Ad Hoc Networks 000 (2018) 1–16

Contents lists available at ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc

Internet congestion control using the power metric: *Keep the pipe just full, but no fuller*

Leonard Kleinrock

UCLA Computer Science Department, United States

ARTICLE INFO

Article history:
Received 8 March 2018
Accepted 21 May 2018
Available online xxx

Keywords:
TCP Congestion control
Bandwidth-delay product
Internet
Optimal power operating point
Universal power profile
Queueing

ABSTRACT

Recently there has been considerable interest in a key paper [1] describing a new approach to congestion control in Internet traffic which has resulted in significant network performance improvement. The approach is based on a 1978 paper [2] and a companion 1979 paper [3] which identified a system operating point that was optimal in that it maximized delivered throughput while minimizing delay and loss. This operating point is simply characterized by the insight that one should *"Keep the pipe just full, but no fuller"* and we show this is equivalent to loading the system so that in many cases (including those relevant to TCP connections) the optimized average number in the pipe is exactly equal to the *Bandwidth-Delay Product*. It is important to understand the reasoning and intuition behind this early insight and why it provides such improved behavior of systems and networks. In this paper, we first develop this insight using purely deterministic reasoning. We then extend this reasoning by examining far more complex stochastic queueing systems and networks using a function called *Power* to mathematically and graphically extract exact and surprising results that support the insight and allow us to identify the optimum operating point for a broad class of systems. These observations allow us to study the impact of Power on networks leading eventually to supporting the statements about steady state congestion and flow control as presented in [1] for today's Internet. We point out that the discussions about the latest congestion control algorithms [1, 4, 5, 6, 7, 8, 9, 10, 11] address the dynamics of tracking flow, dealing with multiple intersecting flows, fairness, and more, and which focus on the dynamic behavior of data networks whereas our work here focuses only on the steady state behavior.

1. Introduction

We begin with the use of deterministic reasoning to develop intuition as regards the proper level of traffic to feed into an Internet connection so as to achieve high performance. This quickly leads us to recommend a level of traffic that translates into the rule of thumb, *"Keep the pipe just full, but no fuller"*[1]. We then consider stochastic systems and seek to gain insight into the same question. To accomplish this, we find we must first establish a quantitative metric that considers the tradeoff between a connection's delay and its throughput; and the metric we choose is Power. Power is first introduced as a very general metric and then specialized for the purposes of an Internet connection as the ratio of system efficiency to normalized response time. The goal is then to find the optimum[2] traffic level that maximizes Power. We provide the solution which exposes some great simplicity that matches the rule of thumb we articulated above. We define and present a Universal Power Profile that works for any system of flow and apply it to some important stochastic systems. We treat networks as stochastic systems for which we adjust the traffic level that optimizes Power. In providing the solution of the Power optimal operating point, we identify the Optimal Power Trajectory. Note, however, that this is an equilibrium (steady state) view which does not address the critical dynamics of traffic flow in networks. The issue of network dynamics is then discussed when we introduce some very recent work on network congestion control. That work focuses on dynamic algorithms that seek to track the network parameters and flows so as to match the rule of thumb we describe above while responding to the network dynamics.

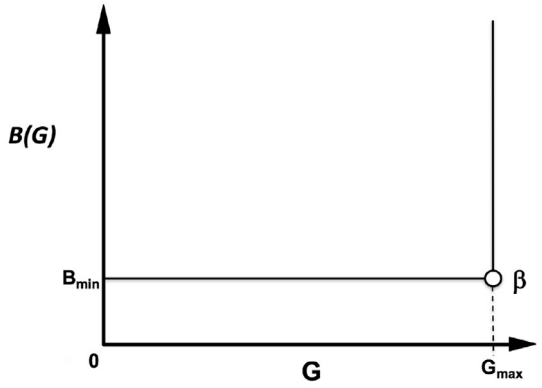
Let us begin with a general model and apply it first to a simple deterministic system. Specifically, consider a "Good" (indepen-

E-mail address: lk@cs.ucla.edu

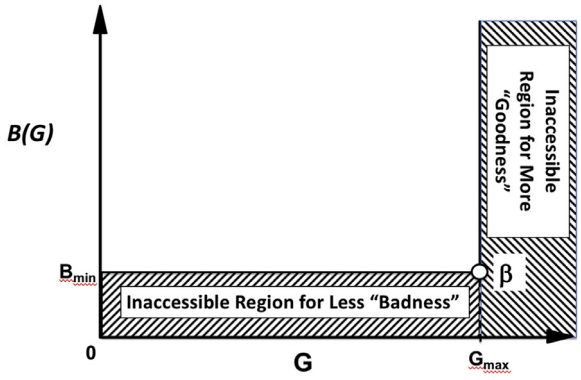
[1] Behind this rule of thumb, we often imply the slightly expanded phrase, *"Keep the pipe's bottleneck just full on average, but no fuller."*

[2] From here on, we use superscript * to denote the (Power) optimized value of a variable.

https://doi.org/10.1016/j.adhoc.2018.05.015
1570-8705/



(a) A Simple Example of the Function B(G).



(b) Inaccessible regions.

Fig. 1. A simple deterministic system.

dent) variable, G in the domain $G \geq 0$ which represents a quantity that we wish to *increase*, while at the same time, we consider a general (dependent) "Bad" function, $B(G)$, whose value we wish to *decrease*. A simple and extreme example of a *deterministic* system of this type is shown in Fig. 1. Specifically, in Fig. 1(a) we show a $B(G)$ that remains constant at its minimum value $B_{\min}$ as G increases from $G = 0$ until the maximum value for G, namely, $G_{\max}$, is reached at which point the system can provide no further increase in G; if we try to gain more G we will simply move vertically up the plot gaining no more G but incurring more $B(G)$[3]. Note further, as shown in Fig. 1(b), that we cannot provide any less "Badness" than $B_{\min}$ and so the horizontal cross-hatched region is inaccessible; similarly we cannot provide any more "Goodness" than $G_{\max}$ and so the vertical cross-hatched region is also inaccessible. To find the operating point of optimal performance in the accessible region(clear white region in Fig. 1(b)), it is clearly at the point β since that is where we achieve maximum Goodness at minimum Badness. No other operating point is better for any sensible definition of optimality.

Later in the paper, we introduce our performance metric, *Power* and use it to mathematically and graphically identify the point of optimal performance (i.e., maximal Power) for this metric in more complex scenarios. Power has some remarkable properties and leads us to the insights about Internet congestion and flow control.

