A Linear System Approach to Serving Gaussian Traffic in Packet-Switching Networks*

Song Chong
Dept. of EECS
KAIST
Taejon, Korea
song@ee.kaist.ac.kr

Minsu Shin
Dept. of EECS
KAIST
Taejon, Korea
msshin@netsys.kaist.ac.kr

Hyun Hee Chong
Data Networks Lab.
Samsung Electronics
Sungnam, Korea
milou@samsung.co.kr

Abstract— **We present a novel service discipline, called *linear service discipline*, to serve multiple QoS queues sharing a resource and analyze its properties. The linear server makes the output traffic and the queueing dynamics of individual queues as a linear function of its input traffic. In particular, if input traffic is Gaussian, the distributions of queue length and output traffic are also Gaussian with their mean and variance being a function of input mean and input power spectrum (equivalently, autocorrelation function of input). Important QoS measures including buffer overflow probability and queueing delay distribution are also expressed as a function of input mean and input power spectrum. This study explores a new direction for network-wide traffic management based on linear system theories by letting us view the queueing process at each node as a linear filter.**

I. Introduction

As link speed increases, the number of flows to be aggregated into a single link increases. This fact naturally makes us invoke the central limit theorem and conjectures that aggregated traffic is more likely to be Gaussian as degree of aggregation becomes higher. In contrast with this conjecture, it is reported that not only the current IP traffic but also video traffic (which might dominate the future IP traffic) are self-similar [1]-[3], which implies that real IP traffic could have a heavy-tailed distribution which are not well characterized by a Gaussian distribution. The question remains as follows: do we really have to give up modeling traffic as a Gaussian process for the sake of accuracy? The practical answer seems to be no as long as a Gaussian process characterizes aggregated traffic with reasonable-degree of accuracy. In particular if we consider the current efforts toward IP QoS such as Diff-Serv [4], QoS differentiation is much more important than exact QoS guarantee which naturally implies that traffic characterization with reasonable-degree of accuracy might be tolerable as long as it does not prohibit us differentiating QoS.

Suppose that input traffic on a link is a stationary Gaussian process. The question we raise in this paper is as follows. Is there a service discipline which is linear such that Gaussian property of traffic stays intact as it travels through the queue and the corresponding queueing performance including queue length distribution, overflow probability and delay distribution can be easily analyzed based on linear system theories? Queueing dynamics in a typical single queue system with a fixed service capacity is nonlinear because the queueing process repeats idle and busy periods. As a result, the analysis of the queueing process and departure process is no longer trivial and the analysis, even if it exists, can hardly be extended to the network-wide queueing problem. A significant number of service disciplines have been proposed to serve multiple queues with QoS guarantee or QoS differentiation [5]-[10]. In turns out that none of them makes the queueing dynamics be linear because each queue inevitably repeats idle and busy periods in them.

We propose a new service discipline, we call it *linear service discipline*, to serve multiple QoS queues. It is adaptive such that the service rate allocated to each queue is a linear function of its input traffic. We show that the linear server makes both queueing process and departure process (output traffic) of individual queues be a linear function of its input traffic. As a consequence, the statistics of queue length, delay and output traffic can be readily expressed as a function of input traffic statistics. In particular, we show that if input traffic is Gaussian, the only statistics to be measured for the calculation of QoS are input mean and input power spectrum (equivalently, autocorrelation function of input).

II. Network Design

A. Queueing Architecture and Gaussian Input

In this subsection we describe our assumptions on queueing structure on an output link and input traffic fed into these output queues. As shown in Figure 1, we assume that there are logically-divided N QoS queues and

*This work was supported by the Ministry of Information & Communication of Korea under Support Project of University Foundation Research in 1999 and Support Project of University Information & Communication Research Center in 2000.

