g(x)}^{(2)})$ is the mean time to serve the jobs which arrive to the system during the service time of the tagged job and which have to be served before it.

Let us describe several properties of the optimal policy.

### D. Properties of the optimal policy

*Property 1:* When the class-2 jobs arrive to the server they are not served immediately, but wait until the high priority queue is empty. The mean sojourn time is the limit $\lim_{g(x)\to 0} T_2(g(x))$. As $\lim_{x\to\theta} g(x) = 0$, then

$$\lim_{g(x)\to 0} T_2(g(x)) = \frac{W_{\theta,0}}{1 - \rho_\theta^{(1)}} = \frac{\lambda_1 \overline{X_\theta^2}^{(1)}}{2(1 - \rho_\theta^{(1)})^2}.$$

Let us notice that

$$\lim_{g(x)\to 0} T_2(g(x)) \neq T_1(\theta) = \frac{\theta + W_{\theta,0}}{1 - \rho_\theta^{(1)}}.$$

Class-2 jobs wait in the system to be served in the low priority queue, the mean waiting time is $\lim_{g(x)\to 0} T_2(g(x))$. Class-1 jobs of size more then $\theta$ also wait in the system to be served in the low priority queue, the mean waiting time for them is $T_1(\theta)$. Property 1 shows that these two mean waiting times are not equal, so class-1 jobs and class-2 jobs wait different times to start to be served in the low priority queue.

*Property 2:* Let us consider the condition of no new arrival. According to the optimal policy structure in the low priority queue jobs are served according to the LAS policy with different rates, which depend on the number of jobs in each class and hazard rate functions. For the case when there are no new arrivals in the low priority queue we can calculate the rates with which the class-1 jobs and class-2 jobs are served in the system at every moment of time. We consider that all the class-1 jobs and all the class-2 jobs already received the same amount of service. Let $n_1$ and $n_2$ be the number of jobs in class-1 and class-2 and let $x_1$ and $x_2$ be the attained services of every job in these classes. Then at any moment $h_1(x_1) = h_2(x_2)$. If the total capacity of the server is $\Delta$, then let $\Delta_1$ and $\Delta_2$ be the capacities which each job of class-1 and class-2 receives. Then

$$n_1\Delta_1 + n_2\Delta_2 = \Delta.$$

Also $h_1(x_1 + \Delta_1) = h_2(x_2 + \Delta_2)$. As $\Delta$ is very small (and so as well $\Delta_1$ and $\Delta_2$) according to the LAS policy, then we can approximate

$$h_i(x + \Delta_i) = h_i(x) + \Delta_i h_i'(x), \quad i = 1, 2.$$

Then from the previous equations we have $\Delta_1 h_1'(x_1) = \Delta_2 h_2'(x_2)$. Then

$$\frac{\Delta_1}{\Delta} = \frac{h_2'(x_2)}{n_1 h_2'(x_2) + n_2 h_1'(x_1)},$$
$$\frac{\Delta_2}{\Delta} = \frac{h_1'(x_1)}{n_1 h_2'(x_2) + n_2 h_1'(x_1)}.$$

This result is true for any two distributions for which the hazard rates are decreasing and never cross. For the case of two Pareto distributions given by (5) we have the following:

$$\frac{\Delta_1}{\Delta} = \frac{c_1}{n_1 c_1 + n_2 c_2}, \quad \frac{\Delta_2}{\Delta} = \frac{c_2}{n_1 c_1 + n_2 c_2}.$$

So, for the case of two Pareto distributions the service rates of class-1 and class-2 jobs do not depend on the current jobs' attained services.

*Property 3:* As one can see from the optimal policy description, the class-1 and class-2 jobs leave the system together if they have the same values of the hazard rate functions of their sizes and if they find each other in the system. According to the definition of the $g(x)$ function we can conclude that the class-1 job of size $x$ and class-2 job of size $g(x)$, if they find each other in the system, leave the system together. But these jobs do not have the same conditional mean sojourn time,

$$T_1(x) \neq T_2(g(x)).$$

This follows from expressions (7) and (8).

