Working Paper

An Analytical Model for Closed Networks with a General Single Server Queue and an Infinite Server Station

Paul Desruelle

WP-94-18
March 1994

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

An Analytical Model for Closed Networks with a General Single Server Queue and an Infinite Server Station

Paul Desruelle

WP-94-18
March 1994

International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

An analytical model for closed networks with a general single server queue and an infinite server station.

by:

Paul Desruelle
University of Wisconsin-Madison
Department of Industrial Engineering
1513, University Avenue
Madison WI 53706
U.S.A.

email: desruell@macc.wisc.edu

Abstract:

In this paper we propose a new analytical model to solve a class of closed queuing networks with general servers and with a small number of customers. This model can be applied to analyze problems such as the machine-repairman problem and simultaneous resource possession. Our algorithm is based on exact mean value analysis and on two-moment approximations of GI/G/1 queues. The model gives good results for a large range of the network parameters.

1. Introduction

1.1 Description of the problem

In this paper, we present a new analytical model for solving a class of closed queuing networks. The networks that we consider consist, first of a first-come-first-serve single server queue with general interarrival and service times (a GI/G/1 queue), and second of a feed-back loop with a delay station. The delay station also has general service times. The delay station is equivalent to a service station where the number of servers is at least as large as the number of customers in the closed network. This type of network is shown in Figure 1.

Solving such a network has practical applications in manufacturing systems analysis, to the so-called machine-repairman problem [Elmaghraby, 1971] and to machine-operator interference analysis [Stecke, 1982; Stecke and Aronson, 1985; Carmichael, 1987; Kamath and Sanders, 1990]. The solution also applies to simultaneous resource possession analysis in computer and telecommunication systems modeling [Jacobson and Lazowska, 1982].

We consider a class of networks with a small number of customers (say 15 or less). This consideration is motivated by current philosophies that advocate the superiority of small human-scale systems over large systems. Practical examples can be found in manufacturing systems where small production entities such as flexible workcells prove to be more competitive than large workshops; similarly more and more computer systems are designed

around small local networks (LAN's) rather than around a large mainframe with an important number of terminals and peripherals.

The core of our system is a queue with general service and interarrival times. To our knowledge there exists no exact analytical solutions to closed networks with this type of queue [Springer and Makens, 1992]. The central problem consists of estimating the mean waiting in the queue. In order to obtain an exact solution, discrete-event simulation [Law, 1986] is the only modeling alternative. However this alternative is not always attractive due to the time it takes to build the model, to simulate it and to analyze the results. A good compromise between speed of getting the solution and its accuracy is to develop and use a "good" analytical model of the system [Suri, 1987] in order to estimate the mean waiting time. This is what we are doing in this paper.

1.2 Organization of the paper

In the second section of this paper we present previous attempts to find approximate analytical solutions. We start by presenting the approaches based on mean value analysis (MVA) which provides exact results for product form networks. Then we present approaches based on network decomposition and two-moment approximation of GI/G/1 queues. We explain why these approaches, when considered independently, are not satisfactory to solve the class of networks that we consider.

In Section 3 we describe our analytical model which combines exact MVA and the analysis of networks of GI/G/1 queues.

In Section 4 we present our experimental results. In order to validate our analytical model, we compare results from discrete event simulation to the analytical results. The experimental design covers a wide range of the input parameters, namely the squared coefficient of variation (scv) of the service times (C_s^2), the scv of the delay times (C_o^2), the ratio of the mean service time over the mean the delay time (E[S]/E[O]), and the number of customers in the network (N). The compared performance measure is the mean waiting time at the queuing station (E[W]).

In the conclusion we discuss the validity of our analytical model. The experimental results show that the accuracy of the results depends on the scv of the delay times. This scv is at this point not incorporated in our model. The effect of the scv of the delay times is discussed and the validity ranges of the analytical model (in terms of the different parameters) are identified. Directions for future research to improve the performance of our model are given.

2. Previous results

In this section we present a short historical overview of the development of analytical solutions for closed queuing networks.

2.1 Product form networks

The product form solution can be considered as the initial breakthrough that enabled to obtain easily satisfactory solutions to queuing networks. The product form solution states that, under certain conditions, the distribution of customers within the network in equilibrium is similar to the product of the states distribution of the queues that compose the network considered independently. Jackson proved the product form solution for open networks [Jackson, 1957] and Gordon and Newell for closed networks [Gordon and Newell, 1967]. Baskett, Chandy, Muntz and Palacios demonstrated the validity of the product form solution for a number of open, closed, and mixed networks [Baskett, Chandy, Muntz and Palacios, 1975]. The so-called BCMP networks are composed of stations with one of the following characteristics: first-come-first-serve (FCFS) stations with exponentially distributed service times and processor-sharing (PS), infinite servers (IS), and last-come-first-serve (LCFS) stations with no restrictions on the distribution of service times, except for the existence of a rational Laplace transform. In addition, if the network is of an open or mixed type, external arrivals to the network should be Poisson. Solving product form networks involves, however, the solution of a normalizing constant which can be quite tedious. For example, for closed networks, this solution requires the summation over all possible states of the network.

Buzen proposed a convolution algorithm to derive efficiently the normalizing constant for closed networks and showed how to compute the network performance characteristics on the basis of the normalizing constants [Buzen, 1973]. An alternative approach for solving closed BCMP networks is the MVA algorithm due to Reiser and Lavenberg [Reiser, 1979; Reiser and Lavenberg, 1980]. MVA's appeal is found in its formulation directly in terms of network performance characteristics: mean queue lengths, sojourns times, and throughput. MVA does not attempt to compute the distributions of customers at individual stations like the product form solution does, but it provides practical performance measures that are really what the analyst is interested in. The original MVA algorithm consists of a recursion in the number of customers in the network. MVA is a widely used tool for solving BCMP networks [Suri and Hildebrant, 1984], and its computation requirements can be reduced as shown by Schweitzer and Bard [Schweitzer, 1979; Bard, 1979, 1981].

2.2 Non product form networks

A important problem with product form networks is that they assume exponentially distributed service times at FCFS stations. This is an highly penalizing assumption when trying to solve real systems. MVA has however been extended into approximate algorithms to solve non product form networks. For example, to take into consideration non exponentially distributed service times, Wijbrands incorporates the variance of the service times in the computation of the sojourn times at the queuing station. But exponential arrivals are assumed [Wijbrands, 1988].

Another avenue to analyze non product form networks is the network decomposition approach. This approach consists of analyzing each queue individually as a queue with

general arrivals and general service times. For example, Whitt uses a two-moment approximation of the interarrival and service times distributions to analyze open and closed networks of GI/G/1 queues with infinite calling population. A major assumption is that arrival processes to each queue are renewal processes [Whitt, 1983]. An interesting application of Whitt's approach is the modeling of closed-loop flexible assembly systems as loops of GI/G/1 queues by Kamath, Suri, and Sanders [Kamath, Suri, and Sanders, 1988]. Two-moment approximations are attractive because they usually achieve a good compromise between mathematical simplicity and result accuracy.

3. Description of the model

To our knowledge there exists no model that is directly applicable to the solution of the closed network described in Section 1. The solution would be trivial if the service times at the single server station were exponentially distributed since our network would have a product form solution (note that the delay station is our network is equivalent to an IS station). This is not the case and an approach solely based on MVA would not be satisfactory. Using solely network decomposition with two-moment approximations of the GI/G/1 queue would not work well either since the network population is small and, therefore, the assumption of an infinite calling population for the queue does not hold. The algorithm derived by Kamath et al. [1988] is not applicable either, since our network is not composed only of GI/G/1 queues, but also of a delay station.

To solve our network, we derive an algorithm that is an extension of the algorithm presented by Kamath et al. [1988]. Our algorithm is a combination of pure MVA applied to the product form equivalent of our network, and of network decomposition with two-moment approximations of the GI/G/1 queue.

In this section, we first introduce the notation used in the paper, then present classical approximations to the mean customer waiting time in a GI/G/1 queue, and finally describe how this mean waiting time can be better approximated in the context of our network.

