Noname manuscript No.
(will be inserted by the editor)

Pseudo conservation for partially fluid, partially lossy queueing systems

Veeraruna Kavitha[1] **Jayakrishnan Nair**[2]
and Raman Kumar Sinha[1]

the date of receipt and acceptance should be inserted later

Abstract We consider a queueing system with heterogeneous customers. One class of customers is eager; these customers are impatient and leave the system if service does not commence immediately upon arrival. Customers of the second class are tolerant; these customers have larger service requirements and can wait for service. In this paper, we establish pseudo-conservation laws relating the performance of the eager class (measured in terms of the long run fraction of customers blocked) and the tolerant class (measured in terms of the steady state performance, e.g., sojourn time, number in the system, workload) under a certain partial fluid limit. This fluid limit involves scaling the arrival rate as well as the service rate of the eager class proportionately to infinity, such that the offered load corresponding to the eager class remains constant. The workload of the tolerant class remains unscaled. Interestingly, our pseudo-conservation laws hold for a broad class of admission control and scheduling policies. This means that under the aforementioned fluid limit, the performance of the tolerant class depends only on the blocking probability of the eager class, and not on the specific admission control policy that produced that blocking probability. Our psuedo-conservation laws also characterize the achievable region for our system, which captures the space of feasible tradeoffs between the performance experienced by the two classes. We also provide two families of complete scheduling policies, which span the achievable region over their parameter space. Finally, we show that our pseudo-conservation laws also apply in several scenarios where eager customers have a limited waiting area and exhibit balking and/or reneging behaviour.

Index terms– Heterogeneous queues, multiclass queues, loss systems, fluid limits, pseudo-conservation, achievable region, complete policies, reneging, balking

[1]IEOR, [2]EE, Indian Institute of Technology Bombay, India

1 Introduction

In this paper, we analyse a queueing system serving two distinct classes of jobs. One class of jobs is eager or impatient – these jobs leave the system if service does not commence (almost) immediately upon arrival. The other class of jobs is tolerant to delay. While the relevant performance metric for the eager class is the blocking probability (i.e., the long run fraction of arrivals that are denied service), the performance of the tolerant class is determined by the delay experienced by these jobs.

The above setup is motivated by modern cellular systems, which carry both voice and data traffic. Voice calls are either admitted or dropped on arrival, whereas data traffic can tolerate delay. Moreover, the service capacity in current cellular systems is made up of disjoint channels or resource blocks (see, for example, Sesia et. al. (2011)) which can be dynamically allocated to either voice or data traffic. This allows the system operator to share the available service capacity across voice and data traffic, resulting in a tradeoff between the quality of service (QoS) experienced by the two classes.

In this paper, we prove 'policy-independent' pseudo-conservation laws relating the blocking probability of the eager jobs and the performance experienced by the tolerant jobs, which can be characterized via the distribution/moments of the steady state queue occupancy, sojourn time, or workload. These conservation laws are derived under a partial fluid scaling, where the arrival and service rates of the eager class are scaled to infinity, such that the offered load of this class remains constant. We refer to this scaling regime as the short-frequent jobs (SFJ) scaling. The tolerant workload is not scaled. In essence, the SFJ scaling corresponds to a timescale separation between the service of the eager and tolerant class; the eager class being served on a faster timescale. Our results are proved under the assumption that the scheduling and admission control policy for the eager class does not depend on the state of the tolerant jobs in the system.[1] In other words, the eager jobs are treated as a higher priority class, that can pre-empt any tolerant jobs in service. This assumption is motivated by the current practice of awarding voice calls a higher priority in cellular systems (Tang et al. (2004); Zhang (2006) etc.).

An important implication of our pseudo-conservation laws is that under the aforementioned partial fluid scaling, the performance experienced by the tolerant class depends only on the blocking probability of the eager class, and not on the specific scheduling (and admission control) policy[2] that produced this blocking probability. Moreover, the pseudo-conservation laws characterize the achievable region for the system, which is the set of performance vectors that can be achieved across all scheduling policies. Note that this is a par-

[1] We categorize the eager customers as those that demand immediate service and at-least at a certain minimum service rate. Since eager customers once admitted have to be allocated a certain minimum service capacity, admission control becomes an important aspect of any scheduling policy for the eager class.

[2] Henceforth, we follow the convention that scheduling for the eager class includes admission control.

tially static achievable region, in the sense that the eager schedulers under consideration are blind to the state of the tolerant customers. Of course, such a characterization of the achievable region is important from the standpoint of a system operator who has to balance the QoS of the eager and tolerant class.

Interestingly, in the pre-limit, i.e., for finite arrival and service rates for the eager class, the achievable region is difficult to characterize. Moreover, it depends on the scheduling policy of the eager class, and not just on the resulting blocking probability. However, the achievable region becomes insensitive to the scheduling policy of the eager class under the SFJ limit, and even admits a simple closed form characterization.

Having characterized the achievable region, we introduce two families of parametrized admission control and scheduling policies for the eager class that are complete, i.e., these policies span the entire achievable region across their parameter space.[3] Note that a complete policy (more precisely, a complete family of policies) enables a system operator to achieve any feasible tradeoff between the QoS of the eager and tolerant classes by tuning the policy parameters. For example, the system operator might be interested in maximizing the revenue from serving eager and tolerant jobs, subject to certain QoS constraints. A complete policy enables the above optimization to be performed over its parameter space with no loss of generality.

The first policy, which we call the PS policy, performs processor sharing between eager jobs (at most K of them) using the entire service capacity. Tolerant jobs are only served when there are no eager jobs in the system. The second policy, which we refer to as the capacity-division (CD) policy, admits upto K concurrent eager jobs, awarding each a fixed fraction of the service capacity. This seeks to capture the current practice in cellular systems, of allocating a channel or resource block having a predefined capacity to an admitted voice call. The residual service capacity is utilized by the tolerant jobs in the system. The above policies were first analysed under the SFJ limit using direct (policy specific) methods in our prior work Kavitha et al. (2017a); the same results follow easily from the general pseudo-conservation laws proved in this paper.

Our results also extend to certain scenarios where eager jobs have limited patience, and can wait for service in a buffer of bounded size. This includes balking models, where customers decide whether or not to join the queue based on the queue occupancy (in a possibly randomised manner), as well as reneging models, where customers abandon the queue if their waiting time exceeds a (possibly random) threshold (e.g., Ancker et al. (1963); Whitt (1999)). So long as we scale the abandonment times appropriately, our pseudo-conservation laws extend to these models. In other words, if the service capacity left unused by the above eager class is utilized by another class with large service requirements that can tolerate delay, our psuedo-conservation laws provide an

[3] Such complete families are well known in the context of multiclass queueing systems with different tolerant classes; see, for example, Mitrani et al. (1977).

approximate closed form expression for the performance of this class (accurate in the fluid limit), once the loss probability of the eager class is computed (as is done, for example, in Whitt (1999)).

In many scenarios, it is relatively straightforward to derive closed form expressions for the stationary measures of the lossy/eager class; this boils down to the analysis of a single-class loss system. However the challenge is to characterize the performance of the tolerant customers. Indeed, the standard performance metrics for the tolerant class cannot be obtained in closed form even when the service process seen by it is Markov modulated (see Mahabhashyam et al. (2005)). However, by our pseudo-conservation laws, the performance of the tolerant class is characterized as a function of the blocking probability of the eager class under a certain fluid limit.

The main feature of our models is a high priority eager class with short jobs and a low priority tolerant class with larger job sizes. There are several other applications that motivate such models. For example, in a super-market, it is common practice to provide fast service to short jobs via dedicated express counters. Alternatively, one could use the same counters to serve both short and long jobs, with selected (controlled) short jobs pre-empting long jobs, to obtain an optimal design. Primary and secondary users in cognitive radio can also be thought of as eager and tolerant classes, respectively. Variations of the models considered in this paper can be applied to these settings (e.g., capturing multi-server, non work-conserving service of tolerant jobs; or when eager job sizes are also significant but are negligible in comparison with the tolerant job sizes) and these models would be of future interest for us.

Related literature

To the best of our knowledge, there has not been much focus in the literature on partially lossy multiclass queueing systems. The only paper we are aware of that analyses a queueing system with impatient and tolerant job classes is Sleptchenko et al. (2003).[4] In this paper, the authors obtain the performance measures for each class of jobs in closed form in a multi-server environment, assuming exponential inter arrival and service times for all classes. The model in Sleptchenko et al. (2003) is similar to the CD policy proposed in this paper, except that we consider a work conserving system wherein the tolerant workload utilizes all the service capacity left unused by the eager jobs. Also, we have more tractable characterization of the performance of the tolerant class, which is accurate under a fluid limit. Additionally, the main contribution of the present paper is a policy independent pseudo-conservation law under a partial fluid scaling, for generally distributed service times.

In terms of application, this paper is also related to the sizeable literature on sharing capacity across voice and data traffic in cellular systems; for example, see Li (2004); Tang et al. (2004); Zhang (2006) etc. In this stream of work, both voice as well as data are assumed to be lossy, but with different priority levels. For various models for capacity sharing between voice and data

[4] Aside from our prior work Kavitha et al. (2017a).

calls, these papers typically propose algorithms for computing the blocking probability of each class, assuming exponential inter-arrival and call holding times. To the best of our knowledge, none of the papers in this space consider the achievable region of performance vectors for cellular systems, which provides a smooth (continuous) trade-off between the two competing performance metrics (e.g., blocking probability of voice calls and mean delay of data traffic).

Some variants of queueing systems in the literature (e.g., de Haan et al. (2009); White et al. (1958)) have connections to special cases of some of our models; we discuss these connections here. White et al. (1958) analyzes an $M/G/1$ queue with Poisson interruptions, caused either by server breakdown or by the arrival of higher priority agents. This system is equivalent to our PS/CD model, assuming only one eager job is served at a time ($K = 1$). The (exponentially distributed) time limited autonomous polling system of de Haan et al. (2009), with a single class, coincides again with the above. Both papers provide complicated expressions of performance metrics, while we have simple closed-form expressions in an appropriate fluid limit.

This paper is also related to the literature on multiclass queueing systems with multiple tolerant classes and a single server (e.g., conservation laws, pioneered by Kleinrock (1965)). The achievable region is well understood for homogeneous classes, when the performance metric of both the classes is expected sojourn time. Coffman et al. (1979) were the first to identify such achievable regions. Multi-class single server queueing systems possess nice geometric structure (polytopes) for the achievable region (e.g., Coffman et al. (1979); Shanthikumar et al. (1992)). These results mainly bank upon the work conservation principle applicable to non-lossy systems with work-conserving scheduling policies. Our attempt in this paper has been to explore if such a conservation is possible for lossy systems, in some partial/limited sense. Indeed, we show that for a partially lossy system with fluid lossy components, there is a conservation of the performance of the non-lossy class, given the blocking probability of the lossy class. However, it is important to note that the achievable region characterization in this paper is over a 'partially static' class of policies, where the scheduling of the eager class is oblivious to the state of the tolerant queue. Interestingly, the achievable region over the class of 'dynamic' policies, where all scheduling decisions can depend on the complete system state, is a strict superset of the 'partially static' achievable region; this is demonstrated in Kavitha et al. (2017a,b) by considering a specific dynamic policy. This presents a contrast from homogeneous systems with multiple tolerant classes, where the static and dynamic achievable regions are known to coincide.

Finally, complete family of schedulers are well known in the context of homogeneous queueing systems, that consider only tolerant classes (e.g., Mitrani et al. (1977)). But we are not aware of such families for queueing systems with heterogeneous classes, even ones that span a sub-achievable region like our partially static region.

Organisation of this paper

The remainder of this paper is organised as follows. In Section 2, we describe our model and define the SFJ scaling regime. The first set of pseudo-conservation laws are obtained in Section 3, for the unused service process of the eager class. Pseudo-conservation laws for performance metrics of the tolerant class which utilizes the above unused service process are derived in Section 4. In Section 5, we discuss the achievable region as well as the complete families of schedulers. Extensions of our model to eager customers with limited patience are considered in Section 6. We present numerical experiments in Section 7 and conclude in Section 8.

2 Model and preliminaries

We consider a single server queueing system serving two classes of jobs: an eager class and a tolerant class. Eager jobs, also referred to as ϵ-jobs, if admitted, require service to commence immediately upon arrival.[5] Tolerant jobs, also referred to as τ-jobs, are always admitted, and can wait in a queue (of infinite capacity) until they are served. Without loss of generality, we assume that the server operates at unit speed.

We assume that the ϵ-customers arrive according to a renewal process with rate λ_ϵ, i.e., the inter-arrival times $\{A_{\epsilon,n}\}_{n\geq 1}$ are independent and identically distributed (IID) with $\mathbb{E}[A_{\epsilon,n}] = 1/\lambda_\epsilon$ for any n. Job sizes (a.k.a. service requirements) of ϵ-jobs are IID, with $\{B_{\epsilon,n}\}_{n\geq 1}$ denoting the job size sequence having mean $1/\mu_\epsilon$. From Section 4 onwards, we will require additional workload-related assumptions; these will be stated in Section 4.

We now state and discuss the assumptions about the scheduling policies. We begin with the ϵ-jobs.

A.1 Admission control and scheduling of ϵ-jobs is τ-insensitive (i.e., does not depend on the state of the τ-queue).

A.2 The admission control and scheduling of ϵ-jobs is based only upon the number of ϵ-jobs present in the system.

A.3 The service of an admitted ϵ-job begins immediately, and each ϵ-job in service receives a service rate of at least $c_{min} > 0$ at all times.

Assumption **A.1** implies that the eager class uses the server as a higher priority class, oblivious to the tolerant class. This means that the tolerant class can only utilize the left-over service capacity unused by the eager class at any time. Assumption **A.2** implies that the decision of whether or not to admit an incoming $\epsilon-$job depends only on the number of $\epsilon-$jobs in service at that time. Finally, Assumption **A.3** implies that there exists $K \leq 1/c_{min}$ such that, at

[5] We consider the case of partially eager customers with limited patience in Section 6.

most K ϵ-jobs derive service at any time.[6] To give a concrete example, one scheduling policy for the eager class satisfying the above assumptions is the following: For some $K \in \mathbb{N}$, an ϵ-arrival is admitted if the number of ϵ-jobs in service is strictly less than K. Each admitted ϵ-job is served at rate $1/K$. We refer to this policy as the capacity division (CD) policy (see Section 5.2).

We also make the following assumptions on the scheduling policy for $\tau-$jobs.

B.1 The $\tau-$scheduler is work-conserving, i.e., it utilizes all the service capacity unused by ϵ-jobs, so long as the τ-queue is non-empty.
B.2 The τ-jobs are served in a serial fashion, i.e., τ jobs cannot pre-empt one another and are served one at a time.
B.3 The $\tau-$scheduler is blind to the sizes of $\tau-$jobs.

Assumption **B.1** implies that the tolerant class experiences an exogenous, time varying, service process. Moreover, note that while the service process of the τ-queue is regenerative, it is not even necessarily Markovian. Assumptions **B.2-3** imply that we consider τ-schedulers which are non-preemptive and non-anticipative. Policies in this class include first-come-first-served (FCFS), last-come-first-served (LCFS), and random-order-of-service (Harchol-Balter 2013, Chapter 29).

For the multiclass queueing system described above, performance evaluation of the higher priority eager class is typically straightforward. Indeed, under Assumption **A.1**, the blocking probability of the eager class can be characterized by analysing a single-class loss system. However, the performance evaluation for the (lower priority) tolerant class, which sees an exogenous and time varying service process, is challenging. For example, even when the service process evolves as a continuous time Markov chain, standard metrics like the steady state mean sojourn time cannot be obtained in closed form (Mahabhashyam et al. 2005). However, it turns out that under a certain fluid limit, the performance of the tolerant class becomes tractable. We describe this scaling regime next.

Short-frequent jobs scaling

We now introduce the partial fluid scaling regime considered in this paper. Under this scaling regime, the arrival rate as well as the service rate of eager class is scaled to infinity, while maintaining a constant offered load. Note however that the workload parameters corresponding to the tolerant class are not scaled.

