Queueing with Heterogeneous Users: Block Probability and Sojourn times

Veeraruna Kavitha and Raman Kumar Sinha
IEOR, Indian Institute of Technology Bombay, India

***Abstract*—We consider a queueing system with heterogeneous agents. One class of agents demand immediate service, and would leave the system if not provided. The agents of the second class have longer job requirements and can wait for their turn. We discuss the achievable region of such a two class system, which is the set of all possible pairs of performance metrics. Blocking probability is the relevant performance metric for impatient class while the expected sojourn time is appropriate for the second tolerant class. We consider static (time invariant and state independent) and dynamic (state dependent) scheduling policies. In queueing systems with homogeneous agents, where expected sojourn time is the performance metric for both the classes, we show that the static and dynamic achievable regions coincide. However, this is not the case with heterogeneous setting. We obtain the achievable region, under static policies. We consider an example dynamic policy and show that the dynamic achievable region is strictly bigger than the static region.**

We conjecture a pseudo conservation law, in a fluid limit for impatient customers, which relates the blocking probability of eager customers with the expected sojourn time of the tolerant customers. We provide a partial proof and use the pseudo conservation law to obtain the static achievable region. We validate the pseudo conservation law using two example families of static schedulers, both of which achieve all the points on the achievable region. Along the way we obtain smooth control (sharing) of resources between voice and data calls.

Index terms– Heterogeneous agents, achievable region, processor sharing, resource sharing, dynamic and static scheduling.

I. Introduction

We consider queueing systems with heterogeneous classes. First one is a class of impatient agents. They reject the system if service is not offered almost immediately. One would require a parallel (upto K servers) service-offer facility to handle this class. Blocking probability, the probability that an agent returns without service, is important metric for this class. The agents of the second class are tolerant, can wait for their turn. However, their satisfaction depends upon the expected sojourn time (total time spent in the system).

An achievable region for an n-class system is the set of relevant performance vectors $(pm_1, \cdots, pm_n)$, obtained by varying all possible scheduling policies (e.g., [12], [15], [16]). In heterogeneous setting, the achievable region is the set of pairs of blocking probability and expected sojourn times.

The achievable region is well understood for homogeneous classes, when the performance metric of both the classes is expected sojourn time (e.g., conservation laws, pioneered by [9]). Coffman and Mitrani [11] were the first to identify such achievable regions. Multi-class single server queueing systems pose nice geometric structure (polytopes) for achievable region (e.g., [11], [12]). The scheduling policies are dynamic if they depend upon the (time varying) system state. Static policies are time invariant and state independent. Our first observation is that the homogeneous achievable region is the same under static as well as dynamic policies.

In this paper we are studying the achievable region of queueing systems with heterogeneous classes.

A parametrized family of scheduling policies is called *complete* by [20], if it achieves all possible performance vectors of the achievable region, average waiting times in their context. A complete scheduling class can be used to find the optimal control policy over all scheduling disciplines. Discriminatory processor sharing (DPS) class of parametrized dynamic priority schedulers is identified as a *complete* family in case of two class $M/G/1$ queue in [21]. Many more families of scheduling policies are identified to be complete.

Our contribution

We conjecture a relation between the expected sojourn time and the blocking probability, which would be valid for any static scheduling policy, and, call it a pseudo conservation law. This pseudo conservation law is valid in a fluid limit for short job impatient agents and we provide a partial proof. We then show that two sets of scheduling (PS and CD) policies satisfy this conservation law and also achieve all the points of the resulting achievable region. In both the policies, the admission of an impatient agent disrupts the service of the ongoing tolerant agent (if any). In the first case the entire system capacity is transferred to the impatient agent, while a fixed fraction of capacity is transferred in the second case. We refer the first system as PS policy, as here the impatient agents derive service in the well known processor sharing mode. At maximum K impatient agents can share the facility. In the second policy, the service of the tolerant agent is continued with the remaining capacity. Admission of another impatient agent results in the transfer of an additional fraction of capacity, while a departure results in the transfer back of the same fraction. The second policy is referred as capacity division-CD policy, because the capacity is divided between the two classes. The tolerant agents are always served in serial fashion, i.e., one at a time in both the policies.

We solve an appropriate set of balance equations to obtain the busy probability. We majorly use a domination technique to study the tolerant class. The actual system is sandwiched in between the two $M/G/1$ queues. The difference in the sojourn time performance between the two queues converges to zero as arrival-departure rates of the impatient agents converge to infinity, while maintaining their ratio fixed. Typically agents with long job requirements form tolerant class, while the impatient ones demand for short jobs (e.g., super markets, data-voice calls etc.). Thus they operate in the precise asymptotic regime, for which our results are accurate.

Applications

The problem of resource allocation, between data-voice calls of a communication network is a loosely related topic (e.g., [5], [6]). Voice calls are dropped if server is not available, while data calls can wait. The voice, data calls respectively form impatient and tolerant agents. In [5] the authors consider a three channel pool scheme to obtain a novel adjustable boundary based channel allocation scheme with pre-emptive priority for integrated data-voice networks. They attain various levels of priority by adjusting the division of total available channels among the three pools. In [6] authors again consider channel allocation scheme for packet level allocations. These papers discuss coarse sharing of resources between data-voice calls. While we provide a smooth control: by varying the admission parameter p (continuously) in the interval $[0, 1]$, one can achieve any pair of performance metrics for data-voice calls on the static achievable region.

Various applications (e.g., computers, communication networks, manufacturing systems) can be modelled as multi-class queueing systems and dynamic control of such systems is an important aspect. One of the main techniques for such problems is to characterize the achievable region, then use optimization methods to obtain optimal control policy (for example [17], [18], [19]). Once the heterogeneous achievable region is known, many relevant optimization problems can be solved in a similar way. If further a good complete family is identified, the optimization problem gets simplified greatly. For example, our CD/PS policies are parametrized by K, the number of parallel service facilities and p the probability with which an impatient job is admitted. Thus the problem of finding the optimal expected sojourn time of data calls in a communication network, given a constraint on blocking probability of voice calls can readily be obtained, by optimally choosing K, p.

In a super market the long jobs wait, while short jobs are provided fast service via dedicated express counters. Alternatively, if one use the same counters to serve both the jobs and if selected (controlled) short jobs pre-empt the long jobs, one could obtain optimal design using our static achievable region. For example, one can obtain the optimal fraction of short jobs lost, given a constraint on the expected sojourn time of long jobs.

Consider a communication system with K orthogonal channels. Initially all the channels are dedicated to data calls. As and when the voice calls arrive, one by one the channels are transferred and data calls use the remaining. Our CD policy captures this scenario precisely. If the voice calls are served at the highest possible rate as with PS policy, it improves the chances of a free server being available to subsequent voice arrivals. The two achievable regions overlap, but PS has a bigger region (when number of maximum parallel calls is fixed) as it attains a smaller blocking probability.

Literature survey, related to Queueing systems

Our paper mainly deals with resource sharing between customers who are willing to wait for the service and the customers who are not willing to wait much. To the best of our knowledge we are not aware of a work that directly studies this type of a heterogeneous achievable region. Some variants of queueing systems (e.g., [23], [13], [14], [7], [4], [10], [2] etc) have some connections to few parts of our models and we brief such papers here.

In [13] authors consider multi-class queueing system with eager and tolerant customers. This is the work that is closest to our work, especially to CD policy. The tolerant customers utilize all the remaining servers (hence are work conserving) in our model, while [13] considers tolerant customers also in multi-server mode. If there is only one tolerant customer and if say $n > 1$ free servers are available, the tolerant customer uses only one server in [13], while with our CD policy the tolerant user is served using the consolidated capacity of all the n servers. The authors in [13] obtain a set of (balance) equations, solving which stationary probabilities (and then the stationary performance) can be derived. We provide a closed form expression for these performance measures in fluid limit for eager customers. We further discuss a 'policy-independent' pseudo conservation law.

In [2] authors analyzed an $M/G/1$ queue with Poisson interruptions, caused either by server breakdown or by the arrival of higher priority agents. This system is similar to our PS/CD policy with single server.

In [23], the authors consider time limited autonomous polling system. Here the server visits a finite number of queues in a periodic manner, while spending a random (exponentially distributed) visit time at each queue, independent of the status of the queues. It spends the designated exponential time, even if the queue gets/is empty. The tolerant agents of our PS model can be studied using their results. However their expressions are complicated, while we derive simple expressions in an appropriate asymptotic limit. Further, similar techniques are used to study the CD model.

In [14] Sleptchenko et. al considered a multi-class $M/M/k$ queueing system with two priority groups, both groups of customers are tolerant and each priority group may consist of multiple customer types. Different customers type having their own arrival and service rates. Upon arrival a high priority customer pre-empts the lower one, if all the servers are busy and some are serving low priority customers. The main contribution is that they obtain the stationary state probabilities, which provides many more performance measure (e.g., moments of type wise waiting customers, mean number of low-priority customers interrupted etc.) beyond the stationary mean values. As already mentioned, this paper only considers tolerant customers.

In [7], authors consider multi-class single-server queue with K classes of customers. Arrival of each class follows independent Poisson process and the service time is generally distributed. The service capacity is shared simultaneously among all customers present in proportion to the respective class-dependent weights. They derived closed form approximation for the mean conditional and unconditional sojourn time of each classes and verified it in light traffic and high traffic modes. In [4], authors considered $M/M/K$ queue with $m > 2$ number of priority classes. They reduce the m-dimensional Markov chain to 1-dimension Markov chain using the busy period of high priority customers. They modelled the one dimensional Markov chain as QBD (Quasi Birth Death) processes and calculate the mean sojourn time for low priorities customers. Both these papers only deal with tolerant customers.

In section II we describe the problem. We conjecture the complete static region and provide a partial proof in section III. PS and CD models are respectively analyzed in sections IV and V and are compared in section V. Using an example dynamic policy for PS model we showed in section VII that the heterogeneous dynamic achievable region is strictly bigger than the static region.

II. Problem Statement and System model

We refer the impatient/**e**ager customers by ϵ-customers while the **t**olerant customers are referred to as τ-customers. The system has a fixed server capacity, that needs to be shared between the two classes of customers. The exact sharing of capacity depends upon the allocation/scheduling policy. For example the system can serve K customers in parallel for some K, by dividing the server capacity among the customers under service. The system can chose to vary K dynamically, e.g., processor sharing. The system can chose to serve one customer with full capacity etc. In this paper we discuss two example sets of scheduling policies. We only consider 'τ-work conserving' policies, wherein the τ-customers utilize all the remaining server capacity.

A. Arrival process and the Jobs

The arrival processes are modelled by independent Poisson processes, with rates λ_ϵ and λ_τ respectively. The job requirements for both the classes are exponentially distributed. The time required to complete a job, depends upon the scheduling policy. If a τ-customer (ϵ-customer) is served with full server capacity, then the service time is exponentially distributed with parameter μ_τ (respectively μ_ϵ).

B. Achievable region

The two classes of users have different goals and hence naturally require different qualities of service (QoS). An eager ϵ-agent would leave the facility without service, if service is not provided almost immediately. Block probability P_B, the probability of such an event is an important performance metric for ϵ-class. The service of a typical τ-customer can possibly be interrupted, possibly to provide required QoS for ϵ-agents, and a typical τ-agent can face several such interruptions during its service. Thus the expected sojourn time $E[S_\tau]$, the expected value of the total time spent by a typical agent would be an appropriate QoS for τ-class. Either of these performance metrics depend upon the scheduler β used.

A scheduling decision is required at every ϵ-arrival instance[1], because the ϵ-agents are impatient. Because of this the service of a τ-agent can be pre-empted several times. Thus *the appropriate QoS for τ-agents is the expected sojourn time, and not the expected waiting time.* Therefore the achievable region is given by:

$$\mathcal{A}_{hetero} = \{(P_B(\beta), E[S_\tau(\beta)]) : \beta \text{ is a scheduler}\}.$$

In this paper we consider (τ) static policies, wherein the ϵ-admission rules do not depend upon the status of the τ-class. The probability of admission, p, is an important parameter of any such scheduling policy. Further, the (maximum) number of ϵ-calls served in parallel and the sharing of resources between ϵ and τ customers is also a part of the scheduling decision. For example, the system may allocate/transfer entire server capacity to the first admitted ϵ-arrival. It may processor-share the capacity among the further admitted ϵ-customers. There may be a limit on the number of ϵ-customers that can simultaneously share the capacity. Alternatively the system may allocate a fixed fraction of the server capacity to each admitted ϵ-arrival and the remaining is allocated to τ-class etc. All these rules are independent of the τ-state (e.g., the number of τ customers in the system, waiting time of them etc). This implies that ϵ-calls pre-empt τ-call when required. In all a τ-static policy implies that an ϵ-arrival is admitted with some probability p, and further admission also depends upon the number of ϵ-calls already in the system, but not on τ-state. Mathematically a static achievable region is defined:

$$\mathcal{A}^{static}_{hetero} = \Big\{ \left(P_B(\beta_p^{CS}), E[S_\tau(\beta_p^{CS})]\right) : \\ 0 \le p \le 1 \text{ and } (CS) \text{ a capacity sharing rule } \Big\}. \quad (1)$$

We primarily analyze the static achievable region. Towards the end, in section VII an example dynamic policy is considered to show that the achievable region with dynamic policies is strictly bigger than the static achievable region.

C. Short-Frequent Job (SFJ) limits

The ϵ-class has short job requirements. If one considers limit $\mu_\epsilon \to \infty$, the impact of ϵ-customers becomes negligible at the limit. To obtain a more general and useful result, we also increase the ϵ-arrival rate while $\mu_\epsilon \to \infty$. That is, every ϵ-agent may utilize the server for a short duration, but the system has to attend the ϵ-agents frequently. Because of this ϵ-agents cause significant impact even in the limit. To be more precise we consider the limits $\mu_\epsilon \to \infty$ and $\lambda_\epsilon \to \infty$ while the load factor $\rho_\epsilon = \lambda_\epsilon / \mu_\epsilon$ is maintained constant. We refer this as "Short-Frequent Job (SFJ) limits".

