Conditional sojourn time of optimal scheduling policy in a multi-class single-server queue [⋆]

K.E. Avrachenkov[1], U. Ayesta[2,3], N. Osipova

[1] INRIA Sophia Antipolis,
2004, route des Lucioles – BP 93 - 06902 Sophia Antipolis
[2] BCAM – Basque Center for Applied Mathematics
Bizkaia Technology Park 500, 48170 Derio, Spain
[3] IKERBASQUE – Basque Foundation for Science, 48011 Bilbao, Spain

Abstract. Based on the groundbreaking result of Gittins on the multi-armed bandit problem we provide a characterization of the optimal non-anticipating scheduling policy in a multi-class single-server queue. We apply Gittins' framework to characterize the optimal policy when the service time distribution of the various classes belong to the set of Decreasing Hazard Rate (DHR) distributions, like Pareto or hyper-exponential. When there is only one class it is known that the Least Attained Service (LAS) 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 and the collective mark method we obtain, for every class, the mean sojourn time conditioned on the service requirement. Numerical computations show that the performance gain of Gittins' policy can be significant.

Keywords: multi-level processor-sharing, gittins index policy, multi-armed bandits, size-based scheduling, collective mark method.
AMS 2000 subject classification: Primary 68M20, Secondary 60K25, 90B22

1 Introduction

In [7], 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 attained service time 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 [18, 20, 19, 17, 5, 6, 2]). However, the optimality result of the Gittins index in the context of an $M/G/1$ queue has not been fully exploited, and it has not received the attention it deserves.

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

Gittins' result generalizes 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 μ_k refer to the holding cost and the inverse of the mean service requirement of class k jobs, respectively. It is known (see for example [3, 16, 13]) 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.

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 modelled with distributions with a decreasing hazard rate [12, 4, 21]. 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), FCFS and LAS. We find that in the particular case of two classes and Pareto-type service time distribution the Gittins policy outperforms LAS by nearly 25%.

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. In [14] we implement the Gittins scheduling policy in the NS-2 simulator and perform experiments to evaluate the achievable performance gain.

The rest of the paper is organized as follows: In Section 2 we review 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. In particular, we derive analytic expressions for the mean conditional sojourn times and study various properties of the optimal policy. At the end of Section 3 we generalize the results to multiple Pareto classes. In Section 4 we study the case when one distribution is exponential and the other distribution is hyper-exponential with two phases. In Section 5 we present numerical results.

2 Gittins policy in a multi-class $M/G/1$ queue

Let Π denote the set of non-anticipating scheduling policies. Popular disciplines such as PS, FCFS and LAS belong to Π. Important disciplines that do not belong to Π are Shortest Remaining Processing First (SRPT) and Shortest Processing Time First.

Let us consider a multi-class $M/G/1$ queue. Let X_i denote the service time with distribution $\mathbb{P}(X_i \leq x) = F_i(x)$ for every class $i = 1, \ldots, N$. The density is denoted by $f_i(x)$ and the complementary distribution by $\overline{F}_i(x) = 1 - F_i(x)$. Class-$i$ jobs arrive according to a Poisson process with rate λ_i, and let $\lambda = \sum_{i=1}^{N} \lambda_i$ denote the total arrival rate.

Definition 1. *For any $a, \varDelta \geq 0$, let*

$$J_i(a, \varDelta) = \frac{\int_0^{\varDelta} f_i(a+t)\mathrm{d}t}{\int_0^{\varDelta} \overline{F}_i(a+t)\mathrm{d}t} = \frac{\overline{F}_i(a) - \overline{F}_i(a+\varDelta)}{\int_0^{\varDelta} \overline{F}_i(a+t)\mathrm{d}t}. \tag{1}$$

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

$$\begin{aligned} J_i(a, 0) &= \frac{f_i(a)}{\overline{F}_i(a)} = h_i(a), \\ J_i(a, \infty) &= \frac{\overline{F}_i(a)}{\int_0^{\infty} \overline{F}_i(a+t)\,\mathrm{d}t} = 1/\mathbb{E}[X_i - a | X_i > a]. \end{aligned}$$

Note further that $J_i(a, \varDelta)$ is continuous with respect to $\varDelta$.

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

$$G_i(a) = \sup_{\varDelta \geq 0} J_i(a, \varDelta), \tag{2}$$

for any $a \geq 0$.

We call $G_i(a)$ the *Gittins index* after the author of book [7], 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 [22] and references therein, in particular the works by Klimov [10, 11].

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

$$\varDelta_i^*(a) = \sup\{\varDelta \geq 0 \mid J_i(a, \varDelta) = G_i(a)\}. \tag{3}$$

By definition, $G_i(a) = J(a, \varDelta_i^*(a))$ for all a.

Definition 4. *The Gittins index policy π_G is the scheduling discipline that at every instant of time gives service to the job in the system with highest $G_i(a)$, where i denotes the class and a is the job's attained service.*

We denote by $T_i^\pi(x)$ the mean conditional sojourn time for the class-i job of size x, $i = 1, \ldots, N$. Let $\overline{T}_i^\pi$ denote the mean (unconditional) sojourn time for class-i users, and let $\overline{T}^\pi$ denote the mean sojourn time in the system. It follows that

$$\overline{T}^\pi = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \overline{T}_i^\pi.$$

We can now state the following result:

Theorem 1. **[7, Theorem 3.28]** *The Gittins index policy minimizes the mean sojourn time in the system among all non-anticipating scheduling policies, i.e. for any $\pi \in \Pi$,*

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

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

Decreasing Hazard Rate. Let the service time distribution of class-i have a decreasing hazard rate. It is possible to show, see [1], that if $h_i(x)$ is non-increasing, the function $J_i(a, \Delta)$ is non-increasing in Δ. Thus

$$G_i(a) = J_i(a, 0) = h_i(a). \qquad (4)$$

As a consequence we obtain the following result.

