# Power of $d$ Choices for Large-Scale Bin Packing: A Loss Model

Qiaomin Xie, Xiaobo Dong, Yi Lu and R. Srikant
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{qxie3,dong6,yilu4,rsrikant}@illinois.edu

## ABSTRACT

We consider a system of $N$ parallel servers, where each server consists of $B$ units of a resource. Jobs arrive at this system according to a Poisson process, and each job stays in the system for an exponentially distributed amount of time. Each job may request different units of the resource from the system. The goal is to understand how to route arriving jobs to the servers to minimize the probability that an arriving job does not find the required amount of resource at the server, i.e., the goal is to minimize blocking probability. The motivation for this problem arises from the design of cloud computing systems in which the jobs are virtual machines (VMs) that request resources such as memory from a large pool of servers. In this paper, we consider power-of-$d$-choices routing, where a job is routed to the server with the largest amount of available resource among $d \geq 2$ randomly chosen servers. We consider a fluid model that corresponds to the limit as $N$ goes to infinity and provide an explicit upper bound for the equilibrium blocking probability. We show that the upper bound exhibits different behavior as $B$ goes to infinity depending on the relationship between the total traffic intensity $\lambda$ and $B$. In particular, if $(B-\lambda)/\sqrt{\lambda} \to \alpha$, the upper bound is doubly exponential in $\sqrt{\lambda}$ and if $(B-\lambda)/\log_d \lambda \to \beta$, $\beta > 1$, the upper bound is exponential in $\lambda$. Simulation results show that the blocking probability, even for small $B$, exhibits qualitatively different behavior in the two traffic regimes. This is in contrast with the result for random routing, where the blocking probability scales as $O(1/\sqrt{\lambda})$ even if $(B-\lambda)/\sqrt{\lambda} \to \alpha$.

## Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Stochastic processes, Queueing theory

## General Terms

Algorithms, Theory, Performance


*SIGMETRICS'15,* June 15–19, 2015, Portland, OR, USA.

http://dx.doi.org/10.1145/2745844.2745849.

## Keywords

Randomized Algorithms; Loss Model; Resource Allocation; Virtual Machine Assignment; Fluid Limit Analysis

## 1. INTRODUCTION

We consider a system of $N$ parallel homogeneous servers. Each server has $B$ units of a resource. There are $J$ different types of jobs. Jobs of type $j$ arrive at the system according to a rate-$N\lambda_j$ Poisson process, and each type-$j$ job requires $b_j$ units of resource and stays in the system for an exponentially distributed amount of time with mean 1. Let $\lambda = \sum_{j=1}^{J} \lambda_j b_j$ denote the total traffic intensity.

Each arriving job is routed to a server according to a routing policy and requires zero-delay service. If the selected server has sufficient resource to accommodate the arriving job, the job will be processed immediately. Otherwise the job is blocked, i.e., it leaves the system immediately without being served. The goal is to understand how to route arriving jobs to the servers to minimize the probability that an arriving job does not find the required amount of resource at the server, i.e., to minimize blocking probability.

For each server $m$, let $n_{j,m}(t)$ denote the number of type-$j$ jobs that the server is serving at time $t$. We use

$$\mathbf{n}_m(t) = (n_{1,m}(t), n_{2,m}(t), ..., n_{J,m}(t))$$

to denote the state of server $m$. Note that $\mathbf{n}_m$ is feasible only if server $m$ has enough resource to accommodate all these jobs. That is,

$$\sum_{j=1}^{J} n_{j,m} b_j \leq B.$$

The model is motivated by the design of cloud computing systems in which the jobs are virtual machines (VMs) that request resources from a large pool of servers. Some examples of cloud computing systems are Amazon EC2 system [1], Google's AppEngine [3] and Microsoft's Azure [5]. Users submit requests for resource in the form of virtual machines (VMs). Each request specifies the amount of physical resources it needs in terms of processor power, memory, I/O bandwidth, disk, etc.

The resource allocation problem for VMs is a stochastic bin-packing problem [6, 9], but with VMs terminating after an application has completed. This motivates our model with jobs arriving and departing the system, which was first considered in [15] and is referred to as a *service* model in [26]. When a user submits VM requests in a cloud computing system, any request that is not immediately fulfilled is typically

rejected [5]. This motivates us to consider a loss model and its blocking probability, in contrast to the models in [15, 26].

In our model, we consider one-dimensional packing constraint for the requests of resources. While VM requests can be modelled as multi-dimensional bin-packing, it has been observed that memory is the dominating bottleneck [11]. Due to the large size of a cloud computing system, we consider asymptotic blocking probability as $N \to \infty$.

We consider the power-of-$d$-choices routing algorithm for this system. An arriving job is routed to the server with the largest amount of available resource among $d \geq 2$ randomly chosen servers. When none of the chosen servers has enough resource to accommodate the job, it is rejected. In this paper, we focus on the fluid limit as $N \to \infty$ and the asymptotic blocking probability. To the best of our knowledge, this is the first work of studying power-of-$d$-choices routing algorithm in a loss model with packing constraints.

## 1.1 Related Work

**VMs packing problem.** Some recent work model the VMs allocation as stochastic bin packing with item departures [11, 17], and focus on improving resource utilization with different packing algorithms. Some other recent work study this problem with different performance objectives, including maximizing system throughput [15], minimizing heavy-traffic queue lengths [16], and minimizing the total energy consumption [29]. In this paper, we are interested in zero-delay service, i.e., a VM is served immediately upon arrival. The recent works in [24, 26, 25] also study zero-delay service. However, their performance objective is to minimize the number of servers occupied, which is different from ours. In particular, they consider the case of infinite number of servers, while we consider finite number of servers and study the blocking probability in the limit as the number of servers goes to infinity.

**The Power-of-$d$-choices algorithm.** Azar et. al. [4] were the first to analyze randomized load balancing schemes using balls-and-bins model. Another line of work focus on the queueing systems, [18, 19, 28, 8, 10, 14, 20, 31]. In particular, a supermarket model has been used widely to analyze the randomized load balancing schemes. Vvedenskaya et.al. [28] and Mitzenmacher [18] showed that when each arriving job is assigned to the shortest $d \geq 2$ randomly chosen queues, the equilibrium queue sizes decay doubly exponentially in the limit as the number of servers goes to infinity. This is a substantial improvement over the $d = 1$ case, where the queue size decays exponentially. While the work in [27] does not address power-of-d choices routing directly, similar analytical techniques have been used there to study the impact of resource pooling in large server farms. However, to the best of our knowledge, the performance of the power-of-$d$-choices algorithm ($d \geq 2$) for a loss model has not been studied previously. Related work has also been done in parallel with our work in [21]

## 1.2 Organization of the Paper

The rest of the paper is organized as follows. Section 1.3 introduces the notation used in this paper. Section 2 states the precise model and main results. The proofs of these main results will be deferred to later sections. We first study the loss model under the power-of-$d$-choices algorithm ($d \geq 2$) for the case when jobs are homogeneous, i.e., all jobs are of the same type, in Section 3-5. Section 3 justifies the use of fluid approximation of sufficiently large finite systems. Section 4 develops an upper bound for the stationary point of the fluid model. Section 5 analyzes the blocking probability in two different limiting regimes. We then extend our analysis to the case with heterogeneous jobs based on an independence ansatz in Section 6.

## 1.3 Notation

We will use bold letters to denote vectors in $\mathbb{R}^B$ or $\mathbb{N}^J$ or $\mathbb{N}^{J\times N}$, and ordinary letters for scalars. Dot product in the vector spaces $\mathbb{R}^J$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$.

Let $\mathbb{N}_+$ be the set of non-negative integers. The following notations will be used throughout the paper:

$$\begin{aligned}
\mathcal{C} &\triangleq \left\{ \mathbf{n} \in \mathbb{N}_+^J : \sum_{j=1}^{J} n_j b_j \leq B \right\}, \\
\mathcal{Q}^{(N)} &\triangleq \{ \mathbf{Q} = \{\mathbf{n}_1, \mathbf{n}_2, ..., \mathbf{n}_N\} : \mathbf{n}_m \in \mathcal{C}, \forall m = 1, 2, ..., N \}, \\
\mathcal{S} &\triangleq \left\{ \mathbf{s} \in [0,1]^{B+1} : 1 = s_0 \geq s_1 \geq ... \geq s_B \geq 0 \right\}, \\
\mathcal{S}^{(N)} &\triangleq \left\{ \mathbf{s} \in \mathcal{S} : s_i = \frac{K_i}{N}, \text{ for some } K_i \in \mathbb{N}_+, \forall i \right\}, \\
\mathcal{P} &\triangleq \left\{ \mathbf{p} \in \mathbb{R}^{|\mathcal{C}|} : \sum_{i=1}^{|\mathcal{C}|} p_i = 1,\ p_i \geq 0,\ \forall i \right\}.
\end{aligned}$$

And we will use the following notation for asymptotic comparisons; here $f$ and $g$ are positive functions:

1. $f(x) \lesssim g(x)$ for $f(x) = \mathcal{O}(g(x))$, and $f(x) \gtrsim g(x)$ for $f(x) = \Omega(g(x))$.
2. $f(x) \sim g(x)$ for $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

# 2. PROBLEM STATEMENT AND MAIN RESULTS

We briefly recap the model that was stated in the introduction. We consider a system with $N$ servers, each of which has $B$ units of a resource, such as CPU, memory, etc. This system is accessed by $J$ different types of jobs, where each type of job is characterized by the number of units of resource that it demands. Jobs of type $j$ arrive according to a Poisson process of rate $N\lambda_j$, each type-$j$ job requests $b_j$ units of the resource, and each job stays in the system for an exponentially distributed amount of time with mean 1. We use $\mathbf{b} = (b_1, b_2, ..., b_J)$ to denote the vector of resource units required by different job types. The arrival processes of the different job types and the job holding times are all independent of each other. We consider two cases separately, $J = 1$ which we call the *homogeneous job* case and $J > 1$ which we call the *heterogeneous job* case. In the homogeneous case, we assume without loss of generality that $b_1 = 1$, i.e., all jobs require one unit of resource.

Our goal is to study the blocking probability of the power-of-$d$-choices routing: under this routing scheme, upon each job arrival, $d$ servers are selected uniformly at random and the job is routed to the least loaded of the servers (the one with the least amount of resource used). If none of the selected servers has sufficient amount of resource, then the arriving job is blocked and lost. The performance in the case $d = 1$ is fundamentally different from the cases where $d > 1$. Therefore, we study these two cases separately. In the

case of $d = 1$, since we are routing an arrival to a randomly selected server, we will call this scheme *the random routing* scheme. We will reserve the use of the term *power-of-d-choices* routing to the case where $d > 1$.

Next, we present the main results of the paper, for the homogeneous job case first followed by the heterogeneous job case.

## 2.1 Homogeneous Jobs

Before we present our main results, we introduce some notation. Consider a system with $N$ servers. Let $S_k^{(N)}(t)$ denote the fraction of servers with at least $k$ jobs in service at time $t$. We use $\boldsymbol{\pi}^{(N)} = (\pi_0^{(N)}, \pi_1^{(N)}, ..., \pi_B^{(N)})$ to denote the equilibrium distribution of $\mathbf{S}^{(N)}(t) = \{S_i^{(N)}(t)\}_{i=0}^B$. We will approximate $\boldsymbol{\pi}^{(N)}$ by the stationary state of the following fluid model in a manner which will be made precise later.