The scaling parameter μ_ϵ corresponds to the service rate of ϵ jobs. Specifically, a generic service requirement for an ϵ-job at scale μ_ϵ is defined as

$$B_\epsilon^{\mu_\epsilon} \stackrel{d}{=} \frac{B_\epsilon^1}{\mu_\epsilon}, \text{ with } \mathbb{E}\left[B_\epsilon^1\right] = 1,$$

[6] We consider exceptions to Assumption **A.3** in Section 6, where we also allow the eager customers to wait in the queue to a limited extent.

and $\stackrel{d}{=}$ representing the equality in distribution. The inter-arrival times of ϵ-jobs at scale μ_ϵ are defined as

$$A_\epsilon^{\mu_\epsilon} \stackrel{d}{=} \frac{A_\epsilon^1}{\mu_\epsilon}, \text{ with } \mathbb{E}\left[A_\epsilon^1\right] = 1/\lambda_\epsilon^1, \text{ i.e., } \lambda_\epsilon^{\mu_\epsilon} = \lambda_\epsilon^1 \mu_\epsilon.$$

Note that as μ_ϵ is scaled to infinity, both the arrival rate as well as the service rate of the eager class are scaled to infinity proportionately, such that the offered load $\rho_\epsilon^{\mu_\epsilon} := \lambda_\epsilon^{\mu_\epsilon} \mathbb{E}\left[B_\epsilon^{\mu_\epsilon}\right] = \rho_\epsilon = \lambda_\epsilon^1$ remains scale invariant. We refer to the above scaling as the short-frequent jobs (SFJ) scaling.

3 Pseudo-conservation of ϵ unused service process/τ service process

In this section, we focus on the service capacity left unused by the (higher priority) eager class. Under Assumption **B.1**, this ϵ unused service (ϵ-US) process is also the service process seen by the (lower priority) tolerant class. We prove a pseudo-conservation law relating the blocking probability of the eager class and the ϵ-US process under the SFJ limit. Specifically, we show under the SFJ limit, that the ϵ-US process has a steady rate, determined purely by the blocking probability of the eager class. In Section 4, we exploit this characterization of the ϵ-US process to develop pseudo-conservation laws for performance metrics of the tolerant class.

Let $\Omega^{\mu_\epsilon}(t)$ represent the total amount of server capacity unused by the (higher priority) ϵ-customers, in the interval $[0, t]$. It is important to note under **A.1**, that the ϵ-US process $\{\Omega^{\mu_\epsilon}(\boldsymbol{.})\}$ depends only on the admission control and scheduling policy of the eager class and is completely oblivious to the τ-system. Our first observation is that the blocking probability of the eager class, as well as the long run average service rate corresponding to the ϵ-US process remain constant under the SFJ scaling for any $\mu_\epsilon > 0$ (i.e., even pre-limit). We require the following assumption, which will be used throughout the paper.[7]

A.0 The interval between the start of two successive busy periods of the the ϵ-class has finite expectation.

Lemma 1 *Assume* **A.0-2**. *The steady state blocking probability of ϵ-jobs is the same for all μ_ϵ under the SFJ scaling regime. Let P_B denote the blocking probability of the eager class. Then for any μ_ϵ,*

$$\lim_{t\to\infty} \frac{\Omega^{\mu_\epsilon}(t)}{t} = \nu_\tau \quad \textit{almost surely, where } \nu_\tau = \nu_\tau(P_B) := 1 - \rho_\epsilon(1 - P_B).$$

Proof: We couple the arrival processes corresponding to the eager class for different μ_ϵ as follows. Consider the sequence of ϵ inter-arrival times and job sizes $\{A_{\epsilon,n}^1, B_{\epsilon,n}^1\}_{n\geq 1}$ for $\mu_\epsilon = 1$. Note that $\mathbb{E}[A_{\epsilon,n}^1] = 1/\rho_\epsilon$ and $\mathbb{E}[B_{\epsilon,n}^1] = 1$ for

[7] However from Section 4 onwards, this assumption is implied by **B.4**.

all n. We define the sequence of inter arrival times and job sizes $\{A^{\mu_\epsilon}_{\epsilon,n}, B^{\mu_\epsilon}_{\epsilon,n}\}_{n\geq 1}$ for any general μ_ϵ as (without loss of generality):

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

Since the scheduling decisions (which include admission control) are independent of τ-customers (by **A.1**), one can identify a renewal process corresponding to the ϵ-system, such that each renewal cycle is composed of an ϵ-busy period and an ϵ-idle period. 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:

$$(1 - P_B) = \frac{\mathbb{E}[X_a]}{\mathbb{E}[X_{tot}]}. \tag{2}$$

The above expectations are finite under **A.0**, as seen below. By **A.2**, the admission control and scheduling policy for the eager class depend only upon the number of ϵ-customers in the system. Thus, if $X^{\mu_\epsilon}(t)$ represents the number of ϵ-customers in the system at time t and scale μ_ϵ, then

$$X^{\mu_\epsilon}(t) = X^1(\mu_\epsilon t). \tag{3}$$

It follows that X_a and X_{tot} are insensitive to the scale parameter μ_ϵ, which implies that P_B remains the same for all μ_ϵ (see (2)).

The length of a typical renewal cycle at scale μ_ϵ can be expressed, under our coupling model, as

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

By Wald's lemma, the expected length of a renewal cycle at scale $\mu_\epsilon = 1$ equals $\mathbb{E}[X_{tot}]/\lambda^1_\epsilon$. Thus by **A.0**, $\mathbb{E}[X_{tot}] < \infty$, which also implies that $\mathbb{E}[X_a] \leq \mathbb{E}[X_{tot}] < \infty$.

The total amount of service which the ϵ-customers utilize over a typical renewal cycle, for any admission control and scheduling policy, stochastically equals

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

whose expected value equals $\mathbb{E}[X_a]/\mu_\epsilon$, again by Wald's lemma[8]. Thus, invoking the RRT, we conclude that the rate corresponding to the ϵ-US process equals:

$$\begin{aligned} \lim_{t\to\infty} \frac{\Omega^{\mu_\epsilon}(t)}{t} &= \frac{\mathbb{E}[X_{tot}]/\lambda^{\mu_\epsilon}_\epsilon - \mathbb{E}[X_a]/\mu_\epsilon}{\mathbb{E}[X_{tot}]/\lambda^{\mu_\epsilon}_\epsilon} \\ &= 1 - \rho_\epsilon \frac{\mathbb{E}[X_a]}{\mathbb{E}[X_{tot}]} = 1 - \rho_\epsilon(1 - P_B) = \nu_\tau \quad \text{(w.p.1)} \end{aligned} \tag{4}$$

[8] Because, $\mathbb{E}[B^{\mu_\epsilon}_{\epsilon,n}; X_a < n] = \mathbb{E}[B^{\mu_\epsilon}_{\epsilon,n}]\mathbb{E}[X_a < n]$.

The above is true for any μ_ϵ and for any admission control and scheduling policy for the eager class. ■

Note that the ϵ-US process $\Omega^{\mu_\epsilon}(t)$ at any fixed time t is not scale invariant. However, Lemma 1 states that its asymptotic (in time) growth rate is scale invariant, i.e., the same for all μ_ϵ. Moreover, this growth rate ν_τ depends only on the blocking probability of the eager class, and not on the admission control and scheduling policy that produced that blocking probability.

The main result of this section is that as $\mu_\epsilon \to \infty$ under the SFJ limit, the ϵ-US process $\Omega^{\mu_\epsilon}(t)$ converges to a linear curve with slope ν_τ. In other words, under the SFJ limit, the ϵ-US process grows at a steady rate, that is determined purely by the blocking probability of the eager class.

Theorem 1 [**ϵ-US process**] *Assume* **A.0-2**. *As $\mu_\epsilon \to \infty$, $\Omega^{\mu_\epsilon}(t)$ converges to a constant slope curve $\nu_\tau t$ almost surely, uniformly over bounded intervals:*

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

and for any initial ϵ-state.

Let $\Upsilon^{\mu_\epsilon} := \inf\{t : \Omega^{\mu_\epsilon}(t) \geq B\}$, represent the time taken to derive B units of service using the ϵ-US process, where the random variable B is independent of the ϵ arrival process. Then Υ^{μ_ϵ} converges to B/ν_τ almost surely as $\mu_\epsilon \to \infty$ for any initial ϵ-state.

Proof: Provided in Appendix A. ■

Theorem 1 states that under the SFJ limit, the ϵ-US process has the same (linear) form, for any admission control and scheduling policy for the eager class that produces the blocking probability P_B. This gives us our first pseudo-conservation law – a relationship between the blocking probability of the eager class and its unused service process (under the SFJ limit), that does not depend on the specifics of the ϵ scheduling policy beyond the resulting blocking probability. Theorem 1 also proves pseudo-conservation of the time required to accumulate B units of service from the ϵ-US process. Under the SFJ limit, this time converges to B/ν_τ almost surely.

In the following section, we explore the performance experienced by the tolerant class, when served by the ϵ-US process.

4 Pseudo-conservation laws for τ performance metrics

In the previous section, we analysed the service process seen by the tolerant class under the SFJ limit. In this section, we focus on the performance experienced by the tolerant class under the same limit. Specifically, we show that various performance measures of the τ-system utilizing the ϵ-US process converge to values corresponding to a steady service process (with rate ν_τ) under the SFJ limit. This implies a pseudo-conservation that characterizes the performance measures of the tolerant class in terms of the blocking probability of the eager class, robust to the specifics of the ϵ-policy that produced

that blocking probability. Moreover, our pseudo-conservation laws hold for any scheduling policy for the tolerant class that is non-preemptive (across τ-jobs) and blind (to τ service requirements).

It is important to note that performance evaluation of the tolerant class is challenging in the pre-limit, since this class experiences an exogenous, time varying, service process. Even in the simplest case where this service process is Markovian, performance measures cannot be expressed in closed form (Mahabhashyam et al. 2005). Thus, the SFJ limit is crucial for analytical tractability of performance measures of the tolerant class.

The results in this section require additional assumptions on the workload processes of the two classes.

B.4 The eager class has Poisson arrivals (of rate $\lambda_\epsilon^{\mu_\epsilon}$). Moreover, sizes $B_\epsilon^{\mu_\epsilon}$ of ϵ-jobs are either exponentially distributed or are bounded by $\bar{b}/\mu_\epsilon$ almost surely.

B.5 The tolerant class has Poisson arrivals with rate λ_τ. Job sizes of the tolerant class are IID, with B_τ denoting a generic job size. Moreover, $\rho_\tau := \frac{\lambda_\tau}{\mu_\tau} < \nu_\tau(P_B)$, where $\mu_\tau := 1/\mathbb{E}[B_\tau]$ (for stability of the τ queue).

By Lemma 6 of Appendix B, **B.4** implies **A.0**. For convergence in expectation of performance measures, we will also require the following assumption (which implies certain uniform integrability conditions, in order to ensure convergence of the required second moments).

B.6 $E[B_\tau^3] < \infty$.

Main Results

We first state our results, the different pseudo-conservation laws. The proofs are discussed later. We begin with the pseudo-conservation law for the stationary number of tolerant customers in the system. Specifically, we show that under the SFJ limit, the stationary number of tolerant jobs in the system converges in distribution as well as in expectation to the stationary number of jobs in a dedicated M/G/1 queue operating at speed ν_τ. As before, note that under the SFJ limit, the steady state number of tolerant jobs in the system depends only on the blocking probability of the eager class, and not on the ϵ-policy that produced that blocking probability.

Theorem 2 [Number in system] *Assume* **A.1-3** *and* **B.1-5**. *As $\mu_\epsilon \to \infty$ under the SFJ scaling, the steady state number of τ-jobs in the system converges in distribution to the steady state number of jobs in an M/G/1 queue with arrival rate λ_τ, service times B_τ and server speed ν_τ. If we also assume* **B.6**, *then we also have convergence of the expectations and the limit equals*

$$\frac{\rho_\tau}{\nu_\tau} + \frac{\lambda_\tau^2 E[B_\tau^2]}{2\nu_\tau^2(1-\rho_\tau/\nu_\tau)}. \qquad \blacksquare$$

It is important to note that Theorem 2 is not a direct consequence of Theorem 1, which characterizes the service process of the tolerant stream under the SFJ limit, as well as the time required to serve a single τ-job. In particular, the time-varying service process introduces dependencies across the service times of different tolerant jobs. Thus, part of the challenge in proving Theorem 2 is in showing that these dependencies get washed out under the SFJ limit.

Our second result is pseudo-conservation of the expected stationary sojourn time, which is a corollary of the first result by Little's law. Note that the sojourn time of a job is the total time spent in the system by the job.

Corollary 1 [Sojourn time] *Assume* **A.1-3** *and* **B.1-6**. *As $\mu_\epsilon \to \infty$ under the SFJ scaling, the steady state sojourn time of τ-jobs converges in expectation to the steady state sojourn time of the limit M/G/1 queue of Theorem 2, i.e.,*

$$\frac{1}{\mu_\tau \nu_\tau} + \frac{\lambda_\tau E[B_\tau^2]}{2\nu_\tau^2(1-\rho_\tau/\nu_\tau)}. \qquad \blacksquare$$

Our final result in this section establishes a pseudo-conservation of the steady state tolerant workload. Note that the workload in the system is defined as the total amount of unfinished work, i.e., the sum total of job sizes of all the waiting customers plus the residual service requirement of the customer in service (if any).

We require two additional assumptions.

W.1 The sizes of τ-jobs are light-tailed, i.e., there exists an $\bar{a} > 1$ such that $\mathbb{E}[\bar{a}^{B_\tau}] < \infty$.

W.2 The embedded Markov chain $\{\zeta_n\}_n$ corresponding to ϵ-queue with $\mu_\epsilon = 1$, obtained by sampling at ϵ-arrival/departure epochs, is ergodic with stationary distribution π^*. Further with $\mathbf{0}$ representing the empty ϵ-state,

$$P(\zeta_n = \mathbf{0}) \to \pi^*(\mathbf{0}) > 0 \text{ as } n \to \infty,$$

and the rate of convergence is uniform across all the initial states.

With exponentially sized ϵ-jobs, the state space describing the ϵ-queue is finite (see Appendix C for further details) and hence the uniform rate of convergence required by **W.2** is readily achieved. With bounded ϵ-jobs, we have a compact state space, which also makes **W.2** non-restrictive (see, for example, the policies in Section 5).

Theorem 3 [Workload] *Assume* **A.1-3**, **B.1-5**, *and* **W.1-2**. *As $\mu_\epsilon \to \infty$ under the SFJ scaling, the steady state workload of the τ-class converges in distribution as well as in expectation to the steady state workload of the limit M/G/1 queue of Theorem 2. The limit of the expected workload equals*

$$\frac{\lambda_\tau E[B_\tau^2]}{2\nu_\tau(1-\rho_\tau/\nu_\tau)}. \qquad \blacksquare$$

Theorem 2 is proved in Appendix B, and Theorem 3 is proved in Appendix C. In the remainder of this section, we provide a sketch of the proof of Theorem 2.

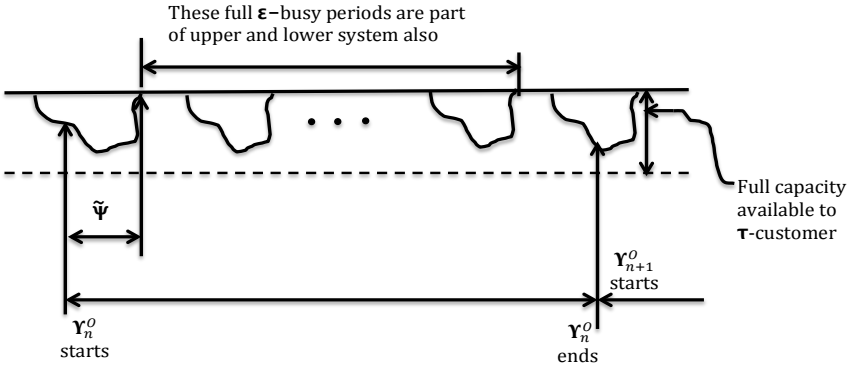

Fig. 1: A typical Υ^O_n spanning over several ϵ-busy periods, starting within one.

Sketch of the proof of Theorem 2

The steady number of τ-customers immediately after a τ-departure matches (in distribution as well as in expectation) the steady state number of τ-customers as seen by a τ-arrival, which by PASTA also matches (in distribution as well as in expectation) the steady state number of τ-customers in the system. We thus study the convergence of X_n, the number of τ-customers in system immediately after n-th τ-departure. For any serial, non-anticipative service policy for τ customers (i.e., under **B.1-3**), the evolution of this number can be represented in terms of ϵ-US process(es) as below:

$$X_{n+1} = (X_n - 1)^+ + A_{n+1}(\Upsilon_n^{\mu_\epsilon}), \text{ with} \tag{5}$$

$$\Upsilon_n^{\mu_\epsilon} = \inf\{t : \Omega_n^{\mu_\epsilon}(t) \geq B_{\tau,n}\}, \text{ where} \tag{6}$$

- $\Omega_n^{\mu_\epsilon}(\boldsymbol{.})$ is the ϵ-US process associated with n-th τ-customer, i.e., $\Omega_n^{\mu_\epsilon}(t)$ for any t equals the total amount of server capacity available for the n-th τ-customer for time duration t after its service has started[9],
- $\Upsilon_n^{\mu_\epsilon}$, as defined in Theorem 1, is the time taken to serve job requirement of size $B_{\tau,n}$ using the n-th ϵ-US process, and
- $A_{n+1}(\Upsilon_n^{\mu_\epsilon})$ is the number of τ-arrivals during the n-th τ-service time $\Upsilon_n^{\mu_\epsilon}$.

