AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

Unintrusive Communication of Status in a Packet Network in Heavy Traffic

By G. J. FOSCHINI*

(Manuscript received October 31, 1983)

We consider a packet switched network in the situation where communication resources are used close to capacity. Such heavy traffic may seem to present a dilemma. On one hand, at each node, the usefulness of status information about queues at other nodes is manifest. On the other hand, since the limitation of transmission resources causes backup, heavy load seems to be the worst situation in which to expend still more communication resources to convey status information. Under extremely general assumptions on interarrivals and services, a scaling appropriate for queueing processes in networks under heavy traffic has been established. Under these assumptions, we demonstrate that the status of the entire network can be communicated throughout the network, perfectly, in real time, without influencing the scaled queueing process. So, within a precise mathematical setting, we see that there is no dilemma: status can be conveyed at a negligible cost in a network operating at heavy load. Most of today's computer networks are designed for light-to-moderate loading. Yet heavy traffic analysis is growing in relevance, as is explained. A brief introduction to the subject of convergence of queueing systems is included.

I. INTRODUCTION AND RESULT

1.1 Status information and the contention for communication resources

The key concern in the mathematical theory of computer communication networks is the stochastic contention for limited network resources and the associated queueing and delay processes. (See Refs.

* AT&T Bell Laboratories.

Copyright © 1985 AT&T. Photo reproduction for noncommercial use is permitted without payment of royalty provided that each reproduction is done without alteration and that the Journal reference and copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free by computer-based and other information-service systems without further permission. Permission to reproduce or republish any other portion of this paper must be obtained from the Editor.

1 through 5 for basic expositions.) Queueing theory has provided the approach to the subject of stochastic resource contention. The waiting lines in a computer network are represented as a vector $Q(t) = (Q_1(t), Q_2(t), \cdots, Q_K(t))$, where $Q_k(t)$ represents the number of items queued for the kth resource. The dynamics of the network are associated with the law of evolution of $Q(t)$.

In this paper, we focus on the contention for communication resources. We assume that the network is a packet switched network where transmission is the bottleneck resource, and we consider the operation of the network in the condition of heavy traffic. By heavy traffic we mean that the demand for communication resources is approaching the network's capability for communication, so that queues are very large.

Status information refers to the information needed to describe queueing (or delay) as it evolves in time at each node. At moderate traffic levels, the issue of communication of status appears formidable. (Some initial investigations are reported in Ref. 6.) How much information about the various components of $Q(t)$ should be communicated to other nodes? Is there a point of diminishing returns beyond which the communication channels become choked with status information, thereby significantly worsening the queueing problem? Questions such as these are very difficult to approach with queueing theory methods. However, under the assumption of heavy traffic, we show that the status issue crystallizes and becomes mathematically tractable.

The circumstance of heavy traffic emphasizes what seems to be the dilemma associated with the communication of nodal status information. On the one hand, during times of heavy traffic, such status information is especially useful. If at each node in the network there is information available about all the queues at every other node, this information could be used to control the flow of packets within the network and to judiciously control access to the network. On the other hand, the limitation of transmission resources causes queueing in the first place; therefore, during heavy loading, it may seem inadvisable to expend still more communication resources to convey status information.

We show that there is no dilemma. In the context of a precise mathematical model, the status information of the entire network can be transmitted throughout the network without any substantial taxing of communication resources. Specifically, one can send enough information to describe the queueing situation as it develops in real time. Yet the utilization of communication resources for status transmission as compared to the utilization of communication resources for transmitting other packets tends to zero as the traffic increases.

Figure 1 illustrates the main result in its simplest form. Two

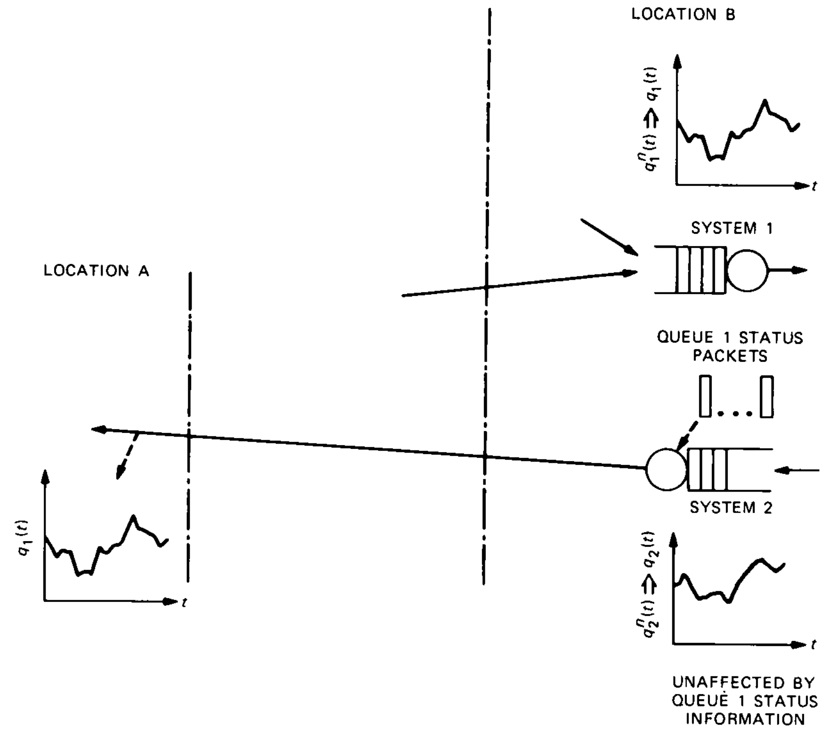

Fig. 1—Simple version of main result.