DEFINITION 1. *(Fluid Model). Given any initial condition $\mathbf{s}^0 \in \mathcal{S}$, a function $\mathbf{s}(t) : [0, \infty) \to \mathcal{S}$ is said to be a solution to the fluid model if:*

1. $\mathbf{s}(0) = \mathbf{s}^0$;
2. $s_0(t) = 1$ *for any* $t \geq 0$;
3. $\mathbf{s}(t)$ *satisfies the following differential equations for any* $t \geq 0$:

$$\frac{ds_k(t)}{dt} = \begin{cases} \lambda(s_{k-1}^d - s_k^d) - k(s_k - s_{k+1}), & 1 \leq k \leq B-1 \\ \lambda(s_{B-1}^d - s_B^d) - Bs_B, & k = B \end{cases} \quad (1)$$

Equation (1) can be written as

$$\dot{\mathbf{s}}(t) = \mathbf{F}(\mathbf{s}),$$

where

$$F_k(\mathbf{s}) = \begin{cases} \lambda(s_{k-1}^d - s_k^d) - k(s_k - s_{k+1}), & 1 \leq k \leq B-1 \\ \lambda(s_{B-1}^d - s_B^d) - Bs_B, & k = B \end{cases}$$

The $k$-th function $F_k(\mathbf{s})$ is the drift of $s_k$ at point $\mathbf{s}(t)$. The stationary point of the differential equation (1), denoted by $\boldsymbol{\pi}$, satisfies

$$\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}. \quad (2)$$

The following theorem presents the main convergence result (in the limit $N \to \infty$) for the homogeneous job case.

THEOREM 1. *The fluid model has a unique stationary distribution $\boldsymbol{\pi}$ and the sequence of stationary distribution $\boldsymbol{\pi}^{(N)}$ converges weakly to $\delta_{\boldsymbol{\pi}}$, which is the Dirac measure concentrated on $\boldsymbol{\pi}$. That is,*

$$\lim_{N\to\infty} \boldsymbol{\pi}^{(N)} = \delta_{\boldsymbol{\pi}}, \text{ in distribution.}$$

Due to the convergence result above and due to the Poisson nature of the arrival process, $\pi_B^d$ is a good approximation to the blocking probability experienced by arriving jobs, denoted by $P_b^{(N)}$, in a system with $N$ servers. This is due to the fact that, in the limit as $N \to \infty$, the servers become independent. However, that is not directly established in the convergence theorem above. But it can be argued as follows: using Little's law, we have

$$N\lambda(1 - P_b^{(N)}) = N\sum_{k=1}^B \pi_k^{(N)}.$$

Summing $F_k(\boldsymbol{\pi}) = 0$ over $1 \leq k \leq B$ yields

$$\lambda(1 - \pi_B^d) = \sum_{k=1}^B \pi_k.$$

Let $P_b = \pi_B^d$. From Theorem 1, we can approximate $P_b^{(N)}$ by $P_b$ when $N$ is sufficiently large.

While $\boldsymbol{\pi}$ can be computed recursively from Eq. (2), we provide a closed-form expression which provides an upper bound on $\pi_B$ for all values of $\lambda$ and $B$ for the case $d \geq 2$. This upper bound is useful later to understand the striking performance difference between the cases $d = 1$ and $d > 1$.

THEOREM 2. *(Upper bound) Let $\boldsymbol{\pi}$ denote the stationary point of the fluid model. Define $\{\bar{\pi}_k\}_{k=0}^B$ as follows:*

$$\bar{\pi}_k = \begin{cases} 1, & 0 \leq k \leq i_0 + 1 \\ \frac{\lambda^{\frac{d^{k-i_0-1}-1}{d-1}}}{(k-1)(k-2)^{d^1}\cdots(i_0+1)^{d^{k-i_0-2}}}, & i_0 + 1 < k \leq B \end{cases} \quad (3)$$

*where $i_0 = \lfloor \lambda \rfloor$.*

*Then $\bar{\boldsymbol{\pi}}$ is an upper bound for $\boldsymbol{\pi}$, i.e., for any $0 \leq k \leq B$,*

$$\bar{\pi}_k \geq \pi_k.$$

Note that in the case $d = 1$, since we are randomly selecting a server, by the property of Poisson processes, the blocking probability is given by the well-known Erlang-B formula for $M/M/B/B$ systems:

$$\mathcal{B}(B, \lambda) = \frac{\lambda^B/B!}{\sum_{k=0}^B (\lambda^k/k!)}. \quad (4)$$

Comparing equations (3) and (4), we can see that the blocking probability goes to zero faster in the case of $d \geq 2$, compared to that for $d = 1$.

To further provide insight into the blocking probability $P_b$ in the case of $d \geq 2$, we consider two limiting regimes: (i) $\frac{B-\lambda}{\sqrt{\lambda}} \to \alpha$ as $B \to \infty$ and (ii) $\frac{B-\lambda}{\log_d \lambda} \to \beta$ as $B \to \infty$. We call the former the heavy-traffic regime and the latter the critically-loaded regime. The heavy-traffic has been studied extensively in the context of $M/M/B/B$ and $G/G/B/B$ systems [7, 30, 23].

THEOREM 3. *Let $\lambda < B$ and $\frac{\lambda}{B} \to 1$ as $B \to \infty$, then*

$$\pi_B \lesssim (e^{-\frac{c^2}{2}})^{\frac{(B-\lambda)^2}{\lambda} d^{(1-c)(B-\lambda)-1}}, \quad (5)$$

*where $c$ is an arbitrary constant satisfying $0 < c < 1$.*

*In particular,*

1. *If $\frac{B-\lambda}{\sqrt{\lambda}} \to \alpha$ as $B \to \infty$, where $\alpha > 0$, then*

$$\log_d \log \frac{1}{P_b} \gtrsim ((1-c)\alpha + o(1))\sqrt{\lambda}.$$

*That is, the blocking probability decays doubly exponentially in $\sqrt{\lambda}$.*

2. *If* $\frac{B-\lambda}{\log_d \lambda} \to \beta$ *as* $B \to \infty$, *where* $\beta > 1$, *then there exists a constant* $\gamma = (1-c)\beta - 1 > 0$ *such that*

$$\log \frac{1}{P_b} \gtrsim \lambda^{\gamma + o(1)}.$$

*That is, the blocking probability decays exponentially in* $\lambda^\gamma$.

**Remark.** Theorem 3 shows that the fluid limit of the equilibrium blocking probability is dominated by an asymptotic upper bound, which exhibits very different behavior depending on the relationship between $\lambda$ and $B$ as $B$ goes to infinity. In particular, if $\frac{B-\lambda}{\sqrt{\lambda}} \to \alpha$, the upper bound is doubly exponential in $\sqrt{\lambda}$ and if $\frac{B-\lambda}{\log \lambda} \to \beta$, $\beta > 1$, the upper bound is exponential in $\lambda^\eta$. This is in contrast with the result for random routing, where the blocking probability scales as $O(\frac{1}{\sqrt{\lambda}})$ even if $\frac{B-\lambda}{\sqrt{\lambda}} \to \alpha$.

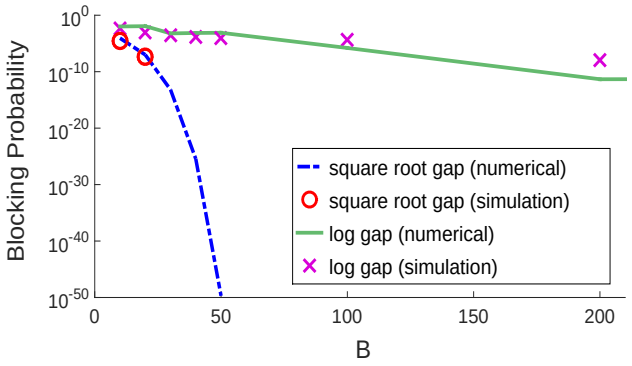


Figure 1: Blocking probability for the power-of-two-choices algorithm with different load gap. Line curves are obtained by solving Eq. (2) numerically. Markers are from simulations with $N = 1000$.

**Numerical Results:** Figure 1 shows the blocking probability for the power-of-two-choices algorithm with $B - \lambda = \sqrt{\lambda}$ and $B - \lambda = 2\log\lambda$, both by solving Eq. (2) numerically and by simulating a finite system with $N = 1000$. Note that the y-axis is in log scale. We can see that even for small $B$, the blocking probability $P_b$ exhibits qualitatively different behavior in these two regions: with $\log\lambda$ load gap, $P_b$ decays exponentially; while for $\sqrt{\lambda}$ load gap, $P_b$ decays much faster. For $B = 30$, $P_b$ is of order $10^{-15}$ with $\sqrt{\lambda}$ load gap. It requires very long simulation time in order to observe blocking event. We simulated around $10^{10}$ arrivals and no job blocking was observed for $B \geq 30$.

To extend the results in this section to the heterogeneous job case, we present a well-known alternative viewpoint of the derivation of $\boldsymbol{\pi}$. Suppose we assume that, in steady-state, the servers become independent of each other and due to symmetry, the number of jobs in each server is given by $\boldsymbol{\pi}$. In this case, let us focus on a particular server, say server 1, and write down the Markov chain corresponding to the number of jobs in the server. To describe the transition rate of this Markov chain, suppose that the server has $k$ jobs currently in service. Then, the arrival rate of jobs to this server (call it $q_k$) is $N\lambda$ times the probability that an arriving job selects this server. It is easy that $q_k$ is given by

$$q_k = N\lambda \cdot \frac{d}{N} \left( \sum_{i=1}^{d} \frac{1}{i} \binom{d-1}{i-1} (\pi_k^{(N)} - \pi_{k+1}^{(N)})^{i-1} (\pi_{k+1}^{(N)})^{d-i} \right),$$

which in the limit as $N \to \infty$ becomes

$$q_k = \lambda \left( \frac{\pi_k^d - \pi_{k+1}^d}{\pi_k - \pi_{k+1}} \right).$$

Thus, the Markov chain can be represented by the transition diagram in Figure 2. It is now easy to see that the steady-state distribution of this Markov chain is given by Eq. (2). This independence ansantz will be used in the next section to derive blocking probability results for the heterogeneous job case.

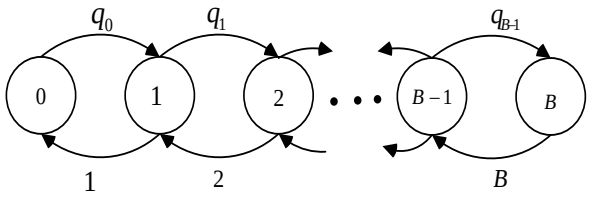


Figure 2: State-transition-rate diagram for server 1 with $B$ units of resource and homogeneous job arrivals.

## 2.2 Heterogeneous Jobs

We use the independence ansatz in the previous subsection as follows. Consider a particular server, say server 1, and let $\mathbf{n} = (n_1, ..., n_J)$ be the number of jobs of different types in this server. Let $\{p_{\mathbf{n}}\}_{\mathbf{n}\in\mathcal{C}}$ denote the asymptotic equilibrium distribution for server 1. Then $p_{\mathbf{n}}$ is also the asymptotic fraction of servers in state $\mathbf{n}$.

Under the asymptotic independence assumption, the arrival process of type $j$ jobs to server 1 is a state-dependent Poisson process with rate $\lambda_j(\mathbf{n})$, which is given by

$$\lambda_j(\mathbf{n}) = \lambda_j \left( \sum_{i=1}^{d} \binom{d}{i} E_{\mathbf{n}}^{i-1} G_{\mathbf{n}}^{d-i} \right), \tag{6}$$

where

$$E_{\mathbf{n}} = \sum_{\substack{\hat{\mathbf{n}}\in\mathcal{C} \\ \langle\hat{\mathbf{n}},\mathbf{b}\rangle = \langle\mathbf{n},\mathbf{b}\rangle}} p_{\hat{\mathbf{n}}} \quad , \quad G_{\mathbf{n}} = \sum_{\substack{\mathbf{n}'\in\mathcal{C} \\ \langle\mathbf{n}',\mathbf{b}\rangle > \langle\mathbf{n},\mathbf{b}\rangle}} p_{\mathbf{n}'}.$$

Let $B_j = \lfloor \frac{B}{b_j} \rfloor$ denote the maximum number of type-$j$ jobs that a server can serve simultaneously. In the case of two job types, the Markov chain is shown in Figure 3. However, it is difficult to analyze the equilibrium distribution of this Markov and obtain a simple expression for the blocking probability. Therefore, we study a one-dimensional recursion as in [12] and [22].

THEOREM 4. *The tail distribution* $\mathbf{r}$ *of the number of occupied resource units satisfies the following equation for any* $k = 0, 1, ...., B$:

$$\sum_{j=1}^{J} \lambda_j b_j (r_{k-b_j}^d - r_{k-b_j+1}^d) \quad = \quad k(r_k - r_{k+1}), \tag{7}$$

*where* $r_x = 1$ *for any* $x \leq 0$ *and* $r_{B+1} = 0$.

