

Performance analysis of a batch service queue arising out of manufacturing system modelling

H. Gold and P. Tran-Gia

*Institute of Computer Science, University of Würzburg, Am Hubland,
D-97074 Würzburg, Germany*

Received 6 January 1992; revised 28 January 1993

In this paper, we present an exact analysis of a queueing system with Poisson arrivals and batch service. The system has a finite number S of waiting places and a batch service capacity b . A service period is initialized when a service starting threshold a of waiting customers has been reached. The model is denoted accordingly by $M/G^{[a,b]}/1-S$. The motivation for this model arises from manufacturing environments with batch service work stations, e.g. in machines for computer components and chip productions. The method of embedded Markov chain is used for the analysis, whereby a representation of the general service time is obtained via a moment matching approach. Numerical results are shown in order to illustrate the dependency of performance measures on special sets of system parameters. Furthermore, attention is devoted to the issues of starting rules, where performance objectives like short waiting time, small blocking probability and minimal amount of work in progress are taken into account.

Keywords: Queueing analysis; performance analysis; manufacturing modelling; batch systems; Markov chain.

1. Introduction

Batch service models are useful to investigate the performance of various processes in production environments. Models arising out of the manufacturing technology for every large scale integrated (VLSI) circuits and the related problems in production planning and control of VLSI manufacturing have been already investigated in the literature in operations research.

Most of the investigations which employed batch service models consider queueing systems having infinite capacities. This assumption stands in contrast to real systems, where e.g. factories' buffer storage capacities of transfer line systems are finite. The starting strategies required for batch servers in production environments seems to be more subtle than those usually regarded in the literature.

There are a few studies in the literature, which deal with batch service systems. The modelling approaches consider various starting strategies with respect to the

following question: when to start a batch processing by observing the number of parts waiting in the input buffer. The most important ones among these are:

- server starts to serve with maximum batch size or fewer jobs according to queue length (Bailey [5], Downton [6], Gnedenko and König [7], Gross and Harris [8]),
- server starts to serve when the number of jobs in the queue reaches a certain threshold a (Neuts [1–3], Chaudhry et al. [4]). This is also considered in this paper.

The major analysis method employed in these studies is the embedded Markov chain technique. The system $M/G^{[1,b]}/1-\infty$ was first analyzed by Bailey [5]. He provided the z -transform of the distribution of the queue length. Detailed calculations of the mean and variance of the queue length and of the mean waiting time were given for the case of χ^2 -distributions with an even number of degrees of freedom of the service time. In Downton [6] the analysis is focused on the waiting time distribution (with non-waiting jobs excluded). The waiting time distribution of the system $M/M^{[a,b]}/1-\infty$ and $M/G^{[1,b]}/1-\infty$ can be found in Gnedenko and König [7]. In some standard queueing literature (e.g. Gross and Harris [8]) results for the state probabilities of some basic batch service systems like $M/M^{[1,b]}/1-\infty$ are given, where the starting strategy was simplified as follows: new arrivals immediately enter service up to the limit b , and finish together with the other jobs being served in the current production phases. The infinite system $M/G^{[a,b]}/1-\infty$ has also already been investigated by some authors. Neuts [1,2] derived results for the queue length, the distribution function of the busy period and a description of the output process of the system. In Neuts [3] he gave a matrix-geometric solution for the steady-state probabilities of the system. Chaudhry et al. [4] showed how to numerically evaluate the steady-state probabilities and moments for the number in system at postdeparture, prearrival and random epochs. Thereby they use the following service time distributions: Deterministic (D), Erlangian (E_k), two-phase hyperexponential (HE_2) and uniform (U). Thus any given service time characteristic given in a representation of its first two moments (mean and coefficient of variation) can roughly be accommodated. We close the gap left by using D, E_k and U for distributions with $c < 1$ and therefore introduce a two-phase distribution consisting of a deterministic and a Markovian period (D+M).

In this study we stress the manufacturing issues of the finite system $M/G^{[a,b]}/1-S$, in particular the dimensioning aspects for the threshold a . A proper choice of this threshold depends on service time characteristics, capacity of the waiting room in front of the batch server and traffic load. The quality criteria are the aims of lean production: short response times, low blocking probability, efficient use of resources, continuous flow through the production line. In order to study the question of the proper choice we devote attention to the finite system $M/G^{[a,b]}/1-S$, where a denotes the server starting threshold, S the waiting room capacity. Since we are interested in practical applicability of our investigations, we attach great importance to the following points:

- numerical tractability of the derived solution,
- easy parameterization of the analytic model according to real world problems particularly with regard to service time characteristics and waiting room capacity,
- the possibility of instationary analysis.

The paper is organized as follows. Details of the model and related parameters are given in section 2. The analysis using an embedded Markov chain is subdivided into the calculation of the Markov chain state probabilities (section 3) and the derivation of the arbitrary time state probabilities (section 4). In section 5 the latter state probability vector is used to obtain the characteristic performance measures. In sections 6 and 7 we explain details concerning the numerical handling of our general solution and finally in section 8 we provide an insight into dimensioning aspects for a common class of production machines by means of diagrams.

