Positive Harris Recurrence and Diffusion Scale Analysis of a Push Pull Queueing Network

Yoni Nazarathy *
yonin@stat.haifa.ac.il

Gideon Weiss *
gweiss@stat.haifa.ac.il

Department of Statistics
The University of Haifa
Mount Carmel 31905, Israel

ABSTRACT

We consider a push pull queueing system with two servers and two types of jobs which are processed by the two servers in opposite order, with stochastic generally distributed processing times. This push pull system was introduced by Kopzon and Weiss, who assumed exponential processing times. It is similar to the Kumar-Seidman Rybko-Stolyar (KSRS) multi-class queueing network, with the distinction that instead of random arrivals, there is an infinite supply of jobs of both types. Thus each server can either process jobs of one of the types, which it pulls from the other server, or jobs of the other type which it pushes out of the infinite supply towards the other server. Unlike the KSRS network, we can find policies under which our push pull network works at full utilization, with both servers busy at all times, and without being congested. We perform an asymptotic analysis of the push pull network under these policies to quantify its behavior: We show that under fluid scaling the fluid model of the network is stable. We adapt the proofs of Dai, to show that as a result the queues of jobs waiting for pull operation are positive Harris recurrent. Finally we obtain the diffusion scale behavior of the network, in which we show that the queues are zero under diffusion scaling, and calculate the Brownian approximation of the output processes of the two types of jobs. The approximation shows that the two output streams are highly negatively correlated.

Categories and Subject Descriptors

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

Keywords

Queueing networks, push pull, infinite virtual queues, fluid models, positive Harris recurrence, diffusion approximations.

*Research supported in part by Israel Science Foundation Grant 454/05 and by European Network of Excellence Euro-NGI.

ValueTools 2008, October 21 – 23, 2008, Athens, GREECE.

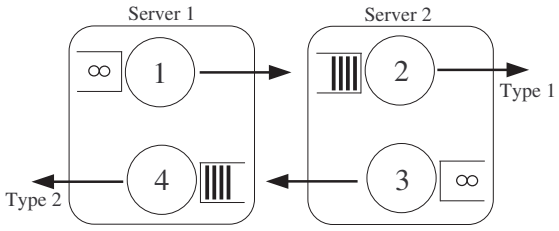

Figure 1: The Push Pull Network.

1. INTRODUCTION

We consider the following queueing network: There are two servers, numbered 1,2 and two types of jobs numbered 1,2 each of which is processed by both servers. Type 1 is processed by server 1 (activity 1) and then by server 2 (activity 2), while type 2 is first processed by server 2 (activity 3) and then by server 1 (activity 4), see Figure 1. Activities 1 and 3 are called *push activities*, in which a server is processing a job and pushing it to queue in front of the other server. Activities 2 and 4 are called *pull activities*, in which a server is processing a job which it pulls from its queue between the servers, and sends it out of the network.

The special feature of this push pull network is that there is no arrival stream. Instead we assume that each server has an infinite supply of jobs available for its push operation. Thus there are two queues in the network: jobs of type 1, waiting for activity 2, are queued at server 2, and jobs of type 2, waiting for activity 4, are queued at server 1.

Infinite supply of work expresses an ability to control the arrivals and is often a reasonable way to model a processing system. In some situations there may indeed be an infinite supply of work — in a communication system a transmitter may have a constant supply of messages generated on the spot in addition to serving messages in transit from other transmitters. In manufacturing systems the supply of parts for processing at an expensive machine may be monitored and not allowed to run out. We refer to this as an infinite virtual queue: it acts like an infinite queue while in fact it only contains a few jobs which are constantly replenished. In standard queueing networks one can regard the input stream as the output of a server which is fed by an infinite supply of work.

We denote by $Q_i(t), i = 2, 4$ the number of jobs in the

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

two queues at time t (including the job in process), and by $D_i(t), i = 1, 2, 3, 4$ the number of jobs that have completed activity i in the time interval $[0, t]$. When $Q_4(t) > 0$, server 1 can either pull, by serving a type 2 job from $Q_4(t)$ or push, by serving a type 1 job from the infinite supply. When $Q_4(t) = 0$ server 1 can still always push jobs of type 1. Hence, server 1 never needs to idle. Similarly for server 2.

What we show in this paper is that it is possible to find policies which never idle and yet keep the queues $Q_i(t)$ stable. Assume that the long term average processing time for activity i is $1/\mu_i$, $i = 1, 2, 3, 4$. Let θ_i, $i = 1, 2, 3, 4$ be the long term fraction of time spent in activity i. If the system never idles then for server 1, $\theta_1 = 1 - \theta_4$, and for server 2, $\theta_3 = 1 - \theta_2$. Furthermore, if $Q_i(t)$ are stable then their input and output rates are equal, so: $\nu_1 = \nu_2 = \theta_1\mu_1 = \theta_2\mu_2$, $\nu_3 = \nu_4 = \theta_3\mu_3 = \theta_4\mu_4$, where ν_i is the long term average rate of the departure process D_i, $i = 1, 2, 3, 4$, and in particular ν_2 (ν_4) is the rate at which jobs of type 1 (type 2) leave the network. Solving the equations we get:

$$\nu_1 = \nu_2 = \frac{\mu_1\mu_2(\mu_3 - \mu_4)}{\mu_1\mu_3 - \mu_2\mu_4}, \quad \nu_3 = \nu_4 = \frac{\mu_3\mu_4(\mu_1 - \mu_2)}{\mu_1\mu_3 - \mu_2\mu_4}.$$

We now specify the policies which we use. We consider preemptive resume head of the line policies. We need to distinguish different cases:

Inherently stable network: When $\mu_1 < \mu_2$ and $\mu_3 < \mu_4$, service of each type of jobs alone, by its second server, is a stable single server queue. In this case the policy which we use is preemptive resume head of the line priority for pull activities 4 and 2 over push activities 1 and 3. We refer to this as *Case 1*, and to the policy as *pull priority policy.*

Inherently unstable network: When $\mu_1 > \mu_2$ and $\mu_3 > \mu_4$, service of each type of jobs alone, by both servers results in an unstable single server queue. In this case priority to pull over push is unstable. A policy that works here is that while $Q_2(t)$ is below some threshold level server 1 will push work to server 2, and server 1 will only pull from $Q_4(t)$ when $Q_2(t)$ is above the threshold, with a similar rule for server 2. We use a linear threshold to determine pull or push preemptive head of the line priority. We define a family of such policies, each determined by a pair of constants κ_1, κ_2 which satisfy $\kappa_1 > \frac{\mu_3}{\mu_1}$, $\kappa_2 > \frac{\mu_1}{\mu_3}$:

Server 1: Priority to pull activity 4 over push activity 1 if $0 < Q_4(t) < \kappa_1 Q_2(t)$,

Server 2: Priority to pull activity 2 over push activity 3 if $0 < Q_2(t) < \kappa_2 Q_4(t)$,

We refer to this as *Case 2*, and to the policy as *linear threshold policy*, see Figure 2.

Unbalanced network: If $\mu_1 > \mu_2$ and $\mu_4 > \mu_3$, then server 2 has more work to do than server 1, for both types of jobs, and the network cannot be stable unless server 1 idles some of the time. Similarly for $\mu_1 < \mu_2$ and $\mu_4 < \mu_3$. We will not consider this case any further in this paper.

Completely balanced network When $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$ it is possible to find policies which work with full utilization of both servers, and which are rate stable, i.e. they satisfy $\nu_1 = \nu_2$ and $\nu_3 = \nu_4$, however these rates are not uniquely determined. We can choose $0 \leq \theta \leq 1$, and specify $\theta_1 = \theta_2 = 1 - \theta_3 = 1 - \theta_4 = \theta$, and use $\nu_i = \mu_i\theta_i$ as nominal rate. As shown in [26], we can use an adaptation of the maximum pressure policy of Dai and Lin [7] to serve jobs of types 1 and 2 at these rates, under full utilization. However, the system will become congested, with expected $O(\sqrt{T})$ jobs in the system at time T. We conjecture that this cannot be improved.

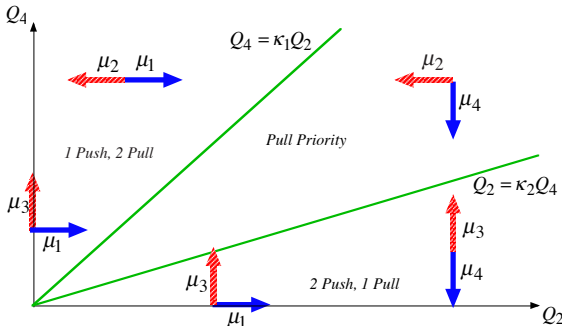

Figure 2: The linear threshold policy for the inherently unstable network (Case 2).

