Approximate Formulae for General Single Server Systems with Single and Batch Arrivals

W. Krämer und M. Langenbach-Belz,
Institut für Nachrichtenvermittlung und Datenverarbeitung, Universität Stuttgart, Seidenstraße 36, 7000 Stuttgart 1

Key-words: G1/G/1 queueing system, 2-moments approximation, mean waiting time, probability of waiting, batch arrivals

Abstract: The general single server system G1/G/1 has been treated manifold, but only for some special cases handy formulae are available. Very often exact calculations are too cumbersome and sophisticated for practical engineering, as well as upper and lower bounds generally are too gross approximations.

Therefore the need was felt to support traffic engineers with simple explicit approximation formulae, based on a 2-moments approximation.

In this paper such formulae are given for the mean waiting time and the probability of waiting, which have been derived in [23].

The quality of the formulae, which have been checked by numerous comparisons with exact and simulation results, is such, that within the most interesting range of server utilizations from 0.2 to 0.9 the error is less than 20 % (typically < 10 %) for all combinations of the arrival and service processes characterized by the following distribution types: D, E2, E3, M, H2.

Besides these types, for validation purposes, also other distribution functions have been investigated, differing in the third and higher moments.

The formulae are also easily applicable with comparable accuracy to batch arrival systems by considering equivalent single arrival arrangements.

By known relations, also simple approximations are provided, e.g. for the variances of the associated output process, and the mean length of an idle and a busy period, respectively.

Deshalb wurde die Notwendigkeit gesehen, den Ingenieur bei verkehrsmäßigen Untersuchungen mit einfachen expliziten Näherungsformeln zu unterstützen, die auf einer 2-Momenten-Approximation beruhen.

In diesem Beitrag werden solche Formeln angegeben für die mittlere Wartezeit und die Wartewahrscheinlichkeit, die in [23] abgeleitet wurden.

Die Güte der Näherungsformeln, die durch zahlreiche Vergleiche mit exakten und mit Simulationsergebnissen geprüft wurden, ist so gut, daß innerhalb des interessierenden Bereichs des Angebots von 0,2 bis 0,9 der Fehler kleiner als 20 % (typisch < 10 %) ist für alle Kombinationen von D, E2, E3, M, H2-Ankufts- und Bedienungsprozessen.

Neben diesen Typen wurden zur Validierung auch andere Verteilungsfunktionen benutzt, die sich im 3. und in den höheren Momenten unterscheiden.

Die Formeln sind ebenfalls leicht anwendbar auf Systeme mit Gruppenankünften, da in diesen Fällen äquivalente Systeme mit Einzelankünften definiert werden können.

Mit Hilfe bekannter Beziehungen stehen auch einfache Näherungen zur Verfügung, z. B. für die Varianz des zugehörigen Ausgangsprozesses und für die mittlere Dauer einer Freizeit- und Arbeitsperiode.

1 Introduction
1.1 General Remarks

In computers and communications systems very often queueing problems may be represented by queueing systems of the type G1/G/1 (general input and general service process, single server). For traffic engineers in particular, the mean waiting time and the probability of waiting are of interest for system analysis or design.

In the literature, exact and explicit solutions for such queueing systems are available only for certain types of arrival or service processes (e.g. M/G/1).

*) Concise version of a paper presented at the 8th International Teletraffic Congress (ITC) Melbourne, 10.–17.11.1976
M. Langenbach-Belz is now with Standard Elektrik Lorenz AG
Very often these different solutions for different traffic assumptions require the numerical evaluation of roots of transcendental equations by the aid of computers. For other types of arrival or service processes implicit solutions are known (e.g. based on Lindley's integral equation). But often these solutions are not straightforward and/or require a lot of evaluation work.

In many applications of traffic engineering, either the procedures and tools for these solutions are not available for the engineers in due time of the amount of evaluation work is not justified for quick estimates. On the other hand, known approximation formulae for the delay are limited strongly in application range (e.g. heavy traffic approximations).

Therefore, the urgent need was felt to support the traffic engineer with simple explicit but general approximation formulae for the mean waiting time and the probability of waiting.

The restriction to the first two moments of the inter-arrival and service time distribution functions (d.f.s) was near at hand, since e.g.:

- in case of Poisson input the mean waiting time only depends on the first two moments of the service time d.f.,
- 2-moments approximations have been proved useful e.g. for overflow systems,
- the models used for system analysis often are approximate models themselves, and often exact d.f.s are not known at all.

