Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

On the stability of polling models with multiple servers

G.D. Down

Department of Operations Reasearch, Statistics, and System Theory

BS-R9605 1996

Report BS-R9605
ISSN 0924-0659

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

On the Stability of Polling Models with Multiple Servers

D.G. Down

CWI

P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

The stability of polling models is examined using associated fluid limit models. Examples are presented which generalize existing results in the literature or provide new stability conditions while in both cases providing simple and intuitive proofs of stability. The analysis is performed for both general single server models and specific multiple server models.

AMS Subject Classification (1991): 60K25, 90B22

Keywords & Phrases: polling model, multiple servers, stability, fluid model

Note: The author's present address is INRIA, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France

1. Introduction

In this paper, we consider a polling model that can be described as a system of several queues, attended by multiple servers. These systems see numerous applications in the areas of distributed computer systems, communications networks, and manufacturing systems. For some specific applications, see Morris and Wang [17] and Gamse and Newell [11]. A good general reference in this area is Levy and Sidi [14].

The system is completely determined by the service policies employed and statistical assumptions on server routing, arrivals to the queues, service times, and the switchover times when a server moves from one queue to another. The stability of the system in terms of these parameters is of fundamental importance and in the case of single server models, this question has been studied by many authors. Among these are Altman, Konstantopoulos and Liu [1], Altman and Spieksma [2], Borovkov and Schassberger [3], Fricker and Jaïbi [9, 10], Georgiadis and Szpankowski [12], and Georgiadis, Szpankowski and Tassiulas [13].

We will consider the problem of stability, but in contrast to the above references, we will use the recent approach which involves the analysis of a fluid model as developed in the work of Dai [6] and Dai and Meyn [5]. These techniques have become very popular in the analysis of the stability of multiclass queueing networks, particularly those with a re-entrant structure. The number of references in this area is too numerous to mention here, some are given in [5]. In the area of polling models, it appears that the only work that applies these techniques is in [5] itself and one part of this paper (the examination of the single server model) is in fact a generalization of work performed there.

The advantages of using a fluid model are threefold. First, it allows us to generalize the results in the literature given at the end of the second paragraph. In particular, the assumption that the arrival stream is Poisson (assumed throughout those references) may be relaxed to general renewal processes, yielding similar stability conditions. Second, new stability conditions may be proved for certain multiple server systems, related to those studied by van der Mei and Borst [19]. Here we prove that versions of their conjectures indeed hold. Third, the simplicity of the stability proofs themselves is appealing. It suggests that these techniques may have wide application, including to other polling models not covered in this work.

The layout of the paper is as follows: In Section 1.1, the model is described in detail. The fluid model is introduced in Section 2 and the stability of various polling models is discussed. In Section 3, the stability proofs themselves are provided. Finally Section 4 contains some concluding remarks.

1.1 Model description

The model considered is a generalization of the single server cyclic polling model considered in Fricker and Jaïbi [9], which in turn covers many of the references given in the Introduction. The polling system is composed of K infinite buffer capacity queues and M servers.

Arrivals occur at queue k with i.i.d. interarrival times $\{\xi_k(n), n \geq 1\}$ (it is a simple modification to incorporate batch arrivals). The service times at queue k are also i.i.d. and are denoted by $\{\eta_k(n), n \geq 1\}$. Let $\lambda_k = 1/\mathsf{E}[\xi_k(1)]$ and $\mu_k = 1/\mathsf{E}[\eta_k(1)]$ be the arrival and service rates, respectively. We assume $\lambda_k > 0$, $k = 1, \ldots, K$. Of course, for each k, $\rho_k := \lambda_k/\mu_k < 1$ is necessary for stability.

There is a set of S possible service policies (at most dependent on the number of customers present at the queue to be serviced), enumerated by $1, \ldots, S$. With x the number of customers at the queue at the beginning of a service period and $N_s(x)$ defined as the number of customers served with policy s during a service period, we assume that

(S1) $\lim_{x\to\infty} \mathsf{E}[N_s(x)] \longrightarrow \bar{N}_s > 0$

(S2) $\mathsf{E}[N_s(x)] \leq \bar{N}_s, \quad \forall x.$

Characterizing which policies satisfy these assumptions is a task we will not undertake here, but it is clear that typical policies such as exhaustive, gated and k-limited policies satisfy the above criteria, as well as the monotone policies of [9].

