# An engineering approximation for the mean waiting time in the $M/H_2^b/s$ queue

Francisco Barceló
Universidad Politécnica de Catalunya
c/ Jordi Girona, 1-3, Barcelona 08034
Email : barcelo@entel.upc.es

## Abstract

Although the exact solution for the mean waiting time in the $M/H_2/s$ queue exists, it requires advanced programming skills and processing capacity. In most engineering problems the offered traffic and its features are known with a certain degree of uncertainty (e.g. traffic is predicted, traffic is not clearly Poisson, etc). In such environments where perfect accuracy is not required, approximations are often used to estimate the performance. For engineering applications, the easiness to compute the estimate is a key factor. The contribution of this paper is a new closed formula to estimate the mean waiting time in the $M/H_2^b/s$ queue. The formula is easy to compute and accurate. In contrast with the already existing approximations, the one presented here keeps its validity for medium-high coefficients of variation.

**Keywords**: Queues, approximate analysis, mean waiting time, multiserver queues.

# 1 Introduction

## *The need for approximations*

In real world evaluation of telecommunications systems, the input data (e.g. offered traffic, mean call duration, etc.) are known within a certain error range, being this range wider for predicted than for measured data. Statistical properties of the arrival and service processes are seldom well known: sometimes they are inferred from traffic theory, others they are measured in the field. In this environment and for practical purposes, the Poisson arrival process is often used when the arrival process is not well known. A sample of this practice is the extensive use of the Poisson process to model the handoff traffic in cellular networks: handoff is clearly carried traffic (i.e. carried in the surrounding cells). The features of the service process are easier to be measured than those of the offered traffic. However, the features of a telecommunications service seldom fit those of a practical probability distribution. With the current trend to integrate several services in the same network, the service time characterization becomes more difficult and perfect matching between measured data and a known distribution (or even with a mix of distributions) is impossible in practice.

In these environments, an exact solution [1] is not necessary, in particular if the cost of this "total accuracy" is high. Approximations can greatly reduce this cost (e.g. needed programming skills and/or time, CPU and memory needs, computing time, etc) while keeping the accuracy of the obtained results within a very narrow error range. In the scenario described above, one could state that a 5% relative error is excellent. Notice that measured traffic data use to have 20-40% relative error while predicted data with less than 50% relative error are difficult to obtain. The fact that most of the telecommunications services are newly launched or upcoming services obviously contributes to these wide error ranges.

## *The $H_2$ holding time*

Multiserver queues are widely used in telecommunications networks, in particular when those networks are circuit-switched. The Erlang-C solution for the M/M/s queue is probably the most widely used traffic formula in telecommunications systems engineering. However, the statistical properties of new telecommunications services are changing and we currently witness the evolution from services with exponential holding time distribution to holding times with squared coefficient of variation (SCV) higher or much higher than unity. Some samples follow:

- Bolotin showed how the current telephone service presents a lognormal holding time distribution with a SCV between 2.8 and 4.6 [2].
- The connection time to a Base Station (BS) in cellular telephony is only a fraction of the overall call holding time and its SCV is between 1.7 and 4 [3, 4].

- Voice at talk spurt level (e.g. to be used in speech interpolation systems) can be modelled as a $H_2$ distribution according to the ITU-T [5, 6] with a SCV of 2.5.
- The transmission time in Private Mobile Radio systems shows SCV around 2.8 [7].
- The size of Internet web pages or electronic mails has been modelled with SCV higher or much higher than unity.
- There is a trend to integrate several services in the same network, which usually leads to holding times with medium or high (sometimes very high) SCV. The fact that many telephone calls finish in the voice-mail causes the holding time to be a mix of two distributions: regular calls with an average duration of some hundreds of seconds and voice-mail calls with a very short average duration of a few seconds. Dial-up connections to the Internet mixed with regular calls also cause high SCV in the access network. As a rough estimate, if 20% of calls are dial-up internet with average duration 30 minutes and 80% are regular calls with average 140 seconds, and assuming that each service is exponentially distributed, we have an overall service time with an $H_2$ distribution and SCV=4.95. Service mixes can be found in most current access networks and integration of services is expected to increase in the near future with upcoming technologies such as UMTS, MPLS, etc.