In addition, very encouraging results have been obtained with 2-moments approximations applied to queues in series [17], considering many different d.f.s of service times.

### 1.2 Some Notations

Let be

- $T_A$ = interarrival time of requests
- $T_W$ = waiting time of a request
- $T_H$ = service or holding time of a request
- $T_F$ = flow time of a request ($= T_W + T_H$)
- $T_{IP}$ = duration of an idle period
- $T_{BP}$ = duration of a busy period
- $T_D$ = interdeparture time of requests

Now, the traffic offered $A$ (or server utilization) is

$$\frac{E(T_H)}{E(T_A)} = \lambda \cdot E(T_H) = \lambda \cdot h, \quad (1.1)$$

with $E(T_H) = h$ as mean service time and $\lambda$ as arrival rate of the requests.

The queue discipline may be arbitrary as long as it is independent of the service times.

The d.f.'s of the interarrival and service times be $F_A(t)$ and $F_H(t)$, with associated coefficients of variation

$$c_A = \sqrt{\text{Var}(T_A)}/\text{E}(T_A), \quad c_H = \sqrt{\text{Var}(T_H)}/\text{E}(T_H). \quad (1.2)$$

### 1.3 Main Results Obtained for Single Arrivals

The key results derived in [23] are simple explicit approximation formulae for the mean waiting time $E(T_W)$ and the probability of waiting $W$ in a GI/G/1 system:

\[
E(T_W) = \frac{A \cdot h}{2(1-A)} \cdot \begin{cases} 
\frac{e^{-\frac{2(1-A)}{3A}} \cdot \frac{(1-c_A^2)^2}{c_A^2 + c_H^2}}{c_A^2 \leq 1} \\
\frac{e^{-\frac{(1-A)}{c_A^2 - 1}} \cdot \frac{c_H^2}{c_A^2 + 4c_H^2}}{c_A^2 > 1}
\end{cases} \quad (1.3)
\]

and

$$W = A + (c_A^2 - 1) \cdot \begin{cases} 
\frac{1 + c_A^2 + A \cdot c_H^2}{1 + A(c_H^2 - 1) + A^2 \cdot (4c_A^2 + c_H^2)} & c_A^2 \leq 1 \\
\frac{4A}{c_A^2 + A^2 \cdot (4c_A^2 + c_H^2)} & c_A^2 > 1
\end{cases} \quad (1.4)$$

The accuracy of these formulae has been tested by comparisons with a large amount of exact and simulation results, including in particular systems with all combinations of $D$, $E_4$, $E_2$, $M$, $H_2$ d.f.s.

From these results, further traffic characteristics can be obtained (e.g. mean queue lengths), also for batch input systems.

### 2 Existing Explicit Results

The substantial amount of publications relating to the general queueing system GI/G/1 demonstrates the state of the art of queueing theory, which nevertheless cannot reveal the gap between exact mathematical results and quick engineering applications.

Several calculation methods have been developed to solve GI/G/1 problems, including

- ERLANG's method of phases, see e.g. [1] - [3],
- LINDLEY’s integral method [4],
- KENDALL’s method of the imbedded Markov chain [5].

These methods and others have been proved very powerful to penetrate into the theoretical depth and to develop many exact results. But for practical applications there are severe disadvantages:

1) Some methods require special traffic assumptions in order to be applicable.
2) Most of the existing exact results include Laplace transforms and generating functions as well as roots of partly very sophisticated equations.
3) For each type of a special GI/G/1 system, the way of obtaining numerical results may be more or less different.

In the essence, simple explicit and exact formulae for the mean waiting time and the probability of waiting exist only for the case of Poisson input (M/G/1). The well-known formula of POLLACZEK and KHINTCHINE (1930/32), cf. e.g. [3], is

\[ E(T_W) = \frac{A(1 + c_A^2)}{2(1 - A)} \cdot h, \]  

(2.1)

here only given for the first moment of the waiting time.

The associated probability of waiting simply is

\[ W = A. \]  

(2.2)

The need for bridging the above mentioned gap has been recognized manifold and resulted e.g. in

- the application of numerical inversion techniques for Laplace transforms,
- fitting observed distributions by step functions [6] or phase-type functions,
- tables for queueing systems of different types, see e.g. [19],
- the derivation of upper and lower bounds for the mean waiting times [9], [10], [11].

KINGMAN [7] has derived an upper bound for the mean waiting time in GI/G/1

\[ E(T_W) \leq \frac{A(c_A^2/A^2) + c_H}{2(1 - A)} \cdot h \]  

(2.3)

being a good approximation for heavy traffic (A → 1).