2. Systems of flow

We consider systems of flow in which a stream of arrivals[4] enter a system requesting service[5] from a network of finite capacity (service) resources[6]. In such systems, the inter-arrival times can be deterministic or stochastic as can be the size of their demands from the resources. The system can contain a single resource, or multiple resources arranged in some configuration through which the arrivals flow.

We begin by defining notation for single resource[7] systems of flow in which arrivals enter the system requesting service from a single server and, if that server is busy, then the arrival joins a queue awaiting its turn for service. These systems of flow are the subject of queueing theory [12] for which we define $\bar{x}$ as the average time a customer spends in service and $\bar{t}$ as the average time between customer arrivals. Often we use the following rate notation for these quantities: $\bar{x} = 1/\mu$ (where μ is the service rate) and $\bar{t} = 1/\lambda$ (where λ is the arrival rate). Further we combine these two quantities and define $\rho = \bar{x}/\bar{t} = \lambda/\mu$ as the system efficiency (also referred to as the utilization factor); in general, stable systems require $\rho < 1$. The notation $A/B/K$ is used for systems in which the interarrival time probability density function is of type A, the service time probability density function is of type B and the system contains K servers in parallel. In the multiple server case, $\rho = \lambda/K\mu$ since the total service rate available to the arrival stream is $K\mu$. In all cases, if $\rho > 1$, then the system is unstable[8] in that the queues grow without limit (assuming the queue has enough storage space, and if not, then overflowing customers are "lost", i.e., forced to leave with no service). To instantiate the A and B types, we use the letter D to refer to a deterministic density, the letter M to denote an exponential density, and the letter G to denote a general density.

One of the most important and general results in the theory of such systems of flow (which applies to stochastic as well as deterministic systems) is Little's Result [12] which states for any such system, that $\overline{N}$, the average of N, the number of customers in the system, is given by

$$\overline{N} = \rho\mu T(\rho) \tag{2.1}$$

where $T(\rho)$ is the mean system response time (time in queue plus time in service[9]) and $\mu T(\rho)$ is referred to as the normalized mean response time. Note that the minimum mean response time is ordinarily at the "no-load" point $\rho = 0$ (when there is no time spent in queue) and for single-server systems is simply equal to $1/\mu$, that is $T(0) = 1/\mu$; this explains why $\mu T(\rho) = T(\rho)/T(0) \geq 1$ is referred to as the *normalized* mean response time.

If we consider flow along a connection for general networks, we identify the familiar *Bandwidth-Delay Product* (*BDP*) as the product of the *BBandwidth* (which is the maximum bandwidth that the pipe can support for the flow in this connection, namely, the bandwidth of the slowest link in this pipe, or, if you will, the *Bottleneck Bandwidth* of the link) times the *NLDelay* (which is the time to traverse the connection when there is no traffic interfering with the

[3] We will interpret this behavior as a deterministic system of flow in Section 3 - and will, in Section 4 and beyond, consider complex stochastic systems that are more realistic than deterministic ones, and for which more sophisticated approaches are necessary.

[4] The (network) systems we consider refer to arrivals as the arrival of data blocks (e.g., bits, bytes, packets, messages, etc.).

[5] Typically transmission.

[6] Typically network links with a finite transmission rate, e.g., bits/sec.

[7] We extend this to networks of resources later.

[8] In queueing systems, it is generally recognized that the system is unstable for $\rho \geq 1$, but the $D/D/1$ system is considered stable for $\rho = 1$ if its initial state has a finite queue (usually assumed to be zero).

[9] We do not explicitly address latency due to speed of light, but assume such latency is included in the service time.

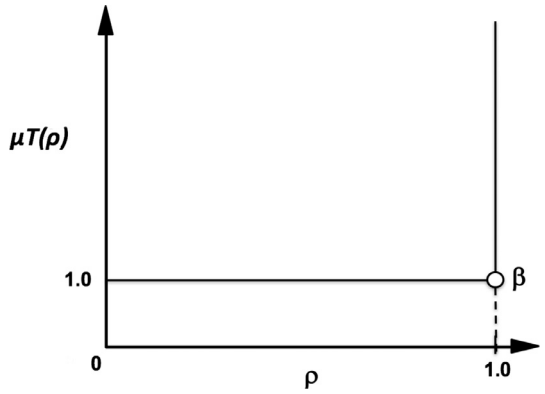


Fig. 2. The D/D/1 deterministic queueing system.

flow, i.e., the *No-Load Delay*)[10]. The *BDP* plays an important role in our optimizations below (e.g., see Theorem 8.1).

For the systems of flow considered below, we set $G = \rho$ and we set $B(G) = \mu T(\rho)$.

3. Deterministic systems of flow & deterministic reasoning

Let us now discuss the issue of deterministic reasoning for deterministic systems. Deterministic reasoning is a useful approach even with stochastic systems since the Law of Large Numbers [12] tells us that in certain limits, systems with stochastic variables behave as if those variables are deterministic. The deterministic approach allows us to develop insights, intuitions and rules of thumb regarding optimal performance that apply for stochastic systems as well.

3.1. *The* D/D/*1 system*

Let us begin with considering the $D/D/1$ system (below in Section 6 we consider more interesting systems such as the classical $M/G/1$ queueing system). So, the system $D/D/1$ is a purely deterministic system wherein a steady stream of arrivals enters, one every $1/\lambda$ seconds, each of which spends exactly $1/\mu$ seconds in service. As long as $\rho \leq 1$, then the previous arrival departs service before (or exactly when) the next arrival occurs; thus the queue is always empty and the server contains a customer a fraction ρ of the time. The response time for each customer is exactly its service time and so the system $D/D/1$ leads to the plot of $B(G) = \mu T(\rho)$ vs $G = \rho$ as shown in Fig. 2.

In this figure, as was the case in Fig. 1(a), for any reasonable definition of optimality, there is little question as to where we should operate for "optimality", and that is exactly at the obvious "knee" of the curve at the point β where $\rho = 1.0$; this achieves the minimum response time and the maximum efficiency. At this point, it is clear from Eq. (2.1), that the number in system, $\overline{N}$, takes on the optimum value $\overline{N^*} = 1$, that is, for $D/D/1$ we note that we have that the *exact* number in system at optimality is equal to 1. Our deterministic reasoning is clear, namely, for optimality, we seek to have the server busy all the time (maximum throughput) and to have customers spend zero time in queue (minimum response time). One can think of the intuition described here as controlling the rate of customer arrivals so as to *"Keep the pipe just full, but no fuller"* where the pipe here has only one space to fill (i.e., the single server with a single customer in service and none in queue). Note further that the *BBandwidth* of this system is the maximum rate of the server (pipe), that is μ customers/sec; moreover, the *NLDelay* for an arrival to move through the pipe is $1/\mu$ sec, and so *BDP* for this system is exactly 1. We see that $BDP = \overline{N^*}$. These themes will repeat throughout this paper.