### E. Two Pareto classes with intersecting hazard rate functions

Now let us consider the case when the hazard rate function cross, then $a^{**} = (c_2 b_1 - c_1 b_2)/(c_1 - c_2) \geq 0$. As we considered $c_1 > c_2$, then $h_1(0) < h_2(0)$ and then class-2 jobs are served in the high priority queue until they receive $\theta^* = (c_2 b_1 - c_1 b_2)/c_1$ amount of service. Here $\theta^*$ is such that $h_2(\theta^*) = h_1(0)$ and $g(\theta^*) = 0$. The $g(x)$ function crosses the $y = x$ function at point $a^{**}$. As we show in Property 2, the rates with which class-1 and class-2 jobs are served in the low priority queue depend only on the $c_1$ and $c_2$ parameters and so class-1 jobs always have priority over class-2 jobs according to the service rates. We can rewrite the expressions of mean conditional sojourn times of Section III, Theorem 2 in the following way.

*Corollary 1:* In the two-class $M/G/1$ queue where the job size distributions are Pareto, given by (5) such that the hazard rate functions cross, and which is scheduled with the Gittins optimal policy, the mean conditional sojourn times for class-1 and class-2 jobs are

$$T_1(x) = \frac{x + W_{x,g(x)}}{1 - \rho_x^{(1)} - \rho_{g(x)}^{(2)}}, \quad x \geq 0,$$

$$T_2(x) = \frac{x + W_{x,0}}{1 - \rho_x^{(2)}}, \quad x < \theta^*,$$

$$T_2(g(x)) = \frac{g(x) + W_{x,g(x)}}{1 - \rho_x^{(1)} - \rho_{g(x)}^{(2)}}, \quad x \geq \theta^*.$$

*Proof:* The proof follows from the previous discussion. ■

### F. Numerical results

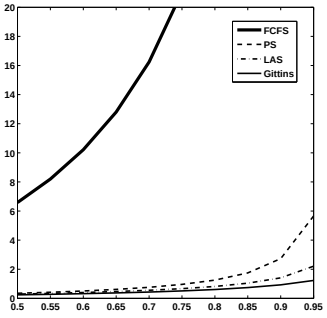


Figure 3. Two Pareto classes, mean sojourn times with respect to the load $\rho$, $V_1$

We consider two classes with parameters presented in Table I and we calculate the mean sojourn time in the system numerically, using the expressions of the mean conditional sojourn time (6), (7) and (8). We provide the results for two different parameters sets, which we call $V_1$ and $V_2$.

It is known that in the Internet most of the traffic is generated by the large files (80%), while most of the files are very small (90%). This phenomenon is referred to as

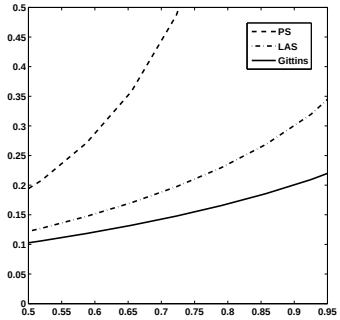


Figure 4. Two Pareto classes, mean sojourn times with respect to the load $\rho$, $V_2$

Table I
TWO PARETO CLASSES, PARAMETERS

| V | $c_1$ | $c_2$ | $m_1$ | $m_2$ | $\rho_1$ | $\rho_2$ | $\rho$ |
|---|---|---|---|---|---|---|---|
| $V_1$ | 25.0 | 2.12 | 0.04 | 0.89 | 0.1 | 0.4..0.85 | 0.5..0.95 |
| $V_2$ | 10.0 | 1.25 | 0.05 | 1.35 | 0.25 | 0.25..0.74 | 0.5..0.99 |

"mice-elephant" effect. Also it is known that the file sizes are well presented by the heavy-tailed distributions like Pareto. Here the class-1 jobs represent "mice" class and class-2 jobs "elephants". We consider that the load of the small files is fixed and find the mean sojourn time in the system according to the different values of the "elephant" class arrival rate.