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

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

The policy presented in Proposition 1 is an optimal policy for the multi-class single-server queue. Let us notice that for the single-class single-server queue the Gittins policy is equivalent to LAS. 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 next.

3 Two Pareto classes

We consider now the case when job sizes are distributed according to Pareto distribution. We thus consider, for $i = 1, 2$,

$$F_i(x) = 1 - \frac{b_i^{c_i}}{(x + b_i)^{c_i}}, \quad x \geq 0. \qquad (5)$$

It follows that $\mathbb{E}[X_i] = b_i/(c_i - 1)$, $i = 1, 2$. The density function is given by $f_i(x) = b_i^{c_i} c_i/(x + b_i)^{c_i+1}$, $i = 1, 2$ and the hazard rate functions are

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

The two hazard rate 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.

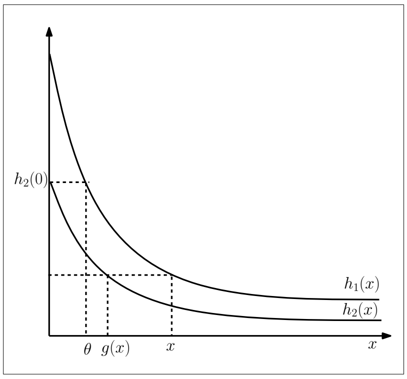

Fig. 1. Two Pareto classes: hazard rate functions that do not cross

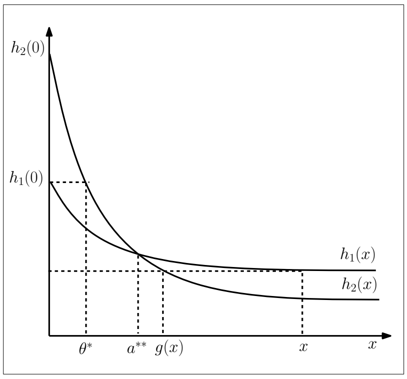

Fig. 2. Two Pareto classes: hazard rate functions cross

Let us first consider the case when the hazard rate function do not cross, so $a^{**} < 0$. This happens when $b_1/b_2 < c_1/c_2$ and it follows $h_1(x) > h_2(x)$, for all $x \geq 0$. Let θ and $g(x)$ be such that:

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

We can see that $g(\theta) = 0$. In the case of Pareto service time distributions we then get that

$$g(x) = \frac{c_2}{c_1}(x + b_1) - b_2, \quad \theta = \frac{c_1 b_2 - c_2 b_1}{c_2}.$$

We observe that the hazard-rate of a class-1 job with attained service x and the hazard rate of a class-2 job with attained service $g(x)$ are the same, see Figure 1.

3.1 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 θ amount of service it is moved to the second low priority queue. Class-2 jobs are put immediately in the low priority queue. The low priority queue is served only when the high priority queue is empty. In both queues, the job with highest hazard-rate is served.

Let us now derive the expressions of the mean conditional sojourn time for class-1 and class-2 jobs.

3.2 Mean conditional sojourn times

Let us denote by indices $[]^{(1)}$ and $[]^{(2)}$ the values for class-1 and class-2, respectively. 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 $F(x)$ for $x \leq y$ and 1 when $x > y$. Let us denote $W_{x,y}$ the mean workload in the system which consists only of class-1 jobs with service times truncated at x and of class-2 jobs with service times truncated at y. According to the Pollaczek-Khinchin formula we have

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

Now let us formulate the following Proposition which we prove in the Appendix.

Proposition 2. *In the two-class $M/G/1$ queue where the job size distributions are Pareto and which is scheduled with the Gittins policy, the mean conditional sojourn times for class-1 and class-2 jobs are*

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

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

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

Proof: The proof is given in the appendix. □

The obtained expressions (6), (7) and (8) can be interpreted using the tagged-job and mean value approach. Let us consider class-1 jobs. Jobs of size $x \leq \theta$ are served in the high priority queue according to the LAS policy, and we have (see [9, Section 4.6]), $T_1(x) = \frac{x+W_{x,0}}{1-\rho_x^{(1)}}$, $x \leq \theta$, where $W_{x,0}$ is the mean workload and $\rho_x^{(1)}$ is the mean load in the system for class-1 jobs with the service time distribution truncated at x.

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

- x is the required service time;

- $W_{x,g(x)}$ is the mean workload which the tagged job finds in the system and that has to be served before its service is completed;
- $T_1(x)(\rho_x^{(1)} + \rho_{g(x)}^{(2)})$ is the mean workload that arrives to the system during the sojourn time of the tagged job and which has to be served before its service is completed.

The expressions of $W_{x,g(x)}$, $\rho_x^{(1)}$ and $\rho_{g(x)}^{(2)}$ are given in the appendix. Equation (8) can be similarly interpreted.

3.3 Properties of the optimal policy

Property 1. When class-2 jobs arrive to the server they are not served immediately, but wait until the high priority queue is empty. The mean waiting 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 θ may also wait in the system to be served in the low priority queue (until the high priority queue is emptied), 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 what happens during a time interval in which no new job arrives. We concentrate on the low priority queue and we consider that all class-1 jobs and all 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

$$h_1(x_1) = h_2(x_2).$$

Consider a very small time interval of length $\mathit{\Delta}$, then let $\mathit{\Delta}_1$ and $\mathit{\Delta}_2$ be the amount of service that each class-1 and class-2 jobs receive. It then follows that

$$n_1\mathit{\Delta}_1 + n_2\mathit{\Delta}_2 = \mathit{\Delta}. \tag{9}$$

In addition it follows that $h_1(x_1+\mathit{\Delta}_1) = h_2(x_2+\mathit{\Delta}_2)$, and since $\mathit{\Delta}$ is very small we can then approximate

$$h_i(x+\mathit{\Delta}_i) = h_i(x) + \mathit{\Delta}_i h_i'(x), \quad i = 1, 2,$$

and as a consequence $\varDelta_1 h_1'(x_1) = \varDelta_2 h_2'(x_2)$. Then using (9) we get

$$\frac{\varDelta_1}{\varDelta} = \frac{h_2'(x_2)}{n_1 h_2'(x_2) + n_2 h_1'(x_1)}, \quad \frac{\varDelta_2}{\varDelta} = \frac{h_1'(x_1)}{n_1 h_2'(x_2) + n_2 h_1'(x_1)}.$$

In the case of two Pareto distributions this becomes we get:

$$\frac{\varDelta_1}{\varDelta} = \frac{c_1}{n_1 c_1 + n_2 c_2}, \quad \frac{\varDelta_2}{\varDelta} = \frac{c_2}{n_1 c_1 + n_2 c_2}.$$

Interestingly, the service rates of class-1 and class-2 jobs do not depend on the current jobs' attained services.