3.1 Notation

c_a^2	Squared coefficient of variation (scv) of the customer interarrival times to the queue,
c_o^2	Scv of the delay times,
c_s^2	Scv of the service times,
E[O]	Mean customer delay time,
E[S]	Mean customer service time,
E[W]	Mean customer waiting time,
$E[W]_{cN}$	Mean customer waiting time with an exponential queue and with N customers in the network,
$E[W]_{c\infty}$	Mean customer waiting time with an exponential queue and an infinite calling population,
$E[W]_N$	Mean customer waiting time with a general queue and with N customers in the

	network,
$E[W]_\infty$	Mean customer waiting time with a general queue and an infinite calling population,
f	Correction factor to account for the small network population (i.e. number of customers),
k	Value of the ratio $(Q_{N-1}/N-1)/(Q_N/N)$,
N	Number of customers in the network,
O	Delay time, the time spent by the customers at the delay station,
Q_N	Expected queue length at the server station, with N customers in the network,
S	Service time at the single server station,
T	Customer interarrival time to the queue,
W	Time that a customer waits in queue for service,
λ	Network throughput rate,
ρ	Queuing station utilization.

3.2 GI/G/1 approximations

In this section, we first recall previous approximations for the mean customer waiting time for service by a FCFS single server station with general service times and general and independent interarrivals (GI/G/1). We then indicate the approximation that we use in our model.

A classical result for approximating the mean customer waiting time in a GI/G/1 queue with infinite calling population is the upper bound derived by Kingman [Kingman, 1962]. With our notation, Kingman's bound is expressed as:

$$E[W] \leq \frac{\lambda.(Var[S]+Var[T])}{2(1-\rho)} \qquad (1)$$

This bound is known as being very good in heavy traffic (i.e. when the server utilization, here ρ, is close to 1) and exact in the deterministic D/D/1 case [Marshall, 1968]. An improved approximation is derived by Krämer and Langenbach-Belz [Krämer and Langenbach-Belz, 1978]. This approximation has been derived heuristically by interpolating between general heavy traffic results and results known for special systems, and then validated by simulation [Kuehn, 1979]. The Krämer and Langenbach-Belz approximation is expressed as follows:

$$E[W]=g(\rho,c_a^2,c_s^2).\frac{c_a^2+c_s^2}{2}.\frac{E[S].\rho}{1-\rho} \qquad (2)$$

where,

$$g(\rho,c_a^2,c_s^2)=\exp(-\frac{-2(1-\rho)}{3\rho}.\frac{(1-c_a^2)^2}{c_a^2+c_s^2}),\quad c_a^2<1 \tag{3}$$

or,

$$g(\rho,c_a^2,c_s^2)=\exp(-(1-\rho).\frac{(c_a^2-1)}{c_a^2+4c_s^2}),\quad c_a^2\geq 1 \tag{4}$$

Typically, the error resulting from the above approximation is less than 10%, and is always less than 20% for specified combinations of distribution functions [Krämer and Langenbach-Belz, 1978]. We use expression (2) as a starting point in our analysis with the modification that we approximate g to the value of one, for the sake of simplicity. Other approximations have been derived for E[W] but, according to Shantikumar and Buzacott, when C_a^2 is larger than one, "any method may be used" [Shantikumar and Buzacott, 1980]. Since in our model we have only values of C_a^2 above one (except in the trivial product form case), we keep the expression for E[W] as simple as possible.

Then, by considering the queuing station independently from the rest of the network, and by assuming infinite calling population, we can approximate the mean customer waiting time by the following expression:

$$E[W]=\frac{c_a^2+c_s^2}{2}.\frac{E[S].\rho}{1-\rho} \tag{5}$$

It is clear from the expression above that we need to know the value of C_a^2 and of ρ in order to estimate E[W]. These values can be determined by iteration algorithms that are not described here, the focus of this paper being on finding a good approximation for E[W]. We therefore assume that we know how to obtain analytically values for C_a^2 and ρ, and compute E[W] as a function of these values. In the model validation phase, we simply use the values of C_a^2 and ρ obtained from the simulation output report.

Up to this point we have an expression to estimate the mean waiting time in a GI/G/1 queue with infinite calling population. In our system, however, the calling population is limited to the number of customers within the closed network, therefore the above expression cannot be applied directly.

3.3 Closed network approximation

To derive a closed network approximation we follow a derivation by Kamath et al. [1988] for closed networks of GI/G/1 queues that relies on the symmetry of the network. Our derivation differs in the sense that our network is not symmetric since it is composed of a

queue and a delay station.

The main idea is to find a correction factor (f) that accounts for the transition from an infinite calling population to a finite calling population in a product form network, and to apply the same correction to our non product form network. The product form network equivalent to our network is obtained by replacing the queue with general services with a queue with exponential services while keeping the same level of station utilization (ρ).

- ***Derivation of the correction factor f for the product form network***

An expression for the mean waiting time in front of the queue can be found using the fundamental elements of MVA [Reiser, 1979]: Little's law and the arrival theorem. Applying Little's law to the queuing station we can write:

$$Q_N=(E[W]_{eN}+E[S]).\lambda \tag{6}$$

where, $E[W]_{eN}$ is the mean waiting time in front of the queuing station if service times are exponential with N customers in the network, E[S] is the mean customer service time, λ is the station throughput rate, and Q_N is the mean number of customers at the station (waiting in queue and being served) if there are N customers in the network. The arrival theorem states that in a closed product form network an arriving customer to a queue sees the queuing station in equilibrium with one less customer in the network. Therefore, and because of the memoryless property of the exponential distribution, we can write:

$$E[W]_{eN}=Q_{N-1}.E[S] \tag{7}$$

In order to continue our derivation, we need to obtain an expression for Q_{N-1} as a function of Q_N. A simple approximation would be to apply (as done in Kamath et al. [1988]) the Schweitzer-Bard heuristic:

$$\frac{Q_{N-1}/N-1}{Q_N/N}=1 \tag{8}$$

which has been used to make MVA computations easier [Suri and Hildebrant, 1984]. This heuristic is exact for closed symmetric exponential networks. With our non-symmetric network, using this heuristic does not give satisfactory results. However it is clear that by applying the MVA algorithm, we can find a value (not necessarily equal to one) to the ratio in equation (8) for any value of N, the number of customers in the network. Therefore, we can write:

$$\frac{Q_{N-1}/N-1}{Q_N/N}=k(N,E[S],E[O]) \tag{9}$$

where E[O] is the mean delay time, the mean time that a customer is delayed at the delay station. Or, for the sake of simplicity:

$$\frac{Q_{N-1}/N-1}{Q_N/N}=k \tag{10}$$

The derivation of the factor k is detailed in appendix.

Then, we have:

$$Q_{N-1}=k.\frac{N-1}{N}.Q_N \tag{11}$$

therefore,

$$E[W]_{eN}=k.\frac{N-1}{N}.Q_N.E[S] \tag{12}$$

or, by applying expression (6),

$$E[W]_{eN}=k.\frac{N-1}{N}.(E[W]_{eN}+E[S]).\lambda.E[S] \tag{13}$$

or, since the station utilization ρ equals (λ. E[S]),

$$E[W]_{eN}=k.\frac{N-1}{N}.(E[W]_{eN}+E[S]).\rho \tag{14}$$

therefore,

$$E[W]_{eN}=k.\frac{N-1}{N}.\frac{\rho}{1-k.\frac{N-1}{N}.\rho}.E[S] \tag{15}$$

It is well known that the mean customer waiting time in an M/M/1 queue with infinite calling population is:

$$E[W]_{e\infty}=\frac{E[S].\rho}{1-\rho} \tag{16}$$

We can now define the correction factor f(E[W]$_{e\infty}$) as:

$$f(E[W]_{e\infty})=\frac{E[W]_{eN}}{E[W]_{e\infty}} \tag{17}$$

Thus,

$$f(E[W]_{e\infty})=k.\frac{N-1}{N}.\frac{\frac{E[W]_{e\infty}}{E[W]_{e\infty}+E[S]}.E[S]}{1-k.\frac{N-1}{N}.\frac{E[W]_{e\infty}}{E[W]_{e\infty}+E[S]}}.\frac{1}{E[W]_{e\infty}} \tag{18}$$

or,

$$f(E[W]_{e\infty})=k.\frac{N-1}{N}.\frac{1}{1+\frac{E[W]_{e\infty}}{E[S]}.(1-k.\frac{N-1}{N})} \tag{19}$$

- ***Application to the original network***

We can now compute the mean waiting time for the queue with general service times of our original closed network. This is done as follows:

$$E[W]_N=f(E[W]_\infty).E[W]_\infty \tag{20}$$

where $E[W]_\infty$ is computed applying equation (5) and where

$$f(E[W]_\infty)=k.\frac{N-1}{N}.\frac{1}{1+\frac{E[W]_\infty}{E[S]}.(1-k.\frac{N-1}{N})} \tag{21}$$

- ***Case when the server utilization ρ is equal to 1***

Note that equation (5) can be applied only if the server utilization is different of one. A general queue with infinite calling population is unstable if the server utilization is equal to one. In a closed network, however, a server utilization can be equal to one and the network still be stable, because the network population is finite. Since expressions (20) and (21) cannot be applied if $\rho=1$, a different expression must be derived, this is done in the following.