III. Entire Static Region: Pseudo Conservation law

In a (homogeneous) multi-class queueing system, with all tolerant classes, a work conservation law holds. The total

[1] On the contrary, in homogeneous setting two or more classes of agents wait at their waiting lines and scheduling epochs are the service completion/departure epochs. The scheduler had to decide which class to be served next. While in heterogeneous setting, at any departure epoch there is only one class of agents possibly waiting and hence no decision is required.

workload in the system remains the same irrespective of the scheduling policy, as long as the server does not idle during busy period. Further by Little's law and Wald's lemma, a linear combination of expected sojourn (or waiting) time of different classes of customers remains the same irrespective of the scheduling policy (e.g., [12]).

The above is obviously true when the incoming workload remains the same. However in our heterogeneous setting, the ϵ-customers depart the system, if service is not offered almost immediately. And this depends upon the scheduling policy. Thus the workload arriving into the system itself changes with different scheduling policies and naturally one may not expect work conservation. However if the amount of work blocked remains the same, one can anticipate a different kind of work conservation. We conjecture that given a probability of blocking, irrespective of the way the ϵ-agents are blocked and irrespective of the way the τ-agents are served, the τ-expected sojourn time remains the same[2]. And this could be conjectured only in SFJ limit and when the policies do not depend upon the τ-state.

In SFJ limit, ϵ-agents will have fluid arrivals and departures. Given the ϵ-load factor (ρ_ϵ) and the probability of blocking (p_B), in the SFJ limit, the ϵ-agents occupy $\rho_\epsilon(1-p_B)$ fraction of system resources at all the times. Hence we *conjecture* that the τ- performance equals that of an $M/M/1$ queue with smaller service rate $\mu_\tau(1-\rho_\epsilon(1-p_B))$, and that the expected sojourn time for any $0 \le p_B \le 1$ equals:

$$\boxed{E_{S_\tau}(p_B) := \frac{1}{\mu_\tau(1-\rho_\epsilon(1-p_B))-\lambda_\tau} \text{ if } \rho_\epsilon(1-p_B)+\rho_\tau < 1. \quad (2)}$$

Conjecture: Static achievable region, in SFJ limit, equals:

$$\boxed{\mathcal{A}^{hetero}_{static} = \Big\{ \big(p_B, E_{S_\tau}(p_B)\big) \Big| \, p_B \in [0,1], \, \rho_\epsilon(1-p_B)+\rho_\tau < 1 \Big\}.}$$

We would like to refer the equation (2) as a *pseudo conservation law,* as it provides the expected sojourn time in terms of the fraction blocked (lost). This would require an explicit proof which is provided immediately.

A. *Proof of Pseudo Conservation Law*

We will prove Pseudo conservation law (Theorem 1 given below) under some of the following assumptions:

A.1) The schedulers are τ-static (do not depend upon τ-state).

A.2) The scheduling policies depend only upon the number of ϵ-customers already in the system.

A.3) The τ-customers are served in serial fashion. The scheduler is work-conserving for τ-customers, i.e., it serves the τ-customers with all the left over server capacity, if there are any.

We begin with some discussions, which form a part of the proof.

We obtain different ϵ-systems with different μ_ϵ using the following special construction (without loss of generality). First we consider a sequence of ϵ inter-arrival, job-requirement sequences $\{A^1_{\epsilon,n}, B^1_{\epsilon,n}\}_n$ for $\mu_\epsilon = 1$. Note that $E[A^1_{\epsilon,n}] = 1/\rho_\epsilon$ and $E[B^1_{\epsilon,n}] = 1$ for all n. Then obtain the sequence $\{A^{\mu_\epsilon}_{\epsilon,n}, B^{\mu_\epsilon}_{\epsilon,n}\}_n$ for any general μ_ϵ, by multiplying the corresponding arrival-departure sequences by $1/\mu_\epsilon$, i.e.,

$$A^{\mu_\epsilon}_{\epsilon,n} = \frac{A^1_{\epsilon,n}}{\mu_\epsilon} \text{ and } B^{\mu_\epsilon}_{\epsilon,n} = \frac{B^1_{\epsilon,n}}{\mu_\epsilon} \text{ for every } n, \mu_\epsilon. \qquad (3)$$

Since the decisions are independent of τ-customers (by **A**.1) one can identify a renewal process, corresponding only to the ϵ-system, alternating between ϵ-busy periods and ϵ-idle periods, both of which depend only upon ϵ-customers. Let X_a, X_{tot} respectively represent the number of ϵ-customers that received service and the number that arrived in one renewal cycle. By renewal reward theorem (RRT) applied twice we obtain the blocking probability:

$$(1 - P_B) = \frac{E[X_a]}{E[X_{tot}]}. \qquad (4)$$

By **A**.1, we are considering τ-static scheduling policies here, wherein the decision do not depend upon the τ-state of the system. By **A**.2 the policies depend only upon the number of ϵ-customers in the system. Thus, if $X^\mu(t)$ represents the number of ϵ-customers in the system at time t with $\mu_\epsilon = \mu$, then

$$X^\mu(t) = X^1(\mu t) \text{ for any } \mu. \qquad (5)$$

Hence with **A**.1-2, P_B remains the same for all μ_ϵ (see (4)).

The length of a typical renewal cycle equals the arrival instance of X_{tot}-th customer, which by the special construction equals:

$$\sum_{n=1}^{X_{tot}} A^{\mu_\epsilon}_{\epsilon,n} = \frac{1}{\mu_\epsilon} \sum_{n=1}^{X_{tot}} A^1_{\epsilon,n},$$

By Wald's Lemma the expected value of the length of a renewal cycle (with μ_ϵ) equals $E[X_{tot}]/\lambda_\epsilon$. The total amount of time for which ϵ-customers utilize the server during one typical renewal cycle, irrespective of the way the service is offered and the way the customers are admitted, stochastically equals

$$\sum_{n=1}^{X_a} B^{\mu_\epsilon}_{\epsilon,n},$$

whose expected value clearly equals $E[X_a]/\mu_\epsilon$. Thus the long run fraction of server time available for τ customers by RRT equals:

$$\begin{aligned} \nu_\tau(\mu_\epsilon) = \nu_\tau(P_B) &= \frac{E[X_{tot}]/\lambda_\epsilon - E[X_a]/\mu_\epsilon}{E[X_{tot}]/\lambda_\epsilon} \\ &= 1 - \rho_\epsilon \frac{E[X_a]}{E[X_{tot}]} = 1 - \rho_\epsilon(1 - P_B). \end{aligned}$$

This is true for any μ_ϵ and for any static scheduler. Hence, we have the following:

Theorem 1: a) Assume **A**.1. For any μ_ϵ and for any static scheduler, the long run fraction of server capacity available to τ-customers depends only upon ρ_ϵ and the probability of blocking P_B of ϵ customers:

$$\nu_\tau(\mu_\epsilon) = \nu_\tau(P_B(\mu_\epsilon)) = 1 - \rho_\epsilon(1 - P_B).$$

[2] With one class of customers, the expected sojourn time by Little's law and Wald's lemma is proportional to the workload in the system

b) Assume **A**.1-2. Then P_B and hence ν_τ remains the same for all μ_ϵ; and
c) When $\mu_\epsilon \to \infty$ the accumulated amount of server capacity available to τ-customers at time t, $R^{\mu_\epsilon}(t)$, converges to a constant curve with value $\mu_\tau \nu_\tau$. This convergence is uniform over any bounded time intervals and almost surely:

$$\sup_{t\in[0,W]} \left|R^{\mu_\epsilon}(t) - \nu_\tau t\right| \to 0 \text{ almost surely, as } \mu_\epsilon \to \infty.$$

for any $0 < W < \infty$.
d) Under **A**.1-3 when $\mu_\epsilon \to \infty$, τ-sojourn time converges to that of a M/M/1 queue with respective parameters λ_τ and $\nu_\tau \mu_\tau = (1 - \rho_\epsilon(1 - P_B))\mu_\tau$:

$$\lim_{\mu_\epsilon \to \infty, \ \ \lambda_\epsilon/\mu_\epsilon = \rho_\epsilon} E[S_\tau] = \frac{1}{\nu_\tau \mu_\tau - \lambda_\tau} \text{ if } \nu_\tau \mu_\tau > \lambda_\tau.$$

Proof: Parts (a) and (b) are already proved. Parts (c)-(d) will be proved below.

We again consider the special construction given by (3) and (5) is true by **A**.1-2. Let $R^\mu(t)$ represent the total amount of residual server time available for τ-customers till time t. By RRT, as depicted by part (a)

$$\frac{R^\mu(t)}{t} \to \nu_\tau(\mu) \text{ a.s. as } t \to \infty \text{ for any } \mu.$$

By part (b), $\nu_\tau(\mu) = \nu_\tau$ is the same for all μ. Now, clearly

$$\frac{R^\mu(t)}{t} = \frac{R^1(\mu t)}{\mu t} \text{ which implies } R^\mu(t) = \frac{R^1(\mu t)}{\mu}.$$

By Theorem 4 of Appendix A, a function version of RRT, we have:

$$\sup_{t\in[0,W]} \left|\frac{R^1(\mu t)}{\mu} - \nu_\tau t\right| \to 0 \text{ almost surely, as } \mu \to \infty.$$

for any $0 < W < \infty$. Hence we proved part (c), i.e.:

$$\sup_{t\in[0,W]} \left|R^\mu(t) - \nu_\tau t\right| \to 0 \text{ almost surely, as } \mu \to \infty, \qquad (6)$$

for any $0 < W < \infty$. Let the time to complete service of a typical τ-customer considering all possible ϵ-interruptions be defined using $R^\mu(t)$ as below:

$$T^\mu = \inf_t \left\{t : R^\mu(t) \geq B_\tau\right\},$$

where B_τ is the job requirement of the τ-customer. By Lemma 3 of Appendix A, using (6), we have that

$$\left|T^\mu - \frac{B_\tau}{\nu_\tau}\right| \text{ almost surely as } \mu \to \infty.$$

Thus we have an exponential random variable in the limit. Hence and further using dominating systems as in CD/PS policies (see sub-section IV-B2 which dominates sojourn times for FCFS service) *we can show that the expected stationary τ-workload with any work conserving policy converges towards that of an MM1 queue with service rate $\mu_\tau \nu_\tau$, as $\mu_\epsilon \to \infty$, under* **A**.*3.* We can first dominate the workload at any time, then time average of the workload till any time, which implies the sand-witching and further converge of the limit of the time average workload and hence that of expected workload.

In exactly similar lines one can sand-witch the number of customers at any time of the original system between that of two dominating systems in almost sure sense, if consider a small modification of the sample paths of the three systems. We assume that the job requirements of n-th departing customers (as opposed to the n-th arriving customer which we considered for workload analysis) for any n is exactly the same in all the three systems. This would provide the domination of corresponding 'effective server times' (T^μ). Using this one can obtain the convergence of expected number in the system towards that of the MM1 queue with job demands as B/ν_τ, when one serves the limiting queue with the same service discipline as used with the original system. One can use this technique as long as the service discipline is non anticipating, i.e., the next customer to be served, does not depend upon the job requirements.

This implies the convergence of the expected sojourn time $E[S_\tau]$ by Little's law. □

Remarks: We are working towards establishing this law under more general conditions. As of now we have the following results by virtue of the above theorem:

1) We have established Pseudo conservation law (2) for subclass of τ-static schedulers which a) serve the τ-class customers serially and when the policies are non anticipative; b) the admission rules for ϵ-class depends upon at maximum upon the number of ϵ-customers in the system;
2) In exactly similar lines, we can extend this result if the admission rules depend upon the time spent by a typical agent, etc, under further assumption that this rule scales appropriately with μ_ϵ. All we need is that equation (5) is true;
3) Further it is clear from the proof of the above theorem, that the expected workload is a deterministic function of blocking probability P_B for all τ-work conserving policies. This establishes a Pseudo conservation law in terms of blocking probability and expected workload, for more general conditions;
4) One can easily verify that the result is true for M/G queues at τ customers and G/G queues for ϵ-customers.
5) With G/G queues for both the classes of the customers, we have that the expected workload of the τ-customer converges towards that of the limit of the the expected workload of the G/G queue with job demands distributed as T^μ. This is true for any work-conserving, non-anticipative policy. We still need to prove that this limit equals the expected workload of the correspond G/G queue with job demands as B/ν_τ, where B is the original τ-job demand.

We now consider two example families of schedulers and illustrate the validity of our pseudo conservation law. Further, using the same sets of schedulers, we achieve all the points of the static region. Such a family is generally *referred to as complete family of schedulers.*

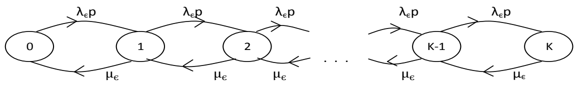


Fig. 1. State transitions for ϵ-agents with $\beta^{PS}_{p,K}$ scheduler.

IV. Processor sharing $PS-(p,K)$ schedulers

Any ϵ-arrival is admitted to the system with probability p, independent of τ-state. Once admitted it will pre-empt the existing τ-agent, if any. We consider K-processor sharing service discipline for ϵ-agents. If there is only one agent of the ϵ-class receiving service, it is served with maximum capacity, i.e., using capacity μ_ϵ. Upon a new (admitted) arrival of the same class, the capacity is shared among the two. Both are served in parallel and independently, each with rate $\mu_\epsilon/2$. Upon a third (admitted) arrival each is served with rate $\mu_\epsilon/3$. This continues up to K ϵ-agents. Any further arrival, leaves without service even after being admitted. When any of the existing ϵ-agents depart, the service rate is readjusted to an appropriate higher value. The τ-service is resumed only after all the ϵ-agents depart. We call this as $\beta^{PS}_{p,K}$ scheduling policy.

Tolerant agents are served in FCFS (first come first serve) basis. They are served in a serial fashion and with full capacity[3] μ_τ. That is, system would serve at maximum one τ-agent, and the service of the next τ-agent begins only after the preceding one departs.