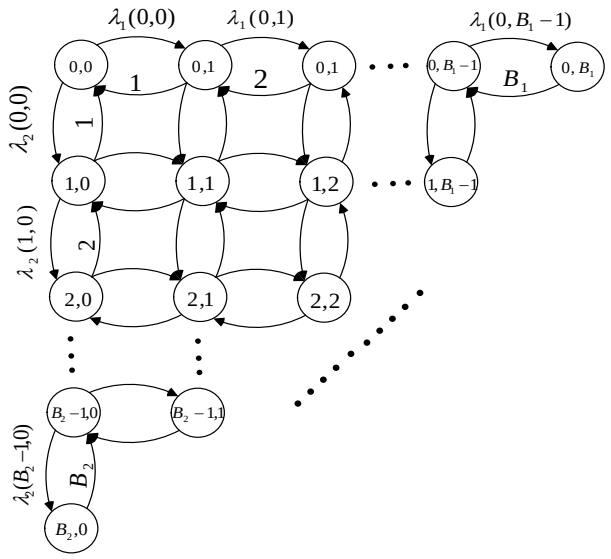


Figure 3: State-transition-rate diagram for server 1 with $B$ units of resource and two types of jobs arrivals.

As for the blocking probability, we obtain analogous results as for homogeneous jobs. Let $b = \max_{j=1,...,J} b_j$, and denote the blocking probability for jobs of type $j$ by $P_{b_j}$. We have the following theorem.

THEOREM 5. *Let $\lambda < B$ and $\frac{\lambda}{B} \to 1$ as $B \to \infty$,*

$$r_{B-b+1} \lesssim (e^{-\frac{c^2}{2}})^{\frac{(B-\lambda)^2}{b\lambda}} d^{(1-c)(\frac{B-\lambda}{b})-1}, \tag{8}$$

*where $c$ is an arbitrary constant satisfying $0 < c < 1$.*
*In particular,*

1. *If $\frac{B-\lambda}{\sqrt{\lambda}} \to \alpha$ as $B \to \infty$, where $\alpha > 0$, then*

$$\log_d \log \frac{1}{P_{b_j}} \gtrsim ((1-c)\frac{\alpha}{b} + o(1))\sqrt{\lambda},$$

*$\forall j \in \{1, 2, ..., J\}$. That is, for any type of jobs, the blocking probability decays doubly exponentially in $\sqrt{\lambda}$.*

2. *If $\frac{B-\lambda}{\log_d \lambda} \to \beta$ as $B \to \infty$, where $\beta > b$, then there exists a constant $\eta = (1-c)\frac{\beta}{b} - 1 > 0$ such that*

$$\log \frac{1}{P_{b_j}} \gtrsim \lambda^{\eta + o(1)},$$

*$\forall j \in \{1, 2, ..., J\}$. That is, for any type of jobs, the blocking probability decays exponentially in $\lambda^\eta$.*

The results in this section are derived under the independence ansatz. The existing technique to establish asymptotic independence depends on monotonicity, which does not hold for our problem. Although we do not have the tools to prove the ansatz without monotonicity, we believe that it is true in terms of the random nature of power-of-$d$-choices algorithm. Alternatively, one can use the fluid approximation: first show convergence of the stochastic system to a differential equation, then show that the differential equation has a unique stationary point to which it converges starting from any initial condition, and finally prove certain tightness results. We have done all of this for the homogeneous case in the next section. In the heterogeneous case, we only have partial results: we can prove convergence to a differential equation and also show that the equation (7) is one of the stationary points of the differential equation. The rest of the steps need to be verified.

# 3. CONVERGENCE RESULTS FOR THE HOMOGENEOUS CASE

In this section, we focus on the convergence results that justify the approximation of the sample paths $\mathbf{S}^{(N)}(t)$ of sufficiently large systems using the solution $\mathbf{s}(t)$ to the fluid model. Before showing the convergence results rigorously, we introduce some notation for system state and provide some interpretation of the fluid model defined in Section 2.1.

## 3.1 Preliminaries

Fix the number of servers $N$. With homogeneous jobs, system state can be represented by $\mathbf{Q}^{(N)}(t) = (n_1^{(N)}(t), n_2^{(N)}(t), ..., n_N^{(N)}(t))$, where $n_m^{(N)}(t)$ is the number of jobs in server $m$ at time $t$. Under the Poisson arrivals and i.i.d exponential service time assumption, the process $\{\mathbf{Q}^{(N)}(t), t \geq 0\}$ is Markov with state space $\mathcal{Q}^{(N)}$. Note that $0 \leq n_m^{(N)}(t) \leq B$ as each server can accommodate at most $B$ jobs simultaneously. Define

$$S_k^{(N)}(t) = \frac{1}{N} \sum_{i=1}^{(N)} \mathbb{I}_{[k,B]}(n_i^{(N)}(t)), \ \forall k \in \{0, 1, 2, ..., B\},$$

where $S_k^{(N)}(t)$ represents the fraction of servers with at least $k$ jobs in service. Note that $S_0^{(N)}(t) = 1$ for all $t$. Since the system is fully symmetric, the evolution of the system can be described by the process $\{\mathbf{S}^{(N)}(t), t \geq 0\}$, which is also Markov. Moreover, the system is stable for any $\lambda \geq 0$, as the amount of resource at each server is finite and there is no extra waiting room for arrivals. Hence the Markov process $\{\mathbf{S}^{(N)}(t), t \geq 0\}$ is positive recurrent. We use $\boldsymbol{\pi}^{(N)} = (\pi_0^{(N)}, \pi_1^{(N)}, ..., \pi_B^{(N)})$ to denote its equilibrium distribution.

**Explanation for the drift of $s_k(t)$ in Eq. (1)**: Consider a system with $N$ servers. We will identify the expected change in the fraction of servers with at least $k$ jobs in service over a small period of time of length $dt$.

**(I).** The first term corresponds to the change caused by the arrivals. When an arriving job is assigned to a server with $k-1$ jobs, $S_k^{(N)}$ increases by $\frac{1}{N}$. Observe that the number of servers with at least $j$ jobs for $j \neq k$, does not change. Thus $S_k^{(N)}$ is increased by $\frac{1}{N}$ if only if an arriving job joins a server with $k-1$ jobs. Note that the probability that all $d$ sampled servers have at least $k-1$ jobs is $s_{k-1}^d$. The difference $s_{k-1}^d - s_k^d$ is the probability that at least one of the sampled servers has $k-1$ jobs. With total arrival rate $N\lambda$, the increment for $S_k^{(N)}$ during this time period due to arrival is hence $dt \times N\lambda \times \frac{1}{N} \times (s_{k-1}^d - s_k^d) = \lambda(s_{k-1}^d - s_k^d)dt$.

**(II).** The second term corresponds to the decrease due to the completion of jobs. The argument is similar to that of the first term.

## 3.2 Convergence Results

We first provide an overview of the convergence results:

First we prove some properties of the fluid model. We will show that there exists a unique solution $\boldsymbol{\pi}$ to the differential equations (1) which is stationary with respect to t, i.e., $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$ (Lemma 1). Moreover, given any finite initial condition, the solution to the fluid equation is unique and converges to the stationary solution as $t \to \infty$ (Lemma 2).

The second step is to show that as $N \to \infty$, the evolution of process $\mathbf{S}^{(N)}(t)$ converges uniformly, over any finite time interval, to the unique solution of the fluid model (Lemma 5). The result is derived by applying Kurtz's theorem ([13, 18]) for density dependent jump Markov processes.

The last step is to prove that the sequence of the steady-state distribution of $\mathbf{S}^{(N)}(t)$ (denoted by $\boldsymbol{\pi}^{(N)}$), concentrates at the unique stationary distribution of the fluid model ($\boldsymbol{\pi}$) as $N \to \infty$ (Theorem 1).

LEMMA 1. *There exists a unique solution $\boldsymbol{\pi} \in \mathcal{S}$ of the differential equation (1) that is invariant with respect to t, i.e., $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$.*

We defer the full proof of Lemma 1 to the appendix and provide an outline below.

**Proof outline of Lemma 1**

**Existence**: The stationary solution $\boldsymbol{\pi}$ satisfies the equation $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$. We construct a continuous mapping $\mathbf{G} : \mathcal{S} \to \mathcal{S}$, such that a fixed point of $\mathbf{G}$ is a solution to $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$. By Brouwer Fixed Point Theorem, $\mathbf{G}$ has at least one fixed point, i.e., there exists $\boldsymbol{\pi} \in \mathcal{S}$ such that $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$.

**Uniqueness**: We prove the uniqueness of stationary solution by contradiction and induction. First we show that if there exist two stationary solutions $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ satisfying $\pi_B = \hat{\pi}_B$, then $\pi_k = \hat{\pi}_k$ for any $k$. Therefore if there exist two different solutions $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$, $\pi_B \neq \hat{\pi}_B$. Assume $\pi_B < \hat{\pi}_B$, by induction, we can show that $\pi_k < \hat{\pi}_k$ for any $k = 0, 1, ..., B$, which contradicts with the fact that $\pi_0 = \hat{\pi}_0 = 1$.

LEMMA 2. *Given any initial condition $\mathbf{s}^0 \in \mathcal{S}$,*

1. *the fluid model has a unique solution $\mathbf{s}(\mathbf{s}^0, t)$ in $\mathcal{S}$,*
2. *as $t \to \infty$, the solution $\mathbf{s}(\mathbf{s}^0, t)$ converges to the unique stationary solution $\boldsymbol{\pi}$.*

We need the following lemmas to prove Lemma 2. The proofs of Lemma 3-4 are provided in the appendix.

LEMMA 3. *Let $\bar{\mathbf{s}}(t)$ and $\mathbf{s}(t)$ be the solutions to differential equations (1) with initial condition $\bar{\mathbf{s}}^0$ and $\mathbf{s}^0$ respectively. If $\bar{s}_k^0 \leq s_k^0$ for $k = 1, 2, ..., B$, then $\bar{s}_k(t) \leq s_k(t)$ for any $t \geq 0$.*

LEMMA 4. *Let $\psi(t) = \sum_{k=0}^{B} |s_k(t) - \pi_k|$, where $\mathbf{s}(t)$ is the solution to differential equations (1) with initial condition $\mathbf{s}^0$ satisfying $s_k^0 \geq \pi_k$ for any $k$ (or $s_k^0 \leq \pi_k$ for any $k$), then $\psi(t)$ converges to 0 as $t \to \infty$.*

**Proof of Lemma 2:** Item 1 follows by the arguments in Theorem 1.(a) of [28].

For any initial values $\mathbf{s}^0 \in \mathcal{S}$, define two initial conditions $\mathbf{s}^u$ and $\mathbf{s}^l$: $s_k^u = \max\{s_k^0, \pi_k\}$, $s_k^l = \min\{s_k^0, \pi_k\}$ for any $k$. Let $\mathbf{s}^u(t)$ and $\mathbf{s}^l(t)$ denote the solutions with initial conditions $\mathbf{s}^u$ and $\mathbf{s}^l$ respectively. From Lemma 3, we have $s_k^u(t) \geq \pi_k \geq s_k^l(t)$ for all $t$ and any $k$. Thus it is sufficient to show that $\lim_{t\to\infty} |\mathbf{s}^u(t) - \pi| = \lim_{t\to\infty} |\mathbf{s}^l(t) - \pi| = 0$, where $|\cdot|$ is $l_1$ norm. The result follows directly from Lemma 4. ■

LEMMA 5. *Consider a sequence of systems with the number of servers $N$ increasing to infinity. Fix any $T > 0$. If the sequence of initial system state $\{\mathbf{S}^{(N)}(0)\}_{N=1}^{\infty}$ concentrates on some $\mathbf{s}^0 \in \mathcal{S}$ as $N \to \infty$, then*

$$\lim_{N\to\infty} \sup_{t\in[0,T]} |\mathbf{S}^{(N)}(t) - \mathbf{s}(\mathbf{s}^0, t)| = 0 \quad a.s. \tag{9}$$

*where $\mathbf{s}(\mathbf{s}^0, t)$ is the solution to the differential equation (1) given initial condition $\mathbf{s}^0$.*

The following lemma is used to prove Lemma 5.