We compare the mean sojourn time for the Gittins policy, PS, FCFS and LAS policies. These policies can be applied either in the Internet routers or in the Web service. The expected sojourn times for these policies are, see [20], $\overline{T}^{PS} = \frac{\rho/\lambda}{1-\rho}$ for the PS policy, $\overline{T}^{FCFS} = \rho/\lambda + W_{\infty,\infty}$ for the FCFS policy, where $W_{\infty,\infty}$ means the total mean unfinished work in the system. For the LAS policy

$$\overline{T}^{LAS} = \frac{1}{\lambda} \int_0^\infty \overline{T}^{LAS}(x) f(x) dx,$$

$$\overline{T}^{LAS}(x) = \frac{x + W_{x,x}}{1 - \rho_x^{(1)} - \rho_x^{(2)}},$$

where $f(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x)$ and $\lambda = \lambda_1 + \lambda_2$.

The mean sojourn times for the parameters sets $V_1$ and $V_2$ are presented in Figures 3,4. For the results of $V_2$ we do not plot the mean sojourn time for the FCFS policy as class-2 has an infinite second moment. As one can see Gittins policy minimizes the mean sojourn time. In particular, it outperforms the LAS policy by almost $25-30\%$ when the system is loaded by around $90\%$. We note that the PS policy produces much worse results than the LAS and Gittins policies.

### G. Simulation results

We implement Gittins policy algorithm for the case of two Pareto distributed classes in NS-2 simulator. The algorithm is implemented in the router queue. In the router we keep the trace of the attained service (by attained service we mean the number of transmitted packets) by every connection in

the system. We keep the trace during some time interval after which there are no more packets from the connection in the queue.

It is possible to select the packet with the minimal sequence number of the connection which has to be served instead of selecting the first packet in the queue. In the current simulation this parameter does not play a big role according to the selected model scheme and parameters. (There are no drops in the system, so there are no retransmitted packets. Then all the packets arrive in the same order as they were sent.)

The algorithm which is used for the simulations is as follows:

**Algorithm**
on packet dequeue
select the flow $f$ with the max $h_i(a_f)$, where
$a_f$ is the flow's attained service
select the first packet $p_f$ of the flow $f$ in the queue
dequeue selected packet $p_f$
set $a_f = a_f + 1$

To compare Gittins policy with the LAS policy we also implemented LAS algorithm in the router queue. According to the LAS discipline the packet to dequeue is the packet from the flow with the least attained service.

The simulation topology is the following: jobs arrive to the bottleneck router in two classes, which represent mice and elephants in the network. Jobs are generated by FTP sources which are connected to TCP senders. The file size distributions are Pareto, $F_i = 1 - b_i^{c_i}/(x+b_i)^{c_i}$, $i = 1, 2$. Jobs arrive according to the Poisson arrivals with rates $\lambda_1$ and $\lambda_2$. For the simulations we selected the scenario described in Subsection III-E.

We consider that all connection have the same propagation delays, $12$ ms. The bottleneck link capacity is $\mu$ =100 Mbit/s. The simulation run time is $2000$ ms. We provide two different versions of parameters selection, which we call $\text{Vs}_1$ and $\text{Vs}_2$. In $\text{Vs}_1$ first class takes $25\%$ of the total bottleneck capacity and in $\text{Vs}_2$ it takes $50\%$.

The parameters we used are given in Table II.

Table II
TWO PARETO CLASSES, SIMULATION PARAMETERS

| Ver. | $c_1$ | $c_2$ | $m_1$ | $m_2$ | $\rho_1$ | $\rho_2$ | $\rho$ |
|---|---|---|---|---|---|---|---|
| $\text{Vs}_1$ | 10.0 | 1.25 | 0.5 | 6.8 | 0.25 | 0.50 | 0.75 |
| $\text{Vs}_2$ | 10.0 | 2.25 | 0.5 | 4.5 | 0.50 | 0.37 | 0.87 |

The results are given in Table III. We provide results for the NS-2 simulations and the values of the numerical mean sojourn times with the same parameters. We calculate the related gain of the Gittins policy in comparison with DropTail and LAS policies, $g_1 = \frac{\overline{T}^{DT} - \overline{T}^{Gitt}}{\overline{T}^{DT}}$ and $g_2 = \frac{\overline{T}^{LAS} - \overline{T}^{Gitt}}{\overline{T}^{LAS}}$.