2. Model description

The basic model is illustrated in fig. 1. The model consists of a finite capacity queue which is served by a single batch server according to a starting rule to be specified below. The arrival stream of jobs constitutes a Poisson process. The server has a maximum capacity of “ b ” jobs and the service time is generally distributed. The server starting scheme is driven by the number of jobs waiting in the queue. When the server is idle and there are less than a number “ a ” of jobs in the queue, the server remains idle until “ a ” jobs have been accumulated. At the end of a service phase, the server will proceed according to the number of waiting jobs. If there are more than a number “ a ” of jobs in the queue at the scheduling time, the server will start the next service immediately by taking up to “ b ” waiting jobs. Jobs seeing upon arrival a full queue are thought of to be blocked.

The following symbols and random variables (r.v.) are used in this paper:

- λ arrival rate,
- H r.v. for the service time distribution for a batch,
- b server capacity,
- S queue capacity,

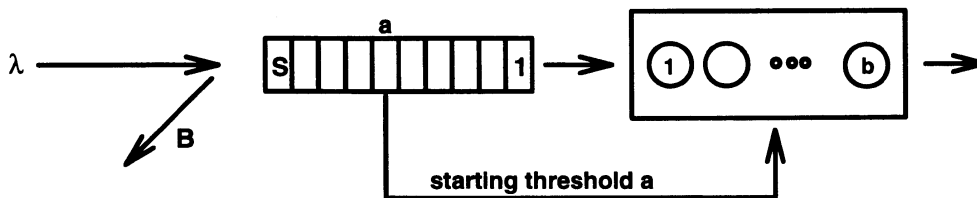


Fig. 1. The basic queueing model.

a	service starting threshold,
X	r.v. for the number of jobs in the queue at embedded points,
X^*	r.v. for the number of jobs in the queue at an arbitrary observation epoch,
Y	r.v. for the number of jobs in the server,
$Y^{(A)}$	r.v. for the number of jobs in the activated server,
B	blocking probability,
D	r.v. for the number of arriving jobs during a service period.

For a random variable (r.v.), e.g. X , we use the following notation:

$f_X(t)$	probability density function (pdf) of r.v. X ,
$F_X(t)$	probability distribution function (PDF) of r.v. X ,
EX	mean of X ,
c_X	coefficient of variation of X .

3. Markov chain state probabilities

We observe the state process in the finite queue. At the end of a service phase, a Markov chain can be embedded. The regeneration points of the embedded Markov chain are chosen to be immediately after departure instants of jobs from service.

We consider for this purpose the two-dimensional stochastic process $(Z(t))$ with $(Z(t)) = (X(t), U(t))$, which is a Markov process, where $X(t)$ denotes the number of jobs in the queue at time t and $U(t)$ the remaining service time for the batch actually in service at time t . Thus we observe now a set of points in time T_N with $Z(T_N - 0) = (X(T_N - 0), 0)$ or $Z^{(N)}(0^-) = (X^{(N)}(0^-), 0)$. The entities $\dots X^{(N-1)}(0), X^{(N)}(0), X^{(N+1)}(0), \dots$ constitute an embedded Markov chain, since the arrival process offered to the queue is Poisson. The sequence $\dots X^{(N-1)}(0), X^{(N)}(0), X^{(N+1)}(0), \dots$ relates to the non-stationary Markov chain state probabilities

$$x^{(n)}(k) = \Pr\{X^{(n)}(0^-) = k\}, \quad k = 0, 1, \dots, S. \quad (1)$$

The steady-state probabilities of the Markov chain under stationary conditions are defined as

$$x(k) = \Pr\{X(0^-) = k\} = \lim_{n \rightarrow \infty} x^{(n)}(k), \quad k = 0, 1, \dots, S. \quad (2)$$

In order to calculate the transition probabilities q_{ij} of the Markov chain,

$$q_{ij} = \Pr\{X^{(n+1)}(0^-) = j / X^{(n)}(0^-) = i\}, \quad (3)$$

we observe the state development of the queue shown in fig. 2. At time t_1 a service period ends and $i < a$ customers are in the queue. Thus a type-4i-interval starts. This interval ends at time t_2 when a number of a customers has accumulated in the

