15th IMACS World Congress EMS, Vol 4, Berlin, August 1997

Mean Waiting Time in the M/H$_2$/*s* Queue: Application to Mobile Communications Systems

Francisco Barceló, Josep Paradells, Mónica Aguilar
E.T.S. Ingenieros de Telecomunicación de Barcelona (U.P.C.)
c/ Jordi Girona 1-3, Mod. C3, Barcelona 08034
e-mail: barcelo@mat.upc.es

Keywords: Priority queue, M/G/*s* queue, Queueing theory.

ABSTRACT

In this paper a procedure to approximately calculate the mean waiting time in the M/H$_2$/*s* queue is presented. The approximation is heuristic although based in the intuitive symmetry between the deterministic and balanced hyperexponential-2 distributions. The three parameters which fully describe the H$_2$ distribution are considered, so the approximation can also be used for the M/G/*s* queue when the first three moments are known. If only the first two moments of the holding time distribution are known, the estimation can also be applied accepting a lesser accuracy. The estimation proposed is a closed formula extremely easy to compute and the results are very accurate. This features makes it helpful in the design of mobile telecommunication systems with more than one channel and queueing allowed (like trunking Private Mobile Radio PMR systems), where holding time distributions with coefficients of variation higher than one may appear.

As a second stage, the possibility of calls owning a certain level of priority is studied. Two service classes are considered according to a non-preemtive priority scheme (also known as Head Of the Line or HOL). This priority feature is often required in mobile telecommunications systems to improve the access delay of some special calls by degrading the delay suffered by the rest. If the proportion of calls owning priority is kept low, the degradation is shared by many calls and then kept small. In this paper a procedure to estimate the mean waiting time in queue for each priority class is presented. This procedure is also very easy to compute.

The environment for which the results of this paper are intended suggests medium or heavy overall load and light priority load (priority proportion is kept low). This is the situation under which the accuracy of the proposed method is checked. Although simulations are necessary in the final phase of the design, the procedure presented here is helpful as a first quick insight into the system performance.

15th IMACS World Congress EMS, Vol 4, Berlin, August 1997

Mean Waiting Time in the M/H$_2$/*s* Queue: Application to Mobile Communications Systems

Francisco Barceló, Josep Paradells, Mónica Aguilar
E.T.S. Ingenieros de Telecomunicación de Barcelona (U.P.C.)
c/ Jordi Girona 1-3, Mod. C3, Barcelona 08034
e-mail: barcelo@mat.upc.es

Keywords: Priority queue, M/G/*s* queue, Queueing theory.

ABSTRACT

In this paper a procedure to approximately calculate the mean waiting time in the M/H$_2$/*s* queue is presented. The approximation is heuristic although based in the intuitive symmetry between the deterministic and balanced hyperexponential-2 distributions. The three parameters which fully describe the H$_2$ distribution are considered, so the approximation can also be used for the M/G/*s* queue when the first three moments are known. This paper introduces a slight modification that improves the approximation given in [1]. Two service classes are considered according to a non-preemtive priority scheme (also known as Head Of the Line or HOL). This priority feature is often required in mobile telecommunications systems. The situation of medium or heavy load under which the accuracy of the proposed method is checked is the natural condition of evaluation for these systems: under light load the system must always work correctly. A way to estimate the mean waiting time or access delay for both type of calls is presented in this paper as an extension of [2, 3].

INTRODUCTION

The M/G/*s* queue is very useful to model mobile telecommunications systems because the general service time allows the designers to include the measured distribution of the channel holding time in the computation of the system Grade of Service (GoS). The distribution of the arrival process is in most cases more difficult to measure, and the Poisson arrival process is a reasonable assumption. As the exact solution for the M/G/*s* queue does not exist the designer must rely on computer simulations to estimate the system GoS. For the M/H$_2$/*s* queue an exact solution exists [4] which can be applied to estimate the delay in the M/G/*s* model when the squared coefficient of variation of the service time distribution is higher than 1; but the degree of complexity to compute this solution is extremely high and it could be quicker to simulate the system than to obtain the mentioned exact solution. This paper is concerned with systems in which the service time is distributed with a coefficient of variation higher than one, giving an approximate formula to estimate the mean waiting time in queue (or the mean queue length related to the waiting time by Little’s formula). The proposed approximation is extremely easy to compute and very precise. Although the results obtained by using the method presented must be completed with simulation results in the final phase of the system design, the approximation is very helpful as a first quick insight into the system performance.