The transitions and evolution of the ϵ-agents is independent of that of τ-agents under a static policy: the arrivals are admitted and the service is provided to the admitted agents immediately, irrespective of the state of τ-agents. Thus one can analyze the ϵ-class independently and we first consider this analysis.

A. Blocking Probability of ϵ-class

Fix $0 \le p \le 1$, K and consider policy $\beta^{PS}_{p,K}$. Blocking probability is the probability with which a new (ϵ-class) arrival leaves the system without service. Blocking can occur in case of two events. Upon arrival, an ϵ-agent is admitted to the system with probability p and is blocked with probability $(1-p)$. Secondly, an admitted agent leaves without service, if the system is already serving K ϵ-agents.

Let $\Phi_\epsilon(t)$ represent the number of ϵ-agents in the system at time t. We claim that the ϵ-class transitions are caused by exponential random events and hence that $\Phi_\epsilon(t)$ is a continuous time Markov jump process (see for example [8]) for the following reasons: a) it is clear that the inter-arrival times are exponentially distributed with parameter $\lambda_\epsilon p$; b) by Lemma 1, given below, the departure times are exponentially distributed with parameter μ_ϵ (i.e., $\sim \exp(\mu_\epsilon)$), irrespective of state $\Phi_\epsilon(t)$.

Lemma 1: Let D^l_ϵ represent the time to first departure among the l ϵ-agents receiving the service, with $1 \le l \le K$. Then for PS policy, $D^l_\epsilon \sim \exp(\mu_\epsilon)$ for any l.

Proof: When l agents are receiving service in parallel, because of processor sharing the service time of each is exponentially distributed with parameter μ_ϵ/l. And the time to first departure, the minimum of these l exponential random variables, is again exponential with parameter $l\mu_\epsilon/l = \mu_\epsilon$. ■

In Figure 1, we depict the transitions of the continuous time Markov jump process $\Phi_\epsilon(t)$. For such processes, well known balance equations are solved to obtain the stationary probabilities (see for example [8]). The stationary probabilities, $\{\pi_0, \pi_1, \cdots, \pi_K\}$, of $\Phi_\epsilon(t)$ are obtained by solving:

$$\pi_0\lambda_\epsilon p = \mu_\epsilon\pi_1, \ \pi_l(\lambda_\epsilon p + \mu_\epsilon) = \lambda_\epsilon p\pi_{l-1} + \mu_\epsilon\pi_{l+1} \text{ for } 1 \le l < K,$$
$$\text{and} \qquad \pi_K\mu_\epsilon = \lambda_\epsilon p\pi_{K-1}.$$

The solution or the stationary probabilities are $(0 \le l \le K)$:

$$\pi_l = \frac{\rho^l_{\epsilon,p}}{a_0} \ \text{ with } \ a_0 := \sum_{j=0}^{K} \rho^j_{\epsilon,p} \text{ and } \rho_{\epsilon,p} := \frac{\lambda_\epsilon p}{\mu_\epsilon} = \rho_\epsilon p.$$

An admitted agent gets blocked, if it finds K ϵ-agents in the system, and, this by PASTA (Poisson Arrivals See Time Averages) equals the stationary probability π_K of K ϵ-agents in the system. The agents are not admitted with probability $(1-p)$ and those admitted are blocked with probability π_K. Therefore the overall blocking probability equals:

$$\boxed{P^{PS}_B(p) = (1-p) + p\pi_K = (1-p) + p\frac{\rho^K_{\epsilon,p}}{a_0}.} \tag{7}$$

B. Expected sojourn time of τ-class

The ϵ-class requires short but frequent jobs (e.g., voice calls). Hence we are looking for a good relevant approximation that facilitates the analysis, and which further allows us to study other important variants (like CD policy of section V). Towards this, we approximately (accurate asymptotically) decouple the evolution of τ-agents from that of ϵ-agents.

We first understand the effective server time (EST), Υ_τ, which is defined as the total time period between the service start and the service end of a typical τ-agent. We refer this as EST of the agent under consideration, as no other τ-agent has access to server during this period. Sojourn time of a typical τ-agent equals the sum of two terms: a) waiting time, the time before the service start; and b) EST Υ_τ, the time after the service start.

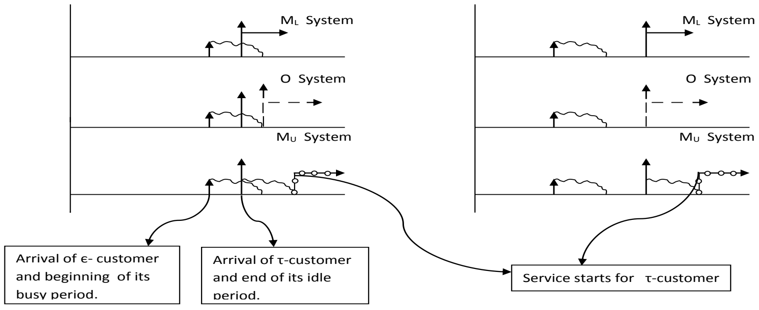
1) Analysis of effective server time (EST) (Υ_τ)*:* This time equals the sum of the actual service time, B_τ, of the τ-agent and the overall time of interruptions caused by ϵ-agents, which is denoted by Υ^e_τ. Let $N(B_\tau)$ represent the total number of the ϵ-class interruptions, that occurred during the service time B_τ. In reality these interruptions would have occurred in disjoint time intervals, the sum of all of which is B_τ. This random number has same stochastic nature as the number of Poisson arrivals that would have occurred in a continuous time interval of length B_τ. This is true because of the memory less property of the exponential service time B_τ and because Poisson process is a counting process. After an ϵ-agent interrupts the ongoing τ-agent, there is a possibility of further admissions. Eventually the service of the τ-class is resumed, where left, when all the ϵ-agents (that were admitted) leave the system.

Thus the time duration for which the service of τ-agent is suspended per interruption, equals a busy period of the ϵ-class, that started with one ϵ-agent. There would be $N(B_\tau)$ (random) number of such interruptions. Hence,

$$\Upsilon_\tau = B_\tau + \Upsilon^e_\tau \text{ with } \Upsilon^e_\tau := \sum_{i=1}^{N(B_\tau)} \Psi_{\epsilon,i}\,, \tag{8}$$

where $\{\Psi_{\epsilon,i}\}_i$ are the IID (independent and identically distributed) copies of ϵ-busy period. We have the following result.

[3] Capacity of the server is such that, it can either serve one tolerant agent at rate μ_τ, or l ϵ-agents each at μ_ϵ/l (where $l \le K$).


Mʟ System
O System
Mᴜ System
Mʟ System
O System
Mᴜ System
Arrival of ϵ- customer and beginning of its busy period.
Arrival of τ-customer and end of its idle period.
Service starts for τ-customer

Case(a): ϵ- customer present, at τ-arrival

Case(b): No ϵ- customer present at τ-arrival

Fig. 2. Example sample paths of the three systems (PS policy)

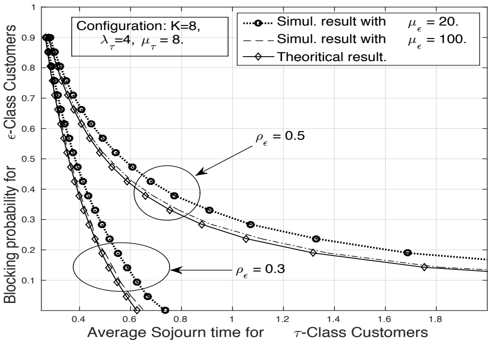


Fig. 3. Achievable region: Simulated and Theoretical results

Lemma 2: The first two moments of the ϵ-busy period and EST Υ_τ are given by:

$$E[\Psi_\epsilon] = \frac{a_1}{\mu_\epsilon} \text{ and } E[\Psi_\epsilon^2] = \frac{1}{\mu_\epsilon^2} \sum_{i=1}^{K} \frac{q^{i-1}\, c_i}{(1-q)^i}, \tag{9}$$

$$E[\Upsilon_\tau] = \frac{1}{\mu_\tau} + \frac{\lambda_\epsilon p}{\mu_\tau} E[\Psi_\epsilon] \ = \ \frac{a_0}{\mu_\tau},$$

$$E[\Upsilon_\tau^2] = \frac{2a_0^2}{\mu_\tau^2} + \frac{\rho_{\epsilon,p}}{\mu_\tau \mu_\epsilon} \sum_{i=1}^{K} \frac{q^{i-1}\, c_i}{(1-q)^i},$$

where the constants q, $\{c_i\}$ and $\{a_i\}$ are defined as:

$$\rho_{\epsilon,p} = \frac{\lambda_\epsilon p}{\mu_\epsilon},\ q = \frac{\rho_{\epsilon,p}}{\rho_{\epsilon,p}+1},\ a_i = \sum_{j=0}^{K-i} \rho_{\epsilon,p}^j \text{ for all } 0 \le i \le K, \tag{10}$$

$$b_i = \sum_{j=K-i+1}^{K-1} (K-j)\rho_{\epsilon,p}^j \text{ for all } 2 \le i \le K, \quad b_1 = 0,$$

$$c_1 = \frac{2\rho_{\epsilon,p}\,(2a_2 + b_2) + 2}{(1+\rho_{\epsilon,p})^2 \mu_\epsilon^2}, \text{ and for all } 1 \le i < K$$

$$c_i = \frac{2\rho_{\epsilon,p}\,((i+1)a_{i+1} + b_{i+1}) + 2\,((i-1)a_{i-1} + b_{i-1}) + 2}{(1+\rho_{\epsilon,p})^2 \mu_\epsilon^2},$$

$$c_K = \frac{2\rho_{\epsilon,p}(Ka_K + b_K) + 2((K-1)a_{K-1} + b_{K-1}) + 2}{(1+\rho_{\epsilon,p})^2 \mu_\epsilon^2}.$$

Proof: The proof is provided in Appendix B. ■

2) Approximate decoupling via Domination: Every τ-agent undergoes similar stochastic behaviour, as below. Each τ-agent will have to wait for the beginning of its service, and has to finish its service in the midst of random interruptions, all of which have identical stochastic nature. Further, evolution of the ϵ-agents during the EST Υ_τ of one τ-agent is independent of that of the other τ-agents. Hence the Υ_τ times corresponding to different τ-agents are independent of each other. Thus the idea is to model the τ-class evolution approximately as an independent process, with that of an $M/G/1$ queue. The arrivals remain the same, but the service times in $M/G/1$ queue are replaced by the sequence of ESTs $\{\Upsilon_\tau^t\}$.

We call this $M/G/1$ queue as $\mathcal{M}_L$ system and the original system as $\mathcal{O}$ system. In fact we will define another $M/G/1$ system $\mathcal{M}_U$ as below and show that: a) the performance (expected sojourn times) of the original system is bounded between the performances of the two $M/G/1$ systems; and b) that the performances of the two sandwiching systems converge towards each other as $\mu_\epsilon \to \infty$ (even with ρ_ϵ fixed).

a) $\mathcal{M}_L$ system: The ESTs are considered as service times of τ-agent in $\mathcal{M}_L$ system. We study the (sample path wise) time evolution of the two systems, original and $\mathcal{M}_L$, to demonstrate the required domination. Towards this, we assume that both the systems are driven by same input (arrival times and service requirements) processes. Consider that both the systems start with same number (greater than 0) of τ-agents and assume that both of them start with service of the first among the waiting ones. Then the trajectories of both the systems evolve in exactly the same manner, until the τ-queue gets empty. There can be a change in the trajectories of the two systems, upon a subsequent new τ-arrival. We can have two scenarios as in Figure 2. If ϵ-agents are absent at the τ-arrival instance in the original $\mathcal{O}$ system (as in sub-figure b), then again, both the systems continue to evolve in the same manner. On the other hand, if an ϵ-agent is deriving service (as in sub-Figure a), the service of τ agent is delayed in the original $\mathcal{O}$ system till the end of the ongoing ϵ-busy period. While the service starts immediately in $\mathcal{M}_L$ system. Then the trajectories in the two systems continue with the same difference, until the end of the next τ-idle period. At this point the difference: a) either gets reduced, if the τ arrival marking the end of τ-idle period occurs after sufficient time and finds no ϵ-agent; b) or can increase, if the τ-arrival occurs again during an ϵ-busy period; c) or can continue with almost previous value, if the τ-arrival occurs immediately and finds no ϵ-agent. And this continues. Thus the sojourn times in $\mathcal{M}_L$ system are lower than or equal to that in $\mathcal{O}$ system in all sample paths. As we notice the difference between the two systems is because of ϵ-busy cycles and this difference may diminish if the later shorten. We will show this indeed is true in coming sections.

b) $\mathcal{M}_U$ system: Consider another $M/G/1$ system whose service times equal $\Upsilon_\tau + \Psi_\epsilon$, where Ψ_ϵ is an additional ϵ-busy period independent of Υ_τ. It is clear that this system dominates the $\mathcal{O}$ system everywhere (see $\mathcal{O}$ and $\mathcal{M}_U$ trajectories in Figure 2). Hence the sojourn times of τ-agent in $\mathcal{O}$ system are upper bounded by that in $\mathcal{M}_U$ system (in all sample paths). Thus the expected sojourn time of $\mathcal{O}$ system is sandwiched as below:

$$E^{\mathcal{M}_L}[S_\tau] \le E^{\mathcal{O}}[S_\tau] \le E^{\mathcal{M}_U}[S_\tau]. \tag{11}$$

3) Performance of $\mathcal{M}_L$ and $\mathcal{M}_U$ systems: In Lemma 2, we obtained the first two moments of the ϵ-busy period and the EST, Υ_τ. Using the well known formula for the expected sojourn time of an $M/G/1$ queue, we have:

$$E^{\mathcal{M}_L}[S_\tau] = E[\Upsilon_\tau] + \frac{\lambda_\tau E[\Upsilon_\tau^2]}{2(1-\rho_\tau^{\mathcal{M}_L})} \text{ with } \rho_\tau^{\mathcal{M}_L} = \lambda_\tau E[\Upsilon_\tau].$$