Further $\Omega_n^{\mu_\epsilon}(t)$, $\Upsilon_n^{\mu_\epsilon}$ and hence A_{n+1}, X_{n+1} depend on state of the ϵ-system at the instance the n-th τ-service starts. This state can be characterized by Y_n the ϵ-number in system, and $\mathbf{R}_n^s$ the vector of residual service times (of dimension K) of the ϵ-customers present, immediately after $(n-1)$-th τ-customer departure. Note that $\mathbf{R}_n^s$, has only Y_n meaningful entries, the rest are taken to be zero. Indeed, $(Y_n, \mathbf{R}_n^s)$ completely determines the n-th US process $\Omega_n^{\mu_\epsilon}(\boldsymbol{.})$ utilized by the nth τ-job.

[9] It is important to note here that we have one ϵ-US process for each τ-customer, which starts at the instance the n-th τ-service starts. This is done to simplify the notations.

Define $Z_n := (X_n, Y_n, \mathbf{R}_n^s)$, and note that $\{Z_n\}_{n\geq 1}$ is a Markov chain. For exponentially distributed ϵ-jobs, $Z_n := (X_n, Y_n)$ is already a Markov chain, as the residual job requirements are again exponentially distributed. This special case is an example of a queueing system with Markov modulated service process. The weak convergence of the sequence $\{X_n\}$ to its stationary distribution is proved for this case in Mahabhashyam et al. (2005). For (generally distributed) bounded ϵ-jobs, by Theorem 6 in Appendix B, $\{Z_n\}$ and hence the sequence $\{X_n\}$ is stationary and ergodic.

The idea of our proof is to approximate the original τ system with an equivalent $M/G/1$ system. However the sequence of service times, the 'actual times' taken to serve τ-customers, $\{\Upsilon_n^{\mu_\epsilon}\}_{n\geq 1}$ is not IID. Note that in general, the service of a tolerant customer may start in the middle of an ϵ-busy cycle, and may also end in a (potentially different) ϵ-busy cycle, which results in correlations between consequent service times (see Figure 1). Hence, we construct two fictitious M/G/1 queues, an upper system and a lower system, such that both these systems see the same τ-arrivals as our original system. However, the service times $\{\Upsilon_n^U\}_{n\geq 1}$ and $\{\Upsilon_{n\geq 1}^L\}_n$ in these systems are defined such that the service times of the original system are sandwiched between these two almost surely:

$$\Upsilon_n^L \leq \Upsilon_n^O := \Upsilon_n^{\mu_\epsilon} \leq \Upsilon_n^U \text{ for all } n.$$

- To achieve this, for any n we construct (using certain coupling rules a busy period $\Theta_n^{\mu_\epsilon}$ of a fictitious infinite buffer and infinite server queuing system in Lemma 6 of Appendix B, which would almost surely (or stochastically) upper bound the residual ϵ-busy cycle starting the n-th τ-service.
- Zero service is offered during Θ^{μ_ϵ} for τ-customer in the upper system, while service at full capacity is offered in the lower system. Then the service is continued identically in all three systems, until service completion in the original system.
- The service in lower system ends before that of original system, while the upper system continues with independent copies of ϵ-busy periods. Further, the residual of the ϵ-busy period in which the original customer departs, is also continued with independent copies of the required random variables.

We thus dominate the number of τ-customers in the original system with the number in two M/G/1 queues in either direction, where both the dominating queues are served by the same τ-service discipline. We further show that the performance metrics of the two dominating queues converge to the same limit under the SFJ scaling, which equals the corresponding performance metrics in an M/G/1 system with constantly reduced service speed ν_τ. The details of the proof are provided in Appendix B.

5 Achievable region and Complete families

The achievable region for an n-class system with common resources is the set of relevant performance vectors $(pm_1, \cdots, pm_n)$, obtained by all possible

scheduling policies (e.g., Shanthikumar et al. (1992); Federgruen et al. (1988); D. Bertsimas et al. (1994)). The two classes of users considered in this paper have different goals and hence naturally have different qualities of service (QoS) metrics. For the eager class, steady state blocking probability P_B is the natural metric. For the tolerant class, the performance metric could be, for example, the steady state average sojourn time, the steady state probability of there being m or more τ-jobs in the system, the steady state average τ-workload, etc.

For simplicity, we will take the steady state average sojourn time to be the performance metric for the tolerant class. In this case, the achievable region is the set of all possible pairs of blocking probability of the eager class and expected sojourn time of the tolerant class, i.e.,

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

If $\rho_\epsilon > 1$, it is not possible to have $P_B = 0$. More generally, it is necessary that $(1 - P_B)\rho_\epsilon < 1$ or equivalently, $P_B > 1 - 1/\rho_\epsilon$. Similarly it follows from Theorem 1 that one requires $\lambda_\tau \mathbb{E}[B_\tau] < \nu_\tau(P_B)$ for τ-stability; see **B.5**. Hence, using Corollary 1, the achievable region under SFJ limit is given by:

$$\boxed{\mathcal{A}_{ps}^{hetero} = \left\{ \left(p_B, \quad \frac{1}{\mu_\tau \nu_\tau(p_B)} + \frac{\lambda_\tau E[B_\tau^2]}{2\nu_\tau^2(1 - \rho_\tau/\nu_\tau(p_B))} \right) \middle| \lambda_\tau < \mu_\tau \nu_\tau(p_B), \right.}$$
$$\boxed{\left. [1 - 1/\rho_\epsilon]_+ < p_B \leq 1 \right\}}$$

Here, $[x]_+ := \max(0, x)$. This is actually a sub-achievable region because it is obtained by the sub-class of schedulers such that the ϵ-scheduling rules are oblivious to τ-state and also satisfy assumptions **A.1-3** and **B.1-6**. We call this the partially static achievable region. If B_τ is exponentially distributed with mean $1/\mu_\tau$, this region is given by:

$$\mathcal{A}_{ps}^{hetero} = \left\{ \left(p_B, \frac{1}{\mu_\tau \nu_\tau(p_B) - \lambda_\tau} \right) \middle| \lambda_\tau < \mu_\tau \nu_\tau(p_B), \ [1 - 1/\rho_\epsilon]_+ < p_B \leq 1 \right\} \quad (7)$$

Complete families

We now demonstrate two families of schedulers which achieve all the points of the above (partially static) achievable region. Such families are generally referred to as *complete families.* The performance of the following complete families is analysed under the SFJ limit using direct methods (i.e., without invoking the general pseudo-conservation laws proved in this paper) in Kavitha et al. (2017a); see also the extended version Kavitha et al. (2017b). These papers additionally characterize the rate of convergence of the performance measures under the SFJ limit.

In the following, we assume exponential service times for both τ as well as ϵ customers.

5.1 Processor sharing PS-(p, K) schedulers

Eager arrivals are admitted into the system with probability p, so long as there are strictly less than K ϵ-jobs in the system at that time. Thus, at most K ϵ-jobs can be served in parallel. The ϵ-jobs present in the system are served in a processor sharing fashion using the complete server capacity, i.e., if there are $1 \leq l \leq K$ ϵ-jobs in the system, each receives service at rate $1/l$. Under this policy, τ-jobs get service only if there are no ϵ-jobs in the system. We refer to this as the $\beta^{PS}_{p,K}$ scheduling policy.

The performance of the $\beta^{PS}_{p,K}$ policy under the SFJ limit can be derived using Corollary 1:

Lemma 2 *The PS achievable region, the possible pairs of* $\{(P_B(p),\ \mathbb{E}[S](p))\}$ *achieved by PS schedulers* $\{(p, K)\}$ *under SFJ limit is given by*

$$\mathcal{A}_{PS} = \left\{ \left((1-p) + p(\rho_{\epsilon,p})^K \nu^{PS}_K,\ \frac{1}{(\mu_\tau \nu^{PS}_K - \lambda\tau)} \right) \Big| \lambda_\tau < \nu^{PS}_K \mu_\tau,\ 0 \leq p \leq 1, K < \infty \right\}$$

with $\nu^{PS}_K := \left(\sum_{j=0}^{K} \rho^j_{\epsilon,p} \right)^{-1}$ *and* $\rho_{\epsilon,p} := \rho_\epsilon p$.

Proof: It is straightforward to characterize the blocking probability of the eager class under the $\beta^{PS}_{p,K}$ policy (see Kavitha et al. (2017a)). Having done this, the statement of the lemma follows easily from Corollary 1. We omit the details. ■

The following lemma proves that the PS family is a complete family.

Lemma 3 *PS schedulers* $\mathcal{F}^{PS} := \left\{\beta^{PS}_{p,K},\ 0 \leq p \leq 1, K < \infty\right\}$ *are a complete family, i.e.,*

$$\mathcal{A}_{PS} = \mathcal{A}^{hetero}_{ps}.$$

Proof: Consider first the case $\rho_\epsilon \leq 1$. For this case, as K increases to ∞, it can be verified that the blocking probability $P^{PS}_B(1)$ with $p = 1$ (see Lemma 2) decreases to zero. Also it is easy to verify that the function, $p \mapsto P^{PS}_B(p)$, is continuous in p for any K. Thus by the intermediate value theorem, all the points of the achievable region can be achieved by PS schedulers.

When $\rho_\epsilon > 1$, it is easy to verify as $K \to \infty$ that (see Lemma 2):

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

Thus $P^{PS}_B(1) \to 1 - 1/\rho_\epsilon$ and all points of $\mathcal{A}^{hetero}_{static}$ are achieved again by intermediate value theorem. ■

It is important to note here that the completeness of the above class of schedulers is achieved by letting $K \to \infty$. However, a larger K also implies a smaller service rate for ϵ-customers.

5.2 Capacity Division (CD) Policies

Under the PS policy, an admitted ϵ-customer pre-empts any ongoing τ-service and the entire system capacity is transferred to the ϵ-customer. In contrast, under the CD policy, only a fixed fraction of the server capacity is allocated to each admitted ϵ-customer. The tolerant class uses the residual service capacity, if any.

The details of the policy are as follows. Eager arrivals are admitted into the system with probability p, so long as there are strictly less than K ϵ-jobs in the system at that time. Thus, at most K ϵ-jobs can co-exist in the system at any time. Moreover, each ϵ-customer is served at rate $1/K$. Thus, if there are $1 \leq l \leq K$ number of ϵ-customers receiving the service, then the tolerant class is served at rate $(K-l)/K$.

The CD-achievable region can also be derived using Corollary 1:

Lemma 4 *The CD achievable region, the possible pairs of $\{(P_B(p),\ \mathbb{E}[S](p))\}$ achieved by CD schedulers $\{(p, K)\}$ under SFJ limit is given by*

$$\mathcal{A}_{CD} = \left\{ \left((1-p) + p\frac{(K\rho_{\epsilon,p})^K}{K!\check{a}_0},\ \frac{1}{(\mu_\tau \nu_K^{CD} - \lambda_\tau)} \right) : \lambda_\tau < \mu_\tau \nu_K^{CD},\ 0 \leq p \leq 1 \right\},\ \textit{with}$$
$$\nu_K^{CD} = \frac{\eta}{\check{a}_0},\quad \check{a}_0 := \sum_{j=0}^{K} \frac{(K\rho_{\epsilon,p})^j}{j!} \quad \textit{and}\ \eta := \sum_{j=0}^{K-1} \frac{(K\rho_{\epsilon,p})^j}{j!}\frac{K-j}{K}.$$

Proof: As before, having characterized the blocking probability of the eager class (see Kavitha et al. (2017a)), the statement of the lemma follows easily from Corollary 1. We omit the details. ■

This family is also a complete family of schedulers:

Lemma 5 *CD schedulers $\mathcal{F}^{CD} := \left\{\beta_{p,K}^{CD},\ 0 \leq p \leq 1, K < \infty\right\}$ are a complete family, i.e.,*

$$\mathcal{A}_{CD} = \mathcal{A}_{ps}^{hetero}.$$

Proof: This result can be proved using the same line of arguments as in the proof of Lemma 3. We omit the details (see Kavitha et al. (2017a,b)). ■

6 Eager customers with limited patience

While the previous sections considered eager customers with zero patience to wait, we now consider the case where ϵ-customers have partial/limited patience (referred to as $p\epsilon$). For example customers may enter the system with probability e_n if the queue length is n as in balking or may wait at maximum for a random time distributed as Γ as in reneging (Whitt (1999)). As before, the $p\epsilon$-system is assumed to be high priority and operates oblivious to the τ-system, with the τ-system utilizing the service capacity left unused by the eager class. In this section, we obtain pseudo-conservation laws for the system with $p\epsilon$ and τ customers under SFJ limit. Towards this, we modify Assumption **A.3** as below:

A.3′ There exists a $K \in \mathbb{N}$ such that, at most K $p\epsilon$-jobs exist in the system. So long as the $p\epsilon$-system is non-empty, at least one $p\epsilon$-job is served. There exists $c_{min} > 0$ such that once service beings for an $p\epsilon$-job, its service rate is at least c_{min} until service completion.

Basically the above assumption ensures that an eager busy period ends, only when the eager queue is empty and then the analysis of the previous sections is applicable again. We obtain performance analysis of this heterogeneous queueing system with $p\epsilon$ customers under one of the following additional assumptions:

A.4(a) The $p\epsilon$ service requirements are IID and there exists an upper bound $\tilde{b}$ on the time spent by a $p\epsilon$ customer in the system, which scales inverse linearly with μ_ϵ (i.e., $\tilde{b}^{\mu_\epsilon} = \tilde{b}^1/\mu_\epsilon$). If reneging is allowed, then the reneging (impatience) times of $p\epsilon$-customers are IID, and also scale inverse linearly with μ_ϵ.

A.4(b) Exponential IID $p\epsilon$-job sizes and IID reneging times which are exponentially distributed with rate α. Further α also scales linearly with μ_ϵ, i.e., $\alpha^{\mu_\epsilon} = \alpha^1 \mu_\epsilon$.

A.4(c) Exponential IID $p\epsilon$-jobs and at any time, upto $K_1 \leq K$ number of customers are served at a fixed rate. In other words, ϵ-customers are served exactly like in an $M/M/K_1/K$ queue.

It is important to note that for the partially patient eager class, the blocking probability P_B, defined as the long run fraction of customers denied service, including those blocked on arrival (as in balking) and those who abandon the queue while waiting (as in reneging). [10] Our main result is the following.

Theorem 4 *Assume* **A.1-2**, **B.1-6.**. *Assume* **A.3′** *in place of* **A.3** *and further assume either* **A.4(a)** *or* **A.4(b)** *or* **A.4(c)**. *Then the conclusions of Theorem 2 and Corollary 1 are true.*

Proof: It is straightforward to prove Lemma 1 when there is no reneging (satisfying **A.4(a)** or **A.4(c)**). For reneging models (satisfying **A.4(a)** or **A.4(b)**), since we assume that reneging times scale inverse linearly with μ_ϵ, we can couple the reneging times such that Equation (3) continues to hold, allowing us to prove Lemma 1.

For the case with partially eager customers, few $p\epsilon$-customers would wait for service, while others would be in service. We would now require additional component in the Markov chains considered in Section 4, Y_n gets split into two components $Y_{s,n}, Y_{w,n}$ which respectively represent the number of $p\epsilon$-customers in service and the number of waiting $p\epsilon$-customers. Vector $\mathbf{R}_n^s$ (of size K) as before represents the vector of residual service times of the $p\epsilon$-customers in service out of which only the first $Y_{s,n}$ of them have non-zero components. It is again easy to verify that the Theorems 6 and 2 go through for this case, once Lemma 6 of Appendix B is true. In Lemma 11 of Appendix D we prove this is true, even for $p\epsilon$ customers, under the given hypotheses. ■

[10] Note that the models considered here allow for reneging/abandonment while waiting, but not during service.

