

DECAY OF TAILS AT EQUILIBRIUM FOR FIFO JOIN THE SHORTEST QUEUE NETWORKS

BY MAURY BRAMSON* YI LU BALAJI PRABHAKAR †

University of Minnesota, Twin Cities

University of Illinois, Champaign-Urbana

Stanford University

In join the shortest queue networks, incoming jobs are assigned to the shortest queue from among a randomly chosen subset of D queues, in a system of N queues; after completion of service at its queue, a job leaves the network. We also assume that jobs arrive into the system according to a rate- αN Poisson process, $\alpha < 1$, with rate-1 service at each queue. When the service at queues is exponentially distributed, it was shown in Vvedenskaya et al. [16] that the tail of the equilibrium queue size decays doubly exponentially in the limit as $N \rightarrow \infty$. This is a substantial improvement over the case $D = 1$, where the queue size decays exponentially.

The reasoning in [16] does not easily generalize to jobs with non-exponential service time distributions. A modularized program for treating general service time distributions was introduced in Bramson et al. [4]. The program relies on an ansatz that asserts, in equilibrium, any fixed number of queues become independent of one another as $N \rightarrow \infty$. This ansatz was demonstrated in several settings in Bramson et al. [5], including for networks where the service discipline is FIFO and the service time distribution has a decreasing hazard rate.

In this article, we investigate the limiting behavior, as $N \rightarrow \infty$, of the equilibrium at a queue when the service discipline is FIFO and the service time distribution has a power law with a given exponent $-\beta$, for $\beta > 1$. We show under the above ansatz that, as $N \rightarrow \infty$, the tail of the equilibrium queue size exhibits a wide range of behavior depending on the relationship between β and D . In particular, if $\beta > D/(D-1)$, the tail is doubly exponential and, if $\beta < D/(D-1)$, the tail has a power law. When $\beta = D/(D-1)$, the tail is exponentially distributed.

1. Introduction. We consider *join the shortest queue* (JSQ) networks, where incoming “jobs” (or “customers”) are assigned to the shortest queue

*Supported in part by NSF Grant CCF-0729537

†Supported in part by NSF Grant CCF-0729537 and by a grant from the Clean Slate Program at Stanford University

AMS 2000 subject classifications: 60K25, 68M20, 90B15

Keywords and phrases: Join the shortest queue, FIFO, decay of tails.

from among D distinct queues, $D \geq 2$, with these queues being chosen uniformly from among the N queues in the system, with $D \leq N$. When two or more of these queues each have the fewest number of jobs, each of the queues is chosen with equal probability. After completion of service at its queue, a job leaves the network. We assume that jobs arrive according to a rate- αN Poisson process, $\alpha < 1$, and that jobs are served independently and at rate 1 at each queue. We are interested in this article in the case where the service discipline at each queue is first-in, first-out (FIFO).

When the service at queues is exponentially distributed, the evolution of the system is given by a countable state Markov chain where a state is given by the number of jobs at each queue. It is not difficult to show that a unique equilibrium distribution exists; this equilibrium is exchangeable with respect to the ordering of the queues. Let $P_k^{(N)}$ denote the probability that there are at least k jobs in equilibrium for the system with N queues. It was shown in Vvedenskaya et al. [16] that

$$(1.1) \quad \lim_{N \rightarrow \infty} P_k^{(N)} = \alpha^{(D^k - 1)/(D - 1)} \quad \text{for } k \in \mathbb{Z}_+;$$

in particular, the right tail of $P_k^{(N)}$ decays doubly exponentially fast in the limit as $N \rightarrow \infty$. This behavior is a substantial improvement over the case $D = 1$, where $P_k^{(N)}$ decays exponentially, and has led to substantial interest in JSQ networks in the literature. For other references, see Azar et al. [1], Graham [8], Luczak-McDiarmid [9],[10], Martin-Suhov [11], Mitzenmacher [12], Suhov-Vvedenskaya [14], Vocking [15] and Vvedenskaya-Suhov [17].

Little work has been done on the behavior of JSQ networks when the service times are not exponentially distributed. In this setting, the underlying Markov process will typically have an uncountable state space, and positive Harris recurrence for the process is no longer obvious. The latter was shown in Foss-Chernova [7], and uniform bounds on the equilibria were shown in Bramson [3]. (Both articles also considered JSQ networks with more general arrivals and routing of jobs.)

This paper builds on previous work [3], [4] and [5] by the authors. Bramson et al. [4] described a modularized program for analyzing the limiting behavior of the equilibria of a family of JSQ networks with general service times, as $N \rightarrow \infty$. An important step is to show that any fixed number of queues become independent of one another, with each converging to a limiting distribution that is the equilibrium for an associated Markov process with a single queue, which is a *cavity process*. This process corresponds, in an appropriate sense, to “setting $N = \infty$ ” in the JSQ network and viewing the corresponding infinite dimensional process at a single queue. We will

refer to this equilibrium as the *equilibrium environment*. In Section 2, we will precisely define this terminology.

Although it seems that this independence should hold in a very general setting, including under a wide range of service disciplines, demonstrating it appears to be a difficult problem. In Bramson et al. [4], this independence and convergence to the equilibrium environment were stated as an ansatz. This ansatz was demonstrated in Bramson et al. [5] in several settings including for networks where the service discipline is FIFO and the service distribution has a decreasing hazard rate.

In this article, we employ the restriction of the above ansatz to FIFO networks. This version of the ansatz will be precisely stated in Section 2. Here, we summarize it for application in the current section:

For a family of networks with the FIFO service discipline that are all in equilibrium, any fixed number of queues become

(1.2) independent in the limit as $N \rightarrow \infty$. Moreover, each marginal distribution converges to the unique associated equilibrium environment.

Although this ansatz has only been demonstrated for service distributions having decreasing hazard rate and for general service distributions when the arrival rate α is sufficiently small, our arguments here do not otherwise require either restriction. Other applications of the ansatz, but for the processor sharing and LIFO service disciplines, are given in [4].

Our goal, in this article, is to investigate the limiting behavior of the right tail of the associated equilibrium environment, under the FIFO service discipline and with the assigned mean-1 service distribution $F(\cdot)$. Denote by P_k the probability that there are at least k jobs in the equilibrium environment. We will show that, when $F(\cdot)$ has a power law tail with exponent $-\beta$, for given $\beta > 1$, the tail of P_k exhibits a wide range of behavior depending on the relationship between β and D . In particular, if $\beta > D/(D-1)$, the tail is doubly exponential and, if $\beta < D/(D-1)$, the tail has a power law; when $\beta = D/(D-1)$, the tail is exponentially distributed. When $\beta \nearrow \infty$, the coefficient $q_D(\beta)$ of k in the doubly exponential tail converges to 1, which is the coefficient of k in (1.1). One obtains the same coefficient of k whether $F(\cdot)$ has an exponential tail or has bounded support. Our main results are Theorems 1.1, 1.2 and 1.3. Theorem 1.1 covers the case $\beta > D/(D-1)$, Theorem 1.2 covers the case $\beta < D/(D-1)$ and Theorem 1.3 covers the case $\beta = D/(D-1)$. We set $\bar{F}(s) = 1 - F(s)$.

THEOREM 1.1. *Consider a family of JSQ networks, with given $D \geq 2$ and $N = D, D + 1, \dots$, where the N th network has Poisson rate- αN input, with $\alpha < 1$, and where service at each queue is FIFO, with distribution $F(\cdot)$ having mean 1. Assume that (1.2) holds and that*

$$(1.3) \quad \lim_{s \rightarrow \infty} \log \bar{F}(s) / \log s = -\beta,$$

with $\beta \in (D/(D-1), \infty)$. Then,

$$(1.4) \quad \lim_{k \rightarrow \infty} (1/k) \log_D \log(1/P_k) = q_D(\beta)$$

for some $q_D(\beta) \in (0, 1)$. Moreover, $q_D(\beta)$ is continuous in β and

$$(1.5) \quad q_D(\beta) \nearrow 1 \quad \text{exponentially fast as } \beta \nearrow \infty.$$

When (1.3) holds with $\beta = \infty$, then (1.4) holds with $q_D(\infty) = 1$.

Theorem 1.1 implies that, when $\bar{F}(s) \sim cs^{-\beta}$ as $s \rightarrow \infty$, for $\beta \in (D/(D-1), \infty)$ and $c > 0$, then $P_k = \exp\{-D^{(1+o(1))q_D(\beta)k}\}$.

THEOREM 1.2. *Consider a family of JSQ networks as in Theorem 1.1, with (1.3) instead holding for $\beta \in (1, D/(D-1))$. Then*

$$(1.6) \quad \lim_{k \rightarrow \infty} \log(1/P_k) / \log k = (\beta - 1) / [1 - (D - 1)(\beta - 1)].$$

Theorem 1.2 implies that, when $\bar{F}(s) \sim cs^{-\beta}$ as $s \rightarrow \infty$, for $\beta \in (1, D/(D-1))$ and $c > 0$, then $P_k = k^{-(1+o(1))\gamma_D(\beta)}$, where $\gamma_D(\beta)$ is the right hand side of (1.6). Note that $\gamma_D(\beta) \searrow 0$ as $\beta \searrow 1$ and $\gamma_D(\beta) \nearrow \infty$ as $\beta \nearrow D/(D-1)$.

THEOREM 1.3. *Consider a family of JSQ networks as in Theorem 1.1, with (1.3) replaced by*

$$(1.7) \quad c_1 \leq \underline{\lim}_{s \rightarrow \infty} s^{D/(D-1)} \bar{F}(s) \leq \overline{\lim}_{s \rightarrow \infty} s^{D/(D-1)} \bar{F}(s) \leq c_2$$

for some $0 < c_1 \leq c_2 < \infty$. Then, for appropriate $r_D(c_2) > 0$ and $s_D(c_1) < \infty$,

$$(1.8) \quad r_D(c_2) \leq \underline{\lim}_{k \rightarrow \infty} (1/k) \log(1/P_k) \leq \overline{\lim}_{k \rightarrow \infty} (1/k) \log(1/P_k) \leq s_D(c_1),$$

where

$$(1.9) \quad \begin{array}{ll} r_D(c_2) \nearrow \infty & \text{as } c_2 \searrow 0, \\ s_D(c_1) \searrow 0 & \text{as } c_1 \nearrow \infty. \end{array}$$

Theorem 1.3 implies that when $\bar{F}(s) \sim cs^{-D/(D-1)}$ as $s \rightarrow \infty$, then P_k decreases exponentially fast in the sense of (1.8). Because of (1.9), the exponent depends strongly on the choice of c .

When $\bar{F}(\cdot)$ satisfies (1.3) for a given $\beta > 1$, the asymptotic behavior of P_k behaves according to (1.4) or (1.6), depending on whether $D > \beta/(\beta - 1)$ or $D < \beta/(\beta - 1)$. In applications where there is a substantial penalty for a moderately large number of jobs at a queue (resulting, for example, in memory overflow), it is therefore important to choose $D > \beta/(\beta - 1)$. This distinction does not occur when $\bar{F}(\cdot)$ has an exponential tail, since any choice of $D \geq 2$ produces a doubly exponential tail for P_k , as in (1.1). (See [4] for more detail.)

We point out that the proofs of Theorems 1.1-1.3 only depend on (1.2) for the existence of an equilibrium environment. Regardless of how the existence of an equilibrium environment is verified, (1.2) will be needed in order to relate the tail behavior of P_k for the equilibrium environment to the tail behavior for the equilibria of the corresponding family of networks as $N \rightarrow \infty$.

We also note that, although the phrase “join the shortest queue network” is widely used in the literature, such systems are not true networks in the sense that, upon the departure of a job from a queue, the job leaves the system instead of being able to return to a different queue. However, such systems have been extended to the setting of Jackson networks (see, e.g., [11] and [14]).

This article is organized as follows. In Section 2, we provide basic background on the properties of the state space and Markov process that underlie the JSQ networks. We then define equilibrium environments and formally state the ansatz. In Sections 3-5, we demonstrate Theorems 1.1, 1.2 and 1.3, respectively. Our approach will be to demonstrate lower bounds and then upper bounds that yield the theorem. In each case, the lower bounds will be considerably easier to show.

Notation. For the reader’s convenience, we mention here some of the notation in the paper. We will employ C_1, C_2, \dots to denote positive constants whose precise value is not of importance to us. For $z \in \mathbb{R}$, $\lfloor z \rfloor$ and $\lceil z \rceil$ will denote the integer part of z , respectively, the smallest integer at least as large as z .

2. Markov process background, equilibrium environments and the ansatz. In this section, we provide a more detailed description of the construction of the Markov processes $X^{(N)}(\cdot)$ that underlie the JSQ networks. We next define the corresponding cavity process and its equilibrium

environment. We then employ these concepts to state the ansatz for JSQ networks. Most of this material is included in Sections 2 and 3 of Bramson et al. [5]. (Related material is also given in [2] and [3].)

We define the state space $S^{(N)}$ to be the set

$$(2.1) \quad (\mathbb{Z} \times \mathbb{R}^2)^N.$$

The first coordinate z^n , $n = 1, \dots, N$, corresponds to the number of jobs at the n th queue; the second coordinate u^n , $u^n \geq 0$, is the amount of time the oldest job there has already been served; and the last coordinate s^n , $s^n > 0$, is the residual service time. When $z^n = 0$, set the other two coordinates equal to 0. The coordinate u^n will not play a role in the evolution of $X^{(N)}(\cdot)$ here; we retain it for comparison with [5], where it was used to demonstrate (1.2) under decreasing hazard rates. (We will employ slightly different notation here than in [5].)

For given $N' \leq N$, $S^{(N')}$ is the *projection* of $S^{(N)}$ obtained by restricting $S^{(N)}$ to the first N' queues; for $x \in S^{(N)}$, $x' \in S^{(N')}$ is thus obtained by omitting the coordinates with $n > N'$. One can also define projections of $S^{(N)}$ onto spaces $S^{(N')}$ corresponding to other subsets of $\{1, \dots, N\}$ analogously, although these are not needed here.

We define the metric $d^{(N)}(\cdot, \cdot)$ on $S^{(N)}$, with $d^{(N)}(\cdot, \cdot)$ given in terms of $d^{(N),n}(\cdot, \cdot)$ by $d^{(N)}(\cdot, \cdot) = (1/N) \sum_{n=1}^N d^{(N),n}(\cdot, \cdot)$. For given $x_1, x_2 \in S^{(N)}$, with the coordinates labelled correspondingly, set

$$(2.2) \quad d^{(N),n}(x_1, x_2) = |z_1^n - z_2^n| + |u_1^n - u_2^n| + |s_1^n - s_2^n|.$$

One can check that the metric $d^{(N)}(\cdot, \cdot)$ is separable and locally compact; more detail is given on page 82 of [2]. We equip $S^{(N)}$ with the standard Borel σ -algebra inherited from $d^{(N)}(\cdot, \cdot)$, which we denote by $\mathcal{S}^{(N)}$.