queueing systems collocated at B represent packets waiting for transmission resources. General independent sequences of positive random variables are assumed to govern interarrival and transmission times. Both systems are assumed to be in heavy traffic. In a sense that we will make exact, the following holds: Packets can be added to those on system 2 such that the recipient of the packets at location A can reconstruct $q_1(t)$ perfectly in real time. The process $q_2(t)$ remains the same after the status packets are added.

1.2 Heavy traffic scaling

Showing that status can be conveyed unintrusively involves analyzing the scaling appropriate for the description of queues (or delays) associated with the condition of heavy load. As will be explained in the following sections, it has been established that in heavy traffic a continuous path process $q(t)$ replaces the jump process $Q(t)$ as the natural descriptor of the number of items queued. Roughly, the normalization involved in obtaining $q(t)$ is a sort of mathematical "lens" used to place the essential features of the $Q(t)$ process in focus as the traffic is increased. If one did not normalize, the queueing process would tend to infinity in the heavy traffic limit. Furthermore, if one does not normalize in just the right way, the resulting process trivializes to be identically infinity or identically zero.

One reason why the communication of the queueing situation is easier in the heavy traffic limit is that it is not necessary to convey $Q(t)$ exactly. The limiting expression of status, $q(t)$, is insensitive to the fine structure of the approximating processes, so that it is not necessary to communicate each arrival and departure and the exact time of occurrence in order to communicate status. Yet the process $q(t)$ possesses sample paths of such elaborate structure that the result that negligible transmission resources are needed to communicate status in the limit may be surprising, or even counterintuitive.

1.3 Content of sequel

In the following we discuss the growing relevance of heavy traffic models. The paper is written so as to make the result accessible to readers unfamiliar with the specialized subject of convergence of queueing systems in heavy traffic and, at the same time, provide a concise introduction to this topic. This introductory material, presented in Sections II and III, gives the basis for a detailed probabilistic analysis in Section IV. This analysis develops the main result that the various effects associated with the heavy traffic limit can be resolved to allow a communication scheme providing unintrusive transmission of status. The literature providing the derivation of the foundation material is cited for readers who would like more detail.

For readers already familiar with the heavy traffic theory of packet networks, we mention at this point a key feature of our main result in Section IV. Namely, from the status information received at a node about a queue at any other network node, an approximation to the queueing process sample path can be constructed so that in the heavy traffic limit the distance between the queueing process sample path and its approximation converges in probability to zero. Moreover, the transmission of the approximation can be accomplished unintrusively in the sense that the limiting queueing process is unchanged in distribution by the additional packets transmitted to convey the approximations.

In Section V, with the status communication issue obviated, we broach the next level of inquiry dealing with use of status information to reduce queueing and delay.

II. THE HEAVY TRAFFIC MODEL AND ITS RELEVANCE

In the heavy traffic model for the network queueing process, the components of $q(t) = (q_1(t), \cdots, q_K(t))$ represent the time evolution of normalized queue sizes in the limit as the rate Λ at which work enters the network approaches U, the ultimate rate at which the network can do work. If $U - \Lambda = \Delta/\sqrt{n}$, where Δ is a positive constant, and if $Q(t)$ is the aforementioned K vector of queue sizes, then $q(t) = \lim_{n\to\infty} Q(nt)/\sqrt{n}$. The process $q(t)$ is an example of a diffusion process. A precise definition of the limit is discussed in Section III.

A mathematically equivalent way to obtain the same limit is to compose $Q(t)/\sqrt{n}$ and replace U and Λ by nU and $n\Lambda$, respectively. The first procedure, involving $Q(nt)/\sqrt{n}$, relates to a situation where, in a given network, operation close to capacity is taking place. The $Q(t)/\sqrt{n}$ situation relates to a sequence of networks, each accommodating a new demand, $n\Lambda$. (There is nothing special about n. Any function of n that increases without bound would yield the same limit.) Such a sequence is of interest if the demand for data communications grows substantially over a period of many years. The transmission capabilities will grow to keep pace with the demand. If processing power and memory are inexpensive relative to transmission cost, transmission will be the bottleneck. The behavior of the limit of $Q(t)/\sqrt{n}$ will be descriptive of long-term tendency of the network queueing processes.

Network design issues that have no crisp resolution when addressed in the context of today's networks under nominal operating conditions can have a clear answer in the context of the heavy traffic limit. The issue of the value of transmitting status information is an example, as we shall see. The result we will prove is a very general one in that the network can be quite arbitrary, as for example Reiman and Harrison's generalization of the Jackson-type network,[7] where each exogenous interarrival process, as well as each service process, is modeled as an independent, identically distributed (i.i.d.) sequence with a general nonnegative distribution.