LEMMA 6. *The drift function $\mathbf{F}(\mathbf{s})$ is Lipschitz, i.e., there exits a constant $M > 0$ such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}$,*

$$|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}|,$$

*where $|\cdot|$ is $l_1$ norm.*

**Proof of Lemma 5:** We prove this lemma by Kurtz's theorem [13].

(a). It is easy to check that $\{\mathbf{S}^{(N)}(t), t \geq 0\}$ is a density dependent jump Markov process with state space $\mathcal{S}^{(N)}$.

(b). When the system is in state $\mathbf{s}$, the possible transitions is given by $\mathcal{L} = \{\pm\mathbf{e}_k : 1 \leq k \leq B\}$, where $\mathbf{e}_k$ are vectors with only the $k-$th element equal to $1/N$ and all other elements zero. The transition rates are given by $q^{(N)}_{\mathbf{s},\mathbf{s}+\mathbf{l}} = N\beta_{\mathbf{l}}(\mathbf{s})$, where $\beta_{\mathbf{e}_k}(\mathbf{s}) = \lambda(s_{k-1}^d - s_k^d)$ and $\beta_{-\mathbf{e}_k}(\mathbf{s}) = k(s_k - s_{k+1})$. Therefore the rate at which jumps occur is bounded above by $\lambda + B$ everywhere.

(c). Lemma 6 states that the differential equation for the limiting deterministic process satisfies the Lipschitz condition.

Then the result follows by Kurtz's Theorem. ■

**Proof of Theorem 1**:

We will use $\Rightarrow$ for weak convergence throughout the proof.

Note that set $\mathcal{S}$ is compact. By a corollary of Prokhorov's Theorem, for any subsequence of $\{N\}$, there exists a subsubsequence $\{N_k\}$ such that $\boldsymbol{\pi}^{(N_k)}$ converges weakly to some probability distribution $\bar{\boldsymbol{\pi}}$. By the Skorokhod's representation theorem, there exist a sequence of random vector $\{\mathbf{X}^{(N_k)}\}$ and a random vector $\bar{\mathbf{X}}$ such that

$$\mathbf{X}^{(N_k)} \overset{d}{=} \boldsymbol{\pi}^{(N_k)} \quad \bar{\mathbf{X}} \overset{d}{=} \bar{\boldsymbol{\pi}},$$

and

$$\mathbf{X}^{(N_k)} \xrightarrow{a.s.} \bar{\mathbf{X}} \text{ as } k \to \infty.$$

Let $\mathbf{S}^{(N_k)}(0) = \mathbf{X}^{(N_k)}$, i.e., start the system with $N_k$ servers at an initial condition specified by its stationary distribution. We use $\bar{\mathbf{S}}(t)$ to denote the random state of the dynamic system with initial condition $\bar{\mathbf{X}}$.

We have the following claim (proven in the appendix):

**Claim 1:** For any $t \geq 0$,

$$\mathbf{S}^{(N_k)}(t) \Rightarrow \bar{\mathbf{S}}(t) \text{ as } k \to \infty.$$

Then the result follows from the arguments in Theorem 5.1 of [2]. ■

**Remark.** Lemma 5 and Theorem 1 state that the behavior of sufficiently large systems can be approximated by that of the deterministic infinite system, which is described by a system of differential equations defined in Eq. (1).

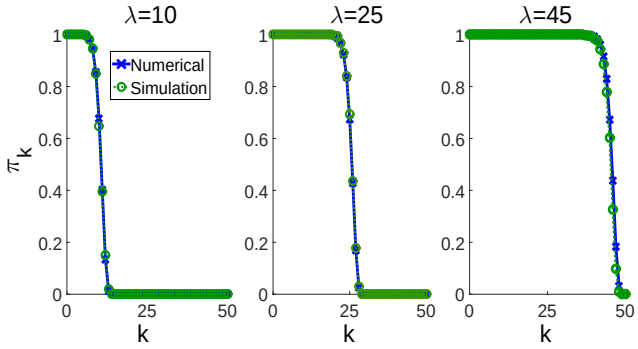


Figure 4: Equilibrium tail distribution for the power-of-two-choices algorithm with $B = 50$ at three different loads. The values for the stationary point are obtained numerically by solving Eq. (2). Simulation results are from a finite system with $N = 500$.

**Numerical Result.** Figure 4 shows the equilibrium tail distributions of the number of jobs at a server under the power-of-two-choices algorithm with $B = 50$ at three different loads, both by solving Eq. (2) numerically and by simulating a finite system with $N = 500$. We can see that the coincidence of the empirical distribution with the stationary point is almost exact. That is, values of the stationary point in the large system limit predict that of a finite system very well.

## 4. AN UPPER-BOUND FOR THE HOMOGENEOUS CASE

Unlike the supermarket model operating under the power-of-$d$-choices policy [18, 28], there is no explicit expression for the stationary point $\boldsymbol{\pi}$ of the loss model. We establish an explicit upper-bound for $\boldsymbol{\pi}$. Observe that the proposed upper-bound $\bar{\boldsymbol{\pi}}$ (defined in Eq. (3)) can be expressed by a recursive formula as follows:

$$\bar{\pi}_k = \begin{cases} 1, & 0 \le k \le i_0 + 1 \\ \frac{\lambda}{k-1}\bar{\pi}^d_{k-1}, & i_0 + 1 < k \le B \end{cases}$$

where $i_0 = \lfloor \lambda \rfloor$.

**Proof of Theorem 2:** We complete the proof in two steps.

(i) First we show that $\pi_k \le \frac{\lambda}{k}\pi^d_{k-1}$ for $1 \le k \le B$ by backward induction. The inequality holds for $k = B$:

$$\pi_B - \frac{\lambda}{B}\pi^d_{B-1} = -\frac{\lambda}{B}\pi^d_B \le 0.$$

Assume that $\pi_{k+1} \le \frac{\lambda}{k+1}\pi^d_k$ hold for $k + 1 \le B$. Then

$$\pi_k - \frac{\lambda}{k}\pi^d_{k-1} = \pi_{k+1} - \frac{\lambda}{k}\pi^d_k \le \pi_{k+1} - \frac{\lambda}{k+1}\pi^d_k \le 0.$$

Hence $\pi_k \le \frac{\lambda}{k}\pi^d_{k-1}$, $\forall k = 1, 2, ..., B$.

(ii) Next we prove the theorem by induction.

For any $k \le i_0 + 1$, $\bar{\pi}_k = 1 \ge \pi_k$.

Assume that $\bar{\pi}_k \ge \pi_k$ hold for some $k \ge i_0 + 1$. Then

$$\bar{\pi}_{k+1} = \frac{\lambda}{k}\bar{\pi}^d_k \ge \frac{\lambda}{k}\pi^d_k \ge \frac{\lambda}{k}\pi^d_k + \pi_k - \frac{\lambda}{k}\pi^d_{k-1} = \pi_{k+1},$$

where the first inequality comes from the assumption and the second one follows by the property of $\boldsymbol{\pi}$ we just proved. ■

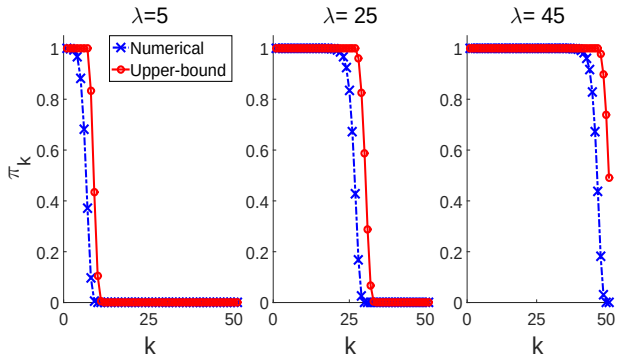


Figure 5: An upper bound for the stationary point.

Figure 5 compares the equilibrium tail distribution of the stationary point and the proposed distribution $\bar{\boldsymbol{\pi}}$ with $B = 50$ at three different loads. We can see that the upper bound always holds. Moreover, the proposed distribution characterizes the steep slope of the stationary point, i.e., $\pi_k$ decreases drastically from 1 to 0 at some $k$.

Since the performance measure of primary interest is the blocking probability $P_b$, we are interested in the tightness of the upper-bound blocking probability.

| $\rho = \lambda/B$ | Fluid limit | Upper-bound |
|---|---|---|
| 0.6 | 0 | 0 |
| 0.8 | 0 | $4.508 \times 10^{-27}$ |
| 0.84 | 0.0000 | $7.003 \times 10^{-7}$ |
| 0.88 | $1.873 \times 10^{-25}$ | 0.0426 |
| 0.9 | $5.229 \times 10^{-13}$ | 0.2419 |
| 0.92 | $8.240 \times 10^{-7}$ | 0.5535 |
| 0.94 | $7.854 \times 10^{-4}$ | 0.8122 |

Table 1: The blocking probability for the power-of-two-choices policy with $B = 50$ at different load.

Table 1 compares the upper-bound blocking probability and values of the stationary fluid limit under the power-of-two-choices policy with $B = 50$ at different loads. The values given by the upper-bound are quite close to that of the stationary fluid limit at low to medium load. With $B$ fixed, as the load increases towards 1, the gap increases. We have seen that the proposed upper-bound resembles a shift of the stationary fluid limit from Fig. 5. At high load, the upper-bound shifts too much that the resulting bound for blocking probability becomes loose. However, if we fix the load $\rho = \frac{\lambda}{B}$ for the system, we can see that the upper-bound blocking probability $\bar{\pi}^d_B$ decays to 0 as $B$ increases. This implies that the upper-bound becomes tight for sufficiently large $B$.

## 5. PROOF OF THEOREM 3

We devote this section to the proof of Theorem 3. We begin by proving the following lemma.

LEMMA 7. *Let $\lambda < B$ and $\frac{\lambda}{B} \to 1$ as $B \to \infty$. If $\frac{k}{B-\lambda} \to \theta$ as $B \to \infty$, where $\theta$ is a constant satisfying $0 \le \theta < 1$, then*

$$\frac{\lambda^{B-i_0-k} \cdot i_0!}{(B-k)!} \sim e^{-\frac{(1-\theta)^2(B-\lambda)^2}{2\lambda}}, \tag{10}$$

*where $i_0 = \lfloor \lambda \rfloor$.*

**Proof.** By Stirling's formula, we have

$$\begin{aligned}
&\frac{\lambda^{B-i_0-k} \cdot i_0!}{(B-k)!} \\
\sim\ & \lambda^{B-i_0-k} \frac{\sqrt{2\pi i_0} \cdot (\frac{i_0}{e})^{i_0}}{\sqrt{2\pi(B-k)} \cdot (\frac{B-k}{e})^{B-k}} \\
\sim\ & \sqrt{\frac{i_0}{B-k}} \cdot e^{B-k-\lambda} \left(\frac{\lambda}{B-k}\right)^{\lambda} \cdot \left(\frac{\lambda}{B-k}\right)^{B-k-\lambda} \\
\sim\ & e^{B-k-\lambda} \left(\frac{\lambda}{B-k}\right)^{\lambda} \cdot \left(\frac{\lambda}{B-k}\right)^{B-k-\lambda} \qquad (11)
\end{aligned}$$

Define $\Delta = B - \lambda$. Note that $\lambda/\Delta \to \infty$ as $B \to \infty$. And $(B-k-\lambda) \sim (1-\theta)\Delta$. Then we have

$$\begin{aligned}
\left(\frac{\lambda}{B-k}\right)^{B-k-\lambda} &\sim \left(\frac{\lambda}{\lambda+(1-\theta)\Delta}\right)^{(1-\theta)\Delta} \\
&\sim \left(1+\frac{(1-\theta)\Delta}{\lambda}\right)^{-(1-\theta)\Delta} \\
&\sim \left[\left(1+\frac{(1-\theta)}{\lambda/\Delta}\right)^{-\lambda/\Delta}\right]^{-(1-\theta)\frac{\Delta^2}{\lambda}} \\
&\sim e^{-(1-\theta)^2\frac{\Delta^2}{\lambda}} \qquad (12)
\end{aligned}$$

Now consider the first two terms in Eq. (11).

$$\begin{aligned}
&\log\left(e^{B-k-\lambda}\left(\frac{\lambda}{B-k}\right)^{\lambda}\right) \\
\sim\ & (1-\theta)\Delta - \lambda\log\left(1+\frac{(1-\theta)\Delta}{\lambda}\right) \\
\sim\ & (1-\theta)\Delta - \lambda\left(\frac{(1-\theta)\Delta}{\lambda} - \frac{1}{2}\left(\frac{(1-\theta)\Delta}{\lambda}\right)^2 + o(\lambda)\right) \\
\sim\ & \frac{(1-\theta)^2\Delta^2}{2\lambda} + o(1)
\end{aligned}$$

That is,

$$e^{B-k-\lambda}\left(\frac{\lambda}{B-k}\right)^{\lambda} \sim e^{\frac{(1-\theta)^2\Delta^2}{2\lambda}} \tag{13}$$

Equations (12)-(13) yield the asymptotic approximation in Eq. (10). ∎

**Proof of Theorem 3:**

From Theorem 2, it is sufficient to show that the upper bound $\bar{\pi}_B$ defined in (3) satisfies Eq. (5). We establish this result using Lemma 7.

