

# Asymptotic Optimality of Maximum Pressure Policies in Stochastic Processing Networks <sup>1</sup>

J. G. Dai

School of Industrial and Systems Engineering  
Georgia Institute of Technology  
Atlanta, Georgia 30332, USA  
dai@gatech.edu

Wuqin Lin

Kellogg School of Management  
Northwestern University  
Evanston, Illinois 60208  
wuqin-lin@kellogg.northwestern.edu

July 21, 2006

## Abstract

We consider a class of stochastic processing networks. Assume that the networks satisfy a complete resource pooling condition. We prove that *each* maximum pressure policy asymptotically minimizes the workload process in a stochastic processing network in heavy traffic. When the objective is to minimize a linear holding cost, for any  $\varepsilon > 0$ , we identify a set of maximum pressure policies and prove that they are asymptotically  $\varepsilon$ -optimal. A key to the optimality proof is to prove a state space collapse result and a heavy traffic limit theorem for the network processes under a maximum pressure policy. We extend a framework of [Bramson \(1998\)](#) and [Williams \(1998b\)](#) from the multiclass queueing network setting to the stochastic processing network setting to prove the state space collapse result and the heavy traffic limit theorem. The extension can be adapted to other studies of stochastic processing networks.

*AMS 2000 subject classification.* Primary 90B15, 60K25 secondary 90B18, 90B22, 68M10, 60J60.

*Key words and phrases.* Stochastic processing networks, maximum pressure policies, back-pressure policies, heavy traffic, Brownian models, diffusion limits, state space collapse, asymptotic optimality.

*Running title.* Maximum Pressure Policies

---

<sup>1</sup>Research supported in part by National Science Foundation grant DMI-0300599, by an IBM Faculty Award

# 1 Introduction

This paper is a continuation of [Dai and Lin \(2005\)](#), in which maximum pressure policies are shown to be throughput optimal for a class of stochastic processing networks. Throughput optimality is an important, first-order objective for many networks, but it ignores some key secondary performance measures like queueing delays experienced by jobs in these networks. In this paper, we show that maximum pressure policies enjoy additional optimality properties; they are asymptotically optimal in minimizing a certain workload or holding cost of a stochastic processing network.

Stochastic processing networks have been introduced in a series of three papers by [Harrison \(2000, 2002, 2003\)](#). In [Dai and Lin \(2005\)](#) and this paper, we consider a special class of Harrison’s model. This class of stochastic processing networks is much more general than multiclass queueing networks that have been a subject of intensive study in the last 20 years; see, for example, [Harrison \(1988\)](#), [Williams \(1996\)](#), and [Chen and Yao \(2001\)](#). The added features in a stochastic processing network allow one to model skills-based routing in call centers ([Gans, Koole and Mandelbaum, 2003](#)), operator-machine interactions in semiconductor wafer fabrication facilities ([Kumar, 1993](#)), and combined input- and output-queued data switches ([Chuang, Goel, McKeown and Prabhakar, 1999](#)) in data networks.

For this general class of stochastic processing networks, [Dai and Lin \(2005\)](#) propose a family of operational policies called maximum pressure policies and prove that they are throughput optimal. For a given set of parameters, the maximum pressure policy associated with the parameter set is specified in Section 3 of [Dai and Lin \(2005\)](#) and will be specified again in Definition 1 of this paper. In this paper, for a stochastic processing network that satisfies a complete resource pooling condition and a heavy traffic condition, we first show in Theorem 2 that a certain workload process is asymptotically minimized under *any* maximum pressure policy. We then identify a set of maximum pressure policies and show in Theorem 3 that they are *asymptotically  $\varepsilon$ -optimal* when the objective is to minimize a linear holding cost, where the holding cost rate is a linear function of buffer contents. By  $\varepsilon$ -optimality we mean, for any given  $\varepsilon > 0$ , one can find a maximum pressure policy that may depend on  $\varepsilon$  such that for each time  $t$  the holding cost rate under the maximum pressure policy is stochastically bounded by  $(1 + \varepsilon)$  multiple of the optimal holding cost rate.

In Theorem 2, except for some nonnegativity requirements, the parameter set that is used to define a maximum pressure policy can be chosen arbitrarily. In particular, the parameter set can be chosen to be independent of network data like arrival or processing rates. In such a case, the corresponding maximum pressure policy does not depend on the arrival rates. This feature is attractive in some applications when network data like arrival rates are sometimes difficult or impossible to be estimated accurately. The maximum pressure policies in Theorem 3 do depend on the arrival rates of the network.

Our asymptotic region is when the stochastic processing network is in *heavy traffic*; at least one server has to be 100% busy in order to handle all the input. A key assumption on our network is that a *complete resource pooling* condition is satisfied. Roughly speaking, the complete resource pooling condition requires enough overlap in the processing capacities of bottleneck servers that these servers form a single, pooled resource or “super server.” As

will be discussed fully in Section 3, the complete resource pooling condition is articulated by the dual problem of a linear program (LP) called the *static planning problem*. For a network satisfying the complete resource pooling condition, the corresponding dual LP has a unique optimal solution and the workload process is defined by this unique optimal solution. [Ata and Kumar \(2005\)](#) develop a discrete-review policy and prove its asymptotic optimality in minimizing the linear holding cost for a class of *unitary* stochastic processing networks that satisfy the complete resource pooling condition and the *balanced* heavy traffic condition. The latter condition requires every server in the network be heavily loaded. This balanced load requirement, combined with the complete resource pooling assumption, rules out some well known networks such as multiclass queueing networks. Our definition of heavy traffic condition is less restrictive than those in [Ata and Kumar \(2005\)](#); our stochastic processing networks include those multiclass queueing networks that have a unique bottleneck server.

The major part of our optimality proof of the maximum pressure policies is a heavy traffic limit theorem. The theorem asserts that when the network is operated under a maximum pressure policy, (a) the one-dimensional workload process converges to a reflecting Brownian motion in diffusion limit, and (b) the multidimensional buffer content process is a constant multiple of the workload process in diffusion limit. The latter result is a form of *state space collapse* for network processes, and its proof is the key to the proof of the limit theorem. We choose to extend a framework of [Bramson \(1998\)](#) and [Williams \(1998b\)](#), from the multiclass queueing network setting to the stochastic processing network setting, to prove the heavy traffic limit theorem. We first show that all solutions to a critically loaded fluid model operating under a maximum pressure policy exhibit some form of state space collapse. Then we translate the state space collapse to the diffusion scaling following [Bramson \(1998\)](#), proving the state space collapse in diffusion limit. Once we have the state space collapse result, we invoke a theorem of [Williams \(1998b\)](#) for perturbed Skorohod problems to establish the heavy traffic limit theorem for our stochastic processing network operating under the maximum pressure policy.

[Stolyar \(2004\)](#) proves that MaxWeight policies asymptotically minimize the workload processes in heavy traffic for a generalized switch model that belongs to one-pass systems in which each job leaves the system after being processed at one processing step. Our Theorem 2 greatly generalizes [Stolyar \(2004\)](#) from one-pass systems to stochastic processing networks. Except for [Ata and Kumar \(2005\)](#), that was discussed earlier in this introduction, most other works that are closely related to our work have focused on parallel server systems. These systems belong to a special class of one-pass systems. All these works assume heavy traffic and complete resource pooling conditions. For a parallel server system that has 2 buffers, 2 processors and 3 activities, [Harrison \(1998\)](#) develops a “discrete-review” policy via the BIGSTEP procedure that was first described in [Harrison \(1996\)](#) for multiclass queueing networks. Furthermore, he proves that the discrete-review policy asymptotically minimizes the expected discounted, cumulative linear holding cost under the restrictive assumption of Poisson arrival processes and deterministic service times. [Harrison and López \(1999\)](#) generalize this work and use the BIGSTEP procedure to produce a family of policies for general parallel systems. For the same 2 buffer, 3 processor and 3 activity parallel server system,

but with general arrival processes and service time distributions, [Bell and Williams \(2001\)](#) develop simple form, buffer priority policies with thresholds and prove their asymptotic optimality under linear holding cost. Since the threshold values are constantly monitored, these policies are termed as “continuous-review” policies. They further generalize their policies to general parallel server systems in [Bell and Williams \(2005\)](#) and prove that they are asymptotically optimal. While all these works deal with the holding cost objective, the proposed asymptotically optimal policies in the literature exploit the special network structures and critically depend on the network data, particularly the arrival rates. [Mandelbaum and Stolyar \(2004\)](#) propose a generalized  $c\mu$  policy for parallel server systems. The policy does not use any arrival rate information. They prove that it is asymptotically optimal in minimizing a strictly convex holding cost.

All asymptotic optimality proofs in the literature involve proving a heavy traffic limit theorem and some form of state space collapse, either explicitly or implicitly. [Ata and Kumar \(2005\)](#) and [Bell and Williams \(2001, 2005\)](#) prove the state space collapse results directly without going through fluid models. [Stolyar \(2004\)](#) and [Mandelbaum and Stolyar \(2004\)](#) mimic the general framework of [Bramson \(1998\)](#) and [Williams \(1998b\)](#). They start with showing a state space collapse result for fluid models, and then prove the optimality directly without proving the state space collapse in diffusion limit as an intermediate step. By choosing to extend Bramson and Williams’s framework in this paper, we are able to provide an optimality proof that is clean and hopefully easy to follow. We expect our extension can be adapted to other studies of stochastic processing networks. Our proof of asymptotic optimality requires that the service times have finite  $2 + \varepsilon$  moments, as in [Ata and Kumar \(2005\)](#). This moment assumption is weaker than the exponential moment assumption that is usually assumed in the literature; see, for example, [Harrison \(1998\)](#) and [Bell and Williams \(2001, 2005\)](#).

Harrison pioneered Brownian control models as a framework to find asymptotically optimal service policies for networks in heavy traffic. The framework was first proposed for multiclass queueing networks in [Harrison \(1988\)](#), and later was extended for stochastic processing networks in [Harrison \(2000\)](#). In his framework, a corresponding Brownian control model of a stochastic processing network is first solved, and then the solution to the Brownian model is used to construct service policies for the original stochastic processing network. Finally, these policies are shown to be asymptotically optimal for the stochastic processing network under a heavy traffic condition. A key step to solving the Brownian control model is to have an *equivalent workload formulation* of the Brownian control model as explained in [Harrison and Van Mieghem \(1997\)](#). The “workload process” of the Brownian control model corresponding to the stochastic processing network in this paper, as well as in [Ata and Kumar \(2005\)](#), [Stolyar \(2004\)](#), and [Bell and Williams \(2001, 2005\)](#), is one-dimensional. Thus, the Brownian control model has a simple solution. Our maximum pressure policies, at least under a special linear holding cost structure, can be considered as another “interpretation” of the solution to the Brownian control model, although this interpretation is not as direct as those in [Bell and Williams \(2005\)](#) and [Ata and Kumar \(2005\)](#). Our paper, together with these papers in the literature, demonstrates that the interpretation of the Brownian

solution is not unique, and proving the optimality of the interpreted policies can be difficult.

Maximum pressure type of policies were pioneered by Tassiulas and his co-authors under various names including “back-pressure policies”; see, for example, [Tassiulas and Ephremides \(1992, 1993\)](#), [Tassiulas \(1995\)](#) and [Tassiulas and Bhattacharya \(2000\)](#). The work of [Tassiulas and Bhattacharya \(2000\)](#) represents a significant advance in finding efficient operational policies for a wide class of networks, and is closely related [Dai and Lin \(2005\)](#). Readers are referred to [Dai and Lin \(2005\)](#) for an explanation of the major differences of these two works. We note that, contrary to the description in [Dai and Lin \(2005\)](#), [Tassiulas and Ephremides \(1992, 1993\)](#) and [Tassiulas \(1995\)](#) do cover network models, not just one-pass systems. For a recent survey of these policies and their applications to wireless networks, see [Georgiadis, Neely and Tassiulas \(2006\)](#).

The remainder of the paper is organized as follows. In [Section 1.1](#), we collect some of the notation used in this paper. In [Section 2](#), we describe a class of stochastic processing networks, and introduce the maximum pressure service policies. We then define the workload process of a stochastic processing network in [Section 3](#), where we also introduce the complete resource pooling condition. The main results of this paper are stated in [Section 4](#). The proofs of the main theorems are outlined in [Section 5](#). A key to the proofs of these theorems is a state space collapse result of the diffusion-scaled network processes under a maximum pressure policy. In [Section 6](#), each fluid model solution under a maximum pressure policy is shown to exhibit a state space collapse. [Section 7](#) applies Bramson’s approach ([Bramson, 1998](#)) to prove the state space collapse of the diffusion-scaled network processes. The state space collapse result is converted into a heavy traffic limit theorem in [Section 8](#). The limit theorem is used in [Section 5](#) to complete the proofs of the main theorems. A number of technical lemmas as well as [Theorem 1](#) are proved in the appendix.

## 1.1 Notation

We use  $\mathbb{R}^d$  to denote the  $d$ -dimensional Euclidean space. Vectors in  $\mathbb{R}^d$  are envisioned as column vectors unless indicated otherwise. The transpose of a vector  $v$  will be denoted as  $v'$ . For  $v, w \in \mathbb{R}^d$ ,  $v \cdot w$  denotes the dot product, and  $v \times w$  denotes the vector  $(v_1 w_1, \dots, v_d w_d)'$ . The max norm in  $\mathbb{R}^d$  is denoted as  $|\cdot|$ , and for a matrix  $A$ , we use  $|A|$  to denote the maximum absolute value among all components. The Euclidean norm  $\|\cdot\|$  in  $\mathbb{R}^d$  is defined by  $\|v\| = \sqrt{v \cdot v}$ . For  $r_1, r_2 \in \mathbb{R}$ , we use  $r_1 \vee r_2$  and  $r_1 \wedge r_2$  to denote the maximum and minimum of  $r_1$  and  $r_2$ , respectively. We use  $\mathbb{D}^d[0, \infty)$  to denote the set of functions  $f : [0, \infty) \mapsto \mathbb{R}^d$  that are right continuous on  $[0, \infty)$  having left limits in  $(0, \infty)$ . For  $f \in \mathbb{D}^d[0, \infty)$ , we let

$$\|f\|_t = \sup_{0 \leq s \leq t} |f(s)|.$$

We endow the function space  $\mathbb{D}^d[0, \infty)$  with the usual Skorohod  $J_1$ -topology ([Ethier and Kurtz, 1986](#)). A sequence of functions  $\{f_r\} \subset \mathbb{D}^d[0, \infty)$  is said to converge to an  $f \in \mathbb{D}^d[0, \infty)$  uniformly on compact (u.o.c.) sets, denoted as  $f_r(\cdot) \rightarrow f(\cdot)$ , if for each  $t \geq 0$ ,  $\lim_{r \rightarrow \infty} \|f_r - f\|_t = 0$ . 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.

## 2 Stochastic Processing Networks

In this section, we describe a general stochastic processing network proposed by Dai and Lin (2005). We follow the notation of Dai and Lin (2005). The network is assumed to have  $\mathbf{I} + 1$  buffers,  $\mathbf{J}$  activities and  $\mathbf{K}$  processors. Buffers, activities and processors are indexed by  $i = 0, \dots, \mathbf{I}$ ,  $j = 1, \dots, \mathbf{J}$  and  $k = 1, \dots, \mathbf{K}$ , respectively. For notational convenience, we define  $\mathcal{I} = \{1, \dots, \mathbf{I}\}$  the set of buffers excluding buffer 0,  $\mathcal{J} = \{1, \dots, \mathbf{J}\}$  the set of activities and  $\mathcal{K} = \{1, \dots, \mathbf{K}\}$  the set of processors. Each buffer, with infinite capacity, holds jobs or materials that await service. Buffer 0 is a special one that is used to model the outside world, where an infinite number of jobs await. Each activity can simultaneously process jobs from a set of buffers. It may require simultaneous possession of multiple processors to be active. Jobs departing from a buffer will go next to other buffers with certain probabilities that depend on the current activity taken.

### 2.1 Resource consumption

Each activity needs one or more processors available to be active. For activity  $j$ ,  $A_{kj} = 1$ , if activity  $j$  requires processor  $k$ , and  $A_{kj} = 0$  otherwise. The  $\mathbf{K} \times \mathbf{J}$  matrix  $A = (A_{kj})$  is the resource consumption matrix. Each activity may be allowed to process jobs in multiple buffers simultaneously. For activity  $j$ , we use the indicator function  $B_{ji}$  to record whether buffer  $i$  can be processed by activity  $j$ . ( $B_{ji} = 1$  if activity  $j$  processes buffer  $i$  jobs.) The set of buffers  $i$  with  $B_{ji} = 1$  is said to be the *constituency* of activity  $j$ . It is denoted by  $\mathcal{B}_j$ . The constituency is assumed to be nonempty for each activity  $j \in \mathcal{J}$ , and may contain more than one buffer. When a processing requirement of an activity is met, a job departs from each one of the constituent buffers. For each activity  $j$ , we use  $u_j(\ell)/\mu_j$  to denote the  $\ell$ th activity  $j$  processing requirement, where  $u_j = \{u_j(\ell), \ell \geq 1\}$  is an i.i.d. sequence of random variables and  $\mu_j$  is a strictly positive real number. We set  $\sigma_j^2 = \text{var}(u_j(1))$ , and assume that  $\sigma_j < \infty$  and  $u_j$  is *unitized*, that is,  $\mathbb{E}[u_j(1)] = 1$ . It follows that  $1/\mu_j$  and  $\sigma_j$  are the mean and coefficient of variation, respectively, for the processing times of activity  $j$ .

An activity  $j$  is said to be an *input activity* if it processes jobs only from buffer 0, i.e.,  $\mathcal{B}_j = \{0\}$ . An activity  $j$  is said to be a *service activity* if it does not process any job from buffer 0, i.e.,  $0 \notin \mathcal{B}_j$ . We assume that each activity is either an input activity or a service activity. We further assume that each processor processes either input activities only or service activities only. A processor that only processes input activities is called an *input processor*, and a processor that only processes service activities is called a *service processor*. The input processors process jobs from buffer 0 (outside) and generate the arrivals for the network. We denote  $\mathcal{J}_I$  to be the set of input activities,  $\mathcal{J}_S$  the set of service activities,  $\mathcal{K}_I$  the set of input processors, and  $\mathcal{K}_S$  the set of service processors.

### 2.2 Routing

Buffer  $i$  jobs, after being processed by activity  $j$ , will go next to other buffers or leave the system. Let  $e_0$  be the  $\mathbf{I}$ -dimensional vector of all 0's, and for  $i \in \mathcal{I}$ ,  $e_i$  is the  $\mathbf{I}$ -dimensional

vector with the  $i$ th component 1 and other components 0. For each activity  $j \in \mathcal{J}$  and each constituent buffer  $i \in \mathcal{B}_j$ , we use an  $\mathbf{I}$ -dimensional binary random vector  $\phi_i^j(\ell) = (\phi_{i' i'}^j(\ell), i' \in \mathcal{I})$  to denote the *routing vector* of the  $\ell$ -th buffer  $i$  job processed by activity  $j$ , where  $\phi_i^j(\ell) = e_{i'}$  if the  $\ell$ -th buffer  $i$  job processed by activity  $j$  goes next to buffer  $i'$ , and  $\phi_i^j(\ell) = e_0$  if the job leaves the system. We assume that the sequence  $\phi_i^j = \{\phi_i^j(\ell), \ell \geq 1\}$  is i.i.d. for each activity  $j \in \mathcal{J}$  and  $i \in \mathcal{B}_j$ . Set  $P_{i' i'}^j = \mathbb{E}[\phi_{i' i'}^j(1)]$ . Then  $P_{i' i'}^j$  is the probability that a buffer  $i$  job processed by activity  $j$  will go next to buffer  $i'$ .

For each  $j \in \mathcal{J}, i \in \mathcal{B}_j$ , the cumulative routing process is defined by the sum

$$\Phi_i^j(\ell) = \sum_{n=1}^{\ell} \phi_i^j(n),$$

and  $\Phi_{i' i'}^j(\ell)$  denotes the number of jobs that will go next to buffer  $i'$  among the first  $\ell$  buffer  $i$  jobs that are processed by activity  $j$ .

The sequences

$$(u_j, \phi_i^j : i \in \mathcal{B}_j, j \in \mathcal{J})$$

are said to be the *primitive increments* of the network. We assume that they are mutually independent and all are independent of the initial state of the network.

## 2.3 Resource allocations

Because multiple activities may require usage of the same processor, not all activities can be simultaneously undertaken at a 100% level. Unless stated otherwise, we assume that each processor's service capacity is infinitely divisible, and processor-splitting of a processor's service capacity is realizable. We use a nonnegative variable  $a_j$  to denote the level at which processing activity  $j$  is undertaken. When  $a_j = 1$ , activity  $j$  is employed at a 100% level. When  $a_j = 0$ , activity  $j$  is not employed. Suppose that the engagement level of activity  $j$  is  $a_j$ , with  $0 \leq a_j \leq 1$ . The processing requirement of an activity  $j$  job is depleted at rate  $a_j$ . (The job finishes processing when its processing requirement reaches 0.) The activity consumes  $a_j A_{kj}$  fraction of processor  $k$ 's service capacity per unit time. The remaining service capacity,  $1 - a_j A_{kj}$ , can be used for other activities.

We use  $a = (a_j) \in \mathbb{R}_+^{\mathbf{J}}$  to denote the corresponding  $\mathbf{J}$ -dimensional allocation (column) vector, where  $\mathbb{R}_+$  denotes the set of nonnegative real numbers. Since each processor  $k$  can decrease processing requirements at the rate of at most 1 per unit of time, we have

$$\sum_{j \in \mathcal{J}} A_{kj} a_j \leq 1 \quad \text{for each processor } k. \quad (1)$$

In vector form,  $Aa \leq e$ , where  $e$  is the  $\mathbf{K}$ -dimensional vector of ones. We assume that there is at least one input activity and that the input processors never idle. Namely,

$$\sum_{j \in \mathcal{J}} A_{kj} a_j = 1 \quad \text{for each input processor } k. \quad (2)$$

We use  $\mathcal{A}$  to denote the set of allocations  $a \in \mathbb{R}_+^{\mathbf{J}}$  that satisfy (1) and (2).

Each  $a \in \mathcal{A}$  represents an allowable allocation of the processors working on various activities. We note that  $\mathcal{A}$  is bounded and convex. Let  $\mathcal{E} = \{a^1, \dots, a^{\mathbf{E}}\}$  be the set of extreme points of  $\mathcal{A}$ , where the total number  $\mathbf{E}$  of extreme points is finite.

