CWI

BS

Centrum voor Wiskunde en Informatica

REPORTRAPPORT

Multiserver queues with impatient customers

O.J. Boxma, P.R. de Waal

Department of Operations Research, Statistics, and System Theory

Report BS-R9319 November 1993

Multiserver queues with impatient customers

O.J. Boxma, P.R. de Waal

Department of Operations Research, Statistics, and System Theory

Report BS-R9319 November 1993

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

Multiserver Queues with Impatient Customers

O.J. Boxma, P.R. de Waal

CWI

P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

We study multiserver queues in which customers leave when their service is not started before the expiration of a stochastic deadline. Insensitive bounds and sharp approximations for the overflow probability are derived. Assigning costs to servers and to the loss of impatient customers, we also consider the problem of determining the number of servers that minimizes a certain cost function.

AMS Subject Classification (1991): 60K25, 68M20.

Keywords & Phrases: M/G/m queue, customer impatience, overflow probability.

Note: The second author was supported by a fellowship of Koninklijke/Shell Laboratorium, Shell Research B.V., Amsterdam.

1. Introduction

In many service systems, customers leave when their service is not started before a certain deadline expires. Some examples of such service systems with impatient customers are:

(i) telecommunication networks where subscribers give up due to impatience before the requested connection is completely established.

(ii) real-time communication systems, in which the content of a message often loses its importance after a certain amount of time.

(iii) packet-switching communication networks in which the switching nodes have a limited buffer capacity: due to the fixed length of a service (transmission), limited buffer capacity translates into a limitation on the waiting time.

(iv) datacommunication networks with a time-out protocol.

(v) inventory systems with storage of perishable goods.

(vi) database systems, in which a query is withdrawn when it has not been handled before a certain deadline. Furthermore, in a parallel database system where several queries can be handled simultaneously on different processors, a partial result of a query may make other queries obsolete.

(vii) repairable systems subject to wear and breakdown. Components of the system may be repaired preventively, but if preventive maintenance is delayed too long because of limited repair capacity, then a breakdown may occur, necessitating corrective maintenance: a customer loses patience and leaves the system. "Impatience" may also occur because a too long delay causes the violation of safety regulations.

Report BS-R9319
ISSN 0924-0659
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

The last example (which led to the present study) describes a situation where overflowing customers still require service - possibly with a different service time distribution. E.g., in an oil platform a repair crew takes care of preventive maintenance jobs; but when a breakdown occurs, a special repair crew may have to be flown in to perform corrective maintenance. Such corrective maintenance may be both considerably longer and considerably more expensive than preventive maintenance. This raises the issue of determining the number of repairmen for preventive maintenance that minimizes total cost. To solve that problem, one needs to know the breakdown probability as a function of the number of repairmen - or more generally, the probability that a customer in a multiserver queue loses its patience.
The present paper is devoted to these issues. It studies the loss probability in a multiserver queueing model with impatient customers. For that loss probability we discuss some exact results (Section 2), insensitive upper and lower bounds (Section 3), and various approximations (Section 4). Extensive tests of these approximations reveal a near-insensitivity of the overflow probability with respect to the service time distribution, and - apart from a small traffic region - a rather weak sensitivity with respect to the patience time distribution. In Section 5 the following cost minimization problem is studied: determine the number of servers, m, that minimizes the cost function

$$c_1 m + c_2 \pi(m), \tag{1.1}$$

where c_1 is the cost per time unit involved in having m regular servers, $\pi(m)$ is the overflow probability, and c_2 is the cost per time unit involved in hiring extra service capacity. Section 6 contains conclusions and some suggestions for further research. In the remainder of this introduction we present a model description and a short review of the literature on multiserver queues with impatience.

Model description
Customers arrive at a service facility according to a Poisson process with rate λ. Service requests of successive customers are independent, identically distributed (i.i.d.) stochastic variables $\mathbf{S}_n, n = 1, 2, ...$ with distribution $B(\cdot)$, with first moment β. The service facility has m servers. The service discipline is First-Come-First-Served (FCFS). Customers have limited patience. If the service of a customer is not started before its patience runs out, then it leaves the system (as far as the loss probability is concerned, an equivalent assumption is that customers whose deadline eventually expires are rejected immediately upon arrival). The patience times of successive customers are i.i.d. stochastic variables $\mathbf{G}_n$, $n = 1, 2, \ldots$, with distribution $F(\cdot)$, with first moment γ. In the sequel, $\bar{F}(t) = 1 - F(t)$. The arrival, service and patience processes are independent stochastic processes. This M/G/m queue with general patience time distribution will be denoted as M/G/m+G.

Literature Review
According to Stanford[16], the reneging phenomenon appears for the first time in the queueing literature in Palm[12], in the context of impatient telephone switchboard customers. An early study of a multiserver queue with impatience is Barrer[4]; he determines the loss probability in the M/M/m+D queue. Gnedenko and Kovalenko[8], Section 1.5, study the M/M/m+D as well as the M/M/m+M model. Haugen and Skogan[9] and Baccelli and Hebuterne[2] analyze the M/M/m+G queue. The first paper presents an approximate analysis, replacing a general patience distribution by a two-point discrete distribution. The latter paper derives a set of

equations for the virtual waiting time distribution and the probability of having $0, 1, \ldots, m-1$ customers; for M/M/m+D and M/M/m+M, this leads to explicit expressions for the loss probability. Bhattacharya and Ephremides[5] show that the number of successful departures and the number of customers lost over a time interval are (stochastically) monotone functions of the arrival, service and deadline processes. An admission control problem in an impatient customer queueing model for a telephone switch is discussed in [19, Chapter 4]. An extension of that model, with an additional arrival stream, is presented in [6]. Finally we mention some fundamental studies concerning *single* server queues with impatience: both Stanford[16] and Baccelli et al.[3] study the GI/G/1+G queue, using an approach based on regenerative processes and deriving stability conditions and results for actual and virtual waiting times; cf. also Stanford[17].

2 Exact results for the overflow probability

Most of the exact results for the M/G/m+G queue have been derived for the case of exponential services (even for the ordinary M/G/m queue hardly any exact results are known). Baccelli and Hebuterne[2] consider the Markov process $\{(\mathbf{N}(t), \eta(t)), t \geq 0\}$ for the M/M/m+G queue; here
$\mathbf{N}(t) = n$ when the number of customers at time t equals n and $0 \leq n \leq m-1$;
$\mathbf{N}(t) = L$ when the number of customers at time t exceeds $m-1$;
$\eta(t)$ is the virtual offered waiting time, i.e. the time that a customer with infinite patience would have to wait for service. It is strictly positive when $\mathbf{N}(t) = L$, and it equals zero otherwise.
Define in the steady-state situation, which exists iff $\lambda \bar{F}(\infty) < m/\beta$ (see [2]):

$$P_j := \lim_{t\to\infty} \Pr\{\mathbf{N}(t) = j, \eta(t) = 0\}, \quad j = 0, \ldots, m-1, \tag{2.1}$$

$$v(x) := \lim_{t\to\infty} \lim_{dx\to 0} \Pr\{\mathbf{N}(t) = L, x < \eta(t) \leq x + dx\}/dx. \tag{2.2}$$

From the Chapman-Kolmogorov equations for $P_j, j = 0, \ldots, m-1$ and $v(x)$ it follows, with offered traffic load $\rho := \lambda\beta$:

$$P_j = \frac{\rho^j}{j!} P_0, \quad j = 0, \ldots, m-1, \tag{2.3}$$

$$v(0) = \lambda P_{m-1}, \tag{2.4}$$

$$v(x) = v(0) exp\left[\lambda \int_{u=0}^{x} \bar{F}(u)du - mx/\beta\right], \quad x > 0. \tag{2.5}$$

The normalizing condition $\sum_{j=0}^{m-1} P_j + \int_0^\infty v(x)dx = 1$ yields:

$$P_0 = \left[1 + \rho + \frac{\rho^2}{2!} + \ldots + \frac{\rho^{m-1}}{(m-1)!}(1 + \lambda J)\right]^{-1}, \tag{2.6}$$

with

$$J := \int_0^\infty exp\left[\lambda \int_0^x \bar{F}(u)du - mx/\beta\right] dx. \tag{2.7}$$

