1981 Winter Simulation Conference Proceedings
T.I. Ören, C.M. Delfosse, C.M. Shub (Eds.)

THE SENSITIVITY OF QUEUEING MODELS SIMULATION TO TIME DISCRETIZATION

Hon-Shiang Lau
College of Business and Economics
Washington State University
Pullman, WA 99164

Ahmed Zaki
School of Business Administration
College of William and Mary
Williamsburg, VA 23185

The time continuum is often appoximated by discrete time units in the simulation and mathematical analysis of queueing systems, but little is known about the effects of this approximation on the final results. This study shows that the effects of time discreteness on some queueing behavior is surprisingly small, and useful implications of this finding are discussed.

1. INTRODUCTION

In computer simulation, it is often necessary to approximate the time continuum of the real world with discrete time units. In queueing problems, two common examples in which time discreteness is introduced are:

(i) In APL, if a programmer uses the expression

(? 1000) ÷ 1000 (1)

to generate uniformly distributed "customer service times" with a mean of 0.5 minute and a range of 0-1 minute, the generated random times are restricted to taking values in multiples of 0.001 minute. That is, the system actually simulated behaves as if time exists in discrete units of 0.001 minute.

(ii) In GPSS, time is explicitly treated as discrete units. To generate uniformly distributed customer service times having a mean of 10 minutes and a range of 5-15 minutes, one can write

Generate 10, 5 (2)

in which case the service times are restricted to taking integer values from 5 to 15. Alternatively, one can designate 1 clock time unit as equivalent to 0.1 minute and write

Generate 100, 50 (3)

in which case the service times are restricted to taking values in discrete steps of 0.1 minute. In spite of this common practice of using discrete time units in simulation to approximate the time continuum, the effects of this approximation have seldom been formally investigated. The purpose of this paper is study the nature and magnitude of any effects of time discreteness in the simulation of simple queues. Examples of practical questions that this study may help to answer are discussed below:

(i) What would be the difference between generating random times by statements such as

(? 100) ÷ 100 (in APL) (4a)

Generate 10, 5 (in GPSS) (4b)

versus statements such as

(? 10000) ÷ 10000 (in APL) (5a)

Generate 100, 50 (in GPSS) (5b)

Of course, every programmer intuitively knows that by writing

(? 1E10) ÷ 1E10 (in APL) (6a)

Generate 100000, 50000 (in GPSS), (6b)

81CH1709-5/81/0000-0341$00.75

the discreteness problem can be safely ignored. However, this does not really answer the original question.

(ii) Relating to the preceding question, many users of GPSS will have encountered situations in which an entire GPSS program can be rewritten to require much shorter computer time if 1 clock time unit can be assigned as equivalent to 0.1 minute instead of, say, 0.001 minute. Clearly, a programmer wants to know whether the greater accuracy of the latter version justifies the additional computer time.

(iii) If time distributions are constructed from empirical data, the times are almost always collected in "classes" for the construction of histograms. Two questions will arise in these situations; firstly, if the times of the process actually follow a continuous density function, how reliable will be the simulation results obtained by using the empirically collected discrete distribution? Secondly, in order to ensure a good approximation of the natural continuous process, what should be the "class width" used to collect the data? Of course, in both questions, the confounding issue of sampling error has to be temporarily ignored.

(iv) Eventually, perhaps the most promising point is to find out the effects of using discrete time units in mathematical modelling. For example, if using discrete time units provides good approximation in simulating a certain system, then using discrete time units must also provide good approximation in the mathematical analysis of that system. One example is the numerical procedure presented in the "Methodology" section of this paper for obtaining a queue's waiting time distribution, and this point will be brought up again later.

2. DEFINITIONS AND OVERVIEW

This study investigates the effects of different levels of time discretization on "customers' waiting time" in single-server queues. The service rate, μ, is fixed at 1 customer per unit time, and queues with different utilization factors ρ are obtained by varying the arrival rate λ.

2.1 Definitions

(i) The "level of discreteness" ℓ is the smallest unit of time permissible in the simulated system. For example, referring to the normalized μ of 1 per unit time, simulating at the discreteness level of $\ell = 0.1$ means that the service and interarrival times can only exist in multiples of 0.1 time units. In GPSS, $\ell = 0.1$ corresponds to representing 1 real time unit by 10 clock (or simulation) time units. The case of continuous time corresponds to $\ell = 0.0$.

(ii) $W(\ell,\rho)$ is the "theoretical" customer's waiting time distribution when the utilization factor is ρ and the simulation is performed at the discreteness level of ℓ. $\bar{W}(\ell,\rho)$ is the corresponding mean waiting time. The corresponding sample waiting time distribution observed from a finite sumulation run is $w(\ell,\rho)$, and $\bar{w}(\ell,\rho)$ is the corresponding sample mean of waiting time.

