THROUGHPUT LIMITATION BY HEAD-OF-LINE BLOCKING

J. Petersen

Universität des Saarlandes, Nachrichten- und Vermittlungstechnik, Im Stadtwald, D6600 Saarbrücken, Germany

In a high capacity packet switch or a multiprocessor system, efficient interconnection between input and output ports or between cooperating processors is required. Even if the interconnection network is nonblocking the transfer of packets or of interprocessor messages, queued at the input ports and served in the order of arrival, is hampered by head-of-line (HOL) blocking. The scope of the paper is to contribute to the understanding of HOL blocking by extending known results to the case of other than constant or neg.-exp. distributed service time. The main result is a formula which approximates the maximum throughput very well if the coefficient of variation of the service time is smaller than or not much larger than 1; the approximation error decreases if the number of ports is increased.

Introduction

In a high capacity packet switch or a multiprocessor system, efficient interconnection between input and output ports or between cooperating processors is required. Even if the interconnection network is nonblocking the transfer of packets or of interprocessor messages, queued at the input ports and served in the order of arrival, is hampered by head-of-line (HOL) blocking. If the first item in the queue is waiting for its output to become free it will prevent the transfer of another item which might be addressed to a free output.

This disadvantage of input queueing is well known; an impressive example of bringing attention to it was the comparison of input and output queueing in /1/. The discussion of input queueing, dating back to 1970 /2/, concentrated on the question of the maximum achievable throughput, the service time mostly being assumed as constant.

The scope of the paper is to contribute to the understanding of HOL blocking: Known results for the maximum throughput are extended to the case of other than constant or neg.-exp. distributed service time. Given is a formula which approximates the maximum throughput very well if the service time follows a phase-type distribution with a coefficient of variation between 0 and 1 and, depending on the type of distribution, also above 1. The approximation can be interpreted in terms of the mean waiting time on the HOL position.

The paper is organized as follows: After a system definition (section I) known results are summarized (section II); their extension to other service time distributions (section III) is the main part of the paper. In section IV, the approximation is compared with analytical or simulation results. The conclusions are followed by an annex containing some remarks on the mean waiting time on the HOL position.

I Systems with input queueing

In the systems under consideration here, K input ports are connected via a nonblocking interconnection network, e.g. a crossbar, to N servers. The service times of all servers follow the same distribution. Since we are interested in the maximum throughput it suffices to assume an arrival process at the input ports such that the HOL positions of the input queues never empty. The requests arriving here are addressed with probability 1/N to one of the servers; the destinations of all requests are independent of one another. The term output queue is used for the set of all HOL requests addressed to the same server. The arrival process to these output queues depends on the service time distribution, the arrival rate depends on the number Y of busy servers.

Throughput is defined here as the mean number of requests served during an interval of time equal to the mean service time, i.e. throughput equals the mean value of Y. The maximum throughput is therefore designated by ymax. pmax designates the corresponding utilization factor of a server, pmax = ymax/N.

II HOL blocking with constant or neg.-exp. distributed service times

We summarize known results here with the intention to show their role as a base for the extension in the following section. We consider first systems with a very large number of input and output ports. Those systems are of special interest because here the N output queues can be regarded as independent of one another (see /1/, /3/, /4/ for the case of constant service times). We follow the approach of /3/.

The authors show that for very large K = N the variance of the number of busy servers, divided by N, approaches zero (constant arrival rate to the output queues):

\[ \lim_{N \to \infty} \frac{\text{var}(Y)}{N} = 0 \]  

They show further in detail for the case of constant service times that the arrival process of each output queue becomes a Poisson process for N \( \to \infty \). Each output queue and its server can, therefore, be modeled as a system M/D/1.
Let Q denote the number of waiting requests, \( q = EQ \), then in a System M/D/1,

\[
q = \frac{Q^2}{2(1 - \rho)} \quad (2)
\]

Here, \( \rho = \rho_{\text{max}} \), and since there are no empty input queues,

\[
\rho_{\text{max}} + Nq = K
\]

\[
\rho_{\text{max}} + q = K/N \quad (3)
\]

and with (2) follows the well known result

\[
\rho_{\text{max}} = 2 - 2^\frac{1}{2}
\]

for \( K = N \) and otherwise

\[
\rho_{\text{max}} = 1 + \frac{K}{N} \left( 1 + \left( \frac{K^2}{N} \right) \right)^{1/2} \quad (4)
\]

If the number of input and output ports is finite and \( K > 1, N > 1 \) the situation is in general much more complicated because the output queues are not independent of one another. Exact results for systems with a finite number of ports and constant service times are due to Bhandarkar /5/.