**Remark.** Harrison (2002) does not have the concept of input processor and input activity for his stochastic processing networks. The input processors and activities introduced in this paper allow us to capture the dynamic routing decisions for external arrivals. However, if we restrict ourselves to the networks without these routing decisions, Harrison’s network models are broader than ours; Harrison describes his models through first-order network data only, leaving the underlying stochastic primitives and system dynamics unspecified. In this sense, our network models are a special class of Harrison’s, which include some stochastic processing networks where the service requirements for jobs from different buffers can be different even if they are processed simultaneously by a single activity. The latter networks are not covered in this paper.

## 2.4 Service policies

Each job in a buffer is assumed to be processed by one activity in its entire stay at the buffer. A processing of an activity can be preempted. In this case, each in-service job is “frozen” by the activity. The next time the activity is made active again, the processing is resumed from where it was left off. In addition to the availability of processors, a (non-preempted) activity can be made active only when each constituent buffer has jobs that are not in service or frozen. We assume that within each buffer *head-of-line* policy is used. When a (non-preempted) activity becomes active with a given engagement level, the leading job in each buffer that is not in service or frozen is processed. If multiple activities are actively working on a buffer, there are multiple jobs in the buffer that are in service. For an allocation  $a$ , if there is an activity  $j$  with  $a_j > 0$  that cannot be made active, the allocation is infeasible. At any given time  $t$ , we use  $\mathcal{A}(t)$  to denote the set of allocations that are *feasible* at that time. A service policy specifies which allocation is being undertaken at each time  $t \geq 0$ , and we use  $\pi = \{\pi(t) : t \geq 0\}$  to denote such a policy. Under the policy  $\pi$ , allocation  $\pi(t) \in \mathcal{A}(t)$  will be employed at time  $t$ .

Dai and Lin (2005) propose a family of service policies called *maximum pressure policies* that are throughput optimal for a large class of stochastic processing networks. To describe these policies for our network, for each buffer  $i = 1, \dots, \mathbf{I}$  and each activity  $j = 1, \dots, \mathbf{J}$  we define

$$R_{ij} = \mu_j \left( B_{ji} - \sum_{i' \in \mathcal{B}_j} P_{i'i}^j \right). \quad (3)$$

The  $\mathbf{I} \times \mathbf{J}$  matrix  $R = (R_{ij})$  is called the input-output matrix in Harrison (2002). One interprets  $R_{ij}$  as the average amount of buffer  $i$  material consumed per unit of activity  $j$ , with a negative value being interpreted to mean that activity  $j$  is a net producer of material in buffer  $i$ . Define  $\mathcal{E}(t) = \mathcal{E} \cap \mathcal{A}(t)$  to be the set of feasible extreme allocations at time  $t$ . Denote  $Z = \{Z(t), t \geq 0\}$  to be the buffer level process with  $Z_i(t)$  being the buffer level of



buffer  $i$  at time  $t$ , including those in service or “frozen” at time  $t$ . Now we are ready to define maximum pressure service policies for our stochastic processing network. Each maximum pressure policy is associated with a pair of vectors  $\alpha, \beta \in \mathbb{R}^I$  with  $\alpha > 0$ . Recall that for two vectors  $\alpha, \beta \in \mathbb{R}^d$ ,  $\alpha \times \beta$  denotes the vector  $(\alpha_1\beta_1, \dots, \alpha_d\beta_d)'$ .

**Definition 1** (Maximum pressure policies). Given a pair of vectors  $\alpha, \beta \in \mathbb{R}^I$  with  $\alpha > 0$ , a service policy is said to be a *maximum pressure policy associated with parameter set*  $(\alpha, \beta)$  if at each time  $t$ , the network chooses an allocation  $a^* \in \operatorname{argmax}_{a \in \mathcal{E}(t)} p_{\alpha, \beta}(a, Z(t))$ , where  $p_{\alpha, \beta}(a, Z(t))$  is called *the network pressure with parameter set*  $(\alpha, \beta)$  *under allocation*  $a$  *and buffer level*  $Z(t)$  and is defined as

$$p_{\alpha, \beta}(a, Z(t)) = (\alpha \times Z(t) - \beta) \cdot Ra. \quad (4)$$

When more than one allocation attains the maximum pressure, a tie-breaking rule is used. Our results are not affected by how ties are broken. However, for concreteness, one can order the extreme allocation set  $\mathcal{E}$ , and always choose the maximum-pressure allocation with the smallest index.

In Section 4, we are going to show that each maximum pressure policy asymptotically minimizes a certain *workload process*, and for each  $\varepsilon > 0$ , we are going to identify a set of maximum pressure policies and show that they are *asymptotically  $\varepsilon$ -optimal* when the stochastic processing network incurs a linear holding cost.

### 3 Workload Process and Complete Resource Pooling

We define the workload process through a linear program (LP) called the static planning problem and its dual problem. For a stochastic processing network with input-output matrix  $R$  and capacity consumption matrix  $A$ , the static planning problem is defined as follows: choose a scalar  $\rho$  and a  $\mathbf{J}$ -dimensional column vector  $x$  so as to

$$\text{minimize} \quad \rho \quad (5)$$

$$\text{subject to} \quad Rx = 0, \quad (6)$$

$$\sum_{j \in \mathcal{J}} A_{kj} x_j = 1 \text{ for each input processor } k, \quad (7)$$

$$\sum_{j \in \mathcal{J}} A_{kj} x_j \leq \rho \text{ for each service processor } k, \quad (8)$$

$$x \geq 0. \quad (9)$$

For each optimal solution  $(\rho, x)$  to (5)-(9), the vector  $x$  is said to be a *processing plan* for the stochastic processing network, where component  $x_j$  is interpreted as the long-run fraction of time that activity  $j$  is undertaken. Since one of the constraints in (8) must be bounding for a service processor,  $\rho$  is interpreted as the long-run utilization of the busiest service processor under the processing plan. With this interpretation, the left side of (6) is

interpreted as the long-run *net flow rates* from the buffers. Equality (6) demands that, for each buffer, the long-run input rate to the buffer is equal to the long-run output rate from the buffer. Equality (7) ensures that input processors never idle, while inequality (8) requires that each service processor's utilization not exceed that of the busiest service processor. The objective is to minimize the utilization of the busiest service processor. For future references, the optimal objective value  $\rho$  is said to be the *traffic intensity* of the stochastic processing network.

The dual problem of the static planning problem is the following: choose an  $\mathbf{I}$ -dimensional vector  $y$  and a  $\mathbf{K}$ -dimensional vector  $z$  so as to

$$\text{maximize} \quad \sum_{k \in \mathcal{K}_I} z_k, \quad (10)$$

$$\text{subject to} \quad \sum_{i \in \mathcal{I}} y_i R_{ij} \leq - \sum_{k \in \mathcal{K}_I} z_k A_{kj}, \text{ for each input activity } j, \quad (11)$$

$$\sum_{i \in \mathcal{I}} y_i R_{ij} \leq \sum_{k \in \mathcal{K}_S} z_k A_{kj}, \text{ for each service activity } j, \quad (12)$$

$$\sum_{k \in \mathcal{K}_S} z_k = 1, \quad (13)$$

$$z_k \geq 0, \text{ for each service processor } k. \quad (14)$$

Recall that  $\mathcal{K}_I$  is the set of input processors, and  $\mathcal{K}_S$  is the set of service processors. Each pair  $(y, z)$  that satisfies (11)-(14) is said to be a *resource pool*. Component  $y_i$  is interpreted as the work dedicated to a unit of buffer  $i$  job by the resource pool, and  $z_k$  is interpreted as the relative capacity of processor  $k$ , measured in fractions of the service capacity of the resource pool; for each input processor  $k$ , the relative capacity  $z_k$  is the amount of work generated by input processor  $k$  per unit of time. Equality (13) ensures that the service capacity of the resource pool equals the sum of service capacities of all service processors. Constraint (12) demands that no service activity can accomplish more work than the capacity it consumes. Recall that  $-R_{ij}$  is the rate at which input activity  $j$  generates buffer  $i$  jobs. For each input activity  $j$ , constraint (11), which can be written as  $\sum_{i \in \mathcal{I}} y_i (-R_{ij}) \geq \sum_{k \in \mathcal{K}_I} z_k A_{kj}$ , ensures that the work dedicated to per unit of the activity is no less than that which it generates. The objective is to maximize  $\sum_{k \in \mathcal{K}_I} z_k$ , which is the total amount of work generated from outside by the input processors per unit of time. A service processor  $k$  is said to be in the resource pool  $(y, z)$  if  $z_k > 0$ .

A *bottleneck pool* is defined to be an optimal solution  $(y^*, z^*)$  to the dual LP (10)-(14). Let  $(\rho^*, x^*)$  be an optimal solution to the primal LP, the static planning problem (5)-(9). From the basic duality theory,  $\sum_j A_{kj} x_j^* = \rho^*$  for any service processor  $k$  with  $z_k^* > 0$ . It says that all service processors in the bottleneck pool  $(y^*, z^*)$  are the busiest servers under any optimal processing plan  $x^*$ .

For a bottleneck pool  $(y^*, z^*)$ , let  $W(t) = y^* \cdot Z(t)$  for  $t \geq 0$ . Then,  $W(t)$  represents the average total work of this bottleneck pool embodied in all jobs that are present at time  $t$  in the stochastic processing network. The process  $W = \{W(t), t \geq 0\}$  is called the

*workload process* of this bottleneck pool. Although the workload process of a non-bottleneck resource pool  $(y, z)$  can also be defined by  $y \cdot Z(t)$ , we will focus on the workload processes of bottleneck pools because bottleneck pools become significantly more important in heavy traffic. In general, the bottleneck pool is not unique. However, we assume all the stochastic processing networks considered in this paper have a unique bottleneck pool; namely, they satisfy the following *complete resource pooling condition*.

**Definition 2** (Complete resource pooling condition). A stochastic processing network is said to satisfy the complete resource pooling condition if the corresponding dual static planning problem (10)–(14) has a nonnegative, unique optimal solution  $(y^*, z^*)$ .

For a processing network that satisfies the complete resource pooling condition, we define the *bottleneck workload process*, or simply the workload process, of the stochastic processing network to be the workload process of its unique bottleneck pool.

The (bottleneck) workload process defined here is different from the workload process defined in Harrison and Van Mieghem (1997). Their workload process is multi-dimensional, with some components corresponding to the non-bottleneck pools; it is defined in terms of what they call “reversible displacements.” For the networks where their workload process has dimension one, these two definitions of the workload process are consistent.

**Remark.** Under certain assumptions including a heavy traffic assumption that requires all servers in the network be critically loaded, Harrison (2000) proposes a “canonical” representation of the workload process for a stochastic processing network through a dual LP that is similar to (10)–(14). There, basic optimal solutions to the dual LP were chosen as rows of the workload matrix which was used to define the workload process. Without his heavy traffic assumption, his “canonical” choice of workload matrix would exclude those non-bottleneck servers. In this case, it is not clear how to define a “canonical” representation of the workload process to include those nonbottleneck servers. Although for some network examples like multiclass queueing networks, we can define the workload matrix such that its rows are the basic solutions to the dual LP, more analysis is required for general stochastic processing networks.

## 4 Asymptotic Optimality and the Main Results

The behavior of the buffer level process and the workload process for a stochastic processing network under any policy is complex. In particular, deriving closed form expressions for performance measures involving these processes is not possible in general. Therefore, we perform an asymptotic analysis for stochastic processing networks operating under maximum pressure policies. Our asymptotic region is when the network is in heavy traffic; i.e., the offered traffic load is approximately equal to the system capacity. Formally, we consider a sequence of stochastic processing networks indexed by  $r = 1, 2, \dots$ ; as  $r \rightarrow \infty$ , the traffic intensity  $\rho^r$  of the  $r$ th network goes to one. We assume that these networks all have the same network topology and primitive increments. In other words, the matrices  $A$  and  $B$ , and the sequences  $(u_j, \phi_i^j : j \in \mathcal{J}, i \in \mathcal{B}_j)$  do not vary with  $r$ . However, we allow the processing

rates to change with  $r$ , and use  $\mu_j^r$  to denote the processing rate of activity  $j$  in the  $r$ th network. Thus, the traffic intensity  $\rho^r$  for the  $r$ th network is the optimal objective value of the static planning problem (5)–(9) with the input-output matrix  $R^r = (R_{ij}^r)$  given by  $R_{ij}^r = \mu_j^r \left( B_{ji} - \sum_{i' \in \mathcal{B}_j} P_{i'i}^j \right)$ . We assume the following *heavy traffic assumption* throughout this paper.

**Assumption 1** (Heavy traffic assumption). There exists a constant  $\mu_j > 0$  for each activity  $j \in \mathcal{J}$  such that as  $r \rightarrow \infty$ ,

$$\mu_j^r \rightarrow \mu_j, \quad (15)$$

and, setting  $R = (R_{ij})$  as in (3) with  $\mu_j$  being the limit values in (15), the static planning problem (5)–(9) with parameter  $(R, A)$  has a unique optimal solution  $(\rho^*, x^*)$  with  $\rho^* = 1$ . Furthermore, as  $r \rightarrow \infty$ ,

$$r(\rho^r - 1) \rightarrow \theta \quad (16)$$

for some constant  $\theta$ .

We define the *limit network* of the network sequence to be the network that has the same network topology and primitive increments as networks in the sequence, and has processing rates equal to the limit values  $\mu_j$ , given in (15). Assumption 1 basically means that in the limit network there exists a unique processing plan  $x^*$  that can avoid inventory buildups over time, and the busiest service processor is fully utilized under this processing plan. Condition (16) requires that the networks' traffic intensities approach to 1 at rate  $r^{-1}$  or faster.

The heavy traffic assumption is now quite standard in the literature; see, for example, [Bramson and Dai \(2001\)](#), [Chen and Zhang \(2000\)](#), [Dai and Kurtz \(1995\)](#), and [Williams \(1996, 1998a\)](#) for heavy traffic analysis of queueing networks and [Ata and Kumar \(2005\)](#), [Harrison \(2000\)](#), [Harrison and López \(1999\)](#), [Harrison and Van Mieghem \(1997\)](#), and [Williams \(2000\)](#) for heavy traffic analysis of stochastic processing networks. However, heavy traffic assumptions in the literature usually assume that, in addition to Assumption 1, *all* service processors are fully utilized. The latter assumption, together with a complete resource pooling assumption to be introduced later in this section, rules out some common networks such as multiclass queueing networks. In our heavy traffic assumption, only the busiest service processor is required to be critically loaded, and some other service processors are allowed to be under-utilized.

The optimal processing plan  $x^*$  given in Assumption 1 is referred to as the *nominal processing plan*. We denote  $T_j(t)$  the cumulative amount of activity  $j$  processing time in  $[0, t]$  for the limit network; let  $T(t)$  be the corresponding  $\mathbf{J}$ -dimensional vector. Then  $T(t)/t$  is the average activity levels over the time span  $[0, t]$ . To avoid a linear buildup of jobs over time in the limit network, the long-run average rate (or activity level) that activity  $j$  is undertaken needs to equal  $x_j^*$ , i.e.,

$$\lim_{t \rightarrow \infty} T(t)/t = x^* \text{ almost surely.} \quad (17)$$

There should be no linear buildup of jobs under a reasonably “good” policy. A policy is said to be efficient for the limit network if (17) holds for the network operating under the policy. Since we consider a sequence of networks, we would like to define an analogous notion of a “good” or efficient policy for the sequence. One can imagine that under a reasonably “good” policy, when  $r$  is large, the average activity levels over long time spans must be very close to the nominal processing plan  $x^*$ . To be specific, we define the notion of *asymptotic efficiency* as follows. Let  $T_j^r(t)$  be the cumulative amount of activity  $j$  processing time in  $[0, t]$  for the  $r$ th network and  $T^r(t)$  be the corresponding  $\mathbf{J}$ -dimensional vector.

**Definition 3** (Asymptotic efficiency). Consider a sequence of stochastic processing networks indexed by  $r = 1, 2, \dots$ , where Assumption 1 holds. A policy  $\pi$  is said to be *asymptotically efficient* if and only if under policy  $\pi$ , with probability 1, for each  $t \geq 0$ ,

$$T^r(r^2t)/r^2 \rightarrow x^*t \quad \text{as } r \rightarrow \infty. \quad (18)$$

Equation (18) basically says that, under an asymptotically efficient policy, the average activity levels over a time span of order  $r^2$  are very close to the nominal processing plan, so that no linear buildup of jobs will occur over the time span of this order.

Asymptotic efficiency is closely related to the *throughput optimality* as defined in Dai and Lin (2005). Fluid models have been used to prove throughput optimality of stochastic processing networks operating under a policy. Similarly, the fluid model corresponding to the limit network can be used to prove the asymptotic efficiency of a policy for the sequence of networks that satisfies Assumption 1. In particular, one can prove that a policy  $\pi$  is asymptotically efficient if the fluid model of the limit network operating under  $\pi$  is weakly stable. To introduce the first result of this paper, we first define the extreme-allocation-available (EAA) condition for a stochastic processing network.

**Definition 4** (EAA condition). A stochastic processing network is said to satisfy the *EAA condition* if, for any vector  $q \in \mathbb{R}_+^{\mathbf{I}}$ , there exists an extreme allocation  $a^* \in \mathcal{E}$  such that  $Ra^* \cdot q = \max_{a \in \mathcal{E}} Ra \cdot q$ , and for each buffer  $i$  with  $\sum_j a_j^* B_{ij} > 0$ , buffer level  $q_i$  is positive.

**Theorem 1.** *Consider a sequence of stochastic processing networks that satisfies Assumption 1. If the limit network satisfies the EAA condition, then any maximum pressure policy is asymptotically efficient.*

Using a fluid model, Dai and Lin (2005) prove in their Theorem 2 that, for a stochastic processing network satisfying the *EAA condition*, a maximum pressure policy is throughput optimal. The proof of Theorem 1 is almost identical to that of Theorem 2 in Dai and Lin (2005), and will be outlined in Appendix B.

Asymptotic efficiency helps to identify reasonably “good” policies, but it is not very discriminating. We would like to demonstrate a certain sense of optimality for maximum pressure policies in terms of secondary performance measures. For this, we will introduce two notions of *asymptotic optimality*. Before giving their definitions, we make the following assumption on the sequence of networks.

**Assumption 2** (Complete resource pooling). All networks in the sequence and the limit network satisfy the complete resource pooling condition defined in Section 3. Namely, the dual static planning problem (10)–(14) of the  $r$ th network has a nonnegative, unique optimal solution  $(y^r, z^r)$ , and the dual static planning problem of the limit network also has a nonnegative, unique optimal solution  $(y^*, z^*)$ .

**Remark.** In Assumption 2, we assume all networks in the sequence satisfy the complete resource pooling condition so that the workload processes  $W^r$  can be uniquely defined as in Section 3 by the first order network data  $(R^r, A^r)$ . This assumption can be removed if one defines the workload process of a network with multiple bottleneck pools to be the workload process of an arbitrarily chosen but prespecified bottleneck pool (with  $y^r$  being any given optimal solution to the dual problem). On the other hand, the complete resource pooling condition for the limit network is crucial for our results to hold.

The first notion of asymptotic optimality is in terms of the workload process introduced in Section 3. Under Assumption 2, we can define the one-dimensional workload process of the  $r$ th network as

$$W^r(t) = y^r \cdot Z^r(t), \quad (19)$$

where  $Z^r = \{Z^r(t), t \geq 0\}$  is the buffer level process of the  $r$ th network.

The second notion of asymptotic optimality is in terms of a linear holding cost structure for the sequence of stochastic processing networks. For each network in the sequence, we assume that there is a linear holding cost incurred at rate  $h_i > 0$  for each job in buffer  $i$ . Let  $h$  be the corresponding vector. For the  $r$ th network, the total holding cost rate at time  $t$  is

$$H^r(t) = h \cdot Z^r(t). \quad (20)$$

Define the diffusion-scaled workload, buffer level, and holding cost rate processes of the  $r$ th network  $\widehat{W}^r = \{\widehat{W}^r(t), t \geq 0\}$ ,  $\widehat{Z}^r = \{\widehat{Z}^r(t), t \geq 0\}$ , and  $\widehat{H}^r = \{\widehat{H}^r(t), t \geq 0\}$  via  $\widehat{W}^r(t) = W^r(r^2t)/r$ ,  $\widehat{Z}^r(t) = Z^r(r^2t)/r$ , and  $\widehat{H}^r(t) = H^r(r^2t)/r$ . Clearly,  $\widehat{W}^r(t) = y^r \cdot \widehat{Z}^r(t)$  and  $\widehat{H}^r(t) = h \cdot \widehat{Z}^r(t)$  for  $t \geq 0$ .

**Definition 5.** Consider a sequence of stochastic processing networks indexed by  $r$ . An asymptotically efficient policy  $\pi^*$  is said to be *asymptotically optimal for workload* if and only if for any  $t > 0, w > 0$ , and any asymptotically efficient policy  $\pi$ ,

$$\limsup_{r \rightarrow \infty} \mathbb{P}(\widehat{W}_{\pi^*}^r(t) > w) \leq \liminf_{r \rightarrow \infty} \mathbb{P}(\widehat{W}_{\pi}^r(t) > w), \quad (21)$$

where  $\widehat{W}_{\pi^*}^r(\cdot)$  and  $\widehat{W}_{\pi}^r(\cdot)$  are the diffusion-scaled workload processes under policies  $\pi^*$  and  $\pi$ , respectively.

**Definition 6.** Consider a sequence of stochastic processing networks indexed by  $r$ . For a given  $\varepsilon > 0$ , an asymptotically efficient policy  $\pi^*$  is said to be *asymptotically  $\varepsilon$ -optimal for holding cost* if and only if for any  $t > 0, w > 0$ , and any asymptotically efficient policy  $\pi$ ,

$$\limsup_{r \rightarrow \infty} \mathbb{P}(\widehat{H}_{\pi^*}^r(t) > w) \leq \liminf_{r \rightarrow \infty} \mathbb{P}((1 + \varepsilon)\widehat{H}_{\pi}^r(t) > w), \quad (22)$$

where  $\widehat{H}_{\pi^*}^r(t)$  and  $\widehat{H}_{\pi}^r(t)$  are the diffusion-scaled total holding cost rates at time  $t$  under policies  $\pi^*$  and  $\pi$ , respectively.

To state our main theorems, we make two more assumptions on the sequence of networks. One is a moment assumption on the unitized service times  $u_j(\ell)$  and the other is an assumption on the initial buffer level processes.

