

Instability of LIFO Queueing Networks

Maury Bramson
University of Minnesota
bramson@math.umn.edu

May 4, 2021

Abstract

Under the last-in, first-out (LIFO) discipline, jobs arriving later at a class always receive priority of service over earlier arrivals at any class belonging to the same station. Subcritical LIFO queueing networks with Poisson external arrivals are known to be stable, but an open problem has been whether this is also the case when external arrivals are given by renewal processes. Here, we show that this weaker assumption is not sufficient for stability by constructing a family of examples where the number of jobs in the network increases to infinity over time.

This behavior contrasts with that for the other classical disciplines: processor sharing (PS), infinite server (IS), and first-in, first-out (FIFO), which are stable under general conditions on the renewals of external arrivals. Together with LIFO, PS and IS constitute the classical symmetric disciplines; with the inclusion of FIFO, these disciplines constitute the classical homogeneous disciplines. Our examples show that a general theory for stability of either family is doubtful.

1 Introduction

Under the preemptive last-in, first-out (LIFO) discipline (or policy), jobs in a queueing network arriving at a class always receive priority of service over earlier arrivals at any class belonging to the same station. Service for the preempted jobs continues after later-arriving jobs have been served. This rule is quite natural, and corresponds to later occurring tasks always being given priority over earlier ones, for instance, new jobs being given priority in a piled stack of work to be done. LIFO is a GAAP accepted accounting method for inventory.

The LIFO discipline is one of the four “classical” disciplines that were analyzed in the famous papers Baskett et al. (1975) and Kelly (1975, 1976), the other disciplines being processor sharing (PS), infinite server (IS), and first-in, first-out (FIFO). In these papers, the stability of these four queueing networks was shown when the Poisson input is subcritical, that is, the corresponding Markov processes are positive recurrent given that work on the average arrives at a slower rate than it would be served if all servers are fully active when there are jobs in the network.

Since these papers, substantial progress has been made in showing the stability of subcritical queueing networks under the PS, IS, and FIFO disciplines when the input is generalized from Poisson to renewal processes. However, little is currently known about the stability of the LIFO discipline in the non-Poisson setting. In this paper, we demonstrate instability for a family of subcritical LIFO queueing networks by showing that the number of jobs in the network increases to infinity over time.

To define this family, we first give the network topology and then its external arrival and service processes. The network consists of four stations, with a total of six classes, and is pictured in Figure 1. Jobs enter the network at either Class 1 or Class 4. The jobs arriving at Class 1 are routed successively through Classes 2 and 3, before leaving the network, and the jobs arriving at Class 4 are routed successively through Classes 5 and 6 before leaving. Classes 1 and 6 together comprise Station I, Classes 3 and 4 together comprise Station IV, and Classes 2 and 5 each form their own single-class stations, Stations II and III. Except for the presence of Classes 2 and 5, the network has the same structure as the well-known Rybko-Stolyar network.

The external arrival and service processes are each symmetrically defined, with jobs entering the upper route following the same rules as those entering the lower route. For each of the two routes, external arrivals are given by independent renewal processes, whose interarrival times are i.i.d. random variables with measure ν given by

$$\begin{aligned}\nu(dt) &= \frac{1}{M} e^{-\beta(t-\gamma M)} dt \quad \text{for } t \in [\gamma M, 2M], \\ \nu(\{\frac{1}{M^2}\}) &= 1 - \frac{1}{M},\end{aligned}\tag{1}$$

where M is assumed to be large, and β and γ are chosen so that ν has both measure and mean 1. (One has $\beta > 1$, $\gamma < 1$, with $\beta \sim 1$, $\gamma \sim 1$ for large M .)

The service laws at Classes 1, 3, 4, and 6 are all deterministic, whereas the service laws at Classes 2 and 5 are exponentially distributed, with the

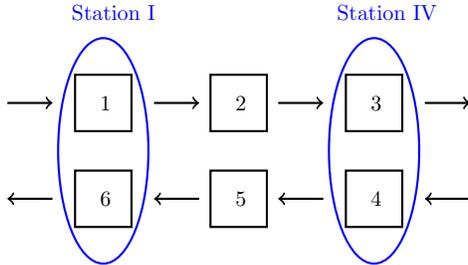


Figure 1: Squares in top row correspond to Classes 1-3, in order of appearance along their route; similarly, squares in lower row correspond to Classes 4-6, in order of appearance. Classes in left oval belong to Station I, classes in right oval belong to Station IV; Classes 2 and 5 belong to the one-class stations, Stations II and III.

means at different classes being given by

$$m_1 = m_4 = \delta^3, \quad m_2 = m_5 = 1 - \delta, \quad m_3 = m_6 = 1 - \delta + \delta^3, \quad (2)$$

where $\delta = 1/M^{1/15}$. We assume that all interarrival and service times are independent of each other. Since, for small δ ,

$$m_1 + m_6 = m_3 + m_4 = 1 - \delta + 2\delta^3 < 1, \quad m_2 = m_5 = 1 - \delta < 1, \quad (3)$$

the system is subcritical.

The system is LIFO, with jobs entering a given class always receiving priority of service over earlier arrivals at any class of its station. (In case of a “tie”, either priority is allowed.) We assume that the network is preemptive resume, with jobs currently in service being interrupted by arrivals, and continuing their service in the absence of more recent arrivals. Jobs originally in the network are assigned an arbitrary ordering for service at their class.

Denoting by $Z(t)$ the total number of jobs in the network at time t , Theorem 1 asserts that $Z(t) \rightarrow \infty$ as $t \rightarrow \infty$.

Theorem 1. *Suppose M is sufficiently large. For any LIFO queueing network with routing as in Figure 1, and external arrival and service processes as in (1)-(2),*

$$Z(t) \rightarrow \infty \quad \text{almost surely as } t \rightarrow \infty. \quad (4)$$

The analog of (4) holds for the total work in the network. We comment on this immediately after the proof of Theorem 3.

When $L = (1 - \delta + \delta^3)/\delta^3$ is an integer, one can create a second family of queueing networks by partitioning the Classes 3 and 6 into L new classes each, Classes 3.1,...,3.L and 6.1,...,6.L, creating in this manner new Stations I and IV that each have $L + 1$ classes (see Figure 2). By immediately continuing service at Classes 3. $(\ell + 1)$ and 6. $(\ell + 1)$ for jobs departing from Classes 3. ℓ and 6. ℓ , the new queueing networks thus defined will also have the LIFO discipline.

The external arrival processes of this second family are again defined as in (1). The service processes for Classes 1, 2, 4, and 5 are as in the first family, and have means given by (2). The service times for Classes 3.1 through 3.L and 6.1 through 6.L are deterministic, and satisfy

$$m_{3.1} = \dots = m_{3.L} = m_{6.1} = \dots = m_{6.L} = \delta^3. \quad (5)$$

Since the mean service times of each of the classes at Stations I and IV is equal to δ^3 , Stations I and IV are of *Kelly type*, that is, the mean service times of the classes at the station are equal. The following analog of (3) holds,

$$m_1 + \sum_{\ell=1}^L m_{6,\ell} = m_4 + \sum_{\ell=1}^L m_{3,\ell} = 1 - \delta + 2\delta^3 < 1, \quad m_2 = m_5 = 1 - \delta < 1, \quad (6)$$

and so the system is subcritical.

Corollary 1 therefore immediately follows from Theorem 1.

Corollary 1. *Suppose M is sufficiently large. For any LIFO queueing network with routing as in Figure 2, external arrival processes as in (1), and service processes for Classes 1, 2, 4, and 5 as in (2) and Classes 3. ℓ and 6. ℓ as in (5),*

$$Z(t) \rightarrow \infty \quad \text{almost surely as } t \rightarrow \infty. \quad (7)$$

1.1 Historical context and some philosophy

The evolution of the theory of multiclass queueing networks has been strongly influenced by explicit results for the four “classical disciplines”, PS, LIFO, IS, and FIFO. The first three of these disciplines are *symmetric* disciplines, whereas FIFO is a member of the more general family of *homogeneous* disciplines. The distinguishing property for homogeneous disciplines is that the distribution of the ordered position assigned to a job arriving at a station and the fraction of service assigned to a job based on its position both

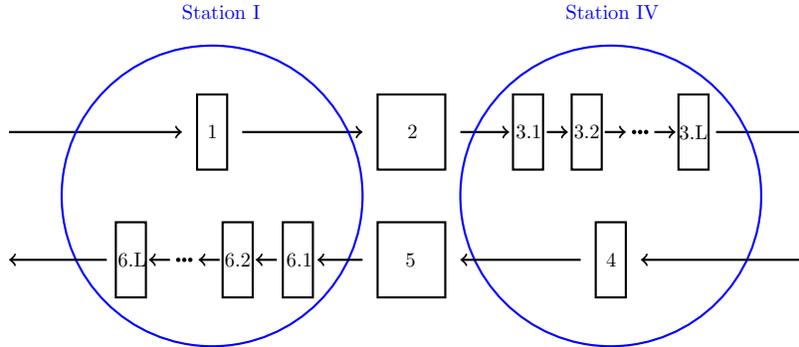


Figure 2: Squares in top row correspond to Classes 1, 2, and 3.1-3.L, in order of appearance along their route; similarly, squares in lower row correspond to Classes 4, 5, and 6.1-6L, in order of appearance. Classes in left circle belong to Station I, classes in right circle belong to Station IV; Class 2 and Class 5 belong to the one-class stations, Stations II and III. All classes at a given station have the same service rule.

do not depend on the class of the job within that station. For symmetric disciplines, the assigned arrival and service distributions are equal to one another at each station. (See the original sources Baskett et al. (1975) and Kelly (1975, 1976, 1979), or the monograph Bramson (2008), for complete definitions.)