2.1 Motivation for diffusion models in computer network theory

Mathematically, diffusions are finite dimensional vectors whose components are continuous random time functions with the property that at each point in time the statistical law concerning future evolution depends only on the present value of the function. (Diffusion evolution can also depend on time, but we shall not use this degree of flexibility. References 8 and 9 are basic references on diffusions.) Diffusion representation of behavior is a natural representation to consider in instances where the randomness stems from a large number of cumulative influences. Diffusion models are used to replace alternative models, which are useful in addressing small-scale phenomena but are too cumbersome for analyzing large-scale behavior. In computer networks the demand for network resources establishes the scale of the queueing processes. The queues can be large enough to require a new scale that no longer tracks each quantum jump (i.e., arrival or departure).

2.2 Weiner process with drift

In this subsection we mention for future reference some diffusions of special interest in the theory of computer networks, namely the reflecting Weiner processes with drift. Although one can point out significant usefulness of other diffusions,[10] the following are the diffusions that are relevant in much of computer network theory.

Let Δ and σ be constants where $\sigma > 0$. By the Wiener process $W(\Delta, \sigma^2)$ with drift Δ (and dispersion σ^2) we mean the continuous path Gaussian process starting at $w(0)$ (a constant) with $E[w(t_1) - w(t_2)] = \Delta(t_1 - t_2)$ and $E[(w(t_1) - Ew(t_1))(w(t_2) - Ew(t_2))] = \sigma^2 \min(t_1, t_2)$. This Wiener process, as it stands, is often unsuited to addressing heavy traffic queueing problems. After all, a queue cannot be negative. The Wiener process can be modified to keep the sample paths positive. This brings us to the subject of reflection. For $\Delta = 0$, the reflecting Wiener process $\underline{W}(0, \sigma^2)$ (reflected about the zero state) is defined to be $\underline{w}(t) = |w(t)|$. For $\Delta \neq 0$, $|w(t)|$ is not appropriate for describing reflection since $|w(t)|$ has drifts of opposite sign on the sets $\{t \mid w(t) > 0\}$ and $\{t \mid w(t) < 0\}$. For the reflecting Wiener process, we want a continuous path process whose incremental behavior matches that of $w(t)$ but with the constraint of nonnegativity. The variant $\underline{w}(t) = w(t) - \min[0, \min_{(0,t)} w(t)]$ is the natural definition for the reflecting Wiener process. As long as $\underline{w}(t) > 0$, then $\underline{w}(t)$ has the same differential properties as $w(t)$. The subtractive term serves to keep the process nonnegative. The notation $\underline{W}(\Delta, \sigma^2)$ is used to denote the reflecting Wiener process.

Vector versions of W and $\underline{W}$ have been defined where Δ is a vector and σ^2 is a positive definite matrix. In the vector case, to complete the description, angles of reflection must be specified[7] at the boundaries that prevent the process components from being negative.

2.3 Importance of diffusion models

In the context of today's computer networks, the condition of heavy traffic represents an extreme condition. We discuss reasons why consideration of this extreme is worthwhile.

First, the maturation of demand for computer network services is not in sight. The accelerating demand coupled with the rapid technological advances associated with network components is likely to foster a rapid evolution of networks accommodating more and more demand. As demand for network resources grows, economies of scale will encourage the operation of networks at higher utilizations. This is most easily seen in the context of a single M/M/1 queue.[11] Recall the mean delay formula $\bar{D} = (\mu - \lambda)^{-1}$. Fix the mean delay $\bar{D}$, and it is apparent that as the demand, λ, increases, μ increases by a smaller amount, so that $\rho = (\lambda/\mu) \to 1$. (See Ref. 12 for a related discussion.)

From the delay formula, we see that if λ increases with slope bounded away from zero, the increase in μ relative to λ that is needed to make $\bar{D}$ negligible goes to zero. Note, however, from the formula for mean queue size $\bar{Q} = (1 - \rho)^{-1}$, that the memory resources needed to queue packets do not enjoy this economy of scale. In packet communication systems where $\bar{D}$ is small, one should consider whether packet voice services should be accommodated. If the answer is yes, λ grows at a still faster rate. In such circumstances, one would anticipate a sequence of queueing processes behaving more like the Wiener process idealization.

Another reason diffusion models are relevant is that for a specific network operating at moderate utilization it is important to understand behavior in crisis situations. Emergencies can arise, for example, if a node crashes and the surviving network attempts to accommodate the resulting overload. Another cause can be a community of customers that suddenly present the network with an unanticipated level of demand. Diffusion methods for tracking the degradation of service and providing the capability of inquiring into which method of operation yields the most graceful degradation are especially useful.

Certain constructions give rise to processes that tend to be characterized by a $\underline{W}$ process. While today the quiescent operation of packet networks is not the condition of heavy load, we have cited above influences that motivate a long-term importance for $\underline{W}$ processes in the theory of computer networks. Sometimes, however, because it is extraordinarily difficult to analyze certain computer networks, a more precise model is replaced by an associated diffusion simply to gain tractability. Such heuristics are usually accompanied by measurements or simulations,[10,13] or else the diffusion can be a very useful "caricature"[14] of the real situation. Looking at the diffusion counterpart of a situation arising in practice in a context of light to moderate load, one can get an answer that, when interpreted in the original loading, gives the correct result.[15] Sometimes diffusion provides special insight[16] so that a new result in the realm of light to moderate traffic is first discovered by obtaining it in the heavy traffic range.