Remark. It appears that (S1) and (S2) may be relaxed. In particular, it may be possible to examine stability if multiple limits exist in (S1), by looking at worst case scenarios. However, it is possible that sharp stability regions could no longer be deduced by the methods introduced here. The assumption (S2) may also be relaxed substantially, in fact it is conjectured that it may be removed completely and stability may hold under (S1) only. Also note that while at first glance restrictive, natural policies satisfy (S2).

The routing of server m and the corresponding service policy implemented are given as follows: At the end of a visit to queue j using service policy s, the server moves to queue j' and uses service policy s' with probability $r^m_{js,j's'}$. At the nth occurrence, this is accompanied by a switchover time of length $\delta^m_{js,j's'}(n)$. There is also a possible limit on the number of servers at j', given by $m_{j'}$. If the number of servers at queue j' is less than $m_{j'}$, then the server in question begins service with policy s'. Otherwise, the system acts as follows: If there are $k > m_{j'}$ servers at queue j' at a particular instant, then $k - m_{j'}$ servers depart, with these being chosen from those present with equal probability (this can be easily generalized to any set of positive probabilities). If a server leaves while in the process of serving a customer, an arriving server continues the service already begun, i.e. there is a hand-off of the customer. Note that this assumption on server routing may not be considered to be the natural one. A possible choice may be for a server arriving at a queue with the maximum number of servers already present to be the one to choose a new queue. The reason for our particular assumption and some consequences of the alternate choice above will be discussed in more detail at the end of Section 3.2. The sequences $\{\delta^m_{js,j's'}(n)\}$ are mutually i.i.d. with mean $\mathsf{E}[\delta^m_{js,j's'}(1)] \geq 0$. Let $\{p^m_{js}\}$ be the unique invariant distribution for the Markov chain with transition matrix $(r^m_{js,j's'})$. Also, define $\delta^*_m := \sum_{js,j's'} p^m_{js} r^m_{js,j's'} \mathsf{E}[\delta^m_{js,j's'}(1)]$.

2. The fluid model: construction and stability conditions

In this section, a fluid model is constructed that may be used to analyze the stability of the polling system introduced in the previous section. In Section 2.1, a Markov process is constructed which describes the system behaviour. This is a necessary precursor for the remaining analysis. In Section 2.2, the fluid model is constructed and the implications of its behaviour for the stability of the system are outlined in Theorem 2.1 and Theorem 2.2, from the work of Dai and Meyn [5] and Meyn [15] respectively. Finally in Section 2.3, stability conditions are given for various models that will be examined in the remainder of the paper.

2.1 Markov process description

Let $Q_k(t)$ be the number of customers at queue k, including any customers in service which originated from that queue. Also, let $A_k(t)$ be the residual arrival time at queue k. For each server m, let $H_m(t)$ be the ordered pair consisting of the queue being serviced (or to be serviced next if the server is switching) and the service policy in use (or to be used if the server is switching). Also, let $B_m(t)$ be the residual service time (set to zero if the server is switching), $B_m^0(t)$ be the residual switchover time (set to zero if the server is currently serving a customer), and $C_m(t)$ indicate the number of customers serviced during the current visit to the queue given in $H_m(t)$. Using these definitions it is straightforward to check that

$$X^T(t) := (Q_k(t), A_k(t), H_m(t), B_m(t), B_m^0(t), C_m(t) : k = 1, \ldots, K; m = 1, \ldots, M), \tag{2.1}$$

is a Markovian state for the polling model under consideration, evolving on the state space $\mathsf{X} = \mathbb{Z}_+^K \times \mathbb{R}_+^K \times (\{1, \ldots, K\} \times \{1, \ldots, S\})^M \times \mathbb{R}_+^M \times \mathbb{R}_+^M \times \mathbb{Z}_+^M$. The residual times $A_k(t)$, $B_m(t)$ and $B_m^0(t)$ are taken to be right continuous and, as in [6], the process $\mathbf{X}$ may be shown to have the strong Markov property.

