Optimal fixed batch sizes for real-time inquiry systems

S. H. Storey

Scientific Control Systems Ltd., Sanderson House, 49-57 Berners Street, London W1

An elementary but useful result is presented which allows the optimal batch size to be found for real time message processing system in which messages are processed in batches of fixed size.

(Received March 1971)

1. Introduction

A standard technique in the design of real-time inquiry systems of many types, is the processing of messages in batches of fixed size. The purpose of this is normally to share out among all the messages in the batch any fixed 'housekeeping' time needed. This sharing then increases the efficiency of *utilisation* of the machine or device doing the processing. The batching of messages has also the effect, initially, of reducing the mean total processing time of the individual messages. As the batch size increases, however, the mean queueing (or total processing) time of the individual messages eventually also increases.

It is the purpose of this note to show that a batch size can be selected which minimises the mean queueing time of individual messages for a reasonably useful model of such systems. The model assumed is that of a single server queue, with a Poisson arrival stream of mean rate λ messages/unit time. The holding time of the server consists of two parts. One part varies from message to message and is described by a general holding time probability distribution. The other part is a constant 'housekeeping' or 'job set up' time.

2. The effect of housekeeping

It is first necessary to specify the effect, on the moments of a holding time distribution, of a constant housekeeping time. If the variable component of the holding time distribution is such that the processing of a given message requires a time in $(t, t + dt)$ with probability $Q(t)\,dt$, the moments of this distribution are given by

$$m_n = \int_0^\infty t^n\, Q(t)\, dt \tag{1}$$

If a constant component τ is added to the holding time, so that the final probability distribution function $P(t)$ of the total holding time has the form of **Fig. 1** then the moments of the new holding time probability distribution are

$$b_n = \int_0^\infty t^n\, P(t)\, dt \tag{2}$$

By changing the integration variable to $t' = t - \tau$, and making use of the binominal theorem and equation (1), one obtains the relation

$$b_n = \sum_{r=0}^{n} {}_nC_r \tau^{n-r} m_r \tag{3}$$

between the two sets of moments. In particular, the first three moments are

$$\begin{aligned} b_0 &= m_0 \ (= 1 \text{ normally}) \\ b_1 &= \tau m_0 + m_1 \\ b_2 &= \tau^2 m_0 + 2\tau m_1 + m_2 \end{aligned} \tag{4}$$

These expressions will be required below, in the form

$$T_s = \tau + T_s' \tag{5}$$

and since $\sigma^2 = b_2 - b_1^2$ by definition

$$\sigma_s^2 = \sigma_s'^2 \tag{6}$$

where it has been assumed that $m_0 = 1$. Here T_s' and T_s are the mean holding times for the holding time distribution with and without the constant housekeeping time, r, respectively. Similarly $\sigma_2'^s$ and σ_s^2 are the variances of the two distributions.

3. The effect of batching

Let the arrival rate of individual messages have a mean of λ messages/unit time. Let the variable portion of the holding time for a single message have a distribution function with a mean of T_s and a variance of σ_s^2. If the messages are processed in batches of fixed size m then (IBM—undated) the mean of the variable portion of the batch waiting time is mT_s and its variance is $m\sigma_s^2$. The batch mean arrival rate is, of course λ/m.

The addition of a constant housekeeping time τ to the batch holding time does not affect its variance (6), but does change the mean holding time to $mT_s + \tau$ (5).

The utilisation rate ρ which equals λT_s for single messages, thus becomes

$$\rho = \left(\frac{\lambda}{m}\right)(mT_s + \tau) \tag{7}$$

for the batches.

