An Algorithmic Solution of Polling Models with Limited Service Disciplines

J. P. C. Blanc

Abstract—**The power-series algorithm, an iterative numerical technique for the evaluation of the joint queue length distributions for a broad class of multiqueue systems, is extended to polling systems with limited service disciplines. Some examples are provided to validate the algorithm.**

I. Introduction

THE power-series algorithm (PSA) is a powerful means for evaluating performance measures of systems consisting of a moderate number of queues. This algorithm is based on power-series expansions of the state probabilities as functions of the load of a system in light traffic. The coefficients of these expansions are computed according to a recursive scheme. The ε-algorithm, cf. Wynn [8], is used to improve the convergence of the series in heavy traffic. An important area for application of the power-series algorithm consists of polling models for computer-communication systems, in which a token is passed for access to a single communication channel. In Blanc [1] the PSA has been discussed and applied to cyclic polling systems with Bernoulli schedules, which include exhaustive and 1-limited service. Here, the PSA will be generalized to polling systems with $(K-)$ limited service. There exist only a few analytical results for models with this service discipline. The most general result is the conservation law for systems without switching times, cf. Kleinrock [5]. Fuhrmann and Wang [4] derived approximations for the mean waiting times and an upper bound for the pseudoconservation law in models with limited service and nonnegligible switching times. Recently, Leung [6] developed an alternative algorithm for these models based on discrete Fourier transforms. We refer to Takagi [7] for a survey on polling systems.

The discussion in this letter will be restricted to models with zero switching times in order to be able to use the conservation law for checking the accuracy of the computations. In a future paper, it will be shown how models with switching times can be handled with the PSA. Further, we will assume Poisson arrival processes, exponential service times, and infinite buffers in order to simplify notations. It should be noted that the PSA can be used to study a much broader class of polling systems, e.g., with Coxian service times, with finite buffers and with general periodic, random or state dependent polling orders, see also Blanc [1]. The letter contains some numerical examples and discusses a heavy traffic property of the mean waiting times.

Paper approved by the Editor for Communication Networks of the IEEE Communications Society. Manuscript received May 10, 1990; revised March 28, 1991.

The author is with the Faculty of Economics, Tilburg University, 5000 LE Tilburg, The Netherlands.

IEEE Log Number 9201101.

II. The Polling Model

The system consists of s queues and a single server. Jobs arrive at queue j according to a Poisson process with rate λ_j, $j = 1, \cdots, s$, so that the total arrival rate to the system is $\Lambda := \sum_{j=1}^{s} \lambda_j$. Each queue may contain an unbounded number of jobs. Service times of jobs at queue j are assumed to be exponentially distributed with mean $1/\mu_j$, $j = 1, \cdots, s$. The server inspects the queues in a cyclic order $(1, 2, \cdots, s, 1, 2, \cdots)$. The number of jobs which are served during a visit of the server to queue j is at most K_j; the server proceeds to the next queue when either K_j jobs have been served or queue j has become empty, $j = 1, \cdots, s$. This discipline is called limited service. The times which are needed for switching from one queue to another will be neglected in the present letter. The load offered at queue j is $\rho_j := \lambda_j/\mu_j$, $j = 1, \cdots, s$, and the total offered load is $\rho := \sum_{j=1}^{s} \rho_j$. The condition for stability of the system is $\rho < 1$. It will be assumed that the system is in steady state. Finally, we introduce $a_j := \lambda_j/\rho$, $j = 1, \cdots, s$, because the PSA is based on power-series expansions of the state probabilities in terms of the load ρ.

III. Balance Equations for Limited Service Discipline

The PSA has been described in Blanc [1] for cyclic polling systems with Bernoulli schedules. Below we will only present the balance equations for the state probabilities of systems with limited service. The recursive scheme of the PSA can be derived from these equations in the same way as in Blanc [1]. Let N_j denote the number of jobs in queue j (waiting or being served), $j = 1, \cdots, s$, and let $\overline{N} := (N_1, \cdots, N_s)$. In order to transform the queue length process into a Markov process we introduce a polling table and a supplementary variable H, indicating the actual position on the table. The polling table is described as follows. Let $L := \sum_{j=1}^{s} K_j$ be the length of the table. The mapping $\ell(h)$ from table entry to queue number is defined by

$$\ell(h) = j, \quad \text{if} \sum_{i=1}^{j-1} K_i < h \leq \sum_{i=1}^{j} K_i, \quad \text{for } j = 1, \cdots, s, \quad h = 1, \cdots, L \tag{1}$$

and it is continued as a periodic function on $\mathbb{Z}$. The value of the variable H is increased by one whenever a service has been completed or when queue $\ell(H)$ is empty, unless the whole system has become empty; in the latter case the value of H is set and kept equal to 1 until a new arrival occurs. The value of $\ell(H)$ determines the queue to which the server is attending. Let $\overline{n} = (n_1, \cdots, n_s)$ be a vector with nonnegative integer entries. The state probabilities are defined as $p(\overline{n}, h) := \Pr\{\overline{N} = \overline{n}, H = h\}$, $\overline{n} \in \mathbb{N}^s$, $h = 1, \cdots, L$.

