A mean value analysis of a closed CP-terminal system with pre-emptive resume priorities and general service time distributions
van Doremalen, J.B.M.

Published: 01/01/1984

Citation for published version (APA):
EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

MEMORANDUM COSOR 84 - 06
A MEAN VALUE ANALYSIS OF A CLOSED CP-TERMINAL SYSTEM WITH PRE-EMPTIVE RESUME PRIORITIES AND GENERAL SERVICE TIME DISTRIBUTIONS
by
Jan van Doremalen

Eindhoven, The Netherlands
May 1984
A MEAN VALUE ANALYSIS OF A CLOSED CP-TERMINAL SYSTEM WITH PREEMPTIVE RESUME PRIORITIES AND GENERAL SERVICE TIME DISTRIBUTIONS

Jan van Doremalen
Eindhoven, 1984

Abstract. This paper deals with the analysis of a queueing system which may be used as a model for a computer system with terminals, the so-called CP-terminal system, or a machine-interference model. The system can be viewed as a single server queue with a number of finite Poisson sources, general service time distributions and a preemptive resume priority scheduling. A recursive scheme will be derived for the computation of performance measures. The technique is based on mean value ideas and renewal arguments.

1. Introduction

This paper deals with the analysis of a queueing system which may be used as a rude model for a real-life system consisting of a number of terminals connected with a central processor, a so-called CP-terminal system. The system is pictured in Figure 1.

![Diagram of CP-terminal system](image)

Figure 1. The CP-terminal system.

There are L terminal groups, denoted as 1, 2, ..., L. The $K_t$ terminals of group $t$ have independent and negative exponentially distributed think times with parameter $\lambda_t$. The service times for jobs at the CP are stochastically independent and distributed according to a distribution function $G_\mathbb{X}$ for a job from terminal group $\mathbb{X}$. A preemptive resume priority scheduling is introduced as follows. A terminal of group $\mathbb{X}$, delivering a job at the CP, interrupts the service of a job from a terminal of group $k$ if $\mathbb{X} < k$. Group $\mathbb{X}$ is said to have a higher priority than group $k$. A terminal is only thinking if it has no job at the CP.
Service for a job is resumed as soon as no higher priority jobs are present anymore. For jobs within the same terminal group one may assume any work-conserving service discipline, for example first-in first-out or processor sharing.

Another application for this queueing network model is the machine-interference problem. An operator has the responsibility for several groups of machines. Based on some criterion, for instance the importance of the machine for the production process or the age of the machine, he gives priority to repairing machines of one group above machines of another group. However, for the sake of clarity we will use the terminology of the CP-terminal system throughout the paper.

We are interested in the consequences of preemptive resume priority rules on the utilization of the CP and the mean response times for the jobs of the groups at the CP. Thus, it is our main purpose to evaluate global performance measures and we are not after a detailed analysis of the stochastic behaviour.

The system can be seen as a single server queue with finite Poisson sources, preemptive resume priority scheduling and generally distributed service times. For the equivalent system with infinite Poisson sources, a M/G/1 priority queue, an elegant analysis based on mean value ideas is possible as we have shown in van Doremalen [1983]. Similar reasonings for example have been used in Wolff [1970], Stidham [1972] and Balachandran [1974]. For the system with finite Poisson sources and generally distributed service times, to our knowledge, only complicated results in terms of Laplace-transforms are known. In the monograph of Jaiswal [1968] an extensive study of such models has been presented. For the model with negative exponentially distributed service times at the CP Veran [1983] has proposed a recursive scheme to evaluate utilizations and mean response times. However, his results are based on a detailed analysis of the steady-state equations of the corresponding continuous-time Markov chain and do not give much insight.

We will derive a recursive scheme to evaluate performance measures for the CP-terminal system with single-terminal groups, i.e. with one terminal at each priority level, in Section 2. The analysis is based on mean value arguments and observations on renewal cycles in the stochastic process describing the behaviour of the queueing network.

In Section 3 we will show how the analysis can be extended to systems with multi-terminal groups. There is a restriction to this extension. The service time for jobs from terminal groups with more than one terminal has to be negative exponentially distributed. The reason for this restriction is that an essential step in
the analysis is the transformation of the service discipline within such a group with more terminals to a preemptive resume priority scheduling. In Section 4 we will draw some conclusions and will give references to other material and further research. Remarks concerning performance questions are made.

2. The analysis of a CP-terminal system with single-terminal groups

2.1 Introduction

In this section we will discuss an analysis of a CP-terminal system with L terminals, each belonging to a different priority class. A recursive scheme will be derived for the computation of the utilization of the CP and the mean response times for the jobs at the CP.