Property 3. According to the definition of the function $g(x)$ we can conclude that the class-1 job of size x and class-2 job of size $g(x)$, if they coincide in the system, will leave the system at the sime time. However on average, the mean sojourn times do not coincide, i.e.,

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

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

3.4 Two Pareto classes with hazard rate functions that cross

Now let us consider the case when the hazard rate functions cross, namely, when $a^{**} = (c_2 b_1 - c_1 b_2)/(c_1 - c_2) \geq 0$, see Figure 2. Assume $c_1 > c_2$, then $h_1(0) < h_2(0)$ and thus 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 θ^* is such that $h_2(\theta^*) = h_1(0)$ and $g(\theta^*) = 0$. Class-2 jobs with attained service than θ and class-1 jobs share the service. If at any given time, more than one job have the highest hazard rate, then they will be served in such a way so the hazard rates remain equal. According to this analysis we have the following corollary of Proposition 2:

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*

$$\begin{aligned} T_1(x) &= \frac{x + W_{x,g(x)}}{1 - \rho_x^{(1)} - \rho_{g(x)}^{(2)}}, \quad x \geq 0, \\ T_2(x) &= \frac{x + W_{0,x}}{1 - \rho_x^{(2)}}, \quad x \leq \theta^*, \\ T_2(g(x)) &= \frac{g(x) + W_{x,g(x)}}{1 - \rho_x^{(1)} - \rho_{g(x)}^{(2)}}, \quad x > \theta^*. \end{aligned}$$

Proof: The proof is similar to that of Proposition 2 and is omitted. □

3.5 Multiple Pareto classes

We now consider a queue with an arbitrary number of classes. Jobs size distributions are Pareto of the form

$$F_i(x) = 1 - \frac{1}{(x+1)^{c_i}}, \quad i = 1, \ldots, N.$$

Then, the hazard rates

$$h_i(x) = \frac{c_i}{(x+1)}, \quad i = 1, \ldots, N,$$

never cross. Without loss of generality, let us consider that $c_1 > c_2 > \ldots > c_N$. We define the values of $\theta_{i,j}$ and $g_{i,j}(x)$, $i, j = 1, \ldots, N$ in the following way

$$\begin{aligned} h_i(\theta_{i,j}) &= h_j(0), \\ h_i(x) &= h_j(g_{i,j}(x)). \end{aligned}$$

Then we get

$$g_{i,j}(x) = \frac{c_j}{c_i}(x+1) - 1, \quad \theta_{i,j} = \frac{c_i}{c_j} - 1.$$

We note that $\theta_{k,i} < \theta_{k,i+1}$ and $\theta_{i,k} > \theta_{i+1,k}$, $k = 1, \ldots, N$, $i = 1, \ldots, N-1$, $i \neq k$, $i \neq k+1$, see Figure 3. Let us denote that $\theta_{i,i} = 0$ for $i = 1, \ldots, N$.

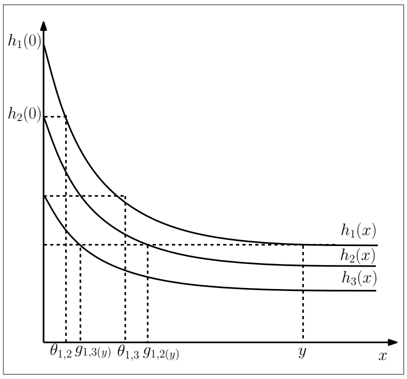

Fig. 3. Multiple Pareto classes: hazard rate functions

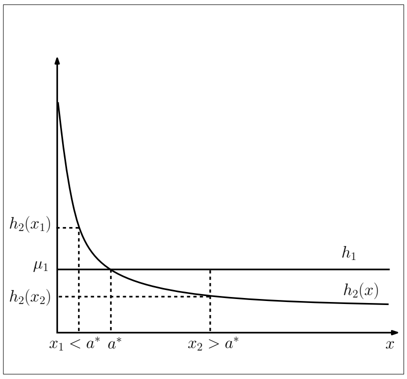

Fig. 4. Hazard rate for exponential and hyperexponential service time distributions.

Optimal policy. There are N queues in the system. Class-1 jobs arrive to the system and go to the first-priority queue-1. There they are served according to the LAS policy until they receive $\theta_{1,2}$ amount of service. Then they are moved to the queue-2, which is served only when the queue-1 is empty. In the queue-2 jobs of class-1 are served together with jobs of class-2. Every moment the service is given to the job with the highest $h_i(a)$, $i = 1, 2$. When jobs of class-1 attain service $\theta_{1,3}$ they are moved to the queue-3 and similarly when class-2 jobs attain service $\theta_{2,3}$ they are also moved to the queue-3. In queue-3 the jobs of class-1, class-2 and class-3 are served together. Every moment of time the service is given to the job with the highest $h_i(a)$, $i = 1, 2, 3$, where a is a jobs attained service. The policy operates similarly for an arbitrary number of classes.

Conditional Sojourn Time. The mean conditional sojourn time for the tagged job belonging to class-k consists of the service requirement, the mean workload in the system upon arrival which has to be served before the tagged job, and the mean workload which arrives during the sojourn time of the tagged job and has to be served before it. Let the tagged job belong to class-1 and let x be its size. 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 $g_{1,i}(x)$.

