Rate Stability and Output Rates in Queueing Networks with Shared Resources

M. Jonckheere [a], R.D. van der Mei [a,b] and *W. van der Weij [a]

[a] CWI, Probability and Stochastic Networks, Amsterdam, The Netherlands
[b] Vrije Universiteit, Faculty of Sciences, Amsterdam, The Netherlands

April 14, 2009

Abstract

Motivated by a variety of applications in information and communication systems, we consider queueing networks in which the service rate at each of the *individual* nodes depends on the state of the *entire* system. The asymptotic behaviour of this type of networks is fundamentally different from classical queueing networks, where the service rate at each node is usually assumed to be independent of the state of the other nodes. We study the per-node rate stability and output rates for a class networks with a general capacity allocation function. More specifically, we derive necessary conditions of per-node rate stability, and give bounds for the per-node output rate and asymptotic growth rates, under mild assumptions on the allocation function. For a set of parallel nodes, we further prove the convergence of the output rates for most parameters and give a sharp characterization of the per-node rate stability. The results provide new intuition and fundamental insight in the stability and throughput behavior of queueing networks with shared resources.

Key words:
queueing networks, state-dependent allocation, rate stability, output rate, growth rate.

AMS 2000 subject classification:
primary 60M20; 60K25, secondary 90B22.

1 Introduction

The analysis of queueing networks has been subject to extensive research for the past few decades and has been successfully applied in many application areas. In a vast majority of papers however, it is assumed that the service rate at each of the nodes of the network is fixed. For example, in FCFS-based single- or multi-server nodes, non-idling servers are usually assumed to be autonomous entities that operate at a fixed rate, independent of the state of the other queues in the network. For the class of so-called Jackson networks [**?**], many stability and performance issues are well understood.

*Corresponding author. CWI, Kruislaan 413, 1098SJ Amsterdam, Netherlands. E-mail: weij@cwi.nl.

In this paper, we study queueing networks in which the service rates at each of the *individual* nodes are not independent, but depend on the state of the *entire* system, according to some general capacity allocation function. For this type of models, exact structural results are rare, and fundamental insight and intuition for seemingly simple questions about stability and throughput are lacking. This motivates an in-depth study of the per-node stability for queueing networks with a general class of capacity allocation functions.
Another source of motivation stems from applications in modern computer-communication systems, in which many heterogeneous applications share parts of the available infrastructure resources. In such environments, different applications compete for access to shared resources, both at the *software* level (e.g., mutex and database locks, thread pools) and at the *hardware* level (e.g., bandwidth, processing power, disk access). For example, many Web-based services are based on multi-tiered system architectures consisting of a client tier to provide an end-user interface, a business logic tier to coordinate information retrieval and processing, and a data tier with legacy systems to store and access customer data. Each end-user initiated Web transaction typically consists of several sub-transactions that have to be processed in some fixed or probabilistic order. To this end, application servers usually implement a number of thread pools, each of which is dedicated to performing a specific sub-transaction. A particular feature of the Web server performance model proposed in [**?**, **?**] is that at any moment in time the active (i.e., non-idling) threads share a common Central Processing Unit (CPU) hardware in a PS fashion. Other examples of performance models in which software resources compete for access to shared hardware resources are presented in [**?**, **?**]. Another interesting line of research in which the service rates among different network nodes are dependent is focused on bandwidth-sharing networks [**?**, **?**], providing a natural modeling framework for describing the dynamic flow-level interaction among elastic data transfers in communication networks. Queueing models with shared resources also occur naturally in the modeling of the flow-level performance in wireline data networks where the capacity of different links are shared among competing flows [**?**], or in wireless networks, where a limited amount of bandwidth is shared among different users, and where customers can communicate via a cascade of intermediate hops (cf. [**?**]).

A considerable amount of work has been dedicated to the stability of queueing networks (see for instance [**?**, **?**, **?**]). Controlling overload situations is essential for the design of communication networks. A well-engineered network should of course avoid to experience overload. However, the traffic fluctuations over time might lead to temporary surges that a well-designed network should deal with. A fine understanding of the behaviour of the network in overload is hence strongly needed. In particular, it is a fundamental issue to characterize, for given traffic conditions, which queues are going to get instable and what are the asymptotic growth rates. Recent results including a sharp characterization of per-node stability for parallel nodes with a decreasing service allocation have been obtained in [**?**]. It clearly emerges from these papers that general results for per-node stability for multi-layered networks (or networks with bandwidth sharing) appear to be very challenging. In particular, if global stability is well known for work-conserving networks, detailed (per-node) stability remains a difficult problem. For general service allocations without monotonicity properties, it is to the best of our knowledge an open problem, even for exponentially distributed services. Instead of focusing on stochastic stability, an alternative approach to tackle stability issues is to weaken the stability definition and to investigate the so-called rate stability of the network [**?**]. Roughly speaking, it consists of characterizing the growth rates as linear or sub-linear. However, because in a great number of practical situations, an overload situation is characterized by a linear asymptotic per node

growth rate, rate stability provides a precious benchmark information in cases where a more detailed stability description is almost hopeless. Using a similar line of thought, Egorova et al. [**?**] give a partial characterization of the overload behavior, for the wide class of so-called α-fair bandwidth sharing strategies defined in [**?**],by examining the fluid limit by suitable scaling the number of flows in the system, and give a fixed-point equation for the corresponding asymptotic growth rates.

In this paper we consider a queueing network with Poisson arrivals, exponential service-time distributions at all nodes, internal feed-forward routing and with a structured work-conserving allocation function driving the service in all nodes, that depend on the state of the entire system. For this general model, we (1) derive necessary conditions of the per-node rate stability, and (2) give bounds for the per-node output rate. We show how to use these conditions on a two-node tandem network to get necessary and sufficient conditions of rate stability. For a set of parallel nodes, we further prove the convergence of the output rates for almost all input parameters and give a sharp characterization of the per-node rate stability. The results provide new intuition and fundamental insight in the stability and throughput behavior of queueing networks with shared resources, which is essential to design effective overload-control mechanisms.

The remainder of this paper is organized as follows. In Section 2 the model is described and the relevant notation and definitions are introduced. In particular, the difference between stochastic and rate stability is rigorously explained. In Section 3, asymptotic values as output rates and growth rates are defined. Using the structure of the considered allocation functions, important properties of these output rates are derived. In Section 4, some traffic inequalities are established leading to necessary conditions for the rate stability of each node. We illustrate the obtained results on a toy example. In Sections 5, we consider the special case parallel nodes (no routing) with monotone allocations, and show that the necessary conditions derived in Sections 3 and 4 are also sufficient, for 'almost' all parameters. Finally, in Section 6 we address a number of challenging topics for further research.

2 Model and stability definitions

2.1 Network model

We consider an open queueing network with N queues. A job present at queue i is said to be of class i ($i = 1, \ldots, N$). External jobs arrive at queue i according to a Poisson process with intensity $\lambda_i \geq 0$. Denote the vector of external arrival rates by $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_N)^\top$. The service times at queue i are exponentially distributed with mean $\beta_i = 1/\mu_i$. Let $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_N)$. The state of the system is described by a vector $\mathbf{x} := (x_1, \ldots, x_N)$, where x_i represents the number of jobs of class i present in the system. Let $\mathbf{x} \in \mathcal{X} := \{0, 1, \ldots, \}^N$. When the system is in state $\mathbf{x}$, jobs of class i receive a service rate $\phi_i(\mathbf{x})$, where the function $\boldsymbol{\phi}(\mathbf{x}) := (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x}))$ is referred to as the system *capacity allocation function* . It is important to note that the various job classes are coupled since their *individual* service rates may depend on the state $\mathbf{x}$ of the *entire* system. The queue discipline is assumed to have no prior knowledge about the actual service requirements.