Although most of the above-mentioned holding times are not $H_2^b$ distributed, they all show a SCV greater than unity. The $H_2^b$ distribution presents clear advantages to model the holding time in those scenarios. In most cases, the designer knows only the first one or two moments of the holding time distribution. The $H_2^b$ distribution is fully described by two parameters (e.g. the first two moments, see Section 2 for notation and difference between $H_2$ and $H_2^b$), it is simple to work with and very intuitive. The SCV of the $H_2^b$ is always greater than unity. This paper proposes to use a two-moment approximation in those cases in which the designer is aware of the fact that the service time presents a SCV greater than unity.

### *The $M/H_2^b/s$ queue*

In the first section of this introduction, we showed how the Markov arrival process suits a wide variety of telecommunications services, even in cases where the Poisson nature of the arrival process has not been verified (e.g. handoff arrivals to a base station). In the second section, the $H_2^b$ distribution has been proposed for a two-moment matching of holding time distributions when they are not accurately known. In addition, the queue investigated here is a Blocked Calls Delayed (BCD) one (i.e. blocked calls are placed in a queue with infinite capacity).

Most telecommunications services can be modelled as BCD. An infinite queue is a very good approximation for systems in which the buffer capacity is high enough to keep loses low, even for medium or high blocking probability such as buffers in switches, queues in call centers, access to Intelligent Network (IN) services, access to database, signalling links, etc. BCD is also a good approximation for pure loss systems in which the customer retries after a short time when blocked: local telephony exchanges and PBX, cellular base stations, etc.

# 2 Review of Existing Approximations

Here some procedures to estimate the mean waiting time in the *M/G/s* queue are reviewed and particularized for the $H_2^b$ case (see Tijms [8] and Kimura [9] for details on approximations for the *M/G/s* queue). The balanced hyperexponential distribution ($H_2^b$) is a particular case of the $H_2$ distribution where the server spends 50% of the service time in each type of call. In the remaining of this paper the following notation is used:

$A$ = offered traffic,

$b(t) = p\mu_1 e^{-\mu_1 t} + (1-p)\mu_2 e^{-\mu_2 t}$ = service time of the $H_2$ p.d.f. It is fully described by three parameters. Only two parameters are needed for the $H_2^b$ since $p/\mu_1 = (1-p)/\mu_2$ .

$1/\mu = p/\mu_1 + (1-p)/\mu_2$ = mean holding time,

$\rho = A/s$ = offered load,

$m_i$= $i$-th ordinary moment ($m_1 = 1/\mu$ = mean),

$c$= coefficient of variation,

$k = \dfrac{m2}{2m_1^{\,2}} = \dfrac{c^2+1}{2}$ = relative 2nd moment,

$W$(M/G/$s$)= mean waiting time in the queue *M/G/s*,

$R_G = \dfrac{W(\mathrm{M/G}/s)}{W(\mathrm{M/M}/s)}$ = relative mean waiting time,

$r = \mu \dfrac{p}{\mu_1}$ time proportion spent in short-type calls ( $\mu_1 \geq \mu_2$ ). (Notice that $r$=0.5 for the $H_2^b$).

The relative mean waiting time ($R_G$) is the ratio of the waiting time in the considered queue with respect to the mean waiting time in a *M/M/s* queue with the same load. The $H_2$ distribution is completely described by three parameters, here $m_1$, $c$ and $r$.