If we combine expressions (20) and (21), we can write:

$$E[W]_N=k.\frac{N-1}{N}.\frac{1}{1+\frac{E[W]_\infty}{E[S]}.(1-k.\frac{N-1}{N})}.E[W]_\infty \tag{22}$$

If $\rho=1$ we know that the mean waiting time for the open queue ($E[W]_\infty$) becomes infinite.

Then,

$$\lim_{\rho \to 1} E[W]_N = \lim_{E[W]_\infty \to \infty} E[W]_N \tag{23}$$

Since,

$$\lim_{E[W]_\infty \to \infty} E[W]_N = \frac{E[S]}{(\frac{N}{N-1} \cdot \frac{1}{k}) - 1} \tag{24}$$

then, when $\rho = 1$,

$$E[W]_N = \frac{E[S]}{(\frac{N}{N-1} \cdot \frac{1}{k}) - 1} \tag{25}$$

4. Experimental results

In this section we describe the experimental design used in order to validate our analytical model, and we present the experimental results.

4.1 Experimental design

A simulation model of the closed network shown on Figure 1 was built and run over the following ranges of network parameters:
- Ratio of the mean service time over the mean delay time: $0.05 \leq E[S]/E[O] \leq 20$,
- Number of customers in the closed network: $2 \leq N \leq 15$,
- Scv of the service times: $1 \leq C_s^2 \leq 20$,
- Scv of the delay times: $0 \leq C_o^2 \leq 5$.

Three experiments were performed; the first one was done varying only the ratio E[S]/E[O], the second one varying only N, and the third one varying only the two scv's. The simulation model was written using the SIMAN language [Pedgen, 1987]. A gamma distribution was chosen to model the general distributions of the service and delay times. The gamma distribution was chosen because it enables with only two parameters (α and β) to set easily desired values for the means and the scv's ($C^2 = 1/\alpha$, mean $= \alpha.\beta$). All simulation runs were 65,000 time units long (1 time unit represents 1 minute of the network actual operation time). In order to eliminate network start-up effects the first thousand time units of each run were truncated. For most of the runs the simulated operation time corresponded to a number of services at the single server station comprised between 18,000 and 24,000. For the first experiment performed with a varying ratio of service times over delay times, the number of services performed varied from approximatively 6,700 (with the smallest ratio) to approximatively 29,000 (with the largest ratio). Since long, the runs were not replicated

[Pedgen, 1987].

The output statistics were directly read from the simulation output report generated by SIMAN. The parameters used in comparing simulation results to analytical results were computed directly from the output report (for example, C_a^2, C_s^2 and ρ, necessary to compute the mean waiting in front of the equivalent GI/G/1 queue (equation (5))). In computations using the input parameters E[S] and C_s^2, output values read from the output report were also considered rather than the input values set through the parameters of the gamma distributions. In general the discrepancy between input values and output values was small (5% or less for E[S] and 6.5% or less for C_s^2).

4.2 Variable ratio of the service times over delay times

The graphs on Figure 2 show the results of the comparison between the analytical and the simulation model when the ratio of the mean service time over the mean delay time (E[S]/E[O]) varies from 0.05 to 20. In this experiment, the following parameters were kept constant: the number of customers in the network (N = 5), the scv of the service times (C_s^2 = 10), the scv of the delay times (C_o^2 = 2), and the mean service time (E[S] = 2.22 min.). The mean delay time (E[O]) was therefore varied from 44.40 to 0.11 min. Figure 2a) shows the mean waiting time E[W] in minutes as computed with the analytical model and as given by the simulation model. Figure 2 b) shows the corresponding relative error in the mean waiting time. The relative error is computed as follows:

Error = (Analytical result - Simulation result)/Simulation result

In Figure 2 b) the vertical scale goes from -0.5 (-50%) to +0.5 (+50%). From the graphs on Figure 2, it is clear that the model is robust within the tested range. For values of the ratio beyond 4, the error is close to zero. For values of the ratio between 0.1 and 4, the absolute value of the error is at most equal to 15% which can be considered good. According to this first set of results, the performance of the analytical model is good for all values of the ratio within the tested range, and excellent when the mean service time is larger than the mean delay time.

4.3 Variable number of customers in the network

The graphs on Figure 3 show the results of the models comparison when N, the number of customers in the closed network, varies from 2 to 15. In this experiment, all other network parameters are kept constant (C_s^2 = 10, C_o^2 = 2, E[S] = 2.22, E[O] = 8.5). Figure 3 a) shows the compared mean waiting times and Figure 3 b) the relative error. These graphs show that the accuracy of the model decreases as the number of customers in the network increases. This is not a surprise or a problem since our approximation is aimed at networks with a small number of customers. Still, even with fifteen customers, the relative error is still less than 20% which can be considered good. Below seven customers, the error is less than 10%.

4.4 Variable service time and delay time scv's

The graphs on Figures 4 and 5 show the results of the models comparison when the scv's of the service times and of the delay times vary ($1 \leq C_s^2 \leq 20$ and $0 \leq C_o^2 \leq 5$) and the rest of the parameters are kept constant (E[S] = 2.22, E[O] = 8.5, N =5). Figures 4 and 5 show the same results presented differently: in Figure 4, the variation in the scv of the delay times is represented on the horizontal axis and the variation in the scv of the service times is represented by the different curves; conversely in Figure 5. We analyze first the results in the product form case ($C_s^2 = 1$) and then in the non product form case.

- ***Product form case***

When the scv of the service times is equal to one, the service times are exponentially distributed and the network has, therefore, a product form solution. In this case the value of the scv of the delay times should not have any effect on the mean waiting time. This is verified on Figure 4 by both the analytical model results and the simulation model results (curves corresponding to $C_s^2 = 1$). Note, however, in Figure 4 c) that the error between the analytical results and the simulation results is not zero. We know that for product form networks, the MVA algorithm gives an exact solution, and since our algorithm is based on MVA, one could expect the results to be exact when C_s^2 equals one. The reason why this is not the case is because our algorithm uses an approximation for the waiting time in front of a general queue with the scv of interarrival times (C_a^2) as a factor (in equation (5)). Within the tested range of the scv of the delay times (C_o^2), the scv of the interarrival times (C_a^2) is less than one. Therefore, our analytical model underestimates the waiting time in front of the queue by a small percentage (up to 6%). Note that our attempt is not to solve the product form case for which MVA is a better tool than our algorithm. Our aim is to solve the non product form case.

- ***Non product form case***

The simulation results shown in Figure 4 a) clearly indicate that, when the network does not have a product form solution ($C_s^2 \neq 1$), the mean waiting time in front of the single server station decreases as the scv of the delay times (C_o^2) increases. In addition, Figure 4 a) confirms that the mean waiting time increases with the scv of the service times (C_s^2). The analytical results on Figure 4 b) also show the increase in mean waiting time as C_s^2 increases. However, it is clear from this figure that the scv of the delay times (C_o^2) has no influence on the mean waiting time computed by the analytical model. This is not surprising since C_o^2 is not a parameter of the analytical model. Figure 4 c) shows the relative error between the analytical and the simulation results. The error decreases as C_o^2 increases.