The Markov process $X^{(N)}(t)$, $t \geq 0$, underlying a given model is defined to be the right continuous process with left limits, taking values x in $S^{(N)}$, whose evolution is determined by the model together with the assigned service discipline. We denote the random values of the coordinates z^n , u^n and s^n taken by $X^{(N)}(t)$, by $Z^n(t)$, $U^n(t)$ and $S^n(t)$. Jobs are allocated service according to the FIFO discipline; during the period a job is being served, $U^n(t)$ increases at rate 1 and $S^n(t)$ decreases at rate 1.

Along the lines of page 85 of [2], a filtration $(\mathcal{F}_t^{(N)})$, $t \in [0, \infty]$, can be assigned to $X^{(N)}(\cdot)$ so that $X^{(N)}(\cdot)$ is a piecewise-deterministic Markov process, and hence is Borel right. This implies that $X^{(N)}(\cdot)$ is strong Markov. (We do not otherwise use Borel right.) The reader is referred to Davis [6] for more detail.

Equilibrium environments and the ansatz. In order to state the ansatz, we require some terminology. We denote by $\mathcal{E}^{(N,N')}$ the projection of the equilibrium measure $\mathcal{E}^{(N)}$ of the N -queue system onto the first N' queues. (Since $X^{(N)}(t)$ is exchangeable when $X^{(N)}(0)$ is, the choice of queues will not matter.)

We wish to describe the evolution of individual queues for the limiting process, as $N \rightarrow \infty$. For this, we construct a strong Markov process $X^{\mathcal{H}}(t)$, $t \geq 0$, on $S^{(1)}$. We will define $X^{\mathcal{H}}(t)$ similarly to $X^{(1)}(t)$, except that only a fraction of incoming potential arrivals at the queue is permitted to arrive at the queue, with the fraction depending on the current number of jobs there, and with the fraction decreasing as the number of jobs increases.

We proceed as follows. Let \mathcal{H} denote a probability measure on $S^{(1)}$, which we refer to as the *environment* of the process $X^{\mathcal{H}}(\cdot)$; we refer to $X^{\mathcal{H}}(\cdot)$ as the associated *cavity process*. We define $X^{\mathcal{H}}(\cdot)$ so that *potential arrivals* arrive according to a rate- $D\alpha$ Poisson process. When such a potential arrival to the queue occurs at time t , $X^{\mathcal{H}}(t-)$ is compared with the states of $D - 1$ independent random variables, each with law \mathcal{H} ; we refer to these $D - 1$ states at a potential arrival as the *comparison states*. Choosing from among these D states, the job is assigned to the state with the fewest number of jobs. (In case of a tie, each of these states is chosen with equal probability.) If the job has chosen the state $X^{\mathcal{H}}(t-)$ at the queue, it then immediately joins the queue; otherwise, the job immediately leaves the system. In either case, the independent $D - 1$ states employed for this purpose are immediately discarded.

We give the following illustrations, denoting by Q_k the probability that the environment \mathcal{H} has at least k jobs. For $D = 2$, if a potential arrival occurs at time t and $X^{\mathcal{H}}(t-) = k$, then the probability that $X^{\mathcal{H}}(t) = k + 1$ is $(Q_k + Q_{k+1})/2$, and so the rate α_k of an arrival at the queue is $\alpha(Q_k + Q_{k+1})$. For general D , in order for a potential arrival to arrive at the queue, it is necessary for all of the $D - 1$ comparison states used at that time to be at least k , in which case the probability of selecting the queue is the reciprocal of the number of states equal to k . This gives the bounds

$$(2.3) \quad \alpha Q_k^{D-1} \leq \alpha_k \leq \alpha D Q_k^{D-1}.$$

We assume that jobs in the cavity process $X^{\mathcal{H}}(\cdot)$ have the same service distribution $F(\cdot)$ as in the queueing network and are served according to the FIFO service discipline. The number of jobs in $X^{\mathcal{H}}(t)$ will be denoted by $Z^{\mathcal{H}}(t)$, the amount of time the oldest job has already been served by $U^{\mathcal{H}}(\cdot)$ and the residual service time by $S^{\mathcal{H}}(t)$; we will employ x , z , u and s for the corresponding terms in the state space.

When a cavity process $X^{\mathcal{H}}(\cdot)$, with environment \mathcal{H} , is stationary with the equilibrium measure \mathcal{H} (i.e., $X^{\mathcal{H}}(t)$ has the distribution \mathcal{H} for all t), we say that \mathcal{H} is an *equilibrium environment*. One can think of an equilibrium environment as being the restriction of an equilibrium measure for the JSQ network, viewed at a single queue, when “the total number of queues N is infinite”. More background on the cavity process is given in [4].

We now state the ansatz. Here, \xrightarrow{v} on $S^{(N')}$ denotes convergence in total variation with respect to the metric $d^{N'}(\cdot, \cdot)$ on $S^{(N')}$.

ANSATZ . *Consider a family of JSQ networks, with given $D \geq 2$ and $N = D, D + 1, \dots$, where the N th network has Poisson rate- αN input, with $\alpha < 1$, and where service at each queue is FIFO, with distribution $F(\cdot)$ having mean 1. Then, (a) for each N' ,*

$$(2.4) \quad \mathcal{E}^{(N, N')} \xrightarrow{v} \mathcal{E}^{(\infty, N')} \quad \text{as } N \rightarrow \infty,$$

where $\mathcal{E}^{(\infty, N')}$ is the N' -fold product of $\mathcal{E}^{(\infty, 1)}$. Moreover, (b) $\mathcal{E}^{(\infty, 1)}$ is the unique equilibrium environment associated with this family of networks.

As was mentioned in the introduction, this ansatz was demonstrated in Bramson et al. [5] when the service time distribution $F(\cdot)$ has a decreasing hazard rate $h(\cdot)$ (i.e., $h(s) = F'(s)/\bar{F}(s)$ is nonincreasing in s) and for general service distributions when the arrival rates are small enough.

In order to demonstrate Theorems 1.1-1.3, we will analyze the cavity process $X^{\mathcal{H}}(\cdot)$ with its unique equilibrium environment $\mathcal{H} = \mathcal{E}^{(\infty, 1)}$. In particular, we will analyze $\mathcal{E}^{(\infty, 1)}$ over a *cycle* starting and ending at the state 0. (The state where the number of jobs z is 0.) Letting ν denote the time at which $X^{\mathcal{H}}(\cdot)$ first returns to 0 after visiting another state, the first cycle is the random time interval $[0, \nu]$. For any $k \geq 1$, we will denote by V_k the *occupation time* at states x , with $z \geq k$, over $[0, \nu]$, that is,

$$V_k = \int_0^\nu \mathbf{1}\{Z^{\mathcal{H}}(t) \geq k\} dt.$$

Setting $m_0 = E[\nu]$, the mean return time to 0, one has

$$(2.5) \quad P_k = m_0^{-1} E[V_k],$$

where P_k is the probability there are at least k jobs in the equilibrium environment.

Letting α_k denote the arrival rate of jobs for $X^{\mathcal{H}}(\cdot)$ when $z = k$, one has

$$(2.6) \quad \alpha P_k^{D-1} \leq \alpha_k \leq \alpha D P_k^{D-1},$$

which is the analog of (2.3). Since the departure of jobs from the queue is deterministic, being a function of the residual service time s , (2.6) gives a reasonably explicit description of the transition rates for $X^{\mathcal{H}}(\cdot)$. Together with (2.5), (2.6) will provide the basis for our demonstration of Theorems 1.1-1.3 and will be used throughout the paper.

3. The case where $\beta > D/(D - 1)$. In this section, we demonstrate Theorem 1.1; we do this by demonstrating lower and upper bounds that are needed for the theorem in Propositions 3.1 and 3.2. Each of these bounds is expressed in terms of a recursion relation for P_k . In order to obtain Theorem 1.1 from these recursions, we employ Proposition 3.3, which analyzes such recursions by utilizing a standard framework involving rational generating functions. The section is organized as follows. After stating Propositions 3.1 and 3.2, we state and prove Proposition 3.3. We next employ the three propositions to demonstrate Theorem 1.1. We then provide the relatively quick proof of Proposition 3.1 and the longer proof of Proposition 3.2, in the following subsections.

In both propositions, we set $k_1 = \lceil k - \beta \rceil$ (or, equivalently, $\lfloor \beta \rfloor = k - k_1$) and $\hat{\beta} = \beta - \lfloor \beta \rfloor$.

PROPOSITION 3.1. *Consider a family of JSQ networks, with given $D \geq 2$ and $N = D, D + 1, \dots$, where the N th network has Poisson rate- αN input, with $\alpha < 1$, and where service at each queue is FIFO, with distribution $F(\cdot)$ having mean 1. Assume that (1.2) holds. Then, for appropriate $C_1 > 0$ and all k ,*

$$(3.1) \quad P_k \geq (C_1/8k)^k \prod_{i=0}^{k-1} P_i^{D-1}.$$

If moreover, for some $s_0 \geq 1$,

$$(3.2) \quad \bar{F}(s) \geq s^{-\beta} \quad \text{for } s \geq s_0,$$

with $\beta \in (D/(D - 1), \infty)$, then, for appropriate $C_1 > 0$ and all k ,

$$(3.3) \quad P_k \geq C_1 3^{-k} \left(\prod_{i=k_1+1}^{k-1} P_i^{D-1} \right) P_{k_1}^{\hat{\beta}(D-1)}.$$

PROPOSITION 3.2. *Consider a family of JSQ networks, with given $D \geq 2$ and $N = D, D + 1, \dots$, where the N th network has Poisson rate- αN input,*

with $\alpha < 1$, and where service at each queue is FIFO, with distribution $F(\cdot)$ having mean 1. Assume that (1.2) holds and that, for some $s_0 \geq 1$,

$$(3.4) \quad \bar{F}(s) \leq s^{-\beta} \quad \text{for } s \geq s_0,$$

with $\beta \in (D/(D-1), \infty)$. If β is not an integer, then, for appropriate C_2 and all k ,

$$(3.5) \quad P_k \leq C_2 k^{\beta+1} \left(\prod_{i=k_1+1}^{k-1} P_i^{D-1} \right) P_{k_1}^{\hat{\beta}(D-1)}.$$

If β is an integer, then, for each $\delta > 0$, appropriate C_2 and all k ,

$$(3.6) \quad P_k \leq C_2 k^{\beta+1} \left(\prod_{i=k_1+2}^{k-1} P_i^{D-1} \right) P_{k_1+1}^{(1-\delta)(D-1)}.$$

To employ the recursions in (3.3) and (3.5)–(3.6) of Propositions 3.1 and 3.2 in the proof of Theorem 1.1, we will analyze the asymptotic behavior of the recursions in (3.7).

PROPOSITION 3.3. *Suppose that R_k satisfies*

$$(3.7) \quad R_k = (D-1) \left(\sum_{i=k-\ell+1}^{k-1} R_i + \eta R_{k-\ell} \right) \quad \text{for } k \geq 1,$$

with $R_k = 1$ for $k = -\ell + 1, \dots, -1, 0$, where $\ell, D \geq 2$ and $\eta \in [0, 1]$. Then, setting $\beta = \ell + \eta - 1$,

$$(3.8) \quad \lim_{k \rightarrow \infty} \frac{1}{k} \log_D R_k = q_D(\beta)$$

for some $q_D(\beta) \in (0, 1)$. Moreover, $q_D(\beta)$ is continuous in β and $q_D(\beta) \nearrow 1$ exponentially fast as $\beta \nearrow \infty$.

PROOF. The recurrence (3.7) is a special case of linear recursions of the form

$$(3.9) \quad R_k + \sum_{i=1}^{\ell} a_i R_{k-i} = 0,$$

with $a_i \in \mathbb{C}$ and general $R_{-\ell+1}, \dots, R_0$. It is well known that (see, e.g., Stanley [13], page 202)

$$(3.10) \quad R_k = \sum_{i=1}^j P_i(k) \gamma_i^k$$

for each k , where γ_i are distinct, $P_i(k)$ is a polynomial in k of degree strictly less than ℓ_i , and

$$(3.11) \quad 1 + \sum_{i=1}^{\ell} a_i x^i = \prod_{i=1}^j (1 - \gamma_i x)^{\ell_i},$$

with $\sum_{i=1}^j \ell_i = \ell$. Moreover the converse holds, that is, if (3.10) and (3.11) both hold, then so does (3.9).

For R_k given by (3.7), it is not difficult to check that there is exactly one value γ_i , say γ_1 , that is real and positive, that γ_1 varies continuously in η , and moreover that γ_1 satisfies $\gamma_1 > 1$, since $a_i < 0$ and $\sum_{i=1}^{\ell} a_i < -1$. (Descartes' rule of signs in fact implies that $1/\gamma_1$ is a simple root.) Also, because $a_i < 0$, and possesses both odd and even indices, $|\gamma_i| < \gamma_1$ for $i \neq 1$. Since the initial data given below (3.7) are all positive, any solution of (3.7) is majorized by this particular solution, up to a multiplicative constant; so, $P_1(\cdot) \not\equiv 0$. The limit in (3.8), with $q_D(\beta) = \log_D \gamma_1 > 0$, follows from these observations.

We still need to examine the limiting behavior of $q_D(\beta)$ as $\beta \rightarrow \infty$. Dividing both sides in (3.7) by R_k , then substituting (3.10) for each of the terms, and letting $k \rightarrow \infty$ implies that

$$\begin{aligned} 1 &= (D-1) \left(x + x^2 + \dots + x^{\ell-1} + \eta x^{\ell} \right) \\ &= (D-1) \left(x - (1-\eta)x^{\ell} - \eta x^{\ell+1} \right) / (1-x) \end{aligned}$$

for $x = 1/\gamma_1 = D^{-q_D(\beta)}$. This again uses $\gamma_1 > |\gamma_i|$ for $i \neq 1$. Hence,

$$(3.12) \quad Dx - 1 = (D-1) \left((1-\eta)x^{\ell} + \eta x^{\ell+1} \right).$$

Note that $x \in (0, 1)$ and that, since $q_D(\beta)$ is increasing in β , x is decreasing in β . Since the right hand side goes to 0 exponentially fast as $\ell \nearrow \infty$, and hence as $\beta \nearrow \infty$, it follows that $x \searrow 1/D$ exponentially fast as $\beta \nearrow \infty$, which also implies $q_D(\beta) \nearrow 1$ exponentially fast, as desired. Note that the precise exponential rate of convergence can be obtained by inserting this limit back into the right hand side of (3.12). \square

Applying Proposition 3.3 to Propositions 3.1 and 3.2, we now demonstrate Theorem 1.1.