We found that with the NS-2 simulations the gain of the Gittins policy in comparison with LAS policy is not so significant when the small jobs do not take a big part of the system load. As one can see in $\text{Vs}_2$ when the class-1 load is $50\%$ the related gain of the Gittins policy in comparison with LAS policy is $10\%$. In both versions the relative gain for the corresponding analytical system is much higher and reaches up to $36\%$. We explain this results with the phenomena related to the TCP working scheme.

Table III
MEAN SOJOURN TIMES

| Ver. | $\overline{T}^{DT}$ | $\overline{T}^{LAS}$ | $\overline{T}^{Gitt}$ | $g_1$ | $g_2$ |
|---|---|---|---|---|---|
| $\text{Vs}_1$ NS-2 | 18.72 | 2.10 | 2.08 | 88.89% | 0.95% |
| $\text{Vs}_1$ theory | PS: 4.71 | 1.58 | 1.01 | 78.56% | 36.08% |
| $\text{Vs}_2$ NS-2 | 6.23 | 2.03 | 1.83 | 70.63% | 9.85% |
| $\text{Vs}_2$ theory | PS: 6.46 | 3.25 | 2.19 | 66.10% | 32.62% |

### *H. Multiple Pareto classes*

We consider a multi-class single server $M/G/1$ queue. Jobs arrive to the system in $N$ classes. Jobs of $i$-th class, $i = 1, \ldots, N$ arrive according to the Poisson arrival processes with rates $\lambda_i$. Jobs size distributions are Pareto, namely $F_i(x) = 1 - \frac{1}{(x+1)^{c_i}}$, $i = 1, \ldots, N$. Then, the hazard rates $h_i(x) = \frac{c_i}{(x+1)}$, $i = 1, \ldots, N$, never cross. Without loss of generality, let us consider that $c_1 > c_2 > \ldots > c_N$. Let us define the values of $\theta_{i,j}$ and $g_{i,j}(x)$, $i, j = 1, \ldots, N$ in the following way $h_i(\theta_{i,j}) = h_j(0)$, $h_i(x) = h_j(g_{i,j}(x))$. Then we get $g_{i,j}(x) = \frac{c_j}{c_i}(x+1)$, $\theta_{i,j} = \frac{c_i}{c_j} - 1$. Let us notice that $\theta_{k,i} < \theta_{k,i+1}$ and $\theta_{i,k} > \theta_{i+1,k}$, $k = 1, \ldots, N$, $i = 1, \ldots, N-1$, $i \neq k$, $i \neq k+1$. Then the optimal policy is the following.

### *I. Optimal policy*

There are $N$ queues in the system. The class-1 jobs arrive to the system and go to the first-priority queue-1. There they are served until they get $\theta_{1,2}$ of service. Then they are moved to the queue-2, which is served only when the queue-1 is empty. In the queue-2 the job of class-1 are served with the jobs of class-2 with the Gittins policy. When the jobs of class-1 attain service $\theta_{1,3}$ they are moved to the queue-3. When the jobs of class-2 attain service $\theta_{2,3}$ they are also moved to the queue-3. And so on.

### *J. Mean sojourn times*

To find the expressions for the mean conditional sojourn times in the system we use the analysis which we used in interpretation of the mean conditional sojourn times expressions in the case of two class system, see Section III.

Let the tagged job be from class-1 of size $x$. The jobs which have the same priority in the system and which have to be served before the tagged job are: class-1 jobs of size less than $x$, class-$i$ jobs of size less than $g_{1,i}(x)$.

We denote $\overline{X_x^n}^{(i)}$ the $n$-th moment and $\rho_x^{(i)}$ the utilization factor for the distribution $F_i(x)$ of the class-$i$, $i = 1, \ldots, N$ truncated at $x$. The mean workload in the system which has to be served before the tagged job is then found with Pollaczek-Khinchin formula and equals to

$$W_{x, g_{1,2}(x), \ldots, g_{1,N}(x)} = \frac{\sum_{i=1}^{N} \lambda_i \overline{X^2_{g_{1,i}(x)}}}{2(1 - \sum_{i=1}^{N} \rho_{g_{1,i}(x)})}.$$

Then we formulate the theorem.

