Stochastic Process Models for Packet/Analytic-Based Network Simulations

Robert G. Cole, George Riley, Derya Cansever and William Yurcik

***Index Terms*—Hybrid Simulation Models, Event Driven Simulations, Network Models, Stochastic Queues and Diffusion Equations.**

Abstract

WE present our preliminary work that develops a new approach to hybrid packet/analytic network simulations for improved network simulation fidelity, scale, and simulation efficiency. Much work in the literature addresses this topic, including [10] [11] [8] [12] [13] and others. Current approaches rely upon models, which we refer to in this paper as *Deterministic Fluid Models* [9] [12], to address the analytic modeling aspects of these hybrid simulations. Instead we draw upon an extensive literature on stochastic models of queues and (eventually) networks of queues to implement a hybrid stochastic model/packet network simulation. We will refer to our approach as *Stochastic Fluid Models* throughout this paper. We outline our approach, present test cases, and present simulation results comparing the measured queue metrics from our approach for hybrid simulation to those of a deterministic fluid model hybrid simulation and a full packet–level simulation. We also discuss plans for future areas of research on this approach.

I. Introduction

Consider a network model as found in Figure 1. The network model is comprised of nodes and communications links. Associated with each communications link is a queue. Traffic can arrive and depart the network model at each node in the network. Each node also carries internal traffic which is forwarded throughout the network between the traffic source and destination nodes.

In hybrid simulation models, some of the traffic is handled via analytic methods and some of the traffic is handled via more CPU and memory intensive discrete-event, packet-level handling. The more traffic modelled analytically, the more efficient the simulation becomes (in the general case), and thus the simulation can scale to larger networks. Typically, there is a distinction drawn between foreground traffic, which is of primary interest, and background traffic, which exists solely to provide competition to the foreground traffic being measured. In the hybrid approach discussed here, analytic models are used is to estimate the impact of the background traffic on the foreground traffic, which is handled explicitly via discrete-event handling with full packet–level detail. In the remainder of this paper, we will assume that the analytically modeled traffic represents, in some sense, background traffic and that the explicitly handled packet traffic is foreground traffic. We will use the terms *background traffic* and *foreground traffic* to distinguish between the analytically modeled versus the explicitly handled event packet traffic. Of course other ways to divide up the analytically modeled and packet handled traffic are possible.

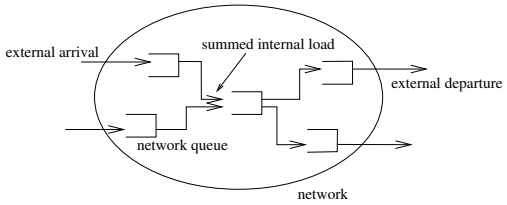

Fig. 1. An example network of queues with internal and external arrival and departure traffic.

There are two interesting aspects of the analytic modeling in the context of a network of queues. One relates to the allocation and estimation of the intermediate load generated by the traffic at the nodes within the network based upon the assumed routing patterns, total estimated external traffic loads, finite queue sizes and associated network internal packet losses. The other aspect relates to methods of mixing, at a given queue, the analytically modeled background traffic with the explicitly handled foreground traffic. Our focus in this paper is the later; hence we concentrate on methods to mix the analytically modeled traffic with the event-driven, packet-level traffic at a given queue. A future paper will concentrate on the network equations.

The remainder of this paper is organized as follows: In the next section we discuss previous methods for hybrid network simulation which rely upon deterministic fluid model approximations to queue dynamics. In Section III we present our approach based upon methods in stochastic queue dynamics. In Section IV we develop our hybrid equations for a specific instance of a stochastic queue model, i.e., a Brownian Motion model based upon solutions to the Fokker-Planck Equation. In Section V we report on our initial simulations investigations comparing three test cases; one based upon pure event-driven packet-level sim-

R. G. Cole is with the Applied Physics Lab and the Department of Computer Science, Johns Hopkins University, Baltimore, MD. Phone: +1 443 778–6951, e-mail: robert.cole@jhuapl.edu

G. Riley is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. Phone: +1 404 463–1774, e-mail: riley@ece.gatech.edu

D. Cansever is the Program Director of Advanced Internetworking, SI International, Inc. Phone: +1 703 234–6960, e-mail: derya.cansever@si-intl.com

W. Yurcik is with the Department of Computer Science, University of Texas-Dallas, Dallas, TX. Phone: +1 309 531–1570, e-mail: wxy081000@utdallas.edu

ulations, one based upon hybrid deterministic fluid/packet simulation and one based upon our hybrid stochastic process/packet simulation. In Section VI we discuss conclusions and future investigations.

II. Deterministic Models

Most current approaches to performance speedup for network simulations involving hybrid event-analytic simulation rely on analytic models based upon deterministic Fluid Flow Approximations (FFAs), e.g., [7] [4] [10] [11] [8] [13]. We refer to these approaches as *Deterministic Models* because the evolution of the queue dynamics modeling the background analytic traffic is assumed to be a deterministic process captured in the form of differential equations. The fluid dynamics is derived from the integration of these equations. In some cases the differential equations are solvable, e.g., fixed arrival rate models, and numerical integration of the differential equations is not necessary. Both cases address variable mean arrival rates; one through time dependent variables in the differential equations [9] [13] and one through discrete event fluid models where rate changes are propagated throughout the network via events [4] [11] [7].

Two approaches to using deterministic analytic models for hybrid simulations are found in the literature. One approach, used in [13], divides the network into a hierarchy consisting of a high speed core and a lower speed edge. In the core, all traffic is modeled analytically and in the edge all traffic is modeled via packet-level, event-driven simulation. Packet traffic traversing the core is converted to fluid load on the core and is then converted back to packet-level traffic at the far edge.

Another approach, used in [4] [11] and [8], is to treat some of the traffic throughout the network analytically. Then develop methods to explicitly mix both analytic and packet traffic at each multiplexing point throughout the network. Our interests are in this later approach.

A. Deterministic/Packet Mixing Equations

Here, we develop the simplest, least assuming method to mix deterministic traffic with packet traffic. There exist methods to improve the fidelity of our example deterministic model, e.g., [4] [11] [8], but these improvements require *a priori* knowledge of the behavior of the foreground traffic, which in general is not known. This choice is made here for several reasons. First, we wish to make an equivalent comparison to our stochastic models presented below in order to access its ability to negate a reliance on *a priori* traffic knowledge. Second, our main focus in this paper is to access aspects of the viability of the stochastic modeling and mixing approach against full packet level simulations.

Assume that λ_d is the average deterministically modeled traffic rate into the queue over a time averaging interval δ. We assume that δ is large compared to a typical packet transmission time but small compared to the total simulated time. Note that we assume that the parameters describing the arrival processes, both deterministic and stochastic, in this paper to be fixed. Let μ be the server rate at the queue, i.e., the inverse of the link bandwidth. Let $w(t)$ be the total workload in the queue at time t, and $w_-(t_i)$ and $w_+(t_i)$ be the total workload in the queue just prior to and just following (respectively) a distinguished time event, e.g., the arrival of packet i at time t_i. Let w_i be the work carried by the ith packet to the queue. Finally, let w_{max} be the maximum amount of work (or backlog) in the queue due to its finite size.

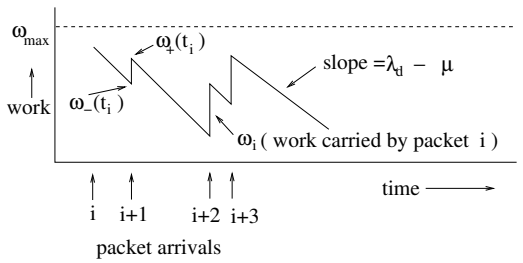

Fig. 2. An illustration of the hybrid deterministic modeling approach for mixing traffic at a queue.

Figure 2 gives a pictorial representation of the method used to mix analytically modeled background traffic with explicitly handled packet traffic. Individual packet arrival events are indicated along the bottom axis of the figure. These packets carry with them a given workload which is a function of their packet size and the link bandwidth serving the queue in question. If the packet is allowed to enter the queue, then the workload in queue jumps from $w_-(t_i)$ to $w_+(t_i)$ where the difference is the workload, w_i, carried by the packet. In between packet arrivals, the workload evolves deterministically. As mentioned, we assume the queue dynamics evolve at a constant rate between packet arrivals (as long as the workload does not exceed the maximum w_{max} or drops below the minimum of zero) given by the difference $\lambda_d - \mu$ which is fixed over the averaging time interval δ.

Then the explicit handling of the packet traffic and the process for updating the impact of the background traffic are as follows:

- If, during the current δ time averaging period, $\lambda_d < \mu$, i.e., the background traffic does not overflow the server rate. For each packet arrival bringing w_i work to the queue, do the following tasks:
  - Increment the packet arrival counter i at the queue.
  - Update the queue backlog just prior to the packet arrival time t_i, as

$$w_-(t_i) \leftarrow max[w_+(t_{i-1}) + (\lambda_d - \mu)(t_i - t_{i-1}), 0] \quad (1)$$

  - If $w_i < w_{max} - w_-(t_i)$, then accept the packet, update queue workload as $w_+(t_i) = w_-(t_i) + w_i$ and schedule the packet departure event at time

$$t_i^{service} = w_-(t_i)/\mu + t_i \quad (2)$$

  - Else, discard the packet and set $w_+(t_i) = w_-(t_i)$.
- Else $\lambda_d \geq \mu$. Here, the background fluid (which is assumed constant) is continually overflowing the queue.

Hence, we have that $w(t) = w_{max}$ and no packet traffic is allowed into the queue. Therefore, the model has the foreground traffic experiencing 100% packet loss. As mentioned, there are ways to overcome this effect, see, e.g., [4], but they rely on having an estimate of the foreground traffic rate.