The overflow probability π is given by

$$\pi = \int_0^\infty F(x)v(x)dx; \tag{2.8}$$

hence (cf. [2], formula (5.9)),

$$\pi = (1 - \frac{m}{\rho})(1 - \sum_0^{m-1} P_j) + P_{m-1}. \tag{2.9}$$

For M/M/m+D this yields:

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{\rho^k}{k!} + \frac{\rho^m}{(\rho - m)m!}\{\rho e^{(\lambda - m/\beta)\gamma} - m\}\right]^{-1}, \tag{2.10}$$

$$\pi = P_0 \frac{\rho^m}{m!} exp[(\lambda - m/\beta)\gamma]. \tag{2.11}$$

For M/M/m+M, with $\alpha := \beta/m\gamma$:

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{\rho^k}{k!} + \frac{\rho^m}{m!}(1 + \frac{\rho/m}{1+\alpha} + \frac{(\rho/m)^2}{(1+\alpha)(1+2\alpha)} + \ldots)\right]^{-1}, \tag{2.12}$$

$$\pi = P_0 \frac{\rho^{m-1}}{(m-1)!}\left[1 + (\rho/m - 1)[\frac{\rho/m}{1+\alpha} + \frac{(\rho/m)^2}{(1+\alpha)(1+2\alpha)} + \ldots]\right]. \tag{2.13}$$

The following observation leads to a simpler representation for M/M/m+M. Assume that customers whose deadline eventually expires are rejected immediately upon arrival. As far as the loss probability is concerned, there is no difference between discarding impatient customers rightaway or only at the expiration of their deadline; they are not served anyway, so they do not influence other customers. The queue length process in the M/M/m+M queue with the above-indicated immediate rejection is a birth-and-death process with death rate j/β $(j < m)$ respectively m/β $(j \geq m)$ and arrival rate λ $(j < m)$ respectively λq_j $(j \geq m)$, with q_j the probability that an arriving customer is accepted when he meets j customers in the system. Clearly $q_0 = \ldots = q_{m-1} = 1$, and $q_j = q_{j-1}/(1+\alpha)$ for $j = m, m+1, \ldots$; hence $q_{m+i} = 1/(1+\alpha)^{i+1}$, $i = 0, 1, \ldots$. The probability $p(i)$ of having i customers in this modified M/M/m+M queue is:

$$p(i) = p(0)\frac{\rho^i}{i!}, \quad i = 0, \ldots, m, \tag{2.14}$$

$$p(m+i) = p(m)\frac{(\lambda\gamma\alpha)^i}{(1+\alpha)^{m(i)}}, \quad i = 0, 1, \ldots, \tag{2.15}$$

with $m(i) := i(i+1)/2$. The overflow probability readily follows:

$$\begin{aligned}\pi &= \sum_{i=0}^{\infty} p(i)(1-q_i) = 1 - \sum_{i=0}^{m-1} p(i) - \sum_{i=0}^{\infty} p(m+i)q_{m+i} \\ &= 1 - p(0)\left[\sum_{i=0}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{m!}\sum_{i=0}^{\infty} \frac{(\lambda\gamma\alpha)^i}{(1+\alpha)^{m(i+1)}}\right]. \end{aligned} \tag{2.16}$$

The infinite sum in the last expression converges faster than the sum in (2.13): the ratio of the $(i+1)$th and the ith term equals $\lambda\gamma\alpha/(1+\alpha)^{i+1}$, which is monotonically decreasing in i.

3 Bounds for the overflow probability

In the previous section we have seen that, apart from the M/M/m+G queue, few exact results are known for multiserver queues with impatient customers. In the present section we shall discuss some - insensitive - bounds for the overflow probability in such multiserver queues. Bhattacharya and Ephremides[5] consider a G/G/m+G queue. They show that the number of lost customers over any time interval decreases stochastically when the patience time becomes stochastically larger (i.e., when the patience time distribution $F(z)$ is replaced by $\tilde{F}(z)$ with $\tilde{F}(z) \leq F(z)$ for all $z \geq 0$). We use their result to derive lower and upper bounds for the overflow probability π in M/G/m+G. *Zero patience* gives

$$\pi \leq \pi_{erl} = \frac{\frac{\rho^m}{m!}}{\sum_{j=0}^{m} \frac{\rho^j}{j!}}, \tag{3.1}$$

with π_{erl} the loss probability in the Erlang loss system M/G/m/0.
Infinite patience gives

$$\pi \geq \pi_{inf} = \max[0, 1 - \frac{m}{\rho}]. \tag{3.2}$$

Indeed, in the case of infinite patience, i.e., an M/G/m system, only a fraction m/ρ of the customers is served when $\rho > m$. Heyman[10] rigorously proves that this same lower bound holds for the overflow probability in a G/G/m/r queue (with r waiting positions), extending an M/G/m/r result of Sobel[14]. In fact Sobel also obtains an upper bound for the overflow probability in M/G/m/r: $\rho/(m+\rho)$. We can easily sharpen this bound by observing that the overflow probability in M/G/m/r is decreasing in r, a result proven by Sonderman[15, Section 3]. Hence an upper bound is obtained by taking $r = 0$, yielding π_{erl}. This is a better lower bound than $\rho/(m+\rho)$, because

$$\frac{\rho}{m+\rho} = \frac{\frac{\rho^m}{m!}}{\frac{\rho^{m-1}}{(m-1)!} + \frac{\rho^m}{m!}} \geq \pi_{erl}. \tag{3.3}$$

Remark
It is interesting to observe that the upper and lower bounds π_{erl} and π_{inf}

1. hold both for M/G/m+G and for M/G/m/r;

2. are insensitive for the patience time distribution and for r;

3. are insensitive for the service time distribution;

4. almost coincide for a large range of ρ values.

Figure 1 shows the overflow probability π for the M/M/m+D queue with $\lambda = 40$, $\beta = 1$ and $\gamma = 0, 0.1, 0.5, 1.0, \infty$. The only m values for which the lower and upper bounds differ by more than 0.05 are $29 \leq m \leq 47$. All figures and tables are placed at the end of the paper.

Let us compare the two systems M/G/m+G and M/G/m/r in some more detail, assuming that arrival intensities and service time distributions are the same in both models, and that an arriving customer in M/G/m+G who will not eventually be served is rejected immediately. Keeping in mind the above-mentioned insensitivity, let us first compare M/M/m+G and M/M/m/r. The queue length process in the latter system is a birth-and-death process; the overflow probability, as given in several textbooks, reads as follows:

$$\pi_{M/M/m/r} = \frac{\frac{\rho^m}{m!}(\rho/m)^r}{\sum_{k=0}^{m-1}\frac{\rho^k}{k!} + \frac{\rho^m}{(\rho-m)m!}\{\rho(\rho/m)^r - m\}}. \tag{3.4}$$

A comparison with (2.11) for M/M/m+D reveals a remarkable structural similarity:

$$\pi_{M/M/m+D} = \pi_{M/M/m/r}, \tag{3.5}$$

when $\rho/m = e^{(\rho-m)\gamma/\beta}$, i.e., when

$$r = \frac{m}{\beta}\gamma\frac{\rho/m - 1}{ln(\rho/m)}. \tag{3.6}$$

In fact even the distributions of the numbers of busy servers coincide for this choice of r (of course r should be integer for M/M/m/r). It should be noted that the two systems do indeed behave very similarly when r is approximately $\gamma m/\beta$. In M/M/m/r there is a sharp distinction between meeting at least r waiting customers upon arrival (rejection) and meeting less than r waiting customers (acceptance). But when an arriving customer in the M/M/m+D system (with immediate rejection) meets $r+s$ waiting customers, than the time until its service could start is the sum of $r+s$ independent exponentially distributed stochastic variables with means β/m, so it has an Erlang-$(r+s)$ distribution. When r is not too small and s is positive (not too small), the almost deterministic character of Erlang-$(r+s)$ shows that the patience time $\gamma = r\beta/m$ will most likely be exceeded. When s is negative, the reverse statement holds.

For the case of a *general* service time distribution and deterministic patience, one might approximate $\pi_{M/G/m+D}$ by $\pi_{M/G/m/r}$ with r given by (3.6). $\pi_{M/G/m/r}$ has been extensively tabulated in the book [13]. Those authors have represented service times by Erlang or hyperexponential distributions with the right mean and coefficient of variation, and have subsequently analysed the M/E_k/m/r and M/H_2/m/r queues exactly. If r as given by (3.6) is non-integer, one could use the table results for the two surrounding integers, and interpolate linearly. In the next section we shall investigate some approximation possibilities for the overflow probability in more detail.