**Assumption 3.** There exists an  $\varepsilon_u > 0$  such that, for all  $j$ ,

$$\mathbb{E}[(u_j(1))^{2+\varepsilon_u}] < \infty. \quad (23)$$

**Assumption 4** (Initial condition). As  $r \rightarrow \infty$ ,

$$\widehat{Z}^r(0) = Z^r(0)/r \rightarrow 0 \text{ in probability.} \quad (24)$$

Assumption 3 requires that the unitized service times have finite  $2 + \varepsilon_u$  moments. It was introduced by [Ara and Kumar \(2005\)](#) and it is stronger than some standard regularity assumptions such as in [Bramson \(1998\)](#). Assumption 3 will be used in Section 7 to prove a state space collapse result for stochastic processing networks operating under maximum pressure policies. Assumption 4 holds if the initial buffer levels of the networks are stochastically bounded, namely,

$$\lim_{\tau \rightarrow \infty} \limsup_{r \rightarrow \infty} \mathbb{P}(|Z^r(0)| > \tau) = 0.$$

Clearly, Assumption 4 implies that as  $r \rightarrow \infty$ ,

$$\widehat{W}^r(0) \rightarrow 0 \text{ in probability.} \quad (25)$$

**Theorem 2.** *Consider a sequence of stochastic processing networks. Assume Assumptions 1–4 and that the limit network satisfies the EAA condition. Each maximum pressure policy is asymptotically optimal for workload.*

**Theorem 3.** *Consider a sequence of stochastic processing networks where Assumptions 1–4 hold and the limit network satisfies the EAA condition. For any given  $\varepsilon > 0$ , there exists a maximum pressure policy  $\pi^*$  that is asymptotically  $\varepsilon$ -optimal for holding cost.*

Theorem 2 says that, at every time  $t$ , asymptotically, the diffusion-scaled workload under any maximum pressure policy  $\pi^*$  is dominated by that under any other asymptotically efficient policy  $\pi$  in the sense of stochastic ordering. Theorem 3 says that, for any  $\varepsilon > 0$ , one can always find a maximum pressure policy such that, for any time  $t$ , the total holding cost rate under the maximum pressure policy is asymptotically dominated by  $(1 + \varepsilon)$  times the total holding cost rate under any other efficient policy in the sense of stochastic ordering.

The proofs of Theorems 2 and 3 will be outlined in Section 5. Throughout this paper, we shall assume Assumptions 1–4 and that the limit network satisfies the EAA condition.

**Remarks. (a)** Let  $\varphi : \mathbb{D}[0, \infty) \mapsto \mathbb{R}$  be a continuous functional. Assume further that  $\varphi$  is nondecreasing in the sense that  $\varphi(f) \leq \varphi(g)$  for  $f, g \in \mathbb{D}[0, \infty)$  whenever  $f(t) \leq g(t)$  for all  $t \geq 0$ . Then using the continuous mapping theorem and following the same argument as

in the proof of Theorem 3, we can obtain a stronger version of Theorem 3 by changing (22) to

$$\limsup_{r \rightarrow \infty} \mathbb{E}[\varphi(\widehat{H}_{\pi^*}^r)] \leq (1 + \epsilon) \liminf_{r \rightarrow \infty} \mathbb{E}[\varphi(\widehat{H}_\pi^r)].$$

One particular example is  $\varphi(f)(t) = \int_0^\infty e^{-\gamma t} (f(t) \wedge M) dt$  for some  $M > 0$ .

(b) For a given  $\epsilon$ , the parameter set  $(\alpha, \beta)$  that is used to define the asymptotically  $\epsilon$ -optimal maximum pressure policy in Theorem 3 depends on the first-order network data  $R$  and  $A$ .

## 5 An Outline of the Proofs

This section outlines the proofs of our main theorems, Theorems 2 and 3. We first derive an asymptotic lower bound on the workload processes under asymptotically efficient policies. Then we state a heavy traffic limit result in Theorem 4, which implies that this asymptotic lower bound is achieved by the workload process under any maximum pressure policy. The heavy traffic limit theorem also implies the  $\epsilon$ -optimality of certain maximum pressure policies under the linear holding cost structure in Theorem 3. At the end, we outline a proof for the heavy traffic limit theorem. The key to the proof is a state space collapse result to be stated in Theorem 5.

We first derive an asymptotic lower bound on the workload processes under asymptotically efficient policies. That is, we search for a process  $W^*$  such that under any asymptotically efficient policy

$$\liminf_{r \rightarrow \infty} \mathbb{P}(\widehat{W}^r(t) > w) \geq \mathbb{P}(W^*(t) > w), \text{ for all } t \text{ and } w.$$

As before, we assume that the sequence of stochastic processing networks satisfies Assumption 2. Hence, the one dimensional workload process  $W^r$  is well defined by (19) for each  $r$ .

We begin the analysis by defining a process  $Y^r = \{Y^r(t), t \geq 0\}$  for the  $r$ th network via

$$Y^r(t) = (1 - \rho^r)t - y^r \cdot R^r T^r(t). \quad (26)$$

Since  $\rho^r$  is interpreted as the traffic intensity of the bottleneck pool, for each  $t \geq 0$ ,  $\rho^r t$  is interpreted as the average total work contributed to the bottleneck pool from the exogenous arrivals in  $[0, t]$ , and  $(1 - \rho^r)t$  represents the average total work that could have been depleted by time  $t$  if the bottleneck pool never idles. Because of the randomness of the processing times, the bottleneck pool will almost surely incur idle time over time, particularly when the system is not overloaded. Under a service policy and its corresponding activity level process  $T^r$ , the average total work that has been depleted by time  $t$  is given by

$$y^r \cdot R^r T^r(t) = \sum_{j \in \mathcal{J}_S} T_j^r(t) \sum_{i \in \mathcal{I}} y_i^r R_{ij}^r - \sum_{j \in \mathcal{J}_I} T_j^r(t) \sum_{i \in \mathcal{I}} y_i^r (-R_{ij}^r).$$

Note that, as in Section 3, for each service activity  $j \in \mathcal{J}_S$ ,  $\sum_{i \in \mathcal{I}} y_i^r R_{ij}^r$  is the average work accomplished per unit of activity  $j$ , and that for each input activity  $j \in \mathcal{J}_I$ ,  $\sum_{i \in \mathcal{I}} y_i^r (-R_{ij}^r)$  is



the average work generated per unit of activity  $j$ . Therefore,  $Y^r(t)$  represents the deviation of the workload depletion in  $[0, t]$  from that under the “best” policy. The following lemma says that this deviation does not decrease over time.

**Lemma 1.** *Consider a sequence of networks satisfying Assumption 2. For each  $r$  and each sample path, the process  $Y^r$  defined in (26) is a nondecreasing function with  $Y^r(0) = 0$ .*

We leave the proof to Appendix B.

**Remark.** Some special stochastic processing networks, such as multiclass queueing networks (Harrison, 1988) and unitary networks (Bramson and Williams, 2003), have no control on the input activities. Then,  $T_j^r(t)$  is fixed for all  $j \in \mathcal{J}_I$  under different policies, and  $\sum_{j \in \mathcal{J}_I} T_j^r(t) \sum_{i \in \mathcal{I}} y_i^r(-R_{ij}^r) = \rho^r t$ . For these networks, one gets

$$Y^r(t) = t - \sum_{j \in \mathcal{J}_S} T_j^r(t) \sum_{i \in \mathcal{I}} y_i^r R_{ij}^r,$$

and  $Y^r(t)$  is interpreted as the cumulative idle time of the bottleneck pool by time  $t$ .

For each activity  $j$ , we define the process  $S_j^r = \{S_j^r(t), t \geq 0\}$  by

$$S_j^r(t) = \max \left\{ n : \sum_{\ell=1}^n u_j(\ell) \leq \mu_j^r t \right\}.$$

Under a head-of-line service policy,  $S_j^r(t)$  is the number of activity  $j$  processing completions in  $t$  units of activity  $j$  processing time for the  $r$ th network. Then we can describe the system dynamics of the  $r$ th network by the following equation:

$$\begin{aligned} Z_i^r(t) &= Z_i^r(0) + \sum_{j \in \mathcal{J}} \sum_{i' \in \mathcal{B}_j} \Phi_{i'i}^j \left( S_j^r(T_j^r(t)) \right) \\ &\quad - \sum_{j \in \mathcal{J}} S_j^r(T_j^r(t)) B_{ji} \text{ for each } t \geq 0 \text{ and } i \in \mathcal{I}. \end{aligned} \tag{27}$$

Since quantity  $T_j^r(t)$  is the cumulative amount of activity  $j$  processing time in  $[0, t]$ ,  $S_j^r(T_j^r(t))$  is the number of activity  $j$  processings completed by time  $t$ , and  $\sum_{j \in \mathcal{J}} S_j^r(T_j^r(t)) B_{ji}$  is the total number of jobs that depart from buffer  $i \in \mathcal{I} \cup \{0\}$  in  $[0, t]$ . For each activity  $j$ ,  $\sum_{i' \in \mathcal{B}_j} \Phi_{i'i}^j \left( S_j^r(T_j^r(t)) \right)$  is the total number of jobs sent to buffer  $i$  by activity  $j$  from its constituent buffers by time  $t$ , so  $\sum_{j \in \mathcal{J}} \sum_{i' \in \mathcal{B}_j} \Phi_{i'i}^j \left( S_j^r(T_j^r(t)) \right)$  is the total number of jobs that go to buffer  $i$  by time  $t$ . Equation (27) says that the number of jobs in buffer  $i$  at time  $t$  equals the initial number plus the number of arrivals subtracted from the number of departures.

From (27), we can write the workload process  $W^r = y^r \cdot Z^r$  as

$$W^r(t) = W^r(0) + \sum_{i \in \mathcal{I}} y_i^r \sum_{j \in \mathcal{J}} \left( \sum_{i' \in \mathcal{B}_j} \Phi_{i'i}^j \left( S_j^r(T_j^r(t)) \right) - B_{ji} S_j^r(T_j^r(t)) \right).$$

Let  $X^r(t) = W^r(t) - Y^r(t)$  for each  $t \geq 0$ . Then one can check that

$$X^r(t) = W^r(0) + \sum_{i \in \mathcal{I}} y_i^r \sum_{j \in \mathcal{J}} \left( \sum_{i' \in \mathcal{B}_j} \Phi_{i'}^j \left( S_j^r(T_j^r(t)) \right) - B_{ji} S_j^r(T_j^r(t)) \right) - (1 - \rho^r)t + y^r \cdot R^r T^r(t).$$

We define the following diffusion-scaled processes:

$$\widehat{S}_j^r(t) = r^{-1} [S_j^r(r^2t) - \mu_j^r r^2t] \quad \text{for each } j \in \mathcal{J}, \quad (28)$$

$$\widehat{\Phi}_{i'}^{j,r}(t) = r^{-1} [\Phi_{i'}^j(\lfloor r^2t \rfloor) - P_{i'}^j r^2t] \quad \text{for each } j \in \mathcal{J} \text{ and each } i' \in \mathcal{B}_j, \quad (29)$$

$$\widehat{X}^r(t) = r^{-1} X^r(r^2t), \quad (30)$$

$$\widehat{Y}^r(t) = r^{-1} Y^r(r^2t). \quad (31)$$

In (29),  $\lfloor t \rfloor$  denotes the greatest integer number less than or equal to the real number  $t$ .

Then the diffusion-scaled workload process  $\widehat{W}^r$  can be written as a sum of two processes

$$\widehat{W}^r(t) = \widehat{X}^r(t) + \widehat{Y}^r(t), \quad (32)$$

and

$$\begin{aligned} \widehat{X}^r(t) = \widehat{W}^r(0) + \sum_{i \in \mathcal{I}} y_i^r \sum_{j \in \mathcal{J}} \left( \sum_{i' \in \mathcal{B}_j} \widehat{\Phi}_{i'}^{j,r} \left( \bar{S}_j^r(\bar{T}_j^r(t)) \right) \right. \\ \left. + \left( \sum_{i' \in \mathcal{B}_j} P_{i'}^j - B_{ji} \right) \widehat{S}_j^r(\bar{T}_j^r(t)) \right) - r(1 - \rho^r)t, \end{aligned} \quad (33)$$

where

$$\bar{T}_j^r(t) = r^{-2} T_j^r(r^2t) \quad \text{and} \quad \bar{S}_j^r(t) = r^{-2} S_j^r(r^2t).$$

The process  $\widehat{X}^r$  depends on the policy only through the fluid-scaled process  $\bar{T}^r$ . In fact, from Lemma 4.1 of Dai (1995) and (18), it follows that, under any asymptotically efficient policy,  $\bar{T}^r \rightarrow x^*(\cdot)$  almost surely, where  $x^*(t) = x^*t$  and  $x^*$  is the optimal solution to the static planning problem (5)-(9) of the limit network. As a consequence,  $\widehat{X}^r$  converges in distribution to a one-dimensional Brownian motion that is independent of policies.

**Lemma 2.** *Consider a sequence of stochastic processing networks operating under an asymptotically efficient policy. Assume Assumptions 1-4. Then  $\widehat{X}^r \Rightarrow X^*$ , where  $X^*$  is a one-dimensional Brownian motion that starts from the origin, has drift parameter  $\theta$  given in (16), and has variance parameter*

$$\sigma^2 = (y^*)' \left( \sum_{j \in \mathcal{J}} x_j^* \mu_j \sum_{i \in \mathcal{B}_j} \Upsilon^{j,i} \right) y^* + \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} (y_i^*)^2 R_{ij}^2 x_j^* \mu_j \sigma_j^2$$

with  $\Upsilon^{j,i}$ ,  $j \in \mathcal{J}$ ,  $i \in \mathcal{B}_j$ , defined by

$$\Upsilon_{i_1, i_2}^{j,i} = \begin{cases} P_{i, i_1}^j (1 - P_{i, i_2}^j), & \text{if } i_1 = i_2, \\ -P_{i, i_1}^j P_{i, i_2}^j, & \text{if } i_1 \neq i_2. \end{cases}$$

*Proof.* First, Lemma 4.1 of Dai (1995) and (18) implies that  $\bar{T}^r(\cdot) \rightarrow x^*(\cdot)$  almost surely under any asymptotically efficient policy. Then, the result in the lemma follows from (33), the functional central limit theorem for renewal processes (cf. Iglehart and Whitt, 1970), the random time change theorem (cf. Billingsley, 1999, (17.9)), and the continuous mapping theorem (cf. Billingsley, 1999, Theorem 5.1). Deriving the expression for  $\sigma^2$  is straightforward but tedious; the derivation is outlined in Harrison (1988) for multiclass queueing networks, so we will not repeat it here.  $\square$

We define the one-dimensional reflection mapping  $\psi : \mathbb{D}[0, \infty) \rightarrow \mathbb{D}[0, \infty)$  such that for each  $f \in \mathbb{D}[0, \infty)$  with  $f(0) \geq 0$ ,

$$\psi(f)(t) = f(t) - \inf_{0 \leq s \leq t} f(s) \wedge 0.$$

Applying diffusion scaling to Lemma 1, we know that  $\widehat{Y}^r(\cdot)$  is a nonnegative, nondecreasing function, so, from (32) and the well-known minimality of the solution of the one-dimensional Skorohod problem (cf. Bell and Williams, 2001, Proposition B.1),

$$\widehat{W}^r(t) \geq \psi(\widehat{X}^r)(t) \quad \text{for every } t \text{ and every sample path;}$$

namely,  $\psi(\widehat{X}^r)(t)$  is a pathwise lower bound on  $\widehat{W}^r$ . It then follows that

$$\liminf_{r \rightarrow \infty} \mathbb{P}(\widehat{W}^r(t) > w) \geq \limsup_{r \rightarrow \infty} \mathbb{P}(\psi(\widehat{X}^r)(t) > w), \text{ for all } t \text{ and } w.$$

Define

$$W^* = \psi(X^*). \tag{34}$$

Then  $W^*$  is a one-dimensional reflecting Brownian motion associated with  $X^*$ . Because  $\widehat{X}^r \Rightarrow X^*$ , by the continuous mapping theorem, we have

$$\psi(\widehat{X}^r) \Rightarrow W^*.$$

Because  $W^*(t)$  has continuous distribution for each  $t$ , we have

$$\lim_{r \rightarrow \infty} \mathbb{P}(\psi(\widehat{X}^r)(t) > w) = \mathbb{P}(W^*(t) > w).$$

Therefore,

$$\liminf_{r \rightarrow \infty} \mathbb{P}(\widehat{W}^r(t) > w) \geq \mathbb{P}(W^*(t) > w) \text{ for each } t \text{ and } w. \tag{35}$$

So far, we have shown that  $W^*$  is an asymptotic lower bound on the workload processes under asymptotically efficient policies. The following heavy traffic limit theorem ensures that the workload processes under any maximum pressure policy converge to  $W^*$ . This completes the proof of Theorem 2.

**Theorem 4.** Consider a sequence of stochastic processing networks operating under a maximum pressure policy with parameter set  $(\alpha, \beta)$ . Assume Assumptions 1–4 and that the limit network satisfies the EAA condition. Then

$$(\widehat{W}^r, \widehat{Z}^r) \Rightarrow (W^*, Z^{*,\alpha}) \text{ as } r \rightarrow \infty, \quad (36)$$

where  $W^*$  is given by (34) and

$$Z^{*,\alpha} = \zeta^\alpha W^* \quad (37)$$

with  $\zeta^\alpha$  being defined as

$$\zeta_i^\alpha = \frac{y_i^*/\alpha_i}{\sum_{i'} (y_{i'}^*)^2/\alpha_{i'}}. \quad (38)$$

The significance of Theorem 4 will be discussed at the end of this section. There, an outline of its proof will also be presented. The full proof of Theorem 4 will be completed in Section 8. With Theorem 4, we now give the proof of Theorem 3.

*Proof of Theorem 3.* First we note that under any policy, for all  $r$ ,

$$\widehat{H}^r(t) = h \cdot \widehat{Z}^r(t) \geq \min_{i \in \mathcal{I}} h_i/y_i^r \widehat{W}^r(t).$$

Let  $\iota \in \operatorname{argmin}_{i \in \mathcal{I}} h_i/y_i^*$ . From (35), we have, for any asymptotically efficient policy  $\pi$ ,

$$\liminf_{r \rightarrow \infty} \mathbb{P}(\widehat{H}_\pi^r(t) > c) \geq \liminf_{r \rightarrow \infty} \mathbb{P}((h_\iota/y_\iota^r) \widehat{W}_\pi^r(t) > c) \geq \mathbb{P}((h_\iota/y_\iota^*) W^*(t) > c).$$

Now consider a maximum pressure policy  $\pi^*$  having the parameter set  $(\alpha, \beta)$  with  $\beta_i = 0$  for all  $i \in \mathcal{I}$  and

$$\alpha_i = \begin{cases} h_i y_i^* \varepsilon / (\mathbf{I} |h \times y^*|), & i = \iota, \\ 1, & \text{otherwise.} \end{cases}$$

Because  $\sum_{i'} (y_{i'}^*)^2/\alpha_{i'} \geq (y_\iota^*)^2/\alpha_\iota = y_\iota^* \mathbf{I} |h \times y^*|/h_\iota \varepsilon$ , we have  $h_i \zeta_i^\alpha \leq h_\iota \varepsilon / (y_\iota^* \mathbf{I})$  for all  $i \neq \iota$  and  $h_\iota \zeta_\iota^\alpha \leq h_\iota / y_\iota^*$ .

From Theorem 4, we have

$$\widehat{H}_{\pi^*}^r \Rightarrow h \cdot Z^{*,\alpha} \text{ as } r \rightarrow \infty.$$

Because

$$\sum_i h_i Z_i^{*,\alpha}(t) = \sum_i h_i \zeta_i^\alpha W^*(t) \leq (h_\iota / y_\iota^*) W^*(t) (1 + \varepsilon),$$

we have

$$\lim_{r \rightarrow \infty} \mathbb{P}(\widehat{H}_{\pi^*}^r(t) > c) \leq \mathbb{P}((h_\iota / y_\iota) W^*(t) (1 + \varepsilon) > c).$$

Then the inequality (22) follows. □

Theorem 4 is known as a *heavy traffic limit theorem*. It states that the scaled buffer level process and the scaled workload process jointly converge in heavy traffic as displayed in (36), and the limit processes exhibit a form of state space collapse: in diffusion limit, the  $\mathbf{I}$ -dimensional buffer level process is a constant vector multiple of the one-dimensional workload process as displayed in (37).

The key to the proof of Theorem 4 is the following state space collapse result.

**Theorem 5** (State Space Collapse). *Consider a sequence of stochastic processing networks operating under the maximum pressure policy with parameter set  $(\alpha, \beta)$ . Assume Assumptions 1–4 and that the limit network satisfies the EAA condition. Then, for each  $T \geq 0$ , as  $r \rightarrow \infty$ ,*

$$\|\widehat{Z}^r(\cdot) - \zeta^\alpha \widehat{W}^r(\cdot)\|_T \rightarrow 0 \text{ in probability.} \quad (39)$$

Recall that  $\|\cdot\|_T$  is the uniform norm over  $[0, T]$ . (The readers should not confuse the symbols  $T$  and  $T(\cdot)$  with one another. We will always include “ $(\cdot)$ ” when dealing with the cumulative activity level process  $T(\cdot)$ .) Theorem 5 states a form of state space collapse for the diffusion-scaled network process: for large  $r$ , the  $\mathbf{I}$ -dimensional diffusion-scaled buffer level process is essentially a constant vector multiple of the one-dimensional workload process.

The proof of Theorem 5 is lengthy. To prove it, we generalize a framework of Bramson (1998) from the multiclass queueing network setting to the stochastic processing network setting. The framework consists of two steps that we will follow in the next two sections: First, in Section 6, we will show that any fluid model solution for the stochastic processing networks under a maximum pressure policy exhibits some type of state space collapse, which is stated in Theorem 6 in that section. Then, in Section 7, we will follow Bramson’s approach to translate the state space collapse of the fluid model into a state space collapse result under diffusion scaling, and thus prove Theorem 5.

Once we have Theorem 5, we apply a perturbed Skorohod mapping theorem from Williams (1998b) to complete the proof of Theorem 4 in Section 8.

## 6 The Fluid Model

In this section, we first introduce the fluid model of a sequence of stochastic processing networks operating under a maximum pressure policy. We then show that any fluid model solution under a maximum pressure policy exhibits a form of state space collapse.

The fluid model of a sequence of stochastic processing networks that satisfies Assump-

tion 1 is defined by the following equations:

$$\bar{Z}(t) = \bar{Z}(0) + R\bar{T}(t), \quad (40)$$

$$\bar{Z}(t) \geq 0, \quad (41)$$

$$\sum_{j \in \mathcal{J}} A_{kj} (\bar{T}_j(t) - \bar{T}_j(s)) = t - s \text{ for each } 0 \leq s \leq t \text{ and each input processor } k, \quad (42)$$

$$\sum_{j \in \mathcal{J}} A_{kj} (\bar{T}_j(t) - \bar{T}_j(s)) \leq t - s \text{ for each } 0 \leq s \leq t \text{ and each processor } k, \quad (43)$$

$$\bar{T} \text{ is nondecreasing and } \bar{T}(0) = 0. \quad (44)$$

Equations (40)–(44) define the fluid model under *any* given service policy. Any quantity  $(\bar{Z}, \bar{T})$  that satisfies (40)–(44) is a *fluid model solution* to the fluid model that operates under a general service policy. Following its stochastic processing network counterparts, each fluid model solution  $(\bar{Z}, \bar{T})$  has the following interpretations:  $\bar{Z}_j(t)$  the fluid level in buffer  $i$  at time  $t$  and  $\bar{T}_j(t)$  the cumulative amount of activity  $j$  processing time in  $[0, t]$ .