The structure of the paper is as follows: In Section 3 we define our stochastic model and primitive assumptions. In Section 4 we analyze the fluid limit model of this network under fluid scaling, and show that the fluid model is stable in both parametric cases under the corresponding policies. In Section 5 we assume i.i.d. processing times and formulate the network and policy as a Markov process. We then follow the proof method of Dai [6] to show that this Markov process is positive Harris recurrent, and so $Q_2(t), Q_4(t)$ possess a stationary limiting distribution. In Section 6 we consider the output processes $D_i(t)$ under diffusion scaling, and obtain a Brownian approximation.

To emphasize the novelty of our results we start with a preliminary discussion in Section 2, in which we outline known results about the well studied Kumar-Seidman Rybko-Stolyar (KSRS) network, and contrast them with the very different behavior of our push pull network.

Bibliographic note: The push pull network was introduced by Kopzon et al. [17, 18] who assumed exponential processing times. Infinite supply of work and infinite virtual queues are discussed in [1, 2, 12, 13, 26, 28].

2. PRELIMINARY DISCUSSION: COMPARING TO THE KSRS NETWORK

The Kumar-Seidman Rybko-Stolyar multi-class queueing network (c.f. Chapter 8 of [5] or Section 2.9 of [20]) differs from our push pull network in that instead of infinite supply of jobs there are two stochastic arrival streams of jobs of type 1 and of type 2, with long term average arrival rates α_1, α_3.

In that case there are 4 queues $Q_i(t)$ of jobs waiting for activities $i = 1, 2, 3, 4$ in the network, and the offered loads

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

for servers 1 and 2 are $\rho_1 = \alpha_1/\mu_1+\alpha_3/\mu_4$ and $\rho_2 = \alpha_3/\mu_3+\alpha_1/\mu_2$ respectively. A necessary condition for stability is $\rho_i < 1$, $i = 1, 2$.

The same two cases of parameters reappear: If $\mu_1 < \mu_2$ and $\mu_3 < \mu_4$ then $\rho_i < 1$, $i = 1, 2$ is sufficient for stability of the network under any work conserving (i.e. any non idling) policy. On the other hand, if $\mu_1 > \mu_2$ and $\mu_3 > \mu_4$ then $\rho_i < 1$, $i = 1, 2$ may not be sufficient for stability. In particular, there exist $\gamma_i < 1$ such that the last buffer first served policy, which gives priority to the pull activities 2 and 4, will not be stable for $\gamma_i < \rho_i < 1$, $i = 1, 2$.

The discovery of this phenomenon by Kumar and Seidman [19] (deterministic processing times) and by Rybko and Stolyar [27] (exponential processing times) revolutionized research on multi-class queueing networks, and it is now realized that stability is not a property of the network, but of the policy in conjunction with the network. In our network, this is exemplified by the need to use the pull priority (last buffer first served) for the inherently stable Case 1, and a different policy for the inherently unstable Case 2.

Nevertheless, if $\rho_i < 1$, $i = 1, 2$ then there are some work conserving (non idling) policies which keep all four queues of the KSRS network stable. However, as ρ_i increase towards 1, either for one of the servers or for both together, the network becomes increasingly congested under any policy.

Of particular interest is the behavior of multi-class queueing networks under balanced heavy traffic conditions (c.f. [14]). Balanced heavy traffic in the KSRS network occurs when $\alpha_1 \to \nu_1$, $\alpha_3 \to \nu_3$. When this happens queues at both servers become congested under any policy.

A diffusion scale analysis of KSRS under balanced heavy traffic considers a sequence $n = 1, 2, \ldots$ of networks, parameterized by α_i^n, $i = 1, 3$ such that $\sqrt{n}(\alpha_i^n - \nu_i)$ converges to some constant as $n \to \infty$. In that case one can hope to show that the diffusion scaled queues,

$$\hat{Q}_i^n(t) = \frac{Q_i^n(nt)}{\sqrt{n}},\ i = 1, 2, 3, 4,$$

will converge to a 4 dimensional Reflected Brownian Motion.

Recent results of Dai and Lin [7, 8] and Ata and Lin [3] show that with the use of a maximum pressure policy, $(\hat{Q}_1^n(t), \ldots, \hat{Q}_4^n(t))$, converges to a 4 dimensional reflected Brownian motion which is actually lifted from a 2 dimensional workload process. Henderson, Meyn and Tadic [16] also considered the KSRS network and obtained stablity. Their policy uses affine switching curves, and is similar to our linear threshold policy for the push pull network.

As the scaling indicates, for the KSRS network under balanced heavy traffic, the diffusion approximation relates to a sequence of networks in which the total number of jobs in the nth network at any time is expected to be of order $\Theta(\sqrt{n})$.

The behavior of the push pull network, as we will show, is of an entirely different nature: Both servers are active all the time, which can be thought of as operating at $\rho_i = 1$, $i = 1, 2$ and jobs leave the network at the rates ν_i, $i = 1, 3$. At the same time, with i.i.d. processing times the network is positive Harris recurrent. Thus in the push pull network with $\rho_i = 1$ the number of jobs in the queues $Q_2(t), Q_4(t)$ is expected to be $O(1)$, and it is 0 under diffusion scaling.

Finally we now compare the behavior of the output processes, $D_i(t)$, $i = 1, 2, 3, 4$ in the KSRS network and in the push pull network, under diffusion scaling. In the KSRS network with $\rho_i < 1$, $i = 1, 2$ the diffusion scaled queue lengths will be 0. Therefore on a diffusion scale, jobs of type 1 have arrivals, departures from queue 1, and departures from queue 2, which are all identical Brownian motions. Similarly for type 2. In particular, the diffusion scaled flow of jobs of type 1 and of jobs of type 2 will be independent. This fully describes the diffusion scale behavior, for fixed $\rho_i < 1$.

Under balanced heavy traffic the behavior of the output processes of the KSRS network seems to be much more complex. The four queue length processes will be reflected Brownian processes, and will affect the diffusion scaled output processes. To the best of our knowledge the behavior of the output processes in that case has not been investigated. We note that even the output process of a single server queue, under balanced heavy traffic, poses some as yet unanswered questions (c.f. [15] and [25]).

In contrast to that, in the push pull network, operated with our policies, under full utilization, the diffusion scaled queue lengths are 0. As a result we can analyze the output processes of the two types of jobs. What we find is that the output processes of jobs of types 1 and 2 that leave the network converge under diffusion scaling to two standard Brownian motions, but these two Brownian motions are highly negatively correlated.

2.1 Notation

We use $\mathbb{R}_+^d$ and $\mathbb{Z}_+^d$ to denote the sets of all d-dimensional non-negative real and integer vectors respectively. For a vector $x \in \mathbb{R}_+^{d_1} \times \mathbb{Z}_+^{d_2}$ we let $|x|$ denote the ℓ_1 norm, given by sum of absolute values of the components. For a metric space $\mathbb{S}$, we denote by $\mathcal{B}(\mathbb{S})$ the Borel sets of $\mathbb{S}$. The transpose of a matrix $\mathbf{A}$ is $\mathbf{A}'$. We use $\mathbb{D}^{\,d}[0,\infty)$ to denote the set of functions $f : [0,\infty) \mapsto \mathbb{R}_+^d$ that are right continuous with left limits. For $f \in \mathbb{D}^{\,d}[0,\infty)$, we let $||f||_t = \sup_{0\le s\le t} |f(s)|$. We endow the function space $\mathbb{D}^{\,d}[0,\infty)$ with the usual Skorohod J_1-topology. For a sequence of stochastic processes $\{X^r\}$ taking values in $\mathbb{D}^{\,d}[0,\infty)$, we use $X^r \Rightarrow X$ to denote that X^r converges to X in distribution as $r \to \infty$. A sequence of functions $\{f_r\} \subset \mathbb{D}^{\,d}[0,\infty)$ is said to converge to $f \in \mathbb{D}^{\,d}[0,\infty)$ uniformly on compact sets (u.o.c.), if for each $t \ge 0$, $\lim_{r\to\infty} ||f_r - f||_t = 0$. A function $f : \mathbb{S} \mapsto \mathbb{R}$ on a metric space $\mathbb{S}$ is called *lower semi-continuous* if the sets $\{x \in \mathbb{S} : f(x) \le c\}, c \in \mathbb{R}$ are closed. In general, when no ambiguity may arise, we omit index subscripts when we refer to vectors. For a vector x we use $|x|$ to denote L_1 norm. We use $I\{\cdot\}$ for indicator function of event $\{\cdot\}$.