PROOF OF THEOREM 1.1. Setting $Q_k = e^{R_k}$, where R_k is given in (3.7), one has

$$(3.13) \quad Q_k = \left(\prod_{i=k-\ell+1}^{k-1} Q_i^{D-1} \right) Q_{k-\ell}^{\eta(D-1)},$$

with $Q_k = e$ for $k = -\ell + 1, \dots, -1, 0$. We proceed to compare Q_k with $1/P_k$, where P_k satisfies one of (3.3), (3.5) and (3.6).

Comparison of Q_k with $1/P_k$, with $\eta = \hat{\beta}$, $\ell = \lfloor \beta \rfloor = k - k_1$ and P_k satisfying (3.3), provides an upper bound on the limit in (1.4). To see this, we first set $\tilde{Q}_k = M^{-k} Q_k$, for given $M > 1$. Since $(D-1)(\beta-1) > 1$, by substituting into (3.13), one can check that, for large enough M and k ,

$$(3.14) \quad \tilde{Q}_k \geq C_3 b^k \left(\prod_{i=k-\ell+1}^{k-1} \tilde{Q}_i^{D-1} \right) \tilde{Q}_{k-\ell}^{\eta(D-1)}$$

for any fixed choice of C_3 and b , in particular, for $C_3 = 1/C_1$ and $b = 3$, where C_1 is chosen as in Proposition 3.1. Moreover, on account of (3.8),

$$(3.15) \quad \lim_{k \rightarrow \infty} (1/k) \log_D \log (\tilde{Q}_k) = q_D(\beta)$$

where, in particular, $q_D(\beta) > 0$, and hence $\tilde{Q}_k \rightarrow \infty$ as $k \rightarrow \infty$.

We observe that $1/P_k$ satisfies the inequality that is analogous to that for P_k in (3.3), but with the inequality reversed and prefactors $3^k/C_1$ instead of $C_1/3^k$. Comparing \tilde{Q}_k with $1/P_k$ therefore implies that, for large enough n not depending on k ,

$$1/P_k \leq \tilde{Q}_{k+n}.$$

The upper bound for (1.4) therefore follows from (3.15) for the same choice of $q_D(\beta)$, which we recall is continuous in β . The limit in (1.5) also follows from Proposition 3.3.

Comparison of Q_k with $1/P_k$ also provides a lower bound on the limit in (1.4). In the case where β is nonintegral, we choose η and ℓ as before, with $\eta = \hat{\beta}$, $\ell = \lfloor \beta \rfloor = k - k_1$; note that P_k satisfies the upper bound in (3.5). We proceed as in the first part, but instead set $\tilde{Q}_k = M^k Q_k$, for given $M > 1$. One can check that, for large enough M and k ,

$$(3.16) \quad \tilde{Q}_k \leq C_3 b^k \left(\prod_{i=k-\ell+1}^{k-1} Q_i^{D-1} \right) Q_{k-\ell}^{\eta(D-1)}$$

for any choice of $C_3 > 0$ and $b > 0$. As before, (3.15) holds.

The terms $1/P_k$ satisfy the inequality that is the analog of (3.5). Also, $1/P_k \rightarrow \infty$ as $k \rightarrow \infty$. Comparing \tilde{Q}_k with $1/P_k$ therefore implies that, for large enough n not depending on k ,

$$(3.17) \quad 1/P_{k+n} \geq \tilde{Q}_k.$$

The lower bound for (1.4) therefore follows from (3.15) when β is nonintegral.

The reasoning in the case where β is integral is similar, but with the difference that we now choose $\eta = 1 - \delta$, $\ell = \beta - 1 = k - k_1 - 1$, where $\delta \in (0, 1)$ is arbitrary. Now, P_k satisfies the upper bound in (3.6). We proceed as in the nonintegral case, once again obtaining (3.16). Comparing $1/P_k$ with \tilde{Q}_k again produces (3.15), except that the limit is now $q_D(\beta - \delta)$ because of our choice of η . By Proposition 3.3, $q_D(\cdot)$ is continuous in its argument. Therefore, letting $\delta \searrow 0$ produces the same limit as in the nonintegral case, and hence implies the lower bound for (1.4) in the case where β is integral.

We still need to demonstrate that when (1.3) holds with $\beta = \infty$, then (1.4) holds with $q_D(\infty) = 1$. The lower bound in (1.4) holds on account of (1.5). The upper bound is not difficult to show and does not require Proposition 3.3; we proceed to show the bound.

We will show by induction that, for all k ,

$$(3.18) \quad P_k \geq (C_1/8k)^{kD^k},$$

where C_1 is as chosen as in (3.1), which we assume WLOG is at most 1. To see (3.18), note that if it holds for all $i = 0, \dots, k - 1$ then this, together with (3.1), implies that

$$\begin{aligned} P_k &\geq (C_1/8k)^k \prod_{i=0}^{k-1} \left[(C_1/8i)^{iD^i} \right]^{D-1} \\ &\geq (C_1/8k)^{(k-1)(D^k-1)+k} \geq (C_1/8k)^{kD^k}. \end{aligned}$$

The upper bound in (1.4), with $q_D(\infty) = 1$, follows immediately from (3.18). \square

Demonstration of Proposition 3.1. The proof of Proposition 3.1 is quick. To obtain the lower bounds in both (3.1) and (3.3), it suffices to construct a path along which $Z^{\mathcal{H}}(t)$ increases from 0 to k within the first cycle. This is done, in both cases, by allocating the same amount of time to each of the first k arrivals, which are also required to occur before the first departure.

PROOF OF PROPOSITION 3.1. Consider the cavity process $X^{\mathcal{H}}(\cdot)$ with $X^{\mathcal{H}}(0) = 0$. In order to show (3.1) and (3.3), we obtain lower bounds on the expected amount of time $E[V_k]$ over which $Z^{\mathcal{H}}(t) \geq k$ before $X^{\mathcal{H}}(\cdot)$ returns to 0. We first show (3.1).

We consider the event A where the first service time S is at least $1/2$ and the first k arrivals occur by time $1/4$. The latter event contains the event where each of the first k arrivals occurs not more than $1/4k$ units of time after the previous arrival, starting at time 0.

Conditioned on there being i jobs in the queue, jobs arrive at rate $\alpha_i \geq \alpha P_i^{D-1}$, and so the probability of such an arrival occurring over an interval of length $1/4k$ is at least $1 - \exp\{-\alpha P_i^{D-1}/4k\}$. So, given that $S \geq 1/2$, the probability that all k of these arrivals occur by time $1/4$ is at least

$$(3.19) \quad \prod_{i=0}^{k-1} \left(1 - \exp\{-\alpha P_i^{D-1}/4k\}\right).$$

The event $S \geq 1/2$ occurs with some positive probability c depending on $F(\cdot)$ and, under the event A , the departure time for the first job occurs at least $1/4$ after the last of the first k arrivals. So, the expected amount of time in $[1/4, 1/2]$, during which $Z^{\mathcal{H}}(t) \geq k$ and before $X^{\mathcal{H}}(\cdot)$ has returned to 0, is at least

$$(3.20) \quad \frac{c}{4} \prod_{i=0}^{k-1} \left(1 - \exp\{-\alpha P_i^{D-1}/4k\}\right),$$

which is therefore a lower bound for $E[V_k]$. It therefore follows from (2.5) that

$$(3.21) \quad P_k \geq \frac{c}{4m_0} \prod_{i=0}^{k-1} \left(1 - \exp\{-\alpha P_i^{D-1}/4k\}\right) \geq \frac{c}{4} (\alpha/8k)^k \prod_{i=0}^{k-1} P_i^{D-1},$$

which implies (3.1) for appropriate C_1 .

We next show (3.3) under the assumption (3.2). For this, we set

$$(3.22) \quad s_1 = 2k/(\alpha P_{k_1}^{D-1}).$$

One can reason analogously as through (3.20), but by replacing the time interval $[0, 1/2]$ by $[0, s_1]$ and employing $s_1/2k$ for the allotted time for each of the k arrivals. One obtains that the expected amount of time in $[s_1/2, s_1]$, during which $Z^{\mathcal{H}}(t) \geq k$ and before $X^{\mathcal{H}}(\cdot)$ has returned to 0, is at least

$$(3.23) \quad \frac{s_1}{2} \bar{F}(s_1) \prod_{i=0}^{k-1} \left(1 - \exp\{-\alpha s_1 P_i^{D-1}/2k\}\right).$$

Choose k large enough so that $s_1 \geq s_0$, where s_0 is as in (3.2) and s_1 is as in (3.22). Since $e^{-x} \leq (1 - x/2) \vee 1/2$ for $x \geq 0$, this is at least

$$\begin{aligned} & 2^{-(k_1+2)} s_1^{-(\beta-1)} (\alpha s_1/4k)^{k-k_1-1} \prod_{i=k_1+1}^{k-1} P_i^{D-1} \\ & \geq 2^{-k} (\alpha/4k)^\beta \left(\prod_{i=k_1+1}^{k-1} P_i^{D-1} \right) P_{k_1}^{\hat{\beta}(D-1)}, \end{aligned}$$

where the inequality follows from (3.22) and $k - k_1 = \beta - \hat{\beta}$. Consequently,

$$E[V_k] \geq 2^{-k} (\alpha/4k)^\beta \left(\prod_{i=k_1+1}^{k-1} P_i^{D-1} \right) P_{k_1}^{\hat{\beta}(D-1)}.$$

Again applying (2.5), it follows that, for large enough k (depending on α and β),

$$P_k \geq 3^{-k} \left(\prod_{i=k_1+1}^{k-1} P_i^{D-1} \right) P_{k_1}^{\hat{\beta}(D-1)},$$

which implies (3.3). □

Demonstration of Proposition 3.2. In order to demonstrate Proposition 3.2, we will employ Lemma 3.1 below; the lemma will also be employed in the demonstration of Propositions 4.2 and 5.2. (A substantially more intricate variant of the proof of Lemma 3.1 will be needed for the proof of Proposition 4.4.) Lemma 3.1 provides upper bounds involving $R(k, s)$, $H(n)$ and $\rho(k, s)$, for $k \geq 1$, $s \geq 0$ and $n \geq 0$, which are defined as follows.

For $s > 0$, $R(k, s)$ is the expected return time of the cavity process $X^{\mathcal{H}}(\cdot)$ (with equilibrium environment \mathcal{H}) to the empty state 0, from $X^{\mathcal{H}}(0)$ with $Z^{\mathcal{H}}(0) = k$ and $S^{\mathcal{H}}(0) = s$. We set $R(k, 0) = \lim_{s \searrow 0} R(k, s)$, which is also the expected return time to 0 just after departure of a job, but without knowledge of the residual service time of the job that is beginning service. The quantity $H(n)$ is the number of jobs, for this process, at the time when the $(n + 1)$ st job has just departed, e.g., $H(0)$ is the number of jobs just after departure of the job originally in service. The stopping time $\rho(k, s)$ is the first time n at which $H(n) = 0$.

We also denote by Y_n the service time of the $(n + 1)$ st job (with $Y_0 = s$ being the service time of the job originally in service), and set $T_\ell = \sum_{n=0}^{\ell} Y_n = \sum_{n=1}^{\ell} Y_n + s$. Note that Y_1, Y_2, \dots are i.i.d. with distribution function $F(\cdot)$, which, as always, is assumed to have mean 1.

LEMMA 3.1. *Let $R(\cdot, \cdot)$ and $\rho(\cdot, \cdot)$ be defined as above. Then, for large enough N_0 ,*

$$(3.24) \quad R(k, s) \leq 2(k + s + N_0)$$

and

$$(3.25) \quad E[\rho(k, s)] \leq 2(k + s/2 + N_0),$$

for all k and s .

PROOF. It is not difficult to see that (3.24) follows from (3.25). By applying Wald's equation to $T(\cdot)$ and $\rho(\cdot, \cdot)$ (with respect to the underlying σ -algebra generated by $X^{\mathcal{H}}(\cdot)$), one obtains

$$R(k, s) = E[T_{\rho(k, s)}] = E \left[\sum_{n=1}^{\rho(k, s)} Y_n \right] + s = E[\rho(k, s)]E[Y_1] + s \leq 2(k + s + N_0),$$

with the inequality following from (3.25) and $E[Y_1] = 1$.

In order to show (3.25), we consider the process

$$(3.26) \quad M(n) = H(n) + n/2 - N_1 \exp\{-\theta(H(n) \wedge k_0)\}.$$

For appropriate choices of $N_1, \theta > 0$ and $k_0 \in \mathbb{Z}_+$, we claim $M(n)$ is a supermartingale, with respect to the filtration $\mathcal{G}_n = \sigma(H(0), \dots, H(n))$, after restricting to times n , with $n \leq \rho(k, s)$, and then stopping the process.

These three constants are chosen as follows. We choose k_0 large enough so that $\alpha DP_{k_0+1}^{D-1} \leq 1/2$. For $H(n) > k_0$, one can check that the supermartingale inequality

$$(3.27) \quad E[M(n+1) | \mathcal{G}_n] \leq M(n)$$

is satisfied – the arrival rate of jobs is at most $1/2$ over the time interval $(T_{n-1}, T_n]$ during which the $(n+1)$ st job is served, which has mean length 1, and so

$$E[H(n+1) | \mathcal{G}_n] \leq H(n) - 1/2.$$

In order to analyze $M(n+1)$ when $H(n) \leq k_0$, we set

$$M_1(n) = -\exp\{-\theta(H(n) \wedge k_0)\}.$$

We choose θ large enough so that, for some $\epsilon > 0$ and all $H(n) \leq k_0$,

$$(3.28) \quad E[M_1(n+1) | \mathcal{G}_n] \leq M_1(n) - \epsilon.$$

This requires a standard computation using the convexity of the exponential function and the upper bound αD on the arrival rate of jobs. (Since $H(\cdot)$ may have positive drift, θ may need to be chosen large.)

We also choose N_1 so that $\epsilon N_1 \geq \alpha D + 1/2$. Together with (3.28), this implies (3.27) also holds for $H(n) \leq k_0$. Consequently, $M(n)$ is a supermartingale, as claimed.