We can write $\bar{\pi}_B$ as

$$\begin{aligned}
\bar{\pi}_B &= \left(\frac{\lambda^{B-i_0-1}\cdot i_0!}{(B-1)!}\right)\cdot\left(\frac{\lambda^{B-i_0-2}\cdot i_0!}{(B-2)!}\right)^{(d-1)\cdot d^0} \\
&\quad \cdot\left(\frac{\lambda^{B-i_0-3}\cdot i_0!}{(B-3)!}\right)^{(d-1)d} \cdots \left(\frac{\lambda\cdot i_0!}{(i_0+1)!}\right)^{(d-1)d^{B-i_0-3}} \qquad (14)
\end{aligned}$$

Note that each term within the bracket in Eq.(14) is no greater than 1. We can obtain an upper bound for $\bar{\pi}_B$ by discarding some terms in Eq. (14). In particular, consider keeping the first $m$ terms, where $m = (1-c)(B-\lambda)$, $c$ is an arbitrary constant satisfying $0 < c < 1$. From Lemma 7, each term we keep here can be approximated by using Eq (10). Define $\Delta = B - \lambda$. Then we have

$$\begin{aligned}
\bar{\pi}_B &\le \left(\frac{\lambda^{B-i_0-1}\cdot i_0!}{(B-1)!}\right)\cdot\left(\frac{\lambda^{B-i_0-2}\cdot i_0!}{(B-2)!}\right)^{(d-1)d^0} \\
&\quad \cdots\left(\frac{\lambda^{B-i_0-m}\cdot i_0!}{(B-m)!}\right)^{(d-1)d^{m-2}} \\
&\sim e^{-\frac{\Delta^2}{2\lambda}\left[(1-\frac{1}{\Delta})^2+(1-\frac{2}{\Delta})^2\cdot(d-1)+\cdots+(1-\frac{m}{\Delta})^2\cdot(d-1)d^{m-2}\right]} \\
&\lesssim e^{-\frac{c^2\Delta^2}{2\lambda}\cdot d^{m-1}} \\
&= \left(e^{-\frac{c^2}{2}}\right)^{\frac{\Delta^2}{\lambda}\cdot d^{(1-c)\Delta-1}}.
\end{aligned}$$

We complete the proof for Eq. (5). As discussed in Section 2.1, we have $P_b = \pi_B^d$. Thus,

$$P_b \lesssim \left(e^{-\frac{c^2}{2}}\right)^{\frac{\Delta^2}{\lambda}\cdot d^{(1-c)\Delta}}$$

Now we can study the blocking probability with various load gap by analyzing the exponent $\frac{c^2}{2}\frac{\Delta^2}{\lambda}\cdot d^{(1-c)\Delta}$.

1. $\frac{B-\lambda}{\sqrt{\lambda}} \to \alpha$: we have:

$$\begin{aligned}
&\log_d\log\frac{1}{P_b} \\
\gtrsim\ & 2\log_d\Delta - \log_d\lambda + \log_d\frac{c^2}{2} + (1-c)\Delta \\
\sim\ & ((1-c)\alpha + o(1))\sqrt{\lambda}.
\end{aligned}$$

2. $\frac{B-\lambda}{\log_d\lambda} \to \beta$: As $\beta > 1$ and $0 < c < 1$ is an arbitrary constant, we can select $c$ to make $\gamma = (1-c)\beta - 1 > 0$. Then we have:

$$\begin{aligned}
&\log_d\log\frac{1}{P_b} \\
\gtrsim\ & ((1-c)\beta-1)\log_d\lambda + 2\log_d\log\lambda + o(1) \\
\sim\ & (\gamma + o(1))\log_d\lambda.
\end{aligned}$$

Hence

$$\log\frac{1}{P_b} \gtrsim \lambda^{\gamma+o(1)}.$$

∎

## 6. HETEROGENEOUS JOBS

In this section, we focus on the heterogeneous job case. In particular, we will employ the ansatz in [8], which asserts that in equilibrium, any finite set of queues in a randomized

load balancing system become asymptotically independent as the number of queues goes to infinity. This will allows us to derive the equilibrium distribution by studying a single server, which has state-dependent Poisson arrivals.

## 6.1 Independence Ansatz

The asymptotic independence for a supermarket model operating under the power-of-$d$ policy with exponentially distributed service time was established by Graham [10] using the propagation of chaos approach. And the independence ansatz for general service time distributions was demonstrated in [8]. A key step of the existing approaches involves standard coupling to establish a monotonicity property for the supermarket model, which is essential to proving the independence ansatz. The monotonicity property states that there exists a coupling such that the evolution of a system with any non-zero initial condition stochastically dominates the evolution of the same system with the all-zeros initial condition. The monotonicity argument is used to demonstrate uniform convergence, i.e., the distance between the two evolutions of the system monotonically decreases with time. This ensures convergence of the system under the arbitrary initial condition to the limiting equilibrium distribution.

We found that it is difficult to establish the independence ansatz using such approach as the loss model with the power-of-$d$ policy does not satisfy the monotonicity property. Consider two copies $\mathbf{X}_1(\cdot)$ and $\mathbf{X}_2(\cdot)$ of the loss model under the power-of-$d$ policy. And assume element-wise dominance of $\mathbf{X}_1(\cdot)$ over $\mathbf{X}_2(\cdot)$. With exponential service times, departures of the two systems can always be coupled. Problem comes from blocking for arrivals. As an arrival is blocked when it is assigned to a server with insufficient resource, it is possible that jobs are blocked in the heavier-loaded system $\mathbf{X}_1(\cdot)$ while enter the lighter-loaded system $\mathbf{X}_2(\cdot)$. This might break the dominance. Therefore monotonicity does not hold for the loss model by standard coupling.

Justification of the independence ansatz for our model remains to be done. However, we believe that it is true considering the randomized nature of power-of-$d$ algorithms. In this following section, we derive some interesting results under the independence ansatz.

## 6.2 Equilibrium Distribution for A Single Queue

We assume asymptotic independence for the loss model with the power-of-$d$ algorithm. Consider server 1 (by symmetry, any server) in the large $N$ limit. Under the asymptotic independence assumption, the arrival process of type $j$ jobs to server 1 is a state-dependent Poisson process with rate $\lambda_j(\mathbf{n})$, which is given in Eq. (6).

We can explain Eq (6) as follows: Assume that server 1 is of state $\mathbf{n}$. When a type $j$ job arrives at the system, it will join server 1 *only if* server 1 is chosen and the state $\mathbf{n}'$ of any other selected server satisfies the condition $\langle \mathbf{n}', \mathbf{b}\rangle \geq \langle \mathbf{n}, \mathbf{b}\rangle$, i.e., server 1 has the largest amount of available resource. Note that server 1 is selected as one of the $d$ sampled servers with probability $\frac{\binom{N-1}{d-1}}{\binom{N}{d}} = \frac{d}{N}$. Consider the case where $i-1$ out of the other $d-1$ selected servers have the same amount of available resource, $i \in \{1, 2, ..., d\}$. Such an event happens with probability $\binom{d-1}{i-1} E_{\mathbf{n}}^{i-1} G_{\mathbf{n}}^{d-i}$, where $E_{\mathbf{n}}$ ($G_{\mathbf{n}}$) represents the fraction of servers with the same (larger) amount of resource occupied. As ties are broken randomly, server 1 is selected with probability $\frac{1}{i}$. Hence the probability that the arrival is routed to server 1 is given by

$$\sum_{i=1}^{d} \frac{d}{N} \cdot \frac{1}{i} \cdot \binom{d-1}{i-1} E_{\mathbf{n}}^{i-1} G_{\mathbf{n}}^{d-i} = \frac{1}{N} \sum_{i=1}^{d} \binom{d}{i} E_{\mathbf{n}}^{i-1} G_{\mathbf{n}}^{d-i}.$$

Multiplying this probability by the arrival rate of type $j$ jobs and letting $N \to \infty$ yield Eq. (6).

Note that queue 1 is a birth-death process with state-dependent arrival and departure rates. The global balance equation is given by:

$$\begin{aligned} &\left[\sum_{j=1}^{J} n_j \delta_j^-(\mathbf{n}) + \sum_{j=1}^{J} \lambda_j(\mathbf{n}) \delta_j^+(\mathbf{n})\right] p_{\mathbf{n}} \\ = \quad &\sum_{j=1}^{J} \lambda_j(\mathbf{n}_j^-) \delta_j^-(\mathbf{n}) p_{\mathbf{n}_j^-} + \sum_{j=1}^{J} (n_j+1) \delta_j^+(\mathbf{n}) p_{\mathbf{n}_j^+}, \end{aligned} \quad (15)$$

where

$$\begin{aligned} \mathbf{n}_j^+ &= (n_1, n_2, ..., n_{j-1}, n_j + 1, n_{j+1}, ..., n_J), \\ \mathbf{n}_j^- &= (n_1, n_2, ..., n_{j-1}, n_j - 1, n_{j+1}, ..., n_J), \\ \delta_j^+(\mathbf{n}) &= \begin{cases} 1, & \text{if } \mathbf{n}_j^+ \in \mathcal{C} \\ 0, & \text{otherwise} \end{cases} \\ \delta_j^-(\mathbf{n}) &= \begin{cases} 1, & \text{if } \mathbf{n}_j^- \in \mathcal{C} \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Moreover, the local balance equation is given by

$$\begin{aligned} &\lambda_j(\mathbf{n}_j^-) \delta_j^-(\mathbf{n}) p_{\mathbf{n}_j^-} = n_j \delta_j^-(\mathbf{n}) p_{\mathbf{n}}, \quad (16) \\ &\forall j \in \{1, 2, ..., J\}, \forall \mathbf{n} \in \mathcal{C}. \end{aligned}$$

**Remark.** We notice that if the local balance equations are satisfied, the global balance equations are satisfied. However, we have not established that the global balance equations have a unique solution. This is normally true for queueing systems where the arrival rate is fixed; however, since the derivation here follows from the independence ansantz, the arrival rate depends on $\mathbf{p}$. Thus, establishing the uniqueness of the solution to Eq. (15) remains to be done.

## 6.3 One-dimensional Recursion

We are interested in the probability $P_{b_j}$ that an arriving job of type $j$ is blocked. Note that

$$P_{b_j} = \left(\sum_{\mathbf{n} \in \mathcal{T}_j^+} p_{\mathbf{n}}\right)^d, \quad (17)$$

where $\mathcal{T}_j^+ = \{\mathbf{n} \in \mathcal{C} : \mathbf{n}_j^+ \notin \mathcal{C}\}$.