Remark
Whitt [20] has developed a heavy-traffic approximation for $\pi_{GI/G/m/r}$ (ρ high but less than m). Approximations for the Erlang loss probability $\pi_{M/G/m/0}$ have also been developed (see, e.g., [11]), usually distinguishing three traffic regions: $\rho \leq m - C\sqrt{m}$, $\rho \geq m + C\sqrt{m}$, with C some positive constant and the (most interesting) intermediate region. Such a traffic region distinction is also natural in our model, cf. Figure 1.

4 Approximations for the overflow probability

In this section we present and test three approximations for the overflow probability π in the M/G/m+G queue.

a. *A simple insensitive approximation*
In Section 2 an exact analysis of the M/M/m+G queue has been presented. The overflow probability π can be determined from (2.9) using (2.3), (2.6) and (2.7). A simple explicit representation is given for the M/M/m+D case (formula (2.11)), and a series representation for the M/M/m+M case (formula (2.13) or (2.16)). In view of the insensitivity of the rather tight bounds of Section 3 and the service time insensitivity of the queue length process in another extreme case, the M/G/∞+G queue (= M/G/∞), we propose to approximate π in the M/G/m+G queue by the overflow probability in the M/M/m+G queue:

$$\pi_A = \pi_{M/M/m+G}. \tag{4.1}$$

And unless the patience time distribution has a large coefficient of variation and m is close to ρ, we recommend to use the even simpler approximation

$$\pi_{A'} = \pi_{M/M/m+D}, \tag{4.2}$$

the overflow probability in the M/M/m+D queue with the same parameters β and γ.

b. *General patience distribution: a weighted integral*
If patience is not deterministic, it might be worthwhile to use the more sophisticated approximation

$$\pi_B = \int_{x=0}^{\infty} \pi_{M/M/m+D(x)} dF(x), \tag{4.3}$$

where $D(x)$ denotes deterministic patience x. Note that $\pi_{M/M/m+D(x)}$ is monotonously decreasing in x, cf. [5], and for $x = 0$ ($x = \infty$) π_B equals the upper (lower) bound of Section 3. This gives some support to the approximation. π_B can easily be evaluated for a discrete patience time distribution. For exponentially distributed patience the evaluation becomes somewhat more complicated; e.g., for $\rho = m$ (2.10), (2.11) and (4.3) lead to a so-called exponential integral (cf. Abramowitz and Stegun [1]).

c. *General patience distribution: a weighted sum*
In practice, patience behaviour will usually be closer to deterministic than to exponential. E.g., in [2] an Erlang-3 patience time distribution is claimed to fit well measurement data obtained for subscriber behaviour in several PABX's in a telephone network. When the coefficient of variation c_f of the patience time distribution (standard deviation divided by

mean) is indeed between 0 and 1, it seems quite natural to approximate π by interpolating between $\pi_{M/G/m+D}$ and $\pi_{M/G/m+M}$. Our numerical experiments suggest that c_f (and in fact also ${c_f}^2$) is a suitable weighing factor:

$$\pi_C = c_f \pi_{M/G/m+M} + (1 - c_f)\pi_{M/G/m+D}, \quad for \quad 0 \leq c_f \leq 1. \tag{4.4}$$

Similar weighted sum approximations, usually with weight factor ${c_f}^2$, have often been used in multiserver queues; cf. Tijms[18, Chapter 4]. In (4.4) $\pi_{M/G/m+M}$ and $\pi_{M/G/m+D}$ are still undetermined. For most purposes it would be sufficient to assume insensitivity w.r.t. the service time distribution, replacing $\pi_{M/G/m+.}$ by $\pi_{M/M/m+.}$. More accuracy may be obtained by employing another idea that has been used before in the context of (mean waiting time approximations for) multiserver queues:

$$\pi_{M/G/m+G} \approx \pi_{M/M/m+G} \frac{\pi_{M/G/1_m+G}}{\pi_{M/M/1_m+G}}. \tag{4.5}$$

Here 1_m indicates that the speed of the single server is m-fold increased. The idea behind this approximation is that

1. it is exact for single server queues;
2. it is exact for exponentially distributed service times;
3. the influence of the service time distribution on the overflow probability should be roughly the same as in an m times faster single server queue (note that in the case of exponential service, both when all m servers in M/M/m are occupied and the single fast server in M/M/1_m is occupied, the next departure occurs after an $exp(m/\beta)$ distributed time);
4. it allows one to use known results for multi- and single server queues with impatience. In [2], [3] Baccelli et al. present an exact analysis of M/G/1+G (π is obtained by solving a Volterra integral equation). The special case M/G/1+M is particularly neat. Tijms[18, Section 4.3.3] obtains simple exact results for $\pi_{M/D/1+D}$ and $\pi_{M/M/1+D}$, and a nice simple approximation for $\pi_{M/G/1+D}$.

Numerical Results

We have seen that the upper and lower bounds in Section 3 are farthest apart in the region around $\rho = m$. In the numerical tests, where we have to choose from a large set of possible distributions and parameter combinations, we therefore restrict ourselves to the region around $\rho = m$. Without loss of generality we take $\beta = 1$, so $\rho = \lambda$.

The approximations are compared to simulation results and a numerical approximation of the overflow probability. The latter, which can be used only for exponential service times, is computed by replacing J in the normalisation constant (2.6)–(2.7) by

$$\int_0^K \exp\left[\lambda \int_0^x \bar{F}(u)du - mx/\beta\right] dx + \int_K^\infty \exp\left[\lambda\gamma - mx/\beta\right] dx \tag{4.6}$$

where K has to be chosen sufficiently large (assuming that $\bar{F}(\infty) = 0$). The first term in (4.6) is integrated numerically.

The first example deals with the approximation a. We have ran a number of simulations of an M/G/m+G queue, in which the deadline distribution was varied between deterministic,

Erlang-2 and exponential, while the service time distribution was varied in such a way that its squared coefficient of variation (c_β^2) ranges from 0 to 2. The simulations were carried out with the AT&T Q+ simulation package and for the service time distributions we used the so-called P2 distribution: deterministic for $c_\beta^2 = 0$, uniform for $c_\beta^2 = 0.25$, gamma for $c_\beta^2 = 0.50$, 0.75, and Hyper-2 exponential for $c_\beta^2 > 1$. The results are depicted in Tables 1–3. For each model we used a run of 40 subsimulations of 16000 arrivals each to get the indicated accuracy. In these tables the values for $c_\beta^2 = 1$ correspond to exponentially distributed service times and for this case the overflow probability is approximated numerically as in (4.6). From these figures we may conclude that $\pi_{M/G/m+G}$ can be approximated quite reasonably by replacing the service time distributions with exponential distributions. Replacing the deadline distributions with deterministic ones on the other hand introduces a substantial deviation in the region around $\rho = m$.

From other experiments, not represented here for lack of space, using the same numerical approximation (4.6), it appeared that the overflow probability is insensitive for the deadline distribution too, provided that the deadline distribution has a coefficient of variation larger than 1; one might replace π by $\pi_{M/G/m+M}$ in such cases.

The next example illustrates the accuracy of approximations b and c. The model that is considered is the M/M/m+D2 queue, where the D2 indicates a discrete distribution on two points. We have chosen four of these distributions such that the squared coefficient of variation ranges from 0.25 to 1. The exact parameters of each distribution are given in Table 4. The mean deadline is varied over $\gamma = 0.1$ (Table 5), $\gamma = 0.5$ (Table 6), $\gamma = 1.0$ (Table 7). It appears that approximation c is considerably better than approximation b, except for case D2-d. Note that for D2-d the corresponding c_β^2 equals 1, so for this distribution approximation c amounts to replacing the D2 distribution by an exponential one. Apparently π is here even quite sensitive to the third moment of the patience time distribution.

From other experiments involving D2 distributions we have gotten the impression that good approximations are difficult to obtain when either of the two points of the D2 distribution becomes close to zero or to infinite patience (in the sense that $\pi_{M/M/m+D}$ on this one point would be close to the upper or lower bound). Consider for instance D2-d for $\gamma = 1.0$, when the distribution is concentrated on the points 0.666666 and 4.