Subcritical networks with a symmetric discipline and Poisson external arrivals are stable irrespective of the distributions of service times at individual classes, and the corresponding Markov processes are positive recurrent with equilibria (i.e., stationary distributions) that have an explicit product form. Subcritical networks with a homogeneous discipline, Poisson external arrivals, and exponentially distributed service times that have the same mean at a given station are also positive recurrent, with equilibria that have a similar product form. These two families of disciplines are among the few disciplines for which the equilibria of multiclass queueing networks are explicitly computable.

These results contributed to overly rosy expectations for the stability of subcritical queueing networks for arbitrary work-conserving disciplines, even though their equilibria were not expected to be explicitly calculable. However, examples in various settings later showed stability need not follow from subcriticality (see, e.g., Bramson (1994), Lu and Kumar (1991), Rybko and Stolyar (1992), Seidman (1994)), including for FIFO networks whose

classes at a given station have unequal mean service times.

On the other hand, in Dai (1995) and Rybko and Stolyar (1992), a machinery was developed that enabled one to show stability of queueing networks in a wide range of settings, where external arrivals were allowed to consist of renewal rather than Poisson processes, and no assumptions on the service times of classes were needed. In one such application, subcritical FIFO networks, with classes at a given station having the same mean service time, were shown to be stable (Bramson (1996a)). Stability for subcritical PS networks can be shown in a similar manner, and is discussed in the appendix. Because of the presence of an infinite number of available servers, IS queueing networks are stable in all settings. Little is currently known about the stability of subcritical multiclass queueing networks with the LIFO discipline.

Theorem 1 of this paper shows that subcritical queueing networks with the LIFO discipline need not be stable. A consequence is the absence of a uniform framework for establishing the stability of symmetric disciplines, in contrast to when input is Poisson. On account of Corollary 1, a subcritical LIFO network being of Kelly type is also not sufficient for stability. So there is no uniform framework for establishing the stability of homogeneous disciplines, again in contrast to the Poisson case.

1.2 Overview of the paper

In Section 2, we construct the state space. Since one needs to keep track of residual times for all jobs except those at Classes 2 and 5, where one does not wish to know the residual times, the state space is somewhat nonstandard.

The demonstration of Theorem 1 involves the construction of “cycles”. One shows, at the end of each cycle, that the state of the process is typically an approximate multiple of the state at the beginning of the cycle, except that the roles of Classes 1-3 and Classes 4-6 have been switched. In Section 3, this induction step, Theorem 2, is stated, and Theorem 1 is demonstrated using Theorem 2.

In Section 4, we provide heuristics for the proof of Theorem 2, avoiding the technical details. We also remark on the behavior of the queueing network under certain modifications. The content in this section is optional, but may be helpful to the reader.

Theorem 2 is demonstrated by dividing each cycle into two random time intervals, $[0, S_1]$ and $[S_1, S_1 + S_2]$. The behavior of the network during the much longer time interval $[0, S_1]$ is analyzed in Section 5, where Theorem 3 is demonstrated; most of the technical work in this paper is devoted to

showing Theorem 3. The behavior of the network during the much shorter time interval $[S_1, S_1 + S_2]$ is analyzed in Section 6.

In Subsection 1.1, we were somewhat vague about known stability results for PS queueing networks. In the appendix, we show stability on a dense set of service time distributions for subcritical PS queueing networks whose external arrivals are given by renewal processes. The result follows quickly from results on the stability of subcritical HLPPS queueing networks in Bramson (1996b). An extension to all service times does not follow in an obvious manner.

2 State space construction

In this section, we construct the state space \mathcal{S} of the Markov process corresponding to the family of LIFO queueing networks in Figure 1. The space \mathcal{S} consists of points x of the form

$$x \in (\mathbb{Z} \times \bar{\mathbb{R}} \times \mathbb{R})^\infty \times \mathbb{Z}^2 \times \mathbb{R}^2, \quad (8)$$

subject to appropriate positivity conditions ($\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$). Only a finite number of the coordinates of $(\mathbb{Z} \times \bar{\mathbb{R}} \times \mathbb{R})^\infty$, indexed by i , are assumed to be nonzero. For each such nonzero triple, the first coordinate k_i is to be interpreted as the current class of a job in the network, selected from among the classes $k_i = 1, 3, 4, 6$. The second coordinate s_i measures how long ago such a job entered the class; we set $s_i = \infty$ for a job i originally at the class. The third coordinate v_i measures the residual service time for a job at the class. The first coordinate k_i is given in descending order, followed by s_i , also in descending order. The coordinates z_2 and z_5 of \mathbb{Z}^2 are to be interpreted as the number of jobs at the two remaining classes, Class 2 and Class 5; since they comprise single class stations, it is not necessary to keep track of arrival times of jobs, and since we wish to preserve the memoryless property of the exponentially distributed service times, we do not include the residual times of jobs in the state space descriptor. The coordinates u_1 and u_4 of \mathbb{R}^2 are the residual interarrival times at Classes 1 and 4.

We equip the state space \mathcal{S} with the metric

$$\begin{aligned} d(x, x') &= \sum_{i=1}^{\infty} ((|k_i - k'_i| + |s_i - s'_i| + |v_i - v'_i|) \wedge 1) \\ &\quad + \sum_{i=1}^2 |z_{a_i} - z'_{a_i}| + \sum_{i=1}^2 |u_{b_i} - u'_{b_i}|, \end{aligned} \quad (9)$$

where $a_1 = 2$, $a_2 = 5$, $b_1 = 1$, and $b_2 = 4$. We denote by \mathfrak{S} the standard Borel σ -algebra inherited from the metric.

The Markov process underlying the LIFO queueing network in Figure 1 is defined to be the stochastic process $X(t)$, $t \geq 0$, whose state at any time is given by a point $x_t \in \mathfrak{S}$ that evolves according to the LIFO rule; the accompanying filtration \mathfrak{F}_t is defined in the usual manner. Although this Markov process is not Feller, it is strong Markov. This is not immediate obvious; for more detail on the construction of the Markov process and its strong Markov property, see Bramson (2008), Chapter 4.5. (In our present setting, the definition of coordinates in (8) is slightly different.)

We denote by z_k , $k = 1, \dots, 6$, the number of jobs in each class and by $z = \sum_{k=1}^6 z_k$ the number of jobs in the network. (z_2 and z_5 are employed in (9).) Denote by w_3 and w_6 the immediate workload at Classes 3 and 6, that is, the sum of the residual service times of all of the jobs currently at these classes. (Only the immediate workloads at these classes is used.)

The random analogs of the quantities z_k , z , w_3 , w_6 , u_1 , and u_4 corresponding to $X(t)$ will be denoted by $Z_k(t)$, $Z(t)$, $W_3(t)$, $W_6(t)$, $U_1(t)$, and $U_4(t)$. We will also employ $A_k(t)$ to denote the total number of arrivals to Class k over times $(0, t]$, and by $D_k(t)$ the total number of departures from Class k over this time interval.

3 The induction step

In this section, we state the induction step, Theorem 2, and then prove Theorem 1 assuming Theorem 2. The theorem asserts that, at the random time T , the number of jobs at Class 5 is a large multiple of the number originally at Class 2 and there are few jobs or work elsewhere, if δ is small (and hence M is large).

