Performance Evaluation 67 (2010) 201–217

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Positive Harris recurrence and diffusion scale analysis of a push pull queueing network

Yoni Nazarathy *, Gideon Weiss

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

ARTICLE INFO

Article history:
Received 1 December 2008
Received in revised form 18 August 2009
Accepted 14 September 2009
Available online 13 October 2009

Keywords:
Queueing networks
Push pull
Infinite virtual queues
Fluid models
Positive Harris recurrence
Diffusion limits
Petite bounded sets

ABSTRACT

We consider a push pull queueing network with two servers and two types of job which are processed by the two servers in opposite order, with stochastic generally distributed processing times. This push pull network 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. 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 fluid and diffusion scale analysis of this network under such policies, to show fluid stability, positive Harris recurrence, and to obtain a diffusion limit for the network. On the diffusion scale the network is empty, and the departures of the two types of job are highly negatively correlated Brownian motions. Using similar methods we also derive a diffusion limit of a re-entrant line with an infinite supply of work.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

We consider the following queueing network: There are two servers, numbered 1, 2 and two types of job numbered 1, 2. Each type of job is processed by both servers. Type 1 is processed first by server 1 and then by server 2, while type 2 is processed in the opposite order, first by server 2 and then by server 1, see Fig. 1. We call the first step of each type a *push activity* and the second step a *pull activity*. We denote push activities of type i by $(i, 1)$ and pull activities by $(i, 2)$.

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, Q_1 and Q_2 indexed by the job type: jobs of type 1, waiting to be pulled by server 2 are in Q_1 and jobs of type 2, waiting to be pulled by server 1 are in Q_2.

Our network operates in continuous time $t \geq 0$. We denote by $Q_i(t)$ the number of jobs in the queue i at time t (including the job in process), and by $D_{i,j}(t)$ the number of jobs that have completed activity (i, j) during the time interval $[0, t]$. Thus $D_{i,2}(t)$ are the numbers of departures from the network of type i up to time t. When $Q_2(t) > 0$, server 1 can either pull, by serving a type 2 job from Q_2 or push, by serving a type 1 job from the infinite supply. When $Q_2(t) = 0$ server 1 can still always push jobs of type 1 into Q_1. Hence, server 1 never needs to idle. Similarly for server 2.

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 (IVQ): it acts like an infinite queue while in fact it only contains a few

* Corresponding author. Tel.: +972 526290317; fax: +972 48253849.
E-mail addresses: yonin@stat.haifa.ac.il, ynazarat@netvision.net.il (Y. Nazarathy), gweiss@stat.haifa.ac.il (G. Weiss).

0166-5316/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.peva.2009.09.010

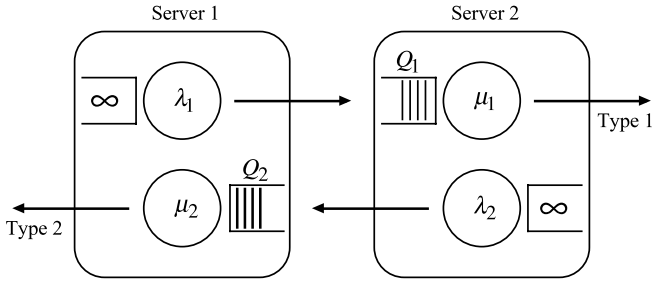

Fig. 1. The push pull network.

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. *A major point of this paper is that it is possible to find policies for the push pull network which never idle and yet keep the queues $Q_i(t)$ stable*. The push pull network was introduced by Kopzon et al. [1,2] who assumed exponential processing times. Infinite supply of work and infinite virtual queues are discussed in [3–8]. A brief survey of these results is in Chapter 2 of [9].

Assume that the long term average processing time of a push activity $(i, 1)$, is $1/\lambda_i$ and that of a pull activity $(i, 2)$, is $1/\mu_i$. Let $\theta_{i,j}$ be the long term fraction of time spent by the server working on activity (i, j). If the servers are never idle then $\theta_{1,1} + \theta_{2,2} = 1$ and $\theta_{2,1} + \theta_{1,2} = 1$. Furthermore, if $Q_i(t)$ are stable then their input and output rates are equal, so:

$$\nu_1 = \lambda_1\theta_{1,1} = \mu_1(1 - \theta_{2,1}), \qquad \nu_2 = \lambda_2\theta_{2,1} = \mu_2(1 - \theta_{1,1}),$$

where ν_i is the long term average rate of the departure process $D_{i,2}(t)$. Solving the equations we get:

$$\nu_1 = \frac{\lambda_1\mu_1(\lambda_2 - \mu_2)}{\lambda_1\lambda_2 - \mu_1\mu_2}, \qquad \nu_2 = \frac{\lambda_2\mu_2(\lambda_1 - \mu_1)}{\lambda_1\lambda_2 - \mu_1\mu_2}.$$

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

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

Inherently unstable network: When $\lambda_i > \mu_i$, $i = 1, 2$, service of each type of jobs alone, by both servers results in an unstable single server queue. In this case the priority to pull over push is unstable. A policy that works here is for each server to push when the queue in the opposite server is below a threshold. Specifically: while Q_1 is below some threshold, server 1 will push work to server 2, and server 1 will only pull from Q_2 when Q_1 is above the threshold, with a similar rule for server 2. We use an affine threshold (switching curve) to determine the pull or push preemptive head of the line priority. We define a family of such policies, each determined by slope constants κ_1, κ_2 and shift constants β_1, β_2, with $\kappa_i > 0$ and $\beta_i \geq 0$, $i = 1, 2$.

Server 1: At time t, priority to pull over push if $0 < Q_2(t) < \beta_1 + \kappa_1 Q_1(t)$.

Server 2: At time t, priority to pull over push if $0 < Q_1(t) < \beta_2 + \kappa_2 Q_2(t)$.

We refer to this as *Case 2*, and to the policy as an *affine threshold policy*, see Fig. 2.

Unbalanced network: If $\lambda_1 > \mu_1$ and $\mu_2 > \lambda_2$, then server 2 has more work to do than server 1, for both types of job, and the network cannot be stable unless server 1 idles some of the time. Similarly for $\lambda_1 < \mu_1$ and $\mu_2 < \lambda_2$. We will not consider this case any further in this paper.

Completely balanced network: When $\lambda_i = \mu_i$, $i = 1, 2$ it is possible to find policies which work with full utilization of both servers, and are rate stable, i.e. the input and output rates of each queue are equal, however these rates are not uniquely determined. We can choose $0 \leq \theta \leq 1$, and specify $\theta_{1,1} = \theta_{1,2} = \theta$, $\theta_{2,1} = \theta_{2,2} = 1 - \theta$ and use $\nu_1 = \mu_1\theta$ as a nominal rate for type 1 and $\nu_2 = \mu_2(1 - \theta)$ as a nominal rate for type 2. As shown in [7], we can use an adaptation of the maximum pressure policy of Dai and Lin [10] to serve jobs of types 1 and 2 at these rates, under full utilization. However, the network will become congested, with expected $O(\sqrt{T})$ jobs in the network at time T. We conjecture that this cannot be improved.

The structure of the paper is as follows: 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. 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 [11] to show that this Markov process is positive Harris recurrent, and so $Q_1(t)$, $Q_2(t)$ possess a stationary limiting distribution. Section 5.1 contains a technical result: we show that for multi-class queueing networks with infinite virtual queues all bounded sets of states are uniformly small. In Section 6 we

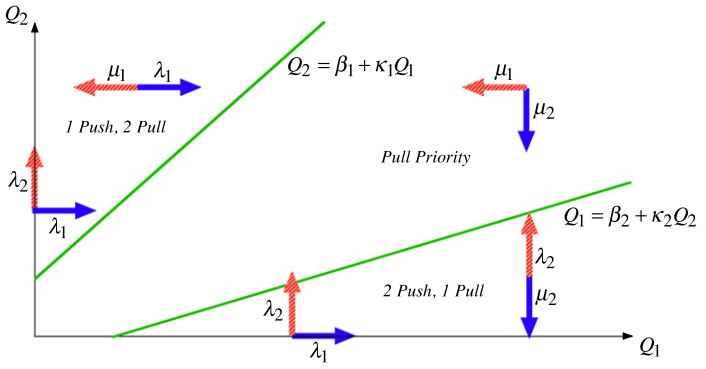

Fig. 2. The affine threshold policy for the inherently unstable network (Case 2).

consider the departure processes under diffusion scaling, and obtain a Brownian limit theorem. The limit result immediately yields asymptotic variance rate parameters for the departure processes and shows that the two departure streams are highly negatively correlated. For comparison, we also present an alternative (less general) derivation of the asymptotic variance rate parameters by means of a renewal-reward approach. Our Brownian limit method is also useful for other models: We exploit this in Section 7 where we present a Brownian approximation result for the departure process of an infinite supply re-entrant line.

Remark. Some of the results of this paper are based on the earlier conference paper [12].

1.1. Notation

In general, when no ambiguity may arise, we omit index subscripts when we refer to vectors. Further, when we do not specify explicit values for i in expressions such as (for example) Q_i, ρ_i, we imply that $i = 1, 2$. Similarly, when we refer to activities (i, j) we mean that $i = 1, 2$ and $j = 1, 2$.

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. We use $I\{\cdot\}$ for the indicator function of event $\{\cdot\}$. 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 $t \ge 0$, $\lim_{r \to \infty} \|f_r - f\|_t = 0$.

2. Preliminary discussion: Comparing to the KSRS network

The Kumar–Seidman Rybko–Stolyar multi-class queueing network (cf. Chapter 8 of [13] or Section 2.9 of [14]) differs from our push pull network in that instead of an 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_2.

