Stability of multi-server polling system with server limits

Nelson Antunes, Christine Fricker, James Roberts

STABILITY OF MULTI-SERVER POLLING SYSTEM WITH SERVER LIMITS

NELSON ANTUNES, CHRISTINE FRICKER, AND JAMES ROBERTS

Abstract. We consider a multi-server polling system with *server limits*, the number of servers that can attend a queue simultaneously is limited. Stability conditions are available when service policies are unlimited. The definition of stability conditions when both server limits and limited service policies apply remains an open problem. We postulate a conjecture for the stability condition in this case that is supported by our simulation results. The study of this particular variant of the multiserver polling system is motivated by the performance evaluation of next generation passive optical access networks.

1. Introduction

Renewed interest in multi-server polling systems is motivated by recent developments in access networks using passive optical network (PON) technology – so-called next generation PONs. A feature of these networks is that the number of servers that can attend any queue is limited, as is the amount of work accomplished on each server visit. These restrictions lead to a polling system whose analysis has so far proved intractable.

PONs are composed of Optical Network Units (ONUs) – which provide high-speed network access to users – connected by fibre links and a passive splitter to an Optical Line Terminal (OLT) [10]. The OLT coordinates both upstream and downstream transmissions of the ONUs. Next generation PONs employ wavelength division multiplexing to augment the capacity of the fibre. For upstream transmissions, depending on declared ONU queue contents, the OLT allocates time slots on specific wavelengths ensuring that no signals collide at the splitter. This dynamic bandwidth allocation procedure can be realized as a multiserver polling system. To bound cycle times, allocated time slots are limited in size leading to limited service policies. In addition, while each ONU can use any wavelength channel, it is typically equipped with only one or a small number of tunable transmitters. The relevant multiserver polling system is therefore both service and server limited.

Multi-server polling systems have received relatively little attention in the vast literature on polling systems. The stability condition of a multi-server polling system with a K-limited service policy (i.e., on each visit, the server takes K customers or empties the queue if number waiting is less than K)[1] was first derived by Dai [3] using fluid limit arguments. The result was generalized independently by Fricker and Jaibi [8], Kovalevskii [9] and Down [5] to Markovian routing of servers with general service policies. Delcoigne

Date: June 20, 2011.

[1] we use the usual polling system terminology (see for example [7]).

and Fayolle [4] proposed a mean field approximation for the queue length distribution of a symmetrical multiserver polling model where the service policy is 1-limited. Foss and Kovalevskii [6] considered the harder problem where servers are heterogeneous obtaining analytical results for a two-queue, two-server system with unlimited (exhaustive) service policy.

Borst and van der Mei introduced the notion of server limits in [11, 2]. Servers visit each queue in cyclic order, moving directly to the next queue and incurring a switch time if the maximum number of servers is already present. They derived approximate formulas for the mean waiting time.

Down studied the stability of a multi-server polling system with server limits and unlimited service policies [5]. However, for tractability reasons, it was necessary to assume a particular service policy: when any server visits a queue that already has a full complement of servers, one server is chosen at random and obliged to move to the next queue on its schedule. With the more natural policy considered here (i.e., no service hand over), simple stability conditions apply when there is unlimited gated or exhaustive service policies [5, 1]. In [1], Antunes *et al.* derived a mean field approximation for multiserver polling systems with server limits and limited service policies under the assumption that both the number of servers and the number of queues are large.

In this paper we introduce a general model of the polling system with server limits motivated by next generation PONs. After recalling easily provable and known stability results for these systems, we postulate stability conditions for the considered multi-server polling system with server limits and limited service policies in the form of a conjecture. We present simulation results that support the validity of this conjecture.

2. Model description

The system is composed of n queues, labeled from 1 to n, each of infinite capacity. Arrivals occur at queue i ($i = 1, \ldots, n$) with independent, identically distributed interarrival times. Let λ_i be the arrival rate at queue i. The service times of customers at queue i are also independent and identically distributed with first moment b_i. We denote the offered traffic at queue i by $\rho_i = \lambda_i b_i$. The total traffic intensity is $\rho = \sum_{i=1}^{n} \rho_i$.