Similarly with $\rho_\tau^{\mathcal{M}_U} = \lambda_\tau E[\Upsilon_\tau + \Psi_\epsilon]$,

$$E^{\mathcal{M}_U}[S_\tau] = E\big[\Upsilon_\tau + \Psi_\epsilon\big] + \frac{\lambda_\tau \left(E[\Upsilon_\tau^2] + E[\Psi_\epsilon^2] + 2E[\Psi_\epsilon]E\Upsilon_\tau]\right)}{2(1-\rho_\tau^{\mathcal{M}_U})}.$$

From Lemma 2 constants $\{c_i\}$, moments of busy period $E[\Psi_\epsilon]$, $E[\Psi_\epsilon^2]$ converge to zero as $\mu_\epsilon \to \infty$, and so the difference $E^{\mathcal{M}_U}[S_\tau] - E^{\mathcal{M}_L}[S_\tau]$ converges to zero. In fact this is true even when $\mu_\epsilon, \lambda_\epsilon$ jointly converge to ∞ while maintaining $\rho_\epsilon = \lambda_\epsilon/\mu_\epsilon$ constant. If $\mu_\epsilon \to \infty$ for a fixed λ_ϵ, then the load factor also decreases to zero in limit. Thus the result would have been true only for low load factors. But by maintaining the ratio ρ_ϵ fixed when $\mu_\epsilon \to \infty$, we ensured that *the approximation is good for any given load factor and for any given admission control* p*, i.e., for any* (ρ_ϵ, p). Under SFJ limit, using Lemma 2:

$$\begin{aligned} E_{PS}[S_\tau(p)] &:= E_{PS}^{\mathcal{O}}[S_\tau(p)] \approx \frac{1}{\tilde{\mu}_{\tau,p}(1-\tilde{\rho}_{\tau,p})}, \qquad (12) \\ \text{with } \tilde{\rho}_{\tau,p} &= \rho_\tau a_0, \ \tilde{\mu}_{\tau,p} = \frac{\mu_\tau}{a_0} \text{ and } \rho_\tau := \frac{\lambda_\tau}{\mu_\tau}. \end{aligned}$$

Thus the achievable region under SFJ limit is given by:

$$\boxed{\mathcal{A}_{PS} = \left\{ \left((1-p) + \frac{p(\rho_{\epsilon,p})^K}{a_0}, \ \frac{a_0}{\mu_\tau\,(1-a_0\rho_\tau)} \right) \ \middle| \text{ with } a_0\rho_\tau < 1, \ 0 \le p \le 1 \right\}.}$$

In the above, condition $a_0\rho_\tau < 1$ ensures stability.

C. Validation of Pseudo conservation law (2), Completeness

By direct substitution[4] one can verify that the performance measures of $\beta_{p,K}^{PS}$ scheduler, for every (p,K), satisfy the pseudo conservation law (2). Further as K increases to ∞, the blocking probability $P_B^{PS}(1)$, given by equation (7), decreases to zero if $\rho_\epsilon \le 1$. When $\rho_\epsilon > 1$, using simple computations[5], one can show that $P_B^{PS}(1) \to 1-1/\rho_\epsilon$ and only $p_B > 1-1/\rho_\epsilon$ can be a part of the $\mathcal{A}_{hetero}^{static}$. Also it is easy to verify that the function, $p \mapsto P_B^{PS}(p)$, is continuous in p for any K. Thus by intermediate value theorem, all the points of $\mathcal{A}_{static}^{hetero}$ can be achieved by these schedulers. And hence the family of schedulers,

$$\mathcal{F}^{PS} := \big\{\beta_{p,K}^{PS}, \ 0 \le p \le 1, K\big\},$$

is complete. It is important to note here that these schedulers achieve the entire static region, nevertheless a larger K implies a larger time spent by ϵ-agents in the system. Thus system may have a restriction on the size of K to be used based on the QoS requirements.

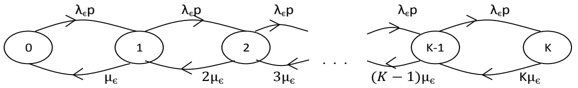


Fig. 4. State transitions for ϵ-agents in CD model.

V. Capacity Division (CD) Policies

In the previous section, when an admitted ϵ-customer preempts the ongoing service of τ-customer the entire system capacity is transferred to ϵ-customer. In this section we analyze a different scheduling policy. Here the capacity is not completely transferred, but rather a fraction of it is used by each ϵ-customer. The τ-customer is continued with the remaining capacity.

Service Discipline: Each ϵ-customer uses $(1/K)$-th part of the capacity, μ_ϵ/K. If the system has only one ϵ-customer, the remaining capacity i.e., $(K-1)/K$-th part of the capacity is utilized by the τ-customer. In other words, τ-class is served with rate $\mu_\tau(K-1)/K$. If there are $0 \le l \le K$ number of ϵ-customers receiving the service, then (l/K)-th part of the capacity is used by the ϵ-customers and the τ-customer is served at rate $((K-l)/K)\mu_\tau$. This continues up to K ϵ-customers, and any further (admitted) ϵ-arrival, departs without service. Whenever an existing ϵ-customer departs, the capacity is *readjusted to an appropriate higher value for* τ*-customer.*

It is clear that ϵ-class evolution is again independent of τ-class evolution and its analysis is considered first.

A. Blocking Probability of ϵ-class

Consider any fixed $0 \le p \le 1$. The ϵ-inter arrival times are exponentially distributed with parameter $\lambda_\epsilon p$. Say there are l ϵ-agents in the system (note $l \le K$). Each one of them receive service at rate μ_ϵ and this happens simultaneously. Thus the first departure time would be exponentially distributed with parameter $l\mu_\epsilon$. This is again a continuous time Markov jump process and its transitions are as shown in Figure 4. In fact the ϵ-agents evolve like the well known, finite capacity and finite buffer queueing system, $M/M/K/K$ queue. The stationary distribution of such a queue is well known and in particular (see for e.g., [8]):

$$\check{\pi}_K = \frac{(K\rho_{\epsilon,p})^K}{K!\check{a}_0} \text{ where } \check{a}_0 := \sum_{j=0}^{K} \frac{\left(K\rho_{\epsilon,p}\right)^j}{j!}.$$

As before, agents are admitted with probability $(1-p)$, and hence the overall blocking probability by PASTA equals:

$$P_B^{CD}(p) = (1-p) + p\check{\pi}_K = (1-p) + p\frac{(K\rho_{\epsilon,p})^K}{K!\check{a}_0} (13)$$

B. Expected sojourn time of τ-class

The idea is once again to approximately decouple the evolution of τ-agents from that of ϵ-agents. The procedure is similar, however the current model is more complicated. Once again EST is denoted as $\check{\Upsilon}_\tau$, has similar meaning as in section (IV-B) and typical τ-sojourn time equals the sum of waiting time and the effective server time (EST).

[4] By (7), $1-\rho_\epsilon(1-P_B^{PS}) = 1/a_0$ and so $\left(\mu_\tau\big[1-\rho_\epsilon(1-P_B^{PS})\big]-\lambda_\tau\right)^{-1}$ (see equation (2)) equals $E_{PS}[S_\tau]$ given by (12).

[5] It is easy to verify as $K \to \infty$ that:

$$\frac{\rho_\epsilon^K}{\sum_{l=0}^K \rho_\epsilon^l} = \frac{1}{\sum_{l=0}^K \rho_\epsilon^{-(K-l)}} = \frac{1}{\sum_{l=0}^K \rho_\epsilon^{-l}} \to \frac{1}{\frac{1}{1-\rho_\epsilon^{-1}}} = 1 - \frac{1}{\rho_\epsilon}.$$

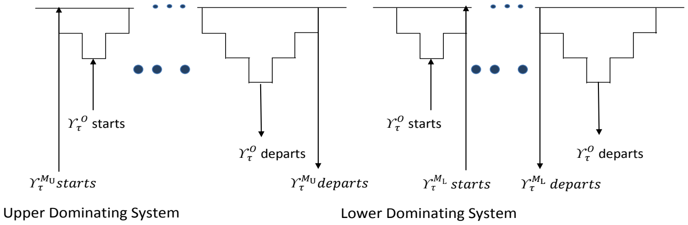

Υ_τ^O starts
Υ_τ^O departs
$\Upsilon_\tau^{M_U}$ starts
$\Upsilon_\tau^{M_U}$ departs
Upper Dominating System
Υ_τ^O starts
Υ_τ^O departs
$\Upsilon_\tau^{M_L}$ starts
$\Upsilon_\tau^{M_L}$ departs
Lower Dominating System

Fig. 5. Lower and Upper dominating systems (CD Model)

1) Analysis of effective server time $(\check{\Upsilon}_\tau)$*:* In the CD model, EST is the total time period between the service start and service end of a typical τ-agent, when the capacity is divided possibly between the two classes. With the arrival+admission of each ϵ-agent the server capacity available for the ongoing τ-agent reduces, with each ϵ-departure it increases, and its service is completed in the midst of such rate changes. In fact, the τ-agent's service is completely pre-empted with the admission of K-th ϵ-agent. The service would again be resumed, where left, when one of the K ϵ-agents depart. *The EST depends upon the number of ϵ-agents in the system at the service start.* Hence we introduce superscript l in the notation of $\check{\Upsilon}$. That is, $\check{\Upsilon}_\tau^l$ represent the EST, when it starts with l ϵ-agents. Thus the analysis of EST for this model is not as easy as in PS model. One can not estimate this using the number of interruptions and time per interruption as in PS model. However, the underlying transitions are Markovian in nature, and hence we obtain the analysis by directly considering the EST's $\{\check{\Upsilon}_\tau^l\}_l$. We have the following, (proof in Appendix C):

Theorem 2: The first two moments of EST $\check{\Upsilon}_\tau^0$ are:

$$E[\check{\Upsilon}_\tau^0] = \frac{\breve{a}_0 + O(1/\mu_\epsilon)}{\eta\mu_\tau + O(1/\mu_\epsilon)}, \ \breve{a}_0 := \sum_{j=0}^{K} \frac{\left(\rho_{\epsilon,p}\right)^j}{j!} \text{ and} \tag{14}$$

$$E\left[\left(\check{\Upsilon}_\tau^0\right)^2\right] = \frac{2\frac{\breve{a}_0^2}{\eta\mu_\tau} + O(1/\mu_\epsilon)}{\eta\mu_\tau + O(1/\mu_\epsilon)} \text{ with } \eta := \sum_{j=0}^{K-1} \frac{\left(\rho_{\epsilon,p}\right)^j}{j!}\frac{K-j}{K},$$

where $f(\mu_\epsilon) = O(1/\mu_\epsilon)$ for any function f implies, $f(\mu_\epsilon)\mu_\epsilon \to constant$ as $\mu_\epsilon \to \infty$, with ρ_ϵ fixed. ■

2) Dominating systems: It was not difficult to obtain the conditional moments of the EST, $\{\check{\Upsilon}_\tau^l\}_l$, when conditioned on the number of ϵ-agents at the service start. However to obtain the unconditional moments, one requires the stationary distribution of the number at service start of a typical τ-agent. And this is not a very easy task. However the various conditional moments differ from each other at maximum in one ϵ-busy period. Hence one can possibly obtain the (approximate) unconditional moments, along with M/G/1 queue approximation, using the idea of dominating fictitious queues.

We again have two dominating systems, which dominate on either side at the beginning of the τ-busy period, exactly as in the previous model. In addition, two partial ϵ-busy periods (if possible) are subtracted from every EST in the lower dominating system. While the upper dominating system is obtained by adding two extra partial ϵ-busy periods as in Figure 5. Using exactly the same logic as in the previous model, one can show that the expected sojourn time of the CD model can also be obtained as limit of the expected sojourn times of M/G/1 queues with service time moments given by that of $\check{\Upsilon}_\tau^0$ of Theorem 2. We again consider the SFT limit.

3) Sojourn time of CD Model: Using $M/G/1$ analysis and Theorem 2, the sojourn time for τ-agent in SFJ limit:

$$\mathcal{A}_{CD} = \left\{ \left((1-p) + p\frac{(K\rho_{\epsilon,p})^K}{K!\breve{a}_0}, \ \frac{1}{\ddot{\mu}_{\tau,p}\left(1 - \ddot{\rho}_{\tau,p}\right)} \right) \right. \tag{15}$$

$$\left. : \ddot{\rho}_{\tau,p} < 1, \ 0 \le p \le 1 \right\}, \text{ with } \ddot{\rho}_{\tau,p} = \frac{\lambda_\tau}{\ddot{\mu}_{\tau,p}},$$

$$\breve{a}_0 := \sum_{j=0}^{K} \frac{\left(K\rho_{\epsilon,p}\right)^j}{j!}, \ \eta := \sum_{j=0}^{K-1} \frac{\left(K\rho_{\epsilon,p}\right)^j}{j!}\frac{K-j}{K}, \text{ and } \ddot{\mu}_{\tau,p} = \frac{\eta\mu_\tau}{\breve{a}_0}.$$

By direct substitution[6] one can verify that the CD policies also satisfy the pseudo conservation law (2). Further they also form a complete family of schedulers, for exactly the same reasons as that for PS policy when $\rho_\epsilon \le 1$ and using Lemma 5 of Appendix E.

VI. Numerical examples

Random system with large μ_ϵ: We conduct Monte-Carlo simulations to estimate the performance of both the policies. We basically generate random trajectories of the two arrival processes, job requirements and study the system evolution when it schedules agents according to PS/CD policy. We estimated the blocking probability and expected sojourn time for ϵ and τ-agents respectively, using sample means, for different values of (p, K).

In Figure 3, we consider an example to compare the theoretical expressions with the ones estimated using Monte-Carlo simulations for PS policy. We consider two different values of ρ_ϵ. We notice negligible difference between the theoretical and simulated values with $\mu_\epsilon = 100$. However even with $\mu_\epsilon = 20$, the difference is about 10-12% for most of the cases.