To get an intuition for the line of reasoning the following observations are made. First, note that the stochastic process describing the behaviour of the first \( k \) terminals is independent of the process describing the behaviour of the last \( L-k \) terminals. Of course, the reverse is not true. So the behaviour of the first \( k \) terminals may be analysed without any knowledge of the last \( L-k \) terminals.

The second observation is that the stochastic process describing the behaviour of terminal \( k \) can be studied by analyzing a closed cyclic queueing network with two single server queues and one customer. The service time at one queue describes the thinktime of the terminal. The other server, the CP, processes the job from the terminal, but is subject to breakdowns being busy periods of the first \( k-1 \) terminals.

These observations indicate the possibility of finding a recursive scheme to evaluate performance measures. This recursive scheme will be derived in Subsections 2.2 and 2.3.

2.2 The basic recursive scheme

The method to evaluate the performance measures is based on a mean value analysis of busy cycles of a special type. Before we present the method the following terminology is introduced to describe these busy cycles,

\[ B_k \text{-period} : \text{a busy period of the first } k \text{ terminals, i.e. an uninterupted period of time during which the CP is processing jobs from the first } k \text{ terminals.} \]
The basic observations are:
1. During a $C_k$-period either one job from terminal $t$ or no job from terminal $t$ is processed and
2. A $C_k$-period forms a renewal cycle in the stochastic process describing the behaviour of the first $k$ terminals. Note that the beginning of a $B_k$-period is a regeneration point in this process also.

Let us introduce the following notations to analyse this renewal process:

$$\Lambda_k := \sum_{i=0}^{i=k} \lambda_i$$ (where $\lambda_0 = 0$)

$u_k$ : fraction of time the CP is processing jobs from terminal $t$.

$$U_k := \sum_{i=0}^{i=k} u_i$$ (where $u_0 = 0$)

$w_k$ : $\int x dG_k(x)$, mean processing time of a job from terminal $t$.

$T_k$ : mean response time of a job from terminal $t$ at the CP, including service and waiting time.

Assume that we have analysed the behaviour of the first $k-1$ terminals and that we are interested in $u_k$. Note that a $B_k$-period may start with a terminal $t$ job, or that a terminal $t$ job may enter during a $B_k$-period, or that during a $B_k$-period no terminal $t$ job is processed at all. Analysing the $C_k$-period and conditioning on the first two events, we find that the probability that during a specific $B_k$-period a terminal $t$ job is processed is given by,

$$\pi_k = \frac{\lambda_k}{\Lambda_k} + \left(1 - \frac{\lambda_k}{\Lambda_k}\right) \pi_k$$

where $\pi_k$ is defined by,

$$\pi_k := \text{probability that during a specific } B_k \text{-period a terminal } t \text{ job is processed, given that the } B_k \text{-period does not start with a terminal } t \text{ job}$$

The evaluation of the probabilities $\pi_k$, $k=1,2,\ldots,L$, is the subject of Subsection 2.3. We now will proceed with the derivation of the basic scheme by introducing a...
a mean value analysis of busy cycles.

The number of $B_\ell$-periods per unit time equals the number of $I_\ell$-periods per unit time. This number is the quotient of the fraction of time no jobs from the first $\ell$ terminals are at the CP, $1 - U_\ell$, and the mean length of an $I_\ell$-period, $\lambda_\ell^{-1}$. The fraction of time the CP is processing jobs from terminal $\ell$ is the product of the mean number of $B_\ell$-periods per unit time, the probability that during a specific $B_\ell$-period a job from terminal $\ell$ is served and the mean service time of a job from terminal $\ell$, or in formula,

\begin{equation}
    u_\ell = (1 - U_\ell) \Lambda_\ell \left( \frac{\lambda_\ell}{\Lambda_\ell} + \left(1 - \frac{\lambda_\ell}{\Lambda_\ell}\right) \pi_\ell \right) w_\ell, \quad \ell = 1, 2, \ldots, L
\end{equation}

Using the relations $U_\ell = U_{\ell-1} + u_\ell$ and $\Lambda_\ell = \Lambda_{\ell-1} + \lambda_\ell$ we find, solving for $U_\ell$,

\begin{equation}
    U_\ell = \frac{U_{\ell-1} + \left(\lambda_\ell + \Lambda_{\ell-1} \pi_\ell\right) w_\ell}{1 + \left(\lambda_\ell + \Lambda_{\ell-1} \pi_\ell\right) w_\ell}, \quad \ell = 1, 2, \ldots, L
\end{equation}

Starting with $U_0 = 0$ and $\Lambda_0 = 0$ Relation 4 gives a recursive scheme to compute the utilization of the CP, provided we will be able to evaluate the probabilities $\pi_\ell$. One may verify that the mean response time $T_\ell$ of a job from terminal $\ell$ at the CP is given by,

\begin{equation}
    T_\ell = \frac{w_\ell}{u_\ell} - \frac{1}{\lambda_\ell}
\end{equation}

In the next subsection we will derive a recursive scheme to evaluate the $\pi_\ell$'s.

