Buffer Occupancy and Delay Asymptotics in Multi-buffered Systems with Generalized Processor Sharing Handling a Large Number of Independent Traffic Streams[*]

Constantinos Kotopoulos
Department of Mathematics
University of Essex
Colchester CO4 3SQ, U.K.

Ravi R. Mazumdar
School of ECE
Purdue University
W. Lafayette, IN 47907-1285, USA

March 6, 2002

Abstract

In this paper we study a multi-buffer Generalized Processor Sharing (*GPS*) scheduled system in the so-called *many sources* regime resulting from multiplexing a large number of independent stationary input streams in each of the buffers. The notion of largeness is via the notion of scaling by which the buffer occupancy and the server speed are scaled in proportion to the number of inputs in such a way as to make the resulting system a scaled version of a nominal system. In particular, we obtain bounds on the tail distribution of the buffer occupancies as well as the virtual delays in each buffer. We provide sufficient conditions on the inputs for the estimates to be asymptotically equivalent as the number of inputs grow. Finally, we provide validation of the accuracy of the estimates via simulation. It is worth noting that the results hold even in the case of heavy-tailed inputs with bounded rates.

[*] This research has been supported in part by grants from the Centre National d'Etudes des Télècommunications, France Telecom under the Consultations Thèmatiques programme (CTI) and from the NSF through grant ANI-0099137.

Introduction

Classical fair queueing schemes have been considered in the context of data networks to achieve some fairness amongst different connections sharing the transmission capacity of a link by implementing the round robin service where the packets of contending customers are served in cyclic order. The proportion of the service offered to a session can further be modulated via the introduction of weight coefficients, giving rise to Weighted Fair Queueing disciplines.

Weighted Fair Queueing, also referred to a Generalized Processor Sharing (GPS), was first introduced in the context of broadband integrated data networks by Demers, Keshav, and Shenker [19], where it is assumed that packets are infinitely divisible (fluid flow approach). It has been further analyzed by Parekh and Gallager [31, 32]. Their work showed that if the input streams are regulated, i.e. they conform to a pre-specified envelope known as a (σ, ρ)-regulator (see [15]), then end-to-end delay bounds can be guaranteed. Yaron and Sidi [43],[42] obtained bounds on the waiting time in a GPS buffer when input processes satisfy an Exponentially Bounded Burtsiness (EBB) property, i.e. the probability of exceeding the envelope is exponentially decaying as t grows. Going further along this line of investigations, Zhang *et al* [45] obtained refinements of these bounds.

The fairness of the GPS schedule also has the nice property of bandwidth protection. It guarantees a minimum bandwidth or rate to each buffer when all buffers have work in them and thus buffers with high loads will not unduly hog all the bandwidth. For this reason, the GPS schedule is gaining acceptance in the context of providing QoS guarantees to traffic streams which require differing QoS such as loss and delay characteristics.

An exact stochastic analysis of buffers with GPS schedules under general assumptions on the inputs is very difficult. This is because the analysis results in boundary value problems in queues [9]. The resolution of such models is via a Riemann-Hilbert technique [22] or Wiener-Hopf technique [7] which is daunting even for the case of 2 buffers. Recently Guillemin *et al* [24] obtained an explicit solution for Laplace transform of the buffer contents in each queue for the 2-queue GPS systems with M/G types of inputs (Poisson arrivals with G service times) via a Wiener-Hopf technique . Extension of the results to more than 2 buffers is extremely difficult.

While the exact analysis of GPS queues is difficult there has been much more success in obtaining the tail distributions of the buffer content via large deviation techniques. The tail distributions are of particular relevance in the context of QoS. This is because the requirements are usually specified in terms of loss probabilities which are required to be very low or in terms of the remote quantiles of the delay distributions. There are two different approaches, the first is the *large buffer* asymptotic and the second is the *many sources* asymptotic. The latter is particularly relevant when a large number of streams are present where each stream can use a small fraction of the total server capacity. The former approach has been better studied due to powerful sample-path large deviations techniques available [18].

Results on the tail of the buffer occupancy have focussed on a two buffer *GPS* system in the large buffer asymptotic regime. Under the assumption that the arrival processes satisfy the Gärtner-Ellis condition and thus an Large Deviation Principle (LDP), DeVeciana and Kesidis in [21] obtain an upper bound for the tail of the workload distribution in each queue. Zhang in [45] generalized the latter result showing that in fact it is exact by imposing stronger assumptions on the input processes (see [18] and [11]) and employing the sample path Large Deviation Principle (sp-LDP). Courcoubetis and Weber [14] derive the decay rate of the queue length distribution by solving an optimal control problem under strong assumptions on the arrival processes. This result was extended by O' Connell in [30] where once again the sp-LDP was used. Recently Bertsimas *et. al.* in

[4] extended the latter results in the case that the service rate is stochastic instead of deterministic solving an optimal control problem under assumptions on the arrival processes similar to those posed in [45]. Assuming that the arrival processes satisfy a sample path large deviation principle, based on a contraction principle Massoulié [28] derived the optimization problem associated with the rate function for the exact logarithmic queue length decay rate in an M-queue *GPS* system when $M \geq 2$. It is then shown that in the particular case where the mean arrival rate to every queue is less than the minimum guaranteed service to them by the *GPS* server, the latter result takes a tractable form.

In this paper we consider the *many sources* case with $M \geq 2$ buffers which are served according to the GPS discipline by a server of rate NC where N is a large integer. The many sources asymptotic is relevant from the point of view of studying multiplexing effects. In a recent paper, [26], we studied the case of a 2 buffer system in detail. In the 2 buffer case the dynamics of the system are much easier to interpret. However, in a system with more than 2 bufers the dynamics are much more complicated and the results for the 2-buffer case cannot be directly extended. Hence, we begin by studying the dynamics of the system in a detailed manner. The key result here is the definition and characterization of an ultimately stable configuration. This allows us to identify regions on which estimations need to be made to obtain tight bounds on the tail distribution of the buffer occupancy in each buffer. We then study some special cases which allows us to draw insights on the so-called critical time scales as well as conditions for the bounds to be asymptotically equivalent.

The organization of this paper is as follows: In section 1 we introduce our model along with all necessary notation and some well known large deviation results for a large sum of independent r.v.'s. Then, in Section 2, we analyze the dynamics of the M-queue *GPS* system in detail which helps us identify the events which play an important role in the estimation of the tail distribution. This is carried out in Section 3 where we upperbound the loss probability by a multiple infinite sum of events. Then we present a way to evaluate the latter events. It is worth noting that sample path conclusions on the most probable way losses occur can be deduced which significantly facilitate our analysis and make our result tractable. These conclusions resemble the ones on the trajectory with the lowest cost leading to overflow presented in theorem 2 in [28]. In Section 4 we show that in fact one of the events in the sum that upperbounds the loss probability dominates exponentially to the rest and it is the one that provides us with the required upper bound. In section 5 the cases where the upper bound becomes asymptotically exact are discussed and some extra sample path results are presented for arrival processes that satisfy a certain assumption. Section 6 discusses the delay distribution case and shows that the delay tails are closely related to the previous results. We conclude with some numerical evidence on the accuracy of the analytic bounds in comparison to simulations.

1 Preliminaries: Model and basic results

We consider a queueing system in discrete-time (slots) with a single server of capacity C attending M buffers of size B_i where $i = 1, 2, \ldots, M$ denotes a given class with its own QoS requirement. It is assumed that each buffer is accessed by n_i sources, $i = 1, 2, \ldots, M$. Thus, the triplet $(C, \{B_i, n_i\}_{i=1}^{M})$ can be thought as the nominal capacity, buffer sizes and number of sources which access the corresponding buffer. We suppose that at any discrete time instant, denoted by $n \in \mathcal{Z}$, (where $\mathcal{Z}$ stands for the set of integers) a finite number of bits $0 \leq \lambda_{i,1,n} \leq K_i$ is transmitted from a typical i-class source into the buffer i, where K_i is referred to as the peak rate. Throughout, the terms buffer and queue will be used interchangeably. The buffers are drained at a rate defined by the *G.P.S.* policy. According to the latter service scheme a weight ϕ_i is associated to each queue guaranteeing a minimum service rate $C\phi_i$ to it even in the worst case when all queues are non-empty. If there are empty queues then their minimum guaranteed service rate is distributed over all the busy queues in proportion to their ϕ's. If now there is insufficient space in the buffer then the corresponding bits which cannot be admitted are lost. It is assumed that the instantaneous cell emission process $\{\lambda_{i,1,n}\}_n$ for each $i = 1, 2, \ldots, M$ form a stationary and ergodic process with $E[\lambda_{i,1,n}] := \Lambda_i$. Moreover, we suppose that the arrival processes are independent among the classes and within them.

Let $r_{i,t}$ be the rate at which queue i, $i = 1, 2 \ldots, M$ is served at the instant t. The GPS scheduling scheme is one in which every queue is associated a weight, denoted by ϕ_i with $0 \leq \phi_i \leq 1$, $i = 1, 2, \ldots, M$ where $\sum_{i=1}^{M} \phi_i = 1$ and (in the continuous-time case),

$$r_{i,t} := C \frac{\phi_i \mathbf{1}_{\{\text{queue } i \text{ is not empty at } t\}}}{\phi_i + \sum_{\substack{j=1 \\ j \neq i}}^{M} \phi_j \mathbf{1}_{\{\text{queue } j \text{ is not empty at } t\}}} \ . \tag{1.1}$$

where $\mathbf{1}_{\{A\}}$ denotes the indicator function of the event A.

In other words, if all the M queues are busy (also called backlogged queues) then queue i receives service at the rate $r_{i,t} = C\phi_i$ for all $i = 1, 2, \ldots, M$, which is called the *minimum guaranteed service rate* to queue i. If some queues are empty then their minimum guaranteed service rates are shared amongst the backlogged queues according to their weights. Thus, in the best case scenario a given buffer can be served at a rate C. In general, the instantaneous service rate offered to a given buffer will lie between its minimum guaranteed rate and the server capacity.

It readily follows that the server is work conserving which means that it never idles when there is backlogged workload in the queues waiting to be served. It is also assumed that the workload in any of the M queues comprising the system is processed according to the FIFO order.

In this paper we consider only discrete time models. This means that the time is divided into discrete slots $(t-1, t)$ for $t \in \mathbf{Z}$. To this end, in the slotted time setting let us denote by $S_{i,t}$ the service queue i receives in the time interval $(t-1, t)$ i.e. one slot, according the GPS scheduler. Then,

$$S_{i,t} := \int_{t-1}^{t} r_{i,t} dt \ . \tag{1.2}$$

The system is depicted in the figure below.

Remark 1.1 *Although we have taken the unit of the input and service rates as bits, the analysis carries through if we consider the units as fixed length packets.*

In this paper we study the particular case when each buffer i is fed by the superposition of a large number of class-i sources, say Nn_i. The parameter N will act as a scaling parameter on the system and will be a measure of size as well as an error control factor. It is also assumed that the service rate is NC and the buffer sizes are NB_i. Therefore, the true system can be considered as an N-fold scaling of a nominal system.

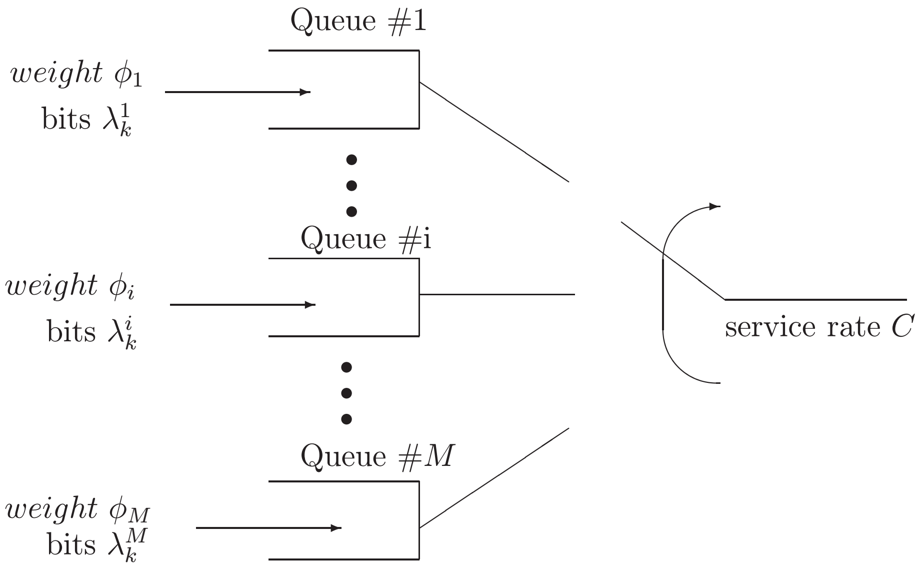

Figure 1: GPS queue.

Let us denote by $\Lambda_{i,1}(s,t)$, the total number of bits which arrive into the buffer B_i (from a typical class-i source) in the interval $[s,t)$ i. e.

$$\Lambda_{i,1}(s,t) = \sum_{n=s}^{t-1} \lambda_{i,1,n}. \tag{1.3}$$

Let $\Lambda_i^N(s,t)$ be the aggregate number of bits produced by the superposition of Nn_i class-i sources in the interval $[s,t)$. Then:

$$\Lambda_i^N(s,t) = \sum_{j=1}^{Nn_i} \Lambda_{i,j}(s,t). \tag{1.4}$$

By assumption $\{\Lambda_i^N\}$ is a stationary increment process. We assume that for all $i = 1,2,...,M$,

$$\Lambda_i n_i < C\phi_i \tag{1.5}$$

and thus the system is stable i. e.,

$$\sum_{i=1}^{M} \Lambda_i n_i < C. \tag{1.6}$$

In the case when $\sum_{i=1}^{M} B_i = \infty$ this condition guarantees the existence of a stationary workload (or queue length) process.

Furthermore, it is assumed that $\sum_{i=1}^{M} K_i n_i > C$ otherwise losses cannot occur.

Let us denote by $\phi_{i,t}(h)$ the moment generating function (m.g.f.) of $\Lambda_{i,1}(0,t)$ defined as

$$\phi_{i,t}(h) = \mathbf{E}[e^{h\Lambda_{i,1}(0,t)}] < \infty \text{ for } h,t < \infty \quad . \tag{1.7}$$

where the existence of the m.g.f. for finite t, h is due to the assumption that the instantaneous arrival rates are bounded.

From now on we will suppress the subscript 1 from the notation of $\Lambda_{i,1}(0,t)$ and $\lambda_{i,1,n}$ for sake of brevity. Instead we will denote these quantities by $\Lambda_i(0,t)$ and $\lambda_{i,n}$ respectively.

The tail distribution of $\Lambda_i^N(0,t)$ plays a fundamental role in determining the buffer asymptotic and hence we state the following results which characterize the density and the asymptotic tail distribution of $\Lambda_i^N(0,t)$ for large N. From the independence of the traffic streams, the key results which are of use are the local limit large deviations result due to Bahadur-Rao [5] and a local limit theorem for densities due to Petrov [35]. See also Korolyuk *et al* [25].

Lemma 1.1 *Let $\Lambda_i^N(0,t)$ as defined in 1.4 and $\phi_{i,t}(h)$ given by 1.7.*

Then as $N \to \infty$ uniformly for $n_i t\Lambda_i < u < n_i tK_i$ we have for the tail of $\Lambda_i^N(0,t)$,

$$P[\Lambda_i^N(0,t) \geq Nu] = \frac{e^{-NJ_{i,t}(u)}}{\tau_{i,t}(u)\sqrt{2\pi N\sigma_{i,t}^2(u)}}\left(1 + O(\frac{1}{N})\right), \tag{1.8}$$

where

$$J_{i,t}(u) = \tau_{i,t}(u)u - n_i \ln \phi_{i,t}(\tau_{i,t}(u)) > 0 \tag{1.9}$$

$\tau_{i,t}(u)$ is the unique solution to:

$$\frac{\phi'_{i,t}(\tau_{i,t}(u))}{\phi_{i,t}(\tau_{i,t}(u))} = \frac{u}{n_i} \tag{1.10}$$

and

$$\sigma_{i,t}^2(u) = n_i\left(\frac{\phi''_{i,t}(\tau_{i,t}(u))}{\phi_{i,t}(\tau_{i,t}(u)} - (\frac{u}{n_i})^2\right) > 0 \quad . \tag{1.11}$$

While for the density of $\Lambda_i^N(0,t)$, as $N \to \infty$ uniformly for any $u \in (0, n_i t\Lambda_i) \cup (n_i t\Lambda_i, tK_i)$,

$$P(\Lambda_i^N(0,t) \in [Nu, Nu + du)) := P^{\Lambda_i^N(0,t)}(Nu) = \frac{e^{-NJ_{i,t}(u)}}{\sqrt{2\pi N\sigma_{i,t}^2(u)}}\left(1 + O(\frac{1}{N})\right) \ , \tag{1.12}$$

where $J_{i,t}(u)$ and $\sigma_{i,t}^2(u)$ are given by 1.9 and 1.11 respectively.

Note that we introduce $P^{\Lambda_i^N(0,t)}(\cdot)$ to denote the density of the random variable $\Lambda_i^N(0,t)$. This notation will be used throughout.

Remark 1.2 *Equations 1.9-1.11 are well known in the theory of large deviations see for instance [18], [36], where $J_{i,t}(u)$ denotes the rate function and $\sigma_{i,t}^2(u)$ stands for the variance of the new exponentially centralized measure. It is also known that $\tau_{i,t}(u)$ is positive (negative) if $u > n_iE[\Lambda_i(0,t)]$ ($u < n_iE[\Lambda_i(0,t)]$) and it is zero if $u = n_iE[\Lambda_i(0,t)]$.*

Remark 1.3 *In the case where the arrival process $\lambda_{i,n}$ has lattice distribution with maximum span H, in formula 1.8 $\tau_{i,T_i}(u)$ in the denominator must be substituted by $(1 - e^{-H\tau_{i,T_i}(u)})/H$.*

The following remark caters for the case when $\tau(\cdot)$ is negative.

Remark 1.4 *In the sequel if $u < n_i\Lambda_i$ we will assume that $P[\Lambda_i^N(t,0) > Nu] = 1$ and thus $J_{i,t}(u) = 0$ which is a simple consequence of the Bahadur-Rao result. To see this write $P[\Lambda_i^N(t,0) > Nu] = 1 - P[\Lambda_i^N(t,0) \leq Nu] = 1 - O(e^{-N\alpha}) \approx 1$ for some positive α and $N \to \infty$.*

Let us denote by $X_{i,t}^N$ the workload present in queue i at the moment t^- for the infinite buffered system. Then the workload process $\{X_{i,t}^N\}_{t\in\mathcal{Z}}$, evolves as,

$$X_{i,t}^N = \max(0, X_{i,t-1}^N + \lambda_{i,t-1}^N - S_{i,t}^N) \tag{1.13}$$

where $S_{i,t}^N$ stands for the service available to queue i during $(t-1,t)$, in the context of the scaled system.

It can be readily shown (via Loynes' theorem [1]) that under the stability assumption 1.6, for the global system the above workload process in each queue converges to a stationary, ergodic workload process. The stationary workload at the queue i is given by:

$$X_{i,0}^N = \sup_{t\in\{1,2,3,\cdots\}} (\Lambda_i^N(-t,0) - S_i^N(-t,0)) \tag{1.14}$$

with $S_i^N(-t,0) := \sum_{s=-t+1}^{0} S_{i,s}^N$ be the amount of service available to queue i during time interval $(-t,0)$.

Our main objective is to estimate the probability that the workload in the M^{th} buffer will hit a predefined level say NB_M assuming it is infinite. This serves as a surrogate of the overflow probability when the buffer is of finite size NB_M.

By symmetry all the results for the M^{th} queue can be extended to any queue $i = 1,2,...,M-1$. Before this we study the dynamics of the M-queue infinite-buffered *GPS* systems which provides with the required insight to choose a tight upper which will exploit the structural properties of the *GPS* scheme. Finally, note that in the sequel the symbols $\succeq$ and $\preceq$ will be used to denote inequality at the corresponding direction in the asymptotic sense. That is for instance $f(x) \succeq g(x)$ must be interpreted as: there exists $\epsilon(x) \to 0^+$ as $x \to \infty$ such that $f(x) \geq g(x)(1-\epsilon(x))$.