The underlying high dimension of the state $\mathbf{n}$ makes it difficult to obtain the equilibrium distribution from Eq. (16). In order to quantify the blocking probability, we will use Kaufman-Roberts recursion [12, 22] to establish a one-dimensional recursion, regardless of the dimensionality of jobs types (Theorem 4). The key idea is to pay attention to the random variable $R(\mathbf{n}) = \sum_{j=1}^{J} n_j b_j$, which denote the amount of occupied resource. We use $\mathbf{r}$ to represent the tail distribution of $R(\mathbf{n})$, i.e.,

$$r_k = \Pr[R \geq k] = \sum_{\mathbf{n} \in \mathcal{C}: \langle \mathbf{n}, \mathbf{b}\rangle \geq k} p_{\mathbf{n}}, \text{ for } k = 0, 1, ..., B.$$

Note that $r_k$ is also the asymptotic *fraction* of servers having at least $k$ units of resource occupied. For ease of exposition, throughout this section, we define $r_x = 1$ for any $x \le 0$, and $r_{B+1} = 0$.

In order to prove Theorem 4, we need the following lemma.

LEMMA 8. *For any $j \in \mathcal{J}$, and $k \in \{0, 1, ..., B\}$,*

$$\lambda_j(r^d_{k-b_j} - r^d_{k-b_j+1}) = \mathbb{E}\left[n_j | \langle \mathbf{n}, \mathbf{b}\rangle = k\right](r_k - r_{k+1}), \quad (18)$$

*where $r_x = 1$ for any $x \le 0$ and $r_{B+1} = 0$.*

**Proof.** Equation (16) can be written as :

$$\lambda_j(\mathbf{n}_j^-)\gamma_j(\mathbf{n})p_{\mathbf{n}_j^-} = n_j p_{\mathbf{n}}, \quad (19)$$

where

$$\gamma_j(\mathbf{n}) = \begin{cases} 1 & \text{if } n_j \ge 1 \\ 0 & \text{if } n_j = 0 \end{cases}$$

For any $k \in \{0, 1, ..., B\}$, define $\mathcal{D}_k = \{\mathbf{n} \in \mathcal{C} : k = \sum_{j=1}^J n_j b_j\}$. Note that for any $\mathbf{n} \in \mathcal{D}_k$,

$$E_{\mathbf{n}} = r_k - r_{k+1},\ G_{\mathbf{n}} = r_{k+1}.$$

Hence $\lambda_j(\mathbf{n})$ depends on $k = \langle \mathbf{n}, \mathbf{b}\rangle$ only.
Summing Eq. (19) over the set $\mathcal{D}_k$, we have

$$\sum_{\mathbf{n}\in\mathcal{D}_k} \lambda_j(\mathbf{n}_j^-)\gamma_j(\mathbf{n})p_{\mathbf{n}_j^-} = \sum_{\mathbf{n}\in\mathcal{D}_k} n_j p_{\mathbf{n}}. \quad (20)$$

Consider the left-hand-side (LHS) of (20).

$$\begin{aligned} LHS &= \sum_{\mathbf{n}\in\mathcal{D}_k} \lambda_j(\mathbf{n}_j^-)\gamma_j(\mathbf{n})p_{\mathbf{n}_j^-} \\ &= \lambda_j \sum_{\mathbf{n}\in\mathcal{D}_k} \left(\sum_{i=1}^d \binom{d}{i} E^{i-1}_{\mathbf{n}_j^-} G^{d-i}_{\mathbf{n}_j^-}\right) \gamma_j(\mathbf{n}) p_{\mathbf{n}_j^-} \\ &= \lambda_j \sum_{\mathbf{n}\in\mathcal{D}_k\cap\{\mathbf{n}:n_j\ge 1\}} \left(\sum_{i=1}^d \binom{d}{i} E^{i-1}_{\mathbf{n}_j^-} G^{d-i}_{\mathbf{n}_j^-}\right) p_{\mathbf{n}_j^-}. \end{aligned}$$

Note that

$$\begin{aligned} &\mathcal{D}_k \cap \{\mathbf{n} : n_j \ge 1\} \\ &= \left\{\mathbf{n} \in \mathcal{C} : \sum_{i\ne j} n_i b_i + (n_j - 1)b_j = k - b_j, n_j \ge 1\right\}. \end{aligned}$$

Let $\hat{\mathbf{n}} = \mathbf{n}_j^-$. Then

$$\begin{aligned} LHS &= \lambda_j \sum_{\hat{\mathbf{n}}\in\mathcal{D}_{k-b_j}} \left(\sum_{i=1}^d \binom{d}{i} E^{i-1}_{\hat{\mathbf{n}}} G^{d-i}_{\hat{\mathbf{n}}}\right) p_{\hat{\mathbf{n}}} \\ &= \lambda_j \left(\sum_{i=1}^d \binom{d}{i} (r_{k-b_j} - r_{k-b_j+1})^{i-1} r^{d-i}_{k-b_j+1}\right) \sum_{\hat{\mathbf{n}}\in\mathcal{D}_{k-b_j}} p_{\hat{\mathbf{n}}} \\ &= \lambda_j \left(\sum_{i=1}^d \binom{d}{i} (r_{k-b_j} - r_{k-b_j+1})^{i} r^{d-i}_{k-b_j+1}\right) \\ &= \lambda_j(r^d_{k-b_j} - r^d_{k-b_j+1}). \end{aligned} \quad (21)$$

The right-hand side (RHS) of (20) can be written as

$$\begin{aligned} RHS &= \sum_{\mathbf{n}\in\mathcal{D}_k} n_j \frac{p_{\mathbf{n}}}{\mathbb{P}[\{\mathbf{n} : \langle \mathbf{n}, \mathbf{b}\rangle = k\}]} \mathbb{P}[\{\mathbf{n} : \langle \mathbf{n}, \mathbf{b}\rangle = k\}] \\ &= \sum_{\mathbf{n}\in\mathcal{D}_k} n_j \mathbb{P}[\mathbf{n} | \langle \mathbf{n}, \mathbf{b}\rangle = k](r_k - r_{k+1}) \\ &= \mathbb{E}\left[n_j | \langle \mathbf{n}, \mathbf{b}\rangle = k\right](r_k - r_{k+1}). \end{aligned} \quad (22)$$

Equation (18) follows from Eq. (21) and (22). ■

**Proof of Theorem 4**: Multiplying Eq. (18) by $b_j$ on both side and summing over $j$ yields

$$\begin{aligned} &\sum_{j=1}^J \lambda_j b_j (r^d_{k-b_j} - r^d_{k-b_j+1}) \\ &= \sum_{j=1}^J b_j \mathbb{E}\left[n_j | k\right](r_k - r_{k+1}) \\ &= \mathbb{E}\left[\sum_{j=1}^J b_j n_j | k\right](r_k - r_{k+1}) \\ &= k(r_k - r_{k+1}) \end{aligned}$$

■

**Remark.** We can write the blocking probability for jobs of type $j$ as

$$P_{b_j} = \left(\sum_{\mathbf{n}\in\mathcal{C}:\langle \mathbf{n},\mathbf{b}\rangle > B - b_j} p_{\mathbf{n}}\right)^d = r^d_{B-b_j+1}.$$

By solving Eq. (7), we can obtain $P_{b_j}$ immediately. Compared with the formula (17), the one-dimensional recursion brings a significant reduction in computation.

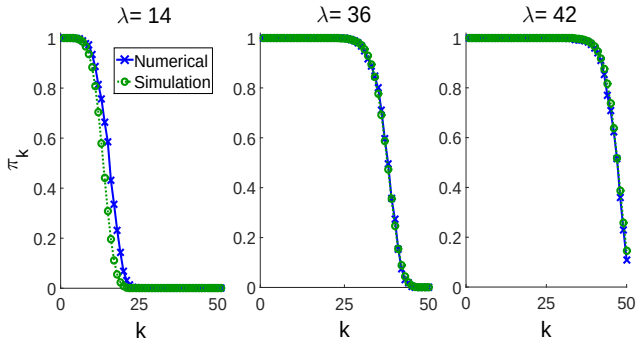


Figure 6: Equilibrium distribution of the number of occupied resource units for the power-of-two-choices algorithm with $B = 50$ and three types of jobs, where $\mathbf{b} = (1, 2, 4)$ and $\lambda_j = \lambda/7$. The values for the stationary point are obtained numerically by solving Eq. (7). Simulation results are from a finite system with $N = 1000$.

**Numerical Results.** Figure 6 compares the empirical distribution from simulation of a finite system with $N = 1000$ with the stationary point at three different loads. Simulation results coincide with the stationary point very well, which also verifies the validness of independence ansatz.

## 6.4 Upper Bound

We first establish an upper bound for the tail distribution $\mathbf{r}$ of the number of occupied resource units. Let $\lambda = \sum_{j=1}^J \lambda_j b_j$ be the total traffic intensity. We have the following theorem.

THEOREM 6. *Define* $\{\bar{r}_k\}_{k=0}^B$ *as follows:*

$$\bar{r}_k = \begin{cases} 1, & 0 \le k \le k_0+1 \\ \frac{1}{k-1}\sum_{j=1}^{J} \lambda_j b_j \bar{r}_{k-b_j}^d, & k_0+1 < k \le B \end{cases} \quad (23)$$

*where* $k_0 = \lfloor \lambda \rfloor$, $\bar{r}_x = 1$ *for any* $x \le 0$ *and* $\bar{r}_{B+1} = 0$.

*Let* $\{r_k\}_{k=0}^B$ *denote the solution to Eq (7). Then for any* $k = 0, 1, ..., B$,

$$\bar{r}_k \ge r_k.$$

Proof of Theorem 6 is essentially the same as that of Theorem 2.

LEMMA 9. *Define* $\{\tilde{r}_k\}_{k=0}^B$ *as follows:*

$$\tilde{r}_k = \begin{cases} 1, & 0 \le k < b(k_0'+2) \\ \frac{\lambda}{(m-1)b}\tilde{r}_{k-b}^d, & mb \le k < (m+1)b, \ k \le B, \\ & m \in \mathbb{N} \text{ and } k_0'+1 < m \le \frac{B}{b} \end{cases} \quad (24)$$

*where* $b = \max_{j=1,...,J} b_j$, *and* $k_0' = \lfloor \frac{\lambda}{b} \rfloor$

*Then* $\tilde{\mathbf{r}}$ *gives an upper bound for* $\bar{\mathbf{r}}$, *i.e., for any* $k = 0, 1, ..., B$,

$$\tilde{r}_k \ge \bar{r}_k.$$

The following corollary follows immediately by Theorem 6 and Lemma 9.

COROLLARY 1. $\tilde{\mathbf{r}}$ *is an upper bound for* $\mathbf{r}$, *i.e.,*

$$\tilde{r}_k \ge r_k, \forall k = 0, 1, ..., B.$$

**Remark.** Although the upper bound $\bar{\mathbf{r}}$ has no explicit expression, the recursion is straightforward and no further iterative calculation is needed here. Lemma 9 provides a further upper bound on $\bar{\mathbf{r}}$ which is used in the analysis of the blocking probability in the heavy-traffic and critically-loaded traffic regimes (Theorem 5).

## 6.5 Proof Outline of Theorem 5

*Proof outline of Theorem 5:*

Note that $b = \max_{j=1,...,J} b_j$. By the monotonicity of the tail distribution $\{r_k\}_{k=0}^B$, the blocking probability $P_{b_j}$ for type $j$ jobs ($\forall j \in \{1, 2, ..., J\}$) satisfies

$$P_{b_j} = r_{B-b_j+1}^d \le r_{B-b+1}^d \le \tilde{r}_{B-b+1}^d.$$

Hence it is sufficient to show that the upper-bound $\tilde{r}_{B-b+1}$ satisfies (8).

From the definition of $\tilde{\mathbf{r}}$, we can see that $\{\tilde{r}_k\}_{k=0}^B$ consists of consecutive subsequences of size $b$, where elements in each subsequence have the same value. That is, $\forall k \in [mb, (m+1)b)$, $m \in \mathbb{N}$, $\tilde{r}_k = \tilde{r}_{mb}$. To analyze its asymptotic behavior, we consider the subsequence $\{\tilde{r}_{mb}\}_{m\in\mathbb{N}}$. Define the scaled arrival rate $\lambda' = \lambda/b$, and resource units $B' = \lfloor B/b \rfloor$. Then the recursion of $\{\tilde{r}_{mb}\}_{m=0}^{B'}$ is the same as $\bar{\pi}$ with arrival rate $\lambda'$ and $B'$ units of resource.