We consider another example of PS model in Table I with, $K = 8$, $\lambda_\tau = 4$, $\mu_\tau = 8$ and $\rho_\epsilon = 0.5$. As the service rate of ϵ-agents increases with fixed load factor (ρ_ϵ), the simulator results are close to the theoretical results. The performance is very close to that of theoretical, for values of μ_ϵ greater than 120. For $\mu_\epsilon = 80$, the difference is at maximum 5%. Even at values as low as 20, the simulator performance is within 10% of the theoretical values for most of the cases. Thus the theoretical results well approximate the simulated ones, in most of the scenarios. Especially in the cases with large μ_ϵ, λ_ϵ.

Achievable region: is also plotted in Figure 3 for different values of ρ_ϵ. Towards this, we plot $E^{PS}[S_\tau(p)]$ versus $P_B^{PS}(p)$, for $p \in \{i\delta : 0 \le i \le 1/\delta\}$ with sufficiently small $\delta > 0$. It is a convex curve. We notice a downward shift (improvement) in the curve with smaller ρ_ϵ, as anticipated. However the formula derived, helps us understand the exact

[6] From equation (13),

$$1 - \rho_\epsilon(1 - P_B^{CD}) = \frac{\sum_{j=0}^{K-1} \frac{\left(K\rho_{\epsilon,p}\right)^j}{j!}\frac{K-j}{K}}{\sum_{j=0}^{K} \frac{\left(K\rho_{\epsilon,p}\right)^j}{j!}} = \frac{\eta}{\breve{a}_0}$$

and so $\left(\mu_\tau\left[1 - \rho_\epsilon(1 - P_B^{PS})\right] - \lambda_\tau\right)^{-1}$ (see equation (2)) equals $E_{PS}[S_\tau]$ given by (15).

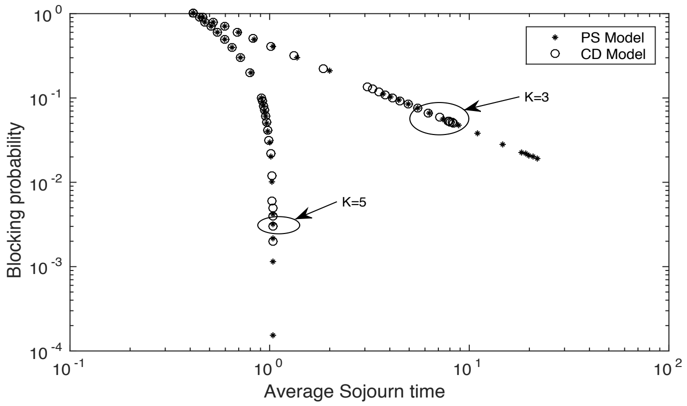


Fig. 6. Achievable regions $\mathcal{A}_{CD}$, $\mathcal{A}_{PS}$: P_B versus $E[S_\tau]$ for different ρ_ϵ, $\rho_\epsilon = 0.9/K$

		Simulation		Theoretical	
p	μ_ϵ	P_B^{PS}	$E[S_\tau]$	P_B^{PS}	$E[S_\tau]$
0	20	1.0000	0.2500	1.0000	0.2500
	80	1.0000	0.2500	1.0000	0.2500
	120	1.0000	0.2500	1.0000	0.2500
0.25	20	0.7500	0.3525	0.7500	0.3333
	80	0.7500	0.3387	0.7500	0.3333
	120	0.7500	0.3373	0.7500	0.3333
0.50	20	0.5000	0.5671	0.5000	0.5000
	80	0.5000	0.5146	0.5000	0.5000
	120	0.5000	0.5105	0.5000	0.5000
0.75	20	0.2502	1.2377	0.2502	0.9993
	80	0.2502	1.0417	0.2502	0.9993
	120	0.2502	1.0367	0.2502	0.9993
1	20	0.0020	110.2218	0.0020	127.750
	80	0.0020	112.1243	0.0020	127.750
	120	0.0020	127.1987	0.0020	127.750

TABLE I
PS MODEL: COMPARISON OF SIMULATED, THEORETICAL METRICS

amount of shift. We plotted the curves only in the τ-stability region, $\{\lambda_\tau : a_0\rho_\tau < 1\}$. *Unlike the case of homogeneous agents, the τ-stability region varies with the scheduling policy. This is because, varying fractions of ϵ-agents are lost with different values of* p*, which can expand or contract the stability region.*

Comparison of the two policies: We compare the achievable regions of PS and CD policies by plotting $\mathcal{A}_{CD}$ and $\mathcal{A}_{PS}$. We set $\rho_\epsilon = 0.9/K$, $\lambda_\tau = 5.6$ $\mu_\tau = 8$ and $K = 3$ or 5.

In Figure 6, we plot the achievable region for both the models/policies, i.e, we plot $E[S_\tau(p)]$ versus $P_B(p)$, for different p. And in Figures 7, we plot the performance measures $P_B(p)$ and $E[S_\tau(p)]$ respectively versus p with $K = 3$. From Figure 6, the two achievable (sub) regions overlap, however we observe from the Figures 7 that the performance measures of the two models are different for the same (p, K). But if we choose a p and $p\prime$ such that $P_B^{CD}(p) = P_B^{PS}(p\prime)$, we observe that the two expected sojourn times are equal. Because of this the two achievable regions overlap in Figure 6. This observation is precisely the pseudo-conservation law. Whatever the policy used, once the blocking probabilities are the same the expected sojourn times are the same.

Now we will discuss a slightly different, yet, a related important aspect. We would compare the two sets of policies, when K (maximum number of parallel calls) is the same. As seen from the figures the sub-achievable region of CD policy, with fixed K, is a strict subset of that of the PS policy. This is because the best possible blocking probability with CD policy,

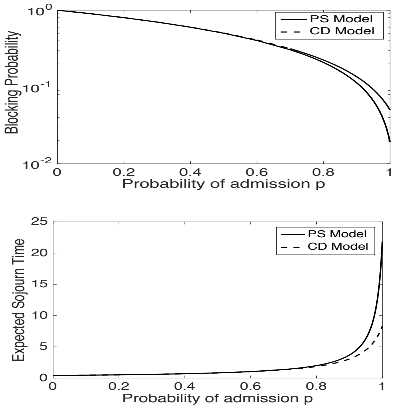


Fig. 7. P_B, $E[S_\tau]$ versus p

$$P_B^{CD}(1) = \frac{(K\rho_\epsilon)^K/K!}{\sum_{j=0}^{K}(K\rho_\epsilon)^j/j!} \geq \frac{(\rho_\epsilon)^K}{\sum_{j=0}^{K}(\rho_\epsilon)^j} = P_B^{PS}(1),$$

is greater than that with the PS policy. In Figure 6 the best P_B with CD and PS models/policies respectively is 0.002 and 0.0002 (0.05 and 0.019) when $K = 5$ ($K = 3$). Thus it appears that the static achievable region would overlap for different policies, however the sub-regions covered by different policies can be different when K is fixed.

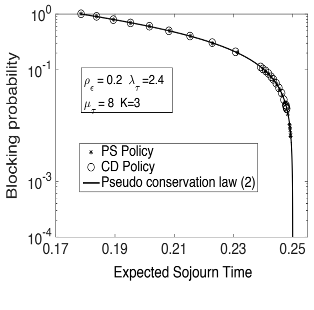


Fig. 8. Static Achievable region

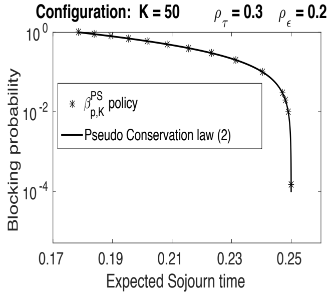


Fig. 9. Completeness of $\mathcal{F}^{PS}$

Completeness: In Figure 8, we plot pseudo-conservation law (2). We also plot the performance of PS/CD policies with $K = 3$ and for varying p. We see that the three curves exactly overlap, again validating (2). For the same configuration we plot performance of PS policies with a bigger $K = 50$, in Figure 9. With $K = 50$ we are able to achieve a bigger part of the achievable region. One can achieve a similar result with CD policy. With even bigger K one can achieve further lower parts of the pseudo-conservation curve. However, as mentioned before, one may not be able to use a larger K because of other QoS restrictions. For example, the ϵ customers may not agree for a very small service rate (μ_ϵ/K) which can prolong their stay in the system. It is in this context that the PS could be better than the CD policies. Even though both the sets of policies are complete, PS policy achieves a bigger sub-region than the CD policy for the same K (see Figures 6 and 8).

VII. A DYNAMIC POLICY

We consider dynamic policies (for PS model) with an aim to demonstrate that the dynamic region is bigger than the static region. Towards this we construct an example dynamic policy and show that the block probability, for the same sojourn time $E[S_\tau]$, is better with the dynamic policy.

The static policy of the previous sections is modified as follows. We refer this as policy β_p^d. When there are no τ-agents in the system, i.e., during the τ-idle period, there is no admission control for ϵ-agents. An arriving ϵ-agent is admitted with probability one. Recall, however that service is offered to an admitted agent only when the number in system is less than K. When the system is in τ-busy period[7], i.e., when the τ-queue is non-empty, we admit the ϵ-agents with probability p. So, this is a dynamic policy which alternates between full and partial admission.

Let Ψ_τ and $\mathcal{I}_\tau$ respectively represent the busy and idle periods of the τ-agents. By stationarity, memoryless property, the consecutive busy, idle periods $\{\Psi_{\tau,i}\}$, $\{\mathcal{I}_{\tau,i}\}_i$ are independent and identically distributed. We have (proof is in Appendix D):

Theorem 3: The block probability, $P_d^B(p)$, for the system with the dynamic policy β_p^d:

$$P_d^B(p) = \frac{E[\mathcal{I}_{\tau,1}]P^B(1)}{E[\Psi_{\tau,1}] + E[\mathcal{I}_{\tau,1}]} + \frac{E[\Psi_{\tau,1}]P^B(p)}{E[\Psi_{\tau,1}] + E[\mathcal{I}_{\tau,1}]}. \quad \blacksquare \quad (16)$$

Using the ideas of dominating systems as in the section IV one can show that the moments of the idle, busy periods of the original system with policy β_p^d converges towards that of the equivalent $M/G/1$ system $\mathcal{M}_L$, as $\mu_\epsilon \to \infty$. Thus we will have for large values of μ_ϵ:

$$\begin{aligned} E[\mathcal{I}_{\tau,1}] &\approx \frac{1}{\lambda_\tau}, \\ E^{\mathcal{O}}[\Psi_{\tau,1}] &\approx E^{\mathcal{M}_L}[\Psi_{\tau,1}] = \frac{E[\Upsilon_\tau]}{1-\lambda_\tau E[\Upsilon_\tau]} \to \frac{a_0}{\mu_\tau - \lambda_\tau a_0}. \end{aligned}$$

The second last equality is obtained using the well known formula for the average busy period of an $M/G/1$ queue. It is easy to see that the sojourn time of the dynamic policy β_p^d is same as that with static policy β_p (asymptotically), while the blocking probability is improved from (7) to (16). Note that $P^B(1) \le P^B(p)$ for any $p \le 1$. Hence the dynamic policy performs better and the dynamic achievable region is bigger. One can obtain similar improvement with CD model.

Numerical comparison of Dynamic and Static regions

In Figure 10, we compare the performance of the dynamic policy β_p^d with the corresponding static policy, for PS model. We notice a good improvement in the curve: blocking probability decreases significantly for the same expected sojourn time. This indicates that the dynamic region is strictly bigger than the static region, unlike the homogeneous case. In future, we would like to obtain complete analysis of dynamic achievable region for this heterogeneous system.

VIII. Conclusions and future work

We consider a queueing system with heterogeneous classes of agents. The impatient class demands immediate service, hence receives the service immediately and if required in parallel with others. There is an admission control to ensure the QoS requirements of the other (tolerant) class. The tolerant class can wait for their turn, however would like to optimize their sojourn time.

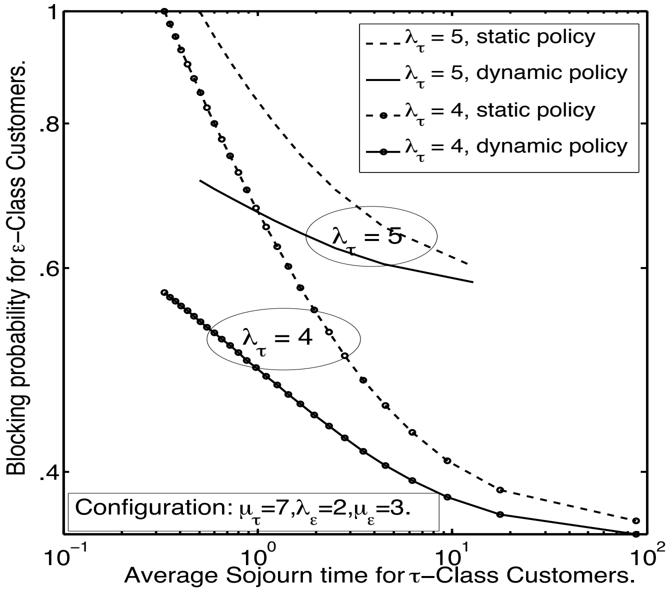


Fig. 10. Static-dynamic policies, $K = 4$.

We conjecture a pseudo conservation law for this lossy queueing system, which relates the blocking probability of impatient agents to the expected sojourn time of the tolerant agents, in a short and frequent job (SFJ) limit-regime for the former. The pseudo conservation law should be satisfied by all the policies, that are static (do not depend upon the state) and work conserving (left over server capacity is completely used when there is a customer) with respect to the tolerant agents.

We consider two families of scheduling policies, which differ in the way the system capacity is shared between the two classes. With processor sharing policy the entire system capacity is transferred to impatient customer, once admitted. In the second policy, which we refer as capacity division policy, only a (fixed) fraction of capacity is transferred to each admitted impatient customer.

We obtain closed form expressions for the asymptotic performance measures, under SFJ limit, for both the families of policies. The two families satisfy the pseudo-conservation law. Further, both the families are complete, i.e., they attain every point of the achievable region given by the pseudo-conservation law. The CD achievable region is a strict subset of the PS region, when restricted to the same number of parallel service possibilities. This demonstrates the limitation of CD model, which could be a more practically used model. The PS model can attain a much smaller blocking probability.