2 The dynamics of the M-queue GPS systems

Consider a queueing system possessing a single server with job processing speed C (i.e. C units of work per unit of time) which serves M queues, figure 1.

Let us first introduce some notation that will be used henceforth in this paper. Denote $\mathcal{M} := \{1,2,\ldots,M\}$ and for any $\mathcal{S} \subseteq \mathcal{M}$ let $\mathcal{S}^* := \mathcal{M}\backslash\mathcal{S}$ and let $\mid \mathcal{S} \mid$ denote the cardinality of the set $\mathcal{S}$. We will only use the calligraphic characters to denote sets with an exception when we are dealing with a set containing a single element. To this end, in the sequel for any $i \in \mathcal{M}$, i^* and i will denote $\{i\}^*$ and $\{i\}$ respectively. The set of all subsets of $\mathcal{S} \subseteq \mathcal{M}$ will be denoted by $\mathcal{F}(\mathcal{S})$ (which includes $\{\emptyset, \mathcal{S}\}$).

Consider an M-queue GPS system which is empty at 0^-. In the slotted time setting arrivals are assumed to take place in the beginning of a slot and thus suppose that at time 0, $\lambda_{i,0} \geq 0$ amount of work arrives at queue $i \in \mathcal{M}$. Define $a_i := C\phi_i - \lambda_{i,0}$, $i \in \mathcal{M}$ and $\lambda := (\lambda_{1,0}, \lambda_{2,0}, \ldots, \lambda_{M,0})$ where in general λ is a random vector. In the sequel we restrict ourselves to a single time slot i.e. $(0,1)$ and we assume that λ is a given constant vector.

Definition 2.1 *Define $\mathcal{S}_\lambda = \cup_{i=1}^{L_\lambda} \mathcal{I}_{i,\lambda}$ with,*

$$\mathcal{I}_{1,\lambda} := \{i \in \mathcal{M} : a_i \geq 0\}$$

$$\mathcal{I}_{k,\lambda} := \left\{ i \in \{\cup_{j<k}\mathcal{I}_{j,\lambda}\}^* : a_i + \frac{\phi_i}{\sum_{m \in \{\cup_{j<k}\mathcal{I}_{j,\lambda}\}^*} \phi_m} \sum_{l \in \cup_{j<k}\mathcal{I}_{j,\lambda}} a_l \geq 0 \right\} \tag{2.1}$$

$$L_\lambda := \min\{j \in \{0,1,\ldots,M\} \ : \ \mathcal{I}_{j+1,\lambda} = \emptyset\} \ .$$

The proof of the next lemma provides an analytic demonstration of the way the GPS scheduler works.

Lemma 2.1 *In an $M-$queue GPS system which is empty at time 0^- with arrivals λ at time 0, the set $\mathcal{S}_\lambda$ contains all queues empty at 1^- and only these. Hence, $\mathcal{S}_\lambda$ defines a unique partition of $\mathcal{M}$ and moreover, $\forall i \in \mathcal{S}_\lambda^*$,*

$$a_i + \frac{\phi_i}{\sum_{m \in \mathcal{S}_\lambda^*} \phi_m} \sum_{l \in \mathcal{S}_\lambda} a_l < 0 \ . \tag{2.2}$$

Proof : We will prove the lemma by showing that the amount of service a queue $i \in \mathcal{I}_{k,\lambda}$, $k = 1, 2, \ldots, L_\lambda$ obtains in $(0,1)$ is given by,

$$C\phi_i + \frac{\phi_i}{\sum_{m \in \{\cup_{j<k}\mathcal{I}_{j,\lambda}\}^*} \phi_m} \sum_{l \in \cup_{j<k}\mathcal{I}_{j,\lambda}} a_l \ . \tag{2.3}$$

The proof is based on induction which mimics the service allocation procedure of a GPS server during the time interval $(0,1)$.

First step

At the first step all queues are offered the minimum service guaranteed to them by the GPS discipline in $(0,1)$ i.e. $C\phi_i$. Now, if there are queues i for which $a_i \geq 0$, these queues trivially would be empty at 0. By definition the latter queues comprise $\mathcal{I}_{1,\lambda}$. If either $\mathcal{I}_{1,\lambda} = \emptyset$ or $\mathcal{I}_{1,\lambda} \equiv \mathcal{M}$ the procedure of service allocation terminates since either there is no extra service to be distributed among the queues with backlogged workload or all queues are already empty. In this case indeed $\mathcal{S}_\lambda$ contains the empty queues at 1^- and only these. Otherwise, we proceed to the second step.

Second step

Now the service left unused in $(0,1)$ by the queues in $\mathcal{I}_{1,\lambda}$ equals $\sum_{l \in \mathcal{I}_{1,\lambda}} a_l \geq 0$ and it will be redistributed among all queues in $\mathcal{I}_{1,\lambda}^*$ in proportion to their weights (see also 1.1). In particular, a queue $i \in \mathcal{I}_{1,\lambda}^*$ receives,

$$\frac{\phi_i}{\sum_{m \in \mathcal{I}_{1,\lambda}^*} \phi_m} \sum_{l \in \mathcal{I}_{1,\lambda}} a_l \geq 0$$

portion from the service left over from the queues in $\mathcal{I}_{1,\lambda}$. Recalling that every queue $i \in \mathcal{I}_{1,\lambda}^*$ has in addition been given $C\phi_i$ service in $(0,1)$ we deduce that in aggregate all the latter queues receive the following amount of service in $(0,1)$,

$$C\phi_i + \frac{\phi_i}{\sum_{m \in \mathcal{I}_{1,\lambda}^*} \phi_m} \sum_{l \in \mathcal{I}_{1,\lambda}} a_l \ . \tag{2.4}$$

Hence, all queues in $\mathcal{I}_{2,\lambda}$ will have no backlogged workload at the instant 1^-. If either $\mathcal{I}_{2,\lambda} = \emptyset$ or $\mathcal{M} \equiv \mathcal{I}_{1,\lambda} \cup \mathcal{I}_{2,\lambda}$ then the procedure terminates and it is apparent that all queues which are empty at 1^- and only these are contained in $\mathcal{S}_\lambda$. Otherwise, we continue with the next step.

Third step

Up to this moment any queue $i \in \{\mathcal{I}_{1,\lambda} \cup \mathcal{I}_{2,\lambda}\}^*$ received $C\phi_i$ amount of service in $(0,1)$ and $\frac{\phi_i}{\sum_{m \in \mathcal{I}_{1,\lambda}^*} \phi_m} \sum_{l \in \mathcal{I}_{1,\lambda}} a_l$ due the service left unused by the queues in $\mathcal{I}_{1,\lambda}$. In addition any queue $i \in \{\mathcal{I}_{1,\lambda} \cup \mathcal{I}_{2,\lambda}\}^*$ receives

$$\frac{\phi_i}{\sum_{m \in \left\{\mathcal{I}_{1,\lambda} \cup \mathcal{I}_{2,\lambda}\right\}^*} \phi_m} \left(\sum_{j \in \mathcal{I}_{2,\lambda}} a_j + \frac{\sum_{j \in \mathcal{I}_{2,\lambda}} \phi_j}{\sum_{m \in \mathcal{I}_{1,\lambda}^*} \phi_m} \sum_{l \in \mathcal{I}_{1,\lambda}} a_l \right) \geq 0$$

amount of service in $(0,1)$, from the queues in $\mathcal{I}_{2,\lambda}$. Thus, the aggregate service offered to a queue $i \in \{\mathcal{I}_{1,\lambda} \cup \mathcal{I}_{2,\lambda}\}^*$ in $(0,1)$ equals,

$$C\phi_i + \frac{\phi_i}{\sum_{m \in \left\{\mathcal{I}_{1,\lambda} \cup \mathcal{I}_{2,\lambda}\right\}^*} \phi_m} \sum_{l \in \mathcal{I}_{1,\lambda} \cup \mathcal{I}_{2,\lambda}} a_l \ . \tag{2.5}$$

Then all queues $i \in \mathcal{I}_{3,\lambda}$ will be empty at 1^-. If either $\mathcal{I}_{3,\lambda} = \emptyset$ or $\mathcal{M} \equiv \mathcal{I}_{1,\lambda} \cup \mathcal{I}_{2,\lambda} \cup \mathcal{I}_{3,\lambda}$ then once again the procedure terminates and $\mathcal{S}_\lambda$ contains these and only these queues which are empty at 1^-. Otherwise we proceed to the next step.

We now turn to the induction hypothesis.

kth step

Any queue $i \in \left\{\bigcup_{j<k} \mathcal{I}_{j,\lambda}\right\}^*$ receives amount of service in $(0,1)$ given by ,

$$C\phi_i + \frac{\phi_i}{\sum_{m \in \left\{\cup_{j<k} \mathcal{I}_{j,\lambda}\right\}^*} \phi_m} \sum_{l \in \cup_{j<k} \mathcal{I}_{j,\lambda}} a_l \tag{2.6}$$

and $L_\lambda > k$ since we are in the induction hypothesis step.

k+1 step

Up to this step any queue $i \in \{\cup_{j<k+1} \mathcal{I}_{j,\lambda}\}^*$ received the amount of service in $(0,1)$ given by,

$$C\phi_i + \frac{\phi_i}{\sum_{m \in \left\{\cup_{j<k} \mathcal{I}_{j,\lambda}\right\}^*} \phi_m} \sum_{l \in \cup_{j<k} \mathcal{I}_{j,\lambda}} a_l$$

by the induction hypothesis formula 2.6. What is more, a queue $i \in \{\cup_{j<k+1} \mathcal{I}_{j,\lambda}\}^*$ receives in $(0,1)$ service due to the queues in $\mathcal{I}_{k,\lambda}$ which equals to,

$$\frac{\phi_i}{\sum_{m \in \left\{\cup_{j<k+1} \mathcal{I}_{j,\lambda}\right\}^*} \phi_m} \left(\sum_{j \in \mathcal{I}_{k,\lambda}} a_j + \frac{\sum_{j \in \mathcal{I}_{k,\lambda}} \phi_j}{\sum_{m \in \left\{\cup_{j<k} \mathcal{I}_{j,\lambda}\right\}^*} \phi_m} \sum_{l \in \cup_{j<k} \mathcal{I}_{j,\lambda}} a_l \right) \geq 0 \ .$$

Hence, the total service a queue $i \in \left\{\bigcup_{j<k+1} \mathcal{I}_{j,\lambda}\right\}^*$ receives in $(0,1)$ is given by ,

$$C\phi_i + \frac{\phi_i}{\sum_{m \in \left\{\cup_{j<k+1} \mathcal{I}_{j,\lambda}\right\}^*} \phi_m} \sum_{l \in \cup_{j<k+1} \mathcal{I}_{j,\lambda}} a_l \ . \tag{2.7}$$

Thus, all queues in $\mathcal{I}_{k+1,\lambda}$ will be empty at 1^-. Once again the procedure terminates if either $\mathcal{M} \equiv \cup_{j\leq k+1}\mathcal{I}_{j,\lambda}$ (then obviously $\mathcal{S}^*_\lambda \equiv \emptyset$) or $\mathcal{I}_{k+1,\lambda} = \emptyset$ (then $L_\lambda = k$). In this case indeed, $\mathcal{S}_\lambda$ contains all empty queues at 1^- and only these. The induction procedure has now completed.

Formula 2.2 follows by noting that expression 2.7 applied for $k = L_\lambda$ is the service offered to any queue $i \in \mathcal{S}^*_\lambda$ during $(0, 1)$ if it is not trivially $\mathcal{S}_\lambda \equiv \mathcal{M}$.

The uniqueness of $\mathcal{S}_\lambda$ is deduced by the property of the queues in the latter set i. e. being empty at 1^- and the fact that a queue is either empty or backlogged. □

From the proof above one infers that any queue belonging to $\mathcal{S}^*_\lambda$ receives extra service from all queues in $\mathcal{S}_\lambda$ and only from these (except possibly from those queues in $\mathcal{I}_{L_\lambda,\lambda}$ for which the corresponding expression in the curly brackets in 2.1 holds as equality see also lemma 2.3 below). Hence, noting from 2.2 that $\phi_i\frac{\sum_{l\in\mathcal{S}_\lambda} a_l}{\sum_{m\in\mathcal{S}^*_\lambda}\phi_m}$ denotes the extra service a queue $i \in \mathcal{S}^*_\lambda$ receives from the queues in $\mathcal{S}_\lambda$ in $(0, 1)$ we have the following remark.

Remark 2.1 *For $\mathcal{S}_\lambda$ as in definition 2.1 and any $\mathcal{S} \in \mathcal{F}(\mathcal{M})$ with $\mathcal{S} \not\equiv \mathcal{S}_\lambda$,*

$$\frac{\sum_{l\in\mathcal{S}_\lambda} a_l}{\sum_{m\in\mathcal{S}^*_\lambda}\phi_m} \geq \frac{\sum_{l\in\mathcal{S}} a_l}{\sum_{m\in\mathcal{S}^*}\phi_m} \; . \tag{2.8}$$

Definition 2.2 *For any given λ we call the set $\mathcal{S}_\lambda$ the λ-**eventually stable** set and its members λ-**eventually stable** queues. Furthermore, a queue $i \in \mathcal{M}$ is said to be λ-**eventually unstable** if and only if it is a member of the set $\mathcal{U}_\lambda := \mathcal{S}^*_\lambda$ which we call the λ-**eventually unstable** set.*

We are now interested in finding a property which uniquely defines the queues in $\mathcal{S}_\lambda$ and more importantly, removes the recursive nature of the definition 2.1.

Definition 2.3 *Let us associate a real number b_i to every $i \in \mathcal{M}$. Define the partition $(\mathcal{B}, \mathcal{M}\backslash\mathcal{B})$ of $\mathcal{M}$ as follows: For any $i \in \mathcal{B}$,*

$$b_i + \frac{\phi_i}{\phi_i + \sum_{m\in\mathcal{B}^*}\phi_m}\sum_{l\in\mathcal{B}\backslash i} b_l \geq 0 \tag{2.9}$$

and for any $i \in \mathcal{B}^*$,

$$b_i + \frac{\phi_i}{\sum_{m\in\mathcal{B}^*}\phi_m}\sum_{l\in\mathcal{B}} b_l < 0 \; . \tag{2.10}$$

Remark 2.2 *Note that the following two way implication holds as long as $\mathcal{B} \not\equiv \mathcal{M}$,*

$$b_i + \frac{\phi_i}{\phi_i + \sum_{m\in\mathcal{B}^*}\phi_m}\sum_{l\in\mathcal{B}\backslash i} b_l \geq 0 \Longleftrightarrow b_i + \frac{\phi_i}{\sum_{m\in\mathcal{B}^*}\phi_m}\sum_{l\in\mathcal{B}} b_l \geq 0 \; .$$

Lemma 2.2 *Definition 2.3 applied for $b_i = a_i$ $(= C\phi_i - \lambda_{i,0})$ with $i \in \mathcal{M}$ and definition 2.1 are equivalent.*

<u>Proof :</u> We first show that the partition $(\mathcal{S}_\lambda, \mathcal{M}\backslash\mathcal{S}_\lambda)$ which satisfies definition 2.1 satisfies definition 2.3 as well. Next, we prove that the partition given by definition 2.3 is unique. The proof is done once we recall that by lemma 2.1 the partition $(\mathcal{S}_\lambda, \mathcal{M}\backslash\mathcal{S}_\lambda)$ is unique.

Let us now prove that the partition $(\mathcal{S}_\lambda, \mathcal{M}\backslash\mathcal{S}_\lambda)$ given by definition 2.1 satisfies definition 2.3. We pick an $i \in \mathcal{I}_{l,\lambda}$ and denote $\mathcal{M}_u = \{\mathcal{I}_{1,\lambda}, \mathcal{I}_{2,\lambda}, \dots, \mathcal{I}_{l-1,\lambda}\}$, $\mathcal{M}_d = \{\mathcal{I}_{l+1,\lambda}, \mathcal{I}_{l+2,\lambda}, \dots, \mathcal{I}_{L_\lambda,\lambda}\}$. Now define,

$$A_u := \sum_{j\in\mathcal{M}_u} a_j \qquad A_d := \sum_{j\in\mathcal{M}_d} a_j \qquad A_s := \sum_{j\in\mathcal{I}_{l,\lambda}} a_j \qquad A_q := \sum_{j\in\mathcal{I}_{l,\lambda}\backslash i} a_j$$

$$\Phi_u := \sum_{j\in\mathcal{M}_u} \phi_j \quad \Phi_d := \sum_{j\in\mathcal{M}_d} \phi_j \quad \Phi_s := \sum_{j\in\mathcal{I}_{l,\lambda}} \phi_j \quad \Phi_q := \sum_{j\in\mathcal{I}_{l,\lambda}\backslash i} \phi_j \quad \Phi_r := \sum_{j\in\mathcal{M}\backslash\mathcal{S}_\lambda} \phi_j$$

and let,

$$c_0 := \frac{\Phi_d + \Phi_q}{\Phi_d + \Phi_r + \Phi_s} A_u \ .$$

Invoking definition 2.1 we have for any $i \in \mathcal{I}_{l,\lambda}$,

$$a_i + \frac{\phi_i}{\Phi_d + \Phi_r + \Phi_s} A_u \geq 0 \tag{2.11}$$

and

$$A_d + A_s + \frac{\Phi_d + \Phi_s}{\Phi_d + \Phi_r + \Phi_s} A_u \geq 0 \ .$$

Write,

$$0 \leq A_d + A_s + \frac{\Phi_d + \Phi_s}{\Phi_d + \Phi_r + \Phi_s} A_u \ = A_d + A_q + c_0 + a_i + \frac{\phi_i}{\Phi_d + \Phi_r + \Phi_s} A_u$$

which implies that,

$$0 \leq \frac{\phi_i}{\phi_i + \Phi_r} \left(A_d + A_q + c_0\right) + a_i + \frac{\phi_i}{\Phi_d + \Phi_r + \Phi_s} A_u \ . \tag{2.12}$$

Simple algebra shows that,

$$\frac{A_u}{\Phi_d + \Phi_r + \Phi_s} + \frac{c_0 + A_d + A_q}{\phi_i + \Phi_r} = \frac{A_u + A_d + A_q}{\phi_i + \Phi_r} \ . \tag{2.13}$$

Hence, combining 2.12 and 2.13 we have,

$$0 \leq a_i + \frac{A_u + A_d + A_q}{\phi_i + \Phi_r} \phi_i \ .$$

This proves that any $i \in \mathcal{S}_\lambda$ satisfies 2.9. It remains to note that for any $i \notin \mathcal{S}_\lambda$, 2.10 must be true due to formula 2.2.