3. THE STOCHASTIC MODEL

We assume that the processing durations of the jobs in activity $i = 1, 2, 3, 4$ are drawn from a sequence of positive random variables: $\xi_i = \{\xi_i^j,\ j = 1, 2, \ldots\}$. The assumptions that we make regarding the processing durations are as follows:

$$(A1)\quad \lim_{n\to\infty} \frac{\sum_{j=1}^n \xi_i^j}{n} = \frac{1}{\mu_i},\ \text{a.s.}$$
$$\text{for some } \mu_i \in (0,\infty),\ i = 1, 2, 3, 4.$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

$$(A2)\quad \begin{cases} (a) & \xi_i,\ i=1,2,3,4 \\ & \text{are mutually independent i.i.d.} \\ (b) & P(\xi_i^1 \geq x) > 0 \text{ for all } x > 0,\ i=1,3. \\ & \exists k_0^i > 0,\ q_i(\cdot) \geq 0 \text{ with } \int_0^\infty q_i(x)dx > 0 : \\ & P(\xi_i^1 + \ldots + \xi_i^{k_0^i} \in dx) \geq q_i(x)dx,\ \ i=1,3. \\ (b') & \text{Compact sets are petite.} \end{cases}$$

$$(A3)\quad \mu_i^2 \mathrm{Var}(\xi_i^1) = c_i^2, \quad \text{for some } c_i^2 \in [0,\infty),\ \ i=1,2,3,4.$$

Assumptions (A1) require that there exist strong laws of large numbers for the sequences of processing times and that the rate of processing of activity i be μ_i. Assumptions (A2) are to be used in a Markov process setting. (a) implies renewal processing. (b) States that the processing times of the push operations are unbounded and spread-out. (b') is a technical assumption to be made precise in Section 5. It is used to prove positive Harris recurrence. We show that under the pull priority policy, (b) implies (b'). Assumptions (A3) require existence of second moments, with squared coefficients of variation c_i^2. We shall make use of Assumptions (A1)-(A3) incrementally.

We associate counting processes with each activity i:

$$S_i(t) = \sup\{n : \sum_{j=1}^{n} \xi_i^j \leq t\}, \qquad t \geq 0.$$

We denote by $T_i(t)$, $i = 1,2,3,4$, the total time that the server allocates to the processing of activity i during the interval $[0,t]$. We require that $T_i(0) = 0$ and that $T_i(\cdot)$ be nondecreasing. Under our policies of full utilization, the servers never idle, thus:

$$T_1(t) + T_4(t) = t, \qquad T_2(t) + T_3(t) = t. \tag{1}$$

Note that $T_i(\cdot)$ are Lipschitz, and are therefore absolutely continuous. Thus their derivative exists almost everywhere with respect to Lebesgue measure on $[0,\infty)$.

The number of jobs that have completed processing of activity i by time t is $D_i(t) = S_i(T_i(t))$. Let $Q_i(0)$, $i = 2,4$ be the initial queue lengths. The number of jobs at time t is:

$$Q_i(t) = Q_i(0) + D_{i-1}(t) - D_i(t), \quad i = 2,4. \tag{2}$$

We further require that $Q_i(\cdot) \geq 0$ for $i = 2,4$.

The policies which we use in the two cases impose additional conditions on the dynamics of the queues. In the inherently stable Case 1, we use pull priority policy. Hence we will not serve activities 1 or 3 (push activities) unless the corresponding Q_4 or Q_2 are empty. This implies that the allocation processes $T(\cdot)$ need to satisfy:

$$\begin{aligned} \textstyle\int_0^t Q_4(s)dT_1(s) &= 0, \\ \textstyle\int_0^t Q_2(s)dT_3(s) &= 0. \end{aligned}$$

In the inherently unstable Case 2, we use a linear threshold policy. The linear threshold for server 1 is the line $Q_4(t) = \kappa_1 Q_2(t)$. Server 1 will give preemptive priority to activity 4 only if $0 < Q_4(t) < \kappa_1 Q_2(t)$, and in that case it will not allocate time to activity 1. On the other hand, if $Q_4(t) \geq \kappa_1 Q_2(t)$ then server 1 will give priority to activity 1, to prevent starvation at the queue of server 2, and will not allocate time to activity 4. A symmetric rule is used by server 2, with the linear threshold given by the line $Q_2(t) = \kappa_2 Q_4(t)$. Hence, for Case 2:

$$\begin{aligned} \textstyle\int_0^t \mathbf{1}\{0 < Q_4(s) < \kappa_1 Q_2(s)\}dT_1(s) &= 0, \\ \textstyle\int_0^t \mathbf{1}\{Q_2(s) \leq \frac{1}{\kappa_1} Q_4(s)\}dT_4(s) &= 0, \\ \textstyle\int_0^t \mathbf{1}\{0 < Q_2(s) < \kappa_2 Q_4(s)\}dT_3(s) &= 0, \\ \textstyle\int_0^t \mathbf{1}\{Q_4(s) \leq \frac{1}{\kappa_2} Q_2(s)\}dT_2(s) &= 0. \end{aligned}$$

Recall that we require $\kappa_1 > \frac{\mu_3}{\mu_1}$, $\kappa_2 > \frac{\mu_1}{\mu_3}$.

4. FLUID LIMITS AND FLUID MODELS

In this section we assume (A1), and consider the behavior of the push pull network under fluid scaling. We use the pull priority policy in Case 1, and the linear threshold policy in Case 2.

To study the network under fluid scaling we consider the six dimensional network process $Y(t) = (Q(t), T(t))$, and parameterize it by $n = 1,2,...$ as follows: For each n set the initial queue lengths as $Q^n(0)$, and let $Y^n(t)$ be the network process starting from this initial condition, where all the Y^n share the same sequences of random processing times $\xi_i, i = 1,2,3,4$. Denote by $Y^n(t,\omega)$ the realization of the n'th network process for some ω in the sample space. We define *fluid scalings* as:

$$\bar{Y}^n(t,\omega) = \frac{Y^n(nt,\omega)}{n}.$$

A function $\bar{Y}(t) = (\bar{Q}(t), \bar{T}(t))$ is said to be a *fluid limit* of our network if there exists a sequence of integers $r \to \infty$ and a sample path ω such that:

$$\bar{Y}^r(\cdot,\omega) \to \bar{Y}(\cdot), \ \text{ u.o.c.}$$

It may now be shown (c.f. Theorem 4.1 of [6] or Appendix A.2 of [7]) that under Assumption (A1), except for a set of ω of measure zero, fluid limits exist for every ω, and every one of them satisfies the following fluid equations:

$$\begin{array}{l} \bar{Q}_i(t) = \bar{Q}_i(0) + \mu_{i-1}\bar{T}_{i-1}(t) - \mu_i \bar{T}_i(t)\,, i = 2,4 \\ \bar{Q}_i(t) \geq 0, \quad i = 2,4 \\ \bar{T}_i(0) = 0,\ \bar{T}_i \text{ is non-decreasing}, \quad i = 1,2,3,4 \end{array} \tag{3}$$

as well as

$$\bar{T}_1(t) + \bar{T}_4(t) = t, \qquad \bar{T}_2(t) + \bar{T}_3(t) = t, \tag{4}$$

and in addition, under pull priority they satisfy:

$$\begin{aligned} \textstyle\int_0^t \bar{Q}_4(s)d\bar{T}_1(s) &= 0, \\ \textstyle\int_0^t \bar{Q}_2(s)d\bar{T}_3(s) &= 0, \end{aligned} \tag{5}$$

and under linear threshold policy they satisfy:

$$\begin{aligned} \textstyle\int_0^t \mathbf{1}\{0 < \bar{Q}_4(s) < \kappa_1 \bar{Q}_2(s)\}d\bar{T}_1(s) &= 0, \\ \textstyle\int_0^t \mathbf{1}\{\bar{Q}_2(s) \leq \frac{1}{\kappa_1} \bar{Q}_4(s)\}d\bar{T}_4(s) &= 0, \\ \textstyle\int_0^t \mathbf{1}\{0 < \bar{Q}_2(s) < \kappa_2 \bar{Q}_4(s)\}d\bar{T}_3(s) &= 0, \\ \textstyle\int_0^t \mathbf{1}\{\bar{Q}_4(s) \leq \frac{1}{\kappa_2} \bar{Q}_2(s)\}d\bar{T}_2(s) &= 0. \end{aligned} \tag{6}$$

Equations (3)-(6) represent a deterministic continuous fluid analog of the stochastic model introduced in the previous section. We shall refer to equations (3)–(5) as the *fluid model of Case 1*. Similarly we shall refer to (3),(4) and (6) as the *fluid model of Case 2*.