The final example presents approximation method c for an M/M/m+E2 queue, i.e. Erlang-2 distributed deadlines. The exact and approximate results for $\gamma = 0.1$, 0.5, 1.0, are depicted in Table 8.

Conclusions

π_A is a very good approximation, due to the near-insensitivity of the overflow probability for the service time distribution. The worst results occur when $\gamma \approx \rho/m$ and small coefficients of c_β^2 (smaller than 0.5). $\pi_{A'}$ is very simple, but only accurate when patience is close to deterministic.

When taking the patience time distribution into account, π_C is both simpler and more accurate than π_B. For the D2-a, D2–b and E2 patience time distributions the error in π_C was at most 0.005.

5 Optimization problem

Consider the M/G/m+G queue with the following cost structure added. The operational costs for a server in the queue are d_1 per time unit. For any customer that abandons the queue because of impatience, an external server is hired, for the duration of the service of that customer only. The cost per time unit for such a server is denoted as d_2. An elementary application of Little's law shows that if the internal queue has m servers, then the longrun average cost $g(m)$ is given by

$$g(m) = d_1 * m + \lambda * \beta^* * d_2 * \pi(m), \tag{5.1}$$

where β^* is the mean service time for customers served by an external server, and $\pi(m)$ denotes the overflow probability as a function of m. We are interested in the value of m that minimizes (5.1).

In this section we investigate the sensitivity of the optimal value of m and the sensitivity of $g(m)$. We illustrate this for the model M/M/m+D, where we take $\beta^* = 1$. From the experiments it appears that $g(m)$ is a convex function of m for all the parameter settings that we tried, although we have not been able to prove this (for M/M/m/r it is known, cf. p. 148 of [7], that $\pi_{M/M/m/r}$ is convex in m and in r; the equivalence between M/M/m/r and M/M/m+D, as noted in Section 3, almost (but not quite) yields the above-mentioned convexity result). As a result there is always a unique m that minimizes $g(m)$.

In Figures 2, 6 and 9 we present the plots of $g(m)$ in the M/M/m+D queue for three values of $d_1 : d_2$ (or equivalently, and more importantly, $\beta d_1 : \beta^* d_2$). We have not included examples with $d_2 \leq d_1$, since for those parameters the optimal m is always 0. In the plots it is remarkable that although the value of m for which $g(m)$ is minimal shows sensitivity with respect to the value of γ, the shape of $g(m)$ is very flat around the minimum. This means that if we deviate from the optimal m^*, then the time average costs may still be close to the optimal $g(m^*)$.

The sensitivity of the average costs on the type of the distribution functions is reported in Figures 3, 4, 5, 7 and 8. In Figure 3 $g(m)$ is plotted for deterministic, Erlang-2 and exponential deadlines. The same observation as for M/M/m+D can be made here: the optimal value m^* shows sensitivity with respect to both the type and the mean of the deadline distribution, but $g(m^*)$ is almost insensitive. Note for instance that m^* ranges from 34 for exponentially distributed deadlines to 37 for deterministic deadlines, but throughout this region the value of $g(m)$ does not deviate more than 2% from $g(m^*)$. This insensitivity suggests that, for M/G/m+G with $c_f \leq 1$, one can accurately solve the minimization problem (5.1) by using the simple approximation $\pi_{A'} = \pi_{M/M/m+D}$.

In Figures 4 and 5 the time average costs are depicted for four different service time distributions and two deadline distributions. In these plots the H_2 distribution is a hyperexponential distribution of order 2 with $c_\beta^2 = 1.5$. All the values of the overflow probability that are needed to compute $g(m)$ were obtained from simulation, except for the models with exponentially distributed service times. These were computed exactly. Again we see that m^* is sensitive with respect to the type of the distribution functions, and that $g(m^*)$ is not sensitive. The same remarks apply to Figures 7 and 8, which shows the same models with a different $d_1 : d_2$ ratio.

Note that we can interpret the extreme cases of $\gamma = 0$ and $\gamma = \infty$ as two degenerate models where no information about the deadline is available. From the monotonicity of π (see [5])

and the convexity of g we can conclude, however, that the optimal m^* must be between m_0^* and m_∞^*, the optimal values for $\gamma = 0$ and $\gamma = \infty$, respectively. If deadlines are tight, then m should be chosen close to m_0^*, while if deadlines are loose, then m around m_∞^* seems more appropriate. Note that (cf. (3.2)), $m_\infty^* = 0$ if $d_1\beta > d_2\beta^*$ and $m_\infty^* = \rho$ if $d_1\beta < d_2\beta^*$.

6 Conclusions and suggestions for further research

Sharp lower and upper bounds for the overflow probability π in the M/G/m+G queue have been derived, and a link between the M/G/m+G and M/G/m/r queue has been established. Several approximations for π are suggested and tested, in the only region where the derived lower and upper bounds differ considerably: $m \approx \rho$. The main conclusions are that π is nearly insensitive for the service time distribution, and that, when $0 \leq c_f \leq 1$, π can be very accurately approximated by interpolating between $\pi_{M/M/m+D}$ find and $\pi_{M/M/m+M}$.
An interesting performance measure in the multiserver model with impatience is the mean waiting time of accepted customers. Approximations for this mean might be obtained along the lines of established mean waiting time approximations for the M/G/m queue, cf. [18].
Another subject that we have not covered is the 'shortest deadline first' service discipline. Bhattacharya and Ephremides[5] have shown (for ./M/1+G) that for that discipline, too, the number of lost customers over any time interval decreases stochastically when the patience time becomes stochastically larger. Consideration of this discipline might be especially interesting in connection to the optimization problem of Section 5.
An interesting optimization problem also arises when a group of m servers should be allocated among N service stations (m_i servers to station i) in such a way that (with an obvious notation, cf. (5.1))

$$\sum_{i=1}^{N} \lambda_i * \beta_i^* * d_{2i} * \pi_i(m_i) \tag{6.1}$$

is minimized. This is similar to a problem mentioned in Buzacott and Shanthikumar [7, p. 149], and can be easily solved numerically using the (assumed) convexity of $\pi_i(.)$.

References

[1] Abramowitz, M., Stegun, I.A. (1970). *Handbook of mathematical functions* (Dover Publications, New York).

[2] Baccelli, F., Hebuterne, G. (1981). On queues with impatient customers. In: *Performance '81*, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 159-179.

[3] Baccelli, F., Boyer, P., Hebuterne, G. (1984). Single-server queues with impatient customers. *Adv. Appl. Probab.* **16**, 887-905.

[4] Barrer, D.Y. (1957). Queuing with impatient customers and ordered service. *Oper. Res.* **5**, 650-656.

[5] Bhattacharya, P.P., Ephremides, A. (1991). Stochastic monotonicity properties of multiserver queues with impatient customers. *Adv. Appl. Probab.* **28**, 673-682.

[6] Blanc, J.P.., De Waal, P.R., Nain, Ph., Towsley, D. (1992). Optimal control of admission to a multiserver queue with two arrival streams. *IEEE Trans. Autom. Control* **37**, 785-797.

[7] Buzacott, J.A., Shanthikumar, J.G. (1992). Design of manufacturing systems using queueing models. *Queueing Systems* **12**, 135-214.

[8] Gnedenko, B.V., Kovalenko, I.N. (1968). *Introduction to Queueing Theory* (Israel Program for Scientific Translations, Jerusalem).

[9] Haugen, R.B., Skogan, E. (1980). Queueing systems with stochastic time out. *IEEE Trans. Commun.* **COM-28**, 1984-1989.

[10] Heyman, D.P. (1980). Comments on a queueing inequality. *Management Science* **26**, 956-959.

[11] Newell, G.F. (1984). *The M/M/∞ Service System with Ranked Servers in Heavy Traffic* (Springer, Berlin).

[12] Palm, C. (1937). Etude des delais d'attente. *Ericsson Technics* **5**, 37-56.

[13] Seelen, L.P., Tijms, H.C., Van Hoorn, M.H. (1985). *Tables for Multi-server Queues* (North-Holland Publ. Cy., Amsterdam).

[14] Sobel, M.J. (1980). Simple inequalities for multiserver queues. *Management Science* **26**, 951-956.

[15] Sonderman, D. (1979). Comparing multi-server queues with finite waiting rooms, I: same number of servers. *Adv. Appl. Probab.* **11**, 439-447.

[16] Stanford, R.E. (1979). Reneging phenomena in single channel queues. *Math. of Oper. Res.* **4**, 162-178.