0-7803-7206-9/01/$17.00 © 2001 IEEE

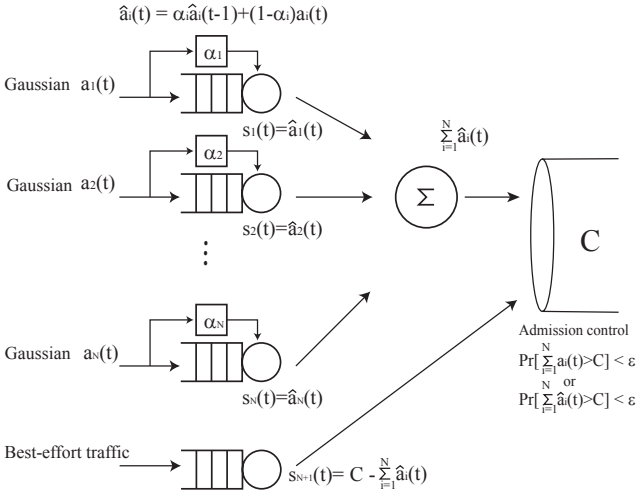

Fig. 1. Network Node Model

a best-effort queue for each output link. Such a queueing structure is fairly standard in modern QoS-capable switch/router design since its per-class queueing structure can scale with increasing number of flows in terms of scheduling complexity. More importantly, by aggregating input flows with identical QoS requirement into a single queue, the aggregated input traffic could exhibit more controllable and measurable statistics. For example, it is more likely to be a Gaussian process by the central limit theorem [11] as the number of flows to be aggregated increases. A representative example of such a per-class queueing structure is that of the DiffServ model for IP QoS [4].

Let $a_i(t)$ be the number of packets (or bytes) arriving at the i-th QoS queue in the time interval $[t, t+1)$. By appealing to the central limit theorem, we assume that $a_i(t)$ is a stationary Gaussian process whose time-independent statistics are simply characterized by two parameters, its mean and variance. The time-dependent statistics of $a_i(t)$ can be characterized by higher-order statistics including the autocorrelation function of the traffic, i.e., $E[a_i(t)a_i(t+\tau)]$.

In practice, the distribution of $a_i(t)$ tends to be more Gaussian as the length of measurement interval $[t, t+1)$ increases but at the cost of filtering out the short-term dynamics of packet arrivals.

B. Service Discipline

Let $s_i(t)$ be the service rate of i-th queue at time t. That is, $s_i(t)$ denotes the number of packets (or bytes) to be served in the time interval $[t, t+1)$. We propose a new service discipline as follows. For the QoS queues, $s_i(t) = \hat{a}_i(t)$, $\forall\ t$, $i = 1, \cdots, N$, where $\hat{a}_i(t)$ is computed by a low-pass filter

$$\hat{a}_i(t) = \alpha_i \hat{a}_i(t-1) + (1-\alpha_i) a_i(t), \quad 0 < \alpha_i < 1, \quad \hat{a}_i(-1) = 0. \tag{1}$$

For the best-effort queue, $s_{N+1}(t) = C - \sum_{i=1}^{N} \hat{a}_i(t)$, $\forall\ t$, where C is the capacity of output link. That is, the best-effort queue consumes the capacity unused by the QoS queues.

The defining feature of the proposed service discipline is that the service rate to be allocated to each queue is adaptive in time and purely a function of the input traffic fed into the queue.

In order to ensure each QoS queue to be served by the amount $s_i(t)$ for all times, it is necessary to have an admission rule such that $\Pr\left[\sum_{i=1}^{N} \hat{a}_i(t) > C\right] < \epsilon$ where ϵ is a sufficiently small positive number. The proposed service discipline, we call it *linear service discipline*, is non-work-conserving. Suppose that there is no best-effort queue. Then, the QoS queues are to be served by the amount of $\sum_{i=1}^{N} \hat{a}_i(t)$ in overall and the remaining capacity $C - \sum_{i=1}^{N} \hat{a}_i(t)$ is to be wasted whether or not the QoS queues are all idle. In general, a work-conserving server serves more packets than a non-work-conserving server for a same capacity whereas by allowing only eligible packets to be served, a non-work-conserving server can control the packet stream more tightly than a work-conserving server. As will be shown in the next section, the proposed linear server is able to control traffic such that Gaussian property of the traffic stays intact as it travels through the node and consequently through the entire network.

Let $q_i(t)$ be the queue length of i-th QoS queue at time t. Then, the dynamics of $q_i(t)$ is given by

$$q_i(t+1) = [q_i(t) + a_i(t) - \hat{a}_i(t)]^+ \tag{2}$$

where $[x]^+ = \max[0, x]$.