The mean waiting time, T_w which for individual messages (without housekeeping) is given by (Takacs, 1962).

$$T_w = \frac{\lambda(\sigma_s^2 - T_s^2)}{2(1 - \rho)} \tag{8}$$

becomes, for the batches treated as oversize messages

$$T_w = \frac{\left(\frac{\lambda}{m}\right)[m\sigma_s^2 + (mT_s + \tau)^2]}{2\left[1 - \left(\frac{\lambda}{m}\right)(mT_s + \tau)\right]} \tag{9}$$

An individual message, however, may be in any one of the m positions in the batch. Assuming a straightforward 'first in, first out' queue discipline, and that the housekeeping is done when the batch begins service (although this does not effect the results we seek), individual messages will also have to wait a mean time of

$$\left(\frac{m-1}{2}\right)T_s + \tau \tag{10}$$

once a batch begins service. Finally, to obtain the queueing time (or total processing time) of the message, the mean holding time T_s for that message must be included.

Thus, the mean queueing time for individual messages, when processed as a batch requiring an additional constant housekeeping time τ per batch, is given by

$$T_q = \frac{\left(\frac{m}{\lambda}\right)[m\sigma_s^2 + (mT_s + \tau)^2]}{2\left[1 - \left(\frac{\lambda}{m}\right)(mT_s + \tau)\right]} + \frac{(m-1)\,T_s}{2} + \tau + T_s \tag{11}$$