2.3 Evaluation of the probabilities $\pi_\ell$

The probability $\pi_\ell$ depends on the thinkrate $\lambda_\ell$ of terminal $\ell$ and the thinkrates and service time distributions at the CP of the first $\ell - 1$ terminals. It does not depend on the service time distribution at the CP of a job from terminal $\ell$. The lower priority terminals have no influence at all.

These observations lead to the idea of introducing auxiliary probabilities $\pi_\ell(x)$, for $\ell = 1, 2, \ldots, L$ and $x > 0$, which are defined by,

\begin{equation}
    \pi_\ell(x) \quad : \text{probability that during a } B_{\ell-1}\text{-period a customer from a Poisson process with parameter } x \text{ arrives.}
\end{equation}

We set $\pi_1(x) = 0$ by definition for $x > 0$. The probabilities $\pi_\ell$, $\ell = 1, 2, \ldots, L$, now correspond with the probabilities $\pi_\ell(\lambda_\ell)$, $\ell = 1, 2, \ldots, L$. 
The sequel of this subsection is devoted to the derivation of a recursive scheme to compute the probabilities $\pi_{g+1}(x)$. The method is based on an analysis of the $B_g$-periods.

The first job in a $B_g$-period is a job from terminal $k$ or a job from a terminal with a higher priority. Conditioning on the first job being from terminal $k$ or not, we find for $g=1,2,\ldots,L-1$ and $x\geq 0$,

$$\pi_{g+1}(x) = \frac{\lambda_k}{\Lambda_g} \psi_g(x) + \left(1 - \frac{\lambda_k}{\Lambda_g}\right) \phi_g(x)$$

where $\psi_g(x)$ and $\phi_g(x)$ may be interpreted as,

$$\psi_g(x) : \text{probability that during a } B_g \text{-period a customer from a Poisson process with parameter } x \text{ arrives, if the } B_g \text{-period starts with a job from terminal } k.$$  

and

$$\phi_g(x) : \text{probability that during a } B_g \text{-period a customer from a Poisson process with parameter } x \text{ arrives, if the } B_g \text{-period does not start with a job from terminal } k.$$  

Let us first have a closer look at the probabilities $\phi_g(x)$. For $g=1$ and $x\geq 0$ we initialize $\phi_1(x)=0$. For $g>1$ the $B_g$-period starts with a $B_{g-1}$-period if it does not start with a job from terminal $k$. At the end of this $B_{g-1}$-period we discern the following three possibilities,

i : with probability $\pi_g(x)$ a customer from a Poisson process with parameter $x$ has arrived during the $B_{g-1}$-period,

ii : with probability $\pi_g(x + \lambda_k) - \pi_g(x)$ a job from terminal $k$ has arrived and no customer from a Poisson process with parameter $x$ has arrived during the $B_{g-1}$-period,

iii : with probability $1 - \pi_g(x + \lambda_k)$ the end of the $B_{g-1}$-period coincides with the end of the $B_g$-period and no customer from a Poisson process with parameter $x$ has arrived during the $B_{g-1}$-period.
Conditioning on these three events we find as a relation for $\psi_\ell(x)$,

$$
\psi_\ell(x) = \pi_\ell(x) + \left( \psi_\ell(x + \lambda_\ell) - \pi_\ell(x) \right) \psi_\ell(x)
$$

Next we will concentrate on the evaluation of the probabilities $\psi_\ell(x)$. Here, for the first time, the explicit service time distributions will play a role in the analysis. To give a clear picture of the line of reasoning, we first will discuss the situation where the service times at the CP are negative exponentially distributed and next we will concentrate on the case with generally distributed service times.

Case 1: Negative exponentially distributed service times.