The coefficient α_i determines the degree of low-pass filtering. It is easy to show that the cutoff frequency of the low-pass filter (1), denoted by f_c in Hertz, satisfies that $\cos(2\pi f_c \Delta) = 1 - \frac{(\alpha_i - 1)^2}{2\alpha_i}$ where Δ denotes the unit time in seconds by which the discrete time in our model is defined. As $\alpha_i \to 1$, $f_c \to 0$, i.e., the passband of the filter becomes narrower.

The admission rule $\Pr\left[\sum_{i=1}^{N} \hat{a}_i(t) > C\right] < \epsilon$ can be loosened to $\Pr\left[\sum_{i=1}^{N} a_i(t) > C\right] < \epsilon$ at the cost of increased flow blocking probability. It is obvious that the latter condition implies the former condition because both $a_i(t)$ and $\hat{a}_i(t)$ are Gaussian with identical mean and the variance of $\hat{a}_i(t)$ is less than or equal to that of $a_i(t)$.

III. Network Analysis

In this section we show that both queue length and output traffic at each QoS queue are Gaussian provided that input traffic is Gaussian. Moreover, the mean and variance of both queue length and output traffic are given explicitly

as a function of the mean and power spectrum (equivalently autocorrelation function) of input traffic. Similarly, the queueing delay distribution and buffer overflow probability at each QoS queue are shown to be an explicit function of the mean and power spectrum of input traffic.

For the sake of simplicity we omit the subscript i to denote i-th QoS queue in the following sections.

A. Queueing Process

We evolve $q(t)$ in time recursively according to (1) and (2), starting with $q(0) = 0$. Then, we get at time t

$$\begin{aligned} q(t) &= [q(t-1) + a(t-1) - \hat{a}(t-1)]^+ \\ &= \alpha^t a(0) + \alpha^{t-1} a(1) + \cdots + \alpha a(t-1). \end{aligned} \quad (3)$$

Note that the queue length at time t is sum of arrivals in the past up to time 0 with geometrically decaying weights. This expression reveals two important facts. First, $q(t)$ is a Gaussian random variable. Second, $q(t) > 0$ for $t > 0$, i.e., the queue never becomes idle. Two immediate implications of the latter is that the departure rate at time t is equal to $\hat{a}(t)$ and the queue dynamics are linear, i.e., (2) becomes

$$q(t+1) = q(t) + a(t) - \hat{a}(t), \quad \forall\, t > 0. \quad (4)$$

Queueing dynamics in a conventional queueing system with a fixed capacity is nonlinear because the queueing process repeats idle and busy periods. As a result, the analysis of the queueing process and the departure process is no longer trivial and the analysis, even if it exists, can hardly be extended to the network-wide queueing problem. In contrast, by employing the linear service discipline, the network node can be viewed as a linear system and provided that input traffic is Gaussian, the queueing process is also Gaussian and the output traffic is simply a low-pass filtered version of input traffic and so Gaussian too. This fact indicates that rich linear system theories can easily be applied to the analysis of network-wide queueing problem.

Next we derive the mean and variance of the Gaussian queueing process. Let $\bar{q}$, σ_q^2 and $\bar{a}$ denote respectively the mean and variance of $q(t)$ and the mean of $a(t)$ in the steady state. Then, $\bar{q} = \lim_{t\to\infty} E[q(t)]$, $\sigma_q^2 = \lim_{t\to\infty} E\left[(q(t) - E[q(t)])^2\right]$ and $\bar{a} = \lim_{t\to\infty} E[a(t)]$. Using these definitions and (4), we can readily get

$$\bar{q} = \frac{\alpha}{1-\alpha}\bar{a} \quad (5)$$

and

$$\sigma_q^2 = \frac{\alpha^2}{\pi(1-\alpha^2)} \int_0^\pi P_a(\Omega) \frac{1-\alpha^2}{1+\alpha^2-2\alpha\cos\Omega} d\Omega - \bar{q}^2 \quad (6)$$

where $P_a(\Omega) = \sum_{\tau=-\infty}^{\infty} E[a(t)a(t+\tau)]\, e^{-j\Omega\tau}$ is the power spectrum of input traffic $a(t)$. Recall that the power spectrum of a discrete-time stationary stochastic process is the Discrete Fourier Transform(DFT) of its autocorrelation function.