For the distribution $F_i(x)$ truncated at y, let $\overline{X_y^n}^{(i)}$ denote the n-th moment and let $\rho_y^{(i)}$ denote the utilization factor, respectively. The mean workload in the system which has to be served before the tagged job finishes its service is then given with Pollaczek-Khinchin formula and equals to

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

We can then state the following result.

Proposition 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*

$$T_1(x) = \frac{x + W_{x,g_{1,2}(x),\ldots,g_{1,p}(x)}}{1 - \rho_x^{(1)} - \rho_{g_{1,2}(x)}^{(2)} - \cdots - \rho_{g_{1,p}(x)}^{(p)}},$$

$$T_k(g_{1,k}(x)) = \frac{g_{1,k}(x) + W_{x,g_{1,2}(x),\ldots,g_{1,p}(x)}}{1 - \rho_x^{(1)} - \rho_{g_{1,2}(x)}^{(2)} - \cdots - \rho_{g_{1,p}(x)}^{(p)}}.$$

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

Proof: The proof, even though notationally cumbersome, is similar to that of Proposition 2. □

4 Hyperexponential and exponential service time distributions

In this section we consider a two class $M/G/1$ queue. The job size distributions for class-1 and class-2 jobs are exponential and hyperexponential:

$$F_1(x) = 1 - e^{-\mu_1 x}, \; F_2(x) = 1 - pe^{-\mu_2 x} - (1-p)e^{-\mu_3 x}. \tag{10}$$

The mean service requirement for class-1 and class-2 jobs are $1/\mu_1$ and $(\mu_3 p + (1-p)\mu_2)/(\mu_2\mu_3)$, respectively.

Note that the hazard rates are

$$h_1(x) = \mu_1,\ h_2(x) = \frac{p\mu_2 e^{-\mu_2 x} + (1-p)\mu_3 e^{-\mu_3 x}}{pe^{-\mu_2 x} + (1-p)e^{-\mu_3 x}},\ x \geq 0.$$

The hazard rate functions are depicted in Figure 4. We note that $h_2(x)$ is decreasing in x. As both hazard rate functions are non-increasing the Gittins (optimal) policy will give service to the job(s) with highest hazard rate.

The possible behaviors of the hazard rate functions, which will depend on the values of μ_1, μ_2, μ_3 and p, 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 class-1 jobs are served with priority to class-2 jobs. This happens when $h_1(x) = \mu_1 > h_2(0)$. Let us assume that $\mu_2 > \mu_3$, then since $h_2(0) = p\mu_2 + (1-p)\mu_3$, we conclude that if $\mu_1 > \mu_2 > \mu_3$ then $\mu_1 > h_2(0)$. In this case the optimal policy is a strict priority policy which serves class-1 jobs with strict priority.

When $\mu_2 > \mu_1 > \mu_3$ and $\mu_1 < p\mu_2 + (1-p)\mu_3$, then there exists a unique point, denoted by a^*, of intersection between $h_2(x)$ and $h_1(x)$. It is easy to see that

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

Optimal policy. There are three queues in the system, which are served with 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. Class-2 jobs arrive to the system are served in the first priority queue according to LAS they get a^* units of service. Afterwards they are moved to the third priority queue. Class-1 jobs arrive to the system and go to the second priority queue. Since $h_1(x) = \mu_1$, class-1 jobs can be served with any non-anticipating scheduling policy.

4.1 Mean conditional sojourn time

Let us recall that the mean workload in the system for 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 have the following result:

Proposition 4. *In the two-class $M/G/1$ queue where the job size distributions are exponential and hyperexponential and which is scheduled with the Gittins* *policy, the mean conditional sojourn times for class-*1 *and class-*2 *jobs are:*

$$T_1(x) = \frac{x + W_{x,a^*}}{1 - \rho_x^{(1)} - \rho_{a^*}^{(2)}}, \quad x \geq 0, \tag{11}$$

$$T_2(x) = \frac{x + W_{0,x}}{1 - \rho_x^{(2)}}, \quad x \leq a^*, \tag{12}$$

$$T_2(x) = \frac{x + W_{\infty,x}}{1 - \rho_\infty^{(1)} - \rho_x^{(2)}}, \quad x > a^*. \tag{13}$$

Proof: The proof is similar to that of Proposition 2 and is therefore omitted. □

4.2 Pareto and exponential service time distributions

We can apply the same analysis when the service time distribution of one class is Pareto instead of hyperexponential. Let $F_1(x) = 1 - e^{-\mu_1 x}$ and $F_2(x) = 1 - b_2^{c_2}/(x + b_2)^{c_2}$. Then $h_1(x) = \mu_1$ 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 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 (11), (12) and (13).

5 Numerical results

We consider two classes with Pareto service time distribution and we compare the mean unconditional sojourn time for various policies. We consider two different set of parameters, which we call V_1 and V_2 (see Table 1). The load of class-1 is kept fixed, and we vary the arrival rate of class-2 in order to change the total load in the system.

Table 1. Two Pareto classes, parameters

V	c_1	c_2	$\mathbb{E}[X_1]$	$\mathbb{E}[X_2]$	ρ_1	ρ_2	ρ
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

For the Gittins policy, we calculate the mean unconditional sojourn time using equations (6), (7) and (8).