In [10], Boxma, Cohen and Huffles give an approximation to the mean waiting time in the M/G/$s$ queue that can be applied to the $H_2^b$ holding time as follows:

$$W(\mathrm{M/H2}/s) = \frac{2kW(\mathrm{M/D}/s)W(\mathrm{M/M}/s)}{2aW(\mathrm{M/D}/s) + (1-a)W(\mathrm{M/M}/s)}, \tag{1}$$

with

$$a = \frac{1}{s-1}\left(\frac{m2}{\gamma_1 m_1} - s - 1\right) \qquad \gamma_1 = \sum_{i=0}^{s} \frac{\binom{s}{i} r^i (1-r)^{s-i}}{i\mu 1 + (s-i)\mu 2}. \tag{2}$$

Eq. (1) can be rewritten in a more compact form as:

$$R_{H2}(r) = \frac{2kR_D}{2aR_D + (1-a)}, \tag{3}$$

where $R_D$ is the relative mean waiting time for the *M/D/s* queue. The dependency of the relative mean waiting time with $r$ is explicit in the notation because not all the approximations include the influence of the third moment. Boxma's approximation considers the three parameters of the $H_2$ distribution and is very accurate when the SCV is not large. The computing complexity of the approximation increases with the number of channels $s$ according to Equation (2).

The linear (with respect to $\rho$ and $k$) approximation [8] is $R_{H2}(r) = (1-\rho)\gamma 1 s\mu + \rho k$ . This approximation is not so accurate as Equation (3) while the computing complexity is almost the same.

The approximation presented by Kimura [9] considers only two moments of the distribution: there is no dependence on $r$. In the notation $R_G$, a G is used to reflect that given the first two moments the formula intends to be valid for any holding time distribution. The approximation is:

$$R_G = \frac{2kR_D}{2c^2 R_D + (1-c^2)}. \tag{4}$$

The simplest approximation was given by Cosmetatos in [11] as $R_G = (1-c^2)R_D + c^2$ . This linear interpolation is exact for $c$=0 and $c$=1 and very inaccurate when $c$>2.

The approximation given by Ma and Mark [12] is very precise for very high values of $s$ and/or *SCV*, but these values are seldom found in actual wireless access systems. Ma and Mark suggest values of $s$ around 200 and *SCV* higher than 20. On the other hand, the algorithm relies on exact figures for a lower number of servers and SCV. The approximation given in [12] is aimed to reduce the huge processing requirements of the exact solution for high $s$ and/or SCV, but the authors assume that the initial solution for a lower s or SCV is an exact solution. This approach is far from being simple to compute.

One problem shared by some of the above-mentioned approximations is the need of $R_D$ in order to compute $R_G$. The need for an exact value, difficult to achieve, to compute an approximation does not seem very practical. Although tables with exact values of $W$(M/D/$s$) are available [13, 14] and the exact solution can be programmed, the following excellent approximation developed by Cosmetatos is helpful:

$$R_D = \frac{1}{2}\left\{1 + (1-\rho)(s-1)\frac{\sqrt{4+5s}-2}{16\rho s}\right\}. \quad (5)$$

In the remaining of this paper the approximation of Equation (5) is always used instead of the exact value of $R_D$. This substitution brings this work nearer to engineering applications, although it slightly reduces the accuracy of the approximations.

# 3 Proposed approximation

The proposed approximation is heuristic and based on numerical tests, thus a mathematical demonstration is not provided. The heuristics that led to this approximation are based on two concepts: a limiting bound for the *M/H₂/s* queue and an intuitive symmetry concept between the $H_2^b$ and the deterministic holding time with respect to the exponential distribution.

If the first two moments of the $H_2$ distribution are given, a limiting case occurs when the shorter calls are extremely short ( $\mu_1 \to \infty$ ). In this case $r$=0, and to maintain the two moments of the distribution:

$$\frac{1}{\mu_2} = km_1, \qquad 1 - p = \frac{1}{k}. \quad (6)$$

The system does not see the shorter calls: only sees calls with a mean duration of $k$ times the mean of the $H_2$ distribution with probability 1/$k$. This is in fact a *M/M/s* process with the same offered traffic $A$ as the *M/H₂/s* and with a mean duration $m'_1 = km_1$. For this extreme case the upper bound when $r$=0 is:

$$W(\mathrm{M/H_2}/s) = C(\rho, s)\frac{km_1}{s-A} = kW(\mathrm{M/M}/s), \qquad R_{H2}(0) = k, \quad (7)$$

where $C(\rho, s)$ represents the Erlang's delay formula.

