# Mean Field Analysis of Join-Below-Threshold Load Balancing for Resource Sharing Servers

ILLÉS ANTAL HORVÁTH, MTA-BME Information Systems Research Group, Hungary
ZIV SCULLY, Carnegie Mellon University, USA
BENNY VAN HOUDT, University of Antwerp, Belgium

Load balancing plays a crucial role in many large scale computer systems. Much prior work has focused on systems with First-Come-First-Served (FCFS) servers. However, servers in practical systems are more complicated. They serve multiple jobs at once, and their service rate can depend on the number of jobs in service. Motivated by this, we study load balancing for systems using Limited-Processor-Sharing (LPS). Our model has heterogeneous servers, meaning the service rate curve and multiprogramming level (limit on the number of jobs sharing the processor) differs between servers. We focus on a specific load balancing policy: Join-Below-Threshold (JBT), which associates a threshold with each server and, whenever possible, dispatches to a server which has fewer jobs than its threshold. Given this setup, we ask: how should we configure the system to optimize objectives such as mean response time? Configuring the system means choosing both a load balancing threshold and a multiprogramming level for each server. To make this question tractable, we study the many-server mean field regime.

In this paper we provide a comprehensive study of JBT in the mean field regime. We begin by developing a mean field model for the case of exponentially distributed job sizes. The evolution of our model is described by a differential inclusion, which complicates its analysis. We prove that the sequence of stationary measures of the finite systems converges to the fixed point of the differential inclusion, provided a unique fixed point exists. We derive simple conditions on the service rate curves to guarantee the existence of a unique fixed point. We demonstrate that when these conditions are not satisfied, there may be multiple fixed points, meaning metastability may occur. Finally, we give a simple method for determining the optimal system configuration to minimize the mean response time and related metrics.

While our theoretical results are proven for the special case of exponentially distributed job sizes, we provide evidence from simulation that the system becomes insensitive to the job size distribution in the mean field regime, suggesting our results are more generally applicable.

**ACM Reference Format:**
Illés Antal Horváth, Ziv Scully, and Benny Van Houdt. 2019. Mean Field Analysis of Join-Below-Threshold Load Balancing for Resource Sharing Servers. *Proc. ACM Meas. Anal. Comput. Syst.* 3, 3, Article 57 (December 2019), 21 pages. https://doi.org/10.1145/3366705

## 1 INTRODUCTION

Most studies on large-scale load balancing systems consider servers that operate in a first-come-first-served (FCFS) manner and have a fixed service rate. In such a setting a simple policy such as the Join-Idle-Queue (JIQ) policy achieves asymptotic zero delay in both homogeneous and

Authors' addresses: Illés Antal Horváth, MTA-BME Information Systems Research Group, Budapest, Hungary, horvath.illes.antal@gmail.com; Ziv Scully, Carnegie Mellon University, 5000 Forbes Ave, Pittsburg, PA 15213, USA, zscully@andrew.cmu.edu; Benny Van Houdt, University of Antwerp, Middelheimlaan 1, Antwerp, B-2020, Belgium, benny.vanhoudt@uantwerpen.be.


2476-1249/2019/12-ART57 $15.00
https://doi.org/10.1145/3366705

heterogeneous clusters [17]. However, servers in practical systems are more complex as they serve multiple jobs at once and need to share resources (such as caches).

Motivated by this we consider a cluster of so-called *resource sharing servers*. In a resource sharing server the throughput tends to increase as more jobs start sharing the server, but starts decreasing beyond some point due to contention and thrashing [10]. Hence, in such a server the total service rate $\mu(k)$ depends on the number of jobs $k$ being processed simultaneously. In addition a limit, called the multi-programming level (MPL), is implemented that dictates how many jobs can equally share the server. When the number of jobs in a server exceeds its MPL, some jobs must wait before their service starts. In other words the server operates in a limited processor sharing (LPS) fashion, except that the total service rate also depends on the number of jobs in service. As stated before the rates $\mu(k)$ tend to increase up to some point $\kappa = \text{argmax}_k \mu(k)$, and decreases afterward. In practice the MPL is often set equal to $\kappa$, though this may not be optimal in case of an isolated resource sharing server (see [10, 18]).

In this paper we study the Join-Below-Threshold (JBT) dispatch policy in a heterogeneous cluster of resource sharing servers. In JBT a threshold is associated with every server and the server informs the dispatcher whenever its queue length either reaches or drops below its threshold. This allows the dispatcher to maintain a list of all the server IDs for which the server's queue length is below its associated threshold. Incoming jobs are assigned by the dispatcher to a random server with a queue length below its threshold, unless no such server exists, in which case the job is assigned to a random server. If we set the threshold of a server equal to its MPL, this policy is a natural generalization of the JIQ policy for resource sharing servers as the dispatcher keeps track of all the servers where the incoming job can immediately start service. It also retains the advantage of JIQ that the communication overhead is upper bounded by one message per job.

The main contributions of this paper are the following:

(1) We develop a mean field model to assess the performance of JBT in a heterogeneous cluster of resource sharing servers assuming exponential job sizes, the evolution of which is described by a differential inclusion (DI).
(2) We prove that as the number of servers tends to infinity, the sample paths of the stochastic systems converge towards the unique solution of the DI and the fixed point of the DI corresponds to the weak limit of the stationary measures of the stochastic systems provided that this fixed point is unique.
(3) Contrary to our initial expectations, we show that for arbitrary service rate curves multiple fixed points can exist (even in a homogeneous cluster), which implies that the system is metastable in such case. We identify simple conditions for the existence of a unique fixed point that we expect to hold for real service rate curves.
(4) We indicate how to determine the thresholds that minimize the mean response time (and related metrics) for a given arrival rate. This turns out to be surprisingly easy provided that we do not necessarily assign the same threshold to all the servers of the same type.
(5) We perform simulation experiments that support our belief that the queue length distribution becomes insensitive to the job size distribution as the system size tends to infinity. This indicates that our results are also relevant in systems with highly variable job sizes.

The paper is structured as follows. The model and main notations are introduced in Section 2. The transient and stationary mean field analysis is presented in Sections 3 and 4, respectively. Fixed points are studied in Section 5. Section 6 indicates how to compute the response time distribution, while Section 7 indicates how to set the thresholds to minimize mean response times. Simulation experiments that suggest asymptotic insensitivity are presented in Section 8. Conclusions are drawn in Section 9.

### 1.1 Related Work

Most prior work on load balancing considers servers that operate in a FCFS manner and a simple policy like JIQ achieves asymptotic zero delay. Under the JIQ policy [13, 15] incoming jobs are distributed among the servers via a dispatcher. Servers inform the dispatcher whenever they become idle. The dispatcher maintains a list of idle server IDs. The dispatcher assigns incoming jobs to idle servers, unless the list of idle server IDs is empty. In the latter case the incoming job is assigned to a random server.

Another load balancing policy that received a lot of attention in this setting is the Join-Shortest-Queue-d (JSQ-d) policy, which assigns a job to a server with the shortest queue length among $d$ randomly selected servers [2, 14, 20]. While JSQ-d uses more messages per job compared to JIQ, it does not require any memory at the dispatcher and thus can be readily implemented in a setting with multiple dispatchers. For a detailed discussion on the trade-off between memory and communication overhead we refer to [4].

Studies on load balancing servers that do not operate in a FCFS manner are quite limited. In [2, 19] the authors studied the JSQ-d policy when the servers use processor sharing (PS). In [2, 19] the authors showed that the fixed point of the cavity map is insensitive to the job size distribution, but to the best of our knowledge a complete proof of asymptotic insensitivity is still missing (while asymptotic independence of the queue lengths was proven for FCFS servers [2], it is still an open problem for PS servers). Propagation of chaos was shown for PS servers in [19] over finite time scales, but in order to establish asymptotic insensitivity one still needs to show that the unique fixed point of the mean field model is a global attractor. In [9] the authors studied load balancing systems where the size of incoming jobs is known and provide guidelines on how to get good performance when the servers use the shortest-remaining-processing-time (SRPT) service discipline.

The heavy-traffic delay optimality of the JBT policy was recently studied in [21] in case of a fixed number of servers $N$, a fixed service rate $\mu$, FCFS service and when the same threshold $K$ is utilized by all $N$ servers. Necessary and sufficient conditions on how the threshold $K$ must scale with the heavy traffic parameter $\epsilon$ were identified such that the system is heavy-traffic delay optimal. The objective of this paper is to study the performance of JBT in a large-scale heterogeneous cluster with $\ell$ server types, where each server type $\ell$ is subject to its own service rate curve $\mu^{(\ell)}(k)$ and uses its own MPL $K^{(\ell)}$.

## 2 MODEL AND NOTATION

The server cluster contains $N' = N\sum_{\ell=1}^{L} M_\ell$ servers and one dispatcher. We allow the cluster to be heterogeneous: each server has a fixed type from the set $\{1, \ldots, L\}$ and we assume that we have exactly $NM_\ell$ servers of type $\ell$ , where $M_\ell, N \geq 1$ are integers. The ratio of servers of type $\ell$ is $\gamma_\ell = M_\ell/M$, where $M = \sum_{\ell=1}^{L} M_\ell$. The service rate curve of a type $\ell$ server is denoted as $\mu^{(\ell)}(k)$ and its MPL as $K^{(\ell)}$. A natural choice would be to set $K^{(\ell)}$ equal to $\text{argmax}_k \mu_k^{(\ell)}$, but we allow any choice for $K^{(\ell)}$ in our model. For further use, denote $\mu_k^{(\ell)} = \mu^{(\ell)}(k)$ for $k \leq K^{(\ell)}$ and $\mu_k^{(\ell)} = \mu_{K^{(\ell)}}^{(\ell)}$ for $k > K^{(\ell)}$. Hence $\mu_k^{(\ell)}$ represents the service rate of a type $\ell$ server containing $k$ jobs when its MPL is set to $K^{(\ell)}$.

We can think of the local state of a server as the number of jobs queued at that server, and the global state of the entire cluster as the collection of all the local states. Instead we define the global state as the collection of the numbers $\left(X_k^{(\ell),N}(t), k \geq 0, \ell = 1, \ldots, L\right)$, where $X_k^{(\ell),N}(t)$ is a multiple of $1/N'$ which denotes the fraction of queues of type $\ell$ with $k$ jobs in the queue at time $t$. We further assume that type $\ell$ servers can store at most $B^{(\ell)} \geq K^{(\ell)}$ jobs and normalization is such that $\sum_{\ell=1}^{L}\sum_{k=0}^{B^{(\ell)}} X_k^{(\ell),N}(t) = 1$. This assumption is quite common (e.g., [5, 6]) and mainly simplifies some

of the proofs. It should have no real impact on the limiting results when the system with an infinite buffer is stable as the buffer size can be set arbitrarily large, e.g., $B^{(\ell)} = 10^{50}$. In fact, we show that if the stationary mean field model has a unique fixed point, the loss probability in any finite buffer $B^{(\ell)} \geq K^{(\ell)}$ tends to zero as $N$ tends to infinity.