In telecommunication systems, service time distributions with coefficient of variation higher than one are often found. Squared coefficients of variation as high as 4.8 are common in fix telephony [5]. When voice is multiplexed at talk-spurt level the squared coefficient of variation reaches 2.5 [6]. In Private Mobile Radio (PMR) systems which integrate dispatch calls (with a typical duration of 20 seconds) and interconnection to public telephone network (120 seconds), the coefficient of variation of the service time can be any, depending on the proportions of the call mixture. The same can be said about the combination or integration of voice (dispatch or interconnection) and data calls (much shorter) in the same PMR system. Some of these systems can be evaluated according to a ‘Blocked Calls Lost’ basis (no queue) and are obviously out of the scope of this paper, but others, primarily PMR systems, should be evaluated under the ‘Blocked Calls Delayed’ basis and the results for the M/G/*s* queue apply. If the mentioned systems are evaluated as a M/M/*s* queueing model which assumes a holding time distribution less disperse than the actual one, the GoS (mean access delay) is underestimated and the system tends to be undersized.

The priority feature is widely spread in PMR systems: it is found in trunking systems like MPT1327 and TETRA. We consider only the non-preemptive priority or Head Of the Line (HOL) case: if the priority call arrives when all the channels are busy, it does not interrupt a call in progress but it is placed in the queue before the non-priority calls. The emergency calls also found in the above mentioned systems are of the preemptive type: a priority call interrupts a non-priority call in progress when there is no channel available. Only priority (HOL) is considered in this paper, being emergency calls ignored.

NOTATION AND REVIEW OF PREVIOUS RESULTS

The following notation is used in the paper:

A	= offered traffic,
$b(t) = a\mu_1 e^{-\mu_1 t} + (1-a)\mu_2 e^{-\mu_2 t}$	= service hyperexponential-2 p.d.function.
$1/\mu = a/\mu_1 + (1-a)/\mu_2$	= mean holding time,
$\rho = A/s$	= offered load,
m_i	= i th ordinary moment ($m_1 = 1/\mu$ = mean),
c	= coefficient of variation,
$k = \frac{m_2}{2m_1^2} = \frac{c^2+1}{2}$	= relative 2nd moment,
$\overline{W}(M/G/s)$	= mean waiting time in the queue M/G/*s*,
$R_G = \frac{\overline{W}(\mathrm{M/G}/s)}{\overline{W}(\mathrm{M/M}/s)}$	= relative mean waiting time,
$r = \mu \frac{a}{\mu_1}$	= proportion of time serving short-type calls.

The three values chosen in this paper to describe the service time are m_1, c and r. In the case of H_2 service time they fully describe the holding time distribution. In the general case (M/G/*s*) the three first moments of the holding time distribution can easily be related to m_1, c and r. Note that the definition of r implies $\mu_1 \geq \mu_2$.

Due to the excessive complexity of the exact solution for the M/H_2/*s* queue, many approximations are described in the literature for the mean waiting time in the queue (see [8] for a very complete list of references). In the environment that we consider with medium or heavy load the approximation introduced by Boxma, Cohen and Huffles [7] takes into account the first three moments and gives excellent results while the coefficient of variation of the holding time keeps reasonably low (i.e. $c^2 < 3$). The drawback of this approximation is that the computation is cumbersome and must be programmed.

The approximation proposed by Kimura [8] is much simpler to compute but only takes into account the first two moments of the service time distribution: it does not depend on *r*. This fact makes it very precise and helpful when the service time is near to be balanced ($r \cong 0.5$) but the approximation is inaccurate for other values of *r*. This approximation is used below in this paper and states that the relative mean waiting time can be estimated as:

$$R_G = \frac{2kR_D}{2c^2 R_D + \left(1 - c^2\right)} \tag{1}$$

where R_D stands for the relative mean waiting time of the M/D/*s* queue.

