Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

On the Singular Behavior of a Queueing System with Random Connectivity

George Michailidis and Nicholas Bambos

Abstract— **In this paper, we study the stationary dynamics of a processing system comprised of several parallel queues and a single server of constant rate. The connectivity of the server to each queue is randomly modulated, taking values 1 (connected) or 0 (severed). At any given time, only the currently connected queues may receive service. A key issue is how to schedule the server on the connected queues in order to maximize the system throughput. We investigate the behavior of two dynamic schedules, when the loading of the system exceeds its capacity. It is shown that unlike many other queueing systems that exhibit a binary behavior -global stability or global instability- the system under consideration exhibits a much richer behavior, with several partial stability modes. These modes are fully determined by the underlying traffic loading. The results are obtained under very general stationary ergodic traffic flows and connectivity modulation.**

I. MOTIVATION

Parallel queueing systems operating in randomly modulated environments have received a lot of attention over the last few years. In a series of papers, allocation of resources for throughput maximization ([2], [4], [6], [11]) and packet loss minimization ([3], [8], [7]) have been studied within a Markovian as well as a stationary ergodic context. In particular, maximum throughput server allocation policies have been proposed and their properties established. These policies can be described as of the *max-weight* variety, where the server's power at any point in time is allocated to the queue with the largest weighted queue length, with the weight given by the prevailing service rate.

In this paper, we would like to investigate the behavior of the queue length/workload process of such a system under a maximum throughput policy, when the loading of the system exceeds the available server's capacity. The following simulation results motivate our interest for studying this issue: in Figure 1 the queue length processes of the model studied in [2], [10], [11] under a maximum throughput policy is shown, when the loading exceeds the capacity of the system. It can be seen that the queue length process increases linearly and the system can be characterized as globally unstable (see Proposition 3.1 in [2]). This behavior of the queue length, as well as the workload, processes operating under maximum throughput policies is consistent for a large class of parallel queueing systems [1].

The work of the first author was supported by the National Science Foundation under grant CCR-0325571

George Michailidis is with the Department of Statistics, The University of Michigan, Ann Arbor, MI 48109-1092 gmichail@umich.edu

Nicholas Bambos is with the Departments of Electrical Engineering and Management Science & Engineering, Stanford University, Stanford, CA 94305-9505 bambos@stanford.edu

In Figure 2 the queue length processes of the model studied in [4] (and described in the next section) are shown, under three different loadings which also exceed the system's capacity. It can be seen that this system exhibits a much richer behavior, since in some scenarios both queues 'blow-up' to infinity, while in some other scenarios only one of the queues increases linearly, while the other queue continues to enter the empty state.

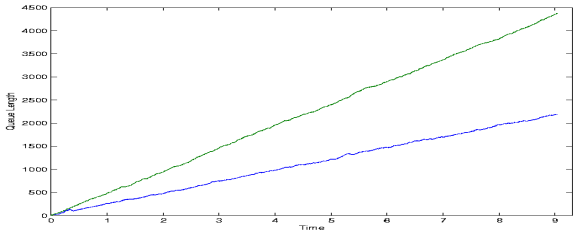

Fig. 1. Queue length process for a two-queue system operating in a random service environment.

In light of the above results, the goal of this paper is to study the singular behavior of the latter model and investigate the dynamics of the workload (queue length) processes. The paper is organized as follows: in section 2, the model is described, while in section 3 a brief overview of the maximum throughput policy and its properties is given. The main results dealing with the fine structure of the instability region and the multiple partial stability modes of the system are presented in section 4.

II. MODEL STRUCTURE AND ASSUMPTIONS

Consider a queueing system comprised of $K \in \mathbb{Z}_+$ first-come-first-served queues and a server of constant service rate $r \in \mathbb{R}_+$. There is a random flow of jobs arriving to the queues with service requests. The queues have infinite capacity buffers where jobs are placed while waiting to be served. At any given time the server is connected (has access) to a subset of the queues and those are the only ones that can receive service. The server-queue connectivities are randomly *modulated*, changing in time according to a stochastic process. Finally, a server *allocation policy* is used to decide which queues to serve among those that are currently connected.

In this paper we study the behavior of the system at large times under general stationary and ergodic job arrival flows and server-queue connectivities. In particular, we are interested in characterizing the system dynamics for a maximum throughput policy, when the loading of the system exceeds its capacity.

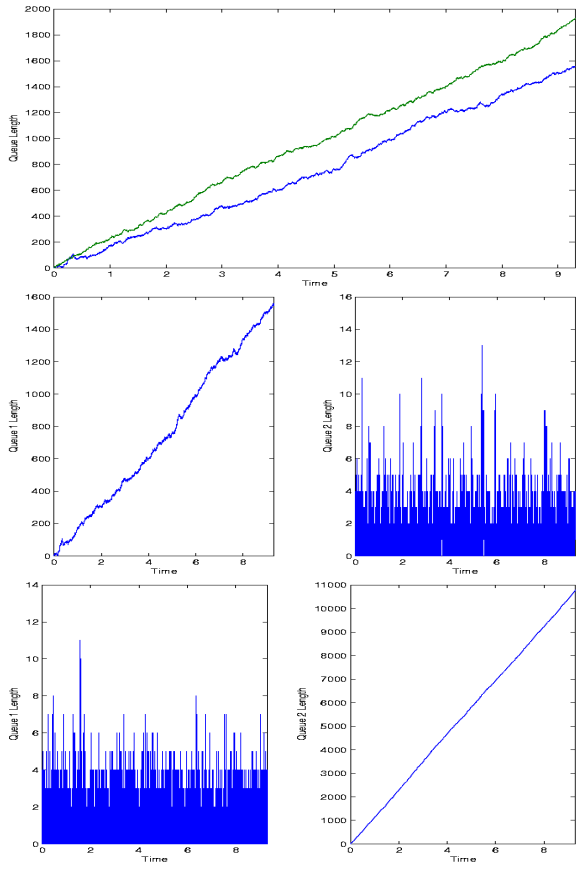

Fig. 2. Queue length process for a two-queue system with random link Connectivities, under different loading conditions.