Our main focus in this paper is to identify the optimum number of customers to have in the system and, in particular, we do not focus on the dynamics and time-dependent behavior of this number. Nevertheless, we point out that the dynamics of deterministic systems are useful to help us gain insight. In that spirit, we point to the material in Section 2.7 of [13] in which we discuss the fluid approximation for queues and describe how to model time-dependent behavior. For example, when a queueing system is temporarily overloaded (as can occur in Internet connections when a bottleneck's bandwidth is temporarily overloaded) then the backlog queue will grow until the load is reduced below the system's capacity at which point the backlogged queue will begin to "drain"; the maximum backlog occurs just when the overload subsides. This concept of needing to drain an overloaded pipe comes up in the algorithms mentioned in Section 7.4.

3.2. *The* D/D/K *system*

We now extend the $D/D/1$ system to include K servers, i.e., $D/D/K$. An arriving customer is assigned to any free server that is available upon its arrival. First we consider the case of *equal rate* deterministic servers, i.e., where a customer spends exactly $1/\mu$ s in service, regardless which server serves that customer. Once again, we have a steady stream of arrivals, one such arrival entering every $1/\lambda$ s. Since we have K servers, the total system service capacity is $K\mu$ customers/s and so, in this deterministic system, we can support a maximum input rate of $K\mu$ arrivals per second, i.e., $\lambda_{\max} = K\mu$ arrivals per second, each of which arrives to find a server just going idle to serve it. The behavior of this system is the same as that shown in Fig. 2, with β being the optimal operating point once again. In this case, we see that each of the K servers is always busy and no customers are in the queue; that is we have kept each of the K bottleneck servers *just full, and no fuller* (i.e., no overflow customers waiting in the queue), resulting in $\overline{N^*} = K$. Note again that the *NLDelay* is $1/\mu$ and the *BBandwidth* is $K\mu$, hence $BDP = K$ which once again gives us $BDP = \overline{N^*}$.

Now consider the $D/D/K$ system with *unequal rates* for each server, namely the kth server has rate μ_k. The total service capacity is now $\sum_{k=1}^{K} \mu_k$ customers/sec. In order to keep (each) bottleneck pipe (i.e., each server) just full, we feed the system with K arrival streams, the kth of which consists of μ_k customers/sec uniformly distributed in time and served by the kth server, and then superimpose these K streams to provide a total input of $\lambda_{\max} = \sum_{k=1}^{K} \mu_k$ arrivals/s. We then draw the same conclusions as for the equal rate case above, namely, that each of the K servers is always busy (i.e., the number of customers in the system is equal to the number of resources - servers) and no customers are in the queue, i.e., $\overline{N^*} = K$ leading to each of them being just full. Let us calculate *BDP* for this system. The *BBandwidth* is $\lambda_{\max}$ and the no-load delay for traffic that flows through the kth server is $1/\mu_k$ so the average no-load delay is the fraction of the traffic served by the kth server, $(\mu_k/\lambda_{\max})$, times that delay summed over all k which gives $NLDelay = K/\lambda_{\max}$; hence we have $BDP = K$ which once again shows that $BDP = \overline{N^*}$.

3.3. K D/D/*1 systems in series*

We investigate a chain of K $D/D/1$ systems in series as shown in Fig. 3.

[10] Often *NLDelay* will be calculated as $T(0)$ which is the no-load delay for the path under consideration.

Fig. 3. K resources in series.

3.3.1. K D/D/*1 systems of equal capacity in series*

We first consider the case where each of the servers has *equal capacity*, i.e., $\mu_k = \mu$ for all $k = 1, 2, \ldots, K$. We drive the system with a deterministic input stream at the rate λ and so each node in the series network sees a utilization factor of $\rho = \lambda/\mu$. Clearly, the time for a customer to pass through the entire series network, $T(\rho)$ is K/μ seconds since there is no queueing in this deterministic system (as before) and each customer spends exactly $1/\mu$ seconds in each of K nodes. The normalization factor for the response time is simply the no-load response time, namely $T(0) = K/\mu$ which, as earlier is the same as $T(\rho)$ for all $\rho \leq 1$. The profile for this case is exactly the same as in Fig. 2 except that the vertical axis should now be labeled $T(\rho)/T(0) = (\mu/K)T(\rho)$ instead of $\mu T(\rho)$, reflecting the fact that customers must now pass through K nodes. Unsurprisingly, we identify β as the optimal operating point again, this being the point where we obtain maximum throughput (μ customers/sec) at minimum response time. We note at β that we have, once again, kept each of the bottleneck pipes (servers) *just full, and no fuller*, and that the number of customers in the system is equal to the number of resources, namely, K, that is, $\overline{N^*} = K$; furthermore, each $D/D/1$ system contains, on average, one customer, i.e., $\overline{N_k^*} = 1$. Calculating *BDP* we see that the *BBandwidth* is μ and the *NLDelay* is K/μ hence $BDP = K$ and so, again, $BDP = \overline{N^*}$.