Jobs arrive at the dispatcher according to a Poisson process with rate $N'\lambda$ and have an exponential job size with mean one. The actual processing time of a job depends on the type of server that executes the job. The dispatcher keeps track of servers of type $\ell$ holding at least $K^{(\ell)}$ jobs for each server type $\ell$. This can be done by simply keeping track of a single bit per server. Note that in our model the choice of the MPL and the threshold of a specific server always coincide. As in the mean field regime we expect to have at most $K^{(\ell)}$ jobs in any type $\ell$ server, increasing the MPL beyond $K^{(\ell)}$ has no impact. In Section 7 we also argue that the mean response time can be minimized by letting the MPLs and the thresholds $K^{(\ell)}$ coincide.

Type $\ell$ servers with at least $K^{(\ell)}$ jobs are considered *full*, while type $\ell$ servers with fewer jobs are deemed *available*. Note the system behaves as the classic JIQ if $\mu_k^{(\ell)}$ does not depend on $k$, $K^{(\ell)} = 1$ for all $\ell$ and jobs are served in FCFS order. Full servers of type $\ell$ with strictly more than $K^{(\ell)}$ jobs are also called *overloaded*. Upon a job arrival, the dispatcher chooses a server uniformly at random from the available servers, regardless of their type, and sends the job there. In case there are no available servers, the dispatcher chooses a server uniformly at random from the entire cluster of $N'$ servers. If the job is assigned to a type $\ell$ server holding $B^{(\ell)}$ jobs, it is lost. We assume that

$$\sum_{\ell=1}^{L} \gamma_\ell \mu_{K^{(\ell)}}^{(\ell)} > \lambda, \tag{1}$$

as the maximum service rate of the system should exceed the arrival rate. Note that this condition also corresponds to the stability condition in case the servers have an infinite queue length.

Clearly $X^N(t) = (X_k^{(\ell),N}(t), k \in \{0, 1, \ldots, B^{(\ell)}\}, \ell \in \{1, \ldots, L\})$ is a finite state Markov process for any fixed $N$ on the state space

$$E^{(N)} = \left\{ (x_k^{(\ell)})_{\ell \in \{1,\ldots,L\}, k \in \{0,\ldots,B^{(\ell)}\}} \middle| 0 \leq x_k^{(\ell)} \leq 1, NMx_k^{(\ell)} \in \mathbb{Z}, \sum_{\ell=1}^{L} \sum_{k=0}^{B^{(\ell)}} x_k^{(\ell)} = 1, \sum_{k=0}^{B^{(\ell)}} x_k^{(\ell)} = \gamma_\ell \right\}.$$

For $x \in E^{(N)}$ denote $a(x) = \sum_{\ell=1}^{L} \sum_{k=0}^{K^{(\ell)}-1} x_k^{(\ell)}$ as the fraction of available servers when the system is in state $x$. The job arrival rate in state $x \in E^{(N)}$ at a given queue of type $\ell$ with queue length $k-1$ is given by

$$\lambda_{k-1}^{(\ell)}(x) = \begin{cases} \lambda/a(x) & \text{if } k \leq K^{(\ell)}; \\ 0 & \text{if } k > K^{(\ell)} \text{ and } a(x) > 0; \\ \lambda & \text{if } k > K^{(\ell)} \text{ and } a(x) = 0, \end{cases} \tag{2}$$

while the service rate in state $x$ at the same queue equals

$$\mu_k^{(\ell)}(x) = \begin{cases} \mu_k^{(\ell)} & \text{if } k < K^{(\ell)}; \\ \mu_{K^{(\ell)}}^{(\ell)} & \text{if } k \geq K^{(\ell)}, \end{cases} \tag{3}$$

meaning the service rate does not depend on $x$.

Define $E$ as

$$E = \left\{ (x_k^{(\ell)})_{\ell \in \{1,\ldots,L\}, k \in \{0,\ldots,B^{(\ell)}\}} \middle| 0 \leq x_k^{(\ell)} \leq 1, \sum_{\ell=1}^{L} \sum_{k=0}^{B^{(\ell)}} x_k^{(\ell)} = 1, \sum_{k=0}^{B^{(\ell)}} x_k^{(\ell)} = \gamma_\ell \right\},$$

then $E^{(N)} \subset E$ and $E$ is a compact and convex subset of $\mathbb{R}^{L+B}$ where $B = \sum_{\ell=1}^{L} B^{(\ell)}$.

## 3 TRANSIENT MEAN-FIELD ANALYSIS

We start by showing that the Markov process $X^N(t) = (X_k^{(\ell),N}(t), \ell \in \{1,\dots,L\}, k \in \{0,1,\dots,B^{(\ell)}\})$ is a density dependent population process with a discontinuous drift $f(x)$. Let $Y^{(N)}(t)$ be a sequence of continuous time Markov chains on $E^{(N)} \subset (MN)^{-1}\mathbb{Z}^d$ with $d \geq 1$ for $N \geq 1$. $Y^{(N)}(t)$ is called a density dependent population process if there exists a set $\mathcal{L} \subset M^{-1}\mathbb{Z}^d$ (with $0 \notin \mathcal{L}$), such that for each $\eta \in \mathcal{L}$ and $x \in E^{(N)}$, the transition rate from state $x$ to state $x + \eta/N$ can be written as $\beta_\eta(x)N \geq 0$, where $\beta_\eta(x)$ does not depend on $N$. The $i$-th component of $Y^{(N)}(t)$ can be seen as the density of individuals of a population of size $MN$ that are in state $i$ and a transition $\eta$ changes the number of individuals in state $i$ by $M\eta_i$. The drift $f(x)$ is defined as $f(x) = \sum_{\eta\in\mathcal{L}} \eta\beta_\eta(x)$. In our case we clearly have $d = B + L$ and the set $\mathcal{L}$ has size $2(B + L)$ as explained below.

Let $e_k^{(\ell)}$ be the $k + 1 + \sum_{v=1}^{\ell-1}(1 + B^{(v)})$-th row of the identity matrix of size $B + L$. There are two types of transitions for this Markov chain: service completions and job arrivals. When a service completion occurs in a type $\ell$ server with length $k > 0$ in state $x$, the state changes to $x + (e_{k-1}^{(\ell)} - e_k^{(\ell)})/(MN)$. Such service completions occur at rate $\mu_k^{(\ell)} x_k^{(\ell)} MN$ which can be written as $\beta_{ser,k}^{(\ell)}(x)N$, where $\beta_{ser,k}^{(\ell)}(x)$ does not depend on $N$. Arrivals in a type $\ell$ server with $k-1$ jobs change the state from $x$ to $(x + e_k^{(\ell)} - e_{k-1}^{(\ell)})/(MN)$. The rate of these transitions is given by $\lambda_{k-1}^{(\ell)}(x)x_{k-1}^{(\ell)}MN$, which we denote as $\beta_{arr,k-1}^{(\ell)}(x)N$. Let $e_{ser,k}^{(\ell)} = (e_{k-1}^{(\ell)} - e_k^{(\ell)})/M$, $e_{arr,k-1}^{(\ell)} = (e_k^{(\ell)} - e_{k-1}^{(\ell)})/M$ and define the drift

$$f(x) = \sum_{\ell=1}^{L}\sum_{k=1}^{B^{(\ell)}} \left( e_{ser,k}^{(\ell)}\beta_{ser,k}^{(\ell)}(x) + e_{arr,k-1}^{(\ell)}\beta_{arr,k-1}^{(\ell)}(x)\right). \tag{4}$$

Due to (2) and (3) and the above discussion the drift $f(x) = (f_1^{(1)}(x),\dots,f_{B^{(\ell)}}^{(\ell)}(x))$ in state $x \in E$ is given by

$$\begin{aligned}
f_k^{(\ell)}(x) &= 1[k > 0](\lambda_{k-1}^{(\ell)}(x)x_{k-1}^{(\ell)} - \mu_k^{(\ell)}x_k^{(\ell)}) - 1[k < B^{(\ell)}](\lambda_k^{(\ell)}(x)x_k^{(\ell)} - \mu_{k+1}^{(\ell)}x_{k+1}^{(\ell)}) \\
&= 1[k < B^{(\ell)}]\mu_{k+1}^{(\ell)}x_{k+1}^{(\ell)} - 1[k > 0]\mu_k^{(\ell)}x_k^{(\ell)} + 1[k > K^{(\ell)}]1[a(x) = 0]\lambda(x_{k-1}^{(\ell)} - 1[k < B^{(\ell)}]x_k^{(\ell)}) \\
&\quad + 1[k = K^{(\ell)}]\lambda\left(1[a(x) > 0]x_{k-1}^{(\ell)}/a(x) - 1[a(x) = 0]x_k^{(\ell)}\right) \\
&\quad + 1[k < K^{(\ell)}]1[a(x) > 0]\lambda(1[k > 0]x_{k-1}^{(\ell)} - x_k^{(\ell)})/a(x).
\end{aligned} \tag{5}$$

These drifts are discontinuous due to the presence of the indicator functions involving $a(x)$. As such the set of ODEs $dx(t)/dt = f(x(t))$ does not have a solution. Instead we define a differential inclusion $dx(t)/dt \in F(x(t))$, where $F$ is the set-valued function defined by

$$F(x) = \text{conv}\left(\left\{f(y)\middle| x = \lim_{n\to\infty} y_n, f(y) = \lim_{n\to\infty} f(y_n)\right\}\right), \tag{6}$$

where conv$(A)$ is the convex closure of the set $A$. A Filippov solution to $dx(t)/dt \in F(x(t))$ is an absolutely continuous function, that is, it is almost everywhere differentiable.

THEOREM 1. *Let $X^N(t)$ be the density dependent population process with drift $f$ specified by (5). If $X^N(0) \to x_0 \in E$ in probability then for any finite $T > 0$ we have*

$$\sup_{t\in[0,T]} ||X^N(t) - x(t)|| \to 0,$$

*in probability, where $x(t)$ is a solution of the differential inclusion $dx(t)/dt \in F(x)$ and $F$ is defined by (6).*

Proof. As $X^N(t)$ is a density dependent population process with $d = L + B$, it suffices to verify the three conditions of Theorem 4 in [7]. We first note that $\sup_{x\in E} \beta^{(\ell)}_{ser,k}(x) \le \mu^{(\ell)}_k M$. Further, the expressions in (2) imply that $\sup_{x\in E} \beta^{(\ell)}_{arr,k}(x) \le \lambda M$ as $x^{(\ell)}_{k-1}/a(x) \le 1$ when $a(x) > 0$. Thus, condition 1, which demands that the transition rates from state $x \in E$ are bounded, holds as

$$\sup_{x\in E} \sum_{\ell=1}^{L} \sum_{k=1}^{B^{(\ell)}} (\beta^{(\ell)}_{ser,k}(x) + \beta^{(\ell)}_{arr,k-1}(x)) \le BM(\lambda + \max_{\ell=1}^{L} \mu^{(\ell)}_{K^{(\ell)}}). \tag{7}$$

Further, condition 2, which demands that $\sum_{\eta\in\mathcal{L}} ||\eta|| sup_x \beta_\eta(x) < \infty$, is verified by noting that

$$\sum_{\ell=1}^{L} \sum_{k=1}^{B^{(\ell)}} \left( ||e^{(\ell)}_{ser,k}|| \beta^{(\ell)}_{ser,k}(x) + ||e^{(\ell)}_{arr,k-1}|| \beta^{(\ell)}_{arr,k-1}(x) \right) \le \sum_{\ell=1}^{L} \sum_{k=1}^{B^{(\ell)}} \frac{\sqrt{2}}{M} M(\mu^{(\ell)}_k + \lambda) \le \sqrt{2}B(\lambda + \max_{\ell=1}^{L} \mu^{(\ell)}_{K^{(\ell)}}). \tag{8}$$

The final condition to apply Theorem 4 in [7] requires that there exists a $c > 0$ with $||f(x)|| \le c(1 + ||x||)$ for all $x \in E$. This condition is in fact redundant as condition 2 implies that there exists a constant $c$ (being $\sqrt{2}B(\lambda + \max_{\ell=1}^{L} \mu^{(\ell)}_{K^{(\ell)}})$ in our case) such that $||f(x)|| \le c$. □

In the next subsection we describe the evolution of any solution $x(t)$ of the DI that remains within $E$.