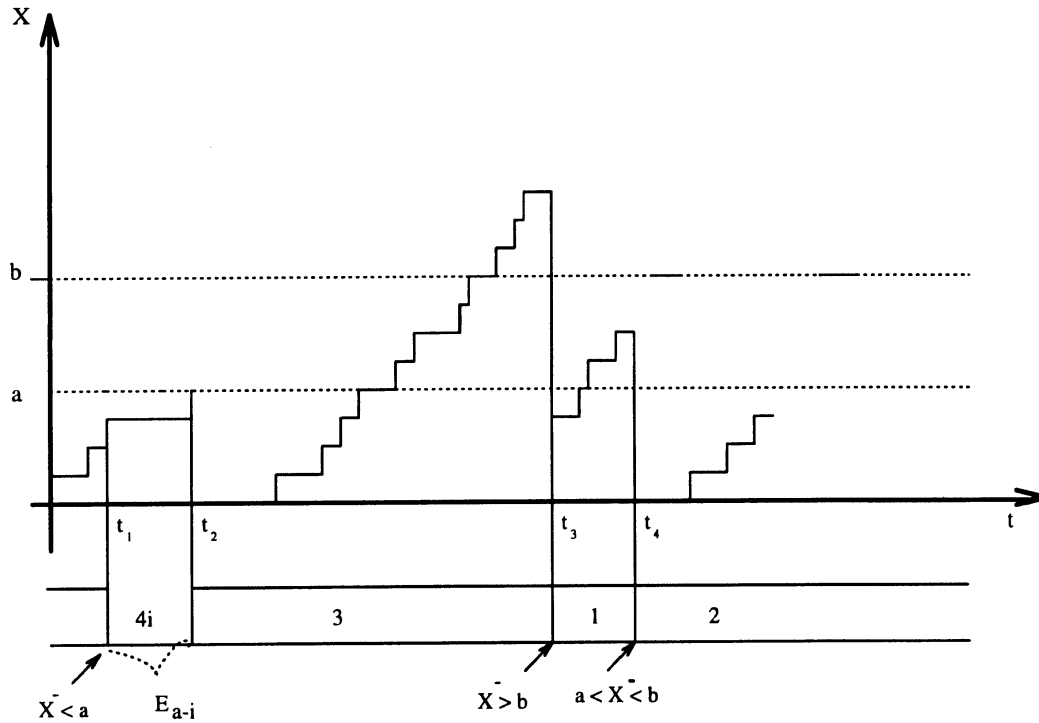


Fig. 2. State space dynamics of the system $M/G^{[a,b]}/1 - S$.

queue. At time t_2 a service period begins and the number in the queue drops to 0. During this service period which ends at time t_3 more than b customers arrive. At time t_3 a number of b customers is taken into service and it is finished at time t_4 . At time t_4 there are more than a but fewer than b customers present in the queue, so all waiting customers are taken into service.

Now we are able to derive the matrix of transition probabilities $Q = \{q_{ij}\}$ by inspecting the state development. In table 1 we show typical entrances of this matrix. Thereby d_i indicates the probability for i arrivals during a service phase:

$$d_i = \int_0^\infty a_i(t) f_H(t) dt \quad \text{with} \quad a_i(t) = \frac{(\lambda t)^i}{i!} e^{-\lambda t}. \quad (4)$$

Finally the Markov chain state probability vector at time $T^{(n+1)}$, denoted by $X^{(n+1)}$, can be determined out of the state probability vector at time $T^{(n)}$ using

$$X^{(n+1)}(0^-) = Q^T X^{(n)}(0^-). \quad (5)$$

To obtain the steady state probabilities of the Markov chain is just to find the eigen vectors of the matrix Q with respect to the normalizing condition

$$\sum_{i=0}^S x(i) = 1. \quad (6)$$

Table 1
Matrix of transition probabilities.

	0	1	$i - b - 1$	$i - b$	a	j	$S - b - 1$	$S - b$	S	
$Q =$	d_0	d_1	d_{i-b-1}	d_{i-b}	d_a	d_j	d_{S-b-1}	d_{S-b}	$\sum_{k=S}^{\infty} d_k$	0
	d_0	d_1	d_{i-b-1}	d_{i-b}	d_a	d_j	d_{S-b-1}	d_{S-b}	$\sum_{k=S}^{\infty} d_k$	1
	d_0	d_1	d_{i-b-1}	d_{i-b}	d_a	d_j	d_{S-b-1}	d_{S-b}	$\sum_{k=S}^{\infty} d_k$	a
	d_0	d_1	d_{i-b-1}	d_{i-b}	d_a	d_j	d_{S-b-1}	d_{S-b}	$\sum_{k=S}^{\infty} d_k$	$b - 1$
	d_0	d_1	d_{i-b-1}	d_{i-b}	d_a	d_j	d_{S-b-1}	d_{S-b}	$\sum_{k=S}^{\infty} d_k$	b
	0	d_0	d_{i-b-2}	d_{i-b-1}	d_{a-1}	d_{j-1}	d_{S-b-2}	d_{S-b-1}	$\sum_{k=S-1}^{\infty} d_k$	$b + 1$
	0	0	0	d_0	d_{a-1+b}	d_{j-i+b}	d_{S-i-1}	d_{S-i}	$\sum_{k=S-i+b}^{\infty} d_k$	i
	0	0	0	0	0	0	0	d_0	$\sum_{k=b}^{\infty} d_k$	S

4. Arbitrary time state probabilities

Using the Markov chain state probability the state probability at an arbitrarily chosen observation epoch can be derived. Recall that X^* is the random variable for the number of jobs in the queue at an arbitrary point in time, i.e.

$$\begin{aligned}
 x^*(k) &= P\{\text{an arbitrary outside observer sees the queue in state } k\} \\
 &= P\{X^* = k\}.
 \end{aligned}$$

Looking at the step-by-step state transitions, as shown in the lower part of fig. 2, we can recognize four types of intervals. Details of these interval types are given in table 2, where the probability for appearance, the length and the number of jobs in service for the different interval types are listed. In this table E_n indicates the Erlang distribution of order n . This leads to the probabilities that an outside observer looking arbitrarily at the system sees an interval of type 1, 2, 3, 4i ($i = 0, \dots, a - 1$):

Table 2
Characteristics of different transition interval types.