In addition to Gittins' policy, we also consider PS, FCFS and LAS. The mean unconditional sojourn time for these policies can be found for example in [9]. For PS we have

$$\overline{T}^{PS} = \frac{\rho/\lambda}{1 - \rho},$$

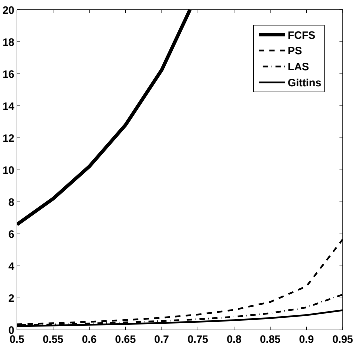

Fig. 5. Two Pareto classes, mean sojourn times with respect to the load ρ, V_1

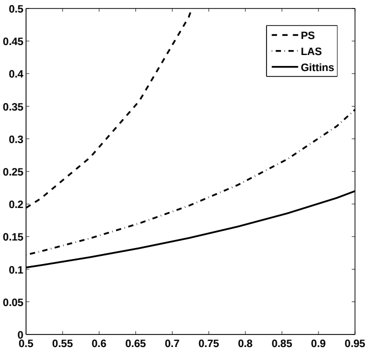

Fig. 6. Two Pareto classes, mean sojourn times with respect to the load ρ, V_2

and for FCFS

$$\overline{T}^{FCFS} = \rho/\lambda + W_{\infty,\infty},$$

where $W_{\infty,\infty}$ means the total mean unfinished work in the system. For LAS we have

$$\overline{T}^{LAS} = \frac{1}{\lambda_1 + \lambda_2} \int_0^\infty \overline{T}^{LAS}(x)(\lambda_1 f_1(x) + \lambda_2 f_2(x))\mathrm{d}x,$$

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

The mean sojourn times for the parameters sets V_1 and V_2 are presented in Figures 5 and 6. 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. The relative gains in mean sojourn time between the Gittins and LAS and Gittins and PS policies are the following. For the set of parameters V_1:$\max_{\rho_2} \frac{\overline{T}^{FCFS}-\overline{T}^{\pi G}}{\overline{T}^{FCFS}} = 0.99$, $\max_{\rho_2} \frac{\overline{T}^{PS}-\overline{T}^{\pi G}}{\overline{T}^{PS}} = 0.78$ and $\max_{\rho_2} \frac{\overline{T}^{LAS}-\overline{T}^{\pi G}}{\overline{T}^{LAS}} = 0.45$. For the set of parameters V_2: $\max_{\rho_2} \frac{\overline{T}^{PS}-\overline{T}^{\pi G}}{\overline{T}^{PS}} = 0.98$ and $\max_{\rho_2} \frac{\overline{T}^{LAS}-\overline{T}^{\pi G}}{\overline{T}^{LAS}} = 0.39$. The maximal gain is achieved when the load in the system is around 0.9. We note that the performance with PS is significantly worse than with LAS and Gittins.

6 Conclusions

In [7], 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 jobs attained service, which job should be served next. Gittins derived this result as a by-product of his groundbreaking results on the multi-armed bandit problem.

In [1], the authors showed that the Gittins policy could be used to characterize the optimal scheduling policy when the hazard rate of the service time distribution is not monotone. In this paper we have used the Gittins policy in order 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-class case.

For several important particular cases we have calculated analytically the mean conditional sojourn time. Our approach relies on the combination of the collective mark method and the tagged-job approach pioneered by Kleinrock. Numerical computations show that the optimal multiclass policy can significantly outperform classical scheduling policies like PS and FCFS.

References

1. Aalto, S., Ayesta, U., and Righter, R.: On the optimal scheduling discipline in a single-server queue. Queueing Systems (special issue The Erlang Centennial), **63** (2009) 437–458.
2. Bertsimas, D., Niño-Mora, J.: Restless bandits, linear programming relaxations, and a primal-dual index heuristic. Operations Research, **48** (2000) 91-113.
3. Buyukkoc, C., Varaya, P., and Walrand, J.: The $c\mu$ rule revisited. Adv. Appl. Prob, **17** (1985) 237–238.
4. Crovella, M.E. and Bestavros, A.: Self-similarity in world wide web traffic: evidence and possible causes. In Proc. ACM SIGMETRICS 1996, pages 160–169, Philadelphia, PA, 1996.
5. Dacre, M., Glazebrook, K., and Niño-Mora, J.: The achievable region approach to the optimal control of stochastic systems. Journal of the Royal Statistical Society. Series B, Methodological, **61** (1996) 747–791.
6. Frostig, E., and Weiss, G.: Four proofs of Gittins' multiarmed bandit theorem. Applied Probability Trust (1999)
7. Gittins, J.C.: Multi-armed Bandit Allocation Indices. Wiley, Chichester, 1989.
8. Kleinrock, L.: Queueing Systems, vol. 1. John Wiley and Sons, 1976.
9. Kleinrock, L. Queueing Systems, vol. 2. John Wiley and Sons, 1976.
10. Klimov, G.P.: Time-sharing service systems. I. Theory of Probability and Its Applications, **19** (1974) 532–551.
11. Klimov, G.P.: Time-sharing service systems. II. Theory of Probability and Its Applications, **23** (1978) 314–321.
12. Nabe, M., Murata, M., and Miyahara, H.: Analysis and modeling of world wide web traffic for capacity dimensioning of Internet access lines. Performance Evaluation, **34** (1998) 249–271
13. Nain, P., and Towsley, T.: Optimal scheduling in a machine with stochastic varying processing rate. IEEE/ACM Transactions on Automatic Control, **39** (1994) 1853–1855.
14. Osipova, N., Ayesta, U. and Avrachenkov, K.E.: Optimal policy for multi-class scheduling in a single server queue. In Proc. of ITC-21, 2009.

15. Sevcik, K., Scheduling for minimum total loss using service time distributions. Journal of the ACM, **21** (1974) 66–75.
16. Shanthikumar, J.G., and Yao, D.: Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. Operations Research, **40** (1992) 293–299.
17. Tsitsiklis, J.N.: A short proof of the Gittins index theorem. In Proc. IEEE CDC 1993, 389–390, 1993.
18. Varaiya, P., Walrand, J., and Buyukkoc, C.: Extensions of the multiarmed bandit problem: the discounted case. IEEE Transactions on Automatic Control, **30** (1985) 426–439.
19. Weber, R.: On the Gittins index for multiarmed bandits. Annals of Applied Probability, **2** (1992) 1024–1033.
20. Whittle, P.: Restless bandits: Activity allocation in a changing world. Journal of Applied Probability, **25** (1998) 287–298.
21. Williamson, C.: Internet traffic measurement. IEEE Internet Computing, **5** (2001) 70–74.
22. Yashkov, S.F.: Mathematical problems in the theory of processor sharing queueing systems. Journal of Soviet Mathematics, **58** (1992) 101–147.