2.1.1 Assumptions on the routing

After receiving service at queue i, a job is routed to queue $j \in \mathcal{I} := \{0, 1, \ldots, N\}$ with probability p_{ij}. We adopt the convention that when $j = 0$, the job simply leaves the network. Denote the N-by-N routing matrix by $P := (p_{ik})$, where $i,\ k = 1, \ldots, N$. We assume that there is *no loop* in the routing, i.e., once a job has been served at a given queue, he never returns to this queue. This type of routing is often referred to as *feed-forward* routing. Consequently, we can order the queues such that $p_{ij} = 0,\ 1 \le j \le i$. The routing matrix P is substochastic, so that $R := (r_{ij}) := (I - P)^{-1}$ exists, where I is the N-by-N identity matrix. Moreover, let $D = (d_{ij})$ be the N-by-N diagonal matrix with diagonal entries, $d_{ii} := \frac{1}{\mu_i}$ $(i = 1, \ldots, N)$. Using these definitions, the load offered to queue i is given by

$$\rho_i := \boldsymbol{\lambda}^\top RD\mathbf{e}_i = \frac{1}{\mu_i} \sum_{j=1}^{N} \lambda_j r_{ji}, \tag{1}$$

where $\mathbf{e}_i$ is the i-th unit vector. Furthermore, denote $\rho = \sum_{i=1}^{N} \rho_i$.

Let $\mathbf{X}(t) := (X_1(t), \ldots, X_N(t))$, where $X_i(t)$ denotes the number of jobs at queue i (i.e., either waiting or being served) at time $t \ge 0$. Then the N-dimensional process $\{\mathbf{X}(t), t \ge 0\}$ can be described as a continuous-time Markov process with state space $\mathcal{X}$. For a subset of indices $\mathcal{S}$, we denote by $\mathbf{x}_\mathcal{S}$ the restriction of the vector $\mathbf{x}$ to queues $\mathcal{S}$, i.e., $\mathbf{x}_\mathcal{S} = (x_i)_{i \in \mathcal{S}}$.

2.1.2 Assumptions on the service rates

Throughout the chapter, the system allocation function $\boldsymbol{\phi}(\mathbf{x})$ satisfies certain assumptions that we describe here.

Assumption 1 (Work-conserving allocation)**.** *Whenever the system is not empty, all capacity is assigned to the queues. For* $\mathbf{x} \neq \mathbf{0} = (0, \ldots, 0)$,

$$\sum_{i=1}^{N} \frac{\phi_i(\mathbf{x})}{\mu_i} = 1, \quad \text{and} \quad \boldsymbol{\phi}(\mathbf{0}) := 0. \tag{2}$$

Without loss of generality, the total capacity of the system is assumed to be equal to 1 in (2).

Assumption 2 (Symmetric uniform limits)**.** *For all subsets of indices* $\mathcal{U} \subseteq \{1, \ldots, N\}$ *and* $\mathcal{S} = \{1, \ldots, N\} \setminus \mathcal{U}$*, there exists a function* $g^\mathcal{U}$ *on* $\{0, 1, \ldots, \}^{N - |\mathcal{U}|}$ *and some strictly positive numbers* $l_i, i \in \mathcal{U}$ *such that*

$$\forall j \in \mathcal{U}, \ \lim_{x_i \to \infty} \frac{\phi_j(\mathbf{x})}{\mu_j} = l_j g^\mathcal{U}(\mathbf{x}_\mathcal{S}). \tag{3}$$

Note that l_j does not depend on the set $\mathcal{U}$. In many applications in computer-communication systems the allocation functions have the following structure which is a special case of work-conserving allocations with symmetric uniform limits

$$\frac{\phi_i(\mathbf{x})}{\mu_i} = \frac{f_i(x_i)}{\sum_{j=1}^{N} f_j(x_j)}, \quad \mathbf{x} \in \mathcal{X},\ \mathbf{x} \neq \mathbf{0}, \tag{4}$$

where $f_i(\cdot)$ is a non-negative function such that $f_i(0) := 0$ and $\lim_{x_i \to \infty} f_i(x_i) =: l_i < \infty$ $(i = 1, \ldots, N)$. Note that in this case, Assumption 2 implies that

$$\forall \mathcal{U} \subset \{1, \ldots, N\}, \quad g^{\mathcal{U}}(\mathbf{x}_{\mathcal{S}}) = \left(\sum_{i \in \mathcal{U}} l_i + \sum_{i \notin \mathcal{U}} f_i(x_i) \right)^{-1}. \tag{5}$$

In the sequel, we refer to these allocations as extended processor-sharing allocations. Here are a few examples that have become classic models in queueing theory and performance evaluation

1. The *limited processor-sharing* allocation defined by

   $$f_i(x_i) = \min\{x_i, l_i\},$$

   where l_i is a positive integer.

2. The *limited discriminatory processor-sharing* allocation defined by

   $$f_i(x_i) = w_i \, \min\{x_i, C_i\},$$

   where C_i is a positive integer and $w_i > 0$ is a weight given to class i. In this case $l_i = w_i C_i$. share of the capacity to each class the classes are weighted.

3. The *coupled processors* allocation defined by

   $$f_i(x_i) = l_i 1_{\{x_i > 0\}},$$

   where $0 < l_i < +\infty$ is a weight associated with class i. In the literature, this allocation is sometimes referred to as the generalized processor-sharing (GPS) allocation.

The Assumptions 1 and 2 are not sufficient in general to get a sharp characterization of the rate stability set of the network. To get more precise results, we may assume the following condition, (which is naturally verified in the context of bandwidth sharing networks):

Assumption 3 (Monotonicity)**.** *ϕ_i is decreasing in x_j, $j \neq i$.*

2.2 Stability definitions

The study of stability of stochastic processes traditionally deals with the question of existence of a measure that is invariant to the transition operator of the process and to which the process converges in distribution or in total variation. We aim here at describing some 'per-queue' stability properties, i.e., properties of the processes $\{X_i(s), s \geq 0\}$, for $i = 1, \ldots, N$. Since the process $\{X_i(s), s \geq 0\}$ is not by itself a Markov process, this is generally a much more ambitious question than describing the global stability (stability of $\mathbf{X}(t)$) which is well-known for work-conserving networks (see Theorem 1 below). To the best of our knowledge, the only per-queue stochastic stability results have been obtained for a set of parallel queues with decreasing allocations and there is no such result available for the general type of networks we consider here. Because the usual definitions of stochastic stability did not lead so far (without stricter assumptions on the allocation function and the topology) to tractable results, we turn our attention to a weaker definition of stability that allows to give practical answers. Hence, we are primarily concerned with the property of the conservation of rates through the network. Roughly speaking, it consists of characterizing the asymptotic growth rates as linear or sub-linear and to characterize

the set of input parameters such that the incoming traffic at a queue is equal to the outgoing traffic. Interesting as a first-order stability property, rate stability turns out to also give useful necessary conditions for stochastic instability. For later reference, we thus define the following two notions of stability: rate stability and the stronger notion of stochastic stability.

From Assumption 2 the allocation functions $\phi_i(.)$, and hence the transition rates are bounded, and thus the process X is non-explosive. Hence we may assume that X and all other stochastic processes treated in the sequel have paths in the space $D = D(\mathcal{R}_+, \mathcal{Z}_+^N)$ of right-continuous functions from $\mathcal{R}_+$ to $\mathcal{Z}_+^N$ with finite left limits.

Definition 1 (Rate stability)**.** *The process $\{X_i(t), t \geq 0\}$ is said to be rate stable if*

$$\liminf_{t\to\infty} \frac{X_i(t)}{t} = 0 \quad a.s.,$$

and the process is called rate unstable if

$$\liminf_{t\to\infty} \frac{X_i(t)}{t} > 0 \quad a.s.$$

Definition 2 (Stochastic stability)**.** *The process $\{X_i(t), t \geq 0\}$ is said to be stochastically stable if*

$$\lim_{r\to\infty} \sup_{t\to\infty} \Pr\{X_i(t) > r\} = 0,$$

and the process is called stochastically unstable if

$$\lim_{r\to\infty} \sup_{t\to\infty} \Pr\{X_i(t) > r\} > 0.$$

Moreover, the N-dimensional process $\{\mathbf{X}(t), t \geq 0\}$ is said to be globally stochastically stable (or stochastically stable) if $\{X_i(t), t \geq 0\}$ is stochastically stable for all $i = 1, \ldots, N$.

