A SIMPLIFIED MODEL OF DISCRETE-TIME RESERVATION SYSTEMS

Miguel A. Díaz Quiñones
Departamento de Investigación y Desarrollo
Telefónica
Madrid, Spain

A method to model slotted time reservation systems is presented. Some simplifying hypotheses are assumed (constant service time, bounded notice interval and bounded transaction space), which lead to simple analysis based on Markov chains and discrete convolutions. Some results and curves are also presented.

1. INTRODUCTION

The analysis of a general reservation system is difficult, and simulations or approximate models must be used in order to gain insight of system behaviour.

A useful approach ([1], [2]) is to assume that the time axis is slotted into a number of fixed size intervals. We present in this paper a simplified model based on the following additional assumptions:

- There is an upper limit M of feasible notice intervals.
- There is also an upper limit L for the number of transactions allowed to be present in the system at a given time slot.
- Service time is assumed to be constant and equal to one time slot.
- The arrival process is supposed to be a Poisson process.

With those assumptions, and taking into account the memoryless property of Poisson process, it is easy to verify that the number of transactions present in the system at a given time slot (either being served or awaiting service in subsequent time slots) is a Markov chain.

2. MODEL DESCRIPTION

In [1], it is proved that, subject to the above hypotheses, service requests can also be described as a Poisson process. It is also observed that the system can be modelled as a queue, identifying arrivals to the queue with requested service times, and taking into account notice intervals by means of a priority scheme.
2.1. Analysis of the Markov Chain

Let $N$ be the number of servers, $\{P_k\}$ the probabilities for the Poisson process and $C$ the Markov chain transition matrix. The elements of $C$ may be computed from the Poisson probabilities in the following way:

$$
C_{ik} = P_k \quad \text{for } 0 \leq i \leq N, 0 \leq k \leq L
$$

$$
C_{ik} = 0 \quad \text{for } N+1 \leq i \leq L, 0 \leq j \leq i-N-1
$$

$$
C_{ik} = P_{k-i-N} \quad \text{for } N+1 \leq i \leq L, i-N \leq k \leq L
$$

The explanation follows easily from the fact that service time is always one slot, and taking into account that at most $N$ servers are released at the end of each time slot.

Let $\{q_n\}, n = 0, 1, ..., L$ be the stationary probabilities of having $n$ transactions present in the system during a certain time slot. Such probabilities can be computed [4] solving the following matrix equation:

$$
Q = QC
$$

where $C$ is the previously defined transition matrix, and $Q$ a row vector whose elements are $q_0, q_1, ..., q_L$.

2.2. Computation of the waiting probabilities

Let us assume that, at a certain time $t$, a transaction arrives to the system (i.e., it requested to be served at time $t$). We will call this transaction "the reference transaction", and we will further assume that it had a notice interval of $k$ time slots. Referring to that particular transaction, we may define the following probabilities:

$\{a_n\}, n=0, 1, ..., M$: $a_n$ is the probability that the reference transaction has a notice interval of $n$ time slots. ($M$ is the maximum allowed notice interval).

$\{u_n\}, n=0, 1, ..., L-N$: $u_n$ is the probability that, when the reference transaction arrives to the system, there are $n$ transactions in the system from previous slots with a higher priority.

$\{v_n\}, n=0, 1, ..., L$: $v_n$ is the probability that $n$ transactions arrive during the same time slot that the reference transaction does it and with the same notice interval, but before it.

$\{w_n\}, n=0, 1, ..., L$: $w_n$ is the probability that $n$ transactions arrive during the same time slot that the reference transaction, but with higher priority. (i.e., they made their request before).

$\{t_n\}, n=1, 2, ..., L$: $t_n$ is the probability that, after having ordered the transactions arrived during the same time slot according with their priorities, the reference transaction is in the position $n$. 

4.4B.2.2
\( r_n \) is the probability that, after having ordered all the transactions, the reference transaction is in position \( n \) in the queue.

First of all, we evaluate the probabilities \( \{w_n\} \). If we call
\[
\mu' = \mu \sum_{m \geq k} a_m
\]
we have
\[
w_n = \frac{e^{-\mu'T} \mu'T^n}{n!}
\]

On the other hand, let us define:
\[
P''(\mu a_n) = \frac{e^{-\mu'T} \mu'T^n}{n!}
\]

In order to compute the probabilities \( \{w_n\} \), we make use of the uniformity property over a fixed size interval of a fixed number of Poisson arrivals. From that, if follows easily:
\[
v_n = \sum_{n=0}^{n+1} P''(\mu a_n), \quad 0 \leq m \leq L
\]

Finally, the probabilities \( \{t_n\} \) may in turn be evaluated by means of a discrete convolution:
\[
t_n = \sum_{m=0}^{n-1} w_m \cdot v_{n-1-m} \quad (n=1, 2, \ldots, L)
\]

In order to compute the probabilities \( \{u_n\} \), we define \( b_k \) as the probability that a transaction which is waiting in the system has a priority lower than that of the reference transaction. This probability depends on the number of slots that the waiting transaction has waited, which in turn depends on the waiting time distribution. Therefore, we must use an approximation; we have used the following one:
\[
b_k = \sum_{m \leq k} a_m
\]

which represents the probability that an incoming transaction has a notice interval lower than that of the reference transaction.