A wide variety of models are captured by our assumptions. Consider the case of balking where $e_n = 0$ for all $n > K$ and say at maximum $K_1 \leq K$ customers are served in parallel each with rate $1/K_1$. If the $p\epsilon$-job sizes are generally distributed and bounded as in **B.4**, then maximum time spent by any $p\epsilon$-customer in the system equals, $(K - K_1 + 1)\bar{b}/(\mu_\epsilon K_1)$. Hence the assumption **A.4 (a)** is satisfied and the pseudo-conservation law is applicable. The blocking probability P_B of such a system is well understood (e.g., see Ancker et al. (1963); Whitt (1999)), and so the τ-performance can be readily estimated in SFJ limit. The analysis is also applicable if the $p\epsilon$-jobs are exponentially distributed; in this case, **A.4(b)** is satisfied. When one considers reneging and balking together with $e_n = 0$ for all $n > K$ then assumption **A.4.(b)** or **A.4(a)** is satisfied depending upon the nature of $p\epsilon$-jobs.

An example

We consider balking and reneging model with mutli-server mode as in Model 1 of Whitt (1999). The system serves at maximum K_1 $p\epsilon$-customers in parallel. The limited patience eager customers are discouraged if they have to wait and hence would balk (i.e., depart without service) with probability δ, when (waiting) queue size is bigger than 0. The system allows at maximum $K - K_1$ customers to wait for service. Even the customers that entered the waiting room, can become impatient if service is not offered quickly. They depart after exponentially distributed (impatience) time with parameter α, if their service has not yet started. This model is easy to analyze (e.g., Whitt (1999)) and the probability of blocking equals:

$$P_B(\mu_\epsilon, \rho_\epsilon, \alpha) \;=\; \frac{\sum_{k=K_1+1}^{K-1} \pi_k(1 - \vartheta_k + \delta\vartheta_k) + \pi_K}{\sum_k \pi_k} \text{ with} \tag{8}$$

$$\vartheta_k = \left(1 - \frac{k\alpha}{K_1\mu_\epsilon + k\alpha}\right)\left(1 - \frac{(k-1)\alpha}{K_1\mu_\epsilon + (k-1)\alpha}\right)\cdots\left(1 - \frac{\alpha}{K_1\mu_\epsilon + \alpha}\right) \text{ and}$$

$$\pi_k = \begin{cases} \frac{\rho_\epsilon^k}{k!}\left(\frac{\rho_\epsilon^{K_1}}{K_1!}\right)^{-1} & \text{if } k \leq K_1 \\ \frac{\rho_\epsilon^k(1-\delta)^k}{K_1^k}\left(\frac{\rho_\epsilon^{K_1}(1-\delta)^{K_1}}{K_1^{K_1}}\right)^{-1} \frac{1}{\left(1+\frac{\alpha}{K_1\mu_\epsilon}\right)\left(1+\frac{2\alpha}{K_1\mu_\epsilon}\right)\cdots\left(1+\frac{(k-K_1)\alpha}{K_1\mu_\epsilon}\right)} & \text{if } k > K_1. \end{cases}$$

The above is the expression if each $p\epsilon$ customer is served at unit rate. But say the system allocates only fraction ξ to $p\epsilon$ customers, then each of them is served at ξ/K_1. In that case we need to replace μ_ϵ with $\mu_\epsilon\xi/K_1$ and ρ_ϵ with $\rho_\epsilon K_1/\xi$ and further α by $\alpha\mu_\epsilon$ as demanded by Assumption **A.4 (b)**. Thus the probability of blocking with scale parameter μ_ϵ is given by

$$P_B(\mu_\epsilon\xi/K_1, \rho_\epsilon K_1/\xi, \alpha\mu_\epsilon). \tag{9}$$

Clearly this is independent of μ_ϵ. Now consider that the system also supports long job tolerant customers using the $p\epsilon$-US process. Because of our Psuedo-conservation laws, for example, the stationary expected workload in the τ-system (with Poisson arrivals and exponential jobs) under SFJ limit equals

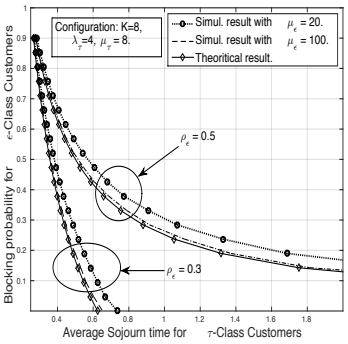

Fig. 2: Achievable region: Simulated and Theoretical results

		Simulation		Theoretical	
p	μ_ϵ	P_B^{PS}	$\mathbb{E}[S_\tau]$	P_B^{PS}	$\mathbb{E}[S_\tau]$
0	20	1.000	0.250	1.000	0.250
	80	1.000	0.250	1.000	0.250
	120	1.000	0.250	1.000	0.250
0.25	20	0.750	0.353	0.750	0.333
	80	0.750	0.339	0.750	0.333
	120	0.750	0.337	0.750	0.333
0.50	20	0.500	0.567	0.500	0.500
	80	0.500	0.515	0.500	0.500
	120	0.500	0.511	0.500	0.500
0.75	20	0.250	1.238	0.250	0.999
	80	0.250	1.042	0.250	0.999
	120	0.250	1.037	0.250	0.999
1	20	0.002	110.22	0.002	127.8
	80	0.002	112.12	0.002	127.8
	120	0.002	127.20	0.002	127.8

Table 1: PS model: Comparison of Simulated, theoretical metrics

(with P_B given by (9)):

$$\mathbb{E}[U_*] = \frac{\lambda_\tau}{\mu_\tau(\mu_\tau\nu_\tau - \lambda_\tau)},\ \nu_\tau = (1 - \rho_\epsilon(1 - P_B)).$$

The system has to provide service to both $p\epsilon$ as well τ customers, it may chose an optimal system configuration (e.g., tune parameters ξ, K_1, K) such that the required trade-off between the two system performances P_B and $\mathbb{E}[U_*]$ is achieved/optimized.

This kind of a system is motivated by super market system with express service counter for short job customers. This also models a cellular system that supports data-voice calls. This system supports at maximum K_1 voice calls, which are provided in total ξ fraction of the serving capacity of the serving base station and where $K - K_1$ voice calls can be queued-up. One does not drop a voice call immediately: a voice call may decide to wait (with probability δ) and when it decides to wait, it might wait for a brief period (exponential waiting time with parameter $\alpha\mu_\epsilon$) before dropping. The data calls use the dedicated $(1 - \xi)$ fraction as well as the server capacity left unused by the voice customers.

7 Numerical examples

7.1 Accuracy of fluid approximation

We conduct Monte-Carlo simulations to estimate the performance of both the scheduler families proposed in Section 5. We 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 estimate the blocking probability and expected sojourn time for ϵ and τ-agents respectively, using sample means, for different values of (p, K).

In Figure 2, we compare the theoretical expressions with the ones estimated using Monte-Carlo simulations for the PS policy. We consider two different values of ρ_ϵ. We notice negligible difference between the theoretical and simulated values when $\mu_\epsilon = 100$. However even with $\mu_\epsilon = 20$, the difference is about 10-12% for most of the cases.

We consider another example of the PS policy in Table 1 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 estimated results are close to the theoretical results. The performance is very close to that of theoretical SFJ approximation, for values of μ_ϵ greater than 120. For $\mu_\epsilon = 80$, the error is at most5%. Even at values as low as 20, the simulator performance is within 10% of the theoretical values for most 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.

The *Achievable region* is also plotted in Figure 2 for different values of ρ_ϵ. Towards this, we plot $\mathbb{E}^{PS}[S_\tau(p)]$ versus $P_B^{PS}(p)$, for $p \in \{i\delta : 0 \leq i \leq 1/\delta\}$ 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 amount of shift.

7.2 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 $\lambda_\tau = 5.6$, $\mu_\tau = 8$, $K = 3$ or 5 and $\rho_\epsilon = 0.9/K$ (i.e, $\rho_\epsilon = 0.3$ when $K = 3$ and $\rho_\epsilon = 0.18$ when $K = 5$).

In Figure 3, we plot the achievable region for both the policies, i.e, we plot $\mathbb{E}[S_\tau(p)]$ versus $P_B(p)$, for different p. In Figures 4-5, we plot the performance measures $P_B(p)$ and $\mathbb{E}[S_\tau(p)]$ respectively versus p with $K = 3$. From Figure 3, the two achievable (sub) regions overlap, however we observe from the Figures 4-5 that the performance measures of the two policies 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 3. This observation is precisely the pseudo-conservation law given by Theorem 2. 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,

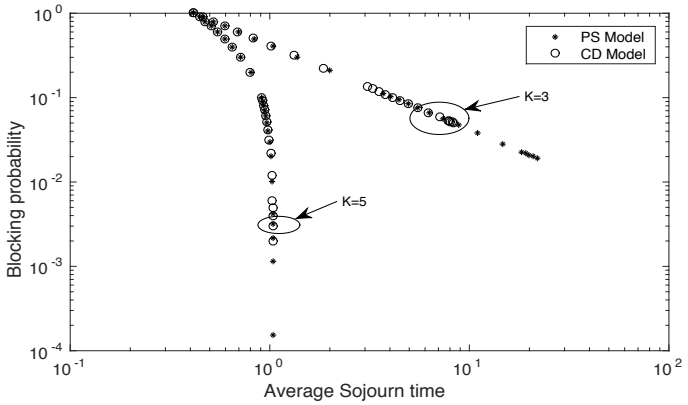

Fig. 3: Achievable regions $\mathcal{A}_{CD}$, $\mathcal{A}_{PS}$: P_B versus $\mathbb{E}[S_\tau]$ for different ρ_ϵ, $\rho_\epsilon = 0.9/K$

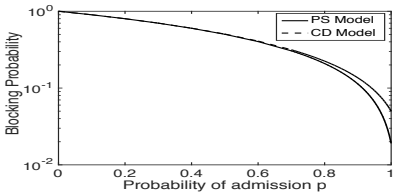

Fig. 4: P_B versus p

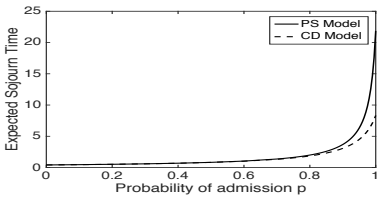

Fig. 5: $\mathbb{E}[S_\tau]$ versus p

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

is greater than that with the PS policy. In Figure 3 the best P_B with CD and PS 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.

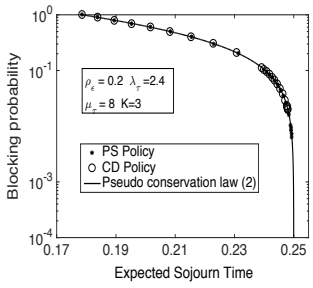

Fig. 6: Static Achievable region

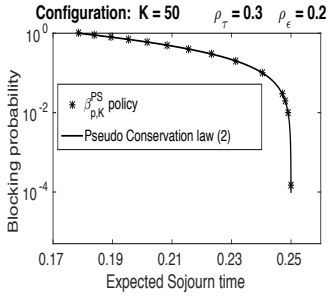

Fig. 7: Completeness of $\mathcal{F}^{PS}$

In Figure 6, we plot the pseudo-conservation law (7). 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 (7). For the same configuration we plot performance of PS policies with a bigger $K = 50$, in Figure 7. 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 3 and 6).

8 Concluding remarks

We analysed a queueing system serving (higher priority) eager customers who require service immediately upon arrival, and (lower priority) tolerant customers who can wait for service. For the class of scheduling policies that admit and serve eager customers oblivious to the state of the tolerant queue, we derive pseudo-conservation laws relating the blocking probability of the eager class and the performance of the tolerant class under a (partially) fluid limit. We also demonstrated two complete families of scheduling policies, which span the space of all performance vectors achievable by the (partially static) class of scheduling policies under consideration. We further extended the pseudo-conservation laws to eager jobs with limited patience, which captures a wide range of balking and/or reneging models.

The main feature of our models is the co-existence of a lossy sub-system of high-priority jobs and a non-lossy sub-system of delay-tolerant, lower-priority jobs. This distinction also implies that the relevant performance metrics of the two classes are different, e.g., blocking probability for the lossy class and average response time for the non-lossy class. To the best of our knowledge, the achievable region for such heterogeneous multi-class systems has not been studied in the literature. While performance evaluation of such systems is highly challenging and policy-specific, we identify a certain partial fluid limit under which the achievable region is policy-independent and has a closed form characterization; albeit with certain restrictions on the class of scheduling policies.

This work motivates generalizations along several dimensions. Immediate generalizations include extending the psuedo-conservation laws derived here to the distribution of τ-sojourn time, and to pre-emption based policies for τ-jobs, like processor sharing and SRPT. One could also analyse the achievable region under a richer class of dynamic policies, which schedule both classes based on the complete system state. Some of our preliminary investigations show that this region is a strict superset of the partially static achievable region analysed

here. Another interesting extension would be to multi-server (and therefore non-work-conserving) settings. Finally, specializing these models to capture particular application scenarios, including supermarkets, cognitive radio, and cloud computing environments, present avenues for future work.

References

Veeraruna Kavitha and Raman Sinha, 'Achievable region with impatient customers', Proceedings of Valuetools 2017.

Veeraruna Kavitha and Raman Sinha, 'Queuing with Heterogeneous Users: Block Probability and Sojourn times' Veeraruna Kavitha, Raman Kumar Sinha', arXiv preprint, http://adsabs.harvard.edu/abs/2017arXiv170906593K.

Moshe Shaked, and J. George Shanthikumar. 'Stochastic orders', Springer Science & Business Media, 2007.

W. Feller. 'An introduction to probability theory and its applications: Volume I,' John Wiley & Sons, 1968.

W. Feller. 'An introduction to probability theory and its applications: Volume II', John Wiley & Sons, 1972.

Sesia, S., Baker, M., and Toufik, I. 'LTE-the UMTS long term evolution: from theory to practice', John Wiley & Sons, 2011.

S. P. Meyn and R. L. Tweedie. 'Markov chains and stochastic stability', Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

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

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

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, 2004.

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.

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

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

Baxendale, Peter H. 'Renewal theory and computable convergence rates for geometrically ergodic Markov chains', The Annals of Applied Probability, 15.1B (2005): 700-738.

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.

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.

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.

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.

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.

Takács, L. 'An Introduction to queueing theory'. (1962).

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

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.
Ancker Jr, C. J., and A. V. Gafarian. 'Some queuing problems with balking and reneging-II', Operations Research 11.6 (1963): 928-937.
Whitt, Ward. 'Improving service by informing customers about anticipated delays', Management science 45.2 (1999): 192-207.
Harchol-Balter, Mor. 'Performance modeling and design of computer systems: queueing theory in action', Cambridge University Press, 2013.
Bin Li, Lizhong Li, Bo Li, K. M. Sivalingam and Xi-Ren Cao, 'Call admission control for voice/data integrated cellular networks: performance analysis and comparative study', IEEE Journal on Selected Areas in Communications, vol. 22, no. 4, pp. 706–718, 2004.
Sai Rajesh Mahabhashyam and Natarajan Gautam. 'On Queues with Markov Modulated Service Rates', Queueing Syst. Theory Appl., vol. 51, no. 1-2, pp. 89–113, 2005.

Appendix A: Proofs related to the pseudo-conservation of the ϵ-US process

This appendix is devoted to the proof of Theorem 1. The proof requires the following functional version of the RRT.

Theorem 5 (Functional RRT) *Let $\Omega^1(t)$ represent a time-monotone non-negative cumulative reward function of the ϵ-renewal process when $\mu_\epsilon = 1$, with ν as its RRT limit:*

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

Then, for any $W < \infty$ and irrespective of the initial condition:

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

Proof: For any $\delta > 0$ there exists a T_δ (for any initial state) such that:

$$\left| \frac{\Omega^1(t)}{t} - \nu \right| \leq \delta \ \forall \ t \geq T_\delta. \tag{10}$$

Now pick $s \in [0, W]$. If $\mu s \geq T_\delta$ we have

$$\left| \frac{\Omega^1(\mu s)}{\mu s} \frac{s}{W} - \nu \frac{s}{W} \right| \leq \frac{s}{W}\delta \leq \delta. \tag{11}$$

If $\mu s < T_\delta$,

$$\left| \frac{\Omega^1(\mu s)}{\mu} - \nu s \right| \leq \ \frac{\Omega^1(\mu s)}{\mu} + \nu s \leq \frac{\Omega^1(T_\delta)}{\mu} + \nu \frac{T_\delta}{\mu} \leq \delta \tag{12}$$

for large enough μ, chosen appropriately for any initial condition.