Let $I\{E\}$ stand for the indicator function of the event E, and let $\bar{e}_j$ be a vector with zero entries except an entry of one at the jth position ($j = 1, \cdots, s$). Noting that a state $(\bar{n},h)$ with $n_{\ell(h)} = 1$ can be entered through an arrival at queue $\ell(h)$ only if $\bar{n} = \bar{e}_{\ell(h)}$ and if h is the first entry on the polling table with the value $\ell(h)$, $h = 1, \cdots, L$, the balance equations for the state probabilities in models with limited service disciplines are readily verified to be, for $\bar{n} \in \mathbb{N}^s$, $h = 1, \cdots, L$,

$$\begin{aligned}\left[\rho\sum_{j=1}^{s} a_j + \mu_{\ell(h)}\right] p(\bar{n}, h) &= \rho\sum_{j=1}^{s} a_j p(\bar{n} - \bar{e}_j, h) \\ &\cdot I\{n_j > 0;\ n_j > 1 \text{ if } j = \ell(h)\} \\ &+ \sum_{i=1}^{L} \mu_{\ell(h-i)} p(\bar{n} + \bar{e}_{\ell(h-i)};\ h - i) I\{n_{\ell(\nu)} = 0, \\ &\nu = h - i + 1, \cdots, h - 1\} \\ &+ a_{\ell(h)} \rho p(\bar{0}, 1) I\{\bar{n} = \bar{e}_{\ell(h)} \wedge \ell(i) \neq \ell(h),\ i = 1, \cdots, \\ &h - 1\},\ n_{\ell(h)} > 0; \end{aligned} \tag{2}$$

$$\rho\sum_{j=1}^{s} a_j p(\bar{0}, 1) = \sum_{h=1}^{L} \mu_{\ell(h)} p(\bar{e}_{\ell(h)}, h). \tag{3}$$

It should be noted that the balance equations (2) and (3) are valid for models with arbitrary periodic polling orders and limited service, i.e., for arbitrary surjective mappings $\ell : \{1, \cdots, L\} \to \{1, \cdots, s\}$. The reader is referred to Blanc [1] for details concerning the derivation of a recursive scheme from the balance equations, the computation of moments of the joint queue length distribution, and the application of the ϵ-algorithm.

IV. Waiting Times

For polling models with Poisson arrival streams and service in order of arrival at each queue the following relations exist between the first two moments of the distributions of the waiting times W_j (excluding service) of jobs in queue j ($j = 1, \cdots, s$) and of the marginal queue-length distributions (Blanc [1]):

$$\begin{aligned} E\{N_j\} &= \lambda_j[E\{W_j\} + 1/\mu_j] \\ E\{N_j^2\} - E\{N_j\} &= \lambda_j^2[E\{W_j^2\} + 2E\{W_j\}/\mu_j + 2/\mu_j^2]. \end{aligned} \tag{4}$$

Hence, the moments of the waiting time distributions can be obtained as soon as those of the marginal queue length distributions have been computed with the aid of the PSA. Let W be the waiting time of an arbitrary job, not depending on the queue at which it arrives. The mean $E\{W\}$ and the standard deviation $\sigma\{W\}$ of the distribution of W are determined by:

$$\begin{aligned} E\{W\} &= \sum_{j=1}^{s} \frac{\lambda_j}{\Lambda} E\{W_j\}, \\ \sigma^2\{W\} &= \sum_{j=1}^{s} \frac{\lambda_j}{\Lambda} E\{W_j^2\} - E^2\{W\}. \end{aligned} \tag{5}$$

We recall that the mean waiting times at the various queues of a polling system with zero switching times satisfy a conservation law, cf. Kleinrock [5]:

$$\sum_{j=1}^{s} \frac{\rho_j}{\rho} E\{W_j\} = \frac{\rho}{1 - \rho} \sum_{j=1}^{s} a_j/\mu_j^2. \tag{6}$$

This section is concluded with a discussion of a conjecture on the heavy traffic behavior of the mean waiting times in cyclic polling systems with limited service. For models with a small number of queues it is possible to compute enough terms of the power-series expansions such that heavy traffic limits of performance measures can be accurately estimated. By doing so we found much convincing evidence for the following property (see also the examples with $s = 3$ and $\rho = 0.95$ in the next section).