(iii) $\Delta_{ijt} = W_t(i,\rho) - W_t(j,\rho)$ is the absolute difference in the waiting time distribution at t between the discreteness levels of i and j. Similarly, $\delta_{ijt} = w_t(i,\rho) - w_t(j,\rho) = \Delta_{ijt}$'s sample estimate.

(iv) $D_{ij} = \bar{W}(i,\rho) - \bar{W}(j,\rho)$, and

$d_{ij} = \bar{w}(i,\rho) - \bar{w}(j,\rho) = D_{ij}$'s sample estimate.

2.2 Overview

In this study, we are primarily interested in the effects of time discreteness as reflected in the values of Δ_{ijt} and D_{ij}. These values can be estimated from their sample values δ_{ijt} and d_{ij} using simulation, they can also be computed via the waiting time distribution $W(\ell,\rho)$ obtained from a numerical procedure. These two procedures are discussed in the next section on "Methodology." Using these procedures, values of Δ_{ijt} and D_{ij} are estimated for queues with different distribution forms of interarrival/services times, the results and an attempt to generalize them are presented in the fourth section. Generally speaking, the effect of time discreteness is found to be small. Finally, the implications of our results are briefly discussed in the concluding section. The third section may be skipped since an understanding of the methodologies is not necessary for the later sections.

3. METHODOLOGY

3.1 Simulation

Obtaining the sample observations Δ_{ijt}'s and d_{ij}'s by a Fortran simulation program is conceptually straightforward; any required discreteness level can be obtained by rounding the generated continuous random numbers to the appropriate numbers of decimal places. However, the actual values of Δ_{ijy}'s and D_{ij}'s turn out to be very small, whereas the variances and autocorrelations of their sample estimates (i.e., δ_{ijt}'s and d_{ij}'s) are comparative large; moreover, the variances and autocorrelations increase rapidly as ρ increases. Therefore, without using excessive computer time, we were only able to consider cases of moderate ρ, and reliable information was obtainable only on the signs, but not on the magnitudes of the D_{ij}'s. No reliable estimates of the Δ_{ijt}'s were obtainable. Fortunately, as will be seen later, these results are adequate for the purpose of this study.

3.2 The Computational Procedure

Our procedure for computing $W_t(\ell,\rho)$ is based on

Lindley's (1952) integral equation for waiting time distributions. Given the density functions of the interarrival times (a) and the service times (s), if f(t) denotes the density function of the random variable (s-a), Lindley has shown that the steady-state waiting time distribution function W(t) for a single-server queue is given by

$$W(t) = \begin{cases} \int_{\infty}^{t} W(t-x)\, f(x)dx & t \geq 0 \\ 0 & t < 0 \end{cases} \qquad (7)$$

It is known that, except when ρ is close to 1, the distribution function W(t) usually increases very rapidly towards 1 as t increases initially, but it then converges very slowly to 1 as t becomes larger and W(t) becomes closer to 1. For most practical purposes, instead of considering the entire domain $(0, \infty)$ for W(t) and the domain $(-\infty, \infty)$ for f(t), "sufficiently accurate" results can be obtained by truncating W(t)'s domain to (0,T) and f(t)'s domain to (-T,T), where T is sufficiently large such that $W(T) \simeq 1$. At this value of T, if time exists in discrete units of ℓ (i.e., discreteness level = ℓ), then by writing

$$M = T/\ell, \qquad (8)$$

and by defining w(t) to be the corresponding discrete density function of waiting time and d(t) to be the discretized counterpart of f(t), it can be seen that eqn. (7) transforms to the system of equations

$$\begin{aligned} w(0) &= w(0)d(0) + w(1)d(-1) + \ldots w(M)d(-M), \\ w(1) &= w(0)d(1) + w(1)d(0) + \ldots\ w(M)d(1-M), \\ &\vdots \\ w(M) &= w(0)d(M) + w(1)d(M-1) + \ldots w(M)d(0), \end{aligned} \qquad (9)$$

or equivalently,

$$\begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(M) \end{bmatrix} = \begin{bmatrix} d(0) & d(-1) & \ldots & d(-M) & w(0) \\ d(1) & d(0) & \ldots & d(1-M) & w(1) \\ \vdots & \vdots & & \vdots & \vdots \\ d(M) & d(M-1) & \ldots & d(0) & w(M) \end{bmatrix}. \qquad (10)$$