A *fluid solution of Case 1* (*Case 2*) is any pair $(\bar{Q}, \bar{T})$ that satisfies the fluid model equations of Case 1 (Case 2). We say that the fluid model of Case 1 (Case 2) is *stable* if there exists a $\delta > 0$ such that for every fluid solution of Case 1 (Case 2), whenever $|\bar{Q}(0)| = 1$ then $\bar{Q}(t) = 0$ for any $t \geq \delta$.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

Our main result in this section is:

THEOREM 1. *Consider the push pull network, assume that Assumption (A1) holds, and use in Case 1 the pull priority policy, and in Case 2 the linear threshold policy. Then the fluid model is stable.*

This theorem will be used to show positive Harris Recurrence in the next section. It also immediately leads to the following corollary, which describes the fluid scale behavior of the push pull network:

COROLLARY 1. *Consider the push pull network with some fixed initial queue lengths, $Q(0)$, under the assumptions of Theorem 1. Then almost surely $Y(nt)/n$ will converge as $n \to \infty$ u.o.c. to a fluid limit $\bar{Y}(t) = (\bar{Q}(t), \bar{T}(t))$ which satisfies:*

$$\bar{T}_i(t) = \theta_i t, \quad \bar{D}_i(t) = \nu_i t, \quad \bar{Q}_i(t) = 0, \quad i = 1, 2, 3, 4.$$

The proof of Theorem 1 is by means of a Lyapounov function, f. As in [9], we shall make use of the following elementary Lemma 1. Recall that $T_i(t)$ are Lipschitz with constant 1. It then follows that $\bar{T}_i$, and also $\bar{Q}_i(t)$, are Lipschitz, for every fluid solution. Hence they are absolutely continuous with derivative defined almost everywhere. We say that t is a regular point of a fluid solution if the derivatives of $\bar{Y}$ exist at t.

LEMMA 1. *Let f be an absolutely continuous nonnegative function, and let $\dot{f}$ denote its derivative whenever it exists.*

(i) If $f(t) = 0$ and $\dot{f}(t)$ exists, then $\dot{f}(t) = 0$.

(ii) Assume that for some $\epsilon > 0$ at regular points $t > 0$, whenever $f(t) > 0$ then $\dot{f}(t) \leq -\epsilon$. Then $f(t) = 0$ for all $t \geq f(0)/\epsilon$. Furthermore, $f(\cdot)$ is non increasing and hence once it reaches 0 it stays there forever.

PROOF OF THEOREM 1. *Case 1*: Define $f(t) = \bar{Q}_2(t) + \bar{Q}_4(t)$. Clearly $f(t) \geq 0$ and $f(t) = 0$ if and only if $\bar{Q}(t) = 0$. Also, if $|\bar{Q}(0)| = 1$ then $f(0)$ is bounded (by $B = 1$). We show that f satisfies the conditions of Lemma 1, for some ϵ, and hence $f(t) = 0$ for $t > f(0)/\epsilon$, and so if $|\bar{Q}(0)| = 1$, $\bar{Q}(t) = 0$ for $t \geq B/\epsilon$ which proves stability of the fluid model.

Define $\epsilon = \min\{\mu_2 - \mu_1,\ \mu_4 - \mu_3\}$. The values of μ_i in Case 1 ensure that $\epsilon > 0$. We now bound $\dot{f}(t)$ by $-\epsilon$ for all regular time points t at which $f(t) > 0$ by analyzing all possible values of $\bar{Q}_i(t)$, $i = 2, 4$:

- Assume $\bar{Q}_2(t),\ \bar{Q}_4(t) > 0$:

  By (5), $\dot{\bar{T}}_1 = \dot{\bar{T}}_3 = 0$ and thus by (4), $\dot{\bar{T}}_2 = \dot{\bar{T}}_4 = 1$. As a consequence, $\dot{\bar{Q}}_i(t) = -\mu_i$ for $i = 2, 4$ and

  $$\dot{f} = -(\mu_2 + \mu_4) \leq -\epsilon.$$

- Assume $\bar{Q}_2(t) > 0$, $\bar{Q}_4(t) = 0$:

  By (5) $\dot{\bar{T}}_3 = 0$ and thus by (4), $\dot{\bar{T}}_2 = 1$. As a consequence,

  $$\dot{f} = \mu_1 \dot{\bar{T}}_1 - \mu_2 - \mu_4 \dot{\bar{T}}_4 =$$
  $$\mu_1 - \mu_2 - (\mu_1 + \mu_4)\dot{\bar{T}}_4 \leq -(\mu_2 - \mu_1) \leq -\epsilon.$$

- Assume $\bar{Q}_2(t) = 0$, $\bar{Q}_4(t) > 0$:

  Similarly to the previous argument,

  $$\dot{f} \leq -(\mu_4 - \mu_3) \leq -\epsilon.$$

This completes the proof for Case 1.

Case 2: We use the same technique as in Case 1. Define:

$$f(t) = \begin{cases} (1+\beta)\bar{Q}_2(t) - (\kappa_2 - \beta)\bar{Q}_4(t) & \text{if } \bar{Q}_2(t) \geq \kappa_2 \bar{Q}_4(t), \\ (1+\beta)\bar{Q}_4(t) - (\kappa_1 - \beta)\bar{Q}_2(t) & \text{if } \bar{Q}_4(t) \geq \kappa_1 \bar{Q}_2(t), \\ \beta(\bar{Q}_2(t) + \bar{Q}_4(t)) & \text{otherwise.} \end{cases}$$

where

$$\beta = \frac{1}{2}\min\{\frac{\kappa_1 - \frac{\mu_3}{\mu_1}}{1 + \frac{\mu_3}{\mu_1}}, \frac{\kappa_2 - \frac{\mu_1}{\mu_3}}{1 + \frac{\mu_1}{\mu_3}}\}.$$

Again, it is easily seen that $f(t) \geq 0$ and $f(t) = 0$ if and only if $\bar{Q}(t) = 0$, and if $|\bar{Q}(0)| = 1$ then $f(0)$ is bounded by some finite value B.

All we need to do is find an ϵ to satisfy the conditions of Lemma 1. We now bound $\dot{f}(t)$ for all regular time points t at which $f(t) > 0$, by analyzing all possible values of $\bar{Q}_i(t)$, $i = 2, 4$:

- Assume $\frac{1}{\kappa_2}\bar{Q}_2(t) < \bar{Q}_4(t) < \kappa_1 \bar{Q}_2(t)$:

  Then $f(t) = \beta(\bar{Q}_2(t) + \bar{Q}_4(t))$, and in this region both servers use pull priority. Hence

  $$\dot{f} = \beta(\mu_1 \dot{T}_1 - \mu_2 \dot{T}_2 + \mu_3 \dot{T}_3 - \mu_4 \dot{T}_4)$$

  and by (6) we have that $\dot{T}_1 = \dot{T}_3 = 0$ and thus $\dot{T}_2 = \dot{T}_4 = 1$. Hence

  $$\dot{f} = -\beta(\mu_2 + \mu_4).$$

- Assume $0 < \bar{Q}_4(t) \leq \frac{1}{\kappa_2}\bar{Q}_2(t)$:

  Then $f(t) = (1+\beta)\bar{Q}_2(t) - (\kappa_2 - \beta)\bar{Q}_4(t)$ and in this region both queues are not empty, and server 1 gives priority to pull while server 2 gives priority to push. Hence

  $$\dot{f} = (1+\beta)(\mu_1 \dot{T}_1 - \mu_2 \dot{T}_2) - (\kappa_2 - \beta)(\mu_3 \dot{T}_3 - \mu_4 \dot{T}_4),$$

  and by (6) we have that $\dot{T}_1 = \dot{T}_2 = 0$ and thus $\dot{T}_3 = \dot{T}_4 = 1$. Hence

  $$\dot{f} = -(\kappa_2 - \beta)(\mu_3 - \mu_4).$$

- Assume $0 < \bar{Q}_2(t) \leq \frac{1}{\kappa_1}\bar{Q}_4(t)$:

  The analysis is symmetric to the previous case, and yields:

  $$\dot{f} = -(\kappa_1 - \beta)(\mu_1 - \mu_2).$$

- Assume $\bar{Q}_2(t) > 0,\ \bar{Q}_4(t) = 0$:

  Again $f(t) = (1+\beta)\bar{Q}_2(t) - (\kappa_2 - \beta)\bar{Q}_4(t)$, and in this region server 2 gives priority to push. With $\bar{Q}_4(t) = 0$ we cannot say where server 1 will work. Hence

  $$\dot{f} = (1+\beta)(\mu_1 \dot{T}_1 - \mu_2 \dot{T}_2) - (\kappa_2 - \beta)(\mu_3 \dot{T}_3 - \mu_4 \dot{T}_4)$$

  and by (6) $\dot{T}_2 = 0$ and as a result $\dot{T}_3 = 1$. Hence:

  $$\begin{aligned} \dot{f} &= (1+\beta)\mu_1 \dot{T}_1 - (\kappa_2 - \beta)(\mu_3 - \mu_4 \dot{T}_4) \\ &= (1+\beta)\mu_1 \dot{T}_1 - (\kappa_2 - \beta)[\mu_3(\dot{T}_1 + \dot{T}_4) - \mu_4 \dot{T}_4] \\ &= -(\kappa_2 - \beta)[(\mu_3 - \frac{1+\beta}{\kappa_2 - \beta}\mu_1)\dot{T}_1 + (\mu_3 - \mu_4)\dot{T}_4] \\ &\leq -(\kappa_2 - \beta)\min\{\mu_3 - \frac{1+\beta}{\kappa_2 - \beta}\mu_1\ ,\ \mu_3 - \mu_4\}. \end{aligned}$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

- Assume $\bar{Q}_4(t) > 0,\ \bar{Q}_2(t) = 0$: The analysis is symmetric to the previous case, and yields:

$$\dot{f} \leq -(\kappa_1 - \beta)\min\{\mu_1 - \frac{1+\beta}{\kappa_1 - \beta}\mu_3\ ,\ \mu_1 - \mu_2\}.$$

All five bounds above are negative, and we choose $-\epsilon$ as their maximum. This completes the proof. □

Remark: So far in this section we assumed that the nth system starts with queue lengths $Q^n(0)$, and that all the jobs in the system had no previous processing, so that the $S_i(t)$ are counting processes, with intervals ξ_i which have long term rate μ_i. A more general model assumes that at time 0 the head of the line job in each queue or infinite supply has received some processing, and let $\xi_{i,0}$ be the residual processing time of this first job. Then the first interval is a residual processing time with a different mean from the other ξ_i^j, $j > 1$. In that case $S_i(t)$ are delayed counting processes. We now associate with the nth system an initial state consisting of $Q_i^n(0), \xi_{i,0}^n, i = 1, 2, 3, 4$. All the results of this section remain valid and unchanged as long as we assume that $\xi_{i,0}^n/n \to 0$ a.s. (see [4]).

5. POSITIVE HARRIS RECURRENCE

In this section we add the set of Assumptions (A2) to Assumption (A1), and use the fluid stability results from the previous section to show that the push pull network under our policies can be described by a positive Harris recurrent Markov chain. To do so we adapt the framework developed by Dai [6], see also [4].

We begin by defining the network state process. Denote by $U_i(t), V_i(t)$ the residual processing times of the head of the line activities which are in process or preempted at the current time t. $U_i(t), i = 2, 4$ is for the pull activities and $V_i(t), i = 1, 3$ is for the push activities. Now denote the *network state process* by,

$$X(t) = (Q(t), U(t), V(t)).$$

The state space is $\mathbb{S} = \mathbb{Z}_+^2 \times \mathbb{R}_+^2 \times \mathbb{R}_+^2$, and $|X(t)|$ is the sum of the components of $X(t)$. Since the evolution of $X(t)$ between arrivals and departures is deterministic, $X(t)$ is piecewise deterministic, and it is not difficult to show that $X(t)$ is a piecewise deterministic strong Markov process (c.f. [10]):

PROPOSITION 1. *Under Assumptions (A1), (A2a), $X = \{X(t),\ t \geq 0\}$ is a strong Markov process with state space $\mathbb{S}$.*

Let $P^t(x, \cdot)$ be the transition probability of X. That is for $x \in \mathbb{S}$, $B \in \mathcal{B}(\mathbb{S})$,

$$P^t(x, B) \equiv P_x\{X(t) \in B\} \equiv P\{X(t) \in B \,|\, X(0) = x\}.$$

A nonzero measure π on $(\mathbb{S}, B(\mathbb{S}))$ is *invariant* for X if π is σ-finite, and for each $t \geq 0$,

$$\pi(B) = \int_{\mathbb{S}} P^t(x, B)\, \pi(dx),\ \ B \in \mathcal{B}(\mathbb{S}).$$

Let $\tau_A = \inf\{t \geq 0 : X(t) \in A\}$. We say that X is *Harris recurrent* if there exists some σ-finite measure ν on $(\mathbb{S}, B(\mathbb{S}))$, such that for all $A \in \mathcal{B}(\mathbb{S})$ with $\nu(A) > 0$ we have $P_x(\tau_A < \infty) = 1$ for all $x \in \mathbb{S}$. If X is Harris recurrent then an essentially (up to a positive scalar multiplier) unique invariant measure π exists. When π is finite (in which case we normalize it to a probability measure) we say that X is *positive Harris recurrent*. Positive Harris recurrence is a common notion of stability since it implies certain ergodicity properties. For example, given $f : \mathbb{S} \mapsto \mathbb{R}_+$, denote

$$\pi(f) = \int_{\mathbb{S}} f(x)\, \pi(dx)$$

whenever the integral makes sense. Then if $\pi(|f|) < \infty$:

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f(X(s))ds = \pi(f) \qquad P_x \text{ a.s. for each } x \in \mathbb{S}.$$

To establish positive Harris recurrence of $X(t)$, we need a further concept: A non-empty set A is said to be *petite* if there exists a probability distribution $\mathbf{a}$ on $(0, \infty)$ and a nontrivial measure ν on $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$, such that for all $x \in A$

$$\int_0^\infty P^t(x, B)\mathbf{a}(dt) \geq \nu(B), \qquad \text{for all } B \in \mathcal{B}(\mathbb{S}).$$

Petiteness of A may be interpreted as the property that all sets B are "equally accessible" from any $x \in A$. For more on Markov processes, positive Harris recurrence and petite sets, see [22] for an introduction and discrete time results, and [23, 24] for continuous time results.

We are now in a position to rigorously define Assumption (A2b'):

$$(A2b') \qquad A = \{x : |x| \leq \sigma\} \text{ is petite for any } \sigma > 0.$$

Our main result in this paper is:

THEOREM 2. *Under Assumptions (A1), (A2a) and (A2b'), the network state process X is Positive Harris Recurrent for Case 1 under the pull priority policy and for Case 2 under the linear threshold policy. Furthermore, for Case 1 we may substitute Assumptions (A2b') with (A2b).*

PROOF. The proof uses the framework of Dai [6]. The main theorem in that paper (Theorem 4.2) states that if the fluid model of a multi-class queueing network (with exogenous arrival streams) is stable then the associated Markov process is positive Harris recurrent. However, our model does not fall in that scope and hence we must adapt the proof.

The following discussion outlines the adaptation. Dai shows that positive Harris recurrence of the network state process follows directly from two statements:

(i) Convergence of a fluid scaled process scaled by its initial state: There exists $\delta > 0$ such that

$$\lim_{|x|\to\infty} \frac{1}{|x|} E_x|X(\delta|x|)| = 0.$$

(ii) Petiteness of closed bounded sets as in our Assumption (A2b').

The arguments of Dai that statements (i) and (ii) imply positive Harris recurrence are valid also for our push pull network, and so to prove the theorem we need to show that (i) and (ii) hold.

The main result of Dai is to show that stability of the fluid model, as defined in the previous Section 4, implies (i). The proof that fluid stability implies (i) needs no changes in our case. Hence, under Assumptions (A1) and (A2a), our Theorem 1, in which we have proved stability of the fluid model, implies (i) for the push pull network.

Hence, if we make Assumption (A2b'), the positive Harris recurrence of the push pull network follows.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

The technical Assumption (ii), that all compact sets are petite is awkward, as it is difficult to check. Thus it is useful instead of Assumption (A2b') to find a sufficient condition which is easier to check. Dai's paper asserts that for multiclass queueing networks with an exogenous input stream the assumption that inter-arrival times have a spread out distribution with unbounded support implies (ii). His proof follows directly from the earlier work of Meyn and Down [21], who proved the same result for generalized Jackson networks. This needs to be extended to the case of infinite supply of work. The difference is that with infinite supply of work the output process from an infinite virtual queue is in general not independent of the state of the other queues. Guo and Zhang [13] have adapted Meyn and Down's ideas to a reentrant line with infinite supply of work where the policy is to give lowest priority to the activity with the infinite supply.

The following Lemma 2 extends these results, and shows that in Case 1, under pull priority, the Assumption (A2b) implies (A2b'), and hence positive Harris recurrence $\square$

LEMMA 2. *For the network state process X, operating with the pull priority policy, under Assumptions (A1) and (A2a), the Assumption (A2b) implies (A2b').*

We present the proof in the Appendix. We were unable to provide a similar result for the more complex linear threshold policy.