Theorem 2. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2). Suppose that, for large M , $\delta = 1/M^{1/15}$, and $N \geq 2M/\delta$,*

$$Z_2(0) = N, \quad W_3(0) \leq \delta^2 N, \quad \sum_{k \neq 2,3} Z_k(0) \leq \delta N. \quad (10)$$

Then there exists a stopping time T , satisfying $T \in [N/3\delta, 3N/\delta]$, such that

$$\begin{aligned} Z_5(T) &\geq N/4\delta, & Z_1(T) + Z_2(T) &\leq 10^3 \delta N, \\ Z_3(T) = Z_4(T) &= 0, & W_6(T) &\leq \delta^3 N, \end{aligned} \quad (11)$$

and

$$Z(t) \geq N/4, \quad \text{for all } t \in [0, T], \quad (12)$$

all hold on a set G_N with $P(G_N) \geq 1 - C_\delta e^{-c_\delta N}$ for some $C_\delta, c_\delta > 0$ depending on δ .

Proof of Theorem 1 assuming Theorem 2. Suppose $X(0)$ satisfies the assumptions of Theorem 2 for a given N . Since the evolution of jobs along the upper and lower routes of the network in Figure 1 is symmetric, one can repeatedly iterate the theorem by switching the roles of Classes 1-3 with those of Classes 4-6, and applying the strong Markov property. One obtains in this manner a sequence of stopping times T_0, T_1, T_2, \dots , with $T_0 = 0$ and $T_n \in [(N/3\delta)^n, (3N/\delta)^n]$ for $n \geq 1$, such that

$$Z(t) \geq (1/4\delta)^n \delta N, \quad \text{for all } t \in [T_{n-1}, T_n] \text{ and } n \in \mathbb{Z}_+, \quad (13)$$

holds on a set G_N^∞ , with

$$P(G_N^\infty) \geq 1 - \sum_{n=1}^{\infty} C_\delta e^{-c_\delta (1/4\delta)^n \delta N}. \quad (14)$$

The right hand side of (14) can be made arbitrarily close to 1 by choosing N sufficiently large. On G_N^∞ , one has $\liminf_{t \rightarrow \infty} Z(t)/t > 0$. So, Theorem 1 will follow by showing, on the set where $\liminf_{t \rightarrow \infty} Z(t) < \infty$, that $X(t)$ satisfying (10) must eventually occur for some $N \geq N_0$ and arbitrarily large N_0 .

Since ν has a positive density on $(\gamma M, 2M)$ with M large, one can check that $B := \{z = 0 \text{ and } u_1 \leq 1/M^2\}$ is accessible, with uniform probability by a fixed time, from any state $x \in \mathcal{S}$ with $z \leq z_0$ and given z_0 . So, off of the set where $\lim_{t \rightarrow \infty} Z(t) = \infty$, B will be revisited at arbitrarily large times.

Set $t_1 = 2(2\delta^3 + 1/M^2)N_0$, where N_0 is large. We claim that, for $X(0) \in B$,

$$P(X(t_1) \text{ satisfies (10), for some } N \geq N_0) \geq e^{-4N_0/M/2}. \quad (15)$$

It follows from (15) and the previous paragraph that, off of the set where $\lim_{t \rightarrow \infty} Z(t) = \infty$, $X(t)$ will eventually satisfy (10) for $N \geq N_0$ and arbitrarily large N_0 . This will complete the proof of the theorem upon demonstration of (15).

To demonstrate (15), restart $X(t)$ at $x \in B$. Set $N' = 2N_0$, and denote by $F_{N'}$ the event on which (a) at Class 1, the first N' non-residual interarrival times are each $1/M^2$ (“short”) and the next $\lceil 2\delta^3 N'/M \rceil$ interarrival times are all at least M (“long”) and (b) at Class 4, the first $\lceil (2\delta^3 + 1/M^2)N'/M \rceil$

non-residual interarrival times are all long. On $F_{N'}$, only a few jobs enter the network over $(0, t_1]$, other than the N' jobs corresponding to short interarrivals. Since short interarrivals occur with probability $1 - 1/M$, long interarrivals with probability $1/M$, and $\delta = 1/M^{1/15}$, one can check that

$$P(F_{N'}) \geq e^{-2N'/M} = e^{-4N_0/M}. \quad (16)$$

The service time of jobs at Class 1 is deterministic with $m_1 = \delta^3 \gg 1/M$. On $F_{N'}$, there will therefore be almost N' jobs at Class 1 at time N'/M^2 , and few jobs elsewhere in the network. The service time of jobs at Class 2 is memoryless with $m_2 = 1 - \delta$. After a further elapsed time of $2\delta^3 N'$, there will therefore be, with high probability, few jobs anywhere in the network except at Class 2, where there will be N jobs with $N \approx N' = 2N_0$. At Class 3, the immediate workload will be much less than $\delta^2 N'$. So, $X(t_1)$ will satisfy the assumptions of (10) for some N , with $N \geq N_0$. Together with (16), this implies (15), which completes the proof of the theorem. \square

4 Basic ideas behind the proof of Theorem 2

The proof of Theorem 2 employs reasoning similar in spirit to that used in Lu and Kumar (1991) and Rybko and Stolyar (1992), where the number of jobs in the network increases proportionately over periodic “cycles”, during which the uneven distribution of jobs “starves” stations for work. The reasoning is trickier for the LIFO discipline, both because of the basic nature of the discipline, and because of the possibility that many partially served jobs will accumulate at some of the classes. Here, we motivate the network topology in Figure 1, and the choice of arrival and service times in (1) and (2) used to demonstrate Theorem 2.

The main reason for the choice of uneven interarrival times and short service times at Class 4 is to in effect create a low priority class there. Because of the large gaps in arrivals caused by the rare interarrival times of length at least γM and the much more common extremely short interarrival times of length $1/M^2$, overwhelmingly most arrivals are tightly bunched together, with large gaps in between. Together with the assumptions on the other classes, this will ensure that most of the arrivals at Class 3 occur during these gaps and so, because of the LIFO discipline, receive a higher priority of service than do the bunched together arrivals at Class 4.

In order for this picture to hold, one needs the flow of jobs to Class 3 to be evenly spaced. This is accomplished by the consistent service of jobs

at Class 2: the service rule there is exponentially distributed and hence memoryless, which ensures an even flow of jobs to Class 3 as long as Class 2 is not empty. Moreover, since the mean service time at Class 3 is only slightly greater than it is at Class 2, work can only accumulate slowly at Class 3. (Many mostly served jobs could conceivably accumulate there.)

The mean service time $m_2 = 1 - \delta$ at Class 2 is only slightly less than the mean time for jobs to arrive at Class 1. So, as long as jobs arriving at Class 1 proceed quickly to Class 2, approximately N/δ jobs will need to be served at Class 2 before Class 2 is first empty, if $Z_2(0) = N$. So, the time S_1 at which Class 2 first empties will be approximately N/δ .

By time S_1 , approximately N/δ jobs will have entered the network at Class 4. The jobs at Class 3 will have priority over most of the jobs arriving at Class 4; since Class 3 will be empty only a small fraction of the time before Class 2 is empty, few jobs arriving at Class 4 over $(0, S_1]$ will have completed service by time S_1 .

As reasoned above, there is comparatively little work remaining at Class 3 at time S_1 . There are also comparatively few jobs at any of the other classes, aside from Class 4: Since few jobs are served at Class 4 up until time S_1 , Class 5 will experience a minimal load over that time and so will have few jobs at time S_1 . For the same reason, few jobs will arrive at Class 6 from Class 5. Since $m_1 = \delta^3 \ll 1$, jobs in Class 1 require little service. Consequently, there will be few jobs at the station that comprises Classes 1 and 6 at time S_1 .

To sum up: At time S_1 , Class 4 has approximately N/δ jobs, Class 3 has comparatively little work remaining, and all other classes have comparatively few jobs. Moreover, over $[0, S_1]$, there will always be at least on the order of N jobs at either Class 2 or Class 4, and so at least this many jobs in the network. These conclusions are stated in Theorem 3, which summarizes the behavior of the queueing network up until time S_1 .

We designate by $T = S_1 + S_2$ the time after S_1 at which the station comprising Classes 3 and 4 is first empty. Because of the relatively little work at time S_1 remaining at Class 3, the short service time δ^3 at Class 4, and the small number of jobs arriving from elsewhere over $(S_1, T]$, S_2 will be of order $\delta^2 N$. Because of the relatively short timespan $[S_1, T]$, nearly all of the order of N/δ jobs arriving at Class 5 from Class 4 will still be at Class 5 at time T . This gives the desired lower bound on $Z_5(T)$ in (11) of Theorem 2 and the lower bound in (12) on $Z(t)$, for $t \in (S_1, T]$. Relatively few jobs will have arrived at Class 1 over $(S_1, T]$, and so $Z_1(T) + Z_2(T)$ will be sufficiently small for (11). By the definition of S_2 , $Z_3(T) = Z_4(T) = 0$.