Since one of the above M+1 equations is redundant, the system can be solved by replacing the first equation with the probability relationship

$$w(0) + (w(1) + \ldots + w(M) = 1, \text{ giving}$$

$$\begin{bmatrix} 1 \\ w(1) \\ \vdots \\ w(M) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 & w(0) \\ d(1) & d(0) & \ldots & d(1-M) & w(1) \\ \vdots & \vdots & & \vdots & \vdots \\ d(M) & f(M-1) & \ldots & d(0) & w(M) \end{bmatrix}, \qquad (11)$$

or

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 & w(0) \\ d(1) & d(0)-1 & \ldots & d(1-M) & w(1) \\ \vdots & \vdots & & \vdots & \vdots \\ d(M) & d(M-1) & \ldots & d(0)-1 & w(M) \end{bmatrix}. \qquad (12)$$

Eqn. (12) is in the form [K] = [D][w], hence

$$[w] = [D]^{-1}[K]. \qquad (13)$$

Observing the elements in matrix [D] of eqn. (12) reveals that the matrix is diagonally dominant except for the first row, therefore, the waiting time density function [w] can be easily obtained by solving eqn. (13) using the standard Gauss-Seidel method. The elements d(k)'s required in matrix D can be obtained in two ways. First, if a convenient theoretical function of f(x) in eqn. (7) is available (e.g., when the service and interarrival times are both normally distributed, and hence f(x) is normal), then d(k) can be simply computed as the area under f(x) in the time interval $(k\ell \pm 0.5\ell)$, where ℓ is the discreteness level. Second, if f(x)'s function cannot be conveniently obtained, d(k) can be computed by a "numerical convolution" procedure:

$$d(k) = \begin{cases} \sum_{i=k}^{L} s(i)\, a(i-k) & \text{for } k \geq 0 \\ \sum_{i=0}^{L+k} s(i)\, a(i-k) & \text{for } k < 0, \end{cases} \qquad (14)$$

where s(k) and a(k) are respectively the probabilities of the service and interarrival times being in the interval $(k\ell \pm 0.5\ell)$, and L is a suitably large value greater than 2M.

For modelling the discrete-time queueing system, the only approximations made in eqn. (13) are the truncation of W(t)'s domain from $(0, \infty)$ to (0,T) in eqn. (8) and (9), and the truncation of the range of s and a from $(0, \infty)$ to $(0, L\ell)$ in eqn. (14). To achieve the accuracies needed in this study, the computations were programmed in "triple precision," and the truncation levels are set to be high enough such that a(L) and s(L) are less than 10E-15 and w(M) is less than 10E-20 in each case. We found that this computational method performs adequately for low ρ and high discreteness level ℓ. When ρ is high, the long tail of W(t) necessitates a high truncation level T, and hence a larger M-value according to eqn. (8). When ℓ is low, both M and L have to be large for the same truncation levels T and $L\ell$. In either case, the computations required to perform eqn. (14) and (13) become excessive.

4: EXPERIMENTAL RESULTS

4.1 Exponentially Distributed Times

The computational method is first used to compute [w] for discreteness levels of 0.2 and 0.1; having [w], the values of $\bar{W}(\ell,\rho)$ and Δ_{ijt} can be easily computed. Table 1 gives the values of $\bar{W}(\ell,\rho)$ for discreteness levels of 0.2, 0.1, 0 and ρ-values of 0.1, 0.2, and 0.3. $\bar{W}(\ell,\rho)$ for $\ell = 0$ is known to be $\rho/1-\rho$.

The effect of time discreteness on the mean waiting time appears to be very small, but there is a consistent pattern: "a higher discreteness

level leads to a slightly higher mean waiting time." Our computed values of Δ_{ijt}'s are all very small (< 0.1%), and no meaningful pattern can be detected.

TABLE 1
Mean Waiting Times for
Exponentially Distributed Times

ℓ \ ρ	0.1	0.2	0.3
0.2	0.111333	0.250581	0.429710
0.1	0.111167	0.250146	0.428866
0	0.111111	0.250000	0.428571

The computational method becomes inadequate for $\rho > 0.3$ or for $\ell < 0.1$, and simulation was used to study the cases of $\rho = 0.4$ to 0.6. While reliable estimates of D_{ij}'s magnitudes were not obtainable, the simulated values of d_{ij}'s do indicate the same consistent pattern: "a higher discreteness level leads to a slightly higher mean waiting time."

For higher ρ-values, the sample observations from simulation become too erratic.

4.2 Normally Distributed Times

Assuming normally distributed service and interarrival times, the computational procedure was used to determine values of $\bar{W}(\ell,\rho)$ for the cases tabulated in Table II. The effect of time discreteness on the mean waiting time is again very small. For the cases below the line AA in Table II, the consistent direction of the time discreteness effect is: "a higher ℓ leads to a slightly higher mean waiting time," which agrees with the earlier observation. However, for cases above the line AA in Table II, the consistent direction of the time discreteness effect is in the opposite direction, i.e., "a higher ℓ leads to a slightly LOWER mean waiting time." The computed values of Δ_{ijt}'s are again very small and no meaningful pattern can be observed.