2.2 Stability of fluid models

For a given initial condition $x \in \mathsf{X}$, a fluid model for the system is developed as follows: Let $Q_k^x(t)$ be the queue length at queue k at time t and $T_{m,k}^x(t)$ be the cumulative service allocation process at queue k for server m. Also, let $T_{m,k}^{x,0}(t)$ be the cumulative time by time t that server m spends in switching from queue k. Now, suppose that the function $(\bar{Q}(\cdot), \bar{T}_m(\cdot), \bar{T}_m^0(\cdot) : m = 1, \ldots, M)$ is a limit point of

$$(\frac{1}{|x|}Q^x(|x|t), \frac{1}{|x|}T_m^x(|x|t), \frac{1}{|x|}T_m^{x,0}(|x|t) : m = 1, \ldots, M)$$

when $|x| \longrightarrow \infty$ and where $Q^x(t)$, $T_m^x(t)$ and $T_m^{x,0}(t)$ are vectors whose kth components are, for $1 \le k \le K$, $Q_k^x(t)$, $T_{m,k}^x(t)$ and $T_{m,k}^{x,0}(t)$ respectively. Then $(\bar{Q}(t), \bar{T}_m(t), \bar{T}_m^0(t) : m = 1, \ldots, M)$ is a (delayed) fluid limit of the system. The set of all possible fluid limits will be called the fluid model. In Section 3, conditions that these fluid limits satisfy will be given in detail. Note that the term *delayed* is used above to indicate that the initial condition may be taken to grow in the direction of the residual times. It is not difficult to see that this creates a finite delay in the fluid model.

At this point, connections between the stability of the fluid model and that of the original system (i.e. of the Markov process $\mathbf{X}$) are given. The work in the rest of this subsection may be found in [5] and [15]. In order to make the desired connections, the following three assumptions are made on the network:

(A1) $\xi_1, \ldots, \xi_K, \quad \eta_1, \ldots, \eta_K, \quad \delta_{js,j's'}^m : j, j' \in \{1, \ldots, K\}, \quad s, s' \in \{1, \ldots, S\}, \quad m = 1, \ldots, M$ are mutually independent and i.i.d. sequences.

(A2) For some integer $p \ge 1$,

$$\mathsf{E}[\xi_k^{p+1}(1)] < \infty, \quad \mathsf{E}[\eta_k^{p+1}(1)] < \infty \quad \text{for} \quad k = 1, \ldots, K \quad \text{and}$$

$$\mathsf{E}[\delta_{js,j's'}^m(1)^{p+1}] < \infty \quad \text{for} \quad j, j' \in \{1, \ldots, K\}, \quad s, s' \in \{1, \ldots, S\}, \quad m = 1, \ldots, M.$$

(A3) For $k = 1, \ldots, K$, there exists some positive function $q_k(x)$ on $\mathbb{R}_+$ and some integer j_k such that

$$\mathsf{P}(\xi_k(1) \ge x) > 0 \quad \text{for all} \quad x > 0, \tag{2.2}$$

$$\mathsf{P}(\xi_k(1) + \cdots + \xi_k(j_k) \in dx) \ge q_k(x)dx \quad \text{and} \quad \int_0^\infty q_k(x)dx > 0.$$

Remark. The independence assumptions in (A1) may be relaxed: see the remark after Proposition 2.1 of Dai [6].
Remark. As in [5], (A3) may be replaced by

(A3') For the Markov process $\mathbf{X}$, every compact subset of X is small.

The assumption (A3) (or (A3')) is a technical one required to deduce ergodicity of the network. The reader is referred to Lemma 3.4 of Meyn and Down [16] for this argument, and Sigman [18] for examples where the condition (2.2) is not necessary.

The fluid model is said to be stable if there exists a fixed time t_0 such that $\bar{Q}(t) = 0, t \geq t_0$ for any fluid limit. In fact, Chen [4] shows that the set of fluid limits that need to be examined to conclude the stability results given below is given by the set of undelayed fluid limits. So, the fluid model may be equivalently said to be stable if $\bar{Q}(t) = 0, t \geq t_0$ for any fluid limit with $|\bar{Q}(0)| = 1$. The following theorems relate stability of the fluid model with that of the original system.

Theorem 2.1 *(Theorem 4.1, [5]) Assume that the fluid model is stable, and that (A1) and (A2) hold. Then*

(i) *For some constant κ_p, and for each initial condition $x \in \mathsf{X}$*

$$\limsup_{t\to\infty} \frac{1}{t} \int_0^t \mathsf{E}_x[|Q(s)|^p] ds \leq \kappa_p,$$

where p is the integer used in (A2).