The number of jobs arriving at Class 6, which is of order of magnitude

$\delta^2 N$, will nevertheless be too great for the recursion argument we wish to employ. However, $m_6 - m_5 = \delta^3$ is sufficiently small so that the amount of work $W_6(T)$ that accumulates at Class 6 over the timespan $\delta^2 N$ of $(S_1, T]$ is less than $\delta^3 N$, which is the desired bound on $W_6(T)$ in (11). With this last bound, we have thus motivated all of the bounds in (11) and (12). This completes our motivation behind the proof of Theorem 2.

We conclude this section with some remarks on the choices of service times we have made in (2) and on the stability of the LIFO discipline for single class networks.

Remark 1 Classes 2 and 5 are stipulated to have exponentially distributed service times. The proofs of Theorems 1 and 2 would be essentially the same if we replaced the deterministic times of Classes 1, 3, 4, and 6 by exponentially distributed service times. However, Corollary 1 would then not follow from Theorem 1 since the distributions resulting by adding the service times at Classes 3.1,...,3.L and 6.1,...,6.L would be gamma and not exponentially distributed.

Remark 2 In this paper, we consider preemptive LIFO, rather than nonpreemptive LIFO, where a job currently in service at a class completes its service before more recent arrivals are served. For nonpreemptive LIFO, Theorem 1 and its corollary continue to hold under the same assumptions. In that setting, one has the option of replacing the exponentially distributed service times at Classes 2 and 5 with deterministic times having the same means, since there can be at most one partially served job at each of these classes at any given time and therefore no sudden arrival of many jobs at Classes 3 and 6. Similarly, one no longer needs to employ the immediate workload rather than the number of jobs for bounds at Classes 3 and 6.

Remark 3 Theorem 1 demonstrates the instability of a family of subcritical multiclass queueing networks with the preemptive LIFO discipline. As stated in Remark 2, an analogous result holds for the nonpreemptive LIFO discipline. Are subcritical single class queueing networks with either of these disciplines necessarily stable? For any nonpreemptive discipline of a single class queueing network, it is easy to see that the order of service of jobs at a station does not affect the stability of the queueing network, provided knowledge of their service times is not used. The FIFO discipline is stable for subcritical single class queueing networks, assuming the external arrivals satisfy (59) and (60) (see, e.g., Bramson (2008) for references). Consequently, so are subcritical single class nonpreemptive LIFO queueing networks. Conditions under which subcritical single class queueing networks with the preemptive LIFO discipline are stable or unstable appear not to be known.

5 Behavior up until time S_1

We demonstrate Theorem 2 by dividing the time interval $[0, T]$ into two subintervals, $[0, S_1]$ and $[S_1, T]$, where S_1 is the first time at which Class 2 is empty. The length of $[0, S_1]$ will be of order N/δ , whereas $[S_1, T]$ will be comparatively short. Most of the effort in showing Theorem 2 will be in analyzing the behavior of the queueing network over $[0, S_1]$, which will be done in this section. The main result is Theorem 3, which gives bounds on $Z_k(S_1)$, for $k \neq 3$, and $W_3(S_1)$, as well as a lower bound on $Z(t)$ over $[0, S_1]$.

Theorem 3. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with initial conditions satisfying those of Theorem 2 for some $N \geq 2M/\delta$. Then there exists a stopping time S_1 , with $S_1 \in [N/2\delta, 2N/\delta]$, such that*

$$\begin{aligned} Z_4(S_1) &\in [N/3\delta, 3N/\delta], & Z_2(S_1) &= 0, \\ W_3(S_1) &\leq 7\delta^2 N, & Z_k(S_1) &\leq 7\delta^2 N \quad \text{for } k = 1, 5, 6, \end{aligned} \quad (17)$$

and

$$Z(t) \geq N/3, \quad \text{for all } t \in [0, S_1], \quad (18)$$

all hold on a set G_{S_1} with $P(G_{S_1}) \geq 1 - C_\delta e^{-c_\delta N}$ for some $C_\delta, c_\delta > 0$.

5.1 Two large deviations lemmas

In this subsection, we state two basic large deviation lemmas. The first, Lemma 1, can be obtained by using the moment generating function and Markov's inequality (see, e.g., Theorem 15 and Lemma 5, in Chapter 3 of Petrov (1975)).

Lemma 1. *Let X_1, X_2, \dots be i.i.d. positive random variables with mean μ and $P(X_1 \geq x) \leq e^{-\alpha x}$ for $x \geq x_0$ and some $\alpha, x_0 > 0$. Set $S_n = \sum_{i=1}^n X_i$ and $\beta^* = \beta(\beta \wedge 1)$. Then there exists $c > 0$ such that, for all $\beta > 0$,*

$$P(|S_n - \mu n|/n \geq \beta) \leq e^{-c\beta^* n} \quad \text{for all } n \geq 0. \quad (19)$$

Setting $N_t = \max\{n : S_n \leq t\}$, one obtains by inverting (19) and a bit of calculation:

$$P(|N_t - \mu^{-1}t|/t \geq \beta) \leq Ce^{-c\beta^* t} \quad \text{for all } t \geq 0, \quad (20)$$

for appropriate $C, c > 0$ not depending on β . This bound will be applied repeatedly in the section to $D_2(t)$ and $A_4(t)$, as well as to other departure

and arrival times. Summing (20) over $t = t_0, t_0 + 1, \dots$ and interpolating in between, one obtains that, for appropriate $C > 0$ and each $t_0 \geq 0$,

$$P(|N_t - \mu^{-1}t|/t \geq \beta \text{ for any } t \geq t_0) \leq C(\beta^* \wedge 1)^{-1}e^{-c\beta^*t_0}. \quad (21)$$

(The explicit dependence on β in the bounds (19)-(21) will only be used in Lemma 2 and Proposition 1. Bounds in other applications will be of the form $C_\delta e^{-c_\delta t}$, where the relationship with β is suppressed.)

For the next lemma, consider a queue at which jobs arrive according to a rate- $(1 + \eta)$ Poisson process and each job requires 1 unit of time to be served before exiting the queue. Let W_t denote the immediate workload at time t , that is, the amount of time required for all jobs then at the queue to be served, provided no other jobs arrive. The lemma will be used in Propositions 1 and 4 to obtain bounds on the state at Classes 3 and 4, until Class 2 is empty.

Lemma 2. *Define W_t as above, with $W_0 = 0$, and set $B = \{t \geq 0 : W_t = 0\}$. Then, for $\eta \in (0, 1]$ and appropriate $C, c > 0$,*

$$P(|B| \geq x) \leq (C/\eta^2)e^{-c\eta^2x} \text{ for all } x \geq 0, \quad (22)$$

and

$$P(W_t \geq 2\eta t \text{ for some } t \geq t_0) \leq (C/\eta^2)e^{-c\eta^3t_0} \text{ for all } t_0 \geq 0. \quad (23)$$

Proof. To obtain (22) and (23), we compare the process W_t with $X_t = Y_t - t$, where Y_t is a rate- $(1 + \eta)$ Poisson process and $X_0 = 0$. Set $B^X = \{t \geq 0 : X_t \leq 0\}$. Coupling the processes W_t and X_t by allowing the same random input for each, $W_t \geq X_t$ for all t . Therefore, (22) will follow by showing its analog,

$$P(|B^X| \geq x) \leq (C/\eta^2)e^{-c\eta^2x} \text{ for } x \geq 0, \quad (24)$$

for appropriate $C, c > 0$. Setting $B_t = \{s \in [0, t] : W_s = 0\}$, one has $W_t - X_t = |B_t| \leq |B|$. Plugging this into (22), one can check that (23) will follow from

$$P(X_t \geq 2\eta t \text{ for some } t \geq t_0) \leq (C/\eta^2)e^{-c\eta^2t_0}, \quad (25)$$

for appropriate $C, c > 0$.

The bound in (25) follows directly from (21). To show (24), note that, by integrating (20),

$$\int_{t_0}^{\infty} P(X_t \leq 0)dt \leq (C/c\eta^2)e^{-c\eta^2t_0}, \quad (26)$$

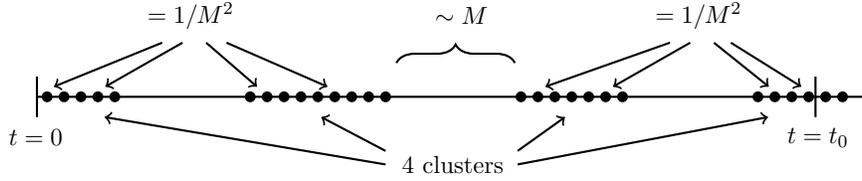


Figure 3: A realization depicting the arrival times of jobs at either Class 1 or at Class 4, with dots indicating these arrival times. In this case, there are four clusters \mathcal{C}_i overlapping $[0, t_0]$. The distance between jobs within a cluster is always $1/M^2$; the distance between clusters is random and approximately M (since $\gamma \sim 1$).

with $C, c > 0$ not depending on η for $\eta \in (0, 1]$. On $|B^X| \geq x$, $X_t \leq 0$ must occur for at least 1 unit of time on $t \geq x - 1$, so

$$P(|B^X| \geq x) \leq \int_{x-1}^{\infty} P(X_t \leq 0) dt \leq (C'/c\eta^2)e^{-c\eta^2 x},$$

for appropriate $C' > 0$, which implies (24) for a new choice of C . □

5.2 Demonstration of Theorem 3

In order to demonstrate Theorem 3, we employ six propositions on the evolution of the queueing network that is due to the continuing service of jobs at Class 2 over $[0, S_1]$. An outline of the reasoning is given in Section 4.