Let $t_j^q \in \mathbb{R}$ be the arrival time of the j^{th} job to arrive to queue $q \in \mathbf{K} = \{1,2,3,...,K\}$, and $\sigma_j^q \in \mathbb{R}_+$ its associated service (processing) time requirement. These are random quantities which we model as elements of a random marked point process (RMPP, [1], [5])

$$\mathcal{N}_q = \{(t_j^q, \sigma_j^q),\ j \in \mathbb{Z}\}, \tag{1}$$

describing the stochastic input to the q^{th} queue. The collection of processes

$$\mathcal{N} = \{\mathcal{N}_q,\ q \in \mathbf{K}\} \tag{2}$$

comprises the overall *input* to the queueing system.

We introduce next the *connectivity process* $\{C_t,\ t \in \mathbb{R}\}$, where C_t is the set of connected queues at time t. Define $\mathbf{C}_o$ to be the set of all values that the $\{C_t\}$ process attains as time evolves. This is some subset of the power set of $\mathbf{K}$, i.e. $C_t \in \mathbf{C} \subseteq 2^{\mathbf{K}}$. Let now $s_k \in \mathbb{R}$ be the time of the k^{th} occurrence of change in the server-queue connectivities and $\mathbf{c}_k \subseteq \mathbf{C}$ the set of connected queues that the system switches to at time s_k. We also model these random quantities as elements of another RMPP

$$\mathcal{M} = \{(s_k, \mathbf{c}_k),\ k \in \mathbb{Z}\}, \tag{3}$$

which we call the *connectivity modulation* process. Based on the switching times s_k we can write

$$C_t = \sum_{k \in \mathbb{Z}} \mathbf{c}_k \mathbf{1}_{\{s_k \le t < s_{k+1}\}}, \tag{4}$$

where $\mathbf{1}_{\{\ \}}$ is the indicator function.

The RMPP's $\mathcal{N}_q$ and $\mathcal{M}$ are defined on some common probability space $(\Omega, \mathcal{F}, P)$ and are assumed to be *stationary* and *ergodic* with respect to time shifts $z \in \mathbb{R}$

$$\theta_z \mathcal{N}_q \overset{D}{=} \{(t_j^q - z, \sigma_j^q),\ j \in \mathbb{Z}\}, \tag{5}$$

$$\theta_z \mathcal{M} \overset{D}{=} \{(s_k - z, \mathbf{c}_k),\ k \in \mathbb{Z}\}, \tag{6}$$

for every $q \in \mathbf{K}$, where D denotes equality in distribution. The numbering of jobs and connectivity switching epochs on every sample path are such that $...t_{-1}^q < t_0^q \le 0 < t_1^q ... < t_j^q < t_{j+1}^q ...$ and $...s_{-1} < s_0 \le 0 < s_1 ... < s_k < s_{k+1} ...$ (simple RMPP's [5]). The processes are assumed to have pathwise a finite number of points in every finite time interval. The *traffic intensity* (average workload per unit time) entering queue $q \in \mathbf{K}$ is given by

$$\rho_q = \lim_{t \to \infty} \left[\frac{1}{t} \sum_{j \in \mathbb{Z}} \sigma_j^q \mathbf{1}_{\{t_j^q \in [0,t)\}} \right]. \tag{7}$$

It is assumed that $\rho_q > 0$ for every $q \in \mathbf{K}$.

The system has to decide on how to allocate the processing power of the server to the queues that are currently connected to it. This is done according to some allocation policy $\mathscr{A} \in \mathbf{A}$, where $\mathbf{A}$ is the set of all such policies. We particularly focus on two simple allocation policies which are shown to exhibit optimal behavior among those in $\mathbf{A}$. The first one, called *Longest Connected Queue* (LCQ) policy and denoted $\mathscr{A}_{LCQ} \in \mathbf{A}$, allocates the server to the connected queue with the largest number of jobs in it at every decision epoch. Such epochs are the times when the job currently being processed completes service, as well as the connectivity switching times s_k. In the case that more than one queue have the same (maximum) number of jobs at a decision epoch (i.e. a tie), one of them may be chosen arbitrarily. The second policy, called *Maximum Connected Workload* (MCW) policy and denoted $\mathscr{A}_{MCW} \in \mathbf{A}$, operates by allocating the server at time t to the connected queue with maximum workload. In the case that the workloads of two or more queues become equal, the MCW policy distributes the processing power of the server equally among these queues. Notice that the LCQ policy pursues the balancing of the queue sizes, while the MCW the balancing of the workloads.

III. Stability Aspects. Some Basic Facts.

We start by defining the *queueing state* of the system $\mathscr{S}$ operating under policy $\mathscr{A} \in \mathbf{A}$. Let $\mathscr{W}_{s,t}^q(\mathscr{A}, \mathbf{x}_o)$ be the *workload* in queue q at time t (i.e. the sum of all residual service time requirements of all jobs present in the buffer) for an initial workload of $\mathbf{x}_o$ and $\vec{\mathscr{W}}_{s,t}(\mathscr{A}, \mathbf{x}_o) = \{\mathscr{W}_{s,t}^q(\mathscr{A}, \mathbf{x}_o),\ q \in \mathbf{K}\}$. Moreover, let $\mathscr{U}_{s,t}^q(\mathscr{A}, \mathbf{x}_o)$ be the number of jobs (queue length) in queue q at time t and $\vec{\mathscr{U}}_{s,t}(\mathscr{A}, \mathbf{x}_o) = \{\mathscr{U}_{s,t}^q(\mathscr{A}, \mathbf{x}_o),\ q \in \mathbf{K}\}$. We make (where appropriate) the technical *assumption* that all stochastic processes

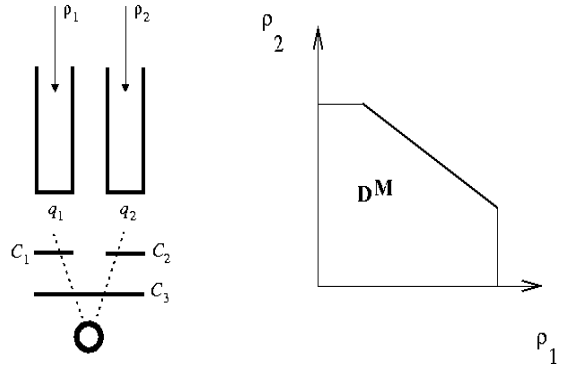

Fig. 3. An example of a simple system of two queues with connectivity set probabilities $P[C_1]=0.3$, $P[C_2]=0.2$, $P[C_3]=0.5$, and its capacity surface.

we are studying are right-continuous and have left limits pathwise (cadlag [1]).

We define next the set

$$\mathbf{D}^{\mathscr{M}}=\left\{\vec{\alpha}\in\mathbb{R}_+^K:\sum_{q\in Q}\alpha_q<r\left[\sum_{C\in\mathbf{C}:\ C\cap Q\neq\emptyset}P[C]\right]\right\}$$

for every $Q\subseteq\mathbf{K},\ Q\neq\emptyset$, which is shown in [4] to be the **Stability Region** of the system. Moreover, we define the topological boundary of $\mathbf{D}^{\mathscr{M}}$ (in the standard Euclidean topology of $\mathbb{R}_+^K$) by $\partial\mathbf{D}^{\mathscr{M}}$, which corresponds to the **Capacity Surface** of this queueing system. The capacity surface of a two-queue system is shown in Figure 3.

The following results (rigorously proved in [4]) describe the long-term behavior of the system under study.

Proposition 3.1: **(The Case of Instability)**
For any stationary and ergodic input and modulation processes $\mathscr{N}$ and $\mathscr{M}$, we have that if

$$\vec{\rho}\notin\mathbf{D}^{\mathscr{M}}\bigcup\partial\mathbf{D}^{\mathscr{M}}=$$
$$\left\{\vec{\alpha}\in\mathbb{R}_+^K:\sum_{q\in Q}\alpha_q\leq r\left[\sum_{C\in\mathbf{C}:\ C\cap Q\neq\emptyset}P[C]\right],\ Q\subseteq\mathbf{K},Q\neq\emptyset\right\}\tag{8}$$

then, for any server allocation policy $\mathscr{A}\in\mathbf{A}$, there exists at least one queue $q\in\mathbf{K}$, such that

$$\lim_{t\to\infty}\mathscr{W}_{s,t}^q(\mathscr{A},\mathbf{x})=\infty\tag{9}$$

almost surely, for every $s\in\mathbb{R}$ and initial state $\mathbf{x}$.. That is, when $\vec{\rho}\notin\mathbf{D}^{\mathscr{M}}\bigcup\partial\mathbf{D}^{\mathscr{M}}$, the system is **unstable** under any policy $\mathscr{A}$, in the sense that the workload of some queues blows up to infinity at large times.

Proposition 3.2: **(Finiteness of the Stationary Workloads under the $\mathscr{A}_{MCW}$ Policy)** For any stationary and ergodic input and modulation processes $\mathscr{N}$ and $\mathscr{M}$, if

$$\vec{\rho}\in\mathbf{D}^{\mathscr{M}}\tag{10}$$

then

$$\widetilde{W}_t^q=\lim_{s\to-\infty}W_{s,t}^q(\mathscr{A}_{MCW},\vec{0})<\infty,\ \ \forall\, t\in\mathbb{R},\quad \textit{for every}\ \ q\in\mathbf{K},\tag{11}$$

almost surely. Under this condition the processes $\{\widetilde{W}_t^q,t\in\mathbb{R}\},\ q\in\mathbf{K}$ form a *finite stationary* operational regime of the system.

Theorem 3.1: **(Stability under the $\mathscr{A}_{MCW}$ Policy)**
For any stationary ergodic input and modulation processes $\mathscr{N}$ and $\mathscr{M}$, if

$$\vec{\rho}\ \in\ \mathbf{D}^{\mathscr{M}}\tag{12}$$

then

$$\lim_{t\to\infty}P[W_{s,t+a_1}^{q_1}(\vec{0})\in B_1,\ W_{s,t+a_2}^{q_2}(\vec{0})\in B_2,\ \ldots,$$
$$W_{s,t+a_n}^{q_n}(\vec{0})\in B_n,\ \ldots,\ W_{s,t+a_N}^{q_N}(\vec{0})\in B_N]=$$
$$\lim_{t\to\infty}P[\mathscr{W}_{s,t+a_1}^{q_1}(\mathscr{A}_{MCW},\mathbf{0})\in B_1,\ \mathscr{W}_{s,t+a_2}^{q_2}(\mathscr{A}_{MCW},\mathbf{0})\in B_2,\ \ldots,$$
$$\mathscr{W}_{s,t+a_N}^{q_N}(\mathscr{A}_{MCW},\mathbf{0})\in B_N]=$$
$$P[\widetilde{W}_{a_1}^{q_1}\in B_1,\ \widetilde{W}_{a_2}^{q_2}\in B_2,\ \ldots,\ \widetilde{W}_{a_n}^{q_2}\in B_n,\ \ldots,\ \widetilde{W}_{a_N}^{q_N}\in B_N]\tag{13}$$

for every $s\in\mathbb{R},\ N\in\mathbb{Z}_+,\ n\in\{1,2,\ldots,N\},\ a_n\in\mathbb{R},\ q_n\in\mathbf{K}$, $B_n\in\mathscr{B}$, where $\mathscr{B}$ is the field of Borel sets of $\mathbb{R}$. That is, given that the system starts empty and operates under the $\mathscr{A}_{MCW}$ policy, the queueing state process $\{\mathscr{W}_{s,t}^q(\mathscr{A}_{MCW},\mathbf{0}),\ q\in\mathbf{K}\}=\{W_{s,t}^q(\vec{0}),\ q\in\mathbf{K}\}$ converges in distribution to the proper stationary regime $\{\widetilde{W}_t^q,\ q\in\mathbf{K}\}$ at large times. Therefore, the system can be characterized as **globally stable**.

Remark: The above show that the $\mathscr{A}_{MCW}$ policy maximizes the set of traffic intensities $\vec{\rho}$ for which the system remains globally stable, i.e. the global stability region.

Remark: Analogous results to those given in Proposition 3.2 and Theorem 3.1 can be established for the Longest Connected Queue $\mathscr{A}_{LCQ}$ policy.

IV. Partial Stability. The Fine Structure of the Instability Region.

A more interesting question that arises is whether the workload of a certain queue is finite or infinite, for given input and modulation processes, under the MCW and the LCQ scheduling policies. It turns out that the answer depends simply on the region (cell) of the rate space $\mathbb{R}_+^K$ where $\vec{\rho}$ lies. We specify below these cells, and determine the queueing dynamics and stability behavior of the system in each one of them. In the remainder, for ease of notation we drop the dependence of the workload process on the policy and write $\mathscr{W}_{s,t}^q(\mathscr{A}_{MCW},\mathbf{x})=W_{s,t}^q(\vec{w}),\quad q\in\mathbf{K}$, and $\vec{\mathscr{W}}_{s,t}(\mathscr{A}_{MCW},\mathbf{x})=\vec{W}_{s,t}(\vec{w}),\ q\in\mathbf{K}$, accordingly.

We start by defining the following two families of sets (parameterized by $E\subseteq\mathbf{K}$), which are used to construct the aforementioned cells. For any given subset of queues $E\subseteq\mathbf{K}-\emptyset$, let

$$\Phi_E^{\mathscr{M}}=\left\{\vec{\alpha}\in\mathbb{R}_+^K:\ \sum_{q\in Q}\alpha_q<r\ \left[\sum_{C\in\mathbf{C}:\ C\cap Q\neq\emptyset,\ Q\subseteq E}P[C]\right]\right\}$$

for every $Q\subseteq E,\ Q\neq\emptyset$ and

$$\widehat{\Phi}_E^{\mathscr{M}}=\left\{\vec{\alpha}\in\mathbb{R}_+^K:\ \sum_{q\in Q}\alpha_q>r\ \left[\sum_{C\in\mathbf{C}:\ C\cap Q\neq\emptyset,\ C\subseteq E}P[C]\right]\right\},$$

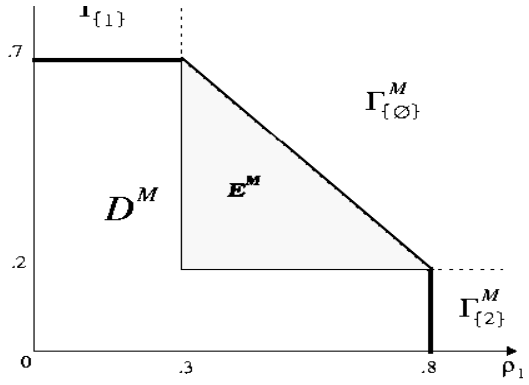

Fig. 4. The stability region and the cells of the instability region of a simple two queue system.

for some $Q \subseteq E,\ Q \neq \emptyset$ and by convention set $\Phi_{\emptyset}^{\mathscr{M}} = \mathbb{R}_+^K$. Note that $\Phi_E^{\mathscr{M}}$ and $\widehat{\Phi}_E^{\mathscr{M}}$ are convex polyhedra in $\mathbb{R}_+^K$ formed by the intersection of the K-dimensional hyperplanes defined by the inequalities in the above expressions. We next define the sets (cells)

$$\Gamma_E^{\mathscr{M}} = \Phi_E^{\mathscr{M}} \bigcap \left\{ \bigcap_{B \subseteq \overline{E},\ B \neq \emptyset} \widehat{\Phi}_{E \cup B}^{\mathscr{M}} \right\}, \tag{14}$$

for any $E \subseteq \mathbf{K}$. It should be noted that for $E = \mathbf{K}$,

$$\Gamma_{\mathbf{K}}^{\mathscr{M}} = \Phi_{\mathbf{K}}^{\mathscr{M}} = \mathbf{D}^{\mathscr{M}}. \tag{15}$$

For every $E \neq \mathbf{K}$, $\Gamma_E^{\mathscr{M}}$ is an *unbounded* convex polyhedron not containing the $\vec{0}$ vector, contrary to the case $E = \mathbf{K}$. Let $\partial\Gamma_E^{\mathscr{M}}$ be the *boundary* of $\Gamma_E^{\mathscr{M}}$ (in the standard topology of $\mathbb{R}_+^K$).

Proposition 4.1: **(Cellulization of the Rate Space $\mathbb{R}_+^K$)** For any stationary modulation process $\mathscr{M}$, we have that

$$\Gamma_E^{\mathscr{M}} \bigcap \Gamma_{E'}^{\mathscr{M}} = \emptyset, \tag{16}$$

for any $E, E' \subseteq \mathbf{K}$ such that $E \neq E'$. Moreover,

$$\bigcup_{E \subseteq \mathbf{K}} \left(\Gamma_E^{\mathscr{M}} \cup \partial\Gamma_E^{\mathscr{M}} \right) = \mathbb{R}_+^K. \tag{17}$$

That is, the family of *cells* (sets) $\{\Gamma_E^{\mathscr{M}}, E \subseteq \mathbf{K}\}$ divides up the rate space into disjoint convex polyhedra, covering it exhaustively.

Proof: The proof is omitted due to space considerations and can be found in [9]. ∎

Before proceeding to examine the fine structure of the instability region, we present in Figure 4 the stability region and the cells that comprise the instability region of a system consisting of two queues with the following connectivity sets: $C_1 = \{$ when only queue 1 is connected $\}$, with $P[C_1] = .3$, $C_2 = \{$ when only queue 2 is connected $\}$, with $P[C_2] = .2$, and $C_3 = \{$ when both queues are connected $\}$, with $P[C_3] = .5$. This would facilitate the presentation that follows.

Denoting by $\mathscr{S}$ the original system under consideration, operating under $\mathscr{A}_{MCW}$, we next introduce two families of modified systems ${}_\star\mathscr{S}|_Q, {}^\star\mathscr{S}|_Q, Q \subseteq \mathbf{K}$, which are needed in the proof of Proposition 4.3.

For any $Q \subseteq \mathbf{K}$, the system ${}_\star\mathscr{S}|_Q$ is derived from the original one $\mathscr{S}$ by imposing on the operation of the latter the following modifications:

1) Arrivals to the queues $q \in \{\mathbf{K} - Q\}$ are blocked and rejected. Therefore, the workloads of queues in $\mathbf{K} - Q$ are always zero.
2) The initial workloads w^q, for $q \in \{\mathbf{K} - Q\}$, are set to zero.
3) Queues in Q receive service according to the $\mathscr{A}_{MCW}$ policy. When the connectivity $C_t = C$ is such that $C \bigcap Q \neq \emptyset$ and $C \cap \{\mathbf{K} - Q\} \neq \emptyset$, the service power is distributed (under $\mathscr{A}_{MCW}$) to queues in $C \bigcap Q$ exclusively, since those in $C \cap \{\mathbf{K} - Q\}$ are permanently empty.

We also define for any $Q \subseteq \mathbf{K}$ the system ${}^\star\mathscr{S}|_Q$ by imposing on $\mathscr{S}$ the same rules 1 and 2, as above, but changing rule 3 to 3′ given below:

3′. For connectivities $C_t = C$ such that $C \cap Q \neq \emptyset$ and $C \cap \{\mathbf{K} - Q\} \neq \emptyset$, the server is forced to idle, even if there is workload in queues $q \in C \cap Q$. Excluding the previous situation, service is provided to queues $q \in Q$ according to the $\mathscr{A}_{MCW}$ policy, while queues in $\mathbf{K} - Q$ are again permanently empty.

The system ${}_\star\mathscr{S}|_Q$ can be viewed as $\mathscr{S}$ being restricted to Q and absorbing all service power when queues in both Q and $\mathbf{K} - Q$ are connected (border connectivity). On the contrary, ${}^\star\mathscr{S}|_Q$ rejects all service power in the border connectivity case. Note that $\mathscr{S}$ is identical to both ${}_\star\mathscr{S}|_{\mathbf{K}}$ and ${}^\star\mathscr{S}|_{\mathbf{K}}$.

Given $\vec{w} = \{w^q,\ q \in \mathbf{K}\}$, let $\vec{w}|_Q = \{w^q \mathbf{1}_{\{q \in Q\}},\ q \in \mathbf{K}\}$ be the restriction of $\vec{w}$ to Q. For any $q \in Q$, define pathwise ${}_\star W_{s,t}^q(Q, \vec{w})$ to be the workload of queue q at time t in ${}_\star\mathscr{S}|_Q$, given that it started operating at time $s < t$ with initial workload $\vec{w}|_Q$. Analogously, define ${}^\star W_{s,t}^q(Q, \vec{w})$ for ${}^\star\mathscr{S}|_Q$. Moreover, set ${}_\star W_{s,t}^q(Q, \vec{w}) = 0$, for every $q \in \mathbf{K} - Q,\ s,t \in \mathbb{R}$. Finally, let ${}_\star\vec{W}_{s,t}(Q, \vec{w}) = \{{}_\star W_{s,t}^q(Q, \vec{w}),\ q \in \mathbf{K}\}$ be the workload vector of the system ${}_\star\mathscr{S}|_Q$, and ${}^\star\vec{W}_{s,t}(Q, \vec{w}) = \{{}^\star W_{s,t}^q(Q, \vec{w}),\ q \in \mathbf{K}\}$ that of ${}^\star\mathscr{S}|_Q$.

Proposition 4.2: **(System Inequalities)** For any system ${}_\star\mathscr{S}|_Q,\ Q \subseteq K$, we have pathwise

$$\left[\sum_{q \in Q} \left\{ w_q + \sum_{j \in \mathbb{Z}} \sigma_j^q \mathbf{1}_{\{t_j \in (s,t]\}} \right\} - r \int_s^t \mathbf{1}_{\{C_z \cap Q \neq \emptyset\}} dz \right]^+ \leq \sum_{q \in Q} {}_\star W_{s,t}^q(Q, \vec{w}) \tag{18}$$

and

$${}_\star W_{s,t}^q(Q, \vec{w}) \leq W_{s,t}^q(\vec{w}) \leq {}^\star W_{s,t}^q(Q, \vec{w}) \tag{19}$$

for any $q \in Q,\ s,t \in \mathbb{R},\ s < t$ and any initial workload $\vec{w}$.

Proof: Relation (18) is proven by simply observing that the integral term in its left hand side (LHS) represents the work that the server can deliver to the queues in Q in the time interval $(s,t]$, while being connected to at least one of them. However, ${}_\star\mathscr{S}|_Q$ may not be able to utilize (absorb) all that work, because under the $\mathscr{A}_{MCW}$ policy the workload in the set of connected queues in some time interval may become zero

(thus, the queues empty and the server idle), while there is workload in others that are not connected. Moreover, the sum in the LHS is the total workload which has entered ${}_{\star}\mathscr{S}|_Q$ in $(s,t]$ plus its initial workload at time s. Inequality (18) follows immediately.

The proof of relation (19) proceeds along similar lines as the proof of Proposition 2.1 in [4] and is omitted due to space considerations, but can be found in [9]. The essence of the proof is that on any fixed sample path of $\mathscr{N}$ and $\mathscr{M}$, we observe the evolution in $(s,t]$ of two copies of the system, $\mathscr{S}$ with initial state $\mathbf{x}_1$, and ${}_{\star}\mathscr{S}|_Q$ with initial state $\mathbf{x}_2$ and establish through appropriate pathwise comparisons the desired relationship. ∎

Proposition 4.3: For any stationary and ergodic input and modulation processes $\mathscr{N}$ and $\mathscr{M}$, and for any $F \subseteq \mathbf{K}$, we have that if

$$\vec{\rho} \in \Gamma_F^{\mathscr{M}}, \tag{20}$$

then

$$\widetilde{W}_t^q < \infty, \quad \forall\, t \in \mathbb{R}, \qquad \text{for every} \qquad q \in F \tag{21}$$

$$\widetilde{W}_t^q = \infty, \quad \forall\, t \in \mathbb{R}, \qquad \text{for every} \qquad q \in \overline{F} = \mathbf{K} - F \tag{22}$$

Under this condition, the processes $\{\widetilde{W}_t^q,\ t \in \mathbb{R}\}$, $q \in \mathbf{K}$ form a *stationary* operational regime for the system, which is finite for queues $q \in F$ and infinite otherwise.

Proof: Note first that $\vec{\rho} \in \Gamma_F^{\mathscr{M}}$ implies that

$$\vec{\rho} \in \Phi_F^{\mathscr{M}} = \left\{ \vec{\alpha} \in \mathbb{R}_+^K :\ \sum_{q \in A} \alpha_q < r \left[\sum_{C \in \mathbf{C}:\ C \cap A \neq \emptyset,\ C \subseteq F} P[C] \right],\ \ A \subseteq F,\ A \neq \emptyset \right\}$$

and, for every nonempty set of queues $B \subseteq \mathbf{K} - F$ there exists a set $Q \subseteq F \cup B$, such that

$$\sum_{q \in Q} \rho_q > \sum_{C \in \mathbf{C}:\ C \cap Q \neq \emptyset,\ C \subseteq F \cup B} P[C]. \tag{24}$$

To prove (21), we consider the system ${}^{\star}\mathscr{S}|_F$ and note that due to Proposition 4.2 we have $W_{s,t}^q(\vec{0}) \leq {}^{\star}W_{s,t}^q(F,\vec{0})$ for any $q \in F,\ s,t \in \mathbb{R},\ s < t$. Notice that we can use a Loynes-type [1] procedure to construct a stationary operational regime for a subset of queues F of the system. Observe that for every $s' < s$, we have $W_{s,t}^q(\vec{0}) \leq W_{s,t}^q(\vec{W}_{s',s}(\vec{0})) = W_{s',t}^q(\vec{0})$. Therefore, since $W_{s,t}^q(\vec{0})$ is increasing as $s \to -\infty$, we can pathwise define the processes

$$\widetilde{W}_t^q = \lim_{s \to -\infty} W_{s,t}^q(\vec{0}) = \lim_{s \to -\infty} \mathscr{W}_{s,t}^q(\mathscr{A}_{MCW}, \mathbf{0}) \tag{25}$$

for every $q \in F$, which are shown below to provide a proper (finite) stationary operational regime of that subset of the system.

Therefore, in view of (25), it is enough to prove that

$$\lim_{s \to -\infty} {}^{\star}W_{s,t}(F,\vec{0}) = {}^{\star}\widetilde{W}_t(F) < \infty. \tag{26}$$

To see this, recall that ${}^{\star}\mathscr{S}|_F$ is the restriction of the actual system $\mathscr{S}$ into the set of queues F, excluding all connectivity sets that reach across to both F and $\mathbf{K} - F$ (i.e. suppressing the sets $\{C \in \mathbf{C}:\ C \cap F \neq \emptyset,\ C \cap \{\mathbf{K} - F\} \neq \emptyset\}$). Hence, ${}^{\star}\mathscr{S}|_F$ operates in isolation from the rest of the queues in $\mathbf{K} - F$. Considering ${}^{\star}\mathscr{S}|_F$ as an isolated system, we see that it falls into the realm of Proposition 3.2 and its global stability region is $\Phi_F^{\mathscr{M}}$. From (23) and Proposition 3.2, (21) follows immediately.

To prove the more intricate case of (22), we argue by contradiction, supposing that there exists some non-empty $B \subseteq \mathbf{K} - F$, such that $\widetilde{W}_t^q < \infty,\ q \in B \cup F$, while $\widetilde{W}_t^q = \infty,\ q \in (\mathbf{K} - F) - B$. Then, from (24)), there must exist some $Q \subseteq F \cup B$, such that

$$\sum_{q \in Q} \rho_q > \sum_{C \in \mathbf{C}:\ C \cap Q \neq \emptyset,\ C \subseteq F \cup B} P[C]. \tag{27}$$

Writing $Q_F = Q \cap F,\ Q_B = Q \cap B$ and $F_B = F \cup B$, and applying (21) for Q_F, we get

$$\sum_{q \in Q_F} \rho_q < \sum_{C \in \mathbf{C}:\ C \cap Q_F \neq \emptyset,\ C \subseteq F} P[C]. \tag{28}$$

Subtracting (28) from (27), we get

$$\sum_{q \in Q_B} \rho_q > \sum_{C \in \mathbf{C}:\ C \cap Q_B \neq \emptyset,\ C \subseteq F_B} P[C] + \sum_{C \in \mathscr{Y}} P[C], \tag{29}$$

where $\mathscr{Y} = \{C \in \mathbf{C} : C \subseteq F_B, C \cap Q \neq \emptyset, C \cap B \neq \emptyset, C \cap Q_B = \emptyset\}$, because $\{C \in \mathbf{C} : C \subseteq F_B, C \cap Q \neq \emptyset\} = \{C \in \mathbf{C} : C \subseteq F, C \cap Q_F \neq \emptyset\} \cap \{C \in \mathbf{C} : C \subseteq F_B, C \cap Q_B \neq \emptyset\} \cap \mathscr{Y}$ is a union of disjoint sets. From (29), we eventually have

$$\sum_{q \in Q_B} \rho_q > \sum_{C \in \mathbf{C}:\ \cup B, C \cap Q_B \neq \emptyset,\ C \subseteq F_B} P[C]. \tag{30}$$

We define next the set of active queues $R_{s,x}$ (receiving service under $\mathscr{A}_{MCW}$) at time $x > s$,

$$R_{s,x} = \{q \in C_x : W_{s,x}^q = \max_{q' \in C_x} \{W_{s,x}^{q'}\} > 0\} \subseteq C_x \tag{31}$$

given that the system has started empty at time s. In case all queues in C_x are empty, we can naturally set $R_{s,x} = \emptyset$. We then define the random time

$$\tau_s(Q_B) = \inf\{z \in [s,t] : R_{s,x} \cap F_B = \emptyset, \text{ when } C_x \cap (\mathbf{K} - F_B) \neq \emptyset,\ \ \forall x \in [z,t]\}, \tag{32}$$

which implies that for every $x \in (\tau_s(Q_B), t]$, no queue in F_B receives service under $\mathscr{A}_{MCW}$, while $C_x \cap (\mathbf{K} - F_B) \neq \emptyset$. In view of the above, we can establish the following pathwise relation

$$\sum_{q \in Q_B} W_{s,t}^q \geq \sum_{q \in Q_B} W_{s,\tau_s(Q_B)^-}^q + \sum_{q \in Q_B} \Sigma^q(\tau_s(Q_B)^-, t) - r \int_{\tau_s(Q_B)}^t \mathbf{1}_{\{C_x \subseteq F_B, C_x \cap Q_B \neq \emptyset\}} dx, \tag{33}$$

where $\Sigma^q(\tau_s(Q_B)^-, t)$ is defined as $\mathscr{V}^q(z,t) = \sum_{j \in \mathbb{Z}} \sigma_j^q \mathbf{1}_{\{t_j^q \in (z,t]\}}$.

We next prove that

$$\lim_{s \to -\infty} \tau_s(Q_B) = -\infty. \tag{34}$$

Indeed, arguing by contradiction, suppose that there exists a decreasing subsequence $\{s_a,\ a \in \mathbb{Z}_+\}$ of $\{s\}$ with $\lim_{a \to \infty} s_a = -\infty$, such that

$$\lim_{a \to \infty} \tau_{s_a}(Q_B) = \tau_* > -\infty. \tag{35}$$

From the definition of $\tau_s(Q_B)$, arguing in the spirit of (**??**), we see that there must exist a decreasing subsequence $\{s_b,\ b\in\mathbb{Z}_+\}$ of $\{s_a\}$ with $\lim_{b\to\infty} s_b=-\infty$, and a queue $q_*\in Q_B$ and another one $q'_*\in\mathbf{K}-F_B$, such that

$$W^{q'_*}_{s_b,\tau^-_{s_b}(Q_B)}\leq W^{q_*}_{s_b,\tau^-_{s_b}(Q_B)} \tag{36}$$

for every $b\in\mathbb{Z}_+$. Observe now that $W^{q_*}_{s_b,\tau^-_{s_b}(Q_B)}\leq W^{q_*}_{s_b,t}-\Sigma^{q_*}(\tau^-_{s_b}(Q_B),t)+r\int_{\tau^-_{s_b}(Q_B)}^{t}\mathbf{1}_{\{C_x\subseteq F_B,q_*\in C_x\}}dx$; hence, taking the limits as $b\to\infty$ and using (35) and the fact that $q_*\in Q_B$ (so that $\widetilde{W}^{q_*}_t<\infty$), we eventually get

$$\limsup_{b\to\infty} W^{q_*}_{s_b,\tau^-_{s_b}(Q_B)}<\infty. \tag{37}$$

Moreover, observe that $W^{q'_*}_{s_b,\tau^-_{s_b}(Q_B)}\geq W^{q'_*}_{s_b,t}-\Sigma^{q'_*}(\tau^-_{s_b}(Q_B),t)$, thus, again taking the limits as $b\to\infty$ and using (35) and the fact that $q'_*\in\mathbf{K}-F_B$ (so that $\widetilde{W}^{q'_*}_t=\infty$), we get

$$\liminf_{b\to\infty} W^{q'_*}_{s_b,\tau^-_{s_b}(Q_B)}=\infty. \tag{38}$$

Finally, taking the limits as $b\to\infty$ in (36) and using (37) and (38) we get a contradiction proving (34).

Based on (34), we can now clinch the proof, by dividing (33) by $(t-\tau_s(Q_B))$, letting $s\to-\infty$, noting that $\widetilde{W}^q_t<\infty$ because $q\in Q_B\subseteq B$, and using (34) and the fact that

$$\lim_{z\to-\infty}\frac{\Sigma^q(z,t)}{t-z}=\rho_q. \tag{39}$$

We then get

$$0\geq\sum_{q\in Q_B}\rho_q-\sum_{C\in\mathbf{C}:\ C\cap Q_B\neq\emptyset,\ C\subseteq F_B}P[C], \tag{40}$$

which is a contradiction to (30). This completes the proof of Proposition 4.3. ∎

Theorem 4.1: **(Partial Stability under the $\mathscr{A}_{MCW}$ Scheduling Policy)**

For any stationary and ergodic input and modulation processes $\mathscr{N}$ and $\mathscr{M}$

$$\vec{\rho}\ \in\ \Gamma^{\mathscr{M}}_F. \tag{41}$$

the following are true:

a) For every queue $q\in\mathbf{K}-F$, we have

$$\lim_{t\to\infty}P[W^q_{s,t}(\vec{0})\leq b]=0 \tag{42}$$

for every $b\in\mathbb{R}$, so the queues in $\mathbf{K}-F$ can be characterized as **unstable**.