*Theorem 3:* For class-1 jobs of size $x$ such as $\theta_{1,p} < x < \theta_{1,p+1}$, $p = 1, \ldots, N$ and corresponding class-$k$ jobs with sizes $g_{1,k}(x)$, $k = 2, \ldots, p$ the mean conditional sojourn times are given by

$$T_1(x) = \frac{x + W(x, g_{1,2}(x), \ldots, g_{1,p}(x))}{1 - \rho_1(x) - \rho_2(g_{1,2}(x)) - \ldots - \rho_p(g_{1,p}(x))},$$

$$T_k(g_{1,k}(x)) = \frac{g_{1,k}(x) + W(x, g_{1,2}(x), \ldots, g_{1,p}(x))}{1 - \rho_1(x) - \rho_2(g_{1,2}(x)) - \ldots - \rho_p(g_{1,p}(x))}.$$

Here we consider that $\theta_{i,N+1} = \infty$, $i = 1, \ldots, N$.

*Proof:* Similar to the proof of Theorem 2. ■

## IV. Hyper-exponential and exponential classes

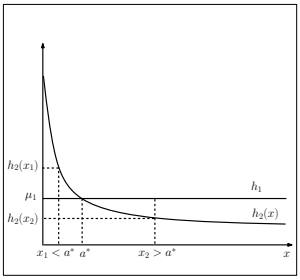


Figure 5. Exponential and HE classes, hazard rates.

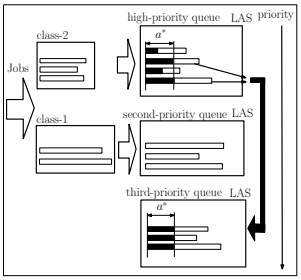


Figure 6. Exponential and HE classes, policy description.

We consider two class $M/G/1$ queue. Jobs of each class arrive with the Poisson arrival process with rates $\lambda_1$ and $\lambda_2$. The job size distribution of class-1 is exponential with mean $1/\mu_1$, and hyper-exponential (HE) with two phases for class-2 with the mean $(\mu_3 p + (1-p)\mu_2)/(\mu_2\mu_3)$. Namely,

$$F_1(x) = 1 - e^{-\mu_1 x}, \; F_2(x) = 1 - pe^{-\mu_2 x} - (1-p)e^{-\mu_3 x}. \quad (9)$$

The hazard rate function of class-1 is a constant and equals to $h_1 = \mu_1$. The hazard rate function of the class-2 $h_2(x) = \frac{p\mu_2 e^{-\mu_2 x} + (1-p)\mu_3 e^{-\mu_3 x}}{pe^{-\mu_2 x} + (1-p)e^{-\mu_3 x}}$, $x \geq 0$. is decreasing in $x$. As both hazard rate functions are non-increasing the optimal policy which minimizes the mean sojourn time is Gittins policy based on the value of the hazard function, which gives service to the jobs with the maximal hazard rate of the attained service.

For the selected job size distributions the hazard rate functions behave in different ways depending on parameters $\mu_1$, $\mu_2$, $\mu_3$ and $p$. The possible behaviors of the hazard rate functions determine the optimal policy in the system. If the hazard rate functions never cross, the hazard rate of class-1 is higher than the hazard rate of class-2, then the class-1 jobs are served with priority to class-2 jobs. This happens when $h_1 > h_2(x)$, $x \in (0, \infty)$. As $h_2(x)$ is decreasing, then this happens when $\mu_1 > h_2(0)$. Let us consider that $\mu_2 > \mu_3$, then as $h_2(0) = p\mu_2 + (1-p)\mu_3$ and $\mu_1 > h_2(0)$ if $\mu_1 > \mu_2 > \mu_3$. For this case it is known that the optimal policy is a strict priority policy, which serves the class-1 jobs with the strict priority with respect to the class-2 jobs. From our discussion it follows that this policy is optimal even if $\mu_2 > \mu_1 > \mu_3$, but still $\mu_1 > p\mu_2 + (1-p)\mu_3$.