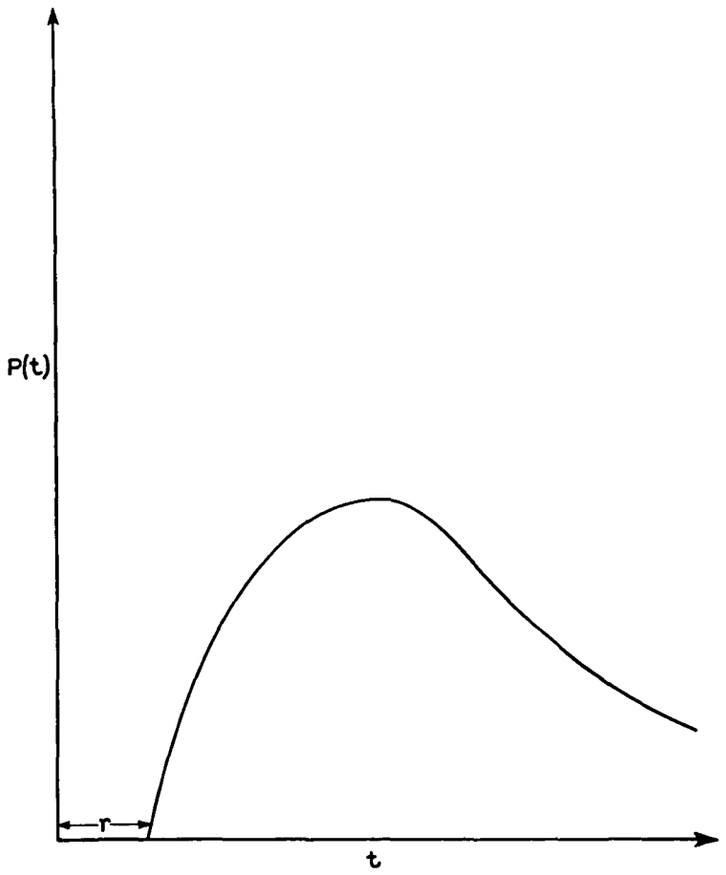

Fig. 1. Total holding time distribution

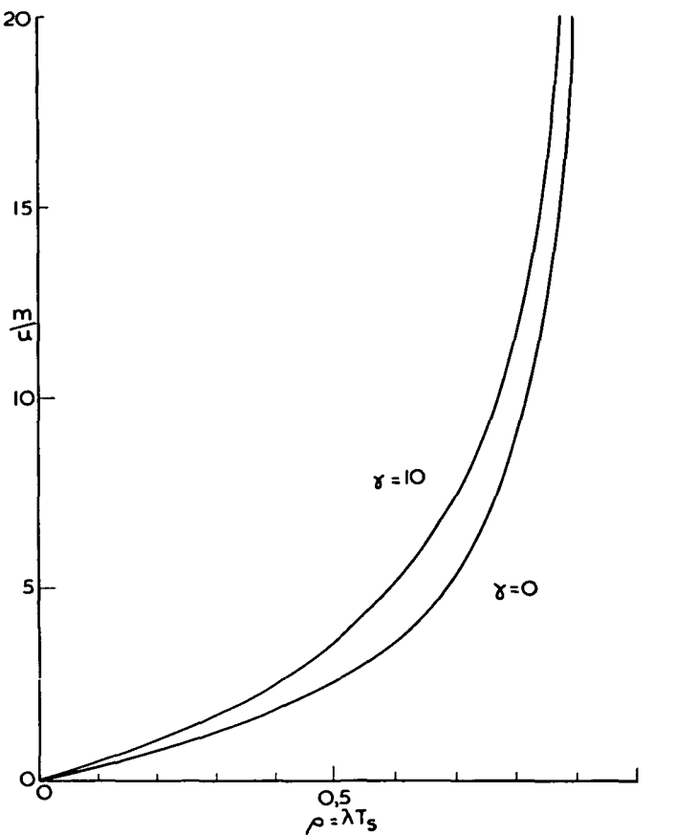

Fig. 2. Variation of optimum with utilisation

4. The optimal batch size

If machine efficiency is the overriding consideration, the optimal batch size is simply the largest possible. If, however, system response is crucial the optimal batch size can be taken to be that value of m which minimises the time taken to process individual messages. The measure of this response which is employed here is the mean queueing time, of (11). m is a discrete variable. It is however sufficient for our purpose here to find that value of m, treated as a continuous variable, which makes

$$\frac{dT_q}{dm} = 0 \tag{12}$$

where T_q is specified by (11). The differentiation of (11) is 'straightforward but tedious' and leads 'after some manipulation' to the requirement that

$$T_s(1 - \lambda T_s)\, m^2 - 2\lambda\tau T_s m - \lambda\tau(\lambda\sigma_s^2 + \tau) = 0 \tag{13}$$

If the quantities

$$\rho = \lambda T_s$$

$$\mu = \frac{\tau}{T_s}$$

and

$$\gamma = \frac{\sigma^2}{\tau T_s}$$

are introduced, the required solution (only one root is positive) is given by

$$m = \left(\frac{\mu\rho}{1-\rho}\right)\left\{1 + \sqrt{\frac{1 + \gamma\rho(1-\rho)}{\rho}}\right\} \tag{14}$$

The actual batch size which can reasonably be chosen, will be the integer nearest the value required by (14). Large batch sizes are thus required if

1. The housekeeping time is large relative to the mean holding time of individual messages.
2. The holding time distribution is very spread.
3. The system, with housekeeping time neglected, is near saturation (i.e. ρ is nearly 1).

As a simple example, consider the case where $\lambda = 2$ messages/sec. $T_s = 0{\cdot}25$ sec, $\tau = 0{\cdot}25$ sec, and $\sigma^2 = 0{\cdot}25$ sec^2. In this case $\rho = 0{\cdot}5$, $\mu = 1$ and $\gamma = 4$, so that substitution in (14) leads to $m = 3$. Thus, processing the messages in batches of three would lead to the best response time. In practice, of course, the value calculated for m would not be found integral, and the actual batch size would be taken to be the nearest integer.

A graph of m/μ against ρ is given in **Fig. 2** for two values of γ(0 and 10). It should be noted that the effect of changes in γ is relatively small.

Acknowledgements

The author would like to thank Mr. R. Davenport and Mr. R. J. Royston for a number of helpful comments.

References

IBM (undated). Analysis of Some Queueing Models in Real-Time Systems, IBM Manual, Data Processing Techniques, F20-0007-1, IBM World Trade Corporation, 821 United Nations Plaza, NY, USA.

TAKACS, L. (1962). A Single Server Queue with Poisson Input, *Operations Research*, Vol. 10, p. 388.