The following result, characterizing the stochastic stability of the process $\{\mathbf{X}(s),\ s \geq 0\}$, is well-known for work-conserving networks. The total number of jobs can indeed be seen as the number of jobs of a single queue with unit service rate and the global stability is then a consequence of Loynes Theorem (cf., e.g., [**?**]).

Theorem 1 (Global stability)**.** *The network is globally stochastically stable if*

$$\sum_{i=1}^{N} \rho_i < 1.$$

The network is globally stochastically unstable if

$$\sum_{i=1}^{N} \rho_i > 1.$$

Definition 3 (Rate stability subsets)**.** *Let $\mathcal{S} := \{i : \{X_i(t), t \geq 0\}$ is rate stable$\}$, and $\mathcal{U} := \{i :$ $\{X_i(t), t \geq 0\}$ is rate unstable$\}$.*

Since each queue is either rate stable or rate unstable, the index set $\{1,\ldots,N\}$ is partitioned into the tuple $\mathcal{P} := (\mathcal{S},\mathcal{U})$, with $\mathcal{S}\cup\mathcal{U} = \{1,\ldots,N\}$, $\mathcal{S}\cap\mathcal{U} = \emptyset$. In case of rate stability, the number of jobs at queue i grows asymptotically 'slower than t' when t goes to infinity, at least on some trajectories. In case of stochastic stability, the process $\{X_i(t), t \geq 0\}$ remains in a finite neighborhood with positive probability. Remark that if $\{X_i(t), t \geq 0\}$ is an irreducible Markov process, then stochastic stability is equivalent to requiring $\{X_i(t), t \geq 0\}$ to be positive recurrent (see for example Theorem 12.25 in [**?**]). Note also that stochastic stability implies rate stability, as it should, but that the converse result is generally not true.

The next result underlines the relation between rate instability and stochastic instability.

Proposition 1. *For $i = 1,\ldots,N$, $\liminf_{t\to\infty} \frac{X_i(t)}{t} > 0$ implies that $X_i(t) \to \infty$ in probability.*

Proof: Suppose that $X_i(t)$ does not diverge to infinity in probability. Then there exists a subsequence $\{t_n, n = 0,1,\ldots\}$ such that $X_i(t_n) \to Z_i$ (in probability) for some honest (i.e., almost surely finite) random variable Z_i. Moreover, there exists another subsequence $\{t'_n, n = 0,1,\ldots\}$ such that $\{X_i(t'_n)\} \to Z_i$ almost surely [**?**]. Hence, $\frac{X_i(t'_n)-Z_i}{t'_n} \to 0$ almost surely and since Z_i is almost surely finite, $\frac{Z_i}{t'_n} \to 0$ and $\frac{X_i(t'_n)}{t'_n} \to 0$ almost surely, which implies that $\liminf_{t\to\infty} \frac{X_i(t)}{t} = 0$, almost surely. □

Remark 1. *Many authors (see for instance [**?**, **?**, **?**]) define rate stability differently, with slightly stronger assumptions. For the purpose of our analysis, we prefer the given definition that allows to link rate instability to the fact that a process is diverging to infinity.*

3 Output rates and growth rates

3.1 Definition

The following notation is useful in the sequel. For a given sample path $\{\mathbf{X}(s), s > 0\}$, we define the *Cesaro mean service rate* at each queue of the network by

$$\varphi_i(t) \;:=\; \frac{1}{t}\int_0^t \phi_i(\mathbf{X}(s))ds, \quad i = 1,\ldots,N, t > 0. \tag{6}$$

The *growth rate* of queue i is defined by

$$Y_i(t) \;:=\; \frac{X_i(t)}{t}, \quad i = 1,\ldots,N, t > 0. \tag{7}$$

Over a given sample path $\{\mathbf{X}(s), s > 0\}$, we can further define the limiting values of the mean service rate

$$\underline{\varphi}_i := \liminf_{t\to\infty}\ \varphi_i(t), \quad \bar{\varphi}_i := \limsup_{t\to\infty} \varphi_i(t), \quad i = 1,\ldots,N,$$

and the *asymptotic growth rate* of the queues

$$\underline{Y}_i = \liminf_{t\to\infty}\ Y_i(t), \quad \text{and} \quad \bar{Y}_i = \limsup_{t\to\infty}\ Y_i(t).$$

From Assumption 2, the random variables $\bar{\varphi}_i$ are bounded, and consequently, we prove in the following section that the $\bar{Y}_i$ are almost surely bounded. We may therefore define the mean values of vectors, for $i = 1,\ldots,N$,

$$O_i := \mathbb{E}[\underline{\varphi}_i], \quad \bar{O}_i := \mathbb{E}[\bar{\varphi}_i], \quad Q_i := \mathbb{E}[\underline{Y}_i], \quad \bar{Q}_i := \mathbb{E}[\bar{Y}_i], \tag{8}$$

and denote the corresponding vectors by $\mathbf{O} := (O_1, \ldots, O_N)^\top$, $\bar{\mathbf{O}} := (\bar{O}_1, \ldots, \bar{O}_N)^\top$, $\mathbf{Q} := (Q_1, \ldots, Q_N)^\top$ and $\bar{\mathbf{Q}} := (\bar{Q}_1, \ldots, \bar{Q}_N)^\top$. Note that rate stability of queue i implies that $\underline{\varphi_i} = 0$ (almost surely) and $Q_i = 0$. Moreover, note that if queue i is stochastically stable, then $\bar{Q}_i = Q_i = 0$ and $\bar{O}_i = O_i$.

3.2 Properties of the asymptotic rates

In this section we derive some properties of the rates of service obtained when a queue is rate unstable. These properties turn out to be crucial when characterizing the rate stability of the network. It is convenient to define, for $i = 1, \ldots, N$,

$$\eta_i := \mu_i l_i.$$

The next result gives a relation between the output rates and the fraction of capacity assigned for rate unstable queues. For a given stability partitioning of the queues $\mathcal{P} = (\mathcal{S}, \mathcal{U})$, denote

$$\bar{Z}_{\mathcal{P}} := \mathbb{E}\left[\limsup_{t\to\infty} \frac{1}{t}\int_0^t g^{\mathcal{U}}(\mathbf{X}_{\mathcal{S}}(s))ds\right].$$

Proposition 2. (Balanced output rates for rate-unstable queues) *Suppose that the allocation has symmetric uniform limits (assumption 2). Then if $i, j \in \mathcal{U}$, then*

$$\eta_j \bar{O}_i = \eta_i \bar{O}_j. \tag{9}$$

In particular if $l_i > 0$ and $l_j > 0$, then

$$\frac{\bar{O}_i}{\eta_i} = \frac{\bar{O}_j}{\eta_j} = \bar{Z}_{\mathcal{P}}. \tag{10}$$

Moreover, if $(\alpha_j)_{j\in\mathcal{U}}$ are positive numbers, then

$$\mathbb{E}\left[\limsup_{t\to\infty}\left(\sum_{j\in\mathcal{U}} \alpha_j \varphi_j(t)\right)\right] = \sum_{j\in\mathcal{U}} \eta_j \alpha_j \bar{O}_j.$$

Proof: For all $i \in \mathcal{U}$, X_i diverges in probability to infinity. As ϕ is bounded, it implies that $\frac{\phi_i(\mathbf{X}(s))}{\mu_i} - l_i g^{\mathcal{U}}(\mathbf{X}_{\mathcal{S}}(s)) \to 0$ (in L^1), which gives that

$$\mathbb{E}\left[\frac{1}{t}\int_0^t \left(\frac{\phi_i(\mathbf{X}(s))}{\mu_i} - l_i g^{\mathcal{U}}(\mathbf{X}_{\mathcal{S}}(s))\right) ds\right] \to 0.$$