Let us now show that the partition given by definition 2.3 is unique. Let $(\mathcal{A}, \mathcal{M}\backslash\mathcal{A})$ and $(\mathcal{B}, \mathcal{M}\backslash\mathcal{B})$ be two arbitrary partitions that satisfy definition 2.3. We will show by contradiction that $\mathcal{B}\backslash\mathcal{A} \equiv \emptyset$ which implies that $\mathcal{A}\backslash\mathcal{B} \equiv \emptyset$ (due to the arbitrary choice of $\mathcal{A}$ and $\mathcal{B}$) and hence $\mathcal{A} \equiv \mathcal{B}$. To this end, suppose that $\mathcal{B}\backslash\mathcal{A} \not\equiv \emptyset$ and define the following quantities:

$$A_u := \sum_{j\in\mathcal{A}\cap\mathcal{B}} a_j \qquad A_d := \sum_{\substack{j\in\mathcal{B}\backslash\mathcal{A} \\ j\neq i}} a_j \qquad A_r := \sum_{j\in\mathcal{M}\backslash(\mathcal{A}\cup\mathcal{B})} a_j$$

$$A_s := \sum_{j\in\mathcal{A}\backslash\mathcal{B}} a_j \qquad A_D := \sum_{j\in\mathcal{B}\backslash\mathcal{A}} a_j$$

and

$$\Phi_u := \sum_{j\in\mathcal{A}\cap\mathcal{B}} \phi_j \qquad \Phi_d := \sum_{\substack{j\in\mathcal{B}\setminus\mathcal{A}\\ j\neq i}} \phi_j \qquad \Phi_r := \sum_{j\in\mathcal{M}\setminus(\mathcal{A}\cup\mathcal{B})} \phi_j$$

$$\Phi_s := \sum_{j\in\mathcal{A}\setminus\mathcal{B}} \phi_j \qquad \Phi_D := \sum_{j\in\mathcal{B}\setminus\mathcal{A}} \phi_j \ .$$

According to definition 2.3 $\forall i \notin \mathcal{A}$ we have,

$$a_i + \frac{\phi_i}{\Phi_D+\Phi_r}\left(A_u + A_s\right) < 0$$

which can be written as,

$$a_i + \frac{\phi_i}{\Phi_D+\Phi_r}\left(A_u - \frac{\Phi_s}{\Phi_D+\Phi_s+\Phi_r}A_u + \frac{\Phi_s}{\Phi_D+\Phi_s+\Phi_r}A_u + A_s\right) < 0 \ . \tag{2.14}$$

Moreover, for every $j \in \mathcal{A}$ (again due to the definition 2.3 and remark 2.2) we have,

$$a_j + \frac{\phi_j}{\Phi_D+\Phi_r}\left(A_u + A_s\right) \geq 0 \ .$$

Summing now for all $j \in \mathcal{A}\setminus\mathcal{B}$,

$$A_s + \frac{\Phi_s}{\Phi_D+\Phi_r}\left(A_u + A_s\right) \geq 0 \Longleftrightarrow A_s + \frac{\Phi_s}{\Phi_D+\Phi_s+\Phi_r}A_u \geq 0 \ .$$

Hence, 2.14 for any $i \notin \mathcal{A}$ and by implication for any $i \in \mathcal{B}\setminus\mathcal{A}$ yields,

$$0 > a_i + \frac{\phi_i}{\Phi_D+\Phi_r}\left(\frac{\Phi_D+\Phi_r}{\Phi_D+\Phi_s+\Phi_r}A_u\right) = a_i + \frac{\phi_i}{\Phi_D+\Phi_s+\Phi_r}A_u \ . \tag{2.15}$$

Then summing the last relation for all elements of $\mathcal{B}\setminus\mathcal{A}$ but the $\{a_i\}$ we obtain,

$$0 > A_d + \frac{\Phi_d}{\Phi_D+\Phi_s+\Phi_r}A_u \ . \tag{2.16}$$

Denoting $\mathrm{c}_1 := \frac{\Phi_d}{\Phi_D+\Phi_r+\Phi_s}A_u$, simple algebra yields,

$$\frac{A_d + A_u}{\phi_i+\Phi_s+\Phi_r} = \frac{A_u}{\phi_i+\Phi_d+\Phi_s+\Phi_r} + \frac{A_d + c_1}{\phi_i+\Phi_s+\Phi_r}$$

which observing that by definition $\Phi_D = \Phi_d + \phi_i$ and in conjunction with 2.15 and 2.16 gives,

$$a_i + \phi_i\frac{A_d + A_u}{\phi_i+\Phi_s+\Phi_r} = a_i + \frac{\phi_i}{\Phi_D+\Phi_s+\Phi_r}A_u + \phi_i\frac{A_d + c_1}{\phi_i+\Phi_s+\Phi_r} < 0 \ .$$

This contradicts the definition of $(\mathcal{B}, \mathcal{M}\setminus\mathcal{B})$ according to which for any $i \in \mathcal{B}$,

$$a_i + \phi_i\frac{A_d + A_u}{\phi_i+\Phi_s+\Phi_r} \geq 0 \ .$$

□

Remark 2.3 *From remark 2.2 formula 2.9 and the last lemma trivially follows that* $\sum_{i\in\mathcal{S}_\lambda} a_i \geq 0$.

The next lemma says that the extra service offered to a queue $i \in \mathcal{S}^*_\lambda$ remains the same if we remove from $\mathcal{S}_\lambda$ any or all those queues for which 2.9 (applied for $b_i = a_i$ for all i) holds as equality. This complements remark 2.1.

Lemma 2.3 *Let us associate a real number* b_i *to every* $i \in \mathcal{M}$ *and assume that a partition* $(\mathcal{B}, \mathcal{M}\backslash\mathcal{B})$ *of* $\mathcal{M}$ *with* $\mathcal{B} \not\equiv \mathcal{M}$*, satisfies definition 2.3. Denote,*

$$\mathcal{G} := \left\{ i \in \mathcal{B} : b_i + \frac{\phi_i}{\sum_{j\in\mathcal{B}^*}\phi_j} \sum_{l\in\mathcal{B}} b_l = 0 \right\} .$$

Then for any $\mathcal{G}_1 \subseteq \mathcal{G}$,

$$\frac{\sum_{l\in\mathcal{B}} b_l}{\sum_{j\in\mathcal{B}^*}\phi_j} = \frac{\sum_{l\in\mathcal{B}\backslash\mathcal{G}_1} b_l}{\sum_{j\in(\mathcal{B}\backslash\mathcal{G}_1)^*}\phi_j} .$$

__Proof :__ Firstly note that,

$$\frac{\sum_{i\in\mathcal{G}_1}\phi_i}{\sum_{j\in(\mathcal{B}\backslash\mathcal{G}_1)^*}\phi_j} \sum_{m\in\mathcal{B}\backslash\mathcal{G}_1} b_m + \sum_{i\in\mathcal{G}_1} b_i = 0 \Leftrightarrow \frac{\sum_{i\in\mathcal{G}_1}\phi_i}{\sum_{j\in\mathcal{B}^*}\phi_j} \sum_{m\in\mathcal{B}} b_m + \sum_{i\in\mathcal{G}_1} b_i = 0 . \qquad (2.17)$$

Hence,

$$\begin{aligned}
\frac{\sum_{l\in\mathcal{B}} b_l}{\sum_{j\in\mathcal{B}^*}\phi_j} &= \frac{1}{\sum_{j\in\mathcal{B}^*}\phi_j}\left(\sum_{l\in\mathcal{B}\backslash\mathcal{G}_1} b_l + \sum_{l\in\mathcal{G}_1} b_l\right) \\
&= \frac{1}{\sum_{j\in\mathcal{B}^*}\phi_j}\left(\frac{\sum_{i\in\mathcal{B}^*}\phi_i}{\sum_{m\in(\mathcal{B}\backslash\mathcal{G}_1)^*}\phi_m}\sum_{l\in\mathcal{B}\backslash\mathcal{G}_1} b_l + \frac{\sum_{i\in\mathcal{G}_1}\phi_i}{\sum_{m\in(\mathcal{B}\backslash\mathcal{G}_1)^*}\phi_m}\sum_{l\in\mathcal{B}\backslash\mathcal{G}_1} b_l + \sum_{l\in\mathcal{G}_1} b_l\right) \\
&= \frac{\sum_{l\in\mathcal{B}\backslash\mathcal{G}_1} b_l}{\sum_{j\in(\mathcal{B}\backslash\mathcal{G}_1)^*}\phi_j}
\end{aligned}$$

where for the last equality we used 2.17. □

Let us now assume that there exist certain queues in the system that (e.g. due to large backlogged workload) exhibit 'greedy' behaviour in the sense that they always use the service offered to them. Then, all the results introduced up the current point can be readily extended to this case. However, we will only present the definition of the corresponding eventual-stable notion and the equivalent result to lemma 2.2.

Throughout this paper $\mathcal{F}$ will denote the set of the 'greedy' queues.

Definition 2.4 *For any set* $\mathcal{F} \subseteq \mathcal{M}$ *and any given* λ *(with* $a_i := C\phi_i - \lambda_{i,0}$ *for all* $i \in \mathcal{M}$*), we define* $\mathcal{I}_{k,\mathcal{F},\lambda}$ *as,*

$$\mathcal{I}_{1,\mathcal{F},\lambda} := \{ i \in \mathcal{F}^* \ : \ a_i \geq 0 \}$$

$$\mathcal{I}_{k,\mathcal{F},\lambda} := \{ i \in (\cup_{j<k} \mathcal{I}_{j,\mathcal{F},\lambda} \cup \mathcal{F})^* \ : \ a_i + \frac{\phi_i}{\sum\limits_{m\in\{\cup_{j<k}\mathcal{I}_{j,\mathcal{F},\lambda}\}^*} \phi_m} \sum_{l\in\cup_{j<k}\mathcal{I}_{j,\mathcal{F},\lambda}} a_l \geq 0 \}$$

$$L_{\mathcal{F},\lambda} := \min\{ j \in \{0,1,\ldots,|\mathcal{M}\backslash\mathcal{F}|\} \ : \ \mathcal{I}_{j+1,\mathcal{F},\lambda} \equiv \emptyset \}$$

and

$$\mathcal{S}_{\mathcal{F},\lambda} := \bigcup_{j=1}^{L_{\mathcal{F},\lambda}} \mathcal{I}_{j,\mathcal{F},\lambda} \,. \tag{2.18}$$

*We call the set $\mathcal{S}_{\mathcal{F},\lambda}$ the $(\mathcal{F},\lambda)$-***eventually stable** *set and its members $(\mathcal{F},\lambda)$-***eventually stable** *queues. Moreover, a queue $i \in \mathcal{F}^*$ is said to be $(\mathcal{F},\lambda)$-***eventually unstable** *if and only if it belongs to the set $\mathcal{U}_{\mathcal{F},\lambda} := \mathcal{M}\backslash(\mathcal{S}_{\mathcal{F},\lambda} \cup \mathcal{F})$ which we call the $(\mathcal{F},\lambda)$-***eventually unstable** *set.*

It is not difficult to see that $\mathcal{S}_\lambda \equiv \mathcal{S}_{\emptyset,\lambda}$ and $\mathcal{U}_{\emptyset,\lambda} \equiv \mathcal{U}_\lambda$. Moreover, following arguments similar to those in lemma 2.1 one can show that $(\mathcal{S}_{\mathcal{F},\lambda},\mathcal{U}_{\mathcal{F},\lambda})$ is a unique partition of $\mathcal{M}\backslash\mathcal{F}$.

Lemma 2.2 can now be extended in a natural way to obtain the next result.

Lemma 2.4 *For any given λ and $\mathcal{F} \subseteq \mathcal{M}$ a queue $i \in \mathcal{S}_{\mathcal{F},\lambda}$ if and only if it satisfies,*

$$a_i + \frac{\phi_i}{\phi_i + \sum_{m \in \mathcal{S}^*_{\mathcal{F},\lambda}} \phi_m} \sum_{l \in \mathcal{S}_{\mathcal{F},\lambda} \backslash i} a_l \geq 0 \tag{2.19}$$

otherwise $i \in \mathcal{U}_{\mathcal{F},\lambda}$.

Remark 2.4 *For any given λ a queue which is λ-eventually unstable in effect behaves like the members of $\mathcal{F}$. Thus, if $\mathcal{F} \subseteq \mathcal{U}_\lambda$ we have $\mathcal{S}_{\mathcal{F},\lambda} \equiv \mathcal{S}_\lambda$ while in general $\mathcal{S}_{\mathcal{F},\lambda} \subseteq \mathcal{S}_\lambda$.*

In the following we will use the results developed above to obtain the asymptotics of the tail of the buffer occupancy.

3 Buffer occupancy asymptotics for M-queue GPS systems

In this section we develop the main results on the tails of the buffer occupancy distributions. Throughout we assume the buffers are infinite and we use the probability of exceeding the level NB_i for bufffer $i; i = 1, 2, \ldots, M$, as the surrogate for the overflow probability when the buffer has size NB_i.

Based on the insight we gained from the last section on the dynamics of the GPS service scheme in this section we obtain an event that always takes place if $X_{M,0} > NB_M$ and exploits in large extend the dynamics of the GPS scheduler.

Lemma 3.1 *If $X^N_{M,0} > NB_M$ there exist moments of time $T_i \geq T_M \geq 1$ integers, for $i = 1, 2, ..., M-1$ and a set $\mathcal{S}_M \subseteq M^*$ such that the following event holds true:*

$$\begin{aligned} E^N_{\mathbf{T}}(\mathcal{S}_M) := \Big\{ & \Lambda^N_M(-T_M,0) > CNT_M\phi_M + NB_M + \phi_M\gamma^N\left(\mathcal{S}_M, \Lambda^N_\cdot(-T_\cdot,0)\right) \\ & \bigcap_{i \in \mathcal{S}_M} CN(T_i - T_M)\phi_i \leq \Lambda^N_i(-T_i,0) \leq CNT_i\phi_i + \phi_i\gamma^N\left(\mathcal{S}_M, \Lambda^N_\cdot(-T_\cdot,0)\right) \\ & \bigcap_{j \in M^* \backslash \mathcal{S}_M} \Lambda^N_j(-T_j,0) > CNT_j\phi_j + \phi_j\gamma^N\left(\mathcal{S}_M, \Lambda^N_\cdot(-T_\cdot,0)\right) \Big\} \end{aligned} \tag{3.1}$$

Proof : Let us assume that $-T_M$ is the most recent moment before 0 where the buffer M was empty. Furthermore, assume that $-T_i$, $i = 1, 2, \dots M-1$ is the last moment before $-T_M + 1$ where queue i was empty. The existence of these moments can be justified via Loynes's theorem. Let us give the following definition.

Definition 3.1 *Let us define the set* $\mathcal{S}_M := \cup_{j=1}^{K_M} \mathcal{I}_{j,M}$ *whith,*

$$\mathcal{I}_{1,M} := \left\{ i \in M^* : \Lambda_i^N \left(-T_i, 0\right) \leq CNT_i \phi_i \right\}$$

$$\mathcal{I}_{k,M} := \Big\{ i \in \left(\cup_{j \leq k-1} \mathcal{I}_{j,M} \cup M\right)^* : CN \left(T_i - T_M\right) \phi_i \leq \Lambda_i^N \left(-T_i, 0\right)$$

$$\leq CNT_i \phi_i + \frac{\phi_i}{\sum_{m \in \left(\cup_{j \leq k-1} \mathcal{I}_{j,M}\right)^*} \phi_m} \sum_{l \in \cup_{j \leq k-1} \mathcal{I}_{j,M}} \left(CNT_l \phi_l - \Lambda_l^N \left(-T_l, 0\right) \right) \Big\} \tag{3.2}$$

$$K_M := \min \left\{ i = 1, 2, ..., M-1 : \mathcal{I}_{i+1,M} = \emptyset \right\}$$

and $\forall i \in \mathcal{S}_M^*$

$$\Lambda_i^N \left(-T_i, 0\right) > CNT_i \phi_i + \frac{\phi_i}{\sum_{m \in \mathcal{S}_M^*} \phi_m} \sum_{l \in \mathcal{S}_M} \left(CNT_l \phi_l - \Lambda_l^N \left(-T_l, 0\right) \right)$$

Then $\mathcal{S}_M$ *is said to be the* M-**Virtually stable set**. *All* $i \in \mathcal{S}_M$, *are said to be* M-**Virtually stable queues**. *Moreover, any* $i \in \mathcal{I}_{k,M}$ *will be called a* k^{th} **order** $M-$**virtually stable queue** *and the set* $\mathcal{I}_{k,M}$ *is the* k^{th} **order** $M-$**virtually stable set**. *We will refer to the number* K_M *as the* **maximum** M-**Virtual stability order**.

The set $\mathcal{S}_M^*$ *is called* M-**Virtually unstable set** *and all its members,* M-**Virtually unstable queues**.

From the definition of $-T_i$, $i = 1, 2, \dots M$ it is not difficult to see that the service left unused by a queue $i \in M^*$ in the interval $(-T_M, 0)$ is at least $CNT_M \phi_i - X_{i,-T_M}^N - \Lambda_i^N(-T_M, 0) + X_{i,0}^N$. Estimating,

$$\begin{aligned} 0 \leq X_{i,-T_M}^N &\leq \Lambda_i^N(-T_i, -T_M) - CN(T_i - T_M)\phi_i \Rightarrow \\ CN(T_i - T_M)\phi_i &\leq \Lambda_i^N(-T_i, -T_M) \leq \Lambda_i^N(-T_i, 0) \end{aligned} \tag{3.3}$$

and trivially $X_{i,0}^N \geq 0$. Thus, we infer that the service left unused in $(-T_M, 0)$ by an $i \in M^*$ is at least $CNT_i \phi_i - \Lambda_i^N(-T_i, 0)$ (at this stage this quantity may be negative but we have not exploited the GPS scheme yet) with the proper lower bound given in 3.3. Now applying lemma 2.1 for $a_i = CNT_i \phi_i - \Lambda_i^N(-T_i, 0)$ one can see that all queues in $\mathcal{S}_M$ leave some unused service in $(-T_M, 0)$ or in the worst case (may occur only if the upper bound in 3.2 equals the first term in the minimum) they use all the service offered to them in $(-T_M, 0)$ and have at time 0 no backlogged workload.

Recalling that $S_M^N(-T_M, 0)$ stands for the service available to queue M during $(-T_M, 0)$ we have,

$$X_{M,0}^N = \Lambda_M^N(-T_M, 0) - S_M^N(-T_M, 0)$$

Observe that by the definition of the GPS,

$$\begin{aligned} S^N_M(-T_M,0) &\geq CNT_M\phi_M + \sum_{i\in M^*}\left(CNT_M\phi_i - X^N_{i,-T_M} - \Lambda^N_i(-T_M,0) + X^N_{i,0}\right) \\ &\geq CNT_M\phi_M + \sum_{i\in M^*}\left(CNT_i\phi_i - \Lambda^N_i(-T_i,0)\right) \qquad (3.4) \end{aligned}$$

For any $\mathcal{S} \in \mathcal{F}(M^*)$ we write,

$$\begin{aligned} \sum_{i\in M^*}(CNT_i\phi_i - \Lambda^N_i(-T_i,0)) &= \frac{\phi_M}{\sum_{m\in\mathcal{S}^*}\phi_m}\sum_{i\in\mathcal{S}}(CNT_i\phi_i - \Lambda^N_i(-T_i,0)) \\ &\quad + \sum_{i\in\mathcal{S}^*}(CNT_i\phi_i - \Lambda^N_i(-T_i,0)) + \frac{\sum_{j\in\mathcal{S}^*\backslash M}\phi_j}{\sum_{m\in\mathcal{S}^*}\phi_m}\sum_{i\in\mathcal{S}}(CNT_i\phi_i - \Lambda^N_i(-T_i,0)) \\ &\leq \frac{\phi_M}{\sum_{m\in\mathcal{S}^*_M}\phi_m}\sum_{i\in\mathcal{S}_M}(CNT_i\phi_i - \Lambda^N_i(-T_i,0)) \qquad (3.5) \end{aligned}$$

where in the last inequality we invoked remark 2.1 and lemmas 2.1 and 2.2. Which suggests that we obtain a tighter bound if we assume that the M^{th} queue receives assistant service from only the queues in $\mathcal{S}_M$. Thus, we obtain,

$$NB_M \leq X^N_{M,0} \leq \Lambda^N_M(-T_M,0) - CNT_M\phi_M - \frac{\phi_M}{\sum_{j\in\mathcal{S}^*_M}\phi_j}\sum_{i\in\mathcal{S}_M}(CNT_i\phi_i - \Lambda^N_i(-T_i,0)) .$$

Invoking lemma 2.2 in conjunction with the definition 3.1 and equation 3.3 we obtain the announced result. □

We can now trivially deduce that

$$\begin{aligned} P\left[X^N_{M,0} > NB_M\right] &\leq \sum_{T_M=1}^{\infty}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_1=T_M}^{\infty} P\left[\cup_{\mathcal{S}\in\mathcal{F}(M^*)}E^N_{\mathbf{T}}(\mathcal{S})\right] \\ &= \sum_{T_M=1}^{\infty}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_1=T_M}^{\infty}\sum_{\mathcal{S}\in\mathcal{F}(M^*)} P\left[E^N_{\mathbf{T}}(\mathcal{S})\right] \qquad (3.6) \end{aligned}$$

recall that the union and by implication the summation above is extended over all possible subsets $\mathcal{S} \subseteq M^*$. What is more observe that the set $\mathcal{S}$ acquires the meaning of the M-eventually stable set once it appears in the arguments of event $E(\cdot)$ hence in the sequel if this is the case we will not carry the subscript M. Finally, note that the equation in the last expression is due to the fact that $E^N_{\mathbf{T}}(\mathcal{S})$ are disjoint events with respect to $\mathcal{S}$ (cf. uniqueness of the partition $(\mathcal{S}, \mathcal{M}\backslash\mathcal{S})$ lemma 2.1).