4.3 Generalization

Other distributions were also used to represent the interarrival and service times, and investigations similar to the ones described in sections 4.1 and 4.2 were conducted. These results (many of them not presented here) clearly suggests the following generalization on the effect of time discreteness:

When ρ and the coefficient of variation of the interarrival/service times are both small, a higher ℓ leads to a lower mean waiting time. When the combination of ρ and the coefficient of variation are "sufficiently large" (as indicated by the region to the right of AA in Table II), a higher ℓ leads to a higher mean waiting time.

This generalization explains very well the observed patterns in Tables I and II. For exponentially distributed times (see Table I), a higher ℓ always leads to a higher mean waiting time because an experimental distribution always has a high coefficient of variation of 1.

5. DISCUSSION AND CONCLUSION

Investigations similar to the ones presented above were also conducted for queues with other distribution forms of interarrival/service times. The data reveal the same pattern of behavior depicted above, and are not presented. One would expect the effect of time discreteness to be small, since this probably has been the implicit justification for the lack of formal investigations on this issue. However, it is probably surprising to see how small the effect actually is. For example, considering situations in which the mean waiting times are sufficiently large (say, more than 10% of the mean service times) to be of practical interest, Table II indicates that very accurate estimates of $\bar{W}(0,\rho)$ can be obtained with $\bar{W}(0.2,\rho)$, i.e., using discrete time units as large as 20% of the mean interarrival time. This leads to very useful answers to the questions raised in the first section of this paper: Perhaps the most promising implication is the suggestion of developing discrete-time queueing models that can be more easily solved than their

TABLE II
Mean Waiting Times for Normally Distributed Times

Coeff. of Variation	Discrete. Level	Load Factor ρ					
		0.1	0.3	0.5	0.7	0.8	0.9
c.v =0.1	ℓ = 0.2	0.	0.275 E-12	0.874 E-7	0.255 E-3	0.363 E-2	0.364 E-1
	ℓ = 0.1	0.	0.428 E-12	1.148 E-7	0.362 E-3	0.444 E-2	0.375 E-1
	ℓ = 0.005	0.	0.482 E-12	1.696 E-7	0.395 E-3	0.466 E-2	0.382 E-1
c.v =0.2	ℓ = 0.2	0.	0.692 E-4	0.192 E-2	0.236 E-1	0.722 E-1	0.258 E-0
	ℓ = 0.1	0.15525 E-3	0.717 E-4	0.201 E-2	0.242 E-1	0.727 E-1	0.255 E-0
	ℓ = 0.05	0.15553 E-3	0.724 E-4	0.203 E-2	0.243 E-1	0.729 E-1	0.254 E-0
c.v. =0.3	ℓ = 0.2	0.12173 E-2	0.473 E-2	0.235 E-1	0.1067 E-0	0.2391 E-0	0.677 E-0
	ℓ = 0.1	0.12193 E-2	0.477 E-2	0.238 E-1	0.1071 E-0	0.2387 E-0	0.672 E-0
	ℓ = 0.05	0.12197 E-2	0.478 E-2	0.238 E-1	0.1072 E-0	0.2386 E-0	0.671 E-0

continuous-time counterparts but are nearly as accurate. One example is the solution of Lindley's equation (eqn. 7) by the computation procedure presented in this paper. Analytical solution of eqn. (7) is known to be difficult for most practical interarrival/service-time distributions. However, if the discreteness level ℓ can be set at a large value, M will be small according to eqn. (8), and the discrete waiting time distribution can be easily obtained through eqn. (14) and (13), using very little computer time; moreover, this computational procedure is applicable to all distribution forms of interarrival/service.

The consistent effect of ℓ on the mean waiting time as generalized at the end of the preceding section is perhaps interesting academically, but we have no theoretical explanation for the observed behavior. Given the smallness of the effect, further investigations do not seem worthwhile.

Besides observing the mean waiting time, we have also observed in these investigations the effects of time discreteness on the waiting time distribution itself and the standard deviation of waiting times, the effects are all negligible for practical purposes, and are therefore not presented.

REFERENCES

APL/360 User's Manual, GH20-0683, IBM Corporation, White Plains, New York.

General Purpose Simulation System V User's Manual, SH20-0851, IBM Corporation, White Plains, New York.

D.V. Lindley (1952), The Theory of Queues with a Single Server, PROCEEDINGS of the Cambridge Philosophical Society, Vol. 48, pp. 227-289.