This result implies two things. First, the Gaussian distribution of the queueing process is completely characterized by mean and power spectrum of the input traffic for a given α. Second, the mean and variance of the queueing process can be controlled as desired by properly choosing α for given input traffic.

Many techniques are available to estimate input mean and input power spectrum from on-line measurements. One of them will be presented in Section IV and used in simulation studies.

B. Overflow Probability

Once we know the distribution of the queueing process, we can readily estimate buffer overflow probability for a given buffer size B. To do this, we invoke the so-called Chernoff bound [11] so that

$$\Pr[q > B] \le e^{-\theta B} E[e^{q\theta}], \quad \theta > 0 \quad (7)$$

where $E[e^{q\theta}]$ is the moment-generating function of q. Since q is a Gaussian random variable, it is easy to show that $\ln E[e^{q\theta}] = \bar{q}\theta + \frac{1}{2}\sigma_q^2\theta^2$. Therefore, the tightest bound is given by

$$\ln \Pr[q > B] \le \min_{\theta>0}\{-\theta B + \bar{q}\theta + \frac{1}{2}\sigma_q^2\theta^2\}. \quad (8)$$

Since the right-hand side of the inequality has the minimum at $\theta = \frac{B-\bar{q}}{\sigma_q^2}$, we have the following large deviations approximation [12] for overflow probability

$$\ln \Pr[q > B] \sim -\frac{(B-\bar{q})^2}{2\sigma_q^2}. \quad (9)$$

Solving this for B, we know that the buffer requirement for a target overflow probability is given by

$$B \sim \bar{q} + K\sigma_q \quad (10)$$

where $K = \sqrt{-2\ln \Pr[q > B]}$.

A refined large deviations approximation [12] gives

$$\ln \Pr[q > B] \sim -\frac{(B-\bar{q})^2}{2\sigma_q^2} + \ln\left(\frac{\sigma_q}{\sqrt{2\pi}(B-\bar{q})}\right). \quad (11)$$

In this case, K can be found by numerically solving $K^2 + 2\ln(\Pr[q > B]\sqrt{2\pi}K) = 0$.

Another approximation found in the literature [13] is

$$\ln \Pr[q > B] \sim -\frac{(B-\bar{q})^2}{2\sigma_q^2} - \ln(\sqrt{2\pi}) \quad (12)$$

and $K = \sqrt{-2\ln \Pr[q > B] - \ln(\sqrt{2\pi})}$.

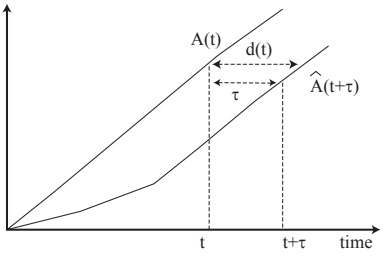

Fig. 2. Delay experienced by an arrival in the system

C. Departure Process

Since the queue becomes never idle, the departure rate at time t is equal to $\hat{a}(t)$. Let $\bar{\hat{a}}$ and $\sigma^2_{\hat{a}}$ denote respectively the mean and variance of $\hat{a}(t)$ in the steady state. Then, $\bar{\hat{a}} = \lim_{t\to\infty} E[\hat{a}(t)]$ and $\sigma^2_{\hat{a}} = \lim_{t\to\infty} E\left[(\hat{a}(t) - E[\hat{a}(t)])^2\right]$. We evolve $\hat{a}(t)$ in time recursively according to (1), starting with $\hat{a}(-1) = 0$. Then, we get at time t

$$\hat{a}(t) = \alpha^t(1-\alpha)a(0) + \alpha^{t-1}(1-\alpha)a(1) + \cdots + (1-\alpha)a(t), \tag{13}$$

which obviously implies that $\hat{a}(t)$ (and hence the departure rate at time t) is Gaussian. By comparing (13) with (3), we immediately get the following relationship

$$\hat{a}(t) = \frac{1-\alpha}{\alpha} q(t+1). \tag{14}$$

Therefore, we conclude from (14), (5), (6) that

$$\bar{\hat{a}} = \frac{1-\alpha}{\alpha}\bar{q} = \bar{a} \tag{15}$$

$$\begin{aligned}\sigma^2_{\hat{a}} &= (\frac{1-\alpha}{\alpha})^2\sigma^2_q \\ &= \frac{(1-\alpha)^2}{\pi(1-\alpha^2)}\int_0^\pi P_a(\Omega)\frac{1-\alpha^2}{1+\alpha^2-2\alpha\cos\Omega}d\Omega - \bar{a}^2.\end{aligned} \tag{16}$$