Without *a priori* knowledge of the foreground traffic behavior, the deterministic model will deny the foreground traffic access to the buffer. This situation can be addressed by estimating foreground traffic load based on previous behavior, as has been done elsewhere.

III. Stochastic Models

In this section we present our approach to hybrid network simulations which relies upon stochastic models of the background traffic and explicit handling of packet events for the foreground traffic. As discussed previously, two aspects of this level of integration need consideration, i.e., methods to mix the analytic background traffic with the explicit event handling of the foreground traffic at a common queue, and methods to estimate the arrival processes of the background traffic at the intermediate queues in the network based upon network routing and external arrival processes. We concentrate on the former in this paper.

A. Stochastic/Packet Mixing Equations

Here we describe our method to mix the stochastic modeled background traffic and the packet traffic at a given queue within the network. In some sense, our approach defines a natural extension to the deterministic/packet hybrid approach described above. However, our approach eliminates the need for ad-hoc methods and assumptions to mixing analytic and packet traffic at common queues, does not require *a priori* knowledge of foreground packet behavior and applies independent of whether the overall load on the server is under or exceeds its capacity.

As above, assume that λ_s is the average background traffic (now treated as a stochastic process, hence the subscript "s") rate into the queue over a time averaging interval δ. This will represent the average rate of traffic associated with the analytically modeled traffic. However, we also characterize the variation in the background arrival process through its Coefficient of Variation (CoV) function, C_v [1]. The CoV is defined as $C_v = \sigma_s/\lambda_s^{-1}$ where σ_s is the standard deviation of the inter-arrival times for the incoming stochastic background traffic. Here, λ_s^{-1} is the mean inter-arrival time for the incoming stochastic background traffic. We assume the characterization of the background traffic is known, i.e., (λ_s, C_v), at each node (queue) in the network. Let μ be the server rate at the queue, i.e., the inverse of the link bandwidth. Let $w(t)$ be the total workload in the queue at time t, and $w_-(t_i)$ and $w_+(t_i)$ be the total workload in the queue just prior to and just following (respectively) a distinguished time event, e.g., the arrival of packet i. Let w_i be the work carried by the ith packet to the queue. Finally, let w_{max} be the maximum amount of work (or backlog) in the queue due to its finite size.

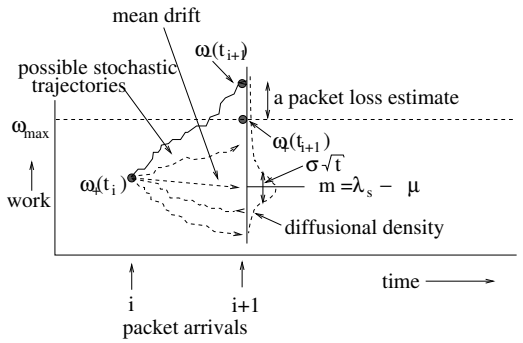

Fig. 3. An illustration of the hybrid stochastic modeling approach for mixing traffic at a queue.

The algorithm to determine a specific $w_-(t_i)$ given a value of $w_+(t_{i-1})$ is based upon the knowledge of the temporal evolution of the Cumulative Distribution Function (CDF) of the stochastic process. The CDF is defined as $F(w, t|w_0, t_0) = Pr[X_t < w_t|X_0 = w_0]$ and satisfies $\int_0^\infty d_w F(w, t|w_0, t_0) = 1$ and $lim_{t\to t_o} F(w, t|w_0, t_0) = \delta(w - w_0)$, the Dirac Delta Function. Our method applies for any chosen model of the CDF of the stochastic processes. Here we outline the mixing algorithm and in Section IV we give an example of one particular stochastic process, e.g., a Brownian Motion process from heavy traffic approximations of queues [3]. We use a CDF to determine probabilistically the evolution of $w_-(t_i)$ given $w_+(t_{i-1})$.

Figure 3 gives a pictorial representation of the method used for mixing analytically modeled stochastic traffic with explicitly handled packet traffic. Individual packet arrival events are indicated along the bottom axis of the figure. These packets carry with them a given workload which is a function of their packet size and the link bandwidth serving the queue in question. If the packet is allowed to enter the queue, then the workload in queue jumps from $w_-(t_i)$ to $w_+(t_i)$ where the difference is the workload carried by the packet. In between packet arrivals, the workload evolves according to a given stochastic process based upon load parameters which characterize the associated arrival process, i.e., λ_s and the coefficient of variation, C_v. Given the stochastic process and its Cumulative Probability Distribution, we can determine an appropriate $w_-(t_{i+1})$ given $w_+(t_i)$. We accomplish this by sampling against the given CDF for each packet interarrival.

The explicit handling of the packet traffic and the process for updating the impact of the stochastic traffic are as follows: For each packet arrival, do the following tasks:

- Increment the packet arrival counter i at the queue.
- Update the queue backlog just prior to the packet arrival time t_i, by sampling against the Cumulative Probability Distribution Function, $F(w, t_i|w_+(t_{i-1}), t_{i-1})$. Specifically, given a Distribution Function $F(X) = Pr[X < w]$, and a random variable Y uniformly distributed between 0 and 1, then we sample against the distribution by the formula $w = F^{-1}(Y)$. This yields a specific value for

$w_-(t_i)$.