### 3.1 Differential inclusion

We now specify the set-valued function $F(x)$ for $f$ defined in (5). Whenever $a(x) > 0$, we clearly have $F(x) = \{f(x)\}$. When $a(x) = 0$ the set $F(x)$ is the convex closure of all the drift vectors $f(y)$ that can be obtained as the limit points of a sequence of drifts $f(y_n)$ where $y_n$ converges to $x$. When applied to (5) we have for any $x$ with $a(x) = 0$ that

$$\begin{aligned} F(x) = \Bigg\{ f(x) \Bigg| & 0 \le \alpha_0, \alpha^{(\ell)}_k \le 1, \alpha_0 + \sum_{\ell=1}^{L} \sum_{k=0}^{K^{(\ell)}} \alpha^{(\ell)}_k = 1, \\ & f^{(\ell)}_k(x) = 1[k < B^{(\ell)}]\mu^{(\ell)}_{k+1} x^{(\ell)}_{k+1} - 1[k > 0]\mu^{(\ell)}_k x^{(\ell)}_k + 1[k < K^{(\ell)}]\lambda(1[k > 0]\alpha^{(\ell)}_{k-1} - \alpha^{(\ell)}_k) \\ & + 1[k = K^{(\ell)}]\lambda(\alpha^{(\ell)}_{K^{(\ell)}-1} - \alpha_0 x^{(\ell)}_k) + 1[k > K^{(\ell)}]\alpha_0\lambda(x^{(\ell)}_{k-1} - 1[k < B^{(\ell)}]x^{(\ell)}_k) \Bigg\}. \end{aligned} \tag{9}$$

Note that the expression for $f^{(\ell)}_k(x)$ is identical to (5) with $1[a(x) = 0]$ replaced by $\alpha_0$ and $1[a(x) > 0]x^{(l)}_k/a(x)$ by $\alpha^{(l)}_k$. We now argue that the differential inclusion

$$\frac{d}{dt} x(t) \in F(x(t)),$$

with $x(0) = x_0$ has a (Filippov) solution for any $x_0 \in E$ such that $x(t) \in E$ for $t \ge 0$.

Define $\mathcal{H} = \{x \in E | a(x) = 0\}$. For any $x \in E \setminus \mathcal{H}$, the drift is fully determined by (5), that is,

$$\begin{aligned} f^{(\ell)}_k(x) = {} & 1[k < B^{(\ell)}]\mu^{(\ell)}_{k+1} x^{(\ell)}_{k+1} - 1[k > 0]\mu^{(\ell)}_k x^{(\ell)}_k + 1[k = K^{(\ell)}]\lambda x^{(\ell)}_{k-1}/a(x) \\ & + 1[k < K^{(\ell)}]\lambda(1[k > 0]x^{(\ell)}_{k-1} - x^{(\ell)}_k)/a(x). \end{aligned} \tag{10}$$

The set of ODEs $dx(t)/dt = f(x(t))$ has a unique local solution for $x(0) \in E \setminus \mathcal{H}$ as the drift $f$ is locally Lipschitz continuous, that is, for any $x \in E \setminus \mathcal{H}$ there exists an open neighborhood $V_x$ of $x$ and a constant $L_x$ such that for any $y, z \in V_x$ we have $||f(y) - f(z)|| \le L_x||y - z||$. Note that the drift $f$ is not globally Lipschitz continuous on $E \setminus \mathcal{H}$.

For $x \in \mathcal{H}$ there are two possibilities. Let

$$\mathcal{H}_s = \left\{ x \in \mathcal{H} \middle| \sum_{\ell=1}^{L} \mu_{K^{(\ell)}}^{(\ell)} x_{K^{(\ell)}}^{(\ell)} \leq \lambda \right\}.$$

If $x \in \mathcal{H}_s$ the rate at which available servers are created does not exceed the arrival rate and $a(x)$ remains zero, which implies that a *sliding motion* on $\mathcal{H}_s$ occurs. In this case the parameters $\alpha_0$ and $\alpha_k^{(\ell)}$ must be such that $a(x)$ remains zero and $x_k^{(\ell)}$ remains larger than or equal to zero. Looking at (9) with $k = 0$ this yields that $\alpha_0^{(\ell)} = 0$ (for $K^{(\ell)} > 1$) as the drift of $x_0^{(\ell)}$ is given by $-\lambda\alpha_0^{(\ell)}$. Similarly for $0 < k < K^{(\ell)} - 1$ we have by induction that $\alpha_k^{(\ell)} = 0$. For $k = K^{(\ell)} - 1$ the drift equals $\mu_{K^{(\ell)}}^{(\ell)} x_{K^{(\ell)}}^{(\ell)} - \lambda\alpha_{K^{(\ell)}-1}^{(\ell)}$, meaning

$$\alpha_{K^{(\ell)}-1}^{(\ell)} = \frac{\mu_{K^{(\ell)}}^{(\ell)} x_{K^{(\ell)}}^{(\ell)}}{\lambda},$$

such that $a(x)$ remains zero and $\alpha_0 = 1 - \sum_{\ell=1}^{L} \alpha_{K^{(\ell)}-1}^{(\ell)}$. During this sliding motion on $\mathcal{H}_s$, the drift is therefore given by

$$f_k^{(\ell)}(x) = 1[k < B^{(\ell)}]\mu_{k+1}^{(\ell)} x_{k+1}^{(\ell)} - \mu_k^{(\ell)} x_k^{(\ell)} + 1[k = K^{(\ell)}]\lambda \left( \frac{\mu_{K^{(\ell)}}^{(\ell)} x_{K^{(\ell)}}^{(\ell)}}{\lambda} - \left( 1 - \sum_{\upsilon=1}^{L} \frac{\mu_{K^{(\ell)}}^{(\ell)} x_{K^{(\ell)}}^{(\ell)}}{\lambda} \right) x_k^{(\ell)} \right)$$
$$+ 1[k > K^{(\ell)}]\lambda \left( 1 - \sum_{\upsilon=1}^{L} \frac{\mu_{K^{(\upsilon)}}^{(\upsilon)} x_{K^{(\upsilon)}}^{(\upsilon)}}{\lambda} \right) (x_{k-1}^{(\ell)} - 1[k < B^{(\ell)}] x_k^{(\ell)}), \tag{11}$$

for $k \geq K^{(\ell)}$ and equals zero for $k < K^{(\ell)}$. The set of ODEs that characterize this sliding motion has a unique local solution as the drift $f$ is globally Lipschitz continuous on $\mathcal{H}_s$. It is worth noting that $\alpha_{K^{(\ell)}-1}^{(\ell)}$ is the probability that an incoming job is assigned to a type $\ell$ server with length $K^{(\ell)} - 1$ during this sliding motion and $\alpha_0$ is the probability that the job is assigned to a full or overloaded server.

When $x \in \mathcal{H} \setminus \mathcal{H}_s$ the rate at which available servers are created exceeds the arrival rate and $a(x)$ becomes positive, meaning the solution leaves the set $\mathcal{H}$. As in the previous case $\alpha_k^{(\ell)} = 0$ for $k < K^{(\ell)} - 1$ (for the solution to stay in $E$),

$$\alpha_{K^{(\ell)}-1}^{(\ell)} = \frac{\mu_{K^{(\ell)}}^{(\ell)} x_{K^{(\ell)}}^{(\ell)}}{\sum_{\upsilon=1}^{L} \mu_{K^{(\upsilon)}}^{(\upsilon)} x_{K^{(\upsilon)}}^{(\upsilon)}},$$

and $\alpha_0 = 0$, which implies that $f_k^{(\ell)}(x) = 0$ for $k < K^{(\ell)} - 1$ and

$$f_k^{(\ell)}(x) = 1[k < B^{(\ell)}]\mu_{k+1}^{(\ell)} x_{k+1}^{(\ell)} - \mu_k^{(\ell)} x_k^{(\ell)} + \left( 1[k = K^{(\ell)}] - 1[k = K^{(\ell)} - 1] \right) \frac{\lambda \mu_{K^{(\ell)}}^{(\ell)} x_{K^{(\ell)}}^{(\ell)}}{\sum_{\upsilon=1}^{L} \mu_{K^{(\upsilon)}}^{(\upsilon)} x_{K^{(\upsilon)}}^{(\upsilon)}}. \tag{12}$$

*Remarks:* (1) For $x_0 \notin \mathcal{H}$, the solution $x(t)$ may be such that $a(x(t^*)) = 0$ for some $t^* > 0$, if this happens $x(t^*) \in \mathcal{H}_s$ and a sliding motion starts.

(2) When $B^{(\ell)} = K^{(\ell)}$ for $\ell = 1, \ldots, L$, then $\mathcal{H}_s = \emptyset$ due to (1) and there are no sliding motions.

(3) When $u(x(t^*)) = \sum_{\ell=1}^{L} \sum_{k>K^{(\ell)}} x_k^{(\ell)}(t^*) = 0$ for some $t^* \geq 0$, then $u(x(t)) = 0$ for any $t \geq t^*$. This can be seen by noting that $u(x(t))$ can only become non-zero if $a(x(t)) = 0$ as well, but this implies that $x_{K^{(\ell)}}^{(\ell)}(t) = \gamma_\ell$ for all $\ell$ and we assumed that $\lambda < \sum_{\ell=1}^{L} \gamma_\ell \mu_{K^{(\ell)}}^{(\ell)}$. Hence, if $u(x(t^*)) = 0$ for

some $t^* \geq 0$, then $a(x(t)) > 0$ for $t \geq t^*$ and $x(t)$ is the unique differentiable solution of (10) on $[t^*, \infty)$.

## 4 STATIONARY MEAN-FIELD ANALYSIS

Let $\pi \in E \setminus \mathcal{H}$ be a fixed point of the DI. Due to (10) we find that $f^{(\ell)}_{B^{(\ell)}}(\pi) = -\mu^{(\ell)}_{B^{(\ell)}}\pi^{(\ell)}_{B^{(\ell)}}$, hence $\pi^{(\ell)}_{B^{(\ell)}} = 0$ and by induction we get $\pi^{(\ell)}_k = 0$ for $k > K^{(\ell)}$. In the next two subsections we provide conditions such that the fixed point is unique and define a set of equations to determine the non-zero components $\pi^{(\ell)}_k$ for $k \leq K^{(\ell)}$.