Using the dominated convergence theorem, which allows us to interchange the expectation and the limit, we obtain that

$$\begin{aligned}&\mathbb{E}\left[\lim_{t\to\infty} \tfrac{1}{t}\int_0^t \left(\tfrac{\phi_i(\mathbf{X}(s))}{\mu_i} - l_i g^{\mathcal{U}}(\mathbf{X}_{\mathcal{S}}(s))\right) ds\right] \\ &= \lim_{t\to\infty} \mathbb{E}\left[\tfrac{1}{t}\int_0^t \left(\tfrac{\phi_i(\mathbf{X}(s))}{\mu_i} - l_i g^{\mathcal{U}}(\mathbf{X}_{\mathcal{S}}(s))\right) ds\right] = 0.\end{aligned}$$

We conclude by observing that

$$\begin{aligned}\mathbb{E}\left[\limsup_{t\to\infty} \tfrac{\varphi_i(t)}{\mu_i}\right] &= \mathbb{E}\left[\lim_{t\to\infty} \tfrac{1}{t}\int_0^t \left(\tfrac{\phi_i(\mathbf{X}(s))}{\mu_i} - l_i g^{\mathcal{U}}(\mathbf{X}_{\mathcal{S}}(s))\right) ds\right] \\ &\quad + l_i \mathbb{E}\left[\limsup_{t\to\infty} \tfrac{1}{t}\int_0^t g^{\mathcal{U}}(\mathbf{X}_{\mathcal{S}}(s))ds\right] \\ &= l_i \bar{Z}_{\mathcal{P}} \\ &= \tfrac{\eta_i}{\mu_i} \bar{Z}_{\mathcal{P}}.\end{aligned}$$

□The next two propositions compare the outputs of rate stable and rate unstable queues for asymptotically decreasing allocations.

Proposition 3. (Unbalanced rates between rate stable and rate unstable queues) *Suppose that the allocation is work conserving and decreasing (assumptions 1 and 3). Then ϕ_i is increasing in x_i and if $i \in \mathcal{S}$ and $j \in \mathcal{U}$, it holds that*

$$\eta_j \bar{O}_i \leq \eta_i \bar{O}_j. \tag{11}$$

Proof: As ϕ_i is increasing in x_i, $\phi_i(x) \leq l_i g(\mathbf{x}_{\mathcal{S}})$. For $i \in \mathcal{S}$ and $j \in \mathcal{U}$, using the convergence esablished in the previous proposition, we have

$$\frac{\bar{O}_i}{\mu_i} \leq l_i \mathbb{E}\left[\limsup_{t\to\infty} \frac{1}{t}\int_0^t g^{\mathcal{U}}(\mathbf{X}(s))ds\right].$$

□

The following proposition explores the structure of the extended processor-sharing allocation.

Proposition 4. (Comparison of output rates for different stability partitioning) *Define the extended processor-sharing allocation as follows, for $i = 1, \ldots, N$,*

$$\phi_i(\mathbf{x}) = f_i(x_i)\left(\sum_{j=1}^N f_j(x_j)\right)^{-1},$$

with $f_i(x_i) \leq l_i$ for all $x_i \geq 0$, $i = 1, \ldots, N$, and consider two rate stability partition sets $\mathcal{P}_1 = (\mathcal{S}_1, \mathcal{U}_1)$ and $\mathcal{P}_2 = (\mathcal{S}_2, \mathcal{U}_2)$ such that $\mathcal{U}_2 = \mathcal{U}_1 \cup \{k\}$, with $k \in \{1, \ldots, N\}$. Then it holds that for $i = 1, \ldots, N$,

$$\eta_j \bar{O}_i^{\mathcal{P}_1} \leq \eta_i \bar{O}_j^{\mathcal{P}_2}. \tag{12}$$

Proof: Using again the lines of the proof of Proposition 2, we get for $i = 1, \ldots, N$,

$$\frac{\bar{O}_i^{\mathcal{P}_1}}{\mu_i} = \mathbb{E}\left[\limsup_{t\to\infty} \frac{1}{t}\int_0^t \frac{f_i(X_i(s))}{f_i(X_i(s)) + \sum_{j\neq i} l_j} ds\right], \tag{13}$$

and

$$\frac{\bar{O}_j^{\mathcal{P}_2}}{\mu_j} = \frac{l_j}{\sum_{j=1}^N l_j}. \tag{14}$$

The proof then follows directly from (13) and (14) by observing that for $i = 1, \ldots, N$,

$$\frac{f_i(X_i(s))}{f_i(X_i(s)) + \sum_{j\neq i} l_j} \leq \frac{l_i}{\sum_{j=1}^N l_j}.$$

□

4 Rate stability necessary conditions

4.1 Traffic inequalities

In the absence of stochastic stability assumptions, it is naturally not possible to define the input rate of the queues as the solutions of the classic traffic equations as in [**?**] for instance. However, we can derive traffic inequalities linking the input rates and the asymptotic output rates of the network. These equations give a mathematical understanding on the common notions of mean output rates and input rates in the network.

Theorem 2 (Traffic inequalities)**.** *The asymptotic output rates* $\mathbf{O}$*,* $\bar{\mathbf{O}}$ *and growth rates* $\mathbf{Q}$*,* $\bar{\mathbf{Q}}$ *are finite and satisfy the following linear inequalities:*

$$Q_i + \bar{O}_i \leq \lambda_i + \sum_j p_{ji} \bar{O}_j, \tag{15}$$

$$\bar{Q}_i + O_i \geq \lambda_i + \sum_j p_{ji} O_j. \tag{16}$$

The work-conserving property (assumption 1) brings the additional inequalities:

$$\sum_{i=1}^{N} \frac{\bar{O}_i}{\mu_i} \geq 1, \quad \textit{and} \quad \sum_{i=1}^{N} \frac{O_i}{\mu_i} \leq 1. \tag{17}$$

In the special case of $\rho > 1$ *and* $\mathcal{U} = \{1, \ldots, N\}$*, we have*

$$\sum_{i=1}^{N} \frac{\bar{O}_i}{\mu_i} = 1. \tag{18}$$

Proof: Because of exponential service times and Poisson arrivals, $X(t)$ is a Markov process. Note again that from Assumption 2 the allocation functions $\phi_i(.)$, and hence the transition rates are bounded. This implies (the departure process from a queue being $D_i(t) = A_i(t) - X_i(t)$, with $A_i(t)$ the arrival process at queue i) that the process $\{M_i(t), t > 0\}$, defined by

$$M_i(t) := X_i(t) - X_i(0) - \int_0^t \{\lambda_i + \sum_j \frac{p_{ji}\phi_j(\mathbf{X}(s))}{\mu_j} - \frac{\phi_i(\mathbf{X}(s))}{\mu_i}\} ds,$$

is a local martingale. And since the transitions are bounded the martingale satisfies $E[M_i^2(t)] < Kt$ for $i = 1, \ldots, N$, $t > 0$ and some $K > 0$ [**?**]. This implies that the process $\{M_i(t)/t, t > 0\}$ is a super-martingale bounded in L^2 and consequently, for $i = 1, \ldots, N$, $\frac{M_i(t)}{t} \to 0 \ \ (t \to \infty)$, a.s. [**?**]. Assuming for simplicity that $\mathbf{X}(0) = \mathbf{0}$, it is readily seen from the definitions (7) and (6) that, for $i = 1, \ldots, N$, $t > 0$,

$$\frac{1}{t} M_i(t) + \lambda_i + \sum_j \frac{p_{ji}\varphi_j(t)}{\mu_j} - Y_i(t) = \frac{\varphi_i(t)}{\mu_i}.$$