References 17 and 18 exhibited situations where state sensitive network behavior can be accommodated by diffusion models. It is well known that G/G models or state dynamic models are not amenable to exact analysis using traditional queueing theory approaches.

III. COMMENTS ON CONVERGENCE TO DIFFUSIONS

In this section we very briefly describe the convergence of $Q(nt)/\sqrt{n}$ to $q(t)$. We aim our presentation to highlight those aspects of the convergence that provide the backdrop for our main result. The material in this section is tutorial. The abstract convergence theory is adapted from Refs. 19 through 23, while the queueing network specific results cited can be found in Refs. 7, and 24 through 26.

We introduce some notation needed in the sequel. Superscripts, unless they are simple fractions, indicate the nth term in a sequence rather than exponentiation. By max $(q(t))$ we mean that the maximum is over the time interval $[0, T]$. We use $\approx$ to denote asymptotic equivalence for large values of a parameter ($f(n) \approx g(n)$ if $g(n)/f(n) \to 1$ as $n \to \infty$). For example, $n^{1/2} + n^{1/3} \log_2 n \approx n^{1/2}$.

3.1 Convergence of random processes

The set D^L of all vector-valued functions on a time interval $[0, T]$, where the components have left-hand limits and are right-continuous, is a very general set of functions for applications. Included are the diffusion sample paths, which are continuous, and the network queueing processes, which have piecewise constant paths. In the set D^L a definition of distance between functions has been given. The metric is too involved to be discussed here, but for our purposes we will only encounter pairs x_1 and x_2 whose distance can be measured by $\rho(x_1, x_2) = \max\| x_1 - x_2 \|$, where $\| \ \|$ is the usual definition of length of an L-dimensioned vector.

By $\mathscr{D}^L$ we mean the set of all vector-valued random processes whose sample paths are in D^L. A notion of distance between processes has been defined so that meaning is given to the convergence ($\Rightarrow$) of a sequence of random processes to a limit process. The metric space $\mathscr{D}^L$ is a highly specialized topic, and we shall not define or discuss it in detail. However, we mention facts about $\mathscr{D}^L$ that are relevant to our result.

$\mathscr{D}^L$ accommodates both network queueing processes and diffusions and provides the setting for demonstrating convergence of queueing processes to diffusions. Since D^L is a metric space, it makes sense to consider continuous real functions defined on D^L. If f is a bounded continuous real function on D^L, then f evaluated on the paths of an element of $\mathscr{D}^L$ is a random variable. Indeed, f maps random functions into random numbers. If $\{x^n\}_1^\infty$ is a sequence of stochastic processes (points in $\mathscr{D}^L$), then $\{f(x^n)\}_1^\infty$ is a sequence of random variables. x^n is said to converge to x if for all bounded-continuous f the distribution functions of $f(x_n)$ converge in the usual sense to the distribution function for x. This definition of convergence of general vector processes is considered to be *the* definition because of its significant practical value. This practical value partially stems from the fact that all the finite dimensional distributions of $q^n(t) = Q(nt)/\sqrt{n}$ converge to those for $q(t)$. Even more important is that the continuous functions on the processes include those of interest in practice, and we have already mentioned that, by definition, the distributions of continuous functions converge. For example, a chief concern in the operation of a network is that queues not exceed certain levels (corresponding to overflow) or that maximum delay not be excessive. Thus, there is interest in $\max_t Q^n(t)$, and the maximum over a time interval can be shown to be a continuous function. Therefore, in situations where $q^n(t)$ is intractable and $q(t)$ is tractable, it is meaningful to use $q(t)$ to approximate the behavior of $q^n(t)$ for large n. Regarding tractability, we note that a great deal is known about Wiener processes, so that properties of general limits of queueing processes can often be easily obtained from previously derived results.

3.2 The G/G/1 queue

We shall need this example in Section IV. For arrivals the mean rate is λ per second, while a denotes the interarrival variance. For the successive service times μ^{-1} is the mean and s is the variance.

Let $A_\#(\tau_1, \tau_2)$ denote the number of arrivals in (τ_1, τ_2), and let $D_\#(\tau_1, \tau_2)$ denote the number of departures. Define

$$a^n(t) = \frac{A_\#(0,nt) - \lambda nt}{\sqrt{n}}.$$

Using the central limit theorem and an asymptotic (large n) analysis, it is not difficult to show that $a^n(t)$ is distributed as $N(0, \lambda^3 at)$. The connection to the central limit theorem stems from the fact that $\{A_\#(0,nt) < k\} = \{A_1 + A_2 + \cdots A_k > nt\}$ (where A_i are the interarrival times).

We can conclude the same sort of result for departures, but that is more difficult. Consider first an imaginary system where the departure process runs forever with no idle times. Then, for each t,

$$d^n(t) = \frac{D_\#(0,nt) - \mu nt}{\sqrt{n}} \text{ is distributed as } N(0, \mu^3 st).$$

For this imaginary system $a^n(t) - d^n(t)$ is distributed as $N(0,(\mu^3 a + \mu^3 s)t)$. These elementary results regarding $a^n(t)$ and $d^n(t)$ are suggestive of much more significant results regarding convergence in $\mathscr{D}^1$.