Appendix: Proof of Proposition 2

Class-1 jobs of size $x \le \theta$ are served in the high priority queue with LAS policy, so the expression for the mean conditional sojourn time for this case is known, see [9, Section 4.6], as is given by (6).

Let us consider class-1 jobs with sizes $x > \theta$ and class-2 jobs, which are served in the low priority queue. There is a strict priority between the queues and the low priority queue is served only when the high priority queue is empty. Then the low priority queue is a queue with batch arrivals. To find the expressions of the mean conditional sojourn times in the system we carry out an analysis similar to the one of Kleinrock for Multi-Level Processor-Sharing queue see [9, Section 4.7].

In the following analysis we consider only class-1 jobs with service requirements smaller than x and class-2 jobs with service requirements smaller than $g(x)$. We have the following Lemma.

Lemma 1. *The mean conditional sojourn times for class-*1 *job of size* $x > \theta$ *and for class-*2 *job of size* $g(x) > 0$ *is given by*

$$T_1(x) = \frac{\theta + W_{\theta,0}}{1-\rho_\theta^{(1)}} + \frac{\alpha_1(x-\theta, g(x))}{1-\rho_\theta^{(1)}}, \tag{14}$$

$$T_2(g(x)) = \frac{W_{\theta,0}}{1-\rho_\theta^{(1)}} + \frac{\alpha_2(x-\theta, g(x))}{1-\rho_\theta^{(1)}}, \tag{15}$$

where $\alpha_1(x-\theta, g(x))$ and $\alpha_2(x-\theta, g(x))$ are the times spent in the low priority queue by class-1 and class-2 jobs respectively and equal to

$$\alpha_1(x-\theta, g(x)) = \frac{x-\theta+A_1(x)+W_b}{1-\rho_b},$$
$$\alpha_2(x-\theta, g(x)) = \frac{g(x)+A_2(g(x))+W_b}{1-\rho_b},$$

where W_b is the mean workload in the low priority queue which the tagged batch sees when arrives to the low priority queue, ρ_b is the load in the low priority queue and $A_i(x)$, $i=1,2$ are the mean workload that arrives together with the tagged job in the batch.

Proof: Let us consider that the tagged job is from class-1 and has a size $x > \theta$. The time it spends in the system consists of the mean time it spends in the high priority queue. This time is $\frac{\theta+W_{\theta,0}}{1-\rho_\theta}$ as it has to be served only with class-1 jobs until it gets θ amount of service. After the tagged job is moved to the low priority queue after waiting while the high priority queue becomes empty. The time $\alpha_1(x-\theta)$ is the time spent by the tagged job in the low priority queue. This time consists of the time spent to serve the job itself, $x-\theta$, of the mean workload in the low priority queue which the tagged job finds, W_b, of the mean work which arrives in the batch with the tagged job, $A_1(x)$ and of the mean work which arrives during the sojourn time of the tagged job, $\alpha_1(x-\theta)\rho_b$.

We use the same analysis for the mean conditional sojourn time of the class-2 job of size $g(x)$. □

Now let us find the expressions for the W_b, ρ_b, $A_1(x)$ and $A_2(x)$. Let us define the truncated distribution $F_{1,\theta,x}(y) = F_1(y), \theta < y < x$ and $F_{1,\theta,x}(y) = 0, y < \theta, y > x$. Let $\overline{X^n_{\theta,x}}^{(i)}$ be the n-th moment and $\rho^{(i)}_{\theta,x}$, $i=1,2$ be the utilization factor for this truncated distribution. We use this notation because class-1 jobs that arrive in the batch have already received θ units of service.

Let N_i be the random variable denoting the number of class-i jobs in a batch. Let $X^{(1)}_{\theta,x}$ and $X^{(2)}_{g(x)}$ denote the service requirement of a class-1 and class-2 jobs in a batch, respectively. Then

$$Y_b = \sum_{i=1}^{N_1} X^{(1)}_{i,\theta,x} + \sum_{i=1}^{N_2} X^{(2)}_{i,g(x)},$$

is the random variable which denotes the total service requirement in a batch. Let us denote as λ_b the batch arrival rate. We know that $\lambda_b = \lambda_1 + \lambda_2$. According to the previous notations we can write

$$\rho_b = \lambda_b \mathbb{E}[Y_b],$$

here $\mathbb{E}[Y_b]$ is the mean work that a batch brings and by Pollaczek-Khinchin

$$W_b = \frac{\lambda_b \mathbb{E}[Y_b^2]}{2(1-\rho_b)}.$$

We note that W_b does not depend on which class the tagged job belongs to. Since we know the first and the second moments of $X^{(1)}_{\theta,x}$, $X^{(2)}_{g(x)}$, in order to find ρ_b and W_b we need to know the first and the second moments of N_i, $i = 1, 2$. To find these values we use the collective marks method see [8, Chapter 7].

Calculation of the Generating Function with the collective mark method. We propose a two dimensional generating function $G(z_1, z_2)$, which we obtain using the collective mark method:

Definition 5. *Let us mark jobs in a batch in the following way. We mark a job of class-1 with probability $1 - z_1$, then z_1 is a probability that a job of class-1 is not marked. Equivalently, class-2 jobs are marked with probability $1 - z_2$. Let p_{n_1,n_2} be the probability that n_1 class-1 and n_2 class-2 jobs arrive in the batch. Then the Generating Function*

$$G(z_1, z_2) = \sum_{n1} \sum_{n2} z_1^{n_1} z_2^{n_2} p_{n_1,n_2}$$

gives the probability that there are no marked jobs in the batch.