Figure 5 shows the same results, this time with the scv of the service times (C_s^2) being on the horizontal axis, and for three different values of the scv of the delay times (C_o^2 = 0, 2, 5). The graphs on Figure 5 a), b), and c) show the mean waiting time comparison, and Figure 5 d), the relative error. It is clear from this figure that the results given by the model are better when C_o^2 is large (larger or equal to 2), the value of the error being then less than

15%, even for large values of C_s^2. When C_o^2 is smaller than 2, the analytical model underestimates the mean waiting time in front of the single server queue (by more than 20% when $C_o^2 = 0$, for value of C_s^2 above 5).

- ***Violation of the independence assumption***

An explanation that can be given to the behavior of our model is the violation of the independence assumption. In a GI/G/1 queue, interarrival times are assumed to be independent of the service completion times. However, when the scv of the delay times decreases, this independence decreases as well. A limit case is when C_o^2 is equal to zero, then every arrival occurs exactly after a fixed time interval from the service completion. Interarrival times are, therefore, more or less correlated to the service times depending on the scv of the delay times. Computation of this correlation from the simulation results shows that, effectively, the value of the correlation is high (close to one) when C_o^2 is equal to zero, but only in this case. When C_o^2 becomes larger than zero (even only equal to 0.5), the value of the correlation drops below 0.5 and remains steady as C_o^2 increases.

- ***Violation of the renewal assumption***

Another cause for the discrepancy with the analytical model results is the violation of the renewal assumption for the interarrival times. For example, when a long service time occurs, all customers can be waiting in front of the station and, therefore, no arrival can occur. Then, when the service is completed, a series of arrivals with short interarrival times might occur. As the scv of the delay times increases, the violation of the renewal assumption becomes less important.

5. Conclusion

5.1 Discussion of the model validity

We have proposed a new analytical model that enables to analyze a class of closed queuing networks composed of a FCFS GI/G/1 queue and of a delay station. The analytical model results are good for large values of the scv of delay times ($C_o^2 \geq 2$) and for a small number of customers in the network ($N \leq 15$). Result quality increases as the number of customers decreases.

The limit on the number of customers is not the most constraining, since we have limited the scope of our model to small systems. The limit on the scv of the delay times is more constraining. For example, our model cannot be applied to study a system in which machines (the customers) would require services from an operator (the single server) at constant or almost constant intervals (C_o^2 small or equal to zero).

5.2 Directions for improvement of the model

The scv of the delay times (C_o^2) is not considered in the analytical model described here.

Therefore our model could be improved by considering C_o^2 as a parameter. A first possible improvement would be to consider the correlation between service times and interarrival times at the single server queue. An analysis similar to the study of correlated GI/G/1 queues by Venkateswar and Sanders [Venkateswar and Sanders, 1990] could be applied to adjust the value of the mean waiting time when correlation exists. Our simulation results show that this correlation, however, is significant only for values of C_o^2 very close to zero, which is only the case when the delay times are deterministic or almost deterministic. This improvement would then have limited application. A second possible improvement is to incorporate C_o^2 in the model by performing a regression analysis over a wide range of the values of C_o^2. The error curves shown in Figure 4 c) seem to have a regular pattern when C_s^2 is not equal to one, and such an analysis could prove successful.

Appendix: Derivation of the factor k

The factor k is derived by applying the original recursive MVA algorithm [Reiser, 1979]. When applying this algorithm, the network performance characteristics are computed as follows:

- Sojourn times:

At the single server queuing station:

$$T_s(n)=E[S].(1+Q_{n-1}) \qquad \textbf{(A.1)}$$

where n is the number of customers in the network at the current iteration of the algorithm (n=1,...,N), and where Q_{n-1} is the mean number of customers at the queuing station (being served or waiting for service) with n-1 customers in the network.

At the delay station:

$$T_o=E[O] \qquad \textbf{(A.2)}$$

- Network throughput rate:

$$\lambda(n)=\frac{n}{T_s(n)+T_o} \qquad \textbf{(A.3)}$$

- Mean number of customers at the queuing station:

$$Q_n=T_s(n).\lambda(n) \qquad \textbf{(A.4)}$$

or, by combining expressions (A.1) through (A.4):

$$Q_n = n.\frac{E[S].(1+Q_{n-1})}{E[S].(1+Q_{n-1})+E[O]} \qquad \textbf{(A.5)}$$

k is the value of the ratio $(Q_{N-1}/N-1)/(Q_N/N)$. In order to derive the value of this ratio we apply the following algorithm:

1) Input:
E[S], E[O], N.

2)Initialization:
$Q_0=0$.

3) Iteration loop:
For n=1 to N:

$$Q_n = n.\frac{E[S].(1+Q_{n-1})}{E[S].(1+Q_{n-1})+E[O]}$$

4) Output:

$$k=\frac{Q_{N-1}\ /(N-1)}{Q_N\ /N}$$

Note that the number of iterations is proportional to N, the number of customers in the network. Since N is small (say 15 or less), the above algorithm described in this appendix is very fast when implemented on a computer.

References

Bard, Y. 1979. Some Extensions to Multiclass Queueing Network Analysis. *Performance of Computer Systems.* pp. 51-62. M. Arato, A. Butrimenko, and E. Gelenbe (Eds.). IIASA and North-Holland Publishing Co.

Bard, Y. 1981. A Simple Approach to System Modeling. *Performance Evaluation.* Vol. 1, pp. 225-248.

Baskett, F., K., M. Chandy, R., R. Muntz and F., G. Palacios. 1975. Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM.* Vol. 22, pp. 248-260.

Buzen, J., P. 1973. Computational Algorithms for Closed Queuing Networks with Exponential Servers. *Communication of the ACM.* Vol. 16, pp. 527-531.

Carmichael, D., G. 1987. Machine Interference with general repair and running times. *Zeitschrift Operations Research.* Vol. 31, pp. B115-B133.

Elmaghraby, S., E. 1971. Operations Research. Chapter in: *Industrial Engineering Handbook, Third Edition.* H. B. Maynard (Ed.-in-Chief). McGraw-Hill. New York.

Gordon, W., J. and G. F. Newell. 1967. Closed Queuing Systems with Exponential Servers. *Operations Research.* Vol. 15, pp. 254-265.

Jackson, J., R. 1957. Networks of Waiting Lines. *Operations Research.* Vol. 5, pp. 518-521

Jacobson, P., A., and E., D. Lazowska. 1982. Analyzing Queueing Networks with Simultaneous Resource Possession. *Communication of the ACM.* Vol. 25, 2. pp. 142-151

Kamath, M., and J., L. Sanders. 1990. Modeling Operator/workstation Interference in Asynchronous Automatic Assembly Systems. Working paper. School of Industrial Engineering and Management. Oklahoma State University.

Kamath, M., R. Suri, and J., L. Sanders. 1988. Analytical Performance Models for Closed-Loop Flexible Assembly Systems. *The International Journal of Flexible Manufacturing Systems.* Vol. 1, pp. 51-84.

Kingman J., F., C. 1962. Some Inequalities for the Queue GI/G/1. *Biometrika.* Vol. 49, 3 and 4. 315-324.

Krämer W., and M. Langenbach-Belz. 1978. Approximate Formulae for General Single Server Systems with Single and Batch Arrivals. *Angewandte Informatik.* 9/1978. pp. 396-402.

Kuehn, P., J. 1979. Approximate Analysis of Queuing Networks by Decomposition. *IEEE Trans. on Communic.* COM-27, 1, pp. 113-126.

Law, A., M. 1986. Introduction to Simulation: A Powerfull Tool for Analyzing Complex Manufacturing Systems. *Industrial Engineering.* May 1986. pp. 46-63.

Marshall, K., T. 1968. Some Inequalities in Queuing. *Operations Research.* Vol. 16, pp. 651-665.