If in addition (A3) holds, then for each initial condition:

(ii) *The transient moments converge to their steady state values:*

$$\lim_{t\to\infty} \mathsf{E}_x[Q_k(t)^r] = \mathsf{E}_\pi[Q_k(0)^r] \leq \kappa_r, \quad \textit{for} \quad r = 1, \ldots, p, \quad k = 1, \ldots, K,$$

where π is the invariant probability for $\mathbf{X}$.

(iii) *The first moment converges at rate t^{p-1}:*

$$\lim_{t\to\infty} t^{p-1} |\mathsf{E}_x[Q(t)] - \mathsf{E}_\pi[Q(0)]| = 0.$$

(iv) *The strong law of large numbers holds:*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t Q_k^r(s) ds = \mathsf{E}_\pi[Q_k^r(0)], \quad \mathsf{P}_x - a.s., \quad \textit{for} \quad r = 1, \ldots, p, \quad k = 1, \ldots, K.$$

For instability of the network, the following result will be used:

Theorem 2.2 *(Theorem 3.2 of [15]) Suppose that the fluid model is unstable in the sense that for some $\varepsilon_0, c_0 > 0$,*

$$|Q(T)| \geq \varepsilon_0 T - c_0, \quad T \geq 0,$$

for all initial conditions $Q(0)$, with $|Q(0)| = 1$. Then, for any $1 \geq q > 0$, there exists $B < \infty$ such that whenever $|x| \geq B$,

$$\mathsf{P}_x\{\mathbf{X} \to \infty\} \geq q.$$

2.3 Stability conditions for particular networks

For the rest of the paper, three particular cases of the polling models given in Section 1.1 will be examined. It will be seen that the use of a fluid model for each of these systems provides for simple stability proofs. For convenience, the results are summarized here.

The single server model. This case corresponds to $M = 1$. Theorem 2.3 is both a precursor to the analysis of multiple server models and itself also generalizes many existing results (in form, it most closely resembles Theorem 2 of [9]). For notational convenience, the dependence on m may be dropped wherever it appears. The proof of Theorem 2.3 is delayed until Section 3.1.

Theorem 2.3 *Consider the following value:*

$$\rho = \sum_{k=1}^{K} \rho_k + \max_{1 \le j \le K} \left(\frac{\lambda_j}{\sum_{s=1}^{S} p_{js} \bar{N}_s} \right) \delta^*.$$

If

(i) $\rho < 1$ *then the network is stable, i.e. Theorem 2.1 holds.*

(ii) $\rho > 1$ *then the network is unstable, i.e. Theorem 2.2 holds.*

The multiple server system with unlimited service. In this case, with $M \ge 1$ servers, the following model is considered: At each queue k, only a maximum number of servers may be present at any particular time, this given by m_k. It is assumed that for each m, if $\sum_{s=1}^{S} p_{js}^m > 0$, then $\sum_{s=1}^{S} p_{js}^m \bar{N}_s = \infty$, i.e. with positive probability there is unlimited service. Stability conditions for this system have not been proved, as far as we know, only conjectures have been provided for a related model, see van der Mei and Borst [19]. The following result holds for this system, with the proof being given in Section 3.2. A server m is said to be *capable* of visiting a queue k if $\sum_{s=1}^{S} p_{ks}^m > 0$.

Theorem 2.4 *If the given multiple server system satisfies the following:*

(i) $\rho_i < m_i, \quad i = 1, \ldots, K$

(ii) *For each set* $I \subseteq \{1, \ldots, K\}$, *the queues* $i \in I$ *are capable of being visited by more than* $\sum_{i \in I} \rho_i$ *servers,*

then the network is stable, i.e. Theorem 2.1 holds.

If either

(iii) $\rho_i > m_i$, *for some* $i \in \{1, \ldots, K\}$ *or*

(iv) *For some set* $I \subseteq \{1, \ldots, K\}$, *the queues* $i \in I$ *are capable of being visited by at most strictly less than* $\sum_{i \in I} \rho_i$ *servers,*

then the network is unstable, i.e. Theorem 2.2 holds.