This implies that $\limsup_{t\to\infty} \frac{X_i(t)}{t} < +\infty$ as well as

$$\begin{aligned} \limsup_{t\to\infty} \frac{\varphi_i(t)}{\mu_i} &= \limsup_{t\to\infty} \left(\lambda_i + \sum_j \frac{p_{ji}\varphi_j(t)}{\mu_j} - Y_i(t) \right) \\ &\leq \lambda_i + \sum_j \frac{p_{ji} \limsup_{t\to\infty} \varphi_j(t)}{\mu_j} - \liminf_{t\to\infty} Y_i(t). \end{aligned}$$

Using the dominated convergence theorem, we get (15). Relations (16) are obtained along the same lines. The inequalities in (17) follow from the dominated convergence theorem as well as the equation

$$1 = \limsup_{t\to\infty} \left(\sum_{i=1}^{N} \frac{\varphi_i(t)}{\mu_i} \right) \leq \sum_{i=1}^{N} \limsup_{t\to\infty} \frac{\varphi_i(t)}{\mu_i}.$$

If $\rho > 1$, the total number of jobs is transient, and hence for all t, almost surely $\sum_{i=1}^{N} \frac{\phi_i(X(t))}{\mu_i} = 1$ and $\sum_{i=1}^{N} \frac{\varphi_i(X(t))}{\mu_i} = 1$. The last assertion thus follows from Proposition 2. □

4.2 Necessary conditions for rate stability for converging rates

In this subsection, we study the case $\bar{\mathbf{O}} = \mathbf{O}$, which serves as a benchmark for finding rate stability conditions in the general case. In many interesting cases, the set of parameters such that the output rates are not converging is negligible, i.e. corresponds to a frontier between two regions in the parameter space. However, proving this statement is difficult in general and is out of the scope of this paper. We however show in the last section that we indeed have the convergence of the asymptotic growth rates outside a negligible set, for a set of parallel queues with monotone allocations.

Definition 4 $(\tilde{\mathbf{O}}, \tilde{\mathbf{Q}})$**.** *For a given stability partitioning* $\mathcal{P} = (\mathcal{S}, \mathcal{U})$ $(\mathcal{U} \neq \emptyset)$*, define* $(\tilde{\mathbf{O}}, \tilde{\mathbf{Q}})$ *as the solution (when it exists) of*

$$o_i + q_i = \lambda_i + \sum_j p_{ji} o_j, \tag{19}$$

$$\sum_{i=1}^{N} \frac{o_i}{\mu_i} = 1, \tag{20}$$

$$\frac{o_i}{\eta_i} = \frac{o_j}{\eta_j} := \tilde{Z}_{\mathcal{P}} \quad (i, j \in \mathcal{U}), \tag{21}$$

$$q_i = 0 \quad (i \in \mathcal{S}). \tag{22}$$

We first prove the existence of a unique solution for equations (19) to (22). We then give conditions for this solution to be positive vectors. To simplify the notations, suppose without loss of generality that the queues are ordered so that the stable ones are the first ones, i.e., there exists an index m such that $\mathcal{S} = \{1, \ldots, m\}$ and $\mathcal{U} = \{m+1, \ldots, N\}$. Define $G^{\mathcal{E}_1\mathcal{E}_2}$ as the truncation of the matrix G to the queues in $\mathcal{E}_1, \mathcal{E}_2$: $G^{\mathcal{E}_1\mathcal{E}_2} = (G)_{i\in\mathcal{E}_1, j\in\mathcal{E}_2}$ and similarly the vector $\mathbf{v}^{\mathcal{E}} = (v_i)_{i\in\mathcal{E}}$. We then write the routing matrix in the following form

$$P = \begin{pmatrix} P^{\mathcal{SS}} & P^{\mathcal{SU}} \\ P^{\mathcal{US}} & P^{\mathcal{UU}} \end{pmatrix}.$$

The vector $\boldsymbol{\eta}$ is defined as $\boldsymbol{\eta} = (l_1\mu_1, \ldots, l_N\mu_N)$. Let $\boldsymbol{\eta}^{\mathcal{S}}$ and $\boldsymbol{\eta}^{\mathcal{U}}$ be the vectors $(\eta_i)_{i\in\mathcal{S}}$ and $(\eta_i)_{i\in\mathcal{U}}$. We further define the vector $\boldsymbol{\omega}^{\mathcal{S}}$, and the positive constants $\kappa_{\mathcal{P}}$ and $\chi_{\mathcal{P}}$ as

$$\begin{aligned} \boldsymbol{\omega}^{\mathcal{S}} &= \boldsymbol{\lambda}^{\mathcal{S}} H^{\mathcal{SS}}, \\ \kappa_{\mathcal{P}} &= \textstyle\sum_{i\in\mathcal{S}} \frac{(\boldsymbol{\eta}^{\mathcal{U}} P^{\mathcal{US}} H^{\mathcal{SS}})e_i}{\mu_i}, \\ \chi_{\mathcal{P}} &= \textstyle\sum_{i\in\mathcal{U}} l_i, \end{aligned}$$

where $H^{\mathcal{SS}} = (I - P^{\mathcal{SS}})^{-1}$. Remark that the matrix $H^{\mathcal{SS}}$ is not in general the restriction of the matrix R.

Proposition 5. *Fix a partition* $\mathcal{P} = (\mathcal{S}, \mathcal{U})$ $(\mathcal{U} \neq \emptyset)$*. There exists a unique solution* $(\tilde{\mathbf{O}}, \tilde{\mathbf{Q}})$ *of Equations* (19) *to* (22)*, characterized by the following equations*

$$\begin{aligned} \tilde{\mathbf{O}}^{\mathcal{S}} &= (\boldsymbol{\lambda}^{\mathcal{S}} + \tilde{Z}_{\mathcal{P}} \boldsymbol{\eta}^{\mathcal{U}} P^{\mathcal{US}}) H^{\mathcal{SS}}, \\ \tilde{\mathbf{O}}^{\mathcal{U}} &= \tilde{Z}_{\mathcal{P}} \boldsymbol{\eta}^{\mathcal{U}}, \\ \tilde{Z}_{\mathcal{P}} &= \tfrac{1 - \sum_{i\in\mathcal{S}} \frac{\omega_i^{\mathcal{S}}}{\mu_i}}{\kappa_{\mathcal{P}} + \chi_{\mathcal{P}}}. \end{aligned}$$

Moreover, the solution $\tilde{\mathbf{O}}, \tilde{\mathbf{Q}}$ *is positive if and only if*

$$\sum_{i \in \mathcal{S}} \frac{\omega_i^{\mathcal{S}}}{\mu_i} \leq 1,$$

and

$$\tilde{Z}_{\mathcal{P}} \boldsymbol{\eta}^{\mathcal{U}} \left(I^{\mathcal{U}\mathcal{U}} - P^{\mathcal{U}\mathcal{U}} - P^{\mathcal{U}\mathcal{S}} H^{\mathcal{S}\mathcal{S}} P^{\mathcal{S}\mathcal{U}} \right) \geq \boldsymbol{\lambda}^{\mathcal{U}} + \boldsymbol{\lambda}^{\mathcal{S}} H^{\mathcal{S}\mathcal{S}} P^{\mathcal{S}\mathcal{U}}.$$