To argue that we cannot have a fixed point on $\mathcal{H}$ under condition (1), first note that such a fixed point must be part of $\mathcal{H}_s$. However on $\mathcal{H}_s$ the system evolves according to the set of ODEs given by (11) and it is not hard to verify that

$$\sum_{\ell=1}^{L}\sum_{k=K^{(\ell)}}^{B^{(\ell)}} k f^{(\ell)}_k(x(t)) = \lambda\left(1 - \sum_{\ell=1}^{L} x^{(\ell)}_{B^{(\ell)}}(t)\right) - \sum_{\ell=1}^{L}\mu^{(\ell)}_{K^{(\ell)}}\gamma_\ell.$$

Thus, under condition (1) the right hand side cannot be zero, while it must be zero in any fixed point $\pi \in \mathcal{H}_s$.

Theorem 2. *Provided that all trajectories of the DI starting in $E$ converge to a unique fixed point $\pi \in E \setminus \mathcal{H}$, the sequence of stationary measures $X^N(\infty)$ of the finite state Markov chains $X^N(t)$ converges weakly to the dirac measure of the fixed point $\pi$.*

Proof. The result is immediate from Theorem 3.5 and Corollary 3.9 in [16], as the Birkhoff center of the DI is the singleton containing the fixed point $\pi$ when all the solutions starting in $E$ converge to $\pi$. □

Thus, if a fixed point $\pi$ is a global attractor, the stationary queue length distribution of a type $\ell$ server weakly converges to $(\pi^{(\ell)}_0, \pi^{(\ell)}_1, \ldots, \pi^{(\ell)}_{K^{(\ell)}}, 0, \ldots, 0)$.

We now show that if the fixed point is unique, then it is necessarily a global attractor. For $x, \tilde{x} \in E$ we state that $x \leq_a \tilde{x}$ if $\sum_{k\geq i} x^{(\ell)}_k \leq \sum_{k\geq i} \tilde{x}^{(\ell)}_k$ for $\ell = 1, \ldots, L$ and $i = 1, \ldots, K^{(\ell)}$.

Theorem 3. *Let $x(t, x^*)$ be the unique solution of the DI with $x(0, x^*) = x^*$. If $x^* \leq_a \tilde{x}^*$, then $x(t, x^*) \leq_a x(t, \tilde{x}^*)$.*

Proof. Denote $y^{(\ell)}_i(t, x^*) = \sum_{k\geq i} x^{(\ell)}_k(t, x^*)$. We need to show that if $x(t, x^*) \leq_a x(t, \tilde{x}^*)$ and $y^{(\ell)}_i(t, x^*) = y^{(\ell)}_i(t, \tilde{x}^*)$ for some $\ell$ and $i$, then $\frac{d}{dt}y^{(\ell)}_i(t, x^*) \leq \frac{d}{dt}y^{(\ell)}_i(t, \tilde{x}^*)$. We start with the case where $x(t, x^*)$ and $x(t, \tilde{x}^*)$ both belong to $E \setminus \mathcal{H}$. Due to (10) we have

$$\frac{d}{dt}y^{(\ell)}_i(t, x^*) = -\mu^{(\ell)}_i(y^{(\ell)}_i(t, x^*) - y^{(\ell)}_{i+1}(t, x^*)) + 1[i \leq K^{(\ell)}]\lambda\frac{y^{(\ell)}_{i-1}(t, x^*) - y^{(\ell)}_i(t, x^*)}{1 - \sum_{\ell=1}^{L} y^{(\ell)}_K(t, x^*)}. \tag{13}$$

Therefore, when $x(t, x^*) \leq_a x(t, \tilde{x}^*)$ and $y^{(\ell)}_i(t, x^*) = y^{(\ell)}_i(t, \tilde{x}^*)$, we have

$$\frac{d}{dt}(y^{(\ell)}_i(t, \tilde{x}^*) - y^{(\ell)}_i(t, x^*)) \geq \mu^{(\ell)}_i(y^{(\ell)}_{i+1}(t, \tilde{x}^*) - y^{(\ell)}_{i+1}(t, x^*)) + 1[i \leq K^{(\ell)}]\lambda\frac{y^{(\ell)}_{i-1}(t, \tilde{x}^*) - y^{(\ell)}_{i-1}(t, x^*)}{1 - \sum_{\ell=1}^{L} y^{(\ell)}_K(t, \tilde{x}^*)} \geq 0,$$

as $-1/(1 - \sum_{\ell=1}^{L} y^{(\ell)}_K(t, x^*)) \geq -1/(1 - \sum_{\ell=1}^{L} y^{(\ell)}_K(t, \tilde{x}^*))$.

Next assume $x(t,x^*)$ and $x(t,\tilde{x}^*)$ both belong to $\mathcal{H}_s$. By (11) we observe that for $i > K^{(\ell)}$

$$\frac{d}{dt}y_i^{(\ell)}(t,x^*) = -\mu_i^{(\ell)}(y_i^{(\ell)}(t,x^*) - y_{i+1}^{(\ell)}(t,x^*)) + \left(\lambda - \sum_{v=1}^{L}\mu_{K^{(v)}}^{(v)}(\gamma_v - y_{K^{(v)}+1}^{(v)}(t,x))\right)(y_{i-1}^{(\ell)}(t,x^*) - y_i^{(\ell)}(t,x^*)), \tag{14}$$

while $\frac{d}{dt}y_i^{(\ell)}(t,x^*) = 0$ for $i \le K^{(\ell)}$. Hence, when $x(t,x^*) \le_a x(t,\tilde{x}^*)$ and $y_i^{(\ell)}(t,x^*) = y_i^{(\ell)}(t,\tilde{x}^*)$

$$\frac{d}{dt}(y_i^{(\ell)}(t,\tilde{x}^*) - y_i^{(\ell)}(t,x^*)) \ge \mu_i^{(\ell)}(y_{i+1}^{(\ell)}(t,\tilde{x}^*) - y_{i+1}^{(\ell)}(t,x^*)) + \left(\lambda - \sum_{v=1}^{L}\mu_{K^{(v)}}^{(v)}(\gamma_v - y_{K^{(v)}+1}^{(v)}(t,\tilde{x}))\right)(y_{i-1}^{(\ell)}(t,x^*) - y_i^{(\ell)}(t,x^*)) \ge 0,$$

for $i > K^{(\ell)}$ and it equals zero for $i \le K^{(\ell)}$.

The case where $x(t,x^*) \in \mathcal{H}_s$ and $x(t,\tilde{x}^*) \in E \setminus \mathcal{H}$ cannot occur when $x(t,x^*) \le_a x(t,\tilde{x}^*)$ as $\sum_{\ell=1}^{L} y_{K^{(\ell)}}^{(\ell)}(t,x) = 1$ and $\sum_{\ell=1}^{L} y_{K^{(\ell)}}^{(\ell)}(t,x) < 1$ in such case. We continue with the case where $x(t,x^*) \in E \setminus \mathcal{H}$ and $x(t,\tilde{x}^*) \in \mathcal{H}_s$. If $y_i^{(\ell)}(t,x^*) = y_i^{(\ell)}(t,\tilde{x}^*)$ for $i \le K^{(\ell)}$, then $y_i^{(\ell)}(t,x^*) = \gamma_\ell$ as $y_i^{(\ell)}(t,\tilde{x}^*) = \gamma_\ell$. This implies that $y_{i-1}^{(\ell)}(t,x^*) = \gamma_\ell$ and by (13) we see that $\frac{d}{dt}y_i^{(\ell)}(t,x^*) = -\mu_i^{(\ell)}(y_i^{(\ell)}(t,x^*) - y_{i+1}^{(\ell)}(t,x^*)) \le 0$, while for $i \le K^{(\ell)}$ we have $\frac{d}{dt}y_i^{(\ell)}(t,\tilde{x}^*) = 0$. When $i > K^{(\ell)}$ and $y_i^{(\ell)}(t,x^*) = y_i^{(\ell)}(t,\tilde{x}^*)$, we immediately see from (13) and (14) that $\frac{d}{dt}y_i^{(\ell)}(t,\tilde{x}^*) - \frac{d}{dt}y_i^{(\ell)}(t,x^*) \ge 0$. We end by noting that we do not need to consider the cases with $x(t,x^*)$ or $x(t,\tilde{x}^*)$ in $\mathcal{H} \setminus \mathcal{H}_s$ as we immediately leave this set and never return to it afterwards. □

Theorem 4. *If the DI has a unique fixed point, it is a global attractor on E.*

Proof. Using the order $\le_a$, we clearly have $x_{min} \le_a x \le_a x_{max}$ for any $x \in E$, with $(x_{min})_0^{(\ell)} = \gamma_\ell$ and $(x_{max})_{B^{(\ell)}}^{(\ell)} = \gamma_\ell$ for $\ell = 1,\dots,L$. Hence, by Theorem 3 it suffices to show that the trajectories staring in $x_{min}$ and $x_{max}$ both converge to $\pi$. Now as $x_{min} = x(0,x_{min}) \le_a x(t-s,x_{min})$, we have $x(s,x_{min}) \le_a x(t,x_{min})$ for any $s < t$. Further, $x(t,x_{min}) \le_a \pi$ as $x_{min} \le_a \pi$ and $x(t,\pi) = \pi$. Hence, $x(t,x_{min})$ converges as $t$ tends to infinity and because $E$ is a compact set, it converges to a fixed point of the DI due to Theorem 1.4 in [11, p248]. By the uniqueness of the fixed point $x(t,x_{min})$ converges to $\pi$. The argument starting in $x_{max}$ proceeds similarly. □

Corollary 1. *If the DI has a unique fixed point $\pi$, the sequence of stationary measures $X^N(\infty)$ of the finite state Markov chains $X^N(t)$ converges weakly to the dirac measure of the fixed point $\pi$.*

### 4.1 Model validation

Before we proceed by identifying conditions for having a unique fixed point, we briefly study the accuracy of the mean field approximation by comparing with simulation. The only difference between the simulation and the mean field analysis is that the number of servers is finite in the simulation.

We use the service rate curves $\mu_k^{(1)}$ and $\mu_k^{(2)}$ depicted in Figure 1. The service rate curve $\mu_k^{(1)}$ is the same as in [10, 18], with maximum 1.25 at $k = 4$ and 5, while the service rate curve $\mu_k^{(2)}$ is similar in shape, but with maximum 1.7 at $k = 9$ and 10. Table 1 contains simulation results for the mean response time for various combinations of $\rho$ and $N'$. The remaining parameters of the server cluster are $\gamma_1 = 0.6$, $\gamma_2 = 0.4$, $K^{(1)} = 4$ and $K^{(2)} = 9$. The system load $\rho$ is defined as $\rho^{-1} = \frac{1}{\lambda}\sum_{\ell=1}^{L}\gamma_k\mu_{K^{(\ell)}}^{(\ell)}$.