D. Delay Process

Let $d(t)$ be the delay experienced by an arrival at time t in the system. Then, as shown in Figure 2, the following relationship holds for a given constant τ.

$$\Pr[d(t) > \tau] = \Pr[A(t) - \hat{A}(t+\tau) > 0] \tag{17}$$

where $A(t) = \sum_{k=0}^{t-1} a(k)$ and $\hat{A}(t) = \sum_{k=0}^{t-1}\hat{a}(k)$. This implies that the probability that a packet arriving at time t leaves the system after time $t+\tau$ is equal to the probability that the random variable $A(t) - \hat{A}(t+\tau)$ is greater than 0. Since $a(t)$ and $\hat{a}(t)$ are Gaussian, $A(t) - \hat{A}(t+\tau)$ is also a Gaussian random variable. Some manipulation gives

$$\begin{aligned}A(t) - \hat{A}(t+\tau) &= q(t) - \sum_{k=t}^{t+\tau-1}\hat{a}(k) &(18)\\ &= q(t+\tau) - \sum_{k=t}^{t+\tau-1} a(k). &(19)\end{aligned}$$

Let $\bar{r}_\tau$ and $\sigma^2_{r_\tau}$ respectively denote the mean and variance of $r_\tau(t) \equiv A(t) - \hat{A}(t+\tau)$ in the steady state. Then, we can readily show that the following relationship holds.

$$\bar{r}_\tau = \bar{q} - \tau\bar{a} = (\frac{\alpha}{1-\alpha} - \tau)\bar{a} \tag{20}$$

and

$$\sigma^2_{r_\tau} = \frac{1}{\pi}\int_0^\pi P_a(\Omega)M(\Omega)d\Omega - \bar{r}^2_\tau \tag{21}$$

where

$$\begin{aligned}M(\Omega) = &\left(\frac{\alpha^2}{1-\alpha^2} - \alpha\frac{1-\alpha^\tau}{1-\alpha}\right)\frac{1-\alpha^2}{1+\alpha^2-2\alpha\cos\Omega} \\ &+ \left(\frac{\alpha}{1-\alpha} - \tau\right)\frac{\cos\Omega\tau - \cos\Omega(\tau-1)}{1-\cos\Omega} \\ &+ \frac{2(1-\cos\Omega)(1-\tau\cos\Omega(\tau-1) + (\tau-1)\cos\Omega\tau)}{3-4\cos\Omega+\cos 2\Omega} \\ &+ \left(\frac{\alpha^{\tau+1}}{1-\alpha}\right)\frac{1-\alpha^2-2\alpha^{-\tau}\cos\Omega\tau + 2\alpha^{-(\tau+1)}\cos\Omega(\tau-1)}{1-2\alpha^{-1}\cos\Omega+\alpha^{-2}}.\end{aligned}$$

By invoking Chenoff bound and using large deviations approximation [12] based on it, we have the following approximation for $\Pr[d > \tau]$.

$$\ln\Pr[d>\tau] = \ln\Pr[r_\tau > 0] \sim -\frac{\bar{r}^2_\tau}{2\sigma^2_{r_\tau}}. \tag{22}$$

Note that in our model d is a discrete-valued random variable. Using (17) and (22) we obtain the following approximation for the distribution of d.

$$\begin{aligned}\Pr[d=\tau] &= \Pr[d>\tau-1] - \Pr[d>\tau] &(23)\\ &= \Pr[r_{\tau-1}>0] - \Pr[r_\tau>0] &(24)\\ &\sim e^{-\frac{\bar{r}^2_{\tau-1}}{2\sigma^2_{r_{\tau-1}}}} - e^{-\frac{\bar{r}^2_\tau}{2\sigma^2_{r_\tau}}}. &(25)\end{aligned}$$

Using Little's theorem and (5),

$$\bar{d} = \frac{\bar{q}}{\bar{a}} = \frac{\alpha}{1-\alpha}. \tag{26}$$

Interestingly, α solely determines mean delay in the system irrespective of input traffic characteristics.