In that case there are 4 queues: $Q_{1,1}$, $Q_{1,2}$ of job type 1 and $Q_{2,1}$, $Q_{2,2}$ of job type 2. The offered loads for servers 1 and 2 are $\rho_1 = \alpha_1/\lambda_1 + \alpha_2/\mu_2$ and $\rho_2 = \alpha_2/\lambda_2 + \alpha_1/\mu_1$ respectively. A necessary condition for stability is $\rho_i < 1$.

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

The discovery of this phenomenon by Kumar and Seidman [15] (deterministic processing times) and by Rybko and Stolyar [16] (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 $\rho_i < 1$ then there is some work conserving (non idling) policies which keeps 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. [17]). Balanced heavy traffic in the KSRS network occurs when $\alpha_1 \to \nu_1$, $\alpha_2 \to \nu_2$. 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 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, $\hat{Q}^n(t) = Q^n(nt)/\sqrt{n}$ will converge to a 4 dimensional reflected Brownian motion.

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 $\rho_i = 1$ and jobs leave the network at the rates ν_i. 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 $Q_1(t)$, $Q_2(t)$ is expected to be $O(1)$, and it is 0 under diffusion scaling.

Finally, compare the behavior of the departure processes, $D_{i,2}(t)$ of the KSRS network and of the push pull network, under diffusion scaling. In the KSRS network with $\rho_i < 1$ 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 departure 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 departure processes. To the best of our knowledge the behavior of the departure processes in that case has not been investigated. We note that even the departure process of a single server queue, under balanced heavy traffic, poses some as yet unanswered questions (c.f. [18,19]).

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 departure processes of the two types of job. What we find is that the departure 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.

3. The stochastic model

We assume that the processing durations of activities (i, j) are drawn from a sequence of positive random variables: $\xi_{i,j} = \{\xi_{i,j}^\ell, \ell = 1, 2, \ldots\}$. The assumptions that we make regarding the processing durations are as follows:

$$\text{(A1)} \quad \lim_{n\to\infty} \frac{\sum_{\ell=1}^{n} \xi_{i,1}^\ell}{n} = \frac{1}{\lambda_i}, \qquad \lim_{n\to\infty} \frac{\sum_{\ell=1}^{n} \xi_{i,2}^\ell}{n} = \frac{1}{\mu_i}, \quad \text{a.s., for some } \lambda_i, \mu_i \in (0, \infty).$$

$$\text{(A2)} \quad \begin{cases} \text{(a)} & \xi_{i,j} \text{ are mutually independent i.i.d. sequences.} \\ \text{(b)} & P(\xi_{i,1}^1 \geq x) > 0 \quad \text{for all } x > 0, \\ & \exists L_0^i > 0,\ q_i(\cdot) \geq 0 \quad \text{with } \int_0^\infty q_i(x)\mathrm{d}x > 0 : P\left(\sum_{\ell=1}^{L_0^i} \xi_{i,1}^\ell \in \mathrm{d}x\right) \geq q_i(x)\mathrm{d}x. \\ \text{(b}') & \text{Compact sets are petite.} \end{cases}$$

$$\text{(A3)} \quad \lambda_i^2 \mathrm{Var}(\xi_{i,1}^1) = c_{i,1}^2, \qquad \mu_i^2 \mathrm{Var}(\xi_{i,2}^1) = c_{i,2}^2, \quad \text{for some } c_{i,1}^2, c_{i,2}^2 \in [0, \infty).$$

Assumptions (A1) require that there exist strong laws of large numbers for the sequences of processing times so that the rate of the push activities is λ_i and the rate of the pull activities is μ_i. Assumptions (A2) are to be used in a Markov process setting to prove positive Harris recurrence. (a) implies renewal processing. A further technical assumption regarding the processing times of the push activities is (b): unbounded and spread-out processing times. Alternatively, we may assume (b′), this assumption is to be made precise in Section 5. We show that under the pull priority policy, (b) implies (b′). Assumptions (A3) require the existence of second moments, with squared coefficients of variation $c_{i,j}^2$. We shall make use of Assumptions (A1)–(A3) incrementally.

We associate counting processes with each activity (i, j):

$$S_{i,j}(t) = \sup\left\{n : \sum_{\ell=1}^{n} \xi_{i,j}^\ell \leq t\right\}, \quad t \geq 0.$$

We denote by $T_{i,j}(t)$ the total time that the server of activity (i, j) allocates to the processing of the activity during the interval $[0, t]$. We require that $T_{i,j}(0) = 0$, $T_{i,j}(\cdot)$ are nondecreasing, and $T_{i,j}(t) - T_{i,j}(s) \leq t - s$ for $s < t$. Under our policies of full utilization, the servers never idle, thus:

$$T_{1,1}(t) + T_{2,2}(t) = t, \qquad T_{2,1}(t) + T_{1,2}(t) = t. \tag{1}$$

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

The number of jobs that have completed processing of activity (i, j) by time t is $D_{i,j}(t) = S_{i,j}(T_{i,j}(t))$. Let $Q_i(0)$ be the initial queue lengths. The number of jobs at time t is:

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

We further require that $Q_i(t) \geq 0$.

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 the push activities $(i, 1)$ unless the corresponding queue of the server is empty. This implies that the allocation processes $T(\cdot)$ need to satisfy:

$$\int_0^t Q_2(s)\mathrm{d}T_{1,1}(s) = 0, \qquad \int_0^t Q_1(s)\mathrm{d}T_{2,1}(s) = 0.$$

In the inherently unstable Case 2, we use an affine threshold policy. The affine threshold for server 1 is the line $Q_2(t) = \beta_1 + \kappa_1 Q_1(t)$. Server 1 will give preemptive priority to the pull activity $(2, 2)$ only if $0 < Q_2(t) < \beta_1 + \kappa_1 Q_1(t)$, and in that case it will not allocate time to the push activity $(1, 1)$. On the other hand, if $Q_2(t) \geq \beta_1 + \kappa_1 Q_1(t)$ then server 1 will give priority to activity $(1, 1)$, to prevent starvation at the queue of server 2 (Q_1), and will not allocate time to activity $(2, 2)$. A symmetric rule is used by server 2, with the affine threshold given by the line $Q_1(t) = \beta_2 + \kappa_2 Q_2(t)$. Hence, for the inherently unstable Case 2:

$$\int_0^t \mathbf{1}\{0 < Q_2(s) < \beta_1 + \kappa_1 Q_1(s)\}\mathrm{d}T_{1,1}(s) = 0, \qquad \int_0^t \mathbf{1}\{Q_1(s) \geq \beta_2 + \kappa_2 Q_2(s)\}\mathrm{d}T_{1,2}(s) = 0$$
$$\int_0^t \mathbf{1}\{Q_2(s) \geq \beta_1 + \kappa_1 Q_1(s)\}\mathrm{d}T_{2,2}(s) = 0, \qquad \int_0^t \mathbf{1}\{0 < Q_1(s) < \beta_2 + \kappa_2 Q_2(s)\}\mathrm{d}T_{2,1}(s) = 0$$

4. Fluid limits and fluid models

In this section we assume (A1), and consider the behavior of the push pull network under fluid scaling. To study the network under fluid scaling we consider the six dimensional network process $Y(t) = (Q(t), T(t))$, and parameterize it by $n = 1, 2, \ldots$ 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 $\xi_{i,j}$. 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:

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

It may now be shown that under Assumption (A1), and assuming

$$\liminf_{n\to\infty} Q^n(0)/n < \infty,$$

(see also the remark at the end of this section) that 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:

$$\begin{aligned} &\bar{Q}_i(t) = \bar{Q}_i(0) + \lambda_i \bar{T}_{i,1}(t) - \mu_i \bar{T}_{i,2}(t), \\ &\bar{Q}_i(t) \geq 0, \\ &\bar{T}_{i,j}(0) = 0, \quad \bar{T}_{i,j}(\cdot) \text{ is non-decreasing} \end{aligned} \tag{3}$$

as well as

$$\bar{T}_{1,1}(t) + \bar{T}_{2,2}(t) = t, \qquad \bar{T}_{2,1}(t) + \bar{T}_{1,2}(t) = t \tag{4}$$

and in addition, under the pull priority they satisfy:

$$\int_0^t \bar{Q}_2(s)\mathrm{d}\bar{T}_{1,1}(s) = 0, \qquad \int_0^t \bar{Q}_1(s)\mathrm{d}\bar{T}_{2,1}(s) = 0. \tag{5}$$

For details, see for example Theorem 4.1 of [11] or Appendix A.2 of [10]. Further, under the affine threshold policy:

$$\begin{aligned} &\int_0^t \mathbf{1}\{0 < \bar{Q}_2(s) < \kappa_1 \bar{Q}_1(s)\}\mathrm{d}\bar{T}_{1,1}(s) = 0, \qquad \int_0^t \mathbf{1}\{\bar{Q}_1(s) \geq \kappa_2 \bar{Q}_2(s)\}\mathrm{d}\bar{T}_{1,2}(s) = 0 \\ &\int_0^t \mathbf{1}\{\bar{Q}_2(s) \geq \kappa_1 \bar{Q}_1(s)\}\mathrm{d}\bar{T}_{2,2}(s) = 0, \qquad \int_0^t \mathbf{1}\{0 < \bar{Q}_1(s) < \kappa_2 \bar{Q}_2(s)\}\mathrm{d}\bar{T}_{2,1}(s) = 0. \end{aligned} \tag{6}$$

Eqs. (3)–(6) represent a deterministic continuous fluid analog of the stochastic model introduced in the previous section. Note that in the fluid scaling the shift constants β_1, β_2 have disappeared. We shall refer to Eqs. (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 $t \geq \delta$.

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 affine 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 $Y(nt)/n$, $D(nt)/n$ will converge as $n \to \infty$ u.o.c. to a fluid limit $\bar{Y}(t) = (\bar{Q}(t), \bar{T}(t)), \bar{D}(t)$ which satisfies: $\bar{Q}_i(t) = 0$, $\bar{T}_{i,j}(t) = \theta_{i,j}t$, $\bar{D}_{i,j}(t) = \nu_i t$.*

The proof of Theorem 1 is by means of a Lyapounov function, f. As in [20], we shall make use of the following elementary Lemma 1. Recall that $T_{i,j}(t)$ are Lipschitz with constant 1. It then follows that $\bar{T}_{i,j}$, 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 $f(t) = \bar{Q}_1(t) + \bar{Q}_2(t)$. 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 $|Q(0)| = 1$, $\bar{Q}(t) = 0$ for $t \geq B/\epsilon$ which proves stability of the fluid model.

Define $\epsilon = \min\{\mu_1 - \lambda_1, \mu_2 - \lambda_2\}$. The rate parameters of Case 1 ensure that $\epsilon > 0$. We bound $\dot{f}(t)$ by $-\epsilon$ for all regular time points t at which $f(t) > 0$. Note that at any regular time point:

$$\dot{\bar{Q}}_i = \lambda_i \dot{\bar{T}}_{i,1} - \mu_i \dot{\bar{T}}_{i,2}. \tag{7}$$

We now analyze all possible values of $\bar{Q}_i(t)$:

- Assume $\bar{Q}_1(t), \bar{Q}_2(t) > 0$:
  By (5), $\dot{\bar{T}}_{1,1} = \dot{\bar{T}}_{3,1} = 0$ and thus by (4), $\dot{\bar{T}}_{1,2} = \dot{\bar{T}}_{2,2} = 1$. As a consequence, $\dot{\bar{Q}}_i(t) = -\mu_i$ and,
  $$\dot{f} = -(\mu_1 + \mu_2) \leq -\epsilon.$$
- Assume $\bar{Q}_1(t) > 0, \bar{Q}_2(t) = 0$:
  By (5) $\dot{\bar{T}}_{2,1} = 0$ and thus by (4), $\dot{\bar{T}}_{1,2} = 1$. As a consequence,
  $$\dot{f} = \lambda_1 \dot{\bar{T}}_{1,1} - \mu_1 - \mu_2 \dot{\bar{T}}_{2,2} = \lambda_1 - \mu_1 - (\lambda_1 + \mu_2)\dot{\bar{T}}_{2,2} \leq -(\mu_1 - \lambda_1) \leq -\epsilon.$$
- Assume $\bar{Q}_1(t) = 0, \bar{Q}_2(t) > 0$:
  Similarly to the previous argument,
  $$\dot{f} \leq -(\mu_2 - \lambda_2) \leq -\epsilon.$$

This completes the proof for Case 1.

Case 2: We use the same technique as in Case 1. First we choose positive constants, d, h and g as follows:

$$d < \frac{\lambda_1}{\lambda_2}, \frac{\lambda_2}{\lambda_1}, \qquad h < d, \frac{1}{\kappa_1}, \frac{1}{\kappa_2}, \qquad g > \frac{\lambda_1}{\lambda_2}, \frac{\lambda_2}{\lambda_1}, \sqrt{\frac{\kappa_1}{\kappa_2}}, \sqrt{\frac{\kappa_2}{\kappa_1}}.$$

Now our Lyapounov function is:

$$f(\bar{Q}_1(t), \bar{Q}_2(t)) = \begin{cases} \eta_1\eta_2(\mathrm{d}\bar{Q}_1(t) - \bar{Q}_2(t)) & \text{if } \bar{Q}_2(t) \leq h\bar{Q}_1(t), \\ \eta_1\eta_2^{-1}(g\bar{Q}_1(t) - \bar{Q}_2(t)) & \text{if } h\bar{Q}_1(t) < \bar{Q}_2(t) \leq \sqrt{\dfrac{\kappa_1}{\kappa_2}}\bar{Q}_1(t), \\ \eta_1^{-1}\eta_2^{-1}(g\bar{Q}_2(t) - \bar{Q}_1(t)) & \text{if } h\bar{Q}_2(t) < \bar{Q}_1(t) < \sqrt{\dfrac{\kappa_2}{\kappa_1}}\bar{Q}_2(t), \\ \eta_1^{-1}\eta_2(\mathrm{d}\bar{Q}_2(t) - \bar{Q}_1(t)) & \text{if } \bar{Q}_1(t) \leq h\bar{Q}_2(t). \end{cases}$$

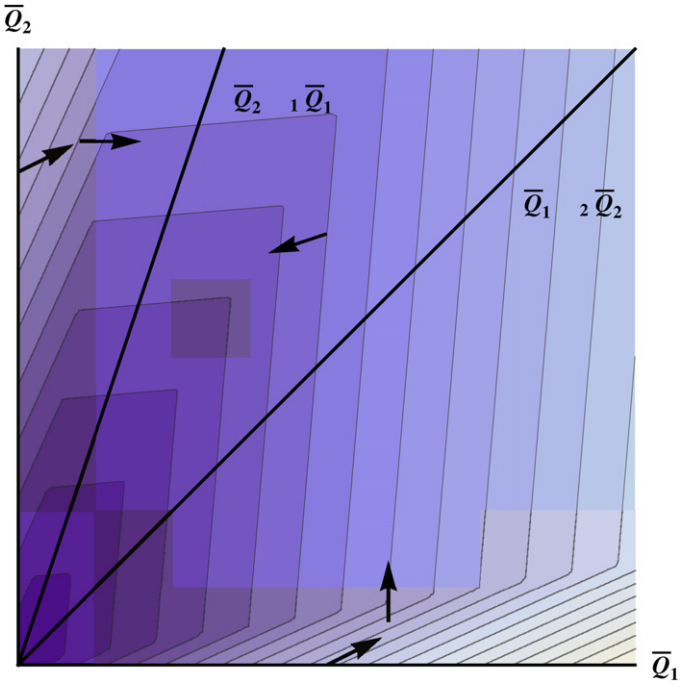

Fig. 3. Illustration of the Lyapunov function for the inherently unstable network (Case 2). Arrows point in the drift directions.

With,

$$\eta_1 = \sqrt{\frac{\sqrt{\kappa_1}g - \sqrt{\kappa_2}}{\sqrt{\kappa_2}g - \sqrt{\kappa_1}}} > 0 \quad \text{and} \quad \eta_2 = \sqrt{\frac{g-h}{d-h}} > 0.$$

The Lyapunov function is illustrated in Fig. 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 $|\bar{Q}(0)| = \underline{1}$ then $f(0)$ is bounded by some finite value B. Furthermore, it is straightforward to see that f is continuous in the values of $\bar{Q}_1(t), \bar{Q}_2(t)$. We now bound $\dot{f}(t)$ for all regular time points t at which $f(t) > 0$, by analyzing all possible values of $\bar{Q}_i(t)$. We again use the dynamics (7):

- Assume $\frac{1}{\kappa_2}\bar{Q}_1(t) < \bar{Q}_2(t) \leq \sqrt{\frac{\kappa_1}{\kappa_2}}\bar{Q}_1(t)$:

  Then $f(t) = \eta_1\eta_2^{-1}(g\bar{Q}_1(t) - \bar{Q}_2(t))$ and by (6) we have that $\dot{\bar{T}}_{1,1} = \dot{\bar{T}}_{2,1} = 0$ and thus $\dot{\bar{T}}_{1,2} = \dot{\bar{T}}_{2,2} = 1$. Hence by (7):

  $$\dot{f} = -\eta_1\eta_2^{-1}(g\mu_1 - \mu_2) < 0.$$

- Assume $0 < \bar{Q}_2(t) \leq \frac{1}{\kappa_2}\bar{Q}_1(t)$:

  By (6) we have that $\dot{\bar{T}}_{1,1} = \dot{\bar{T}}_{1,2} = 0$ and thus $\dot{\bar{T}}_{2,1} = \dot{\bar{T}}_{2,2} = 1$. Now look at two cases: If $(h\bar{Q}_1(t) \leq \bar{Q}_2(t))$ then $f(t) = \eta_1\eta_2^{-1}(g\bar{Q}_1(t) - \bar{Q}_2(t))$ and by (7):

  $$\dot{f} = -\eta_1\eta_2^{-1}(\lambda_2 - \mu_2) < 0.$$

  Alternatively, if $\bar{Q}_2(t) < h\bar{Q}_1(t)$ then $f(t) = \eta_1\eta_2(\mathrm{d}\bar{Q}_1(t) - \bar{Q}_2(t))$ and:

  $$\dot{f} = -\eta_1\eta_2(\lambda_2 - \mu_2) < 0.$$

- Assume $\bar{Q}_1(t) > 0, \bar{Q}_2(t) = 0$:

  By (6) we have that $\dot{\bar{T}}_{1,2} = 0$ and thus $\dot{\bar{T}}_{2,1} = 1$. Note that we can not use (6) to explicitly obtain $\dot{\bar{T}}_{1,1}$ and $\dot{\bar{T}}_{2,2}$. In this region, $f(t) = \eta_1\eta_2(\mathrm{d}\bar{Q}_1(t) - \bar{Q}_2(t))$ and thus by (7):

  $$\begin{aligned}
  \dot{f} &= \eta_1\eta_2(d\lambda_1\dot{\bar{T}}_{1,1} - (\lambda_2 - \mu_2\dot{\bar{T}}_{2,2})) \\
  &= \eta_1\eta_2(d\lambda_1\dot{\bar{T}}_{1,1} - (\lambda_2(\dot{\bar{T}}_{1,1} + \dot{\bar{T}}_{2,2}) - \mu_2\dot{\bar{T}}_{2,2})) \\
  &= -\eta_1\eta_2((\lambda_2 - d\lambda_1)\dot{\bar{T}}_{1,1} + (\lambda_2 - \mu_2)\dot{\bar{T}}_{2,2}) \\
  &\leq -\eta_1\eta_2 \min\{\lambda_2 - d\lambda_1, \lambda_2 - \mu_2\} \\
  &< 0.
  \end{aligned}$$

The remaining cases of $\bar{Q}_1(t) < \sqrt{\frac{\kappa_2}{\kappa_1}}\bar{Q}_2(t)$ are symmetric and yield similar bounds. All the bounds above are negative constants, and we choose $-\epsilon$ as their maximum. This completes the proof. ■

Remark. So far in this section we assumed that the nth network starts with queue lengths $Q^n(0)$, and that all the jobs in the network had no previous processing, so that the $S_{i,j}(t)$ are counting processes, with intervals $\xi_{i,j}$ which have long term rates as specified in assumption (A1). 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 $\tilde{\xi}_{i,j}$ be the residual processing time of this first job. Then the first interval is a residual processing time with a different distribution from the other $\xi^{\ell}_{i,j}$, $\ell > 1$. In that case $S_{i,j}(t)$ are delayed counting processes. We now associate with the nth network an initial state consisting of $Q^n_i(0)$, $\tilde{\xi}^n_{i,j}$. All the results of this section remain valid and unchanged as long as we assume that $\tilde{\xi}^n_{i,j}/n \to 0$ a.s. (see [21]).