Pedgen, C., D. 1987. *Introduction to SIMAN.* Systems Modelling Corp.

Reiser, M. 1979. Mean Value Analysis of Queuing Networks, a New Look at an Old Problem. *Performance of Computer Systems.* pp. 63-77. M. Arato, A. Butrimenko, and E. Gelenbe (Eds.). IIASA and North-Holland Publishing Co.

Reiser, M., and S., S. Lavenberg, 1980. Mean-Value Analysis of of Closed Multichain Queueing Network. *Journal of the ACM.* Vol. 27, pp. 313-332.

Schweitzer, P., J. 1979. Approximate Analysis of Multiclass Closed Networks of Queues. Proceedings. *International Conference on Stochastic Control and Optimization.* Free University. Amsterdam.

Shantikumar, J., G., and J., A. Buzacott. 1980. On the Approximations to the Single Server Queue. *International Journal of Production Research.* Vol. 18, pp. 761-773.

Springer, M., C., and P., K. Makens. 1992. Queuing Models for Performance Analysis: Selection of Single Station Models. *European Journal of Operational Research.* Vol. 58, pp. 123-145.

Stecke, K., E. 1982. Machine Interference: Assignment of Machines to Operators. Chapter in: *Handbook of Industrial Engineering.* pp. 3.5.1-3.5.43.

Stecke, K. E., and J. E. Aronson. 1985. Review of Operator/Machine Interference Models. *International Journal of Production Research.* Vol. 23, 1, pp. 129-151.

Suri, R., and R., R. Hildebrant. 1984. Modelling Flexible Manufacturing Systems Using Mean-Value Analysis. *Journal of Manufacturing Systems.* Vol. 3, pp.27-38.

Suri, R. 1987. Rough-Cut Modeling: An Alternative to Simulation. *CIM Review.* Winter 1987. pp. 25-32.

Venkateswar, K., and J., L. Sanders. 1990. Optimization of Assembly Systems. Technical Report 90-7. Department of Industrial Engineering. University of Wisconsin-Madison.

Whitt, W. 1983. The Queuing Network Analyzer. *Bell Systems Technical Journal.* Vol. 62, pp. 2779-2815.

Wijbrands, R., J. 1988. Queuing Network Models and Performance Analysis of Computer Systems. PhD Thesis. Dept. of Math. and Computer Science. Eindhoven University of Technology. Eindhoven.

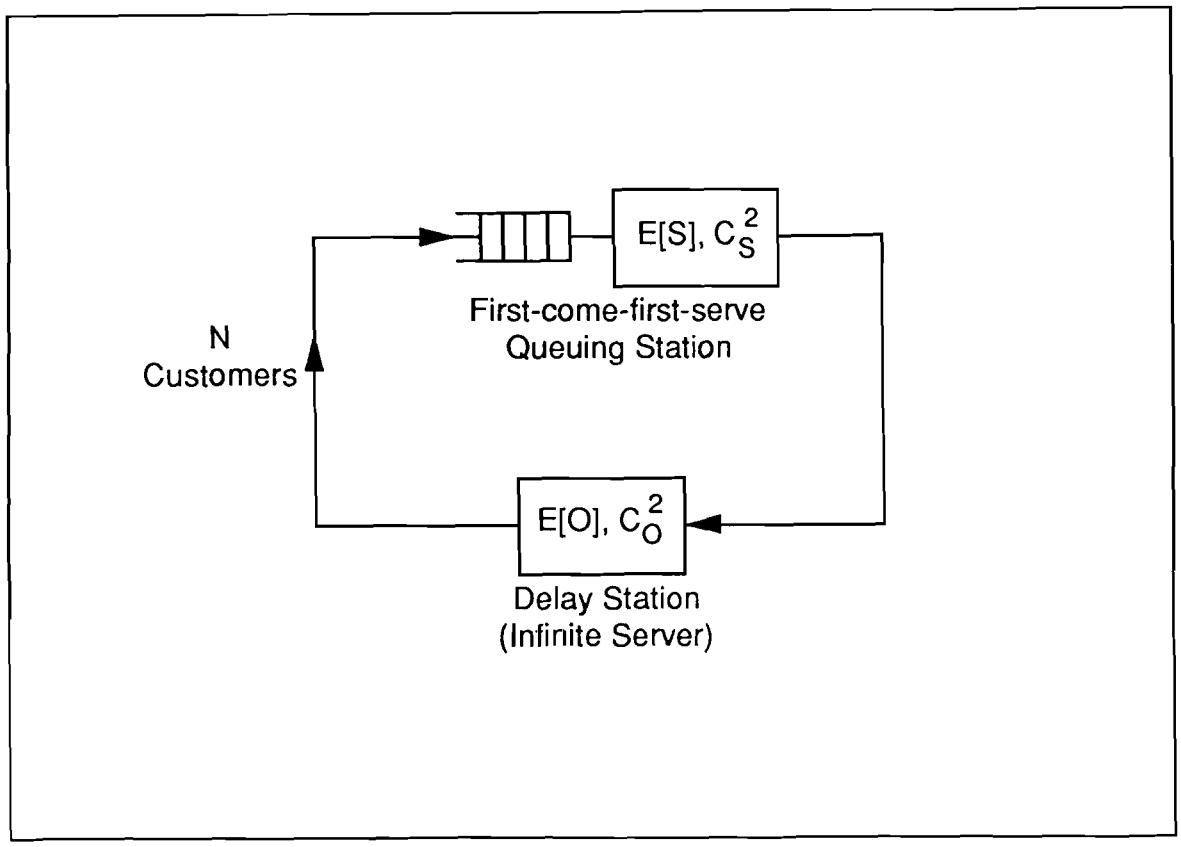

Figure 1. The analyzed network

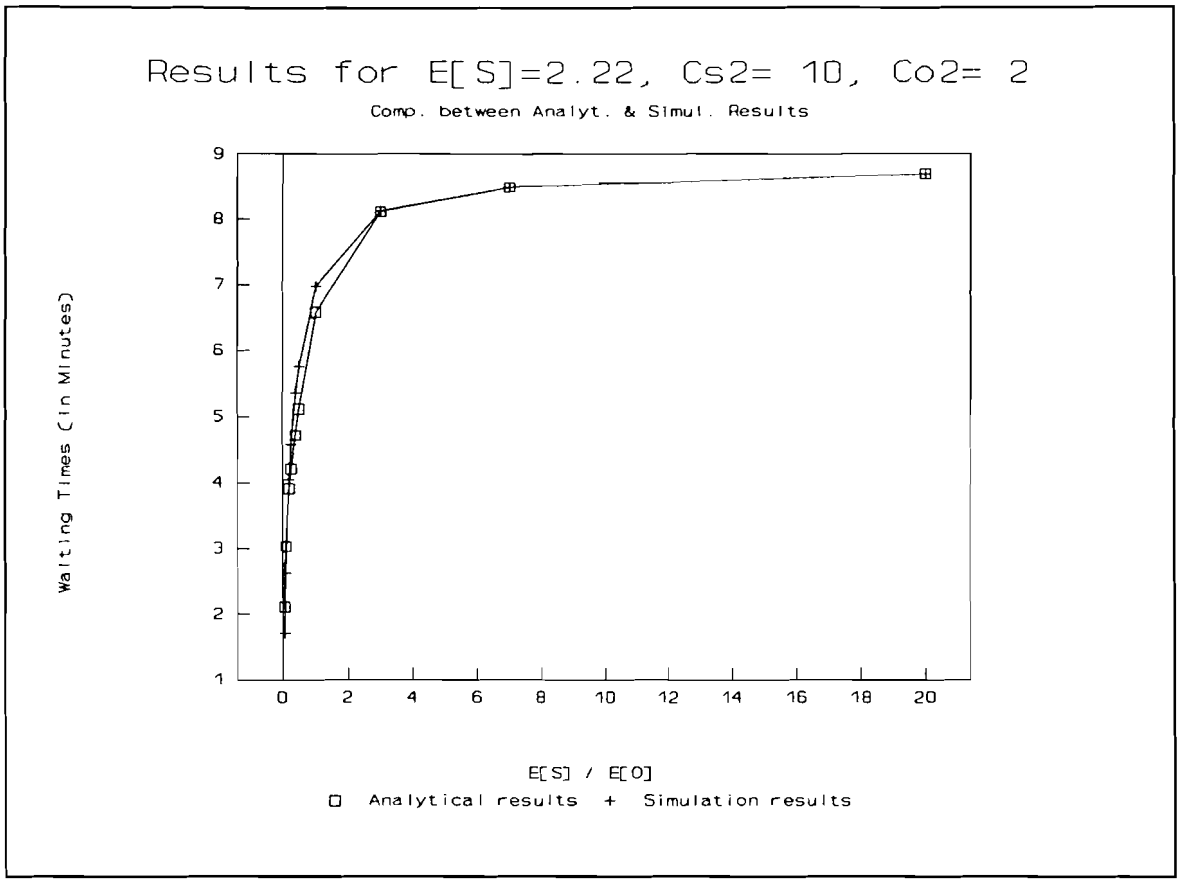

Figure 2 a)