E. Admission Control

In order to ensure each QoS queue to be served by the amount $\hat{a}_i(t)$ for all times, it is necessary to have an admission rule such that $\Pr\left[\sum_{i=1}^N \hat{a}_i(t) > C\right] < \epsilon$ where ϵ is a sufficiently small positive number. The admission rule $\Pr\left[\sum_{i=1}^N \hat{a}_i(t) > C\right] < \epsilon$ can be loosened to $\Pr\left[\sum_{i=1}^N a_i(t) > C\right] < \epsilon$ at the cost of increased flow blocking probability. It is obvious that the latter condition implies the former condition because both $a_i(t)$ and

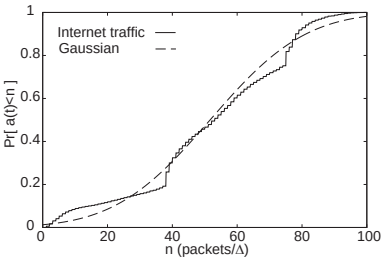

Fig. 3. Distribution of Internet traffic

$\hat{a}_i(t)$ are Gaussian with identical mean and the variance of $\hat{a}_i(t)$ is less than or equal to that of $a_i(t)$.

Let $\sum_{i=1}^{N} a_i = S$ and $\sum_{i=1}^{N} \hat{a}_i = \hat{S}$. By invoking Chernoff bound and using a refined large deviations approximation [12] based on it, the admission rules can be approximated by

$$\ln \Pr\left[S > C\right] \sim -\frac{(C-\bar{S})^2}{2\sigma_S^2} + \ln\left(\frac{\sigma_S}{\sqrt{2\pi}(C-\bar{S})}\right) < \ln \epsilon \tag{27}$$

and

$$\ln \Pr\left[\hat{S} > C\right] \sim -\frac{(C-\bar{\hat{S}})^2}{2\sigma_{\hat{S}}^2} + \ln\left(\frac{\sigma_{\hat{S}}}{\sqrt{2\pi}(C-\bar{\hat{S}})}\right) < \ln \epsilon. \tag{28}$$

Note that $\bar{S}$ and σ_S^2 are directly measurable from input traffic and $\bar{\hat{S}} = \bar{S}$ and if $\alpha_i = \alpha$ for all i,

$$\sigma_{\hat{S}}^2 = \sigma_S^2 - \frac{1}{\pi}\int_0^{\pi} P_S(\Omega)\frac{2\alpha(1-\cos\Omega)}{1+\alpha^2-2\alpha\cos\Omega}d\Omega \tag{29}$$

where $P_S(\Omega)$ is power spectrum of $\sum_{i=1}^{N} a_i$, which is also measurable from input traffic.

IV. Traffic Measurement

In the previous sections we found that the only statistics to be measured for the calculation of QoS measures of interest are input mean and input power spectrum. In this section we present an estimator that we used in simulation studies for the estimation of these statistics.

A. Input Mean

For M samples, we estimate input mean by

$$\bar{a} = \frac{1}{M}\sum_{n=0}^{M-1} a(n). \tag{30}$$

B. Input Power Spectrum

For the power spectrum estimation, we use Welch's method [14] instead of well-known periodogram method to acquire a more consistent estimate. For M samples,

$$P_a(\Omega) = \frac{1}{KLU}\sum_{i=0}^{K-1}\left|\sum_{n=0}^{L-1} w(n)a(n+iD)e^{-j\Omega n}\right|^2 \tag{31}$$