The queues are attended by m servers, labeled from 1 to m, which visit the queues according to independent irreducible Markov chains on $\{1, \ldots, n\}$ with respective probability transition matrices (r_{ij}^k), $k = 1, \ldots, m$. We assume that the number of servers that can simultaneously attend queue i is limited to m_i. Under this restriction, a server arrival at queue i is called effective if there are less than m_i other servers already at the queue at this time. If an arrival at queue i is not effective, then the server moves immediately to the next queue determined by its Markov chain. If an arrival is effective at queue i, the server applies the service policy attached to the queue. After an effective visit to queue i, a server incurs in a switch-over time with first moment s_i before moving to the next queue. We assume $m \leq \sum_{i=1}^{n} m_i$, so that a server can always find a queue with less than the maximum number of servers in attendance. The arrival times, service times, and switch-over times are assumed to be mutually independent.

3. Stability Conditions

Let l_i be the expected number of customers served on an effective visit to queue i when there are infinitely many customers waiting at the queue. The service policy at queue i is said to be of limited type if $l_i < \infty$ and of unlimited type otherwise. Policies such as gated or exhaustive K-limited and Bernoulli are limited service policies [7].

Unlimited service policies. The stability condition in the case of server limits and unlimited service policies can be readily derived using the same fluid limit arguments as in [5, 1]. We omit the proof since it closely follows the proof of Proposition 3 in [1]. In this case, the polling system is stable if and only if

$$\rho_i < m_i, \quad \text{for} \quad i = 1, \ldots, n \qquad \text{and} \qquad \rho < m.$$

Limited service policies without server limits. First suppose that $m_i = m$, for $i = 1, \ldots, n$, so that all server visits are effective. In addition, assume that the service policy in each queue is of limited type and the servers visit the queues according to the same Markov chain (r_{ij}). The stability condition is then given by the general results for multi-server polling systems in [8] and [5]. The polling system is stable if and only if

$$\rho + \max_{1 \le i \le n} \frac{\lambda_i}{p_i l_i} S < m, \tag{1}$$

where p_i is the *long run proportion of visits* to queue i and $S = \sum_{i=1}^{n} p_i s_i$ is the expected per-visit switch-over time. In this case, (p_i) is given by the unique invariant measure of the Markov chain (r_{ij}).

When service policy is limited, some queues remain stable even when condition (1) is violated. Such queues are said to be locally stable. Suppose, without loss of generality, that the n queues are renumbered such that

$$\frac{\lambda_1}{p_1 l_1} \ge \frac{\lambda_2}{p_2 l_2} \ge \cdots \ge \frac{\lambda_n}{p_n l_n}. \tag{2}$$

The polling system of queues $i, \ldots, n$ obtained by saturating queues $1, \ldots, i-1$ is stable if and only if

$$\hat{\rho}_i + \frac{\lambda_i}{p_i l_i} S^i < m \tag{3}$$

where $\hat{\rho}_i = \sum_{j=i}^{n} \rho_j$ and $S^i = S + \sum_{j=1}^{i-1} l_j p_j b_j$. Alternatively, suppose that stability condition (1) is violated and we want to determinate which queues are stable. The subset of stable queues is given by

$$\mathcal{S} = \{j : \hat{\rho}_j + \frac{\lambda_j}{p_j l_j} S^j < m\}. \tag{4}$$

Note that the above results were derived assuming $r_{ij}^k = r_{ij}$ for each server k. It may be possible to relax this requirement as long as the Markov chain of each server has the same invariant measure (e.g., if each server has its own visit cycle).