Some authors also derived lower bounds or low traffic approximations ([8], [9]) for single arrival GI/G/1 queues.

Also diffusion approximation methods have been applied [12], to obtain heavy traffic formulae even for transient conditions. The mean virtual delay for stationary conditions has been approximated in [12] by

\[ E(T_W)_{\text{virtual}} = \frac{A(c_A^2 + c_H^2)}{2(1 - A)} \cdot h. \]  

(2.4)

Unfortunately, the upper and lower bounds for real mean waiting times are not very useful for the major interesting range of utilizations. E.g. the heavy traffic formula (2.3) applied to A = 0.7 overestimates the mean waiting times up to 100%.

Therefore, the goal was set to derive purely heuristically an extension of the Pollaczek-Khintchine formula, allowing to give quick answers with a reasonable accuracy, though being restricted to the first 2 moments of interarrival and service time d.f.s.

3 Systems with Single Arrivals

3.1 Approximation of the Mean Waiting Time

Preliminary investigations had shown, that a good starting point for the heuristic approximation of the mean waiting time was the formula

\[ E(T_W) = \frac{A \cdot h}{2(1 - A)} \cdot (c_A^2 + c_H^2) \cdot g(A, c_A^2, c_H^2). \]  

(3.1)

containing a refinement function g(·). This function has been derived in [23] by using exact boundary conditions and by a heuristical determination of free parameters, which led to the final result (1.3).

Fig. 3.1 is an example of comparisons between the approximate results according to (1.3) and results of simulation runs.

The coincidence between approximate and simulation results is sufficiently good. This has also been shown for many other cases, cf. [23].

![Fig. 3.1 Mean Waiting Times for D/G/1 Systems (1 Simulation results with 95% confidence intervals. H2 is a hyperexponential d.f. with 2 alternative branches being utilized equally.)](image)

3.2 Approximation of the Probability of Waiting

Similarly as in the previous case for the mean waiting time, the frame of the formula for the probability of waiting W had to satisfy different boundary conditions. As a reasonable base it was selected

\[ W = A + (c_A^2 - 1) \cdot A \cdot (1 - A) \cdot f(A, c_A^2, c_H^2) \]  

(3.2)

with f(·) as another refinement function. This function has been derived analogously in [23], leading to (1.4).

Fig. 3.2 shows as an example some approximation results for the probabilities of waiting for systems with E4 input processes, compared with simulation results (small confidence intervals omitted).
3.3 Error Considerations

Since only the first two moments of both interarrival and service time d.f. are considered, certain errors have been tolerated from the start.

First of all, as a dedicated statement, for all GI/G/1 systems with arbitrary combinations of the d.f.'s

\[ D, E_4, E_2, M, H_2 \ (c^2 = 2), H_2 \ (c^2 = 4) \]

(here called "standard" types), the error in the mean waiting time and the probability of waiting is less than \( \approx 20\% \) for traffics offered \( A = 0.2 \) up to 0.9. Typically, the error is less than 10%.

The same error limits are obtained even if higher variances of \( H_2 \) d.f.'s are included, only restricted by the condition

\[ c_A^2 + c_H^2 \leq 12, \]

which seems to be uncritical for practical applications.

To judge the influence of the third and higher moments, many other types of d.f.'s have been investigated in [23], confirming the wide application range of the approximation formulae.

4 Systems with Batch Arrivals

In this chapter it is shown that the formulae for the mean waiting time and the probability of waiting also can be applied to system with batch arrivals by defining equivalent single arrival systems.

4.1 General Remarks

In communications systems, e.g., often the arrivals at centralized units occur in batches or groups of arbitrary size \( K \), now being assumed to be an independent random variable. Let \( p_K \) (\( K = 0, 1, \ldots \)) be the probability that at a possible arrival instant of a batch (characterized by a d.f. for the interarrival times \( T_{AB} \) of batches) a batch of size \( K \) arrives. For reasons of generality, let also "batches" of size \( K = 0 \) be included. This is useful especially in so-called sampled systems, where the batches only arrive at equidistantly distributed clock instants (see e.g. [13]).

Now, by standard calculations, general relations between characteristic values of the first request of a batch (indexed by an additional 1) and an arbitrary member of it can be established. With respect to the mean waiting time it results

\[ E(T_W) = E(T_{W1}) + \left( \frac{\text{Var}(K)}{E(K)} + E(K) - 1 \right) \cdot \frac{h}{2}. \]  

(4.1)

This formula can be found e.g. in [10], [13] and [14].