3.3.2. K D/D/*1 systems of dissimilar capacity in series*

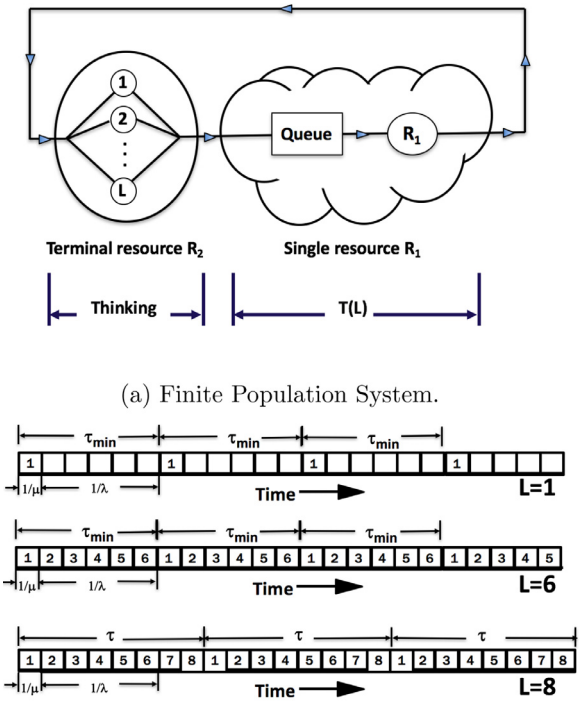
Now consider the non-uniform case where each server has its own constant service rate, namely, the kth server has a rate μ_k. This being a series network, all customers must visit each of the K servers, so we must limit the input rate, λ, to assure that no server has a utilization, ρ_k that exceeds unity. Let us identify the service rate of the slowest server (and there may be more than one with the same slowest rate) and label it μ_s ($\mu_s \leq \mu_k$ for all k); this node is clearly the bottleneck node of the network. Since we require that $\rho_k = \lambda/\mu_k \leq 1$ for each node, then $\lambda \leq \mu_s$. We seek the optimum operating point, i.e., to maximize the throughput, λ, and so we set $\lambda = \mu_s$. We see that nodes with service rates greater than the minimum μ_s will not be serving at their full capacity and so will be busy only μ_s/μ_k of the time, thereby reducing the number of customers in the system to less then K as opposed to the never-idle case for the optimized uniform case. Importantly, the optimum number of customers in the system has now been reduced from K to $\sum_{k=1}^{K} \mu_s/\mu_k$, that is,

$$\overline{N^*} = \sum_{k=1}^{K} \mu_s/\mu_k \tag{3.1}$$

Once again, the same deterministic intuition applies, namely, that we must *"Keep the pipe just full, but no fuller"* where the bottleneck is the slowest node(s) in the series chain; the other nodes are not bottlenecks and therefore are not the critical pipes about which to be concerned. (The profile for this case, once again is exactly the same as that shown in Fig. 2 except that the vertical axis should now be labeled $T(\rho)/T(0) = T(\rho)/\sum_{k=1}^{K} 1/\mu_k$ instead of $\mu T(\rho)$ and the maximum achievable value for the average utilization, $\rho_{\max}$, instead of reaching $\rho_{\max} = 1$ is $\rho_{\max} = \frac{\sum_{k=1}^{K} \mu_s/\mu_k}{K}$.) Calculating *BDP* we see that the *BBandwidth* is μ_s and *NLDelay* $= T(0) = \sum_{k=1}^{K} 1/\mu_k$, hence $BDP = \mu_s \sum_{k=1}^{K} 1/\mu_k = \overline{N^*}$ again.

3.3.3. The deterministic single resource finite population model

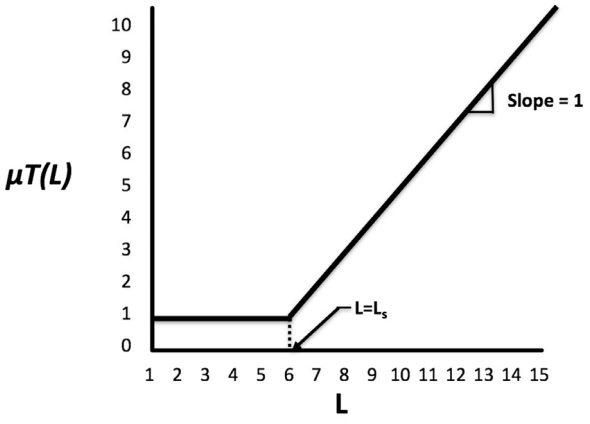
Another manifestation that exposes the value of deterministic reasoning is evident in the extension we now consider. The model, shown in Fig. 4(a), is that of a finite population of L users accessing a single server resource (denoted as R_1) in a cyclic fashion, as in Section 4.11 of [13] as well as in [14]. We assume $1/\mu$ seconds is the deterministic service time a customer spends being serviced in the single server and that the deterministic time each user spends in the "Thinking Resource" (which we denote by R_2) thinking up a new request for the single server (i.e., the classic notion of *thinking time*), is $1/\lambda$ seconds. The system response time, $T(L)$, is defined as the time spent by a user in the cloud waiting for and using the server in this L-user system after that user has finished thinking and has just requested service. Referring to Fig. 4(b) we see in the top row the behavior for a single user ($L = 1$) denoted as "1" cycling through the system. We assume the cycle time for a user is his/her thinking plus service time, i.e. $1/\lambda + 1/\mu$ which we denote by $\tau_{\min}$. Note that if we begin to increase L, then we can insert 5 more users (for a total of $L = 6$ users) without "bumping into" the first user when he comes back for his next service, as can be seen in the middle row of Fig. 4(b). That is, whenever a user requests service, the server is always available to him, as if that server was his private resource; this is a perfect fit. If we increase L beyond 6 as in the bottom row of Fig. 4(b), we will cause users to wait in the queue until the now extended cycle time τ, ends. In this case, the critical number of users, which we denote as the *saturation number*, L_s is 6. It is easy to see that L_s is simply the minimum cycle time, $\tau_{\min}$ divided by the service time $1/\mu$, that is $L_s = 1 + \mu/\lambda$. In general, we see that this deterministic system model performs as if the first L_s users appear as if they were collectively just one user and for each user beyond L_s, the system response time increases by exactly one service time (i.e., by $1/\mu$ seconds) and that user completely interferes with all the other users. As in our earlier observations, our *"Keep the pipe just full, but no fuller"* intuition suggests that we drive the system with exactly L_s customers (giving an always busy server and an always empty queue), thus achieving maximum throughput and no time wasted queueing (i.e., minimum $T(L)$).



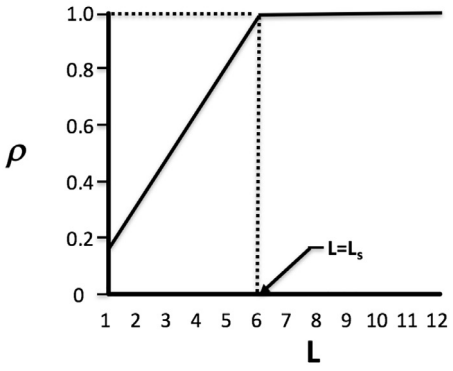
(a) Finite Population System.

(b) Cycle time.

Fig. 4. The finite population with a single resource deterministic model.



(a) The Saturation Number L_s.



(b) Efficiency ρ.

Fig. 5. Performance of the finite resource deterministic model.