Only the minimum number of states are distinguished, namely

\[
J(K, N) = \sum_{m=1}^{N} p(K, m)
\]

where \( p(K, m) \) denotes the number of partitions /7/ of \( K \) (HOL requests) in \( m \) positive integers (busy servers). A state is described by the vector of output queue lengths. \( \rho_{\text{max}} \) is computed by setting up a time discrete Markov chain and determining numerically the state probabilities.

Several closed form approximations for \( \rho_{\text{max}} \) are known, notably the one by Rau /6/:

\[
R(K, N) = \frac{\sum_{m=1}^{L} m \cdot 2^{m-1} \binom{K-1}{m-1} \binom{N}{m}}{\sum_{m=1}^{L} 2^{m-1} \binom{K-1}{m-1} \binom{N}{m}} \quad (5)
\]

where \( L = \min(K, N) \). With the usual convention

\[
\binom{n}{m} = 0 \quad \text{for } m > n
\]

one can use \( N \) as the upper limit in the summation. \( R(K, N) \) approximates \( \rho_{\text{max}}(K, N) \) closely,

\[
0 \leq R(K, N)/N - \rho_{\text{max}} < 0.0017
\]

where equality holds for \( L = 2 \) and for \( K \) or \( N \to \infty \). Rau gives also a computationally efficient recursive formula,

\[
R(K, N) = K \frac{2N - R(K-1, N)}{2N - R(K-1, N) + K - 1} \quad (6)
\]

which leads, with \( R(K-1, N) \to R(K, N) \) for \( K \to \infty \), to (4).

If the service times are neg.-exp. distributed, the output queues can be regarded as a closed queueing system; the key to the solution is the observation that "all the states of the system are equally likely" /5/. The number of states distinguished here is

\[
G(K, N) = \sum_{m=1}^{N} \binom{K-1}{m-1} \binom{N}{m} \quad (7)
\]

The first factor in the summation is the number of compositions /7/ of \( K \) (HOL requests) on \( m \) (not distinguished servers); the second factor is the number of possible assignments of \( m \) to \( N \). Using e.g. the Vandermonde convolution /8/ leads to

\[
G(K, N) = \binom{N + K - 1}{N - 1} \quad (8)
\]

and finally to

\[
\rho_{\text{max}} = K \frac{N}{N + K - 1}
\]

III HOL blocking with phase-type distributed service times – an approximation

We return to systems with a very large number of input and output ports. If the service time distribution is such that the arrival process of an output queue consists of a superposition of \( N \) independent arrival processes, each with a very low arrival rate \( \mu/N \), then the superposition approaches a Poisson process /9/. Independence of the arrival processes is given if the service times terminate independently and the destinations of fresh HOL requests are independent of one another. The reasoning which leads to (1) is also valid here, so that the output queue and its server may be modeled as a system M/G/1 with

\[
q = \frac{Q^2}{2(1 - \rho)}
\]

\( c^2 \) the squared coefficient of variation of the service time. With \( \rho = \rho_{\text{max}} \) and (3), we get

\[
\rho_{\text{max}} = \begin{cases} 
\frac{K/N}{1 + K/N} & c^2 = 1 \\
\frac{1}{1 - c^2} \left[ 1 + \frac{K}{N} - \left(1 + \frac{K}{N} \right)^2 - 2\frac{K}{N}(1 + c^2) \right]^{1/2} & c^2 \neq 1 
\end{cases}
\]
Turning now to systems with finite K and N, we note that the number of states in the derivation of Rau's approximation for constant service time (5) and in the derivation of (8) for neg.-exp. distributed service time is the same. Comparing (5) and an expanded version of (8), namely

\[ \text{ymax} = \frac{\sum_{m=1}^{N} m \binom{K-1}{m-1} \binom{N}{m}}{\sum_{m=1}^{N} \binom{K-1}{m-1} \binom{N}{m}}, \quad c = 1 \]

indicates that

\[ \hat{y}(K, N) = \frac{\sum_{m=1}^{N} m \left( \frac{2}{1+c^2} \right)^{m-1} \binom{K-1}{m-1} \binom{N}{m}}{\sum_{m=1}^{N} \left( \frac{2}{1+c^2} \right)^{m-1} \binom{K-1}{m-1} \binom{N}{m}} \] \tag{10}

may serve as an approximation to y_{\text{max}} for \( c^2 \neq 1 \). (In fact, (10) also contains an approximation for the state probabilities.)

The corresponding recursive formula is

\[ \hat{y}(K, N) = \frac{K}{1 + \frac{\hat{y}(K-1, N) \left( 1 - c^2 \right)}{N}} \] \tag{11}

The proof that (10) follows from (11), starting from \( \hat{y}(1, N) = 1 \), is straightforward, but lengthy and omitted here. With \( c = 0 \), (11) leads to (6), with \( c = 1 \) to (8) and with \( \hat{y}(K-1, N) \to \hat{y}(K, N) \) for \( K \to \infty \) to (9). See fig.1 for \( K = N = 2 \) and \( K = N = 4 \).