Proof: The system of Equations (19) to (22) can be rewritten as

$$\tilde{\mathbf{O}}^{\mathcal{S}} = \boldsymbol{\lambda}^{\mathcal{S}} + \tilde{\mathbf{O}}^{\mathcal{S}} P^{\mathcal{S}\mathcal{S}} + \tilde{Z}_{\mathcal{P}} \boldsymbol{\eta}^{\mathcal{U}} P^{\mathcal{U}\mathcal{S}}, \tag{23}$$

$$\tilde{\mathbf{Q}}^{\mathcal{U}} = \boldsymbol{\lambda}^{\mathcal{U}} + \tilde{\mathbf{O}}^{\mathcal{S}} P^{\mathcal{S}\mathcal{U}} + \tilde{Z}_{\mathcal{P}} \boldsymbol{\eta}^{\mathcal{U}} (P^{\mathcal{U}\mathcal{U}} - I^{\mathcal{U}\mathcal{U}}), \tag{24}$$

$$\sum_{i \in \mathcal{S}} \frac{\tilde{O}_i^{\mathcal{S}}}{\mu_i} = 1 - \chi_{\mathcal{P}} \tilde{Z}_{\mathcal{P}}, \tag{25}$$

since from the definition it follows that $\tilde{Q}_i^{\mathcal{S}} = 0$ if $i \in \mathcal{S}$, and thus Equation (19) reduces to Equation (23). Similarly Equation (24) can be obtained. Using Equations (21) and (22) the Equation (25) can be derived. The proposition follows from the fact that the matrices $I^{\mathcal{E}} - P^{\mathcal{E}}$, $\mathcal{E} = \mathcal{S}\mathcal{S}$, $\mathcal{U}\mathcal{U}$ are invertible with positive inverse matrices. Then, $\tilde{\mathbf{O}} \geq \mathbf{0}$ and $\tilde{Z}_{\mathcal{P}} \geq 0$, if and only if

$$\sum_{i \in \mathcal{S}} \frac{\omega_i^{\mathcal{S}}}{\mu_i} \leq 1.$$

Moreover, $\tilde{\mathbf{Q}} \geq 0$ if and only if

$$\tilde{Z}_{\mathcal{P}} (I^{\mathcal{U}\mathcal{U}} - P^{\mathcal{U}\mathcal{U}}) \boldsymbol{\eta}^{\mathcal{U}} \geq \boldsymbol{\lambda}^{\mathcal{U}} + \tilde{\mathbf{O}}^{\mathcal{S}} P^{\mathcal{S}\mathcal{U}},$$

which follows from Equation (24). □

It is remarkable that the conditions of positivity of the output rates are not sufficient to characterize the stability set. In the case of parallel queues, we show that the additional conditions underlined in Section 3 are indeed needed to sharply characterize, for given input parameters, the rate stability set.

4.3 Necessary conditions for rate stability

To derive necessary conditions for a given rate stability partitioning, we bound the output rates, taking into account the assumption of feed-forward routing. The bounds are obtained by comparing the maximum output rates with the outputs previously obtained in a (virtual) network where $\bar{O}_i = O_i$, for all i.

Lemma 1. *For* $i = 1, \ldots, N$, *we have*

$$\bar{O}_i \leq \omega_i,$$

where the vector $\boldsymbol{\omega} = \boldsymbol{\lambda} R$ *is the solution of the usual traffic equations*

$$\boldsymbol{\omega} = \boldsymbol{\lambda} + \boldsymbol{\omega} P.$$

Proof: Remark first that $\boldsymbol{\omega}$ exists and is unique because $R = (I-P)^{-1}$ is a well-defined positive matrix since $I - P$ is substochastic. Define the degree of queue i in the following way; $d_i = 0$ if $p_{ji} = 0$, for all $j = 1, \ldots, N$,. Otherwise, $d_i = \max_{j:\ p_{ji}>0}\{d_j\}$. Because of the absence of loops in the network, there exists at least one queue i_0 of degree 0 (a source). Using the traffic inequalities of the previous section, we get for all queues i_0 of degree 0

$$\bar{O}_{i_0} \leq \lambda_{i_0} = \omega_{i_0}.$$

We further proceed by induction on the degree of queues. Suppose the assertion is true for all degrees less than m. Consider a queue of degree $m+1$. It is receiving traffic from queues of lower degree. Using the traffic inequalities, the induction assumption and the definition of $\boldsymbol{\omega}$, we get

$$\bar{O}_i \leq \lambda_i + \sum_{j:d_j \leq m} p_{ji}\bar{O}_j \leq \lambda_i + \sum_{j=1}^{N} p_{ji}\omega_j = \omega_i.$$

□

We now derive the lemma leading to the main result of this section.

Lemma 2. *Let $\bar{Z}_{\mathcal{P}} := \frac{\bar{O}_i}{\eta_i}, \forall i \in \mathcal{U}$. For each partitioning $\mathcal{P} = (\mathcal{S},\mathcal{U})$ $(\mathcal{U} \neq \emptyset)$, we have*

$$\bar{Z}_{\mathcal{P}} \geq \tilde{Z}_{\mathcal{P}}.$$

If moreover $P^{\mathcal{US}} = 0$, then

$$\forall i \in \mathcal{S}, \bar{O}_i \leq \tilde{O}_i.$$

Note that this result holds without restriction on the routing policy, and is not limited to feed-forward routing.

Proof: Using Lemma 1 and the traffic inequalities given in Theorem 2, we can write that

$$\bar{\mathbf{O}}^{\mathcal{S}} \leq \boldsymbol{\omega}^{\mathcal{S}} + \bar{Z}_{\mathcal{P}}\boldsymbol{\eta}^{\mathcal{S}}P^{\mathcal{US}}H^{\mathcal{SS}},$$

since from Equation (15) we have that $Q_i = 0$ for $i \in \mathcal{S}$. Next, using Equation (17) and $\chi_{\mathcal{P}} = \sum_{i\in\mathcal{U}} l_i$, it follows that

$$1 \leq \chi_{\mathcal{P}}\bar{Z}_{\mathcal{P}} + \sum_{i\in\mathcal{S}} \frac{\bar{O}_i}{\mu_i}.$$

Hence, combining these equations we have

$$(\chi_{\mathcal{P}} + \kappa_{\mathcal{P}})\bar{Z}_{\mathcal{P}} + \sum_{i\in\mathcal{S}} \omega_i^{\mathcal{S}} \quad \geq \quad \chi_{\mathcal{P}}\bar{Z}_{\mathcal{P}} + \sum_{i\in\mathcal{S}} \frac{\bar{O}_i}{\mu_i} \geq 1,$$

which gives $\bar{Z}_{\mathcal{P}} \geq \tilde{Z}_{\mathcal{P}}$. If $P^{\mathcal{US}} = 0$, the second assertion follows from Proposition 5. □

We can now derive necessary conditions for the partitioning $\mathcal{P} = (\mathcal{S},\mathcal{U})$ to hold. We make use here of Lemma 1 and we therefore need the assumption of feed-forward routing.

Theorem 3. *Assume a given partitioning $\mathcal{P} = (\mathcal{S},\mathcal{U})$. Then for all $i \in \mathcal{U}$*

$$\frac{\omega_i}{\eta_i} > \tilde{Z}_{\mathcal{P}}.$$

Proof: For an unstable queue i, Q_i can be written as strictly positive for all $i \in \mathcal{U}$, which gives, using the traffic inequalities

$$\sum_{j=1}^{N} p_{ji}\bar{O}_j + \lambda_i - \bar{O}_i \geq Q_i > 0.$$

Using the two previous lemmas, it leads to

$$\forall i \in \mathcal{U},\ \sum_{j=1}^{N} p_{ji}\omega_j + \lambda_i - \eta_i \tilde{Z}_{\mathcal{P}} \geq \sum_{j=1}^{N} p_{ji}\bar{O}_j + \lambda_i - \bar{O}_i \geq Q_i > 0,$$

which gives using Lemma 1 that $-\eta_i \tilde{Z}_{\mathcal{P}} \geq \bar{O}_i$ and thus $\frac{\omega_i}{\eta_i} > \tilde{Z}_{\mathcal{P}}$. □

So far, only necessary conditions for a given rate stability partition of the queues follow from Theorem 3. We illustrate the obtained results on two examples, where the obtained necessary conditions turn out to be sufficient in the first example and not sufficient in the second.