- If $w_i < w_{max} - w_-(t_i)$, then accept the packet, update queue workload as $w_+(t_i) = w_-(t_i) + w_i$ and schedule the packet departure event at time

$$t_i^{service} = w_-(t_i)/\mu + t_i \tag{3}$$

- Else, discard the packet and set $w_+(t_i) = w_-(t_i)$.

The approach outlined above leads to a few observations. First, this approach is a natural extension to the methods previously discussed. Second, this approach explicitly accounts for background traffic loss and event packet loss for all arrival load conditions, not just for background loads less than the line rate. Specifically, the impact of background traffic on the packet traffic is handled at each packet arrival, while the impact of packet traffic on background traffic (and hence background traffic loss) is handled by resetting the initial queue state following each packet arrival. Third, unlike for deterministic background models where it is often necessary to know *a priori* the foreground load in order to get good estimates of foreground packet loss, this approach requires no such information or estimation. Instead, the foreground packet loss is a natural result of the fluctuations in the background traffic and the current load placed on the system by the foreground traffic.

The above procedure provides a better estimation of aspects of the foreground and background losses and their cross interactions.

IV. An Example Stochastic Model

Our approach to analytic/packet level mixing at a given queue within the network was described independently of the underlying stochastic process describing the background traffic. Hence, it is possible to rely upon a number of potential stochastic processes. We describe one here and rely upon it in our simulation results in Section V. Further, it is certainly possible to use several stochastic processes to model the background traffic based upon which model best represents the results as a function of system load and other aspects of the system model. This is a topic for future research.

The stochastic model we investigate here plays a prominent role in the work on heavy traffic limits in queuing models and is based upon the Brownian Motion stochastic process [3].

A. Brownian Motion Model

For our initial investigations, we assume that the stochastic process is well approximated by a Brownian Motion model which is strictly appropriate for heavy traffic limits, see, e.g., [3]. Given this, and assuming a *GI/G/1/K* queueing system, we then have that the Cumulative Distribution Function, $F(w_i, t_i | w_+(t_{i-1}), t_{i-1})$ satisfies the Fokker-Planck Equation [3] with appropriate boundary conditions to be discussed below, i.e.,

$$\frac{\partial}{\partial t}F = -m\frac{\partial F}{\partial w} + \frac{1}{2}\sigma^2\frac{\partial^2 F}{\partial w^2} \tag{4}$$

where $m = \lambda_s - \mu$, $\sigma^2 = \lambda_s \times C_v^2 + \mu \times C_{v,\mu}^2$, $F(w, t=0) = H(w - w_0)$, and $H(x)$ is the Heavy-side Function. For the purpose of this paper, we assume that all packets have the same size, hence we model a *GI/D/1/K* queueing system and have that the CoV for the service process is zero. This assumption implies that $\sigma^2 = \lambda_s \times C_v^2$.

The Fokker-Planck equation has an analytic solution derived in [3]

$$F(w,t|w_0,t=0) = \alpha \times \Phi(\frac{w - w_0 - mt}{\sigma\sqrt{t}}) + \beta \times e^{2mw/\sigma^2}\Phi(\frac{-w - w_0 - mt}{\sigma\sqrt{t}}) \tag{5}$$

$$F(w, t=0) = H(w - w_0) \tag{6}$$

$$\Phi(\frac{\pm w - w_0 - mt}{\sigma\sqrt{t}}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{\pm w - w_0 - mt}{\sigma\sqrt{t}}} e^{-x^2/2}dx \tag{7}$$

The multipliers, α and β, are determined by Boundary Conditions (BCs) of the system.

The most natural BCs found in the literature are

$$\lim_{w\to 0} F(w,t|w_0,t=0) = 0 \tag{8}$$

and

$$\lim_{w\to w_{max}} F(w,t|w_0,t=0) = 1 \tag{9}$$

where the first BC addresses the lower limit of zero on the work in the system and the later BC ensures that the finite size queue limits the work in the system to the maximum work allowed. Solving for α and β using these BCs, we get

$$\alpha^{-1} = \Phi(w_{max,+}) - e^{2mw_{max}/\sigma^2}\Phi(w_{max,-}) \tag{10}$$

and

$$\beta = -\alpha \tag{11}$$

where we have used the abbreviations

$$\Phi(w_{max,\pm}) = \Phi(\frac{\pm w_{max} - w_0 - mt}{\sigma\sqrt{t}}) \tag{12}$$