(11) is of interest not only because it is efficient in computing \( \hat{y}(K, N) \) but also because it shows the maximum throughput in terms of the mean waiting time on the HOL position. Let \( b \) denote the mean service time and \( w \) the mean waiting time, beginning with the point in time, when a request reaches the HOL position. Then

\[ \text{ymax} = \frac{1}{K} \left( 1 + \frac{w}{b} \right) \]

(11) approximates this expression:

\[ \frac{\hat{y}(K, N)}{K} = \frac{1}{1 + \frac{\hat{w}(K, N)}{b}} \]

where \( \hat{w}(K, N) \) is an approximation to \( w \),

\[ \frac{\hat{w}(K, N)}{b} = \frac{(K-1)}{N} \cdot \frac{\left( 1 - c^2 \right)/2}{1 - \frac{\hat{y}(K-1, N)}{N} \left( 1 - c^2 \right)/2} \]

IV Comparison with analytical and simulation results.

As a criterion whether or where \( \hat{y}(K, N) \) as in (10) is a sufficiently good approximation to \( y_{\text{max}} \) we will use \( \Delta \rho \),

\[ \Delta \rho = \frac{\hat{y}(K, N)}{N} - \rho_{\text{max}} \] \tag{13}

We consider systems with \( N = K \) and a service time distribution as in fig.2. \( \rho_{\text{max}} \) can be determined by setting up and solving a system of linear equations for the stationary state probabilities. Since the number of states increases considerably with \( N \) and with the number \( k \) of phases, we used discrete-time simulation to determine \( \rho_{\text{max}} \) for \( N > 2 \). The confidence interval was approximately \( \pm 0.0007 \) for \( c < 1 \) and \( \pm 0.0009 \) for \( c > 1 \).

Given \( c \) and the parameters \( k, p \) and \( u \) (see fig.2), \( \Delta \rho \) depends on \( N \). \( \Delta \rho \) takes on its maximum value for \( N = 2 \) except for values of \( c \) close to 0 where \( N = 3 \) yields the maximum value and \( \Delta \rho \) decreases roughly with \( 1/N \) (see table 1). Since we know that \( \hat{y}(K, N) \) approximates \( y_{\text{max}} \) closely for \( c = 0.6/6 \), the following discussion concentrates on systems with \( N = 2 \).
Phase k-l  k = max \left( 2, \left\lceil \frac{1}{c^2} \right\rceil \right)

Fig.2: Phase type distribution of the service time (c = coefficient of variation)

For c < 1, \( \Delta \rho \) depends only weakly on p and u (in general, many pairs of parameters p and u lead to a given c). We found that (see table 1)

\[
0 < \Delta \rho < 0.0076 \quad 0 < c < 1.
\]

For c > 1, p and u influence the result strongly. With a two-phase service time distribution, the exact solution for \( \rho_{\text{max}} \) is

\[
\rho_{\text{max}} = 1 - \frac{2pu(1 + u) + 2 + u}{2pu^2 + (3 + 4pu)(2 + u)} \quad (14)
\]

Inserting (14) into (13) gives, after some manipulation,

\[
\Delta \rho = \frac{(c^2 - 1)(c^2 + \frac{pu-1}{pu+1})(-1)}{6(c^2 + 2)(c^2 + \frac{pu-1}{pu+1} + \frac{4}{3}pu(c^2 + 2))} \quad (15)
\]

From (15), a simple upper bound for |\( \Delta \rho \)| can be derived, namely

\[
|\Delta \rho| < \frac{1}{6} \cdot \frac{c^2 - 1}{c^2 + 2} \quad c > 1
\]

saying that for c very close to 1, \( \tilde{y} \) is a close approximation to \( y_{\text{max}} \), irrespective of pu. This has to be expected, since \( \Delta \rho = 0 \) for c = 1.

For c not very close to 1, pu has to be taken into account if \( \tilde{y} \) shall be used. Another upper bound derived from (15) is

\[
|\Delta \rho| < \frac{1}{8pu} \cdot \frac{c^4 - 1}{(c^2 + 2)^2} \quad c > 1
\]

and especially for very large c

\[
|\Delta \rho| \approx \frac{1}{8pu + 6} \quad c \gg 1
\]

both indicating that the approximation is useful only if pu is large enough. This product equals the contribution of the second phase to the mean service time, divided by the contribution of the first phase, and it can assume arbitrary large values for a given c since

\[
u_{\min} \leq u < \infty, \quad u_{\min} = c^2 + \sqrt{c^4 - 1}
\]

Conclusion

For a service system with N different servers which are addressed with equal probability by the requests waiting in K input queues and being served in the order of arrival we have given an approximate formula for the maximum utilization of the servers. For small N and K, the approximation is good for subexponential phase type distributions of the service time; the approximation is good also for a coefficient of variation c greater 1 if certain conditions on the parameters determining c are satisfied. With increasing N or K, the approximation error decreases.