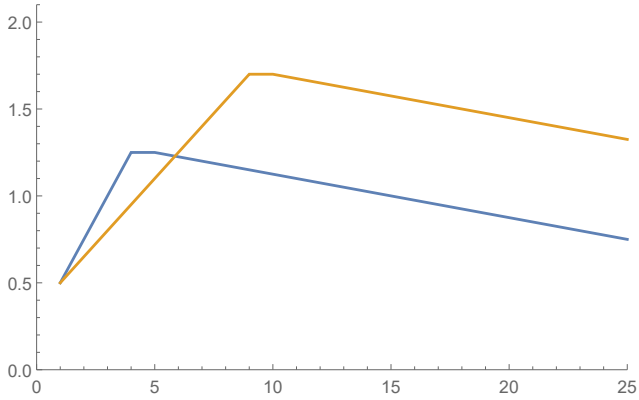



Fig. 1. Typical service rate curves

| $\rho$ | 0.6000 | 0.7000 | 0.8000 | 0.9000 |
|---|---|---|---|---|
| $N' = 10$ | 3.5337 | 3.7642 | 3.9771 | 4.3590 |
| $N' = 20$ | 3.5261 | 3.7550 | 3.9498 | 4.1752 |
| $N' = 40$ | 3.5225 | 3.7509 | 3.9449 | 4.1094 |
| $N' = 80$ | 3.5207 | 3.7490 | 3.9434 | 4.0955 |
| $N' = 160$ | 3.5199 | 3.7478 | 3.9426 | 4.0942 |
| mean field | 3.5190 | 3.7467 | 3.9417 | 4.0942 |

Table 1. Mean response time for various system loads and number of servers

In Table 1, the mean field mean response time is calculated from (24), while the other elements are obtained using simulations, with a deviation of less than $10^{-4}$. Table 1 shows that for the mean response time, the mean field approximation is already very accurate even for $N' = 10$ when the load is not too high. Convergence to the mean field limit can also be observed as $N'$ increases. Dispite the non-smoothness of the drift of the mean field model, it is also possible to develop a refined mean field approximation using the method presented in [8]. Numerical experiments (not reported in the paper) indicate that this further improves the accuracy of the mean field model for small $N'$ values as long as the fixed point is not too close to the boundary area $\mathcal{H}$ where the drift is discontinuous.

## 5 FIXED POINT ANALYSIS

### 5.1 Homogeneous server cluster

First we present the special case when the server cluster is homogeneous, that is, $\ell = 1$, to display some of the ideas with some easier calculations. In this case, we drop $\ell$ entirely from the notation and just have a single service rate curve $(\mu_k, k = 0, \ldots, K)$ for all servers and stationary mean-field limit $(\pi_k, k = 0, 1, \ldots, K)$.

The stability condition is now simply

$$\mu_K > \lambda. \tag{15}$$

The corresponding version of (2) and (3) is that the arrival rate for a queue with $k-1$ jobs is

$$\lambda_{k-1}(x) = \begin{cases} \lambda/a(x) & \text{if } k \le K; \\ 0 & \text{if } k > K \text{ and } a(x) > 0; \\ \lambda & \text{if } k > K \text{ and } a(x) = 0. \end{cases} \tag{16}$$

while the service rate at the same queue is

$$\mu_k(x) = \begin{cases} \mu_k & \text{if } k < K; \\ \mu_K & \text{if } k \ge K. \end{cases} \tag{17}$$

The fixed point equations can then be derived from (10) by setting $dx(t)/dt = 0$ to get

$$\mu_k \pi_k = \frac{\lambda \pi_{k-1}}{1-\pi_K}, \quad \text{for} \quad k = 1, \dots, K, \quad \text{and} \quad \pi_0 + \cdots + \pi_K = 1. \tag{18}$$

The next theorem gives a sufficient condition for the existence of a unique fixed point.

Theorem 5. *If $\mu_K > \lambda$, (18) has an odd number of positive solutions. Further, if there exists an $s \ge 1$ such that $\mu_1, \dots, \mu_{s-1} \le \lambda \le \mu_s, \dots, \mu_{K-1}$, then the solution is unique.*

Proof. Rearranging (18) to express all $\pi_k$ in terms of $\pi_K$ gives

$$\pi_k = \left( \prod_{j=k+1}^{K} \frac{\mu_j}{\lambda} \right) (1-\pi_K)^{K-k} \pi_K \qquad k = 0, \dots, K-1$$

and

$$\sum_{k=0}^{K-1} \left( \prod_{j=k+1}^{K} \frac{\mu_j}{\lambda} \right) (1-\pi_K)^{K-k} \pi_K = 1 - \pi_K,$$

which is a polynomial in $x = 1 - \pi_K$:

$$\sum_{k=0}^{K-1} \left( \prod_{j=k+1}^{K} \frac{\mu_j}{\lambda} \right) x^{K-k}(1-x) = x$$

After simplifying by $x$ and re-indexing with $i = K-1-k$ we have

$$\sum_{i=0}^{K-1} \left( \prod_{j=K-i}^{K} \frac{\mu_j}{\lambda} \right) x^{i}(1-x) = 1.$$

which is equivalent to

$$\left( \frac{\mu_K}{\lambda} - 1 \right) + \sum_{i=1}^{K-1} \left( \prod_{j=K-i+1}^{K} \frac{\mu_j}{\lambda} \right) \left( \frac{\mu_{K-i}}{\lambda} - 1 \right) x^i - \left( \prod_{j=1}^{K} \frac{\mu_j}{\lambda} \right) x^K = 0.$$

The constant term is positive as $\mu_K > \lambda$, the coefficient of $x^K$ is negative, which implies there exists an odd number of positive solutions. Further if $s$ is such that $\mu_1, \dots, \mu_{s-1} \le \lambda \le \mu_s, \dots, \mu_{K-1}$, the middle coefficients change sign exactly once, so by Descartes' rule of signs, the equation has a unique positive solution. Putting in $x = 0$ and $x = 1$ also shows that the solution must be in $(0, 1)$. □

Due to Corollary 1 and Theorem 5, we have the following:

Corollary 2. *If $\mu_K > \lambda$ and there exists an $s \ge 1$ such that $\mu_1, \dots, \mu_{s-1} \le \lambda \le \mu_s, \dots, \mu_{K-1}$, then the sequence of stationary measures $X^N(\infty)$ of the finite state Markov chains $X^N(t)$ converges weakly to the dirac measure of the fixed point $\pi$.*

| $\pi_0$ | $\pi_1$ | $\pi_2$ | $\pi_3$ | type |
|---|---|---|---|---|
| 0.000012 | 0.000236 | 0.017035 | 0.982717 | stable |
| 0.005934 | 0.013912 | 0.122323 | 0.857831 | unstable |
| 0.154791 | 0.101726 | 0.250698 | 0.492785 | stable |

Table 2. Three fixed points for a homogeneous system with $K = 3$, $\lambda = 1$, $\mu_1 = 3$, $\mu_2 = 0.8$ and $\mu_3 = 1.003$.

*5.1.1 Multiple fixed points example.* We now present an example that illustrates that there can be multiple fixed points if we do not put any restrictions on the rates $\mu_k$ (apart from demanding that $\mu_K > \lambda$). Consider the system with $K = 3$, $\lambda = 1$, $\mu_1 = 3$, $\mu_2 = 0.8$ and $\mu_3 = 1.003$. This system has three fixed points that are listed in Table 2. By linearization one finds that two of these fixed points are stable (meaning any trajectory starting sufficiently close to the fixed point converges towards this fixed point) and one is unstable.

For a finite $N'$, the stability regions of the stable fixed points correspond to quasi-stationary distributions concentrated around the stable fixed point. The system spends very long periods of time near one of the stable fixed points, but switches to another fixed point eventually. As $N'$ is increased, the time spent near a fixed point (before switching to another fixed point) gets longer and longer, but remains finite and random for any finite $N'$.

Also, for larger values of $K$, it is possible to construct examples that have more than 3 fixed points (e.g., 3 stable and 2 unstable for $K = 5$), but we do not pursue this direction.

### 5.2 Heterogeneous server cluster

In general, the balance equations for $\pi_k^{(\ell)}$ are

$$\mu_k^{(\ell)} \pi_k^{(\ell)} = \frac{\lambda \pi_{k-1}^{(\ell)}}{1 - \sum_{\ell=1}^{L} \pi_{K^{(\ell)}}^{(\ell)}}, \tag{19}$$

for $k = 0, \ldots, K^{(\ell)}$, $\ell = 1 \ldots, L$ and

$$\pi_0^{(\ell)} + \cdots + \pi_{K^{(\ell)}}^{(\ell)} = \gamma_\ell,$$

for $\ell = 1 \ldots, L$.

Theorem 6. *Assuming the stability condition (1) holds, (19) has an odd number of positive solutions. Further, if $\mu_1^{(\ell)} \le \cdots \le \mu_{K^{(\ell)}}^{(\ell)}$ holds for all $1 \le \ell \le L$, this solution is unique.*

Proof. Similarly to Theorem 5, from (19) we get

$$\pi_k^{(\ell)} = \left( \prod_{j=k+1}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) \left( 1 - \sum_{\ell=1}^{L} \pi_{K^{(\ell)}}^{(\ell)} \right)^{K^{(\ell)}-k} \pi_{K^{(\ell)}}^{(\ell)} \tag{20}$$

for $k = 0, \ldots, K^{(\ell)} - 1$, $\ell = 1, \ldots, L$ and

$$\sum_{k=0}^{K^{(\ell)}-1} \left( \prod_{j=k+1}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) \left( 1 - \sum_{\ell=1}^{L} \pi_{K^{(\ell)}}^{(\ell)} \right)^{K^{(\ell)}-k} \pi_{K^{(\ell)}}^{(\ell)} + \pi_{K^{(\ell)}}^{(\ell)} = \gamma_\ell$$

for $\ell = 1, \ldots, L$, from which

$$\pi_{K^{(\ell)}} = \frac{\gamma_\ell}{1 + \sum_{k=0}^{K^{(\ell)}-1} \left( \prod_{j=k+1}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) \left( 1 - \sum_{\ell=1}^{L} \pi_{K^{(\ell)}}^{(\ell)} \right)^{K^{(\ell)}-k}}, \tag{21}$$

for $\ell = 1, \ldots, L$. We sum (21) for $\ell = 1, \ldots, L$ to get

$$\sum_{\ell=1}^{L} \pi_{K^{(\ell)}} = \sum_{\ell=1}^{L} \frac{\gamma_\ell}{1 + \sum_{k=0}^{K^{(\ell)}-1} \left( \prod_{j=k+1}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) \left(1 - \sum_{\ell=1}^{L} \pi_{K^{(\ell)}}^{(\ell)}\right)^{K^{(\ell)}-k}},$$

which only depends on $x = 1 - \sum_{\ell=1}^{L} \pi_{K^{(\ell)}}$:

$$1 - x = \sum_{\ell=1}^{L} \frac{\gamma_\ell}{1 + \sum_{k=0}^{K^{(\ell)}-1} \left( \prod_{j=k+1}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) x^{K^{(\ell)}-k}}. \tag{22}$$

If we subtract 1 from both sides, divide by $-x$ and use the fact that $\sum_\ell \gamma_\ell = 1$, we get

$$1 = \sum_{\ell=1}^{L} \frac{\gamma_\ell \sum_{k=0}^{K^{(\ell)}-1} \left( \prod_{j=k+1}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) x^{K^{(\ell)}-k-1}}{1 + x \sum_{k=0}^{K^{(\ell)}-1} \left( \prod_{j=k+1}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) x^{K^{(\ell)}-k-1}} = \sum_{\ell=1}^{L} \frac{\gamma_\ell \sum_{i=0}^{K^{(\ell)}-1} \left( \prod_{j=K^{(\ell)}-i}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) x^{i}}{1 + x \sum_{i=0}^{K^{(\ell)}-1} \left( \prod_{j=K^{(\ell)}-i}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \right) x^{i}}. \tag{23}$$

This can be written as $1 = P(x) = \sum_\ell \gamma_\ell p_\ell(x)/(1 + x p_\ell(x))$, where $p_\ell(x) = \sum_{i=0}^{K^{(\ell)}-1} a_i^{(\ell)} x^i$ and $a_i^{(\ell)} = \prod_{j=K^{(\ell)}-i}^{K^{(\ell)}} \mu_j^{(\ell)}/\lambda$. By (1) we have $P(0) > 1$ and as $\sum_\ell \gamma_\ell = 1$ we further have that $P(1) < 1$. Hence, there exists an odd number of solutions $x \in (0, 1)$ such that $P(x) = 1$ (as $P(x)$ is continuous).