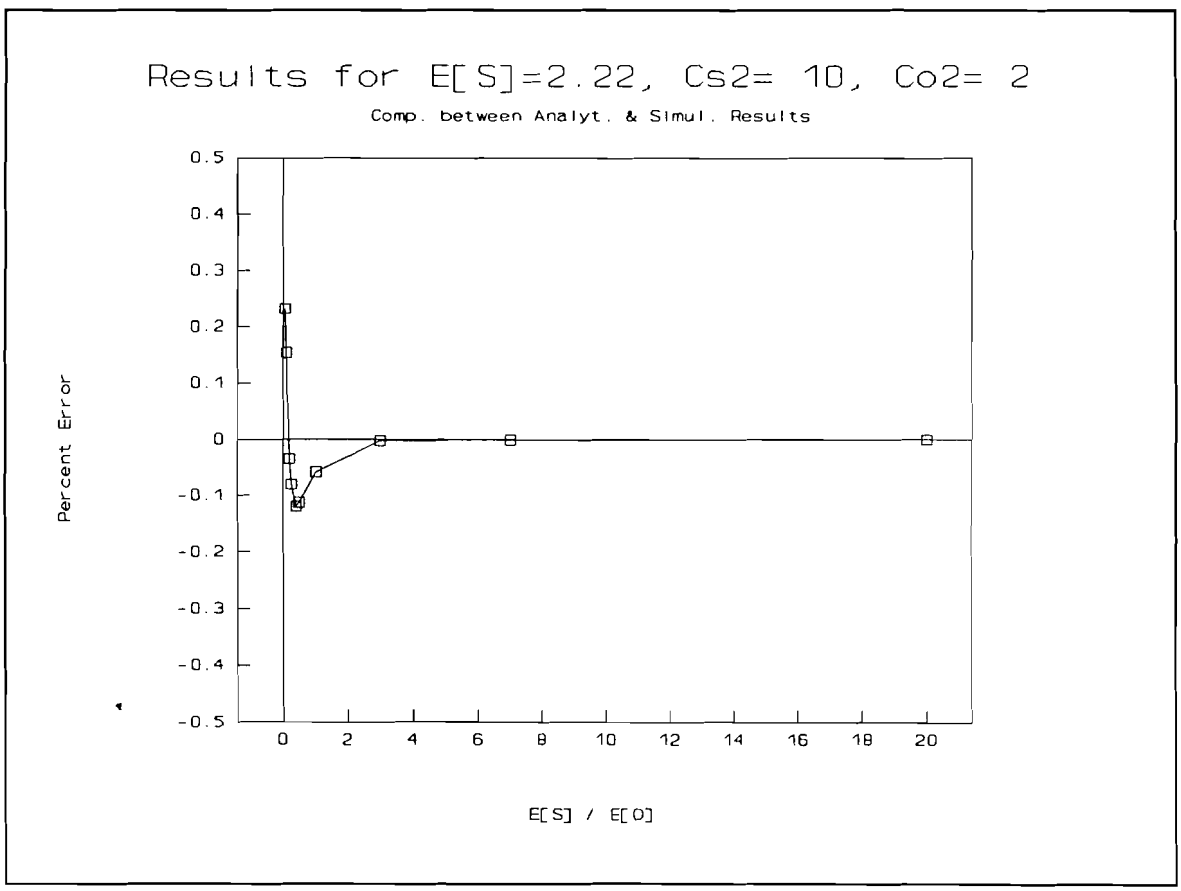

Figure 2 b)

Figure 2. Experimental results for different values of E[S]/E[O] ratio

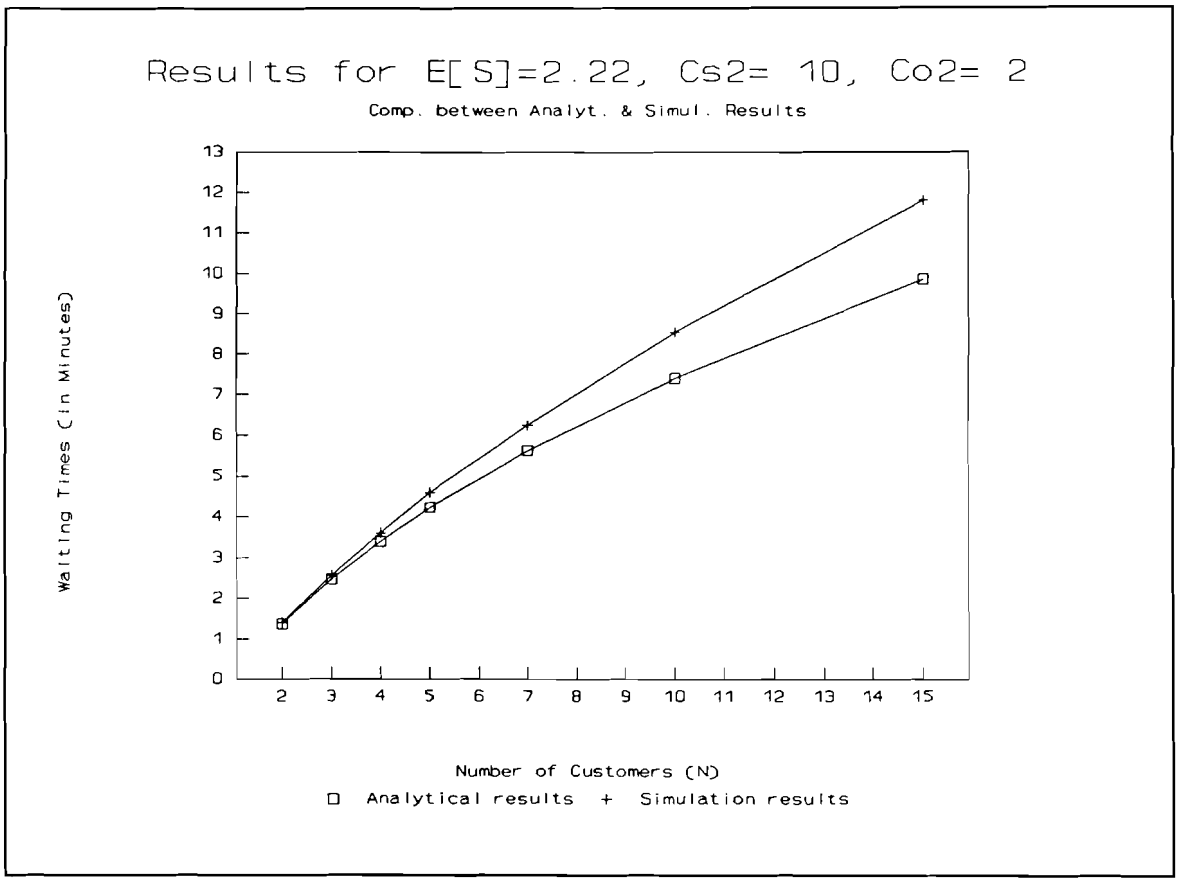

Figure 3 a)