It follows that

$$\sup_{s\in[0,W]} \left| \frac{\Omega^1(\mu s)}{\mu} - \nu s \right| \leq \delta$$

for large enough μ. This completes the proof. ■

Proof of Theorem 1: As in the proof of Lemma 1, we couple the arrival processes corresponding to the eager class for different μ_ϵ as per (1). Under this construction, it is easy to see that

$$\frac{\Omega^{\mu_\epsilon}(t)}{t} = \frac{\Omega^1(\mu_\epsilon t)}{\mu_\epsilon t}, \text{ which implies } \Omega^{\mu_\epsilon}(t) = \frac{\Omega^1(\mu_\epsilon t)}{\mu_\epsilon}.$$

Thus by Theorem 5, for any $W > 0$,

$$\sup_{t\in[0,W]} \left|\Omega^{\mu_\epsilon}(t) - \nu_\tau t\right| = \sup_{t\in[0,W]} \left|\frac{\Omega^1(\mu_\epsilon t)}{\mu_\epsilon} - \nu_\tau t\right| \to 0 \text{ a.s., as } \mu_\epsilon \to \infty. \quad (13)$$

This proves the statement about the asymptotic (in time) growth rate of the ϵ-US process.

We now prove the claim regarding the time required to obtain B units of service from the ϵ-US process. By continuity of probability measure, one can assume that (13) is satisfied together for a sequence of $\{W_n\}$ with $W_n \to \infty$, almost surely. Define the event A as follows.

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

For any outcome $\omega \in A$, consider a $W_n > (B(\omega)+1)/\nu_\tau$. For every $\delta \in (0,1)$, there exists $\bar{\mu} > 0$ such that:

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

Thus $\nu_\tau t - \delta \leq \Omega^{\mu_\epsilon}(t) \leq \nu_\tau t + \delta$ for all $t \leq W_n$. Now in particular for $t = (B(\omega)+\delta)/\nu_\tau < W_n$, we have:

$$\Omega^{\mu_\epsilon}(B(\omega)/\nu_\tau + \delta/\nu_\tau) \geq \nu_\tau((B(\omega)+\delta)/\nu_t) - \delta = B(\omega),$$

which implies $\Upsilon^{\mu_\epsilon} \leq (B(\omega)+\delta)/\nu_\tau$. Similarly, for $t = (B(\omega)-\delta)/\nu_\tau < W_n$, we have:

$$\Omega^{\mu_\epsilon}(B(\omega)/\nu_\tau - \delta/\nu_\tau) \leq \nu_\tau((B(\omega)-\delta))/\nu_t) + \delta = B(\omega).$$

Thus, $\Upsilon^{\mu_\epsilon} \geq (B(\omega)-\delta)/\nu_\tau$. We conclude that

$$\left|\Upsilon^{\mu_\epsilon} - B(\omega)/\nu_\tau\right| \leq \frac{\delta}{\nu_\tau} \quad \forall\ \mu_\epsilon \geq \bar{\mu}.$$

This completes the proof. Note that the above argument applies for any initial condition of the ϵ-state, as Theorem 5 is true for any initial state. ■

Appendix B: Proof of Theorem 2

This section is devoted to the proof of Theorem 2.

Stochastic upper bound on residual ϵ-busy periods

The first step in the proof is the construction of a stochastic upper bound on the residual ϵ-busy period $\tilde{\Psi}$ starting from any ϵ-state. Let s_ϵ denote the state of the ϵ-subsystem. For the case of bounded eager service requirements, $s_\epsilon = (y, \mathbf{r}) \in \{0, 1, \cdots, K\} \times [0, \bar{b}/\mu_\epsilon]^K$, for any μ_ϵ. Here, y is the number of ϵ-jobs present, and the vector $\mathbf{r}$ holds the residual service requirements of these ϵ-jobs at the start of $\tilde{\Psi}$. Note that $\mathbf{r}$ has exactly y positive entries; the remaining $K - y$ entries are set to zero. For the case of exponential eager service requirements, the ϵ-state $s_\epsilon = y$, the number of ϵ-jobs in the system.

Lemma 6 *Assume* **A.1-3** *and* **B.4**. *Let $\tilde{\Psi}^{\mu_\epsilon}(s_\epsilon)$ denote the residual ϵ-busy period starting with ϵ-state s_ϵ. One can construct an upper bound Θ^{μ_ϵ} independent s_ϵ that almost surely dominates $\tilde{\Psi}^{\mu_\epsilon}$, i.e.,*

$$\tilde{\Psi}^{\mu_\epsilon}(s_\epsilon) \leq_{\text{a.s.}} \Theta^{\mu_\epsilon}.$$

uniformly over s_ϵ. Further, as $\mu_\epsilon \to \infty$ under the SFJ scaling,

$$\Theta^{\mu_\epsilon} \to 0 \quad \textit{almost surely}, \tag{14}$$

$$\mathbb{E}\left[\left(\tilde{\Psi}^{\mu_\epsilon}(s_\epsilon)\right)^k\right] \leq \mathbb{E}\left[(\Theta^{\mu_\epsilon})^k\right] \to 0 \quad \forall\ k \in \mathbb{N}. \tag{15}$$

Proof: We begin with the case of bounded service requirements. Let $\Theta = \Theta^1$, be the random time distributed as the busy period in a fictitious M/G/∞ system with infinite buffer space (no loss system), when started with exactly K users each demanding $\bar{b}$ amount of service (see assumption **B.4**), and seeing the same ϵ-arrival stream as in our model at scale $\mu_\epsilon = 1$. Each server in this fictitious system operates at a service rate that equals the lowest possible service rate $c_{min} > 0$ of ϵ-jobs under the given ϵ-scheduling policy. By the way of our special construction, $\Theta^{\mu_\epsilon} := \Theta/\mu_\epsilon$ will be the busy period of the same system for any arbitrary μ_ϵ, with the understanding that the initial residual service requirement of each job equals $\bar{b}/\mu_\epsilon$. This immediately implies the statement about almost sure convergence of Θ^{μ_ϵ} under the SFJ limit. The statement about convergence of kth moment holds, so long as $\mathbb{E}[(\Theta^1)^k] < \infty$. To show that $\mathbb{E}[(\Theta^1)^k] < \infty$, divide time into blocks of length $\frac{\bar{b}}{c_{min}}$. Since $\frac{\bar{b}}{c_{min}}$ is the maximum time spent by a job in our fictitious system, the busy period ends if there is no arrival in any block of time. Thus, for $m \geq 2$,

$$P\left(\Theta^1 \geq m\frac{\bar{b}}{c_{min}}\right) \leq (1-p)^{m-1}, \tag{16}$$

where $p := e^{-\lambda_\epsilon \frac{\bar{b}}{c_{min}}}$ is the probability of no arrival in any block. Upper bound (16) implies that all moments of Θ^1 are bounded.

It is not hard to see that Θ dominates $\tilde{\Psi}^1$; indeed, the fictitious system has, at any time, a larger number of ϵ-jobs, each with a greater residual service requirement, as compared to the original system. While the original system may

have losses, the fictitious system does not. One can obtain the required dominance in almost sure sense when the following coupling[11] rules are employed: a) the ϵ-arrival epochs in the beginning of the Θ^{μ_ϵ} system are exactly the same as those that evolved $\tilde{\Psi}$ in the original system; b) the job demands of the subsequent customers of Θ^{μ_ϵ} system are exactly the same as the ones corresponding to the original system; and c) we complete construction of Θ^{μ_ϵ}, using independent copies of ϵ-job requirements and inter arrival times as required. By above construction, it is clear that Θ^{μ_ϵ} is independent of ϵ-initial state $s_\epsilon = (y, \mathbf{r})$.

For exponential ϵ-jobs, similar arguments are applicable, except that we further couple the service times of the first y customers with the (residual) service times of the y customers of the original system. The remaining $(K-y)$ customer in upper bound system are independent copies of exponential distribution with parameter μ_ϵ. In fact we replace the (residual) service times of the y customers of the original system with IID copies, and the same realizations are used by Θ^{μ_ϵ} customers. This does not change the stochastic description of the original system. Thus for exponential ϵ-jobs, Θ^{μ_ϵ} almost surely upper bounds $\tilde{\Psi}(s_\epsilon)$ (for any $s_\epsilon = y$) and is independent of the residual service times of the ϵ customers present at the start of $\tilde{\Psi}$. ■

Ergodicity

The next step in the proof of Theorem 2 is to establish ergodicity of the number of jobs in the tolerant sub-system. For the case of exponential ϵ-service requirements, ergodicity follows from Mahabhashyam et al. (2005). Thus, we focus only on the case of bounded ϵ-service requirements. As discussed in Section 4, we analyse the evolution of the system across departure instants of tolerant jobs, via the Markov process $Z_n := (X_n, Y_n, \mathbf{R}_n^s)$. The following theorem states that this Markov process is ergodic for μ_ϵ large enough.

Theorem 6 *Assume* **A.1-3**, **B.1-6** *and general bounded ϵ-jobs. There exists a $\bar{\mu} < \infty$ such that Z_n is positive recurrent and aperiodic for every $\mu_\epsilon \geq \bar{\mu}$. Hence it has a stationary distribution and the state at time n converges to the stationary distribution in total variation norm. Thus we have the convergence of the marginals also, i.e., with X_* representing the stationary quantity*

$$\sup_j |\mathbb{E}^{\mu_\epsilon}[X_n = j] - \mathbb{E}^{\mu_\epsilon}[X_* = j]| \to 0 \text{ as } n \to \infty \text{ for all } \mu_\epsilon \geq \bar{\mu}.$$

Proof: Let $P(z, .)$ represent the probability transition Kernel of the Markov chain. Consider the Lyaponuv function $V(z) = x$ for all $z = (x, y, \mathbf{r}) = (x, s_\epsilon)$. Note that $PV(z) = \mathbb{E}[V(Z_1)|Z_0 = z]$. For any $x \geq 1$ it is clear from (5),

[11] To compare two stochastic systems, we use realizations of some random quantities of one system in defining the other system so as to ensure the required dominance in almost sure sense. Where required we use independent copies of some other random quantities.

$$\begin{aligned}\Delta V(z) = PV(z) - V(z) &= (x-1) + \mathbb{E}[A_{n+1}(\Upsilon_n^{\mu_\epsilon})|s_\epsilon] - x \\ &= \mathbb{E}_{s_\epsilon}[A_{n+1}(\Upsilon_n^{\mu_\epsilon})] - 1, \qquad \text{and by conditioning on } \Upsilon_n^{\mu_\epsilon}, \\ &= \lambda_\tau \mathbb{E}_{s_\epsilon}[\Upsilon_n^{\mu_\epsilon}] - 1 \;\le\; \lambda_\tau \sup_{s'_\epsilon} \mathbb{E}_{s'_\epsilon}[\Upsilon_n^{\mu_\epsilon}] - 1.\end{aligned}$$

By Lemma 8 given below, with $\delta := (1 - \lambda_\tau(\mathbb{E}[B_\tau]/\nu_\tau)/2$ (which is greater than 0 by **B.5**) there exists a $\bar{\mu} < \infty$ such that, for any $\mu_\epsilon \ge \bar{\mu}$:

$$\sup_{s'_\epsilon} \mathbb{E}_{s'_\epsilon}[\Upsilon_n^{\mu_\epsilon}] \le \mathbb{E}[B_\tau]/\nu_\tau + \delta/\lambda_\tau, \text{ hence } \lambda_\tau \sup_{s'_\epsilon} \mathbb{E}_{s'_\epsilon}[\Upsilon_n^{\mu_\epsilon}] - 1 \le -\delta < 0.$$

Thus one can choose a $\delta > 0$ and a $\bar{\mu} < \infty$, such that for all $\mu_\epsilon \ge \bar{\mu}$

$$\Delta V(z) < -\delta \text{ when } x \ge 1. \tag{17}$$

The above negative drift is true for all $z \in C^c$ where

$$C := \{(x, y, \mathbf{r}) :\; x = 0\}. \tag{18}$$

For any $z \in C$, i.e., when $x = 0$ and for any (s_ϵ), again using Lemma 8

$$\Delta V(z) = PV(z) - V(z) = \mathbb{E}_{s_\epsilon}[A_{n+1}(\Upsilon_n^{\mu_\epsilon})] \;\le\; \lambda_\tau \sup_{s'_\epsilon} \mathbb{E}_{s'_\epsilon}[\Upsilon_n^{\mu_\epsilon}] < \infty.$$

By Lemma 7 of Appendix B, C is a small set and is aperiodic. Thus, by (Meyn et. al. 1993, Theorem 13.0.1, pp. 313, equation 13.4), we have convergence of the stationary distribution in total variation norm and hence the theorem. ■

Lemma 7 *Set C of (18) is a small set and the Markov chain is aperiodic.*

Proof: Let $P(z, .)$ represent the probability transition kernel of the Markov chain. For any event A and any $z \in C$ (i.e., when $z = (0, s_\epsilon)$) it is clear to see that

$$P(z, A) \ge P(z,\; A \cap \{(0, 0, \mathbf{0})\}) = Prob(z, (0, 0, \mathbf{0}))\mathbb{1}_{\{(0,0,\mathbf{0}) \in A\}}. \tag{19}$$

Starting from any $z \in C$ one can uniformly lower bound $Prob(z, (0, 0, \mathbf{0}))$ by the probability of the following event: a) there are no τ or ϵ arrivals for a time $\frac{\bar{b}}{\mu_\epsilon c_{min}}$, by which time all existing ϵ-customers would have been served (note that ϵ-jobs have a maximum size of $\bar{b}/\mu_\epsilon$, and are served at a minimum rate of c_{min}) ; b) the first arrival after time $\frac{\bar{b}}{\mu_\epsilon c_{min}}$ is a τ-job; c) there are no τ or ϵ arrival for time B_τ, which is the service requirement of this newly arrived τ-job. Under this event, it is clear that state of the system after the next τ-departure would be $(0, 0, \mathbf{0})$. Thus for any $z \in C$,

$$Prob(z, (0, 0, \mathbf{0})) \ge e^{-\frac{(\lambda_\tau + \lambda_\epsilon)\bar{b}}{\mu_\epsilon c_{min}}} \left(\frac{\lambda_\tau}{\lambda_\tau + \lambda_\epsilon}\right) \int e^{-(\lambda_\epsilon + \lambda_\tau)s} dG_\tau(s) =: \kappa, \tag{20}$$

where G_τ is the distribution of the τ-service times B_τ. Thus from (19),

$$P(z, A) \ge P(z,\; A \cap \{(0, 0, \mathbf{0})\}) \ge \kappa \mathbb{1}_{\{(0,0,\mathbf{0}) \in A\}} \text{ for any } z \in C. \tag{21}$$

Hence the one step transition probability measure (19), is lower bounded uniformly for all $z \in C$, with a non-trivial measure (need not be probability) that concentrates only on $\{(0,0,\mathbf{0})\}$ and hence C is a small set (see (Meyn et. al. 1993, Equation (5.14), pp.109) for definition of small set). This also shows that the 'significant' atom $\{(0,0,\mathbf{0})\}$ (see (Meyn et. al. 1993, page 103) for definition of atom) is aperiodic and hence that the Markov chain is aperiodic. ■

Lemma 8 *Assume* **A.1-3**, **B.1-6**. *Let Υ^{μ_ϵ} denote the time required for the tolerant class to receive service B_τ, starting with ϵ-state s_ϵ. There exists a uniform (over initial ϵ-states) almost sure upper bound $\bar{\Upsilon}^{\mu_\epsilon}$ of Υ^{μ_ϵ}, such that $\bar{\Upsilon}^{\mu_\epsilon} \to \frac{B_\tau}{\nu_\tau}$ almost surely as $\mu_\epsilon \to \infty$. Moreover,*

$$\begin{array}{lll} \sup_{s_\epsilon} \mathbb{E}\left[\Upsilon^{\mu_\epsilon} \middle| s_\epsilon\right] & \leq \mathbb{E}\left[\bar{\Upsilon}^{\mu_\epsilon}\right] & \text{and } \lim_{\mu_\epsilon \to \infty} \mathbb{E}\left[\bar{\Upsilon}^{\mu_\epsilon}\right] = \frac{\mathbb{E}[B_\tau]}{\nu_\tau}, \\ \sup_{s_\epsilon} \mathbb{E}\left[\left(\Upsilon^{\mu_\epsilon}\right)^2 \middle| s_\epsilon\right] & \leq \mathbb{E}\left[\left(\bar{\Upsilon}^{\mu_\epsilon}\right)^2\right] & \text{and } \lim_{\mu_\epsilon \to \infty} \mathbb{E}\left[\left(\bar{\Upsilon}^{\mu_\epsilon}\right)^2\right] = \frac{\mathbb{E}[B_\tau^2]}{\nu_\tau^2}. \end{array}$$

Proof of Lemma 8: Let $\Upsilon^{\mu_\epsilon}(s_\epsilon)$ represent the time taken to serve a τ-customer with job requirement B_τ (see (6)) when ϵ-state at τ-service start equals s_ϵ. We construct a fictitious upper system such that $\Upsilon^{U,\mu_\epsilon}$, the time taken to serve the same (realization of the) job B_τ in the upper system, upper bounds $\Upsilon^{\mu_\epsilon}(s_\epsilon)$ of original system uniform over all s_ϵ and almost surely. The starting residual ϵ-busy period $\tilde{\Psi}$ of the original system (for any s_ϵ) is replaced by Θ^{μ_ϵ} of Lemma 6 in the upper system, so that,

$$\tilde{\Psi}(s_\epsilon) \leq \Theta^{\mu_\epsilon} \text{ almost surely for any } s_\epsilon.$$

Zero τ-service is offered during Θ^{μ_ϵ}. And after $\tilde{\Psi}$ in the original system (respectively Θ^{μ_ϵ} in upper system), the remaining B_τ (entire B_τ in upper system) is completed using the server capacity available to the τ-class when the ϵ-jobs in both the systems are driven with same realizations of the random quantities like further arrival times, ϵ-job requirements etc. Further, both the systems use the same ϵ-scheduling policy. This ensures required almost sure domination[12]. Thus $\Upsilon^{\mu_\epsilon}(s_\epsilon) \leq \Upsilon^{U,\mu_\epsilon}$ for all s_ϵ a.s. and hence

$$\sup_{s_\epsilon} \mathbb{E}\left[\Upsilon^{\mu_\epsilon} \middle| s_\epsilon\right] \leq \mathbb{E}\left[\Upsilon^{U,\mu_\epsilon}\right] \text{ for any } \mu_\epsilon \text{ and observe that } \Upsilon^{U,\mu_\epsilon} \stackrel{d}{=} \Theta^{\mu_\epsilon} + \Upsilon^{\mu_\epsilon}(\mathbf{0}).$$

It now follows from Theorem 1 and Lemma 6 that $\Upsilon^{U,\mu_\epsilon} \to \frac{B_\tau}{\nu_\tau}$ almost surely as $\mu_\epsilon \to \infty$.