We now show that $P(x)$ is decreasing on $(0, 1)$ if the condition $\mu_1^{(\ell)} \leq \cdots \leq \mu_{K^{(\ell)}}^{(\ell)}$ holds and therefore there exists a unique $x \in (0, 1)$ such that $P(x) = 1$. $P(x)$ is clearly decreasing if $p_\ell(x)/(1 + x p_\ell(x))$ is decreasing in $x$ for $\ell = 1, \ldots, L$. Taking the derivative we see that $p_\ell(x)/(1 + x p_\ell(x))$ decreases if $p_l'(x) < p_\ell(x)^2$. Both $p_l'(x)$ and $p_\ell(x)^2$ are polynomials in $x$ and by comparing the coefficients, we see that $p_l'(x) < p_\ell(x)^2$ if $(i + 1)a_{i+1}^{(\ell)} \leq \sum_{s=0}^{i} a_s^{(\ell)} a_{i-s}^{(\ell)}$ for $i = 0, \ldots, K^{(\ell)} - 2$. Now,

$$\sum_{s=0}^{i} a_s^{(\ell)} a_{i-s}^{(\ell)} = \sum_{s=0}^{i} \prod_{j=K^{(\ell)}-s}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \prod_{j'=K^{(\ell)}-(i-s)}^{K^{(\ell)}} \frac{\mu_{j'}^{(\ell)}}{\lambda} \geq \sum_{s=0}^{i} \prod_{j=K^{(\ell)}-s}^{K^{(\ell)}} \frac{\mu_j^{(\ell)}}{\lambda} \prod_{j'=K^{(\ell)}-i-1}^{K^{(\ell)}-s-1} \frac{\mu_{j'}^{(\ell)}}{\lambda} = (i + 1)a_{i+1}^{(\ell)}.$$

Having found $\sum_{\ell=1}^{L} \pi_{K^{(\ell)}}^{(\ell)}$, we immediately have $\pi_{K^{(\ell)}}^{(\ell)}$ from (21) and the remaining entries $\pi_k^{(\ell)}$ with $k < K^{(\ell)}$ follow from (20). □

Due to Corollary 1 and Theorem 6, we have the following:

Corollary 3. *If $\mu_K > \lambda$ and if $\mu_1^{(\ell)} \leq \cdots \leq \mu_{K^{(\ell)}}^{(\ell)}$ holds for all $1 \leq \ell \leq L$, then the sequence of stationary measures $X^N(\infty)$ of the finite state Markov chains $X^N(t)$ converges weakly to the dirac measure of the fixed point $\pi$.*

## 6 RESPONSE TIME DISTRIBUTION

In case we have a unique fixed point $\pi$, we can also compute the limiting response time distribution. Let $R(s)$ be the Laplace transform of the response time distribution of a job and $R_k^{(\ell)}(s)$ the Laplace transform of the response time distribution of a job that is assigned to a type $\ell$ server that contained $k - 1$ jobs on arrival. Clearly we have

$$R(s) = \sum_{\ell=1}^{L} \sum_{k=1}^{K^{(\ell)}} \frac{\pi_{k-1}^{(\ell)}}{1 - \sum_{\ell=1}^{L} \pi_{K^{(\ell)}}^{(\ell)}} R_k^{(\ell)}(s),$$

| $N'$ | $q(0.1)$ | $q(0.2)$ | $q(0.3)$ | $q(0.4)$ | $q(0.5)$ | $q(0.6)$ | $q(0.7)$ | $q(0.8)$ | $q(0.9)$ | $q(0.99)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.370 | 0.789 | 1.270 | 1.831 | 2.502 | 3.334 | 4.425 | 5.996 | 8.770 | 18.690 |
| 20 | 0.370 | 0.788 | 1.267 | 1.827 | 2.497 | 3.328 | 4.416 | 5.982 | 8.747 | 18.636 |
| 40 | 0.370 | 0.788 | 1.266 | 1.825 | 2.494 | 3.324 | 4.411 | 5.975 | 8.736 | 18.611 |
| 80 | 0.369 | 0.787 | 1.266 | 1.825 | 2.493 | 3.322 | 4.408 | 5.972 | 8.731 | 18.600 |
| 160 | 0.369 | 0.787 | 1.266 | 1.824 | 2.493 | 3.322 | 4.408 | 5.971 | 8.729 | 18.595 |
| mean field | 0.369 | 0.787 | 1.265 | 1.824 | 2.492 | 3.321 | 4.406 | 5.969 | 8.727 | 18.592 |

Table 3. Response time distribution quantiles

as all the incoming jobs are randomly assigned among the servers that have a queue length below the threshold. Denote $\lambda^* = \lambda/(1 - \sum_{\ell=1}^{L} \pi_{K^{(\ell)}}^{(\ell)})$ as the arrival rate of a server with a queue length below its threshold. Then, due to the exponential service time and the fact that the servers use processor sharing, we obtain the following set of equations:

$$R_1^{(\ell)}(s) = \frac{\mu_1^{(\ell)}}{s + \lambda^* + \mu_1^{(\ell)}} + \frac{\lambda^*}{s + \lambda^* + \mu_1^{(\ell)}} R_2^{(\ell)}(s),$$

$$R_k^{(\ell)}(s) = \frac{\mu_k^{(\ell)}}{s + \lambda^* + \mu_k^{(\ell)}} \left( \frac{1}{k} + \frac{k-1}{k} R_{k-1}^{(\ell)}(s) \right) + \frac{\lambda^*}{s + \lambda^* + \mu_k^{(\ell)}} R_{k+1}^{(\ell)}(s) \qquad (1 < k < K^{(\ell)}),$$

$$R_{K^{(\ell)}}^{(\ell)}(s) = \frac{\mu_{K^{(\ell)}}^{(\ell)}}{s + \mu_{K^{(\ell)}}^{(\ell)}} \left( \frac{1}{K^{(\ell)}} + \frac{K^{(\ell)} - 1}{K^{(\ell)}} R_{K^{(\ell)}-1}^{(\ell)}(s) \right).$$

Hence, we obtain $\ell$ linear systems of equations, where the $\ell$-th system is of size $K^{(\ell)}$. For a given $s$ we can numerically solve these linear systems of equations and therefore compute the response time distribution by numerically inverting the Laplace transform [1, 12].

Table 3 uses the same two class setup as in Table 1, and displays some quantiles of the response time distribution for $N' = 10, 20, 40, 80, 160$ as well as the mean field limit. The system load is $\rho = 0.7$.

The mean field mean response time is calculated from numerical Laplace inversion using concentrated matrix exponential distributions of order 100 which guarantee a precision of $10^{-4}$ [12], while the other elements are obtained using simulations with a deviation of less than $10^{-3}$. $q(0.1)$ denotes the 10% quantile of the response time distribution etc.

Note that the mean response time $E[R]$ can be obtained directly from the fixed point $\pi$ using Little's law, that is,

$$E[R] = \frac{1}{\lambda} \sum_{\ell=1}^{L} \sum_{k=1}^{K^{(\ell)}} k \pi_k^{(\ell)}. \tag{24}$$

## 7 OPTIMAL THRESHOLD SETTING

We give a method for configuring the thresholds in the system to minimize mean response time and related metrics. Specifically, we consider cost functions $\bar{c}$ that can be described by a single-server cost function $c$ in the following way:

$$\bar{c} = \sum_{\ell,k} \pi_k^{(\ell)} c_k^{(\ell)}.$$

Some examples of possible objectives are:

- Minimize mean number in system, and thereby mean respone time, with $c_k^{(\ell)} = k$.
- Minimize energy consumption by setting

$$c_k^{(\ell)} = r_k^{(\ell)} + \mu_k^{(\ell)} s_k^{(\ell)},$$

where $r_k^{(\ell)}$ and $s_k^{(\ell)}$ are the energy costs of running the server and switching between jobs, respectively.

The key idea is the following: within a group of servers of the same type there is no need to enforce all servers within that group to use the same threshold $K^{(\ell)}$. If we demand that all servers of the same type must use the same threshold, the optimization problem would be much harder. Note that our model allows us to analyze systems where we split some server types into a finite number of subtypes, as this is equivalent to a system with an increased number of server types. We will show that the optimal threshold settings will be such that *all servers belonging to the same type use the same threshold, except for (at most) one server type that must be split into two server subtypes.*

Using this key idea and assuming the service rate curves are increasing up to the point $K_*^{(\ell)} = \operatorname{argmax}_k \mu_k^{(\ell)}$, the system can have *any* limiting distribution $\pi \geq 0$ as long as it satisfies the constraints

$$\sum_{\ell=1}^{L} \sum_{k=1}^{K_*^{(\ell)}} \pi_k^{(\ell)} \mu_k^{(\ell)} = \lambda, \tag{25}$$

$$\sum_{k=1}^{K_*^{(\ell)}} \pi_k^{(\ell)} = \gamma^{(\ell)}, \tag{26}$$

for $\ell = 1, \ldots, L$. Indeed, all we need to do is to set the threshold and MPL of a fraction $\pi_k^{(\ell)}/\gamma^{(\ell)}$ of the type $\ell$ servers to $k$. In this manner, in the mean field regime, all the servers will have a constant queue length equal to their threshold/MPL (due to the increasing nature of the service rate curve until it reaches its maximum).

Maximizing $\bar{c}$ subject to the above constraints is a linear program. In the rest of this section we show that the linear program has a specific structure that makes it especially easy to solve, yielding a simple description of the optimal threshold configuration.

### 7.1 Homogeneous Servers

We start with the case of homogeneous servers, meaning every server has the same type, so we drop the superscripts from our notation.

Define the point $p_k = (\mu_k, c_k)$, for $k = 1, \ldots, K$, and $p_0 = (0, 0)$, and let $P = \{p_0, \ldots, p_K\}$. We can visualize the optimization problem by plotting the points in $P$ on a plane. The point

$$\bar{p} = \sum_k \pi_k p_k$$

describes the overall system behavior: its first coordinate is the average service rate $\sum_k \pi_k \mu_k$, which is constrained to be $\lambda$, and its second coordinate is the average cost $\bar{c} = \sum_k \pi_k c_k$ (i.e., queue length if we set $c_k = k$).