By following the proof for Theorem 3, we can establish the asymptotic behavior of $\tilde{r}_{B'b}$ in large $B'$ limit, which gives Eq. (8). The analysis for the two limiting regimes is the same as that in Theorem 3.

**Remark.** Theorem 5 states that for the general case with multiple types of jobs, the blocking probability for jobs of any type under the power-of-$d$ algorithm has exactly the same asymptotic behavior as that of homogeneous job case.

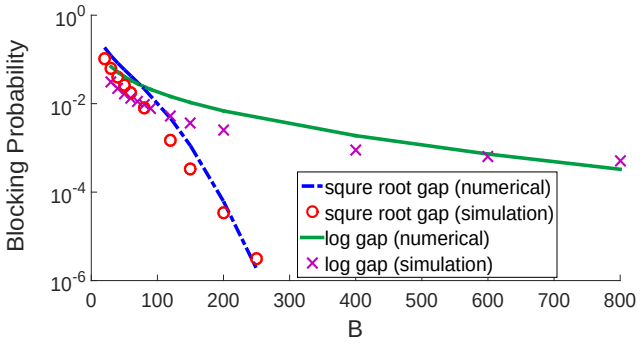


Figure 7: Blocking probability for the power-of-two-choices algorithm with different load gap. There are three types of jobs with $\mathbf{b} = (1, 2, 4)$. Line curves are obtained by solving Eq. (7) numerically. Markers are from simulations with $N = 1000$.

**Numerical Results.** We simulate a system of $N = 1000$ servers under the power-of-two-choices algorithm with different load gap. We consider three types of jobs with same arrival rate, i.e., $\lambda_1 = \lambda_2 = \lambda_3$, and $\mathbf{b} = (1, 2, 4)$. For each $B$, we simulate this system with different load gap $B - \lambda = \sqrt{\lambda}$ and $B - \lambda = 2\log\lambda$, where $\lambda = \sum_j \lambda_j b_j$ is the total traffic intensity. Figure 7 compares the blocking probability for jobs that require the maximum amount of resource, i.e., type 3 jobs, with different load gap, both by solving Eq. (7) numerically and by simulation. Note that the y-axis is in log scale. Observe that the blocking probability for jobs of type 3 exhibits similar behavior as that of the homogeneous job case (Fig. 1). That is, $P_{b_3}$ decays exponentially with $\log\lambda$ load gap, while it decays much faster with $\sqrt{\lambda}$ load gap. Similar behavior can be observed for the blocking probability of the other two types of jobs.

# 7. CONCLUSION AND FUTURE WORK

This paper considered a loss model for the VM assignment problem in a cloud system. The overall goal is to study how to route arriving jobs to the servers in order to minimize the probability that an arriving job does not find the required number of resources in the system. Using the fluid model approach, we showed that when arrivals are routed to the least utilized of $d \ge 2$ randomly selected servers, the blocking probability decays exponentially or doubly exponentially. This is a substantial improvement over the random policy. In addition, we developed an explicit upper-bound for the stationary fluid limit. The analysis of the upper-bound revealed significant insight into the asymptotic behavior of large systems with the power-of-$d$-choices ($d \ge 2$) algorithm.

We have seen that for a fixed $B$, the gap between the proposed upper-bound and the stationary fluid limit increases with the load. For future work, we are interested in characterizing the gap and establishing an approximation with higher accuracy. Some of current model assumptions could be relaxed to make the model closer to the real system, including the assumption of exponential service times and the constraint on the one-dimensionality of requested resources.

## 8. ACKNOWLEDGMENT

Qiaomin Xie and Yi Lu were supported by NSF grant CNS-1150080. The work of Xiaobo Dong and R. Srikant was supported in part by NSF Grant ECCS-1202065 and ARO MURI W911NF-12-1-0385. We thank Prof. Ravi Mazumdar for pointing out a gap in the proof of Lemma 1 in an earlier version of the paper.

## 9. REFERENCES

[1] Amazon EC2. `http://aws.amazon.com/ec2/`.

[2] V. Anantharam and M. Benchekroun. A technique for computing sojourn times in large networks of interacting queues. *Probability in the Engineering and Informational Sciences*, 7(04):441–464, 1993.

[3] Google App Engine. `https://cloud.google.com/appengine/docs?csw=1`.

[4] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal. Balanced allocations. *SIAM J. Comput.*, 29(1):180–200, Sept. 1999.

[5] Azure. `http://azure.microsoft.com/en-us/`.

[6] N. Bansal, A. Caprara, and M. Sviridenko. A new approximation method for set covering problems, with applications to multidimensional bin packing. *SIAM Journal on Computing*, 39(4):1256–1278, 2010.

[7] A. A. Borovkov. *Stochastic Processes in Queueing Theory*. Springer, 1976.

[8] M. Bramson, Y. Lu, and B. Prabhakar. Asymptotic independence of queues under randomized load balancing. *Queueing Systems*, 71(3):247–292, 2012.

[9] J. Csirik, D. S. Johnson, C. Kenyon, J. B. Orlin, P. W. Shor, and R. R. Weber. On the sum-of-squares algorithm for bin packing. *J. ACM*, 53(1):1–65, Jan. 2006.

[10] C. Graham. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability*, 37(1):198–211, 2000.

[11] V. Gupta and A. Radovanovic. Online stochastic bin packing, 2012.

[12] J. Kaufman. Blocking in a shared resource environment. *Communications, IEEE Transactions on*, 29(10):1474–1481, Oct 1981.

[13] T. G. Kurtz. *Approximation of Population Processes*. Society for Industrial and Applied Mathematics, 1981.

[14] M. Luczak and C. McDiarmid. On the maximum queue length in the supermarket model. *The Annals of Probability*, 34(2):493–527, 2006.

[15] S. Maguluri, R. Srikant, and L. Ying. Stochastic models of load balancing and scheduling in cloud computing clusters. In *Proc. of IEEE INFOCOM*, pages 702–710, Mar 2012.

[16] S. T. Maguluri, R. Srikant, and L. Ying. Heavy traffic optimal resource allocation algorithms for cloud computing clusters. In *Proc. of the 24th International Teletraffic Congress*, pages 25:1–25:8, 2012.

[17] X. Meng, V. Pappas, and L. Zhang. Improving the scalability of data center networks with traffic-aware virtual machine placement. In *Proc. of IEEE INFOCOM*, pages 1154–1162, Piscataway, NJ, USA, 2010.

[18] M. Mitzenmacher. *The power of two choices in randomized load balancing*. PhD thesis, UC Berkeley, 1996.

[19] M. Mitzenmacher. Studying balanced allocations with differential equations. *Combinatorics, Probability and Computing*, 8(5):473–482, Sept. 1999.

[20] A. Mukhopadhyay and R. R. Mazumdar. Analysis of load balancing in large heterogeneous processor sharing systems. ArXiv preprint arXiv:1311.5806, 2013.

[21] A. Mukhopadhyay, R. R. Mazumdar, and F. Guillemin. Static versus dynamic user assignment to cloud resources. 2015.

[22] J. W. Roberts. A service system with heterogeneous user requirement. In G. Pujolle, editor, *Performance of Data Communications Systems and Their Applications*, 1981.

[23] R. Srikant and W. Whitt. Simulation run lengths to estimate blocking probabilities. *ACM Trans. Model. Comput. Simul.*, 6(1):7–52, Jan. 1996.

[24] A. L. Stolyar. An infinite server system with general packing constraints. ArXiv preprint arXiv:1205.4271, 2012.

[25] A. L. Stolyar and Y. Zhong. An infinite server system with general packing constraints: Asymptotic optimality of a greedy randomized algorithm. In *Proc. 53th Annu. Allerton Conf. Commun., Control Comput.*, pages 575–582, Oct 2013.

[26] A. L. Stolyar and Y. Zhong. A large-scale service system with packing constraints: Minimizing the number of occupied servers. *SIGMETRICS Perform. Eval. Rev.*, 41(1):41–52, June 2013.

[27] J. N. Tsitsiklis and K. Xu. On the power of (even a little) resource pooling. *Stochastic Systems*, 2(1):1–66, 2012.

[28] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Probl. Peredachi Inf.*, 32(1):20–34, 1996.

[29] L. Wang, F. Zhang, A. V. Vasilakos, C. Hou, and Z. Liu. Joint virtual machine assignment and traffic engineering for green data center networks. *SIGMETRICS Perform. Eval. Rev.*, 41(3):107–112, Jan. 2014.

[30] W. Whitt. Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal*, 63(5):689–708, 1984.

[31] L. Ying, R. Srikant, and X. Kang. The power of slightly more than one sample in randomized load balancing. In *Proc. of IEEE INFOCOM*, 2015.

# Appendix

## Proof of Lemma 1

The case $\lambda = 0$ is trivial with a unique stationary solution $\boldsymbol{\pi} = (1, \underbrace{0, 0, ..., 0}_{B})$. We focus on the case $\lambda > 0$.

**Existence**: For ease of exposition, throughout the proof we define $x_{B+1} = 0$ for any $\mathbf{x} \in \mathcal{S}$.

**Step 1.** Define $\mathbf{G}(\mathbf{x}),\ \forall \mathbf{x} \in \mathcal{S}$.

For any $\mathbf{x} \in \mathcal{S}$, let $\mathbf{G}(\mathbf{x}) = (G_0(\mathbf{x}), G_1(\mathbf{x}), ..., G_B(\mathbf{x}))$, where $G_0(\mathbf{x}) = 1$, and $\forall k = 1, 2, ..., B$, $G_k(\mathbf{x}) \geq 0$ satisfies

$$\lambda G_k^d(\mathbf{x}) + kG_k(\mathbf{x}) - \lambda x_{k-1}^d - kx_{k+1} = 0. \tag{25}$$

We will show that $\mathbf{G}$ is uniquely determined by $\mathbf{x}$. Consider a sequence of functions $\{H_k(y_k)\}_{k=1}^B$, where

$$H_k(y_k) = \lambda y_k^d + ky_k - \lambda x_{k-1}^d - kx_{k+1}.$$

Since $\mathbf{x} \in \mathcal{S}$,

$$\begin{aligned} H_k(x_{k-1}) &= kx_{k-1} - kx_{k+1} \geq 0, \\ H_k(x_{k+1}) &= \lambda x_{k+1}^d - \lambda x_{k-1}^d \leq 0. \end{aligned}$$

Note that $H_k(y_k)$ is strictly increasing in $y_k \in [0, \infty)$. Hence there exists a unique $y_k^* > 0$ such that $H_k(y_k^*) = 0$. By the definition of $G_k$ in (25), $H_k(G_k) = 0$. Hence $G_k = y_k^*$ is determined by $\mathbf{x}$ uniquely, and

$$x_{k+1} \leq G_k(\mathbf{x}) \leq x_{k-1}. \tag{26}$$

**Step 2.** Show that $\mathbf{G}(\cdot)$ is mapping $\mathcal{S}$ into $\mathcal{S}$.

We will verify that $\forall \mathbf{x} \in \mathcal{S}$, $\mathbf{G}(\mathbf{x}) \in \mathcal{S}$, i.e., $1 = G_0(\mathbf{x}) \geq G_1(\mathbf{x}) \geq ... \geq G_B(\mathbf{x}) \geq 0$. For any $\mathbf{x} \in \mathcal{S}$, inequality in (26) ensures that $G_k \in [0, 1]$ for all $k$. To prove that $G_k \geq G_{k+1}$, consider a function

$$\varphi_k(z) = \lambda z^d + kz,$$

which is strictly increasing in $[0, 1]$. Hence it is sufficient to show that $\varphi_k(G_k) \geq \varphi_k(G_{k+1})$.

$$\begin{aligned} & \varphi_k(G_k) - \varphi_k(G_{k+1}) \\ = \ & \lambda G_k^d + kG_k - \lambda G_{k+1}^d - (k+1)G_{k+1} + G_{k+1} \\ \overset{(a)}{=} \ & \lambda x_{k-1}^d + kx_{k+1} - \lambda x_k^d - (k+1)x_{k+2} + G_{k+1} \\ = \ & \lambda(x_{k-1}^d - x_k^d) + k(x_{k+1} - x_{k+2}) + G_{k+1} - x_{k+2} \\ \overset{(b)}{\geq} \ & G_{k+1} - \pi_{k+2} \\ \overset{(c)}{\geq} \ & 0, \end{aligned}$$

where the equality $(a)$ comes from the definition of $G_k, G_{k+1}$ in (25), and the inequality $(b)$ follows by the fact that $\mathbf{x} \in \mathcal{S}$, and the inequality $(c)$ results from the property of $G_k$ in (26).