For the $\mathscr{D}^1$ convergence, we can let μ and λ be functions of n (and write μ^n and λ^n) so long as μ^n and λ^n converge to constants and to each other so that

$$\lim_{n\to\infty} \sqrt{n}\,(\lambda^n - \mu^n) = \Delta$$

is a constant. This flexibility in μ^n will be useful in Section 4.2. For each n, we have an arrival sequence $\{A_{jn}\}_1^\infty$ and a departure sequence $\{B_{nj}\}_1^\infty$ of i.i.d. random variables. For the imaginary system

$$\hat{q}^n(t) = \frac{\hat{Q}(nt)}{\sqrt{n}} = \frac{A_{\#}(0,nt) - D_{\#}(0,nt)}{\sqrt{n}}$$

$$= \frac{(A_{\#}(0,nt) - \lambda^n nt) + (D_{\#}(0,nt) - \mu^n nt)}{\sqrt{n}} + \frac{(\mu^n - \lambda^n)}{\sqrt{n}} nt,$$

we have that $\hat{q}^n(t)$ converges to $N(\Delta t, \lambda^3 a + \mu^3 s)$. The convergence is for each t. But much more is true, namely, it has been shown that $\hat{q}^n(t)$ converges to $W(\Delta t, \lambda^3 a + \mu^3 s)$ in $\mathscr{D}^1$. Now for $q^n(t)$ one needs to account for departures not occurring when the queue is empty. This is a substantial complication that has been dealt with in Refs. 7, and 24 through 26. The upshot is that $q^n(t) \Rightarrow \underline{W}(\Delta, \lambda^3 a + \mu^3 s)$.

For $\underline{W}$ (or W) one can write and solve a partial differential equation (called the Fokker-Planck equation) for the probability transition density associated with the system being in state q' at time t given that it was in state q_0' at an earlier time t_0.[27] So long as $\Delta < 0$ the process $\underline{W}$ has a statistical equilibrium for which the state density is characterized by taking the limit of the probability transition density as $t \rightarrow \infty$.

In higher dimensions (networks) there is a vector version of the construction[7] in which the boundaries present even more difficulties than in the one-dimensional case.

3.3 Two results for future reference

We next cite two additional results* from the theory that we will use to derive our result on communication of status.

Lemma 1: If $\{x^n\}_1^\infty$ and $\{y^n\}_1^\infty$ are each positive sequences of random processes in $\mathscr{D}^L$ with $x^n \Rightarrow x$, then if, for each $\epsilon > 0$,

$$\lim_{n\to\infty} Pr\{\max_t \| x^n(t) - y^n(t) \| \geq \epsilon\} = 0,$$

we conclude that $y^n \Rightarrow x$.

Lemma 2: If $\{x^n\}_{n=1}^\infty$ is a sequence of random processes in $\mathscr{D}^L$ with continuous paths, convergent to a process with continuous paths, then for each positive ϵ and η there exist a δ, $0 < \delta < T$, and an integer n_0 such that

$$P\{\max_{|s-t|<\delta} \| x^n(s) - x^n(t) \| \geq \epsilon\} \leq \eta,\ n \geq n_0.$$

---

* Those familiar with the subject of convergence of probability measures will notice that Lemma 1 is a weakened form of the Converging Together Theorem (4.1 in Ref. 20), where the Skorokhod metric ρ has been replaced by a simpler metric that dominates ρ. Lemma 2 is a tightness consequence (8.2 in Ref. 20).

IV. DERIVATION OF THE MAIN RESULT

4.1 Approach

Refer to Fig. 1. Were it not for the complication of the status packets on system 2, we would simply have a pair of independent G/G/1 systems. The queueing process (q_1^n, q_2^n) would converge in $\mathscr{D}^2$ to a two-dimensional $\underline{W}$ process with each independent component a $\underline{W}$ process in $\mathscr{D}^1$, as already described in Section 3.2. We want to show that in the limit the status packets adequate to describe $q_1(t)$ in real time can be communicated on system 2 without perturbing the $\mathscr{D}^1$ limit, $q_2(t)$, from that for the case with no status packets.

So far we considered T to be fixed, and consequently, we have written $\mathscr{D}^L$ rather than use the more complete notation $\mathscr{D}^L$ [0, T]. The variable T is needed to express mathematically the meaning of communication of status in real time. We mean that for each time T, from the status packets communicated up to time T, we can construct a sequence, $r^n(t)$, so that for each $\epsilon > 0$

$$\lim_{n \to \infty} P\{\rho(r^n(t), q_1^n(t)) > \epsilon\} = 0$$

(i.e., the distance between the process $q_1^n(t)$ and its approximation converges in probability to zero). From Lemma 1 we see that for each T the convergence of $r^n(t)$ to $q_1(t)$ in $\mathscr{D}^1$ [0, T] is implied.

In Section 4.2 we derive a channel for status information that is asymptotically, for large n, of negligible rate in comparison to the total information rate of the communication resource. In the remaining two subsections, it is shown that the changes in $q_1^n(t)$ are such that there is enough rate to convey a sampled, clipped version of $q_1^n(t)$ that satisfies the above requirement for $r^n(t)$. The purpose of the clipping is to limit the number of bits per sample so as to meet the capacity limitation on the status channel.