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 [11], see also [22].

We begin by defining the network state process. Denote by $U_i(t)$, $V_i(t)$, $i = 1, 2$ the residual processing times of the head of the line activities which are in process or preempted at the current time t. $U_i(t)$ is for the pull activity of type i and $V_i(t)$ is for the push activity of type i. 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. [23]):

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 $x \in \mathbb{S}$, $B \in \mathcal{B}(\mathbb{S})$,

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

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

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

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 $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$, 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 $\pi(f) = \int_{\mathbb{S}} f(x)\, \pi(\mathrm{d}x)$ whenever the integral makes sense. Then if $|\pi(f)| < \infty$:

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(s))\mathrm{d}s = \pi(f) \quad P_x \text{ a.s. for each } x \in \mathbb{S}. \tag{8}$$

Ergodicity of the process X is a stronger property: A positive Harris recurrent process $X(t)$ is *ergodic* if $P^t(x, \cdot)$ converges to π in total variation norm:

$$\lim_{t\to\infty} \sup_{B\in\mathcal{B}(\mathbb{S})} |P^t(x, B) - \pi(B)| = 0 \quad \text{for all } x \in \mathbb{S}.$$

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

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

Petiteness of A may be interpreted as the property that all sets B are "equally accessible" from any $x \in A$. If for some closed set A, the recurrence time τ_A satisfies $P_x(\tau_A < \infty) = 1$ and A is petite then $X(t)$ is Harris recurrent (see Theorem 4.1 in [22]).

A non-empty set A is said to be *small* with respect to ν if there exists a fixed n such that for all $x \in A$:

$$P^n(x, B) \geq \nu(B), \quad \text{for all } B \in \mathcal{B}(\mathbb{S}).$$

A non-empty set A is said to be *uniformly small* with respect to ν if this holds for all $n \in [s_1, s_2]$ for some $s_1 < s_2$. If $X(t)$ is positive Harris recurrent and if for some closed set A the recurrence time τ_A satisfies $P_x(\tau_A < \infty) = 1$ and A is uniformly small then $X(t)$ is ergodic (see Theorem 4.3 in [22]).

For more on Markov processes, positive Harris recurrence and petite or small sets, see [24] for an introduction and discrete time results, and [25,26] for continuous time results. In the context of queueing networks the lecture notes of Bramson [22] give an excellent summary.

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

(A2b′) $A = \{x : |x| \leq m\}$ is petite for any $m > 0$.

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 an affine threshold policy. Furthermore, for Case* 1 *we may substitute Assumptions* (A2b′) *with (A2b), which implies in that case that the process X is ergodic.*

Proof. The proof uses the framework of Dai [11]. 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 into 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.

The technical Assumption (ii), that all closed bounded 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 multi-class queueing networks with an exogenous input stream the assumption (A2b), 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 [27], who proved the same result for generalized Jackson networks. This needs to be extended to the case of an infinite supply of work. The difference is that with an infinite supply of work the departure process from an infinite virtual queue is in general not independent of the state of the other queues. Guo and Zhang [6] have adapted Meyn and Down's ideas to a reentrant line with an infinite supply of work where the policy is to give lowest priority to the activity with the infinite supply.

The following Lemma 2 extends the results of Guo and Zhang [6], and shows that in Case 1, under pull priority, the Assumption (A2b) implies (A2b′), and hence positive Harris recurrence. In fact it is shown in Lemma 2 that Assumption (A2b) implies that all closed bounded sets $A = \{x : |x| \leq m\}$ are uniformly small. This implies that under (A2b) X is not only positive Harris recurrent, but is actually ergodic (see Theorem 4.3 in [22]). ■

We note that so far we have not been able to prove the equivalent of Lemma 2 for the affine threshold policies, nevertheless we believe it to be true. This would imply that assumption (A2b) can replace assumptions (A2b′) also for case 2.

5.1. Uniformly small property of bounded sets under pull priority policy

In this section we show that for pull priority policies, if the push activities have distributions which are spread out with unbounded support, then all closed and bounded sets of states are uniformly small. We prove this result not just for the push pull network but for a wider class of multi-class queueing network with infinite virtual queues.

We consider a multi-class queueing network with nodes $k \in \mathcal{K} = \{1, \ldots, K\}$, where node k serves one or more classes, and has one class which has an infinite virtual queue. There are no exogenous arrivals. We assume that routing of jobs between classes is deterministic, and processing times of jobs are independent, with those of each class being identically distributed. For the j job produced by the IVQ class of node k, we denote by $\xi_k(j)$ its first step processing time at node k, and by $\zeta_{kk'}(j)$ its subsequent processing time along the route, at node k'. Note that because routing is deterministic all the $\zeta_{kk'}(j)$ are independent.

Service to each class (push and pull) is head of the line preemptive resume (HL). For a given policy the state of the network $x \in \mathbb{S}$ consists of the queue lengths, the residual processing times of the head of the line job in each class, and some additional information on jobs in the network, which is needed by the policy (such as service completion times along the route for each of the jobs currently in the network), so that under this policy the state of the network at time t, $X(t)$, is a Markov process. We define a norm $|x|$ to be the sum of the queue lengths and the residual processing times of all the classes and denote the Borel sets of states by $\mathcal{B}(\mathbb{S})$.

A pull priority non-idling policy is a policy which keeps each node busy at all times, and which works on the IVQ class of each node only when all other queues at that node are empty.

A weak pull priority non-idling policy is a policy which keeps each node busy at all times, and which is processing some jobs from a non-IVQ class at all times at which not all of them are empty.

Lemma 2. *Consider a multi-class network with IVQs as above, under weak pull priority non-idling policy. Assume that the distributions of $\xi_k(1)$, $k \in \mathcal{K}$ have unbounded support and are spread out (Assumption* A2b *). Then every closed bounded set of states $A = \{x : |x| \leq m\}$ is uniformly small.*

Proof. Our proof is patterned on the proof of Proposition 4.7 in the Lecture Notes of Bramson [22].

Let V_k denote the total amount of processing on node k which is needed to process all the jobs initially in the network. The initial state x of the network includes residual processing times with total duration $\leq m$ and up to m jobs in the various queues. We index previous jobs which were generated by IVQ k (including the residual job) as $0, -1, \ldots, -m$. Then: $V_k \leq \tilde{V}_k = m + \sum_{k' \in \mathcal{K}} \sum_{j=0}^{m} \zeta_{k'k}(-j)$. Note that $\tilde{V}_k$ are independent. We can find $M, \epsilon_1 > 0$ such that $P(\tilde{V}_k \leq M) > \epsilon_1$.

By the spread out assumption, for each k there exists L_0^k such that $\sum_{j=1}^{L_0^k} \xi_k(j)$ has a continuous component. It follows that there exists $\tilde{L}_k$, $\delta_k > 0$ and some interval of length $M + 3$ such that $\Xi_k(\tilde{L}_k) = \sum_{j=1}^{\tilde{L}_k} \xi_k(j)$ satisfies: $P(\Xi_k(\tilde{L}_k) \in [t_1, t_2]) > \delta_k(t_2 - t_1)$ for all $[t_1, t_2]$ contained in the interval. Let $L = \max \tilde{L}_k$, and define $W_k(L) = \Xi_k(L) + \sum_{k' \in \mathcal{K}} \sum_{j=1}^{L} \zeta_{k'k}(j)$. Then $W_k(L)$ is at least as spread out as $\Xi_k(L)$. Hence, there exist some $a_k \geq M$ and an $\epsilon_2 > 0$ such that $P(W_k(L)) \in [t_1, t_2] > \epsilon_2(t_2 - t_1)$, for all $[t_1, t_2] \subseteq [a_k - M, a_k + 3]$.

Let $N = \sum_{k = \in \mathcal{K}} a_k + K(M + 3)$. By the assumption of unbounded support, there exist $b_k > N$ and ϵ_3 such that $P(\xi_k(L + 1) \in [b_k, b_k + 1]) > \epsilon_3$.

Define now the events:

$$
\begin{aligned}
&G_{1,k} = \{\omega : V_k \leq M\},\\
&G_{2,k} = \{\omega : V_k + W_k \in [a_k, a_k + 3]\},\\
&G_{3,k}(t_{1,k}, t_{2,k}) = \{\omega : V_k + W_k + \xi_k(L + 1) \in [t_{1,k}, t_{2,k}]\},\\
&G_1 = \bigcap_{k \in \mathcal{K}} G_{1,k}, \qquad G_2 = \bigcap_{k \in \mathcal{K}} G_{2,k}, \qquad G_3(\mathbf{t}_1, \mathbf{t}_2) = \bigcap_{k \in \mathcal{K}} G_{3,k}(t_{1,k}, t_{2,k}),\\
&G = G_1 \cap G_2 \cap G_3(\mathbf{t}_1, \mathbf{t}_2)
\end{aligned}
$$

where $\mathbf{t}_i = (t_{i,k}, k \in \mathcal{K})$. Also define the intervals $\mathcal{I}_k = [a_k + b_k + 1, a_k + b_k + 3]$.