6. DIFFUSION SCALE ANALYSIS

In this section we add the assumption on existence of second moments, (A3), to the Assumptions (A1,A2), and consider the behavior of the push pull network under diffusion scaling. We find that the queues are 0 on the diffusion scale, and the output processes $D_i(t)$ converge under diffusion scaling to Brownian motions. We calculate the parameters of these, including the covariances between the output streams.

We now define diffusion scalings for $n = 1, 2, \ldots$. First denote $\bar{S}(t) = \lim_{n\to\infty} \bar{S}^n(t) = \lim_{n\to\infty} \frac{S(nt)}{n} = \mu t$, where the limit exists a.s. u.o.c. by Assumption (A1). Further use the fluid limit processes of Section 4, Corollary 1. The diffusion scalings are:

$$\begin{array}{ll} \hat{S}_i^n(t) = \frac{S_i(nt)-\bar{S}_i(nt)}{\sqrt{n}}, & \hat{T}_i^n(t) = \frac{T_i(nt)-\bar{T}_i(nt)}{\sqrt{n}}, \\ \hat{D}_i^n(t) = \frac{D_i(nt)-\bar{D}_i(nt)}{\sqrt{n}}, & \hat{Q}_i^n(t) = \frac{Q_i(nt)}{\sqrt{n}}. \end{array} \tag{7}$$

Note that in this analysis we use a fixed $Q(0)$, which does not change with n. Define the 10 dimensional diffusion scaled process:

$$\hat{X}^n(t) = (\hat{D}^n(t), \hat{T}^n(t), \hat{Q}^n(t))$$

The following theorem describes the diffusion limit for our model.

THEOREM 3. *Consider the Push Pull network, under Assumptions (A1–A3), for Case 1 under pull priority policy, and for Case 2 under linear threshold policy. Then as $n \to \infty$, $\hat{X}^n \Rightarrow \hat{X}$, where $\hat{X}(t)$ is a 10 dimensional driftless Brownian motion. Furthermore,*

$$\begin{array}{lll} \hat{D}_1^n(t) - \hat{D}_2^n(t) = \hat{Q}_2^n(t) & \Rightarrow & 0, \\ \hat{D}_4^n(t) - \hat{D}_3^n(t) = \hat{Q}_4^n(t) & \Rightarrow & 0, \end{array} \tag{8}$$

$$\hat{T}_1^n(t) + \hat{T}_4^n(t) = \hat{T}_3^n(t) + \hat{T}_2^n(t) = 0, \tag{9}$$

and the variances and covariances of the limiting Brownian motions are given by:

$$Var(\hat{D}_2(1)) = \frac{\mu_1\mu_2}{(\mu_1\mu_3 - \mu_2\mu_4)^3} \times \tag{10}$$
$$[\mu_1\mu_2\mu_3\mu_4(c_3^2 + c_4^2)(\mu_1 - \mu_2) + (\mu_1^2\mu_3^2c_2^2 + \mu_2^2\mu_4^2c_1^2)(\mu_3 - \mu_4)],$$

$$Cov(\hat{D}_2(1), \hat{D}_4(1)) = -\frac{\mu_1\mu_2\mu_3\mu_4}{(\mu_1\mu_3 - \mu_2\mu_4)^3} \times \tag{11}$$
$$[(\mu_1\mu_3c_4^2 + \mu_2\mu_4c_3^2)(\mu_1 - \mu_2) + (\mu_1\mu_3c_2^2 + \mu_2\mu_4c_1^2)(\mu_3 - \mu_4)],$$

with a symmetric expression for $Var(\hat{D}_4(1))$. Similar expressions for variances and covariances of $\hat{T}_2(\cdot)$, $\hat{T}_4(\cdot)$ may be read off from (16).

PROOF. The equalities (8) and (9) follow immediately from (2) and (1). The convergence to 0 in (8) follows from Theorem 2, since $Q_i(t)$ has a limiting stationary distribution, therefore $Q_i(nt)$ converges to this limiting distribution as $n \to \infty$, and dividing by $\sqrt{n}$ implies converges to 0 in probability and therefore also weakly.

Also, by Corollary 1, $\bar{T}^n(t) \to \bar{T}(t) = \theta t$ and $\bar{D}^n(t) \to \bar{D}(t) = \nu t$ u.o.c as $n \to \infty$.

The rest of the proof and the calculations are straightforward:

$$\begin{array}{rcl} \hat{D}_i^n(t) & = & \frac{D_i(nt)-\bar{D}_i(nt)}{\sqrt{n}} \\ & = & \frac{S_i(n\bar{T}_i^n(t))-\bar{S}_i(n\bar{T}_i^n(t))}{\sqrt{n}} + \frac{\bar{S}_i(n\bar{T}_i^n(t))}{\sqrt{n}} - \frac{\bar{D}_i(nt)}{\sqrt{n}} \\ & = & \hat{S}_i^n(\bar{T}_i^n(t)) + \mu_i\frac{T_i(nt)-\bar{T}_i(nt)}{\sqrt{n}} + \mu_i\frac{\bar{T}_i(nt)}{\sqrt{n}} - \frac{\bar{D}_i(nt)}{\sqrt{n}} \\ & = & \hat{S}_i^n(\bar{T}_i^n(t)) + \mu_i\hat{T}_i^n(t) + \theta_i\mu_i\sqrt{n}t - \theta_i\mu_i\sqrt{n}t, \end{array}$$

where all we did is to add and subtract quantities, use the definitions (7), and use $\bar{S}_i(t) = \mu_i t$ (by Assumption (A1)), and $\bar{T}_i(t) = \theta_i t$, $\bar{D}_i(t) = \nu_i t = \mu_i\theta_i t$ (from Corollary 1).

Define $\hat{P}_i^n(t) = \hat{S}_i^n(\bar{T}_i^n(t))$, $i = 1, 2, 3, 4$, then summarizing the above and also using similar calculations (for (13) and (14)) we obtain:

$$\hat{D}_i^n(t) = \hat{P}_i^n(t) + \mu_i\hat{T}_i^n(t), \quad i = 1, 2, 3, 4, \tag{12}$$
$$\hat{Q}_i^n(t) = \hat{D}_{i-1}^n(t) - \hat{D}_i^n(t), \quad i = 2, 4, \tag{13}$$
$$\hat{T}_2^n(t) = -\hat{T}_3^n(t), \qquad \hat{T}_4^n(t) = -\hat{T}_1^n(t). \tag{14}$$

Now using (12)–(14):

$$\begin{bmatrix} \hat{D}_2^n(t) \\ \hat{D}_4^n(t) \\ \hat{T}_2^n(t) \\ \hat{T}_4^n(t) \end{bmatrix} = \mathbf{A}\,\hat{P}^n(t) + \mathbf{B}\begin{bmatrix} \hat{Q}_2^n(t) \\ \hat{Q}_4^n(t) \end{bmatrix}, \tag{15}$$

where

$$\mathbf{A} = \frac{1}{\mu_1\mu_3 - \mu_2\mu_4}\begin{bmatrix} -\mu_2\mu_4 & \mu_1\mu_3 & \mu_1\mu_2 & -\mu_1\mu_2 \\ \mu_3\mu_4 & -\mu_3\mu_4 & -\mu_2\mu_4 & \mu_1\mu_3 \\ -\mu_4 & \mu_4 & \mu_1 & -\mu_1 \\ \mu_3 & -\mu_3 & -\mu_2 & \mu_2 \end{bmatrix},$$

and

$$\mathbf{B} = \frac{1}{\mu_1\mu_3 - \mu_2\mu_4}\begin{bmatrix} \mu_2\mu_4 & -\mu_1\mu_2 \\ -\mu_3\mu_4 & \mu_2\mu_4 \\ \mu_4 & -\mu_1 \\ -\mu_3 & \mu_2 \end{bmatrix}.$$

By the functional central limit theorem for renewal processes and the continuous mapping theorem (c.f. [11]) we have

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

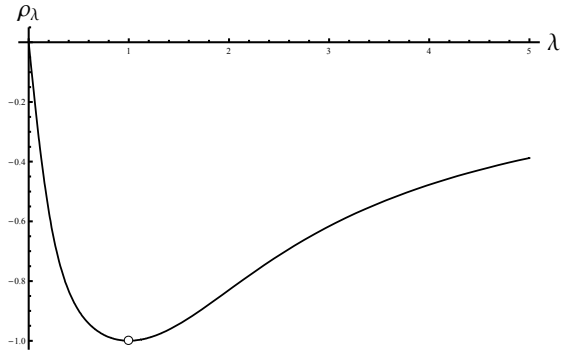

Figure 3: The correlation between outputs of a symmetric push pull network.