We also investigate the use of alternative BCs. Consider the system starting out with a very large initial value for the work, w_0. If we are interested in the predicted work in the system after a very short time interval, we would expect that a lower limit on the work in queue would be given by $w_0 - \mu t$. Hence, an alternative set of BCs, which we refer to BCs with a moving lower BC, is

$$\lim_{w\to max[0, w_0-\mu t]} F(w,t|w_0,t=0) = 0 \tag{13}$$

and

$$\lim_{w\to w_{max}} F(w,t|w_0,t=0) = 1 \tag{14}$$

Solving for α and β using these BCs, we get

$$\alpha^{-1} = \Phi(w_{max,+}) - \frac{\Phi(w_{max,+})\Phi(w_{max,-})e^{2m(w_{max}-w_{min})/\sigma^2}}{\Phi(w_{min,-})} \tag{15}$$

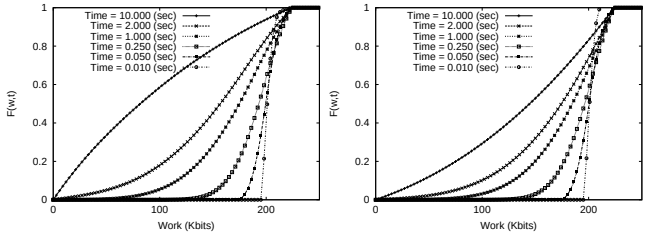

Fig. 4. Evolution of the CDF for two cases of load equal to 0.98 (left) and 1.02 (right).

and

$$\beta = \frac{-\alpha\Phi(w_{min,+})e^{2mw_{min}/\sigma^2}}{\Phi(w_{min,-}} \tag{16}$$

where $w_{min}(t) = max[0, w_0 - \mu t]$ and we have used the further abbreviations

$$\Phi(w_{min,\pm}) = \Phi(\frac{\pm w_{min} - w_0 - mt}{\sigma\sqrt{t}}) \tag{17}$$

These sets of equations for the CDF completely determine the time evolution of the stochastic system. The CDF with BCs at $w = 0$ and $w = w_{max}$ are well studied and are discussed in [5] and [6]. In Figure 4 we present some results for the shape of the CDF for the moving lower BCs, for two server utilization of 0.98 and 1.02 and various times. For these examples we used a buffer size of 50 packets of 560 Bytes each, an initial queue workload of 179200 Bytes representing a buffer at 80% occupancy and a server rate of 500 Kbps. We see that for both the cases, where loads are less than and greater than server capacity, there is a finite probability that the buffer is not fully occupied. This fact allows the packet level traffic access to the buffer even in server overload conditions, unlike the deterministic case.

V. Results

In this section we present the results for our investigations into the accuracy of the stochastic models in predicting queuing behavior in a single queue system. We compare the results of this method to the comparable event driven simulation model of the equivalent queuing system and of the equivalent deterministic background traffic model as described above. We use the Georgia Tech Simulation tool (GTNetS) [2] for our simulation studies of the base case (packet–level detail only) as well as the hybrid cases. For our simulation results, we developed two new modules within GTNetS. The two modules developed are a) A Hybrid Interface Module (HIM) which handled the mixing of the analytic background traffic with the foreground packet traffic, and b) A new traffic generation application, the Hyper-Exponential Arrival (HEA) module, which generates application level UDP packets according to a Hyper-Exponential Process [1]. This process models a source which alternates between a high rate and a low rate traffic source, where each rate period is exponentially distributed. Hence, by setting the rates equal, we reproduce a simple and relatively smooth Poisson traffic pattern. Our use of the Hyper-Exponential Process allows us to model bursty traffic sources, which are well known to exist in the Internet, by setting the rates to different values.

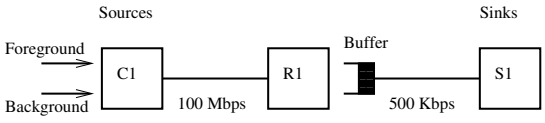

Fig. 5. The reference connection used for the simulation runs.

Our reference topology used for the simulation runs is shown in Figure 5. Here node C1 is the source for two different applications; one which is a 50Kbps Poisson packet stream representing our foreground UDP traffic case and one representing our Hyper-Exponential background UDP packet stream which can run at different rates and burstiness factors. Both application traffic, foreground and background, travel to node R1 over a 100 Mbps link and then onto node S1 over a 500 Kbps link. The link between R1 and S1 represents the queueing point, or bottleneck in our simulations. For each reference case studied, we ran three different simulation models; first is the base case which is a full event-driven packet level simulation of both the foreground and the background traffic, second is the deterministic hybrid case where the foreground traffic is treated at the packet level while the background traffic is treated analytically as a fixed rate fluid, and third is the stochastic hybrid case where the foreground traffic is treated at the packet level while the background traffic is treated analytically as a stochastic fluid with fixed mean and variance. To investigate impacts of different foreground packet flows on the hybrid systems, we also ran these three cases when the foreground traffic is a TCP-Reno stream from nodes C1 to S1 through R1, where we set the slow-start threshold to 20 KBytes.