Type	Probability γ_{type}	Interval length T_{type}	Mean ET_{type}	Number in service
1	$\sum_{i=b+1}^{\infty} x(i)$	H	EH	b
2	$\sum_{i=a}^b x(i)$	H	EH	x^-
3	$\sum_{i=0}^{a-1} x(i)$	H	EH	a
4i	$x(i) (i < a)$	E_{a-i}	$(a - i)/\lambda$	0

$$\pi_{\text{type}} = \frac{\gamma_{\text{type}} * ET_{\text{type}}}{\sigma}, \tag{7}$$

where the normalization factor σ is given by

$$\sigma = \sum_{\text{all types}} \gamma_{\text{type}} * ET_{\text{type}}. \tag{8}$$

The time interval from the last scheduling epoch until the observation time in the case of interval type 1, 2 or 3 is the recurrence time of the service time with the probability density function

$$f_H^r(t) = \frac{1 - F_H(t)}{EH}. \tag{9}$$

Thus, the arrival probabilities during the forward recurrence time is

$$d_i^* = \int_0^\infty a_i(t) f_H^r(t) dt. \tag{10}$$

We define T as the random variable for the type of interval seen at an arbitrary epoch. Considering all four types of intervals and combining the above results, the arbitrary time state probabilities can be written as follows:

$$j = 0, \dots, a - 1,$$

$$x^*(j) = \sum_{i=b+1}^S P(X(0^-) = i, T = 1) d_{j-i+b}^* + (\pi_2 + \pi_3) d_j^* + \sum_{i=0}^j \frac{\pi_{4i}}{a-i}, \tag{11}$$

$$j = a, \dots, S - 1,$$

$$x^*(j) = \sum_{i=b+1}^S P(X(0^-) = i, T = 1) d_{j-i+b}^* + (\pi_2 + \pi_3) d_j^*, \tag{12}$$

$$j = S,$$

$$x^*(j) = \sum_{i=b+1}^S P(X(0^-) = i, T = 1) \sum_{k=S-i+b}^\infty d_k^* + (\pi_2 + \pi_3) \sum_{k=S}^\infty d_j^*, \tag{13}$$

where the joint probability $P(X = i, T = \text{type})$ is given by

$$P(X(0^-) = i, T = \text{type}) = \frac{x(i) * \pi_{\text{type}}}{\gamma_{\text{type}}}. \tag{14}$$

5. System characteristics

As the arrivals follow a Poisson process having the Markov property (cf. PASTA: Poisson arrivals see time averages), the arbitrary time state probability from eq. (13) can be used directly to yield the blocking probability B :

$$B = x^*(S). \quad (15)$$

The mean waiting time EW in the queue can be derived applying the Little's theorem:

$$EW = \frac{EX^*}{\lambda(1 - B)}, \quad (16)$$

where

$$EX^* = \sum_{i=0}^S ix^*(i) \quad (17)$$

is the mean queue length. The amount of accepted traffic is $\lambda(1 - B)$. Hence, again with Little's theorem, we get the formula for the average number of jobs in the server:

$$EY = \lambda(1 - B)EH. \quad (18)$$

In particular, for use in production environments, the number of jobs the server takes for each start is of interest. It indicates the efficiency of machine handling and starting rules. This measure, the mean number of jobs per start, is given by

$$EY^{(A)} = a \sum_{i=0}^{a-1} x(i) + \sum_{i=a}^{b-1} ix(i) + b \sum_{i=b}^S x(i). \quad (19)$$

6. Substitute service time distribution functions

The service time of the batch server can be arbitrarily specified. However, in order to have a parametric representation of the service time in the numerical results discussed below, we adopt the two-moment substitution as proposed in Kuehn [9] as well as in Tran-Gia and Raith [10]. The r.v. H be now characterized by only two parameters: mean and coefficient of variation, where the following substitute distribution functions $F(t)$ are used:

Case 1: $0 \leq c_H \leq 1$

$$F_H(t) = \begin{cases} 0, & 0 \leq t \leq t_1, \\ 1 - e^{-(t-t_1)/t_2}, & t \geq t_1, \end{cases} \quad (20)$$

where $t_1 = EH(1 - c_H)$ and $t_2 = EHc_H$.