The point $\bar{p}$ is a convex combination of the points in $P$, so the polytope $\operatorname{conv}(P)$ is the achievable region of $(\lambda, \bar{c})$ pairs. Moreover, given the arrival rate $\lambda$, the optimal point $\bar{p}$ lies on the lower boundary of $\operatorname{conv}(P)$, meaning it is on a line segment between two vertices of the polytope (see Figure 2 for an illustration). This means that the optimal server configuration is given by the following procedure:

(1) Find the set $\Xi$ of values of $k$ such that $p_k$ is on the lower boundary of $\operatorname{conv}(P)$. This step needs

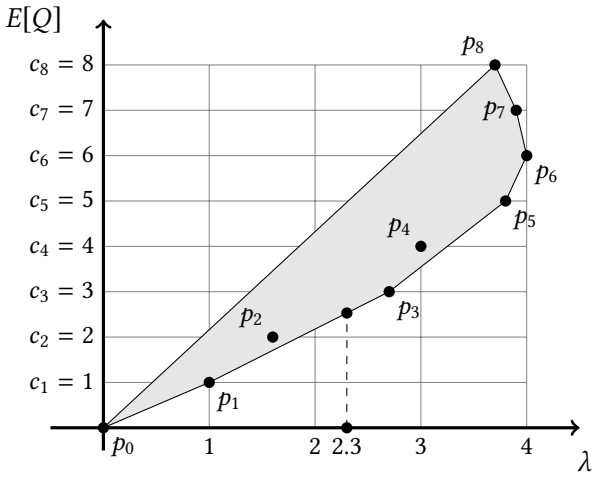



Fig. 2. Illustration of polytope $conv(P)$ for $\mu_1 = 1, \mu_2 = 1.6, \mu_3 = 2.7, \mu_4 = 3, \mu_5 = 3.8, \mu_6 = 4, \mu_7 = 3.9$ and $\mu_8 = 3.7$. The mean queue length is minimized for $\lambda = 2.3$ when a fraction $4/17$ of the servers set their threshold at 1 and the remaining servers set their threshold at 3.

to be done just once, as the result can be reused for every arrival rate $\lambda$.
(2) Given an arrival rate $\lambda$, find the values of $k_1, k_2 \in \Xi$ with service rates immediately below and above $\lambda$. That is, $k_1 = \arg\max_{k\in\Xi}\{\mu_k | \mu_k \le \lambda\}$ and $k_2 = \arg\min_{k\in\Xi}\{\mu_k | \mu_k > \lambda\}$. Set every server's threshold to either $k_1$ or $k_2$ such that the overall service rate is $\lambda$:

$$\pi_{k_1} = \frac{\mu_{k_2} - \lambda}{\mu_{k_2} - \mu_{k_1}},$$
$$\pi_{k_2} = \frac{\lambda - \mu_{k_1}}{\mu_{k_2} - \mu_{k_1}}.$$

Note that the lower boundary of the polytope $\mathrm{conv}(P)$ is the same as the lower bound of the polytope formed by the points $P = \{p_0, \ldots, p_{K_*}\}$ with $K_* = \arg\max_k \mu_k$. In Figure 2 the set $\Xi = \{0, 1, 3, 5, 6\}$ and for $\lambda = 2.3$ we find $\pi_1 = 4/17$ and $\pi_3 = 13/17$. An immediate corollary is the following:

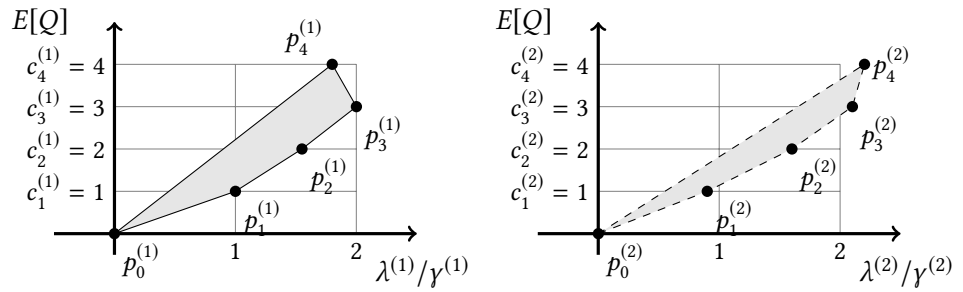
Corollary 4. *The minimization objective $\bar{c}$ under the constraints (25) and (26) as a function of $\lambda$ is convex, nondecreasing, and piecewise linear if we assume that $c_k$ is nondecreasing in $k$ and $\mu_k$ is increasing up to $K_* = \arg\max_k \mu_k$.*

### 7.2 Heterogeneous Servers

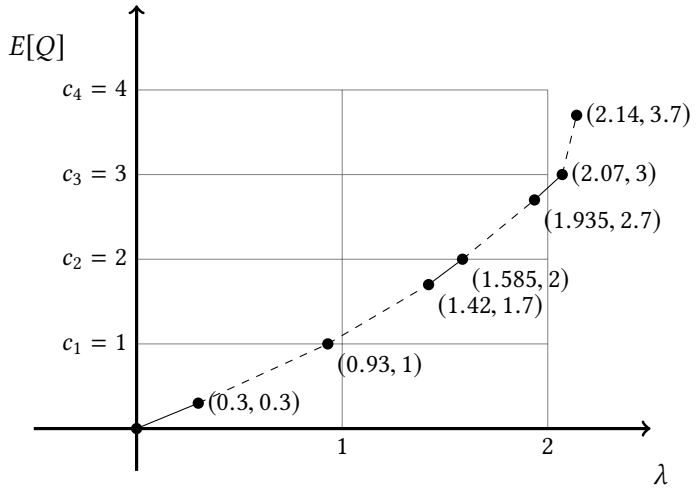
We now turn to the case of heterogeneous servers. Roughly speaking, we split the problem into homogeneous subproblems, solve each one as in the previous subsection, then combine the solutions.

Assume that the arrival rate to the type $\ell$ servers is $\lambda^{(\ell)}$ with $\sum_\ell \lambda^{(\ell)} = \lambda$. In order to find the optimal threshold settings for the type $\ell$ servers, given $\lambda^{(\ell)}$, we can simply study the optimal setting in a homogeneous system where all the servers are of type $\ell$ and the overall arrival rate equals $\lambda^{(\ell)}/\gamma^{(\ell)}$. Thus it suffices to determine the optimal arrival rates $\lambda^{(\ell)}$ at which jobs are dispatched to type $\ell$ servers and to use the results of the homogeneous setting.

Recall from the homogeneous case that the optimal cost is a convex, nondecreasing, piecewise linear function of the arrival rate. So if we increase the overall arrival rate $\lambda = \sum_\ell \lambda^{(\ell)}$ by a sufficiently small amount $\delta$, it is optimal to "allocate" that $\delta$ to whichever server type $\ell^*$ has the




(a) Optimal mean queue length for type 1 and type 2 servers when $\mu_1^{(1)} = 1, \mu_2^{(1)} = 1.55, \mu_3^{(1)} = 2, \mu_4^{(1)} = 1.8$ and $\mu_1^{(2)} = 0.9, \mu_2^{(2)} = 1.6, \mu_3^{(2)} = 2.1, \mu_4^{(2)} = 2.2$.




(b) Stitched optimal mean queue length when $\gamma^{(1)} = 0.3$ and $\gamma^{(2)} = 0.7$

Fig. 3. Illustration of achievable region polytopes for heterogeneous system with two server types

least marginal cost, that is, for which the slope in $\lambda^{(\ell)}/\gamma^{(\ell)}$ is the smallest (where ties can be broken arbitrarily).

The resulting optimal cost curve has a simple geometric interpretation: take the optimal cost curves for each type $\ell$ (see Figure 3a), sort all the line segments by slope, then arrange them end-to-end to create one big curve (see Figure 3b). This means that in general, all but one server type will have a single threshold, and that last server type $\ell^*$ is split between two thresholds such that $\lambda = \sum_\ell \lambda_\ell$. In Figure 3b the slopes coming from type 1 are solid, while those coming from type 2 are dashed. Thus, if we want to minimize $E[Q]$ for $\lambda \in (0.93, 1.42)$ we should set the threshold of the type 1 servers to one and split the type 2 servers in two groups where one group has threshold 1 and the other threshold value 2 (as $\lambda$ falls into the second dashed slope and there is one solid slope that precedes the second dashed one). If we modify the fractions $\gamma_\ell$ we can still use the same slopes in the exact same order, but only need to modify how long they last.

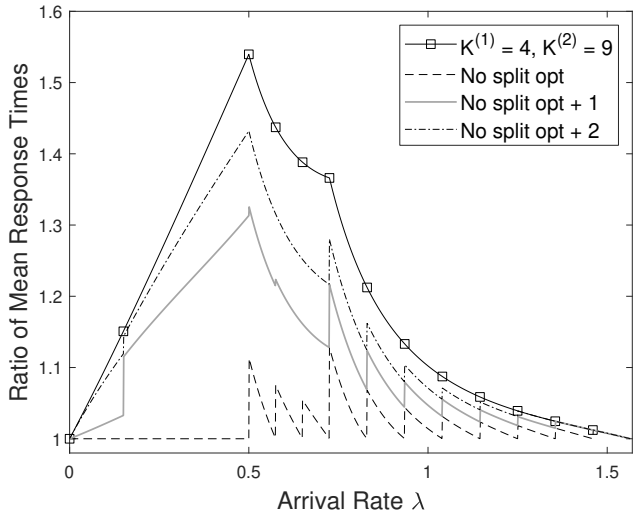



Fig. 4. Illustration of the sensitivity of the thresholds settings for the service rate curves shown in Figure 1 with $\gamma^{(1)} = 0.3$ and $\gamma^{(2)} = 0.7$.

### 7.3 Guaranteed Mean Response Time

In a real system the arrival rate $\lambda$ changes over time and as such one needs to reset the thresholds whenever such a change occurs in order to minimize the mean response time. Further the exact service rate curves can also be hard to estimate. In addition, the mean response time may be quite sensitive to the exact settings of the thresholds as illustrated in Figure 4. In this Figure we used the service rate curves of Figure 1 and depict the relative increase in the mean response time if the thresholds are not set in an optimal manner. The *No split opt* curve corresponds to the setting where we do not split one of the homogeneous groups of servers into two subgroups, which is needed to achieve a minimum mean response time. The *No split opt + a* curves are identical, but in addition we also increased all thresholds by $a$ (with an upper bound of $\arg\max_k \mu_k^{(\ell)}$ for a type $\ell$ server). The curve with the square markers corresponds to using fixed thresholds equal to $\arg\max_k \mu_k^{(\ell)}$. If the thresholds are not set in the optimal manner, but somewhat larger, some jobs could be served more efficiently by another server. Surprisingly, even a little bit of this type of inefficiency degrades performance significantly as the *No split opt + a* curves with $a = 1$ and $a = 2$ quickly approach the fixed threshold curve. Thus, if a system designer uses a conservative estimate of the service rate curves or a loose upper bound for the arrival rate $\lambda$, performance may be significantly worse than optimal.