For each of the above cases, we varied the mean and the CoV of the background Hyper-Exponential UDP stream to cover a range of traffic loads and loss rates. Table V lists the various mean background rates and example parameters representing the extreme CoV investigated for each mean rate. For all the results presented, we averaged the metric values over ten independent simulation runs with different random generator seedings for each. Each simulation run consisted of an initialization period of 50 seconds where just the background traffic processes run (in order to allow the system to equalibriate) and then we initiated the foreground process. We continued each simulation run for an additional 1000 seconds.

We first present our results for the case of a UDP foreground packet-based stream. For this foreground process we measure the packet level delay and loss. In Figure 6 we show the results for mean delay estimates for the three cases of base, deterministic and stochastic traffic for a background stream with a CoV equal to unity. The rates simulated in the plots are 300, 350, 400, 450 and 500 Kbps. Note, however, that these are the application-level rates and due to packet overhead of 48 bytes per packet, the

TABLE I
PARAMETERS FOR THE HYPER-EXPONENTIAL BACKGROUND PROCESS.

λ_s^{-1} (Kbps)	C_v	t_H (msec)	r_H (Kbps)	t_L (msec)	r_l (Kbps)
300	1.000	200	300	200	300
300	1.511	200	500	200	100
400	1.000	200	400	200	400
400	1.753	200	700	200	100
410	1.000	200	410	200	410
410	1.753	200	710	200	110
420	1.000	200	420	200	420
420	1.753	200	720	200	120
430	1.000	200	430	200	430
430	1.753	200	730	200	130
440	1.000	200	440	200	440
440	1.753	200	740	200	140
450	1.000	200	450	200	450
450	1.340	200	700	200	200
500	1.000	200	500	200	500
500	1.413	200	700	200	300

respective line rates are 328, 383, 438, 492 and 530 Kbps. Here we see mean delay of the foreground stream increasing as the mean rate of the background traffic increases for all three cases. However, the stochastic case does a better job at tracking the delays recorded for the base simulation case, while the deterministic case underestimates the delays at lower loads and overestimates the delay at higher loads.

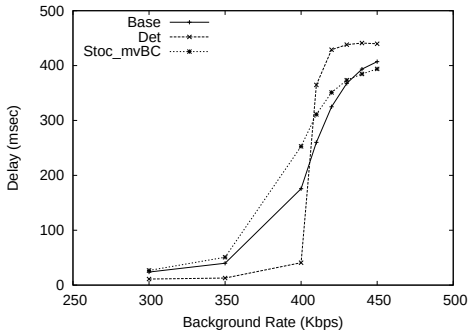

Fig. 6. Results for the UDP foreground delay versus background traffic rate.

Figure 7 presents results on the foreground packet loss probability for the different background traffic rates where the CoV is unity. The loss metric is a much more sensitive and challenging test of models due to its dependence upon the high percentiles of the work-in-queue probability distributions. We see extremely low packet losses for all cases for rates less than 400 Kbps, as expected. Above 400 Kbps, where the link is approaching a load near unity, the loss begin to increase. The losses for the deterministic model rapidly approach unity, because it has no mechanism to allow packet traffic entry into the buffer when the fluid input rate equals or exceeds the service rate of the system. While the results for the stochastic model are quite close to the base case loss results. As mentioned previously, the stochastic model explicitly accounts for the fluctuations in

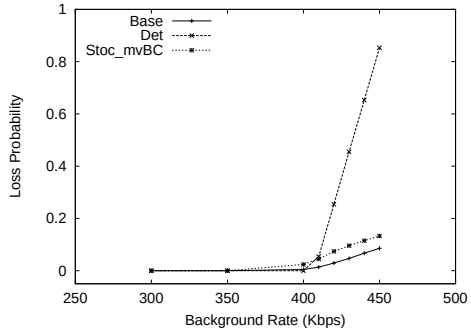

Fig. 7. Results for the UDP foreground loss probability versus background traffic rate.

the background traffic and this is what allows the packet traffic lossy entry into the buffer.

In the plots in Figure 8 we investigate the impact of increased CoV in the background stream on the foreground packet metrics. Figure 8 present the results for mean rates of 300, 400, 430 and 450 Kbps in the background streams. We see that the mean delay increases as the CoV increases for lower loads, while for higher loads the mean delay actually decreases for increasing CoV. The deterministic model does not account for variations in the dynamics of the background stream so it is incapable of tracking these effects. We show results for the moving lower BC discussed above for the stochastic Brownian Motion Model. The stochastic model captures the trends in the delay versus load and CoV. However, the trends are not as pronounced as in the base simulation case. This result will be addressed in future studies.

We now present our results for the case where the foreground traffic is a TCP packet stream. Due to TCP's adaptive windowing policy, the foreground traffic in these simulation runs will attempt to utilize the remaining bandwidth on the link between R1 and S1. Hence, the foreground loads vary for different cases of background traffic loads. We first show a set of results comparing the TCP goodput versus rate for the various simulation cases. These results are shown in Figure 9. We see that the goodput generally decreases with increasing background rate. Further, the deterministic model underestimates the TCP goodput, due to its over estimation of the foreground packet losses (from previous results above). While the stochastic models do a more reasonable job in tracking TCP throughput. Further, the estimates of the stochastic models improve for higher loads as expected. We show the results for both BCs discussed above. That the TCP goodput as modeled by the stochastic hybrid case tracks well the base case is a tribute to the models ability to estimate packet loss probability, as TCP goodput is highly non-linear with respect to packet loss.