[17] Stanford, R.E. (1990). On queues with impatience. *Adv. Appl. Probab.* **22**, 768-769.

[18] Tijms, H.C. (1986). *Stochastic Modelling and Analysis* (Wiley, New York).

[19] De Waal, P.R. (1990). *Overload Control of Telephone Exchanges* (Ph.D. Thesis, Tilburg University, Tilburg).

[20] Whitt, W. (1984). Heavy-traffic approximations for service systems with blocking. *AT&T Bell Labs. Techn. J.* **63**, 689-708.

Overflow probability in $M/G/m+D$			
c_β^2	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$
.00	.071 ± .001	.021 ± .001	.013 ± .001
.25	.075 ± .002	.028 ± .002	.014 ± .002
.50	.077 ± .002	.031 ± .001	.016 ± .002
.75	.078 ± .002	.033 ± .002	.019 ± .002
1.00	.079	.035	.021
1.25	.078 ± .002	.037 ± .002	.022 ± .002
1.50	.081 ± .002	.038 ± .003	.024 ± .002
1.75	.078 ± .003	.038 ± .002	.024 ± .002
2.00	.081 ± .002	.039 ± .003	.027 ± .003

TABLE 1. Overflow probability and 5% confidence interval in $M/G/m+D$ ($\lambda = 40$, $\beta = 1$, $m = 40$)

Overflow probability in $M/G/m+E_2$			
c_β^2	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$
.00	.085 ± .001	.050 ± .001	.037 ± .001
.25	.088 ± .002	.054 ± .001	.041 ± .001
.50	.088 ± .002	.058 ± .002	.043 ± .002
.75	.089 ± .002	.057 ± .002	.045 ± .002
1.00	.089	.061	.047
1.25	.089 ± .002	.063 ± .002	.048 ± .002
1.50	.089 ± .002	.065 ± .002	.049 ± .002
1.75	.092 ± .003	.063 ± .002	.053 ± .002
2.00	.091 ± .002	.063 ± .002	.051 ± .002

TABLE 2. Overflow probability and 5% confidence interval in $M/G/m+E_2$ ($\lambda = 40$, $\beta = 1$, $m = 40$)

Overflow probability in $M/G/m+M$			
c_β^2	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$
.00	.091 ± .001	.064 ± .001	.054 ± .001
.25	.094 ± .002	.069 ± .002	.057 ± .001
.50	.094 ± .002	.071 ± .002	.060 ± .002
.75	.095 ± .002	.071 ± .002	.062 ± .002
1.00	.095	.074	.063
1.25	.095 ± .002	.074 ± .002	.062 ± .002
1.50	.096 ± .002	.074 ± .002	.065 ± .002
1.75	.096 ± .002	.075 ± .003	.065 ± .002
2.00	.095 ± .002	.075 ± .002	.068 ± .002

TABLE 3. Overflow probability and 5% confidence interval in $M/G/m+M$ ($\lambda = 40$, $\beta = 1$, $m = 40$)

Parameters of D_2 distributions				
	δ_1	δ_2	$P(D=\delta_1)$	c_β^2
D_2-a	$0.5*\gamma$	$1.5*\gamma$	0.5	0.25
D_2-b	$0.2909*\gamma$	$1.7091*\gamma$	0.5	0.50
D_2-c	$0.1*\gamma$	$1.9*\gamma$	0.5	0.81
D_2-d	$0.666666*\gamma$	$4.0*\gamma$	0.9	1.00

TABLE 4. Parameters of D_2 distributions

Overflow probability in $M/M/m+D_2$												
	D_2-a			D_2-b			D_2-c			D_2-d		
m	exact	π_B	π_C	exact	π_B	π_C	exact	π_B	π_C	exact	π_B	π_C
30	.271	.269	.271	.275	.271	.273	.279	.275	.275	.272	.271	.276
31	.249	.247	.250	.253	.249	.252	.259	.253	.254	.251	.249	.255
32	.228	.225	.229	.233	.228	.231	.238	.232	.233	.230	.228	.235
33	.208	.204	.208	.213	.207	.211	.219	.210	.213	.209	.207	.215
34	.188	.184	.188	.193	.187	.191	.199	.190	.194	.189	.187	.195
35	.169	.164	.169	.174	.167	.172	.181	.170	.175	.170	.167	.176
36	.150	.146	.151	.156	.148	.154	.163	.152	.157	.152	.148	.158
37	.133	.128	.133	.139	.130	.137	.146	.134	.140	.134	.131	.141
38	.117	.111	.117	.123	.114	.120	.130	.117	.123	.118	.114	.125
39	.101	.096	.102	.108	.098	.105	.115	.101	.108	.102	.098	.110
40	.087	.081	.087	.093	.084	.091	.100	.086	.094	.088	.084	.095
41	.074	.068	.074	.080	.070	.077	.087	.073	.080	.075	.071	.082
42	.062	.057	.062	.068	.059	.065	.074	.061	.068	.063	.059	.070
43	.052	.046	.052	.057	.048	.055	.063	.051	.057	.052	.049	.059
44	.042	.037	.042	.047	.039	.045	.053	.042	.047	.042	.040	.049
45	.034	.030	.034	.039	.031	.036	.044	.034	.039	.034	.032	.040
46	.027	.023	.027	.031	.025	.029	.036	.027	.031	.027	.025	.032
47	.021	.018	.021	.025	.019	.023	.029	.021	.025	.021	.020	.026
48	.016	.014	.016	.019	.015	.018	.023	.017	.019	.016	.015	.020
49	.012	.010	.012	.015	.011	.014	.018	.013	.015	.012	.011	.016
50	.009	.007	.009	.011	.008	.010	.014	.010	.011	.009	.008	.012

TABLE 5. Approximations B and C for the overflow probability in $M/M/m+D_2$ ($\lambda=40$, $\beta=1$, $\gamma=0.1$)

Overflow probability in $M/M/m + D_2$												
	D_2-a			D_2-b			D_2-c			D_2-d		
m	exact	π_B	π_C	exact	π_B	π_C	exact	π_B	π_C	exact	π_B	π_C
30	.252	.252	.254	.257	.255	.256	.268	.264	.258	.251	.251	.258
31	.228	.227	.231	.234	.231	.233	.246	.241	.235	.227	.227	.236
32	.205	.203	.207	.211	.208	.210	.226	.218	.213	.203	.203	.214
33	.182	.179	.185	.189	.184	.188	.205	.195	.191	.180	.179	.193
34	.159	.156	.163	.168	.162	.167	.186	.173	.171	.157	.156	.173
35	.138	.133	.141	.148	.139	.146	.167	.151	.151	.134	.134	.153
36	.117	.111	.121	.129	.118	.127	.149	.130	.132	.113	.112	.135
37	.098	.090	.102	.111	.097	.108	.132	.109	.115	.094	.092	.118
38	.081	.071	.084	.094	.078	.091	.115	.090	.098	.075	.074	.102
39	.065	.054	.068	.079	.061	.076	.100	.073	.083	.059	.057	.087
40	.051	.040	.054	.065	.046	.062	.086	.058	.070	.045	.043	.074
41	.039	.028	.043	.053	.035	.050	.073	.046	.058	.034	.032	.062
42	.029	.020	.033	.042	.026	.040	.062	.037	.047	.024	.023	.051
43	.021	.014	.025	.033	.020	.032	.051	.030	.038	.017	.016	.041
44	.015	.009	.019	.025	.015	.025	.042	.024	.031	.011	.010	.033
45	.011	.006	.015	.019	.011	.020	.034	.019	.024	.007	.007	.026
46	.007	.004	.011	.014	.008	.015	.027	.016	.019	.005	.004	.021
47	.005	.003	.008	.010	.006	.012	.021	.012	.014	.003	.003	.016
48	.003	.002	.006	.007	.004	.009	.016	.010	.011	.002	.002	.012
49	.002	.001	.005	.005	.003	.006	.012	.007	.008	.001	.001	.009
50	.001	.001	.003	.003	.002	.005	.009	.006	.006	.001	.001	.007

TABLE 6. Approximations B and C for the overflow probability in $M/M/m + D_2$ ($\lambda = 40$, $\beta = 1$, $\gamma = 0.5$)