Thus, defining

$$E^N_{\mathbf{T}} := \bigcup_{\mathcal{S}\in\mathcal{F}(M^*)} E^N_{\mathbf{T}}(\mathcal{S})$$

and recalling the independence of the arrival processes $\{\lambda_{i,t}\}^{\infty}_{t=0}$, for all i, we have:

$$\begin{aligned} P[E^N_{\mathbf{T}}] &= \sum_{\mathcal{S}\in\mathcal{F}(M^*)} P[E^N_{\mathbf{T}}(\mathcal{S})] \\ &= \sum_{\mathcal{S}\in\mathcal{F}(M^*)} N^{|\mathcal{S}|}\int_{\mathcal{D}_{\mathbf{T},\mathcal{S}}}\prod_{i\in\mathcal{S}}P^{\Lambda^N_i(-T_i,0)}(Nu_i)\prod_{i\in\mathcal{S}^*\backslash M}P[\Lambda^N_i(-T_i,0) > N(CT_i\phi_i + \phi_i\gamma(\mathcal{S},u.))] \\ &\qquad \cdot\ P[\Lambda^N_M(-T_M,0) > N(CT_M\phi_M + B_M + \phi_M\gamma(\mathcal{S},u.))]\,d\,\mathbf{u}_{\mathcal{S}} \qquad (3.7) \end{aligned}$$

where the region $\mathcal{D}_{\mathbf{T},\mathcal{S}}$ is defined as

$$\begin{aligned} \mathcal{D}_{\mathbf{T},\mathcal{S}} := \{\mathbf{u}_{\mathcal{S}} \in \Re^{|\mathcal{S}|} \ : \ C\left(T_i - T_M\right)\phi_i \quad &\leq \quad u_i \leq \min\left\{CT_i\phi_i + \phi_i\gamma\left(\mathcal{S}, u.\right), K_iT_i\right\} \quad \forall i \in \mathcal{S} \\ CT_j\phi_j + B_M\mathbf{1}_{\{j=M\}} + \phi_j\gamma\left(\mathcal{S}, u.\right) \quad &\leq \quad K_jT_j \ \ \forall j \in \mathcal{S}^* \ \} \ . \end{aligned} \qquad (3.8)$$

From the definition of $\mathbf{u}_{\mathcal{S}}$ and for sake of brevity we will drop the subscript $\mathcal{S}$ from $\mathbf{u}_{\mathcal{S}}$ whenever it is clear that a vector $\mathbf{u} \in \mathcal{D}_{\mathbf{T},\mathcal{S}}$. For the same reason we will write $\mathcal{D}_{\mathcal{S}}$ instead of $\mathcal{D}_{\mathbf{T},\mathcal{S}}$ if there will be no ambiguity about the time vector $\mathbf{T}$ which the latter region is associated to.

Invoking the results of lemma 1.1 we obtain,

$$P[E_{\mathbf{T}}^N] = \sum_{\mathcal{S} \in \mathcal{F}(M^*)} \int_{\mathcal{D}_{\mathcal{S}}} D_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}, N\right) \ e^{-N \ J_{\mathbf{T},\mathcal{S}}(\mathbf{u})} \ d\,\mathbf{u} \left(1 + O\left(\frac{1}{N}\right)\right) \qquad (3.9)$$

where we will refer to $J_{\mathbf{T},\mathcal{S}}(\mathbf{u})$ as the rate function of the event $E_{\mathbf{T}}^N\left(\mathcal{S}\right)$ given by

$$\begin{aligned} &J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}\right) \\ &= \ \sum_{i \in \mathcal{S}} J_{i,T_i}\left(u_i\right) + \sum_{k \in \mathcal{S}^* \setminus M} R_{k,T_k}\left(CT_k\phi_k + \phi_k\gamma\left(\mathcal{S}, u.\right)\right) + R_{M,T_M}\left(CT_M\phi_M + B_M + \phi_M\gamma\left(\mathcal{S}, u.\right)\right) \end{aligned}$$

where $R_{k,T_k}(x) = J_{k,T_k}(x)$ if $x > n_k\Lambda_k$ and zero otherwise (cf. remark 1.4). However, by the assumption that for all $i = 1, 2, ..., M$, $n_i\Lambda_i < C\phi_i$ the latter rate function reads,

$$\begin{aligned} &J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}\right) \\ &:= \ \sum_{i \in \mathcal{S}} J_{i,T_i}\left(u_i\right) + \sum_{k \in \mathcal{S}^* \setminus M} J_{k,T_k}\left(CT_k\phi_k + \phi_k\gamma\left(\mathcal{S}, u.\right)\right) + J_{M,T_M}\left(CT_M\phi_M + B_M + \phi_M\gamma\left(\mathcal{S}, u.\right)\right) \end{aligned} \qquad (3.10)$$

and

$$D_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}, N\right) := \frac{N^{|\mathcal{S}| - M/2}}{\left(2\pi\right)^{M/2} \sqrt{\prod_{i \in \mathcal{S}} \sigma_{i,T_i}^2\left(u_i\right) \prod_{i \in \mathcal{S}^*} \sigma_{i,T_i}^2\left(\mathbf{u}\right)} \prod_{i \in \mathcal{S}^*} \tau_{i,T_i}\left(\mathbf{u}\right)} \qquad (3.11)$$

where we adopted the following notation (see lemma 1.1): For all $i \in \mathcal{S}^*$

$$\tau_{i,T_i}(\mathbf{u}) = \tau_{i,T_i}(CT_i\phi_i + B_M\mathbf{1}_{\{i=M\}} + \phi_i\gamma(\mathcal{S}, u.))$$

and similarly

$$\sigma_{i,T_i}^2(\mathbf{u}) = \sigma_i^2(CT_i\phi_i + B_M\mathbf{1}_{\{i=M\}} + \phi_i\gamma(\mathcal{S}, u.)) \ .$$

This notation will be used throughout if there will be no ambiguity about the arguments of τ and σ^2 which nevertheless are the same as the arguments in the corresponding rate function.

For once more we should point out that the dimension of $\mathbf{u}$ appearing in $J_{\mathbf{T},\mathcal{S}}(\mathbf{u})$ is $|\mathcal{S}|$ defined by the subscript $\mathcal{S}$ of $J_{\mathbf{T},\mathcal{S}}\left(\cdot\right)$ while recall that the numbering of $\mathbf{u}$'s coordinates will be driven by the members of $\mathcal{S}$ as stated earlier.

Now we turn to the evaluation of the multiple integral appear in equation 3.9. This is known in the literature as a Laplace type multiple integral.

The following theorem provides us with all necessary results to estimate the latter integral and can be found in [41] and [6].

Theorem 3.1 *Consider the following integral*

$$I(N) = \int_{\mathcal{D}} g(\mathbf{x})\, e^{\{-N\, f(\mathbf{x})\}} d\mathbf{x} \qquad \mathbf{x} = (x_1, x_2, ..., x_n) \tag{3.12}$$

where n is any positive integer and $\mathcal{D}$ is a possibly unbounded domain in $\Re^n$. Assume that f and g are infinitely differentiable throughout the closure of the integration domain $\overline{\mathcal{D}}$. And let us denote Γ the boundary of $\mathcal{D}$. Then, we have the following results:

<u>**Case 1**</u> *Let $\nabla f(\mathbf{x}) \neq 0$ in $\overline{\mathcal{D}}$ so that the minimum of f is achieved on a $\mathbf{x}_0 \in \Gamma$ and we further suppose that it is unique.Then, as $N \to \infty$*

$$I(N) = \frac{g(\mathbf{x}_0)}{2\pi\sqrt{|\mathbf{J}|}} e^{\{-N\, f(\mathbf{x}_0)\}} \left(\frac{2\pi}{N}\right)^{(n+1)/2} \left(1 + O\left(\frac{1}{N}\right)\right) . \tag{3.13}$$

Where $\mathbf{J} = \sum_{p=1}^{n} \sum_{q=1}^{n} f_{x_p} f_{x_q}$ cof$[f_{x_p x_q} - K h_{x_p x_q}]$ for K a constant such that $\nabla f(\mathbf{x}_0) = K \nabla h(\mathbf{x}_0)$ and $h(\mathbf{x}) = 0$ represents the boundary Γ of the integration domain $\mathcal{D}$ in a neighborhood of $\mathbf{x}_0$. Moreover, $f_{x_\cdot}$ stands for the partial derivative of f wrt $x_\cdot$ while $f_{x_p x_q}$ represents the partial derivative of f wrt x_p and x_q. The symbol cof$[a_{i,j}]$ denotes the cofactor of the element $a_{i,j}$ in the matrix $(a_{i,j})$, i.e. , the determinant of the sub-matrix obtained after eliminating the i^{th} row and the j^{th} column from the original matrix $(a_{i,j})$ multiplied by $(-1)^{i+j}$.

<u>**Case 2**</u> *Suppose the minimum of f occurs at an interior point $\mathbf{x}_0$ of $\mathcal{D}$ i.e.$\nabla f(\mathbf{x}_0) = 0$ and it is unique. Then, as $N \to \infty$,*

$$I(N) = \frac{g(\mathbf{x}_0)}{\sqrt{det(\, f_{i,j}(\mathbf{x}_0))}} e^{\{-N\, f(\mathbf{x}_0)\}} \left(\frac{2\pi}{N}\right)^{n/2} \left(1 + O\left(\frac{1}{N}\right)\right) \tag{3.14}$$

where $det(\, f_{i,j}(\mathbf{x}_0)) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)\Big|_{\mathbf{x}=\mathbf{x}_0}$ is the determinant of the Hessian matrix of f which is positive due to the fact that the point $\mathbf{x}_0$ is a minimizing point for f.

<u>**Case 3**</u> *Suppose f attains its minimum at $\mathbf{x}_0$ on Γ with $\nabla\ f(\mathbf{x}_0) = 0$ and it is unique. Then, as $N \to \infty$,*

$$I(N) = \frac{g(\mathbf{x}_0)}{2\sqrt{det(\, f_{i,j}(\mathbf{x}_0))}} e^{\{-N\, f(\mathbf{x}_0)\}} \left(\frac{2\pi}{N}\right)^{n/2} \left(1 + O\left(\frac{1}{\sqrt{N}}\right)\right) . \tag{3.15}$$

The above theorem can be stated in a more general form where in Case 2 it suffices to assume that $g(\mathbf{x})$ is continuous and $f(\mathbf{x})$ has continuous second order partial derivatives in a neighborhood of the point $\mathbf{x}_0$.

Remark 3.1 *Theorem 3.1 states that a Laplace type integral over a domain $\mathcal{D}$ is asymptotically equivalent to the integral evaluated in an arbitrarily small neighborhood of a point $\mathbf{x}_0$ where the function in the exponent of the integrand attains its minimum in $\mathcal{D}$.*

Turning to our objective which is the evaluation of 3.9 and in conjunction with the last theorem it becomes apparent that 2^{M-1} optimizations are required. One for each $\mathcal{S} \in \mathcal{F}(M^*)$. However, the next two lemmas significantly simplify things suggesting that it suffices to carry out only one minimization which corresponds to the case where $\mathcal{S} = M^*$. The proofs of both lemmas stated below are deferred to the appendix.

Lemma 3.2 *Let $\mathcal{S} \in \mathcal{F}(M^*)$ and denote $\mathcal{S}_{\mathcal{Q}} := \mathcal{S} \cup \mathcal{Q}$. Then for any $\mathcal{Q} \subseteq \mathcal{S}^* \backslash M$ we have,*

$$\inf_{\mathbf{u} \in \mathcal{D}_{\mathcal{S}_{\mathcal{Q}}}} J_{\mathbf{T},\mathcal{S}_{\mathcal{Q}}}(\mathbf{u}) \leq \inf_{\mathbf{u} \in \mathcal{D}_{\mathcal{S}}} J_{\mathbf{T},\mathcal{S}}(\mathbf{u}) \tag{3.16}$$

and thus,

$$\inf_{\mathbf{u} \in \mathcal{D}_{M^*}} J_{\mathbf{T},M^*}(\mathbf{u}) = \inf_{\mathcal{S} \in \mathcal{F}(M^*)} \inf_{\mathbf{u} \in \mathcal{D}_{\mathcal{S}}} J_{\mathbf{T},\mathcal{S}}(\mathbf{u}) \tag{3.17}$$

Lemma 3.2 suggests that, invoking theorem 3.1 it is apparent that the exponent of $P[E_{\mathbf{T}}^N]$ equals to the exponent of $P[E_{\mathbf{T}}^N(M^*)]$ and hence only one minimization is required. This lemma resembles the cost reduction achieved by Massoulié in [28] theorem 2. However, in order to obtain the right coefficient of $P[E_{\mathbf{T}}^N]$ we should take into account all (if any) $P[E_{\mathbf{T}}^N(\mathcal{S})]$ with the same exponent as $P[E_{\mathbf{T}}^N(M^*)]$ since they contribute significantly. Lemma 3.3 below caters for this case. But before we proceed let us define for every $\mathcal{S} \subseteq \mathcal{M}$ the vector $\mathbf{u}_{\mathcal{S}}^*$ as the unique minimizer of $J_{\mathbf{T},\mathcal{S}}(\mathbf{u})$. Note that the uniqueness of $\mathbf{u}_{\mathcal{S}}^*$ is guaranteed by the fact that $J_{\mathbf{T},\mathcal{S}}(\mathbf{u})$ is a strictly convex function as a sum of strictly convex functions. Thus,

$$\mathbf{u}_{\mathcal{S}}^* := \left(u_{i_1,\mathcal{S}}^*, u_{i_2,\mathcal{S}}^*, ..., u_{i_{|\mathcal{S}|},\mathcal{S}}^*\right) := \arg \inf_{\mathbf{u} \in \mathcal{D}_{\mathcal{S}}} J_{\mathbf{T},\mathcal{S}}(\mathbf{u}) \tag{3.18}$$

which obviously depends on the time vector $\mathbf{T}$ but we drop this index for sake of brevity.

Lemma 3.3 *Denote $\mathcal{G} := \left\{i \in M^* : u_{i,M^*}^* = CT_i\phi_i + \phi_i\gamma\left(M^*, u_{\cdot,M^*}^*\right)\right\}$. Then for $\mathcal{S} \subseteq M^*$ the next two statements are equivalent:*

1) $J_{\mathbf{T},M^*}(\mathbf{u}_{M^*}^*) = J_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*)$

2) $\mathcal{S}^* \backslash M \subseteq \mathcal{G}$

Moreover, if any of the latter statements is true then

3) $u_{i,M^*}^* = u_{i,\mathcal{S}}^*$ for all $i \in \mathcal{S}$.

Hence, knowing $\mathbf{u}_{M^*}^*$ the coefficient of $P[E_{\mathbf{T}}^N]$ can be directly calculated without the need of carrying additional minimizations even if there are some $\mathcal{S} \subseteq M^*$ such that $J_{\mathbf{T},M^*}(\mathbf{u}_{M^*}^*) = J_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*)$. The next corollary provides us with the expression of $P[E_{\mathbf{T}}^N]$. We remind the reader that as it follows by the last proposition $\mathbf{u}_{\mathcal{S}}^*$ below is such that $u_{i,\mathcal{S}}^* = u_{i,M^*}^*$ for all $i \in \mathcal{S}$ and once more recall that $u_{\mathcal{S}}^* \in \Re^{|\mathcal{S}|}$.

Corollary 3.1 *For $\mathcal{G}$ as defined in lemma 3.3 and $\mathcal{S} \subseteq M^*$, we have as $N \to \infty$,*

$$P[E_{\mathbf{T}}^N] = e^{-N\, J_{\mathbf{T},M^*}(\mathbf{u}_{M^*}^*)} \sum_{\mathcal{S}^* \backslash M \subseteq \mathcal{G}} C_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*, N)\left(1 + O\left(\frac{1}{N}\right)\right)$$

where $C_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*, N)$ is the term $D_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*, N)$ multiplied by a proper constant depending on which case of theorem 3.1 applies. The term $C_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*, N)$ is either $O\left(N^{(|\mathcal{S}|-M)/2}\right)$ or $O\left(N^{(|\mathcal{S}|-M-1)/2}\right)$ and $\sum_{\mathcal{S}^* \backslash M \subseteq \mathcal{G}} C_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*, N)$ is either $O\left(N^{-1/2}\right)$ or $O\left(N^{-1}\right)$.

Proof The proof is trivially deduced by applying lemma 3.3 and lemma 3.2 to 3.9 after invoking theorem 3.1. We only note that $C_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*, N)$ are determined by $D_{\mathbf{T},\mathcal{S}}(\mathbf{u}_{\mathcal{S}}^*, N)$ with the proper amedments due to the evaluation of the Laplace's integral (cf. theorem 3.1) and thus it is either

$O\left(N^{(|\mathcal{S}|-M)/2}\right)$ if we are in case 2 or case 3 of theorem 3.1 or $O\left(N^{(|\mathcal{S}|-M-1)/2}\right)$ if case 1 applies. Now since $(M^*)^* \setminus M = \emptyset \subseteq \mathcal{G}$ we infer that $C_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}, N\right)$ is a member of $\sum_{\mathcal{S}^* \setminus M \subseteq \mathcal{G}} C_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}^*_{\mathcal{S}}, N\right)$ thus the latter sum is either $O\left(N^{-1/2}\right)$ or $O\left(N^{-1}\right)$. □

Thus, invoking relation 3.6 we obtain,

$$P\left[X^N_{M,0} > NB_M\right] \le \sum_{T_M=1}^{\infty} \sum_{T_{M-1}=T_M}^{\infty} \cdots \sum_{T_1=T_M}^{\infty} \left[\sum_{\mathcal{S}^* \setminus M \subseteq \mathcal{G}} C_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}^*_{\mathcal{S}}, N\right)\right] e^{-N\, J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right)} \left(1 + O\left(\frac{1}{N}\right)\right) \tag{3.19}$$

Note that the rate function of $E^N_{\mathbf{T}}$ takes the form,

$$J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right) = \sum_{i \in M^*} J_{i,T_i}\left(u^*_{i,M^*}\right) + J_{M,T_M}\left(\sum_{i=1}^{M-1}\left(CT_i\phi_i - u^*_{i,M^*}\right) + CT_M\phi_M + B_M\right). \tag{3.20}$$

The following lemma confines the region over which $\mathbf{u}^*_{M^*}$ should be searched. Apart from the technical implication (smaller run time in the search of the optimum of $J_{\mathbf{T},M^*}\left(\mathbf{u}_{M^*}\right)$ by a computer) it simplifies some proofs in the sequel and palys an important role in the pursuit of a lower bound for the tail of the queue length.