In order to demonstrate (3.25), we will apply the optional sampling theorem to $M(\cdot)$ stopped at times $\rho_n(k, s) = \rho(k, s) \wedge n$. First note that

$$(3.29) \quad E[M(0)] \leq E[H(0)] \leq k + s/2$$

for $k \geq k_0$, since the arrival rate of jobs is bounded above by $1/2$. Also, for given s , $E[H(0)]$ is increasing as a function of k , the number of jobs in the cavity process at time 0. Together with (3.29), this implies that, for all k ,

$$(3.30) \quad E[M(0)] \leq (k \vee k_0) + s/2 \leq k + s/2 + k_0.$$

Since the supermartingale $M(\cdot)$ is bounded from below, application of the optional sampling theorem to $\rho_n(k, s)$ implies that

$$E[M(\rho_n(k, s))] \leq E[M(0)] \leq k + s/2 + k_0,$$

and hence

$$0 \leq E[H(\rho_n(k, s))] \leq k + s/2 + k_0 + N_1 - E[\rho_n(k, s)]/2.$$

Solving for $E[\rho_n(k, s)]$ implies

$$E[\rho_n(k, s)] \leq 2(k + s/2 + k_0 + N_1) = 2(k + s/2 + N_0)$$

for $N_0 = k_0 + N_1$. Letting $n \rightarrow \infty$ implies (3.25). \square

Lemma 3.1 provides an upper bound on the expected time over a cycle during which there are at least k jobs, provided such a state has already been attained. Below, we will obtain an upper bound on the probability of attaining such a state and combine this with (3.24).

In order for $X^{\mathcal{H}}(\cdot)$, starting at 0, to attain a state with k jobs, it must first attain states with $k_1 + 1, k_1 + 2, \dots, k - 1$ jobs, where k_1 has been specified in the previous subsection. (It turns out that including states with fewer jobs in this sequence will not improve our bounds.) We let $\sigma_{k_1+1}, \dots, \sigma_k$ denote the number of jobs that have already departed when such a state is first attained (e.g., $\sigma_i = 0$ means that the first job is still being served at the time t when $Z^{\mathcal{H}}(t) = i$ first occurs).

One trivially has

$$0 \leq \sigma_{k_1+1} \leq \sigma_{k_1+2} \leq \dots \leq \sigma_k.$$

Partition $\{k_1 + 1, k_1 + 2, \dots, k\}$ so that $i \neq i'$ are in the same subset if $\sigma_i = \sigma_{i'}$, i.e., the times t_i and $t_{i'}$ at which $Z^{\mathcal{H}}(t_i) = i$ and $Z^{\mathcal{H}}(t_{i'}) = i'$ first occur are in the same service time interval. One can write such a partition as

$$(3.31) \quad \|i_0 + 1, \dots, i_1 \| i_1 + 1, \dots, i_2 \| \dots \| i_{m-1} + 1, \dots, i_m \|,$$

with $i_0 = k_1$ and $i_m = k$, when the partition consists of m sets (where m is random). We denote by Π_k the set of all such partitions and by $\pi \in \Pi_k$ an element in the set, with the notation $i_0(\pi), i_1(\pi), \dots, i_m(\pi)$ being used when convenient. We will say that a partition π occurs during a cycle when the corresponding sequence of events occurs, and denote by A_π the event associated with the partition.

For each of the sets in (3.31) except the last, there is a corresponding service interval, $[T_{n_{\ell-1}}, T_{n_\ell})$, with $\ell = 1, \dots, m-1$, at the beginning of which there are strictly less than $i_{\ell-1}$ jobs and at the end exactly i_ℓ jobs. (Since such an interval ends with a departure, the number of jobs at the beginning of the next service interval must be one less, which requires the cavity process to “retrace some of its steps” before the number of jobs reaches i_ℓ again.) For $\ell = m$, there may be strictly more than k jobs at T_{n_ℓ} ; instead, we consider the restricted interval $[T_{n_{m-1}}, \tau_k]$, where τ_k is the first time at which there are at least k jobs. Unlike at the end of the other intervals $[T_{n_{\ell-1}}, T_{n_\ell})$, the residual service time s will not be 0. When s is large, this will increase the occupation time where $Z^{\mathcal{H}}(t) \geq k$, which will require us to exercise some care with our computations.

Since $k - k_1 \leq \beta$, the number of distinct partitions in (3.31) is at most 2^β . In Proposition 3.4 below, we compute an upper bound on P_k using an upper bound on the expected occupation time corresponding to each partition, and then by multiplying by 2^β . The upper bound in (3.34) includes a factor k^β obtained by employing Lemma 3.1 repeatedly. The form of the bounds in (3.34) and (3.35) varies in different ranges of s ; we will therefore find it useful to employ the notation

$$(3.32) \quad L_\ell(s) = \prod_{i=i_{\ell-1}}^{i_\ell-1} [(\alpha DP_i^{D-1} s) \wedge 1].$$

($L_\ell(\cdot)$ implicitly depends on the partition π through $i_{\ell-1}$ and i_ℓ .) We will employ $L(s)$ when i goes from k_1 to $k-1$, which corresponds to the trivial partition in (3.31) consisting of a single set.

In the proof of Proposition 3.4, we will use the following elementary Chebyshev integral inequality, which states that, if $f(s)$ and $g(s)$ are both integrable functions that are increasing in s , then, for any distribution function $F(\cdot)$,

$$(3.33) \quad \int_{-\infty}^{\infty} f(s)g(s) F(ds) \geq \int_{-\infty}^{\infty} f(s) F(ds) \cdot \int_{-\infty}^{\infty} g(s) F(ds).$$

PROPOSITION 3.4. *Consider a family of JSQ networks, with the same assumptions holding as in Proposition 3.2, except that (3.4) is not assumed. Then, for large enough k ,*

$$(3.34) \quad P_k \leq 3m_0^{-1}(6k)^\beta \int_0^\infty (k+s)L(s) F(ds).$$

PROOF. We first claim that the probability of the cavity process $X^{\mathcal{H}}(\cdot)$, with $Z^{\mathcal{H}}(0) \leq i_{\ell-1}$ and $S^{\mathcal{H}}(0) = s$, attaining i_ℓ jobs before time s is at most

$$(3.35) \quad \prod_{i=i_{\ell-1}}^{i_\ell-1} \left(1 - \exp\{-\alpha DP_i^{D-1}s\}\right) \leq \prod_{i=i_{\ell-1}}^{i_\ell-1} \left[(\alpha DP_i^{D-1}s) \wedge 1\right] \\ = L_\ell(s).$$

Under this event, arrivals must occur sequentially over $[0, s]$ at times t_i when $Z^{\mathcal{H}}(t_i-) = i$, for $i = i_{\ell-1}, \dots, i_\ell - 1$, and the rate of such arrivals is at most αDP_i^{D-1} . Since there is at most time s for each arrival, multiplying the corresponding upper bounds on the probability of an arrival at each step gives the first bound in (3.35). The following inequality is then obtained by applying the inequality $1 - e^{-x} \leq x \wedge 1$.

Recall that V_k denotes the occupation time over a cycle when $Z^{\mathcal{H}}(t) \geq k$. In order for $V_k > 0$, the event A_π must occur for some $\pi \in \Pi_k$; hence $E[V_k] = \sum_{\pi \in \Pi_k} E[V_k; A_\pi]$. We claim that, for any partition $\pi \in \Pi_k$ and large enough k ,

$$(3.36) \quad E[V_k; A_\pi] \leq (3k)^{m_\pi} \prod_{\ell=1}^{m_\pi-1} \left(\int_0^\infty L_\ell(s) F(ds) \right) \times \\ \times 3 \int_0^\infty (k+s)L_{m_\pi}(s) F(ds).$$

To obtain (3.36), we argue by induction, applying (3.35) at each step. It suffices to show that, for each step with $\ell < m_\pi$, one obtains an additional factor $3i_{\ell-1} \int_0^\infty L_\ell(s) F(ds)$ and, for $\ell = m_\pi$, one obtains the factor

$9(i_{m_\pi-1}) \int_0^\infty (k+s)L_{m_\pi}(s) F(ds)$. For $\ell \geq 2$, the factor $3i_{\ell-1}$ is obtained by applying (3.25), with $s = 0$, which gives an upper bound on the expected number of service intervals occurring over the remainder of the cycle, after the service interval corresponding to the $(\ell - 1)$ st step ends; also, $i_0 \geq m_0$, which equals the expected number of service intervals at the beginning of the cycle. The other factor is obtained from (3.35) by integrating against $F(\cdot)$ and, for $\ell = m_\pi$, by employing (3.24) to provide an upper bound on the expected occupation time V_k , again employing (3.35) and then integrating against $F(\cdot)$.

On the other hand, by repeatedly applying the Chebyshev integral inequality (3.33) to (3.36), it follows that, for an arbitrary partition in (3.31), (3.36) is maximized for the trivial partition. That is, for any partition $\pi \in \Pi_k$, the quantity in (3.36) is bounded above by

$$(3.37) \quad 3(3k)^\beta \int_0^\infty (k+s)L(s) F(ds).$$

Since $|\Pi_k| \leq 2^\beta$, it follows from (3.36) and (3.37) that

$$\begin{aligned} P_k &= m_0^{-1} E[V_k] = m_0^{-1} \sum_{\pi \in \Pi_k} E[V_k; A_\pi] \\ &\leq 3m_0^{-1} (6k)^\beta \int_0^\infty (k+s)L(s) F(ds), \end{aligned}$$

which implies (3.34) □

We now complete the proof of Proposition 3.2.

PROOF OF PROPOSITION 3.2. We employ the upper bound for P_k given by (3.34) for large enough k . The integral in (3.34) is bounded above by

$$(3.38) \quad \begin{aligned} &2ks_0 \int_0^{s_0} L(s) F(ds) + 2k \int_{s_0}^\infty sL(s) F(ds) \\ &\leq 2\beta(s_0^{\beta+1} + 1)k \int_1^\infty s^{-\beta} L(s) ds \end{aligned}$$

by integrating by parts and absorbing the first term into the second; note that $L(s)$ is increasing in s on account of (3.32). We decompose this last integral using intervals of the form $[1/\alpha DP_{k-1}^{D-1}, \infty)$, $[1/\alpha DP_{i-1}^{D-1}, 1/\alpha DP_i^{D-1})$, for $i = k_1 + 1, \dots, k - 1$, and $[1, 1/\alpha DP_{k_1}^{D-1})$; we need to consider the cases where β is and is not an integer separately.

Suppose that β is not an integer. Applying (3.32) to the above integral over $[1/\alpha DP_{k-1}^{D-1}, \infty)$, one has the upper bound

$$(3.39) \quad \int_{1/\alpha DP_{k-1}^{D-1}}^{\infty} s^{-\beta} ds = \frac{1}{\beta-1} (\alpha DP_{k-1})^{(D-1)(\beta-1)}.$$

For $i = k_1 + 1, \dots, k-1$, one has, over $[1/\alpha DP_{i-1}^{D-1}, 1/\alpha DP_i^{D-1})$, the upper bounds

$$(3.40) \quad \begin{aligned} & \int_{1/\alpha DP_{i-1}^{D-1}}^{1/\alpha DP_i^{D-1}} (\alpha Ds)^{k-i} (P_{k-1} \cdots P_i)^{D-1} s^{-\beta} ds \\ & \leq \frac{(\alpha D)^{\beta-1}}{\beta+i-k-1} \left(P_{k-1} \cdots P_i P_{i-1}^{\beta+i-k-1} \right)^{D-1}. \end{aligned}$$

For the last interval $[1, 1/\alpha DP_{k_1}^{D-1})$, one has the upper bound

$$(3.41) \quad \begin{aligned} & \int_1^{1/\alpha DP_{k_1}^{D-1}} (\alpha Ds)^{k-k_1} (P_{k-1} \cdots P_{k_1})^{D-1} s^{-\beta} ds \\ & \leq \frac{(\alpha D)^{\beta-1}}{1-\hat{\beta}} \left(P_{k-1} \cdots P_{k_1+1} P_{k_1}^{\hat{\beta}} \right)^{D-1}, \end{aligned}$$

where we recall that $\hat{\beta} = \beta - k + k_1$. Note that the lower limits of integration supply the dominant term in (3.39) and (3.40), whereas the upper limit supplies the dominant term in (3.41), because of the choice of k_1 .

Since P_i is decreasing in i , if one ignores the coefficients not involving powers of P_i on the right hand sides of (3.39)–(3.41), the largest bounds in (3.39)–(3.41) are given in (3.40), with $i = k_1 + 1$, and in (3.41), in each case by the powers of P_i ,

$$(3.42) \quad \left(P_{k-1} \cdots P_{k_1}^{\hat{\beta}} \right)^{D-1}.$$

The coefficients of these powers are bounded above by terms not involving k . Employing (3.34) of Proposition 3.4, together with (3.38), one obtains the bound (3.5) for P_k , for appropriate C_2 and all k .

When β is an integer, the computations are similar. The inequalities in (3.39) and (3.41) are the same as before, as are all of the cases in (3.40) except for $i = k_1 + 1$. Rather than (3.40), one obtains the following inequality when $i = k_1 + 1$:

$$(3.43) \quad \begin{aligned} & \int_{1/\alpha DP_{k_1}^{D-1}}^{1/\alpha DP_{k_1+1}^{D-1}} (\alpha Ds)^{\beta-1} (P_{k-1} \cdots P_{k_1+1})^{D-1} s^{-\beta} ds \\ & \leq (D-1) (\alpha D)^{\beta-1} (P_{k-1} \cdots P_{k_1+1})^{D-1} \log(P_{k_1}/P_{k_1+1}). \end{aligned}$$

By comparing terms involving P_i and ignoring the other coefficients, one can check that the largest bound is given in (3.43). Since the logarithm term there is dominated by $P_{k_1+1}^{-\delta(D-1)}$, for given $\delta > 0$ and small enough P_{k_1+1} , it follows that (3.6) holds for P_k , for appropriate C_2 and all k . \square

4. The case where $\beta \in (1, D/(D-1))$. In this section, we demonstrate Theorem 1.2. We do this by demonstrating the lower and upper bounds needed for the theorem in Propositions 4.1 and 4.2. Here, we set $\nu_\beta = (\beta - 1)/[1 - (D - 1)(\beta - 1)]$.