The multiple server system with limited service. Consider the model of Theorem 2.4, with the following modifications: First, let $m_k = M$, for $k = 1, \ldots, K$ and relax the condition that unlimited service occurs, i.e. allow for the possibility that $\sum_{s=1}^{S} p_{js}^m \bar{N}_s < \infty$ for $j = 1, \ldots, K$. We will further simplify the model by assuming that all servers are *identical*, by which we mean that the quantities that are m-dependent (routing and switching times) are statistically the same for each server, while maintaining the mutual independence assumptions. Thus, for notational convenience, we may drop the dependence on m throughout. Once again, stability conditions for a simpler version of this system are conjectured in [19]. The following theorem gives general conditions. The proof is postponed to Section 3.3.

Theorem 2.5 *Consider the value*

$$\rho = \sum_{k=1}^{K} \rho_k + \max_{1 \le j \le K} \left(\frac{M \lambda_j}{\sum_{s=1}^{S} p_{js} \bar{N}_s} \right) \delta^*.$$

If

(i) $\rho < M$ *then the network is stable, i.e. Theorem 2.1 holds.*

(ii) $\rho > M$ *then the network is unstable, i.e. Theorem 2.2 holds.*

Remark. Note that the case when $m_k < M$ for at least one k is not covered in Theorem 2.5. It is not even clear that stability in this case is determined by first order statistics. The interaction between servers appears to be quite complicated and seemingly defies a simple analysis.

3. Stability Proofs

It is easy to check (see [6]) that the fluid model satisfies (but is not necessarily determined by) the following set of equations:

$$\begin{aligned}
\bar{Q}_k(t) &= \bar{Q}_k(0) + \lambda_k t - \sum_{m=1}^{M} \mu_k \bar{T}_{m,k}(t) \quad \text{for } k = 1, \ldots, K, && (3.1)\\
\bar{Q}_k(t) &\ge 0 \quad \text{for } k = 1, \ldots, K,\\
\bar{T}_{m,k}(0) &= 0\\
\bar{T}_{m,k}(\cdot) &\quad \text{is nondecreasing for } k = 1, \ldots, K, \quad m = 1, \ldots, M,\\
\sum_{k=1}^{K} \bar{T}^0_{m,k}(t) + \bar{T}_{m,k}(t) &= t \quad \text{for} \quad m = 1, \ldots, M. && (3.2)
\end{aligned}$$

This set of equations will be the basis for the proofs that follow. In Section 3.1, the proof of Theorem 2.3 is given, Section 3.2 contains the proof of Theorem 2.4 and Section 3.3 has the proof of Theorem 2.5.

3.1 The single server model

In this section, the proof of Theorem 2.3 is provided. Note that the work in this section can be seen as a generalization of the work in Section 4.3 of [5].

Let G be the set of queues for which $\sum_{s=1}^{S} p_{js} \bar{N}_s = \infty$. If $G = \{1, \ldots, K\}$, then Theorem 2.3 trivially holds. Note that in determining ρ, the term in the maximum expression will be zero for queues in G. In the following lemma it will be seen that either the network is unstable (Theorem 2.3 (ii) holds) or after some finite time, the queues in G for the fluid model will become empty and remain so.

Lemma 3.1 *If* $\sum_{k \in G} \rho_k < 1$ *then for some finite* T *and for all* $t \ge T$, $\sum_{k \in G} \bar{Q}_k(t) = 0$. *Otherwise, if* $\sum_{k \in G} \rho_k > 1$, *then* $\rho > 1$ *and Theorem 2.3 (ii) holds.*

Proof Consider the work for the fluid model restricted to the queues in G, i.e.

$$W_G(t) = \sum_{k \in G} \frac{\bar{Q}_k(t)}{\mu_k}.$$

Clearly, if $\sum_{k \in G} \rho_k < 1$ and $\sum_{k \in G} \bar{Q}_k(t) > 0$ then

$$\dot{W}_G(t) = -1 + \sum_{k \in G} \rho_k < 0,$$

which implies that at $T = (\sum_{k \in G} \bar{Q}_k(0))/(1 - \sum_{k \in G} \rho_k)$, $W_G(T) = 0$. The first part of the lemma then follows from the positivity of $\bar{Q}_k(t)$.

If $\sum_{k \in G} \rho_k > 1$, the second part of the lemma follows from the fact that $\dot{W}_G(t) > 0$. ■

So, now assume $\sum_{k \in G} \rho_k < 1$, as otherwise from Lemma 3.1, Theorem 2.3 (ii) holds. For the remainder of the proof, it suffices to look at the fluid model with initial condition $\sum_{k \in G^c} \bar{Q}_k(0) = 1$ and $\sum_{k \in G} \bar{Q}_k(0) = 0$, as if this is not the case, from Lemma 3.1 we can look at the system beginning at time T, when the queues in G have emptied. Also, Lemma 3.1 and (3.1) imply directly that, for the initial condition under consideration,

$$\dot{\bar{T}}_k(t) = \rho_k, \quad \text{for all } t \geq 0,\ k \in G. \tag{3.3}$$

The following is a straightforward generalization of Proposition 4.2 (v) and (vi) of [5].

Lemma 3.2 *For the single server model under consideration, let* $\mathsf{E}[N_s(x)] = \bar{N}_s, \forall x$. *Then*

(i) *For all* $k,\ j \in G^c$

$$\frac{\bar{T}^0_k(t)}{\delta_k} = \frac{\bar{T}^0_j(t)}{\delta_j}$$

where $\delta_k = \sum_{s,j's'} p_{ks} r_{ks,j's'} \mathsf{E}[\delta_{ks,j's'}(1)]$.

(ii) *For all* $k \in G^c$,

$$\mu_k \dot{\bar{T}}_k(t) = (\sum_{s=1}^{S} p_{js} \bar{N}_s) \dot{\bar{T}}^0_k(t),$$

whenever $\bar{Q}_k(t) > 0$.

■

Now, examining the work in the system for the fluid model, $W(t) = \sum_{k=1}^{K} \frac{\bar{Q}_k(t)}{\mu_k}$, following the same procedure as in Section 4.3 of [5], we find that $\dot{W}(t) < 0$ for any $\bar{Q}(t) \neq 0$ iff $\rho < 1$ and $\dot{W}(t) > 0$ for any $\bar{Q}(t)$ if $\rho > 1$. This immediately implies the results of Theorem 2.3. We now outline the proof. The intermediate details are left to the reader, however for the most part these are simple algebraic manipulations.

Using (3.1),

$$\dot{W}(t) = \sum_{k=1}^{K} \rho_k - \sum_{k=1}^{K} \dot{\bar{T}}_k(t). \tag{3.4}$$

Given (S1) and (S2), it may be seen that the stability of the system under consideration is equivalent to one in which for the queues in G^c, $\mathsf{E}[N_s(x)] = \bar{N}_s, \forall x$. So, from Lemma 3.2 (i), (3.2) and (3.3) we have

$$\sum_{k=1}^{K} \dot{\bar{T}}_k(t) = 1 - \delta^* \dot{v}(t), \tag{3.5}$$

where $v(t)$ is the common value of $\bar{T}_k^0(t)/\delta_k$. Now, let H be the set of queues for which $\bar{Q}_k(t) > 0$ (note that $H \subset G^c$). Using the positivity of $\bar{Q}_k(t)$ and Lemma 3.2 (ii) we may write

$$\left[\delta^* + \sum_{k\in H} \frac{\sum_{s=1}^S p_{ks}\bar{N}_s}{\mu_k}\right] \dot{v}(t) \quad = \quad 1 - \sum_{k\in H^c} \rho_k.$$

Substituting this expression for $\dot{v}(t)$ in (3.5) and then for $\sum_{k=1}^K \dot{\bar{T}}_k(t)$ in (3.4) yields

$$\dot{W}(t) = \frac{\sum_{k\in H}\left[\delta^*\rho_k - (1-\sum_{j=1}^K \rho_j)\frac{\sum_{s=1}^S p_{ks}\bar{N}_s}{\mu_k}\right]}{\delta^* + \sum_{k\in H}\frac{\sum_{s=1}^S p_{ks}\bar{N}_s}{\mu_k}}. \tag{3.6}$$

Now, a simple rearrangement of terms yields the result, as the right side of (3.6) is negative if and only if $\rho < 1$ and positive if $\rho > 1$.

3.2 The multiple server system with unlimited service

The proof of Theorem 2.4 is straightforward, which shows the power of the fluid model to provide simple, intuitive proofs.