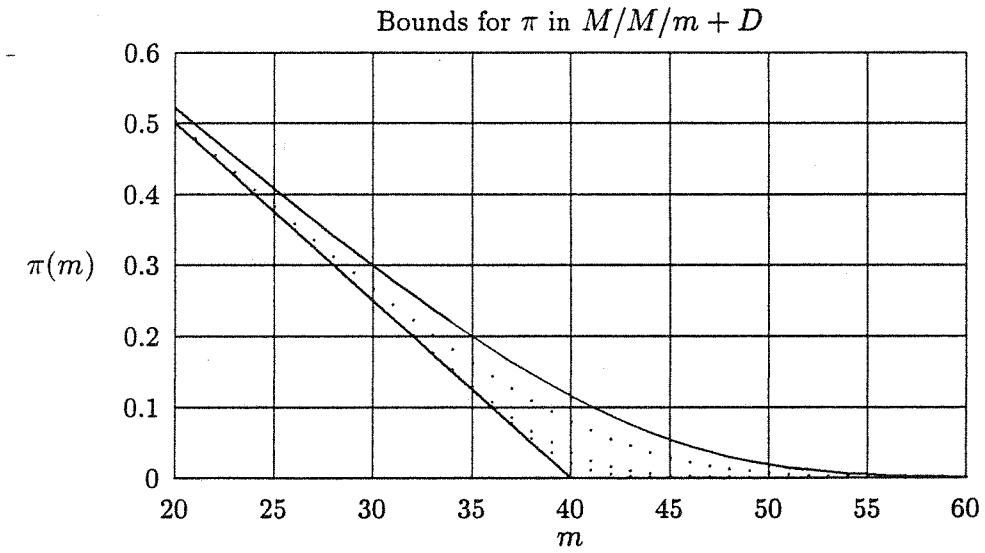

FIGURE 1. Overflow probability in $M/M/m + D$.
In decreasing order: $\gamma = 0$ (upper bound), 0.1, 0.5, 1.0, ∞ (lower bound).

Overflow probability in $M/M/m + D_2$												
	D_2-a			D_2-b			D_2-c			D_2-d		
m	exact	π_B	π_C	exact	π_B	π_C	exact	π_B	π_C	exact	π_B	π_C
30	.250	.250	.252	.252	.251	.252	.261	.258	.253	.250	.250	.253
31	.225	.225	.227	.227	.227	.228	.238	.235	.229	.225	.225	.230
32	.201	.200	.203	.203	.202	.205	.216	.211	.206	.200	.200	.207
33	.176	.176	.180	.180	.178	.182	.195	.188	.184	.175	.175	.185
34	.152	.151	.157	.157	.154	.160	.175	.166	.162	.151	.150	.163
35	.128	.127	.134	.135	.131	.138	.155	.143	.141	.127	.126	.143
36	.106	.103	.113	.114	.108	.117	.137	.122	.122	.103	.102	.124
37	.084	.080	.092	.095	.086	.098	.119	.100	.104	.081	.078	.107
38	.065	.059	.073	.077	.065	.080	.103	.080	.087	.060	.055	.091
39	.048	.040	.056	.061	.046	.064	.087	.061	.072	.043	.034	.076
40	.034	.025	.042	.047	.031	.051	.073	.046	.059	.028	.017	.063
41	.023	.015	.031	.035	.021	.039	.061	.035	.047	.017	.009	.051
42	.014	.008	.023	.026	.014	.031	.050	.028	.038	.010	.005	.042
43	.009	.005	.018	.018	.010	.024	.040	.022	.030	.005	.003	.033
44	.005	.003	.013	.013	.007	.019	.032	.018	.023	.003	.001	.026
45	.003	.002	.010	.008	.005	.014	.025	.014	.018	.001	.001	.020
46	.002	.001	.008	.006	.003	.011	.019	.011	.014	.001	.000	.015
47	.001	.000	.006	.004	.002	.008	.014	.008	.010	.000	.000	.012
48	.000	.000	.004	.002	.001	.006	.011	.006	.008	.000	.000	.009
49	.000	.000	.003	.001	.001	.004	.008	.005	.006	.000	.000	.006
50	.000	.000	.002	.001	.000	.003	.005	.003	.004	.000	.000	.005

TABLE 7: Approximations B and C for the overflow probability in $M/M/m + D2$ ($\lambda = 40$, $\gamma = 1.0$)

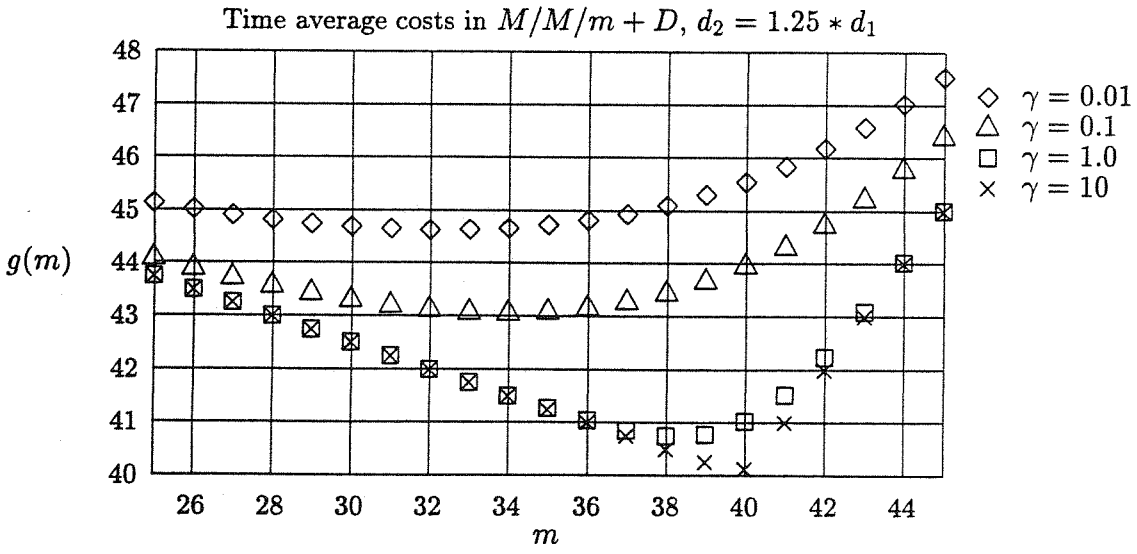

FIGURE 2. Time average cost in $M/M/m + D$
($\lambda = 40$, $\beta = \beta^* = 1$, $d_2 = 1.25$, $d_1 = 1$)

Overflow probability in $M/M/m+E_2$						
	$\gamma = 0.1$		$\gamma = 0.5$		$\gamma = 1.0$	
m	exact	π_C	exact	π_C	exact	π_C
30	0.273	0.273	0.254	0.256	0.251	0.252
31	0.251	0.252	0.230	0.233	0.226	0.228
32	0.230	0.231	0.207	0.210	0.202	0.205
33	0.210	0.211	0.185	0.188	0.178	0.182
34	0.190	0.191	0.164	0.167	0.155	0.160
35	0.171	0.172	0.143	0.146	0.133	0.138
36	0.153	0.154	0.124	0.127	0.112	0.117
37	0.136	0.137	0.106	0.108	0.093	0.098
38	0.120	0.120	0.089	0.091	0.076	0.080
39	0.104	0.105	0.074	0.076	0.060	0.064
40	0.090	0.091	0.061	0.062	0.047	0.051
41	0.077	0.077	0.049	0.050	0.036	0.039
42	0.065	0.065	0.039	0.040	0.027	0.031
43	0.054	0.055	0.030	0.032	0.020	0.024
44	0.045	0.045	0.023	0.025	0.014	0.019
45	0.036	0.036	0.018	0.020	0.009	0.014
46	0.029	0.029	0.013	0.015	0.006	0.011
47	0.023	0.023	0.010	0.012	0.004	0.008
48	0.018	0.018	0.007	0.009	0.002	0.006
49	0.014	0.014	0.005	0.006	0.001	0.004
50	0.010	0.010	0.003	0.005	0.000	0.003

TABLE 8. Approximation C for the overflow probability in $M/M/m+E_2$ ($\lambda = 40$, $\beta = 1$)