Assume that the service time of a job from terminal $\ell$ at the CP is negative exponentially distributed with parameter $\mu_\ell$, $\ell=1,2,\ldots,L$. If the $B_{\ell-1}$-period starts with a job from terminal $\ell$, then three things may happen: 1. the job of terminal $\ell$ is interrupted by a $B_{\ell-1}$-period, or 2. the job is interrupted by an arrival of the Poisson process with parameter $x$, or 3. the job is processed without interrupts. Conditioning on these three events we obtain for $\psi_\ell(x)$ the following relation,

$$
\psi_\ell(x) = \frac{\Lambda_{\ell-1}}{\Lambda_{\ell-1} + \mu_\ell + x} \left( \psi_\ell(x) + (1 - \pi_\ell(x)) \psi_\ell(x) \right) + \frac{x}{\Lambda_{\ell-1} + \mu_\ell + x}
$$

which may be transformed into

$$
\psi_\ell(x) = \frac{x + \Lambda_{\ell-1} \pi_\ell(x)}{x + \pi_\ell + \Lambda_{\ell-1} \pi_\ell(x)}
$$

Given the proposed initializations, the Relations 4, 7, 10 and 12 form a recursive scheme to compute the utilization of the CP. It is interesting to note that one may proof the equivalence of this scheme and the scheme proposed by Veran [1983:295].

Case 2: Generally distributed service times.

With $G_\ell$ being the service time distribution of a job from terminal $\ell$ at the CP we have for $\psi_\ell(x)$, conditioning on the service time duration of a job from ter-
minimally \( \ell \),

\[
\psi_\ell(x) = \int_0^\infty p_{\ell,y}(x) \, dG_\ell(y)
\]

where \( p_{\ell,y}(x) \) may be interpreted as,

\[
p_{\ell,y}(x) = \text{probability that during a specific } B_{\ell} \text{-period a customer}
\]

\[
of a Poisson process with parameter } x \text{ arrives, if this}
\]

\[
B_{\ell} \text{-period starts with a job from terminal } \ell \text{ of length } y.
\]

The service time of a job from terminal \( \ell \) at the \( CP \) is interrupted by jobs from a

higher priority level forming \( B_{\ell-1} \)-periods. Conditioning on the number of \( B_{\ell} \)-peri-

ods interrupting the service time of the job from terminal \( \ell \) of length \( y \), we

find,

\[
p_{\ell,y}(x) = \sum_{k=0}^{\infty} \frac{(A_{\ell-1}y)^k}{k!} \exp(-A_{\ell-1}y) p_{\ell,y,k}(x)
\]

where \( p_{\ell,y,k}(x) \) may be interpreted as,

\[
p_{\ell,y,k}(x) = \text{probability that during a specific } B_{\ell} \text{-period a customer}
\]

\[
of a Poisson process with parameter } x \text{ arrives, if this}
\]

\[
B_{\ell} \text{-period starts with a job from terminal } \ell \text{ of length } y
\]

which is interrupted by \( k B_{\ell-1} \)-periods.

Using renewal arguments one may verify that the following holds,

\[
p_{\ell,y,k}(x) = 1 - \exp(-xy) (1 - \psi_\ell(x))^k
\]

The Relations 13, 15 and 17 lead to the following relation for \( \psi_\ell(x) \),

\[
\psi_\ell(x) = \int_0^\infty \left( 1 - \exp(-(x + A_{\ell-1}\psi_\ell(x))y) \right) \, dG_\ell(y)
\]

If we write the Laplace-Stieltjes transform of \( G_\ell \) as \( \mathcal{G}_\ell^* \), i.e.

\[
\mathcal{G}_\ell^*(\sigma) = \int_0^\infty \exp(-\sigma y) \, dG_\ell(y) , \quad \sigma > 0
\]
then we can write $\psi_l(x)$ as,

$$
(20) \quad \psi_l(x) = 1 - G_x^l(x + \lambda_{l-1}^l \psi_l(x)).
$$

This relation is interesting because of the fact that it shows that for service time distributions with a rational Laplace-Stieltjes transform, the evaluation of $\psi_l(x)$ reduces to solving a system of linear equations. In van Doremalen [1984] we show how this linear system can be found by using the same technique as in the case of negative exponentially distributed service times, for the case of phase-type service time distributions.

3. The analysis of a CP-terminal system with multi-terminal groups

The exact analysis discussed in the previous section can be extended to cover the CP-terminal system with multi-terminal groups under certain restriction which will become apparent. A multi-terminal group consists of a number of terminals with the same specifications, i.e. the thinkrates and service time distributions are equal, and they have the same priority level. For jobs from terminals within a group the scheduling of the service at the CP is governed by another discipline.