We once again examine the work in the system for the fluid model, which as a reminder is given by

$$W(t) = \sum_{k=1}^K \frac{\bar{Q}_k(t)}{\mu_k}.$$

From (3.1),

$$\dot{W}(t) = \sum_{k=1}^K (\rho_k - \sum_{l=1}^M \dot{\bar{T}}_{l,k}(t)). \tag{3.7}$$

Let G be the set of queues for which $\bar{Q}_k(t) > 0$ or $\dot{W}_k(t) := \dot{\bar{Q}}_k(t)/\mu_k > 0$ when it exists. In this case (3.7) and the positivity of $\bar{Q}_k(t)$ imply

$$\dot{W}(t) = \sum_{k\in G} (\rho_k - \sum_{l=1}^M \dot{\bar{T}}_{l,k}(t)).$$

Now, the server routing scheme, (i) and (ii) imply that $\dot{W}(t) < 0$ for any choice of G. Thus, by ranging over all possible initial conditions, we have the first part of Theorem 2.4. If (iii) holds, the second part of Theorem 2.4 follows upon defining $W_i(t)$ as the work in the fluid model at queue i (where i satisfies (iii)) and noting that $\dot{W}_i(t) > 0, \forall t$. If (iv) holds, the second part of Theorem 2.4 follows in a similar manner, by taking $W_I(t) := \sum_{i\in I} W_i(t)$ and noting that $\dot{W}_I(t) > 0, \forall t$.

Remark. With the assumptions on server routing, the proof above is extremely simple. However, if the server routing is changed such that any server arriving to a queue with the maximum number of servers already in attendance is forced to move away, then there is a difficulty which arises in the analysis. Consider the following example: Let $K = 2$ and $\lambda_j = 2$, $\mu_j = 3$, $j = 1, 2$. Also, let $M = 2$ with server 1 capable of serving both queues, but server 2 capable of serving only queue 2. We also let $m_j = 1$, $j = 1, 2$. It is easy to see that if $\bar{Q}_1(0) = 0$, $\bar{Q}_2(0) = 1$ and server 1 is at queue 2 using an exhaustive policy, then $\dot{W}(0) = 1/3 > 0$. This situation is avoided with the routing considered in this paper, as in the fluid model for this situation, at time 0 server 2 is at queue 2 and server 1 at queue 1. Then, $\dot{W}(0) = -1/3$.

Upon looking at the first case, one may argue that after some finite time, queue 2 empties for the fluid model and from that point $\dot{W}(t) < 0$ until the fluid model empties. This argument may be extended to the general two queue case, however it is difficult to see how this could extend to a system with more than two queues, as the possibility of idle servers arises in a very complex way. One has to confront the problem that at *every* switching time (rather than just at time 0), the above problem may occur. Nevertheless, it is conjectured that the same stability conditions hold in this more involved case.

Of course, even with the assumptions that we have employed here, a stability proof without resorting to a fluid model analysis appears to be difficult.

3.3 The multiple server system with limited service

In this section, Theorem 2.5 will be proved. This is not difficult, in fact, once one realizes that the restriction $m_k = M$, $k = 1, \ldots, K$ implies that the servers operate independently. Thus the results of Section 3.1 may be employed for each server in turn.

To see this, create M copies of the polling model, but with the arrival streams thinned by a factor of $1/M$ and with the mth copy being served only by server m. Otherwise, the properties of the model are the same. Denote the work in the system for the fluid model for each of these systems by $W_m(t)$. If we sum $W_m(t)$ over m, the result is identical to the work in the system for the fluid model for the original multiple server system. We have already examined the single server system in Section 3.1, so can apply the results there to each of the M copies.

The expression (3.6) is valid for each $\dot{W}_m(t)$, i.e.

$$\dot{W}_m(t) = \frac{\sum_{k\in H_m}\left[\delta^*\frac{\rho_k}{M} - (1-\sum_{k=1}^{K}\frac{\rho_k}{M})\frac{\sum_{s=1}^{S}p_{ks}\bar{N}_s}{\mu_k}\right]}{\delta^* + \sum_{k\in H_m}\frac{\sum_{s=1}^{S}p_{ks}\bar{N}_s}{\mu_k}} \tag{3.8}$$

where H_m in (3.8) is the analogue of H in (3.6) for the particular copy of the polling model. Now, summing $\dot{W}_m(t)$ yields

$$\dot{W}(t) = \sum_{m=1}^{M}\left(\frac{\sum_{k\in H_m}\left[\delta^*\frac{\rho_k}{M} - (1-\sum_{k=1}^{K}\frac{\rho_k}{M})\frac{\sum_{s=1}^{S}p_{ks}\bar{N}_s}{\mu_k}\right]}{\delta^* + \sum_{k\in H_m}\frac{\sum_{s=1}^{S}p_{ks}\bar{N}_s}{\mu_k}}\right)$$

which, as in the single server case is negative if and only if $\rho < M$ and positive when $\rho > M$.