With respect to the probability of waiting, it can be simply stated that with the probability \( q \) that an arbitrary request is the first request within its batch

\[ q = \frac{1}{E(K|K > 0)} = \frac{1 - p_0}{E(K)}. \]  

(4.2)

it holds

\[ W = q \cdot W_1 + (1 - q) \cdot 1. \]  

(4.3)

This means

\[ W = 1 - \left( 1 - \frac{p_0}{E(K)} \right) \cdot (1 - W_1). \]  

(4.4)

The two formulæ (4.1) and (4.4) allow to determine values for single requests, if the corresponding ones of the first of a batch are known, see 4.2.

4.2 Equivalent Systems with Single Arrivals

The waiting time of the first request of a batch can be calculated by considering an equivalent system, defining a whole batch \( (> 0) \) as a "super-request" [2].

Then the new interarrival times are the times between the arrivals of batches \( > 0 \), the service times are the total times to serve a whole batch \( (> 0) \).

If we denote the characteristics of the equivalent system by an additional asterisk*, then

\[ E(T_W^o) = E(K|K > 0) \cdot E(T_H). \]  

(4.5)

With

\[ \text{Var}(T_W^o) = E(T_H)^2 \cdot \text{Var}(K|K > 0) + E(K|K > 0) \cdot \text{Var}(T_H). \]  

(4.6)

it can be shown that the resulting equivalent (squared) variation coefficient of service time is

\[ c_{H^*}^2 = \frac{1 - p_0}{E(K)} \left( \frac{\text{Var}(K)}{E(K)} + c_H^2 \right) - p_0. \]  

(4.7)

Let now the arrival process be characterized by

- the interarrival times \( T_{AB} \) of batches (of size \( K \geq 0 \)), i.e. the times between successive closings of an "input stack", with mean \( E(T_{AB}) \) and variation coefficient \( c_{AB} \),
- the batch size probabilities \( p_K \) (\( K \geq 0 \)).
Since the equivalent system is based on batches of size greater than zero, it holds
\[ E(T_A^E) = \frac{E(T_{AB})}{1 - p_0} \]  \hspace{1cm} (4.8)

and
\[ c_A^2 = (1 - p_0) \cdot c_{AB}^2 + p_0. \]  \hspace{1cm} (4.9)

Summarizing, the equivalent system is characterized by:
- the mean service time
\[ E(T_H^E) = \frac{E(K)}{1 - p_0} \cdot E(T_H), \]  \hspace{1cm} (4.10)
- the same traffic offered as in the original batch input system
\[ A^E = \frac{E(T_H^E)}{E(T_A^E)} = \frac{E(K) \cdot E(T_H)}{E(T_A)} = A, \]  \hspace{1cm} (4.11)
- and the variation coefficients \( c_A^E \) and \( c_H^E \) according to (4.9) and (4.7).

These values have to be calculated before using the GI/G/1 approximation formulae for the equivalent single arrival system.

### 4.3 Examples for Batch Arrival Systems

Since for M/G/1 systems the formulae (1.3) and (1.4) are exact, all batch arrival systems rendering equivalent M/G/1 systems will be calculated exactly.

This holds for negative exponentially distributed interarrival times between batches of arbitrary size \( K \geq 0 \) and distribution (i.e. \( c_{AB}^2 = 1, p_0 \equiv 0 \)).

Combining (4.1) with (1.3) and (4.7) to (4.11), it results for negative exponentially distributed interarrival times between batches of arbitrary size \( K \geq 0 \)
\[ E(T_W) = \frac{h}{2(1 - A)} \left[ A(1 + c_A^2) + \frac{\text{Var}(K)}{E(K)} + E(K) - 1 \right]. \]  \hspace{1cm} (4.12)

For the probability of waiting from (4.4) with \( W_1 = A \)
\[ W = 1 - \frac{1 - p_0}{E(K)} \cdot (1 - A). \]  \hspace{1cm} (4.13)

Eq. (4.12) is identical with a result of GAVER [20], who derived the generating function of the state probabilities of systems with compound Poisson arrival processes (K > 0).

Since the formulae are exact for compound Poisson arrival processes, for validity purposes, it remains to show examples, where the times between two batch arrivals are not negative exponentially distributed.

To select an extreme but nevertheless important case, equidistantly distributed interarrival times are selected, i.e. so-called sampled systems. These systems may be conceived as having an input switch being closed periodically.