The symmetry between the deterministic and the balanced $H_2^b$ distribution ($r$=0.5), applied to the approximation for the mean waiting time, leads to the following proposed formula:

$$R_{H2}(0.5) = \frac{k}{1 + \frac{k(1-\rho)(s-1)\left(\sqrt{4+5s}-2\right)}{16\rho s}} \quad (8)$$

Note how the symmetry concept leads to the use of the Cosmetatos approximation for the M/D/s queue in a "reverse sense": note the similarities between Equations (8) and (5). There is only one deterministic but infinite $H_2^b$ distributions with the same mean (depending on the SCV), thus the normalization factor $k$ is needed.

Unlike the others, the new approximation is not asymptotically exact when the SCV tends to one, giving inconsistent results: the approximated waiting time is higher for the $H_2$ than for the exponential. Note that Eq. (5) on which our proposal is based also gives inconsistencies for very light loads. Hence, the proposed approximation should not be used with low SCV (i.e. near to unity) or light loads. If the SCV is close to one any other of the approximations mentioned in Section 2 can be used; we suggest Kimura's for its accuracy and simplicity. Light load is not a frequent scenario of evaluation.

# 4 Numerical results

In Section 2, five approximations for the *M*/H₂/*s* queue have been reviewed. The linear and Cosmetato's approximations give less accurate results than Equations (3) and (4). In addition, the accuracy of those approximations is poor for SCV>2. Kimura's approximation needs two parameters (e.g. two first moments) and is precise and simple to compute. Boxma's approximation introduces a third parameter ($r$) at the cost increasing the required computing effort. For practical purposes, this is interesting when the designer knows more than two moments of the holding time distribution. When only two moments are known, the extra complexity is not compensated by extra accuracy in front of the Kimura and new approximation (as shown below).

In Table I, exact values of the balanced $R_{H2}$ [13, 14] are compared with Equations (3), (4) and (8) (i.e. Boxma's, Kimura's and proposed approximations respectively). Light and medium loads are considered and the most accurate result is highlighted in each case. Simplicity has been used as a second criterion when two or more results have the

same accuracy. In all cases the mean holding time is one time unit, hence the mean waiting time is normalized to the mean holding time. Waiting times is lower than 0.01 have not been displayed since they are not interesting for practical engineering problems. Notice how Kimura`s approximation gives the best result for low SCV while the new approximation is better for medium and high SCV. Kimura and Boxma results are very similar when the blocking probability is medium-low (i.e. light-medium load and/or many servers). Since Boxma includes a third parameter (*r*) and Kimura does not (Kimura presents another excellent approximation for this last case in [9]), one should use Boxma's when the first three moments of the holding time distribution are known (i.e. for improved accuracy) and Kimura's or the new approximation when only two parameters are known (i.e. simpler to compute). Table II presents similar results for heavy load. Notice how the most accurate result changes from Kimura's to new approximation as the SCV increases.