Statements about convergence in expectation follow once we show uniform integrability of $\{\Upsilon^{U,\mu_\epsilon}\}_{\mu_\epsilon}$. Define $\hat{\hat{\Upsilon}}^{\mu_\epsilon}(\mathbf{0})$ as the time taken to serve B_τ amount of work for the tolerant class, when the τ-customers are served only during ϵ-idle periods. Note this also implies that the tolerant class service has not started till the first ϵ-busy period is over in $\hat{\hat{\Upsilon}}^{\mu_\epsilon}(\mathbf{0})$. Clearly,

[12] The τ-customer receives some (respectively zero) service in original (respectively upper) system during $\tilde{\Psi}$ (respectively during bigger Θ^{μ_ϵ}) and after $(\tilde{\Psi}, \Theta^{\mu_\epsilon})$ τ-service is received in exactly the same way, in both the systems.

$\Upsilon^{\mu_\epsilon}(\mathbf{0}) \leq_{\text{a.s.}} \hat{\hat{\Upsilon}}^{\mu_\epsilon}(\mathbf{0})$ (dominated almost surely). Via exactly the same analysis as in Theorem 1, one can show that

$$\hat{\hat{\Upsilon}}^{\mu_\epsilon}(\mathbf{0}) \to \frac{B}{\bar{\nu}_\tau} \text{ almost surely, where } \bar{\nu}_\tau = \frac{1/\rho_\epsilon}{1/\rho_\epsilon + \mathbb{E}^1[\Psi_1]} = \frac{1}{1+\rho_\epsilon \mathbb{E}^1[\Psi_1]},$$

equals the long run fraction of time the ϵ-queue is idle. Also, because of the simplistic τ-server availability rules and because of memoryless ϵ-arrivals, one can express

$$\hat{\hat{\Upsilon}}^{\mu_\epsilon}(\mathbf{0}) = B_\tau + \sum_{i=1}^{N(B_\tau)} \Psi_i,$$

where $N(B_\tau)$ is the number ϵ-arrivals (or ϵ-interruptions) during time period B and $\{\Psi_i\}$ are resulting ϵ-busy cycles. Note these are IID. By the special construction (1) and from Lemma 6 (Ψ_1 is a partial busy period with $s_\epsilon = \mathbf{0}$):

$$\mathbb{E}^{\mu_\epsilon}[(\Psi_1)^k] = \frac{\mathbb{E}^1[(\Psi_1)^k]}{\mu_\epsilon^k} \text{ for any } k.$$

Further by conditioning first on B_τ,

$$\begin{aligned}
\mathbb{E}\left[\hat{\hat{\Upsilon}}^{\mu_\epsilon}(\mathbf{0})\right] &= \mathbb{E}[B_\tau] + \lambda_\epsilon \mathbb{E}[B_\tau]\mathbb{E}^{\mu_\epsilon}[\Psi_1] = \mathbb{E}[B_\tau] + \rho_\epsilon \mathbb{E}[B_\tau]\mathbb{E}^1[\Psi_1] = \mathbb{E}[B_\tau]/\bar{\nu}_\tau, \text{ for any } \mu_\epsilon \text{ and} \\
\mathbb{E}\left[\left(\hat{\hat{\Upsilon}}^{\mu_\epsilon}(\mathbf{0})\right)^2\right] &= \mathbb{E}[B_\tau^2] + 2\mathbb{E}[B_\tau N(B_\tau)]\mathbb{E}^{\mu_\epsilon}[\Psi_1] + \mathbb{E}[N(B_\tau)]\mathbb{E}^{\mu_\epsilon}[(\Psi_1)^2] \\
&\qquad + \mathbb{E}[N(B_\tau)(N(B_\tau)-1)]\left(\mathbb{E}^{\mu_\epsilon}[\Psi_1]\right)^2 \\
&= \mathbb{E}[B_\tau^2] + 2\mathbb{E}[B_\tau^2]\rho_\epsilon \mathbb{E}^1[\Psi_1] + \rho_\epsilon \mathbb{E}[B_\tau]\frac{\mathbb{E}^1[(\Psi_1)^2]}{\mu_\epsilon} + \rho_\epsilon^2 \mathbb{E}[B_\tau^2]\left(\mathbb{E}^1[\Psi_1]\right)^2 \\
&\to \mathbb{E}[B_\tau^2] + 2\mathbb{E}[B_\tau^2]\rho_\epsilon \mathbb{E}^1[\Psi_1] + 0 + \rho_\epsilon^2 \mathbb{E}[B_\tau^2]\left(\mathbb{E}^1[\Psi_1]\right)^2 = \mathbb{E}\left[\frac{B_\tau^2}{\bar{\nu}_\tau^2}\right], \text{ as } \mu_\epsilon \to \infty.
\end{aligned}$$

Thus and further using (15) of Lemma 6, there exists a $\bar{\mu}_\epsilon$ large enough such that:

$$\sup_{\mu_\epsilon \geq \bar{\mu}_\epsilon} \mathbb{E}^{\mu_\epsilon}\left[\left(\Upsilon^{U,\mu_\epsilon}\right)^2\right] \leq \mathbb{E}\left[\frac{B^2}{\bar{\nu}_\tau^2}\right] + \delta.$$

Thus $\{\Upsilon^{U,\mu_\epsilon}\}_{\mu_\epsilon}$ are uniformly integrable and hence we obtain L^1 convergence. Using analogous arguments, it can be shown that

$$\sup_{\mu_\epsilon \geq \bar{\mu}_\epsilon} \mathbb{E}^{\mu_\epsilon}\left[\left(\Upsilon^{U,\mu_\epsilon}\right)^3\right] < \infty,$$

which implies that $\{(\Upsilon^{U,\mu_\epsilon})^2\}_{\mu_\epsilon}$ are uniformly integrable and hence we also obtain L^2 convergence. ■

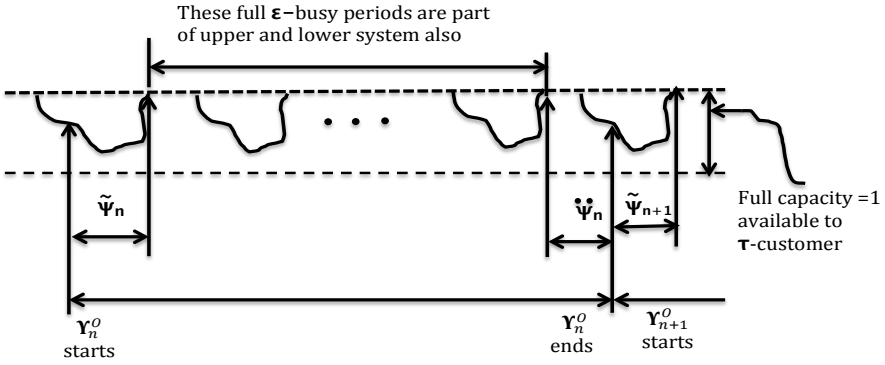

Fig. 8: Υ_n^O spanning over several ϵ-busy periods, starting with $\tilde{\Psi}_n$ and ending with $\ddot{\Psi}_n$.

Proof of Theorem 2

We are now ready to prove Theorem 2. The main idea of the proof is that the number of tolerant customers in the system at any time can be bounded from above and below by two M/G/1 systems, such that the performance metrics for the two bounding systems converge to the same value under the SFJ limit.

We begin with construction of the dominating queues for the scenario with bounded ϵ-jobs. The case with exponential ϵ-job requirements is considered at the end. The key challenge is to ensure IID service times in the bounding systems. Note that τ-service times in the original system are not in general independent (since successive τ-jobs might begin service within the same ϵ-busy period).

- Let Υ^O be defined as the total time period between the service start and the service end of a typical τ-agent (in original system) and we refer this as the effective server time (EST). A typical EST (say Υ_n^O) begins with a residual ϵ-busy period (call this $\tilde{\Psi}_n$) and then might span over multiple ϵ-busy periods, before ending with another partial ϵ-busy period (call this $\ddot{\Psi}_n$) (for e.g., see Figure 8). Thus $\ddot{\Psi}_n$ and $\tilde{\Psi}_{n+1}$ together form the ϵ-busy period in between which the n-th τ-customer departs (if it does not leave behind an τ-queue empty), which leads to correlations in original system.
- We construct $\Theta_n^{\mu_\epsilon}$ (for any arbitrary μ_ϵ, n), which almost surely dominates $\tilde{\Psi}_n$ (for any ϵ-state s_ϵ at τ-service start) as described in Lemma 6 of Appendix B.
- The τ-service in original system begins with $\tilde{\Psi}_n$, while that in the two bounding systems begins with $\Theta_n^{\mu_\epsilon}$. During this period the τ customer in lower system has access to the complete system capacity (i.e., it is served at rate 1) while in upper system it has access to zero capacity. After $\Theta_n^{\mu_\epsilon}$ in bounding systems (respectively after $\tilde{\Psi}_n$ in original system) the two bounding systems continue with τ-service exactly as in original system.

This continues at maximum for the full-ϵ-busy periods of the original system, over which the τ-service of the customer under consideration in the original system, spans completely (Figure 8).

- We continue the same for the 'last' partial busy period $\ddot{\Psi}_n$, till which the service of the current τ-customer continues (in the original system). From this point onwards subsequent ϵ-inter-arrival times and the service times of the subsequent new ϵ-arrivals are continued with IID copies. We however couple the departure epochs of the ϵ-customers (Y_{n+1} of them) deriving service at the (n-th) τ-departure epoch in the original system.
- This ensures that the modified 'last' ϵ-busy period used by the dominating systems is stochastically the same as the usual ϵ-busy period, because of memoryless property of ϵ-Poisson arrivals. And further this period is independent of $\Theta^{\mu_\epsilon}_{n+1}$ (constructed by coupling the subsequent ϵ-inter-arrival times and service times in Lemma 6) which dominates $\tilde{\Psi}_{n+1}$ of the next τ-customer.
- The span of the time during which the τ-job is completed, in each system, becomes the service time of the corresponding customer in the respective dominating systems.
- The $\tilde{\Psi}_{n+1}$ of any customer can be correlated with $\ddot{\Psi}_n$ of the previous customer in original system. However after $\ddot{\Psi}_n$, ϵ-busy period of previous customer is continued using independent quantities in the dominating systems, except for departure epochs of the ϵ-customers existing at τ-departure, and $\Theta^{\mu_\epsilon}_{n+1}$ dominating $\tilde{\Psi}_{n+1}$ of the next customer is independent of these ϵ-departure epochs. Thus correlations are avoided in the dominating systems and hence they are M/G/1 queues.
- To summarize, by way of the construction the IID service times in the dominating queues equal:

$$\begin{aligned} \Upsilon^U &= \Theta^{\mu_\epsilon} + \Upsilon(\mathbf{0}, B_\tau), \\ \Upsilon^O &= \Upsilon(s_\epsilon, B_\tau) \text{ and} \\ \Upsilon^L &= \min\left\{\Theta^{\mu_\epsilon}, B_\tau\right\} + \Upsilon(\mathbf{0}, [(B_\tau - \Theta^{\mu_\epsilon})\mathbb{1}_{\{B_\tau > \Theta^{\mu_\epsilon}\}}]), \end{aligned} \tag{22}$$

  where $\Upsilon(s_\epsilon, b)$ represents the time taken to complete job requirement b, when ϵ-state at τ-service start is s_ϵ.
- The service time Υ^L in the lower system (with same realization of B_τ) is less or equal to Υ^O, because τ-customer in lower system receives service at the maximum rate during the (bigger) interval Θ^{μ_ϵ}, as compared to the time varying rate (which is less than or equal to the maximum rate) with which the service is offered in original system during the (smaller) interval $\tilde{\Psi}$. After $\tilde{\Psi}$ and Θ^{μ_ϵ} (respectively), the service rate available for τ-customer evolves in exactly the same way for both the systems. Thus, the customer of lower system leaves before that in the original system.
- Similarly, the service time Υ^U in the upper system is greater than or equal to Υ^O. There is a possibility that service is not completed in the same ϵ-busy period (which was modified for bounding system) in the upper system

as in the original system. If so, the rest of the service is completed using independent copies of ϵ-busy periods as required.

Thus departure instances of τ-customers are delayed in the upper system in comparison with that of the original system almost surely, while they occur sooner in lower system almost surely. Hence the number customers at time t in the three systems:

$$X^L(t) \leq X^O(t) \leq X^U(t) \text{ almost surely, for all } t. \tag{23}$$

This implies the following dominance at τ-departure epochs

$$X^L_n \leq X^O_n \leq X^U_n \text{ almost surely and for all } n. \tag{24}$$

Stationary Performance: Let X^m_* (with $m = U$ or L or O) be the random variable distributed as the stationary number in the system, if the corresponding stationary distribution exists. The upper system may not be stable for all μ_ϵ because of Θ^{μ_ϵ}. By Lemma 9 of Appendix B and **B.5**, the upper system is stable (i.e., $\lambda_\tau \mathbb{E}[\Upsilon^U] < 1$) for all large enough μ_ϵ. Consider larger μ_ϵ, if required, for which the original system is also stationary as given by Theorem 6. Thus, for large enough μ_ϵ, we have the weak convergence

$$X^L_n \implies X^L_*, \quad X^O_n \implies X^O_*, \quad X^U_n \implies X^U_*.$$

The three weak convergence results along with stochastic domination[13] as given by (24) gives the following for any large enough $\mu_\epsilon < \infty$ by (Shaked et al. 2007, Theorem 1.A.3(c), pp. 6):

$$X^L_* \leq_{\text{st}} X^O_* \leq_{\text{st}} X^L_*. \tag{25}$$

Convergence in distribution: It follows from (25) that the moment generating functions (MGFs) corresponding to X^L_*, X^O_*, and X^U_* satisfy the following relation[14], for $z \in (0, 1)$.

$$\mathbb{E}^{\mu_\epsilon}\left[z^{X^L_*}\right] \geq \mathbb{E}^{\mu_\epsilon}\left[z^{X^O_*}\right] \geq \mathbb{E}^{\mu_\epsilon}\left[z^{X^U_*}\right] \text{ (for large enough } \mu_\epsilon). \tag{26}$$

Now, as $\mu_\epsilon \to \infty$, Lemma 9 below implies that the $\Upsilon^U, \Upsilon^L \xrightarrow{d} \frac{B_\tau}{\nu_\tau}$. It follows from the continuity theorem for Laplace transforms (see (Feller 1972,

[13] Almost sure dominance given by (24) easily implies the stochastic dominance:

$$P(X^U_n \geq x) \geq P(X^O_n \geq x) \geq P(X^L_n \geq x) \text{ for all } x, n.$$

[14] Stochastic dominance implies the dominance of expected values of any increasing function (e.g., (Shaked et al. 2007, Equation (1.A.7), pp. 4)), i.e.,:

$$\mathbb{E}[\phi(X^U_n)] \geq \mathbb{E}[\phi(X^O_n)] \geq \mathbb{E}[\phi(X^L_n)] \text{ for any increasing function, } \phi.$$

Theorem 2, Chapter 13.1)) that the corresponding Laplace transforms satisfy, for $s > 0$,

$$\mathbb{E}^{\mu_\epsilon}\left[e^{-s\Upsilon^U}\right], \mathbb{E}^{\mu_\epsilon}\left[e^{-s\Upsilon^L}\right] \to \mathbb{E}\left[e^{-s\frac{B_\tau}{\nu_\tau}}\right] \quad \text{as } \mu_\epsilon \to \infty.$$

Given the representation of the MGF of the steady state number in system of an M/G/1 queue in terms of the Laplace transform of the job size distribution (see (Harchol-Balter 2013, Chapter 26)), it follows that

$$\lim_{\mu_\epsilon \to \infty} \mathbb{E}^{\mu_\epsilon}\left[z^{X_*^U}\right] = \lim_{\mu_\epsilon \to \infty} \mathbb{E}^{\mu_\epsilon}\left[z^{X_*^L}\right] = \mathbb{E}\left[z^{X_*}\right],$$

where X_* denotes the stationary number in system corresponding to an M/G/1 queue with arrival rate λ_τ and job sizes $\frac{B_\tau}{\nu_\tau}$. We now conclude from (26) that

$$\lim_{\mu_\epsilon \to \infty} \mathbb{E}^{\mu_\epsilon}\left[z^{X_*^O}\right] = \mathbb{E}\left[z^{X_*}\right],$$

which implies, from the continuity theorem for MGFs (see (Feller 1968, Section 11.6) that $X_*^O \xrightarrow{d} X_*$.

