Simulation-Based Estimation of Variability in Queueing Systems

Kimmo E. E. Raatikainen

University of Helsinki, Department of Computer Science, P. O. Box 26 (Teollisuuskatu 23), FIN-00014 University of Helsinki, Finland

Queueing systems are popular and widely used tools for analyzing the performance of computer and communication systems. When traffic rate or queue length is examined, the variability during some period of time is of great practical importance. In this paper we propose a way to examine variability in queueing systems through simulation. Our approach is based on the following idea. The original traffic rates or queue lengths are classified into low and high values. The resulting binary sequence of indicators characterizes the traffic or load level of the system over time. The variation over time is summarized through estimating the histograms of level durations. The rationale of translating the problem into estimation of histograms is that we can use an existing method for assessing the simultaneous precisions of the estimated proportions.

Key words: variability, queueing, traffic rate, precision control

1 Introduction

When queueing systems are used as a tool for analyzing computer and communication systems, the usual performance measures examined have been *steady-state* means. The most popular indices include mean waiting time, mean throughput, and mean queue length. However, *variability* of traffic rate or queue length during some period of time is of great practical importance even if we assume that the process examined is in stochastic equilibrium. In this paper we propose a way to examine variability in a queueing system through simulation.

When variability is examined, the fundamental problem is the characterization: Which set of indices properly summarizes the variability? The second problem is, how we can assess the precision of estimated indices. As noted in [5], a simulation study of any stochastic process is only a programming exercise if the output analysis is not properly carried out.

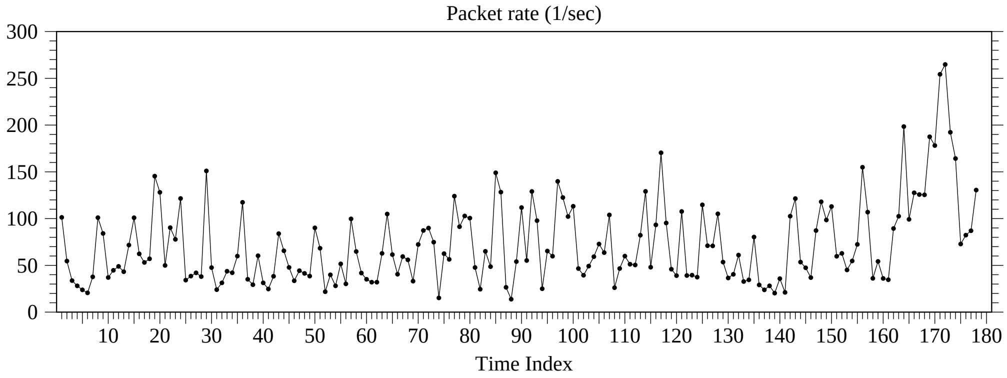

Fig. 1. Measured Packet Rate

We propose one possible way to characterize and estimate variability in a queueing system. We recognize that several other possibilities exist but we believe that the approach we propose can be used in many situations. The approach is based on the following idea: *The original output sequence* $\{X_j\}_{j=1}^N$ *is classified into low and high values. The resulting binary sequence of indicators* $\{I_j\}_{j=1}^N$ *characterizes the levels in the output sequence. The variation of the level is summarized through estimating the two histograms of level durations.* The rationale of translating the problem into simultaneous estimation of two histograms is that we can use the method for assessing the precisions of estimated proportions introduced in [7].

The rest of the paper is organized as follows. In Section 2 we introduce our approach to summarize variability in the original output sequence. In Section 3 we briefly describe the method for estimating the proportions. In Section 4 we report empirical results.

2 An Approach to Summarize Variability

In this section we use an example to describe our approach to summarize variability in the output sequence. We need three successive transformations to obtain the sequences of indicators from which the cell proportions of the histograms are estimated. Our example uses a sequence of measured packet rates to and from a file server shown in Fig. 1. The measured sequence is a realization of $\{X_j\}_{j=1}^N$.

The first step in our approach is to define a suitable threshold value (x_*) to classify the observations into *low* and *high* values. Usually the analyst has already done some pilot simulations which help in selecting a suitable threshold value. Alternatively, the goals of the simulation study determine an appropriate value. In our example we use $x_* = 100$. Fig. 2 shows the sequence of

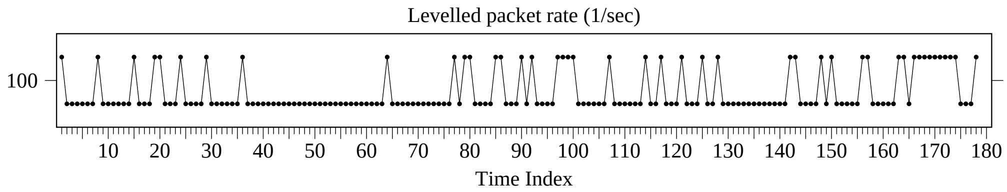

Fig. 2. Levelled Packet Rate

dichotomized packet rates which corresponds to the resulting binary sequence $\{I_j\}_{j=1}^{N}$:

$$I_j = \begin{cases} 0, \text{ if } X_j \leq x_* \, , \\ 1, \text{ if } X_j > x_* \, . \end{cases}$$

The second step in our approach is to transform the binary sequence, $\{I_j\}_{j=1}^{N}$, into a bivariate sequence of durations: $\{Y_{0,k}, Y_{1,k}\}_{k=1}^{M}$ where $Y_{0,k}$ is the number of consecutive zeros in the kth cycle and $Y_{1,k}$ is the number of consecutive ones in the kth cycle. A cycle consists of one consecutive sequence of zeros and ones (or ones and zeros). If the first observation is one (zero), we must discard all leading ones (zeros) since we do not know how many ones (zeros) precede the first observation. At the end of the binary sequence we always have an incomplete cycle. Fig. 3 shows the sequence of durations corresponding to the dichotomized packet rate in Fig. 2.

$$\begin{array}{rl} \text{Binary sequence:} & \overbrace{1\,\underbrace{0000001}_{\text{cycle 1}}\,\underbrace{0000001}_{\text{cycle 2}}\,\underbrace{00011}_{\text{cycle 3}}\,0\,.........\,1\,\underbrace{0111111111}_{\text{cycle } M}\,0001}^{\text{sequence length } N} \\ \text{Sequence of durations:} & (6,1)\,,\ (6,1)\,,\ (3,2)\,,\,.........\,,\ (1,9) \end{array}$$

Fig. 3. Binary Sequence and Sequence of Durations Based on the Dichotomized Packet Rate in Fig. 2

The third step in our approach is to transform this sequence, $\{Y_{0,k}, Y_{1,k}\}_{k=1}^{M}$, into a multidimensional sequence of indicators that enables us to estimate the cell proportions in histograms. The construction of any histogram requires the analyst to specify the number of classes and the class boundaries. This may be a difficult task. However, pilot simulations are usually executed. They provide a rough impression of the duration distributions. If the pilot simulations indicate (or the analyst believes) that the distribution may be long-tailed, then boundaries that are powers of 2 are usually a good guess.

Suppose that we have specified that C_0 and C_1 are the number of classes for low and high traffic durations, respectively. Let $\{c_{0,i}\}_{i=0}^{C_0}$ and $\{c_{1,i}\}_{i=0}^{C_1}$ be the

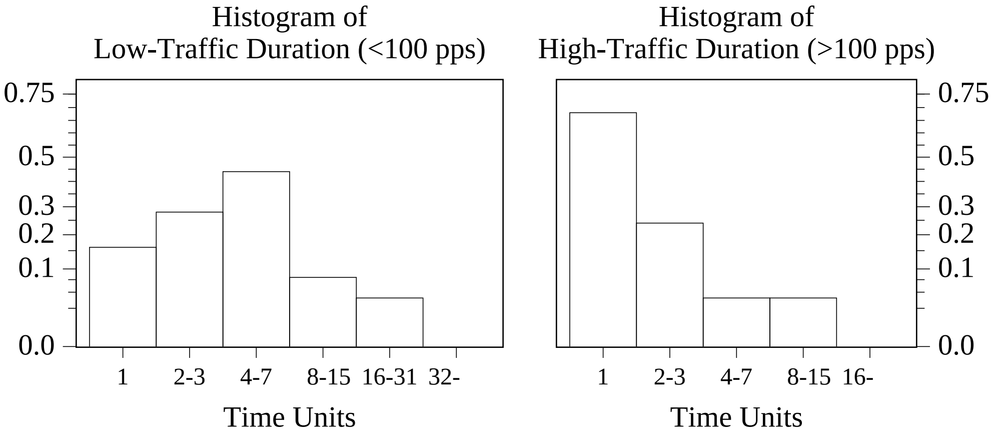

Fig. 4. Histograms of Low and High Traffic Durations

selected class boundaries; usually (but not necessarily) $c_{0,0} = c_{1,0} = 0$ and $c_{0,C_0} = c_{1,C_1} = \infty$.

Now we can transform $\{Y_{0,k}, Y_{1,k}\}_{k=1}^{M}$ into a $(C_0 + C_1)$-dimensional sequence of indicators: for Y_0 define

$$I_{i,k} = \begin{cases} 1, \text{ if } c_{0,i-1} \leq Y_{0,k} < c_{0,i} \\ 0, \text{ otherwise} \end{cases}, \; i = 1, \ldots, C_0 \, , \tag{1}$$

and for Y_0 define

$$I_{C_0+i,k} = \begin{cases} 1, \text{ if } c_{1,i-1} \leq Y_{1,k} < c_{i,j} \\ 0, \text{ otherwise} \end{cases}, \; i = 1, \ldots, C_1. \tag{2}$$

The estimated cell proportions, $\hat{p}_i$, are:

$$\hat{p}_i = \frac{1}{M} \sum_{k=1}^{M} I_{i,k} \, , \; i = 1, \ldots, C_0 + C_1 \, . \tag{3}$$

Fig. 4 shows the histograms in our example when the number of classes for the low-traffic duration is six ($C_0 = 6$) and for the high-traffic duration it is five ($C_1 = 5$). In both histograms we use powers of 2 as the class boundaries. Note that in the figure the scale of the x-axis is ordinal.

3 Assessing the Precision of Estimated Cell Proportions

As noted in the Introduction, a simulation study is only a programming exercise if the output analysis is not properly carried out. In this section we briefly describe the method for assessing the precision of the estimated proportions introduced in [7]. In addition, we present a way to visualize the estimated histograms with precision assessment information.

The precision of $\hat{p}_i$ as an estimate for the true unknown p_i is measured by the $1-\alpha_i$ confidence interval $[\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}]$:

$$\Pr\{p_i \in [\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}]\} = 1 - \alpha_i \,, \tag{4}$$

where δ_{L_i} and δ_{U_i} are the desired lower and upper half-widths of the confidence interval. The joint precision of estimates $\hat{p}_1, \ldots, \hat{p}_m$ is measured by the simultaneous $1-\alpha$ confidence interval:

$$\Pr\{p_i \in [\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}] \text{ for every } i = 1, \ldots, m\} = 1 - \alpha \,. \tag{5}$$

When we apply the Bonferroni inequality, see e.g. [4, Ch. 9.7], we obtain an upper bound for α, namely $\alpha \leq \sum_{i=1}^{m} \alpha_i$, and a lower bound for the simultaneous confidence:

$$\Pr\{p_i \in [\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}] \text{ for every } i = 1, \ldots, m\} \geq 1 - \sum_{i=1}^{m} \alpha_i \,, \tag{6}$$

where the α_i's are as in Eq. (4).

In order to apply Eq. (6), we need to know the *marginal* distributions of the $\hat{p}_i$'s. When the indicator sequences $\{I_{i,k}\}$ in Eq. (1) and Eq. (2) are covariance stationary and satisfy certain fairly general regularity conditions, see e.g. [1, p. 478], then the $\hat{p}_i$'s in Eq. (3) have normal limiting distributions, provided $0 < p_i < 1$. Nevertheless, the confidence intervals based on the normal approximation are not always appropriate. To improve the confidence intervals for p_i the so-called arcsin-transform, $\hat{y}_i = 2\sin^{-1}\sqrt{\hat{p}_i}$, is one of the methods proposed; see e.g. [2, p. 699].

In addition to the (approximate) distributions of the $\hat{p}_i$'s, we need to specify the desired half-widths. Typically, the closer to 0 or 1 the proportion is, the narrower the desired confidence interval is. We specify the desired confidence interval $[\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}]$ using the arcsin-transform as follows. The desired confidence interval for $y_i = 2\sin^{-1}\sqrt{p_i}$ is symmetric around $\hat{y}_i$ and independent of i: $[\hat{y}_i - \delta_y, \hat{y}_i + \delta_y]$, so we need to specify only a single precision

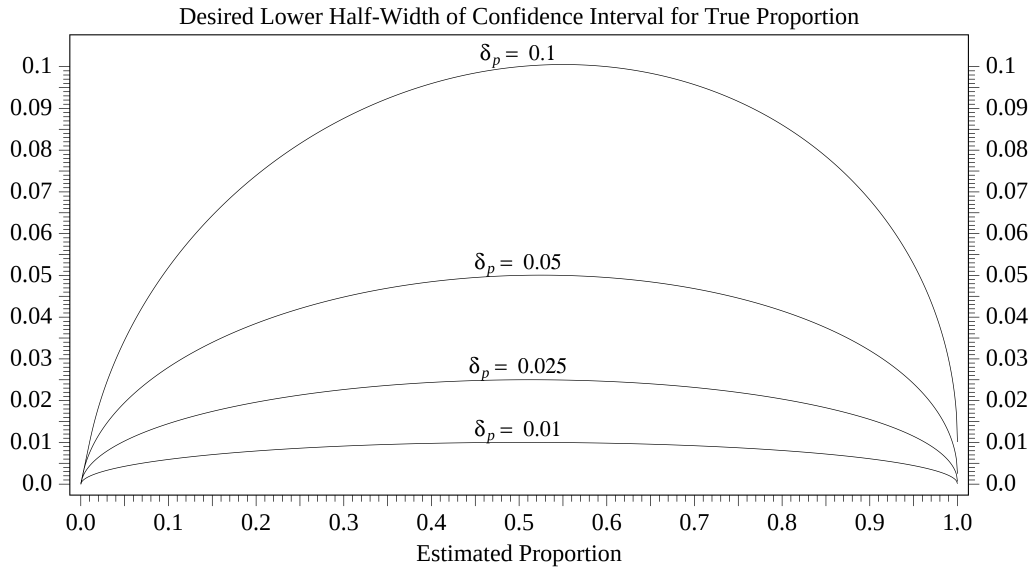

Fig. 5. δ_{L}, Desired Lower Half-Width of Confidence Interval for Proportion p, when the Desired Confidence Interval for $y = 2\sin^{-1}\sqrt{p}$ is $[\hat{y} - \delta_y, \hat{y} + \delta_y]$ and $\delta_y = \sin^{-1}(2\delta_p)$.

parameter δ_y. A common δ_y is indeed appropriate since the arcsin-transform suitably scales the desired confidence intervals for the p_i's; see Fig. 5. The desired upper half-widths are mirror images of the desired lower half-widths.

The analyst may find it difficult to choose a suitable δ_y. One solution is to solve δ_y from Eq. (7) or Eq. (8) given below for a specific $\hat{p}_i$ and δ_p; for details see [7].

$$\delta_{\mathrm{L}_i} = \begin{cases} \sqrt{\hat{p}_i(1-\hat{p}_i)}\,\sin(\delta_y) & \\ -(0.5-\hat{p}_i)(1-\cos(\delta_y)) & \text{, when } \hat{p}_i \geq (\sin(\delta_y/2))^2 \\ \hat{p}_i & \text{, when } \hat{p}_i < (\sin(\delta_y/2))^2 \end{cases} \tag{7}$$

$$\delta_{\mathrm{U}_i} = \begin{cases} \sqrt{\hat{p}_i(1-\hat{p}_i)}\,\sin(\delta_y) & \\ -(\hat{p}_i-0.5)(1-\cos(\delta_y)) & \text{, when } \hat{p}_i \leq 1-(\sin(\delta_y/2))^2 \\ 1-\hat{p}_i & \text{, when } \hat{p}_i > 1-(\sin(\delta_y/2))^2 \end{cases} \tag{8}$$

When the indicator sequences are long, that is large M in Eq. (3), we can use the Student t-distribution to approximate the marginal distribution of $(\hat{y}_i - y_i)/s_{\hat{y}_i}$, where $s^2_{\hat{y}_i}$ is the estimated variance of $\hat{y}_i$. Based on this approximation, we obtain $\hat{\alpha}_i$, the estimated confidence level of the confidence interval for p_i;

see Eq. (4):

$$\hat{\alpha}_i = 1 - \widehat{\Pr}\{p_i \in [\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}]\} = 2F_{\nu_i}\left(-\delta_y\sqrt{\hat{p}_i(1-\hat{p}_i)}\Big/ s_{\hat{p}_i}\right) , \quad (9)$$

where δ_{L_i} and δ_{U_i} depend on δ_y as in Eq. (7) and Eq. (8), F_{ν_i} is the cumulative distribution function of the Student t-distribution with ν_i degrees of freedom, and $s^2_{\hat{p}_i}$ is the estimated variance of $\hat{p}_i$ with ν_i degrees of freedom. Since we cannot assume that the indicator sequences are independent, we use the spectral method introduced in [3] to obtain the $s_{\hat{p}_i}$'s.

Eq. (9) and the Bonferroni's inequality imply that an estimated lower bound of the simultaneous confidence level is

$$\begin{aligned}&\widehat{\Pr}\{p_i \in [\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}] \text{ for every } i = 1,\ldots,m\} \geq 1 - \hat{\alpha}\\ &= 1 - 2\sum_{i=1}^{m} F_{\nu_i}\left(-\delta_y\sqrt{\hat{p}_i(1-\hat{p}_i)}\Big/ s_{\hat{p}_i}\right) . \qquad (10)\end{aligned}$$

When we wish to control the precision of estimated cell proportions, we let the simulation continue until $1 - \hat{\alpha}$ is at least the prespecified desired joint confidence level.

If the true proportion p_i is close to 0 (or close to 1), it will be highly probable that $\hat{p}_i(1-\hat{p}_i) = 0$, which implies that Eq. (9) gives $\hat{\alpha}_i = 2F_{\nu_i}(0) = 1$, so the simulation should be continued forever. Moreover, the sequence $\{0,\ldots,0\}$ (or $\{1,\ldots,1\}$) does not provide any information about $\mathrm{Var}(\hat{p}_i)$. Therefore, we propose the following practical treatment of small proportions: We specify a tolerance of small proportion p_{small}. If $\hat{p}_i < p_{\text{small}}$ or $\hat{p}_i > 1 - p_{\text{small}}$, then we do not use Eq. (9) but we set

$$\hat{\alpha}_i = \begin{cases} 0, \text{ if } Mp_{\text{small}} \geq 16 ,\\ 1, \text{ if } Mp_{\text{small}} < 16 .\end{cases}$$

By setting $\hat{\alpha}_i$ to 1, we force the simulation to be continued so that reliable estimates for all $p_i \in [p_{\text{small}}, 1 - p_{\text{small}}]$ may be obtained. The constant 16 is based the results in [7].

Fig. 6 visualizes the estimated histograms, including precision assessment information. The diamond mounted on the top of the bar presents the desired confidence interval $[\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}]$: the top is $\hat{p}_i + \delta_{\mathrm{U}_i}$, the middle is $\hat{p}_i$, and the bottom is $\hat{p}_i - \delta_{\mathrm{L}_i}$. The number above the diamond is the estimated marginal confidence level. When $\hat{p}_i < p_{\text{small}}$ or $\hat{p}_i > 1 - p_{\text{small}}$, then the diamond is not shown since then we have not estimated the precision.

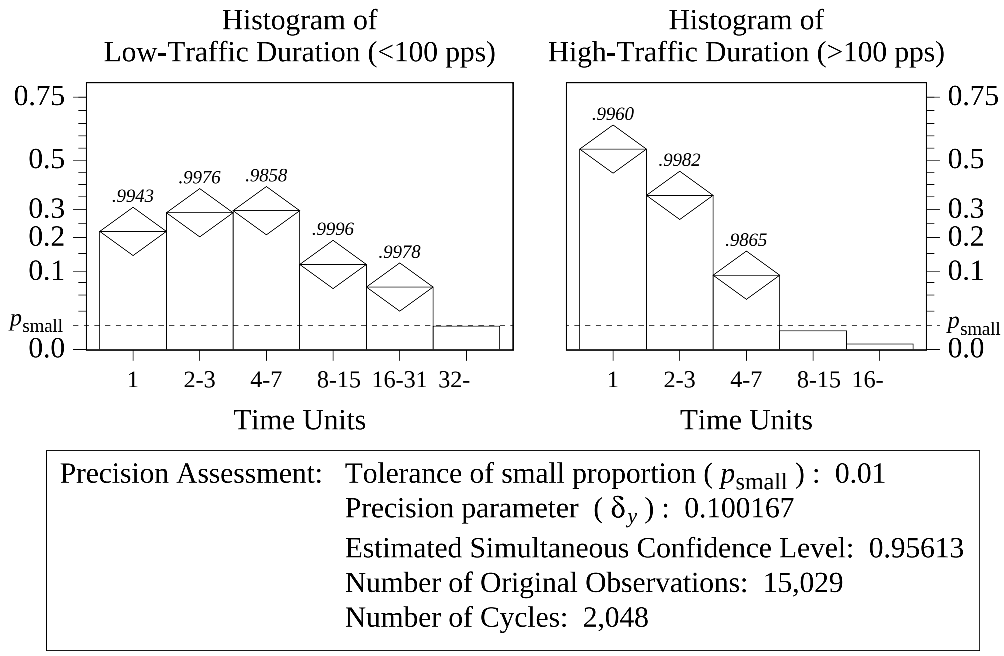

Fig. 6. Example of Estimated Histograms With Precision Assessment Information

4 Example Applications

As examples of estimating variability in queueing systems, we simulated three different models: an M/M/1 queue, an M/D/1 queue, and an embedded autoregressive process.

In the M/M/1 queue with arrival rate $\lambda = 1$ and service rate $\mu = 1.25$ we examined the queue length ($X_j \in \{0, 1, 2, \ldots\}$). The threshold level ($x_*$) was 0.5, which implies that we are interested in the durations of the server's idle and busy periods. The sampling interval was 1.

In the M/D/1 queue with $\lambda = 125$ and a constant service time $S = 0.002$ we examined the server utilization. The threshold level was 0.25, which is the steady-state utilization. The sampling interval was 1.

Our embedded AR process models the packet rate to and from a file server that was used as the example in Section 2. The threshold level was 100 packets per second, which is quite close to the steady-state mean; see Fig. 1.

With each model we executed 101 independent simulation runs. The parameters controlling the precision were $\alpha = 0.1$ (desired simultaneous confidence level of 0.90) and $\delta_y = \sin^{-1}(0.1)$, implying that at $\hat{p}_i = 0.5$ the desired confidence interval is $[0.45, 0.55]$; see Fig. 5. The tolerance of small proportion (p_{small}) was 0.01. In the histograms (see vertical axis in Figs 7–9) we used a

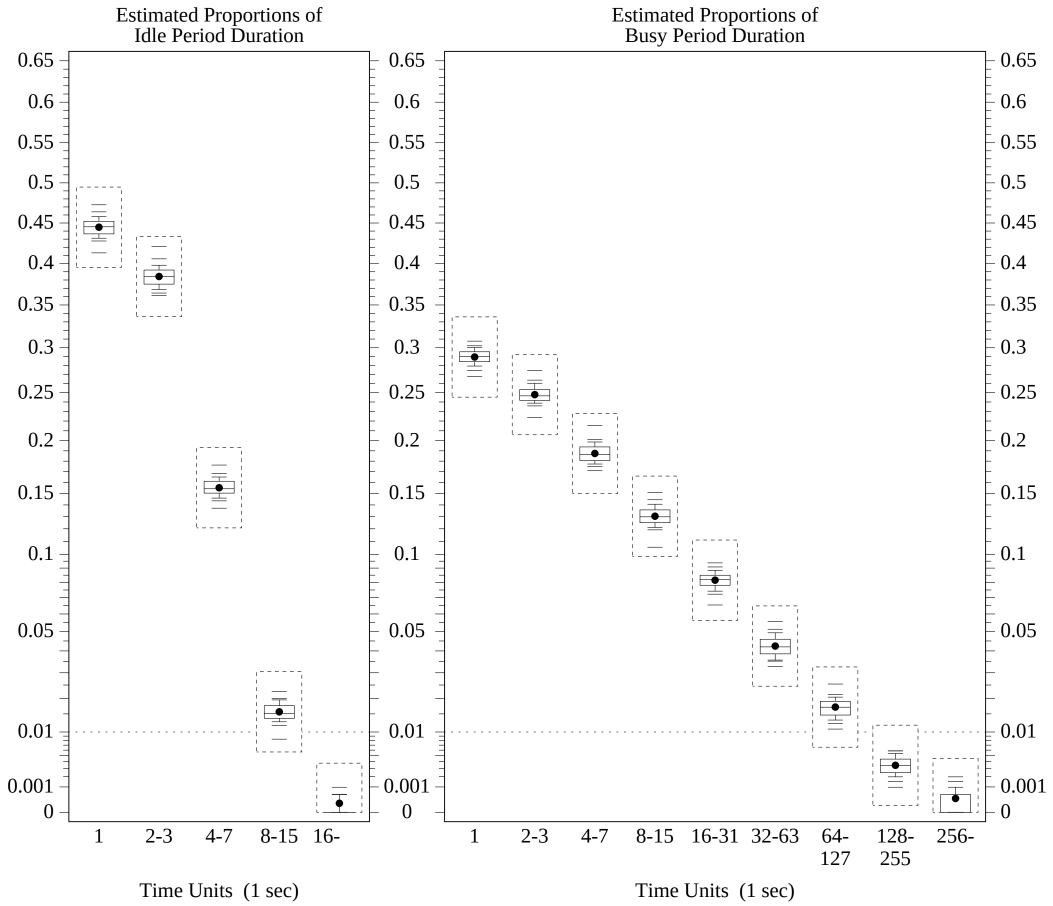

Fig. 7. Boxplots of Estimated Proportions for Histograms in M/M/1 Queue; Target Precision: 90% of proportions simultaneously inside the dashed boxes; $\alpha = 0.1,\ \delta_y = \sin^{-1}(0.1),\ p_{\text{small}} = 0.01$

$\log_2$ related classification in which the boundaries are powers of 2.

The estimated proportions for the histograms are summarized in Figures 7–9, which provide the *boxplots* of estimated proportions. In these boxplots a bullet ‘ • ’ denotes the mean of the 101 estimates. The three quartiles are visualized by a box ‘ ’. The whiskers ‘ ——| ’ reach from the 25^{th} to the 10^{th} percentile and from the 75^{th} to the 90^{th} percentile. The smallest and largest estimate together with the 5^{th} and 95^{th} percentile are marked by ticks ‘ | ’. The dashed boxes denotes the values of $\hat{p}_i$, for which $p_i \in [\hat{p}_i - \delta_{\mathrm{L}_i}, \hat{p}_i + \delta_{\mathrm{U}_i}]$. In other words the dashed boxes show the range of estimates that fulfil the precision requirement.

As the figures indicate, all the estimated proportions satisfy the target precision. This is essentially more than requested, but not surprising:

(i) Bonferroni's inequality provides a lower bound only (which may not be tight) for the confidence level;

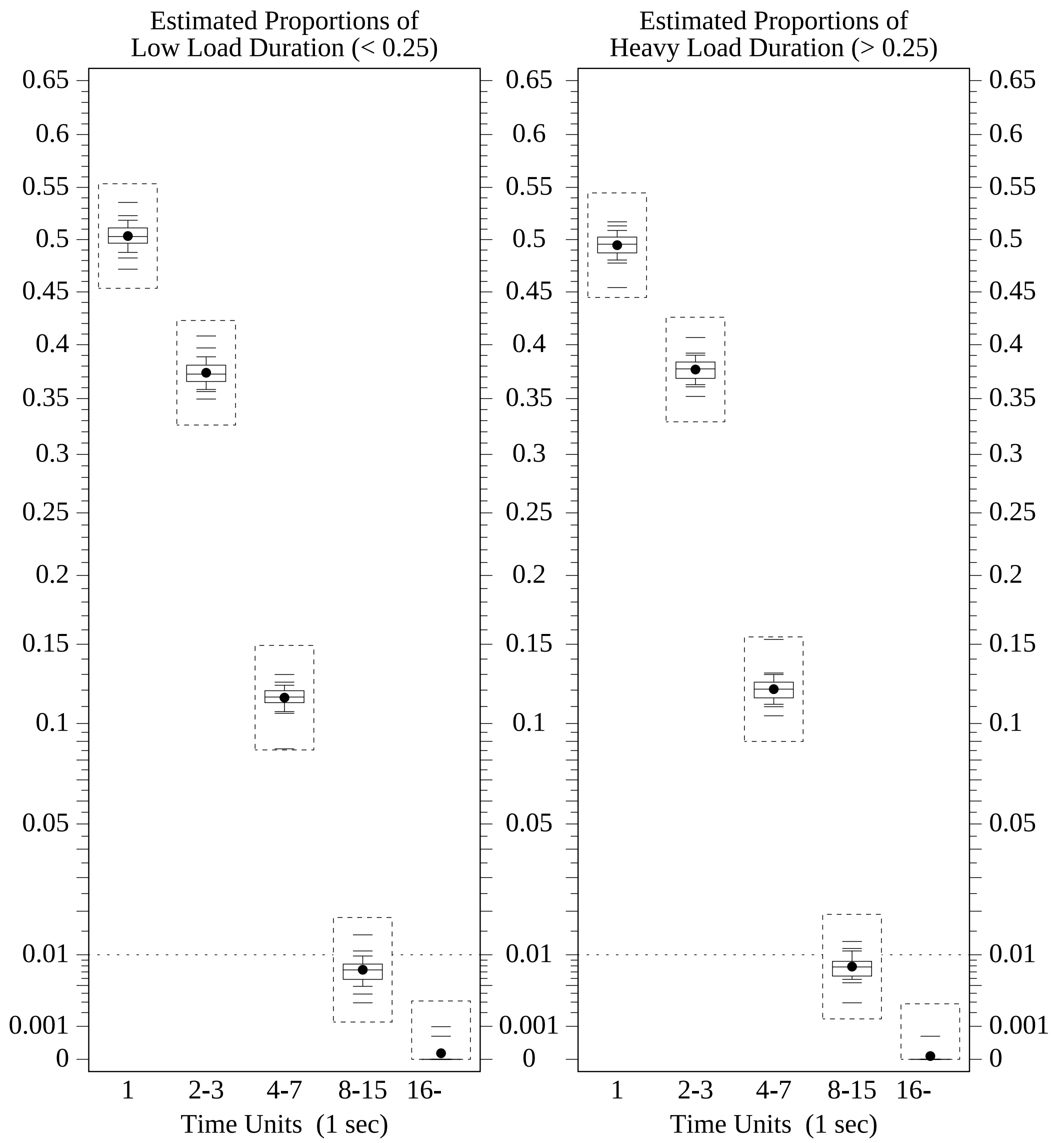

Fig. 8. Boxplots of Estimated Proportions for Histograms in M/D/1 Queue; Target Precision: 90% of proportions simultaneously inside the dashed boxes; $\alpha = 0.1,\ \delta_y = \sin^{-1}(0.1),\ p_{\text{small}} = 0.01$

(ii) with $p_{\text{small}} = 0.01$ the minimum number of cycles is 2,048 if at least one estimated proportion is less than 0.01, and in each model we indeed have estimated proportions less than 0.01.

In order to assess the amount of computing required in estimating the proposed histograms for summarizing the variability, we also estimated the corresponding means: the mean queue length in the M/M/1 queue, the mean utilization of the server in the M/D/1 queue, and the mean packet rate in the embedded AR process. The objective of estimating these means was to learn how hard the simultaneous estimation of two histograms is compared to the estimation of a single mean. When the mean $\hat{\theta}$ was estimated, the parameters controlling

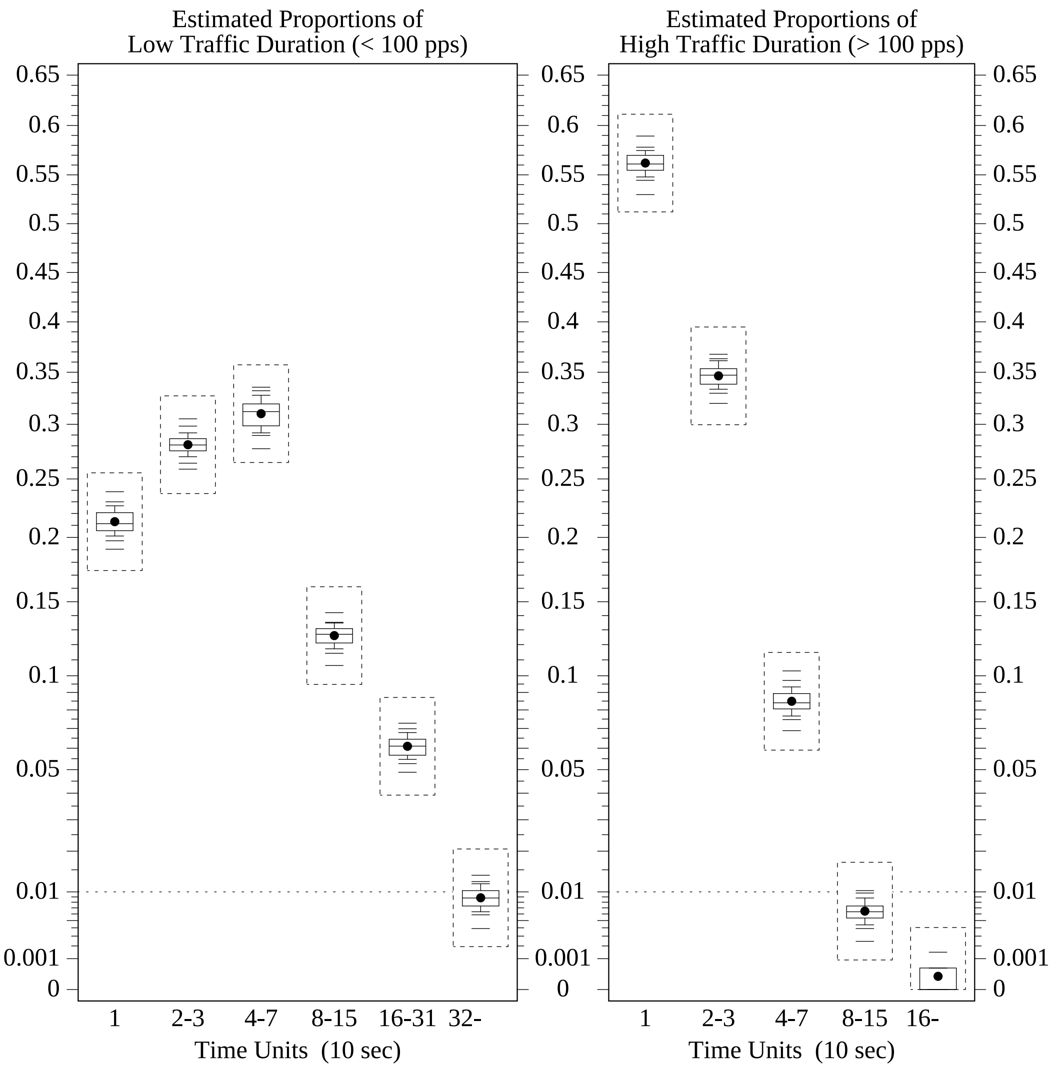

Fig. 9. Boxplots of Estimated Proportions for Histograms in Embedded AR Process; Target Precision: 90% of proportions simultaneously inside the dashed boxes; $\alpha = 0.1,\ \delta_y = \sin^{-1}(0.1),\ p_{\text{small}} = 0.01$

the precision were $\alpha = 0.1$ and $\varepsilon = 0.1$: the simulation was continued until our desired precision is $\Pr\{\hat{\theta}(1-\varepsilon) \le \theta \le \hat{\theta}(1+\varepsilon)\} = 1-\alpha$, which implies $s_{\hat{\theta}}/\hat{\theta} \le \varepsilon/t_\nu(1-\alpha/2)$.

The *run lengths* are compared in Table 1. The table indicates that the estimation of histograms can be either considerably cheaper or more expensive than the estimation the corresponding mean, when $\alpha = 0.1$, $\varepsilon = 0.1$, and $\delta_y = \sin^{-1}(0.1)$. It should be noted that the settings $\varepsilon = 0.1$ and $\delta_y = \sin^{-1}(0.1)$ are not necessarily equivalent. However, the comparison shows that the estimation of histograms does not necessarily require simulation runs of great lengths.

If we compare the *computer times*, then histograms are more expensive. This is due to the variance estimation: for each proportion in the two histograms we

Table 1
Comparison of run lengths when histograms and means are estimated
histograms estimated: $\alpha = 0.1$, $\delta_y = \sin^{-1}(0.1)$
mean estimated: $\alpha = 0.1$, $\varepsilon = 0.1$

	M/M/1 queue		**M/D/1 queue**		**Embedded AR model**	
Run Length	Histo-gram	Mean	Histo-gram	Mean	Histo-gram	Mean
Mean	23,978	46,070	8,194	256	15,925	1,212
St.dev.	4,544	29,137	77	0	3,536	637
Min	20,801	16,384	7,963	256	14,480	256
Max	46,809	131,072	8,360	256	30,519	4,096

need an estimated variance. However, the spectral method can be implemented so that multiple variance estimates can be calculated efficiently.

5 Summary

We have proposed a way to characterize and estimate variability in a queueing system. The method is based on translating the problem into simultaneous estimation of several proportions. The examples demonstrate that the method controls the run length so that the estimated proportions have the precision requested. We have not presented a theoretical analysis since that analysis can be found in [7].

We have recognized that the number of parameters in our approach is quite high. We have observed that especially the threshold level (x_*) significantly affects the estimated variability. Moreover, suitable numbers of classes and suitable class boundaries may be found only through several pilot simulations. To overcome these difficulties we have integrated the approach into our prototype of interactive simulation tool described in [6]. This tool provides three windows when variability is estimated; Fig. 10 shows these windows: 1) the Time Series Window displays a part of the original output sequence and the selected threshold level, 2) the Control Window allows the user to easily change the parameters, and 3) the Precision Window displays the two histograms with the precision assessment information.

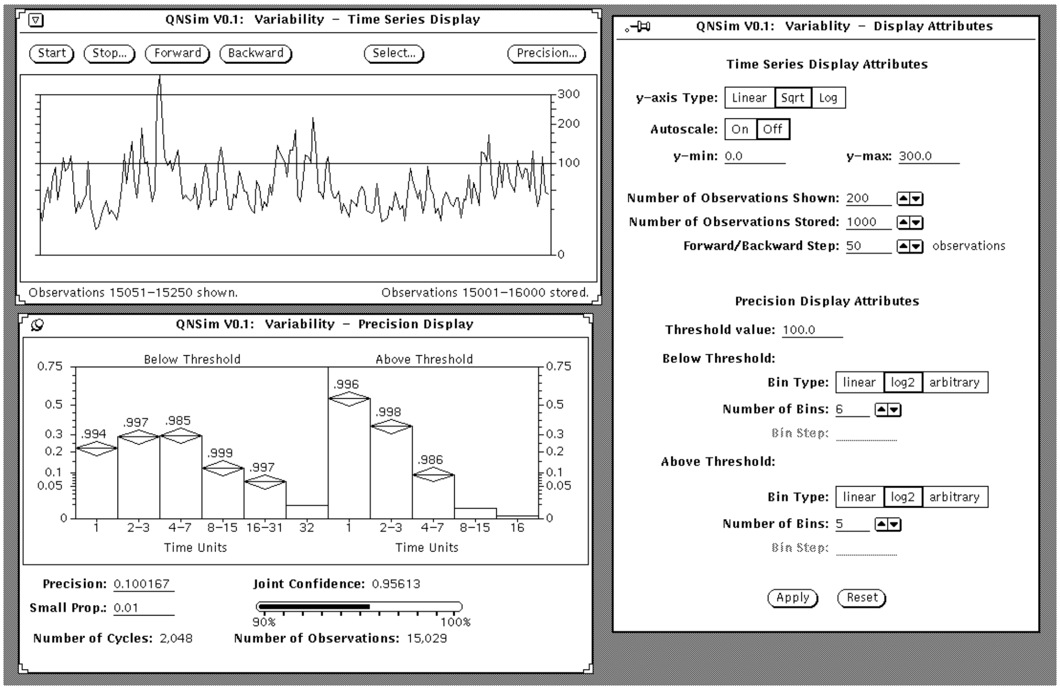

Fig. 10. Graphical User Interface for the Approach Proposed

References

[1] T. W. Anderson, *The Statistical Analysis of Time Series* (John Wiley & Sons, New York, 1971).

[2] A. Hald, *Statistical Theory with Engineering Applications* (John Wiley & Sons, New York, 1952).

[3] P. Heidelberger and P. D. Welch, A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM* 24 (4) (1981) 233–245.

[4] A. M. Law and W. D. Kelton, *Simulation, Modeling and Analysis*, 2nd Ed. (McGraw-Hill, New York, 1991).

[5] K. Pawlikowski, Steady-state simulation of queueing processes: A survey of problems and solutions, *ACM Computing Surveys* 22 (2) (1990) 123–170.

[6] K. E. E. Raatikainen, Controlling the precision of estimated means in interactive simulation, *Annals of Operations Research* 54 (1994) 481–505.

[7] K. E. E. Raatikainen, Simulation-based estimation of proportions, *Management Science* 41 (7) (1995) 1202–1223.