**Table I. Comparative results for light and medium load.**

| | | $\rho$=0.5 | | | | $\rho$=0.7 | | | | $\rho$=0.8 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *scv* | s | *Ex* | Box | Kim | New | *Ex* | Box | Kim | New | *Ex* | Box | Kim | New |
| | 2 | *0.412* | **0.417** | 0.415 | 0.399 | *1.209* | 1.217 | 1.215 | **1.195** | *2.254* | 2.263 | **2.261** | 2.239 |
| | 4 | *0.103* | 0.105 | 0.095 | **0.104** | *0.440* | 0.444 | **0.444** | 0.426 | *0.932* | 0.939 | **0.938** | 0.915 |
| 1.563 | 8 | *0.017* | 0.017 | **0.017** | 0.014 | *0.135* | 0.137 | **0.137** | 0.127 | *0.351* | 0.355 | **0.355** | 0.339 |
| | 15 | | | | | *0.036* | 0.037 | **0.037** | 0.033 | *0.128* | 0.130 | **0.130** | 0.121 |
| | 25 | | | | | | | | | *0.049* | 0.050 | **0.050** | 0.045 |
| | 2 | *0.738* | 0.766 | 0.722 | **0.734** | *2.261* | 2.312 | 2.248 | **2.270** | *4.285* | 4.345 | 4.272 | **4.298** |
| | 4 | *0.167* | **0.166** | 0.160 | 0.162 | *0.781* | 0.783 | 0.767 | **0.780** | *1.718* | 1.719 | 1.697 | **1.718** |
| 4 | 8 | *0.024* | **0.023** | 0.023 | 0.023 | *0.223* | 0.219 | 0.218 | **0.222** | *0.620* | 0.609 | 0.607 | **0.617** |
| | 15 | | | | | *0.055* | 0.055 | 0.054 | **0.055** | *0.213* | 0.209 | 0.209 | **0.213** |
| | 25 | | | | | *0.013* | 0.013 | **0.013** | 0.013 | *0.077* | 0.076 | 0.076 | **0.077** |
| | 2 | *1.381* | 1.465 | 1.180 | **1.310** | *4.365* | 4.529 | 4.063 | **4.302** | *8.387* | 8.582 | 8.027 | **8.323** |
| | 4 | *0.285* | **0.262** | 0.259 | 0.222 | *1.447* | 1.369 | 1.240 | **1.383** | *3.283* | 3.154 | 2.952 | **3.186** |
| 9 | 8 | *0.036* | 0.029 | **0.033** | 0.028 | *0.388* | 0.324 | 0.316 | **0.365** | *1.142* | 0.986 | 0.969 | **1.086** |
| | 15 | | | | | *0.088* | 0.073 | 0.072 | **0.084** | *0.374* | 0.309 | 0.308 | **0.354** |
| | 25 | | | | | *0.019* | 0.017 | 0.017 | **0.019** | *0.128* | 0.106 | 0.105 | **0.123** |

# 5 Summary and Conclusion

The *M/G/s* queue with SCV higher than one is often found in performance evaluation of telecommunications systems. In the design phase, most of the times the probability distribution of the channel holding time is not well known (e.g. only two or three moments are known). In such cases, the $M/H_2^b/s$ queue is a convenient model to obtain estimates of the mean waiting time.

After the comparative results presented in this paper, it is recommended that when the SCV of the channel holding time is lower than 2, Kimura's approximation should be used if only the first two moments of the holding time distribution are known. When the first three moments are known, Boxma's approximation adds extra computing complexity and accuracy. For SCV greater than 2, the above-mentioned approximations loose accuracy and the approximation presented in this paper is recommended for simplicity and accuracy.

# Acknowledgement

This research was funded by the Spanish Government through project CICYT TIC2000-1041-C03-01.