As a system designer may be mostly interested in guaranteeing a certain performance, it may suffice to make sure that the mean response time is below some predefined target, instead of necessarily minimizing the mean response time. The result in the previous two subsections can also be leveraged for exactly this purpose.

Suppose we know that the arrival rate lies somewhere between 0 and some $\lambda_{max}$. Given a feasible target mean response time $E[R]^{(target)}$, we propose to set the thresholds in the following manner. Let $E[Q]^{(target)} = E[R]^{(target)}/\lambda_{max}$ and determine the arrival rate $\lambda^*$ for which the minimum mean queue length equals $E[Q]^{(target)}$. If $\lambda^* < \lambda_{max}$, the target mean response time $E[R]^{(target)}$ is not feasible. Otherwise, if $\lambda^* > \lambda_{max}$, we suggest to use the threshold setting that minimizes the mean response time when arrival rate equals $\lambda^*$. This should typically suffice to guarantee that

| $l$ | $\mu_0^{(\ell)}$ | $\mu_1^{(\ell)}$ | $\mu_2^{(\ell)}$ | $\mu_3^{(\ell)}$ | $\mu_4^{(\ell)}$ | $\mu_5^{(\ell)}$ | $\mu_6^{(\ell)}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.5 | 0.6 | 0.8 | 0.85 | 1 | 1.2 | 1.2 |
| 2 | 0.5 | 0.65 | 0.8 | 0.9 | 1.05 | 1.2 | 1.2 |

Table 4. Service rate curves used in simulation experiments with non-exponential job sizes.

| $N'$ | $SCV$ | $\ell$ | $\pi_0^{(\ell)}$ | $\pi_1^{(\ell)}$ | $\pi_2^{(\ell)}$ | $\pi_3^{(\ell)}$ | $\pi_4^{(\ell)}$ | $\pi_5^{(\ell)}$ | $\pi_6^{(\ell)}$ | $\pi_7^{(\ell)}$ | $\pi_8^{(\ell)}$ | $\pi_9^{(\ell)}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 1 | 1 | .0007 | .0024 | .0073 | .0177 | .0448 | .1148 | .3399 | .0783 | .0325 | .0147 |
| | | 2 | .0280 | .1388 | .0410 | .0271 | .0196 | .0148 | .0115 | .0092 | .0075 | .0060 |
| | 10 | 1 | .0007 | .0026 | .0078 | .0190 | .0460 | .1118 | .3034 | .0709 | .0338 | .0189 |
| | | 2 | .0269 | .0926 | .0223 | .0148 | .0112 | .0093 | .0081 | .0071 | .0065 | .0059 |
| 102 | 1 | 1 | .0002 | .0008 | .0034 | .0116 | .0402 | .1326 | .4680 | .0086 | .0011 | .0002 |
| | | 2 | .0368 | .2845 | .0092 | .0020 | .0005 | .0002 | .0001 | .0000 | .0000 | .0000 |
| | 10 | 1 | .0002 | .0008 | .0034 | .0116 | .0403 | .1321 | .4638 | .0112 | .0022 | .0006 |
| | | 2 | .0368 | .2588 | .0183 | .0075 | .0040 | .0024 | .0015 | .0010 | .0007 | .0005 |
| 1020 | 1 | 1 | .0001 | .0003 | .0020 | .0089 | .0385 | .1453 | .4715 | 0 | 0 | 0 |
| | | 2 | .0383 | .2950 | .0000 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 10 | 1 | .0001 | .0004 | .0021 | .0090 | .0386 | .1454 | .4711 | 0 | 0 | 0 |
| | | 2 | .0384 | .2950 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| mean field limit | | 1 | .0000 | .0003 | .0018 | .0085 | .0383 | .1470 | .4708 | 0 | 0 | 0 |
| | | 2 | .0384 | .2950 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 5. Simulation results that suggest asymptotic insensitivity of the queue length distributions with respect to the job size distribution for a system with a load $\rho = 27/29$.

the mean response time is below the target $E[R]^{(target)}$ for any $\lambda < \lambda_{max}$ as the mean response time typically increases with the arrival rate. However for exotic service rate curves the mean response time may not be monotone in the arrival rate. We do know that this threshold setting guarantees a mean queue length below $E[Q]^{(target)}$ for any $\lambda < \lambda^*$, as all queues have a length that is upper bounded by the threshold value for $\lambda < \lambda^*$ and the mean threshold value is by design exactly $E[Q]^{(target)}$.

We could also set the thresholds such that the mean response time is minimized when the arrival rate equals $\lambda_{max}$. This however would imply that we use smaller threshold values and therefore we expect that the queue lengths at the servers are typically closer to the threshold, which yields a higher communication overhead. Further, if we use the thresholds associated with $\lambda_{max}$ the system load can get arbitrarily close to 1, which means that a very large finite system may be needed for the mean field model to be highly accurate.

## 8 ASYMPTOTIC INSENSITIVITY

The mean field model presented in this paper assumed an exponential job size distribution, while in real systems jobs tend to be more variable. We believe however that the mean field limit obtained and studied in this paper is also the proper limit in case of non-exponential job sizes. Thus, we believe that the queue length distribution is asymptotically insensitive to the job size distribution. Note this type of asymptotic insensitivity is known to hold for JIQ [17] and is believed to hold for JSQ-d with processor sharing servers [2, 19].

The main reason why we believe that asymptotic insensitivity holds is that in the limit we expect that the fraction of jobs that is assigned at random tends to zero (if the load is below one) and

therefore each server behaves as an M/G/1 processor sharing server with queue length dependent arrival and service rates. The queue length distribution of such queues is known to be insensitive to the job size distribution [3]. In our case the arrival rate is fixed for any queue length below the threshold and zero otherwise.

To support this belief we also performed simulation experiments where we compare the queue length distribution of a system with exponential job sizes to that with hyperexponential job sizes with balanced means and a squared coefficient of variation (SCV) equal to 10. We used an arbitrarily selected system with 2 types of servers with $\gamma_1 = 2/3$, $K^{(1)} = 6$, $K^{(2)} = 1$ and $\lambda = 0.9$. The service rate curves are shown in Table 4. Table 5 strengthens our believe as the queue length distributions for both job size distributions become nearly indistinguishable from each other as the number of servers $N'$ becomes large. For $N'$ small we see that the tail of the distribution for the more variable job sizes decays more slowly as expected. Note that the load in this experiment is quite high, being 27/29. Additional experiments (not reported in the paper) suggest that the job size distribution matters less for lower loads.

Note that by Little's law insensitivity with respect to the job size distribution also implies that the mean response time is insensitive. The response time distribution clearly depends on the complete job size distribution. As we expect that all arriving jobs in the mean field regime can immediately commence service and a minimal service rate is guaranteed, the tail behavior of the response time distribution should be the same as that of the job size distribution.

## 9 CONCLUSIONS

In this paper we studied the Join-Below-Threshold (JBT) load balancing policy in a heterogeneous system of resource sharing servers in the mean field regime under exponential job sizes. JBT is a natural generalization of the Join-Idle-Queue policy for resource sharing servers. We proved convergence of the stationary measures to the fixed point of the mean field model, provided that it is unique. We derive simple sufficient conditions for the existence of a unique fixed point and illustrated that for arbitrary service rate curves metastability can occur. We showed how to optimize the mean response time and presented some simulation experiments that suggest that the queue length distribution is asymptotically insensitive to the job size distribution.

A possible extension to this work is developing a mean field model for more general service times and proving asymptotic insensitivity. In addition in a heterogeneous cluster one can also consider variations of JBT that take the server speeds into account when assigning jobs (instead of blindly picking a server with a queue length below its threshold).

## REFERENCES

[1] J. Abate and W. Whitt. Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on computing*, 7(1):36–43, 1995.

[2] M. Bramson, Y. Lu, and B. Prabhakar. Randomized load balancing with general service time distributions. In *ACM SIGMETRICS 2010*, pages 275–286, 2010.

[3] S. L. Brumelle. A generalization of Erlang's loss system to state dependent arrival and service rates. *Mathematics of Operations Research*, 3(1):10–16, 1978.

[4] D. Gamarnik, J. N. Tsitsiklis, and M. Zubeldia. Delay, memory, and messaging tradeoffs in distributed service systems. *SIGMETRICS Perform. Eval. Rev.*, 44(1):1–12, June 2016.

[5] A. Ganesh, S. Lilienthal, D. Manjunath, A. Proutiere, and F. Simatos. Load balancing via random local search in closed and open systems. *SIGMETRICS Perform. Eval. Rev.*, 38(1):287–298, June 2010.

[6] N. Gast and B. Gaujal. A mean field model of work stealing in large-scale systems. *SIGMETRICS Perform. Eval. Rev.*, 38(1):13–24, June 2010.

[7] N. Gast and B. Gaujal. Markov chains with discontinuous drifts have differential inclusion limits. *Performance Evaluation*, 69(12):623 – 642, 2012.

[8] N. Gast and B. Van Houdt. A refined mean field approximation. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(2):33:1–33:28, December 2017.
[9] I. Grosof, Z. Scully, and M. Harchol-Balter. Load balancing guardrails: Keeping your heavy traffic on the road to low response times. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(2):42:1–42:31, June 2019.
[10] V. Gupta and M. Harchol-Balter. Self-adaptive admission control policies for resource-sharing systems. *SIGMETRICS Perform. Eval. Rev.*, 37(1):311–322, June 2009.
[11] M.W. Hirsch and H. Smith. Monotone dynamical systems. In *Handbook of differential equations: ordinary differential equations*, volume 2, pages 239–357. Elsevier, 2006.
[12] G. Horváth, I. Horváth, S. Al-Deen Almousa, and M. Telek. High order low variance matrix-exponential distributions. *The Tenth International Conference on Matrix-Analytic Methods in Stochastic Models (MAM10)*, 02 2019.
[13] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg. Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Perform. Eval.*, 68:1056–1071, 2011.
[14] M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distrib. Syst.*, 12:1094–1104, October 2001.
[15] M. Mitzenmacher. Analyzing distributed join-idle-queue: A fluid limit approach. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 312–318, Sept 2016.
[16] G. Roth and W.H. Sandholm. Stochastic approximations with constant step size and differential inclusions. *SIAM Journal on Control and Optimization*, 51(1):525–555, 2013.
[17] A.L. Stolyar. Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems*, 80(4):341–361, 2015.
[18] M. Telek and B. Van Houdt. Response time distribution of a class of limited processor sharing queues. *SIGMETRICS Perform. Eval. Rev.*, 45(3):143–155, March 2018.
[19] T. Vasantam, A. Mukhopadhyay, and R. R. Mazumdar. The mean-field behavior of processor sharing systems with general job lengths under the sq(d) policy. *Performance Evaluation*, 127-128:120 – 153, 2018.
[20] N.D. Vvedenskaya, R.L. Dobrushin, and F.I. Karpelevich. Queueing system with selection of the shortest of two queues: an asymptotic approach. *Problemy Peredachi Informatsii*, 32:15–27, 1996.
[21] X. Zhou, J. Tan, and N. Shroff. Heavy-traffic delay optimality in pull-based load balancing systems: Necessary and sufficient conditions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(3):41, 2018.

Received August 2019; revised September 2019; accepted October 2019