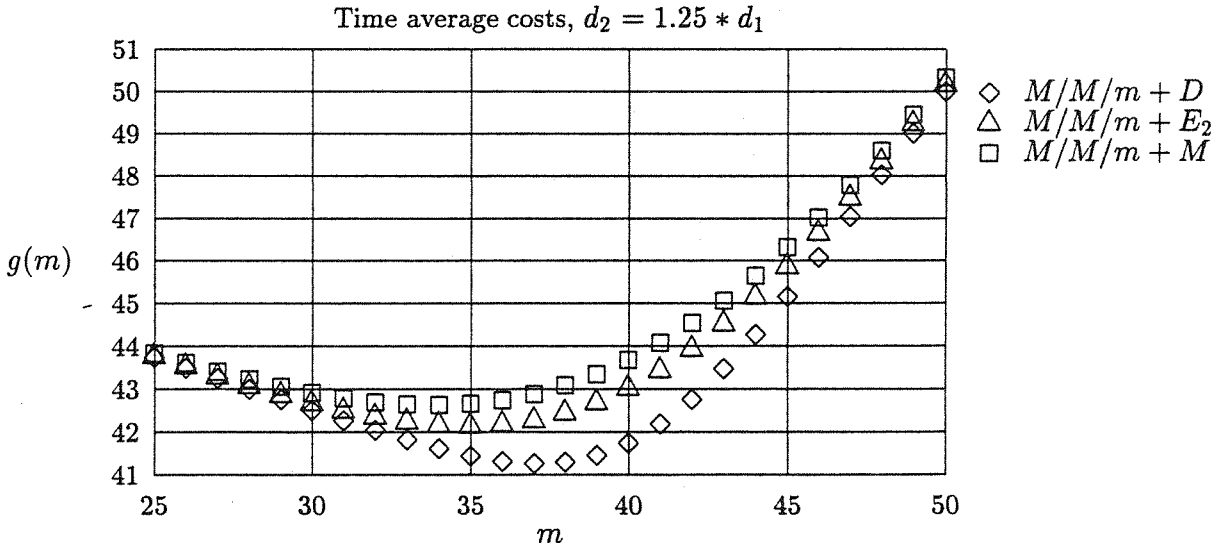

FIGURE 3. Time average costs in $M/M/m+D$, $M/M/m+E_2$, $M/M/m+M$ ($\lambda = 40$, $\beta = \beta^* = 1$, $\gamma = 0.5$, $d_2 = 1.25$, $d_1 = 1$)

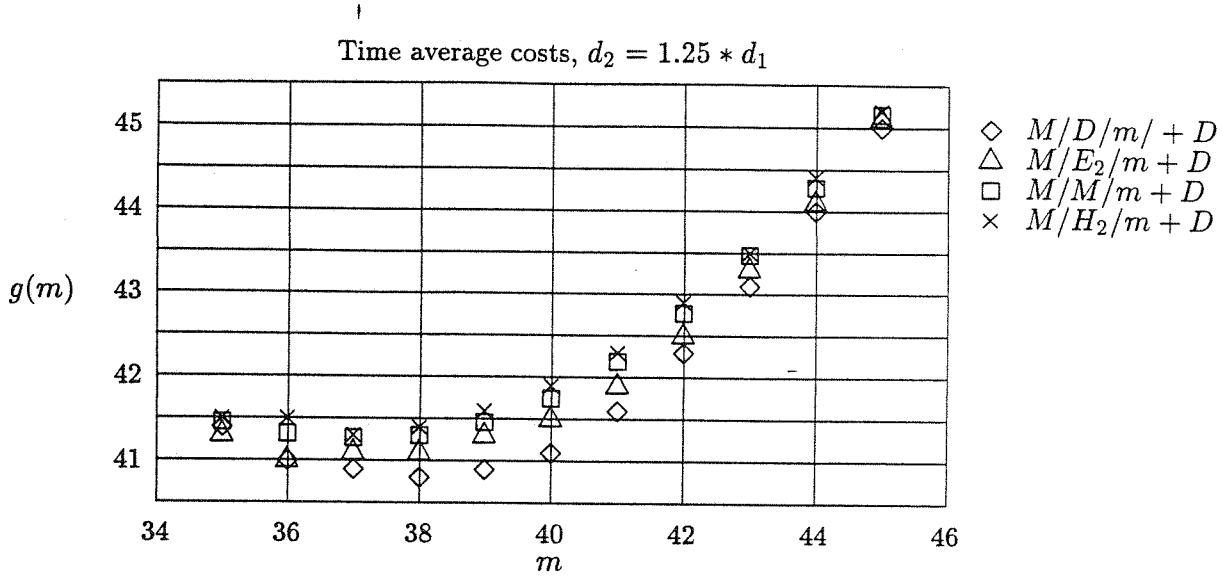

FIGURE 4: Time average costs in $M/D/m+D$, $M/E_2/M+D$, $M/M/m+D$, $M/H_2/m+D$
($\lambda = 40$, $\beta = \beta^* = 1$, $\gamma = 0.5$, $d_2 = 1.25$, $d_1 = 1$)

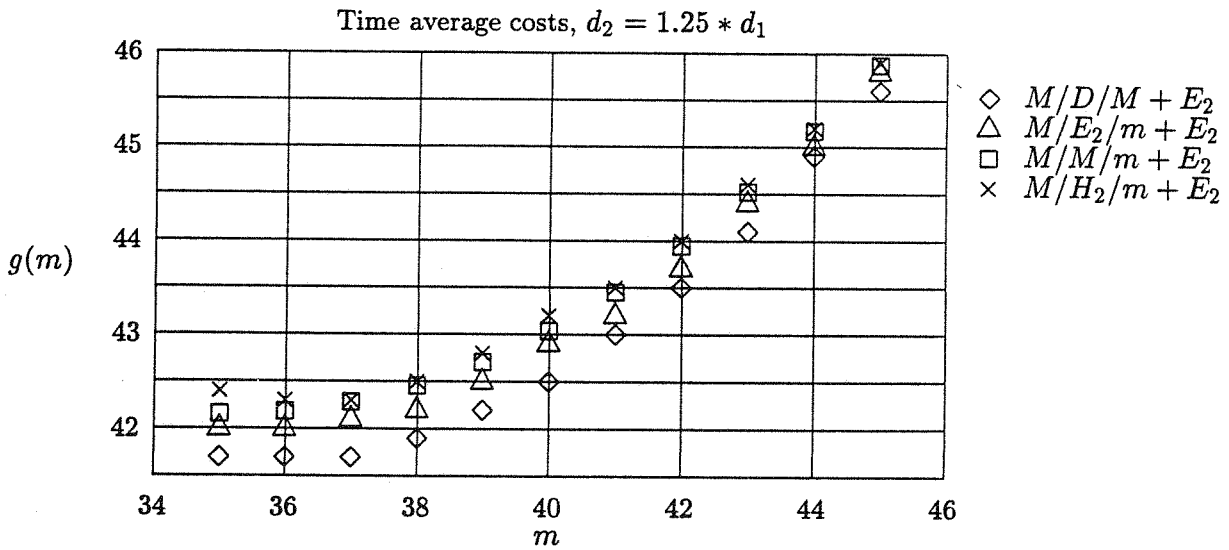

FIGURE 5: Time average costs in $M/D/m+E_2$, $M/E_2/M+E_2$, $M/M/m+E_2$, $M/H_2/m+E_2$
($\lambda = 40$, $\beta = \beta^* = 1$, $\gamma = 0.5$, $d_2 = 1.25$, $d_1 = 1$)