PROPOSED APPROXIMATION FOR THE MEAN WAITING TIME

When *c* is high and assuming that the first three moments of the service distribution are known, the approximation given by Boxma is more precise than Eq. (1), but much harder to compute. For values of $c^2 > 3$ both approximations start to be inaccurate. In this case the approximation given in [1] provides an excellent trade-off between accuracy and simplicity as it is extremely easy to compute. Here a slight modification to the formula given in [1] is proposed which improves the accuracy of the approximation. The proposed estimated value of the relative mean waiting time in queue is:

$$R_H(r) = k\left(1 - \frac{2(k-0.5)(1-\rho)(s-1)\left(\sqrt{4+5s}-2\right)}{16\rho s + \left(k - 1/k\right)(1-\rho)(s-1)\left(\sqrt{4+5s}-2\right)}\, r\right) \tag{2}$$

In Table I some numerical results of $R_H(0.5)$ are shown for different values of load, coefficient of variation and number of channels; Kimura approximation of Eq. (1) is used for comparison. In Figure 1 the proposed formula is compared with simulation results and the approximation proposed by Boxma for two different values of $r = 0.25$ and $r = 0.75$.

Table I: Numerical results for the relative mean waiting time R_H in the M/H$_2$/s queue

	s	5			12			20		
	Load	Exact	Kim.	New	Exact	Kim.	New	Exact	Kim.	New
c^2=1.5	0.6	1.17	1.19	1.19						
	0.7	1.20	1.22	1.21	1.16	1.17	1.17	1.13	1.15	1.14
	0.8	1.22	1.23	1.23	1.19	1.20	1.20	1.17	1.19	1.19
	0.9	1.23	1.24	1.24	1.22	1.23	1.22	1.21	1.22	1.22
c^2=2.5	0.6	1.51	1.52	1.54						
	0.7	1.58	1.59	1.61	1.44	1.47	1.49	1.35	1.39	1.40
	0.8	1.64	1.65	1.67	1.55	1.56	1.59	1.48	1.50	1.53
	0.9	1.70	1.70	1.71	1.65	1.66	1.67	1.62	1.63	1.65
c^2=4	0.6	1.97	1.91	2.02						
	0.7	2.12	2.09	2.17	1.84	1.81	1.91	1.66	1.62	1.74
	0.8	2.26	2.23	2.30	2.06	2.02	2.11	1.92	1.89	1.99
	0.9	2.39	2.37	2.40	2.29	2.25	2.31	2.21	2.17	2.24

THE M/H$_2$/*s* QUEUE WITH PRIORITY

It is common in mobile telecommunication systems the use of priority schemes to improve the performance for some special calls. Obviously this improvement will cause a longer access delay to regular calls, but if the proportion of priority calls is kept small, the longer access delay is shared among many calls and the degradation is insignificant for practical purposes. Here an easy way to estimate the mean waiting time is presented as an extension of [2, 3] where the deterministic service time and coefficients of variation lower than one are considered.

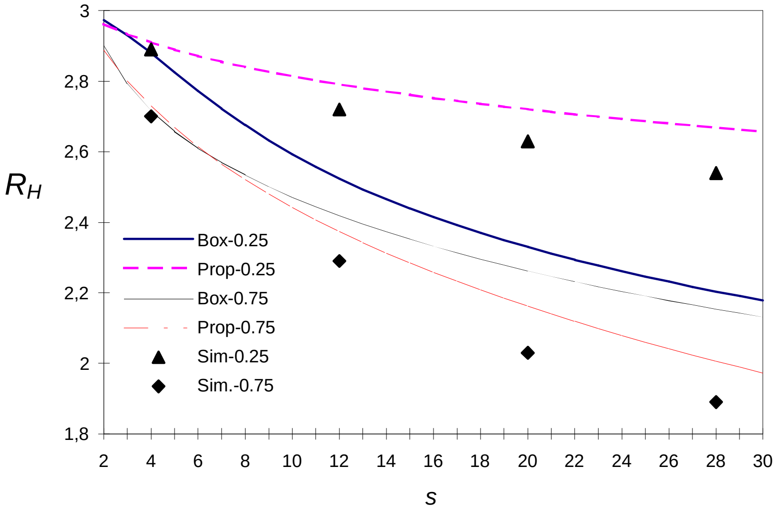

Figure 1: Comparative results for load = 85% and c^2=5. The proportions of time spent in short-type calls are r = 0.25 and r = 0.75.