We now show, for $\mathbf{t}_1, \mathbf{t}_2$ such that $[t_{1,k}, t_{2,k}] \subseteq \mathcal{I}_k$, $k \in \mathcal{K}$:

$$P(G) \geq (\epsilon_1 \epsilon_2 \epsilon_3)^K \prod_{k \in K} (t_{2,k} - t_{1,k}).$$

Let $\tilde{G}_{1,k} = \{\omega : \tilde{V}_k \leq M\}$, then $G_{1,k} \supseteq \tilde{G}_{1,k}$. We calculate:

$$
\begin{aligned}
P(\tilde{G}_{1,k} \cap G_{2,k} \cap G_{3,k}(t_{1,k}, t_{2,k})) &= \int_{b_k}^{b_{k+1}} \int_0^M P(W_k(L) \in [t_{1,k} - r - s, t_{2,k} - r - s]) P(\tilde{V}_k \in \mathrm{d}r) P(\xi_k(L + 1) \in \mathrm{d}s)\\
&> \epsilon_1 \epsilon_2 \epsilon_3 (t_{2,k} - t_{1,k})
\end{aligned}
$$

where we use the independence of V_k, $W_k(L)$, $\xi_k(L + 1)$ to write the integral, and we use:

$$
\begin{aligned}
&[t_{1,k} - r - s, t_{2,k} - r - s] \subseteq [a_k + b_k + 1 - r - s, a_k + b_k + 3 - r - s]\\
&\quad \subseteq [a_k - r, a_k + 3 - r] \subseteq [a_k - M, a_k + 3]
\end{aligned}
$$

to obtain the inequality. Furthermore, $(\tilde{V}_k, W_k(L), \xi_k(L + 1))$ for $k \in \mathcal{K}$ are independent, and so:

$$P(G) \geq P\left(\bigcap_{k \in \mathcal{K}} \tilde{V}_k \cap G_2 \cap G_3(\mathbf{t}_1, \mathbf{t}_2)\right) = \prod_{k \in \mathcal{K}} P(\tilde{G}_{1,k} \cap G_{2,k} \cap G_{3,k}(t_{1,k}, t_{2,k})).$$

Let $J(k, j)$ denote the jth job generated by the kth IVQ. We now argue that conditional on G, at the time N, all the initial jobs $J(k, j)$, $j \leq 0$ and all the jobs $J(k, j)$ with $1 \leq j \leq L$ will have completed all of their processing at all the nodes, and each node k will be processing the first step of $J(k, L + 1)$, at the IVQ class. First we note that because $\xi_k(L + 1) > b_k > N$, all the nodes will work only on the first jobs $J(k, j)$, $j \leq L$ somewhere along their routes, or on the first processing step of $J(k, L + 1)$ until at least the time N. Let T be the earliest time at which all the jobs $J(k, j)$, $j \leq L$ have completed their whole processing route. Assume that at some time t all the nodes k are processing the first step of $J(k, L + 1)$ simultaneously. Because we are using a weak pull priority policy, all the queues which are not IVQs must be empty at t, and also, because we use HL policy, each of the first operations of the jobs $J(k, L)$ must be completed. Hence $t \geq T$. Therefore at all $t < T$ at least one node k is not working on the first operation of $J(k, L + 1)$. Because our policy is non idling this implies that for all $t < T$ at least one node k is working on some job $J(k', j)$, $j \leq L$, $k' \in \mathcal{K}$. But this implies that $T < \sum_{k \in \mathcal{K}} (V_k + W_k(L)) < N$.

We have seen that $P_x(G) \geq (\epsilon_1 \epsilon_2 \epsilon_3)^K \prod_{k \in K} (t_{2,k} - t_{1,k})$ for every $|x| \leq m$, and $[t_{1,k}, t_{1,k}] \subseteq \mathcal{I}_k$. Take any time $s \in [N, N + 1]$. The state at that time is $X(s)$, which includes the vector $Z(s)$ of queue lengths of the non-IVQ classes, and $V_k(s)$ which is the

residual service time of the HL IVQ job of node k. Conditional on G, we have seen that $Z(s) = 0$ and $V_k(s) \in [t_{1,k} - s, t_{2,k} - s]$. Hence:

$$P_x Z(s) = 0, \qquad V_k(s) \in [t_{1,k} - s, t_{2,k} - s] \geq (\epsilon_1\epsilon_2\epsilon_3)^K \prod_{k\in K}(t_{2,k} - t_{1,k}).$$

Consider now the measure ν which is concentrated on states $Z(s) = 0$ and $V_k(s)$ in the rectangle:

$$\prod_{k\in\mathcal{K}}[a_k + b_k + 1 - N, a_k + b_k + 2 - N] \subseteq \prod_{k\in\mathcal{K}}[a_k + b_k + 1 - s, a_k + b_k + 3 - s]$$

and is proportional to K-dimensional Lebesgue measure on this rectangle, with proportionality constant $(\epsilon_1\epsilon_2\epsilon_3)^K$. Then for all $s \in [N, N + 1]$ the set $A = \{|x| \leq m\}$ is small with respect to the measure ν. ■

Unfortunately this proof does not work for our affine threshold policies, since they require both servers to push when $Q_1 = 0, Q_2 > \beta_1$ or when $Q_2 = 0, Q_1 > \beta_2$.

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 departure processes, $D_{i,j}(t)$ converge under diffusion scaling to Brownian motions. We calculate the asymptotic variance parameters of these, including the covariances between the departure streams.

We now define diffusion scalings for $n = 1, 2, \ldots$. First denote

$$\bar{S}(t) = \lim_{n\to\infty} \bar{S}^n(t) = \lim_{n\to\infty} \frac{S(nt)}{n}.$$

By Assumption (A1), the limit exists a.s. u.o.c. and $\bar{S}_{i,1}(t) = \lambda_i t$ and $\bar{S}_{i,2}(t) = \mu_i t$. Further, use the fluid limit processes of Section 4, Corollary 1. The diffusion scalings are:

$$\begin{aligned} \hat{S}^n_{i,j}(t) &= \frac{S_{i,j}(nt) - \bar{S}_{i,j}(nt)}{\sqrt{n}}, \qquad \hat{T}^n_{i,j}(t) = \frac{T_{i,j}(nt) - \bar{T}_{i,j}(nt)}{\sqrt{n}}, \\ \hat{D}^n_{i,j}(t) &= \frac{D_{i,j}(nt) - \bar{D}_{i,j}(nt)}{\sqrt{n}}, \qquad \hat{Q}^n_{i,j}(t) = \frac{Q_{i,j}(nt)}{\sqrt{n}}. \end{aligned} \tag{9}$$

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 an affine 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,*

$$\hat{D}^n_{i,1}(t) - \hat{D}^n_{i,2}(t) = \hat{Q}^n_i(t) \Rightarrow 0, \tag{10}$$

$$\hat{T}^n_{1,1}(t) + \hat{T}^n_{2,2}(t) = \hat{T}^n_{2,1}(t) + \hat{T}^n_{1,2}(t) = 0, \tag{11}$$

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

$$\mathrm{Var}(\hat{D}_{1,2}(1)) = \frac{\lambda_1\mu_1}{(\lambda_1\lambda_2 - \mu_1\mu_2)^3} \times [\lambda_1\lambda_2\mu_1\mu_2(c^2_{2,1} + c^2_{2,2})(\lambda_1 - \mu_1) + (\lambda^2_1\lambda^2_2 c^2_{1,2} + \mu^2_1\mu^2_2 c^2_{1,1})(\lambda_2 - \mu_2)], \tag{12}$$

$$\begin{aligned} \mathrm{Cov}(\hat{D}_{1,2}(1), \hat{D}_{2,2}(1)) \;=\; & -\frac{\lambda_1\lambda_2\mu_1\mu_2}{(\lambda_1\lambda_2 - \mu_1\mu_2)^3} \times [(\lambda_1\lambda_2 c^2_{2,2} + \mu_1\mu_2 c^2_{2,1})(\lambda_1 - \mu_1) \\ & + (\lambda_1\lambda_2 c^2_{1,2} + \mu_1\mu_2 c^2_{1,1})(\lambda_2 - \mu_2)], \end{aligned} \tag{13}$$

with a symmetric expression for $\mathrm{Var}(\hat{D}_{2,2}(1))$. *Similar expressions for variances and covariances of* $\hat{T}_{1,2}(\cdot)$, $\hat{T}_{2,2}(\cdot)$ *may be read off from* (19).

Proof. The equalities (10) and (11) follow immediately from (2) and (1). The convergence to 0 in (10) follows from Theorem 2, since $Q(t)$ has a limiting stationary distribution, therefore $Q(nt)$ 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, $\bar{T}^n_{i,j}(t) \to \bar{T}_{i,j}(t) = \theta_{i,j}t$ and $\bar{D}^n_{i,j}(t) \to \bar{D}_{i,j}(t) = \nu_i t$ u.o.c as $n \to \infty$.