We obtain the performance with an example dynamic policy and illustrate that the dynamic region is strictly bigger than the static region. This is in contrast to the homogeneous case (all tolerant classes), where the two regions coincide. However, under a dynamic policy the impatient agents may experience non-stationary blocking. Some systems may not prefer this and hence static policies have importance of their own.

[7] Normally a busy period begins immediately with an arrival to an empty queue. However, in our system we say a τ-busy period starts with the service start of that τ-agent, which arrives to an τ-empty queue. If ϵ-agents were present at the τ-arrival instance, the service of the τ-agent is deferred till the end of the ongoing ϵ-busy period.

The results are asymptotic and are accurate when the arrival-departure rates of the impatient class is large. Usually such customers have short frequent job requirements and hence this is an useful asymptotic result. Further, we have an upper and lower bound for the sojourn time performance, even when the rates are not large.

References

[1] Ayesta, Urtzi. "A unifying conservation law for single-server queues." Journal of Applied Probability 44.4 (2007): 1078-1087.

[2] White, Harrison, and Lee S. Christie, 'Queuing with pre-emptive priorities or with breakdown', *Operations research*, 6.1, pp. 79-95, 1958.

[3] L. Kleinrock, 'A delay dependent queue discipline', *Naval Research Logistics Quarterly*, vol. 11, pp. 329–341, September-December 1964.

[4] Harchol-Balter, Mor, Takayuki Osogami, Alan Scheller-Wolf, and Adam Wierman. 'Multi-server queueing systems with multiple priority classes', *Queueing Systems*, 51(3-4), pp.331-360, 2005.

[5] S. Tang and Wei Li, 'A Channel Allocation Model with Preemptive Priority for Integrated Voice/Data Mobile Networks', *Proceedings of the First International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks*, *IEEE*, 2004.

[6] Yan Zhang, Boon-Hee Soong and Miao Ma, 'A dynamic channel assignment scheme for voice/data integration in GPRS networks', *Elsevier Computer communications,* 29, pp. 1163–1163, 2006.

[7] Izagirre, Ane, Urtzi Ayesta, and Ina Maria Verloop. 'Sojourn time approximations in a multi-class time-sharing server', *In IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 2786-2794. IEEE, 2014.

[8] Hoel, Paul G., Sidney C. Port, and Charles J. Stone. 'Introduction to stochastic processes', *Waveland Press*, 1986.

[9] L. Kleinrock, 'A conservation law for wide class of queue disciplines', *Naval Research Logistics Quarterly*, vol. 12, pp. 118–192, 1965.

[10] Choudhury, G.L., Leung, K.K. and Whitt, W. 'An algorithm to compute blocking probabilities in multi-rate multi-class multi-resource loss models', *Advances in Applied Probability*, pp.1104-1143, 1995.

[11] E. G. Coffman and I. Mitrani, 'A characterization of waiting time performance realizable by single server queues', *Operations Research*, vol. 28, pp. 810 – 821, 1979.

[12] J. G. Shanthikumar and D. D. Yao, 'Multiclass queueing systems: Polymatroidal structure and optimal scheduling control', *Operations Research*, vol. 40, no. 3-supplement-2, pp. S293–S299, 1992.

[13] Sleptchenko, A., A. van Harten, and M. C. van der Heijden. 'An Exact Analysis of the Multi-class M/M/k Priority Queue with Partial Blocking', pp. 527-548, 2003.

[14] Sleptchenko, A., A. van Harten, and M. C. van der Heijden. 'An exact solution for the state probabilities of the multi-class, multi-server queue with preemptive priorities', *Queueing Systems*, 50.1, 2005.

[15] A. Federgruen and H. Groenevelt, 'M/G/c queueing systems with multiple agent classes: Characterization and control of achievable performance under nonpre-emptive priority rules', *Management Science*, vol. 9, pp. 1121– 1138, 1988.

[16] D. Bertsimas, I. Paschalidis, and J. N. Tistsiklis, 'Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance', *The Annals of Applied Probability*, vol. 4, pp. 43–75, 1994.

[17] D. Bertsimas, 'The achievable region method in the optimal control of queueing systems; formulations, bounds and policies', *Queueing systems*, vol. 21, no. 3-4, pp. 337–389, 1995.

[18] D. Bertsimas and J. Niño-Mora, 'Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems', *Mathematics of Operations Research*, vol. 21, no. 2, pp. 257–306, 1996.

[19] C.-p. Li and M. J. Neely, 'Delay and rate-optimal control in a multi-class priority queue with adjustable service rates', in *INFOCOM, Proceedings of IEEE*, pp. 2976–2980, 2012.

[20] I. Mitrani and J. Hine, 'Complete parametrized families of job scheduling strategies', *Acta Informatica*, vol. 8, pp. 61– 73, 1977.

[21] R. Hassin, J. Puerto, and F. R. Fernández, 'The use of relative priorities in optimizing the performance of a queueing system', *European Journal of Operational Research*, vol. 193, no. 2, pp. 476–483, 2009.

[22] SF Yashkov, 'Processor sharing queues: Some progress in analysis', *Queueing Systems*, 2, 117, 1987.

[23] Roland de Haan, , Richard J. Boucherie, and Jan-Kees van Ommeren. 'A polling model with an autonomous server', *Queueing Systems*, 62.3, pp. 279-308, 2009.

Proofs and Appendices are in the next page

Appendix : Static and dynamic regions coincide in homogeneous systems

We discuss the case without pre-emption. By the well known work conservation principle ([9]), for any (work conserving) scheduling policy the expected waiting times of the two classes satisfy:

$$\rho_1 E[W_1] + \rho_2 E[W_2] = c \implies E[W_1] = \frac{c}{\rho_1} - \frac{\rho_2}{\rho_1} E[W_2], \tag{17}$$

where c is an appropriate constant. A family of schedulers is called complete (e.g., [20], [21]), if every performance vector of the achievable region is obtained by one of its schedulers. It is well known that the priority schedulers (see for e.g., [3]) form a complete family for homogeneous classes. Each scheduler is represented by a priority factor b, say meant for class 1, and class 1 is scheduled if b times its longest waiting time $\tilde{w}_1$ is greater than $\tilde{w}_2$. By varying b from 0 to infinity, we obtain all the performance pairs of the dynamic region.

Consider the set of static policies parametrized[8] by $p \in (0,1)$. Class 1 is scheduled with probability p, now independent of everything else. As $p \to 1$, class 1 gets maximum priority. It is not difficult to see that the limit performance of the two classes with $p \to 1$ equals that obtained with dynamic priority scheduler as $b \to \infty$. Both the limits are obtained by giving absolute priority to class 1 and hence are equal. Similarly the performance under static policy with $p \to 0$, equals the performance under dynamic policy with $b = 0$. Further, it is easy to show that the performance of the two queues is continuous in parameter p of the static policy. Thus by intermediate value theorem all the values between the two extremes are achieved as p varies between 0 and 1. And these pairs satisfy (1). Hence the static achievable region coincides with dynamic region in homogeneous setting. ■

Appendix A: Results related to Pseudo Conservation Law

Lemma 3: Let $\tau^\mu := \inf_t \{t : R^\mu(t) \geq B\}$, be any random time defined using any random variable B which is independent of $\{A^1_{\epsilon,n}, B^1_{\epsilon,n}\}_n$. Since $R^\mu(t)$ satisfies (6) we have:

$$\left|\tau^\mu - \frac{B}{\nu_\tau}\right| \to 0 \text{ a.s. as } \mu \to \infty$$

Proof: By continuity of probability measure, one can assume that (6) is satisfied together for a sequence of $\{W_n\}$ with $W_n \to \infty$, almost surely. Let

$$A := \left\{ w : \sup_{t \in [0, W_n]} \left|R^\mu(t) - \nu_\tau t\right| \to 0 \text{ for all } n \right\} \text{ and } P(A) = 1.$$

For any $w \in A$, consider a $W_n > (B(w)+1)/\nu_\tau$. For every $1 > \varepsilon > 0$, there exists an $\mu_\varepsilon < \infty$ such that:

$$\sup_{t \in [0, W_n]} \left|R^\mu(t) - \nu_\tau t\right| \leq \varepsilon \text{ for all } \mu \geq \mu_\varepsilon.$$

$$\text{Thus } \nu_\tau t - \varepsilon \leq R^\mu(t) \leq \nu_\tau t + \varepsilon \text{ for all } t \leq W_n.$$

Now in particular for $t = (B(w)+\varepsilon)/\nu_\tau < W_n$, we have:

$$R^\mu(B(w)/\nu_\tau + \varepsilon/\nu_\tau) \geq \nu_\tau((B(w)+\varepsilon)/\nu_t) - \varepsilon = B(w),$$

which implies

$$\tau^\mu \leq (B(w)+\varepsilon)/\nu_\tau.$$

Further, trivially $(B(w)-\varepsilon)/\nu_\tau < W_n$ and hence we also have:

$$R^\mu(B(w)/\nu_\tau - \varepsilon/\nu_\tau) \leq \nu_\tau((B(w)-\varepsilon))/\nu_t) + \varepsilon = B(w).$$

Thus for any $0 < \varepsilon < 1$,

$$\tau^\mu \geq (B(w)-\varepsilon)/\nu_\tau \text{ and so } \left|\tau^\mu - B(w)/\nu_\tau\right| \leq \frac{\varepsilon}{\nu_\tau}. \qquad \square$$

Theorem 4: (Functional RRT) Let $R^1(.)$ be any monotone time-reward function of a renewal process with parameter 1. Let a be the RRT limit, i.e, say

$$\frac{R^1(t)}{t} \to a \text{ almost surely as } t \to \infty.$$

Then, for any $W < \infty$:

$$\sup_{s \in [0,W]} \left|\frac{R^1(\mu s)}{\mu} - as\right| \to 0 \text{ as } \mu \to \infty \text{ a.s..}$$

[8] The system is unstable with $p = 1$ or 0, these values are not considered.

Proof: For any ϵ there exists a T_ϵ such that:

$$\left|\frac{R^1(t)}{t} - a\right| \leq \epsilon \;\forall\; t \geq T_\epsilon \tag{18}$$

Consider s such that $\mu s \geq T_\epsilon$, it follows that

$$\left|\frac{R^1(\mu s)}{\mu s}\frac{s}{W} - a\frac{s}{W}\right| \leq \frac{s}{W}\epsilon \leq \epsilon. \tag{19}$$

If s is such that $\mu s < T_\epsilon$,

$$\begin{aligned}\left|\frac{R^1(\mu s)}{\mu s}\frac{s}{W} - a\frac{s}{W}\right| &= \left|\frac{R^1(\mu s)}{\mu W} - a\frac{s}{W}\right| \\ &\leq \frac{R^1(\mu s)}{\mu W} + a\frac{s}{W} \\ &\leq \frac{R^1(T_\epsilon)}{\mu W} + a\frac{T_\epsilon}{\mu W} < \epsilon,\end{aligned} \tag{20}$$

by choosing μ further large to get the required upper bound, if required. Equations (19) and (20) are true for any arbitrary s. Hence the result follows. □

Appendix B: Proof of Lemma 2

By conditioning on B_τ, one can verify that

$$\begin{aligned} E[N(B_\tau)] &= \frac{\lambda_\epsilon p}{\mu_\tau},\; E[B_\tau N(B_\tau)] = \frac{2\lambda_\epsilon p}{\mu_\tau^2}, \\ E[(N(B_\tau))^2] &= \frac{\lambda_\epsilon p}{\mu_\tau} + \frac{2(\lambda_\epsilon p)^2}{\mu_\tau^2}. \end{aligned}$$

By conditioning on $N(B_\tau)$ we obtain the first moment:

$$\begin{aligned} E[\Upsilon_\tau] &= E[B_\tau] + E\left[\sum_{i=1}^{N(B_\tau)} \Psi_{\epsilon,i}\right] \\ &= E[B_\tau] + E\left[E\left[\sum_{i=1}^{N(B_\tau)} \Psi_{\epsilon,i}\middle| N(B_\tau)\right]\right] = \frac{1}{\mu_\tau} + \frac{\lambda_\epsilon p E[\Psi_\epsilon]}{\mu_\tau}. \end{aligned} \tag{21}$$

Note that the busy periods $\{\Psi_{\epsilon,i}\}_i$ are IID. From (8) we have:

$$E[\Upsilon_\tau^2] \;=\; E[B_\tau^2] + 2E\big[B_\tau \Upsilon_\tau^e\big] + E\big[(\Upsilon_\tau^e)^2\big]. \tag{22}$$

By first conditioning on $(B_\tau,\, N(B_\tau))$ and then on B_τ:

$$\begin{aligned} E\big[B_\tau \Upsilon_\tau^e\big] &= E\left[E\left[B_\tau \sum_{i=1}^{N(B_\tau)} \Psi_{\epsilon,i}\middle| B_\tau, N(B_\tau)\right]\right] \\ &= \lambda_\epsilon p E\big[\Psi_\epsilon\big] E\big[B_\tau B_\tau\big] = \frac{2\lambda_\epsilon p E\big[\Psi_\epsilon\big]}{\mu_\tau^2}. \end{aligned} \tag{23}$$

Conditioning as before and because of independence:

$$\begin{aligned} E\big[(\Upsilon_\tau^e)^2\big] &= E\left[E\left[\Big(\sum_{i=1}^{N(B_\tau)} \Psi_{\epsilon,i}\Big)^2 \middle| N(B_\tau)\right]\right], \\ &= \frac{\lambda_\epsilon p E\big[\Psi_\epsilon^2\big]}{\mu_\tau} + \frac{2(\lambda_\epsilon p)^2}{\mu_\tau^2}\big(E\big[\Psi_\epsilon\big]\big)^2, \end{aligned}$$

which simplifies to (9).

Busy period of ϵ-class: Busy period of any class is defined as the time till the first epoch at which all the customers of that class have departed. Let Ψ_k, represent the busy period of ϵ-class, when it begins with k number of customers. Note that $\Psi_\epsilon = \Psi_1$. In all the discussions below, an arrival is meant an admitted arrival.