$\hat{P}^n(t) \Rightarrow \hat{P}(t)$ where $\hat{P}(t)$ is a 4 dimensional driftless Brownian motion with a diagonal covariance matrix $\mathbf{\Lambda}$, having entries

$$\mathrm{Var}(\hat{P}_i(1)) = \mu_i c_i^2 \theta_i,\ i = 1, 2, 3, 4.$$

Incorporating the above with the weak convergence of $\hat{Q}^n$ to 0, we have that $(\hat{D}_2^n(t), \hat{D}_4^n(t), \hat{T}_2^n(t), \hat{T}_4^n(t))$ converges to a driftless Brownian motion process with covariance matrix:

$$\mathbf{\Gamma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}'. \tag{16}$$

□

The following two subsections show some surprising facts about the diffusion scale behavior of the push pull network.

6.1 Insensitivity to the Policy

The proof of Theorem 3 does not depend on the exact policy which was used. All that is needed is $\hat{Q}_2^n(t) \Rightarrow 0$ and $\bar{T}^n(t) \to \theta t$ u.o.c. In particular, the calculations for Case 1 and Case 2 are the same. In fact, any policy which achieves full utilization and which achieves $\hat{Q}^n(t) \Rightarrow 0$ will automatically satisfy the convergence in Corollary 1. Hence the analysis and the results are valid for the push-pull network operating under any full utilization policy which satisfies $\hat{Q}^n(t) \Rightarrow 0$.

We reach the surprising conclusion that the diffusion scale output processes $\hat{D}(t)$ do not depend on the policy, so long as it is fully utilizing and stabilizing. In a sense this means that all these policies are optimal on the diffusion scale.

6.2 Negative Covariance of Outputs

It is evident from (11) that $\mathrm{Cov}(\hat{D}_2(t), \hat{D}_4(t)) < 0$. Also, when all activity processing times have the same squared coefficient of variation c^2, then both the variance and the covariance in (10,11) are linear in c^2.

In Figure 3 we illustrate the negative correlation between the output processes of our network. We plot as a function of λ:

$$\rho_\lambda = \frac{\mathrm{Cov}(\hat{D}_2(1), \hat{D}_4(1))}{\sqrt{\mathrm{Var}(\hat{D}_2(1))\mathrm{Var}(\hat{D}_4(1))}}, \tag{17}$$

for symmetric push pull networks with parameters $c_i^2 = c^2, i = 1, 2, 3, 4$, $\mu_2 = \mu_4 = 1$, $\mu_1 = \mu_3 = \lambda$.

Our analysis applies to all $\lambda \neq 1$. When $\lambda = 1$ we have a completely balanced network (as defined in Section 1) and with our policies, under diffusion scaling the queues do not converge to 0, so the analysis in this paper does not apply.

Note that for $1/2 < \lambda < 2$, i.e when the ratio of processing times for each type of job on the two servers is not too far from 1, we get $-1 < \rho_\lambda < -0.8$, so the negative correlation is very high. Most surprisingly, as $\lambda \to 1$ the correlation approaches -1, and we are close to complete resource pooling [8].

When λ is very small or very large the correlation approaches zero. This is intuitively clear, since each server is now spending almost all of its time on just one type of job, and so the fluctuations in D_2 depend mostly on the processing times of jobs of type 1, and the fluctuations of D_4 will depend mostly on the processing times of jobs of type 2, and hence they will be almost independent.

7. ACKNOWLEDGMENTS

We would like to thank Serguei Foss for useful discussions on stability of Markov chains, and fluid and diffusion approximations of queueing networks.

8. REFERENCES

[1] I. Adan and G. Weiss. A two node Jackson network with infinite supply of work. *Probability in the Engineering and Informational Sciences*, 19(2):191–212, 2005.

[2] I. Adan and G. Weiss. Analysis of a simple Markovian re-entrant line with infinite supply of work under the LBFS policy. *Queueing Systems*, 54(3):169–183, 2006.

[3] B. Ata and W. Lin. Heavy traffic analysis of maximum pressure policies for stochastic processing networks with multiple bottlenecks. *Preprint*.

[4] M. Bramson. Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems*, 28(1-3):7–31, 1998.

[5] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks, Performance, Asymptotics and Optimization*. Springer, 2003.

[6] J. G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *The Annals of Applied Probability*, 5(1):49–77, 1995.

[7] J. G. Dai and W. Lin. Maximum pressure policies in stochastic processing networks. *Operations Research*, 53(2), 2005.

[8] J. G. Dai and W. Lin. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Preprint*, 2006.

[9] J. G. Dai and G. Weiss. Stability and instability of fluid models for re-entrant lines. *Mathematics of Operations Research*, 21(1):115–134, 1996.

[10] M. H. A. Davis. Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of Royal Statistical Society. Series B.*, 46(3):353–388, 1984.

[11] P. W. Glynn. Diffusion approximations. *In Handbooks in Operations Research, Vol 2, D.P. Heyman and M.J. Sobel (eds.), North-Holland, Amsterdam*, pages 145–198, 1990.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

[12] Y. Guo and H. Zhang. On the stability of a simple re-entrant line with infinite supply. *Preprint*, 2006.
[13] Y. Guo and H. Zhang. Positive Harris recurrence of re-entrant lines with infinite supply. *Preprint*, 2007.
[14] M. J. Harrison. Brownian models of queueing networks with heterogeneous customer populations. *In Stochastic Differential Systems, Stochastic Control Theory and Applications (W. Fleming and P.-L. Lions, eds.)*, pages 147–186, 1988.
[15] M. J. Harrison. and R. J. Williams Brownian models of queueing networks with heterogeneous customer populations. *Ann. Appl. Probab*, 2(2):263–193, 1992.
[16] S. G. Henderson, S. P. Meyn, and V. B. Tadic. Performance evaluation and policy selection in multiclass networks. *Discrete Event Dynamic Systems*, 13(1-2):149–189, 2003.
[17] A. Kopzon and G. Weiss. A push pull queueing system. *Operations Research Letters*, 30(6):351–359, 2002.
[18] A. Kopzon, Y. Nazarathy and G. Weiss. A push pull system with infinite supply of work. *Preprint*, 2008.
[19] P. Kumar and T. Seidman. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control*, AC-35(3):289–298, 1990.
[20] S. P. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, 2008.
[21] S. P. Meyn and D. Down. Stability of generalized Jackson networks. *The Annals of Applied Probability*, 4(1):124–148, 1994.
[22] S. P. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.
[23] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes II: Continuous-time processes and sampled chains. *Advances in Applied Probability*, 25(3):487–517, 1993.
[24] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993.
[25] Y. Nazarathy and G. Weiss. The asymptotic variance rate of finite capacity birth-death queues. *Queueing Systems. To Appear.*, 2008.
[26] Y. Nazarathy and G. Weiss. Near optimal control of queueing networks over a finite time horizon. *Annals of Operations Research. To Appear.*, 2008.
[27] A. Rybko and A. Stolyar. On the ergodicity of random processes that describe the functioning of open queueing networks. *Probl. Pereda. Inf.*, 28(3):3–26, 1992.
[28] G. Weiss. Jackson networks with unlimited supply of work. *Journal of Applied Probability*, 42(3):879–882, 2005.

APPENDIX

A. PROOF OF LEMMA 2

The proof requires some more concepts (c.f. [23]): We say that X is *ψ-irreducible*, if there exists a measure ψ on $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$ such that, whenever $\psi(A) > 0$, we have $P_x\{\tau_A < \infty\} > 0$ for all $x \in \mathbb{S}$.

Let $\mathbf{a}$ be a probability distribution on $\mathbb{R}_+$. Define the *Markov transition function* $K_{\mathbf{a}}$ as

$$K_{\mathbf{a}}(x, \cdot) = \int_0^\infty P^t(x, \cdot)\, \mathbf{a}(dt).$$

A *continuous component* of $K_{\mathbf{a}}$ is a non-negative function $T(x, A)$ which is lower semi-continuous in x, and satisfies

$$K_{\mathbf{a}}(x, A) \geq T(x, A), \qquad x \in \mathbb{S}, \quad A \in \mathcal{B}(\mathbb{S}),$$

We say that X is a *T-process* if there exists a distribution $\mathbf{a}$ such that $K_{\mathbf{a}}$ possesses a continuous component T, with $T(x, \mathbb{S}) > 0$ for all $x \in \mathbb{S}$. The following proposition (c.f. Theorem 4.1(i) of [23]), connects ψ-irreducible T-processes and petiteness of compacts.