Proposition 3.1 *For all $i = 1, 2, ..., M-1$, $u^*_{i,M^*} > n_i\Lambda_i$.*

Proof Recall that the region over which,

$$J_{\mathbf{T},M^*}\left(\mathbf{u}\right) = \sum_{i=1}^{M-1} J_{i,T_i}\left(u_i\right) + J_{M,T_M}\left(CT_M\phi_M + B_M + \sum_{m=1}^{M-1}\left(CT_m\phi_m - u_m\right)\right)$$

is minimised with respect to $\mathbf{u}$ is,

$$\mathcal{D}_{M^*} := \left\{\mathbf{u} \in \Re^{M-1} \;:\; C\left(T_i - T_M\right)\phi_i \le u_i \le \min\left\{CT_i\phi_i + \phi_i\gamma\left(M^*, u.\right), K_iT_i\right\} \quad \forall i \in M^*\right\}. \tag{3.21}$$

It is easily seen that if $C\left(T_i - T_M\right)\phi_i \le n_i\Lambda_iT_i$ for all $i = 1, 2, ..., M-1$ then the vector $\mathbf{\Lambda_T} := \left(n_1\Lambda_1T_1, n_2\Lambda_2T_2, ..., n_M\Lambda_MT_M\right) \in \mathcal{D}_{M^*}$ (recall that $C\phi_i > n_i\Lambda_i$ for all i). Now note that for all $i \in M^*$,

$$\begin{aligned} \left.\frac{\partial J_{\mathbf{T},M^*}\left(\mathbf{u}\right)}{\partial u_i}\right|_{\mathbf{u}=\mathbf{\Lambda_T}} &= \left.\frac{\partial\left(J_{i,T_i}\left(u_i\right) + J_{M,T_M}\left(CT_M\phi_M + B_M + \sum_{m=1}^{M-1}\left(CT_m\phi_m - u_m\right)\right)\right)}{\partial u_i}\right|_{\mathbf{u}=\mathbf{\Lambda_T}} \\ &= -\tau_{M,T_M}\left(CT_M\phi_M + B_M + \sum_{m=1}^{M-1}\left(C\phi_m - n_m\Lambda_m\right)T_m\right) < 0 \end{aligned}$$

Hence, invoking the strict convexity of $J_{\mathbf{T},M^*}\left(\mathbf{u}\right)$ we deduce that $\mathbf{u}^*_{M^*} > \mathbf{\Lambda_T}$ in the sense that all coordinates of $\mathbf{u}^*_{M^*}$ must be strictly greater than the corresponding coordinates of $\mathbf{\Lambda_T}$.

If on the other hand there exists an $i = 1, 2, ..., M-1$ such that $C\left(T_i - T_M\right)\phi_i > n_i\Lambda_iT_i$ then the result follows trivially from the definition of $\mathcal{D}_{M^*}$. □

4 The tail of the M^{th} queue length in M-queue infinite-buffered GPS systems

In this subsection we introduce an upper bound for the probability the workload in the M^{th} queue to reach the level NB_M. In the next subsection we will discuss when this bound becomes asymptotically exact.

We will show that under very general assumptions on the rate functions of the arrival processes, the multiple sum in 3.19 converges providing us with the required upper bound.

Assumption 1 *For $n_i\Lambda_i < v_i \in \Re$, and for $i = 1, 2, ..., M$,* $\lim_{T\to\infty} \frac{J_{i,T}(Tv_i)}{\ln T} > 0$.

The following proposition describes the consequences of the assumption 1 on the rate function of the event $E^N_{\mathbf{T},M^*}$. Its proof can be found in the appendix.

Proposition 4.1 *Let $n_i\Lambda_i < v_i \in \Re$, $i = 1, 2, ..., M$ and $\mathbf{T} := (T_1, T_2, ..., T_M) \in \mathcal{N}^M$ then*

$$\forall i = 1, 2, ..., M \qquad \lim_{T_i\to\infty} \frac{J_{i,T_i}(T_i v_i)}{\ln T_i} > 0 \Longrightarrow \lim_{\sum_{i=1}^M T_i\to\infty} \frac{J_{\mathbf{T},M^*}(\mathbf{u}^*_{M^*})}{\ln\left(\sum_{i=1}^M T_i\right)} > 0$$

Assumption 2 *Let $\mathbf{T} \in \mathcal{N}^M$ then,*

$$\mathbf{T}^0 := \left(T^0_1, T^0_2, ..., T^0_M\right) := \inf\left\{J_{\mathbf{T},M^*}(\mathbf{u}^*_{M^*}) \text{ for } 1 \leq T_M \leq T_i \in \mathcal{N}\ \forall i \in M^*\right\} \tag{4.1}$$

is unique and finite.

Theorem 4.1 *If assumptions 1 and 2 hold then,*

$$P[X^N_{0,M} \geq NB_M] \leq e^{-N\, J_{\mathbf{T}^0,M^*}(\mathbf{u}^*_{M^*})} \sum_{\mathcal{S}^*\setminus M\subseteq\mathcal{G}} C_{\mathbf{T}^0,\mathcal{S}}(\mathbf{u}^*_{\mathcal{S}}, N)\left(1 + O\left(\frac{1}{N}\right)\right) \tag{4.2}$$

for $\mathcal{S} \subseteq M^*, \mathcal{G}$ as defined in lemma 3.3 and $C_{\mathbf{T}^0,\mathcal{S}}(\mathbf{u}^*_{\mathcal{S}}, N)$ as defined in corollary 3.1.

Proof From the assumptions of the theorem and proposition 4.1 we can choose $T > T^0_i\ \forall i \in \mathcal{M}$ and $a > 0$ such that whenever $\sum_{i=1}^M T_i > T$, $J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{\mathbf{T},M^*}\right) > a\ln\left(\sum_{i=1}^M T_i\right) > J_{\mathbf{T}^0,M^*}\left(\mathbf{u}^*_{\mathbf{T}^0,M^*}\right)$. We write,

$$\begin{aligned} P\left[X^N_{M,0} > NB_M\right] &\leq \sum_{T_M=1}^{\infty}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_1=T_M}^{\infty} P\left[E^N_{\mathbf{T}}\right] \\ &= \sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_1=T_M}^{\infty} P\left[E^N_{\mathbf{T}}\right] + \sum_{T_M=T}^{\infty}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_1=T_M}^{\infty} P\left[E^N_{\mathbf{T}}\right] \\ &= \sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{T-1}\cdots\sum_{T_1=T_M}^{T-1} P\left[E^N_{\mathbf{T}}\right] + \sum_{T_M=1}^{T-1}\sum P\left[E^N_{\mathbf{T}}\right] + \sum_{T_M=T}^{\infty}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_1=T_M}^{\infty} P\left[E^N_{\mathbf{T}}\right] \end{aligned} \tag{4.3}$$

where the second term above consists of $M-1$ terms of the type

$$\sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_{i+1}=T_M}^{\infty}\sum_{T_i=T}^{\infty}\sum_{T_{i-1}=T_M}^{T-1}\cdots\sum_{T_1=T_M}^{T-1}P\left[E_{\mathbf{T}}^{N}\right] \tag{4.4}$$

where in the i^{th} term ($1\leq i<M$) we have for $1\leq k\leq i-1$, $T_M\leq T_k\leq T-1$, for T_i, $T\leq T_i\leq\infty$ and for $i<j<M$, $T_M\leq T_j\leq\infty$.

Let us call 'Term i' $i=1,2,3$ each of the terms appear in expression 4.3 according to the order they appear from the left to the right. We will evaluate each one separately.

First of all notice that,

$$\frac{P\left[E_{\mathbf{T}}^{N}\right]}{P\left[E_{\mathbf{T}^0}^{N}\right]}=L\left(N\right)O\left(e^{-N}\right) \tag{4.5}$$

where $O\left(N^{-1/2}\right)\preceq L\left(N\right)\preceq O\left(N^{1/2}\right)$ which is due to the fact that for any $\mathbf{T}$, $\sum_{\mathcal{S}^*\setminus M\subseteq\mathcal{G}}C_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}_{\mathcal{S}}^*,N\right)$ is either $O\left(N^{-1/2}\right)$ or $O\left(N^{-1}\right)$ (cf. corollary 3.1)

Term 1

Since this multiple summation has a finite number of terms it is readily seen that,

$$\sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{T-1}\cdots\sum_{T_1=T_M}^{T-1}P\left[E_{\mathbf{T}}^{N}\right]=P\left[E_{\mathbf{T}^0}^{N}\right]\left(1+O\left(e^{-N}\right)L\left(N\right)\right) \tag{4.6}$$

Term 2

We will evaluate the typical say i^{th}, term given by 4.4. Then since $\sum_{k=1}^{M}T_k>T$ and for $N>\frac{M}{a}$ we can write,

$$\begin{aligned}
&\sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_{i+1}=T_M}^{\infty}\sum_{T_i=T}^{\infty}\sum_{T_{i-1}=T_M}^{T-1}\cdots\sum_{T_1=T_M}^{T-1}P\left[E_{\mathbf{T}}^{N}\right]\\
\leq\;& O\left(N^{-1/2}\right)\sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_{i+1}=T_M}^{\infty}\sum_{T_i=T}^{\infty}\sum_{T_{i-1}=T_M}^{T-1}\cdots\sum_{T_1=T_M}^{T-1}e^{-Na\ln\left(\sum_{i=1}^{M}T_i\right)}\\
=\;& O\left(N^{-1/2}\right)\sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_{i+1}=T_M}^{\infty}\sum_{T_i=T}^{\infty}\sum_{T_{i-1}=T_M}^{T-1}\cdots\sum_{T_1=T_M}^{T-1}\left(\sum_{i=1}^{M}T_i\right)^{-Na}\\
\leq\;& O\left(N^{-1/2}\right)T^{i-1}\sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_{i+1}=T_M}^{\infty}\sum_{T_i=T}^{\infty}\left(\sum_{k=i}^{M}T_k\right)^{-Na}\\
=\;& \frac{O\left(N^{-1/2}\right)}{Na-1}T^{i-1}\sum_{T_M=1}^{T-1}\sum_{T_{M-1}=T_M}^{\infty}\cdots\sum_{T_{i+1}=T_M}^{\infty}\left(T+\sum_{k=i+1}^{M}T_k\right)^{-Na+1}\\
=\;& \frac{O\left(N^{-1/2}\right)}{\left[Na-1\right]_{M-i}}T^{i-1}\sum_{T_M=1}^{T-1}\left(T+\left(M-i\right)T_M\right)^{-Na+M-i}=\frac{O\left(N^{-1/2}\right)}{\left[Na-1\right]_{M-i}}T^{-Na+M}
\end{aligned}$$

where $\left[X\right]_l:=X\left(X-1\right)\cdots\left(X-l+1\right)$.

Thus,

$$\sum_{T_M=1}^{T-1} \sum P\left[E_{\mathbf{T}}^N\right] \leq T^{-Na+M} \sum_{i=1}^{M-1} \frac{O\left(N^{-1/2}\right)}{[Na-1]_{M-i}} = T^{-Na} O\left(N^{-1/2}\right) \qquad (4.7)$$

Now applying 4.5 we have,

$$\sum_{T_M=1}^{T-1} \sum P\left[E_{\mathbf{T}}^N\right] \leq P\left[E_{\mathbf{T}^0}^N\right] O\left(N^{-1/2}\right) e^{-N\left(a \ln T - J_{\mathbf{T}^0,\mathbf{M}^*}\left(\mathbf{u}_{M^*}^*\right)\right)}$$

by the choice of T and a we deduce that $a \ln T - J_{\mathbf{T}^0,M^*}\left(\mathbf{u}_{M^*}^*\right) > 0$ and hence,

$$\sum_{T_M=1}^{T-1} \sum P\left[E_{\mathbf{T}}^N\right] = P\left[E_{\mathbf{T}^0}^N\right] O\left(N^{-1/2} e^{-N}\right) \qquad (4.8)$$

Term 3

For $N > \frac{M}{a}$,

$$\begin{aligned}
&\sum_{T_M=T}^{\infty} \sum_{T_{M-1}=T_M}^{\infty} \cdots \sum_{T_1=T_M}^{\infty} P\left[E_{\mathbf{T}}^N\right] \\
&\leq O\left(N^{-1/2}\right) \sum_{T_M=T}^{\infty} \sum_{T_{M-1}=T_M}^{\infty} \cdots \sum_{T_1=T_M}^{\infty} \left(\sum_{i=1}^{M} T_i\right)^{-Na} \\
&= \frac{O\left(N^{-1/2}\right)}{Na-1} \sum_{T_M=T}^{\infty} \sum_{T_{M-1}=T_M}^{\infty} \cdots \sum_{T_2=T_M}^{\infty} \left(T_M + \sum_{i=2}^{M} T_i\right)^{-Na+1} \\
&= \frac{O\left(N^{-1/2}\right)}{(Na-1)(Na-2)} \sum_{T_M=T}^{\infty} \sum_{T_{M-1}=T_M}^{\infty} \cdots \sum_{T_3=T_M}^{\infty} \left(2T_M + \sum_{i=3}^{M} T_i\right)^{-Na+2} \\
&= \frac{O\left(N^{-1/2}\right)}{(Na-1)(Na-2)\cdots(Na-M)} (MT)^{-Na+M} \qquad (4.9)
\end{aligned}$$

As in term 2 applying 4.5 to 4.9 we obtain,

$$\begin{aligned}
\sum_{T_M=T}^{\infty} \sum_{T_{M-1}=T_M}^{\infty} \cdots \sum_{T_1=T_M}^{\infty} P\left[E_{\mathbf{T}}^N\right] &= P\left[E_{\mathbf{T}^0}^N\right] O\left(N^{\frac{1}{2}-M}\right) e^{-N\left(a \ln T - J_{\mathbf{T}^0,\mathbf{M}^*}\left(\mathbf{u}_{M^*}^*\right)\right)} \\
&= P\left[E_{\mathbf{T}^0}^N\right] O\left(N^{\frac{1}{2}-M} e^{-N}\right) . \qquad (4.10)
\end{aligned}$$

Substituting terms 4.6, 4.8 and 4.10 in 4.3 we obtain,

$$P\left[X_{M,0}^N > NB_M\right] \leq P\left[E_{\mathbf{T}^0}^N\right] \left(1 + L(N) O\left(e^{-N}\right)\right)$$

The proof completes by recalling corollary 3.1 and noting that the dominant error term is $O\left(\frac{1}{N}\right)$. □

Proposition 4.2 *If the arrival processes $\Lambda_i(-t,0)$ $i=1,2$, satisfy:*

For every $t>0$, $\lambda_{i,t}\le K_i$ and $\exists\ \alpha_i>0$ and t_i such that $\forall\ t\ge t_i$,

$$\mathbf{P}[\Lambda_i(-t,0)>C_i t]\le \widetilde{C}_i t^{-\alpha_i} \tag{4.11}$$

where $n_i\Lambda_i<C_i<v_i$ with C_i, $\widetilde{C}_i$ constants then $\lim_{T\to\infty}\frac{J_{i,T}(Tv_i)}{\ln T}>0$ for $v_i>n_i\Lambda_i$ and thus assumption 1 is satisfied.

The proof can be found in [27]. The condition 4.11 corresponds to the source being long-tailed with bounded support. Hence, our results in theorems 4.1 hold for this kind of inputs as well.

5 Discussion for the lower bound

An exact approximation of the queue length decay rate in an M queue GPS system is only reported in [28] to the best of our knowledge. In that paper the same queueing system is studied under the large buffer asymptotic regime. Sample path large deviation techniques are used and thus assumptions like those introduced in [11] and [17] are in place.

Since the many sources results developed so far have drawn upon large deviations for random variables, it is in general difficult to relate the conditions to the process case as is done in the large buffer asymptotics. Nevertheless, it is of interest to know if the bounds are tight. For this we need to derive lower bounds. We show that under certain readily verifiable conditions, the lowerebound coincides with the upperbound.

Assumption 3 *Let $\epsilon_i:=C\phi_i-n_i\Lambda_i>0$ and d some finite constant. Then for $i=1,2,...,M$, $J_{s,i}\left(Cs\phi_i+a+\epsilon_i\left(s-t\right)\right)>J_{t,i}\left(Ct\phi_i+a\right)$ for $\left(C(s-t)\phi_i-n_i\Lambda_i s\right)^+\le a\le\epsilon_1 t+d$ and $t,s\in\mathcal{N}$ with $0<t<s$, $t<\infty$. Here, $x^+=x$ if $x>0$ and 0 otherwise.*

Proposition 5.1 *Under assumption 3, $T_i^0=T_M^0$ for all $i=1,2,...,M-1$.*

Proof Recall that $\mathbf{T}=(T_1,T_2,...,T_M)$. We want to show that for any $\mathbf{T}':=\left(T_1',T_2'...,T_{M-1}',T_M\right)$ such that $T_i'\ge T_M$, $i=1,2,...,M-1$, $J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{\mathbf{T},M^*}\right)<J_{\mathbf{T}',M^*}\left(\mathbf{u}^*_{\mathbf{T}',M^*}\right)$ if for at least one $i=1,2,...,M-1$, $T_i'>T_i$ and $T_i'=T_i$ for the rest.

Define $\mathbf{T}^i_k:=(T_1,...,T_{i-1},T_i+k,T_{i+1},...,T_M)$ for finite$k\in\mathcal{N}$ (by assumption 2 it suffices to restrict our attention to finite times) and $T_i\ge T_M$, $i=1,2,...,M-1$.

We will show that $J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{\mathbf{T},M^*}\right)<J_{\mathbf{T}^i_k,M^*}\left(\mathbf{u}^*_{\mathbf{T}^i_k,M^*}\right)$ for an arbitary $i=1,2,...,M-1$. Then this result can be directly generalized to obtain $J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{\mathbf{T},M^*}\right)<J_{\mathbf{T}',M^*}\left(\mathbf{u}^*_{\mathbf{T}',M^*}\right)$.

Let us for the needs of this proof, denote by $\mathcal{D}_{\mathbf{T},M^*}$ the region defined in 3.8 associated with the time vector $\mathbf{T}$ and by $\mathcal{D}_{\mathbf{T}^i_k,M^*}$ the one that corresponds to the time vector $\mathbf{T}^i_k$. We will denote by $\mathbf{u}:=(u_1,u_2,...,u_{M-1})$ all members of $\mathcal{D}_{\mathbf{T}^i_k,M^*}$ and by $\mathbf{v}:=(v_1,v_2,...,v_{M-1})$ all elements of $\mathcal{D}_{\mathbf{T},M^*}$. Then combining proposition 3.1 and expression 3.8 we see that for all $l=1,2,...,M-1$, it suffices to consider $v_l\in I^l_{\mathbf{T},M^*}$ where,

$$I^l_{\mathbf{T},M^*}:=\left(\max\left\{n_l\Lambda_l T_l, C\left(T_l-T_M\right)\phi_l\right\},\min\left\{CT_l\phi_l+\frac{\phi_l}{\phi_l+\phi_M}\sum_{\substack{m=1\\m\ne M,l}}^{M-1}\left(CT_m\phi_m-v_m\right),K_lT_l\right\}\right] \tag{5.1}$$

where the left endpoint is closed if $n_l\Lambda_l T_l < C\left(T_l - T_M\right)\phi_l$.

Similarly, it is not restrictive to assume that $u_i \in I^i_{\mathbf{T}^i_k,M^*}$ where,

$$I^i_{\mathbf{T}^i_k,M^*} := \Big(\max\left\{n_i\Lambda_i\left(T_i+k\right), C\left(T_i+k-T_M\right)\phi_i\right\}, \\ \min\Big\{C\left(T_i+k\right)\phi_i + \frac{\phi_i}{\phi_i+\phi_M}\sum_{\substack{m=1\\ m\neq M,i}}^{M-1}\left(CT_m\phi_m - u_m\right), K_i\left(T_i+k\right)\Big\}\Big] \quad (5.2)$$

and for $i \neq l \in M^*$, $u_l \in I^l_{\mathbf{T}^i_k,M^*}$ where,

$$u_l \in I^l_{\mathbf{T}^i_k,M^*} := \Big(\max\left\{n_l\Lambda_l T_l, C\left(T_l-T_M\right)\phi_l\right\}, \\ \min\Big\{CT_l\phi_l + \frac{\phi_l}{\phi_l+\phi_M}(Ck\phi_i + \sum_{\substack{m=1\\ m\neq M,l,i}}^{M-1}\left(CT_m\phi_m - u_m\right)), K_l T_l\Big\}\Big] \quad (5.3)$$

where for the left endpoint of the last two intervals the same observation holds as for 5.1.