This deterministic performance curve is shown in Fig. 5(a). Note, however, that this is not quite a $B(G)$ vs G curve since L is not really a G function. Indeed, recall that in this Section 2, we have chosen $G = \rho$, i.e., system utilization. The measure, ρ, for this finite population model, is simply the fraction of the service capacity of the service resource, R_1 that is utilized by our population. We have already established that L_s is the maximum number of users that the system can support with no interference, and so we see that the relative efficiency (ρ) of the server resource R_1 is the fraction of time the resource is being used in a cycle, which is simply $\rho = L/L_s$. If we plot ρ vs L, we obtain the curve in Fig. 5(b). Note that at the point where $L = L_s$ we have that there is exactly one user in service (i.e., in the "system" - the cloud) at all times, and none in queue (all the rest are thinking) showing again that we are keeping the pipe just full, but no fuller. Clearly, $\overline{N^*} = 1$. Furthermore, the *BBandwidth* is simply μ and the *NLDelay* in the cloud is $1/\mu$, hence, $BDP = 1$. Once again we have $BDP = \overline{N^*} = 1$.

Looking at Figs. 5(a) and (b), we can create a single plot eliminating L and mapping $\mu T(\rho)$ directly vs ρ. This produces Fig. 6 below and we note that this is the same Fig. 2 that we saw in Sections 3.1 and 3.3 where β is again the optimal operating point (i.e., at the point of minimal $\mu T(\rho)$ and maximum ρ or, more generally, at the point of minimum $B(G)$ and maximum G). This process of eliminating an intermediate variable (in this case, L) will be used again when we discuss congestion control in the Internet in Section 7.4 below.

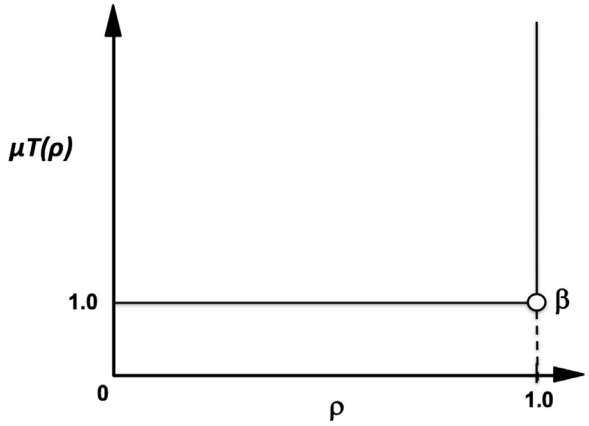


Fig. 6. The finite resource deterministic queueing system.

4. Stochastic systems of flow

In Section 3, we have been considering deterministic systems of flow. These considerations have led us to the dominant insight that we should *"Keep the pipe just full, but no fuller"*. This resulted in operating the systems at their minimum $B(G)$ and simultaneously at their maximum (G), which is the best we could hope for. However, few systems are truly deterministic and so we now ask what insights apply to stochastic systems of flow. Indeed, we find the remarkable and satisfying result that the deterministic insight holds very well (exact in some cases and approximate in others). Stochastic behavior leads us to consider queueing systems [12] in which the arrival process and/or the service process is random. The key observation here is that we cannot drive the system to utilizations that are as high as for the deterministic systems. This is because the uncertainties in the arrival times and the service times (and even in the path followed through the more complex networks we consider below) create unpredictable bunching of arrivals and variations in service times; this causes interference among the objects moving through the system and increases waiting times even when the system is not fully loaded. As a result, we find that the loads must be backed-off from the maximum so as to reduce the additional waiting times (reducing $B(G)$) due to stochastic behavior while at the same time lowering the efficiency (G). This suggests that we need a more sophisticated balancing of $B(G)$ and G.

Our journey here begins, as in Section 1, with the consideration of a plot of $B(G)$ vs G. The key observation is that the typical performance function for stochastic systems is not as simple as that shown in Fig. 1(a) but rather typically looks like that shown in Fig. 7. Here we plot the generic performance curve $B(G)$ vs G (instead of $\mu T(\rho)$ vs ρ) in order to prove a theorem (Theorem 5.1) with great applicability.

As earlier, we seek an "optimum" operating point for the profile in Fig. 7. Looking at this Figure, one wonders if it is better to operate at the point α where we get lots of "good" G while paying the price of lots of "bad" $B(G)$, or conversely, at the point γ where the reverse is true, i.e., getting little "good" G and incurring only a little "bad" $B(G)$. Somehow, we would like to identify the intuitive "knee" of the curve to help us with this trade-off when the knee is not clearly evident. This tradeoff was not in question for Fig. 1(a) since the "knee" of the curve was readily apparent at the point β in that figure. So how can we handle this tradeoff for more general cases?

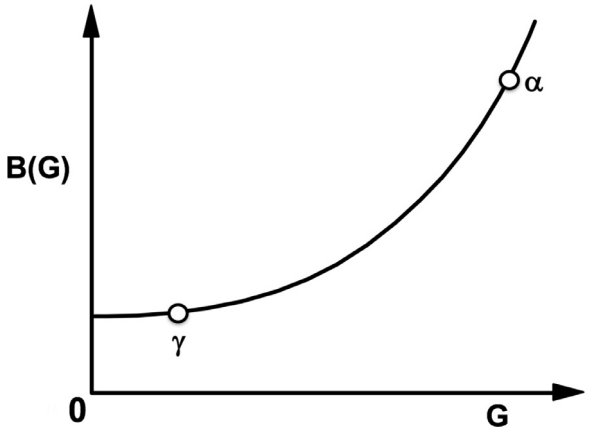
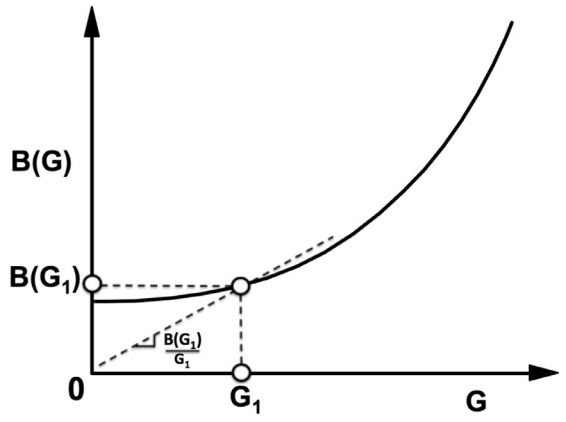
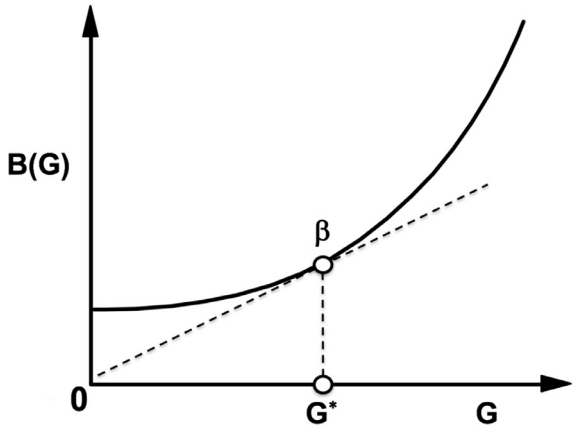


Fig. 7. Stochastic systems.