We show the results for background streams with increasing Coefficients of Variation, where we are presenting the average goodput of the TCP foreground process. Figure 10 show the results for a mean background traffic rate of 300, 400, 450 and 500 Kbps. We observe that the TCP

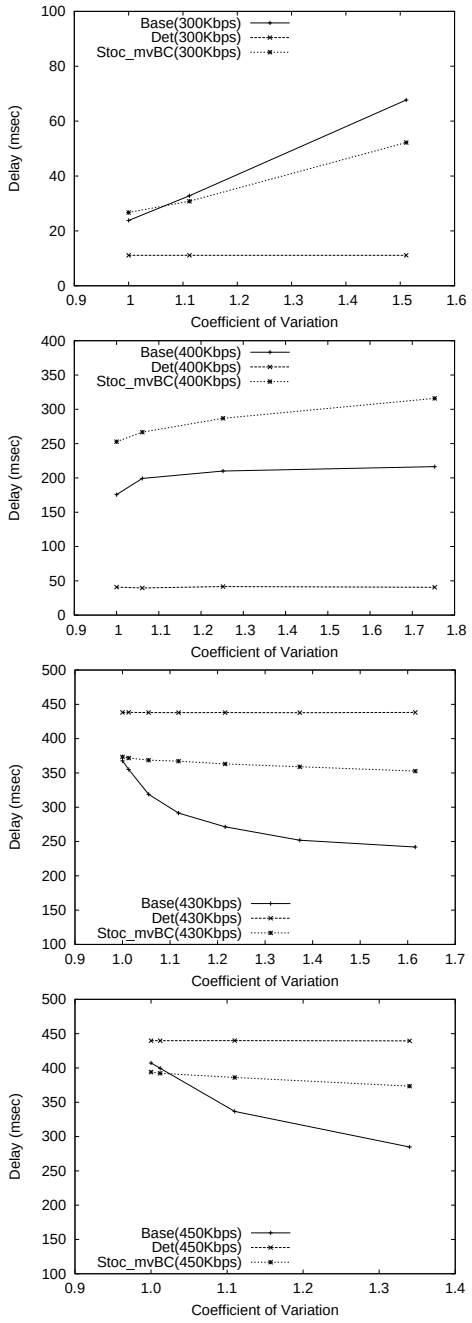

Fig. 8. Results for the UDP foreground delay versus CoV for various background traffic rates.

goodput for the base case shows a weak dependence on the CoV for lower loads and a greater, increasing dependence on CoV at higher loads. Further, the analytic models result in an underestimation of the TCP goodput for all cases. We present the results for both stochastic models, i.e., the models resulting from different lower BCs. The stochastic models capture the qualitative tends shown in the base case. However, quantitatively they underestimate the TCP goodput. As before, we see that the deterministic model results are independent of the CoV of the background traffic stream.

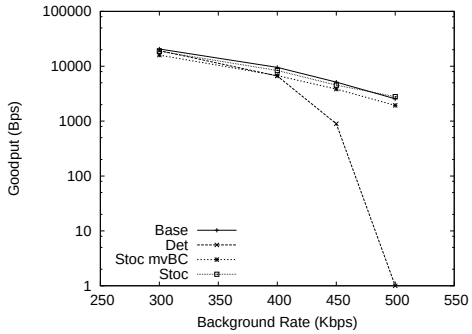

Fig. 9. Results for the TCP foreground goodput versus background traffic rate.

VI. Conclusions and Future Studies

We have presented the initial results of an investigation into the use of stochastic queueing models for the purpose of developing hybrid analytic and packet level, event driven network simulation tools. Our ultimate objective is to improve the scalability of network simulations, decrease their run-times and maintain high fidelity of their results. There are numerous aspects of this topic which require investigations, including packet/analytic model mixing at network queues and estimations of stochastic model parameters at interior queues in the network. In this paper, we have only begun investigating the initial question of how to mix stochastic analytic fluids with packet level events and performed some initial investigations into the fidelity of these methods. However, we are encouraged by these initial results. We believe the use of stochastic background models lead to a natural, unforced means of mixing packet and analytic traffic in common queues. This is not the case when relying upon deterministic models which force designers to make ad-hoc assumptions and decisions with respect to how to mix packet and analytic traffic and which often have to rely upon *a priori* knowledge of foreground traffic behavior.