Let us define as a “starter” or S a tagged job. We distinguish the cases when the starter S belongs to class-1 or class-2 and denote by $G_1(z_1, z_2)$ and $G_2(z_1, z_2)$ the probabilities that there are no marked jobs in the batch when the starter is from the class-1 and class-2, respectively. We consider first the case when the starter belongs to class 1. We consider two cases depending on whether the service requirement of the starter is less or larger than θ. Conditioning on which class the starter belongs to we have:

$$G(z_1, z_2) = \frac{\lambda_1}{\lambda_b}([G_1(z_1, z_2), S \le \theta] + [G_1(z_1, z_2), S > \theta]) + \frac{\lambda_2}{\lambda_b} G_2(z_1, z_2).$$

We have the following characterization of $G(z_1, z_2)$:

Lemma 2. *The Generating function satisfies*

$$\begin{aligned} G(z_1, z_2) = \frac{\lambda_1}{\lambda_b}(\int_0^\theta e^{-\lambda_1 x(1-G_1(z_1,z_2)) - \lambda_2 x(1-z_2)} \mathrm{d}F_1(x) + \\ + z_1 e^{-\lambda_1 \theta(1-G_1(z_1,z_2)) - \lambda_2 \theta(1-z_2)} \overline{F}_1(\theta)) + \frac{\lambda_2}{\lambda_b} z_2. \end{aligned} \tag{16}$$

Proof: When a class-1 job arrives to the system it creates the busy period. Until this job does not receive θ amount of service the low priority queue will not be served. Thus, jobs that arrive to the low priority queue and jobs which are already in the low priority queue make up the batch. The probability that there are no marked job in this batch is $G_1(z_1, z_2)$.

Let the class-1 job of size x arrives to the system. Let $x \le \theta$. The probability that k_1 class-1 jobs arrive in the period $(0, x)$ is $P_1(x) = e^{-\lambda_1 x}(\lambda_1 x)^{k_1}/k_1!$. The probability that all the batches generated by these k_1 jobs is $G_1(z_1, z_2)^{k_1}$, because each of them generates a batch which does not have any marked jobs with probability $G_1(z_1, z_2)$. During time $(0, x)$ the probability that k_2 class-2 jobs arrive to the system is $P_2(x) = e^{-\lambda_2 x}(\lambda_2 x)^{k_2}/k_2!$. The probability that this jobs are not marked is not included in $G_1(z_1, z_2)$ and equals to $z_2^{k_2}$. Then we sum over k_1 and k_2 and we uncondition on x to get:

$$
\begin{aligned}
[G_1(z_1, z_2), S \le \theta] &= \int_0^\theta \left(\sum_{k_1=0}^{\infty} P_1(x) G_1(z_1, z_2)^{k_1} P_2(x) z_2^{k_1} \right) \mathrm{d}F_1(x) = \\
&= \int_0^\theta e^{-\lambda_1 x(1-G_1(z_1, Z_2)) - \lambda_2 x(1-z_2)} \mathrm{d}F_1(x).
\end{aligned}
$$

Let us consider now the case $x > \theta$. The class-1 job is first served in the high priority queue until it gets θ of service. Then it is moved to the low priority queue. The probability that k_1 class-1 jobs arrive during the service time $(0, \theta)$ is $P_1(\theta) = e^{-\lambda_1 \theta}(\lambda_1 \theta)^{k_1}/k_1!$. The probability that there are no marked jobs in all the batches generated by these k_1 jobs is $G_1(z)^{k_1}$. The probability that k_2 class-2 jobs arrive to the system during this interval is $P_2(\theta) = e^{-\lambda_2 \theta}(\lambda_2 \theta)^{k_2}/k_2!$, and the probability that none of these jobs is marked is z^{k_2}.

Finally we have to take into account the “starter” itself. The probability that the starter is not marked is z_1. Then summing over k_1 and k_2 and unconditioning on x we get

$$
\begin{aligned}
[G_1(z_1, z_2), S > \theta] &= \int_\theta^\infty \left(\sum_{k_1=0}^{\infty} P_1(\theta) G_1(z_1, z_2)^{k_1} z_1 P_2(\theta) z_2^{k_1} \right) \mathrm{d}F_1(x) = \\
&= z_1 e^{-\lambda_1 \theta(1-G_1(z_1, z_2)) - \lambda_2 \theta(1-z_2)} \overline{F}_1(\theta).
\end{aligned}
$$

We focus now on $G_2(z_1, z_2)$. When a class-2 job arrives to the system it generates a batch of size one, then the probability that jobs of this batch are not marked is simply z_2, and thus $G_2(z_1, z_2) = z_2$.

Combining all the expressions we get (16). □

We can now calculate $\mathbb{E}[N_1]$, $\mathbb{E}[N_2]$, and derive the expressions for ρ_b and W_b.

Lemma 3.

$$
\begin{aligned}
\rho_b &= 1 - \frac{1 - \rho_x^{(1)} - \rho_{g(x)}^{(2)}}{1 - \rho_\theta^{(1)}}, \\
W_b &= W_{x, g(x)} - W_{\theta, 0}(1 + \rho_b) - \theta \frac{\rho_x^{(1)} - \rho_\theta^{(1)}}{1 - \rho_\theta^{(1)}}.
\end{aligned}
$$

Proof: For $i = 1, 2$ we have

$$\mathbb{E}[N_i] = \frac{\partial G(z_1, z_2)}{\partial z_i}|_{1,1},$$
$$\mathbb{E}[N_i(N_i - 1)] = \mathbb{E}[N_i^2] - \mathbb{E}[N_i] = \frac{\partial^2 G(z_1, z_2)}{\partial z_i^2}|_{1,1},$$
$$\mathbb{E}[N_1 N_2] = \frac{\partial^2 G(z_1, z_2)}{\partial z_1 \partial z_2}|_{1,1}.$$

Using $b_i = \frac{\mathbb{E}[N_i^2]}{\mathbb{E}[N_i]} - 1$ and after some straightforward calculations we obtain the result. □

Now let us find expressions for $A_1(x)$ and $A_2(x)$.