PROPOSITION 4.1. *Consider a family of JSQ networks, with given $D \geq 2$ and $N = D, D + 1, \dots$, where the N th network has Poisson rate- αN input, with $\alpha < 1$, and where service at each queue is FIFO, with distribution $F(\cdot)$ having mean 1. Assume that (1.2) holds and that*

$$(4.1) \quad \bar{F}(s) \geq s^{-\beta} \quad \text{for } s \geq s_0,$$

with $\beta \in (1, D/(D-1))$ and some $s_0 \geq 1$. Then, for appropriate $C_4 > 0$ and all k ,

$$(4.2) \quad P_k \geq C_4 k^{-\nu_\beta}.$$

PROPOSITION 4.2. *Consider a family of JSQ networks, with given $D \geq 2$ and $N = D, D + 1, \dots$, where the N th network has Poisson rate- αN input, with $\alpha < 1$, and where service at each queue is FIFO, with distribution $F(\cdot)$ having mean 1. Assume that (1.2) holds and that*

$$(4.3) \quad \bar{F}(s) \leq s^{-\beta} \quad \text{for } s \geq s_0,$$

with $\beta \in (1, D/(D-1))$ and some $s_0 \geq 1$. Then, for each $\delta > 0$, appropriate $C_5 > 0$, and all k ,

$$(4.4) \quad P_k \leq C_5 k^{-(1-\delta)\nu_\beta}.$$

Theorem 1.2 follows immediately from Propositions 4.1 and 4.2 upon letting $\delta \searrow 0$ in (4.4).

As in Section 3, the demonstration of the lower bound is much quicker than that of the upper bound. We first demonstrate the lower bound, Proposition 4.1, and then, in the remainder of the section, derive the upper bound, Proposition 4.2.

Demonstration of Proposition 4.1. As in Section 3, when we considered the case where $\beta > D/(D - 1)$, for the lower bound, it suffices to construct a path along which $Z^{\mathcal{H}}(t)$ increases from 0 to k within the first cycle. As before, we allocate the same amount of time for each of the first k arrivals, which are also required to occur before the first departure.

PROOF OF PROPOSITION 4.1. Consider the cavity process $X^{\mathcal{H}}(\cdot)$ with $X^{\mathcal{H}}(0) = 0$. We obtain a lower bound on the expected amount of time over which $Z^{\mathcal{H}}(t) \geq k$ before $X^{\mathcal{H}}(\cdot)$ returns to 0, assuming that $k \geq s_0$.

We consider the event where the first service time is at least $s_1 = 4k/(\alpha P_k^{D-1})$ and the first k arrivals occur by time $s_1/2$. We note that the probability of the latter event occurring is greater than the probability of at least k events occurring by time $s_1/2$ for a rate- αP_k^{D-1} Poisson process, which, by a simple large deviations estimate, is at least

$$1 - e^{-C_6 k} \geq 1/2$$

for large enough k and an appropriate constant C_6 . Together with (4.1), this implies that the expected amount of time in $[s_1/2, s_1]$, during which $Z^{\mathcal{H}}(t) \geq k$ and before $X^{\mathcal{H}}(\cdot)$ has returned to 0, is at least

$$(4.5) \quad \frac{1}{2} \cdot \frac{s_1}{2} \cdot \bar{F}(s_1) \geq \frac{1}{4} (4k/(\alpha P_k^{D-1}))^{-(\beta-1)}.$$

Inequality (4.5) implies that

$$P_k \geq \frac{\alpha}{16m_0} k^{-(\beta-1)} P_k^{(D-1)(\beta-1)},$$

where m_0 is the mean return time to 0. Solving for P_k , it follows from this that, for large k ,

$$P_k \geq \frac{\alpha}{16m_0} k^{-\nu\beta},$$

which implies (4.2) for all k . □

Demonstration of Proposition 4.2. The demonstration of the upper bound (4.4) for Theorem 1.2 is considerably more involved than is the lower bound. The basic idea is to consider two cases, depending on whether or not there is a service time s with $s > s_1$, for preassigned $s_1 \geq 1$, before a state x with $z = k$ is reached in the first cycle, and to obtain upper bounds for each case. The two bounds are given in Propositions 4.3 and 4.4, which are then combined in Corollary 4.1. Employing Corollary 4.1, the proof of Proposition 4.2 provides an iteration scheme where a sequence of values

$s_1(n), n = 0, 1, 2, \dots$, for s_1 are given that provide successively better upper bounds for P_k , and that yield (4.4) in the limit. The demonstration of Proposition 4.4 involves the construction of a supermartingale, whose details are postponed until the end of the section.

Let τ_k , for given $k \in \mathbb{Z}_+$, denote the first time t in the first cycle at which $Z^{\mathcal{H}}(t) = k$. For Propositions 4.3 and 4.4, we denote by $B_{s_1, k}$ the set of realizations on which some service time that is strictly greater than s_1 , with $s_1 \geq 1$, occurs up to and including the service time interval that contains τ_k . Proposition 4.3 considers the case where $B_{s_1, k}$ occurs; the demonstration of the proposition is quick, using Lemma 3.1. As in Sections 2 and 3, we denote by V_k the occupation time at states x , with $z \geq k$.

PROPOSITION 4.3. *Consider a family of JSQ networks with the same assumptions holding as in Proposition 4.2. Then, for appropriate C_7 and all k ,*

$$(4.6) \quad E[V_k; B_{s_1, k}] \leq C_7 s_1^{-\beta} (k + s_1).$$

PROOF. We apply Lemma 3.1 at the beginning of the first service time that is greater than s_1 . Since there are less than k jobs under $B_{s_1, k}$ then, it follows that, for appropriate C_8 and large enough k ,

$$(4.7) \quad \begin{aligned} E[V_k; B_{s_1, k}] &\leq 3(P(B_{s_1, k})/\bar{F}(s_1)) \int_{s_1}^{\infty} (k + s) F(ds) \\ &\leq C_8 \int_{s_1}^{\infty} (k + s) F(ds). \end{aligned}$$

For the latter inequality, note that there are only a finite expected number of service times in the first cycle, and that, by Wald's equation, the expected number of such times that are at most s , for given $s \geq 0$, is proportional to $F(s)$. Since $k + s$ is increasing in s , integration by parts together with (4.3) implies that the last quantity in (4.7) is at most $C_7 s_1^{-\beta} (k + s_1)$, for appropriate C_7 . \square

In order to consider the behavior of $X^{\mathcal{H}}(\cdot)$ on $B_{s_1, k}^c$, we find it convenient to employ the service time distribution $F^{s_1}(\cdot)$ that is given by

$$(4.8) \quad \begin{aligned} F^{s_1}(s) &= F(s) && \text{for } s < s_1, \\ &= 1 && \text{for } s \geq s_1. \end{aligned}$$

We define $X_{s_1}^{\mathcal{H}}(\cdot)$ analogously to $X^{\mathcal{H}}(\cdot)$, but where the service time distribution of the process is $F^{s_1}(\cdot)$ up to and including the service time interval

containing τ_k , and is given by $F(\cdot)$ afterwards; $Z_{s_1}^{\mathcal{H}}(\cdot)$ and $S_{s_1}^{\mathcal{H}}(\cdot)$ are defined analogously. One has

$$(4.9) \quad E[V_k; B_{s_1, k}^c] \leq E[V_k^{s_1}],$$

where $V_k^{s_1}$ is the occupation time at states x with $z \geq k$ for $X_{s_1}^{\mathcal{H}}(\cdot)$. Note that the mean of $F^{s_1}(\cdot)$ is at most 1.

In contrast to Proposition 4.3, Proposition 4.4 requires us to restrict our choice of s_1 in terms of k . For this, we set $k_1 = \lfloor k/3 \rfloor$ and introduce the abbreviation

$$(4.10) \quad p = p_{k_1} = \alpha D P_{k_1}^{D-1}.$$

The required restriction on s_1 is that

$$(4.11) \quad s_1 \leq k^{1-\eta}/p,$$

where $\eta \in (0, 1/2)$. In the proof of Proposition 4.2, we will introduce an iterative scheme that involves explicit choices of s_1 based on our knowledge of P_{k_1} at each step.

Proposition 4.4 gives us the following upper bound for $E[V_k; B_{s_1, k}^c]$.

PROPOSITION 4.4. *Consider a family of JSQ networks with the same assumptions holding as in Proposition 4.2. Suppose that $\delta > 0$ and $\eta \in (0, 1/2)$ are given, and that s_1 satisfies (4.11). Then, for appropriate C_9 and all k ,*

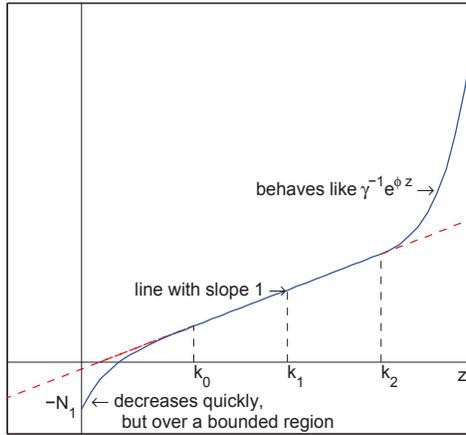
$$(4.12) \quad E[V_k; B_{s_1, k}^c] \leq C_9(k + s_1) \exp\{-\delta k^\eta\}.$$

The demonstration of Proposition 4.4 depends on an appropriate supermartingale. In order to construct the supermartingale, we employ the following notation. We fix $k_0 \in \mathbb{Z}_+$, which will not depend on k as k increases, and set $k_2 = 2k_1$, where k_1 is as defined earlier. We set

$$(4.13) \quad f(z) = (z \wedge k_2) - N_1 \exp\{-\theta(z \wedge k_0)\} + \gamma^{-1} \exp\{\phi(z \vee k_2)\} - \gamma^{-1} \exp\{\phi k_2\},$$

where $N_1, \theta > 0$, $\phi = \delta k^{\eta-1}$ and $\gamma = \phi e^{\phi k_2}$, and where $\delta > 0$ and $\eta \in (0, 1/2)$ are as in Proposition 4.4; the function $f(\cdot)$ is sketched in Figure 1. The terms P_k will continue to refer to the probabilities defined at the beginning of the paper with respect to the cavity process with the original service distribution $F(\cdot)$ (not $F^{s_1}(\cdot)$).

We let $H(n)$, with $n \geq 1$, denote the number of jobs for the process $X_{s_1}^{\mathcal{H}}(\cdot)$, with $X_{s_1}^{\mathcal{H}}(0) = 0$, at the time when the n th job has just departed; we set

FIG 1. graph of $f(z)$

$H(0) = 1$, and we let ρ denote the first time n at which either $H(n) = 0$ or $H(n) \geq k - 1$. Using this notation, we define the analog of $M(\cdot)$ in (3.26),

$$(4.14) \quad M(n) = f(H(n \wedge \rho)).$$

Note that, unlike for $M(\cdot)$ in (3.26), $M(\cdot)$ here depends strongly on the choice of k . Also, unlike $M(\cdot)$ in (3.26), it was not necessary to wait until the first departure in defining $H(0)$, since $X_{s_1}^{\mathcal{H}}(0) = 0$, and hence there is no initial residual service time; in both cases, $H(1) - H(0)$ is the change in the number of jobs during the service time of the first job that begins service when $t > 0$.

PROPOSITION 4.5. *Consider a family of JSQ networks with the same assumptions holding as in Proposition 4.2. Suppose that $\delta > 0$ and $\eta \in (0, 1/2)$ are given, and that $M(\cdot)$ is defined as above. Also, assume that s_1 satisfies (4.11). Then, for large enough k , $M(\cdot)$ is a supermartingale, with respect to the filtration $\mathcal{G}_n = \sigma(H(0), \dots, H(n))$, for small enough $\delta > 0$, and appropriate $\theta, N_1 > 0$, with δ, θ , and N_1 not depending on k .*

The demonstration of Proposition 4.5 will be given at the end of the section. Employing Proposition 4.5, we now demonstrate Proposition 4.4.

PROOF OF PROPOSITION 4.4. We suppose that the terms δ, θ and N_1 are chosen so that, for large enough k , $M(\cdot)$ is a supermartingale. Set $\sigma_L = \min\{n : M(n) \geq L\}$, for given $L > 0$, which will depend on k . Since $M(\cdot)$ is

bounded below by $-N_1$ and $M(0) \leq 1$, by the optional sampling theorem,

$$(4.15) \quad P(\sigma_L < \infty) \leq \frac{1}{L}(1 + N_1).$$

On the other hand, denoting by n_k the service interval during which $Z_{s_1}^{\mathcal{H}}(t) = k$ first occurs and by T_{n_k} the end of that interval, $H(n_k) = Z_{s_1}^{\mathcal{H}}(T_{n_k}) \geq k - 1$. Substituting this into (4.13)-(4.14) and recalling that $\phi = \delta k^{\eta-1}$, one obtains

$$M(n_k) \geq -N_1 + \gamma^{-1} \exp\{\phi(k-1)\} - \gamma^{-1} \exp\{2\phi k/3\} \geq \exp\{\delta k^{\eta}\}/2\gamma,$$

for large k . Let $\tau_k^{s_1}$ denote the first time t , during the first cycle, at which $Z_{s_1}^{\mathcal{H}}(t) = k$. Plugging $L = \exp\{\delta k^{\eta}\}/2\gamma$ into (4.15), substituting in for γ and recalling that $k_2 = 2\lfloor k/3 \rfloor$, it follows that, for large k ,

$$(4.16) \quad \begin{aligned} P(\tau_k^{s_1} < \infty) &\leq P(\sigma_L < \infty) \leq \exp\{-\delta k^{\eta}\} \cdot \exp\{2\delta k^{\eta}/3\} \\ &= \exp\{-\delta k^{\eta}/3\}. \end{aligned}$$

Lemma 3.1 applied to $F(\cdot)$, which is the service distribution of new service times after $\tau_k^{s_1}$, provides the upper bound

$$E[V_k^{s_1} | \mathcal{F}_{\tau_k^{s_1}}] \leq 2(k + s + N_0),$$

given that $S_{s_1}^{\mathcal{H}}(\tau_k^{s_1}) = s$. Since the residual service time for $X_{s_1}^{\mathcal{H}}(t)$ is at most s_1 for $t \leq \tau_k^{s_1}$, it therefore follows from (4.16) that, for large k ,

$$(4.17) \quad E[V_k^{s_1}] \leq 3(k + s_1) \exp\{-\delta k^{\eta}/3\}.$$

The inequality in (4.12) follows upon applying (4.9) to (4.17) and substituting in a smaller choice of η . \square

We combine the upper bounds given in Propositions 4.3 and 4.4 for $E[V_k; B_{s_1}]$ and $E[V_k; B_{s_1}^c]$ to obtain the following upper bound on $E[V_k]$. Since we will always assume $s_1 \leq k^{\nu\beta+1}$ in our application of the corollary, this allows us to omit the exponential term inherited from (4.12).