(a) Slope for Straight Line Out of the Origin.



(b) The Best Tangent.

Fig. 8. Minimum slope is maximum Power.

5. Power functions

To resolve this tradeoff, we introduce a performance metric, called *Power*, which we will use to mathematically (and therefore precisely) identify the knee of the curve. Specifically, we define Power, $P(G)$, as the ratio of G divided by the function $B(G)$, namely, goodness divided by badness[11]. Our objective is to find that value of G which achieves maximum Power, i.e., to optimize the tradeoff between maximizing G while minimizing the risks that come due to the system behavior $B(G)$[12].

Our Power definition below has the attractive property that it leads to intuitive rules of thumb that are totally consistent with the deterministic reasoning we explored in Section 3. Specifically, Power leads to the same intuition that the optimal load on the system is to drive it to *"Keep the pipe just full, but no fuller"* by choosing it to be the *BDP*, i.e., such that the average number in the system should be less than or equal to the number of resources in the pipe.

5.1. The basic form for Power

We define Power, P(G), as

$$P(G) = \frac{G}{B(G)} \tag{5.1}$$

First, let us assume that $B(G)$ is differentiable and convex with respect to G and that $B(G) > 0$ for $G \geq 0$. To obtain maximum Power, we differentiate to find

$$\frac{dP(G)}{dG} = \frac{G\frac{dB(G)}{dG} - B(G)}{B^2(G)}$$

Setting this to zero we find the condition for maximum Power to be:

$$\frac{dB(G)}{dG} = \frac{B(G)}{G} \tag{5.2}$$

Let us interpret this condition. We first note that a straight line out of the origin of the $[G, B(G)]$ plane that passes through any point, say $[G_1, B(G_1)]$, has a slope equal to $B(G_1)/G_1$ as shown in Fig. 8(a). The value of the slope to any point $[G_1, B(G_1)]$ is thus seen to be $1/P(G_1)$, and so to find the value of G which maximizes $P(G)$, we need simply to find that point on the function $B(G)$ for which a line out of the origin to $B(G)$ has a slope which is *minimized*. This optimum point occurs at $G = G^*$ where the line out of the origin to the point $[G^*, B(G^*)]$ is tangent to $B(G)$ as shown in Fig. 8(b). We also observe that this satisfies the optimality condition given in Eq. (3), i.e., that the slope of $B(G)$ at G^* is equal to the slope of a line out of the origin to the point $[G^*, B(G^*)]$.

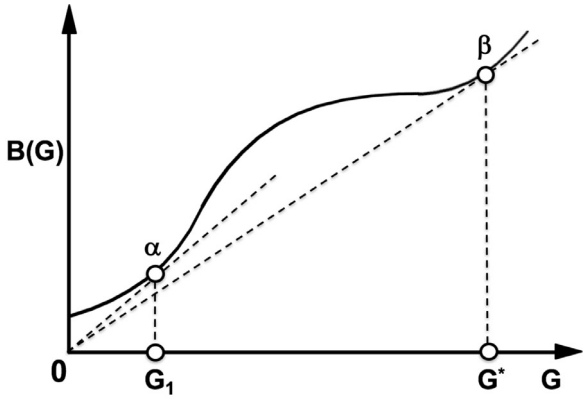
If, however, we drop the requirement that $B(G)$ be convex, it is possible for this last condition (i.e., Eq. 5.2) to hold at some point G_1 and not maximize Power; an example is shown in Fig. 9(a) where there are two points G_1 and G^* that satisfy Eq. 5.2; in this case the point β at G^* with minimum slope identifies the optimum.

Let us now drop the requirement that $B(G)$ be differentiable and convex. In fact, $B(G)$ need not have any properties beyond $B(G) > 0$; that is, it need not be differentiable nor continuous nor convex, etc. In this case, our key observation above still holds, namely, that the slope of a line out of the origin to any point G_1 is seen to be $1/P(G_1)$, and so to find the value of G which maximizes $P(G)$, we need simply to find that point on the function $B(G)$ for which the slope of this line, $1/P(G)$, is minimized. An example of such a situation is shown in Fig. 9(b), where G^* is the optimal power point.

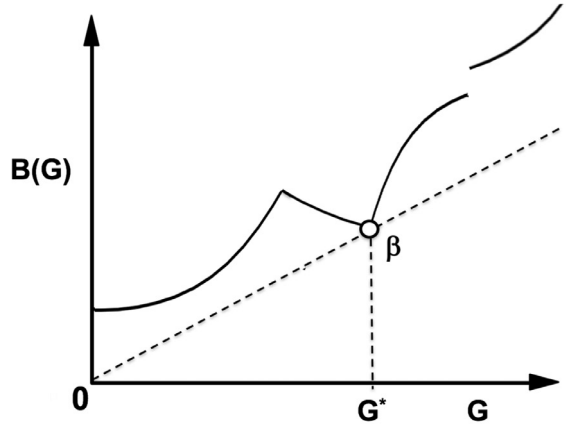
Now let us consider the limiting case of $B(G)$ as $G \to \infty$. If, in this limit, $B(G) < \infty$, then the optimum G^* occurs for $G \to \infty$ since a line out of the origin touching this finite limiting value of B(G) will have slope $\to 0$ and the limiting value of $P(G)$ will approach infinity.

[11] We explore a more generalized definition of Power in the Appendix.

[12] Note that optimizing Power has application to any field of study well beyond those addressed herein.



(a) Two Multiple Tangents.



(b) Non-differentiable, Non-convex, Discontinuous $B(G)$.

Fig. 9. Finding the optimum operating Point G^*.