Limited service policies with server limits. The question that arises is whether the stability condition (1) and local stability results (3)-(4) also hold under the general assumptions of Section 2 with $m_i < m$ for at least one queue i. The p_i now have to be interpreted as the ***long run proportion of effective visits*** to queues by the servers. These proportions are unknown in general and depend on the whole system. We have conducted a large number of simulations for different system configurations. Results suggest that stability conditions (1) and (3), where the p_i are derived by measurement, do indeed apply. On the strength of this empirical evidence, we postulate the following.

Conjecture 1. *Under the assumptions of Section 2, the polling model with server limits and limited service policies is stable if and only if (1) holds where p_i is the long run proportion of effective server visits to queue i. The (p_i) exist for all system loads. Moreover, if the queues are renumbered according to (2) then the local stability results (3)-(4) remain valid.*

Of course, it is not in general easy to derive the effective visit frequencies p_i. Even when this is possible (e.g., in the case of perfectly symmetrical systems where $p_i = 1/n$), the associated fluid limit model does not readily yield the stability conditions.

The conjecture, if true, would only constitute a partial solution to the issue of evaluating the performance of general multiserver polling systems with both server limits and limited service policies. Note finally, that the large system approximation derived in [1] does yield stability conditions that conform to the conjecture and the (p_i) can be derived explicitly.

4. Simulation

In this section we present simulation results to illustrate the conjecture in the case of a particular asymmetrical polling system with server limits and limited service policies. We consider a polling system with 4 queues and 3 servers. The interarrival times at queue i ($i = 1, \ldots, 4$) follow a gamma distribution function, $f_i(x) = x^{\gamma_i - 1} e^{-x/\alpha_i} / (\alpha_i{}^{\gamma_i} \Gamma(\gamma_i))$, $x, \gamma_i, \alpha_i > 0$, where Γ is the gamma function. We fix $\gamma_i = 2$ and vary α_i to cover different arrival rates $\lambda_i = 1/(2\alpha_i)$. Service times in queue i have the Weibull distribution with density $g_i(x) = \beta_i/\theta_i (x/\theta_i)^{\beta_i - 1} e^{-(x/\theta_i)^{\beta_i}}$, $x, \beta_i, \theta_i > 0$. We let $\beta_i = 1.5$ and $\theta_i = i$ (the mean service time at queue i is $b_i = i\Gamma(5/3)$). We consider asymmetrical traffic loads with $\rho_1 = 2\rho_2 = 3\rho_3 = 4\rho_4$.

Servers 1, 2 and 3 visit the queues in a cyclic order with transition probabilities $r^1_{12} = r^1_{23} = r^1_{34} = r^1_{41} = 1$, $r^2_{43} = r^2_{32} = r^2_{21} = r^2_{14} = 1$ and $r^3_{14} = r^3_{42} = r^3_{23} = r^3_{31} = 1$, respectively. The number of servers that can simultaneously attend queue i is $m_i = \max(4 - i, 1)$. The service policy at queue i is K_i-limited exhaustive, i.e., a server continues serving until K_i customers have been served or the queue is empty. We set $K_i = i$ so that l_i in the formulas is also equal to i. In each queue, customers are served in the order they arrive. The switch-over time after an effective visit to queue i is uniformly distributed on $(0, i/10)$.

Figure 1 shows estimates of the (long run) proportion of effective visits to each queue as a function of the load in queue 1. Queue 1 has the highest