Under a specific service policy, there are additional fluid model equations. For a given parameter set  $(\alpha, \beta)$  with  $\alpha > 0$ , we are to specify the fluid model equation associated with the maximum pressure policy with parameter set  $(\alpha, \beta)$ . To motivate the fluid model equation, we note that for each fluid model solution  $(\bar{Z}, \bar{T})$ , it follows from equations (42)–(43) that  $\bar{T}$ , and hence  $\bar{Z}$ , is Lipschitz continuous. Thus, the solution is absolutely continuous and has derivatives almost sure everywhere with respect to Lebesgue measure on  $[0, \infty)$ . A time  $t > 0$  is said to be a regular point of the fluid model solution if the solution is differentiable at time  $t$ . For a function  $f : \mathbb{R}_+ \rightarrow \mathbb{R}^d$ , where  $d$  is some positive integer, we use  $\dot{f}(t)$  to denote the derivative of  $f$  at time  $t$  when the derivative exists. From (42)–(43), one has  $\dot{\bar{T}}(t) \in \mathcal{A}$  at each regular time  $t$ . Thus,

$$R\dot{\bar{T}}(t) \cdot (\alpha \times \bar{Z}(t)) \leq \max_{a \in \mathcal{A}} Ra \cdot (\alpha \times \bar{Z}(t)) = \max_{a \in \mathcal{E}} Ra \cdot (\alpha \times \bar{Z}(t)).$$

Assume that the limit network satisfies the EAA condition. The fluid model equation associated with the maximum pressure policy with parameter set  $(\alpha, \beta)$  takes the following form:

$$R\dot{\bar{T}}(t) \cdot (\alpha \times \bar{Z}(t)) = \max_{a \in \mathcal{E}} Ra \cdot (\alpha \times \bar{Z}(t)) \quad (45)$$

for each regular time  $t$ .

Each fluid model equation will be justified through a fluid limit procedure. Two types of fluid limits are considered in this paper. One will be introduced in Section 7 and the other in Appendix B. They both satisfy the fluid model equations (40)–(45). Note that the fluid model equation (45) does not depend on the parameter  $\beta$ . This is because the parameter disappears in fluid limits. Equation (45) says that, under the maximum pressure policy with parameter set  $(\alpha, \beta)$ , the instantaneous activity allocation  $\dot{\bar{T}}(t)$  in the fluid model maximizes the network pressure,  $Ra \cdot (\alpha \times \bar{Z}(t))$ , at each regular time  $t$ . Any fluid model solution that satisfies fluid model equations (40)–(45) is called a *fluid model solution under the maximum pressure policy* with parameter set  $(\alpha, \beta)$ .

**Theorem 6.** Consider a sequence of stochastic processing networks where Assumptions 1 and 2 hold. For any parameter set  $(\alpha, \beta)$  with  $\alpha > 0$ , there exists some finite  $\tau_0 > 0$ , which depends on just  $\alpha$ ,  $\mathbf{I}$ ,  $R$ , and  $A$ , such that, for any fluid model solution  $(\bar{Z}, \bar{T})$  under the maximum pressure policy with parameter set  $(\alpha, \beta)$ , which satisfies equations (40)–(45), if  $|\bar{Z}(0)| \leq 1$ , then

$$|\bar{Z}(t) - \zeta^\alpha \bar{W}(t)| = 0, \quad \text{for all } t \geq \tau_0, \quad (46)$$

where  $\bar{W} = y^* \cdot \bar{Z}$  is the workload process of the fluid model and  $\zeta^\alpha$  is given by (38). Furthermore, if

$$|\bar{Z}(\tau_1) - \zeta^\alpha \bar{W}(\tau_1)| = 0, \quad \text{for some } \tau_1 \geq 0,$$

then

$$|\bar{Z}(t) - \zeta^\alpha \bar{W}(t)| = 0, \quad \text{for all } t \geq \tau_1.$$

Theorem 6 says that the fluid model under the maximum pressure policy exhibits a form of state space collapse: after some finite time  $\tau_0$ , the  $\mathbf{I}$ -dimensional buffer level process  $\bar{Z}$  equals a constant vector multiple of the one-dimensional workload process  $\bar{W}$ ; if this happens at time  $\tau_1$ , it happens at all times after  $\tau_1$ . In particular, if  $\bar{Z}(0) = \zeta^\alpha \bar{W}(0)$ , then  $\bar{Z}(t) = \zeta^\alpha \bar{W}(t)$  for all  $t \geq 0$ .

The rest of this section is to prove Theorem 6. We first define

$$\bar{Z}^*(t) = \zeta^\alpha \bar{W}(t),$$

and we shall prove  $\bar{Z}(t) - \bar{Z}^*(t) = 0$  for  $t$  large enough. The following lemmas will be used in the proof.

**Lemma 3.** For any  $t \geq 0$ ,

$$(\bar{Z}(t) - \bar{Z}^*(t)) \cdot y^* = 0,$$

and for each regular time  $t$ ,

$$(\bar{Z}(t) - \bar{Z}^*(t)) \cdot (\alpha \times \dot{\bar{Z}}^*(t)) = 0.$$

*Proof.*

$$\bar{Z}^*(t) \cdot y^* = \bar{W}(t) \zeta \cdot y^* = \bar{W}(t) = \bar{Z}(t) \cdot y^*.$$

Because  $\dot{\bar{Z}}^*(t) = \zeta^\alpha \dot{\bar{W}}(t)$ , we have

$$\alpha \times \dot{\bar{Z}}^*(t) = \dot{\bar{W}}(t) y^* / \left( \sum_{i \in \mathcal{I}} (y_i^*)^2 / \alpha_i \right).$$

Therefore,

$$(\bar{Z}(t) - \bar{Z}^*(t)) \cdot (\alpha \times \dot{\bar{Z}}^*(t)) = \dot{\bar{W}}(t) (\bar{Z}(t) - \bar{Z}^*(t)) \cdot y^* / \left( \sum_{i \in \mathcal{I}} (y_i^*)^2 / \alpha_i \right) = 0.$$

□

The following lemma will be used repeatedly later. It follows directly from Lemma 12 to be presented in Appendix A.

**Lemma 4.** *Suppose  $(\hat{y}, \hat{z})$  is the unique optimal solution to the dual problem (10)–(14) with objective value  $\rho$ . Then  $\hat{y}$  is the unique  $\mathbf{I}$ -dimensional vector that satisfies*

$$\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} \hat{y}_i R_{ij} a_j = 1, \quad (47)$$

and

$$\max_{a \in \mathcal{A}} \hat{y} \cdot Ra = 1 - \rho. \quad (48)$$

Define  $V = \{Ra : a \in \mathcal{A}\}$ . Recall that  $\mathcal{A}$  is the set of all possible allocations and the vector  $Ra$  is the average rate at which material is consumed from all buffers under allocation  $a$ , so  $V$  is the set of all possible flow rates out of buffers in the limit network. It is obvious that  $V$  is a polytope containing the origin because  $Rx^* = 0$ , where, as before,  $x^*$  is the optimal solution to the static planning problem of the limit network. Furthermore, from Lemma 4 and the fact that  $\rho^* = 1$  for the limit network, we have

$$\max_{v \in V} y^* \cdot v = 0. \quad (49)$$

Then,  $V$  is in a half space separated by the  $(\mathbf{I} - 1)$  dimensional hyperplane

$$V^\circ = \{v \in \mathbb{R}^I : y^* \cdot v = 0\}. \quad (50)$$

The hyperplane  $V^\circ$  is orthogonal to  $y^*$  and passes through the origin, so  $V \cap V^\circ$  is not empty. Furthermore, in the hyperplane  $V^\circ$ , there exists some  $(\mathbf{I} - 1)$  dimensional neighborhood of the origin that is a subset of  $V$ . This is stated in the following lemma.

**Lemma 5.** *There exists some  $\delta > 0$  such that  $\{v \in V^\circ : \|v\| \leq \delta\} \subset V$ .*

*Proof.* First, if the statement is not true, then we can find a  $v_0 \in V^\circ$  such that  $\kappa v_0 \notin V$  for all  $0 < \kappa \leq 1$  because of the convexity of  $V$ . Denote  $V_0 = \{\kappa v_0, 0 < \kappa < 1\}$ . Because any  $v$  in  $V_0$  is not in  $V$ ,  $V_0 \cap V = \emptyset$ . It is easy to see that  $V_0$  is open and convex. Therefore there exists a hyperplane separating  $V$  and  $V_0$  (cf. Rudin, 1991, Theorem 3.4). In other words, there exists a vector  $\hat{y}$  and a constant  $b$  such that  $\hat{y} \cdot v \leq b$  for all  $v \in V$  and  $\hat{y} \cdot v > b$  for all  $v \in V_0$ . We notice that  $b$  must be zero. To see this, first we have  $b \geq 0$  because the origin is in  $V$ . Moreover, for any  $\epsilon > 0$ , we can choose  $\kappa$  arbitrarily small such that  $\kappa \hat{y} \cdot v_0 < \epsilon$ . Because  $\kappa v_0 \in V_0$ , we have  $b < \kappa \hat{y} \cdot v_0 < \epsilon$ . This implies  $b = 0$ , therefore the origin is in the separating hyperplane and  $\max_{a \in \mathcal{A}} \hat{y} \cdot Ra = 0$ . Obviously,  $\hat{y} \neq y^*$  because  $y^* \cdot v = 0 > \hat{y} \cdot v$  for  $v \in V_0$ . We consider two cases:

**Case 1:**  $\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} \hat{y}_i R_{ij} a_j > 0$ . For this case, without loss of generality, we select  $\hat{y}$  such that  $\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} \hat{y}_i R_{ij} a_j = 1$ . Then  $\hat{y}$  satisfies both (47) and (48) with  $\rho = 1$ . On the other hand, from Lemma 4,  $y^*$  is the unique vector that satisfies both (47) and (48). This is a contradiction.



**Case 2:**  $\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} \hat{y}_i R_{ij} a_j = 0$ . For this case, one can verify that  $y^* + \hat{y}$  satisfies both (47) and (48) with  $\rho = 1$ . This is also a contradiction. □

Let  $V^* = V \cap V^o$ . Then from (49) and (50),  $V^* = \operatorname{argmax}_{v \in V} y^* \cdot v$ ; namely,  $V^*$  is the set of all possible flow rates that maximize  $y^* \cdot v$ .

*Proof of Theorem 6.* We consider the Lyapunov function

$$f(t) = \left( \alpha \times (\bar{Z}(t) - \bar{Z}^*(t)) \right) \cdot (\bar{Z}(t) - \bar{Z}^*(t)).$$

Let  $v(t) = R\dot{T}(t)$  denote the net flow rate out of the system at time  $t$  (*total departure rate minus total arrival rate*). Then  $\dot{\bar{Z}}(t) = -v(t)$ , and we have

$$\dot{f}(t) = 2 \left( \alpha \times (\bar{Z}(t) - \bar{Z}^*(t)) \right) \cdot \left( \dot{\bar{Z}}(t) - \dot{\bar{Z}}^*(t) \right) = 2 \left( \alpha \times (\bar{Z}(t) - \bar{Z}^*(t)) \right) \cdot (-v(t)). \quad (51)$$

The second equality in (51) follows from Lemma 3.

Because  $\dot{T}(t) \in \mathcal{A}$ , we have  $v(t) \in V$ . This implies that  $y^* \cdot v(t) \leq 0$  and hence  $(\alpha \times \bar{Z}^*(t)) \cdot v(t) \leq 0$ . Furthermore, under the maximum pressure policy, it follows from the fluid model equation (45) that  $v(t)$  satisfies

$$(\alpha \times \bar{Z}(t)) \cdot v(t) = \max_{v \in V} (\alpha \times \bar{Z}(t)) \cdot v.$$

Therefore, the last term in (51) is bounded from above as follows:

$$2 \left( \alpha \times (\bar{Z}(t) - \bar{Z}^*(t)) \right) \cdot (-v(t)) \leq -2 (\alpha \times \bar{Z}(t)) \cdot v(t) = -2 \max_{v \in V} (\alpha \times \bar{Z}(t)) \cdot v. \quad (52)$$

Since  $V^* \subset V$ , we have

$$\max_{v \in V} (\alpha \times \bar{Z}(t)) \cdot v \geq \max_{v \in V^*} (\alpha \times \bar{Z}(t)) \cdot v = \max_{v \in V^*} (\alpha \times (\bar{Z}(t) - \bar{Z}^*(t))) \cdot v. \quad (53)$$

The second equality in (53) holds because  $y^* \cdot v = 0$  for all  $v \in V^*$  and

$$\alpha \times \bar{Z}^*(t) = \bar{W}(t) y^* / \left( \sum_{i \in \mathcal{I}} (y_i^*)^2 / \alpha_i \right).$$

If  $f(t) > 0$ , let

$$v^* = \frac{\delta (\bar{Z}(t) - \bar{Z}^*(t))}{\|\bar{Z}(t) - \bar{Z}^*(t)\|}.$$

Then  $\|v^*\| = \delta$  and  $y^* \cdot v^* = 0$ . It follows from Lemma 5 that  $v^* \in V^*$ . Therefore

$$\max_{v \in V^*} \left( \alpha \times (\bar{Z}(t) - \bar{Z}^*(t)) \right) \cdot v \geq \left( \alpha \times (\bar{Z}(t) - \bar{Z}^*(t)) \right) \cdot v^* = \delta f(t) / \|\bar{Z}(t) - \bar{Z}^*(t)\|. \quad (54)$$

Combining (51)–(54), we have

$$\dot{f}(t) \leq -2\delta f(t)/\|\bar{Z}(t) - \bar{Z}^*(t)\| = -2\delta \sqrt{\min_{i \in \mathcal{I}} \alpha_i} \sqrt{f(t)}. \quad (55)$$

Therefore,  $f(t) = 0$  for  $t \geq \sqrt{f(0)}/\delta$ . Set  $\tau_0 = \sqrt{\mathbf{I}}/(\delta\sqrt{\min_{i \in \mathcal{I}} \alpha_i})$ . Then  $f(t) = 0$  for  $t \geq \tau_0$ , because

$$f(0) = (\bar{Z}(0) - \bar{Z}^*(0)) \cdot (\bar{Z}(0) - \bar{Z}^*(0)) = (\bar{Z}(0) - \bar{Z}^*(0)) \cdot Z(0) \leq \|\bar{Z}(0)\|^2 \leq \mathbf{I}.$$

Here  $\tau_0$  depends only on  $\alpha$ ,  $R$ ,  $A$ , and  $\mathbf{I}$ , because the set  $V$  is completely determined by  $R$  and  $A$  and so is  $\delta$ .

Equation (55) also implies that for any  $\tau_1$ , if  $f(\tau_1) = 0$  then  $f(t) = 0$  for all  $t \geq \tau_1$ .  $\square$

## 7 State Space Collapse

In this section, we translate the state space collapse result of the fluid model into a state space collapse result under the diffusion scaling, thus proving Theorem 5. We apply Bramson's approach (Bramson, 1998) to prove that for each  $T \geq 0$ , as  $r \rightarrow \infty$ ,

$$\|\widehat{Z}^r(\cdot) - \zeta^\alpha \widehat{W}^r(\cdot)\|_T \rightarrow 0 \text{ in probability.}$$

In Bramson's approach, the following fluid scaling plays an important role in connecting Theorem 6 with Theorem 5: For each  $r = 1, 2, \dots$ , and  $m = 0, 1, \dots$ ,

$$\begin{aligned} S_j^{r,m}(t) &= \frac{1}{\xi_{r,m}} (S_j^r(rm + \xi_{r,m}t) - S_j^r(rm)), \quad \text{for each } j \in \mathcal{J}, \\ \Phi_i^{j,r,m}(t) &= \frac{1}{\xi_{r,m}} \left( \Phi_i^j(S_j^r(rm) + \lfloor \xi_{r,m}t \rfloor) - \Phi_i^j(S_j^r(rm)) \right), \quad \text{for each } j \in \mathcal{J}, i \in \mathcal{B}_j, \\ T_j^{r,m}(t) &= \frac{1}{\xi_{r,m}} (T_j^r(rm + \xi_{r,m}t) - T_j^r(rm)), \quad \text{for each } j \in \mathcal{J}, \\ Z_i^{r,m}(t) &= \frac{1}{\xi_{r,m}} Z_i^r(rm + \xi_{r,m}t), \quad \text{for each } i \in \mathcal{I}, \end{aligned}$$

where  $\xi_{r,m} = \lfloor Z^r(rm) \rfloor \vee r$ . Recall that  $\lfloor t \rfloor$  denotes the integer part of  $t$ .

Here scaling the processes by  $\xi_{r,m}$  ensures  $|Z^{r,m}(0)| \leq 1$ , which is needed for compactness reasons. Using index  $(r, m)$  allows the time scale to expand; we will examine the processes over  $[0, L]$  for  $m = 0, 1, \dots, \lceil rT \rceil - 1$ , where  $L > 0$  and  $T > 0$  are fixed. Thus, the diffusion-scaled time  $[0, T]$  is covered by  $\lceil rT \rceil$  fluid scaled time pieces, each with length  $L$ . Here  $\lceil t \rceil$  denotes the smallest integer greater than or equal to  $t$ . For future references, we assume that  $L > \tau_0 \vee 1$ , where  $\tau_0$  is the time when state space collapse occurs in the corresponding fluid model as determined in Theorem 6.

We outline the proof of Theorem 5 as follows. First, in Proposition 1 below, we give a probability estimate on the upper bound of the fluctuation of the stochastic network

processes  $\mathbb{X}^{r,m}(\cdot) = (T^{r,m}(\cdot), Z^{r,m}(\cdot))$ . The estimates on the service processes  $S_j^{r,m}(\cdot)$  and the routing processes  $\Phi_i^{j,r,m}(\cdot)$  are also given. From Proposition 1, a so-called “good” set  $\mathcal{G}^r$  of sample paths can be defined, where the processes  $\mathbb{X}^{r,m}$  perform nicely for  $r$  large enough. On this “good” set, for large enough  $r$ , the processes  $\mathbb{X}^{r,m}$  can be uniformly approximated by so-called *Lipschitz cluster points*. These cluster points will be shown to be fluid model solutions under the maximum pressure policy. Since the state space collapse result holds for the fluid model under the maximum pressure policy by Theorem 6, the network processes  $\mathbb{X}^{r,m}$  asymptotically have the state space collapse. The latter result will be translated into the state space collapse result for diffusion-scaled processes  $\hat{\mathbb{X}}^r$  as  $r$  approaches infinity.

Note that in Theorem 6 the state space collapse of the fluid model does not happen instantaneously after time 0 if the initial state does not exhibit a state space collapse. The fluid-scaled processes  $\mathbb{X}^{r,m}$  start from time  $rm$  in the original network processes. Hence, for  $m \geq 1$ ,  $\mathbb{X}^{r,m}$  do not automatically have the state space collapse at the initial point. Such collapse can only be expected to occur in the interval  $[\tau_0, L]$ . However, for  $m = 0$ , the state space collapse happens at time 0 because of the initial condition (24), therefore the collapse remains in the whole interval  $[0, L]$ . For this reason, we separate the proof into two parts according to the two intervals in the diffusion-scaled time:  $[0, L\xi_{r,0}/r^2]$  and  $[\tau_0\xi_{r,0}/r^2, T]$ . Note that the two intervals overlap because of our assumption on the choice of  $L$ .

Propositions 2–5 develop a version of state space collapse for the interval  $[\tau_0\xi_{r,0}/r^2, T]$ . Proposition 2 shows that, on  $\mathcal{G}^r$ , the scaled processes  $\mathbb{X}^{r,m}(\cdot)$  are uniformly close to Lipschitz cluster points for large  $r$ . Proposition 3 shows that the above cluster points are solutions to the fluid model equations. In Proposition 4, we use Propositions 2 and 3 and Theorem 6 to prove the state space collapse of the fluid scaled processes  $\mathbb{X}^{r,m}(\cdot)$  in the time interval  $[\tau_0, L]$  on the “good” set  $\mathcal{G}^r$ . The result in Proposition 4 is then translated into diffusion scaling in Proposition 5 to give a version of state space collapse for the diffusion-scaled process  $\hat{\mathbb{X}}^r(t)$  in the interval  $[\tau_0\xi_{r,0}/r^2, T]$ .

In Propositions 6–8, we show that the state space collapse occurs for the diffusion-scaled process  $\hat{\mathbb{X}}^r(t)$  in the interval  $[0, L\xi_{r,0}/r^2]$ . The basic idea is the same as the one described in the preceding paragraph except that now we only consider the scaled processes with  $m = 0$ . The corresponding network processes start from time 0, and by assuming the state space collapse happens at time 0 in Assumption 4, we have a stronger result for these types of processes: the state space collapse holds during the whole time interval  $[0, L]$  instead of just in  $[\tau_0, L]$ . In fact, the scaled processes  $\mathbb{X}^{r,0}(\cdot)$  are proved to be uniformly close to some cluster points for which the state space collapse starts at time 0. These facts are stated in Propositions 6 and 7. In Proposition 8, we summarize the state space collapse result for the fluid-scaled process  $\mathbb{X}^{r,0}(\cdot)$  on  $[0, L]$  and translate it into a result for the diffusion-scaled process in  $[0, L\xi_{r,0}/r^2]$ .

The results to be obtained in Propositions 5 and 8 are actually *multiplicative state space collapse*, as called in Bramson (1998). To obtain the state space collapse result stated in Theorem 5, we will prove that  $\xi_{r,m}/r$  are stochastically bounded at the end of this section.

## 7.1 Probability estimates

In this section, we give probability estimates on the service processes  $S_j^{r,m}(\cdot)$ , the routing processes  $\Phi_i^{j,r,m}(\cdot)$ , and the upper bound of the fluctuation of the stochastic network processes  $\mathbb{X}^{r,m}(\cdot)$ .

**Proposition 1.** *Consider a sequence of stochastic processing networks where the moment assumption, Assumption 3, is assumed. Fix  $\epsilon > 0$ ,  $L$  and  $T$ . Then, for large enough  $r$ ,*

$$\mathbb{P}(\max_{m < rT} \|S_j^{r,m}(T_j^{r,m}(t)) - \mu_j^r T_j^{r,m}(t)\|_L > \epsilon) \leq \epsilon, \text{ for each } j \in \mathcal{J}, \quad (56)$$

$$\mathbb{P}(\max_{m < rT} \|\Phi_i^{j,r,m}(S_j^{r,m}(T_j^{r,m}(t))) - P_i^j \mu_j^r(T_j^{r,m}(t))\|_L > \epsilon) \leq \epsilon, \text{ for each } j \in \mathcal{J} \text{ and } i \in \mathcal{B}_j, \quad (57)$$

$$\mathbb{P}(\sup_{0 \leq t_1 \leq t_2 \leq L} |\mathbb{X}^{r,m}(t_2) - \mathbb{X}^{r,m}(t_1)| > N|t_2 - t_1| + \epsilon \text{ for some } m < rT) \leq \epsilon, \quad (58)$$

where  $N$  is some constant that depends on just the bounds of  $\{R^r\}$ .