Let us consider the case when $\mu_2 > \mu_1 > \mu_3$ and $\mu_1 < p\mu_2 + (1-p)\mu_3$. Then it exists the unique point of intersection of $h_2(x)$ and $h_1$. Let us denote $a^*$ the point of this intersection. The value of $a^*$ is the solution of $h_2(x) = \mu_1$. Solving this equation, we get that

$$a^* = \frac{1}{\mu_2 - \mu_3} \ln\left(\frac{p}{1-p}\frac{\mu_2 - \mu_1}{\mu_1 - \mu_3}\right).$$

The hazard rate function scheme is given in Figure 5. Then, the optimal policy is the following.

### A. Optimal policy.

There are three queues in the system, which are served with the strict priority between them. The second priority queue is served only when the first priority queue is empty and the third priority queue is served only when the first and second priority queues are empty. The class-2 jobs arrive to the system are served in the first priority queue with the LAS policy until they get $a^*$ amount of service. After they get $a^*$ amount of service they are moved to the third priority queue, where they are served according to the LAS policy. The class-1 jobs arrive to the system and go to the second priority queue, where they are served with LAS policy. Since $h_1(x) = \mu_1$, class-1 jobs can be served with any non-anticipating scheduling policy. The scheme of the optimal policy is given in Figure 6.

According to this optimal policy we find the expressions of the expected sojourn times for the class-1 and class-2 jobs.

### B. Expected sojourn times

Let us recall that the mean workload in the system for the class-1 jobs of size less than $x$ and class-2 jobs of size less than $y$ is $W_{x,y}$ and is given by (6). We prove the following Theorem.

*Theorem 4:* The mean conditional sojourn times in the $M/G/1$ queue with job size distribution given by (9) under Gittins optimal policy described in Subsection IV-A are given by

$$T_1(x) = \frac{x + W_{x,a^*}}{1 - \rho_x^{(1)} - \rho_{a^*}^{(2)}}, \quad x \in [0, \infty], \quad (10)$$

$$T_2(x) = \frac{x + W_{0,x}}{1 - \rho_x^{(2)}}, \quad x \in [0, a^*], \quad (11)$$

$$T_2(x) = \frac{x + W_{\infty,x}}{1 - \rho_\infty^{(1)} - \rho_x^{(2)}}, \quad y \in (a^*, \infty). \quad (12)$$

*Proof:* Similar to the proof of the Theorem 2. To find expressions of the mean conditional sojourn times we use the mean-value analysis and tagged job approach. ■

### C. Pareto and exponential classes

We can apply the same analysis for the case when class-1 job size distribution is exponential and class-2 job size distribution is Pareto. Let us consider the case when the hazard rate functions of class-1 and class-2 cross at one point.

Let $F_1(x) = 1 - e^{-\mu_1 x}$ and $F_2(x) = 1 - b_2^{c_2}/(x + b_2)^{c_2}$. Then $h_1 = \mu_1$ and $h_2(x) = c_2/(x + b_2)$. The crossing point is $a^* = c_2/\mu_1 - b_2$. When $a^* \leq 0$ the hazard rate functions do

not cross and then the optimal policy is to give strict priority to the class-1 jobs. If $a^* > 0$ then the hazard rate functions cross at one point and the optimal policy is the same as in the previous section. Then the expressions of the mean conditional sojourn timed of class-1 and class-2 are also (10), (11) and (12).

## V. Conclusions

In [1] Gittins considered an $M/G/1$ queue and proved that the so-called Gittins index rule minimizes the mean delay. The Gittins rule determines, depending on the service times of jobs, which job should be served next. Gittins derived this result as a by-product of his groundbreaking results on the multi-armed bandit problem. Gittins' results on the multi-armed bandit problem have had a profound impact and it is extremely highly cited. However, and in despite of the big body of literature on scheduling disciplines in single server queues, Gittins work in the $M/G/1$ context has not received much attention.

In [19] authors showed that Gittins' policy could be used to characterize the optimal scheduling policy when the hazard rate of the service time distribution is not monotone. In the current work we have used Gittins' policy to characterize the optimal scheduling discipline in a multi-class queue. Our results show that, even though all service times have a decreasing hazard rate, the optimal policy can significantly differ from LAS, which is known to be optimal in the single-server case. We demonstrate that in particular cases PS has much worse performance than Gittins policy.