The limit of $E\{W_j\}, j = 1, \cdots, s$, as $\rho \uparrow 1$, keeping the relative arrival rates a_j, $j = 1, \cdots, s$, and all other parameters fixed, is infinite if and only if

$$a_j/K_j = \max_{i=1,\cdots,s} \{a_i/K_i\}. \tag{7}$$

That only the arrival rates, and not the service rates, play a role in this heavy traffic property can be explained by the fact that a certain (integer) number of jobs is served during each cycle of the server along the queues according to the limited service discipline. Queues for which (7) holds have the largest probability of behaving temporarily during some cycles of the server as if they were overloaded, when there is much work in the system.

As a consequence of property (7) the ϵ-algorithm which is being used to accelerate the convergence of the power series should not be modified as described in Blanc [1] for moments of the marginal queue length distributions at queues where (7) does not hold. Another implication is that approximations for mean waiting times in polling systems, which do not possess property (7), will behave poorly under heavy traffic circumstances. For instance, the approximations for the mean waiting times in Boxma and Meister [3] have the right heavy traffic limits in case of exhaustive service [3, eq. (17)], but they do not have proper heavy traffic behavior in case of 1-limited service [3, eq. (20)].

V. Validation—Examples

In order to validate the PSA data generated with the aid of this algorithm are compared in Table I with the known right-hand side of the conservation law (6) and with simulation results. Table I concerns a model with one relatively heavily loaded queue. The parameters of the system are: $\mu_1 = 1, \mu_j = 2,\ j = 2, \cdots, s;\ a_1 = 2a_j = \frac{4}{s+3},\ j = 2, \cdots, s$. The examples concern only service disciplines with $K_j = K_2$, $j = 3, \cdots, s$. In the cases of 6 and 12 queues the values of $E\{W_3\}, \cdots, E\{W_{s-1}\}$ lie, in this order, in between those of $E\{W_2\}$ and $E\{W_s\}$, according to the PSA results (simulation results are often not accurate enough to confirm or to refute this statement).

TABLE I
VALIDATION OF THE POWER-SERIES ALGORITHM

s	ρ	K_1	K_2		$E\{W_1\}$	$E\{W_2\}$	$E\{W_s\}$	Cons.law[a]	$E\{W\}$	$\sigma\{W\}$	meth.[b]	M[c]	cpu[d]
3	0.95	1	1		22.39	2.70	2.75	15.83344	12.56	20.1	PSA	60	7
					24.92	2.72	2.77	17.53	13.83	23.5	SIM		23
				+/−	4.37	0.07	0.04	(95/6)					
3	0.95	2	1		15.33	16.83	16.86	15.83334	16.09	20.0	PSA	60	6
					15.08	16.82	16.67	15.63	15.91	19.8	SIM		17
				+/−	1.88	2.29	1.63	(95/6)					
3	0.95	8	1		2.90	41.67	41.74	15.83193	22.30	41.8	PSA	60	10
					2.93	42.03	42.47	16.04	22.59	41.0	SIM		17
				+/−	0.07	4.32	5.83	(95/6)					
3	0.95	8	8		21.09	5.25	5.45	15.84596	13.20	19.3	PSA	60	21
					21.08	5.24	5.44	15.83130	13.21	19.3	PSA	73	50
					19.61	5.27	5.45	14.86	12.49	17.1	SIM		16
				+/−	2.29	0.10	0.11	(95/6)					
6	0.75	1	1		2.955	1.507	1.573	2.168	1.942	3.10	PSA	20	14
					3.048	1.522	1.596	2.219	1.982	3.22	SIM		21
				+/−	0.108	0.042	0.035	(13/6)					
6	0.90	1	1		11.210	2.837	2.936	6.587	5.154	8.70	PSA	20	0*
					11.510	2.798	2.999	6.721	5.354	9.38	SIM		24
				+/−	0.838	0.067	0.091	(13/2)					
6	0.75	2	1		1.996	2.263	2.345	2.16662	2.215	3.47	PSA	20	15
					1.955	2.279	2.354	2.159	2.217	3.46	SIM		28
				+/−	0.078	0.081	0.083	(13/6)					
6	0.90	2	1		5.926	6.909	6.998	6.496	6.664	9.44	PSA	20	0*
					5.887	6.824	6.821	6.470	6.637	9.44	SIM		20
				+/−	0.483	0.419	0.551	(13/2)					
12	0.75	1	1		2.509	1.622	1.652	1.861	1.806	2.96	PSA	8	7
					2.622	1.600	1.653	1.886	1.772	2.82	SIM		29
				+/−	0.097	0.051	0.050	(19/10)					
12	0.75	2	1		1.701	1.927	2.038	1.903	1.931	3.09	PSA	8	7
					1.678	1.908	2.001	1.898	1.932	3.13	SIM		30
				+/−	0.046	0.072	0.062	(19/10)					
12	0.90	2	1		5.089	5.860	6.076	5.720	5.794	8.31	PSA	8	0*
					5.103	5.909	6.096	5.787	5.892	8.68	SIM		44
				+/−	0.302	0.280	0.280	(57/10)					