Case 2: $c_H > 1$

$$F_H(t) = 1 - pe^{-t/t_1} - (1 - p)e^{t/t_2}, \quad (21)$$

where

$$t_{1,2} = EH \left(1 \pm \sqrt{\frac{c_H^2 - 1}{c_H^2 + 1}} \right)^{-1} \quad \text{and} \quad p = EH/2t_1, \quad pt_1 = (1 - p)t_2.$$

With regard to the corresponding probability density functions, eq. (4) yields:

Case 1: $0 \leq c_H \leq 1$

$$d_j = \frac{(\lambda t_2)^j}{(1 + \lambda t_2)^{j+1}} e^{-\lambda t_1} \sum_{k=0}^j \frac{((t_1/t_2)(\lambda t_2 + 1))^k}{k!}. \tag{22}$$

Case 2: $c_H > 1$

$$d_j = p \frac{(\lambda t_1)^j}{(1 + \lambda t_1)^{j+1}} + (1 - p) \frac{(\lambda t_2)^j}{(1 + \lambda t_2)^{j+1}}. \tag{23}$$

Based on the arrival probabilities in eqs. (22) and (23) the Markov chain state probabilities are numerically calculated. Analogously, the arrival probabilities during the recurrence time of the service time can be derived (cf. Tran-Gia and Raith [10]):

Case 1: $0 \leq c_H \leq 1$

$$d_j^* = \frac{1}{\lambda(t_1 + t_2)} \left(1 - \sum_{k=0}^j \frac{(\lambda t_1)^k}{k!} e^{-\lambda t_1} \right) + \frac{t_2}{t_1 + t_2} d_j, \tag{24}$$

with d_j given in eq. (22).

Case 2: $c_H > 1$

$$d_j^* = \frac{(\lambda t_1)^j}{2(1 + \lambda t_1)^{j+1}} + \frac{(\lambda t_2)^j}{2(1 + \lambda t_2)^{j+1}}. \tag{25}$$

7. Numerical calculation of Markov chain state probabilities

In section 3 a recursive calculation scheme for the Markov chain state probabilities has been stated. Recall that the stationary Markov chain state probabilities are defined by

$$X(0^-) = \lim_{n \rightarrow \infty} X^{(n)}(0^-) \tag{26}$$

assuming the limit exists. This property is guaranteed by the normalization condition in eq. (6) in conjunction with the Bolzano–Weierstrass theorem which indicates that if A is a bounded set containing infinitely many points in a metric space S ,

then A has at least one limit point. As discussed, the calculation of the stationary Markov chain state probabilities is reduced to the determination of eigen vectors of the state transition matrix Q . Since the matrix Q for models depicted from a real production environment is normally very large, in the numerical result section to follow we calculate $X(0^-)$ iteratively according to eqs. (5) and (6). Subsequently the arbitrary time state probabilities according to section 4 and the performance measures exposed in section 5 are evaluated.

8. Numerical results

In this section, we present numerical results for various classes of service processes, different service starting or batch collection rules and under various load conditions. In the discussion of the results we stress the influence of, firstly, the variation of the service process, secondly, the service starting threshold dimensioning and, finally, the traffic intensity on the mean waiting time and on the average number of jobs per start, keeping in mind that these are essential aspects in models considered in production environments.

In accordance with the substitute distributions discussed in section 7 we use a $D+M$ distribution and an H_2 distribution to obtain service time distributions with $0 \leq c_H \leq 1$ and $c_H > 1$, respectively. Note that this parametric representation though an appropriate means for our purposes arises not quite as naturally as e.g. the use of the negative-binomial distribution for a parametric representation of stochastic processes in discrete-time domain. Since the time variables are standardized by $EH = 1$, the offered traffic intensity is just $\rho = \lambda/b$. In the following we stick to the case of a $M/G^{[a,b]}/1-S$ system with $b = 32$ and $S = 64$.

Figure 3 shows the mean waiting time as a function of the traffic intensity. This figure includes a family of curves for different values of the service starting threshold ($a = 4, a = 16$) and coefficient of variation of the service time ($c_H = 0, c_H = 1$). As can be seen, with traffic intensity very low the mean waiting time is very high especially when the service starting threshold “ a ” is much larger than one. With increasing traffic intensity waiting time decreases, takes its minimum and finally tends to $EW = S/b = 2$ as the limit for $\rho \rightarrow \infty$. For the case of deterministic service time the best batch collection rule is not to collect batches at all but to start the server even with only a single job in the queue. In the case of $c_H > 0$ there is a crossover of the waiting time diagrams for service starting thresholds $a = 1$ and $a > 1$. This is due to the fact that normally during shorter service periods less jobs will arrive and thus the server is often caused to work inefficiently. The reduction of waiting time gained by choosing the service starting threshold “ a ” appropriately gets larger with growing coefficient of variation of the service time and diminishes slightly with higher traffic intensity (fig. 4). The discontinuous behavior of the curves in fig. 4 at the change of substitute distribution types ($c_H = 1$) is due to the unnatural element of the representation of the service time