PROPOSITION 2. *If X is a ψ-irreducible T-process then every compact set in $\mathcal{B}(\mathbb{S})$ is petite.*

We say that a state x^* is reachable if $\int_0^\infty P^t(x, O)dt > 0$ for every open neighborhood O of x^* and every $x \in \mathbb{S}$. It can be shown (c.f [13]) that if X is a T-process with a reachable point x^* then it is also ψ-irreducible with $\psi(\cdot) = T(x^*, \cdot)$.

Returning to our push pull queueing network with pull priority, it is easy to see, by Assumption (A2b), that the state $Q(t) = 0, U(t) = 0, V(t) = 0$ is reachable.

Thus the main part of the proof is to show that X is a T-process: We need to construct a lower semi-continuous function $T(\cdot, A)$ and a transition kernel $K_{\mathbf{a}}(\cdot, A)$, so that $K_{\mathbf{a}}(x, A) \geq T(x, A)$, for all $(x, A) \in (\mathbb{S}, \mathcal{B}(\mathbb{S}))$.

Following Meyn and Down [21] the construction is in several steps. The crucial step in the construction of T is to consider the initial state in a bounded rectangle, the set of states to be reached is an empty system with both servers engaged in push activity, and to then bound the probability of reaching this set after a deterministic integer time by a continuous function.

For an integer ℓ define $R_\ell = \{0, \ldots, \ell\}^2 \times [0, \ell)^2 \times [0, \ell)^2$. Now take the initial state at time 0 as $x_0 \in R_\ell$.

Define $Z(t) = (Q_2(t), Q_4(t), U_2(t), U_4(t))$. Then the network state process is $X(t) = (Z(t), V(t))$. Let $A_1, A_3 \in \mathcal{B}(\mathbb{R}_+)$. The set to be reached is the set $\{Z = 0, V \in A_1 \times A_3\}$. For an integer time n_l we will bound $P_{x_0}(Z(n_\ell) = 0, V(n_\ell) \in A_1 \times A_3)$ from below by a function $T'_l(x_0, A_1, A_3)$, which is continuous in x_0.

Define two events:

$$D_\ell = \{\sum_{j=1}^{k_0^i} \xi_i^j \leq \frac{n_\ell}{4}, \xi_i^{k_0^i+1} \geq 2n_\ell \text{ for } i = 1, 3\},$$

for large n_l it has a positive probability, since we assume that the distribution of ξ_1^1, ξ_3^1 has infinite support.

$$E_{L,\ell} = \{\xi_i^j \leq L, j = 1, \ldots, \ell + k_0^i \text{ for } i = 2, 4\},$$

where L is taken large enough such that,

$$\epsilon_{L,\ell} = P(E_{L,\ell}) > 0.$$

If we require that

$$n_\ell > 4\ell + 2(\ell - 1)L + 2\frac{n_\ell}{4} + (k_0^1 + k_0^3)L, \tag{18}$$

that is set n_ℓ to

$$n_\ell > 8\ell + 4(\ell - 1)L + 2(k_0^1 + k_0^3)L,$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355

then we have that the event $D_\ell \cap E_{L,\ell}$ implies that at time n_ℓ, $Z(n_\ell) = 0$ and server 1 (server 3) is engaged in push activity 1 (push activity 3) with the long $k_0^1 + 1$st ($k_0^3 + 1$st) job from the infinite supply. To see this, recall that our policy is head of the line with low priority to push activities. Therefore prior to the first time that the servers are both working on the long push activities, at least one of them is working on pull activities or on the first k_0^i push activities. The expression (18) is an upper bound on the total amount of work that has to be done, and it will therefore be completed by time n_l. The long push activities will of course not be complete by time n_l.

With the above definitions in hand,

$$\begin{aligned} &P_{x_0}(Z(n_\ell) = 0, V(n_\ell) \in A_1 \times A_3) \\ &\quad \geq P_{x_0}(Z(n_\ell) = 0, V(n_\ell) \in A_1 \times A_3, D_\ell, E_{L,\ell}) \\ &\quad = P_{x_0}(V(n_\ell) \in A_1 \times A_3, D_\ell, E_{L,\ell}) \\ &\quad = \epsilon_{L,\ell} P_{x_0}(V(n_\ell) \in A_1 \times A_3, D_\ell | E_{L,\ell}). \end{aligned}$$

The number of jobs to be processed by activity $i = 2, 4$ by time n_ℓ, apart from the residuals, is

$$\ell_i = Q_i(0) - I\{Q_i(0) > 0\} + k_0^{i-1}.$$

Now define the truncation $\zeta_i^j = I\{\xi_i^j \leq L\}\, \xi_i^j$ for $i = 2, 4$, and observe that when D_ℓ occurs and conditional on $E_{L,\ell}$,

$$\begin{aligned} V_1(n_\ell) &= V_1(0) + U_4(0) + \sum_{j=1}^{k_0^1} \xi_1^j + \sum_{j=1}^{\ell_4} \xi_4^j + \xi_1^{k_0^1+1} - n_\ell \\ &= V_1(0) + U_4(0) + \sum_{j=1}^{k_0^1} \xi_1^j + \sum_{j=1}^{\ell_4} \zeta_4^j + \xi_1^{k_0^1+1} - n_\ell, \end{aligned}$$

with a similar expression for $V_3(n_\ell)$.

Denote the distribution of ξ_i^1 by η_i and the k_0^i fold convolutions of these distributions by $\eta_i^{*k_0^i}$ for $i = 1, 3$. Also, for $i = 2, 4$, use η_i' to denote the distribution of $\sum_{j=1}^{\ell_i} \zeta_i^j$.

We now have

$$\begin{aligned} &P_{x_0}(V(n_\ell) \in A_1 \times A_3, D_\ell | E_{L,\ell}) = \\ &\int I_{s_1,s_3,t_1,t_3,r_2,r_4} \eta_1^{*k_0^1}(ds_1) \eta_3^{*k_0^3}(ds_3) \eta_1(dt_1) \eta_3(dt_3) \eta_2'(dr_2) \eta_4'(dr_4) \end{aligned}$$

where the integral is on the range $(s_1, s_3, t_1, t_3, r_2, r_4) \in [0, \infty)^6$, and the integrand is the indicator function

$$\begin{aligned} &I_{s_1,s_3,t_1,t_3,r_2,r_4} = \\ &I\{V_1(0) + U_4(0) + s_1 + r_4 + t_1 - n_\ell \in A_1\} \cdot \\ &I\{V_3(0) + U_2(0) + s_3 + r_2 + t_3 - n_\ell \in A_3\} \cdot \\ &I\{s_1 \leq \tfrac{n_\ell}{4}\} I\{s_3 \leq \tfrac{n_\ell}{4}\} I\{t_1 \geq 2n_\ell\} I\{t_3 \geq 2n_\ell\}. \end{aligned} \tag{19}$$

We now use Assumption (A2b) to get,

$$\begin{aligned} &P_{x_0}(V(n_\ell) \in A, G_{n_\ell} | E_{L,\ell}) \geq \\ &\int I_{s_1,s_3,t_1,t_3,r_2,r_4} q_1(s_1) ds_1 q_3(s_3) ds_3 \eta_1(dt_1) \eta_3(dt_3) \eta_2'(dr_2) \eta_4'(dr_4) \end{aligned} \tag{20}$$

We define the function $T_\ell'(x_0, A)$ as $\epsilon_{L,\ell}$ multiplied by the integral in (20). It is evident that T_ℓ' is continuous in each of the coordinates $V_1(0), V_3(0), U_2(0), U_4(0)$ and hence it is continuous in x_0. It is also strictly positive, as required.

For every x_0 this $T_\ell'(x_0, A_1 \times A_3)$ is now defined for $A_1 \times A_3 \in \mathcal{B}(\mathbb{R}^2)$. We can extend it to a measure on the whole $\mathcal{B}(\mathbb{R}^2)$, so that $T_\ell'(x_0, A) > 0$ for every $A \in \mathcal{B}(\mathbb{R}^2)$ with positive Lebesgue measure, and so that $T_\ell'(x_0, A)$ is continuous in x_0 and satisfies

$$P_{x_0}(Z(n_\ell) = 0, V(n_\ell) \in A) \geq T_\ell'(x_0, A).$$

The remainder of the construction of the continuous component T follows exactly the steps of Meyn and Down [21].

Remark: The above proof can be extended to a proof for petiteness of compacts of the network state process of a multi-class queueing network with infinite virtual queues (c.f Section 3 of [26]) operating under a policy that gives lowest priority to the infinite virtual queues. Writing this statement and proof does not require any further ideas than those presented here.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4355
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4355