b) For any $q\in F$, we have that

$$\lim_{t\to\infty}P[W^{q_1}_{s,t+a_1}(\vec{0})\in B_1,\ W^{q_2}_{s,t+a_2}(\vec{0})\in B_2,$$

$$...,\ W^{q_n}_{s,t+a_n}(\vec{0})\in B_n,\ ...,\ W^{q_N}_{s,t+a_N}(\vec{0})\in B_N]=$$

$$\lim_{t\to\infty}P[\mathscr{W}^{q_1}_{s,t+a_1}(\mathscr{A}_{MCW},\mathbf{0})\in B_1,\ \mathscr{W}^{q_2}_{s,t+a_2}(\mathscr{A}_{MCW},\mathbf{0})\in B_2,$$

$$...,\ \mathscr{W}^{q_N}_{s,t+a_N}(\mathscr{A}_{MCW},\mathbf{0})\in B_N]=$$

$$P[\widetilde{W}^{q_1}_{a_1}\in B_1,\ \widetilde{W}^{q_2}_{a_2}\in B_2,\ ...,\ \widetilde{W}^{q_2}_{a_n}\in B_n,\ ...,\ \widetilde{W}^{q_K}_{a_N}\in B_N] \tag{43}$$

for every $s\in\mathbb{R}$, $N\in\mathbb{Z}_+$, $n\in\{1,2,...,N\}$, $a_n\in\mathbb{R}$, $B_n\in\mathscr{B}$, where $\mathscr{B}$ is the set of Borel sets of $\mathbb{R}$. That is, given that the system starts empty and operates under the $\mathscr{A}_{MCW}$ policy, the workload processes $\mathscr{W}^q_{s,t}(\mathscr{A}_{MCW},\mathbf{0})=W^q_{s,t}(\vec{0})$ of queues $q\in F$ converge in distribution to proper stationary regime $\widetilde{W}^q_t$ (almost surely finite) at large times. Therefore, the queues in F can be characterized as **stable**.

Proof: The proof is analogous to that of Theorem 3.1, using now Proposition 4.3. ∎

Remark 4.1: Using the forward argument in Proposition 4.3 we can actually show that $\lim_{t\to\infty}W^q_{s,t}(\vec{0})=\infty$ almost surely for $q\in\mathbf{K}-F$.

Remark 4.2: **(Crossing the Global Stability Region Boundary. The Transition to Instability).** Note that the system switches from a global strong stability mode, when operating under $\mathscr{A}_{MCW}$, to an at least partial instability mode under any policy $\mathscr{A}\in\mathbf{A}$ (including $\mathscr{A}_{MCW}$), as $\vec{\rho}$ crosses from cell $\Gamma^{\mathscr{M}}_{\mathbf{K}}$ to cell $\Gamma^{\mathscr{M}}_F$, where F is any proper subset of $\mathbf{K}$. However, under the $\mathscr{A}_{MCW}$ policy, if $\vec{\rho}\in\Gamma^{\mathscr{M}}_F$ then all queues in F remain strongly stable, while all queues in $\overline{F}$ become unstable.

Remark 4.3: **(The Critical Case. Stability on the Boundaries.)** There is a final issue to be discussed concerning the stability mode of the system in the case that $\vec{\rho}$ belongs to $\partial\Gamma^{\mathscr{M}}_F$ for some $F\subseteq\mathbf{K}$. In that case, the system can not be characterized in terms of stability in an *almost surely* manner. Indeed, Proposition 4.3 and Theorem 4.1 collapse. Therefore, the system may exhibit distinct behaviors on different sample paths. No almost sure characterization of system behavior can be established in this case.

REFERENCES

[1] F. Baccelli and P. Bremaud, *Elements of Queueing Theory*, 1994, Springer: New York

[2] N. Bambos and G. Michailidis, Queueing and Scheduling in Random Environments, *Adv. Applied Prob.*, vol. 36, 2004, pp 293-317.

[3] N. Bambos and G. Michailidis, On Parallel Queueing with Random Server Connectivity and Routing Constraints, *Prob. Engin. Info. Sciences*, vol. 16, 2002, pp 185-203.

[4] N. Bambos and G. Michailidis, Queueing Networks of Random Link Topology: Stationary Dynamics Of Maximal Throughput Schedules, *Queueing Sys. Theory & Appl.*, 50, 2005, pp 5-52.

[5] A. Brandt, P. Franken and B. Lisek, *Stationary Stochastic Models*, 1990, Wiley: Chichester

[6] R.T. Buche and H.J. Kushner, Control of Mobile Communication Systems with Time-Varying Channels via Stability Methods, *IEEE Trans. Aut. Control*, vol. 49, 2004, pp. 1954-1962

[7] A. Ganti, E. Modiano and J.N. Tsitsiklis, Optimal Transmission Scheduling in Symmetric Communication Models with Intermittent Connectivity, Technical Report, LIDS, MIT, 2004

[8] C. Lott and D. Teneketzis, On the Optimality of an Index Rule in Multichannel Allocation for Single-hop Mobile Networks with Multiple Service Rates, *Prob. Engin. Info. Sciences*, vol. 14, 2000, pp 259-297.

[9] G. Michailidis and N. Bambos, On the Singular Behavior of a Queueing System with Random Connectivity, Technical Report, Dept. of Statistics, The University of Michigan, 2005

[10] L. Tassiulas, Scheduling and Performance Limits of Networks with Constantly Changing Topology, *IEEE Trans. Info. Theory*, vol. 39, 1997, pp 466-478

[11] K. Wasserman and T.L. Olsen, On Mutually Interfering Parallel Servers subject to External Disturbances, *Operat. Res.*, vol. 49, 2001, pp 700-709