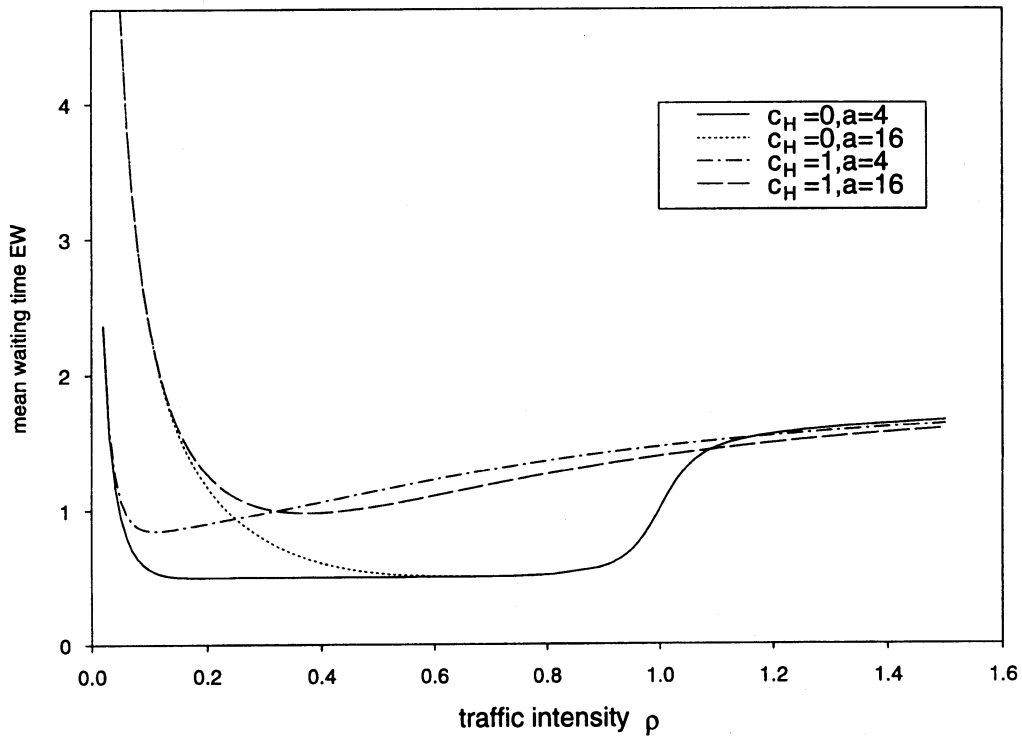


Fig. 3. Waiting time behavior.

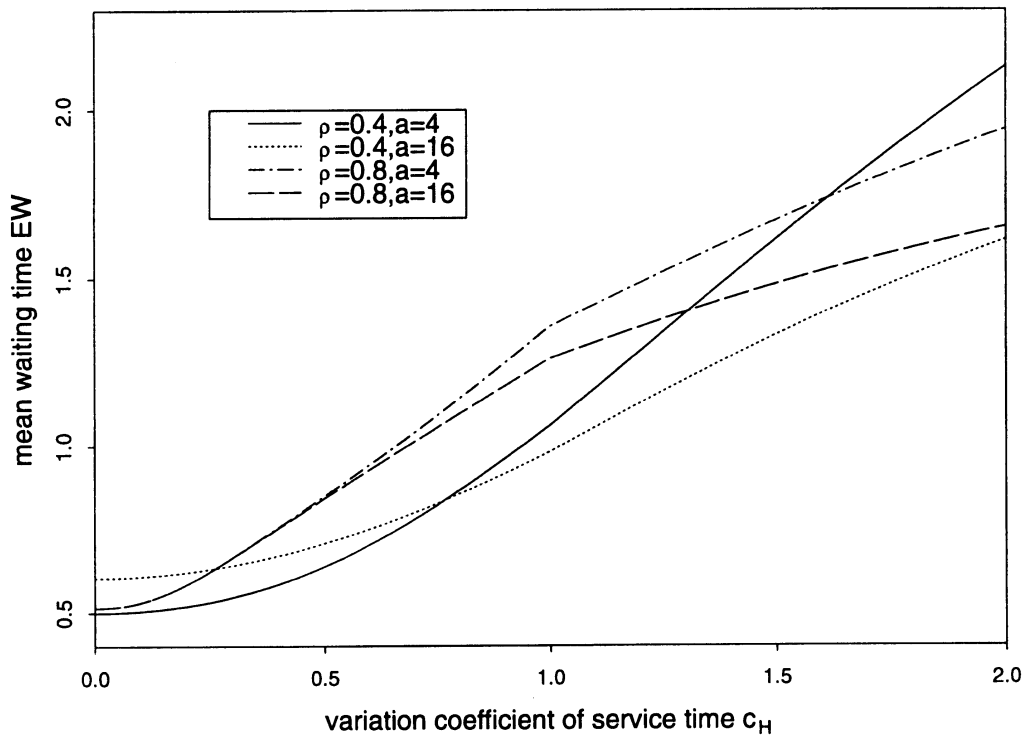


Fig. 4. Influence of service process.

distribution mentioned above. We point out that there doesn't exist a batch collection rule which is generally valid. Instead we have to analyze the model for some given configuration and have to look for the optimal value of the service starting threshold "a".

Figure 5 shows that the optimal choice of "a" is not always as simple as for the case with $c_H = 0$ but becomes more critical for higher variation of service time. The superposition of the diverse dependencies of the waiting time leads to a special appreciation for the choice of the service starting threshold "a" for each set of parameters c_H and ρ . In fig. 6 we show the blocking probability as a function of the traffic intensity for different values of the service starting threshold "a" and coefficient of variation of the service time c_H . As expected, the blocking probability is smaller for the large service starting threshold and becomes higher for larger coefficient of variation of the service time.