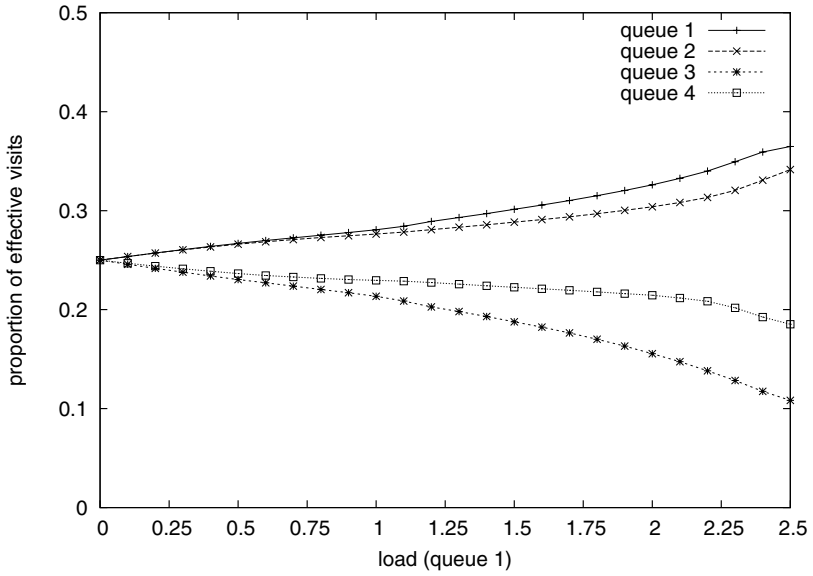

FIGURE 1. Proportion of effective visits to queues by servers.

proportion of effective visits since all servers can attend this queue at the same time. Note that, while queues 3 and 4 have the same server limit, queue 4 has a higher proportion of effective visits because of its lower load.

Figure 2 depicts on the left y-axis the left-hand side (lhs) of stability condition (1) using the estimated proportions of effective visits. In addition, we plot on the right y-axis the mean queue delay in queue 1. As the load increases, we observe that when the lhs of stability condition (1) reaches the value $m = 3$, the mean delay of queue 1 goes to infinity. The mean delay in the other queues remains stable in the considered load range.

We now consider the polling system of queues 2, 3 and 4 when queue 1 is saturated. First note that the initial ordering of the queues satisfies (2) for the proportions of effective visits given by Figure 1. Figure 3 plots on the left y-axis the lhs of local stability condition (3) with $i = 2$ and, on the right y-axis, the mean delay in queue 2. We observe that, when local stability condition (3) is violated, the mean delay in queue 2 goes to infinity. Queues 3 and 4 remain stable.

We have tested the conjecture in this way for various system configurations using other limited service policies. The same behavior was consistently observed.

REFERENCES

[1] N. Antunes, C. Fricker, P. Robert, and J. Roberts. Traffic capacity of large WDM passive optical networks. In *22th International Teletraffic Congress (ITC 22), September, Amestardan*, 2010.

[2] S. C. Borst and R. D. van der Mei. Waiting time approximations for multiple-server polling systems. *Performance Evaluation*, 31:163–182, 1998.

[3] J. G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability*, 5(1):49–77, 1995.

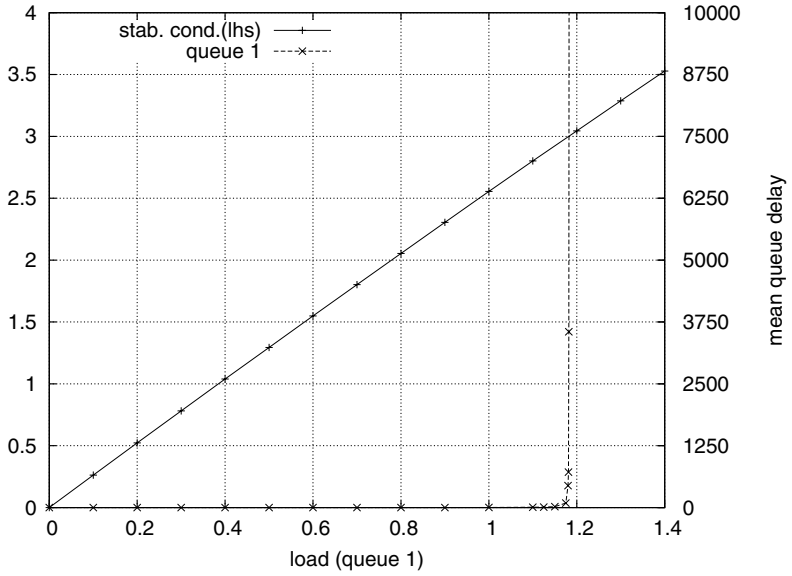

FIGURE 2. Left-hand side (lhs) of the stability condition (1) (left y-axis) and mean queue delay in queue 1 (right y-axis).