Let us define the interval $\widetilde{I}^i_{\mathbf{T}^i_k,M^*} \subseteq I^i_{\mathbf{T}^i_k,M^*}$ as,

$$\widetilde{I}^i_{\mathbf{T}^i_k,M^*} := \Big[C\phi_i k + \max\left\{n_i\Lambda_i T_i, C\left(T_i-T_M\right)\phi_i\right\}, \\ \min\Big\{C\left(T_i+k\right)\phi_i + \frac{\phi_i}{\phi_i+\phi_M}\sum_{\substack{m=1\\ m\neq M,i}}^{M-1}\left(CT_m\phi_m - u_m\right), K_i\left(T_i+k\right)\Big\}\Big] \quad (5.4)$$

As long as $u_i \in \widetilde{I}^i_{\mathbf{T}^i_k,M^*}$ the vector $\mathbf{v}$ which differs from $\mathbf{u}$ to the i^{th} coordinate which is $u_i - C\phi_i k$ instaed of u$_i$ is such that $\mathbf{v} \in \mathcal{D}_{\mathbf{T},M^*}$. Thus, we have that $u_l = v_l$ for all $l \neq i$ and $u_i - Ck\phi_i = v_i$. Trivially then,

$$\begin{aligned} & J_{M,T_M}\Big(CT_M\phi_M + B_M + \sum_{\substack{m=1\\ m\neq M,i}}^{M-1}\left(CT_m\phi_m - u_m\right) + C\left(T_i+k\right)\phi_i - u_i\Big) \\ = \ & J_{M,T_M}\Big(CT_M\phi_M + B_M + \sum_{\substack{m=1\\ m\neq M,i}}^{M-1}\left(CT_m\phi_m - v_m\right) + CT_i\phi_i - v_i\Big) \end{aligned}$$

and $J_{l,T_l}(u_l) = J_{l,T_l}(v_l)$ for all $l \neq i$. Rewrite v_i as $v_i = n_i\Lambda_i T_i + a$ for some $a \geq (C(T_i-T_M)\phi_i - n_i\Lambda_i T_i)^+$ (to be precice $\alpha \neq 0$) and smaller than the length of the interval $I^l_{\mathbf{T},M^*}$. Hence, $J_{i,T_i}\left(v_i\right) = J_{i,T_i}\left(n_i\Lambda_i T_i + a\right)$ and at the same time

$$J_{i,T_i+k}\left(u_i\right) = J_{i,T_i+k}\left(Ck\phi_i + v_i\right) = J_{i,T_i+k}\left(n_i\Lambda_i\left(T_i+k\right) + \left(C\phi_i - n_i\Lambda_i\right)k + a\right)\ .$$

Thus by assumption 3, $J_{i,T_i+k}\left(u_i\right) > J_{i,T_i}\left(v_i\right)$ implying that $J_{\mathbf{T},M^*}\left(\mathbf{v}\right) < J_{\mathbf{T}^i_k,M^*}\left(\mathbf{u}\right)$ and hence,

$$\inf_{\mathbf{v}\in\mathcal{D}_{\mathbf{T},M^*}} J_{\mathbf{T},M^*}\left(\mathbf{v}\right) < \inf\Big\{J_{\mathbf{T}^i_k,M^*}\left(\mathbf{u}\right) : \mathbf{u} \in \mathcal{D}_{\mathbf{T}^i_k,M^*} \text{ and } u_i \in \widetilde{I}^i_{\mathbf{T}^i_k,M^*}\Big\} \quad (5.5)$$

Let us now assume that $\mathbf{u} \in \mathcal{D}_{\mathbf{T}^i_k, M^*}$ with $u_i \in I^i_{\mathbf{T}^i_k, M^*} \backslash \widetilde{I}^i_{\mathbf{T}^i_k, M^*}$ i.e.,

$$\max\left\{n_i \Lambda_i \left(T_i + k\right), C\left(T_i + k - T_M\right)\phi_i\right\} < u_i < C\phi_i k + \max\left\{n_i \Lambda_i T_i, C\left(T_i - T_M\right)\phi_i\right\} \qquad (5.6)$$

which is empty if $n_i \Lambda_i T_i < C\left(T_i - T_M\right)\phi_i$. Hence we assume that $\max\left\{n_i \Lambda_i T_i, C\left(T_i - T_M\right)\phi_i\right\} = n_i \Lambda_i T_i$. Then let us consider the vector $\mathbf{v} \in \mathcal{D}_{\mathbf{T}, M^*}$ such that $v_i = n_i \Lambda_i T_i + \epsilon$ for arbitrarily small positive constant ϵ. We identify two cases:

a) The vector $\mathbf{v}$ with all $i \neq l \in M^*$, $u_l = v_l$ and v_i defined above is such that $\mathbf{v} \in \mathcal{D}_{\mathbf{T}, M^*}$. It is trivially then seen that,

$$\sum_{\substack{m=1 \\ m \neq M, i}}^{M-1} \left(CT_m \phi_m - v_m\right) + CT_i \phi_i - v_i < \sum_{\substack{m=1 \\ m \neq M, i}}^{M-1} \left(CT_m \phi_m - u_m\right) + C\left(T_i + k\right)\phi_i - u_i$$

since from the interval 5.6 $C\left(T_i + k\right)\phi_i - u_i > CT_i \phi_i - n_i \Lambda_i T_i$ while by assumtion on $\mathbf{v}$

$$\sum_{\substack{m=1 \\ m \neq M, i}}^{M-1} \left(CT_m \phi_m - u_m\right) = \sum_{\substack{m=1 \\ m \neq M, i}}^{M-1} \left(CT_m \phi_m - v_m\right) \ .$$

This implies that,

$$\begin{aligned} & J_{M, T_M}\Big(CT_M \phi_M + B_M + \sum_{\substack{m=1 \\ m \neq M, i}}^{M-1} \left(CT_m \phi_m - v_m\right) + CT_i \phi_i - v_i\Big) \\ < \quad & J_{M, T_M}\Big(CT_M \phi_M + B_M + \sum_{\substack{m=1 \\ m \neq M, i}}^{M-1} \left(CT_m \phi_m - u_m\right) + C\left(T_i + k\right)\phi_i - u_i\Big) \end{aligned}$$

and thus $J_{\mathbf{T}, M^*}\left(\mathbf{v}\right) < J_{\mathbf{T}^i_k, M^*}\left(\mathbf{u}\right)$ since $J_{T_l, l}\left(v_l\right) = J_{T_l, l}\left(u_l\right)$ for all $i \neq l \in M^*$ and $J_{i, T_i}\left(v_i\right) = J_{i, T_i}(n_i \Lambda_i T_i + \epsilon) \leq J_{T_i + k, i}\left(u_i\right)$ for some ϵ sufficiently small positive constant.

b) Let us now consider the case where the vector $\mathbf{v}$ defined in (a) above does not belong to $\mathcal{D}_{\mathbf{T}, M^*}$. That suggests that there exist(s) v_l for $i \neq l \in M^*$ such that $v_l \notin I^l_{\mathbf{T}, M^*}$ and particularly these elements are such that,

$$v_l > \min\Big\{CT_l \phi_l + \frac{\phi_l}{\phi_l + \phi_M}(Ck\phi_i + \sum_{\substack{m=1 \\ m \neq M, l, i}}^{M-1} \left(CT_m \phi_m - v_m\right)), K_l T_l\Big\} \qquad (5.7)$$

We now consider the vector $\mathbf{v}' := \left(v'_1, v'_2, ..., v'_{M-1}\right)$ which has all its elements equal to the corresponding $\mathbf{v}$ above but the ones for which 5.7 holds. The latter components are substituted by,

$$v'_l = \min\Big\{CT_l \phi_l + \frac{\phi_l}{\phi_l + \phi_M}(Ck\phi_i + \sum_{\substack{m=1 \\ m \neq M, l, i}}^{M-1} \left(CT_m \phi_m - v'_m\right)), K_l T_l\Big\}$$

Then it is apparent that $\mathbf{v}' \in \mathcal{D}_{\mathbf{T}, M^*}$. Furthermore,

$$0 = \sum_{\substack{m=1 \\ m \neq M, i}}^{M-1} \left(CT_m \phi_m - v'_m\right) + CT_i \phi_i - v'_i \leq \sum_{\substack{m=1 \\ m \neq M, i}}^{M-1} \left(CT_m \phi_m - u_m\right) + C\left(T_i + k\right)\phi_i - u_i$$

implying that

$$
\begin{aligned}
& J_{M,T_M}\Big(CT_M\phi_M + B_M + \sum_{\substack{m=1\\ m\neq M,i}}^{M-1} (CT_m\phi_m - v'_m) + CT_i\phi_i - v'_i\Big) \\
\leq \quad & J_{M,T_M}\Big(CT_M\phi_M + B_M + \sum_{\substack{m=1\\ m\neq M,i}}^{M-1} (CT_m\phi_m - u_m) + C\left(T_i + k\right)\phi_i - u_i\Big)
\end{aligned}
$$

At the same time $J_{i,T_i}\left(v_i\right) = J_{i,T_i}(n_i\Lambda_i + \epsilon) \leq J_{i,T_i}\left(u_i\right)$ for sufficienlty small positive ϵ; for all l for which $v'_l = u_l$, $J_{T_l,l}\left(v'_l\right) = J_{T_l,l}\left(u_l\right)$ and $J_{T_l,l}\left(v'_l\right) < J_{T_l,l}\left(u_l\right)$ for the rest coordinates. Hence $J_{\mathbf{T},M^*}\left(\mathbf{v}'\right) < J_{\mathbf{T}^i_k,M^*}\left(\mathbf{u}\right)$. Thus, from (a) and (b) we deduce that

$$
\inf_{\mathbf{u}\in\mathcal{D}_{\mathbf{T},M^*}} J_{\mathbf{T},M^*}\left(\mathbf{u}\right) < \inf\Big\{J_{\mathbf{T}^i_k,M^*}\left(\mathbf{u}\right) : \mathbf{u}\in\mathcal{D}_{\mathbf{T}^i_k,M^*} \text{ and } u_i \in I^i_{\mathbf{T}^i_k,M^*}\backslash\widetilde{I}^i_{\mathbf{T}^i_k,M^*}\Big\} \tag{5.8}
$$

Combining inequalities 5.5 and 5.8 we obtain the announced result. □

We now have all necessary results to obtain a lower bound for the tail of the queue length which under the proper assumption on the arrival processes coincides asymptotically to the upper one.

Lemma 5.1 *Let us assume that $\mathbf{u}^*_{\mathbf{T}^0,M^*}$ lies in the interior of the domain $\mathcal{D}_{\mathbf{T}^0,M^*}$. If assumption 3 holds and in addition there exist ϵ_i with $u^*_{i,M^*} + T^0_M\epsilon_i < CT^0_M\phi_i + \phi_i\gamma\left(\mathbf{T}^0, M^*, u^*_{\cdot,M^*} + T^0_M\epsilon.\right)$ such that $J_{i,T^0_M}\left(u^*_{i,M^*} + \delta\right) < J_{i,1}\left(\frac{u^*_{i,M^*}}{T^o_M} + \epsilon_i\right)$ for all $i \in M^*$ and some constant $\delta > 0$ arbitrarily small; then the upper bound given in 4.2 is asymptotically exact.*

Proof Since $\mathbf{u}^*_{M^*}$ (we drop the subscript $\mathbf{T}^0$ for sake of brevity) lies in the interior of the domain $\mathcal{D}_{\mathbf{T}^0,M^*}$we are able to find ϵ_i's such that $r_i \leq C\phi_i + \phi_i\gamma\left(\mathbf{1}, M^*, r.\right)$, $i = 1, 2, ..., M-1$ with $r_i := u^*_{i,M^*}/T^0_M + \epsilon_i$. From [45] relation (55) we can write

$$
X^N_{M,0} \geq \Lambda^N_M\left(-T^0_M, 0\right) - CNT^0_M + \sum_{i=1}^{M-1} \inf_{0\leq t_i\leq T^0_M} \Big\{\Lambda^N_i\left(-T^0_M, -t_i\right) + r_i t_i\Big\} \tag{5.9}
$$

The latter relation has been shown in [45] for r_i's the *feasible rates* (for the definition of *feasible rate* see relation(20) in [45]). One can easily check that $r_i \leq C\phi_i + \phi_i\gamma\left(\mathbf{1}, M^*, r.\right)$ implies that r_i's can play the role of the *feasible rates.*

Hence, from 5.9 we obtain within distribution equivalence,

$$
X^N_{M,0} \geq \Lambda^N_M\left(-T^0_M, 0\right) - CNT^0_M + \sum_{i=1}^{M-1} \inf_{0\leq t_i\leq T^0_M} \Big\{\Lambda^N_i\left(-t_i, 0\right) + r_i\left(T^0_M - t_i\right)\Big\}
$$

Therefore,

$$
\begin{aligned}
& P\left[X^N_{M,0} \geq NB_M\right] \\
\geq \quad & P\Big[\Lambda^N_M\left(-T^0_M, 0\right) + \Big(\sum_{i=1}^{M-1} r_i - CN\Big)T^0_M + \sum_{i=1}^{M-1} \inf_{0\leq t_i\leq T^0_M} \Big\{\Lambda^N_i\left(-t_i, 0\right) - r_i t_i\Big\} \geq NB_M\Big]
\end{aligned} \tag{5.10}
$$

Let us now define $A = \cap_{i=1}^{M-1} \{A_i\}$ with

$$A_i = \left\{ \cap_{s=1}^{T_M^0} \Lambda_i^N(-s,0) < r_i s \right\}$$

then we have

$$\begin{aligned} & P\left[X_{M,0}^N \geq NB_M\right] \\ \geq \quad & P\Big[\Lambda_M^N\left(-T_M^0,0\right) + \Big(\sum_{i=1}^{M-1} r_i - CN\Big)T_M^0 + \sum_{i=1}^{M-1}\left(\Lambda_i^N\left(-T_M^0,0\right) - r_i T_M^0\right) \geq NB_M \;\mid\; A\Big]\, P[A] \end{aligned} \tag{5.11}$$

Invoking the fact that $r_i > n_i \Lambda_i$ (see proposition 3.1) and the independence among the arrival processes $\{\lambda_{i,n}\}_{n=0}^{\infty}$, of different classes for all i, it is readily seen that,

$$\begin{aligned} P[A] &= \prod_{i=1}^{M-1} P\left[\cap_{s=1}^{T_M^0} \Lambda_i^N(-s,0) < r_i s\right] \\ &\geq \prod_{i=1}^{M-1}\left(1 - \sum_{s=1}^{T_M^0} P\left[\Lambda_i^N(-s,0) \geq r_i s\right]\right) \\ &= 1 - O\left(e^{-N}\right) \end{aligned}$$

Then combining the last two realations we obtain,

$$\begin{aligned} & P\left[X_{M,0}^N \geq NB_M\right] \\ \succeq \quad & P\Big[\Lambda_M^N\left(-T_M^0,0\right) - CNT_M^0 + \sum_{i=1}^{M-1}\Lambda_i^N\left(-T_M^0,0\right) \geq NB_M \;\mid\; A\Big] \\ \geq \quad & \int_{\mathcal{D}^\delta} P\Big[\Lambda_M^N\left(-T_M^0,0\right) - CNT_M^0 + \sum_{i=1}^{M-1} u_i \geq NB_M\Big] \prod_{i=1}^{M-1} P\left[\Lambda_i^N\left(-T_M^0,0\right) = u_i \;\mid\; A_i\right] d\mathbf{u} \end{aligned} \tag{5.12}$$

where

$$\mathcal{D}^\delta := \left\{\mathbf{u} \in \mathcal{D}_{\mathbf{T}^0,M^*} : u_{i,M^*}^* - \delta \leq u_i \leq u_{i,M^*}^* + \delta, \quad i = 1,2,...,M-1\right\}$$

for arbitarily small $\delta > 0$ which is not empty since u_{i,M^*}^* for $i = 1,2,...,M-1$ belong to an open interval (see statement of the theorem and proposition 3.1).

Applying the total probability law for $i = 1,2,...,M-1$

$$P\left[\Lambda_i^N\left(-T_M^0,0\right) = u_i\right] \leq P\left[\Lambda_i^N\left(-T_M^0,0\right) = u_i \mid\; A_i\right] + \sum_{s=1}^{T_M^0} P\left[\Lambda_i^N(-s,0) \geq r_i s\right] \tag{5.13}$$

Invoking the well known result that the most likely time scale to overflow in a bufferless system is one (cf. proposition 3 [27], [8] and [13]) we see that

$$\sum_{s=1}^{T_M^0} P\left[\Lambda_i^N(-s,0) \geq r_i s\right] = O\left(P\left[\Lambda_i^N(-1,0) \geq r_i\right]\right) = O\left(e^{-N\,J_{i,1}(r_i)}\right)$$

Morover,

$$P\left[\Lambda_i^N\left(-T_M^0,0\right)=u_i\right]=O\Big(e^{-N\,J_{i,T_M^0}(u_i)}\Big)$$

and thus by assumtpion and inequality 5.13 we have for all $u_i\in\mathcal{D}^\delta$,

$$P\left[\Lambda_i^N\left(-T_M^0,0\right)=u_i\ \mid\ A\right]\geq P\left[\Lambda_i^N\left(-T_M^0,0\right)=u_i\right]\left(1-O\left(e^{-N}\right)\right) \tag{5.14}$$

Therefore, from 5.12 we deduce that ,

$$\begin{aligned}&P\left[X_{M,0}^N\geq NB_M\right]\\&\succeq\int_{\mathcal{D}^\delta}P\Big[\Lambda_M^N\left(-T_M^0,0\right)-CNT_M^0+\sum_{i=1}^{M-1}u_i\geq NB_M\Big]\prod_{i=1}^{M-1}P\left[\Lambda_i^N\left(-T_M^0,0\right)=u_i\right]d\mathbf{u}\end{aligned} \tag{5.15}$$

which asymptotically coincides with the upper bound if we recall theorem 3.1. □

6 The Virtual delay

In this section we present an upper bound for the probability a packet entering say the M^{th} buffer has to wait at least t time slots until service. The quantity is said to be the virtual delay and we denote it $V_i(-t)$, $i=1,2,\ldots,M$ if the virtual customer enters queue i at the beginning of the time slot $-t$.

The virtual delay problem has been treated by Paschalidis in [33] in the large buffer context. Our approach closely follows [33].

Let us denote by $V_{i,t}^N$ the time elapsed from the arrival moment t (precisely the beginning of the time slot $(t,t+1)$) of a packet at queue $i\in\mathcal{M}$ until the moment the server attends to this packet. We assume that the packet arrived before any other in the burst at t i.e. it sees the workload in the buffer corresponding to the end of the time instant $t-1$.. Then $\left\{V_{i,t}^N\right\}_t$ is called the virtual delay process in queue i. It is not hard to show that (see [33]):

$$P[V_{i,-t}^N\geq t]=P[W_{i,0}^N\geq\Lambda_i^N(-t,0)]\ . \tag{6.1}$$

Alternatively, one may assume that the packet whose delay we want to compute arrived after all others in the burst that contains this packet. Then expression similar to 6.1 can be established for the sojourn time of the packet in the system. In this paper by virtual delay we will refer to the situation described by formula 6.1.