Convergence in expectation:

Similarly, it follows from (25) that

$$\mathbb{E}^{\mu_\epsilon}(X_*^L) \le \mathbb{E}^{\mu_\epsilon}(X_*^O) \le \mathbb{E}^{\mu_\epsilon}(X_*^U). \tag{27}$$

From Lemma 9 below, we have, as $\mu_\epsilon \to \infty$, the job requirements of the two bounding queues, satisfy

$$\mathbb{E}^{\mu_\epsilon}[\Upsilon^U] \to \frac{\mathbb{E}[B_\tau]}{\nu_\tau},\ \mathbb{E}^{\mu_\epsilon}[\Upsilon^L] \to \frac{\mathbb{E}[B_\tau]}{\nu_\tau} \text{ and}$$
$$\mathbb{E}^{\mu_\epsilon}[(\Upsilon^U)^2] \to \frac{\mathbb{E}[(B_\tau)^2]}{\nu_\tau^2},\ \mathbb{E}^{\mu_\epsilon}[(\Upsilon^L)^2] \to \frac{\mathbb{E}[(B_\tau)^2]}{\nu_\tau^2}.$$

Thus the stationary expected number in the system with μ_ϵ in the two bounding systems converges to the stationary performance $\mathbb{E}[X_*]$ of the M/G/1 system with the same arrival rate and with service times B_τ/ν_τ:

$$\mathbb{E}^{\mu_\epsilon}(X_*^L) \to \mathbb{E}(X_*) \text{ and } \mathbb{E}^{\mu_\epsilon}(X_*^U) \to \mathbb{E}(X_*).$$

Hence the performance of both the dominating systems converge to the same constant and hence the sand-witched performance of the original system from (27) also converges to

$$\mathbb{E}^{\mu_\epsilon}(X_*^O) \to \mathbb{E}\left[X_*\right].$$

This completes the proof for bounded ϵ-jobs, when we observe that $X(t)$ performance of M/G/1 system with service times B_τ/ν_τ and unit rate server is same as that in the M/G/1 system with service times B_τ and ν_τ-rate server.

For exponential ϵ-jobs, the construction is almost the same, except that we need not couple the departure epochs of ϵ-customers present at the end of $\ddot{\Psi}$ in the dominating systems. This is because these departure epochs are after

exponentially distributed times (by memoryless property). More importantly we replace these departure epochs in $\tilde{\Psi}$ corresponding to the next customer of the original system by an IID copy and further couple the same with Θ^{μ_ϵ} construction. This does not change stochastic description of the original system. At the start of each τ-service at state $s_\epsilon = y$ in the original system, we construct Θ^{μ_ϵ} (see Lemma 6) using K (fresh) IID exponential random variables to capture the residual service requirements. The residual service requirement of the y ϵ-jobs in the original system is taken to as the first y of the same random variables. This coupling ensures that the Θ^{μ_ϵ} almost surely dominates the residual ϵ-busy period in the original system, and is additionally independent of any previous upper bound constructions in the same ϵ-busy period. This ensures that the service requirements in the two bounding systems remain IID. The rest of the proof follows along exactly similar lines. ■

Lemma 9 *For the upper and lower bounding system constructed in the proof of Theorem 2, as* $\mu_\epsilon \to \infty$,

$$\Upsilon^U,\ \Upsilon^L \xrightarrow{d} \frac{B_\tau}{\nu_\tau},$$
$$\mathbb{E}^{\mu_\epsilon}\left[\Upsilon^U\right],\ \mathbb{E}^{\mu_\epsilon}\left[\Upsilon^L\right] \to \frac{\mathbb{E}\left[B_\tau\right]}{\nu_\tau}$$
$$\mathbb{E}^{\mu_\epsilon}\left[(\Upsilon^U)^2\right],\ \mathbb{E}^{\mu_\epsilon}\left[(\Upsilon^L)^2\right] \to \frac{\mathbb{E}\left[B_\tau^2\right]}{(\nu_\tau)^2}$$

Proof of Lemma 9: For any ϵ-state at the start of τ-service, almost sure convergence of $\Upsilon^U,\ \Upsilon^L$ under the SFJ limit follow from Theorem 1 and Lemma 6.

Convergence of the first and second moment of $\Upsilon^U,\ \Upsilon^L$ under the SFJ limit follow from the uniform integrability arguments in the proof of Lemma 8. ■

Appendix C: Proofs for Theorem 3, workload pseudo-conservation

Proof of Theorem 3: By PASTA the workload seen by the arriving customers equals the stationary workload. Hence we consider an alternate embedded Markov chain of the original system, that observed at τ-arrival instances to study the τ-workload. We also consider an additional Markov chain, that defined in assumption **W.2**. This chain is updated at every ϵ-arrival/departure epoch and hence evolves at a much faster rate than the former τ state representing Markov chain. Further the $\{\zeta_n\}_n$ chain is defined with $\mu_\epsilon = 1$, however by Equations (1) and (3) also represents the Markov chain for any μ_ϵ, when the residual ϵ-job size component of the ϵ-state is scaled down by $1/\mu_\epsilon$. With exponential ϵ-jobs its state space is finite, i.e., $\zeta \in \{0, 1, \cdots, K\}$ and hence the uniform rate of convergence required by **W.2** is immediate. With bounded ϵ-jobs $\zeta \in \{0, 1, \cdots, K\} \times [0, \bar{b}]^K$, i.e., we have a compact state space.

Theorem 1 is again applicable. We obtain the ergodicity as well as the convergence in distribution/mean, both by establishing geometric ergodicity

of the alternate embedded Markov chain, just discussed. Let U_n represent the workload seen by n-th τ-arrival and redefine $(Y_n, \mathbf{R}_n^s)$ to again represent the ϵ-number and the ϵ-residual service times vector respectively, but now at τ-arrival instances and we study the redefined Markov chain $Z_n = (U_n, Y_n, \mathbf{R}_n^s)$. Firstly

$$U_{n+1} = \left(U_n + B_{\tau,n} - \Omega_n^{\mu_\epsilon}(A_{\tau,n+1})\right)^+,$$

where the start of $\Omega_n^{\mu_\epsilon}$ (the total server capacity available to τ-customer for a duration $A_{\tau,n+1}$) is dictated by $(Y_n, \mathbf{R}_n^s)$, $B_{\tau,n}$ is the service time of n-th τ customer and $A_{\tau,n+1}$ is the inter-arrival time before the $(n+1)$-th τ arrival.

Lyaponuv Function: By Lemma 10 given below, for any given $\delta > 0$, one can chose $\bar{\mu}, \bar{u}$ such that for any $\mu_\epsilon \geq \bar{\mu}$, $u \geq \bar{u}$ and for any s_ϵ:

$$\mathbb{E}_{s_\epsilon}\left[a^{(u+B_{\tau,1}-\Omega_n^{\mu_\epsilon}(A_{\tau,1}))^+}\right] \leq \mathbb{E}\left[a^{(u+B_{\tau,1}-\nu_\tau A_{\tau,1})}\right] + a^u\delta. \tag{28}$$

The above is true for any $1 < a \leq \bar{a}$ by **W.1**, however $\bar{u}$, $\bar{\mu}$ can depend upon a, δ. Consider $V(z) = V(u) = a^u$ and we will choose a suitable $\bar{a} \geq a > 1$, to obtain the appropriate Lyapunov function. Consider any such a, any $\delta > 0$, and $u > \bar{u}(a,\delta)$, $\mu_\epsilon \geq \bar{\mu}(a,\delta)$. Then, with $P(\cdot,\cdot)$ being the transition probability kernel and $PV(z) = \mathbb{E}_z^{\mu_\epsilon}[V(U_1, Y_1, \mathbf{R}_1^s)] = \mathbb{E}_z^{\mu_\epsilon}[V(U_1)]$, we have

$$\begin{aligned} PV(z) - V(z) \;&\leq\; a^u(g(a) + \delta - 1) = V(z)(g(a) + \delta - 1) \text{ with} \\ g(a) &:= \mathbb{E}[a^{B_\tau - \nu_\tau A_{\tau,1}}]. \end{aligned}$$

One can chose an a in neighbourhood of 1 such that $g(a) < 1$ because of the following reasons:

$$g(1) = 1 \text{ for } a = 1, \quad \text{and } \frac{dg}{da}\Big|_{a=1} = \mathbb{E}[B_\tau - \nu_\tau A_\tau] < 0.$$

Basically the derivative is less than 0, is continuous in a, and hence would remain negative in some neighbourhood of $a = 1$. Thus $g(\cdot)$ is decreasing beyond $a = 1$ in neighbourhood of $a = 1$. Thus choose an $a \leq \bar{a}$, for which $g(a) < 1$ and further choose a $\delta > 0$ such that $g(a) < 1 - \delta$. By Lemma 10 choose $\bar{u}, \bar{\mu}$ for this (δ, a) and then the geometric drift condition is satisfied with $\beta := g(a) + \delta < 1$ because,

$$\begin{aligned} \Delta V(z) = PV(z) - V(z) \;&=\; V(z)(\beta - 1) \text{ for all } u > \bar{u} \text{ and} \\ \Delta V(z) \;&\leq\; K_\Delta \text{ for all } u \leq \bar{u} \text{ with} \\ K_\Delta &:= \sup_{u \leq \bar{u}} PV(u) = \sup_{u \leq \bar{u}} \mathbb{E}_{s_\epsilon}^{\mu_\epsilon}\left[a^{(u+B_{\tau,1}-\Omega_n^{\mu_\epsilon}(A_{\tau,1}))^+}\right] \\ &\leq\; a^{\bar{u}}\mathbb{E}[a^{B_\tau}] < \infty. \end{aligned}$$

Note that K_Δ and β are the same for every $\mu_\epsilon \geq \bar{\mu}$.

Small set: We prove that one of the following sets is a small set:

$$\begin{aligned} \mathcal{C} &:= [0, \bar{u}] \times \{0, 1, \cdots, K\} \times [0, \bar{b}/\mu_\epsilon]^K \text{ for bounded } \epsilon\text{-jobs} \\ \mathcal{C} &:= [0, \bar{u}] \times \{0, 1, \cdots, K\} \text{ for exponential } \epsilon\text{-jobs.} \end{aligned} \tag{29}$$

The embedded ϵ-Markov chain $\{\zeta_n\}_n$ changes at every ϵ-arrival/departure epoch. Thus if the number of arrivals increases, the number of transitions increase to infinity, and then it converges to the stationary distribution π^* given by **W.2**. Further as already mentioned, by equations (1) and (3), it also represents the Markov chain for any μ_ϵ. Let S^{μ_ϵ} represent the state of ϵ-Markov process, immediately after the last ϵ-change before τ-arrival instance, A_τ. Note that $S^{\mu_\epsilon} = (Y_1, \mathbf{R}_1^s)$. As $\mu_\epsilon \to \infty$, the number of ϵ-arrivals within one τ-inter arrival time, $N^\epsilon(A_\tau) \to \infty$. Hence by **W.2** as $\mu_\epsilon \to \infty$ (rate of convergence independent of s_ϵ),

$$\mathbb{P}^{\mu_\epsilon}(S^{\mu_\epsilon} = (0,\mathbf{0})|s_\epsilon) \to \pi^*(0,\mathbf{0}) \text{ for any } s_\epsilon.$$

Choose further large $\bar{\mu}$ such that:

$$\mathbb{P}^{\mu_\epsilon}(S^{\mu_\epsilon} = (0,\mathbf{0})|s_\epsilon) \geq \pi^*(0,\mathbf{0})/2 \text{ for all } \mu_\epsilon \geq \bar{\mu} \text{ and all } s_\epsilon.$$

For all such μ_ϵ, when $z = (u, s_\epsilon) \in \mathcal{C}$ (see (29)), i.e., when $u \leq \bar{u}$:

$$\begin{aligned} P(z,A) &\geq P(z,\ A \cap \{(0,0,\mathbf{0})\}) = P(z,(0,0,\mathbf{0}))\mathbb{1}_{\{(0,0,\mathbf{0})\in A\}} \\ &\geq \kappa^{\mu_\epsilon}\mathbb{1}_{\{(0,0,\mathbf{0})\in A\}}, \text{ where} \\ \kappa^{\mu_\epsilon} &:= \mathbb{P}^{\mu_\epsilon}_{s_\epsilon}(\bar{u} + B_\tau < \Omega^{\mu_\epsilon}(A_\tau))\pi^*(0,\mathbf{0})/2 \\ &= \mathbb{P}^{\mu_\epsilon}_{s_\epsilon}(\Upsilon^{\mu_\epsilon}(\bar{u}+B_\tau) < A_\tau)\pi^*(0,\mathbf{0})/2 = \mathbb{E}^{\mu_\epsilon}_{s_\epsilon}\left[e^{-\lambda_\tau \Upsilon^{\mu_\epsilon}(\bar{u}+B_\tau)}\right]\pi^*(0,\mathbf{0})/2. \end{aligned}$$

As in Lemma 8 one can show that

$$\mathbb{E}^{\mu_\epsilon}_{s_\epsilon}\left[e^{-\lambda_\tau \Upsilon^{\mu_\epsilon}(\bar{u}+B_\tau)}\right] \to \mathbb{E}^{\mu_\epsilon}\left[e^{-\lambda_\tau(\bar{u}+B_\tau)/\nu_\tau}\right] \text{ as } \mu_\epsilon \to \infty \text{ uniformly over all } s_\epsilon,$$

and hence, if required by choosing even larger $\bar{\mu}$, we have

$$\kappa^{\mu_\epsilon} \geq \mathbb{E}^{\mu_\epsilon}\left[e^{-\lambda_\tau(\bar{u}+B_\tau)/\nu_\tau}\right]\pi^*(0,\mathbf{0})/4 := \kappa^\infty \text{ for all } \mu_\epsilon \geq \bar{\mu}.$$

Thus in all, $\mathcal{C}$ is a small set with uniform lower bound measure as below for all $\mu_\epsilon \geq \bar{\mu}$:

$$P(z,A) \geq \kappa^\infty \gamma(A) \text{ for all } z \in \mathcal{C},$$

where measure $\gamma(.)$ is a Dirac measure at $(0,0,\mathbf{0})$, i.e.,

$$\gamma(A) = \mathbb{1}_{\{(0,0,\mathbf{0})\in A\}} \text{ and for all events } A.$$

Without loss of generality start with $U_0 = 0$ and note $V(0) = 1$. By (Baxendale 2005, Theorem 1.1), with $\mathcal{P}$ representing the transition function of $\{Z_n\}_n$: a) there exists a unique stationary distribution π^*_τ; and b)

$$||\mathcal{P}^n\phi - \pi^*_\tau\phi|| := \left|\mathbb{E}^{\mu_\epsilon}_{s_\epsilon}[\phi(U_{n+1})] - \mathbb{E}^{\mu_\epsilon}[\phi(U_*)]\right| \leq r^n C \text{ for all } n, s_\epsilon,$$

where constants $r < 1$ and $C < \infty$ are same for all $\mu_\epsilon \geq \bar{\mu}$ (because constants κ^∞, β, K_Δ are the same for all such μ_ϵ), and for any function ϕ that is dominated by V:

$$|\phi(u)| \leq V(u) \text{ for all } u \text{ or equivalently } \varrho|\phi(u)| \leq V(u) \text{ for some constant } \varrho > 0 \text{ and for all } u. \tag{30}$$