Using NS-2 simulator we implemented the Gittins optimal policy in the router queue and provided simulations for several particular schemes. With the simulation results we found that Gittins policy can achieve $10\%$ gain in comparison with LAS policy and provides much better performance than DropTail policy.

In future research we may consider other types of service time distributions. The applicability of our results in real systems like the Internet should also be more carefully evaluated. We also would like to investigate the conditions under which Gittins policy gives significantly better performance than LAS policy.

## Acknowledgement

This research work is partially funded by the European Commission through the ECODE project (INFSO-ICT-223936) of the European Seventh Framework Programme (FP7).

## References

[1] J. Gittins, "Multi-armed Bandit Allocation Indices," *Wiley, Chichester*, 1989.

[2] P. Varaiya, J. Walrand, and C. Buyukkoc, "Extensions of the multiarmed bandit problem: the discounted case," *IEEE Transactions on Automatic Control*, vol. 30, pp. 426–439, 1985.

[3] P. Whittle, "Restless bandits: activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, pp. 287–298, 1988.

[4] R. Weber, "On the Gittins index for multiarmed bandits," *Annals of Applied Probability*, vol. 2, no. 4, pp. 1024–1033, 1992.

[5] J. Tsitsiklis, "A short proof of the Gittins index theorem," in *IEEE CDC*, 1993, pp. 389–390.

[6] M. Dacre, K. Glazebrook, and J. Niño-Mora, "The achievable region approach to the optimal control of stochastic systems," *Journal of the Royal Statistical Society. Series B, Methodological*, vol. 61, no. 4, pp. 747–791, 1996.

[7] E. Frostig and G. Weiss, "Four proofs of Gittins' multiarmed bandit theorem," *Applied Probability Trust*, 1999.

[8] D. Bertsimas and J. Niño-Mora, "Restless bandits, linear programming relaxations and a Primal-Dual index heuristic," *Operations Research*, vol. 48, pp. 80–90, 2000.

[9] C. Buyukkoc, P. Varaya, and J. Walrand, "The $c\mu$ rule revisited," *Adv. Appl. Prob.*, vol. 17, pp. 237–238, 1985.

[10] J. Shanthikumar and D. Yao, "Multiclass queueing systems: Polymatroidal structure and optimal scheduling control," *Operations Research*, vol. 40, no. 2, pp. 293–299, 1992.

[11] P. Nain and D. Towsley, "Optimal scheduling in a machine with stochastic varying processing rate," *IEEE/ACM Transactions on Automatic Control*, vol. 39, pp. 1853–1855, 1994.

[12] M. Nabe, M. Murata, and H. Miyahara, "Analysis and modeling of World Wide Web traffic for capacity dimensioning of Internet access lines," *Perform. Eval.*, vol. 34, no. 4, pp. 249–271, 1998.

[13] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 835–846, 1997.

[14] C. Williamson, "Internet traffic measurement," *IEEE Internet Computing*, vol. 5, pp. 70–74, 2001.

[15] K. Sevcik, "Scheduling for minimum total loss using service time distributions," *Journal of the ACM*, vol. 21, pp. 66–75, 1974.

[16] S. Yashkov, "Mathematical problems in the theory of shared-processor systems," *Journal of Mathematical Sciences*, vol. 58, pp. 101–147, 1992.

[17] G. Klimov, "Time-sharing service systems. i," *Theory of Probability and Its Applications*, vol. 19, pp. 532–551, 1974.

[18] ——, "Time-sharing service systems. ii." *Theory of Probability and Its Applications*, vol. 23, pp. 314–321, 1978.

[19] S. Aalto and U. Ayesta, "Mean delay optimization for the $M/G/1$ queue with Pareto type service times," in *Extended abstract in ACM SIGMETRICS 2007, San Diego, CA*, 2007, pp. 383–384.

[20] L. Kleinrock, *Queueing systems*. John Wiley and Sons, 1976, vol. 2.