Hence, an upper bound for $P[V_M^N(-t)\geq t]$ can be obtained by applying the analysis carried out for the tail of the workload in the M^{th} queue substituting NB_M by $\Lambda_M^N(-t,0)$. Then the total rate function reads:

$$\begin{aligned}&J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}\right)\\&:=\sum_{i\in\mathcal{S}}J_{i,T_i}\left(u_i\right)+\sum_{k\in\mathcal{S}^*\setminus M}J_{k,T_k}\left(CT_k\phi_k+\phi_k\gamma\left(\mathcal{S},u.\right)\right)+J_{M,T_M-t}\left(CT_M\phi_M+\phi_M\gamma\left(\mathcal{S},u.\right)\right)\end{aligned} \tag{6.2}$$

and

$$D_{\mathbf{T},\mathcal{S}}\left(\mathbf{u},N\right):=\frac{N^{|\mathcal{S}|-M/2}}{(2\pi)^{M/2}\sqrt{\prod_{i\in\mathcal{S}}\sigma_{i,T_i}^2\left(u_i\right)\prod_{i\in\mathcal{S}^*}\sigma_{i,T_i-t\mathbf{1}\{j=M\}}^2\left(\mathbf{u}\right)}\prod_{i\in\mathcal{S}^*}\tau_{i,T_i-t\mathbf{1}\{j=M\}}\left(\mathbf{u}\right)} \tag{6.3}$$

where for all $i \in \mathcal{S}^*$,

$$\tau_{i,T_i - t\mathbf{1}\{j=M\}}(\mathbf{u}) = \tau_{i,T_i}(CT_i\phi_i + \phi_i\gamma(\mathcal{S}, u.))$$

and similarly,

$$\sigma^2_{i,T_i - t\mathbf{1}\{j=M\}}(\mathbf{u}) = \sigma^2_{i,T_i}(CT_i\phi_i + \phi_i\gamma(\mathcal{S}, u.)) \ .$$

Furthermore, the region $\mathcal{D}_{\mathbf{T},\mathcal{S}}$ now reads,

$$\begin{aligned} \mathcal{D}_{\mathbf{T},\mathcal{S}} := \{\mathbf{u}_{\mathcal{S}} \in \Re^{|\mathcal{S}|} \ : \ C\left(T_i - T_M\right)\phi_i \quad &\leq \quad u_i \leq \min\left\{CT_i\phi_i + \phi_i\gamma\left(\mathcal{S}, u.\right), K_iT_i\right\} \quad \forall i \in \mathcal{S} \\ CT_j\phi_j + \phi_j\gamma\left(\mathcal{S}, u.\right) \quad &\leq \quad K_j(T_j - t\mathbf{1}_{\{j=M\}}) + \quad \forall j \in \mathcal{S}^* \ \} \ . \end{aligned} \tag{6.4}$$

Moreover, note that $J_{i,T}(Tv_i) \leq J_{i,T'}(Tv_i)$ for $T > T'$ and thus assumption 1 implies that $\lim_{T\to\infty} \frac{J_{M,T-t}(Tv)}{T} > 0$ for $T > t \geq 1$, $t \in \mathcal{N}$. Then we obtain the following theorem whose proof differs from theorem 4.2 to minor technicalities and thus it is omitted.

Theorem 6.1 *If assumption 2 (with the infimum taken over $T_i \geq T_M \geq t+1$, $i = 1, 2, \ldots, M-1$) and assumption 1 hold then for the tail of the virtual delay in the M^{th} queue we have for $t \in \mathcal{N}$ finite,*

$$P[V^N_M(-t) \geq t] \leq e^{-N\, J_{\mathbf{T}^0,M^*}(\mathbf{u}^*_{M^*})} \sum_{\mathcal{S}^*\setminus M \subseteq \mathcal{G}} C_{\mathbf{T}^0,\mathcal{S}}\left(\mathbf{u}^*_{\mathcal{S}}, N\right)\left(1 + O\left(\frac{1}{N}\right)\right) \tag{6.5}$$

for $\mathcal{S} \subseteq M^*, \mathcal{G}$ as defined in lemma 3.3 and $C_{\mathbf{T}^0,\mathcal{S}}\left(\mathbf{u}^*_{\mathcal{S}}, N\right)$ as defined in corollary 3.1.

Under the proper assumptions introduced in section 5 one can show that the upper bound given by the theorem above becomes asymptotically exact.

Generally the virtual delay is an upperbound on the actual delay. Results in the same spirit for a two-queue Head-of-Line (HOL) priority system appear in [39].

7 Numerical results

We conclude the paper with some numerical evidence of the tightness of the bounds that have been presented in this paper. We compare the analytical results given by Theorem 4.1 with those obtained via simulations for different scalings N. We report results for a GPS system with 3 buffers.

We consider a three-queue GPS system where the inputs are assumed to be deterministic ON/OFF with uniformly distributed shift. We assume that the server's capacity is 10 and $B_i = 2$ for all $i = 1, 2, 3$. The weights allocated to each queue by the GPS server is $\phi_1 = .35$, $\phi_2 = .45$ and $\phi_3 = .2$ while the traffic mix accessing the buffers is $n_1 = 20$, $n_2 = 5$ and $n_3 = 2.9$. The period of the inputs is taken to be 60. The precise characteristics of each traffic class are the following:

Class 1: $\lambda_{1,t} = 5$ for $t = 0, 1$ and $\lambda_{1,t} = 0$ for $2 \leq t \leq 59$.

Class 2: $\lambda_{2,t} = 10$ for $t = 0, 1, 2, 3, 4$ and $\lambda_{2,t} = 0$ for $5 \leq t \leq 59$.

Class 3: $\lambda_{3,t} = 10$ for $t = 0, 1, 2, 3$ and $\lambda_{3,t} = 1$ for $4 \leq t \leq 59$.

Figures 2 3 and 4 provide a numerical verification of the performance of the upper bound obtained in theorem 4.1 in the context of a three-queue GPS system.

Figures $2 - 4$ depict a comparison of the upper bound obtained by theorem 4.1 for the tail of the buffer occupancy with the corresponding simulation 90% confidence intervals for each buffer in the system. The results clearly indicate that the theoretical bound improves as N increases while it can be accurate even for small values of N.

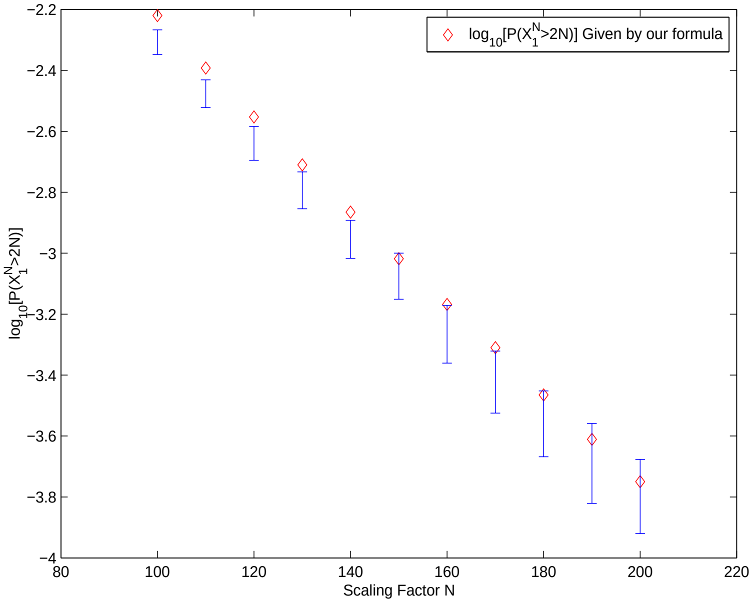

Figure 2: Three-queue GPS system. $P[W_{1,0}^N > NB_1]$ **vs** the scaling factor N against the corresponding simulation 90% confidence interval.

Concluding remarks

In this paper we have obtained bounds for the tail distribution of the buffer content in a multi-buffer system with GPS service when they handle many sources. Our first result was a detailed analysis of the GPS schedule and identifying the so-called eventually stable states over which the estimations are done. We have provided sufficient conditions for the bound to be an asymptotic equivalent and also derived the asymptotics for the virtual delay. We have also clearly exposed the relevance of the critical time-scales for overflow of each buffer and have shown that the upperbound is exact when the time-scales coalesce.

References

[1] F. Bacceli and P. Brémaud : *Elements of queueing theory*, Applications of Mathematics, Vol. **26**, Springer-Verlag, New York, 1995

[2] S. Borst, O. Boxma and P. R. Jelenkovic, Generalized Processor Sharing with Long-Tailed Traffic Sources, In:*Teletraffic Engineering in a Competitive World*, Proc. ITC-16, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), pp. 345-354, 1999.

[3] D. Bertsimas, I.Ch. Paschalidis, J.N. Tsitsiklis : Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach, IEEE Trans. Automat. Control 43 (1998), no. 3, 315–335.

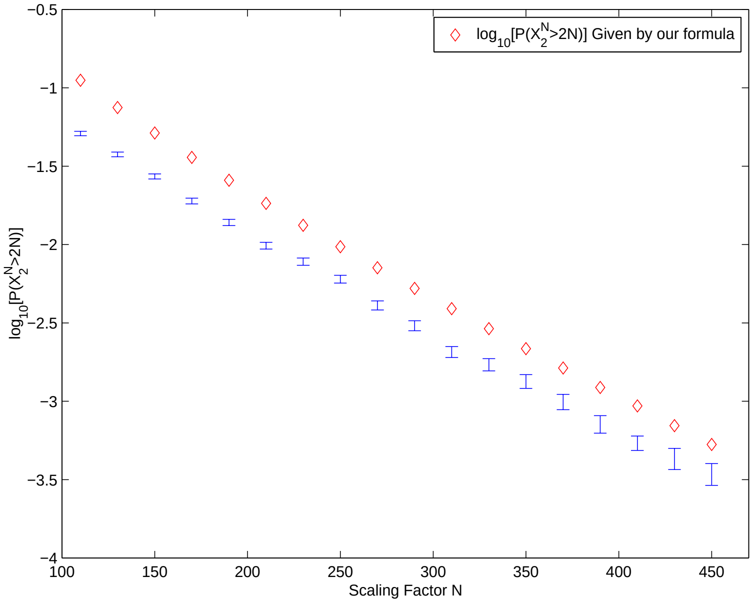

Figure 3: Three-queue GPS system. $P[W^N_{2,0} > NB_2]$ **vs** the scaling factor N against the corresponding simulation 90% confidence interval.

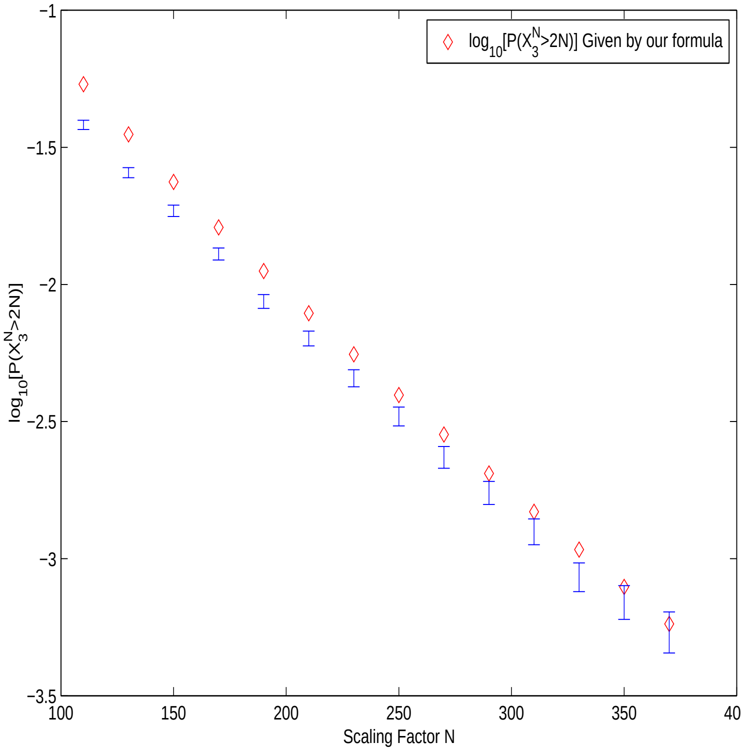

Figure 4: Three-queue GPS system. $P[W^N_{3,0} > NB_3]$ **vs** the scaling factor N against the corresponding simulation 90% confidence interval.

[4] D. Bertsimas, I.Ch. Paschalidis, J.N. Tsitsiklis : Large deviations analysis of the generalized processor sharing policy, Queueing Systems, 32:319-349, 1999.

[5] R.R. Bahadur and R. Rao : On deviations of the sample mean, *Ann. Math. Statist.*, **31**, pp: 1015-1027.

[6] N. Bleistein, A.R. Handelsman : Asymptotic expansions of integrals, Holt, Rinehart and Winston, 1975 USA.

[7] A. A. Borovkov : *Stochastic processes in queueing theory*, Springer-Verlag, N. Y., 1976

[8] D.D. Botvich, N.G. Duffield : Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. Queueing Systems Theory Appl. 20 (1995), no. 3-4, 293–320.

[9] O. Boxma and J. W. Cohen : *Boundary value problems in queueing system analysis*,North-Holland Mathematics Studies, 79. North-Holland Publishing Co., Amsterdam, 1983.

[10] C. S. Chang : Stability, queue length, and delay of deterministic and stochastic queueing networks *IEEE Trans. Autom. Contr.*, Vol **39**, No. 5, May 1994.

[11] C. S. Chang, P. Heidelberger, S. Juneja, P. Shahabuddin: Effective bandwidth and fast simulation of ATM in intree networks, *Performance Evaluation*, Vol 20, pp.45-56, 1994.

[12] D. Clark, S. Shenker, and L. Zhang. Supporting real time applications in an integrated services packet network : architecture and mechanism. In *Proc. Sigcomm'92*, Baltimore, 1992.

[13] C. Courcoubetis, R. Weber : Buffer overflow asymptotics for a buffer handling many traffic sources, J. Appl. Probab. 33 (1996), no. 3, 886–903.

[14] C. Courcoubetis, R. Weber : Estimation of overflow probabilities for state-dependent service of traffic streams with dedicated buffers, Talk at the RSS Workshop in Stochastic Networks, Endiburgh, U.K., 1995.

[15] R.L. Cruz: A calculus of network delay, Part I : network elements in isolation, *IEEE Trans. Information Theory 37(3)*, January 1991.

[16] Delas S., Mazumdar R. R. and Rosenberg C. : Cell loss asymptotics for buffers handling a large number of independent stationary sources with HOL service priorities, *Queueing Systems*, Vol. 40, 2002, pp. 205-226.

[17] A. Dembo, T. Zajic: Large deviations: From Empirical mean and measure to partial sums process, *Stoc. Proc. and Appl.*, Vol 57, pp. 191-224, 1995.

[18] A. Dembo and O. Zeitouni: *Large deviation techniques and applications*, Jones and Bartlett, U.S.A. 1993.

[19] A. Demers, S. Keshav, and S. Shenker: Analysis and simulation of a Fair Queueing algorithm, in *ACM Sigcom'89*, 1989.

[20] G. de Veciana, C. Courcoubetis, J. Walrand : Decoupling bandwidths for networks: a decomposition approach to resource management, IEEE INFOCOM '94, 1994.

[21] G. de Veciana and G. Kesidis : Bandwidth allocation for multiple qualities of service using generalized processor sharing, *IEEE Trans. Inform. Theory*, vol **42**, pp.268-272, 1996.

[22] G. Fayolle, P. J. B. King and I. Mitrani; The solution of certain two-dimensional markov models, *Adv. in Appl. Prob.*, 14, 1982, pp. 295-308

[23] P. W. Glynn, W. Whitt : Logarithmic asymptotics for stady-state tail probabilities, *J. Appl. Prob* vol 31A, pp. 131-159, 1994.

[24] F. M. Guillemin, R. R. Mazumdar, P. Dupuis and J. Boyer: Analysis of the fluid weighted fair queueing system, pre-print June 2001, submitted for publication.

[25] V. Korolyuk, N. Portenko, A. Skorohod and A. Turbin; *Handbook on Probability Theory and Mathematical Statistics*, (in Russian), Nauka, Moscow, 1985

[26] C. Kotopoulos and R. R. Mazumdar; Many sources asymptotics for a 2-buffer system with generalized processor sharing, submitted Feb. 2002.

[27] N. Likhanov N. and R. R. Mazumdar : Cell loss asymptotics in buffers fed with a large number of independent stationary sources, *Journal of App. Prob.*, March 1999, Vol. 36, No. 1, 1999, pp. 86-96.

[28] L. Massoulié : Large deviation estimates for polling and weighted fair queueing service systems, submitted to *Advan. Perf. Anal.*, Vol. 2, (1999), No. 3, 103-128

[29] T. Nakatsuka : The substability and ergodicity of complicated queueing systems, J. Appl. Prob., **23**, pp.193-200, 1986.

[30] O'Connell N.: Large deviations for queue lengths at a multi-buffered resource. J. Appl. Probab. 35 (1998), no. 1, 240–245

[31] A. K. Parekh and R. G. Gallager : A generlized processor sharing approach to floe control in integrated services networks: The single node case, *IEEE/ACM Trans. Networking*, Vol. 1, No. 3, pp.344-357, June 1993.

[32] A. K. Parekh R. G. Gallager : A generalized processor sharing approach to flow control in integrated services networks: The multiple node case, *IEEE/ACM Trans. Networking*, Vol. 2, No. 2, pp.137-150, April 1994. .

[33] I. Ch.Paschalidis: Class specific quality of service guarantees in multimedia communication networks, Automatica, 35 (12), 1951-1969, 1999.

[34] V. V. Petrov V. V. : *Sums of independent random variables*, Springer-Verlag, Berlin.

[35] V. V. Petrov : On the probabilities of large deviations for sums of independent random variables, Theor.Prob.Applic., **10**,287-298.

[36] A. Shwartz and A. Weiss : Large deviations for performance analysis, Chapman&Hall, 1995.

[37] A. Simonian, J. Guibert : Large deviations approximation for fluid queues fed by a large number of on-off sources, *Proceedings of ITC 14, Antibes*, pp. 1013-1022, 1994

[38] F. Toomey :Bursty traffic and finite capacity queues. Queueing networks with blocking. Ann. Oper. Res. 79 (1998), 45–62.

[39] S. Shakottai and R. Srikant : Many-sources delay asymptotics with applications to priority queues. Queueing Syst. Theory Appl. 39 (2001), no. 2-3, 183–200.

[40] W. Whitt : Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues, *Telecommunications Systems* Vol. 2, pp. 71-107, 1993.

[41] R. Wong R. : *Asymptotic approximations of integrals*, Academic Press, Boston 1989.

[42] O. Yaron, M. Sidi, Generalized processor sharing networks with exponentially bounded burstiness arrivals : *Proc. IEEE INFOCOM '94*, June 1994.

[43] O. Yaron and M. Sidi : Performance and stability of communication networks via robust exponential bounds, *IEEE/ACM Transactions on Networking*, 1(3):372–385, 1993.

[44] Z.-L. Zhang, D. Towsley, J. Kurose : Statistical analysis of generalized processor sharing scheduling scheme, *IEEE J.S.A.C.*, Vol. 13, No. 6, pp. 1071-1080, August 1995.

[45] Z-L. Zhang : Large deviations and the generalized processor sharing scheduling for a two-queue system. Queueing Systems Theory Appl. 26 (1997), no. 3-4, 229–254.

Appendix

A Proof of lemma 3.2

Proof By definition,

$$\begin{aligned} J_{\mathbf{T},\mathcal{S}_\mathcal{Q}}\left(\mathbf{u}^*_{\mathcal{S}_\mathcal{Q}}\right) &= \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}_\mathcal{Q}}} J_{\mathbf{T},\mathcal{S}_\mathcal{Q}}(\mathbf{u}) \\ &= \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}_\mathcal{Q}}} \Bigg\{ \sum_{i\in\mathcal{S}_\mathcal{Q}} J_{i,T_i}(u_i) + \sum_{j\in\mathcal{S}^*_\mathcal{Q}\setminus M} J_{j,T_j}\left(CT_j\phi_j + \phi_j\gamma\left(\mathcal{S}_\mathcal{Q}, u.\right)\right) \\ &\qquad + J_{M,T_M}\left(CT_M\phi_M + \phi_M\gamma\left(\mathcal{S}_\mathcal{Q}, u.\right) + B_M\right)\Bigg\} \qquad \text{(A.1)} \end{aligned}$$

Invoking 3.8,

$$\begin{aligned} \mathcal{D}_{\mathcal{S}_\mathcal{Q}} = \{\mathbf{u}\in\Re^{|\mathcal{S}_\mathcal{Q}|} \;:\; & C\left(T_i - T_M\right)\phi_i \le u_i \le \min\left\{CT_i\phi_i + \phi_i\gamma\left(\mathcal{S}_\mathcal{Q}, u.\right), K_iT_i\right\} \text{ for all } i\in\mathcal{S}_\mathcal{Q} \\ & CT_j\phi_j + B_M\mathbf{1}_{\{j=M\}} + \phi_j\gamma\left(\mathcal{S}_\mathcal{Q}, u.\right) \le K_jT_j \;\; \forall j\in\mathcal{S}^*_\mathcal{Q} \;\} \\ \supset \{\; \mathbf{u}\in\Re^{|\mathcal{S}_\mathcal{Q}|} \;:\; & C\left(T_i - T_M\right)\phi_i \le u_i \le \min\left\{CT_i\phi_i + \phi_i\gamma\left(\mathcal{S}_\mathcal{Q}, u.\right), K_iT_i\right\} \text{ for } i\in\mathcal{S} \\ & \textbf{and } u_l = CT_l\phi_l + \phi_l\gamma\left(\mathcal{S}_\mathcal{Q}, u.\right) \le K_lT_l \;\forall l\in\mathcal{Q} \\ & \textbf{and } CT_j\phi_j + B_M\mathbf{1}_{\{j=M\}} + \phi_j\gamma\left(\mathcal{S}_\mathcal{Q}, u.\right) \le K_jT_j \;\; \forall j\in\mathcal{S}^*_\mathcal{Q}\} \end{aligned}$$

which is a particular partition of $\mathcal{S}_{\mathcal{Q}}$ suggesting that $\mathcal{Q}$ contains only $K_{\mathcal{S}_{\mathcal{Q}}}$-order M-eventually stable queues (not necessarily all of them see definition 3.1). Now lemma 2.3 entails that $\gamma\left(\mathcal{S}_{\mathcal{Q}}, u.\right) = \gamma\left(\mathcal{S}, u.\right)$.