The rest of the proof and the calculations are straightforward:

$$\begin{aligned}\hat{D}^n_{i,2}(t) &= \frac{D_{i,2}(nt) - \bar{D}_{i,2}(nt)}{\sqrt{n}} \\ &= \frac{S_{i,2}(n\bar{T}^n_{i,2}(t)) - \bar{S}_{i,2}(n\bar{T}^n_{i,2}(t))}{\sqrt{n}} + \frac{\bar{S}_{i,2}(n\bar{T}^n_{i,2}(t))}{\sqrt{n}} - \frac{\bar{D}_{i,2}(nt)}{\sqrt{n}} \\ &= \hat{S}^n_{i,2}(\bar{T}^n_{i,2}(t)) + \mu_i \frac{T_{i,2}(nt) - \bar{T}_{i,2}(nt)}{\sqrt{n}} + \mu_i \frac{\bar{T}_{i,2}(nt)}{\sqrt{n}} - \frac{\bar{D}_{i,2}(nt)}{\sqrt{n}} \\ &= \hat{S}^n_{i,2}(\bar{T}^n_{i,2}(t)) + \mu_i \hat{T}^n_{i,2}(t) + \theta_{i,2}\mu_i\sqrt{n}t - \theta_{i,2}\mu_i\sqrt{n}t \\ &= \hat{S}^n_{i,2}(\bar{T}^n_{i,2}(t)) + \mu_i \hat{T}^n_{i,2}(t),\end{aligned}$$

where all we did is to add and subtract quantities, use the definition (9), and use $\bar{S}_{i,2}(t) = \mu_i t$ (by Assumption (A1), and $\bar{T}_{i,j}(t) = \theta_{i,j}t$, $\bar{D}_{i,j}(t) = \nu_i t = \mu_i\theta_{i,2}t$ (from Corollary 1).

Define $\hat{P}^n_{i,j}(t) = \hat{S}^n_{i,j}(\bar{T}^n_{i,j}(t))$, then summarizing the above and also using similar calculations for (15)–(17) we obtain:

$$\hat{D}^n_{i,2}(t) = \hat{P}^n_{i,2}(t) + \mu_i \hat{T}^n_{i,2}(t), \tag{14}$$

$$\hat{D}^n_{i,1}(t) = \hat{P}^n_{i,1}(t) + \lambda_i \hat{T}^n_{i,1}(t), \tag{15}$$

$$\hat{Q}^n_i(t) = \hat{D}^n_{i,1}(t) - \hat{D}^n_{i,2}(t) \tag{16}$$

$$\hat{T}^n_{1,1}(t) = -\hat{T}^n_{2,2}(t), \qquad \hat{T}^n_{2,1}(t) = -\hat{T}^n_{1,2}(t). \tag{17}$$

Now using (14)–(17):

$$\begin{bmatrix}\hat{D}^n_{1,2}(t)\\ \hat{D}^n_{2,2}(t)\\ \hat{T}^n_{1,2}(t)\\ \hat{T}^n_{2,2}(t)\end{bmatrix} = \mathbf{A}\begin{bmatrix}\hat{P}^n_{1,1}(t)\\ \hat{P}^n_{1,2}(t)\\ \hat{P}^n_{2,1}(t)\\ \hat{P}^n_{2,2}(t)\end{bmatrix} + \mathbf{B}\begin{bmatrix}\hat{Q}^n_1(t)\\ \hat{Q}^n_2(t)\end{bmatrix}, \tag{18}$$

where

$$\mathbf{A} = \frac{1}{\lambda_1\lambda_2 - \mu_1\mu_2}\begin{bmatrix}-\mu_1\mu_2 & \lambda_1\lambda_2 & \lambda_1\mu_1 & -\lambda_1\mu_1\\ \lambda_2\mu_2 & -\lambda_2\mu_2 & -\mu_1\mu_2 & \lambda_1\lambda_2\\ -\mu_2 & \mu_2 & \lambda_1 & -\lambda_1\\ \lambda_2 & -\lambda_2 & -\mu_1 & \mu_1\end{bmatrix},$$

and

$$\mathbf{B} = \frac{1}{\lambda_1\lambda_2 - \mu_1\mu_2}\begin{bmatrix}\mu_1\mu_2 & -\lambda_1\mu_1\\ -\lambda_2\mu_2 & \mu_1\mu_2\\ \mu_2 & -\lambda_1\\ -\lambda_2 & \mu_1\end{bmatrix}.$$

By the functional central limit theorem for renewal processes and the continuous mapping theorem (c.f. [28]) we have $\hat{P}^n(t) \Rightarrow \hat{P}(t)$ where $\hat{P}(t)$ is a 4 dimensional driftless Brownian motion with a diagonal covariance matrix $\boldsymbol{\Lambda}$, having entries $\mathrm{Var}(\hat{P}_{i,1}(1)) = \lambda_i\theta_{i,1}c^2_{i,1}$ and $\mathrm{Var}(\hat{P}_{i,2}(1)) = \mu_i\theta_{i,2}c^2_{i,2}$.

Incorporating the above with the weak convergence of $\hat{Q}^n$ to 0, we have that $(\hat{D}^n_{1,2}(t), \hat{D}^n_{2,2}(t), \hat{T}^n_{1,2}(t), \hat{T}^n_{2,2}(t))$ converges to a driftless Brownian motion process with covariance matrix:

$$\boldsymbol{\Gamma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}'. \quad \blacksquare \tag{19}$$

The above theorem gives us the asymptotic variance rate of departures:

$$\bar{V}_i = \lim_{t\to\infty}\frac{\mathrm{Var}(D_{i,2}(t))}{t} = \mathrm{Var}(\hat{D}_{i,2}(1)).$$

The following three subsections highlight some surprising facts about the diffusion scale behavior of the push pull network and the asymptotic variance rate of departures.

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 $\hat{Q}^n(t) \Rightarrow 0$ 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.

We reach the surprising conclusion that the diffusion scale departure processes $\hat{D}(t)$ do not depend on the policy, so long as it is fully utilizing and stabilizing. In Section 7 we encounter the same phenomena for general infinite supply re-entrant lines.

6.2. Departures variability increases with balancing

When the system approached complete balance ($\lambda_i \approx \mu_i, i = 1, 2$) it was shown in [1] that for exponential processing times the network becomes increasingly congested in both the inherently stable and inherently unstable case. This seems to be the case also for general processing times: We confirm this in Section 6.5 for exponential pull and general push activities in the inherently stable case, and it is also observed in simulations for general processing times.

It turns out that this congestion in the queues is accompanied by increasing variance of the diffusion scaled departure processes. For illustration let us evaluate (12) for the symmetric case: $c_i^2 = c^2$, $\mu_i = 1$, $\lambda_i = \lambda$. In this case the asymptotic variance rate of departures is:

$$\bar{V}_1 = \bar{V}_2 = \frac{\lambda}{\lambda + 1}\left(\frac{\lambda^2 + 1}{\lambda^2 - 1}\right)^2.$$

The departure rate in this case is: $\nu_1 = \nu_2 = \lambda/(\lambda + 1)$. So the limiting index of dispersion of counts ($\bar{V}_{i,j}/\nu_i$) grows to infinity as $\lambda \to \mu = 1$. Thus the departures of the push pull network become more variable in the sense of limiting index of dispersion of counts as the system becomes more congested. This behavior is unusual. It is different from the behavior of a stable GI/G/1 queue in which the limiting index of dispersion of counts is constant for any congestion level. It is also different from the BRAVO effect observed in a finite buffer single server queue [19], where it was seen that balancing reduced the asymptotic variance of departures.

6.3. Variance of combined departures

Using (18) or using (12), (13) one can obtain after simple manipulation the asymptotic variance rate of the combined departure process, $D_{1,2}(t) + D_{2,2}(t)$. We have:

$$\begin{aligned}\mathrm{Var}(\hat{D}_{1,2}(1) + \hat{D}_{2,2}(1)) = {} & \nu_1{}^3\left[\sigma_{1,1}^2\left(\frac{\mu_2}{\mu_1}\right)^2\left(\frac{\lambda_2 - \mu_1}{\lambda_2 - \mu_2}\right)^2 + \sigma_{1,2}^2\left(\frac{\lambda_2}{\lambda_1}\right)^2\left(\frac{\lambda_1 - \mu_2}{\lambda_2 - \mu_2}\right)^2\right] \\ & + \nu_2{}^3\left[\sigma_{2,1}^2\left(\frac{\mu_1}{\mu_2}\right)^2\left(\frac{\lambda_1 - \mu_2}{\lambda_1 - \mu_1}\right)^2 + \sigma_{2,2}^2\left(\frac{\lambda_1}{\lambda_2}\right)^2\left(\frac{\lambda_2 - \mu_1}{\lambda_1 - \mu_1}\right)^2\right]\end{aligned}$$

where $\sigma_{i,j}$ is the variance of the (i, j) processing time.

For the symmetric case, with $\lambda_1 = \lambda_2 = \lambda$, $\mu_1 = \mu_2 = \mu$, and with $\nu_1 = \nu_2 = \nu$, the result is quite surprising:

$$\mathrm{Var}(\hat{D}_{1,2}(1) + \hat{D}_{2,2}(1)) = \nu^3(\sigma_{1,1}^2 + \sigma_{1,2}^2 + \sigma_{2,1}^2 + \sigma_{2,2}^2)$$

which is the sum of the variances of two renewal processes: The process of producing jobs of type 1, with the sum of the processing times of the two activities of type 1, and the process of producing jobs of type 2, with the sum of the processing times of the two activities of type 2.

6.4. Negative covariance of departures

It is evident from (13) that $\mathrm{Cov}(\hat{D}_1(t), \hat{D}_2(t)) < 0$. Also, when all activity processing times have the same squared coefficient of variation c^2, then both the variance and the covariance in (12) and (13) are linear in c^2. In Fig. 4 we illustrate the negative correlation between the departure processes of our network. We plot as a function of λ:

$$\rho_\lambda = \frac{\mathrm{Cov}(\hat{D}_1(1), \hat{D}_2(1))}{\sqrt{\mathrm{Var}(\hat{D}_1(1))\mathrm{Var}(\hat{D}_2(1))}}, \tag{20}$$

again for symmetric push pull networks with parameters $c_{i,j}^2 = c^2$, $\mu_i = 1$, $\lambda_i = \lambda$.

Our analysis applies to all $\lambda \neq 1$. When $\lambda = 1$ we have a completely balanced network 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 [29].