After computing \( b_k \), the probabilities \( u_n \) can be computed as follows:
\[
\begin{align*}
U_0 &= \sum_{m=0}^{L-N} q_{m} + \sum_{m=1}^{L-N} q_{m} b_k, \\
U_n &= \sum_{m=n}^{L-N} \left[ \begin{array}{c} m \\ n \end{array} \right] q_{N+m} (1-b_k) b_k \quad (n=1, \ldots, L-N)
\end{align*}
\]

where \( \left[ \begin{array}{c} m \\ n \end{array} \right] \) is a combinatorial number.

Provided that the probabilities \( \{u_n\} \) are known, the probabilities \( \{r_n\} \) can also be computed by means of a discrete convolution:
\[
r_n = \sum_{m=0}^{n-L-N} \sum_{n=1}^{L-N} \left[ \begin{array}{c} m \\ n \end{array} \right] q_{N+m} (1-b_k) b_k \quad (n=1, 2, \ldots, L)
\]

From the definitions of the above probabilities, it easy to compute the waiting probability for a transaction with notice interval \( k \):
\[
P_w = 1 - \sum_{m=1}^{\infty} r_m \quad \text{and therefore: } P_w = \sum_{k=1}^{M} (k r_1, \ldots, r_M) \frac{k}{a_k}
\]

4.4B.2.3
2.3. Computation of the mean waiting times

In order to compute mean waiting times, let us define the following set of probabilities:

\( \{d_n\}, \quad n=1,2,\ldots \): \( d_n \) is the mean waiting time, measured as number of time slots, of the reference transaction when, at the beginning of the next slot it is at the \( n \)-th position, assuming that it had a notice interval of \( k \) time slots.

To compute the probabilities \( \{d_n\} \), we use a recurrent scheme:

For \( k = M \) (the maximum allowed notice interval), the expected waiting time, for a transaction located in position \( n \), will be:

- 0 slots, if \( 1 \leq n \leq N \);
- 1 slot, if \( N+1 \leq n \leq 2N \), etc.

or, equivalently \( d_n = I (n-1)/N \), where \( I \{x\} \) stands for the integer modulus.

For \( k < M \) we have: If \( 1 \leq n \leq N \), \( d_n = 0 \). If \( n \geq N \), we define:

\[
\hat{S}_m = \frac{e^{-\mu'T} (\mu'T)^m}{m!}, \quad \text{with} \quad \mu' = \mu \sum_{m \geq k} a_j
\]

\( \hat{S}_m \) represents the probability of \( m \) transactions arriving during the same time slot than the reference transaction but with a higher notice interval. Besides, if a transaction has a mean waiting time of \( d_k \) slots at the beginning of a slot, at the beginning of the next one its mean waiting time will be \( d_{n-N} \) where \( m \) is the number of newly arrived transactions with a higher priority. Therefore, averaging over \( m \) we have (for \( n > N \)):

\[
d_n = 1 + d_{n-N} S_0 + d_{n-N} S_1 + \ldots + d_{n-N} S_{n-N}.
\]

Thus, we can compute \( \{d_n\} \), then \( \{d_{n+1}\} \), etc. The overall mean waiting time may be computed as follows:

\[
E[\hat{d}^k] = \sum_{m=1}^{k} r_m d_m
\]

and the mean waiting time for the delayed transactions will be:

\[
E[\hat{d}/\text{delayed}] = E[\hat{d}^k] / \sum_{m=1}^{k} P_m
\]

The overall mean waiting time for all notice intervals may be computed the same way as in the case of waiting probabilities, by averaging over the index \( k \):

\[
E[d] = \sum_{k=1}^{M} a_k E[\hat{d}^k] \quad \text{and} \quad E[d/\text{delayed}] = \sum_{k=1}^{M} a_k E[\hat{d}/\text{delayed}]
\]