4.2 Deriving a channel of negligible rate for status

The normalization parameter n can have different interpretations, as discussed in Section II. For definiteness in the presentation of the proof, we choose the perspective that the demand for service, Λ^n, is increasing and the capacity of the resource, M^n, is likewise increasing:

$$\Lambda^n = n\left(\mu - \frac{\Delta}{\sqrt{n}}\right)$$

and

$$M^n = n\mu,$$

where μ and Δ are constants.

From this viewpoint we see countervailing aspects of the nature of the limit. On the one hand, with M^n increasing indefinitely, we see the opportunity of deriving a status channel of some significance that is vanishingly small relative to M^n. On the other hand, despite the fact that the limit $q_1(t)$ is insensitive to considerable changes in $Q_1^n(t)$, the limit has an extremely intricate structure that must be conveyed on the status channel. In fact, the $w(t)$ constituent of $q_1(t)$ is the epitome of a chaotic continuous random function—it is the indefinite integral of white Gaussian noise.[28]

Proposition: There exists a channel of negligible rate for status.

Proof: The units of Λ^n and M^n are packets per second. Say that there are on the average B bits per packet so that the communication resource has a capacity of M^nB bits per second. Since M^n/n and Λ^n/n correspond to μ^n and λ^n, respectively, as indicated in Section 3.2, we can alter M^n and not perturb the $q_1(t)$ diffusion if we preserve the asymptotic behavior:

$$M^n/n \to \mu \quad \text{and} \quad \frac{M^n - \Lambda^n}{\sqrt{n}} \to \Delta$$

for large n.

This flexibility in M^n enables us to derive a status channel. Specifically, we choose a channel for status that has rate

$$n^{1/3}\left(\frac{1}{2} + \theta\right)\log_2 n$$

bits per second, where θ is any positive number. We shall see the adequacy of this rate. If the transmission resources for channel number 2 provide the status channel, then

$$M^n = n\mu - \{n^{1/3}[(1/2) + \theta]\log_2 n\}/B$$

packets per second, and still $M^n \approx n\mu$ and

$$M^n - \Lambda^n = n\mu - n^{1/3}\frac{\left(\frac{1}{2} + \theta\right)\log_2 n}{B} - n\left(\mu - \frac{\Delta}{\sqrt{n}}\right)$$

$$= \sqrt{n}\,\Delta - \frac{n^{1/3}\left(\frac{1}{2} + \theta\right)\log_2 n}{B}.$$

From Section 3.2 we have that $q_2^n(t)$ converges to the same limit whether or not the server provides the derived channel.

We use the derived channel as follows to transmit information for the required approximation to $Q_1^n(t)$: first we clip the process $Q_1^n(t)$ at the upper threshold $n^{\theta+(1/2)}$, and then we sample the result every $n^{-1/3}$ seconds. So, the set $\{0, 1, \cdots [n^{\theta+(1/2)}]\}$ is the range of the samples (where $[n^{\theta+(1/2)}]$ means the largest integer less than $n^{\theta+(1/2)}$).

We now show that in the limit of large n the clipped and sampled replica of $Q_1^n(t)$ has the desired convergence property. Partition $[0, T]$ letting $I(i, n)$ denote the interval $[(i-1)/n^{1/3}, i/n^{1/3}]$, $i = 1, 2, \cdots$. In what follows, let $v^n(t)$ denote the piecewise constant process defined by $q_1^n(i/\sqrt{n})$ on $I(i, n)$. Let $r^n(t)$ be defined in precisely the same way as $v^n(t)$ except that $r^n(t)$ is truncated at n^θ so that $r^n(t) = \min(n^\theta, v^n(t))$. (Of course, if $q_1^n(t)$ has an upward or downward drift, it would seem advisable to define $v^n(t)$ and $r^n(t)$ to have a linear slope between samples matching the trend in $q_1^n(t)$. However, the piecewise constant $r^n(t)$ and $v^n(t)$ are adequate for us and connect easily to existing results we want to use.) The successive values of $r^n(t)$ are transmitted over the derived channel. The function $v^n(t)$ is introduced to help establish the convergence of $r^n(t)$ to $q_1(t)$.

Using the above proposition we can now establish convergence. We must demonstrate the following.

Theorem: For each $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\{\max_t |q_1^n(t) - r^n(t)| > \epsilon\} = 0. \tag{1}$$

Proof: It is convenient to note two ways in which $r^n(t)$ can fail to track $q_1^n(t)$ to within ϵ:

A. $q_1^n(t)$ could take on large values sufficiently beyond the peak value n^θ of the tracking process.

B. $q_1^n(t)$ could, on some $I(i, n)$, depart too much from the sample value approximation.

We shall see that both ways occur with sufficiently small probability.

It is evident that

$$\{\max_t |q_1^n(t) - r^n(t)| > \epsilon\}$$

is contained in

$$\{\max_t q_1^n(t) > n^\theta\} \cup \{\max_t |q_1^n(t) - v^n(t)| > \epsilon\} = A^n \cup B^n,$$

where A^n and B^n are defined in the obvious way. We next show $\lim_{n\to\infty} P\{A^n\} = \lim_{n\to\infty} P\{B^n\} = 0$. From these limits (1) follows.