[a] The exact values of the right-hand side of the conservartion law (6) are given between brackets.
[b] In this column it is indicated whether the data have been generated with the algorithm (PSA) or with simulation (SIM). All simulations had a run length of 500 000 time units. Below the simulation results for the mean waiting times 95% confidence intervals are shown.
[c] M stands for the number of terms of the power-series expansions which have been computed. If this value has been underlined, then it is the maximal number of terms which can be computed with a storage capacity of about 14 MB.
[d] Approximate cpu-time in minutes on a VAX-8700; we have observed large variations in cpu-time for the same model, i.e., for the same amount of computations with the PSA as well as with simulation. For instance, in the system with $s = 6$, $K_1 = 2$, $K_2 = 1$, 20% more jobs were generated in the simulation run for $\rho = 0.90$ than in that for $\rho = 0.75$; still the cpu-time was longer in the case of $\rho = 0.75$. These counter-intuitive observations are probably due to the internal management and variations in the workload of the computer system. Some cpu-times have been indicated with 0* to stress that once the coefficients of the power-series expansions have been computed for some model, it requires usually less than 1 s of cpu-time to compute performance measures for various values of the load ρ.

Generally, results obtained with the PSA compare favorably with simulation results when the number of queues s and/or the size of the supplementary space L are moderate. Limitation of memory space is usually the main restriction on the applicability of the PSA. The accuracy of the results obtained with the PSA for a given model and a given number of terms (M) may vary strongly depending on the service discipline (see the examples with $s = 3$). Generally, the accuracy is highest with disciplines for which (7) holds for each queue, and it decreases when equality in (7) is approximated for some queue, or when the service limits K_j become large. In these cases the behavior of the systems in heavy traffic is quite different from that in light traffic, so that the coefficients of the power-series expansions at $\rho = 0$ do not contain enough information to compute performance measures accurately for ρ close to 1. For instance, in the example with $s = 3$, $K_1 = K_2 = 8$, it holds that $E\{W_1\} < E\{W_2\} < E\{W_3\}$ for ρ up to around 0.75 (corresponding to exhaustive service: $K_1 = K_2 = \infty$), while $E\{W_2\} < E\{W_3\} < E\{W_1\}$ for ρ larger than 0.80 [in agreement with (7)]. Because many parameters are involved in polling models and because of the above mentioned features, it is difficult to predict the accuracy of the PSA for a given model. The interested reader is referred to Blanc [2] for more data concerning cyclic polling systems with limited service, and for a comparison with systems with Bernoulli schedules.

REFERENCES

[1] J. P. C. Blanc, "A numerical approach to cyclic-service queueing models," *Queueing Syst.* vol. 6, pp. 173–188, 1990.
[2] ——, "Cyclic polling systems: limited service versus Bernoulli schedules," Rep. FEW 422, Dept. Economics, Tilburg University, 1990.

[3] O. J. Boxma, and B. W. Meister, "Waiting-time approximations in multi-queue systems with cyclic service," *Perform. Eval.* vol. 7, pp. 59–70, 1987.
[4] S. W. Fuhrmann, and Y. T. Wang, "Analysis of cyclic service systems with limited service: bounds and approximations," *Perform. Eval.*, vol. 9, pp. 35–54, 1988.
[5] L. Kleinrock, *Queueing Systems*, vol. 2. New York: Wiley, 1976.
[6] K. K. Leung, "Waiting time distributions for token-passing systems with limited-K service via discrete Fourier transforms," in *Performance '90*, P. J. B. King, I. Mitrani, and R. J. Pooley, Eds. Amsterdam: North-Holland, 1990, pp. 333–347.
[7] H. Takagi, "Queueing analysis of polling models: an update," in *Stochastic Analysis of Computer Communication Systems*, H. Takagi Ed. Amsterdam: North-Holland, 1990, pp. 267–318.
[8] P. Wynn, "On the convergence and stability of the epsilon algorithm," *SIAM J. Numer. Anal.* vol. 3, pp. 91–122, 1966.