4.4 Example : two-queue tandem model

We derive *necessary and sufficient* conditions for rate stability for a tandem system with a monotone, work-conserving allocation with symmetric uniform limits (assumptions 1, 2, 3). Consider the system of two queues illustrated in Figure 1. The routing matrix is given by

$$P = \begin{pmatrix} 0 & p \\ 0 & 0 \end{pmatrix}. \tag{26}$$

Thus, a fraction p of the output rate of the first queue is sent as input rate to the second queue. The following traffic equations and inequalities hold (Theorem 2)

$$\begin{aligned} Q_1 + \bar{O}_1 &= \lambda_1, \\ Q_2 + \bar{O}_2 &\leq p\bar{O}_1, \\ \frac{\bar{O}_1}{\mu_1} + \frac{\bar{O}_2}{\mu_2} &\geq 1. \end{aligned}$$

For the corresponding virtual model satisfying $\bar{\mathbf{O}} = \mathbf{O}$, the traffic equations are

$$\begin{aligned} \tilde{O}_1 &= \lambda_1 - \tilde{Q}_1, \\ \tilde{O}_2 &= p\tilde{O}_1^{\mathcal{P}} - \tilde{Q}_2, \\ \frac{\tilde{O}_1}{\mu_1} &+ \frac{\tilde{O}_2}{\mu_2} = 1, \\ \frac{\tilde{O}_i}{\eta_i} &= \tilde{Z}_{\mathcal{P}}, \quad \text{for } i \in U. \end{aligned}$$

By $\mathcal{P}$ we denote the partition of queues according to their rate stability. $\mathcal{P}$ can thus be $\mathcal{U} = \emptyset$, and $\mathcal{S} = \{1\}$, $\mathcal{U} = \{2\}$, and $\mathcal{S} = \{2\}$, $\mathcal{U} = \{1\}$, and $\mathcal{U} = \{1,\ 2\}$. The solutions of $\tilde{\mathbf{O}}$ and $\tilde{\mathbf{Q}}$ are given in Table 1 for each stability subset $\mathcal{P}$.

According to Theorem 3 the network is globally stochastically stable if and only if $\rho < 1$ which reads

$$\lambda_1 < \frac{\mu_1\mu_2}{p\mu_1 + \mu_2}.$$

Note that in this case $\bar{\mathbf{O}} = \tilde{\mathbf{O}} = \mathbf{O}$ and $\bar{\mathbf{Q}} = \tilde{\mathbf{Q}} = \mathbf{Q}$.

$\mathcal{P}$	$\tilde{Q}_1$	$\tilde{Q}_2$	$\tilde{O}_1$	$\tilde{O}_2$
$\mathcal{S}=\{1,2\}, \mathcal{U}=\emptyset$	0	0	λ_1	$\lambda_1 p$
$\mathcal{S}=\{1\},\ \mathcal{U}=\{2\}$	0	$p\lambda_1-(1-\frac{\lambda_1}{\mu_1})\mu_2$	λ_1	$(1-\frac{\lambda_1}{\mu_1})\mu_2$
$\mathcal{S}=\{2\},\ \mathcal{U}=\{1\}$	$\lambda_1-\frac{\mu_1\mu_2}{p\mu_1+\mu_2}$	0	$\frac{\mu_1\mu_2}{p\mu_1+\mu_2}$	$\frac{p\mu_1\mu_2}{p\mu_1+\mu_2}$
$\mathcal{S}=\emptyset, \mathcal{U}=\{1,\ 2\}$	$\lambda_1-\frac{l_1\mu_1}{l_1+l_2}$	$\frac{pl_1\mu_1}{l_1+l_2}-\frac{l_2\mu_2}{l_1+l_2}$	$\frac{l_1\mu_1}{l_1+l_2}$	$\frac{l_2\mu_2}{l_1+l_2}$

Table 1: Output rates for the stability subsets.

Figure 1: Two queues in tandem.

4.4.1 Necessary conditions for $\mathcal{U}=\{1,\ 2\}$

For the partition $\mathcal{U}=\{1,2\}$, given that $\tilde{Z}_{\mathcal{P}}=\frac{1}{l_1+l_2}\,\boldsymbol{\omega}=(\lambda_1,p\lambda_1)$, the following conditions given by Theorem 3 are necessary

$$p>\frac{l_2\mu_2}{l_1\mu_1},$$

$$\lambda_1>\frac{\mu_1 l_1}{l_1+l_2}.$$

Using the last assertion in Theorem 2, we further obtain that $\bar{Z}_{\mathcal{P}}=\tilde{Z}_{\mathcal{P}}$ which implies $\bar{\mathbf{O}}=\tilde{\mathbf{O}}=\mathbf{O}$ and $\bar{\mathbf{Q}}=\tilde{\mathbf{Q}}=\mathbf{Q}$.

4.4.2 Necessary conditions for $\mathcal{S}=\{1\},\ \mathcal{U}=\{2\}$

For the partitions $\mathcal{U}=\{1\},\mathcal{S}=\{2\}$ and $\mathcal{S}=\{1\},\mathcal{U}=\{2\}$, the necessary conditions stated in Theorem 3 lead to the already known condition $\rho>1$. Using Theorem 3 ($\bar{Z}_{\mathcal{P}}>\tilde{Z}_{\mathcal{P}}$), the first traffic equation and the additional inequalities given by Theorem 2 and Proposition 3, we obtain

$$\frac{\lambda_1}{l_1\mu_1}=\frac{\tilde{O}_1^{(\mathcal{S}=\{1\},\mathcal{U}=\{2\})}}{\eta_1}=\frac{\bar{O}_1^{(\mathcal{S}=\{1\},\mathcal{U}=\{2\})}}{\eta_1}<\frac{\bar{O}_1^{(\mathcal{U}=\{1,2\})}}{\eta_1}=\frac{\tilde{O}_1^{(\mathcal{U}=\{1,2\})}}{\eta_1}=\frac{1}{l_1+l_2},$$

which leads to the necessary condition $\lambda_1<\frac{\mu_1 l_1}{l_1+l_2}$.

4.4.3 Necessary conditions for $\mathcal{U}=\{1\},\ \mathcal{S}=\{2\}$

$$\frac{\mu_1 p\mu_2}{(\mu_1 p+\mu_2)l_2\mu_2}=\frac{\tilde{O}_2^{(\mathcal{U}=\{1\},\mathcal{S}=\{2\})}}{\eta_2}\leq\frac{\bar{O}_2^{(\mathcal{U}=\{1\},\mathcal{S}=\{2\})}}{\eta_2},$$

and

$$\frac{\bar{O}_2^{(\mathcal{U}=\{1\},\mathcal{S}=\{2\})}}{\eta_2}<\frac{\bar{O}_2^{(\mathcal{U}=\{1,2\})}}{\eta_2}=\frac{\tilde{O}_2^{(\mathcal{U}=\{1,2\})}}{\eta_2}=\frac{1}{l_1+l_2},$$

and this leads to the necessary condition that $p < \frac{l_2\mu_2}{l_1\mu_1}$.

The obtained necessary conditions are easily seen to lead to a complete partitioning of the parameter set, which gives a sharp characterization of the stability set. As a consequence, the obtained conditions are *both necessary and sufficient*, except on a boundary set of input parameters.

In Figure 2 the stability set is depicted for two different sets of input parameters.

Figure 2: Stability regions with $(\mu_1, l_1, \mu_2, l_2) = (3, 1, 1, 1)$ in the left figure, and with $(\mu_1, l_1, \mu_2, l_2) = (1, 1, 3, 1)$ in the right figure.

5 Parallel queues

In this section, we consider parallel queues and thus suppose that there is no internal routing, i.e., $p_{ij} = 0$, for all i, j and that the allocation is monotone. In that case, we can derive a sharp characterization of the per-queue rate stability. To this end, we first show that in that case, the traffic inequalities are actually a set of traffic equations (Proposition 6). This allows to prove that the output rates and asymptotic growth rates converge. Using the results of Sections 3 and 4, we then derive a characterization of the per-queue stability (Theorem 4).

5.1 Extended traffic equations

In this subsection, we specify the traffic inequalities obtained in the general case by deriving traffic *equations* linking the input rates and the asymptotic output rates of the network. service rates.