The busy period Ψ_1 starts with the arrival of one ϵ-customer. If the customer leaves before the next arrival, the busy period ends. On the other hand, if an arrival occurs before the departure of the existing customer, it marks the beginning of a busy period with two customers, Ψ_2. As seen in section IV-A (see Fig. 1), a departure time is memoryless, i.e., exponential random variable with parameter μ_ϵ irrespective of the number of customer sharing the service. Let D represent the departure time. The inter arrival time, A, is exponential with parameter $\lambda_\epsilon p$. Let $W := \min\{D, A\}$ represent the minimum of the two. With these definitions:

$$\Psi_1 = 1_{\{D<A\}}\, 0 + 1_{\{A<D\}}\, \Psi_2 + W. \tag{24}$$

The busy period Ψ_2 starts with two ϵ-customers. If one of the two customers leave before the next arrival, it marks the beginning of the busy period Ψ_1, and an early arrival marks the beginning of a busy period with three customers, Ψ_3. From Lemma 1 of section IV-A, the departure time of the earliest customer among the two is again exponential random variable with parameter μ_ϵ. Thus this departure time is also distributed as D, defined above. The inter arrival time A obviously remains the same as in the previous paragraph. Thus:

$$\Psi_2 = 1_{\{D<A\}}\, \Psi_1 + 1_{\{A<D\}}\, \Psi_3 + W.$$

Continuing using similar logic we have:

$$\begin{aligned} \Psi_K &= 1_{\{D<A\}}\, \Psi_{K-1} + 1_{\{A<D\}}\, \Psi_K + W \text{ and} \\ \Psi_i &= 1_{\{D<A\}}\, \Psi_{i-1} + 1_{\{A<D\}}\, \Psi_{i+1} + W \;\; \forall 1 < i < K. \end{aligned} \tag{25}$$

In the first equation of (25) the two Ψ_K are different, independent of each other, but they are identically distributed. For ease of notation, we represent them by the same symbol. Note, in all, that the random variables W, D and A have same stochastic nature and are correlated. Further, if an arrival occurs before departure when the system already has K customers, the arrival is dropped. By memoryless property of exponential distributions, we again have busy period Ψ_K. Taking expectation of equations (24) - (25) and solving backward recursively ($\{a_i\}$, $\{b_i\}$ given in (10)):

$$E[\Psi_i] = \frac{i a_i + b_i}{\mu_\epsilon} \text{ for all } 1 \le i \le K. \tag{26}$$

Squaring and taking the expectation of (24) we obtain:

$$\begin{aligned} E\big[\Psi_1^2\big] &= c_1 + qE\big[\Psi_2^2\big] \text{ where } q := E\big[A < D\big] \\ c_1 &= 2E\left[W 1_{\{A<D\}}\right] E\big[\Psi_2\big] + E\big[W^2\big] \\ &= \frac{2\lambda_\epsilon p}{(\lambda_\epsilon p + \mu_\epsilon)^2} E\big[\Psi_2\big] + \frac{2}{(\lambda_\epsilon p + \mu_\epsilon)^2}. \end{aligned}$$

Terms c_1, q simplify as in (10). Similarly from (25) we have

$$\begin{aligned} E[\Psi_i^2] &= c_i + qE\big[\Psi_{i+1}^2\big] + (1-q)E\big[\Psi_{i-1}^2\big] \text{ with} \\ c_i &= 2E\left[W 1_{\{A<D\}}\right] E[\Psi_{i+1}] + 2E[W 1_{\{D<A\}}]E[\Psi_{i-1}] \\ &\quad + E[W^2] \quad \text{for any } 2 \le i < K. \end{aligned}$$

Constant c_i simplifies as in (10). Now squaring Ψ_K of (25):

$$E\big[\Psi_K^2\big] \;\; = \;\; E\big[\Psi_{K-1}^2\big] + \frac{c_K}{(1-q)}.$$

Solving the expressions backward recursively we obtain:

$$\begin{aligned} & E[\Psi_\epsilon^2] = E[\Psi_1^2] \\ & = \frac{q^{K-1}\, c_K}{(1-q)^K} + \frac{q^{K-2}\, c_{K-1}}{(1-q)^{K-1}} + ... + \frac{q\, c_2}{(1-q)^2} + \frac{c_1}{1-q}. \quad \blacksquare \end{aligned} \tag{27}$$

Appendix C: Proof of Theorem 2

Let $I_l^m := [l, \cdots, m]$ represent the interval of integers, the big O notations are shortly represented by

$$O_\epsilon := O(1/\mu_\epsilon),\ O_\epsilon^{(2)} := O(1/\mu_\epsilon^2),\ \tilde{\lambda} := \lambda_\epsilon p \text{ and } [i] := K - i. \tag{28}$$

Let $\check{\Upsilon}_l = \check{\Upsilon}_\tau^l$, with $l \in I_0^{K-1}$, represent a simpler notation for EST when it begins with l ϵ-agents. Let us begin with Łthe analysis of $\check{\Upsilon}_0$. If τ-agent leaves before next ϵ-arrival, the EST ends. If instead an ϵ-agent arrives before, it marks the beginning of EST, $\check{\Upsilon}_1$.Ł Let D_τ represent the departure time of τ-agent and $D_\tau \sim \exp(\mu_\tau)$. It equals B_τ since the service is offered at full capacity. Let $A \sim \exp(\tilde{\lambda})$ represent the exponential inter arrival time of admitted ϵ-agent. Let $\bar{W}_0 := \min\{D_\tau, A\}$ represent the minimum. Then,

$$\check{\Upsilon}_0 = 1_{\{D_\tau < A\}}\, 0 + 1_{\{A < D_\tau\}}\, \check{\Upsilon}_1 + \bar{W}_0. \tag{29}$$

Let $D_\epsilon \sim \exp(\mu_\epsilon)$ represent the departure time of an ϵ-agent. EST, $\check{\Upsilon}_l$, starts with l ϵ-agents. If one of the ϵ-agents depart before the next ϵ-arrival or τ-departure, it marks the beginning of EST $\check{\Upsilon}_{l-1}$ and an early ϵ-arrival begins $\check{\Upsilon}_{l+1}$. The τ-departure ends the EST. Let D_ϵ^l represent the departure time of the earliest among l ϵ-agents, and note $D_\epsilon^l \sim \exp(l\mu_\epsilon)$. Let D_τ^l represent the departure time of τ-agent, when the capacity is shared with l ϵ-agents, then $D_\tau^l \sim \exp([l]\mu_\tau/K)$. Inter arrival time, A, remains the same. Let $\bar{W}_l := \min\{A, D_\epsilon^l, D_\tau^l\}$ represent the minimum of three, which is again exponentially distributed. Thus for any $l \in I_1^{K-1}$ we have:

$$\check{\Upsilon}_l = 1_{\{D_\epsilon^l = \bar{W}_l\}}\, \check{\Upsilon}_{l-1} + 1_{\{A = \bar{W}_l\}}\, \check{\Upsilon}_{l+1} + \bar{W}_l. \tag{30}$$

For the case with K ϵ agents, if one of them leave before next arrival, an EST with $(K-1)$ ϵ-agents begins and an early arrival begins another EST with K ϵ-agents. Further, any arrival of ϵ-agent is dropped in this case. By memoryless property, we again have busy period $\check{\Upsilon}_K$ ($\bar{W}_K := \min\{D_\epsilon^K, A\}$):

$$\check{\Upsilon}_K = 1_{\{D_\epsilon^K < A\}}\, \check{\Upsilon}_{K-1} + 1_{\{A < D_\epsilon^K\}}\, \check{\Upsilon}_K + \bar{W}_K. \tag{31}$$

In the above the two $\check{\Upsilon}_K$ are different, but are identically distributed. For ease of notation, we represent them by the same symbol. Taking expectation of (29)-(31) ($[i] := K - i$):

$$\begin{aligned} E\big[\check{\Upsilon}_0\big] &= \frac{\tilde{\lambda}}{\alpha_0} E\big[\check{\Upsilon}_1\big] + \frac{1}{\alpha_0}, \\ E\big[\check{\Upsilon}_{[i]}\big] &= \frac{[i]\mu_\epsilon}{\alpha_{[i]}} E\big[\check{\Upsilon}_{[i]-1}\big] + \frac{\tilde{\lambda}}{\alpha_{[i]}} E\big[\check{\Upsilon}_{[i]+1}\big] + \frac{1}{\alpha_{[i]}},\ \forall\ i \in I_1^{K-1} \\ E\big[\check{\Upsilon}_K\big] &= \frac{K\mu_\epsilon}{\alpha_K} E\big[\check{\Upsilon}_{K-1}\big] + \frac{\tilde{\lambda}}{\alpha_K} E\big[\check{\Upsilon}_K\big] + \frac{1}{\alpha_K},\ \text{where} \\ \alpha_i &:= \tilde{\lambda} + i\mu_\epsilon + \frac{[i]\mu_\tau}{K}\ \text{ for any } i \in I_0^K. \end{aligned} \tag{32}$$

Solving the equations backward recursively (start with K):

$$\begin{aligned} E\big[\check{\Upsilon}_K\big] &= m_K + E\big[\check{\Upsilon}_{K-1}\big], \\ E\big[\check{\Upsilon}_{K-1}\big] &= m_{K-1} + n_{K-1} E\big[\check{\Upsilon}_{K-2}\big] \text{ and} \\ E\big[\check{\Upsilon}_{[i]}\big] &= m_{[i]} + n_{[i]} E\big[\check{\Upsilon}_{[i]-1}\big]\ \forall\ i \in I_2^{K-1}, \end{aligned} \tag{33}$$

where the coefficients are defined recursively as below:

$$\begin{aligned} m_K &= \frac{1}{\gamma_k},\ m_{K-1} = \frac{1 + m_K\tilde{\lambda}}{\gamma_{K-1}},\ n_{K-1} = \frac{(K-1)\mu_\epsilon}{\gamma_{K-1}}, \\ m_{[i]} &= \frac{1 + m_{[i]+1}\tilde{\lambda}}{\alpha_{[i]} - n_{[i]+1}\tilde{\lambda}},\ n_{[i]} = \frac{([i])\mu_\epsilon}{\alpha_{[i]} - n_{[i]+1}\tilde{\lambda}},\ \forall\ i \in I_2^{K-1} \\ \gamma_i &= i\mu_\epsilon + \frac{[i]\mu_\tau}{K},\ \forall\ i \in I_0^K. \end{aligned} \tag{34}$$

Using the first equation of (32) and $E[\check{\Upsilon}_1]$ of equation (33) we obtain:

$$E[\check{\Upsilon}_0] = \frac{1 + m_1\tilde{\lambda}}{\alpha_0 - n_1\tilde{\lambda}} = \frac{1 + m_1\tilde{\lambda}}{\tilde{\lambda} + \mu_\tau - n_1\tilde{\lambda}}. \tag{35}$$

Squaring and taking the expectation of (31) we get[9]

$$\begin{aligned} E\big[\big(\check{\Upsilon}_K\big)^2\big] &= \frac{\tilde{\lambda}}{\alpha_K} E\big[\big(\check{\Upsilon}_K\big)^2\big] + \frac{K\mu_\epsilon}{\alpha_K} E\big[\big(\check{\Upsilon}_{K-1}\big)^2\big] \\ &\quad + \frac{2}{\alpha_K^2} + \frac{2\tilde{\lambda}}{\alpha_K^2} E\big[\check{\Upsilon}_K\big] + \frac{2K\mu_\epsilon}{\alpha_K^2} E\big[\check{\Upsilon}_{K-1}\big]. \end{aligned}$$

[9] Since the product of the two indicators is zero, we will not have cross correlation terms like $E\big[\check{\Upsilon}_{K1}\check{\Upsilon}_{K2}\big]$ etc. Furthe note that the indicators, $\{\bar{W}_l\}$ are independent of the ESTs $\{\check{\Upsilon}_l\}_l$ on the right hand side of the equations (29)-(31).

Simplifying we obtain:

$$E\big[\big(\check{\Upsilon}_K\big)^2\big] = r_K + E\big[\big(\check{\Upsilon}_{K-1}\big)^2\big] \text{ with } \delta_K = \gamma_K = K\mu_\epsilon \text{ and} \tag{36}$$
$$r_K = \frac{2}{\delta_K \alpha_K} + \frac{\sigma_K}{\delta_K},\ \sigma_K = \frac{2\delta_K}{\alpha_K} E\big[\check{\Upsilon}_{K-1}\big] + \frac{2\tilde{\lambda}}{\alpha_K} E\big[\check{\Upsilon}_K\big].$$

Similarly from (30) we obtain for any $i \in I_1^{K-1}$,

$$\begin{aligned} E\big[\big(\check{\Upsilon}_{[i]}\big)^2\big] &= r_{[i]} + \frac{[i]\mu_\epsilon}{\delta_{[i]}} E\big[\big(\check{\Upsilon}_{[i]-1}\big)^2\big] \text{ with} \\ r_{[i]} &= \frac{1}{\delta_{[i]}}\left[\frac{2}{\alpha_{[i]}} + \tilde{\lambda} r_{[i]+1} + \sigma_{[i]}\right], \\ \delta_{[i]} &= \frac{\alpha_{[i]}\delta_{[i]+1} - \tilde{\lambda}([i]+1)\mu_\epsilon}{\delta_{[i]+1}}, \\ \sigma_{[i]} &= \frac{2\tilde{\lambda}}{\alpha_{[i]}} E[\check{\Upsilon}_{[i]+1}] + \frac{2([i])\mu_\epsilon}{\alpha_{[i]}} E[\check{\Upsilon}_{[i]-1}]. \end{aligned} \tag{37}$$

From (29),
$$E\big[\big(\check{\Upsilon}_0\big)^2\big] = \frac{\tilde{\lambda}}{\alpha_0} E\big[\big(\check{\Upsilon}_1\big)^2\big] + \frac{2}{\alpha_0^2} + \frac{2\tilde{\lambda} E[\check{\Upsilon}_1]}{\alpha_0^2}.$$