Up to now we judged the batch collection rule from the viewpoint of minimizing the waiting time. Clearly, if our main point lies in utilizing the server in an efficient way, the optimization issues are somewhat different. Figure 7 shows that for constant service time and low traffic intensity the average number of customers per start with small service starting threshold is significantly smaller than with larger service starting threshold. In the case of high traffic intensity the service starting threshold has no influence on the average number of customers per start. For the coefficient of variation of service time $c_H = 1$ the service starting threshold has an

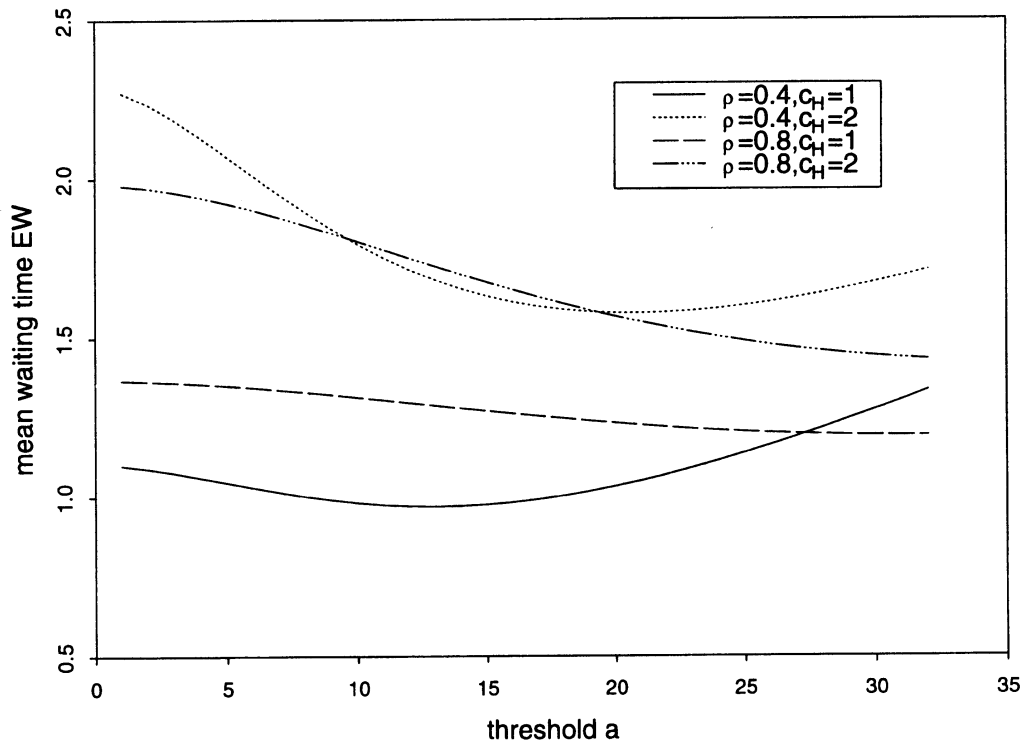


Fig. 5. Threshold dimensioning aspects.

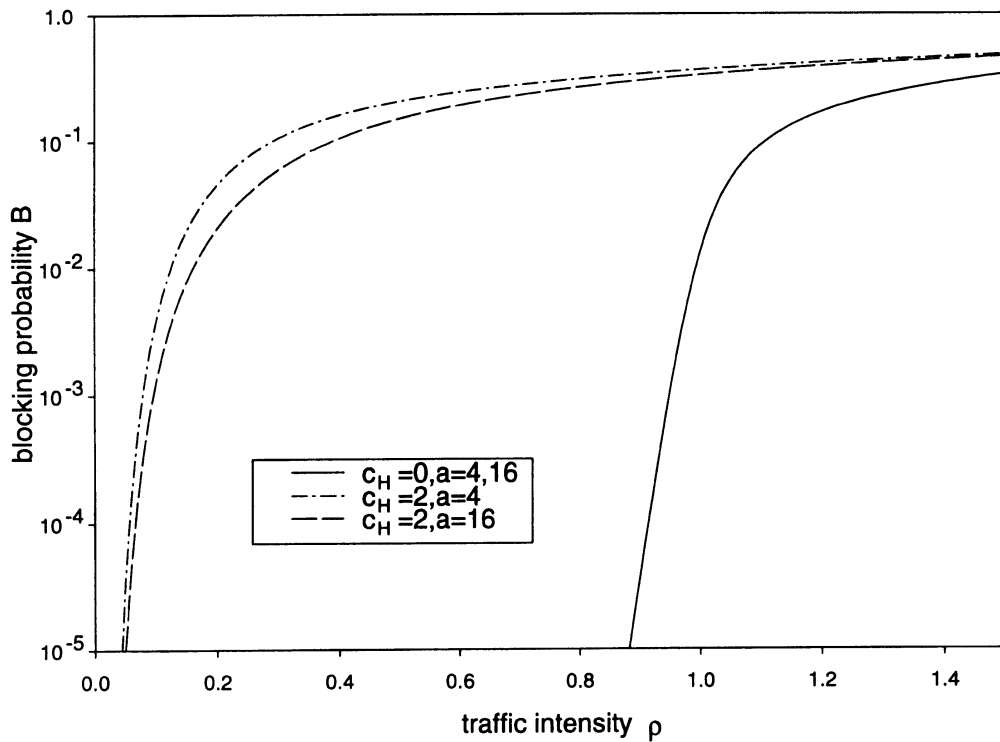


Fig. 6. Blocking probability.