2.4. Computation of the waiting time distribution

Let \( f_k(n) \) be the probability density function (p. d. f.) for the waiting time \( n \) (measured in time slots), for a transaction which had a notice interval of \( k \) time slots and which, after reordering as explained before, was placed in position \( m \) in the queue. We have in fact a family of p. d. f. for the waiting time, which may be computed recurrently.
For $k = M$ (the maximum allowed notice interval) we have:

$$f_m(n) = 1 \quad \text{if} \quad I^{m-1}_N = n, \quad f_m(n) = 0 \quad \text{otherwise.}$$

For $k < M$ and $1 \leq m \leq N$:

$$f_m(0) = 1, \quad f_m(n) = 0, \quad n > 0$$

For $k < M$ and $m > N$:

$$f_m(n) = \sum_{j=0}^{k-1} g_j f_{m+j}(n-1)$$

since a transaction which at a given time slot will wait $n$ time slots, at the
following one will wait only $n-1$ time slots.

The overall p.d.f. may be computed averaging over the indexes $k$ and $m$:

$$f(n) = \sum_{k=0}^{M} a_k \sum_{j=1}^{k} r_j f_j(n)$$

3. APPLICATION OF THE MODEL

3.1. Analysis of truncation errors

We have analized the effect over model performance of the limitations on the
number of transactions allowed to be present in the system. The values obtained
for the waiting probabilities are shown in the following table, for notice
intervals with geometric distribution. The table was computed for 6 servers, a
maximum notice interval of 34 slots and a traffic per server of 0.9 erlang.

<table>
<thead>
<tr>
<th>No. transactions</th>
<th>Waiting Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.290204</td>
</tr>
<tr>
<td>30</td>
<td>0.304244</td>
</tr>
<tr>
<td>40</td>
<td>0.306147</td>
</tr>
<tr>
<td>50</td>
<td>0.306399</td>
</tr>
</tbody>
</table>

3.2. Binomial distribution

We have studied a system with binomial notice probabilities, as described in
[11], pg. 61. The results obtained for a system with $N = 4$ servers, $\mu T = 3, p = 0.5$ are the following:

<table>
<thead>
<tr>
<th>M</th>
<th>$P_w$</th>
<th>$E [d]$</th>
<th>$E [d / delayed]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2359</td>
<td>0.2669</td>
<td>1.1325</td>
</tr>
<tr>
<td>3</td>
<td>0.1942</td>
<td>0.2803</td>
<td>1.2764</td>
</tr>
<tr>
<td>0</td>
<td>0.1904</td>
<td>0.2862</td>
<td>1.2501</td>
</tr>
</tbody>
</table>

which compare well with the equivalent results from [11]:

<table>
<thead>
<tr>
<th>M</th>
<th>$P_w$</th>
<th>$E [d]$</th>
<th>$E [d / delayed]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2359</td>
<td>0.2697</td>
<td>1.1434</td>
</tr>
<tr>
<td>5</td>
<td>0.2007</td>
<td>0.2720</td>
<td>1.3550</td>
</tr>
<tr>
<td>0</td>
<td>0.1914</td>
<td>0.2723</td>
<td>1.4229</td>
</tr>
</tbody>
</table>

3.2. Three-point probability distributions

We have obtained results for systems with notice intervals which follow a three-
point distribution. To characterize such a distribution, we use two parameters: the variance of the notice interval and the ratio between maximum and minimum probabilities, which we call R.

![Graph](image)

**FIGURE 1**
Waiting probabilities for 3 points of notice interval. R=2.5, variance=4, N=4 servers.

In Figure 1, the waiting probabilities for systems with 4 and 6 servers are shown. Those results were obtained for a ratio R = 2.5, and variance of 4. In Figure 2, we show the mean waiting time, also with R = 2.5, but for a system with 4 servers. In addition to the overall mean waiting time, the mean waiting time for transactions with the maximum and minimum notice interval are also shown on the upper, middle and lower curves respectively. On the other hand, in Figure 3, the mean waiting time for the delayed transactions is shown, as well as the mean waiting time for transactions with the minimum notice interval.

![Graph](image)

**FIGURE 2**
Mean waiting time for 3 points of notice interval. R=2.5, variance=4, N=4 servers.
Mean waiting time for the delayed transactions, for 3 points of notice interval. Variance=4, N=3 servers, μT/N=0.9. Up. curve: overall mean waiting time. Lower curve: mean waiting time for transactions with minimum notice time.

4. SUMMARY

A model has been presented to analyze discrete time reservation systems. The model has been applied to different systems and notice time interval distributions.

REFERENCES


ACKNOWLEDGEMENTS

A previous version of the model was presented in the COST 214 project. We acknowledge the most useful comments received from COST 214 chairman, Mr. James W. Roberts, as well as from the other members of the project.