Further using equation (37) with $i = K-1$ or $[i] = 1$:

$$E\big[\big(\check{\Upsilon}_0\big)^2\big] = \frac{1}{\delta_0}\left[\frac{2}{\alpha_0} + \tilde{\lambda} r_1 + \frac{2\tilde{\lambda} E[\check{\Upsilon}_1]}{\alpha_0}\right],\ \delta_0 = \frac{\alpha_0 \delta_1 - \tilde{\lambda}\mu_\epsilon}{\delta_1}. \tag{38}$$

SFT Limit: From (35) and using (41) of Lemma 4 (see (28)):

$$E\big[\check{\Upsilon}_0\big] = \frac{\check{a}_0 + O_\epsilon}{\eta\mu_\tau + O_\epsilon}. \tag{39}$$

Thus and considering the limit (forward) recursively in (32)

$$E\big[\check{\Upsilon}_l\big] = \lim E\big[\check{\Upsilon}_0\big] + O_\epsilon = \frac{\check{a}_0}{\eta\mu_\tau} + O_\epsilon.$$

Hence for all $i \in I_0^{K-1}$ from (37)

$$\sigma_{[i]} = \theta + O_\epsilon, \quad r_K \tilde{\lambda} = \frac{\rho_{\epsilon,p}\theta}{K} + O_\epsilon \text{ where } \theta := \frac{2\check{a}_0}{\eta\mu_\tau}.$$

Again considering limits (backward) recursively in (37), while using the above two equations and equation (42) of Lemma 4 and backward induction (as in Lemma 4) we obtain:

$$\begin{aligned} \tilde{\lambda} r_{[i]} &= \left(\frac{\rho_{\epsilon,p}}{[i]} + O_\epsilon\right)\left(\tilde{\lambda} r_{[i]+1} + \theta + O_\epsilon\right) \\ &= \left(\frac{\rho_{\epsilon,p}}{[i]} + O_\epsilon\right)\left(\sum_{j=0}^{i-1} \frac{\rho_{\epsilon,p}^{j+1}\theta}{([i]+1)\cdots([i]+1+j)} + \theta + O_\epsilon\right) \\ &= \sum_{j=0}^{i} \frac{\rho_{\epsilon,p}^{j+1}\theta}{[i]\cdots([i]+j)} + O_\epsilon \text{ for all } i \in I_0^{K-1}. \end{aligned}$$

Simplifying we obtain:
$$\tilde{\lambda} r_1 + \frac{2\tilde{\lambda} E[\check{\Upsilon}_1]}{\alpha_0} = \frac{2\check{a}_0^2}{\eta\mu_\tau} + O_\epsilon.$$

Further using δ_0 of (42) of Lemma 4 we obtain the asymptotic limit of the second moment (38). ■

Lemma 4: We have the following asymptotic results for the coefficients defined in the proof of Theorem 2:

$$\begin{aligned} n_{[i]} &= 1 - \frac{\mu_\tau \omega_{[i]}}{\mu_\epsilon} + O_\epsilon^{(2)},\ \ i \in I_1^{K-1} \text{ with} \\ \omega_{[i]} &:= \frac{1}{K}\sum_{j=0}^{i-1} \frac{(i-j)\rho_{\epsilon,p}^j}{([i]+j)([i]+j-1)\cdots([i])}, \end{aligned} \tag{40}$$

$$\tilde{\lambda} + \mu_\tau - n_1\tilde{\lambda} = \eta + O_\epsilon, \qquad 1 + m_1\tilde{\lambda} = \check{a}_0 + O_\epsilon, \tag{41}$$

$$\delta_{[i]} = [i]\mu_\epsilon + [i]\omega_{[i]}\mu_\tau + O_\epsilon\ \forall\, i \in I_1^{K-1} \text{ and } \delta_0 = \eta\mu_\tau + O_\epsilon. \tag{42}$$

Proof: We begin with terms $\{n_{[i]}\}$ and prove the required result by backward mathematical induction. From (34),

$$n_{K-1} = 1 - \frac{\mu_\tau \omega_{K-1}}{\mu_\epsilon} + O_\epsilon^{(2)},\ \text{ with } \omega_{K-1} := \frac{1}{K(K-1)}.$$

Assume the statement holds for $i = l-1$, i.e., say:

$$\begin{aligned} n_{K-l+1} &= 1 - \frac{\mu_\tau \omega_{K-l+1}}{\mu_\epsilon} + O^{(2)}_\epsilon \text{ and} \\ \omega_{K-l+1} &:= \frac{1}{K}\sum_{j=0}^{l-2} \frac{(l-1-j)\rho^j_{\epsilon,p}}{(K-l+1+j)\cdots(K-l+1)}. \end{aligned}$$

We need to prove the result for $i = l$. From (34) and substituting the above

$$\begin{aligned} n_{K-l} &= \frac{(K-l)\mu_\epsilon}{\alpha_{K-l} - n_{K-l+1}\tilde{\lambda}} = \frac{(K-l)\mu_\epsilon}{(K-l)\mu_\epsilon + \frac{l\mu_\tau}{K} + \mu_\tau\rho_{\epsilon,p}\omega_{K-l+1} + O_\epsilon} \\ &= 1 - \frac{\mu_\tau\omega_{K-l}}{\mu_\epsilon} + O^{(2)}_\epsilon \text{ as } \mu_\epsilon \to \infty \text{ with } \rho_\epsilon \text{ constant.} \end{aligned}$$

This proves (40). It is easy to see that

$$\omega_1 = \sum_{j=0}^{K-2} \frac{K-j}{K}\frac{\rho^j_{\epsilon,p}}{j!} \text{ and hence that } \omega_1\rho_{\epsilon,p} + 1 = \eta,$$

where η is defined in the hypothesis of the Theorem 2. Using this we obtain the first part of (41):

$$\tilde{\lambda} + \mu_\tau - n_1\tilde{\lambda} = (\omega_1\rho_{\epsilon,p} + 1)\mu_\tau + O_\epsilon = \eta\mu_\tau + O_\epsilon.$$

Using (40), for all $i \in I_2^{K-1}$ (note $\lambda_\epsilon = \rho_\epsilon\mu_\epsilon$ with ρ_ϵ fixed),

$$\alpha_{[i]} - n_{[i]+1}\tilde{\lambda} = [i]\mu_\epsilon + \left(\omega_{[i]+1}\rho_{\epsilon,p} + \frac{i}{K}\right)\mu_\tau + O_\epsilon \tag{43}$$

and hence

$$\frac{\tilde{\lambda}}{\alpha_{[i]} - n_{[i]+1}\tilde{\lambda}} = \frac{\rho_{\epsilon,p}}{[i]} + O_\epsilon.$$

From above and using a similar backward induction on $\{m_i\}$ of (34) we obtain,

$$m_{[i]}\tilde{\lambda} = \sum_{j=1}^{i+1} \frac{\rho^j_{\epsilon,p}}{([i])([i]+1)\cdots([i]+j)} + O_\epsilon, \forall\ i \in I_2^{K-1}. \tag{44}$$

Thus we get the second part of (41):

$$1 + m_1\tilde{\lambda} = \sum_{j=0}^{K} \frac{\rho^j_{\epsilon,p}}{j!} = \breve{a}_0 + O_\epsilon.$$

Let $\zeta_{[i]} := [i]\mu_\epsilon/\delta_{[i]}$, for all $i \in I_1^{K-1}$. Using the recursive definition of δ_i, as given in (37), $\zeta_{[i]}$ satisfies the following recursive equation,

$$\zeta_{[i]} = \frac{[i]\mu_\epsilon}{\alpha_{[i]} - \tilde{\lambda}\zeta_{[i]+1}}, \tag{45}$$

just like the recursive definition of $\{n_i\}$ given in (34). Further,

$$\begin{aligned} \zeta_{K-1} &= (K-1)\mu_\epsilon/\gamma_{K-1} = n_{K-1} \text{ and hence using (40)} \\ \zeta_{[i]} &= n_{[i]} = 1 - \frac{\mu_\tau\omega_{[i]}}{\mu_\epsilon} + O^{(2)}_\epsilon \text{ for all } i \in I_1^{K-1}. \end{aligned}$$

Thus we have the first part of (42):

$$\delta_{[i]} = [i]\mu_\epsilon + \mu_\tau[i]\omega_{[i]} + O_\epsilon \text{ for any } i \in I_1^{K-1}.$$

And from (38),

$$\delta_0 = \mu_\tau + \mu_\tau\rho_{\epsilon,p}\omega_1 + O_\epsilon = \eta\mu_\tau + O_\epsilon. \qquad \blacksquare$$

Appendix D: Proof of Theorem 3

For any static policy β_p, the processor sharing system with ϵ-agents is ergodic. With $L_p(T)$ representing the number of ϵ agents lost in time T, when the arrivals are admitted at rate p, we have:

$$\lim_{T\to\infty} \frac{L_p(T)}{T} = P^B(p) \text{ almost surely,}$$

where $P^B(p)$ is given by equation (7).

Let $L^d_p(T)$ represent the number of ϵ-agents lost in time T with dynamic policy. Let $\mathcal{I}_\tau(T)$, $\Psi_\tau(T)$ respectively represent the total τ-idle time and total τ-busy period until time T. These are basically the sum of all the busy/idle periods that elapsed till the time T. Note that $\mathcal{I}_\tau(T) + \Psi_\tau(T) = T$. In this case,

$$L^d_p(T) = L^d_p(\mathcal{I}_\tau(T)) + L^d_p(\Psi_\tau(T)).$$

At the end epoch of any busy period ($\Psi_{\tau,i}$ for some i), the system is completely empty. That is, agents of both the classes are absent. Also a τ-busy period starts only once the system is free from all of its ϵ-agents. Thus there are no ϵ-agents in the system at both start and end epochs of a τ-busy period. Hence the evolution of the ϵ-class loss counting process that occurred during disjoint time intervals (of τ-busy periods) constituting $\Psi_\tau(T)$ is stochastically equivalent to the ϵ-class loss counting process that would have evolved in a continuous time interval of length exactly $\Psi_\tau(T)$. This is because of the memoryless property associated with Poisson arrival process. Thus we have:

$$\frac{L_p^d(\Psi_\tau(T))}{\Psi_\tau(T)} \to P^B(p) \text{ almost surely .}$$

In a similar way at the start/end epoch of any τ-idle period the system is free of ϵ-agents. Using similar arguments, and because all ϵ-agents are admitted during idle periods we have:

$$\frac{L_p^d(\Psi_\tau(T))}{\Psi_\tau(T)} \to P^B(1) \text{ almost surely .}$$

Further using renewal reward theorem, one can show that the following happens almost surely:

$$\frac{\Psi_\tau(T))}{T} \to \frac{E[\Psi_{\tau,1}]}{E[\Psi_{\tau,1}] + E[\mathcal{I}_{\tau,1}]}, \quad \frac{\mathcal{I}_\tau(T))}{T} \to \frac{E[\mathcal{I}_{\tau,1}]}{E[\Psi_{\tau,1}] + E[\mathcal{I}_{\tau,1}]}.$$

Using all the results established so far, equation (16) follows because:

$$P_d^B(p) = \lim_{T\to\infty} \frac{L_p^d(T)}{T}. \qquad \blacksquare$$

Appendix E

Lemma 5: For any $\rho_\epsilon > 1$,

$$P_B^{CD}(1) \to 1 - \frac{1}{\rho_\epsilon} \text{ and } P_B^{CD}(1) \to 0 \text{ if } \rho_\epsilon \leq 1.$$

Proof: When $\rho_\epsilon \leq 1$ we have:

$$P_B(1) = \frac{(K\rho_\epsilon)^K}{K! \sum_{j=0}^{K} \frac{\left(K\rho_\epsilon\right)^j}{j!}} = \frac{\frac{1}{\sum_{j=0}^{K} \frac{\left(K\rho_\epsilon\right)^j}{j!}}}{\frac{(K\rho_\epsilon)^K}{K!}} = \frac{1}{\sum_{j=0}^{K} \frac{K(K-1)\cdots(K-j+1)}{K^j}\rho_\epsilon^{j-K}} \leq \frac{1}{\sum_{j=0}^{K} \frac{K(K-1)\cdots(K-j+1)}{K^j}}.$$

Let

$$f(K) := \sum_{j=0}^{K} \frac{K(K-1)\cdots(K-j+1)}{K^j} = 1 + (1-\frac{1}{K}) + (1-\frac{1}{K})(1-\frac{2}{K}) + \cdots + \Pi_{i=1}^{K}(1-\frac{i}{K}),$$

and note that

$$f(K) \geq N\Pi_{i=1}^{N}(1-\frac{i}{K}) \text{ for any } N \leq K.$$

Fix any $\varepsilon > 0$. For any N there exist a large K_N such that for all $K \geq K_N$,

$$((1-\frac{1}{K+1})\cdots(1-\frac{N}{K+1}) \geq (1-\varepsilon) \text{ which implies } f(K) \geq N(1-\varepsilon) \text{ and hence } f(K) \to \infty \text{ as } K \to \infty.$$

Thus $P_B^{CD}(1) \to 0$. When $\rho_\epsilon > 1$, it is clear after redefining that:

$$f(K) := \sum_{j=0}^{K} \frac{K(K-1)\cdots(K-j+1)}{K^j}\rho_\epsilon^{j-K} \leq \sum_{j=0}^{K} \rho_\epsilon^{j-K} = \sum_{j=0}^{K} \rho_\epsilon^{-j} \text{ hence } \lim_{K\to\infty} f(K) \leq \frac{1}{1-\rho_\epsilon^{-1}}.$$

On the other hand for any $\varepsilon > 0$, as before for any N for all $K > K_N$ we have:

$$f(K) \geq \sum_{j=K-N}^{K} \rho_\epsilon^{j-K}(1-\varepsilon) = \sum_{j=0}^{N} \rho_\epsilon^{-j}(1-\varepsilon).$$

By first letting $N \to \infty$ we have

$$\lim_{K\to\infty} f(K) \geq (1-\varepsilon)\left(\frac{1}{1-\rho_\epsilon^{-1}}\right) \text{ and then with } \varepsilon \to 0 \lim_{K\to\infty} f(K) \geq \frac{1}{1-\rho_\epsilon^{-1}}. \qquad \square$$