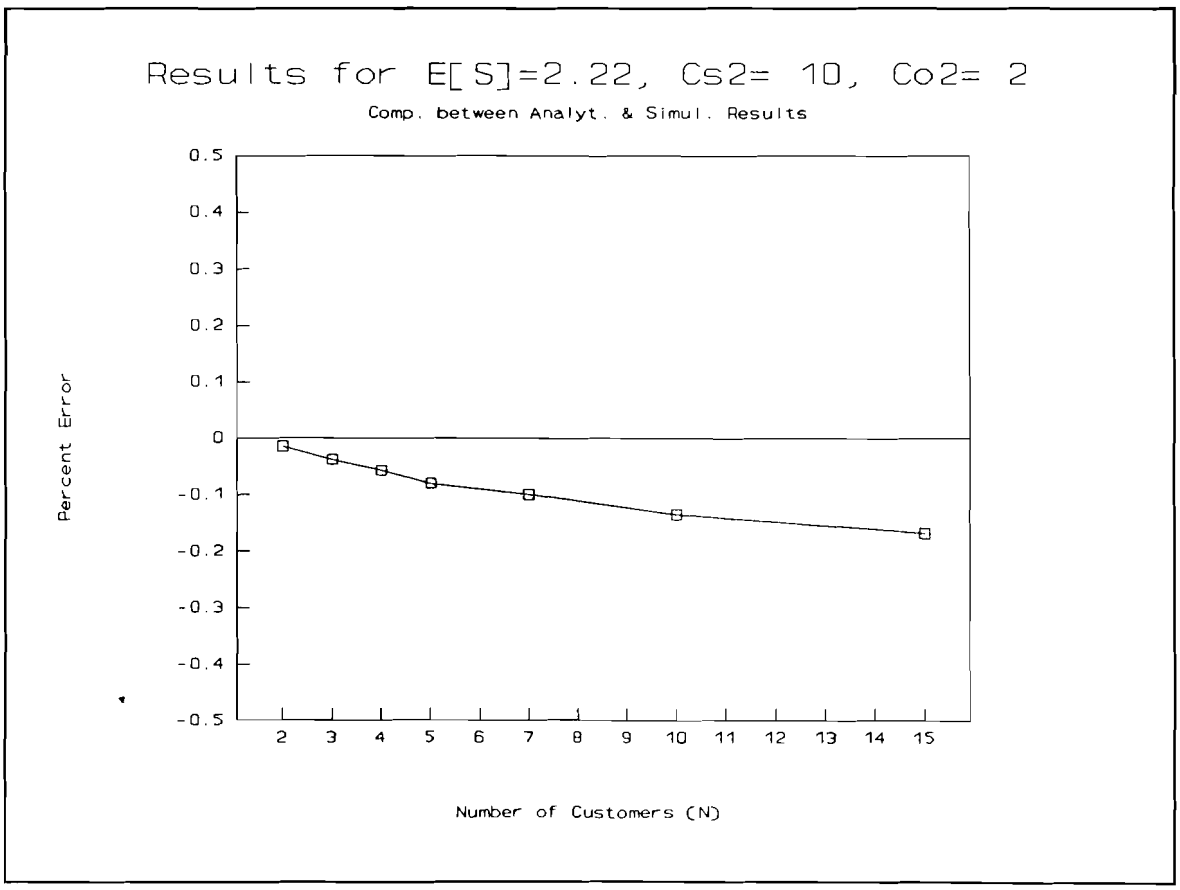

Figure 3 b)

Figure 3. Experimental results for different numbers of customers in the network (N)

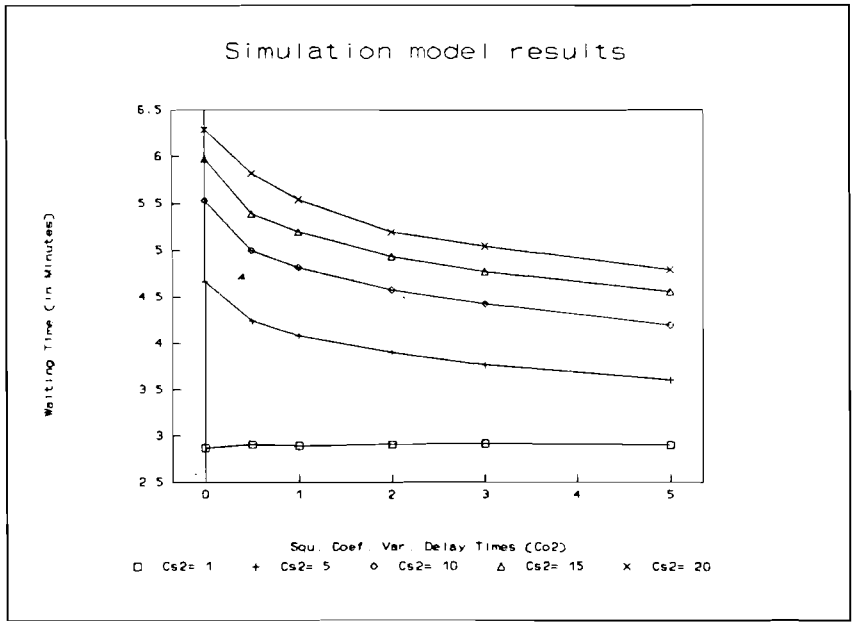

Figure 4 a)

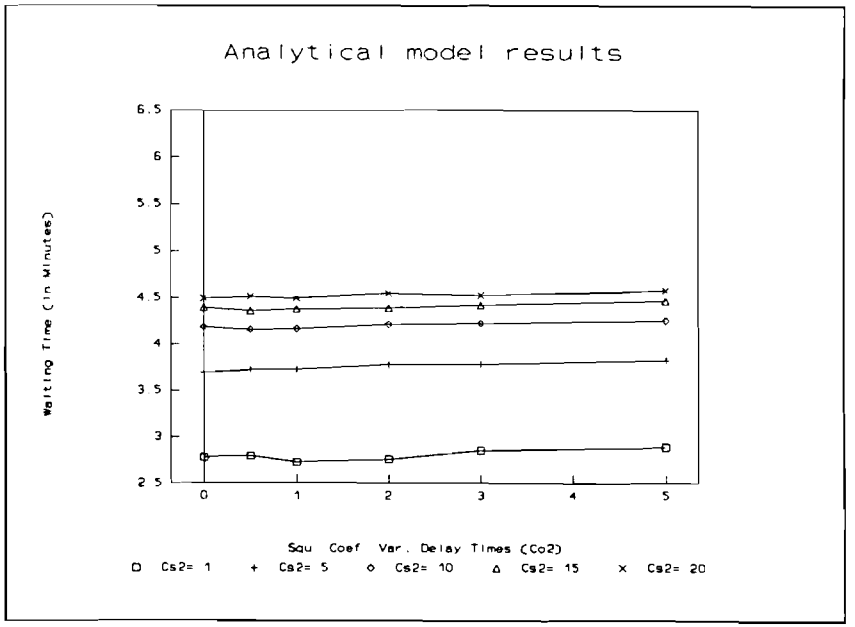

Figure 4 b)

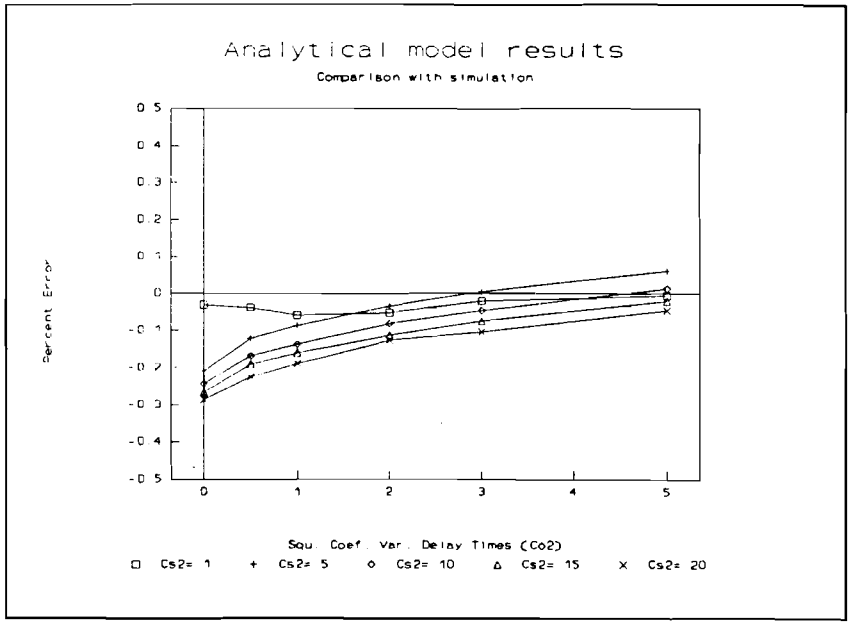

Figure 4 c)