Further, and trivially, we see that the optimum operating point we found for the deterministic systems in Section 3 also corresponds to optimal power (the slope of a line out of the origin is minimized at the point $\rho^* = 1$).

We can now state the following:

Theorem 5.1 (Basic Power Theorem)**.** *For a convex and differentiable $B(G) > 0$ defined for $G \geq 0$, the Power, $P(G)$, is maximized at that value of G, namely, G^*, for which a straight line out of the origin is tangent to $B(G)$. The analytic condition for finding this point is simply Eq. (5.2) above, namely,*

$$\left.\frac{dB(G)}{dG}\right|_{G=G^*} = \left.\frac{B(G)}{G}\right|_{G=G^*}$$

More generally, for any $B(G)$ for which $B(G) > 0$ in the range $G \geq 0$, then $P(G)$ is maximized at that value of G, namely, G^*, for which the slope of a straight line out of the origin to $B(G^*)$ is minimized.

Now what does the metric Power have to say about our intuitive result, *"Keep the pipe just full, but no fuller"*. We address this by studying some specific queueing systems as examples of stochastic systems of flow in the next sections.

6. Using the power metric for queueing systems

In this section we determine the optimal operating point for a number of queueing system configurations. The optimization metric we use is Power. We show for all $M/G/1$ systems that $BDP = \overline{N^*} = 1$ at optimization. For some other systems, we show that the optimized average number in system, $\overline{N^*}$, is typically less than or equal to the number of resources in the pipe.

Once again we set $G = \rho$ and $B(G) = \mu T(\rho)$. In this case we see that Power is expressed as the ratio of efficiency to normalized response time, i.e.,

$$P(\rho) = \rho/\mu T(\rho) \tag{6.1}$$

We will use this definition throughout the rest of this paper (and will introduce its generalization in the Appendix).[13]

Since $P(\rho) = \rho/\mu T(\rho)$ and $\overline{N} = \rho\mu T(\rho)$, we see that

$$P(\rho) = \rho^2/\overline{N} \tag{6.2}$$

which offers another expression for Power.

Furthermore, since $\rho \leq 1$ and $\mu T(\rho) \geq 1$ we conclude from Eq. (6.1) that

$$P(\rho) \leq 1 \tag{6.3}$$

for all stable queueing systems.

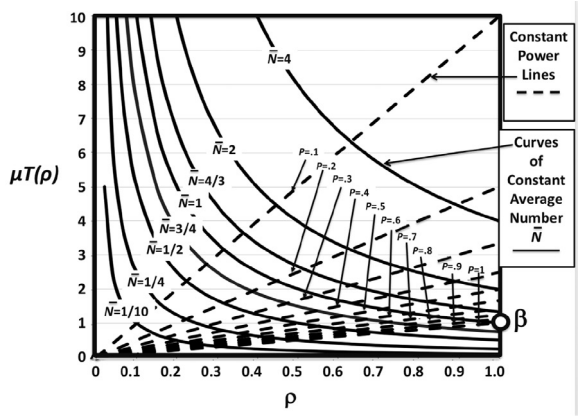
6.1. The universal power profile

As we have said, the plot of $\mu T(\rho)$ vs ρ is the common performance plot for queueing systems. Now that we have introduced $P(\rho)$ as our important optimization metric, we find from Eqs. (6.1) and (2.1) that, *independent of the queueing system involved*, we can easily plot curves of constant power, $P(\rho)$, as well as curves of constant average number in system, $\overline{N}$, on the $\mu T(\rho)$ vs ρ axes as shown in Fig. 10(a). On this plot we note that a curve of constant power, say P_0, is simply a (dashed) straight line out of the origin of slope $1/P_0$ since from Eq. (6.1) we have $\mu T(\rho) = \rho/P_0$; these are shown in Fig. 10(a) for the sample values P_0= 1.0, 0.9, 0.8, ... , 0.1. In addition, since for any particular average number in system, say $\overline{N_0}$, we note from Eq. (2.1) that $\mu T(\rho) = \overline{N_0}/\rho$ allowing us to plot the family of hyperbola as (solid) curves in Fig. 10(a); we show these for a sample set of values, namely, $\overline{N_0}$=1/10, 1/4, 1/2, 3/4, 1, 4/3, 2 and 4.

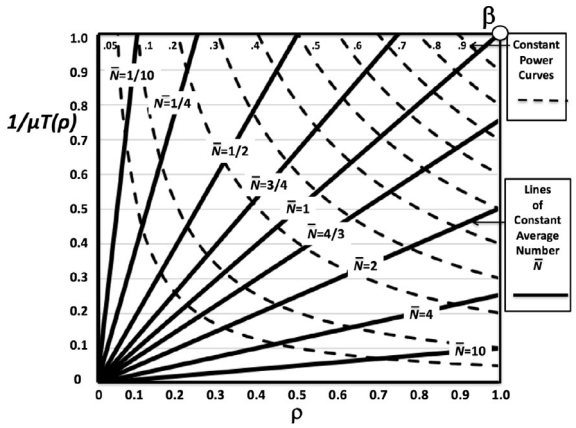
We now introduce the *inverse* of the normalized response time, namely, the function $T(0)/T(\rho)$ (which we often write as $1/\mu T(\rho)$ when there is no ambiguity). When plotted against ρ, we conveniently find that the range of this function is fully contained in the $[1 \times 1]$ unit square as shown in Fig. 10(b) where we have plotted curves of constant Power and curves of constant $\overline{N}$ for essentially the same set of values as in Fig. 10(a). We refer to this canonical plot of $1/\mu T(\rho)$ versus ρ as *The Universal Power Profile*. As above, these curves are independent of the queueing system involved. In this case we note the dual situation to that of Fig. 10(a) in that the curves of constant power are now hyperbola (since for any P_0, $1/\mu T(\rho) = P_0/\rho$ shown as dashed lines) and curves of constant $\overline{N}$ are now straight lines out of the origin (since for any particular average number in system, say $\overline{N_0}$, $1/\mu T(\rho) = \rho/\overline{N_0}$ shown as solid lines).

For consistency, in both parts of Fig. 10 we have shown the constant Power curves as dashed lines and the constant $\overline{N}$ curves as solid lines. Let us observe in Fig. 10(a), that at $\rho = 1$, the constant Power curves intersect the vertical axis at $1/P_0$ and the constant $\overline{N}$ curves intersect this vertical axis at $\overline{N_0}$. This situation is reversed for the Universal Power Profile in Fig. 10(b) in that at $\rho = 1$ the constant Power curves intersect the vertical axis at P_0

[13] The ratio throughput to response time was first introduced as a measure of power by Giessler, et al.[15]; however note our Power definition in Eq. (5.1) is far more general and that the more specific version of Power we introduced in Eq. (6.1) is a ratio of *normalized* quantities which provides a metric that lends itself better to optimization [16].