To estimate the access delay for each type of calls (priority and non-priority) the following steps must be followed (index 1 stands for priority calls and 2 for non-priority):

Step1: The access delay for priority calls should be calculated according to Eq. (1). It is assumed that the priority proportion *p* is low so Eq. (2) is not valid as it is heuristically based in an approximation which is inconsistent for low load (see [1, 2, 3] for further details). The value of R_D in Eq. (1) must be calculated according to the linear approximation which is not so accurate as the below mentioned Eq. (5) but keeps its validity for low loads (is asymptotically exact when $\rho \to 0$):

$$R_{D1} = \frac{\overline{W}_1(M/D/s)}{\overline{W}_1(M/M/s)} = \frac{(1-p\rho)s}{s+1} + \frac{p\rho}{2} \tag{3}$$

The mean waiting time for the priority calls in the M/M/*s* equivalent queue can be computed according to the exact value given in [9]:

$$\overline{W}_1(M/M/s) = Erlang - c(A,s)\frac{1/\mu}{s(1-p\rho)} \tag{4}$$

Step 2: The access delay for all calls can be calculated according to Eq. (2) as the overall load is heavy or medium. R_D should now be calculated according to the following excellent approximation:

$$R_D = \frac{\overline{W}(M/D/s)}{\overline{W}(M/M/s)} = \frac{1}{2}\left\{1 + \frac{(1-\rho)(s-1)\left(\sqrt{4+5s}-2\right)}{16\rho s}\right\} \tag{5}$$

and $\overline{W}(M/M/s)$ can be calculated according to the Erlang-c formula.

Step 3: The mean access time for non-priority calls can be computed from the following average waiting time:

$$\overline{W}(M/H_2/s) = p\overline{W}_1(M/H_2/s) + (1-p)\overline{W}_2(M/H_2/s) \tag{6}$$

where the only unknown is the mean waiting time for non-priority calls.

NUMERICAL RESULTS

The proposed procedure can only be validated by numerical tests, comparing simulation results with the values obtained using steps 1 to 3 of the previous section; to this purpose extensive numerical tests have been performed under different conditions of load (overall and priority), number of channels, coefficient of variation and third moment (described through *r*). In Figure 2 a system consisting of 5 channels 90% loaded is assumed with priority proportion *p* varying between 0 and 0.5 and average call duration normalised to one time unit. Results of the mean waiting time obtained by simulation and by using the procedure proposed here to approximate the mean waiting time are plotted for both priority (Figure 1a) and non-priority calls (Figure 2b). The values calculated according to the Erlang-c formula for the M/M/*s* queue are also plotted for reference.

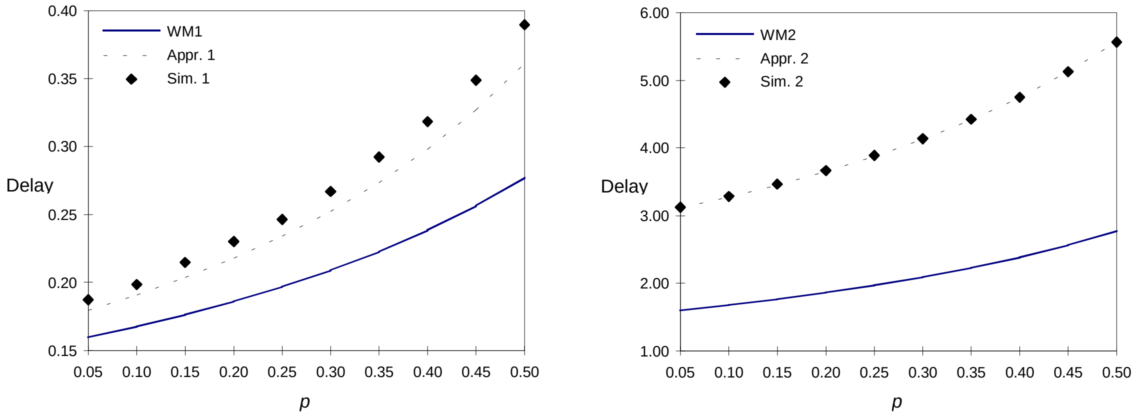

Figure 2: Mean access delay in a system with 5 channels 90% loaded for c^2=3
a) Priority calls ; b) Regular calls

In table II values of the mean waiting time for priority calls obtained by simulation are shown along with the values obtained by using the proposed approximation for different values of the squared coefficient of variation, load, priority proportion and number of available channels.

Table II: Mean access delay for priority calls; mean call duration normalised to 1,000 time units.