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 $D_{i,2}$ depend mostly on the long activity (push if $\mu \gg \lambda$, pull if $\lambda \gg \mu$), but this means that one server will essentially produce the departure stream of jobs of type 1 while the other server will produce the departure stream of jobs of type 2. Hence the two departure processes will be almost independent.

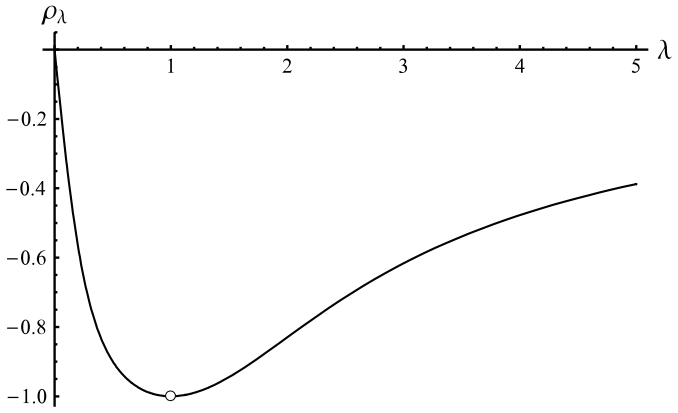

Fig. 4. The correlation between departures of a symmetric push pull network.

6.5. Alternative derivation of asymptotic variance rate

We now present another derivation of (12) under some more restrictive assumptions. It illustrates an alternative approach and provides some additional insights. We assume that the network is inherently stable (Case 1), and that the push activity processing durations are exponentially distributed. Our derivation uses a renewal-reward approach. In this case the behavior of the network is like two randomly alternating M/G/1 single server queues and at every time that the system empties there is a regeneration epoch. We get first[1]:

Proposition 2. *Consider the inherently stable push pull network with preemptive pull priority. Assume exponential processing times for the push activities and general processing times for the pull activities with Laplace-Stieltjes transforms $G_1^*(s)$, $G_2^*(s)$. Further let $\rho_1 = \frac{\lambda_1}{\mu_1}$, $\rho_2 = \frac{\lambda_2}{\mu_2}$.*

(i)

$$\mathbb{E}\left[z_1^{Q_1}, z_2^{Q_2}\right] = \left(1 + \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}\right)^{-1} \times \left(\frac{G_1^*(\lambda_1(1-z_1))(1-z_1)}{G_1^*(\lambda_1(1-z_1)) - z_1} + \frac{G_2^*(\lambda_2(1-z_2))(1-z_2)}{G_2^*(\lambda_2(1-z_2)) - z_2} - 1\right)$$

(ii)

$$\mathbb{E}\left[Q_i\right] = \nu_i \frac{1}{\mu_i} \frac{2 + \rho_i(\dot{c}_{i,2}^2 - 1)}{2(1-\rho_i)}.$$

Proof. After a finite duration, the system enters a regime in which at least one of the two queues is empty. Denote by P_0 the steady state probability that the network is empty and by P_i the probability of having a positive number of jobs in queue i. Using renewal reward considerations: $P_i = \lambda_i \mathbb{E}\left[B_i\right]/(1 + \lambda_1 \mathbb{E}\left[B_1\right] + \lambda_2 \mathbb{E}\left[B_2\right])$, where B_i is a busy period duration of an $M/G/1$ queue with arrival rate λ_i and service mean μ_i^{-1}. Now the condition on the queue that is being served to obtain,

$$\mathbb{E}\left[z_1^{Q_1}, z_2^{Q_2}\right] = P_0 + P_1 \mathbb{E}\left[z^{\tilde{Q}_1} | \tilde{Q}_1 > 0\right] + P_2 \mathbb{E}\left[z^{\tilde{Q}_2} | \tilde{Q}_2 > 0\right],$$

where $\tilde{Q}_i$ is distributed as the steady state number of jobs in an $M/G/1$ system with arrival rate λ_i and service mean μ_i^{-1}. Application of the P–K formula yields (i). To obtain (ii) either directly use (i) or observe that it is an application of Littles law and the P–K formula for the mean sojourn time. ■

The calculation of the asymptotic variance rate of departures is based on an embedding of the departure process $D_{i,2}(t)$ in renewal-reward process where the rewards are counts of job type i departures.

We consider type 1 departures, type 2 is analogous. Let $\{(X_n, Y_n), n = 1, 2, \ldots\}$ denote an i.i.d sequence where X_n denotes the times between returns to an empty system (both servers are working on an exponential push) and Y_n denotes the number of jobs processed on activity $(1, 2)$ between successive empty times. Define a renewal reward process $C(t)$:

$$C(t) = \sum_{i=1}^{N(t)-1} Y_i \quad \text{where } N(t) = \sup\left\{n : \sum_{k=1}^{n} X_k \le t\right\}.$$

[1] This result is of independent interest and is not needed for the derivation of the asymptotic variance rate.

Denote by (X, Y) a generic random variable from the sequence $\{(X_n, Y_n)\}$. If $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$ then the asymptotic variance rate of $C(t)$ is computable (c.f. [30]):

$$\lim_{t\to\infty} \frac{\mathrm{Var}(C(t))}{t} = \frac{\mathbb{E}[X^2]\mathbb{E}[Y]^2}{\mathbb{E}[X]^3} - 2\frac{\mathbb{E}[XY]\mathbb{E}[Y]}{\mathbb{E}[X]^2} + \frac{\mathbb{E}[Y^2]}{\mathbb{E}[X]}. \tag{21}$$

Now $C(t)$ counts the number of departures of type 1 during a time interval $[0, \tau]$ where $\tau \le t$ is the last regeneration time. As a result,

$$\lim_{t\to\infty} \frac{\mathrm{Var}(C(t))}{t} = \bar{V}_1.$$

Thus to obtain (12) (for this special M/G pull-priority case) the problem reduces to computation of moments of (X, Y): Consider two M/G/1 queues ($i = 1, 2$) with arrival rates λ_i and service means μ_i^{-1}. Let B_i, I_i and N_i denote random variables that are distributed as the busy period, idle period and number of customer served during a busy period respectively. Also denote by χ an indicator random variable for the event of having the first push operation to complete in a cycle to be of type 1. Then we have the following equalities in distribution:

$$\begin{aligned} &X = \tilde{I} + \chi B_1 + (1-\chi)B_2, \qquad X^2 = \tilde{I}^2 + \chi(B_1^2 + 2\tilde{I}B_1) + (1-\chi)(B_2^2 + 2\tilde{I}B_2), \\ &Y = \chi N_1, \qquad Y^2 = \chi^2 N_1^2 \\ &XY = \chi(N_1\tilde{I} + N_1 B_1). \end{aligned} \tag{22}$$

Evaluation of expectation of the above quantities is based on the first two moments of M/G/1 busy periods and the number of customer served during a busy period as well as the covariance of the busy period duration and the number of customers served. All of these quantities are well known (c.f. [31]) and when plugged into (21) we obtain (12) with $c_{i,1}^2 = 1$.

7. Re-entrant lines with infinite supply of work

In this section we consider a re-entrant line with infinite supply of work, and perform the same diffusion scale analysis as in Section 6. A re-entrant line [32] is a multi-class queueing network with a single deterministic job route. Servers are $k = 1, \ldots, K$ and classes are the ordered processing steps $i = 1, \ldots, I$, partitioned into $C_1, \ldots, C_K$, with $i \in C_k$ if step i is at server k. We let $1 \in C_1$ and assume that there is an infinite virtual queue of class 1 jobs. We assume independent sequences of i.i.d. processing times, with expected processing times m_i, processing rates $\mu_i = 1/m_i$, processing time variances σ_i^2. We also assume that the processing time distribution of step 1 is spread out with infinite support, as in (A2b). If server 1, with the infinite supply of work, is working all the time, and if the network is stable, then the departure rate is:

$$\nu = \left(\sum_{i\in C_1} m_i\right)^{-1}.$$

We will assume that server 1 is the single bottleneck in the network, by assuming

$$\rho_k = \nu \sum_{i\in C_k} m_i < 1, \quad k \neq 1.$$

Guo and Zhang [6] considered this infinite supply re-entrant line under, a policy in which class 1 is only served when there are no other jobs at server 1, and in addition the service to the buffers of all other classes is last buffer first served (LBFS) or first buffer first served (FBFS). They have shown that the network process is positive Harris recurrent.

As in the previous sections of this paper we use $S_i(\cdot)$, $Q_i(\cdot)$, $T_i(\cdot)$, $D_i(\cdot)$ to denote the service completion counting, queue length, time allocation and departure processes associated with class i. We let $Q_i^+(t) = \sum_{j=i+1}^{I} Q_j(t)$ denote the total number of jobs in the network which have completed step i (downstream of i). As before we have:

$$D_i(t) = S_i(T_i(t)), \qquad Q_i(t) = D_{i-1}(t) - D_i(t), \qquad D_i(t) = Q_i^+(t) + D_I(t).$$

For any non-idling policy, server 1 will work all the time and we have:

$$\sum_{i\in C_1} T_i(t) = t.$$

Under non-idling pull priority LBFS or FBFS, or any other non-idling policy for which the network process is positive Harris recurrent we can perform a fluid scale analysis of the network as in Section 4. Let $\bar{S}(\cdot)$, $\bar{T}(\cdot)$, $\bar{D}(\cdot)$ denote limiting fluid scaled processes. Then:

$$\bar{S}_i(t) = \mu_i t, \qquad \bar{T}_i(t) = \theta_i t, \qquad \theta_i = \nu/\mu_i, \qquad \bar{D}(t) = \bar{S}(\bar{T}(t)) = \nu t.$$

As in Section 6 we let $\hat{S}^n(t)$, $\hat{T}^n(t)$, $\hat{D}^n(t)$, $\hat{Q}^n(t)$ be the diffusion scaled processes defined analogously to (9), from which we define the $3I - 1$ dimensional diffusion scaled process $\hat{X}^n(t) = (\hat{D}^n(t), \hat{T}^n(t), \hat{Q}^n(t))$. We then have:

Theorem 4. *For the re-entrant line with infinite supply of work, under non-idling pull priority that has a positive Harris recurrent network process, as $n \to \infty$ the process $\hat{X}^n(t) \Rightarrow \hat{X}(t)$ where $\hat{X}(t) = (\hat{D}(t), \hat{T}(t), \hat{Q}(t))$ is a $3I-1$ dimensional Brownian motion.*

Furthermore:

$$\hat{Q}_i(t) = 0, \qquad \hat{D}_i(t) = \hat{D}_l(t), \qquad \sum_{i \in C_1} \hat{T}_i(t) = 0, \tag{23}$$

and the variance of the departure process from the line is:

$$\mathrm{Var}(\hat{D}_l(1)) = \frac{\sum_{i \in C_1} \sigma_i^2}{\left(\sum_{i \in C_1} m_i\right)^3}. \tag{24}$$

Variances of various $\hat{T}_i$ and various covariances can be read from (30).