For Proposition 1, we will employ Lemmas 3 and 4. We begin by observing that the interarrival times at Class 4 are either very long or very short. This allows us to decompose the sequence of arrivals into *clusters*, with an individual cluster \mathcal{C} consisting of a finite sequence of jobs where the interarrival times between members of the cluster are each of length $1/M^2$, and these jobs are preceded and followed by interarrival times of length $t \geq \gamma M$. (The first arrival after time 0 is assumed to begin a cluster.) We denote successive clusters by \mathcal{C}_i , $i \in \mathbb{Z}_+$ (see Figure 3). Since $\gamma \sim 1$, the following lemma is immediate.

Lemma 3. *The number of clusters at Class 4 overlapping the time interval $(0, t_0]$ is at most $\lceil 2t_0/M \rceil$ for any $t_0 > 0$.*

We denote by \mathcal{L}_i the number of jobs in \mathcal{C}_i and by U_i the time of arrival of the first job of this cluster. Also denote by Y_i the amount of time on the

interval $(U_i + \mathcal{L}_i/M^2, U_{i+1}] \cap (0, t_0]$ that Class 3 does not have any jobs that arrived after $U_i + \mathcal{L}_i/M^2$. Over $(U_i + \mathcal{L}_i/M^2, U_{i+1}]$, all Class 3 jobs arriving after $U_i + \mathcal{L}_i/M^2$ will have priority over the jobs in Class 4 because of the LIFO rule.

We also introduce \tilde{Y}_i , which is defined in the same way as Y_i , but for a *modified process* where the arrival stream of jobs at Class 3 is now given by a rate- $(1 - \delta)^{-1}$ Poisson process instead of by actual departures from Class 2, and the time interval $(U_i + \mathcal{L}_i/M^2, U_{i+1}]$ corresponding to \tilde{Y}_i is not intersected by $(0, t_0]$. One can couple the original and modified processes so that an arrival at Class 3 for the modified process always occurs whenever an arrival for the original process occurs. The sequence $(Y_i)_{i \in \mathbb{Z}_+}$ is not i.i.d. However, setting

$$A_{t_0} = \{Z_2(t) > 0 \text{ for all } t \in [0, t_0)\}, \quad (27)$$

it is easy to check the following:

Lemma 4. *The sequence $(\tilde{Y}_i)_{i \in \mathbb{Z}_+}$ is i.i.d. For $\omega \in A_{t_0}$ and $U_{i+1} \leq t_0$, $Y_i = \tilde{Y}_i$; for $\omega \in A_{t_0}$ and all i , $Y_i \leq \tilde{Y}_i$.*

Proposition 1 provides an upper bound on the number of jobs that can depart from Class 4 on the event A_{t_0} . The bound is due to “most” arriving jobs at Class 3 having higher priority than “most” jobs at Class 4, and Class 3 seldom being empty. This reasoning will employ the relatively low number of clusters of jobs at Class 4 together with there being “few” jobs in each cluster that can have higher priority than the jobs in Class 3. In the proposition, the choice of the term δ^3 is somewhat arbitrary – we require at least this power, but δ^n could instead be used if we defined $\delta = M^{1/(n+12)}$.

We denote by $D_4^o(t)$ the number of jobs originally in Class 4 that have departed the class by time t

Proposition 1. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with any initial condition. Then, for appropriate $C_\delta, c_\delta > 0$,*

$$P(D_4(t_0) \geq \delta^3 t_0 + D_4^o(t_0); A_{t_0}) \leq C_\delta e^{-c_\delta t_0}, \quad (28)$$

where A_{t_0} is as in (27).

Proof. By time $U_i + \mathcal{L}_i/M^2$, all jobs in \mathcal{C}_i have arrived at Class 4, and jobs arriving at Class 3 after then have priority over these jobs. Consequently, the time over $(U_i, U_{i+1}]$ that is available for service of jobs in Class 4 is at most

$\mathcal{L}_i/M^2 + Y_i$. Because of Lemma 4, the dominating sequence $(\mathcal{L}_i/M^2 + \tilde{Y}_i)_{i \in \mathbb{Z}_+}$ is i.i.d.

Each job at Class 4 requires δ^3 amount of service. Applying Lemmas 3 and 4, it follows that, on A_{t_0} ,

$$D_4(t_0) - D_4^o(t_0) \leq \sum_{i=1}^{\lceil 2t_0/M \rceil} (\mathcal{L}_i/M^2 + Y_i)/\delta^3 \leq \sum_{i=1}^{\lceil 2t_0/M \rceil} (\mathcal{L}_i/M^2 + \tilde{Y}_i)/\delta^3. \quad (29)$$

It follows quickly from (1) that

$$P(\mathcal{L}_i \geq \ell) \leq e^{-\ell/M} \quad \text{for any } \ell \geq 0.$$

On the other hand, by (22) of Lemma 2, with $\eta = \delta^3/2$,

$$P(\tilde{Y}_i \geq x) \leq (C/\delta^6)e^{-c\delta^6 x} \quad \text{for } x \geq 0,$$

and some $C, c > 0$. It follows from these two displays (the main contribution is from the latter) that the summands on the right hand side of (29) are dominated by random variables

$$V_i := (2/c\delta^9)(R_i + \log[2C/\delta^6]),$$

where R_i are independent and mean-1 exponentially distributed.

By (19), for $y \geq y_0 > 0$ and some $C', c' > 0$ depending on y_0 ,

$$P\left(\sum_{i=1}^{b_t} R_i \geq b_t(1+y)\right) \leq C'e^{-c'b_t y}.$$

Substituting in V_i for R_i and applying $\delta = M^{1/15}$, it follows from the above two displays and a bit of computation that

$$\begin{aligned} P\left(\sum_{i=1}^{\lceil 2t_0/M \rceil} V_i \geq \delta^3 t_0\right) &\leq C' \exp\left\{-c'\delta^{12}t_0\left(\frac{c}{2} - 2\delta^3(\log[2C/\delta^6] + 1)\right)\right\} \\ &\leq C'e^{-cc'\delta^2 t_0/4}, \end{aligned}$$

with the second inequality holding since $c/2$ is the dominant term inside of the large parentheses on the right hand side. The desired inequality (28) follows from this display and (29). \square

We denote by S_1 the stopping time at which $Z_2(t) = 0$ first occurs. In the remainder of this section and in Section 6, we will examine the behavior of $X(S_1)$, and then restart the process there. Our first result provides an elementary upper bound on S_1 .

Proposition 2. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with initial conditions satisfying $Z_1(0) + Z_2(0) \leq 2N$ for some N . Then*

$$P(S_1 \geq 2N/\delta) \leq C_\delta e^{-c_\delta N} \quad (30)$$

for appropriate $C_\delta, c_\delta > 0$.

Proof. For any time t , $A_2(t) \leq A_1(t) + Z_1(0)$; adding $Z_2(0)$ to this gives an upper bound on the number of jobs to have visited Class 2 by time t . On the other hand, since the interarrival distribution satisfies (1), with mean 1, and $m_2 = 1 - \delta$, it follows from (20) that

$$\begin{aligned} P(D_2(2N/\delta) \leq A_1(2N/\delta) + Z_1(0) + Z_2(0) ; Z_2(s) > 0 \text{ for all } s \in [0, t]) \\ \leq C_\delta e^{-c_\delta N}, \end{aligned}$$

for appropriate $C_\delta, c_\delta > 0$. Off of the exceptional set in the display, $S_1 < 2N/\delta$, which implies (30). □

The following result is a quick consequence of Propositions 1 and 2. It provides an upper bound on the number of jobs ever to visit Classes 5 and 6 over $t \in [0, S_1]$ and is important for establishing the long-term cyclical growth of $Z(t)$.