Hence, the last expression implies,

$$\mathcal{D}_{\mathcal{S}_{\mathcal{Q}}} \supset \{\ \mathbf{u} \in \Re^{|\mathcal{S}_{\mathcal{Q}}|}\ : C\left(T_i - T_M\right)\phi_i \leq u_i \leq \min\left\{CT_i\phi_i + \phi_i\gamma\left(\mathcal{S}, u.\right), K_iT_i\right\} \text{ for } i \in \mathcal{S}$$
$$\textbf{and } CT_j\phi_j + B_M\mathbf{1}_{\{j=M\}} + \phi_j\gamma\left(\mathcal{S}, u.\right) \leq K_jT_j \ \ \forall j \in \mathcal{S}^*_{\mathcal{Q}}$$
$$\textbf{and } u_l = CT_l\phi_l + \phi_l\gamma\left(\mathcal{S}, u.\right) \leq K_lT_l \text{ for } l \in \mathcal{Q}\ \}$$

Thus, if we write $\mathbf{u} := \left(\mathbf{u}_{\mathcal{S}}, \mathbf{u}_{\mathcal{Q}}\right)$ (recall that $\mathbf{u}_{\mathcal{S}} \in \Re^{|\mathcal{S}|}$ and $\mathbf{u}_{\mathcal{Q}} \in \Re^{|\mathcal{Q}|}$) then the last relation yields,

$$\mathcal{D}_{\mathcal{S}_{\mathcal{Q}}} \supset \mathcal{D}'_{\mathcal{S}_{\mathcal{Q}}} := \{\left(\mathbf{u}_{\mathcal{S}}, \mathbf{u}_{\mathcal{Q}}\right) \in \Re^{|\mathcal{S}_{\mathcal{Q}}|} : \mathbf{u}_{\mathcal{S}} \in \mathcal{D}_{\mathcal{S}} \quad \textbf{and} \quad u_l = CT_l\phi_l + \phi_l\gamma\left(\mathcal{S}, u.\right) \text{ for } l \in \mathcal{Q}\ \}$$

Now A.1 can be written as,

$$\begin{aligned}
J_{\mathbf{T},\mathcal{S}_{\mathcal{Q}}}\left(\mathbf{u}^*_{\mathcal{S}_{\mathcal{Q}}}\right) &= \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}_{\mathcal{Q}}}} J_{\mathbf{T},\mathcal{S}_{\mathcal{Q}}}\left(\mathbf{u}\right) \leq \inf_{\mathbf{u}\in\mathcal{D}'_{\mathcal{S}_{\mathcal{Q}}}} J_{\mathbf{T},\mathcal{S}_{\mathcal{Q}}}\left(\mathbf{u}\right) \\
&= \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}}} \left\{ \sum_{i\in\mathcal{S}} J_{i,T_i}\left(u_i\right) + \sum_{j\in\mathcal{S}^*\backslash M} J_{j,T_j}\left(CT_j\phi_j + \phi_j\gamma\left(\mathcal{S}, u.\right)\right) \right. \\
&\qquad \left. + J_{M,T_M}\left(CT_M\phi_M + \phi_M\gamma\left(\mathcal{S}, u.\right) + B_M\right) \right\} \\
&= \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}}} J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}\right) \qquad\qquad \text{(A.2)}
\end{aligned}$$

Thus, we showed that for any $\mathcal{Q} \subseteq \mathcal{S}^*\backslash M$, $\inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}_{\mathcal{Q}}}} J_{\mathbf{T},\mathcal{S}_{\mathcal{Q}}}\left(\mathbf{u}\right) \leq \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}}} J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}\right)$. This proves 3.16.

To complete the proof note that

$$\inf_{\mathcal{S}\in\mathcal{F}(M^*)} \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}}} J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}\right) \leq \inf_{\mathbf{u}\in\mathcal{D}_{M^*}} J_{\mathbf{T},M^*}\left(\mathbf{u}\right) \leq \inf_{\mathcal{S}\in\mathcal{F}(M^*)} \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}}} J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}\right)$$

where the first inequality trivially follows since $M^* \in \mathcal{F}\left(M^*\right)$ while the second one is due to fomula 3.16. □

B Proof of lemma 3.3

Proof We first prove that (2) implies (1) and (3). Let us assume that $\mathcal{G} \neq \emptyset$ for otherwise the result follows trivially ($\mathcal{S} \equiv M^*$). We write,

$$J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right) = \sum_{i\in\mathcal{S}} J_{i,T_i}\left(u^*_{i,M^*}\right) + \sum_{j\in\mathcal{S}^*\backslash M} J_{j,T_j}\left(u^*_{j,M^*}\right) + J_{M,T_M}\left(CT_M\phi_M + \phi_M\gamma\left(M^*, u^*_{\cdot,M^*}\right) + B_M\right)$$

But since $\mathcal{S}^*\backslash M \subseteq \mathcal{G}$,

$$\begin{aligned}
&J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right) \\
&\quad = \sum_{i\in\mathcal{S}} J_{i,T_i}\left(u^*_{i,M^*}\right) + \sum_{j\in\mathcal{S}^*\backslash M} J_{j,T_j}\left(CT_j\phi_j + \phi_j\gamma\left(M^*, u^*_{\cdot,M^*}\right)\right) \\
&\qquad\qquad + J_{M,T_M}\left(CT_M\phi_M + \phi_M\gamma\left(M^*, u^*_{\cdot,M^*}\right) + B_M\right)
\end{aligned}$$

Invoking lemma 2.3 we deduce that $\gamma\left(M^*, u^*_{\cdot,M^*}\right) = \gamma\left(\mathcal{S}, u^*_{\cdot,M^*}\right)$. Hence, we write,

$$\begin{aligned}
&J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right)\\
&= \sum_{i\in\mathcal{S}} J_{i,T_i}\left(u^*_{i,M^*}\right) + \sum_{j\in\mathcal{S}^*\backslash M} J_{j,T_j}\left(CT_j\phi_j + \phi_j\gamma\left(\mathcal{S}, u^*_{\cdot,M^*}\right)\right)\\
&\quad + J_{M,T_M}\left(CT_M\phi_M + \phi_M\gamma\left(\mathcal{S}, u^*_{\cdot,M^*}\right) + B_M\right)
\end{aligned}$$

Let $\widehat{\mathbf{u}}^*_{\mathcal{S}} \in \Re^{|\mathcal{S}|}$ with $u_i = u^*_{i,M^*}$ $\forall i \in \mathcal{S}$, the last equality reads,

$$J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right) = J_{\mathbf{T},\mathcal{S}}\left(\widehat{\mathbf{u}}^*_{\mathcal{S}}\right)$$

But by the definition of $\mathbf{u}^*_{\mathcal{S}}$,

$$J_{\mathbf{T},\mathcal{S}}\left(\widehat{\mathbf{u}}^*_{\mathcal{S}}\right) \geq J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}^*_{\mathcal{S}}\right)$$

and from lemma 3.2,

$$J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}^*_{\mathcal{S}}\right) \geq J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right)$$

The last three expessions yield,

$$J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right) = J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}^*_{\mathcal{S}}\right) = J_{\mathbf{T},\mathcal{S}}\left(\widehat{\mathbf{u}}^*_{\mathcal{S}}\right)$$

which proves (1) and since $J_{\mathbf{T},\mathcal{S}}(\cdot)$ is strictly convex it possesses a unique infimum suggesting that $\mathbf{u}^*_{\mathcal{S}} = \widehat{\mathbf{u}}^*_{\mathcal{S}}$ which proves (3).

Let us now show that (1) implies (2). Let us denote $\mathcal{G}_1 := \mathcal{S}^* \cap \mathcal{G}$. Suppose that (1) holds and at the same time $\mathcal{S}^*\backslash\{\mathcal{G}_1 \cup M\} \neq \emptyset$. We will end up to contradiction.

Let us define,

$$\begin{aligned}
\mathcal{D} := \Big\{\mathbf{u} \in \Re^{M-1} : C\left(T_i - T_M\right)\phi_i \leq u_i \leq \min\left\{CT_i\phi_i + \phi_i\gamma\left(M^*, u.\right), K_iT_i\right\} \forall i \in \mathcal{S}\\
\textbf{and } u_l = CT_l\phi_l + \phi_l\gamma\left(M^*, u.\right) \leq K_lT_l \ \forall l \in \mathcal{S}^*\backslash M\\
\textbf{and } CT_M\phi_M + B_M + \phi_M\gamma\left(M^*, u.\right) \leq K_MT_M\Big\}
\end{aligned} \tag{B.1}$$

For all $k \in \mathcal{S}^*\backslash(\mathcal{G}_1 \cup M) \neq \emptyset$ (i.e. $k \notin \mathcal{G}$) we have (see statement of the lemma) that $u^*_{k,M^*} < CT_k\phi_k + \phi_k\gamma\left(M^*, u^*_{\cdot,M^*}\right)$ implying that $\mathbf{u}^*_{M^*} \notin \mathcal{D}$. Hence, due to the uniqueness of $\mathbf{u}^*_{M^*}$ we have,

$$J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right) < \inf_{\mathbf{u}\in\mathcal{D}} J_{\mathbf{T},M^*}\left(\mathbf{u}\right) \tag{B.2}$$

Moreover, invoking lemma 2.3 we see that if $\mathbf{u} \in \mathcal{D}$ we have $\gamma\left(M^*, u.\right) = \gamma\left(\mathcal{S}, u.\right)$ and hence $\mathcal{D}$ now reads,

$$\begin{aligned}
\mathcal{D} = \Big\{\mathbf{u} \in \Re^{M-1} : C\left(T_i - T_M\right)\phi_i \leq u_i \leq \min\left\{CT_i\phi_i + \phi_i\gamma\left(\mathcal{S}, u.\right), K_iT_i\right\} \forall i \in \mathcal{S}\\
\textbf{and } u_l = CT_l\phi_l + \phi_l\gamma\left(\mathcal{S}, u.\right) \leq K_lT_l \ \forall l \in \mathcal{S}^*\backslash M\\
\textbf{and } CT_M\phi_M + B_M + \phi_M\gamma\left(\mathcal{S}, u.\right) \leq K_MT_M\Big\}
\end{aligned} \tag{B.3}$$

Thus, if we write $\mathbf{u} := \left(\mathbf{u}_{\mathcal{S}}, \mathbf{u}_{\mathcal{S}^*\backslash M}\right)$ (recall that $\mathbf{u}_{\mathcal{S}} \in \Re^{|\mathcal{S}|}$ and $\mathbf{u}_{\mathcal{S}^*\backslash M} \in \Re^{|\mathcal{S}^*\backslash M|}$) then the last relation implies that,

$$\mathcal{D} \equiv \{\left(\mathbf{u}_{\mathcal{S}}, \mathbf{u}_{\mathcal{S}^*\backslash M}\right) \in \Re^{M-1} : \mathbf{u}_{\mathcal{S}} \in \mathcal{D}_{\mathcal{S}} \quad \textbf{and} \quad u_l = CT_l\phi_l + \phi_l\gamma\left(\mathcal{S}, u.\right) \leq K_lT_l \ \text{ for } l \in \mathcal{S}^*\backslash M\}$$

and B.2 yields,

$$\begin{aligned} J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right) &< \inf_{\mathbf{u}\in\mathcal{D}_{\mathcal{S}}}\Bigg\{\sum_{i\in\mathcal{S}} J_{i,T_i}\left(u_i\right)+\sum_{j\in\mathcal{S}^*} J_{j,T_j}\left(CT_j\phi_j+\phi_j\gamma\left(\mathcal{S},u.\right)\right) \\ &\qquad +J_{M,T_M}\left(CT_M\phi_M+\phi_M\gamma\left(\mathcal{S},u.\right)+B_M\right)\Bigg\} \\ &= J_{\mathbf{T},\mathcal{S}}\left(\mathbf{u}^*_{\mathcal{S}}\right) \end{aligned}$$

which contradicts the assumption that (1) holds. □

C Proof of proposition 4.1

Proof Let us define $C\phi_i - n_i\Lambda_i := \epsilon_i > 0$ and pick $a < \min_{i=1,2\ldots,M}\epsilon_i$. We consider two cases according to whether $\lim_{\sum_{i=1}^M T_i\to\infty}\frac{T_M}{\sum_{i=1}^M T_i}$ is zero or strictly positive.

Let us firstly assume that $\lim_{\sum_{i=1}^M T_i\to\infty}\frac{T_M}{\sum_{i=1}^M T_i}=0$. Then there exist a set $\mathcal{I}\subseteq M^*$ such that for all $i\in\mathcal{I}$, $\lim_{\sum_{i=1}^M T_i\to\infty}\frac{T_i}{\sum_{i=1}^M T_i}>0$ and thus $\lim_{\sum_{i=1}^M T_i\to\infty}T_i=\infty$.

Now from proposition 3.1 we see that $u^*_{i,M^*} > n_iT_i\Lambda_i$ for all i. Consider the case where there exists at least one $i\in\mathcal{I}$such that $u^*_{i,M^*}\geq\left(n_i\Lambda_i+\alpha\right)T_i$. Then (due to the non-negativity of the rate function) we write,

$$J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right)\geq J_{i,T_i}\left(u^*_{i,M^*}\right)\geq J_{i,T_i}\left(\left(n_i\Lambda_i+\alpha\right)T_i\right)$$

and hence by assumption 1,

$$\lim_{\sum_{i=1}^M T_i\to\infty}\frac{J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right)}{\ln\left(\sum_{i=1}^M T_i\right)} > \lim_{\sum_{i=1}^M T_i\to\infty}\frac{\ln T_i}{\ln\left(\sum_{i=1}^M T_i\right)}\frac{J_{i,T_i}\left(\left(n_i\Lambda_i+\alpha\right)T_i\right)}{\ln T_i}>0$$

Let us now consider the case where for all $i\in\mathcal{I}$, $n_i\Lambda_iT_i < u^*_{i,M^*} < \left(n_i\Lambda_i+\alpha\right)T_i$. Due to the bounded support of the arrival processes both u^*_{j,M^*} and $\Lambda_j\left(-T_j,0\right)$ are less than or equal K_jT_j for all $j\in\mathcal{M}$. Then recalling that $C\phi_M > n_M\Lambda_M$ it is readily deduced that $J_{M,T_M}(\mathbf{u})$ is decreasing wrt any coordinate of $\mathbf{u}\in\Re^{M-1}$ as long as $u_i < C\phi_i$, we write,

$$\begin{aligned} &\lim_{\sum_{i=1}^M T_i\to\infty}\frac{J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right)}{\ln\left(\sum_{i=1}^M T_i\right)} \\ &\geq \lim_{\sum_{i=1}^M T_i\to\infty}\frac{J_{M,T_M}\left(CT_M\phi_M+B_M+\sum_{j\notin\mathcal{I}}\left(CT_j\phi_j-u^*_{j,M^*}\right)+\sum_{i\in\mathcal{I}}\left(C\phi_i-n_i\Lambda_i-a\right)T_i\right)}{\ln\left(\sum_{i=1}^M T_i\right)} \\ &\geq \lim_{\sum_{i=1}^M T_i\to\infty}\Bigg(\frac{CT_M\phi_M+B_M+\sum_{j\notin\mathcal{I}}\left(CT_j\phi_j-u^*_{j,M^*}\right)+\sum_{i\in\mathcal{I}}\left(C\phi_i-n_i\Lambda_i-a\right)T_i}{\ln\left(\sum_{i=1}^M T_i\right)} \\ &\qquad -\frac{n_M\ln\left(E\left[e^{\Lambda_M\left(-T_M,0\right)}\right]\right)}{\ln\left(\sum_{i=1}^M T_i\right)}\Bigg) \\ &\geq \lim_{\sum_{i=1}^M T_i\to\infty}\frac{\sum_{i\in\mathcal{I}}\left(C\phi_i-n_i\Lambda_i-a\right)T_i}{\ln\left(\sum_{i=1}^M T_i\right)}>0 \end{aligned}$$

Now we turn to the case where $\lim_{\sum_{i=1}^{M} T_i \to \infty} \frac{T_M}{\sum_{i=1}^{M} T_i} > 0$ and thus for all $i \in \mathcal{M}$, $\lim_{\sum_{i=1}^{M} T_i \to \infty} \frac{T_i}{\sum_{i=1}^{M} T_i} > 0$ which in turn implies that $\lim_{\sum_{i=1}^{M} T_i \to \infty} T_i = \infty$ for all $i \in \mathcal{M}$.

Once again we firstly consider the case where there exists at least one $i \in M^*$ such that $u^*_{i,M^*} \geq (n_i \Lambda_i + \alpha) T_i$. Then the announced result follows in an identical way to the corresponding case above.

It remains the case where for all $i \in M^*$, $n_i \Lambda_i T_i < u^*_{i,M^*} < (n_i \Lambda_i + \alpha) T_i$. For this we write,

$$\lim_{\sum_{i=1}^{M} T_i \to \infty} \frac{J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right)}{\ln\left(\sum_{i=1}^{M} T_i\right)} \geq \lim_{\sum_{i=1}^{M} T_i \to \infty} \frac{J_{M,T_M}\left(CT_M \phi_M + B_M + \sum_{i \in M^*} \left(C\phi_i - n_i \Lambda_i - a\right) T_i\right)}{\ln\left(\sum_{i=1}^{M} T_i\right)} \tag{C.1}$$

Note that

$$\begin{aligned} & CT_M \phi_M + B_M + \sum_{i \in M^*} \left(C\phi_i - n_i \Lambda_i - a\right) T_i \\ = & \left(\sum_{i \in M^*} \left(C\phi_j - n_j \Lambda_j - a\right) + C\phi_M\right) T_M + B_M + \sum_{i \in M^*} \left(C\phi_i - n_i \Lambda_i - a\right)\left(T_i - T_M\right) \\ = & \left(n_M \Lambda_M + \sum_{j=1}^{M} \epsilon_j - a\left(M-1\right)\right) T_M + B_M + \sum_{i \in M^*} \left(C\phi_i - n_i \Lambda_i - a\right)\left(T_i - T_M\right) \end{aligned} \tag{C.2}$$

where by choice $\sum_{j=1}^{M} \epsilon_j - a\left(M-1\right) > 0$. Then from C.1 and the assumption 1,

$$\begin{aligned} & \lim_{\sum_{i=1}^{M} T_i \to \infty} \frac{J_{\mathbf{T},M^*}\left(\mathbf{u}^*_{M^*}\right)}{\ln\left(\sum_{i=1}^{M} T_i\right)} \\ \geq & \lim_{\sum_{i=1}^{M} T_i \to \infty} \frac{\ln T_M}{\ln\left(\sum_{i=1}^{M} T_i\right)} \frac{J_{M,T_M}\left(\left(n_M \Lambda_M + \sum_{j=1}^{M} \epsilon_j - a\left(M-1\right)\right) T_M\right)}{\ln T_M} > 0 \end{aligned}$$

The proof is now complete. □