Proof. Since the network process is positive Harris recurrent we have that $\hat{Q}_i^n(t) \Rightarrow 0$. From the dynamics of the network the diffusion scalings satisfy:

$$\hat{D}_i^n(t) = \hat{Q}_i^{+^n}(t) + \hat{D}_l^n(t), \quad i = 1, \ldots, I, \tag{25}$$

and therefore $\hat{D}_i^n(t) - \hat{D}_l^n(t) \Rightarrow 0$. From the non-idling we have, for the diffusion scaling:

$$\sum_{i \in C_1} \hat{T}_i^n(t) = 0 \tag{26}$$

and therefore: $\sum_{i \in C_1} \hat{T}_i(t) = 0$.

We now calculate the variance of the limiting diffusion scaled departure process $\hat{D}_l(t)$. Using the exact same calculations as in the proof of Theorem 3 we have:

$$\hat{D}_i^n(t) = \hat{S}_i^n(\bar{T}_i^n(t)) + \mu_i \hat{T}_i^n(t) \quad i = 1, \ldots, I. \tag{27}$$

Summing (27) over the classes $i \in C_1$, and using (26), we obtain:

$$\sum_{i \in C_1} \frac{\hat{D}_i^n(t)}{\mu_i} - \sum_{i \in C_1} \frac{\hat{P}_i^n(t)}{\mu_i} = 0, \tag{28}$$

where as in the proof of Theorem 3, $\hat{P}_i^n(t) = \hat{S}_i^n(\bar{T}_i^n(t))$. Substituting the Eqs. (25) in (28) and solving for $\hat{D}_l^n(t)$ we obtain:

$$\hat{D}_l^n(t) = \nu \sum_{i \in C_1} m_i \hat{P}_i^n(t) + \sum_{i=1}^{I} b_i \hat{Q}_i^n(t),$$

where b_i are some constants (expressions of m_i).

Now as in Theorem 3, we have $\hat{P}_i^n(t), i = 1, \ldots, n$ converge to independent drift-less Brownian motions with $\mathrm{Var}(\hat{P}_i(1)) = \nu\sigma_i^2/m_i^2$. At the same time $\hat{Q}_i^n(t) \Rightarrow 0$. Hence $\hat{D}_l^n(t)$ converges weakly to a Brownian motion, The expression for the variance of $\hat{D}_l(1)$, (24), follows.

The diffusion scaled time allocations can be expressed similarly as:

$$\hat{T}_i^n(t) = m_i \nu \sum_{j \in C_1} m_j \hat{P}_j^n(t) - m_i \hat{P}_i^n(t) + \sum_{i=1}^{I} c_i \hat{Q}_i^n(t), \quad i = 1, \ldots, I, \tag{29}$$

where c_i are some constants (expressions of m_i). Let $\zeta_1(t), \ldots, \zeta_I(t)$ be independent standard Brownian motions. Then the joint distribution of $\hat{D}_l(t)$, $\hat{T}_i(t)$, $i = 1, \ldots, I$ can be obtained from the representation:

$$\begin{aligned} \hat{D}_l(t) &= (\nu)^{3/2} \sum_{i \in C_1} \sigma_i \zeta_i(t) \\ \hat{T}_i(t) &= m_i\, (\nu)^{3/2} \sum_{j \in C_1} \sigma_j \zeta_j(t) - (\nu)^{1/2} \sigma_i \zeta_i(t), \quad i = 1, \ldots, I. \end{aligned} \tag{30}$$

■

Note that if $C_1 = \{1, \ldots, I\}$ (the system is re-entrant through a single server) then for pull priority policy each job undergoes processing from step 1 to step I before the next job is introduced. Hence the departure process is actually a renewal process with inter-departure times having mean $\sum_{i \in C_1} m_i$ and variance $\sum_{i \in C_1} \sigma_i^2$. In this case, the asymptotic variance rate (24) immediately follows. Theorem 4 shows, surprisingly, that when the departure process is not renewal (as is the case when there is more then one server) then the asymptotic variance rate of the departures still depends only on the first server (which is the bottleneck) and is equal to that of the renewal departures case.

Acknowledgements

We would like to thank Serguei Foss for useful discussions on stability of Markov chains, and fluid and diffusion approximations of queueing networks. The author's research was supported in part by Israel Science Foundation Grant 249/02 and 454/05 and by European Network of Excellence Euro-NGI.

References

[1] A. Kopzon, Y. Nazarathy, G. Weiss, A push pull system with infinite supply of work, Preprint.
[2] A. Kopzon, G. Weiss, A push pull queueing system, Operations Research Letters 30 (6) (2002) 351–359.
[3] I. Adan, G. Weiss, A two node Jackson network with infinite supply of work, Probability in the Engineering and Informational Sciences 19 (2) (2005) 191–212.
[4] I. Adan, G. Weiss, Analysis of a simple Markovian re-entrant line with infinite supply of work under the LBFS policy, Queueing Systems 54 (3) (2006) 169–183.
[5] Y. Guo, H. Zhang, On the stability of a simple re-entrant line with infinite supply, OR Transactions 10 (2) (2006) 75–85.
[6] Y. Guo, H. Zhang, Positive Harris recurrence of re-entrant lines with infinite supply, Preprint.
[7] Y. Nazarathy, G. Weiss, Near optimal control of queueing networks over a finite time horizon, Annals of Operations Research (in press).
[8] G. Weiss, Jackson networks with unlimited supply of work, Journal of Applied Probability 42 (3) (2005) 879–882.
[9] Y. Nazarathy, On control of queueing networks and the asymptotic variance rate of outputs, Ph.D. Thesis, The University of Haifa, 2008.
[10] J.G. Dai, W. Lin, Maximum pressure policies in stochastic processing networks, Operations Research 53 (2) (2005) 197–218.
[11] 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) (1995) 49–77.
[12] Y. Nazarathy, G. Weiss, Positive Harris recurrence and diffusion scale analysis of a push pull queueing network, in: Proceedings of Valuetools, 2008.
[13] H. Chen, D. Yao, Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization, Springer, 2001.
[14] S.P. Meyn, Control Techniques for Complex Networks, Cambridge University Press, 2008.
[15] P. Kumar, T. Seidman, Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems, IEEE Transactions on Automatic Control AC-35 (3) (1990) 289–298.
[16] A. Rybko, A. Stolyar, Ergodicity of stochastic processes describing the operation of open queueing networks, Problems of Information Transmission 28 (3) (1992) 3–26.
[17] J.M. Harrison, Brownian models of queueing networks with heterogeneous customer populations, in: W. Fleming, P.-L. Lions (Eds.), Stochastic Differential Systems, Stochastic Control Theory and Applications, 1988, pp. 147–186.
[18] J.M. Harrison, R. Williams, Brownian models of feedforward queueing networks: Quasireversibility and product form solutions, Ann. Appl. Probab. 2 (2) (1992) 263–293.
[19] Y. Nazarathy, G. Weiss, The asymptotic variance rate of finite capacity birth-death queues, Queueing Systems 59 (2) (2008) 135–156.
[20] J.G. Dai, G. Weiss, Stability and instability of fluid models for re-entrant lines, Mathematics of Operations Research 21 (1) (1996) 115–134.
[21] M. Bramson, Stability of two families of queueing networks and a discussion of fluid limits, Queueing Systems 28 (1–3) (1998) 7–31.
[22] M. Bramson, Stability of Queueing Networks, Springer, 2008.
[23] 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) (1984) 353–388.
[24] S.P. Meyn, R. Tweedie, Markov Chains and Stochastic Stability, Springer-Verlag, 1993.
[25] S.P. Meyn, R.L. Tweedie, Stability of Markovian processes II: Continuous-time processes and sampled chains, Advances in Applied Probability 25 (3) (1993) 487–517.
[26] S.P. Meyn, R.L. Tweedie, Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes, Advances in Applied Probability 25 (3) (1993) 518–548.
[27] S.P. Meyn, D. Down, Stability of generalized Jackson networks, The Annals of Applied Probability 4 (1) (1994) 124–148.
[28] P.W. Glynn, in: D.P. Heyman, M.J. Sobel (Eds.), Diffusion Approximations, in: Handbooks in Operations Research, vol. 2, North-Holland, Amsterdam, 1990, pp. 145–198.
[29] J.G. Dai, W. Lin, Asymptotic optimality of maximum pressure policies in stochastic processing networks, Preprint.
[30] M. Brown, H. Solomon, A second order approximation for the variance of a renewal reward process, Stochastic Processes and their Applications 3 (1974) 301–314.
[31] N. Prabhu, Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Communication, Springer, 1998.
[32] P. Kumar, Re-entrant lines, Queueing Systems 13 (1) (1993) 87–110.

Yoni Nazarathy

Gideon Weiss