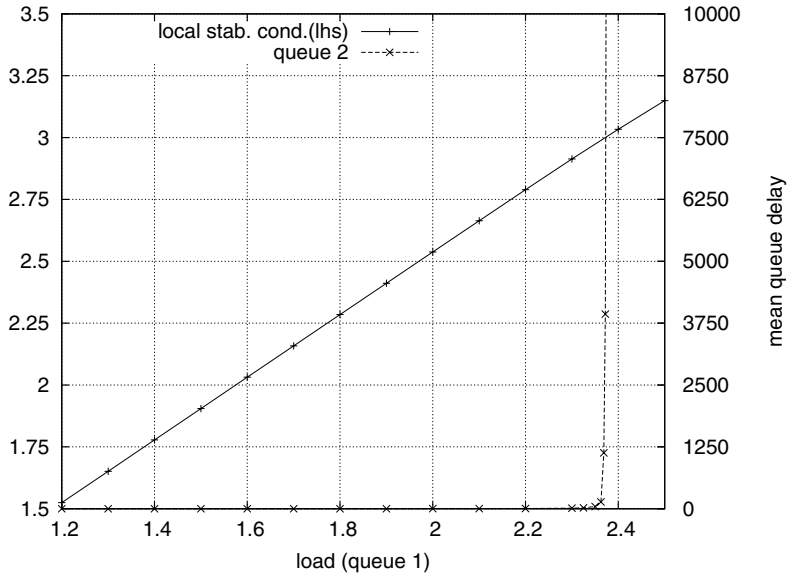

FIGURE 3. Left-hand side (lhs) of the local stability condition (3) (left y-axis) and mean queue delay in queue 2 (right y-axis).

[4] F. Delcoigne and G. Fayolle. Thermodynamical limit and propagation of chaos in polling systems. *Markov Proceses and Related Fields*, 5(1):89–124, 1999.
[5] D. Down. On the stability of polling models with multiple server. *Journal of Applied Probability*, 35(4):925–935, 1998.
[6] S. Foss and A. Kovalevskii. A stability criterion via fluid limits and its application to a polling system. *Queueing Systems*, 32:131–168, 1999.

[7] C. Fricker and M. R. Jaibi. Monotonicity and stability of periodic polling models. *Queueing Systems*, 15:211–238, 1994.
[8] C. Fricker and M. R. Jaibi. Stability of multi-server polling models. INRIA Research report, No. 3347, January 1998 (available at ftp://ftp.inria.fr/INRIA/publication/dienst/RR-3347.pdf).
[9] A. P. Kovalevskii. Positive recurrence and optimization of polling systems with several servers. *Aktual'nye Problemy Sovremennoi Matematiki*, 3:75–86, 1997.
[10] M. Maier. WDM passive optical networks and beyond: the road ahead [invited]. *IEEE/OSA Journal of Optical Communications and Networking*, 1(4):C1–C16, 2009.
[11] R. D. van der Mei and S. C. Borst. Analysis of multiple-server polling systems by means of the power-series algorithm. *Stochastic Models*, 13(2):339–369, 1997.

(N. Antunes) University of Algarve/CEMAT and INRIA Paris-Rocquencourt
Current address: Campus de Gambelas, 8005-139 Faro, Portugal.
E-mail address: nantunes@ualg.pt

(C. Fricker) INRIA Paris-Rocquencourt
Current address: Domaine de Voluceau, 78153 Le Chesnay, France.
E-mail address: Christine.Fricker@inria.fr

(J. Roberts) INRIA Paris-Rocquencourt
Current address: Domaine de Voluceau, 78153 Le Chesnay, France.
E-mail address: James.Roberts@inria.fr