Therefore $\mathbf{G}(\mathbf{x}) \in \mathcal{S}$.

**Step 3.** Show that $\mathbf{G}(\cdot)$ is continuous.

Consider any point $\mathbf{x} \in \mathcal{S}$. For every $\epsilon > 0$, set $\delta = \frac{\epsilon}{\lambda d + 1}$. Let $\mathbf{y}$ be any point in $\mathcal{S}$ such that $|\mathbf{x} - \mathbf{y}| < \delta$. By the definition of $\mathbf{G}(\cdot)$, $\forall k = 1, 2, ..., B$,

$$\begin{aligned} & \lambda(G_k^d(\mathbf{x}) - G_k^d(\mathbf{y})) + k(G_k(\mathbf{x}) - G_k(\mathbf{y})) \\ = \ & (G_k(\mathbf{x}) - G_k(\mathbf{y})) \left( \lambda \sum_{i=0}^{d-1} G_k^{d-1-i}(\mathbf{x}) G_k^i(\mathbf{y}) + k \right) \\ = \ & \lambda(x_{k-1}^d - y_{k-1}^d) + k(x_{k+1} - y_{k+1}) \\ = \ & \lambda(x_{k-1} - y_{k-1}) \left( \sum_{i=0}^{d-1} x_{k-1}^{d-1-i} y_{k-1}^i \right) + k(x_{k+1} - y_{k+1}) \end{aligned}$$

Then we have

$$\begin{aligned} & |G_k(\mathbf{x}) - G_k(\mathbf{y})| \\ = \ & \frac{|\lambda(x_{k-1} - y_{k-1}) \left( \sum_{i=0}^{d-1} x_{k-1}^{d-1-i} y_{k-1}^i \right) + k(x_{k+1} - y_{k+1})|}{\lambda \sum_{i=0}^{d-1} G_k^{d-1-i}(\mathbf{x}) G_k^i(\mathbf{y}) + k} \\ \leq \ & \frac{\lambda d |x_{k-1} - y_{k-1}| + k|(x_{k+1} - y_{k+1})|}{k} \\ \leq \ & \lambda d |x_{k-1} - y_{k-1}| + |(x_{k+1} - y_{k+1})|, \end{aligned}$$

which implies that

$$\begin{aligned} & |\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{y})| \\ = \ & \sum_{k=0}^{B} |G_k(\mathbf{x}) - G_k(\mathbf{y})| \\ \leq \ & \sum_{k=1}^{B} (\lambda d |x_{k-1} - y_{k-1}| + |(x_{k+1} - y_{k+1})|) \\ \leq \ & (\lambda d + 1) \sum_{k=0}^{B} |x_k - y_k| \\ < \ & (\lambda d + 1)\delta \\ = \ & \epsilon. \end{aligned}$$

Therefore $\mathbf{G}$ is continuous at any point $\mathbf{x} \in \mathcal{S}$.

**Step 4.** Show that a fixed point of $\mathbf{G}$ in $\mathcal{S}$ is a stationary point.

Note that set $\mathbf{S}$ is compact and convex. Step 1-3 ensures that there exists a fixed point of $\mathbf{G}$ in $\mathcal{S}$, denoted by $\hat{\mathbf{x}}$. That is, $\hat{\mathbf{x}} = \mathbf{G}(\hat{\mathbf{x}})$. From the definition of $\mathbf{G}$ in (25), we have

$$F_k(\hat{\mathbf{x}}) = \lambda \hat{x}_k^d + k\hat{x}_k - \lambda \hat{x}_{k-1}^d - k\hat{x}_{k+1} = 0.$$

That is, $\hat{\mathbf{x}}$ is a stationary point.

**Uniqueness**: We prove the uniqueness of stationary solution by contradiction.

Assume that there exists two different solutions $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$. We claim that $\pi_B \neq \hat{\pi}_B$. Otherwise, we have

$$\pi_{B-1} = \sqrt[d]{\pi_B^d + \frac{B}{\lambda}\pi_B} = \hat{\pi}_{B-1}.$$

Note that

$$\pi_k = \sqrt[d]{\pi_{k+1}^d + \frac{k+1}{\lambda}(\pi_{k+1} - \pi_{k+2})}.$$

Hence by induction, we can show that $\pi_k = \hat{\pi}_k$ for any $k = 0, 1, ..., B$.

Consider the case that $\pi_B < \hat{\pi}_B$. Similarly, we can establish that $\pi_k < \hat{\pi}_k$ for any $k = 0, 1, ..., B$ by induction. Therefore, $\pi_0 < \hat{\pi}_0$, which contradicts with the fact that $\pi_0 = \hat{\pi}_0 = 1$. This completes the proof for the uniqueness. ■

## Proof of Lemma 3

Due to continuous dependence of a solution on the initial values, it is sufficient to show that if $\bar{s}_k^0 < s_k^0$ for any $k \geq 1$,

$\bar{s}_k(t) \le s_k(t)$ for all $t \ge 0$ and any $k$. Assume that strict inequalities hold for $t < t_1$ and are broken at $t = t_1$. Consider two cases:

(i) $\bar{s}_k(t_1) = s_k(t_1)$ for any $k$.

The uniqueness of solution ensures that $\bar{s}_k(t) = s_k(t)$ for all $t \ge t_1$ and any $k$. Hence the claim holds.

(ii) $\exists k^* \ge 1$ such that $\bar{s}_{k^*}(t_1) < s_{k^*}(t_1)$.

Then there exists $k \ge 1$ such that $\bar{s}_k(t_1) = s_k(t_1)$, and at least of one following conditions hold: $\bar{s}_{k-1}(t_1) < s_{k-1}(t_1)$, $\bar{s}_{k+1}(t_1) < s_{k+1}(t_1)$. If $k < B$, we have

$$\begin{aligned}\frac{d\bar{s}_k}{dt}(t_1) - \frac{ds_k}{dt}(t_1) &= \lambda(\bar{s}^d_{k-1} - s^d_{k-1}) + k(\bar{s}_{k+1} - s_{k+1}) \\ &\quad -\lambda(\bar{s}^d_k - s^d_k) - k(\bar{s}_k - s_k) \\ &< 0,\end{aligned}$$

where the inequality comes from the definition of $k$. Similarly, we can verify that $\frac{d\bar{s}_k}{dt}(t_1) - \frac{ds_k}{dt}(t_1) < 0$ if $k = B$.

Since $\bar{\mathbf{s}}(t)$ and $\mathbf{s}(t)$ are continuous functions of $t$, there exists $t_0 < t_1$ such that $\bar{s}_k(t_0) < s_k(t_0)$ and

$$\frac{d\bar{s}_k}{dt}(t) - \frac{ds_k}{dt}(t) < 0$$

for any $t \in (t_0, t_1)$. Thus

$$\bar{s}_k(t_1) - s_k(t_1) = \bar{s}_k(t_0) - s_k(t_0) + \int_{t_0}^{t_1} \left( \frac{d\bar{s}_k}{dt}(t) - \frac{ds_k}{dt}(t) \right) dt < 0,$$

which contradicts with the assumption that $\bar{s}_k(t_1) = s_k(t_1)$. ■

## Proof for Lemma 4

We will show that $d\psi(t)/dt \le -\psi$. Then $\psi(t) \le \psi(0)e^{-t}$, which implies that $\psi(t)$ converges to 0 exponentially fast.

Consider the case where $s^0_k \ge \pi_k$ for any $k$. From Lemma 4, $s_k(t) \ge \pi_k$ for any $t \ge 0, \forall k \in \{0, 1, ..., B\}$. We can rewrite $\psi(t)$ as $\psi(t) = \sum_{k=0}^{B}(s_k(t) - \pi_k)$. Since $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$, $\dot{\mathbf{s}} = \mathbf{F}(\mathbf{s})$, we have

$$\begin{aligned}\frac{d\psi(t)}{dt} &= \sum_{k=0}^{B} \frac{ds_k(t)}{dt} \\ &= \sum_{k=1}^{B} F_k(\mathbf{s}(t)) - \sum_{k=1}^{B} F_k(\boldsymbol{\pi}) \\ &= \left( \lambda(s^d_0(t) - s^d_B(t)) - \sum_{k=1}^{B} s_k(t) \right) \\ &\quad - \left( \lambda(\pi^d_0 - \pi^d_B) - \sum_{k=1}^{B} \pi_k \right) \\ &= -\lambda(s^d_B(t) - \pi^d_B) - \psi(t) \\ &\le -\psi(t),\end{aligned}$$

where the last inequality follows by the fact that $s^d_B(t) \ge \pi^d_B$.

The other case where $s^0_k \le \pi_k$ for any $k$ can be proved similarly. ■

## Proof of Lemma 6

Since $\mathbf{x}, \mathbf{y} \in \mathcal{S}$, for any $0 \le k \le B$

$$0 \le x_k \le 1,\ 0 \le y_k \le 1.$$

Then we have:

$$\begin{aligned}&|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \\ =\ & \sum_{k=1}^{B-1} |\lambda(x^d_{k-1} - x^d_k) - k(x_k - x_{k+1}) \\ & -\lambda(y^d_{k-1} - y^d_k) + k(y_k - y_{k+1})| \\ & +|\lambda(x^d_{B-1} - x^d_B) - Bx_B - \lambda(y^d_{B-1} - y^d_B) + By_B| \\ \le\ & 2\sum_{k=0}^{B} k|x_k - y_k| + 2\sum_{k=0}^{B} \lambda|x^d_k - y^d_k| \\ \le\ & 2B\sum_{k=0}^{B} |x_k - y_k| + 2\lambda \sum_{k=0}^{B} \left( |x_k - y_k| \sum_{i=0}^{d-1} x^{d-1-i}_k y^i_k \right) \\ \le\ & 2(B + d\lambda)\sum_{k=0}^{B} |x_k - y_k| \\ =\ & M|\mathbf{x} - \mathbf{y}|,\end{aligned}$$

where $M = 2(B + d\lambda)$. ■

## Proof of Claim 1

From Lemma 5, we have

$$\mathbf{S}^{(N_k)}(t) \Rightarrow \bar{\mathbf{S}}(t) \text{ as } k \to \infty.$$

By the definition of weak convergence, for a bounded continuous function $f$,

$$\mathbb{E}\left[ f(\mathbf{S}^{(N_k)}(t)) | \mathbf{S}^{(N_k)}(0) \right] \overset{n\to\infty}{\longrightarrow} \mathbb{E}\left[ f(\bar{\mathbf{S}}(t)) | \bar{\mathbf{S}}(0) \right],$$

if $\mathbf{S}^{(N_k)}(0) \to \bar{\mathbf{S}}(0)$ as $k \to \infty$.

As $\mathbf{S}^{(N_k)}(0) = \mathbf{X}^{(N_k)}$ and $\bar{\mathbf{S}}(0) = \bar{\mathbf{X}}$, by Skorokhod's representation theorem,

$$\mathbf{S}^{(N_k)}(0) \to \bar{\mathbf{S}}(0).$$

Define

$$\mathbf{Y}_k = \mathbb{E}\left[ f(\mathbf{S}^{(N_k)}(t)) | \mathbf{X}^{(N_k)} \right], \quad \mathbf{Y} = \mathbb{E}\left[ f(\bar{\mathbf{S}}(t)) | \bar{\mathbf{X}} \right].$$

Since $f$ is bounded, $\mathbf{Y}_k$ and $\mathbf{Y}$ are bounded. By the bounded convergence theorem, we have

$$\mathbb{E}[\mathbf{Y}_k] \to \mathbb{E}[\mathbf{Y}].$$

This holds for all bounded, continuous $f$. Thus again by the definition of weak convergence,

$$\mathbf{S}^{(N_k)}(t) \Rightarrow \bar{\mathbf{S}}(t) \text{ as } k \to \infty$$

. ■