Corollary 2. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with initial conditions satisfying $Z_1(0) + Z_2(0) \leq 2N$ for some N . Then*

$$P(D_4(S_1) \geq 2\delta^2 N + Z_4(0)) \leq P(D_4(S_1) \geq 2\delta^2 N + D_4^o(S_1)) \leq C_\delta e^{-c_\delta N} \quad (31)$$

for appropriate $C_\delta, c_\delta > 0$. Hence, denoting by \mathcal{V}_{S_1} the total number of jobs ever to be in either Class 5 or Class 6 over $[0, S_1]$,

$$P\left(\mathcal{V}_{S_1} \geq 2\delta^2 N + \sum_{k=4}^6 Z_k(0)\right) \leq C_\delta e^{-c_\delta N}. \quad (32)$$

We employ Corollary 2 to obtain the following upper bound on $Z_1(t)$, for $t \in [0, S_1]$, and the following lower bound on S_1 .

Proposition 3. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with initial conditions satisfying those of Theorem 2 for some $N \geq 2M/\delta$. Then*

$$P(Z_1(t) \geq 7\delta N \text{ for some } t \in [0, S_1]) \leq C_\delta e^{-c_\delta N} \quad (33)$$

and

$$P(S_1 \leq N/2\delta) \leq C_\delta e^{-c_\delta N} \quad (34)$$

for appropriate $C_\delta, c_\delta > 0$.

Proof. The demonstration of (33) is rather long; by employing (33), the demonstration of (34) will be quick.

Demonstration of (33). For given s , set

$$\begin{aligned} \mathcal{A}_s &= \{Z_1(t) = 0 \text{ for some } t \in [s, s + 3\delta N]\}, \\ \mathcal{B}_s &= \{Z_1(t) \geq 7\delta N \text{ for some } t \in [s + 3\delta N, s + 6\delta N]\}. \end{aligned}$$

We claim that

$$P(\mathcal{A}_s \cap \mathcal{B}_s) \leq C_\delta e^{-c_\delta N} \quad \text{and} \quad P(\mathcal{A}_s^c) \leq C_\delta e^{-c_\delta N} \quad (35)$$

for $s \leq S_1 - 3\delta N$ and appropriate $C_\delta, c_\delta > 0$. On \mathcal{A}_s , there is at most time $6\delta N$ for at least $7\delta N$ jobs to arrive at empty Class 1 in order for \mathcal{B}_s to hold, so the first inequality in the display follows from (1) and (20) applied to $A_1(s + 6\delta N) - A_1(s)$.

For the second inequality, note that, on account of (10) and (32) of Corollary 2, there are typically at most $2\delta N$ jobs in Class 6 to interfere with the service of Class 1 jobs over times $[s, s + 3\delta N]$, and they require only time $2\delta N$ to be served. Also, the service time of Class 1 jobs is δ^3 , $Z_1(0) \leq \delta N$ and, by (30) of Proposition 2, the number of arrivals by time $s + 3\delta N$ is at most $A_1(2N/\delta)$. Applying (20), the time required to serve all Class 1 jobs arriving by time S_1 is therefore typically at most $3\delta^2 N$. Since $2\delta N + 3\delta^2 N < 3\delta N$, the second inequality follows.

By (35),

$$P(\mathcal{B}_s) \leq 2C_\delta e^{-c_\delta N}.$$

Denoting by I the set of s that are multiples of $3\delta N$ and $s \leq S_1 - 3\delta N$, it follows that

$$P\left(\bigcup_{s \in I} \mathcal{B}_s\right) \leq C_\delta e^{-c_\delta N} \quad (36)$$

for another choice of $C_\delta, c_\delta > 0$. The bound in (33), but restricted to $t \in [3\delta N, S_1]$, follows from (36). The extension to $t \in [0, S_1]$ follows from (10), (32), and (20) applied to $A_1(3\delta N)$.

Demonstration of (34). Since $N \geq 2M/\delta$, one has $U_1(0) \leq \delta N$. Therefore, by (1), (2), and (21),

$$P(A_1(t) \leq (1 - \delta^2)t - 2\delta N \text{ for any } t \geq 0) \leq C_\delta e^{-c_\delta N}, \quad (37)$$

$$P(D_2(t) \geq (1 + \delta + 2\delta^2)t + \delta N \text{ for any } t \geq 0) \leq C_\delta e^{-c_\delta N}, \quad (38)$$

for appropriate $C_\delta, c_\delta > 0$. Together with $Z_2(0) \geq N$ and (33), (37) and (38) imply that $Z_2(t) > 0$ for $t \leq N/2\delta$, off of the exceptional sets in (33), (37), and (38). This implies (34). \square

We employ the preceding three propositions to obtain the following upper bound on $W_3(S_1)$ and lower bound on $Z_4(t)$, for $t \in [N/3, S_1]$. We require a bound on $W_3(S_1)$ rather than on $Z_3(S_1)$ because of the possible presence of many mostly served jobs at Class 3.

Proposition 4. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with initial conditions satisfying those of Theorem 2 for some $N \geq 2M/\delta$. Then*

$$P(W_3(S_1) \geq 7\delta^2 N) \leq C_\delta e^{-c_\delta N} \quad (39)$$

and

$$P(Z_4(t) \leq (1 - 5\delta)t \text{ for some } t \in [N/3, S_1]) \leq C_\delta e^{-c_\delta N} \quad (40)$$

for appropriate $C_\delta, c_\delta > 0$.

Proof. Since $N \geq 2M/\delta$, one has $U_4(0) \leq \delta N$. Together with (1) and (21), this implies

$$P(A_4(t) \leq (1 - 4\delta)t \text{ for some } t \geq N/3) \leq C_\delta e^{-c_\delta N}$$

for appropriate $C_\delta, c_\delta > 0$. Together with (28) of Proposition 1, this implies (40).

Since $m_2 = 1 - \delta$, $m_3 - m_2 = \delta^3$, and $W_3(0) \leq \delta^2 N$, it follows from (23) of Lemma 2 that, for given t_0 ,

$$P(W_3(t) \geq 3\delta^3 t + \delta^2 N \text{ for some } t \geq t_0) \leq C_\delta e^{-c_\delta t_0}, \quad (41)$$

for appropriate $C_\delta, c_\delta > 0$. By (30) of Proposition 2 and (34) of Proposition 3, $S_1 \in (N/2\delta, 2N/\delta)^c$ occurs only on an exceptional set. Inequality (39) follows by applying these bounds together with that in (41). \square

The following corollary of Proposition 4 allows us to improve on Corollary 2 by bounding more precisely the number of jobs leaving Class 4 after time $N/3$ and the total number of jobs to be in Classes 5 or 6 over $[N, S_1]$.

Corollary 3. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with initial conditions satisfying those of Theorem 2 for some $N \geq 2M/\delta$. Then*

$$P(D_4^o(S_1) \neq D_4^o(N/3)) \leq C_\delta e^{-c_\delta N}, \quad (42)$$

and

$$P(D_4(S_1) - D_4(N/3) \geq 2\delta^2 N) \leq C_\delta e^{-c_\delta N} \quad (43)$$

for appropriate $C_\delta, c_\delta > 0$. Denoting by $\mathcal{V}_{S_1}^N$ the total number of jobs ever to be in either Class 5 or Class 6 over $[N, S_1]$,

$$P(\mathcal{V}_{S_1}^N \geq 2\delta^2 N) \leq C_\delta e^{-c_\delta N}. \quad (44)$$

for a new choice of $c_\delta > 0$.

Proof. By (40) of Proposition 4,

$$P(Z_4(t) \leq Z_4(0) \text{ for some } t \in [N/3, S_1]) \leq C_\delta e^{-c_\delta N/3}$$

for appropriate $C_\delta, c_\delta > 0$. So, off of the exceptional set, no job originally at Class 4 can depart over $[N/3, S_1]$ because of the LIFO property, and (42) follows. Inequality (43) follows from (31) of Corollary 2.

For (44), note that, by (32) of Corollary 2 and the initial conditions of Theorem 2,

$$P(\mathcal{V}_{S_1} \geq 3\delta N) \leq C_\delta e^{-c_\delta N}$$

for appropriate $C_\delta, c_\delta > 0$. It therefore follows after applying (20) to $A_1(N)$ and $D_5(N)$ that, for some stopping time $S_0 \in [N/3, N]$,

$$P(Z_k(S_0) = 0 \text{ for } k = 1, 5, 6) \geq 1 - C_\delta e^{-c_\delta N}$$

for a new choice of $C_\delta, c_\delta > 0$. Display (44) follows from this and (43). \square

Employing (44) of Corollary 3, we obtain the following stronger version of (33) of Proposition 3.