The following lemma is essential to the proof of Proposition 1.

**Lemma 6.** *Assume that the moment assumption, Assumption 3, holds. Then for given  $T$ , and each  $j \in \mathcal{J}$*

$$u_j^{r,T,max}/r \rightarrow 0 \text{ as } r \rightarrow \infty \text{ with probability 1,} \quad (59)$$

where  $u_j^{r,T,max} = \max\{u_j(\ell) : 1 \leq \ell \leq S_j^r(r^2T) + 1\}$ . Furthermore, for any given  $\epsilon > 0$ ,

$$\mathbb{P}(\|\Phi_i^j(\ell) - P_i^j \ell\|_n \geq \epsilon n) \leq \epsilon/n, \text{ for each } j \in \mathcal{J} \text{ and } i \in \mathcal{B}_j, \text{ and large enough } n, \quad (60)$$

and for large enough  $t$ ,

$$\mathbb{P}(\|S_j^r(\tau) - \mu_j^r \tau\|_t \geq \epsilon t) \leq \epsilon/t, \text{ for all } j \in \mathcal{J}, \text{ and all } r. \quad (61)$$

We delay the proof of Lemma 6 to Appendix B, and we now prove Proposition 1.

*Proof of Proposition 1.* The proof here essentially follows the same reasoning as in Propositions 5.1 and 5.2 of Bramson (1998). We first investigate the processes with index  $m$ , and then multiply the error bounds by the number of processes in each case,  $\lceil rT \rceil$ . We start with (56). From (59), we have, for large enough  $r$ ,

$$\mathbb{P}(u_j^{r,T,max}/r \geq \epsilon) \leq \epsilon/2. \quad (62)$$

Denote  $\mathcal{M}^r$  to be the complement of the events in (62). Then, for large enough  $r$ ,

$$\mathbb{P}(\mathcal{M}^r) \geq 1 - \epsilon/2. \quad (63)$$

Let  $\tau$  be the time that the first activity  $j$  service completion occurs after time  $rm$ . We consider two cases. First suppose that  $\tau > rm + \xi_{r,m} T_j^{r,m}(t)$ . Then  $\xi_{r,m} T_j^{r,m}(t) < \tau - rm \leq u_j^{r,T,max}/\mu_j^r$  and  $S_j^r(rm + \xi_{r,m} T_j^{r,m}(t)) = S_j^r(rm)$ . Thus,

$$\|S_j^r(rm + \xi_{r,m} T_j^{r,m}(t)) - S_j^r(rm) - \mu_j^r \xi_{r,m} T_j^{r,m}(t)\|_L \leq u_j^{r,T,max}.$$

Or if  $\tau \leq rm + \xi_{r,m}T_j^{r,m}(t)$ , then by restarting the process  $S_j^r$  at time  $\tau$ , we have, by (61), that for large enough  $r$ ,

$$\mathbb{P}\left(\|S_j^r(rm + \xi_{r,m}T_j^{r,m}(t)) - S_j^r(\tau) - \mu_j^r(rm + \xi_{r,m}T_j^{r,m}(t) - \tau)\|_L \geq \epsilon L \xi_{r,m}\right) \leq \epsilon/Lr;$$

we use the fact that  $T_j^{r,m}(t) \leq t$  and  $\xi_{r,m} \geq r$ . Because  $S_j^r(\tau) = S_j^r(rm) + 1$  and  $\tau - rm \leq u_j^{r,T,max}/\mu_j^r$ , we have

$$\mathbb{P}\left(\|S_j^r(rm + \xi_{r,m}T_j^{r,m}(t)) - S_j^r(rm) - \mu_j^r \xi_{r,m} T_j^{r,m}(t)\|_L \geq |1 - u_j^{r,T,max}| + \epsilon L \xi_{r,m}\right) \leq \epsilon/Lr.$$

In both cases, we have

$$\mathbb{P}(\|S_j^{r,m}(T_j^{r,m}(t)) - \mu_j^r T_j^{r,m}(t)\|_L \geq 2\epsilon L \mid \mathcal{M}^r) \leq \epsilon/Lr, \text{ for large enough } r. \quad (64)$$

Therefore, multiplying the lower bound by  $\lceil rT \rceil$ , we have

$$\mathbb{P}(\max_{m \leq rT} \|S_j^{r,m}(T_j^{r,m}(t)) - \mu_j^r T_j^{r,m}(t)\|_L \geq 2\epsilon L \mid \mathcal{M}^r) \leq \epsilon T.$$

Then enlarging  $\epsilon$  by a factor of  $2(L \vee T)$ , we have

$$\mathbb{P}(\max_{m \leq rT} \|S_j^{r,m}(T_j^{r,m}(t)) - \mu_j^r T_j^{r,m}(t)\|_L \geq \epsilon \mid \mathcal{M}^r) \leq \epsilon/2.$$

This, together with (63), implies (56).

Let  $\mu^{max}$  be an upper bound on  $|\mu^r|$ . Then from (64), for large enough  $r$ ,

$$\mathbb{P}\left(S_j^{r,m}(T_j^{r,m}(L)) \geq (\mu^{max} + 2)L \mid \mathcal{M}^r\right) \leq \epsilon/Lr. \quad (65)$$

From (60), we have for each  $i \in \mathcal{B}_j$ ,

$$\begin{aligned} \mathbb{P}\left(\|\Phi_i^{j,r,m}\left(S_j^{r,m}(T_j^{r,m}(t))\right) - P_i^j S_j^{r,m}(T_j^{r,m}(t))\|_L > \epsilon(\mu^{max} + 2)L \mid \mathcal{M}^r\right) \\ \leq \epsilon/(\mu^{max} + 2)Lr + \epsilon/Lr. \end{aligned} \quad (66)$$

Let  $P^{max} = \max_{j \in \mathcal{J}, i \in \mathcal{B}_j} |P_i^j|$ . Then from (64), we have

$$\mathbb{P}(\|P_i^j S_j^{r,m}(T_j^{r,m}(t)) - P_i^j \mu_j^r T_j^{r,m}(t)\|_L \geq 2P^{max} \epsilon L \mid \mathcal{M}^r) \leq \epsilon/Lr. \quad (67)$$

It follows from (66) and (67) that

$$\mathbb{P}\left(\|\Phi_i^{j,r,m}\left(S_j^{r,m}(T_j^{r,m}(t))\right) - P_i^j \mu_j^r(T_j^{r,m}(t))\|_L > \epsilon(2P^{max} + \mu^{max} + 2)L \mid \mathcal{M}^r\right) \leq 5\epsilon/2Lr.$$

Enlarging  $\epsilon$  by a factor of  $(2P^{max} + \mu^{max} + 2)L \vee 5T$  and multiplying the error bound by  $\lceil rT \rceil$ , one gets

$$\mathbb{P}\left(\max_{m \leq rT} \|\Phi_i^{j,r,m}\left(S_j^{r,m}(T_j^{r,m}(t))\right) - P_i^j \mu_j^r(T_j^{r,m}(t))\|_L > \epsilon \mid \mathcal{M}^r\right) \leq \epsilon/2,$$

and then (57) follows.

Now we are going to show (58). First, it is easy to see that, for each  $j \in \mathcal{J}$  and each  $r$ ,

$$T_j^r(t) - T_j^r(s) \leq t - s \text{ for } 0 \leq s \leq t$$

along any sample path. Therefore, the bounds in (58) on components  $T_j$  are obtained with  $N = 1$ . For components  $Z_i, i \in \mathcal{I}$ , scaling (27) and applying (56) and (57) gives

$$\begin{aligned} & \mathbb{P} \left( \sup_{0 \leq t_1 \leq t_2 \leq L} |Z_i^{r,m}(t_2) - Z_i^{r,m}(t_1)| \right. \\ & \quad \left. > R_{ij}^r |T_j^{r,m}(t_2) - T_j^{r,m}(t_1)| + 2\mathbf{J}(\mathbf{I} + 2)\epsilon \text{ for some } m \leq rT \right) \leq 2\mathbf{J}(\mathbf{I} + 2)\epsilon. \end{aligned} \quad (68)$$

Then, enlarging  $\epsilon$  in (68) by a factor of  $2\mathbf{J}(\mathbf{I} + 2)$  and setting  $N = 1 \vee \sup_r |R^r|$ , one gets (58).  $\square$

Now for each large enough  $r$ , we let  $\epsilon(r)$  be the smallest  $\epsilon$  such that (56)-(58) are satisfied. By Proposition 1, it is easy to see that

$$\lim_{r \rightarrow \infty} \epsilon(r) = 0.$$

With  $\epsilon(r)$  replacing  $\epsilon$ , we call each of the events in (56)-(58) a “bad” event. We let  $\mathcal{G}^r$  denote the intersection of the complements of these “bad” events, and it is referred to as a “good” event in the rest of this section. Obviously

$$\lim_{r \rightarrow \infty} \mathbb{P}(\mathcal{G}^r) = 1.$$

## 7.2 State space collapse on $[\xi_{r,0}\tau_0/r^2, T]$

Again, let  $L > 1 \vee \tau_0$  and  $T > 0$  be fixed. We divide the diffusion-scaled time interval  $[0, T]$  into two overlapping intervals:  $[0, \xi_{r,0}L/r^2]$  and  $[\xi_{r,0}\tau_0/r^2, T]$ . In this section, we show a state space collapse result on the time interval  $[\xi_{r,0}\tau_0/r^2, T]$ . The state space collapse on  $[0, \xi_{r,0}L/r^2]$  will be presented in the next section.

In order to connect the fluid-scaled processes with the fluid model, we first introduce the notion of *cluster point*. Let  $F = \mathbb{D}^d[0, L]$  be the space of right continuous functions with left limits from  $[0, L]$  to  $\mathbb{R}^{\mathbf{I}+\mathbf{J}}$ , where  $d = \mathbf{I} + \mathbf{J}$ . Let  $\mathcal{F} = \{F_r\}$  be a sequence of subsets of  $F$ . A point  $f$  in  $F$  is said to be a *cluster point* of  $\mathcal{F}$  if for each  $\epsilon > 0$  and  $r_0 > 0$ , there exist  $r \geq r_0$  and  $g \in F_r$  such that  $\|f - g\|_L < \epsilon$ . The sequence  $\{F_r\}$  is said to be *asymptotically Lipschitz* if there exist a constant  $N > 0$  and a sequence of positive numbers  $\{\epsilon(r)\}$  with  $\epsilon(r) \rightarrow 0$  as  $r \rightarrow \infty$  such that for each  $r$  all elements  $f \in F_r$  satisfy both

$$|f(0)| \leq 1 \quad (69)$$

and

$$|f(t_2) - f(t_1)| \leq N|t_2 - t_1| + \epsilon(r) \text{ for all } t_1, t_2 \in [0, L]. \quad (70)$$

Let  $F'$  denote those  $f \in F$  satisfying both (69) and

$$|f(t_2) - f(t_1)| \leq N|t_2 - t_1| \text{ for all } t_1, t_2 \in [0, L]. \quad (71)$$

The following lemma is due to Bramson (1998). We state it here for completeness.

**Lemma 7** (Bramson (1998), Proposition 4.1). *Assume that  $\mathcal{F}$  is asymptotically Lipschitz. For each  $\epsilon > 0$ , there exists an  $r_0$ , so that for each  $r \geq r_0$  and  $g \in F_r$ , one has  $\|f - g\|_L < \epsilon$  for some cluster point  $f$  of  $\mathcal{F}$  with  $f \in F'$ .*

Lemma 7 says that the cluster points are “rich”: for large  $r$ , all elements in  $F_r$  can be uniformly approximated by cluster points.

We set

$$F_g^r = \{\mathbb{X}^{r,m}(\cdot, \omega), \quad m < rT, \omega \in \mathcal{G}^r\} \text{ for each } r$$

and

$$\mathcal{F}_g = \{F_g^r\}. \quad (72)$$

From the choice of our fluid scale  $\xi_{r,m}$  and the definition of  $T^{r,m}(\cdot)$ , it follows that

$$|\mathbb{X}^{r,m}(0)| \leq 1.$$

It follows from (58) in Proposition 1 that the sequence of sets of scaled stochastic processing network processes  $\mathbb{X}^{r,m}(\cdot)$  is asymptotically Lipschitz. Lemma 7 immediately implies the following proposition which says that for large  $r$ ,  $\mathbb{X}^{r,m}(\cdot)$  are uniformly close to cluster points that are Lipschitz continuous.

**Proposition 2.** *Fix  $\epsilon > 0$ ,  $L$  and  $T$ , and choose  $r$  large enough. Then, for  $\omega \in \mathcal{G}^r$  and any  $m < rT$ ,*

$$\|\mathbb{X}^{r,m}(\cdot, \omega) - \tilde{\mathbb{X}}(\cdot)\|_L \leq \epsilon$$

*for some cluster point  $\tilde{\mathbb{X}}(\cdot)$  of  $\mathcal{F}_g$  with  $\tilde{\mathbb{X}}(\cdot) \in F'$ .*

The next proposition says that if the stochastic processing networks operate under a maximum pressure policy, then each cluster point of  $\mathcal{F}_g$  satisfies fluid model equations (40)–(45), and thus is a fluid model solution to the fluid model operating under the maximum pressure policy.

**Proposition 3.** *Consider a sequence of stochastic processing networks operating under a maximum pressure policy. Assume Assumptions 1 and 3 and that the limit network satisfies the EAA condition. Fix  $L > 0$  and  $T > 0$ . Then all cluster points of  $\mathcal{F}_g$  are solutions to the fluid model equations (40)–(45) on  $[0, L]$ .*

*Proof.* The idea is to approximate each cluster point of  $\mathcal{F}_g$  with some  $\mathbb{X}^{r,m}$  on  $\mathcal{G}^r$  and show that the equations (40)–(45) are asymptotically satisfied by  $\mathbb{X}^{r,m}$ . We will only demonstrate equations (40) and (45); the others are quite straightforward and can be verified similarly.

Let  $\tilde{\mathbf{X}} = (\tilde{Z}, \tilde{T})$  be a cluster point. We first verify equation (40). For any  $\epsilon > 0$  and all  $r$  that are large enough, there are  $\omega \in \mathcal{G}^r$  and  $m < rT$  such that

$$\|S_j^{r,m}(T_j^{r,m}(t)) - \mu_j^r T_j^{r,m}(t)\|_L \leq \epsilon \text{ for each } j \in \mathcal{J}, \quad (73)$$

$$\|\Phi_i^{j,r,m}(S_j^{r,m}(T_j^{r,m}(t))) - P_i^j \mu_j^r T_j^{r,m}(t)\|_L \leq \epsilon, \text{ for each } j \in \mathcal{J}, i \in \mathcal{B}_j, \quad (74)$$

$$\|\tilde{Z}(\cdot) - Z^{r,m}(\cdot)\|_L \leq \epsilon, \quad (75)$$

$$\|\tilde{T}(\cdot) - T^{r,m}(\cdot)\|_L \leq \epsilon, \quad (76)$$

$$|R^r - R| \leq \epsilon. \quad (77)$$

Scaling (27) and plugging in the bounds in (73) and (74), we have

$$|Z^{r,m}(t) - Z^{r,m}(0) - R^r T^{r,m}(t)| \leq 2\epsilon \text{ for all } t \in [0, L].$$

From (76) and (77), we have, for each  $t \in [0, L]$ ,

$$|R^r T^{r,m}(t) - R\tilde{T}(t)| \leq |R^r(T^{r,m}(t) - \tilde{T}(t))| + |(R^r - R)\tilde{T}(t)| \leq N\mathbf{J}\epsilon + \epsilon N L \mathbf{J}.$$

Recall that  $N \geq \sup_r |R^r|$ . It then follows that, for each  $t \leq L$ ,

$$\begin{aligned} |\tilde{Z}(t) - \tilde{Z}(0) - R\tilde{T}(t)| &\leq |\tilde{Z}(t) - Z^{r,m}(t)| + |Z^{r,m}(0) - \tilde{Z}(0)| + |R^r T^{r,m}(t) - R\tilde{T}(t)| \\ &\quad + |Z^{r,m}(t) - Z^{r,m}(0) - R^r T^{r,m}(t)| \\ &\leq (4 + NL\mathbf{J} + N\mathbf{J})\epsilon. \end{aligned}$$

Thus, equation (40) is satisfied by  $\tilde{\mathbf{X}}$  because  $\epsilon$  can be arbitrarily small.

To show equation (45), first observe that for any allocation  $a \in \mathcal{A}$ ,

$$\begin{aligned} |p_{\alpha, \epsilon_0}(a, \tilde{Z}(t)) - p_{\alpha, \epsilon_0}^r(a, Z^{r,m}(t))| &= |\alpha \times \tilde{Z}(t) \cdot Ra - \alpha \times Z^{r,m}(t) \cdot R^r a| \\ &\leq |\alpha| \left( |\tilde{Z}(t) \cdot (Ra - R^r a)| + |(\tilde{Z}(t) - Z^{r,m}(t)) \cdot R^r a| \right) \\ &\leq |\alpha| ((NL + 1)\mathbf{J}\epsilon + \epsilon\mathbf{J}N) = (NL + N + 1)|\alpha|\mathbf{J}\epsilon. \end{aligned} \quad (78)$$

In (78),  $p_{\alpha, \beta}^r(a, q) \equiv (\alpha \times q - \beta) \cdot R^r a$  is the network pressure for the  $r$ th network, associated with the parameter set  $(\alpha, \beta)$ , under allocation  $a$  when the queue length is  $q$ .

Denote  $\mathcal{E}^* = \operatorname{argmax}_{a \in \mathcal{E}} p_{\alpha, \epsilon_0}(a, \tilde{Z}(t))$  as the set of maximum extreme allocations under buffer size  $\tilde{Z}(t)$ . Because the limit network satisfies the EAA condition, we can choose an  $a^* \in \mathcal{E}^*$  such that  $\tilde{Z}_i(t) > 0$  for each constituent buffer  $i$  of  $a^*$ . Denote  $\mathcal{I}(a^*)$  the set of constituent buffers. Namely,

$$\mathcal{I}(a^*) = \left\{ i : \sum_j a_j^* B_{ji} > 0 \right\}.$$

Then  $\tilde{Z}_i(t) > 0$  for all  $i \in \mathcal{I}(a^*)$ .



Suppose  $a \in \mathcal{E} \setminus \mathcal{E}^*$ . Then  $p_{\alpha, e_0}(a, \tilde{Z}(t)) < \max_{a' \in \mathcal{E}} p_{\alpha, e_0}(a', \tilde{Z}(t))$ . Since  $p_{\alpha, e_0}(a^*, \tilde{Z}(t)) = \max_{a' \in \mathcal{E}} p_{\alpha, e_0}(a', \tilde{Z}(t))$  and  $\min_{i \in \mathcal{I}(a^*)} \tilde{Z}_i(t) > 0$ , by the continuity of  $\tilde{\mathbf{X}}(\cdot)$ , there exist  $\epsilon_1 > 0$  and  $\delta > 0$  such that for each  $\tau \in [t - \delta, t + \delta]$  and  $i \in \mathcal{I}(a^*)$ ,

$$p_{\alpha, e_0}(a, \tilde{Z}(\tau)) + \epsilon_1 \leq p_{\alpha, e_0}(a^*, \tilde{Z}(\tau)) \quad \text{and} \quad \tilde{Z}_i(\tau) \geq \epsilon_1. \quad (79)$$

For sufficiently large  $r$ , we can choose  $\epsilon$  small enough such that  $(NL + N + 1)|\alpha| \mathbf{J} \epsilon \leq \epsilon_1/3$ . It follows from (75), (78) and (79) that for all  $\tau \in [t - \delta, t + \delta]$ ,

$$\begin{aligned} p_{\alpha, e_0}^r(a, Z^{r,m}(\tau)) + \epsilon_1/3 &\leq p_{\alpha, e_0}^r(a^*, Z^{r,m}(\tau)), \\ Z_i^{r,m}(\tau) &\geq \epsilon_1/2 \quad \text{for each } i \in \mathcal{I}(a^*). \end{aligned}$$

Choosing  $r > (2 \vee 6|\beta| \mathbf{N} \mathbf{I}) \mathbf{J} / \epsilon_1$ , for each  $\tau \in [rm + \xi_{r,m}(t - \delta), rm + \xi_{r,m}(t + \delta)]$ , we have

$$p_{\alpha, \beta}^r(a, Z^r(\tau)) < p_{\alpha, \beta}^r(a^*, Z^r(\tau)), \quad (80)$$

$$Z_i^r(\tau) \geq \mathbf{J} \quad \text{for each } i \in \mathcal{I}(a^*). \quad (81)$$

Condition (81) implies that  $a^*$  is a feasible allocation at any time  $\tau \in [rm + \xi_{r,m}(t - \delta), rm + \xi_{r,m}(t + \delta)]$ , i.e.,  $a^* \in \mathcal{E}(\tau)$ . Following (80) and the definition of a (preemptive-resume) maximum pressure policy, the allocation  $a$  will not be employed during time interval  $[rm + \xi_{r,m}(t - \delta), rm + \xi_{r,m}(t + \delta)]$ . Therefore, only the allocations in  $\mathcal{E}^*$  will be employed during this interval. For each  $a \in \mathcal{E}$ , denote  $(T^a)^r(t)$  to be the cumulative amount of time allocation  $a$  has been employed by time  $t$ . It is easy to see that, for each  $r$  and all  $t \geq 0$ , under the maximum pressure policy,

$$\begin{aligned} \sum_{a \in \mathcal{E}} (T^a)^r(t) &= t, \\ T^r(t) &= \sum_{a \in \mathcal{E}} a (T^a)^r(t). \end{aligned}$$

Then it follows that

$$\begin{aligned} &(\alpha \times \tilde{Z}(t)) \cdot \left( R \left[ T^r(rm + \xi_{r,m}(t + \delta)) - T^r(rm + \xi_{r,m}(t - \delta)) \right] \right) \\ &= \sum_{a \in \mathcal{E}} \left( (\alpha \times \tilde{Z}(t)) \cdot Ra \right) \left[ (T^a)^r(rm + \xi_{r,m}(t + \delta)) - (T^a)^r(rm + \xi_{r,m}(t - \delta)) \right] \\ &= \sum_{a \in \mathcal{E}^*} \left( (\alpha \times \tilde{Z}(t)) \cdot Ra \right) \left[ (T^a)^r(rm + \xi_{r,m}(t + \delta)) - (T^a)^r(rm + \xi_{r,m}(t - \delta)) \right] \\ &= \left( \max_{a \in \mathcal{E}} (\alpha \times \tilde{Z}(t)) \cdot Ra \right) \sum_{a \in \mathcal{E}^*} \left[ (T^a)^r(rm + \xi_{r,m}(t + \delta)) - (T^a)^r(rm + \xi_{r,m}(t - \delta)) \right] \\ &= \left( \max_{a \in \mathcal{E}} (\alpha \times \tilde{Z}(t)) \cdot Ra \right) \sum_{a \in \mathcal{E}} \left[ (T^a)^r(rm + \xi_{r,m}(t + \delta)) - (T^a)^r(rm + \xi_{r,m}(t - \delta)) \right] \\ &= 2\xi_{r,m}\delta \left( \max_{a \in \mathcal{E}} (\alpha \times \tilde{Z}(t)) \cdot Ra \right). \end{aligned} \quad (82)$$

The second and fourth equalities in (82) hold because only allocations in  $\mathcal{E}^*$  will be employed during  $[rm + \xi_{r,m}(t - \delta), rm + \xi_{r,m}(t + \delta)]$ ; the third holds because every allocation  $a \in \mathcal{E}^*$  has the same network pressure equal to  $\max_{a \in \mathcal{E}}(\alpha \times \tilde{Z}(t)) \cdot Ra$ . From (82), we have

$$(\alpha \times \tilde{Z}(t)) \cdot R(T^{r,m}(t + \delta) - T^{r,m}(t - \delta))/2\delta = \max_{a \in \mathcal{E}}(\alpha \times \tilde{Z}(t)) \cdot Ra.$$

Because  $\epsilon$  in (76) can be arbitrarily small, we have

$$(\alpha \times \tilde{Z}(t)) \cdot R(\tilde{T}(t + \delta) - \tilde{T}(t - \delta))/2\delta = \max_{a \in \mathcal{E}}(\alpha \times \tilde{Z}(t)) \cdot Ra,$$

and by letting  $\delta \rightarrow \infty$ , (45) is verified.  $\square$

Since every fluid model solution under the maximum pressure policy satisfies (46) by Theorem 6, Proposition 3 implies that any cluster point  $\tilde{\mathbb{X}}$  of  $\mathcal{F}_g$  satisfies (46). That is,

$$|\tilde{Z}(t) - \zeta^\alpha \tilde{W}(t)| = 0 \text{ for } t \geq \tau_0,$$

where  $\tilde{W}(t) = y^* \cdot \tilde{Z}(t)$ . Define

$$W^{r,m} = y^r \cdot Z^{r,m}.$$

Because the fluid-scaled stochastic processing network processes can be uniformly approximated by cluster points, it leads to the following proposition.