Figure 4. Experimental results for different values of the scv's (C_o^2 on the horizontal axis)

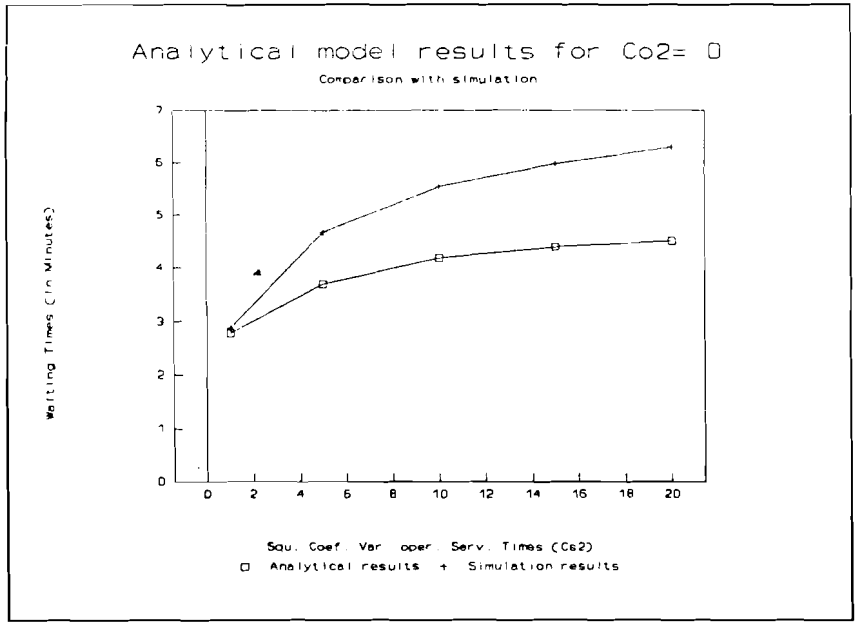

Figure 5 a)

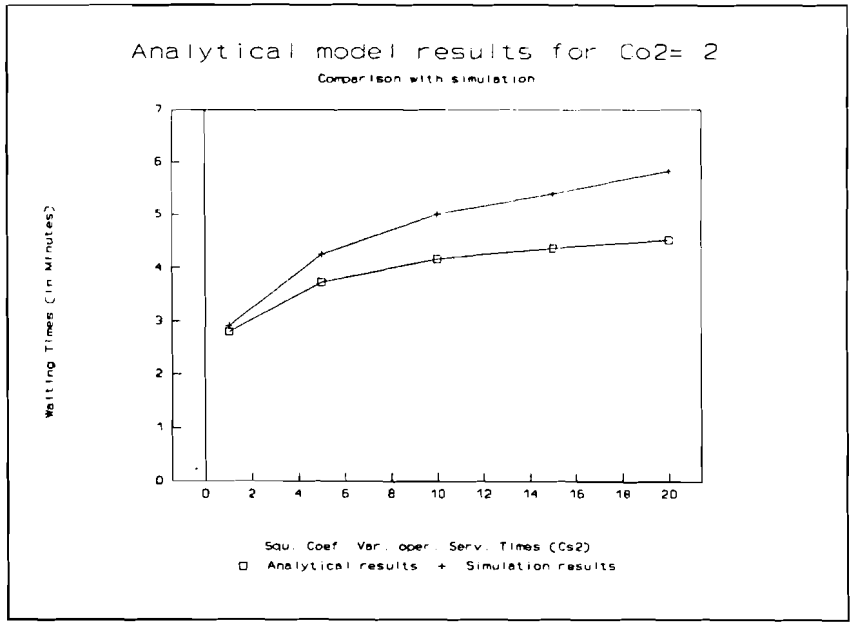

Figure 5 b)

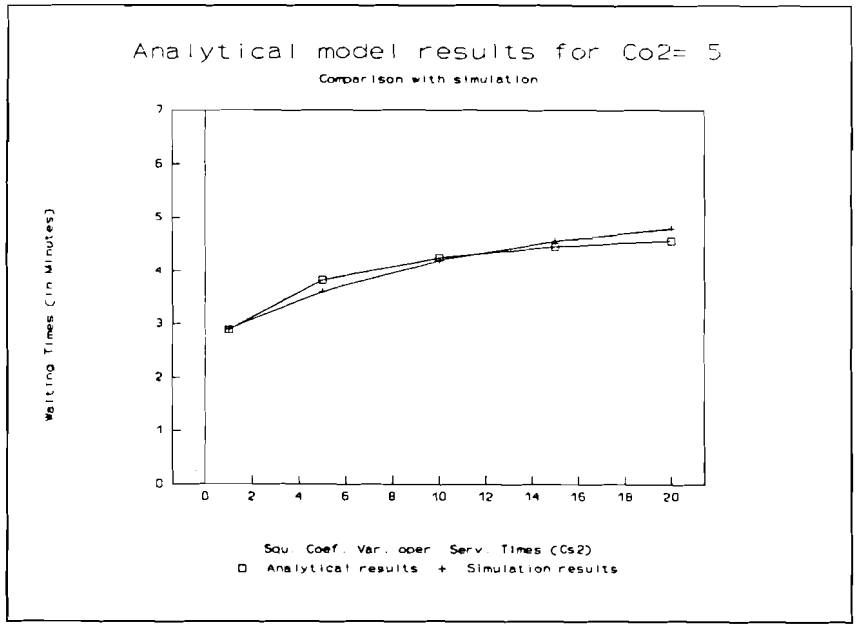

Figure 5 c)

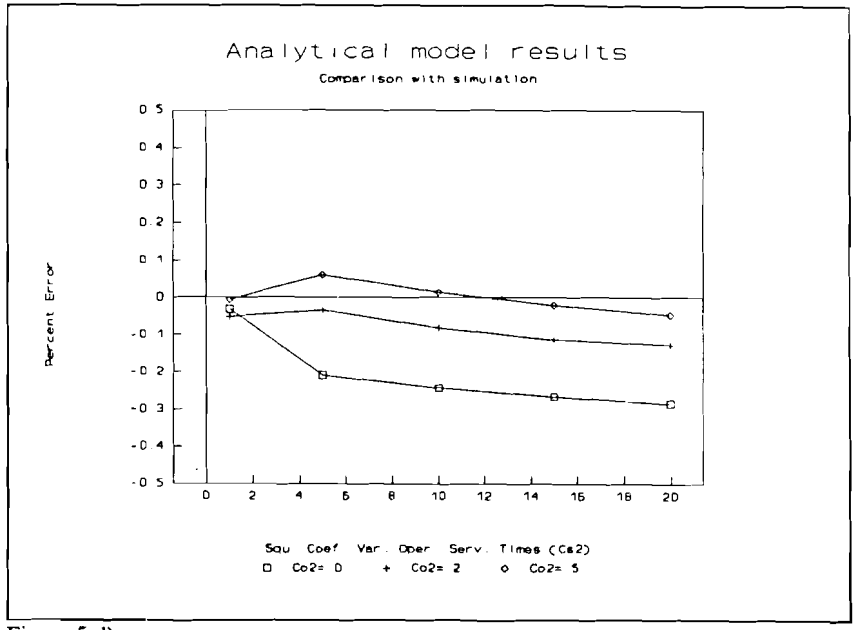

Figure 5 d)

Figure 5. Experimental results for different values of the scv's
(C_s^2 on the horizontal axis)