To see that $P\{A^n\} \to 0$, use the fact that $q_1^n(t) \Rightarrow w(t)$ so that $\max_t q_1^n(t) \to \max_t \underline{w}(t)$. In terms of distributions we have that for each number y

$$\lim_{n\to\infty} P\{\max_t q_1^n(t) > y\} = P\{\max_t \underline{w}(t) > y\}.$$

Now

$$\lim_{n\to\infty} P\{\max_t q_1^n > n^\theta\} < \lim_{n\to\infty} P\{\max_t q_1^n(t) > C\}$$

$$= P\{\max_t \underline{w}(t) > C\}, \quad (2)$$

where C is any number. So the right-hand side of (2) can be made arbitrarily small. Thus we have obtained the desired result for A^n.

Next we show that $\lim_{n\to\infty} P(B^n) = 0$; that is, for each $\epsilon > 0$

$$\lim_{n\to\infty} P\{\max_t |q_1^n(t) - v^n(t)| \geq \epsilon\} = 0. \quad (3)$$

First let $\hat{q}_1^n(t)$ denote the continuous variant of $q_1^n(t)$ formed by connecting the consecutive points of discontinuity in the graph of $q_1^n(t)$ by line segments. By construction $\max_t |\hat{q}_1^n(t) - q_1^n(t)| \leq (1/\sqrt{n})$. So by Lemma 1, $\hat{q}_1^n(t) \Rightarrow q_1(t)$. Employing Lemma 2, we have that for each $\eta > 0$ there exists a $\delta > 0$, so that for

$$P\{\max_{|s-t|<\delta} |\hat{q}_1^n(t) - \hat{q}_1^n(s)| \geq \epsilon\} < \eta$$

for sufficiently large n. Therefore,

$$P\{\max_{|s-t|\leq n^{-1/3}} |\hat{q}_1^n(t) - \hat{q}_1^n(s)| \geq \epsilon\} < \eta$$

for sufficiently large n, or what is the same, the limit of this sequence of probabilities is zero. Rewriting the maximum in terms of $q_1(t)$, we have

$$P\{e^n + \max_{|s-t|\leq n^{-1/3}} |q_1^n(t) - q_1^n(s)| \geq \epsilon\} \to 0,$$

where the magnitude of the error term, e^n, cannot exceed $2/\sqrt{n}$. Since

$$\lim_{n\to\infty} P\{\max_{|s-t|\leq n^{-1/3}} |q_1^n(t) - q_1^n(s)| \geq \epsilon - e_n\} = 0,$$

by set containment it follows that for each fixed $\epsilon' > \epsilon$

$$P\{\max_t |q_1^n(t) - v^n(t)| \geq \epsilon'\} \to 0.$$

The containment stems from taking the maximum over a smaller set and from the fact that the threshold ϵ' eventually exceeds $\epsilon - e^n$. Now (3) follows because ϵ is arbitrary.

V. DISCUSSION

5.1 Immediate extension

We have proven a simple form of the result that in heavy traffic, status can be conveyed unintrusively. The result extends immediately to much more general network settings. It is not difficult to go beyond the simple form of the result and conclude that status can be conveyed unintrusively throughout an entire network such as Reiman and Harrison's generalization of a Jackson network. One simply derives as many parallel status channels as is necessary on each communication link.

5.2 Similarities

While heavy traffic analysis of computer communication networks is more intricate than established analytical techniques for electrical networks, there are some striking parallels. For example, as with electrical networks, one writes a differential equation[7] (the aforementioned intricacy stems from the fact that in computer networks it is a partial-differential equation: the Fokker-Planck equation). There is interest in both transient and steady-state analysis. As with voltage and current there is an extremely simple steady-state relationship between the two dependent variables of most interest, queue size, and delay.[7,29] The main result of this paper can be considered to add to this list of similarities by deriving what amounts to a sampling theorem.

5.3 Future work

The availability of status information that was established here is only one aspect of the broader subject of network control. This availability prompts the question of how should status information be used to optimally control $q(t)$ under certain performance criteria?

Examples already appear in the literature[17,18] that demonstrate significant improvements using status information to guide the evolution of a queueing network in heavy traffic. One of the established examples involves that of the case of a Poisson arrival process where each arrival has the option of joining one of K queues. The K queueing systems have i.i.d. exponential service times. Two systems are contrasted. In the first, status information is used and the arrival joins the queues offering the least expected delay at the time of arrival. In the second system the arrival is blind to status information and randomly selects a queue. In heavy traffic the performance of the first system is superior to that of the second by a factor of K in mean queue size and delay and, more importantly, a factor of K in the tail exponent of the equilibrium distribution for queue size and delay.

The published examples seem part of a more general theory that would allow more relaxed assumptions on arrivals and services and would address the question of how status information should be transferred and used. The result of this paper needs to be shown in the context of $q(t)$ being controlled with the status information. We remark that the proof we have given utilizes very little of the structure of the approximating queueing processes and appears to allow great flexibility of form for $Q^n(t)$.

For a Reiman-Harrison-type of network, but with state dynamic routing, the problem of finding the optimum control to minimize the maximum (over source-sink pairs) average delay is a very challenging one.

Another layer of difficulty is introduced if one includes fixed delays in status information to account for processing or propagation.