Beyond its use in the analysis of systems with input queuing the result will help to understand HOL blocking by showing the dependence of \( \rho_{\text{max}} \) on the mean waiting time on the HOL position. The exact calculation of this waiting time and of \( \rho_{\text{max}} \) remains an open question.

<table>
<thead>
<tr>
<th>( c^2 )</th>
<th>Service time distribution</th>
<th>( \Delta \rho ) for K=N=2</th>
<th>( \Delta \rho ) for K=N=3</th>
<th>( \Delta \rho ) for K=N=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>constant /6/</td>
<td>0</td>
<td>0.0017</td>
<td>0.0010</td>
</tr>
<tr>
<td>0.025</td>
<td>Erlang-k</td>
<td>0.0063</td>
<td>0.0047</td>
<td>0.0028</td>
</tr>
<tr>
<td>0.05</td>
<td>&quot;</td>
<td>0.0071</td>
<td>0.0049</td>
<td>0.0036</td>
</tr>
<tr>
<td>0.1</td>
<td>&quot;</td>
<td>0.0075</td>
<td>0.0046</td>
<td>0.0031</td>
</tr>
<tr>
<td>0.25</td>
<td>&quot;</td>
<td>0.0067</td>
<td>0.0031</td>
<td>0.0020</td>
</tr>
<tr>
<td>0.5</td>
<td>&quot;</td>
<td>0.0043</td>
<td>0.0022</td>
<td>0.0011</td>
</tr>
<tr>
<td>1</td>
<td>neg.-exp.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.5</td>
<td>fig.2, pu = 1</td>
<td>-0.0058</td>
<td>-0.0036</td>
<td>-0.0027</td>
</tr>
<tr>
<td>2</td>
<td>&quot;</td>
<td>-0.0114</td>
<td>-0.0073</td>
<td>-0.0048</td>
</tr>
<tr>
<td>3</td>
<td>&quot;</td>
<td>-0.0207</td>
<td>-0.0139</td>
<td>-0.0094</td>
</tr>
<tr>
<td>4</td>
<td>&quot;</td>
<td>-0.0278</td>
<td>-0.0203</td>
<td>-0.0137</td>
</tr>
</tbody>
</table>

Table 1: \( \Delta \rho \) (Examples)
Annex: Mean waiting time on the HOL position

We will give examples in which \( w(K,N) = w \), the mean waiting time on the HOL position of an input queue. A simple example is that of a system with \( K = 2 \) and constant service time; we consider the HOL request of input queue 1. Since the probability that the HOL request of input queue 2 is addressed to the same server is \( 1/N \) and, assuming a random selection, the probability of not being served is \( 1/2N \), we have a mean sojourn time \( s \):

\[
s = \frac{1}{1 - \frac{1}{2N}} \cdot b
\]

and a mean waiting time

\[
w = s - b = \frac{b}{2N} \cdot \frac{1}{1 - \frac{1}{2N}} = \hat{w}(2, N), \quad c = 0, K = 2
\]

since \( \hat{q}(1, N) = 1/N \) and \( r = b/2 \).

Turning now to more general cases, we have to find the mean waiting time \( w \) of a request after attaining the HOL position of some input queue, destined to output \( n \). In this output queue, we have - from the remaining input queues - a mean number of \( (K-1)/N \) requests, \( q \) of them waiting, and a mean virtual waiting time \( v \),

\[
v = \left( \frac{K-1}{N} \right) r + q b.
\]

Replacing \( q \) by the product of mean waiting time of these requests, \( w(K-1) \), and arrival rate \( \rho(K-1)/b \) yields

\[
v = \frac{K-1}{N} r + w(K-1) \rho(K-1) \cdot \left( 1 - \frac{r}{b} \right)
\]

For \( c = 1 \), we have \( w = v \) and \( b = r \), leading to

\[
w = \frac{K-1}{N} r = \hat{w}(K, N) \quad c = 1
\]

For \( K, N \to \infty \),

\[
w(K-1) \to w(K) = w,
\]

and since the arrival process becomes a Poisson process,

\[
v \to w
\]

and finally, as in (12),

\[
w = \frac{K-1}{N} \cdot \frac{r}{1 - \rho(K-1)(1 - \frac{1}{N})}
\]

References:


