https://dx.doi.org/10.1109/UKSIM.2009.73

Mean and Standard Deviation Profiles for a Single-Server FIFO Queuing Node Carrying Heterogeneous Poisson and Pareto Traffic during a Bandwidth Probing Event: Analysis vs. Computer Simulation

M. J. Tunnicliffe, M. Hosseinpour
Faculty of Computing, Information Systems and Mathematics, Kingston University, Kingston-on-Thames, Surrey, KT1 2EE. +20-85472000+62674 M.J.Tunnicliffe@king.ac.uk

Abstract

Analytical expressions were obtained for the mean and standard deviation virtual waiting time profiles of a single-server FIFO queuing buffer loaded with heterogeneous (non-uniform packet-size) Poisson traffic disturbed by the passage of a bandwidth probing packet. The predictions were compared with the results of a computer simulation, using both Poisson and batch-Pareto (ON/OFF) cross-traffic arrival mechanisms. The model accurately predicted the simulated behaviour under Poisson traffic, though it showed a slight but consistent tendency to under-predict the standard deviation. For the scenarios reported, the Pareto results also showed reasonably close agreement with the model, though without the tendency towards under-prediction.

List of Symbols

w	Virtual waiting time (VWT) (s).
λ	Arrival rate (packets/s).
t_s	Packet service time (s).
$S_1, S_2 ... S_P$	Cross-traffic packet sizes (bits).
α_j	Proportion of the total traffic carried by packets of size S_j.
β_i	Proportion of the total packet population being of size S_i.
S_c	Mean cross-traffic packet size (bits) weighted by α_j.
γ	Coefficient of variation for cross-traffic packet size weighted by α_j.
S_p	Probe packet size (bits).
c	Cross-traffic rate (bits/s).
l	Link bandwidth (bits/s).
ρ	Server utilization ($= c/l$).
$w(t)$	Mean VWT at time t (s).
$\sigma^2(t)$	VWT variance at time t.
$\bar{f}(w)$	Equilibrium distribution for w.
$\bar{w}$, $\bar{\sigma}^2$	Mean and variance of $\bar{f}(w)$.
$\tilde{f}(w,t)$	Transient component distribution for w at time t.
$\tilde{w}(t)$, $\tilde{\sigma}^2(t)$	Mean and variance of the transient VWT distribution *after* the probe-packet arrival, but *before* there is any significant chance of the queue being empty.
$\tilde{E}_t(w^n)$	n'th moment of the transient component distribution at time t.
$\bar{E}_t(w^n)$	n'th moment of the equilibrium component distribution at time t.

1. Introduction

Several algorithms (e.g. [1]) have been devised to monitor the properties of network links by injecting probe packets and measuring the resulting disturbances. Analysis of such information requires an accurate model of the behavior of transient queuing systems. Though an accurate model was devised by Park et al. [2], this was based on the *M/D/1* queuing system which is only applicable under homogeneous (constant packet-size) traffic. A simpler approximate model was proposed by one of the authors [3] and later extended to heterogeneous (variable packet-size) scenarios [4]. However, these models primarily addressed the *mean* queuing behavior whilst ignoring the statistical variance.

Here we devise a more generalized model, which yields expressions for the mean and standard deviation profiles of the queue's waiting time. The model's predictions are compared with the results of computer simulation.

2. Equilibrium Queue Behavior

We assume firstly that the probed network link can be modelled as a FIFO (first in first out) queuing system which begins in a state of statistical equilibrium. In the generalised *M/G/1* system with arrival rate λ and utilisation ρ, the first and second moments of the equilibrium virtual waiting-time[1] (VWT) are given by

$$E(w) = \frac{\lambda E\left(t_s^{\,2}\right)}{2(1-\rho)} \tag{1a}$$

$$E\left(w^2\right) = 2E^2(w) + \frac{\lambda E\left(t_s^{\,3}\right)}{3(1-\rho)} \tag{1b}$$

where t_s is the packet service time [5]. If α_j is the ratio of the total traffic (bits/s) carried by packets of size S_j then $\lambda = c\sum_{j=1..P}\alpha_j/S_j$. Now if β_i is the ratio of the number of packets of size S_i to the total population, then the *n*'th moment of service-time is given by $E\left(t_s^{\,n}\right) = \left(1/l^n\right)\cdot\sum_{i=1..P}\beta_i S_i^{\,n}$. Since $\beta_i = (\alpha_i/S_i)/\sum_{j=1..P}\alpha_j/S_j = (c/\lambda)(\alpha_i/S_i)$, the mean waiting time can easily be shown to be

$$\overline{w} = E(w) = \frac{c}{2l^2(1-\rho)}\sum_{i=1}^{P}\alpha_i S_i \tag{2}$$

and the corresponding variance

$$\begin{aligned}\mathrm{var}(w) &= E\left(w^2\right) - E^2(w)\\ &= E^2(w) + \frac{c}{3l^2(l-c)}\sum_{i=1}^{P}\alpha_i S_i^{\,2}\end{aligned} \tag{3}$$

Now if the packet size has a weighted mean $S_c = \sum_{i=1..P}\alpha_i S_i$ with a coefficient of variation $\gamma = \sqrt{\sum_{i=1..P}\alpha_i S_i^{\,2}/S_c^{\,2} - 1}$ then the equilibrium mean and variance become:

$$\overline{w} = \frac{cS_c}{2l(l-c)} \tag{4}$$

$$\overline{\sigma}^2 = \overline{w}^2\left[1 + \frac{4}{3}\left(\frac{l}{c} - 1\right)\left(1+\gamma^2\right)\right]. \tag{5}$$

This equilibrium is abruptly disturbed by the first probe packet of a sequence, and the residue of this disturbance carries information about the link's properties. The modelling of this disturbance is considered next.

3. Non-Equilibrium Queue Behavior

The probe packet creates transient profiles for both the mean and variance of the VWT. For simplicity we denote the packet's arrival time $t = 0$, at which instant the queue acquires S_p bits (in addition to the workload it already contains). Thus at $t = 0$ the mean VWT becomes $S_p/l + \overline{w}$ with a standard deviation $\overline{\sigma}$. For $t > 0$ the server feeds upon the queue, whilst further packets arrive; eventually the mean rates of arrival and departure equalise and the equilibrium condition is restored.

During this transient phase, the waiting-time can be divided into two components: The set of possibilities where the queue has not yet completely emptied, and the set of possibilities where the queue *has* fully emptied and is returning to its equilibrium state. Following the conventions adopted in the earlier papers [2,3] we refer to these as the *transient component* and the *equilibrium component* respectively.

First let us consider the transient component: According to the assumed Poisson arrival model, the mean number of packets of size S_i arriving during an interval t seconds is $\alpha_i ct/S_i$ with a variance equal to this value. Thus the total mean number of arrived *bits* is $\sum_{i=1..P}\alpha_i ct = ct$ and the corresponding variance is $\sum_{i=1..P}\alpha_i cS_i t = cS_c t$. Since during the same interval *lt* bits enter the server, the mean and variance waiting-time profiles are

$$\tilde{w}(t) = \overline{w} + \frac{S_p}{l} - \left(1 - \frac{c}{l}\right)t \tag{6}$$

$$\tilde{\sigma}^2(t) = \overline{\sigma}^2 + \frac{cS_c}{l^2}t \tag{7}$$

before there is any significant chance of the queue becoming empty. With its combination of discrete

[1] "Virtual waiting time" at time *t* is the time a hypothetical packet entering the system at time *t* would take to reach the server.

and continuous elements, the probability distribution will have a complex shape. However, we can approximate it using a continuous Gaussian distribution for all positive w, and a delta-function for $w = 0$ (the empty queue):

$$\tilde{f}(w,t) = f_{\tilde{w}(t),\tilde{\sigma}(t)}(w) + \int_{-\infty}^{0} f_{\tilde{w}(t),\tilde{\sigma}(t)}(w)dt \cdot \delta(w) \qquad (8)$$

where $f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$

for which the first and second moments are:

$$\tilde{E}_t(w) = \int_0^{\infty} w \cdot \tilde{f}(w,t)dw = \frac{\tilde{\sigma}(t)}{\sqrt{2\pi}} \exp\left[-\frac{\tilde{w}^2(t)}{2\tilde{\sigma}^2(t)}\right] + \frac{\tilde{w}(t)}{2}\left[1 + \operatorname{erf}\left(\frac{\tilde{w}(t)}{\tilde{\sigma}(t)\sqrt{2}}\right)\right] \qquad (9)$$

$$\tilde{E}_t(w^2) = \int_0^{\infty} w^2 \cdot \tilde{f}(w,t)dw = \frac{\tilde{w}(t)\tilde{\sigma}(t)}{\sqrt{2\pi}} \exp\left(-\frac{\tilde{w}^2(t)}{2\tilde{\sigma}^2(t)}\right) + \frac{\tilde{w}^2(t)+\tilde{\sigma}^2(t)}{2}\left[1 + \operatorname{erf}\left(\frac{\tilde{w}(t)}{\tilde{\sigma}(t)\sqrt{2}}\right)\right] \qquad (10)$$

Although the empty queue recovers its equilibrium behaviour gradually, this is difficult to model analytically. We therefore use an approximation employed in earlier papers [3,4] whereby equilibrium is restored abruptly Δt seconds after the queue becomes empty. Thus the equilibrium components of the first and second moments of waiting time are given by:

$$\bar{E}_t(w) = \bar{w}\int_{-\infty}^{0} \tilde{f}(w,t-\Delta t)dw = \frac{\bar{w}}{2}\left[1 - \operatorname{erf}\left(\frac{\tilde{w}(t-\Delta t)}{\tilde{\sigma}(t-\Delta t)\sqrt{2}}\right)\right] \qquad (11)$$

$$\bar{E}_t(w^2) = \left(\bar{\sigma}^2 + \bar{w}^2\right)\int_{-\infty}^{0} \tilde{f}(w,t-\Delta t)dw = \frac{\bar{\sigma}^2 + \bar{w}^2}{2}\left[1 - \operatorname{erf}\left(\frac{\tilde{w}(t-\Delta t)}{\tilde{\sigma}(t-\Delta t)\sqrt{2}}\right)\right] \qquad (12)$$

for all $t > \Delta t$. The overall mean profile $w(t)$ for $t \geq 0$ is the sum of the first moments of both distributions

$$w(t) = E_t(w) = \begin{cases} \tilde{E}_t(w) & 0 \leq t \leq \Delta t \\ \tilde{E}_t(w) + \bar{E}_t(w) & t > \Delta t \end{cases} \qquad (13)$$

and the corresponding variance

$$\sigma^2(t) = E_t(w^2) - E_t^2(w) \qquad (14)$$

where

$$E_t(w^2) = \begin{cases} \tilde{E}_t(w^2) & 0 \leq t \leq \Delta t \\ \tilde{E}_t(w^2) + \bar{E}_t(w^2) & t > \Delta t \end{cases}. \qquad (15)$$

The one remaining problem is to find an expression for the time-lag Δt. A convenient scheme (devised by one of the authors [3]) is to consider the waiting-time profile $w_0(t)$ for a queue released from zero-size at time $t = 0$. Replacing this with an abrupt step function $\bar{w} \cdot H(t-\Delta t)$ is equivalent to replacing the first derivative $w'(t)$ with $\bar{w} \cdot \delta(t-\Delta t)$. Aligning the centroids of the two functions yields the following expression for computing Δt from $w_0(t)$:

$$\Delta t = \frac{1}{\bar{w}}\int_0^{\infty} t \cdot w_0'(t)dt = \lim_{T\to\infty}\left[T - \frac{1}{\bar{w}}\int_0^{T} w_0(t)dt\right]. \qquad (16)$$

This formula was been successfully applied to homogeneous traffic in an earlier paper [3], where Δt was found to follow the empirical formula

$$\Delta t = \frac{S_c}{l} \cdot \frac{0.333}{(1-\rho)^{2.19}}. \qquad (17)$$

Here we investigate the validity of this formula given the heterogeneous traffic profiles used in [4][2].

[2] A less accurate technique of establishing Δt was used in [4] which preceded the final published version of [3].

Numerical estimations of $w_0(t)$ were obtained by averaging the results of 1,000 independent simulations and computing the integral of Eqn. 16 from zero to the earliest instant at which $w_0(t) > \overline{w}$. Figure 1 compares the results with the prediction of Eqn.16, showing that the model is still valid under heterogeneous traffic. (Throughout this paper, simulation was performed using a queue class-library written in C++.)

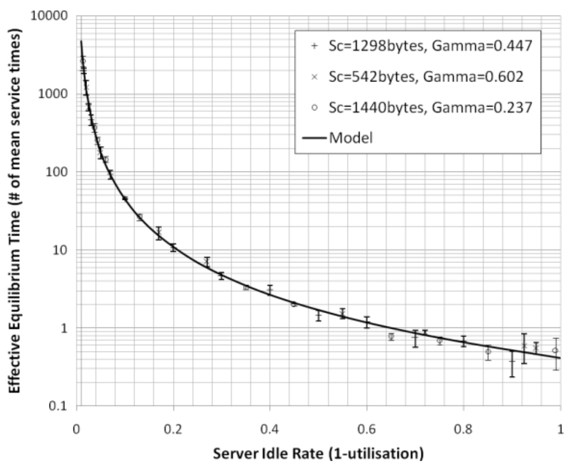

Figure 1. Simulated equilibrium time lag measured for three traffic profiles compared with the model. (Data-points represent the average of 5 runs of 1,000 independent simulations. Error bars indicate 95% confidence intervals.)

4. Comparison with Simulation Data

Figure 2 shows mean virtual waiting time profiles for four simulation scenarios. Figure 2(a) shows a homogeneous traffic scenario, in which S_p=1500bytes, S_c=100bytes and $\gamma = 0$. Figure 2(b) shows the same experiment performed using heterogeneous traffic in which 48% of packets contained 40bytes, 24% 80bytes, 16% 120bytes and 12% 160bytes. (This maintains S_c=100bytes while making γ=0.447.) Figure 2(c) and (d) illustrate the effect of changing the probe packet size (reducing it from 1500 to 500 bytes) and increasing utilization respectively. In all four scenarios, the measured and predicted mean VWT's are almost identical.

Figure 3 shows the standard deviation profiles corresponding to the mean profiles of Figure 2. While the model prediction is generally close to the data and follows a similar profile, there is a slight but persistent tendency towards under-prediction, even when the queue is in equilibrium. Though the authors have no explanation for this anomaly, it is too small to have any major practical significance.

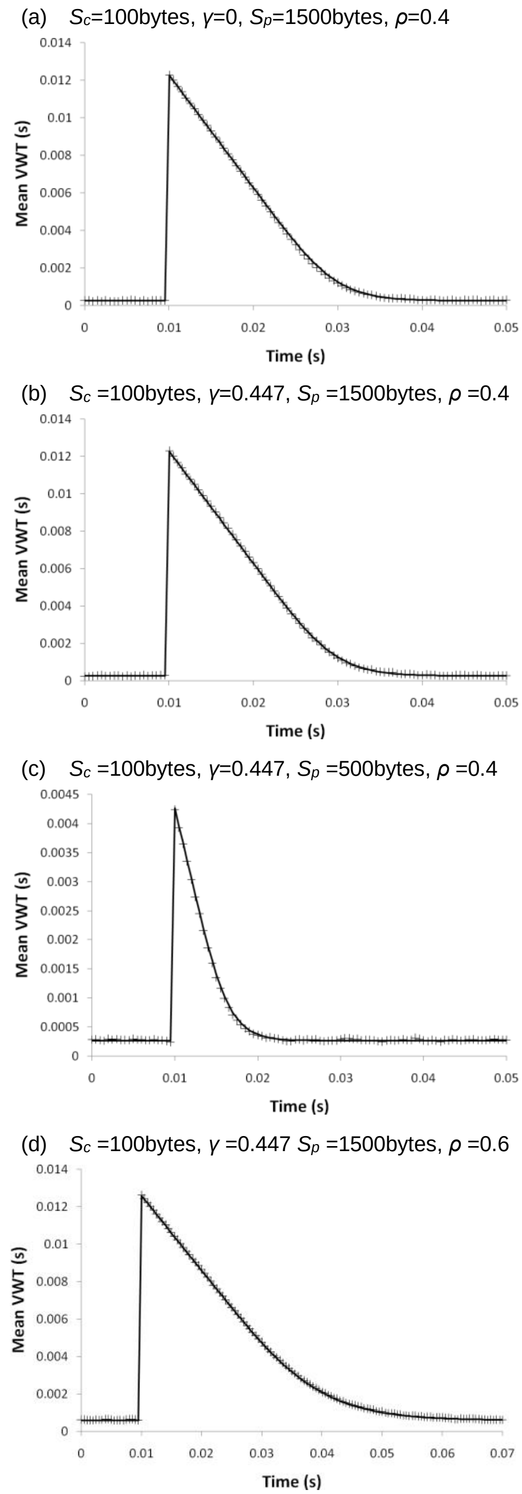

Figure 2. Mean VWT profiles. The link bandwidth in all four experiments was 1Mbit/s. (Data points represent the average of 2000 simulations, solid lines represent model predictions.)

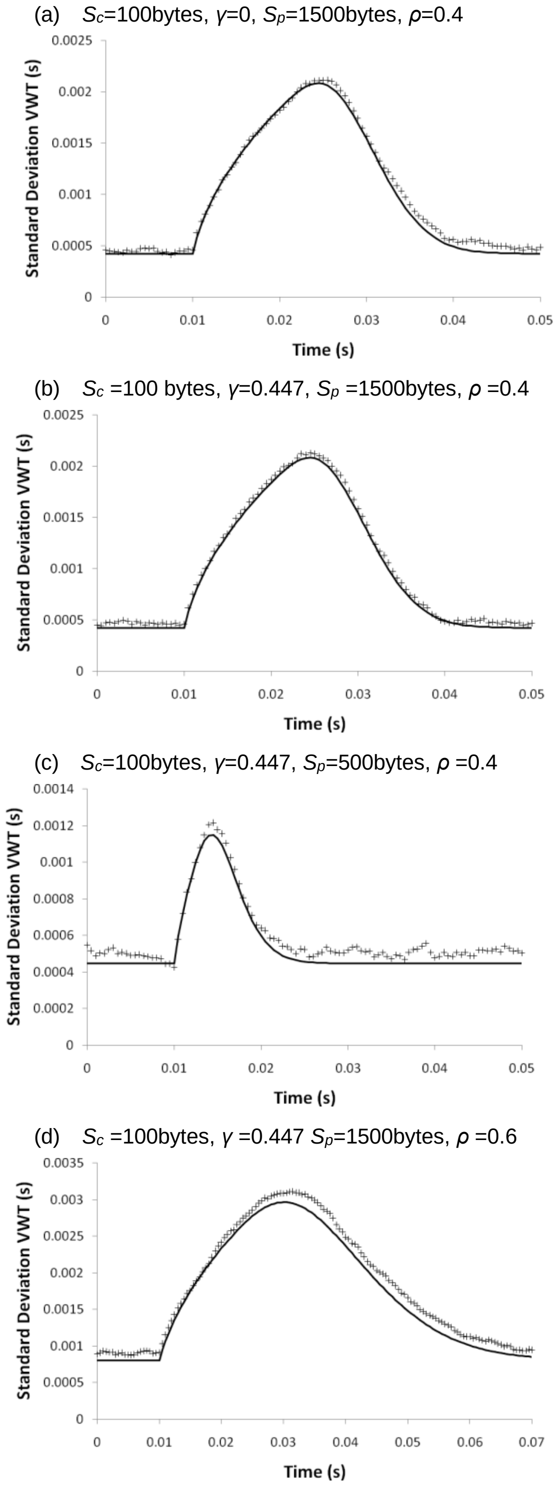

Figure 3. Standard deviation virtual waiting time profiles. The link bandwidth in all three examples was 1Mbit/s. (Data points represent the average of 2000 independent simulations, solid lines represent model predictions.)

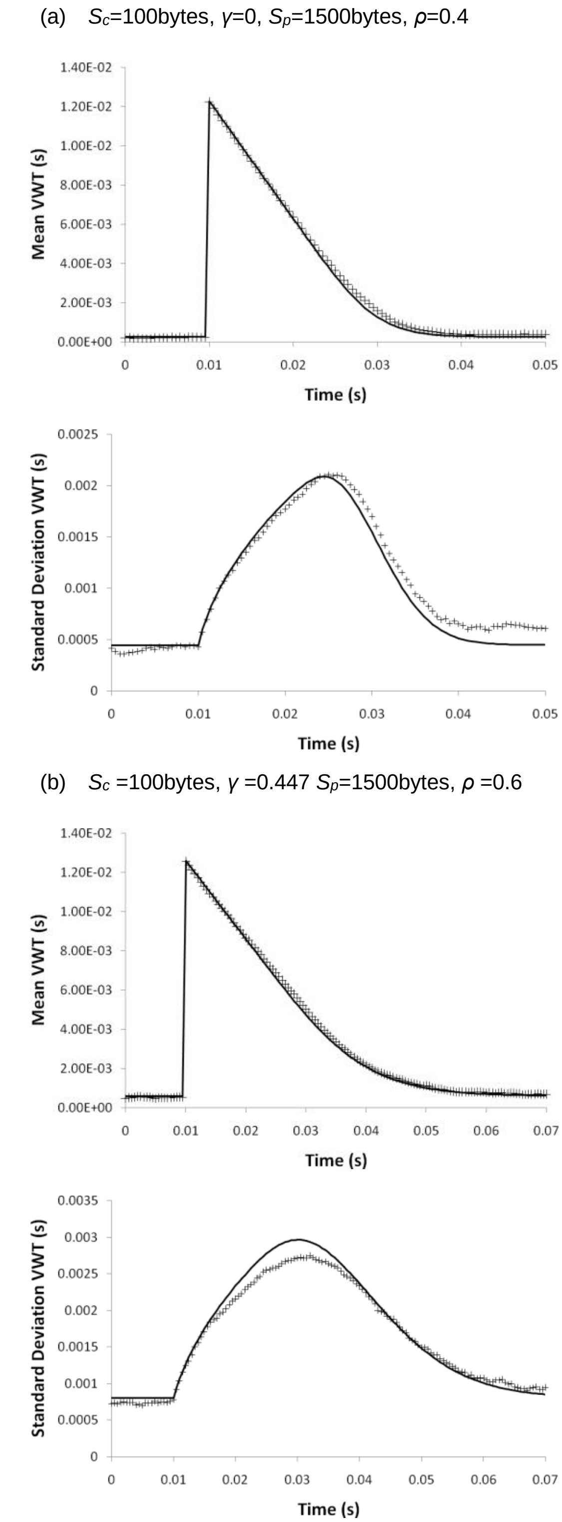

Figure 4. Mean and standard deviation virtual waiting time profiles for the system carrying Pareto ON/OFF traffic. The link bandwidth in all three examples was 1Mbit/s. (Data points represent the average of 2000 independent simulations, solid lines represent model predictions.)

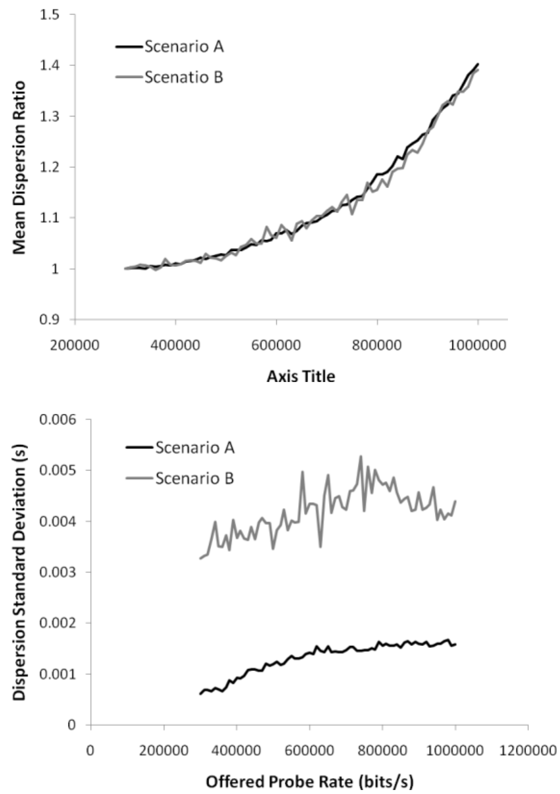

Figure 5. Probing results obtained from two network scenarios: (A) A two-hop path carrying 50byte packets and (B) a one-hop network path carrying 800byte packets. The mean dispersion ratios are practically identical, but the standard deviations significantly different. (Offered probe rate=probe packet size/packet separation)

A possible weakness of the model is its assumption of a Poisson arrival mechanism. However, an interesting outcome of [2] was that a model based on this assumption nonetheless provides a reasonable approximation of a queue loaded with inherently bursty Pareto ON-OFF traffic. We therefore test our model against such a scenario. Pareto traffic-generation is based on a scheme outlined by Pitts and Schormans [6]: Pareto-distributed ON times (during which packets were transmitted) were interspersed with Poisson-distributed OFF times, mimicking the arrival of self-similar packet-batches. Mean ON and OFF times were 0.03 and 0.05 seconds respectively, with a minimum on-time of 0.01 seconds.

Figure 4 shows some typical results. While the the profiles still agree fairly well with the model predictions, they are not so close as the corresponding Poisson results. Also, there is no longer a consistent tendency towards underestimation of the standard deviation.

5. Conclusions

In this paper we have extended an earlier model for mean dynamic queuing response to predict the variance profiles of virtual waiting time, and showed that the results are approximately consistent with simulation data obtained using Poisson and Pareto ON/OFF traffic. We have also showed that the empirical formula for the queue's empty-to-equilibrium time-lag developed for homogeneous traffic in [3] remains accurate when the packet-size is non-uniform.

The importance of variance information in a network probing experiment is illustrated in Figure 5. In this experiment (based on [1]), the residue of a probe-packet's disturbance profile is detected as an additional latency by one or more further probe-packets, and hence an increase in the packets' temporal separation (or dispersion). Mean dispersion ratios are plotted against the offered probing rate (probe packet size/packet separation) for two scenarios: A single hop path carrying large packets and a two-hop path carrying much smaller packets. The simulations produced near-identical *mean* dispersion curves, suggesting (incorrectly) that the underlying scenarios were also identical. However, the standard deviation profiles are significantly different. We hope that further developments of the model presented in this paper will help to resolve such ambiguities and allow underlying network infrastructure and traffic to be correctly inferred from probing data.

6. References

[1] Melander,B Björkman,M, Gunningberg,P, "A New End-to-End Probing and Analysis Method for Traffic Estimating Bandwidth Bottlenecks", *Proc. IEEE Globecom'00, San Francisco*, CA, USA, Nov. 2000.

[2] Park, K-J, Lim,H, Choi,C-H, "Stochastic Analysis of Packet-Pair Probing for Network Bandwidth Estimation", *Computer Networks*, 50, 2006, pp. 1901-15.

[3] Tunnicliffe,M.J., Winnett,M, "An Approximate Stochastic Analysis of the Packet-Pair Probing Technique for Available Bandwidth Estimation", accepted for publication in *International J. of Comm. Systems*, 2008.

[4] Tunnicliffe,M.J., "Stochastic Analysis of the Packet-Pair Bandwidth Probing Event under Heterogeneous Cross-Traffic", *Proc. 10th International Conference on Computer Modelling and Simulation*, IEEE Computer Society, 2008, pp.691-6

[5] Kleinrock,L, "Queuing Systems: Volume 1: Theory", Wiley, NY, 1975, p.201.

[6] Pitts,J.M. Schormans,J.A. “Introduction to IP and ATM Design and Performance”, Wiley, 2000, p.293.