Proposition 5. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with initial conditions satisfying those of Theorem 2 for some $N \geq 2M/\delta$. Then*

$$P(Z_1(t) \geq 7\delta^2 N \text{ for some } t \in [N, S_1]) \leq C_\delta e^{-c_\delta N}. \quad (45)$$

Proof. The argument is the same as that for (33) except for minor changes. For given s , we now set

$$\begin{aligned}\mathcal{A}_s &= \{Z_1(t) = 0 \text{ for some } t \in [s, s + 5\delta^2 N]\}, \\ \mathcal{B}_s &= \{Z_1(t) \geq 7\delta^2 N \text{ for some } t \in [s + 5\delta^2 N, s + 6\delta^2 N]\}.\end{aligned}$$

The events $\mathcal{A}_s \cap \mathcal{B}_s$ and \mathcal{A}_s^c each occur with low probability: On \mathcal{A}_s , there is at most time $6\delta^2 N$ for at least $7\delta^2 N$ jobs to arrive at empty Class 1 in order for \mathcal{B}_s to hold. On the other hand, because of (44), there are typically at most $2\delta^2 N$ jobs in Class 6 to interfere with the service of Class 1 jobs over times $[s, s + 5\delta^2 N]$. Also, the service time of Class 1 jobs is δ^3 and the time required to serve all of these jobs is typically strictly less than $3\delta^2 N$. So the same reasoning as for (33) implies that \mathcal{A}_s will typically occur.

The remainder of the argument for (45) follows that of the proof of Proposition 3. □

The following proposition estimates $Z_4(S_1)$ and gives a lower bound on $Z(t)$ over $[0, S_1]$; it follows quickly from previous results.

Proposition 6. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2), with initial conditions satisfying (10) of Theorem 2 for some $N \geq 2M/\delta$. Then*

$$P(Z_4(S_1) \in [N/3\delta, 3N/\delta]^c) \leq C_\delta e^{-c_\delta N} \quad (46)$$

and

$$P(Z(t) \leq N/3 \text{ for any } t \in [0, S_1]) \leq C_\delta e^{-c_\delta N} \quad (47)$$

for appropriate $C_\delta, c_\delta > 0$.

Proof. The inequality in (46) follows from the upper and lower bounds on S_1 given in (30) and (34), the lower bound on $Z_4(t)$ in (40), and by applying (20) to $A_4(t)$.

Up until time $N/2$, the inequality for (47) follows from $Z_2(0) = N$ and by applying (20) to $D_2(t)$; on $[N/2, S_1]$, the inequality follows from (31) and by again applying (20) to $A_4(t)$. □

Combining the preceding results, we obtain Theorem 3.

Proof of Theorem 3. The assertion for S_1 follows from (30) of Proposition 2 and (34) of Proposition 3. The assertion for $Z_4(S_1)$ follows from (46) of Proposition 6, and that for $W_3(S_1)$ follows from (39) of Proposition 4. The assertion for $Z_k(S_1)$, for other k , follows from (44) of Corollary 3, (45) of Proposition 5, and $Z_2(S_1) = 0$. The assertion for $Z(t)$ follows from (47) of Corollary 6. □

6 Demonstration of Theorem 2

Proposition 7. *Let $X(t)$ be the Markov process associated with the queueing network in Figure 1 satisfying (1)-(2). Suppose that, for some N and $a \in [1, N]$,*

$$Z_4(0) = N, \quad W_3(0) \leq a\delta^3 N, \quad \sum_{k \neq 3,4} Z_k(0) \leq a\delta^3 N. \quad (48)$$

Then there exists a stopping time S_2 , with $S_2 \leq 4a\delta^2 N$, such that

$$\begin{aligned} Z_5(S_2) &\geq (1 - 5a\delta^2)N, & Z_1(S_2) + Z_2(S_2) &\leq 5a\delta^2 N, \\ Z_3(S_2) = Z_4(S_2) &= 0, & W_6(S_2) &\leq 10a\delta^5 N, \end{aligned} \quad (49)$$

and

$$Z(t) \geq (1 - 5a\delta^2)N, \quad \text{for all } t \in [0, S_2], \quad (50)$$

all hold on a set G_{S_2} with $P(G_{S_2}) \geq 1 - C_\delta e^{-c_\delta N}$ for some $C_\delta, c_\delta > 0$.

Proof. By showing that, off of an exceptional set, the amount of work ever to be present at Station IV over $[0, 4a\delta^2 N]$ is strictly less than $4a\delta^2 N$, it will follow that

$$Z_3(t) = Z_4(t) = 0 \quad \text{for some } t \leq 4a\delta^2 N, \quad (51)$$

which implies the first part of the claim by setting S_2 equal to the first such t .

By (20),

$$P(A_1(4a\delta^2 N) \geq 4a\delta^2(1 + \delta^2)N) \leq C_\delta e^{-c_\delta N} \quad (52)$$

for appropriate $C_\delta, c_\delta > 0$. By (48), off of this exceptional set, the number of jobs arriving at Class 3 over $(0, 4a\delta^2 N]$ is at most $4a\delta^2(1 + \delta/4 + \delta^2)N$. Since $m_3 = 1 - \delta + \delta^3$, it follows from this that the amount of work arriving at Class 3 over $(0, 4a\delta^2 N)$ is at most

$$4a\delta^2(1 + \delta/4 + \delta^2)(1 - \delta + \delta^3)N \leq 4a\delta^2(1 - 3\delta/4 + \delta^2)N; \quad (53)$$

together with $W_3(0) \leq a\delta^3 N$, this implies that the total amount of work ever to be at Class 3 over $[0, 4a\delta^2 N]$ is at most $4a\delta^2(1 - \delta/2 + \delta^2)N$.

The analog of (52), but for Class 4, together with $m_4 = \delta^3$ and $Z_4(0) = N$, implies that the total amount of work ever to be at Class 4 over $[0, 4a\delta^2 N]$ is at most

$$\delta^3(N + 4a\delta^2(1 + \delta^2)N).$$

Adding this bound to the bound in (53) shows that the total amount of work ever to be at Station IV over $[0, 4a\delta^2 N]$ is at most

$$4a\delta^2(1 - \delta/4 + 2\delta^2)N < 4a\delta^2 M,$$

which implies (51).

By applying (20) to $A_1(S_2)$ and $Z_1(0) + Z_2(0) \leq a\delta^3 N$,

$$P(Z_1(S_2) + Z_2(S_2) \geq 5a\delta^2 N) \leq C_\delta e^{-c_\delta N}$$

and, by applying (20) to $D_5(S_2)$ and the definition of S_2 ,

$$P(Z_5(S_2) \leq (1 - 5a\delta^2)N) \leq C_\delta e^{-c_\delta N}, \quad (54)$$

for some $C_\delta, c_\delta > 0$. Moreover, because of $m_1 = m_6 - m_5 = \delta^3$ and the above two bounds on $A_1(S_2)$ and $D_5(S_2)$, the total amount of work ever to be at Station I over $[0, S_2]$, and hence at Class 6, is at most

$$2\delta^3 \cdot 5a\delta^2 N = 10a\delta^5$$

off an exceptional set. Since $Z_3(S_2) = Z_4(S_2) = 0$ by the definition of S_2 , this completes the demonstration of (49). The same reasoning as for (54) implies (50), which completes the proof of the proposition. \square

The proof of Theorem 2 follows quickly from Theorem 3 and Proposition 7 by setting $N = Z_4(S_1)$ and $a = 63$ in Proposition 7. (The factor 63 is used because of the possible range of 9 for $Z(S_1)$ in (49) and the coefficient 7 in the bounds on $Z_k(S_1)$.)

Proof of Theorem 2. Setting $T = S_1 + S_2$, the lower bound on T is immediate from Theorem 3 and the upper bound also follows because $S_2 \leq N$ off of the exceptional set in Proposition 7. The bound on $Z_5(T)$ follows from the lower bound on $Z_4(S_1)$ in (17) and the lower bound on $Z_5(S_2)$ in (49). The bound on $Z_1(T) + Z_2(T)$ follows from that in (17) with $a = 63$; since Station IV is empty at time T , $Z_3(T) = Z_4(T) = 0$. The bounds on $W_3(S_1)$

in (17) and $W_6(S_2)$ in (49) imply the bound on $W_6(T)$. The lower bound on $Z(t)$ over $[0, T]$ follows from the corresponding bounds in (18) and (50). \square

Remark 4 We commented after Theorem 1 that the analog of (4) holds for the total amount of work in the network. More precisely, we denote by $\mathcal{W}(t)$ the *total workload* in the network at time t , that is, the sum of the immediate workload due to the residual service times of jobs currently at their respective classes, together with the service times of these jobs at all classes they will visit before leaving the network. Then

$$\mathcal{W}(t) \rightarrow \infty \quad \text{almost surely as } t \rightarrow \infty. \quad (55)$$

We sketch here the argument for (55), using bounds from the demonstration of Theorem 2. Under the assumptions in (10) of Theorem 2, the reasoning for (47) of Proposition 6 implies that

$$P(\text{both } Z_2(t) \leq N/3 \text{ and } Z_4(t) \leq N/3 \text{ for any } t \in [0, S_1]) \leq C_\delta e^{-c_\delta N} \quad (56)$$

and the reasoning for (54) of Proposition 7 implies that

$$P(Z_4(t) + Z_5(t) \leq (1 - 5a\delta^2)N \text{ for any } t \in [S_1, T]) \leq C_\delta e^{-c_\delta N}, \quad (57)$$

where, in each case, $C_\delta, c_\delta > 0$. In particular, (56) was obtained by bounding the number of jobs at Class 2, at time 0, that can leave Class 2 over $[0, N/2]$, and the number of jobs at Class 4, at time $N/2$, that can leave Class 4 over $[N/2, S_1]$; and (57) was obtained by bounding the number of jobs that can leave Class 5 over $[S_1, T]$.

The service laws at Classes 2 and 5 are each exponentially distributed with mean $1 - \delta$, and jobs currently at Class 4 must pass through Class 5 before leaving the network. Together with the previous paragraph and elementary large deviations estimates, these observations imply that

$$P(\mathcal{W}(t) \leq N/6 \text{ for any } t \in [0, T]) \leq C_\delta e^{-c_\delta N} \quad (58)$$

for appropriate $C_\delta, c_\delta > 0$. The limit (55) follows from (58) and the same induction argument as for (14) in the proof of Theorem 1.

7 Appendix

As mentioned in Subsection 1.1, the processor sharing (PS) discipline is stable for all subcritical queueing networks with Poisson input and exponentially distributed service times; its equilibrium distribution can be written

explicitly in its famous “product form”. As with symmetric queueing networks in general, the exponentially distributed service law can be relaxed to a mixture of Erlang distributions. (An Erlang distribution is the distribution of a sum of i.i.d. exponentially distributed random variables.) This generalization employs the *method of stages* (see, e.g., Bramson (2008), Kelly (1979)), and the resulting equilibria are again explicit. It is not difficult to check that mixtures of Erlang distributions are dense in the weak topology in the set of distribution functions (see, e.g., Exercise 3.3.3 in Kelly (1979).)

Consider a subcritical queueing network with Poisson input, but arbitrary service distribution. Because of the explicit nature of the equilibria for mixtures of Erlang distributions, one can choose a sequence of queueing networks whose service distributions converge to the service distributions of the given network and the corresponding sequence of equilibria is tight. Because of this, the above representation of equilibria extends to all subcritical PS queueing networks with Poisson input (Barbour (1976)). In particular, these queueing networks with general service distributions are stable.

This explicit representation no longer holds when external arrivals are generalized to renewal processes. However, for networks with general interarrival distributions (assuming only (59) and (60) below) but where the service times are exponentially distributed, one can compare PS networks with networks with the head-of-the-line processor sharing (HLPPS) discipline. The service rule for the HLPPS discipline is the same as that for PS, except that all service a class receives is devoted to the earliest arriving job at that class, rather than being spread out uniformly among jobs at that class. When the service distributions are exponentially distributed, the specific rule assigning service within a class does not affect the rate at which jobs leave the class, and so processes with the PS and HLPPS disciplines and exponentially distributed service times have the same law.

The stability of subcritical HLPPS networks can be shown under interarrival distributions satisfying (59) and (60), and general service distributions by employing the standard machinery of fluid limits (Corollary 1 of Theorem 1, in Bramson (1996b)). This technique unfortunately produces little qualitative information about the nature of the equilibria for the corresponding queueing networks.

In this appendix, we make two observations. First, that the connection between the PS and HLPPS disciplines, together with the method of stages, enables one to quickly show, using known results, the stability of subcritical PS networks, with interarrival times satisfying (59) and (60), for a dense family of service distributions. This is done in Proposition 9 below.

The second observation is that it nevertheless appears to be difficult to

extend this result to all service distributions. One cannot use the above argument that was applied for Poisson input without somehow first showing the tightness of the equilibria for the corresponding sequence of queueing networks. Because of lack of a direct characterization of these equilibria, it is not clear how to proceed. Nevertheless, based on “obvious” intuition, such stability should hold for subcritical PS queueing networks with both general renewal input and service distributions.

For Proposition 9, we first state Corollary 1 of Theorem 1 (Bramson (1996b)) when the service times are exponentially distributed. Two assumptions on the interarrival times are required: The interarrival time distributions are unbounded, that is, denoting by $\xi_k(i)$, $i \in \mathbb{Z}_+$, the i.i.d. interarrival times at a class k with external arrivals,

$$P(\xi_k(1) \geq x) > 0 \quad \text{for all } x. \quad (59)$$

Moreover, for some $\ell_k > 0$ and non-negative $q_k(\cdot)$, with $\int_0^\infty q_k(x)dx > 0$,

$$P(\xi_k(1) + \cdots + \xi(\ell_k) \in dx) \geq q_k(x)dx, \quad (60)$$

that is, the above sum dominates Lebesgue measure in an appropriate sense. We note that both properties are only needed to ensure that all states communicate with one another; they are not needed to show that the total number of jobs in the network $Z(t)$ is tight (without these conditions, the residual interarrival times could synchronize in some manner).

Proposition 8. *Any subcritical HLPPS queueing network with exponentially distributed service times and whose external interarrival times satisfy (59) and (60) is stable.*

In Section 1, LIFO queueing networks, with routing given by Figure 1, were reinterpreted as LIFO networks, with routing given by Figure 2, by decomposing the service time at a class into service times at successive classes at the same station. Since the PS discipline is symmetric, the same reasoning can be applied to it as well; in fact, since service is assigned uniformly to all jobs within a station, it is easy to see that any reclassification of classes within a station will not affect service.

This reasoning can be applied to service times that are mixtures of Erlang distributions, as in the method of stages. One thus extends results on the equilibria for subcritical PS networks with renewal external arrivals and exponentially distributed service times to subcritical PS networks with renewal external arrivals and service times that are mixtures of Erlang distributions. This reasoning was applied for Poisson external arrivals (see,

e.g., Kelly (1979)). In the current setting, one does not obtain an explicit formula for equilibria, but stability nevertheless follows:

Proposition 9. *Any subcritical PS queueing network whose external interarrival times satisfy (59) and (60) and whose service times are mixtures of Erlang distributions is stable.*

Proof. By Proposition 8, any HLPPS network whose external interarrivals satisfy (59) and (60) and whose service times are exponentially distributed is stable. By the above reasoning, the same queueing network, but with the PS rather than the HLPPS discipline, is also stable. On account of the PS discipline, the evolution of a queueing network will be the same when classes at a given same station are combined into a single class. It follows from this that a subcritical PS network whose external interarrival times satisfy (59) and (60) and whose service times are mixtures of Erlang distributions will be stable.

□

References

- Barbour AD (1976) Networks of queues and the method of stages. *Advances in Appl Probability* 8(3):584–591, DOI 10.2307/1426145, URL <https://doi-org.ezp3.lib.umn.edu/10.2307/1426145>
- Baskett F, Chandy K, Muntz R, Palacios F (1975) Open, closed, and mixed networks of queues with different classes of customers. *J ACM* 22(2):248–260
- Bramson M (1994) Instability of FIFO queueing networks. *The Annals of Applied Probability* 4(2):414–431, DOI 10.1214/aoap/1177005066
- Bramson M (1996a) Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems Theory Appl* 22(1-2):5–45
- Bramson M (1996b) Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems Theory Appl* 23(1-4):1–26, DOI 10.1007/BF01206549, URL <https://doi-org.ezp3.lib.umn.edu/10.1007/BF01206549>
- Bramson M (2008) Stability of queueing networks. *Probab Surv* 5:169–345, DOI 10.1214/08-PS137

- Dai J (1995) On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability* 5(1):49–77
- Kelly F (1975) Networks of queues with customers of different types. *Journal of Applied Probability* 12(3):542–554
- Kelly F (1979) *Reversibility and Stochastic Networks*. Wiley, Chicester
- Kelly FP (1976) Networks of queues. *Advances in Appl Probability* 8(2):416–432, DOI 10.2307/1425912, URL <https://doi-org.ezp2.lib.umn.edu/10.2307/1425912>
- Lu SH, Kumar P (1991) Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control* 36(12):1406–1416
- Petrov VV (1975) *Sums of independent random variables*. Springer-Verlag, New York-Heidelberg, translated from the Russian by A. A. Brown, *Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82*
- Rybko A, Stolyar A (1992) Ergodicity of stochastic processes describing the operation of open queueing networks. *Problemy Peredachi Informatsii* 28(3):3–26
- Seidman TI (1994) ” first come, first served” can be unstable! *IEEE Transactions on Automatic Control* 39(10):2166–2171