In particular we are interested in stationary average, i.e., when $\phi(u) = u$. Consider the function

$$h(u) := u - V(u) = u - a^u \text{ and note that } h'(u) = 1 - a^u(ln(a)).$$

It is clear that $h'(u) > 0$ only for all $0 \leq u \leq \bar{u}$ where $a^{\bar{u}} = 1/ln(a)$ and $h'(u) < 0$ for $u > \bar{u}$. Thus eventually $V(u)$ dominates and hence $\phi(u) = u$ satisfies (30) and hence we have

$$|\mathbb{E}^{\mu_\epsilon}_{0,s_\epsilon}[U_{n+1}] - \mathbb{E}^{\mu_\epsilon}[U_*]| \leq r^n C V(0) = r^n C.$$

Note that the above upper bounds are uniform for all $\mu_\epsilon \geq \bar{\mu}$ and for all s_ϵ and we always start with $u = 0$. And then we have for any s_ϵ and n

$$|\mathbb{E}^{\mu_\epsilon}[U_*] - \mathbb{E}^{\infty}[U_*]| \leq \left|\mathbb{E}^{\mu_\epsilon}_{s_\epsilon}[U_n] - \mathbb{E}^{\mu_\epsilon}[U_*]\right| + \left|\mathbb{E}^{\infty}_{s_\epsilon}[U_n] - \mathbb{E}^{\infty}[U_*]\right| + \left|\mathbb{E}^{\mu_\epsilon}_{s_\epsilon}[U_n] - \mathbb{E}^{\infty}_{s_\epsilon}[U_n]\right|.$$

Given an $\varepsilon > 0$ choose n_ε large enough such that the sum of the first two terms is less than $\varepsilon/2$ (for any $\mu_\epsilon \geq \bar{\mu}$ and for any s_ϵ), further using ergodicity of the limit system represented by evolution:

$$U^\infty_{n+1} = \left(U^\infty_n + B_{\tau,n+1} - \nu_\tau A_{\tau,n+1}\right)^+.$$

Finally for this $n = n_\varepsilon$,

$$\begin{aligned}
\left|\mathbb{E}^{\mu_\epsilon}_{s_\epsilon}[U_n] - \mathbb{E}^{\infty}[U_n]\right| &= \left|\mathbb{E}_{s_\epsilon}[U^{\mu_\epsilon}_n] - \mathbb{E}[U^\infty_n]\right| \\
&= \left|\mathbb{E}^{\mu_\epsilon}_{s_\epsilon}\left[(U^{\mu_\epsilon}_{n-1} + B_{\tau,n} - \Omega^{\mu_\epsilon}(A_{\tau,n}))^+ - (U^\infty_{n-1} + B_{\tau,n} - \nu_\tau A_{\tau,n})^+\right]\right| \\
&\leq \mathbb{E}^{\mu_\epsilon}_{s_\epsilon}\left[\left|(U^{\mu_\epsilon}_{n-1} + B_{\tau,n} - \Omega^{\mu_\epsilon}(A_{\tau,n})) - (U^\infty_{n-1} + B_{\tau,n} - \nu_\tau A_{\tau,n})\right|\right] \\
&\leq \mathbb{E}^{\mu_\epsilon}_{s_\epsilon}\left[\left|\left(U^{\mu_\epsilon}_{n-1} - U^\infty_{n-1}\right)\right| + \left|\left(\Omega^{\mu_\epsilon}(A_{\tau,n})\right) - \nu_\tau A_{\tau,n}\right|\right] \\
&\quad\vdots \\
&\leq \sum_{k \leq n_\varepsilon} \mathbb{E}^{\mu_\epsilon}_{s_\epsilon}\left[\left|\Omega^{\mu_\epsilon}(A_{\tau,k}) - \nu_\tau A_{\tau,k}\right|\right].
\end{aligned}$$

Using similar arguments to those in the proof of Theorem 1, it can be shown that $\Omega^{\mu_\epsilon}(A_{\tau,k}) \to \nu_\tau A_{\tau,k}$ almost surely as $\mu_\epsilon \to \infty$, for each k. Moreover, $\Omega^{\mu_\epsilon}(A_{\tau,k}) \leq A_{\tau,k}$, with $\mathbb{E}[A_{\tau,k}] < \infty$, implying we also have uniform integrability:

$$\mathbb{E}^{\mu_\epsilon}_{s_\epsilon}\left[\left|\Omega^{\mu_\epsilon}(A_{\tau,k}) - \nu_\tau A_{\tau,k}\right|\right] \to 0 \quad \text{as } \mu_\epsilon \to \infty.$$

It follows that for μ_ϵ large enough,

$$\left|\mathbb{E}^{\mu_\epsilon}_{s_\epsilon}[U_n] - \mathbb{E}^{\infty}[U_n]\right| \leq \varepsilon/2.$$

Since ε is arbitrary, this proves the convergence in expectation.

Choosing $\phi(u) := z^u$ (for any $0 < z < 1$) in (30) we obtain the convergence of MGF and hence the convergence in distribution as in the proof of Theorem 2. ■

Lemma 10 *Consider any $1 < a < \bar{a}$ of* **W.1**. *Let $\Omega^{\mu}_{n+1} := \Omega^{\mu}_{n}(A_{\tau,n+1})$, be the server capacity totally available to τ-customer for time duration $A_{\tau,n+1}$. Let $\Omega^{\infty}_{n+1} := A_{\tau,n+1}\nu_\tau$, be the total server capacity available to τ-customer in time $A_{\tau,n+1}$ for the limit system. For any given $\varepsilon > 0$, there exists a $\bar{\mu} < \infty$ and $\bar{u} < \infty$ such that for any $\mu_\epsilon \geq \bar{\mu}$ and for any $u \geq \bar{u}$ we have*

$$\left|\mathbb{E}_{s_\epsilon}\left[a^{\left(u+B_{\tau,n}-\Omega^{\mu}_{n+1}\right)^+}\right] - \mathbb{E}\left[a^{\left(u+B_{\tau,n}-\Omega^{\infty}_{n+1}\right)}\right]\right| \leq a^u\varepsilon \text{ for any } s_\epsilon.$$

Proof: By Theorem 5, equation (13) and exactly as in the proof of Theorem 1 one can prove that $\Omega^{\mu}_{n+1} \to \Omega^{\infty}_{n+1}$ almost surely as $\mu \to \infty$.
Further as in Lemma 8 of Appendix B, for any given $\varepsilon > 0$, there exists a $\bar{\mu} < \infty$ such that for any $\mu \geq \bar{\mu}$ (we use shorter notation $\mu = \mu_\epsilon$ and it is only meant for ϵ-customers) and any u

$$\begin{aligned}
&\sup_{s_\epsilon}\left|\mathbb{E}_{s_\epsilon}\left(u + B_{\tau,n} - \Omega^{\mu}_{n+1}\right)^+ - \mathbb{E}\left(u + B_{\tau,n} - \Omega^{\infty}_{n+1}\right)^+\right| \\
&\qquad \leq \sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left|\left(u + B_{\tau,n} - \Omega^{\mu}_{n+1}\right)^+ - \left(u + B_{\tau,n} - \Omega^{\infty}_{n+1}\right)^+\right| \\
&\qquad \leq \sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left|\Omega^{\mu}_{n+1} - \Omega^{\infty}_{n+1}\right| \leq \varepsilon/2. \qquad (31)
\end{aligned}$$

Define $\xi^\mu := \left(u + B_{\tau,n} - \Omega^{\mu}_{n+1}\right)^+$, $\xi^\infty := \left(u + B_{\tau,n} - \Omega^{\infty}_{n+1}\right)^+$ and then for any $C < \infty$:

$$\begin{aligned}
\sup_{s_\epsilon}\left|\mathbb{E}_{s_\epsilon}a^{\xi^\mu} - \mathbb{E}a^{\xi^\infty}\right| &\leq \sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left[\left|a^{\xi^\mu} - a^{\xi^\infty}\right|; |\Omega^\mu - \Omega^\infty| > C\right] \\
&\quad + \sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left[\left|a^{\xi^\mu} - a^{\xi^\infty}\right|; |\Omega^\mu - \Omega^\infty| \leq C\right] \\
&\leq \sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left[\left|a^{\xi^\mu} - a^{\xi^\infty}\right|; |\Omega^\mu - \Omega^\infty| > C\right] \\
&\quad + \sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left[a^{\xi^\infty}\left|a^{\xi^\mu - \xi^\infty} - 1\right|; |\Omega^\mu - \Omega^\infty| \leq C\right]
\end{aligned}$$

because $|a^x - 1|$ is uniformly Lipschitz on bounded interval $[0, C]$

$$\begin{aligned}
&\leq 2\sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left[a^{u+B_\tau}; |\Omega^\mu - \Omega^\infty| > C\right] \\
&\quad + \tilde{C}(C)\sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left[a^{u+B_\tau}\left|\xi^\mu - \xi^\infty\right|; |\Omega^\mu - \Omega^\infty| \leq C\right].
\end{aligned}$$

Thus and using Markov inequality,

$$\begin{aligned}
&a^{-u}\sup_{s_\epsilon}\left|\mathbb{E}_{s_\epsilon}a^{\xi^\mu}-\mathbb{E}a^{\xi^\infty}\right| \\
&\quad\le 2\mathbb{E}[a^{B_\tau}]\sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left[|\Omega^\mu-\Omega^\infty|>C\right]+\tilde{C}(C)\sup_{s_\epsilon}\mathbb{E}\left[a^{B_\tau}\right]\mathbb{E}_{s_\epsilon}\left[\left|\Omega^\mu-\Omega^\infty\right|\right] \\
&\quad\le\left(\frac{2}{C}+\tilde{C}(C)\right)\mathbb{E}\left[a^{B_\tau}\right]\sup_{s_\epsilon}\mathbb{E}_{s_\epsilon}\left[\left|\Omega^\mu-\Omega^\infty\right|\right]\le\varepsilon/2,\ \text{for all } \mu\ge\bar{\mu},
\end{aligned}$$

if sufficiently large $\bar{\mu}$ is chosen by (31). Now

$$\begin{aligned}
a^{-u}\sup_{s_\epsilon}&\left|\mathbb{E}_{s_\epsilon}a^{\left(u+B_{\tau,n}-\Omega^\mu_{n+1}\right)^+}-\mathbb{E}a^{\left(u+B_{\tau,n}-\Omega^\infty_{n+1}\right)}\right| \\
&\le a^{-u}\sup_{s_\epsilon}\left|\mathbb{E}_{s_\epsilon}a^{\left(u+B_{\tau,n}-\Omega^\mu_{n+1}\right)^+}-\mathbb{E}a^{\left(u+B_{\tau,n}-\Omega^\infty_{n+1}\right)^+}\right| \\
&\quad+a^{-u}\left|\mathbb{E}a^{\left(u+B_{\tau,n}-\Omega^\infty_{n+1}\right)^+}-\mathbb{E}a^{\left(u+B_{\tau,n}-\Omega^\infty_{n+1}\right)}\right| \\
&\le\varepsilon/2+\varepsilon/2=\varepsilon.
\end{aligned}$$

In the above the last term is so upper bounded because, as $u\to\infty$, by integrability:

$$\begin{aligned}
a^{-u}&\left|\mathbb{E}a^{\left(u+B_{\tau,n}-\Omega^\infty_{n+1}\right)^+}-\mathbb{E}a^{\left(u+B_{\tau,n}-\Omega^\infty_{n+1}\right)}\right| \\
&=a^{-u}\mathbb{E}\left(a^{A_{\tau,n+1}\nu_\tau-B_{\tau,n}-u};\ u<A_{\tau,n+1}\nu_\tau-B_{\tau,n}\right) \\
&\le a^{-u}\mathbb{E}\left(a^{A_{\tau,n+1}\nu_\tau-B_{\tau,n}};\ u<A_{\tau,n+1}\nu_\tau-B_{\tau,n}\right)\to 0. \quad\blacksquare
\end{aligned}$$

Appendix D: Proofs for eager customers with limited patience

Lemma 11 *When* **A.3** *is replaced by* **A.3′** *and if additionally one of the assumptions of* **A.4** *is satisfied, one can construct an upper bound* Θ^{μ_ϵ} *which uniformly dominates any partial* ϵ*-busy period uniformly over all* ϵ*-states* $s_\epsilon=(y_s,y_w,\mathbf{r})$ *and converges to zero as* $\mu_\epsilon\to\infty$ *exactly as in Lemma 6.*

Proof: The construction of Θ^{μ_ϵ} is the same for the first two conditions and is similar[15] to that in Lemma 6. We first define Θ^1 and then $\Theta^{\mu_\epsilon}=\Theta^1/\mu_\epsilon$. Consider a fictitious $M/G/\infty$ (infinite server) queue when started with K initial customers, whose service times are given by: a) $\tilde{b}^1$, the upper bound on the time spent by any ϵ-customer in the system (with $\mu_\epsilon=1$), when **A.4(a)** is satisfied; b) the sum Γ^1+B_ϵ, where Γ^{μ_ϵ} is the reneging time, when **A.4(b)** is satisfied. Then Θ^1 is the busy period of the above fictitious queue. By using appropriate coupling rules as in Lemma 6, Θ^{μ_ϵ} almost surely dominates $\tilde{\Psi}(s_\epsilon)$ for any initial state s_ϵ. This is because the fictitious $M/G/\infty$ again accepts

[15] For the case with **A.4(a)** assumption, the proof follows exactly as in Lemma 6. We provide an alternate proof which also works for assumption **A.4(b)**.

all the arrivals, any customer accepted in original system spends longer time in the fictitious $M/G/\infty$ queue etc.

We first observe few simple facts. The busy period $\mathcal{B}_K$ of an $M/G/\infty$ when started with K customers and with general IID service times $\{B_{G,n}\}_n$ is almost surely larger than that of an $M/G/\infty$ queue ($\hat{\mathcal{B}}_1$) when started with 1 customer and whose IID service times are given by the maximum of K independent service times of the original $M/G/\infty$ queue, i.e., if service time of n-th customer in upper $M/G/\infty$ queue equals $\hat{B}_{G,n} := \max_{1\le j\le K} B_{G,(n-1)K+j}$. Therefore (Takács (1962)),

$$\mathbb{E}[\mathcal{B}_K] \le \mathbb{E}[\hat{\mathcal{B}}_1] = \frac{e^{\lambda_\epsilon \mathbb{E}[\hat{B}_{G,1}]} - 1}{\lambda_\epsilon} \le \frac{e^{K\lambda_\epsilon \mathbb{E}[B_{G,1}]-1}}{\lambda_\epsilon} < \infty.$$

In a similar way

$$\mathbb{E}[\mathcal{B}_K^2] \le \mathbb{E}[\hat{\mathcal{B}}_1^2] < \infty.$$

Thus the first two moments of Θ^1 are bounded. Hence the conclusions of Lemma 6 are true under **A.4(a)** and **A.4(b)**.

When A.4(c) is satisfied: Here Θ^{μ_ϵ} is the busy period of a fictitious $M/G/K_1/2K$ queue, and when started with K customers. We would couple the inter arrival times, job sizes etc., as before in both the systems.

System with $2K$ servers (each of same capacity as before), when started with K ϵ-customers and: a) if y_s number of ϵ-customers are deriving service at the beginning of $\tilde{\Psi}(s_\epsilon)$, the service times of those y_s customers also equal the service time requirements of the first y_s customers of the $2K$ system and these are independent of residual service times[16]; b) the service times of the remaining $(K_1 - y_s)$ ϵ-customers are independent copies of the exponential random variable with the same parameter; c) further inter arrival times and service times of all the new ϵ-customers coincide with that in the original system; and d) if a customer is not accepted in original system, we consider an independent service time for that customer. With this construction, an ϵ-customer departure during $\tilde{\Psi}(s_\epsilon)$, of the original system definitely marks a departure in $2K$ system also, any customer accepted in original system is also accepted in the $2K$ system (it has double $2K$ holding capacity). Thus the busy period Θ^{μ_ϵ} of the $2K$ system dominates the residual ϵ-busy period $\tilde{\Psi}(s_\epsilon)$, irrespective of the state $s_\epsilon = (y_s, y_w)$ of the original system at the start of $\tilde{\Psi}(s_\epsilon)$.

Rest of the arguments are as in Lemma 6. ■

[16] As before this replacement with independent copies does not change the original system stochastically.