There remains much work to further explore and develop a high fidelity hybrid stochastic and packet-level network simulation capability. Future work includes:

- Investigate the use of improved stochastic models and approximations from the field of Heavy Traffic results in queuing theory to improve the fidelity of the predictions.
- Extend our methods to networks of queues.
- Further explore the fidelity of these methods in the presence of other background traffic types and loads and their impact of other foreground application traffic.
- Investigate the benefits of using hybrid stochastic/packet simulation techniques when modeling processes with long tailed distributions.

VII. Acknowledgements

We would like to thank the reviewers for their detailed and insightful comments and in pointing out Reference [7].

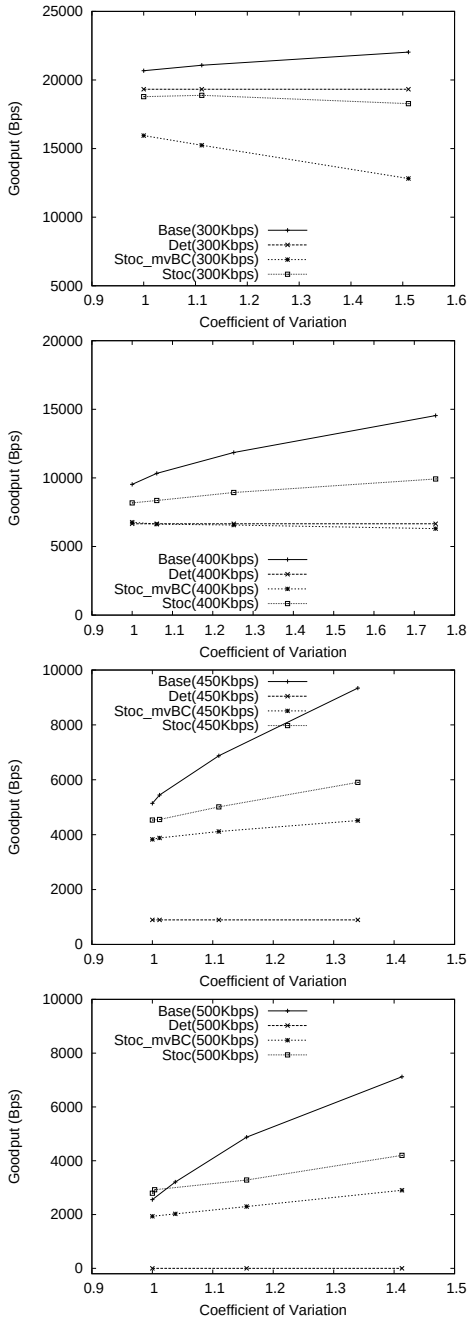

Fig. 10. Results for the TCP foreground goodput versus Cv for various background traffic rates.

References

[1] Cooper, R., *Introduction to Queueing Theory*, North-Holland, Oxford, 1981.

[2] Riley, G., *The Georgia Tech Simulation Tool*, http://maniacs.ece.gatech.edu/ 2008.

[3] Heyman, D., *A Diffusion Model Approximation for the GI/G/1 Queue in Heavy Traffic*, Bell Labs Technical Journal, Vol. 54, No. 9, November 1975.

[4] Kiddle, C., Simmonds, R., Williamson, C. and B. Unger, *Hybrid Packet/Fluid Flow Network Simulation*, 17th Workshop on Parallel and Distributed Simulation, June 2003.

[5] Kobayashi, H., *Applications of the Diffusion Approximation to Queueing Networks, I. Equilibrium Queue Distributions*, Journal of the Association for Computing Machinery, Vol. 21, No. 2, 1974.

[6] Kobayashi, H., *Applications of the Diffusion Approximation to Queueing Networks, I. Nonequilibrium Distributions and Computer Modeling*, Journal of the Association for Computing Machinery, Vol. 21, No. 3, 1974.

[7] Liu, B., Guo, Y., Kurose, J., Towsley, D. and W. Gong, *Fluid Simulation of Large Scale Networks: Issues and Tradeoffs*, PDPTA, 1999.

[8] Liu, J., *Parallel Simulation of Hybrid Network Traffic Models*, IEEE ACM PADS'07, San Diego, CA, USA, June 2007.

[9] Mitra, D., *Stochastic theory of a fluid model of producers and consumers coupled by a buffer*, Adv. in Appl. Prob., Vol. 20, 1988.

[10] Nicol, D. and G. Yan, *Simulation of Network Traffic at Coarse Time-scales*, IEEE ACM PADS'05, Monterrey, CA, USA, June 2005.

[11] Liljenstam, M., Nicol, D., Yuan, Y. and J. Liu, *RINSE: the real-time Interactive Network Simulation Environment for Network Security Exercises*, IEEE ACM PADS'05, Monterrey, CA, USA, June 2005.

[12] Misra, V., Gong, W.B. and D. Towsley, *Fluid-based Analysis of a Network of AQM Routers Supporting TCP Flows With an Application to RED*, Proceedings of ACM SIGCOMM'2000, 2000.

[13] Gu,Y., Liu, Y. and D. Towsley, *On Integrating Fluid Models with Packet Simulation*, IEEE INFOCOM'04, 2004.