Lemma 4. *The mean workload that comes with the tagged job of class-*1 *of size* x *in the batch and that has to be served before its service is completed equals:*

$$A_1(x) = 2(W_{\theta,0} + \theta)\rho_b - \theta \frac{\rho^{(2)}_{g(x)}}{1 - \rho^{(1)}_\theta}.$$

Proof: Since the tagged job arrives from class-1 only when the batch is started by a class-1 job, the calculations now will depend on $G_1(z_1, z_2)$. We denote $b_{1|1}$ and $b_{2|1}$ the mean number of jobs of class-1 and class-2 which arrive in the batch with the tagged job of class-1 when the batch is initiated by a class-1 job. Then

$$A_1(x) = b_{1|1}\mathbb{E}[X^{(1)}_{\theta,x}] + b_{2|1}\mathbb{E}[X^{(2)}_{g(x)}] - \mathbb{E}[X^{(1)}_{\theta,x}].$$

Here

$$b_{1|1} = \sum_{n_1} n_1 \frac{n_1 \mathbb{P}(n_1)}{\mathbb{E}[N_{1|1}]} = \frac{\mathbb{E}[N^2_{1|1}]}{\mathbb{E}[N_{1|1}]},$$

where $N_{1|1}$ is the random variable which corresponds to the number of jobs of class-1 in the batch when the batch is initiated by the class-1 job. So the number of class-1 jobs that arrive in addition to the tagged job is $\left(\frac{\mathbb{E}[N^2_{1|1}]}{\mathbb{E}[N_{1|1}]} - 1\right)$. Note that since we condition on the fact that the starter is a class-1 job, $N_{1|1}$ is now calculated from $G_1(z_1, z_2)$ so:

$$\mathbb{E}[N_{1|1}] = \frac{\partial G_1(z_1, z_2)}{\partial z_1}|_{1,1},$$
$$\mathbb{E}[N_{1|1}(N_{1|1} - 1)] = \frac{\partial^2 G_1(z_1, z_2)}{\partial z_1 \partial z_1}|_{1,1}.$$

Then we can find $(b_{1|1} - 1)$. Now we need to calculate $b_{2|1}$, that is, the mean number of class-2 jobs that the tagged job of class-1 job see. We have that from the Generating function $G_1(z_1,z_2)$ by conditioning on the number of class-1 jobs:

$$G_1(z_1,z_2) = \sum_{n1}\sum_{n2} z_1^{n_1} z_2^{n_2} p_{n_1,n_2} = \sum_{n1}\sum_{n2} z_1^{n_1} z_2^{n_2} p_{n_2|n_1} p_{n_1},$$

$$\frac{\partial^2 G_1(z_1,z_2)}{\partial z_1 \partial z_2}|_{1,1} = \mathbb{E}[N_1]\sum_{n1}\sum_{n2} n_2 p_{n_2|n_1}\frac{n_1 p_{n_1}}{\mathbb{E}[N_1]} = \mathbb{E}[N_1] b_{2|1}.$$

Then we can calculate $b_{2|1}$

$$b_{2|1} = \frac{1}{\mathbb{E}[N_{1|1}]}\frac{\partial^2 G_1(z_1,z_2)}{\partial z_1 \partial z_2}|_{(1,1)}.$$

Finally we find the expression for $A_1(x)$. □

Lemma 5. *The mean workload that arrives in the batch together with a class-*2 *job with service requirement* $g(x)$*, and that has to be served before its service is completed equals:*

$$A_2(g(x)) = 2(W_{\theta,0}+\theta)\rho_b - \theta\frac{\rho^{(2)}_{g(x)}}{1-\rho^{(1)}_\theta} - \theta\rho_b.$$

Proof: When the tagged job arrives from class-2 the batch can be started by a class-1 or by a class-2 job, so the calculations depend on $G(z_1,z_2)$. We denote $b_{1|2}$ and $b_{2|2}$ the mean number of jobs of class-1 and class-2 which arrive in the batch with the tagged job of class-2. Then

$$A_2(g(x)) = b_{1|2}\mathbb{E}[X^{(1)}_{\theta,x}] + b_{2|2}\mathbb{E}[X^{(2)}_{g(x)}] - \mathbb{E}[X^{(2)}_{g(x)}] =$$
$$= b_{1|2}\mathbb{E}[X^{(1)}_{\theta,x}] + (b_{2|2}-1)\mathbb{E}[X^{(2)}_{g(x)}].$$

As the tagged job is from class-2, then $b_{2|2} = b_2$. We need to find the value of $b_{1|2}$. We use the fact that jobs of class-1 and class-2 arrive independently from each other.

$$G(z_1,z_2) = \sum_{n1}\sum_{n2} z_1^{n_1} z_2^{n_2} p_{n_1,n_2} = \sum_{n1}\sum_{n2} z_1^{n_1} z_2^{n_2} p_{n_1|n_2} p_{n_2}$$

$$\frac{\partial^2 G(z_1,z_2)}{\partial z_1 \partial z_2}|_{1,1} = \mathbb{E}[N_2]\sum_{n1}\sum_{n2} n_1 p_{n_1|n_2}\frac{n_2 p_{n_2}}{\mathbb{E}[N_2]} = \mathbb{E}[N_2] b_{1|2}.$$

Then

$$b_{1|2} = \frac{1}{\mathbb{E}[N_2]}\frac{\partial^2 G(z_1,z_2)}{\partial z_1 \partial z_2}|_{1,1}.$$

From here we get the expression for $A_2(g(x))$. □

The result of Proposition 2 now follows by observing that after substitution of all the terms, (14) and (15) become (7) and (8), respectively.