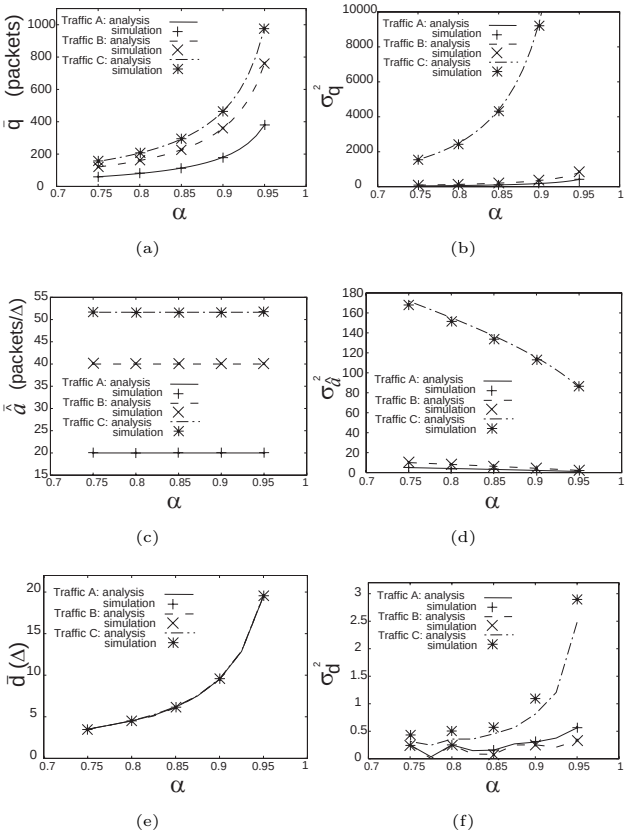

Fig. 4. Comparison of simulation results and analysis: (a) mean queue length, (b) queue length variance, (c) output mean, (d) output variance, (e) mean delay, (f) delay variance

where $M = L + D(K-1)$, $U = \frac{1}{M}\sum_{n=0}^{M-1}|w(n)|^2$ and we use Hanning window for $w(n)$.

V. Simulation Results

In order to verify the analytic results, we simulate a single queue system served by the linear server. We assume that the buffer size is infinite and the packet size is constant. Three types of input traffic are used. Traffic A is the superposition of 100 homogeneous, discrete-time ON-OFF sources whose source activity factor is 0.5 and the rate in ON state is 0.4 [packets/Δ]. Then, one can easily show that this traffic has a binomial distribution with mean and variance being 20 [packets/Δ] and 16 [packets^2/Δ^2], respectively. Recall that if the number of ON-OFF sources superposed is large, the binomial distribution tends to be a Gaussian distribution [11]. Traffic B is the aggregation of Traffic A and another superposition of 400 homogeneous ON-OFF sources whose activity factor is 0.4 and the rate in ON state is 0.125 [packets/Δ]. Then, the distribution of Traffic B is a convolution of two different binomial distributions, which again tends to be a

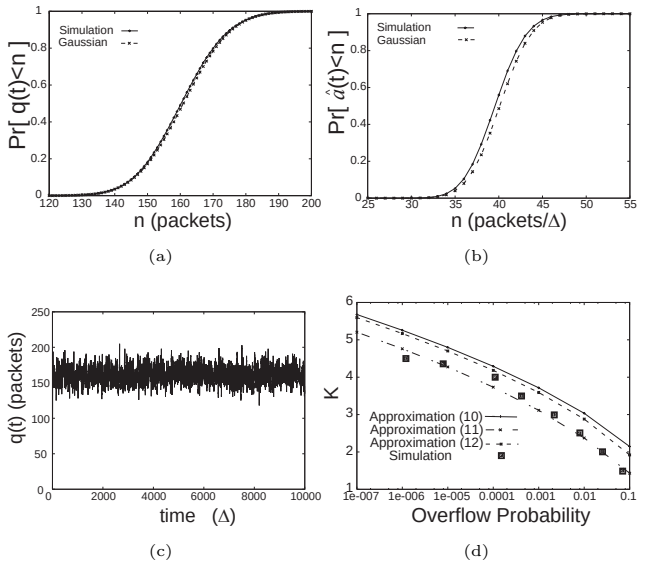

Fig. 5. Comparison of simulation results and analysis (Traffic B, α=0.80): (a) queue length distribution, (b) output distribution, (c) queue length trace, (d) buffer requirement

Gaussian distribution by the central limit theorem if the number of sources superposed is large. Traffic C is 10-second long real Internet traffic collected by NLANR [15]. For Traffic C, we let Δ=1 [ms] and packet size be 40 bytes which is its average packet size. Then, the mean and variance of the Internet traffic appears to be 51.6 [packets/Δ] and 533.0 [packets^2/Δ^2]. The distribution of the Internet traffic is compared with a Gaussian distribution with the same mean and variance in Figure 3.