COROLLARY 4.1. *Consider a family of JSQ networks with the same assumptions holding as in Proposition 4.2. Fix $\eta \in (0, 1)$ and assume that*

$$(4.18) \quad s_1 \leq [(\alpha D)^{-1} k^{1-\eta} P_{k_1}^{1-D}] \wedge k^N,$$

for some $N > 0$. Then, for appropriate C_{10} and all k ,

$$(4.19) \quad E[V_k] \leq C_{10} s_1^{-\beta} (k + s_1).$$

PROOF. It follows from Propositions 4.3 and 4.4 that

$$E[V_k] \leq C_7 s_1^{-\beta} (k + s_1) + C_9 (k + s_1) \exp\{-\delta k^\eta\}$$

for appropriate C_7 and C_9 . The assumption $s_1 \leq k^N$ allows us to absorb the second term into the first. \square

The following elementary lemma will be employed in the proof of Proposition 4.2.

LEMMA 4.1. *Suppose that $R(n)$ satisfies*

$$(4.20) \quad R(n) = aR(n-1) + b \quad \text{for } n \geq 1,$$

with $R(0) = c$, for $a \in (0, 1)$ and $b, c \in \mathbb{R}$. Then,

$$(4.21) \quad \lim_{n \rightarrow \infty} R(n) = b/(1-a).$$

If $R(0) < b/(1-a)$, then the sequence $R(n)$ is increasing, and if $R(0) > b/(1-a)$, then the sequence is decreasing.

PROOF. Setting $\tilde{R}(n) = R(n) - b/(1-a)$, it follows from (4.20) that

$$(4.22) \quad \tilde{R}(n) = a\tilde{R}(n-1) \quad \text{for } n \geq 1,$$

with $\tilde{R}(0) = c - b/(1-a)$. All of the claims follow by iterating (4.22). \square

We will employ the lemma in the following multiplicative format.

COROLLARY 4.2. *Suppose that $Q_k(n)$ satisfies*

$$(4.23) \quad Q_k(n) = \left(k^{-(1-2\eta)} Q_k(n-1)^{D-1} \right)^{\beta-1} \quad \text{for } n \geq 1,$$

with $Q_k(0) = k^{1-\beta+2\eta\beta}$, for $(D-1)(\beta-1) \in (0, 1)$ and $\eta \in (0, 1/2)$. Then, $Q_k(n)$ satisfies $Q_k(n) = k^{-R(n)}$, where the sequence $R(n)$ is increasing in n and

$$(4.24) \quad \lim_{n \rightarrow \infty} R(n) = (1-2\eta)\nu_\beta,$$

with $\nu_\beta = (\beta-1)/[1-(D-1)(\beta-1)]$.

PROOF. The limit in (4.24) follows from (4.21) upon setting $a = (D - 1)(\beta - 1)$, $b = (1 - 2\eta)(\beta - 1)$ and $c = \beta - 1 - 2\eta\beta$. The sequence $R(n)$ is increasing since $R(0) < (1 - 2\eta)\nu_\beta$. \square

We now employ Corollaries 4.1 and 4.2 to demonstrate Proposition 4.2.

PROOF OF PROPOSITION 4.2. For given k and $\eta \in (0, 1/2)$, we define $Q_k(n)$ as in Corollary 4.2 and set

$$(4.25) \quad \begin{aligned} s_1(n) &= (\alpha D)^{-1} k^{1-\eta} && \text{for } n = 0, \\ &= (\alpha D Q_{k_1}(n-1)^{D-1})^{-1} k^{1-\eta} && \text{for } n \geq 1, \end{aligned}$$

where $k_1 = \lfloor k/3 \rfloor$. Using $s_1(n)$, we will inductively show that, for large k (depending on η),

$$(4.26) \quad P_k \leq Q_k(n) \quad \text{for all } n \geq 0.$$

Letting $n \rightarrow \infty$, it therefore follows from the corollary that

$$(4.27) \quad P_k \leq k^{-(1-2\eta)\nu_\beta}.$$

This implies (4.4) in Proposition 4.2, with $\delta < 2\eta$.

To show (4.26) holds for $n = 0$, we note that $s_1(0)$ satisfies (4.18). Therefore, by (2.5) and Corollary 4.1, for large k ,

$$(4.28) \quad P_k \leq 2C_{10}(m_0)^{-1} s_1(0)^{-\beta} k \leq k^{-(\beta-1)+2\eta\beta} = Q_k(0),$$

where the constants in the second expression are absorbed in the third expression by using the 2η term. Note that, in this application of (4.19), $s_1(0) \leq k$. In the application of (4.19) given next, $s_1(n) \geq k$ for all $n \geq 1$.

Suppose that (4.26) holds with $n - 1$ in place of n . Choosing $s_1(n)$ as in (4.25) and employing the lower bound for $Q_k(n)$ given in (4.24), one can check that $s_1(n)$ satisfies (4.18), with $N = \nu_\beta + 1$. Also note that, by Corollary 4.2,

$$Q_{k_1}(n) \leq 3^{\nu_\beta} Q_k(n)$$

for large k and all n . Applying (2.5) and Corollary 4.1 again, we therefore obtain that, for large k ,

$$(4.29) \quad P_k \leq 2C_{10}(m_0)^{-1} s_1(n)^{-(\beta-1)} \leq \left(k^{-(1-2\eta)} Q_k(n-1)^{D-1} \right)^{\beta-1} = Q_k(n).$$

This demonstrates (4.26). \square

In order to complete the demonstration of Proposition 4.2, we need to prove Proposition 4.5, which asserts that $M(\cdot)$, given by (4.14), is a supermartingale.

PROOF OF PROPOSITION 4.5. We need to show the supermartingale inequality (3.27) for $H(n) \in (0, k-1)$. We do this separately over the intervals $(0, k_1]$ and $(k_1, k-1)$. The basic idea for the first interval will be to show that, on $(0, k_1]$, (3.27) will be satisfied for the same reasons as was $M(\cdot)$, for $M(\cdot)$ given by (3.26), the point being that, since $k_2 - k_1 = \lfloor k/3 \rfloor$ is large, the role played by the additional terms $\gamma^{-1} \exp\{\phi(z \vee k_2)\} - \gamma^{-1} \exp\{\phi k_2\}$ in (4.13) is negligible. On the second interval $(k_1, k-1)$, the strong negative drift of $Z_{s_1}^H(\cdot)$ will be enough to compensate for both the $z \wedge k_2$ and $\gamma^{-1} \exp\{\phi(z \vee k_2)\} - \gamma^{-1} \exp\{\phi k_2\}$ terms. We do the latter interval first.

We claim that for large k and $H(n) \geq k_1$,

$$(4.30) \quad E[\exp\{\phi H(n+1)\} | \mathcal{G}_n] \leq E[\exp\{\phi H(n)\}].$$

We first note that, because of (4.10), for $H(n) \geq k_1$, the number of arrivals over the $(n+1)$ st service interval is dominated by a mixture of Poisson rate- ps random variables, with s being distributed according to $F^{s_1}(\cdot)$. Therefore,

$$E[\exp\{\phi(H(n+1) - H(n))\} | \mathcal{G}_n] \leq e^{-\phi} \int_0^{s_1} \exp\{ps(e^\phi - 1)\} F^{s_1}(ds).$$

Since the integrand is convex and the mean of $F^{s_1}(\cdot)$ is at most 1, the right hand side is at most

$$(4.31) \quad e^{-\phi} \left[\left(1 - \frac{1}{s_1}\right) + \frac{1}{s_1} \exp\{ps_1(e^\phi - 1)\} \right].$$

On account of the definitions of ϕ and p given between (4.10) and (4.14), both ϕ and $ps_1\phi$ are at most δ . Using $e^z \sim 1 + z$ for z close to 0, one can therefore check that, for given $\epsilon > 0$ and small enough $\delta > 0$, (4.31) is at most

$$1 + \phi[(1 + \epsilon)p - (1 - \epsilon)].$$

For $p \leq (1 - \epsilon)/(1 + \epsilon)$, the above quantity is at most 1, which holds here since $p \rightarrow 0$ as $k \rightarrow \infty$. This implies (4.30).

For $H(n) > k_2$, it is easy to see that (3.27) follows from (4.30), since

$$(4.32) \quad \begin{aligned} f(z) - \gamma^{-1} e^{\phi z} &= b && \text{for } z \geq k_2, \\ &\leq b && \text{for } z < k_2, \end{aligned}$$

where $b \stackrel{\text{def}}{=} f(k_2) - \gamma^{-1}e^{\phi k_2}$. For $H(n) \in (k_1, k_2]$, (3.27) follows from (4.30) with a bit more work. In place of (4.32), one uses

$$(4.33) \quad g(z) \stackrel{\text{def}}{=} f(z) - \gamma' e^{\phi z} \leq f(H(n)) - \gamma' e^{\phi H(n)}$$

for all z , where $\gamma' \stackrel{\text{def}}{=} (\phi e^{\phi H(n)})^{-1} = \gamma^{-1}e^{\phi(k_2 - H(n))}$. To check (4.33), note that equality holds for $z = H(n)$; we claim that the maximum of $g(\cdot)$ is taken there. One has $g'(H(n)) = 0$ because of our definition of γ' ; $g'(z) \geq 0$ for $z \leq H(n)$ and $g'(z) \leq 0$ for $z \in [H(n), k_2)$ because of the concavity of $g(\cdot)$ there; and since $\gamma' \geq 1$, for $z > k_2$, it is easy to see that $g'(z) \leq 0$ there. This shows (4.33) and hence (3.27) for $H(n) \in (k_1, k_2]$ as well.

We still need to show (3.27) for $H(n) \in (0, k_1]$. For this, we compare $M(\cdot)$ with $\tilde{M}(\cdot)$, where

$$\tilde{f}(z) = z + n/2 - N_1 \exp\{-\theta(z \wedge k_0)\}$$

and

$$\tilde{M}(n) = \tilde{f}(H(n \wedge \rho)).$$

Set $R(n) = M(n) - \tilde{M}(n)$. For $H(n) \in (0, k_1]$, one has

$$(4.34) \quad \begin{aligned} R(n+1) - R(n) + 1/2 &= 0 && \text{for } H(n+1) \leq k_2, \\ &\leq \gamma^{-1}e^{\phi H(n+1)} && \text{for } H(n+1) > k_2. \end{aligned}$$

Since $\tilde{M}(\cdot)$ is the supermartingale in (3.26), except with a different initial state, $\tilde{M}(\cdot)$ satisfies (3.27) if θ and N_1 are chosen as in (3.26). In a moment, we will show that

$$(4.35) \quad E[e^{\phi H(n+1)} \mathbf{1}\{H(n+1) > k_2\} | \mathcal{G}_n] \leq \gamma/2$$

for $H(n) \leq k_1$ and large k . Using (4.34) and (4.35), (3.27) therefore also follows for $M(\cdot)$ for $H(n) \leq k_1$.

It suffices to show (4.35) for $H(n) = k_1$. To do this, we need to control the right tail of $H(n+1)$. The number of arrivals over the $(n+1)$ st service interval for the cavity process is dominated by a mixture of Poisson mean- ps_1 random variables, with the mixture distributed according to F^{s_1} . This mixture is in turn dominated by a Poisson mean- s_1 random variable. Therefore, the left hand side of (4.35) is at most

$$(4.36) \quad \sum_{k'=k_2}^{\infty} \left[e^{-ps_1} (ps_1)^{k'-k_1} / (k'-k_1)! \right] e^{\phi k'}.$$

Setting $\ell = k' - k_2$, one has

$$(k' - k_1)! \geq \ell!(k_2 - k_1)! \geq \ell!((k_2 - k_1)/e)^{k_2 - k_1},$$

where the last inequality follows from Stirling's formula. Substituting ℓ into (4.36), applying this bound, and employing $\exp\{e^\phi ps_1\} = \sum_{\ell=0}^{\infty} (e^\phi ps_1)^\ell / \ell!$, it follows that (4.36) is at most

$$(4.37) \quad \left(\frac{eps_1}{k_2 - k_1} \right)^{k_2 - k_1} \exp \left\{ ps_1(e^\phi - 1) + \phi k_2 \right\} \leq C_{11} k^{-\eta k/3} e^{4\phi k}$$

for appropriate C_{11} , where the inequality employs (4.11) and $e^\phi - 1 \leq 2\phi$, for small ϕ . As $k \rightarrow \infty$, the right hand side of (4.37) goes to 0. It follows that the left hand side of (4.35), with $H(n) = k_1$, goes to 0 as $k \rightarrow \infty$. This implies (4.35) holds for $H(n) \leq k_1$ and large k , which completes the proof of the proposition. \square

5. The case where $\beta = D/(D - 1)$. In this section, we demonstrate Theorem 1.3. We do this by demonstrating the lower and upper bounds needed for the theorem, in Propositions 5.1 and 5.2.

PROPOSITION 5.1. *Consider a family of JSQ networks, with given $D \geq 2$ and $N = D, D + 1, \dots$, where the N th network has Poisson rate- αN input, with $\alpha < 1$, and where service at each queue is FIFO, with distribution $F(\cdot)$ having mean 1. Assume that (1.2) holds and that*

$$(5.1) \quad \bar{F}(s) \geq c_1 s^{-D/(D-1)} \quad \text{for } s \geq s_0,$$

for some $c_1 > 0$ and $s_0 \geq 1$. Then, for appropriate $C_{12} > 0$ and $s_D(c_1) < \infty$,

$$(5.2) \quad P_k \geq C_{12} e^{-s_D(c_1)k} \quad \text{for all } k,$$

where

$$(5.3) \quad s_D(c_1) \searrow 0 \quad \text{as } c_1 \nearrow \infty,$$

PROPOSITION 5.2. *Consider a family of JSQ networks, with given $D \geq 2$ and $N = D, D + 1, \dots$, where the N th network has Poisson rate- αN input, with $\alpha < 1$, and where service at each queue is FIFO, with distribution $F(\cdot)$ having mean 1. Assume that (1.2) holds and that*

$$(5.4) \quad \bar{F}(s) \leq c_2 s^{-D/(D-1)} \quad \text{for } s \geq s_0,$$

for some $c_2 < \infty$ and $s_0 \geq 1$. Then, for appropriate C_{13} and $r_D(c_2) > 0$,

$$(5.5) \quad P_k \leq C_{13} e^{-r_D(c_2)k} \quad \text{for all } k,$$

where

$$(5.6) \quad r_D(c_2) \nearrow \infty \quad \text{as } c_2 \searrow 0.$$

Theorem 1.3 follows immediately from Propositions 5.1 and 5.2.

As in the previous two sections, the demonstration of the lower bound is substantially quicker than that of the upper bound. We first demonstrate the lower bound, Proposition 5.1 and then, in the remainder of the section, derive the upper bound Proposition 5.2.

Demonstration of Proposition 5.1. As in Sections 3 and 4, where we considered the cases $\beta > D/(D-1)$ and $\beta \in (1, D/(D-1))$, for the lower bound, it suffices to construct a path along which $Z^{\mathcal{H}}(t)$ increases from 0 to k within the first cycle. In contrast to the previous two settings, we allocate geometrically increasing amounts of time to the sequence of arrivals, up through the k th arrival; as before, these arrivals are required to occur before the time of the first departure.