		C	5		12		20	
c^2	Load	p	Appr.	Sim.	Appr.	Sim.	Appr.	Sim.
1.5	0.90	0.1	176	179	60	61	31	31
	0.95	0.1	204	205	76	76	42	42
	0.90	0.2	198	201	67	67	35	35
	0.95	0.2	230	230	87	87	48	48
2.5	0.90	0.1	187	191	62	63	32	32
	0.95	0.1	217	221	79	80	44	44
	0.90	0.2	213	216	71	71	36	36
	0.95	0.2	249	258	90	90	50	50
4.0	0.90	0.1	196	214	64	64	32	33
	0.95	0.1	228	252	81	83	44	44
	0.90	0.2	226	253	73	74	37	38
	0.95	0.2	264	294	94	95	51	51

In Figure 3 the dependency of the mean waiting time for priority and non-priority calls with respect to the squared coefficient of variation of the holding time distribution is plotted. A system with 7 channels loaded 85% with 10% of priority calls is considered.

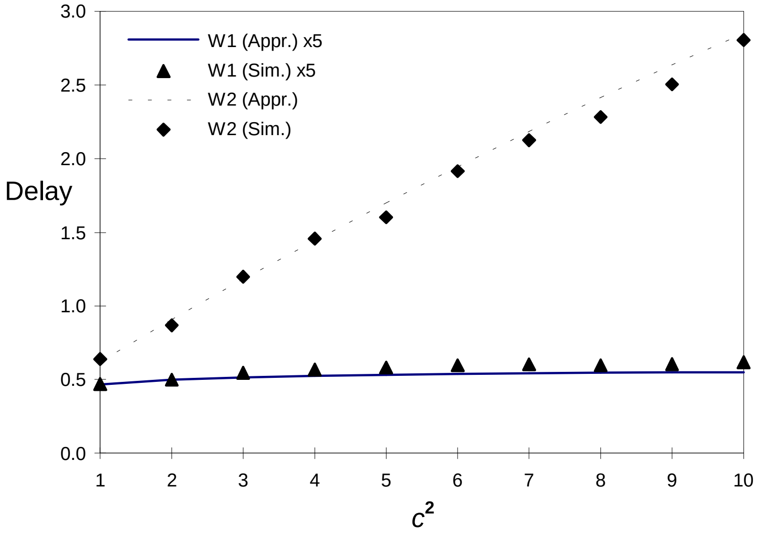

Figure 3: Mean access delay versus squared coefficient of variation

CONCLUSIONS

Due to the complexity of the exact solution for the waiting time distribution in the M/H_2/*s* queue, approximated results are helpful when they are accurate and easy to compute. In this paper a closed formula to estimate the mean waiting time in queue is given. The approximation is checked in the environment of mobile communications (medium or heavy load and multi-channel) proving to be precise. A procedure to estimate the access delay in the case that the system features some priority scheme is also presented. This latter estimation is also precise and simple to compute. All the results presented are useful for quick evaluation of communication systems with a model more realistic than the conventional M/M/*s*.

REFERENCES

[1] F. Barceló, J. Paradells, *The M/H2/s queue in mobile communications: Approximation of the mean waiting time,* 14th U.K. Teletraffic Symposium, IEE (1997).

[2] F. Barceló, V. Casares, J. Paradells, *The M/D/C Queue with Priority: Application to trunked Mobile Radio Systems* IEE Electronics Letters, Vol 32, No. 18 , p. 1644 (1996).

[3] F. Barceló, J. Paradells, *Performance Evaluation of Public Access Mobile Radio (PAMR) Systems with Priority Calls,* 5th Int. Conf. on Telecommunication Systems, p. 360 (1997).

[4] J. H. A. De Smit, *A Numerical Solution for the Multi-Server Queue with Hyper-Exponential Service Times,* Operations Research Letters, Vol. 2, No. 5, p. 217 (1983).

[5] V. Bolotin, *Telephone Circuit Holding Time Distributions,* Proc. 14th Inernational Teletraffic Congress, p. 125 (1994).

[6] H. H. Lee, C.K. Un, *A study of On-Off Characteristics of Conversational Speech*, IEEE Trans. on Communications, COM-34, num. 6, p. 630 (1986).

[7] O.J. Boxma, J.W. Cohen, N. Huffles, *Approximations of the Mean Waiting Time in an M/G/s Queueing System*, Operations Research, Vol. 27, p. 1115 (1979).

[8] T. Kimura, *Approximations for multi-server queues: system interpolation,* Queuing Systems 17, p. 347 (1994).

[9] D. Gross, C. M. Harris, Fundamentals of Queueing Theory, John Wiley & Sons (1974).