If the single requests arrive in front of the sample switch with negative exponentially distributed interarrival times, the distribution of the batch sizes is Poisson, having \( \text{Var}(K)/E(K) = 1 \).

For the examples shown in fig. 4.1 the clock time was chosen to be equal to the mean service time \( E(T_{AB}) = E(T_H) \), rendering \( A = E(K) \) and thus determining the \( p_0 \) values via the Poisson distribution.

Then from (4.9) \( c_A^2 = p_0 \), whereas \( c_H^2 \) is determined with (4.7).

The solid lines are exact results, the curves for \( M \) and \( E_q \) have been calculated with a program of WEISSCHUH and WIGGALL [15]. For constant service times it simply holds (LANGENBACH-BELZ [13])
\[ E(T_W) = \frac{A}{2(1 - A)} \cdot h. \]  \hspace{1cm} (4.14)

The associated curves for the probabilities of waiting \( W \) nearly have been identical for all 3 d.f.s with the line \( W = A \), both simulation and approximate calculation. Therefore they are not shown here.

Remember that these approximation results have been obtained with very low calculation effort. Further curves can be found in [19] and [23].

### 5 Determination of Further Traffic Values

Up to this chapter only the expected waiting time \( E(T_W) \) and the probability of waiting \( W \) have been considered, both for single arrival and batch arrival systems. Based on relations for a wide class of stationary single server systems (MARSHALL [9], RICE [16]), the approximation formulae (1.3) and (1.4) can also be used to calculate approximately further systems characteristics.
5.1 Output Variance

The following relation between the variance of the output process and the mean waiting time can be found in [9]:

$$\text{Var}(T_D) = \text{Var}(T_A) + 2 \text{Var}(T_H) - \frac{2}{\lambda} (1 - A) \cdot E(T_W).$$

(5.1)

Using coefficients of variation and $E(T_D) = E(T_A)$:

$$c_D^2 = c_A^2 + 2A^2 \cdot c_H^2 - 2A (1 - A) \cdot \frac{E(T_W)}{h}.$$  

(5.2)

With $E(T_W)$ according to (1.3) this is a simple and explicit approximation for the variation coefficient $c_D$ of the output or departure process.

This formula specializes for Poisson input to the well-known form [21]

$$c_D^2 = 1 - A^2 \cdot c_H^2.$$  

(5.3)

KÜHN [18] has used (5.2) for the output processes in general queueing networks and took over (1.3) (see also output figures in [18]).

5.2 Idle and Busy Period

According to an exact formula from [16]

$$(1 - W) \cdot E(T_{IP}) = \frac{1}{\lambda} - h,$$

(5.4)

also the mean time $E(T_{IP})$ of an idle period can be approximated, using (1.4):

$$E(T_{IP}) = \frac{1 - A}{A(1 - W)} \cdot h.$$  

(5.5)

There is also a possibility to approximate the mean length of a busy period $T_{BP}$ via the relation [16]

$$E(T_{BP}) = \frac{1 - P_0}{P_0} \cdot E(T_{IP}),$$

(5.6)

where $P_0$ is the absolute probability of an empty system. For pure waiting systems, as being considered here, $P_0 = 1 - A$, such that with (5.5)

$$E(T_{BP}) = \frac{1}{1 - W} \cdot h.$$  

(5.7)

In RIORDAN [22] eqns. (5.4) and (5.6) are derived without special assumptions concerning the input process. Thus the equations for the mean idle period (5.5) and the mean busy period (5.7) are also applicable to batch input systems with $W$ according to (4.4).

6 Summary and Conclusion

For the general single server system GI/G/1 simple 2-moments approximations have been given for the mean waiting time and the probability of waiting. The stimulation have been the gap between many complex exact results and a quick numerical calculation for engineering purposes.

With the restriction to 2 moments, the formulae and application should be quick and simple, naturally thus inducing certain errors.

The accuracy of the formulae has been investigated and proved to be very useful within a wide range of applications or traffic assumptions, also easily including systems with batch arrivals.

In addition, also simple formulae are available for the variances of the output processes as well as the mean values for the idle and busy periods.

The usefulness of the approximation of the mean waiting times and of the output variances has been already demonstrated by KÜHN [18] in context with queueing networks. Also tables including these approximations are provided [19].

It is hoped that these heuristic approximation, which cannot and will not replace GI/G/1 investigations with detailed reflection of the interarrival and service time d.f.'s, will be helpful for the traffic engineer to obtain very quickly and simply useful estimates for the delay in his special single server models.

References


Eingegangen am 17.10.1977