**Proposition 4.** *Fix  $L > \tau_0$ ,  $T > 0$  and  $\epsilon > 0$ . For  $r$  large enough,*

$$|Z^{r,m}(t) - \zeta^\alpha W^{r,m}(t)| \leq \epsilon \text{ for all } 0 \leq m \leq rT, \tau_0 \leq t \leq L, \omega \in \mathcal{G}^r.$$

To prove Proposition 4, we first need to show that  $y^r \rightarrow y^*$  as  $r \rightarrow \infty$ .

**Lemma 8.** *Assume Assumptions 1 and 2. Then  $(y^r, z^r) \rightarrow (y^*, z^*)$  as  $r \rightarrow \infty$ , where, as before,  $(y^r, z^r)$  is the unique optimal solution to the static planning problem (10)–(14) for the  $r$ th network.*

The proof of Lemma 8 will be provided in Appendix B.

*Proof of Proposition 4.* From Proposition 2, for each  $\epsilon > 0$ , large enough  $r$ , and each  $0 \leq m \leq rT$  and  $\omega \in \mathcal{G}^r$ , we can find a cluster point  $\tilde{\mathbb{X}}$  such that

$$\|\tilde{Z}(\cdot) - Z^{r,m}(\cdot)\|_L \leq \epsilon.$$

From Proposition 3 and Theorem 6, we have

$$|\tilde{Z}(t) - \zeta^\alpha \tilde{W}(t)| = 0 \text{ for } t \geq \tau_0.$$

Then, for each  $t \geq \tau_0$ ,

$$|Z^{r,m}(t) - \zeta^\alpha W^{r,m}(t)| \leq |Z^{r,m}(t) - \tilde{Z}(t)| + |\zeta^\alpha \tilde{W}(t) - \zeta^\alpha W^{r,m}(t)|. \quad (83)$$

Because  $y^r \rightarrow y^*$  as  $r \rightarrow \infty$ , for large enough  $r$ ,

$$|y^r - y^*| < \epsilon.$$

Let  $\kappa = (\sup_r |y^r|) \vee |\zeta^\alpha| < \infty$ . Then we have

$$\begin{aligned} |\tilde{W}(t) - W^{r,m}(t)| &\leq |y^* \cdot \tilde{Z}(t) - y^r \cdot \tilde{Z}(t)| + |y^r \cdot \tilde{Z}(t) - y^r \cdot Z^{r,m}(t)| \\ &\leq \mathbf{I}(\epsilon |\tilde{Z}(t)| + \kappa \epsilon) \leq (NL + \kappa) \mathbf{I} \epsilon \end{aligned}$$

Note that  $\tilde{Z}(t) \leq NL$  for all  $t \leq L$ . One also gets  $\tilde{W}(t) \leq \kappa NL$  since  $|y^*| \leq \kappa$ . From (83), we have

$$|Z^{r,m}(t) - \zeta^\alpha W^{r,m}(t)| \leq \epsilon + \kappa(NL + \kappa) \mathbf{I} \epsilon.$$

The proposition follows by rechoosing  $\epsilon$ .  $\square$

We need to translate the results in Proposition 4 into the state space collapse results for diffusion-scaled processes. First we can express  $Z^{r,m}(t)$  by  $\hat{Z}^r$  via

$$Z^{r,m}(t) = \frac{r}{\xi_{r,m}} \hat{Z}^r\left(\frac{t\xi_{r,m} + rm}{r^2}\right) = \frac{1}{\bar{\xi}_{r,m}} \hat{Z}^r\left(\frac{t\bar{\xi}_{r,m} + m}{r}\right),$$

where  $\bar{\xi}_{r,m} = \xi_{r,m}/r$ . The time interval  $[\tau_0, L]$  for the fluid-scaled process  $Z^{r,m}$  corresponds to the time interval  $[(m + \bar{\xi}_{r,m}\tau_0)/r, (m + \bar{\xi}_{r,m}L)/r]$  for the diffusion-scaled process  $\hat{Z}^r$ . Proposition 4 immediately leads to the following.

**Proposition 5.** *Fix  $L > \tau_0$ ,  $T > 0$  and  $\epsilon > 0$ . For  $r$  large enough and each  $m < rT$ ,*

$$|\hat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t)| \leq \bar{\xi}_{r,m} \epsilon \text{ for all } (m + \bar{\xi}_{r,m}\tau_0)/r \leq t \leq (m + \bar{\xi}_{r,m}L)/r, \quad \omega \in \mathcal{G}^r. \quad (84)$$

Proposition 5 gives estimates on each small interval for  $|\hat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t)|$ . We shall obtain the estimate on the whole time interval  $[0, T]$ , and then show that  $\bar{\xi}_{r,m}$  are stochastically bounded. The following lemma ensures that for large enough  $L$ ,  $L > 3NT \vee \tau_0 \vee 1$ , the small intervals in Proposition 5 are overlapping, and therefore the estimate on  $[\bar{\xi}_{r,0}\tau_0/r, T]$  is obtained.

**Lemma 9.** *For a fixed  $T > 0$  and large enough  $r$ ,*

$$\bar{\xi}_{r,m+1} \leq 3N\bar{\xi}_{r,m}$$

for  $\omega \in \mathcal{G}^r$  and  $m < rT$ .

*Proof.* By the definition of  $\mathcal{G}^r$ ,

$$|Z^{r,m}(t_2) - Z^{r,m}(t_1)| \leq N|t_2 - t_1| + 1$$

for  $t_1, t_2 \in [0, L]$  and  $m < rT$ . Set  $t_1 = 0$  and  $t_2 = 1/\bar{\xi}_{r,m}$ , we have

$$|\hat{Z}^r((m+1)/r)| - |\hat{Z}^r(m/r)| \leq N + \bar{\xi}_{r,m}.$$

Therefore,

$$\bar{\xi}_{r,m+1} \leq |\hat{Z}^r((m+1)/r)| \vee 1 \leq \left(|\hat{Z}^r(m/r)| + N + \bar{\xi}_{r,m}\right) \vee 1 \leq N + 2\bar{\xi}_{r,m} \leq 3N\bar{\xi}_{r,m}.$$

$\square$

### 7.3 State space collapse on $[0, \xi_{r,0}L/r^2]$

Now we shall estimate  $|\widehat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t)|$  on the interval  $[0, \bar{\xi}_{r,0}L/r] = [0, \xi_{r,0}L/r^2]$ . This will be given by the initial condition (24) and the result in the second part of Theorem 6.

Condition (24) implies that

$$|Z^{r,0}(0) - \zeta^\alpha W^{r,0}(0)| \rightarrow 0 \quad \text{in probability.}$$

Then, for each  $r > 0$ , we let

$$\epsilon_1(r) = \min\{\epsilon : \mathbb{P}(|Z^{r,0}(0) - \zeta^\alpha W^{r,0}(0)| > \epsilon) \leq \epsilon\}. \quad (85)$$

It follows from (85) that

$$\lim_{r \rightarrow \infty} \epsilon_1(r) \rightarrow 0.$$

Now let  $\mathcal{L}^r$  be the intersection of  $\mathcal{G}^r$  and the event

$$|Z^{r,0}(0) - \zeta^\alpha W^{r,0}(0)| \leq \epsilon_1(r).$$

Obviously,  $\lim_{r \rightarrow \infty} \mathbb{P}(\mathcal{L}^r) = 1$ .

We define

$$\mathcal{F}_o = \{F_o^r\}$$

with

$$F_o^r = \{\mathbb{X}^{r,0}(\cdot, \omega), \omega \in \mathcal{L}^r\}.$$

Parallel to Proposition 2, we have the following proposition which states that  $\mathcal{F}_o$  can be asymptotically approximated by cluster points of  $\mathcal{F}_o$ .

**Proposition 6.** *Fix  $\epsilon > 0$ ,  $L > 0$  and  $T > 0$ , and choose  $r$  large enough. Then, for  $\omega \in \mathcal{L}^r$ ,*

$$\|\mathbb{X}^{r,0}(\cdot, \omega) - \tilde{\mathbb{X}}(\cdot)\|_L \leq \epsilon$$

for some cluster point  $\tilde{\mathbb{X}}(\cdot)$  of  $\mathcal{F}_o$  with  $\tilde{\mathbb{X}}(\cdot) \in F'$ .

*Proof.* Since both (69) and (71) are satisfied by  $\mathbb{X}^{r,0}$ , the result follows from Lemma 7.  $\square$

**Proposition 7.** *Fix  $L > 0$ . Then for any cluster point  $\tilde{\mathbb{X}}(\cdot)$  of  $\mathcal{F}_o$ ,*

$$\tilde{Z}(t) = \zeta^\alpha \tilde{W}(t) \quad \text{for } t \in [0, L].$$

*Proof.* Since any cluster point of  $\mathcal{F}_o$  is automatically a cluster point of  $\mathcal{F}_g$ , therefore it satisfies all the fluid model equations. Now we show that

$$|\tilde{Z}(0) - \zeta^\alpha \tilde{W}(0)| = 0, \quad (86)$$

which, together with Theorem 6, implies the result.

For any given  $\delta > 0$ , one can choose  $r$  large enough and  $\omega \in \mathcal{L}^r$  such that

$$\begin{aligned} \|\mathbb{X}^{r,0}(\cdot) - \tilde{\mathbb{X}}(\cdot)\|_L &\leq \delta, \\ |Z^{r,0}(0) - \zeta^\alpha W^{r,0}(0)| &\leq \delta. \end{aligned}$$

Let  $\kappa = \sup_r |y^r|$ . Then

$$|W^{r,0}(0)| = |y^r \cdot Z^{r,0}(0)| \leq \kappa \mathbf{I}$$

and

$$\begin{aligned} |\tilde{W}(0) - W^{r,0}(0)| &= |y^* \cdot \tilde{Z}(0) - y^r \cdot Z^{r,0}(0)| \\ &\leq |y^* \cdot (\tilde{Z}(0) - Z^{r,0}(0)) + (y^* - y^r) \cdot Z^{r,0}(0)| \\ &\leq \mathbf{I}(|y^*| + 1)\delta. \end{aligned}$$

It follows that

$$\begin{aligned} |\tilde{Z}(0) - \zeta^\alpha \tilde{W}(0)| &\leq |\tilde{Z}(0) - Z^{r,0}(0)| + |Z^{r,0}(0) - \zeta^\alpha W^{r,0}(0)| \\ &\quad + |\zeta^\alpha (W^{r,0}(0) - \tilde{W}(0))| \\ &\leq 2\delta + |\zeta^\alpha| \mathbf{I}(|y^*| + 1)\delta. \end{aligned}$$

Because  $\delta$  can be arbitrarily small, we have (86) and the proposition follows.  $\square$

Propositions 6 and 7 immediately lead to the following proposition, which is parallel to Propositions 4 and 5.

**Proposition 8.** *Fix  $L$  and  $\epsilon > 0$ . For large enough  $r$ ,*

$$|Z^{r,0}(t) - \zeta^\alpha W^{r,0}(t)| \leq \epsilon \text{ for all } 0 \leq t \leq L, \omega \in \mathcal{L}^r,$$

and

$$|\hat{Z}^r(t) - \zeta^\alpha \hat{W}^r(t)| \leq \bar{\xi}_{r,0} \epsilon \text{ for all } 0 \leq t \leq \bar{\xi}_{r,0} L/r, \omega \in \mathcal{L}^r. \quad (87)$$

## 7.4 Proof of Theorem 5

Propositions 5 and 8 together give the multiplicative state space collapse of the stochastic processing network processes. To prove the state space collapse result stated in Theorem 5, it is enough to prove that  $\bar{\xi}_{r,m}$  are stochastically bounded. We first give an upper bound on  $\bar{\xi}_{r,m}$  in terms of  $\hat{W}^r$ .

**Lemma 10.** *If  $|\hat{Z}^r(m/r) - \zeta^\alpha \hat{W}^r(m/r)| \leq 1$ , there exists some  $\kappa \geq 1$  such that*

$$\bar{\xi}_{r,m} \leq \kappa (\hat{W}^r(m/r) \vee 1).$$

*Proof.* Because  $\xi_{r,m} = |Z^r(rm)| \vee r$ , we have

$$\bar{\xi}_{r,m} = |\hat{Z}^r(m/r)| \vee 1 \leq |\zeta^\alpha \hat{W}^r(m/r) + 1| \leq 2(|\zeta^\alpha| \hat{W}^r(m/r) \vee 1).$$

The result then follows by choosing  $\kappa = 2(|\zeta^\alpha| \vee 1)$ .  $\square$

The following proposition will be used to derive an upper bound on the oscillation of  $\widehat{W}^r$ .

**Proposition 9.** *Consider a sequence of stochastic processing networks operating under the maximum pressure policy with parameter set  $(\alpha, \beta)$ . There exists  $\epsilon_0 > 0$  such that for large enough  $r$ , and any  $0 \leq t_1 < t_2$ , if  $\widehat{W}^r(t) \geq \mathbf{J}/(r\epsilon_0)$  and  $|\widehat{Z}^r(t)/\widehat{W}^r(t) - \zeta^\alpha| \leq \epsilon_0$  for all  $t \in [t_1, t_2]$ , then*

$$\widehat{Y}^r(t_2) = \widehat{Y}^r(t_1).$$

*Proof.* First, we can choose  $0 < \epsilon_0 \leq \min\{\zeta_i^\alpha/2 : \zeta_i^\alpha > 0\}$  so that for all  $i$  with  $\zeta_i^\alpha > 0$  and all  $t \in [t_1, t_2]$ , we have

$$\widehat{Z}_i^r(t) \geq \widehat{W}^r(t)(\zeta_i^\alpha - \epsilon_0) \geq \widehat{W}^r(t)\epsilon_0 \geq \mathbf{J}/r.$$

Namely,  $Z_i^r(\tau) \geq \mathbf{J}$  for all  $i$  with  $\zeta_i^\alpha > 0$  and all  $\tau \in [r^2t_1, r^2t_2]$ . Define

$$\mathcal{E}^* = \operatorname{argmax}_{a \in \mathcal{E}} y^* \cdot Ra = \operatorname{argmax}_{a \in \mathcal{E}} (\alpha \times \zeta^\alpha) \cdot Ra.$$

Notice that  $\alpha \times \zeta^\alpha = y^*/\sum_{i \in \mathcal{I}}((y_i^*)^2/\alpha_i)$ . Because the limit network satisfies the EAA condition, there exists an allocation  $a^* \in \mathcal{E}^*$  such that for each  $i \in \mathcal{I}$ ,  $\zeta_i^\alpha > 0$  if  $\sum_j B_{ji}a_j^* > 0$ . It follows that  $Z_i^r(\tau) \geq \mathbf{J}$  for all  $\tau \in [r^2t_1, r^2t_2]$  if  $\sum_j B_{ji}a_j^* > 0$ . This implies that  $a^*$  is a feasible allocation during  $[r^2t_1, r^2t_2]$ .

Next, we will show that if  $\epsilon_0$  is chosen sufficiently small, only allocations in  $\mathcal{E}^*$  can be employed under a maximum pressure policy during  $[r^2t_1, r^2t_2]$ . Let

$$\epsilon_1 = (y^* \cdot Ra^* - \max_{a \in \mathcal{E} \setminus \mathcal{E}^*} y^* \cdot Ra) > 0.$$

Set  $\kappa_0 = \sup_r \max_{a \in \mathcal{E}} |R^r a|$  and  $\kappa_1 = \sum_{i \in \mathcal{I}}((y_i^*)^2/\alpha_i)$ . Choose  $r$  large enough such that  $|(y^*)'(R^r - R)| \leq \epsilon_0$ . Then for each allocation  $a \in \mathcal{E} \setminus \mathcal{E}^*$  and each  $\tau \in [r^2t_1, r^2t_2]$ ,

$$\begin{aligned} & \left( (\alpha \times Z^r(\tau)) \cdot R^r a^* - (\alpha \times Z^r(\tau)) \cdot R^r a \right) / W^r(\tau) = \left( \frac{\alpha \times Z^r(\tau)}{W^r(\tau)} - \alpha \times \zeta^\alpha \right) \cdot R^r a^* \\ & + (y^* \cdot R^r a^* - y^* \cdot Ra^*) / \kappa_1 + (y^* \cdot Ra^* - y^* \cdot Ra) / \kappa_1 \\ & + (y^* \cdot Ra - y^* \cdot R^r a) / \kappa_1 + \left( \alpha \times \zeta^\alpha - \frac{\alpha \times Z^r(\tau)}{W^r(\tau)} \right) \cdot R^r a \\ & \geq -|\alpha| \epsilon_0 \kappa_0 \mathbf{I} - \epsilon_0 \mathbf{J} / \kappa_1 + \epsilon_1 / \kappa_1 - \epsilon_0 \mathbf{J} / \kappa_1 - |\alpha| \epsilon_0 \kappa_0 \mathbf{I} \\ & = (\epsilon_1 - 2(|\alpha| \kappa_0 \kappa_1 \mathbf{I} + \mathbf{J}) \epsilon_0) / \kappa_1. \end{aligned} \tag{88}$$

Meanwhile, because  $W^r(\tau) \geq \mathbf{J}/\epsilon_0$ , we have

$$|\beta \cdot R^r a^* - \beta \cdot R^r a| / W^r(\tau) \leq 2|\beta| \mathbf{I} \kappa_0 \epsilon_0 / \mathbf{J} \leq 2|\beta| \mathbf{I} \kappa_0 \epsilon_0. \tag{89}$$

Combining (88) and (89), we have

$$p_{\alpha,\beta}^r(a^*, Z^r(\tau)) - p_{\alpha,\beta}^r(a, Z^r(\tau)) \geq \left( \epsilon_1 - 2(|\alpha| + |\beta|)\kappa_0\kappa_1\mathbf{I} + \mathbf{J} \right) \epsilon_0 / \kappa_1.$$

Set  $\epsilon_0 = \epsilon_1 / \left( 3(|\alpha| + |\beta|)\kappa_0\kappa_1\mathbf{I} + \mathbf{J} \right) \wedge \min_{i \in \mathcal{I}} \{ \zeta_i^\alpha / 2 : \zeta_i^\alpha > 0 \}$ , then

$$p_{\alpha,\beta}^r(a^*, Z^r(\tau)) - p_{\alpha,\beta}^r(a, Z^r(\tau)) > 0. \quad (90)$$

Obviously,  $\epsilon_0$  depends only on  $\alpha, \beta, R$  and  $A$ . Equation (90) implies that the pressure under allocation  $a^*$  is strictly larger than that under allocation  $a$ . It follows that only the allocations in  $\mathcal{E}^*$  can be employed during  $[r^2t_1, r^2t_2]$ . Therefore, for large enough  $r$ ,

$$\begin{aligned} y^r \cdot R^r (T^r(r^2t_2) - T^r(r^2t_1)) &= \sum_{a \in \mathcal{E}} y^r \cdot R^r a ((T^a)^r(r^2t_2) - (T^a)^r(r^2t_1)) \\ &= \sum_{a \in \mathcal{E}^*} y^r \cdot R^r a ((T^a)^r(r^2t_2) - (T^a)^r(r^2t_1)) \\ &= \sum_{a \in \mathcal{E}^*} (1 - \rho^r) ((T^a)^r(r^2t_2) - (T^a)^r(r^2t_1)) \\ &= (1 - \rho^r)(r^2t_2 - r^2t_1), \end{aligned} \quad (91)$$

where the third equality follows from Lemma 11 below. Equations (91) and (26) imply that  $\widehat{Y}^r(t_2) - \widehat{Y}^r(t_1) = 0$ .  $\square$

**Lemma 11.** *For each  $a^* \in \mathcal{E}^* = \operatorname{argmax}_{a \in \mathcal{E}} y^* \cdot Ra$  and large enough  $r$ ,*

$$y^r \cdot R^r a^* = \max_{a \in \mathcal{E}} y^r \cdot R^r a = 1 - \rho^r.$$

The proof of Lemma 11 will be provided in Appendix B. We now complete the proof of Theorem 5.