**Table II. Comparative results for heavy load.**

| | | $\rho$=0.9 | | | | $\rho$=0.95 | | | | $\rho$=0.99 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *scv* | *s* | *Ex* | Box | Kim | New | *Ex* | Box | Kim | New | *Ex* | Box | Kim | New |
| | 2 | *5.436* | 5.446 | **5.444** | 5.420 | *11.83* | 11.84 | **11.84** | 11.82 | *63.08* | 63.09 | **63.08** | 63.06 |
| | 4 | *2.495* | 2.503 | **2.502** | 2.475 | *5.680* | 5.689 | **5.688** | 5.659 | *31.29* | 31.30 | **31.30** | 31.27 |
| 1.563 | 8 | *1.102* | 1.107 | **1.107** | 1.084 | *2.679* | 2.684 | **2.684** | 2.659 | *15.47* | 15.48 | **15.48** | 15.45 |
| | 15 | *0.499* | 0.503 | **0.503** | 0.487 | *1.325* | 1.329 | **1.329** | 1.309 | *8.136* | 8.140 | **8.140** | 8.116 |
| | 25 | *0.250* | 0.252 | **0.252** | 0.241 | *0.732* | 0.735 | **0.735** | 0.719 | *4.806* | 4.810 | **4.810** | 4.790 |
| | 2 | *10.48* | 10.55 | 10.47 | **10.50** | *22.96* | 23.03 | 22.94 | **22.98** | *122.9* | 123.0 | 122.9 | **123.0** |
| | 4 | *4.745* | 4.743 | 4.713 | **4.744** | *10.95* | 10.94 | 10.91 | **10.95** | *60.91* | 60.91 | 60.87 | **60.91** |
| 4 | 8 | *2.054* | 2.030 | 2.026 | **2.048** | *5.116* | 5.081 | 5.076 | **5.106** | *30.07* | 30.02 | 30.01 | **30.05** |
| | 15 | *0.909* | 0.891 | 0.890 | **0.904** | *2.501* | 2.471 | 2.470 | **2.492** | *15.78* | 15.73 | 15.73 | **15.76** |
| | 25 | *0.442* | 0.432 | 0.432 | **0.439** | *1.365* | 1.342 | 1.341 | **1.357** | *9.300* | 9.259 | 9.259 | **9.284** |
| | 2 | *20.76* | 20.98 | 20.34 | **20.69** | *45.71* | 45.93 | 45.25 | **45.63** | *245.7* | 245.9 | 245.2 | **245.6** |
| | 4 | *9.298* | 9.098 | 8.800 | **9.156** | *21.69* | 21.44 | 21.09 | **21.52** | *121.6* | 121.3 | 120.9 | **121.4** |
| 9 | 8 | *3.965* | 3.632 | 3.597 | **3.844** | *10.06* | 9.597 | 9.547 | **9.893** | *59.94* | 59.34 | 59.28 | **59.72** |
| | 15 | *1.719* | 1.501 | 1.497 | **1.642** | *4.879* | 4.507 | 4.501 | **4.745** | *31.40* | 30.85 | 30.84 | **31.20** |
| | 25 | *0.817* | 0.692 | 0.691 | **0.774** | *2.635* | 2.371 | 2.369 | **2.538** | *18.48* | 18.02 | 18.02 | **18.31** |

# References


[1] J.H.A. de Smit, “A numerical solution for the multi-server queue with hyper-exponential service times”, *Operation Research Letters*, Vol. 2, pp. 217-224, Dec. 1983.

[2] V. Bolotin, "Telephone Circuit Holding Time Distributions", *Proc. 14th International Teletraffic Congress* pp. 125-134, North Holland, 1994.

[3] F. Barceló, J. Jordán, “Channel Holding Time Distribution in Public Telephony Systems (PAMR and PCS)” *IEEE Trans. on Vehicular Technology*, Vol 49, no 5, pp. 1615-1625, Sept. 2000.

[4] C.Jedrzycki, V.C.M.Leung, “Probability Distribution of Channel Holding Time in Cellular Telephony Systems”, *IEEE Vehicular Technology Conference* (*VTC’96)*, pp. 247-251, 1996.

[5] H.H. Lee, C.K. Un, "A study of On-Off Characteristics of Conversational Speech", *IEEE Trans. on Communications*, COM-34, num. 6, pp. 630-637, June 1986.

[6] ITU-T Rec. P.84: Subjective listening test method for evaluating digital circuit multiplication and packetized voice systems.

[7] P.Cohen, H.H. Hoang, D. Haccoun, "Traffic Characterization and Classification of Land Mobile Communications Channels", *IEEE Trans. on Vehicular Technology*, VT-33, no. 4, pp. 276-284, 1984.

[8] H.C. Tijms, *Stochastic Modelling and Analysis: A Computational Approach*, John Wiley & Sons, 1986.

[9] T. Kimura, "Approximations for multi-server queues: system interpolations", *Queueing Systems*, Vol.17, pp. 347-382, 1994.

[10] 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, pp. 1115-1127, 1979.

[11] G.P. Cosmetatos, "Some Approximate Equilibrium Results for the Multi-Server Queue (M/G/r)", *Operational Research Quarterly*, vol 27, num.3, pp. 615-620, 1976.

[12] B. Ma, J. Mark, “Approximation of the mean queue length of an M/G/c queueing system”, *Operations Research*, Vol 43, pp. 158-165, 1995.

[13] L.P. Seelen, H.C. Tijms, M.H. Van Horn, *Tables for Multi-Server Queues*, North Holland, 1985.

[14] T. Kimura, *Exact data for waiting time in M/PH/s queues.*