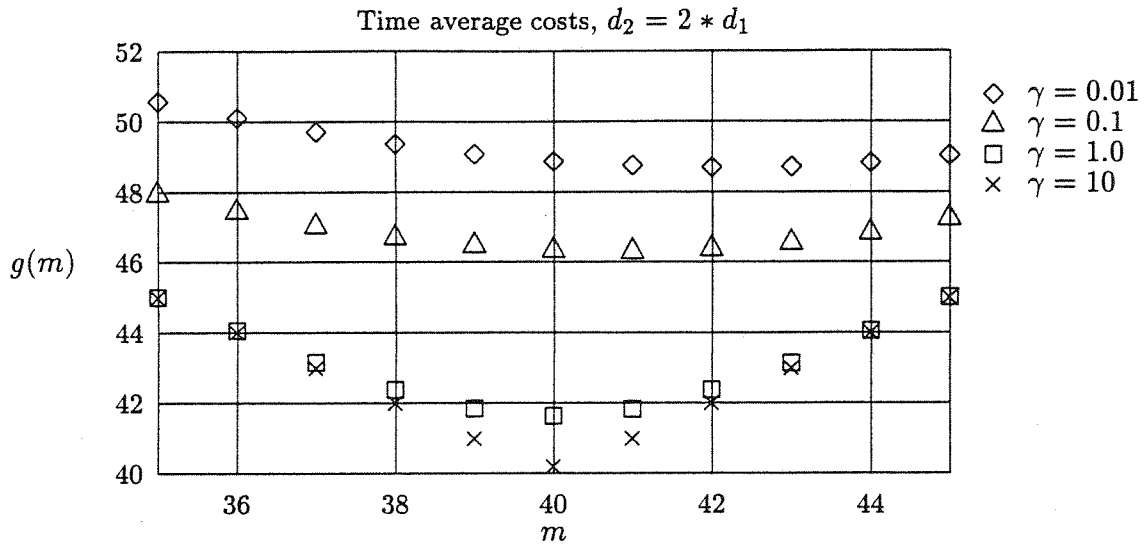

FIGURE 6. Time average costs in $M/M/m + D$
($\lambda = 40$, $\beta = \beta^* = 1$, $d_2 = 2$, $d_1 = 1$)

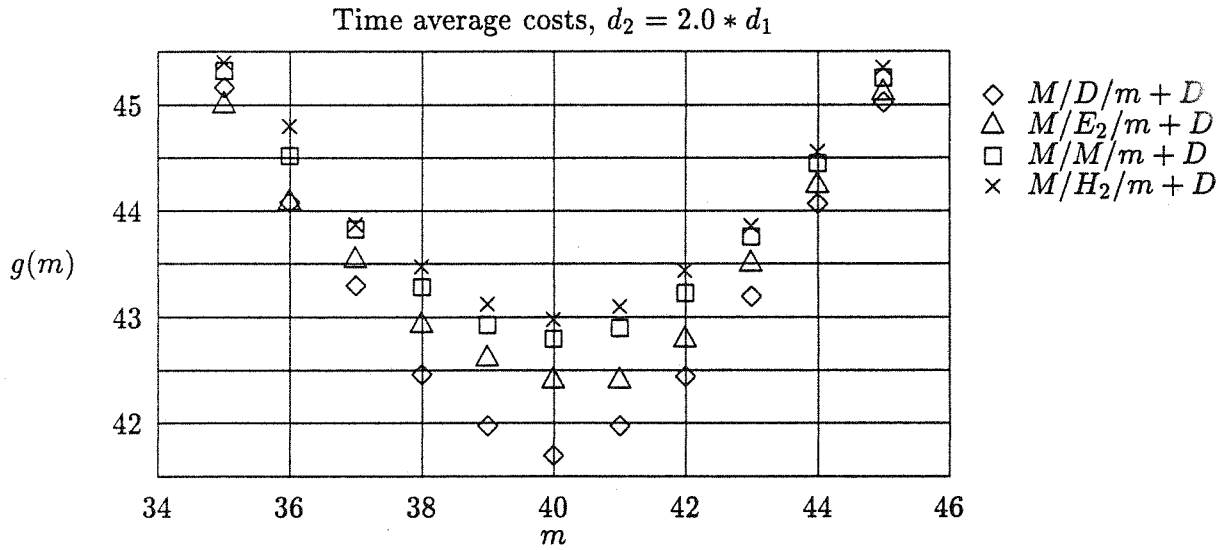

FIGURE 7: Time average costs in $M/D/m + D$, $M/E_2/m + D$, $M/M/m + D$, $M/H_2/m + D$
($\lambda = 40$, $\beta = \beta^* 1$, $\gamma = 0.5$, $d_2 = 2$, $d_1 = 1$)

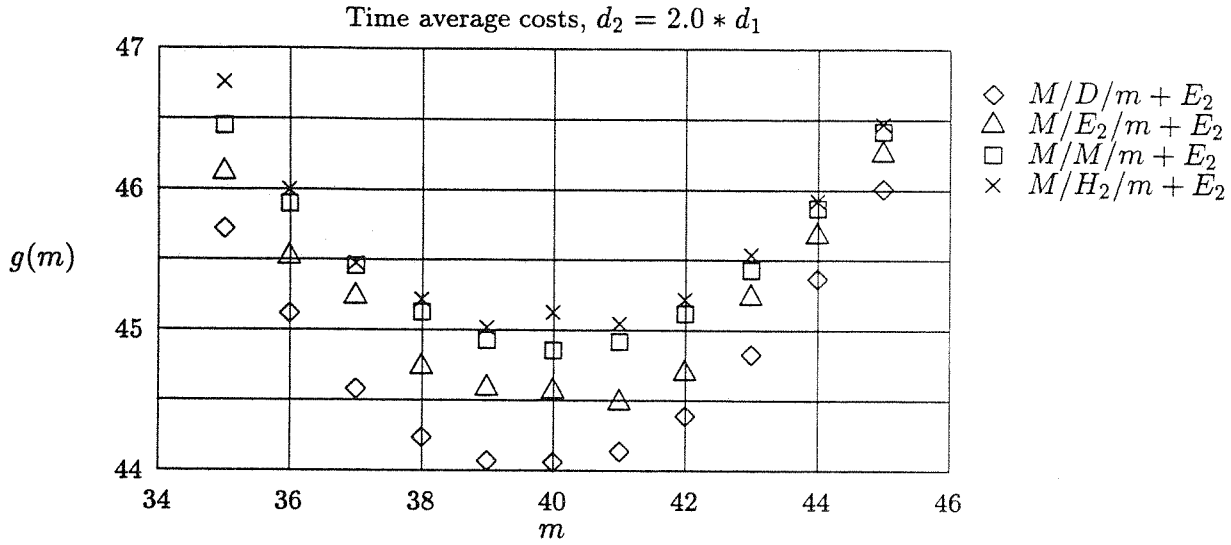

FIGURE 8: Time average cost in $M/D/m+E_2$, $M/E_2/m+E_2$, $M/M/m+E_2$, $M/H_2/m+E_2$
($\lambda = 40$, $\beta = \beta^*1$, $\gamma = 0.5$, $d_2 = 2$, $d_1 = 1$)

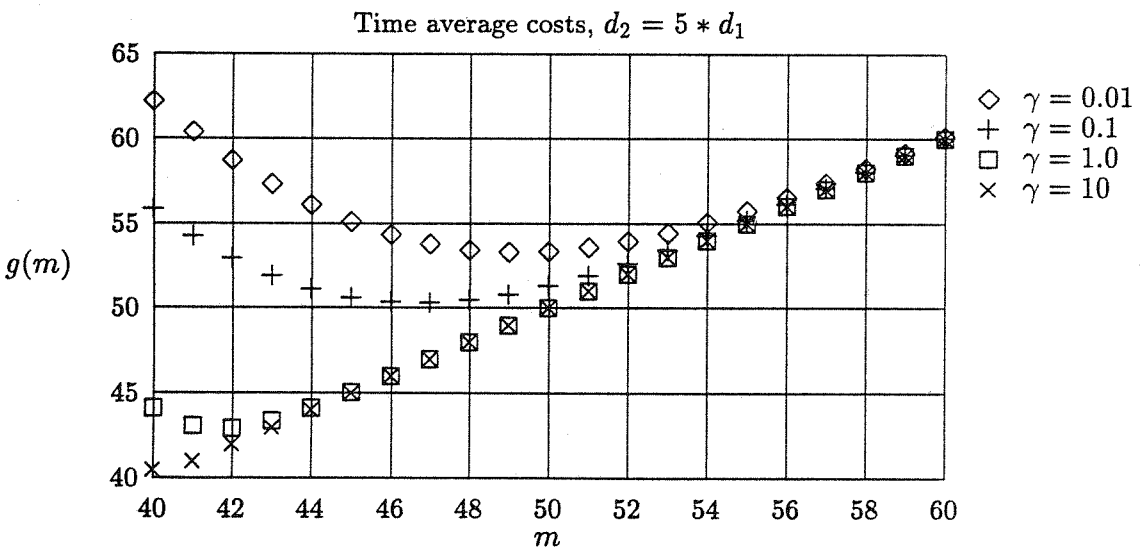

FIGURE 9. Time average cost in $M/M/m + D$
$\lambda = 40$, $\beta = \beta^* = 1$, $d_2 = 5$, $d_1 = 1$.