PROOF OF PROPOSITION 5.1. The argument is similar to that for Proposition 4.1 in that we examine the cavity process $X^{\mathcal{H}}(\cdot)$ with $X^{\mathcal{H}}(0) = 0$, and obtain a lower bound on the expected amount of time V_k over which $Z^{\mathcal{H}}(t) \geq k$ before $X^{\mathcal{H}}(\cdot)$ returns to 0. Here, we argue by induction, and assume that

$$(5.7) \quad P_i \geq C_{12} e^{-a_1 i} \quad \text{for } i = 0, \dots, k-1,$$

for given k , where $C_{12} \leq [(a_1 \vee 1)s_0]^{-1}$, and $a_1 > 0$ will be specified later.

We consider the following event A that leads to a lower bound on P_k that is compatible with (5.7). We stipulate that the first service time is at least

$$(5.8) \quad s_1 \stackrel{\text{def}}{=} C_{14} e^{a_1(D-1)k},$$

where $C_{14} = 4(\alpha a_1)^{-1} C_{12}^{-(D-1)}$. Note that $C_{14} \geq s_0$. We also assume that the interarrival time for the $(i+1)$ st arrival at the queue, $i = 0, \dots, k-1$, is at most

$$(5.9) \quad \alpha^{-1} C_{12}^{-(D-1)} \exp\{\frac{1}{2} a_1 (D-1)(k+i)\}.$$

A little estimation shows that the sum of the terms in (5.9), over $i = 0, \dots, k-1$, is bounded above by

$$(5.10) \quad \begin{aligned} & \alpha^{-1} C_{12}^{-(D-1)} \exp\{a_1(D-1)k\} / (\exp\{\frac{1}{2}a_1(D-1)\} - 1) \\ & \leq (2/\alpha a_1) C_{12}^{-(D-1)} \exp\{a_1(D-1)k\}, \end{aligned}$$

which is one-half of (5.8).

On account of the induction hypothesis in (5.7), the probability that the $(i+1)$ st arrival occurs within the interarrival time in (5.9) is at least

$$1 - \exp\{-e^{\frac{1}{2}a_1(D-1)(k-i)}\}.$$

So, the probability that the corresponding events for $i = 0, \dots, D-1$ all occur within the allotted time is at least

$$\prod_{i=1}^k \left(1 - \exp\left\{-e^{\frac{1}{2}a_1(D-1)i}\right\}\right) \geq \psi(a_1),$$

where $\psi(a_1) > 0$ for $a_1 > 0$ and does not depend on k or D , with $\psi(a_1) \rightarrow 1$ as $a_1 \rightarrow \infty$; the inequality requires a little computation.

It follows from the previous two paragraphs that the event A , given by the service time and interarrival times restricted as in (5.8) and (5.9), has probability at least

$$\psi(a_1) \bar{F}(C_{14} \exp\{a_1(D-1)k\}).$$

On A , $Z^{\mathcal{H}}(t) \geq k$ over the interval $[s_1/2, s_1]$, which has length $\frac{1}{2}C_{14} \exp\{a_1(D-1)k\}$. So,

$$E[V_k] \geq \frac{1}{2} C_{14} \psi(a_1) \exp\{a_1(D-1)k\} \bar{F}(C_{14} \exp\{a_1(D-1)k\}).$$

By substituting the bound in (5.1) for $\bar{F}(s)$ and employing $P_k = m_0^{-1} E[V_k]$, one obtains

$$\begin{aligned} P_k & \geq \frac{1}{2m_0} \psi(a_1) c_1 C_{14} \exp\{a_1(D-1)k\} (C_{14} \exp\{a_1(D-1)k\})^{-D/(D-1)} \\ & = \frac{1}{2m_0} \psi(a_1) c_1 (C_{14})^{-1/(D-1)} e^{-a_1 k} \geq \frac{c_1}{4m_0} \psi(a_1) (\alpha a_1)^{1/(D-1)} C_{12} e^{-a_1 k}. \end{aligned}$$

For given c_1 and large enough a_1 , the last quantity in the above display is at least $C_{12} e^{-a_1 k}$. This implies the induction hypothesis in (5.7) for k and this choice of a_1 . Since (5.7) obviously holds for $i = 0$, (5.2) follows, with $s_D(c_1) = a_1$. Similarly, for given a_1 , one obtains the lower bound $C_{12} e^{-a_1 k}$, if c_1 is chosen large enough, which implies (5.3). This completes the proof. \square

Demonstration of Proposition 5.2. The demonstration of the upper bound (5.5) is substantially more involved than is the lower bound. The basic idea is similar to that employed for the upper bound in Section 3, where we classified different paths for attaining $Z^{\mathcal{H}}(t) + k$, for given k and some t , in terms of partitions π given by (3.31). There, the probability of the event associated with the trivial partition dominated the probabilities for the other partitions. Computing an upper bound for the probability for the trivial partition and multiplying by the upper bound 2^β for the total number of partitions gave us our desired upper bounds on P_k .

The details of our setup here will be different. The partitions we consider will be defined somewhat differently, and we will need to be more careful in summing up probabilities – we will compute the probability of the event associated with the trivial partition separately, and will then sum up the probabilities for the other partitions, which will be negligible in comparison. We will also require an upper bound on P_k from Proposition 4.2, at the beginning of our argument. On the other hand, the computations of these upper bounds will be substantially easier here than the corresponding bounds were in Section 3. The key difference is that here the probabilities P_k will decrease sufficiently slowly in k so that, for our estimates, not too much will be lost if we consider P_i to be approximately the same for $i = k_1, \dots, k - 1$, which will simplify our computations.

In order to show (5.5) and (5.6) of Proposition 5.2, we will argue by induction, assuming that, for preassigned $a_2, C_{13} > 0$ and $k_0, h_T \in \mathbb{Z}_+$,

$$(5.11) \quad P_i \leq C_{13} e^{-a_2 i} \quad \text{for } i = k_0, \dots, k - 1,$$

for given k with $k \geq k_0 + h_T$. For appropriate choices of these preassigned values, we will show that the inequality in (5.11) holds with $i = k$. We set

$$(5.12) \quad h_T = \lceil 700D^2c_2 \rceil^{D-1} \vee 6$$

and

$$(5.13) \quad a_2 = (h_T)^{-1} \vee \frac{1}{6} \log((220D^2c_2)^{-1}),$$

where c_2 is as in (5.4). These particular choices of h_T and a_2 are not needed for most of the argument, and will only be inserted at the very end.

In order to specify C_{13} and k_0 , we note that, since (4.3) is satisfied for every $\beta < D/(D-1)$ because of (5.4) and since $\nu_\beta \nearrow \infty$ as $\beta \nearrow D/(D-1)$, it follows from Proposition 4.2 that, for any N , $\lim_{k \rightarrow \infty} k^N P_k = 0$. Here, we set $N = h_T + 1$. We choose k_0 large enough so that $P_{k_0} \leq (DM^2k_0^N)^{-1}$, $(1 + 1/k_0)^N \leq e^{a_2}$,

$$(5.14) \quad k_0 \geq D(c_2 \vee (1/c_2)) s_0^{2h_T} h_T^{h_T+1}$$

and $k_0 \geq N_0$ all hold, where $M = e^{a_2 h_T}$, s_0 is as in (5.4) and N_0 is as in Lemma 3.1. Setting $C_{13} = M e^{a_2 k_0} P_{k_0}$ implies (5.11) holds for $k = k_0, \dots, k_0 + h_T$, which we will need in order to begin our induction argument.

It follows from the definition of C_{13} and the first two conditions on k_0 that

$$C_{13} D M e^{-a_2 k} \leq k^{-N} \quad \text{for all } k \geq k_0.$$

Setting $q_k = \alpha D (C_{13} M e^{-a_2 k})^{D-1}$, it follows from this that

$$(5.15) \quad q_k \leq k^{-(D-1)N} \quad \text{for all } k \geq k_0,$$

which we will use throughout the induction argument for (5.11). In order to follow the basic induction argument, the reader should keep in mind (5.11) and (5.15), without worrying much about the other inequalities.

In order to demonstrate the inequality in (5.11) with $i = k$, we proceed as outlined in the beginning of the subsection, employing the partitions π given in (3.31) and the events A_π , on which a sequence of arrivals and departures occurs in the first cycle that induces the partition π . We define Π_k , as before, as the set of all partitions with final element $i_m = k$; here, the first element will be $i_0 + 1$, with $i_0 = k_1$, where $k_1 = k - h_T$. In the present setting, we will pay more attention than in Section 3 to the length of each of the sets in a partition π , setting $h_\ell = |G_\ell|$, for $\ell = 1, \dots, m_\pi$, for the number of elements in the ℓ th set G_ℓ of the partition; one has $h_T = \sum_{\ell=1}^{m_\pi} h_\ell$.

An important step in computing an upper bound for P_k is Proposition 5.3, which is the analog of Proposition 3.4. Rather than employing $L_\ell(s)$ as in the proof of Proposition 3.4 for the upper bound for a set in the partition, we employ

$$(5.16) \quad J_{k,h}(s) \stackrel{\text{def}}{=} e^{-q_k s} \sum_{i=h}^{\infty} (q_k s)^i / i!.$$

The quantity $J_{k,h}(s)$ is the probability of at least h events occurring for a mean- $q_k(s)$ Poisson random variable, and dominates the probability that, over the time interval $(0, s]$, at least h arrivals occur for a cavity process $X^{\mathcal{H}}(\cdot)$ with $Z^{\mathcal{H}}(0) \geq k_1$ and $S^{\mathcal{H}}(0) \geq s$. This bound follows from the upper bound in (2.6), together with the induction hypothesis (5.11) and our definition of M .

PROPOSITION 5.3. *Consider a family of JSQ networks with the same assumptions holding as in Proposition 5.2, except that (5.4) is not assumed.*

Suppose that the induction assumption (5.11) holds for given h_T and for $k_0 \geq N_0$, where N_0 is as in Lemma 3.1. Then,

$$(5.17) \quad P_k \leq 3 \sum_{\pi \in \Pi_k} (3k)^{m_\pi-1} \prod_{\ell=1}^{m_\pi-1} \left(\int_0^\infty J_{k,h_\ell}(s) F(ds) \right) \times \\ \times \int_0^\infty (k+s) J_{k,h_{m_\pi}}(s) F(ds).$$

PROOF. One can reason similarly to the argument for (3.36), in the proof of Proposition 3.4, by computing an upper bound on $E[V_k; A_\pi]$. Summation over $\pi \in \Pi_k$ and application of (2.5) will then imply (5.17). The assumption $k_0 \geq N_0$ is needed only to absorb the term N_0 when applying Lemma 3.1.

One argues inductively, repeating the argument for (3.36), except for the substitution of $J_{k,h_\ell}(s)$ for $L_\ell(s)$ and a minor change involving the factors of $3k$. For each step with $\ell < m_\pi$, one obtains an additional factor $i_{\ell-1}^* \int_0^\infty J_{k,h_\ell}(s) F(ds)$ and, for $\ell = m_\pi$, one obtains the factor $3i_{m_\pi-1}^* \int_0^\infty (k+s) J_{k,h_{m_\pi}}(s) F(ds)$, where $i_{\ell-1}^* = 3i_{\ell-1}$ for $\ell \geq 2$ and $i_0^* = m_0$, with m_0 being the mean return time to 0 for $X^{\mathcal{H}}(\cdot)$. For $\ell < m_\pi$, the integral part of the factor is obtained by employing the comparison given directly before the statement of the proposition, comparing $J_{k,h_\ell}(s)$ with the probability of at least h arrivals over a service time of at least s , and then by integrating against s ; for $\ell = m_\pi$, one also employs (3.24) to provide an upper bound on the expected occupation time V_k .

For $\ell \geq 2$, the factor $i_{\ell-1}^*$ is obtained by applying (3.25), with $s = 0$, which gives an upper bound on the expected number of service intervals occurring over the remainder of the cycle, after the service interval corresponding to the $(\ell - 1)$ st step ends. For ℓ_0^* , instead of the factor $3i_0$, one can employ m_0 , since this is the expected number of service intervals over an entire cycle, and no conditioning is needed for this first step. Since each of the remaining factors is at most $3k$, the product of all of the factors is at most $m_0(3k)^{m_\pi-1}$, and since $P_k = (m_0)^{-1}E[V_k]$, the m_0 factors cancel, and one obtains the $(3k)^{m_\pi-1}$ factor in (5.17). (The improved bound just obtained by removing a factor of $3k$ will only be needed when bounding the right hand side of (5.17) for the trivial partition, in Proposition 5.4.) \square

In Propositions 5.4 and 5.5, we provide upper bounds for the summands on the right hand side of (5.17), which we denote by $Q_k(\pi)$. In Proposition 5.4, we do this for the trivial partition consisting of a single set, for which we write π_1 . In Proposition 5.5, we do this for each of the other partitions. The sum over $\Pi - \{\pi_1\}$ of the bounds for $Q_k(\pi)$ that are obtained in Proposition

5.5 will be negligible in comparison with the bound obtained for $Q_k(\pi_1)$ in Proposition **5.4**. This last bound will therefore dominate the upper bound for P_k that will be obtained by inserting these bounds into **(5.17)** of the preceding proposition.

Both Propositions **5.4** and **5.5** employ the elementary upper bounds for $J_{k,h}(s)$,

$$(5.18) \quad \begin{aligned} J_{k,h}(s) &\leq (4(q_k s)^h/h!) \wedge 1 && \text{for } s \leq h/4q_k, \\ &\leq 1 && \text{for } s > h/4q_k, \end{aligned}$$

which one obtains by dominating the series in **(5.16)** by the geometric series $((q_k s)^h/h!) \sum_{i=0}^{\infty} (3/4)^i$, for $s \leq h/4q_k$.

PROPOSITION 5.4. *Suppose that*

$$(5.19) \quad Q_k(\pi_1) = \int_0^{\infty} (k+s) J_{k,h_T}(s) F(ds),$$

where $F(\cdot)$ satisfies **(5.4)** and $J_{k,h_T}(s)$ is chosen as above, with $h_T \geq 6$, and suppose that $k \geq k_0$, with **(5.14)** and **(5.15)** both holding. Then,

$$(5.20) \quad Q_k(\pi_1) \leq 55Dc_2(q_k/h_T)^{1/(D-1)}.$$

PROOF. Throughout the proof, we will abbreviate by setting $h_T = h$. We begin the argument by decomposing the integral into the three parts, \int_0^k , $\int_k^{h/27q_k}$ and $\int_{h/27q_k}^{\infty}$, which we analyze separately.