Proposition 6 (Extended traffic equations)**.** *The asymptotic output rates* $\mathbf{O}$, $\bar{\mathbf{O}}$ *and growth rates* $\mathbf{Q}$, $\bar{\mathbf{Q}}$ *are finite and satisfy the following linear equations. For* $i = 1, \ldots, N$,

$$Q_i + \bar{O}_i = \lambda_i, \tag{27}$$

$$\bar{Q}_i + O_i = \lambda_i. \tag{28}$$

Proof: We follow the same lines as in Theorem 2,

$$M_i(t) := X_i(t) - X_i(0) - \int_0^t \{\lambda_i - \phi_i(\mathbf{X}(s))\} ds, \tag{29}$$

is a martingale that satisfies $E[M_i^2(t)] < Kt$ for $i = 1, \ldots, N$, $t > 0$ and some $K > 0$. This implies that $\limsup_{t\to\infty} \frac{X_i(t)}{t} < +\infty$ and $\liminf_{t\to\infty} Y_i(t) = \lambda_i - \limsup_{t\to\infty} \varphi_i(t)$. Using the dominated convergence theorem, we get Equations (27) and (28). □

5.2 Output rates convergence

We fix $\mathcal{P}$ to be a partition of queues such that queues in $\mathcal{S}$ are rate stable while queues in $\mathcal{U}$ are rate unstable. In the following proposition, we prove that the output rates of the different queues converge which further allows a complete description of the rate stability set.

Proposition 7. *Consider a set of parallel queues with a decreasing allocation satisfying the Assumptions* 1 *and* 2*. Then, outside a negligible set of input parameters*[1]*, for* $t \to \infty$*,*

$$\frac{X_i(t)}{t} \to Q_i, \textit{ in probability}, \tag{30}$$

$$\varphi_i(t) \to O_i, \textit{ in probability}, \tag{31}$$

with

$$O_i = \lambda_i \quad (i \in \mathcal{S}), \qquad O_i = Z_{\mathcal{P}}\eta_i \quad (i \in \mathcal{U}),$$

$$Q_i = 0 \quad (i \in \mathcal{S}), \quad Q_i = \lambda_i - Z_{\mathcal{P}}\eta_i \quad (i \in \mathcal{U}),$$

where

$$Z_{\mathcal{P}} := \frac{1 - \sum_{j\in\mathcal{S}} \frac{\lambda_j}{\mu_j}}{\sum_{j\in\mathcal{U}} \ l_j} = \frac{1 - \sum_{j\in\mathcal{S}} \rho_j}{\sum_{j\in\mathcal{U}} \ l_j}.$$

Proof:

We first prove that the set of parameters such that $Q_i = 0$ and $\bar{Q}_i > 0$ is 'small' in the sense, that increasing one of the $\lambda_j, j \neq i$ parameter of any $\epsilon > 0$ will force $Q_i > 0$. Because the allocation is decreasing, if λ_j is replaced by $\lambda_j + \epsilon$, then it can be proven that the number of customers $X_j^{\epsilon}(t)$ is stochastically increased for all t. We refer to [**?**] for notions and proofs on stochastic comparisons for parallel coupled queues. It implies that $\bar{O}_i^{\epsilon} > \bar{O}_i$. Hence, we obtained that $Q_i^{\epsilon} = \lambda_i - \bar{O}_j^{\epsilon} > \lambda_i - \bar{O}_j = 0$.

We can now restrict our attention to the case where $i \in \mathcal{S}$ implies $Q_i = \bar{Q}_i > 0$. Let us now prove the convergence of the rates in that case. Consider (q, o) any limiting point of the vector $(\mathbf{Y}(t), \varphi(t))$. Using the arguments used in Proposition 2 and Proposition 5, we obtain the set of equations (27) and (28) for (q, o) together with $q_i = 0, i \in \mathcal{S}$. Using the work-conserving property at each time t, we further obtained that $\sum_i \frac{o_i}{\mu_j} = 1$. Hence there is a unique limiting point and the rates are converging. □

Remark 2. *It appears plausible to prove an almost sure convergence for these processes even without the assumption of exponential service times and Poisson arrivals. This result is out of the scope of this study but we refer to the method presented in [**?**] and further used for a set of discriminatory processor-sharing queues (DPS) in [**?**] for such a derivation. These techniques, jointly used with the traffic conservation used here would prove the stated convergence in the context of stationary marked point processes.*

5.3 Characterization of the per-queue rate stability

We assume without loss of generality that the queues are ranked in decreasing order of the loads $\zeta_i := \frac{\lambda_i}{l_i\mu_i}$, in the sense that

$$\zeta_1 \leq \cdots \leq \zeta_N. \tag{32}$$

The following result shows the relation between the ordering of the queues and the per-queue rate stability.

Proposition 8. *If queue* i *is rate stable and* $j < i$*, then queue* j *is also rate stable.*

[1] i.e., of dimension strictly less that d if d is the dimension of the space in which the input parameters are chosen.

Proof: Suppose $j \in \mathcal{U}$, $i \in \mathcal{S}$ and $j < i$. From Proposition 3, we get $\frac{O_i}{\eta_i} < \frac{O_j}{\eta_j}$. From Theorem 2, it follows that $O_i = \lambda_i$ and that $O_j \leq \lambda_j$. We thus find that

$$\zeta_i = \frac{O_i}{\eta_i} < \frac{O_j}{\eta_j} \leq \frac{\lambda_j}{\eta_j}. \tag{33}$$

This contradicts $\zeta_j \leq \zeta_i$. □

Denote $Z(m) = Z_{\{1,\ldots,m\}} = \frac{1-\sum_{i\leq m}\rho_i}{\sum_{i>m} l_i}$. The following result shows that the partitioning $\mathcal{P} = (\mathcal{S},\mathcal{U})$ has a simple structure.

Theorem 4 (Structure of stability partitioning)**.** *Consider a set of parallel queues with a decreasing allocation verifying the Assumptions* 1, 2 *and* 3. *The stability partitioning* $\mathcal{P} = (\mathcal{S},\mathcal{U})$ *is characterized as follows* $\mathcal{P} = (\mathcal{S},\mathcal{U})$ *with* $\mathcal{S} = \{1,\ldots,m\}$ *and* $\mathcal{U} = \{m+1,\ldots,N\}$ *if and only if*

$$\zeta_m \leq Z(m) < \zeta_{m+1}. \tag{34}$$

Proof: Using Proposition 8, there exists a k such that $\mathcal{S} = \{1,\ldots,k\}$ and $\mathcal{U} = \{k+1,\ldots N\}$. Theorem 3 combined with Proposition 7 gives that $Z(k) < \zeta(k+1)$. Proposition 8 gives

$$\frac{\bar{O}_k}{\eta_k} \leq \frac{\bar{O}_{k+1}}{\eta_{k+1}},$$

which in combination with the traffic equations leads to

$$\zeta_k \leq Z(k).$$

As $Z(\cdot)$ is a decreasing function, we conclude that $m = k$. □

We emphasize that Theorem 4 gives a complete characterization of the rate stability partitioning for model instances that satisfy Assumptions 1, 2, and 3. Typical examples of such allocations are the coupled-processors allocation (defined in Section 2.1), and utility-based allocations on some tree topology (see [**?**]).

6 Conclusion and topics for further research

The results presented in this study provide new intuition and fundamental insight in the stability and throughput behavior of queueing models in which resources are shared among different queues. These results should be viewed as a first step in understanding the behavior of this type of queueing networks, and open up a wealth of challenging open research questions. Some of these questions will be briefly touched upon below.

In the context of stability and throughput characteristics, several interesting questions remain to be answered. First, when X is a continuous-time Markov process, it actually remains an open and crucial question to know for which input parameters, rate instability of queue i coincides to the convergence of X_i to infinity (either in probability or in law). In [**?**], per-queue stochastic stability is established for parallel queues with *monotone* allocation functions. It is remarkable that, except possibly on the boundary of the stability sets, the conditions for rate instability (and thus stochastic instability) that we have derived here coincide with the sharp characterization of the stochastic instability set given in [**?**]. This encouraging observation calls for a generalization of this result to more complex topologies.

References