In Figure 4, we compare simulation results with analysis with respect to the mean and variance of queue length, output traffic and packet delay in the system. In the analysis of Traffic A and B we use the true input mean and input power spectrum derived from the traffic model whereas in the analysis of Traffic C we use the measured input mean and input power spectrum obtained by the estimators (30) and (31). In overall, the simulation results agree with the analysis almost perfectly irrespective of the value of α, which confirms that all three types of traffic we used in the simulation tend to be Gaussian so that the queue length and output traffic tend to be Gaussian. Figures 4 a, b show that the mean and variance of the queue length increase as α increases, i.e., the passband of the server becomes narrower, as indicated by (5) and (6). Figures 4 c, d show that the output mean is equal to the input mean (20 [packets/Δ] for Traffic A, 40 [packets/Δ] for Traffic B and 51.6 [packets/Δ] for Traffic C) irrespective of α whereas the output variance decreases as α increases, i.e., the passband of the server becomes narrower, as indicated by (15) and (16). Finally, Figures 4 e, f show that the mean delay is independent of traffic types and is increasing with respect to α as indicated by (26), whereas the delay variance depends on traffic types.

The queue length distribution and output distribution obtained from the simulation are shown in Figures 5 a, b for Traffic B when α=0.80. As expected, both distributions are very close to a Gaussian distribution with the same mean and variance. The trace of the corresponding queue length is also shown in Figure 5 c, which confirms that the linear server never makes the queue idle. Finally, through simulations we find the buffer requirement $B = \bar{q} + K\sigma_q$ as a function of target overflow probability. Note that $\bar{q}$ and σ_q are fixed for given input traffic and α. As shown in Figure 5 d, the simulation results agree quite well with the analytical approximations given by (10), (11) and (12).

References

[1] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)", *IEEE/ACM Trans. on Networking*, vol. 2, no. 1, Feb. 1994, pp. 1-15.

[2] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", *IEEE/ACM Trans. on Networking*, vol. 5, no. 6, Dec. 1997, pp. 835-846.

[3] M. W. Garret and W. Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic", *Proc. ACM SIGCOMM '94*, Sept. 1994, pp. 269-280.

[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", IETF-RFC2475, December 1998.

[5] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks - the Single Node Case", *IEEE/ACM Trans. on Networking*, vol. 1, no. 3, June 1993, pp. 344-357.

[6] A. Dmers, S. Keshav and S. Shenker, "Analysis and Simulation of a Fair Queueing Algorithm", *Internetworking: Reserach and Experience*, vol. 1, no. 1, 1990, pp. 3-26.

[7] L. Zhang, "VirtualClock: A New Traffic Control Algorithm for Packet Switching Networks", *ACM Trans. on Comp. Sys.*, vol. 9, May 1991, pp. 101-124.

[8] S. Golestani, "A Self-Clocked Fair Queueing Scheme for Broadband Applications", *Proc. IEEE INFOCOM '94*, Apr. 1994, pp. 636-646.

[9] D. Stiliadis and A. Varma, "Rate-Proportional Servers: A Design Methodology for Fair Queueing Algorithm", *IEEE/ACM Trans. on Networking*, vol. 6, no. 2, April 1998, pp. 164-174.

[10] M. Shreedhar and G. Varghese, "Efficient Fair Queueing Using Deficit Round Robin", *Proc. ACM SIGCOMM '95*, Sept. 1995, pp. 231-242.

[11] A. Papoulis, **Probability, Randdom Variables, and Stochastic Processes**, 3rd Ed., McGraw-Hill, 1991.

[12] A. Elwalid, D. Mitra and R. H. Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node", *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, Aug. 1995, pp. 1115-1127.

[13] R. Guerin, H. Ahmadi and M. Nagshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks", *IEEE J. Select. Areas Commun.*, vol. 9, 1991, pp.968-981.

[14] M. Hayes, **Statistical Digital Signal Processing and Modeling**, 1st Ed., John Wiley & Sons, Inc., 1996.

[15] http://moat.nlanr.net/Traces/