Since $k+s \leq 2k$ for $s \in [0, k]$, it is easy to check that

$$(5.21) \quad \int_0^k (k+s) J_{k,h}(s) F(ds) \leq 8k^{h+1} q_k^h/h!.$$

One has $k \geq s_0$ for s_0 in **(5.4)**. Applying **(5.4)** and $k+s \leq 2s$, and substituting $t = q_k s/h$, one sees that the second integral is bounded above by

$$(5.22) \quad (8D/(D-1))c_2 \int_0^{1/27} q_k^{1/(D-1)} \frac{(t^{2/3}h)^h}{h!} t^{(\frac{h}{3}-\frac{D}{D-1})} dt.$$

Since $h \geq 6$, one can check that $(t^{2/3}h)^h/h! \leq 3^{-h}$ and $t^{(\frac{h}{3}-\frac{D}{D-1})} \leq 1$ for $t \leq 1/27$. Therefore, **(5.22)** is bounded above by

$$(5.23) \quad (8/27)(D/(D-1))c_2 3^{-h} q_k^{1/(D-1)} \leq c_2 3^{-h} q_k^{1/(D-1)}.$$

Applying (5.4), the third integral is at most

$$(5.24) \quad 2(D/(D-1))c_2 \int_{h/27q_k}^{\infty} s^{-D/(D-1)} ds \leq 54Dc_2(q_k/h)^{1/(D-1)}.$$

On account of (5.15) and $q_k \leq c_2$, the bound for the third integral is clearly the dominant term. Combining the bounds for the three integrals therefore implies that

$$Q_k(\pi_1) \leq 55Dc_2(q_k/h)^{1/(D-1)},$$

which is the bound in (5.20). \square

PROPOSITION 5.5. *Suppose that*

$$(5.25) \quad Q_k(\pi) = (3k)^{m_\pi-1} \prod_{\ell=1}^{m_\pi-1} \left(\int_0^\infty J_{k,h_\ell}(s) F(ds) \right) \times \\ \times \int_0^\infty (k+s) J_{k,h_{m_\pi}}(s) F(ds),$$

where $F(\cdot)$ satisfies (5.4) and $J_{k,h_\ell}(s)$ is chosen as above, with $h_T \geq 5$, and suppose that $k \geq k_0$, with (5.14) and (5.15) both holding. Then,

$$(5.26) \quad Q_k(\pi) \leq 81D^2(c_2+1)^2 s_0^{2h_T} h_T^{h_T} (3k)^{h_T} q_k^{D/(D-1)}$$

for each $\pi \in \Pi_k - \{\pi_1\}$.

In order to demonstrate Proposition 5.5, we will categorize each partition in $\Pi_k - \{\pi_1\}$ as one of three types, based on the sizes and indices of its constituent sets G_ℓ , $\ell = 1, \dots, m_\pi$. We will say G_ℓ is *large* if $h_\ell \geq 3$ and *small* if $h_\ell \leq 2$; we will also distinguish between sets G_ℓ with $\ell < m_\pi$ and $\ell = m_\pi$. We will say that a partition π is of type (I) if at least one of its sets G_ℓ , with $\ell < m_\pi$, is large; that it is of type (II) if G_{m_π} is large, but all of the other sets are small; and that it is of type (III) if none of its sets is large, but at least two sets G_{ℓ_1} and G_{ℓ_2} , with $\ell_1 < \ell_2 < m_\pi$ are small. It is easy to check that, for any $h_T \geq 5$, the three types of sets partition $\Pi_k - \{\pi_1\}$.

PROOF OF PROPOSITION 5.5. We will show separately that (5.26) holds when π is a member of any of the above three types of partitions. We will first bound the above integrals for the large and small sets G_ℓ , for both $\ell = m_\pi$ and $\ell < m_\pi$, and will then apply these bounds to the three types of partitions. When convenient, we abbreviate by setting $h_\ell = h$.

Applying almost the same reasoning as in the proof of Proposition 5.4, one obtains, for large G_{m_π} ,

$$(5.27) \quad \int_0^\infty (k+s) J_{k,h_{m_\pi}}(s) F(ds) \leq 2Dc_2 h_{m_\pi}^{h_{m_\pi}} q_k^{1/(D-1)}.$$

One decomposes the integral into the parts \int_0^k , \int_k^{h/q_k} and \int_{h/q_k}^∞ . A bound for the first integral is again given by the right hand side of (5.21) and a bound for the third integral is given by $2Dc_2(q_k/h)^{1/(D-1)}$. For the second integral, one obtains the bound $c_2 h^h q_k^{1/(D-1)}$, after substituting $t = q_k s/h$ as before. Instead of (5.22), one employs

$$(5.28) \quad 8(D/(D-1))c_2 \int_0^1 q_k^{1/(D-1)} \frac{h^h}{h!} t^{(h-\frac{D}{D-1})} dt$$

as an intermediate bound for the second integral, to which one applies $t^{(h-\frac{D}{D-1})} \leq 1$; the acquired factor h^h will not cause difficulties in the present context. For $k \geq k_0$, the bound in (5.27) follows from the bounds on the three integrals, on account of (5.15) and $q_k \leq c_2$.

Similar reasoning can be applied for large G_ℓ , with $\ell < m_\pi$, to obtain the upper bound

$$(5.29) \quad \int_0^\infty J_{k,h_\ell}(s) F(ds) \leq 2c_2 h_\ell^{h_\ell} q_k^{D/(D-1)}.$$

One decomposes the integral into the parts $\int_0^{s_0}$, $\int_{s_0}^{h/q_k}$ and \int_{h/q_k}^∞ . The first integral is at most $s_0^h q_k^h \leq s_0^h q_k^3$ and the third integral is at most $c_2 q_k^{D/(D-1)}$. For the second integral, one obtains the upper bound $c_2 h^h q_k^{D/(D-1)}$, after substituting $t = q_k s/h$. Instead of (5.22) or (5.28), one employs

$$(5.30) \quad 4(D/(D-1))c_2 \int_0^1 q_k^{D/(D-1)} \frac{h^{h-1}}{h!} t^{(h-\frac{D}{D-1}-1)} dt,$$

as an intermediate bound for the second integral, to which one applies $t^{(h-\frac{D}{D-1}-1)} \leq 1$. Since $1/q_k \geq s_0^h$, the bound in (5.27) follows from the bounds on the three integrals.

For small G_ℓ with $\ell < m_\pi$, one obtains the upper bound

$$(5.31) \quad \int_0^\infty J_{k,h_\ell}(s) F(ds) \leq 9D(c_2+1)s_0^{h_\ell} q_k.$$

As in the previous case, one decomposes the integral into the parts $\int_0^{s_0}$, $\int_{s_0}^{h/q_k}$ and \int_{h/q_k}^∞ . The same estimates show that the first integral is at most

$s_0^h q_k^h \leq s_0^h q_k$ and the third integral is at most $c_2 q_k^{D/(D-1)}$. For the second integral, one obtains the upper bounds

$$(5.32) \quad 4(D/(D-1))c_2 \int_{s_0}^{h/q_k} q_k \frac{h^{h-1}}{h!} s^{-D/(D-1)} ds \leq 8Dc_2 s_0 q_k,$$

with the inequality using $h \leq 2$. The bound in (5.31) follows from the bounds on the three integrals.

For small G_{m_π} , the upper bound

$$(5.33) \quad \int_0^\infty (k+s) J_{k, h_{m_\pi}}(s) F(ds) \leq k+1 \leq 2k$$

follows from $J_{k, h_{m_\pi}}(s) \leq 1$, since $F(\cdot)$ has mean 1.

We also note that, for G_ℓ with $\ell < m_\pi$,

$$(5.34) \quad \int_0^\infty J_{k, h_\ell}(s) F(ds) \leq 1$$

trivially holds.

We now combine the upper bounds in (5.27), (5.29), (5.31) (5.33) and (5.34) to obtain upper bounds for the right hand side of (5.25), for large k . When π is a type (I) partition, it follows from (5.29), (5.33) and (5.34) that

$$(5.35) \quad Q_k(\pi) \leq 2c_2 h_T^{h_T} (3k)^{m_\pi} q_k^{D/(D-1)};$$

when π is a type (II) partition, it follows from (5.27), (5.31) and (5.34) that

$$(5.36) \quad Q_k(\pi) \leq 18D^2(c_2+1)^2 s_0^{h_T} h_T^{h_T} (3k)^{m_\pi-1} q_k^{D/(D-1)};$$

and when π is a type (III) partition, it follows from (5.31), (5.33) and (5.34) that

$$(5.37) \quad Q_k(\pi) \leq 81D^2(c_2+1)^2 s_0^{2h_T} (3k)^{m_\pi} q_k^2.$$

The right hand side of (5.26) is greater than each of the quantities in (5.35)–(5.37). Consequently, (5.26) holds for all $\pi \in \Pi_k - \{\pi_1\}$, as desired. \square

Employing Propositions 5.3, 5.4 and 5.5, and the induction hypothesis (5.11), we now complete the proof of Proposition 5.2.

PROOF OF PROPOSITION 5.2. We will demonstrate that the inequality in (5.11) holds for $i = k$, provided it holds for $i = k_0, \dots, k-1$, for h_T and

a_2 satisfying (5.12) and (5.13), and for k_0 satisfying the inequalities in (5.14) and on each side. By induction, it will follow that

$$(5.38) \quad P_k \leq C_{13}e^{-a_2k} \quad \text{for all } k \geq k_0.$$

By Proposition 5.3,

$$(5.39) \quad P_k \leq 3 \sum_{\pi \in \Pi_k} Q_k(\pi) \leq 3Q_k(\pi_1) + 3 \cdot 2^{h_T} \max_{\pi \in \Pi_k - \{\pi_1\}} Q_k(\pi).$$

On account of (5.14) and (5.15), it follows from the bounds in (5.20) and (5.26), for $Q_k(\pi_1)$ and for $Q_k(\pi)$, $\pi \in \Pi_k - \{\pi_1\}$, that the first term on the right hand side of (5.39) dominates the second term, and therefore

$$(5.40) \quad P_k \leq 220Dc_2(q_k/h_T)^{1/(D-1)}.$$

Substituting for q_k and then for M , this is at most

$$(5.41) \quad \left(220D^2c_2h_T^{-1/(D-1)}e^{a_2h_T}\right) C_{13}e^{-a_2k}.$$

Upon substitution of the value for h_T in (5.12) and $a_2 = 1/h_T$, the quantity inside the parentheses in (5.41) is less than 1. Also, by replacing the term $h_T^{-1/(D-1)}$ by 1, it is easy to see that the quantity inside the parentheses is again less than 1, for $a_2 = \frac{1}{6} \log((220D^2c_2)^{-1})$. So, in either case, the inequality in (5.11) holds for $i = k$. This implies (5.38).

With a large enough choice of C_{13} , (5.38) extends to all $k \geq 0$. This implies (5.5) of Proposition 5.2 with $r_D(c_2) = a_2$, for this choice of C_{13} . Moreover, as $c_2 \searrow 0$, one has $a_2 \nearrow \infty$, and so (5.6) also holds. This completes the proof of Proposition 5.2. \square

REFERENCES

- [1] AZAR, Y., BRODER, A., KARLIN, A. and UPFAL, E. (1994). Balanced allocations. *Proc. 26th ACM Symp. Theory Comp.*, 593-602.
- [2] BRAMSON, M. (2008). *Stability of Queueing Networks*. École d'Été de Probabilités de Saint-Flour XXXVI - 2006, Lecture Notes in Mathematics 1950. Springer-Verlag, Berlin.
- [3] BRAMSON, M. (2010). *Stability of join the shortest queue networks*. Submitted to *Ann. Appl. Probab.*
- [4] BRAMSON, M., LU, Y. and PRABHAKAR, B. (2010). Randomized load balancing with general service time distributions. To appear in *Proc. ACM SIGMETRICS 2010*.
- [5] BRAMSON, M., LU, Y. and PRABHAKAR, B. (2011). Asymptotic independence of queues under randomized load balancing. Submitted to *Queueing Systems*.
- [6] DAVIS, M.H.A. (1993). *Markov Models and Optimization*. Chapman & Hall, London.

- [7] FOSS, S. and CHERNOVA, N. (1998). On the stability of a partially accessible multi-station queue with state-dependent routing. *Queueing Systems* **29**, 55-73.
- [8] GRAHAM, C. (2000). Chaoticity on path space for a queueing network with selection of the shortest queue among several. *J. Appl. Prob.* **37**, 198-211.
- [9] LUCZAK, M. and MCDIARMID, C. (2005). On the power of two choices: Balls and bins in continuous time. *Ann. Appl. Probab.* **15**, 1733-1764.
- [10] LUCZAK, M. and MCDIARMID, C. (2006). On the maximum queue length in the supermarket model. *Ann. Probab.* **34**, 493-527.
- [11] MARTIN, J.B. and SUHOV, Y.M. (1999). Fast Jackson networks. *Ann. Appl. Probab.* **9**, 840-854.
- [12] MITZENMACHER, M. (2001). The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems* **12:10**, 1094-1104.
- [13] STANLEY, R.P. (1997). *Enumerative Combinatorics, Volume 1*. Cambridge University Press, Cambridge.
- [14] SUHOV, Y.M. and VVEDENSKAYA, N.D. (2002). Fast Jackson networks with dynamic routing. *Problems Inform. Transm.* **39**, 136-153.
- [15] VOCKING, B. (1999). How asymmetry helps load balancing. *IEEE Symp. Found. Comp. Sci.*, 131-140.
- [16] VVEDENSKAYA, N.D., DOBRUSHIN, R.L. and KARPELEVICH (1996). Queueing system with selection of the shortest of two queues: an asymptotic approach. *Problems of Information Transmission* **32 (1)**, 15-29.
- [17] VVEDENSKAYA, N.D. and SUHOV, Y.M. (1997). Dobrushin's mean-field approximation for a queue with dynamic routing. *Markov Proc. Rel. Fields* **3**, 493-527.

UNIVERSITY OF MINNESOTA
TWIN CITIES CAMPUS
SCHOOL OF MATHEMATICS
206 CHURCH STREET S.E.
MINNEAPOLIS, MINNESOTA 55455
E-MAIL: bramson@math.umn.edu

UNIVERSITY OF ILLINOIS
COORDINATED SCIENCE LAB
1308 W. MAIN STREET
URBANA, ILLINOIS 61801
E-MAIL: yilu4@illinois.edu

STANFORD UNIVERSITY
DAVID PACKARD BUILDING, ROOM 269
STANFORD, CA 94305
E-MAIL: balaji@stanford.edu