VI. ACKNOWLEDGMENT

The author gratefully acknowledges helpful discussions with Marty Reiman, Jack Salz, Ward Whitt, and Clark Woodworth. Marty Reiman suggested that using Lemma 2 might simplify an earlier (more cumbersome and restrictive) derivation of the main result, and indeed he was correct.

REFERENCES

1. L. Kleinrock, *Queueing Systems, Computer Applications,* Vol. 2, New York: Wiley, 1976.
2. H. Kobayashi, *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology,* Reading, MA: Addison-Wesley, 1978.
3. M. Schwartz, *Computer-Communication Network Design and Analysis,* Englewood Cliffs, NJ: Prentice-Hall, 1977.
4. K. M. Chandy and C. H. Sauer, "Approximate Methods for Analyzing Queueing Network Models of Computing Systems," Comput. Surv., *10,* No. 3 (September 1978), pp. 281–317.
5. H. Kobayashi and A. G. Konheim, "Queueing Models for Computer Communication System Analysis," IEEE Trans. Commun., *COM-25,* No. 1 (January 1975), pp. 2–29.
6. J. M. Holtzman, "Routing Updates in Packet Switched Networks," in *Computer Performance,* K. M. Chandy and M. Reiser, Eds., New York: North-Holland, pp. 537–46, 1977.
7. M. Reiman, "Open Queueing Networks in Heavy Traffic," Math. Oper. Res., *9* (August 1984), pp. 441–58.
8. S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes,* New York: Academic Press, 1975.
9. S. Karlin and H. M. Taylor, *A Second Course in Stochastic Processes,* New York: Academic Press, 1981.
10. J. P. Lehoczky and D. P. Gaver, "Gaussian Approximation to Service Problems: A Communications System Example," J. Appl. Probab., *13,* No. 4 (December 1976), pp. 768–80.
11. L. Kleinrock, Computer Networks Seminar, Washington, D. C., August 1976.
12. S. Halfin and W. Whitt, "Heavy-Traffic Limits for Queues With Many Exponential Servers," Oper. Res., *29,* No. 3 (May–June 1981), pp. 567–88.
13. E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems,* New York: Academic Press, 1980.
14. F. B. Knight, *"Essentials of Brownian Motion and Diffusion,"* Math. Surv., No. 18, American Mathematical Society, RI, 1981.
15. D. P. Heyman, "An Approximation for the Busy Period of the M/G/1 Queue Using a Diffusion Model," J. Appl. Probab., *11,* No. 1 (March 1974), pp. 159–69.
16. G. J. Foschini and B. Gopinath, "Sharing Memory Optimally," IEEE Trans. Commun. Technol., *COM-31,* No. 3 (March 1983), pp. 352–60.
17. G. J. Foschini and J. Salz, "A Basic Dynamic Routing Problem and Diffusion," IEEE Trans. Commun. Technol., *COM-26* (March 1978), pp. 320–7.
18. G. J. Foschini, "On Heavy Traffic Diffusion Analysis and Dynamic Routing in Packet Switched Networks," in *Computer Performance,* K. M. Chandy and M. Reiser, Eds., New York: North-Holland, 1977, pp. 499–514.

19. Yu. V. Prokhorov, "Convergence of Random Processes and Limit Theorems in Probability Theory," Theory Probab. Appl., *1,* No. 2 (1956), pp. 157–214.
20. P. Billingsly, *Convergence of Probability Measures,* New York: Wiley, 1968.
21. P. Billingsly, "Weak Convergence of Measures Applications in Probability," SIAM, Philadelphia, 1971.
22. K. R. Parthasarathy, *Probability Measures in Metric Spaces,* New York: Academic Press, 1967.
23. S. R. S. Varadhan, *Stochastic Processes,* The Courant Institute, New York University, 1968.
24. M. Reiman and M. Harrison, "Reflected Brownian Motion in an Orthant," Ann. Probab., *9,* No. 2 (1981), pp. 302–8.
25. A. A. Borovkov, "Some Limit Theorems in the Theory of Mass Service, Theory Probab. Appl., *10,* No. 4 (1965), pp. 550–65.
26. D. L. Iglehart and W. Whitt, "Multiple Channel Queues in Heavy Traffic, II," Advanc. Appl. Probab., *2,* No. 1, pp. 355–69.
27. D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes,* London: Chapman and Hall, 1972.
28. E. Wong, *Stochastic Processes in Information and Dynamical Systems,* New York: McGraw-Hill, 1971.
29. G. J. Foschini, "Equilibria for Diffusion Models for Pairs of Communicating Computers-Symmetric Case," IEEE Trans. Inform. Theory, *IT-28,* No. 2 (March 1982), pp. 273–84.

AUTHOR

Gerard J. Foschini, B.S.E.E., 1961, Newark College of Engineering, Newark, NJ; M.E.E., 1963, New York University, New York; Ph.D. (Mathematics), 1967, Stevens Institute of Technology, Hoboken, NJ; AT&T Bell Laboratories, 1961—. Mr. Foschini initially worked on real-time program design. For many years he worked in the area of communication theory. In the spring of 1979 he taught at Princeton University. Mr. Foschini has supervised planning the architecture of data communications networks. Currently, he is involved with digital radio and optical communication research. Member, Sigma Xi, Mathematical Association of America, IEEE, New York Academy of Sciences.