The basic idea we would like to use in the analysis of systems with multi-terminal groups is a conversion of the service discipline at the CP within a specific group. Each terminal group is splitted in as many priority classes as there are terminals. The resulting model is a CP-terminal system with single-terminal groups, which can be analysed by the techniques developed in Section 2. An aggregation step provides the results for the original system with multi-terminal groups.

The essential point in the justification of the proposed analysis of multi-terminal systems is that it must be verified whether the utilization of the CP is influenced by the transformation of the service discipline or not. In general one may show that the service discipline indeed has an influence on the utilization of the CP. However, if the service times at the CP are negative exponentially distributed within a certain multi-terminal group and if the original service discipline is work-conserving, then the transformation will have no influence on the utilization of the CP. A service discipline is said to be work-conserving if it does not affect the total time spent in service of any job (cf. Wolff [1970]). Examples of such rules are the first-in first-out discipline, the processor sharing discipline and, to us a very important one, the preemptive resume priority scheduling. To
understand that the utilization is equal for all work-conserving service disciplines, note that the jobs of a terminal group with negative exponentially distributed service times at the CP are in a sense undistinguishable and therefore interchangeable at the CP.

We will conclude this section with some remarks on the evaluation of the utilizations and mean response times.

Let the tuple \((t, k)\) denote terminal \(k\) in group \(t\), where \(t\) is in the range of 1 to \(L\) and where, for given \(t\), \(k\) is in the range of 1 to \(K_t\). An auxiliary priority rule is introduced as follows. Terminal \((t_1, k_1)\) has a higher priority than terminal \((t_2, k_2)\) if \(t_1 < t_2\) or, if \(t_1 = t_2\), if \(k_1 < k_2\). Now the technique of Section 2 may be used to obtain the utilizations for each specific terminal \((t, k)\) say \(u_{t, k}\).

In the aggregation step the utilization for terminal group \(t\), say \(u_t\), is set to

\[
(21) \quad u_t = \sum_{k=1}^{K_t} u_{t, k}
\]

and the mean response time of a job from a terminal of group \(t\), say \(T_t\), is given by,

\[
(22) \quad T_t = \frac{K_t \mu_t}{u_t} - \frac{1}{\lambda_t}.
\]

4. Conclusions and remarks

Based on mean value ideas and renewal arguments, a recursive scheme has been developed for the evaluation of relevant behavioral characteristics in a CP-terminal system with preemptive resume priorities. It should be observed that the evaluation of the recursive scheme is not trivial. Of course, it is possible to implement the recursion by brute force in a programming language with a recursion feature like ALGOL 68. For larger values of \(L, K_1, \ldots, K_L\) this strongly has to be dissuaded, as the complexity of the algorithm and the storage requirements of the resulting implementation would be enormous. In van Doremalen [1984] we have discussed an elegant algorithm, which is rather efficient.

Using the same line of argument as in Section 2, the analysis of the CP-terminal system with single-terminal groups can be extended to other priority rules, for example the preemptive repeat and the preemptive identical repeat priority rules.
The analysis of the queueing network as presented in this paper is interesting as it gives an extension of the exact analysis of known models. But we have to make some remarks with respect to the validity of the model and its merits in studying the performance of computer network models.

For the CP-terminal system with negative exponentially distributed service times at the CP, some nice results have been derived for the nature of optimal priority schedulings. Van der Wal [1982] and Courcoubetis, Varaiya and Walrand [1982] give schedulings, which are optimal in a sense that they maximize the CP utilization. For the model presented in this paper such schedulings are not known. However, a comparison of schedulings is possible and the influence of the distributions of the service times on a specific scheduling can be studied.

A more profound problem in using the proposed model is that the preemptive priority scheduling rule is static in a sense that it only depends on fixed priority labels. It will be clear that, in general, a dynamic scheduling will have to be recommended. One may think of a scheduling using elapsed service times or a more detailed description of the state of the CP. Furthermore, the model of the network is very simple and the influence of the priority rules on the behavior of the CP as a more complex system consisting of several central processor units and I/O devices might have to be taken into account.

This brings us to our last remarks. For more complicated models and more advanced schedulings an exact analysis, if possible at all, will be beyond the possibilities of nowadays computational techniques and, therefore, approximate methods will have to be developed.

The interpretation of the recursive scheme leads to the development of such approximation methods for the CP-terminal system and more complex queueing networks with priority schedulings in general. Some promising numerical results have been obtained so far and further research is in progress.

5. References


J. van der Wal [1982], The maximization of CP-utilization in an exponential CP-terminal system with different think times and different job sizes. Memorandum COSOR 82-14, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., Eindhoven.