*Proof of Theorem 5.* We only need to show that  $\bar{\xi}_{r,m}$  are stochastically bounded. We first fix an  $\epsilon > 0$ . Because  $\widehat{W}^r(0)$  is stochastically bounded there exists a  $\kappa_1$  such that for  $r$  large enough,

$$\mathbb{P}(\widehat{W}^r(0) \leq \kappa_1) \geq 1 - \epsilon.$$

By the property of Brownian motion, there exists a  $\kappa_2$  such that

$$\mathbb{P}(\operatorname{Osc}(X^*, [0, T + L]) \leq \kappa_2/2) \geq 1 - \epsilon/2,$$

where  $X^*$  is the Brownian motion defined in Lemma 2, and

$$\operatorname{Osc}(f, [0, t]) = \sup_{0 \leq t_1 < t_2 \leq t} |f(t_2) - f(t_1)|.$$

Since  $\widehat{X}^r$  converges to  $X^*$  in distribution, for large enough  $r$ ,

$$\mathbb{P}(\operatorname{Osc}(\widehat{X}^r, [0, T + L]) \leq \kappa_2) \geq 1 - \epsilon.$$

Meanwhile, because

$$|\widehat{Z}^r(0) - \zeta^\alpha \widehat{W}^r(0)| \rightarrow 0 \text{ in probability,}$$

we have

$$\mathbb{P}(|\widehat{Z}^r(0) - \zeta^\alpha \widehat{W}^r(0)| \leq \epsilon) \geq 1 - \epsilon$$

for  $r$  large enough.

Define

$$\mathcal{H}^{r,\epsilon} = \{\omega : \widehat{W}^r(0) \leq \kappa_1, \text{Osc}(\widehat{X}^r, [0, T + L]) \leq \kappa_2, \text{ and } |\widehat{Z}^r(0) - \zeta^\alpha \widehat{W}^r(0)| \leq \epsilon\}.$$

Then for  $r$  large enough,

$$\mathbb{P}(\mathcal{H}^{r,\epsilon}) \geq 1 - 3\epsilon.$$

Furthermore, we can choose  $r$  large enough such that Propositions 5 and 8 hold with  $\epsilon$  replaced by  $\epsilon/\kappa(\kappa_1 + \kappa_2 + 1)$  and  $\mathbb{P}(\mathcal{L}^r) \geq 1 - \epsilon$ , where  $\kappa$  is given as in Lemma 10. Note that  $\mathbb{P}(\mathcal{L}^r) \rightarrow 1$  as  $r \rightarrow \infty$ .

Denote  $\mathcal{N}^{r,\epsilon} = \mathcal{L}^r \cap \mathcal{H}^{r,\epsilon}$ , then for all  $r$  large enough,

$$\mathbb{P}(\mathcal{N}^{r,\epsilon}) \geq 1 - 4\epsilon.$$

Now we are going to show that if  $\epsilon \leq \epsilon_0$ ,  $\bar{\xi}_{r,m} \leq \kappa(\kappa_1 + \kappa_2 + 1)$  on  $\mathcal{N}^{r,\epsilon}$  for all  $r$  large enough and  $m \leq rT$ . In fact, we are going to show the following is true on  $\mathcal{N}^{r,\epsilon}$  for all  $r$  large enough and  $m \leq rT$ :

$$|\widehat{Z}^r(m/r) - \zeta^\alpha \widehat{W}^r(m/r)| \leq \epsilon \tag{92}$$

$$\bar{\xi}_{r,m} \leq \kappa(\kappa_1 + \kappa_2 + 1) \tag{93}$$

$$|\widehat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t)| \leq \epsilon \text{ for all } 0 \leq t \leq (m + \bar{\xi}_{r,m}L)/r \tag{94}$$

$$\int_0^{(m + \bar{\xi}_{r,m}L)/r} 1_{(\widehat{W}^r(s) > 1)} d\widehat{Y}^r(s) = 0. \tag{95}$$

This will be shown by induction. When  $m = 0$ , (92) obviously holds on  $\mathcal{N}^{r,\epsilon}$ , and

$$\bar{\xi}_{r,0} \leq \kappa(\widehat{W}^r(0) \vee 1) \leq \kappa(\kappa_1 \vee 1) \leq \kappa(\kappa_1 + \kappa_2 + 1).$$

Meanwhile, from (87), replacing  $\epsilon$  by  $\epsilon/\kappa(\kappa_1 + \kappa_2 + 1)$ , we have

$$|\widehat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t)| \leq \epsilon \text{ for all } t \in [0, \bar{\xi}_{r,0}L/r].$$

Then from Proposition 9, with  $r \geq 2\mathbf{J}/\epsilon_0$ , we have

$$\int_0^{\bar{\xi}_{r,0}L/r} 1_{(\widehat{W}^r(s) > 1)} d\widehat{Y}^r(s) = 0.$$



Now we assume that (92) - (95) hold up to  $m$ , and we shall show they also hold for  $m + 1$ . First, (94) directly implies (92) because  $\bar{\xi}_{r,m}L > 1$ . Next, because (32) and (95) hold, and by Theorem 5.1 of Williams (1998b), we have

$$\text{Osc}(\widehat{W}^r, [0, (m + \bar{\xi}_{r,m}L)/r]) \leq \text{Osc}(\widehat{X}^r, [0, (m + \bar{\xi}_{r,m}L)/r]) + 1 \leq \kappa_2 + 1.$$

Note that, because  $m \leq rT$  and (93), we have  $(m + \bar{\xi}_{r,m}L)/r \leq T + L$  for  $r$  large enough. It then follows that

$$\begin{aligned} \bar{\xi}_{r,m+1} &\leq \kappa(\widehat{W}^r((m+1)/r) \vee 1) \\ &\leq \kappa(\widehat{W}^r(0) + \text{Osc}(\widehat{W}^r, [0, (m+1)/r]) \vee 1) \\ &\leq \kappa(\widehat{W}^r(0) + \text{Osc}(\widehat{W}^r, [0, (m + \bar{\xi}_{r,m}L)/r]) \vee 1) \\ &\leq \kappa(\kappa_1 + \kappa_2 + 1). \end{aligned}$$

This fact and (84) imply that

$$|\widehat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t)| \leq \epsilon \text{ for all } t \in [(m+1 + \bar{\xi}_{r,m+1}\tau_0)/r, (m+1 + \bar{\xi}_{r,m+1}L)/r].$$

Since  $L \geq 3N \vee \tau_0 \vee 2$  by our choice, Lemma 9 implies  $\bar{\xi}_{r,m}L/r \geq (1 + \bar{\xi}_{r,m+1}\tau_0)/r$ . It follows that

$$|\widehat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t)| \leq \epsilon \text{ for all } t \in [0, (m+1 + \bar{\xi}_{r,m+1}L)/r].$$

Then again Proposition 9 gives

$$\int_0^{(m+1+\bar{\xi}_{r,m+1}L)/r} 1_{(\widehat{W}^r(s) > 1)} d\widehat{Y}^r(s) = 0.$$

Therefore, we can conclude that  $\bar{\xi}_{r,m} \leq \kappa(\kappa_1 + \kappa_2 + 1)$  for all  $0 \leq m \leq rT$  which implies

$$\|\widehat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t)\|_T \leq \epsilon \text{ on } \mathcal{N}^{r,\epsilon}$$

for all large enough  $r$ . Theorem 5 follows because  $\epsilon$  can be chosen arbitrarily small.  $\square$

## 8 Heavy Traffic Limit Theorem

In this section, we prove the heavy traffic limit theorem, Theorem 4. Our proof employs an invariance principle developed in Williams (1998b) for semimartingale reflecting Brownian motions (RBMs), specialized to the one-dimensional case.

Recall that the constant  $\epsilon_0$  is defined in Proposition 9. For each  $t \geq 0$ , let

$$\varepsilon^r(t) = \widehat{Z}^r(t) - \zeta^\alpha \widehat{W}^r(t) \quad \text{and} \quad \delta^r(t) = (|\varepsilon^r(t)| \vee (2\mathbf{J}/r))/\epsilon_0.$$

It follows from the state space collapse theorem, Theorem 5, that for each  $t \geq 0$ ,  $\varepsilon^r(t) \rightarrow 0$ , and hence  $\delta^r(t) \rightarrow 0$  in probability as  $r \rightarrow \infty$ . Now, for each  $t \geq 0$ , define

$$\gamma^r(t) = \widehat{W}^r(t) \wedge \delta^r(t) \quad \text{and} \quad \tilde{W}^r(t) = \widehat{W}^r(t) - \gamma^r(t).$$

It is easy to see that  $\tilde{W}^r(t) = (\widehat{W}^r(t) - \delta^r(t)) \vee 0 \geq 0$  for each  $t \geq 0$ . Now we claim that

$$\int_0^\infty \tilde{W}^r(s) d\widehat{Y}^r(s) = 0. \quad (96)$$

To see this, it is enough to show that for any  $0 \leq t_1 < t_2$ ,  $\widehat{Y}^r(t_2) = \widehat{Y}^r(t_1)$  whenever  $\tilde{W}^r(t) > 0$  for all  $t \in [t_1, t_2]$ ; see, for example, Lemma 3.1.2 of [Dai and Williams \(2003\)](#). Suppose that  $\tilde{W}^r(t) > 0$  for all  $t \in [t_1, t_2]$ . Then  $\widehat{W}^r(t) > \delta^r(t)$  for all  $t \in [t_1, t_2]$ . The latter condition implies that  $\widehat{W}^r(t) \geq 2\mathbf{J}/(r\epsilon_0)$  and  $|\widehat{Z}^r(t)/\widehat{W}^r(t) - \zeta^\alpha| \leq \epsilon_0$  for all  $t \in [t_1, t_2]$ . It follows from Proposition 9 that  $\widehat{Y}^r(t_2) = \widehat{Y}^r(t_1)$ , thus proving (96).

It follows that, for each  $r$ , the triple  $(\widehat{W}^r, \widehat{X}^r, \widehat{Y}^r)$  is a solution to the following one-dimensional, perturbed Skorohod problem: with probability 1,

$$\begin{aligned} \widehat{W}^r &= \widehat{X}^r + \widehat{Y}^r, \\ \widehat{W}^r &= \tilde{W}^r + \gamma^r, \\ \tilde{W}^r(t) &\geq 0 \text{ for all } t \geq 0, \\ \widehat{Y}^r(0) &= 0 \text{ and } \widehat{Y}^r(\cdot) \text{ is non-decreasing,} \\ \int_0^\infty \tilde{W}^r(s) d\widehat{Y}^r(s) &= 0. \end{aligned}$$

For each  $t \geq 0$ , since  $\delta^r(t) \rightarrow 0$  in probability as  $r \rightarrow \infty$ ,  $\gamma^r(t) \rightarrow 0$  in probability as  $r \rightarrow \infty$ . By Theorem 1 the maximum pressure policy is asymptotically efficient, and thus, by Lemma 2,  $\widehat{X}^r$  converges in distribution to the Brownian motion  $X^*$  as  $r \rightarrow \infty$ . It follows from the invariance principle for SRBMs, Theorem 4.1 of [Williams \(1998b\)](#), that  $\widehat{W}^r$  converges in distribution to the reflecting Brownian motion  $W^* = \psi(X^*)$ . The convergence of  $\widehat{Z}^r$  to  $\zeta^\alpha W^*$  follows from the state space collapse result (39). Thus, we have completed the proof of Theorem 4.

## Acknowledgement

We thank Ruth Williams for improving the exposition of the paper.

## References

- ATA, B. and KUMAR, S. (2005). Heavy traffic analysis of open processing networks with complete resource pooling: asymptotic optimality of discrete review policies. *Ann. Appl. Probab.*, **15** 331–391.
- BELL, S. L. and WILLIAMS, R. J. (2005). Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electronic J. of Probability*, **10** 1044–1115.

- BELL, S. L. and WILLIAMS, R. L. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Annals of Applied Probability*, **11** 608–649.
- BILLINGSLEY, P. (1999). *Convergence of probability measures*. 2nd ed. John Wiley & Sons Inc., New York.
- BRAMSON, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems: Theory and Applications*, **30** 89–148.
- BRAMSON, M. and DAI, J. G. (2001). Heavy traffic limits for some queueing networks. *Annals of Applied Probability*, **11** 49–90.
- BRAMSON, M. and WILLIAMS, R. J. (2003). Two workload properties for Brownian networks. *Queueing Syst.*, **45** 191–221.
- CHEN, H. and YAO, D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag, New York.
- CHEN, H. and ZHANG, H. (2000). Diffusion approximations for some multiclass queueing networks with FIFO service disciplines. *Mathematics of Operations Research*, **25** 679–707.
- CHUANG, S.-T., GOEL, A., MCKEOWN, N. and PRABHAKAR, B. (1999). Matching output queueing with a combined input output queued switch. *IEEE Journal of Selected Areas in Communications*, **17** 1030–1039.
- DAI, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability*, **5** 49–77.
- DAI, J. G. and KURTZ, T. G. (1995). A multiclass station with Markovian feedback in heavy traffic. *Mathematics of Operations Research*, **20** 721–742.
- DAI, J. G. and LIN, W. (2005). Maximum pressure policies in stochastic processing networks. *Oper. Res.*, **53** 197–218.
- DAI, J. G. and WILLIAMS, R. J. (2003). Stochastic networks, fluids and diffusions. Unpublished Draft, URL [www.isye.gatech.edu/faculty/dai/](http://www.isye.gatech.edu/faculty/dai/).
- ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.
- GANS, N., KOOLE, G. and MANDELBAUM, A. (2003). Telephone call centers: tutorial, review, and research prospects. *Management Science and Operations Management*, **5** 79–141.
- GEORGIADIS, L., NEELY, M. J. and TASSIULAS, L. (2006). Resource allocation and cross-layer control in wireless networks. *Foundations and Trends in Networking*, **1** 1–144.

- HARRISON, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Their Applications* (W. Fleming and P. L. Lions, eds.), vol. 10 of *The IMA volumes in mathematics and its applications*. Springer, New York, 147–186.
- HARRISON, J. M. (1996). The BIGSTEP approach to flow management in stochastic processing networks. In *Stochastic Networks: Theory and Applications* (S. Z. F. P. Kelly and I. Ziedins, eds.), vol. 4 of *Lecture Note Series*. Oxford University Press, 57–90.
- HARRISON, J. M. (1998). Heavy traffic analysis of a system with parallel servers: asymptotic analysis of discrete-review policies. *Annals of Applied Probability*, **8** 822–848.
- HARRISON, J. M. (2000). Brownian models of open processing networks: canonical representation of workload. *Annals of Applied Probability*, **10** 75–103.
- HARRISON, J. M. (2002). Stochastic networks and activity analysis. In *Analytic Methods in Applied Probability* (Y. Suhov, ed.). In Memory of Fridrik Karpelevich, American Mathematical Society, Providence, RI.
- HARRISON, J. M. (2003). A broader view of Brownian networks. *Annals of Applied Probability*, **13** 1119–1150.
- HARRISON, J. M. and LÓPEZ, M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems: Theory and Applications*, **33** 339–368.
- HARRISON, J. M. and VAN MIEGHEM, J. A. (1997). Dynamic control of Brownian networks: state space collapse and equivalent workload formulations. *Annals of Applied Probability*, **7** 747–771.
- IGLEHART, D. L. and WHITT, W. (1970). Multiple channel queues in heavy traffic I. *Adv. Appl. Probab.*, **2** 150–177.
- KUMAR, P. R. (1993). Re-entrant lines. *Queueing Systems: Theory and Applications*, **13** 87–110.
- MANDELBAUM, A. and STOLYAR, A. L. (2004). Scheduling flexible servers with convex delay costs: heavy traffic optimality of the generalized  $c\mu$ -rule. *Operations Research*, **52** 836–855.
- MANGASARIAN, O. (1979). Uniqueness of solution in linear programming. *Linear Algebra and its Applications*, **25** 151–162.
- RUDIN, W. (1991). *Functional Analysis*. 2nd ed. McGraw-Hill.
- STOLYAR, A. L. (2004). Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Annals of Applied Probability*, **14** 1–53.

- TASSIULAS, L. (1995). Adaptive back-pressure congestion control based on local information. *IEEE Transactions on Automatic Control*, **40** 236–250.
- TASSIULAS, L. and BHATTACHARYA, P. B. (2000). Allocation of interdependent resources for maximal throughput. *Stochastic Models*, **16** 27–48.
- TASSIULAS, L. and EPHREMIDES, A. (1992). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, **37** 1936–1948.
- TASSIULAS, L. and EPHREMIDES, A. (1993). Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory*, **39** 466–478.
- WILLIAMS, R. J. (1996). On the approximation of queueing networks in heavy traffic. In *Stochastic Networks: Theory and Applications* (F. P. Kelly, S. Zachary and I. Ziedins, eds.). Royal Statistical Society, Oxford University Press.
- WILLIAMS, R. J. (1998a). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems: Theory and Applications*, **30** 27–88.
- WILLIAMS, R. J. (1998b). An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Systems: Theory and Applications*, **30** 5–25.
- WILLIAMS, R. J. (2000). On dynamic scheduling of a parallel server system with complete resource pooling. In *Analysis of Communication Networks: Call Centres, Traffic and Performance* (D. R. McDonald and S. R. E. Tuners, eds.), vol. 8 of *Fields Institute Communications*. American Mathematical Society, 49–71.
- WRIGHT, S. (1997). *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia, PA.

## A An Equivalent Dual Formulation for the Static Planning Problem

In this section we describe an equivalent dual formulation for the static planning problem (5)-(9). The equivalent dual formulation will be useful for our proofs in Appendix B. Throughout this section, we will consider an arbitrary stochastic processing network, so the results developed here can be applied to each of the  $r$ th network in the network sequence that we discussed in Section 4 and they can also be applied to the limit network. The main result of this section is the following lemma.

**Lemma 12.** Suppose  $\rho$  is the optimal objective value to the static planning problem (5)-(9). Then  $(y^*, z^*)$  is optimal to the dual problem (10)-(14) if and only if  $y^*$  satisfies

$$\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_I} y_i^* R_{ij} a_j = -\rho \quad (97)$$

and

$$\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} y_i^* R_{ij} a_j = 1, \quad (98)$$

and  $\{z_k^*, k \in \mathcal{K}_I\}$  and  $\{z_k^*, k \in \mathcal{K}_S\}$  are, respectively, optimal solutions to

$$\min - \sum_{k \in \mathcal{K}_I} z_k \quad (99)$$

$$\text{s.t. } \sum_{i \in \mathcal{I}} y_i^* R_{ij} \leq - \sum_{k \in \mathcal{K}_I} z_k A_{kj}, \text{ for each input activity } j; \quad (100)$$

and

$$\min \sum_{k \in \mathcal{K}_S} z_k \quad (101)$$

$$\text{s.t. } \sum_{i \in \mathcal{I}} y_i^* R_{ij} \leq \sum_{k \in \mathcal{K}_S} z_k A_{kj}, \text{ for each service activity } j, \quad (102)$$

$$z_k \geq 0, \text{ for each service processor.} \quad (103)$$

*Proof.* For the “only if” part, we assume  $(x^*, \rho)$  and  $(y^*, z^*)$  are an optimal dual pair for the static planning problem and its dual problem. We will show that  $y^*$  satisfies (97)-(98) and  $z^*$  is optimal to (99)-(100) and (101)-(103). We first show that

$$\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_I} y_i^* R_{ij} a_j \geq -\rho, \quad (104)$$

and

$$\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} y_i^* R_{ij} a_j \geq 1. \quad (105)$$

For this, we construct a feasible allocation  $a^*$  with  $a_j^* = \begin{cases} x_j^*, & j \in \mathcal{J}_I \\ x_j^*/\rho, & j \in \mathcal{J}_S \end{cases}$ . By the complementary slackness on the constraints (11) and (12), we have

$$\sum_{i \in \mathcal{I}, j \in \mathcal{J}_I} y_i^* R_{ij} x_j^* = - \sum_{k \in \mathcal{K}_I, j \in \mathcal{J}_I} z_k^* A_{kj} x_j^*, \quad (106)$$

and

$$\sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} y_i^* R_{ij} x_j^* = \sum_{k \in \mathcal{K}_S, j \in \mathcal{J}_S} z_k^* A_{kj} x_j^*. \quad (107)$$

By the complementary slackness on the constraints (7) and (8), we have

$$\sum_{j \in \mathcal{J}_I, k \in \mathcal{K}_I} z_k^* A_{kj} x_j^* = \sum_{k \in \mathcal{K}_I} z_k^* = \rho, \quad (108)$$

and

$$\sum_{j \in \mathcal{J}_S, k \in \mathcal{K}_S} z_k^* A_{kj} x_j^* = \rho \sum_{k \in \mathcal{K}_S} z_k^* = \rho. \quad (109)$$

The last equality in (108) is from the strong duality theorem; the optimal objective value of the dual problem equals the optimal objective value of the primal problem. The last equality in (109) follows from (13). Then from the definition of  $a^*$  and (106)-(109), it immediately follows that

$$\sum_{i \in \mathcal{I}, j \in \mathcal{J}_I} y_i^* R_{ij} a_j^* = \sum_{i \in \mathcal{I}, j \in \mathcal{J}_I} y_i^* R_{ij} x_j^* = -\rho$$

and

$$\sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} y_i^* R_{ij} a_j^* = \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} y_i^* R_{ij} x_j^* / \rho = 1,$$

which imply (104) and (105) because  $a^* \in \mathcal{A}$ .

Next we shall show that

$$\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_I} y_i^* R_{ij} a_j \leq -\rho,$$

and

$$\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} y_i^* R_{ij} a_j \leq 1.$$

For any  $a \in \mathcal{A}$ , we have

$$\sum_{i \in \mathcal{I}, j \in \mathcal{J}_I} y_i^* R_{ij} a_j \leq - \sum_{k \in \mathcal{K}_I, j \in \mathcal{J}_I} z_k^* A_{kj} a_j = - \sum_{k \in \mathcal{K}_I} z_k^* = -\rho.$$

The first inequality above follows from (11), and the non-negativity of  $a$ ; the second inequality holds since  $a \in \mathcal{A}$  and therefore  $\sum_{j \in \mathcal{J}_I} A_{kj} a_j = 1$  for each  $k \in \mathcal{K}_I$ ; the third is due to the strong duality theorem. Similarly, we have

$$\sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} y_i^* R_{ij} a_j \leq \sum_{k \in \mathcal{K}_S, j \in \mathcal{J}_S} z_k^* A_{kj} a_j \leq \sum_{k \in \mathcal{K}_S} z_k^* = 1.$$

The first inequality above follows from (12); the second follows from  $\sum_{j \in \mathcal{J}_S} A_{kj} a_j \leq 1$  and  $z_k \geq 0$  for each  $k \in \mathcal{K}_S$ ; and the third is due to (13).