4. Concluding remarks

We have examined the stability of certain polling models using an approach involving associated fluid models that provides for simple intuitive proofs while extending the existing work in the area. It would be of interest to see whether these techniques may be extended to other models, in particular the multiple server system with limited service where there is a restriction on the number of servers that can be at any one queue.

Generalizing the conditions on server routing to allow state dependence is also of interest. Work in this area has been done by Foss and Last [7, 8]. It appears that the results there may be generalized slightly, as while they do not use a fluid model approach, their analysis is of a similar flavour and seems to yield conditions under which fluid limits would exist.

It also would be of interest to see whether these techniques provide any insight into approximations for performance analysis, however, this is at first glance not an easy task.

Acknowledgements. This work was supported by an ERCIM Postdoctoral Fellowship. The author would like to thank Onno Boxma for helpful suggestions upon reviewing an earlier draft.

References

1. E. Altman, P. Konstantopoulos, and Z. Liu. Stability, monotonicity and invariant quantities in general polling systems. *Queueing Systems*, 11:35–57, 1992.
2. E. Altman and F. Spieksma. Polling systems–moment stability of station times and central limit theorems. Technical Report R 92/9, University of Leiden, 1992.
3. A. A. Borovkov and R. Schassberger. Ergodicity of a polling network. *Stochastic Processes and their Applications*, 50:253–262, 1994.
4. H. Chen. Fluid approximations and stability of multiclass queueing networks I: Work conserving disciplines. *Adv. Appl. Probab.*, 5:637–665, 1995.
5. J. G. Dai and S. P. Meyn. Stability and convergence of moments for multiclass queueing networks via fluid models. *IEEE Trans. Automat. Control*, 40:1889–1904, 1995.
6. J.G. Dai. On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.*, 5:49–77, 1995.
7. S. Foss and G. Last. Stability of polling systems with exhaustive service policies and state dependent routing. Technical Report 94/19, TU Braunschweig, 1994. Also, to appear, *Ann. Appl. Prob.*
8. S. Foss and G. Last. On the stability of greedy polling systems with general service policies. Technical Report 95/6, TU Braunschweig, 1995. Also, submitted, *Stoch. Proc. Applns.*
9. C. Fricker and M.R. Jaïbi. Monotonicity and stability of periodic polling models. *Queueing Systems*, 15:211–238, 1994.
10. C. Fricker and M.R. Jaïbi. Stability of a polling model with a Markovian scheme. Technical Report RR2278, INRIA, 1994.
11. B. Gamse and G.F. Newell. An analysis of elevator operation in moderate height buildings - ii. multiple elevators. *Transp. Res.*, B 16:321–335, 1982.
12. L. Georgiadis and W. Szpankowski. Stability of token passing rings. *Queueing Systems*, 11:7–33, 1992.
13. L. Georgiadis, W. Szpankowski, and L. Tassiulas. Stability analysis of quota allocation access protocols in ring networks with spatial reuse. Technical report CSD-TR-94-047, Purdue University, 1994.
14. H. Levy and M. Sidi. Polling systems: Applications, modeling, and optimization. *IEEE Trans. Comm.*, 38:1750–1760, 1990.
15. S. P. Meyn. Transience of queueing networks via fluid limit models. In *Proceedings of the Third SIAM Conference on Control and its Applications*, St. Louis, MO, 1994.
16. S. P. Meyn and D. Down. Stability of generalized Jackson networks. *Ann. Appl. Probab.*, 4:124–148, 1994.
17. R.J.T. Morris and Y.T. Wang. Some results for multi-queue systems with multiple cyclic servers. In W. Bux and H. Rudin, editors, *Performance of Computer-Communication Systems*, pages 245–248. North-Holland, Amsterdam, 1984.
18. K. Sigman. The stability of open queueing networks. *Stoch. Proc. Applns.*, 35:11–25, 1990.
19. R.D. van der Mei and S.C. Borst. Analysis of multiple-server polling systems by means of the power-series algorithm. Technical report BS-R9410, CWI, 1994.