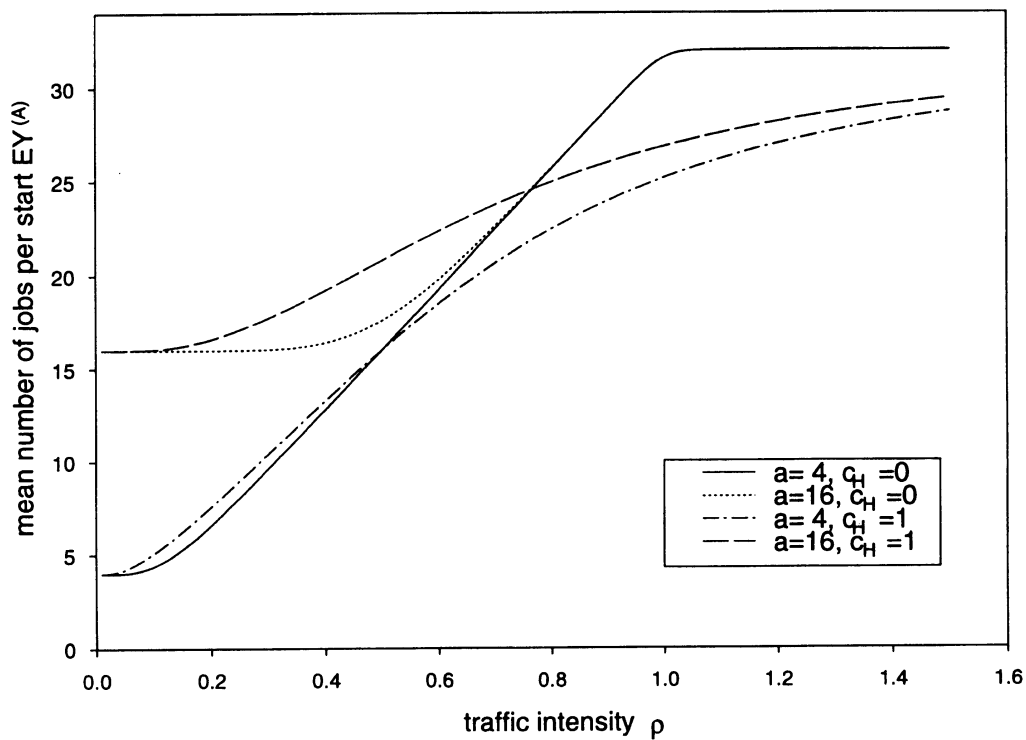


Fig. 7. Server utilization.

influence on the average number of customers per start even when traffic intensity is high, but fortunately the resulting appreciation for the choice of the service starting threshold has the same tendency as under the optimization issue of minimizing the waiting time. Regardless of the value of c_H , in the case of low traffic intensity the two different viewpoints of optimization impose contradictory consequences as far as the choice of the service starting threshold “ a ” is concerned. Depending on the question which viewpoint is more important, individual choice has to be taken into account.

Acknowledgements

The authors would like to thank Heidrun Grob for the stimulating discussions during the course of this work and for her valuable programming efforts. The inspiring cooperations and supports of the “German Manufacturing Technology Center, IBM Germany GmbH” are greatly appreciated.

References

- [1] M.F. Neuts, A general class of bulk queues with Poisson input, *Ann. Math. Stat.* 38 (1967) 759–770.
- [2] M.F. Neuts, Queues solvable without Rouchés, *Oper. Res.* 27 (1979) 767–781.
- [3] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models – An Algorithmic Approach* (Johns Hopkins University Press, Baltimore, 1981).
- [4] M.L. Chaudhry, B.R. Madill and G. Brière, Computational Analysis of steady-state probabilities of $M/G^{a,b}/1$ and related nonbulk queues, *Queueing Systems* 2 (1987) 93–114.
- [5] N.T.J. Bailey, On queueing process with bulk service, *J. Roy. Stat. Soc. B16* (1954) 80–87.
- [6] F. Downton, Waiting time in bulk service queues, *J. Roy. Stat. Soc. B17* (1955) 256–261.
- [7] B.W. Gnedenko and D. König, *Handbuch der Bedienungstheorie II* (Akademie-Verlag Berlin, 1984) pp. 203–204.
- [8] D. Gross and C.M. Harris, *Fundamentals of queueing theory* (Wiley, New York, 1985) pp. 163–170.
- [9] P.J. Kuehn, Approximate analysis of general queueing networks by decomposition, *IEEE Trans. Comm. COM-27* (1979) 113–126.
- [10] P. Tran-Gia and T. Raith, Performance analysis of finite capacity polling systems with nonexhaustive service, *Performance Evaluation* 9 (1988) 1–16.