Hence  $y^*$  satisfies (97) and (98). To see  $z^*$  is an optimal solution to (99)-(100) and (101)-(103), we consider the following problems:

$$\max \left\{ \sum_{i \in \mathcal{I}, j \in \mathcal{J}_I} y_i^* R_{ij} a_j : \sum_{j \in \mathcal{J}_I} A_{kj} a_j = 1 \text{ for each input processor } k \right\}, \quad (110)$$

and

$$\max \left\{ \sum_{i \in \mathcal{I}, j \in \mathcal{J}_S} y_i^* R_{ij} a_j : \sum_{j \in \mathcal{J}_S} A_{kj} a_j \leq 1 \text{ for each service processor } k \right\}. \quad (111)$$

It is easy to see that the above two problems are equivalent to (97) and (98). Furthermore they are the dual problems of (99)-(100) and (101)-(103). This implies that the optimal objective values to (99)-(100) and (101)-(103) are  $-\rho$  and 1 respectively. Because  $(y^*, z^*)$  is an optimal solution to (10)-(14), (100) and (102) are satisfied by  $z^*$ ,  $\sum_{k \in \mathcal{K}_I} z_k^* = \rho$ , and  $\sum_{k \in \mathcal{K}_S} z_k^* = 1$ . This implies that  $z^*$  is feasible to (99)-(100) and (101)-(103) with respective objective values  $-\rho$  and 1. Therefore,  $z^*$  is optimal to (99)-(100) and (101)-(103).

For the “if” part, let  $(y^*, z^*)$  be such that  $y^*$  satisfies (97) and (98) and  $z^*$  is optimal to (99)-(100) and (101)-(103). Because (99)-(100) and (101)-(103) are dual problems of the equivalent formulation of (97) and (98),

$$\sum_{k \in \mathcal{K}_I} z_k^* = \rho, \quad (112)$$

and

$$\sum_{k \in \mathcal{K}_S} z_k^* = 1. \quad (113)$$

The fact that  $z^*$  is feasible to (99)-(100) and (101)-(103), together with (113), implies that  $(y^*, z^*)$  is feasible to the dual problem (10)-(14). Furthermore, the corresponding objective value is  $\rho$  because of (112). This implies that  $(y^*, z^*)$  is optimal to (10)-(14).  $\square$

Lemma 12 immediately implies Lemma 4 in Section 6 because the problem  $\max_{a \in \mathcal{A}} y^* \cdot Ra$  can be decomposed into problems (110) and (111).

## B Proofs of Lemmas and Theorem 1

In this section, we provide the proofs for Lemmas 1, 6, 8, and 11, and Theorem 1.

*Proof of Lemma 1.* From Lemma 4 and Assumption 2, we have

$$\max_{a \in \mathcal{A}} y^r \cdot Ra = 1 - \rho^r.$$

On the other hand,  $T^r(t)/t \in \mathcal{A}$ . Hence  $(1 - \rho^r)t \geq y^r \cdot RT^r(t)$ , which implies  $Y^r$  is nonnegative. Similarly, for any  $0 \leq t_1 \leq t_2$ ,  $(T^r(t_2) - T^r(t_1))/(t_2 - t_1) \in \mathcal{A}$ , and we have

$$y^r \cdot R(T^r(t_2) - T^r(t_1))/(t_2 - t_1) \leq (1 - \rho^r).$$

It follows that  $Y^r(t_2) - Y^r(t_1) \geq 0$ .  $\square$



*Proof of Lemma 6.* It is natural to work in a more general setting. Consider an i.i.d. sequence of nonnegative random variables  $\{v_\ell, \ell = 1, 2, \dots\}$  with mean  $1/\mu_v$ . Assume  $v_\ell$  have finite  $2 + \epsilon_v$  moments for some  $\epsilon_v > 0$ . That is, there exists some  $\hat{\sigma} < \infty$  such that  $\mathbb{E}(v_\ell^{2+\epsilon_v}) = \hat{\sigma}$ . Let  $V(n) = \sum_{\ell=1}^n v_\ell, n \in \mathbb{Z}^+$ . Define the renewal process associated with  $V(n)$  as  $R(t) = \max\{n : V(n) \leq t\}$ . Let  $v^{r,T,max} = \max\{v_\ell : 1 \leq \ell \leq R(r^2T) + 1\}$ . We first show that

$$v^{r,T,max}/r \rightarrow 0 \text{ with probability 1.} \quad (114)$$

By the strong law of large numbers, with probability 1,

$$R(t)/t \rightarrow \mu_v \text{ as } t \rightarrow \infty, \quad (115)$$

and

$$\sum_{\ell=1}^n v_\ell^{2+\epsilon_v}/n \rightarrow \hat{\sigma} \text{ as } n \rightarrow \infty. \quad (116)$$

Choose a sample path such that both (115) and (116) are satisfied. Then

$$(v^{r,T,max})^{2+\epsilon_v} = \max_{1 \leq \ell \leq R(r^2T)+1} v_\ell^{2+\epsilon_v} \leq \sum_{1 \leq \ell \leq R(r^2T)+1} v_\ell^{2+\epsilon_v}.$$

Because  $R(r^2T) \rightarrow \infty$  as  $r \rightarrow \infty$ , we have

$$\lim_{r \rightarrow \infty} \sum_{1 \leq \ell \leq R(r^2T)+1} v_\ell^{2+\epsilon_v} / (R(r^2T) + 1) = \hat{\sigma}.$$

Therefore,

$$\limsup_{r \rightarrow \infty} \frac{(v^{r,T,max})^{2+\epsilon_v}}{r^2} \leq \lim_{r \rightarrow \infty} \left( \sum_{1 \leq \ell \leq R(r^2T)+1} \frac{v_\ell^{2+\epsilon_v}}{R(r^2T) + 1} \right) \frac{R(r^2T) + 1}{r^2} = \hat{\sigma} \mu_v T.$$

This implies that

$$\lim_{r \rightarrow \infty} \frac{(v^{r,T,max})^{2+\epsilon_v}}{r^{2+\epsilon_v}} = 0.$$

The result follows.

We next show that for any fixed  $\epsilon > 0$  and large enough  $n$ ,

$$\mathbb{P}(\|V(\ell) - \ell/\mu_v\|_\ell \geq \epsilon n) \leq \epsilon/n. \quad (117)$$

Because  $v_\ell$  have finite  $2 + \epsilon_v$  moments, one gets

$$\mathbb{E}\left(|V(\ell) - \ell/\mu_v|^{2+\epsilon_v}\right) \leq \kappa_v \ell^{1+\epsilon_v/2} \quad \text{for all } \ell \leq n,$$

where  $\kappa_v$  is some constant that depends just on  $\hat{\sigma}$  and  $\epsilon_v$  (cf. [Ata and Kumar, 2005](#), Lemma 8). Then from Chebyshev's inequality, we have, for each  $\ell \leq n$ ,

$$\mathbb{P}(|V(\ell) - \ell/\mu_v| \geq \epsilon n) \leq \kappa_v \ell^{1+\epsilon_v/2} / (\epsilon n)^{2+\epsilon_v} \leq \kappa_v / \epsilon^{2+\epsilon_v} n^{1+\epsilon_v/2}.$$

Choosing  $n$  large enough,

$$\mathbb{P}(|V(\ell) - \ell/\mu_v| \geq \epsilon n) \leq \epsilon/n.$$

Let

$$\tau = \min\{\ell : |V(\ell) - \ell/\mu_v| \geq n\epsilon\}.$$

Then, restarting the process at time  $\tau$ ,

$$\mathbb{P}(|V(n) - n/\mu_v| \leq \epsilon n/2 \mid \tau \leq n) \leq \mathbb{P}(|V(n) - V(\tau) - (n - \tau)/\mu_v| \geq \epsilon n/2 \mid \tau \leq n) \leq \epsilon/2n.$$

On the other hand,

$$\mathbb{P}(|V(n) - n/\mu_v| \leq \epsilon n/2) \geq 1 - \epsilon/2n.$$

It then follows that  $\mathbb{P}(\tau \leq n) \leq \epsilon/(2n - \epsilon) \leq \epsilon/n$ . This implies the result.

Finally, the inversion of (117) gives

$$\mathbb{P}(\|R(s) - \mu_v s\|_t \geq \epsilon t) \leq \epsilon/t \quad \text{for large enough } t. \quad (118)$$

Applying (114) and (118) to the utilized service times  $u_j(\ell)$ , one gets (59) and (61) in Lemma 6. Using (117) for each component of the routing vector  $\phi_i^j(\ell)$  yields (60).  $\square$

*Proof of Lemma 8.* We first show that  $\{y^r, z^r\}$  is bounded. Since  $z^r \geq 0$ ,  $\sum_{k \in \mathcal{K}_S} z_k^r = 1$ , and  $\sum_{k \in \mathcal{K}_I} z_k^r = \rho^r$  for all  $r$ , we have  $|z^r| \leq 1$ . To show  $\{y^r\}$  is bounded, we consider the following primal-dual pair:

$$\text{minimize} \quad \rho \quad (119)$$

$$\text{subject to} \quad Rx \geq be, \quad (120)$$

$$\sum_{j \in \mathcal{J}} A_{kj} x_j = 1 \text{ for each input processor } k, \quad (121)$$

$$\sum_{j \in \mathcal{J}} A_{kj} x_j \leq \rho \text{ for each service processor } k, \quad (122)$$

$$x \geq 0, \quad (123)$$

and

$$\text{maximize} \quad \sum_{k \in \mathcal{K}_I} z_k + b \sum_{i \in \mathcal{I}} y_i, \quad (124)$$

$$\text{subject to} \quad \sum_{i \in \mathcal{I}} y_i R_{ij} \leq - \sum_{k \in \mathcal{K}_I} z_k A_{kj}, \text{ for each input activity } j, \quad (125)$$

$$\sum_{i \in \mathcal{I}} y_i R_{ij} \leq \sum_{k \in \mathcal{K}_S} z_k A_{kj}, \text{ for each service activity } j, \quad (126)$$

$$\sum_{k \in \mathcal{K}_S} z_k = 1, \quad (127)$$

$$y \geq 0; \text{ and } z_k \geq 0, \text{ for each service processor } k. \quad (128)$$

The dual LP (124)–(128) is obtained by perturbing the objective function coefficients of the dual static planning problem (10)–(14). Because the dual static planning problem (10)–(14) has a unique optimal solution, for sufficiently small  $b > 0$ , the optimal solution of the dual LP (124)–(128) equals  $(y^*, z^*)$  (cf. Mangasarian, 1979, Theorem 1). Therefore, the primal problem (119)–(123) has an optimal solution  $(\hat{\rho}, \hat{x})$ . Now choose  $r$  large enough such that  $|R^r \hat{x} - R\hat{x}| < be/2$ . Then  $R^r \hat{x} \geq be/2$ . Consider the problem (119)–(123) with  $b$  replaced by  $b/2$  and  $R$  replaced by  $R^r$ . Because  $(\hat{\rho}, \hat{x})$  is a feasible solution, the optimal objective value  $\tilde{\rho}^r \leq \hat{\rho}$ . The corresponding dual problem of this new LP is the dual problem (124)–(128) with  $b$  in the objective function coefficients replaced by  $b/2$  and  $R$  replaced by  $R^r$ . The optimal objective value of this new dual LP equals  $\tilde{\rho}^r \leq \hat{\rho}$ . Because  $(y^r, z^r)$  is a feasible solution to the new dual LP,

$$\sum_{k \in \mathcal{K}_I} z_k^r + b/2 \sum_{i \in \mathcal{I}} y_i^r \leq \tilde{\rho}^r \leq \hat{\rho}.$$

This implies that  $\sum_i y_i^r \leq 2\hat{\rho}/b$  for large enough  $r$ , so  $\{y^r\}$  is bounded.

Then we only need to show that every convergent subsequence of  $\{(y^r, z^r)\}$  converges to  $(y^*, z^*)$ . Let  $(\hat{y}, \hat{z})$  be a limit point of any subsequence  $\{(y^{r_n}, z^{r_n})\}$ . We will verify that  $(\hat{y}, \hat{z})$  is an optimal solution to the dual static planning problem (10)–(14) of the limiting network. First, we show that they are feasible. Since  $\{(y^{r_n}, z^{r_n})\} \rightarrow (\hat{y}, \hat{z})$  as  $n \rightarrow \infty$ , for any  $\epsilon > 0$ , for large enough  $n$ ,  $|\hat{y} - y^{r_n}| < \epsilon$ ,  $|\hat{z} - z^{r_n}| < \epsilon$  and  $|R - R^{r_n}| < \epsilon$ . For each input activity  $j \in \mathcal{J}_I$ ,

$$\begin{aligned} \sum_{i \in \mathcal{I}} \hat{y}_i R_{ij} &\leq \sum_{i \in \mathcal{I}} y_i^{r_n} R_{ij}^{r_n} + \mathbf{I}\epsilon(|\hat{y}| + \sup_r |R^r|) \\ &\leq - \sum_{k \in \mathcal{K}_I} A_{kj} z_k^{r_n} + \mathbf{I}\epsilon(|\hat{y}| + \sup_r |R^r|) \\ &\leq - \sum_{k \in \mathcal{K}_I} A_{kj} \hat{z}_k + \mathbf{I}\epsilon(|\hat{y}| + \sup_r |R^r|) + \mathbf{K}\epsilon. \end{aligned}$$

Since  $\epsilon$  can be arbitrarily small, we have

$$\sum_{i \in \mathcal{I}} \hat{y}_i R_{ij} \leq - \sum_{k \in \mathcal{K}_I} A_{kj} \hat{z}_k \text{ for each input activity } j \in \mathcal{J}_I.$$

Similarly, one can verify that

$$\sum_{i \in \mathcal{I}} \hat{y}_i R_{ij} \leq \sum_{k \in \mathcal{K}_S} A_{kj} \hat{z}_k \text{ for each service activity } j \in \mathcal{J}_S,$$

and

$$\sum_{k \in \mathcal{K}_S} \hat{z}_k = 1.$$

Furthermore, because  $(y^{r_n}, z^{r_n})$  are optimal solutions,  $\sum_{k \in \mathcal{K}_I} z_k^{r_n} = \rho^{r_n}$ . Again we can show that

$$\sum_{k \in \mathcal{K}_I} \hat{z}_k = 1.$$

Therefore,  $(\hat{y}, \hat{z})$  is an optimal solution to the dual problem (10)-(14) of the limit network. Then by the uniqueness of the optimal solution, we conclude that  $(\hat{y}, \hat{z}) = (y^*, z^*)$ . Since the subsequence is arbitrary, we have  $(y^r, z^r) \rightarrow (y^*, z^*)$  as  $r \rightarrow \infty$ .  $\square$

*Proof of Lemma 11.* Similar to the proof of Lemma 8 above, we can prove that  $x^r \rightarrow x^*$  as  $r \rightarrow \infty$ . From the strict complementary theorem (cf. Wright, 1997), every pair of primal and dual LPs has a strict complementary optimal solution if they have optimal solutions. Hence we have the following relations: for the limit network,

$$\sum_{j \in \mathcal{J}_S} A_{kj} x_j^* = 1 \quad \text{is equivalent to} \quad z_k^* > 0, \quad \text{for all } k \in \mathcal{K}_S; \quad (129)$$

$$\sum_{i \in \mathcal{I}} y_i^* R_{ij} = \sum_{k \in \mathcal{K}_S} A_{kj} z_k^* \quad \text{is equivalent to} \quad x_j^* > 0, \quad \text{for all } j \in \mathcal{J}_S; \quad (130)$$

and for each  $r$ ,

$$z_k^r > 0 \quad \text{implies} \quad \sum_{j \in \mathcal{J}_S} A_{kj} x_j^r = 1 - \rho^r \quad \text{for all } k \in \mathcal{K}_S; \quad (131)$$

$$x_j^r > 0 \quad \text{implies} \quad \sum_{i \in \mathcal{I}} y_i^r R_{ij} = \sum_{k \in \mathcal{K}_S} A_{kj} z_k^r \quad \text{for all } j \in \mathcal{J}_S. \quad (132)$$

Since  $x^r \rightarrow x^*$  as  $r \rightarrow \infty$ , for large enough  $r$ ,  $x_j^r > 0$  if  $x_j^* > 0$ . This, together with (130) and (132), implies that for each  $j \in \mathcal{J}$ ,

$$y_i^* R_{ij} = \sum_{k \in \mathcal{K}_S} A_{kj} z_k^* \quad \text{implies} \quad \sum_{i \in \mathcal{I}} y_i^r R_{ij} = \sum_{k \in \mathcal{K}_S} A_{kj} z_k^r \quad \text{for large enough } r. \quad (133)$$

Suppose  $z_k^* = 0$  for some  $k \in \mathcal{K}_S$ , then  $\sum_{j \in \mathcal{J}_S} A_{kj} x_j^* < 1$ . There exists an  $\epsilon > 0$  such that  $\sum_{j \in \mathcal{J}_S} A_{kj} x_j^* = 1 - \epsilon$ . For large enough  $r$ , we have  $\sum_{j \in \mathcal{J}_S} A_{kj} x_j^r \leq \sum_{j \in \mathcal{J}_S} A_{kj} x_j^* + \epsilon/2 \leq 1 - \epsilon/2$  because  $x^r \rightarrow x^*$  as  $r \rightarrow \infty$ . This implies  $z_k^r = 0$  for large enough  $r$ . Therefore,

$$z_k^* = 0 \quad \text{implies} \quad z_k^r = 0 \quad \text{for large enough } r. \quad (134)$$

Because  $(y^*, z^*)$  is the optimal solution to the dual static planning problem (10)-(14), we have for each  $a \in \mathcal{E}$ ,

$$\sum_{j \in \mathcal{J}} \left( \sum_{i \in \mathcal{I}} y_i^* R_{ij} \right) a_j \leq \sum_{j \in \mathcal{J}_S} \left( \sum_{k \in \mathcal{K}_S} A_{kj} z_k^* \right) a_j - \sum_{j \in \mathcal{J}_I} \left( \sum_{k \in \mathcal{K}_I} A_{kj} z_k^* \right) a_j \leq 0. \quad (135)$$

Since  $a^* \in \mathcal{E}^*$ ,  $y^* \cdot Ra^* = \max_{a \in \mathcal{E}} y^* \cdot Ra = 0$ . It follows that both inequalities in (135) are equalities and therefore,

$$\sum_{i \in \mathcal{I}} y_i^* R_{ij} = \sum_{k \in \mathcal{K}_S} A_{kj} z_k^* \quad \text{for all } j \in \mathcal{J}_S \text{ with } a_j^* > 0, \quad (136)$$

$$\sum_{j \in \mathcal{J}_S} A_{kj} a_j^* = 1 \quad \text{for all } k \in \mathcal{K}_S \text{ with } z_k^* > 0. \quad (137)$$

From (133) and (136), we have for large enough  $r$ ,

$$\sum_{i \in \mathcal{I}} y_i^r R_{ij}^r = \sum_{k \in \mathcal{K}_S} A_{kj} z_k^r \text{ for all } j \in \mathcal{J} \text{ with } a_j^* > 0.$$

Therefore,

$$\sum_{j \in \mathcal{J}} a_j^* \sum_{i \in \mathcal{I}} y_i^r R_{ij}^r = \sum_{j \in \mathcal{J}} a_j^* \sum_{k \in \mathcal{K}_S} A_{kj} z_k^r \text{ for large enough } r. \quad (138)$$

From (134) and (137), it follows that for large enough  $r$ ,

$$\sum_{j \in \mathcal{J}_s} a_j^* A_{kj} = 1 \text{ for all } k \in \mathcal{K}_S \text{ with } z_k^r > 0.$$

Therefore,

$$\sum_{k \in \mathcal{K}_S} z_k^r \sum_{j \in \mathcal{J}_s} a_j^* A_{kj} = \sum_{k \in \mathcal{K}_S} z_k^r \text{ for large enough } r. \quad (139)$$

It follows from  $\sum_{j \in \mathcal{J}_I} a_j^* A_{kj} = 1$  for each  $k \in \mathcal{K}_I$ ,  $\sum_{j \in \mathcal{J}_I} z_k^r = \rho^r$ , (138) and (139) that

$$\sum_{j \in \mathcal{J}} a_j^* \sum_{i \in \mathcal{I}} y_i^r R_{ij}^r = \sum_{j \in \mathcal{J}} a_j^* \sum_{k \in \mathcal{K}_S} A_{kj} z_k^r = 1 - \rho^r \text{ for large enough } r.$$

Then Lemma 11 follows from the fact that  $\max_{a \in \mathcal{E}} y^r \cdot R^r a = 1 - \rho^r$ .  $\square$

*Proof of Theorem 1.* We define the scaled process  $\bar{\bar{\mathbb{X}}}^r$  via

$$\bar{\bar{\mathbb{X}}}^r(t) = r^{-2} \mathbb{X}(r^2 t) \text{ for each } t \geq 0.$$

Fix a sample path that satisfies the strong law of large numbers for  $u_j$  and  $\phi_j^i$ . Let  $(\bar{\bar{Z}}, \bar{\bar{T}})$  be a fluid limit of  $(\bar{Z}^r, \bar{T}^r)$  along the sample path. Following the arguments in Section A.2 in Dai and Lin (2005), such a limit exists and satisfies the fluid model equations (40)–(44), presented in Section 6. Under the maximum pressure policy with parameter set  $(\alpha, \beta)$ , each fluid limit  $(\bar{\bar{Z}}, \bar{\bar{T}})$  also satisfies the fluid model equation (45). The justification of fluid model equation (45) is similar to Lemma 4 in Dai and Lin (2005), with the scaling  $r$  replaced by  $r^2$ . Therefore,  $(\bar{\bar{Z}}^r, \bar{\bar{T}}^r)$  is a fluid model solution under the maximum pressure policy. Similar to the proof of Theorem 4 in Dai and Lin (2005), using a different Lyapunov function  $f(t) = (\alpha \times \bar{\bar{Z}}(t)) \cdot \bar{\bar{Z}}(t)$ , one can easily prove that the fluid model under the maximum pressure policy with a general parameter set  $(\alpha, \beta)$  is weakly stable; namely, any fluid model solution  $(\bar{\bar{Z}}^r, \bar{\bar{T}}^r)$  under the maximum pressure policy satisfies  $\bar{\bar{Z}}(t) = 0$  for each  $t \geq 0$  given  $\bar{\bar{Z}}(0) = 0$ . As a consequence, we have for any  $t > 0$ ,  $\bar{\bar{T}}(t)/t$  satisfies (6)–(9) with  $\rho = 1$ . Because  $x^*$  is the unique optimal solution to the static planning problem (5)–(9) with objective value equal to 1,  $\bar{\bar{T}}(t)/t = x$  and  $\bar{\bar{T}}(t) = xt$  for each  $t \geq 0$ . Since this is true for any fluid limit, we have  $\bar{\bar{T}}^r(t) \rightarrow x^*t$  for each  $t$  with probability 1, which implies asymptotic efficiency.  $\square$