(a) Power and $\overline{N}$ for Any Queueing System



(b) The Universal Power Profile.

Fig. 10. Performance curves for any single server queueing system.

and the constant $\overline{N}$ curves intersect the vertical axis at $1/\overline{N_0}$. However, in this normalized inverse case of the Universal Power Profile, we have in addition that the constant Power curves intersect the line $1/\mu T(\rho) = 1$ at $\rho = P_0$ and the constant $\overline{N}$ curves intersect the line $1/\mu T(\rho) = 1$ at $\rho = \overline{N_0}$. Another advantage of the Universal Power Profile is that we can see the full range of P and $\overline{N}$ curves in the compact region of the [1x1] plot whereas in the ordinary plot of $\mu T(\rho)$ vs ρ, the upper limit of the vertical axis shown will limit the visibility of large values of $\overline{N}$ (note that for these queueing systems, we need only consider $P(\rho) \leq 1$ as seen in Eq. (6.3)).

Given our discussion earlier for deterministic systems, we note that β, the optimal deterministic operating point for our systems, is easily located on both plots of Fig. 10. Specifically, β is identified with the point $\overline{N^*} = 1$ and $\rho^* = 1$ (where also $P(\rho) = 1$ and $\mu T(\rho) = 1$) as shown in both parts of the Figure. In addition, for all single resource systems, we have that $BDP = 1$.

Once we apply both plots in Fig. 10 to a given class of queueing systems (as for example in Section 6.3 for $M/G/1$), we can plot the actual $1/\mu T(\rho)$ vs ρ curves to investigate the behavior of that class.

6.2. The M/M/1 queueing system

We begin by applying the Power metric to the classic queueing system $M/M/1$ [12].

For $M/M/1$, we know that $\mu T(\rho) = 1/(1-\rho)$. Thus, $d\mu T(\rho)/d\rho = 1/(1-\rho)^2$. Applying Eq. (5.2), we see that optimal Power occurs for that ρ which satisfies $\rho = 1 - \rho$, i.e., $\rho = 0.5$. That is, the maximum Power occurs at the point G^*, where $G^* = \rho^* = 0.5$. In addition, at maximum Power, $\mu T(0.5) = 2 = 2\mu T(0)$. Thus, for $M/M/1$, the optimum Power point occurs at half the maximum efficiency and twice the minimum normalized response time. Moreover, the maximum Power is 1/4. Furthermore, we know for $M/M/1$ that $\overline{N}$, the average number in system, is given by $\overline{N} = \rho/(1-\rho)$. *Hence, at optimality, we see that* $\overline{N^*} = 1$. Thus, we have the key result for $M/M/1$

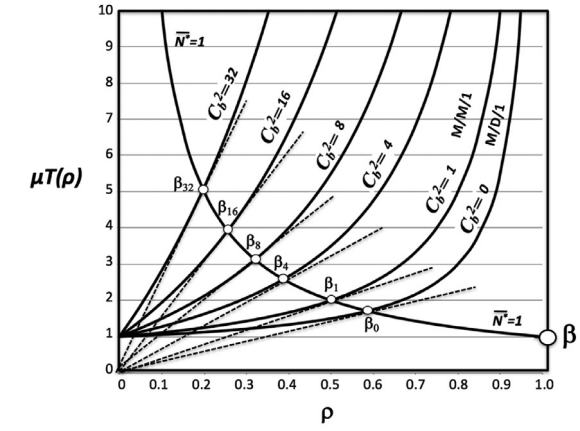
$$\overline{N^*} = 1 \tag{6.4}$$

and $\rho^* = 0.5$. Furthermore, the *BBandwidth* is simply μ and the *NLDelay* is the average service time $1/\mu$; hence, $BDP = 1$. Once again we have $BDP = \overline{N^*} = 1$. This result in Eq. (6.4) is especially pleasing since, as we saw from Section 3, our deterministic reasoning of *"Keep the pipe just full, but no fuller"* suggests that we keep *exactly* one person in the system in order to maximize efficiency (the single server is always busy) while minimizing response time (no one is on queue wasting time). However, we cannot control the $M/M/1$ system deterministically (it is a stochastic system), and so this optimum Power result says that for $M/M/1$, control the input rate so as to keep one person in the system *on average*; occasionally, there will be more than one in system which adds additional (wasted) response time and occasionally there will be no one in the system which reduces efficiency, *but by setting the average number in system = 1, we are doing the best possible*. From now on, we will imply, but usually omit, the additional phrase *on average* to our intuitive rule *"Keep the pipe just full on average, and no fuller"*. These results for $M/M/1$ were first shown by the author [2].

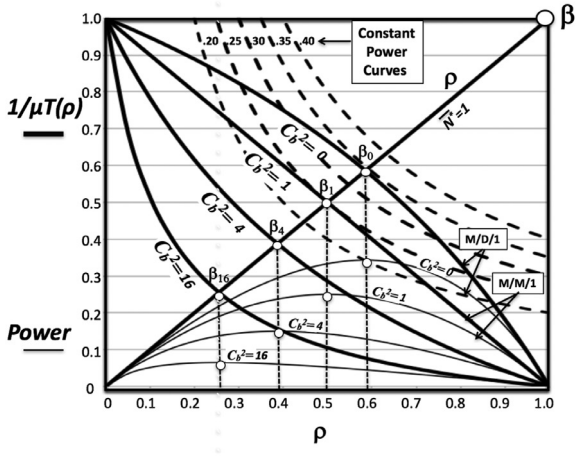
Another way to think about these results is as follows. We recognize that in a pure deterministic system, we keep *exactly* one person in the system in order to maximize Power (i.e. $\rho = 1$ giving 100% utilization of the server and no one ever in the queue wasting time). However, in a stochastic system, we must account for fluctuations which cause queues to form, and to ameliorate the waste due to these queues, we allocate some residual system capacity to absorb the random fluctuations (this is the "Balance of Power Principle" for Pareto optimal power as articulated by Yemini [17]). In the case of $M/M/1$ we just found it optimal (with regard to Power) to load the server at only 50% efficiency, leaving the other 50% to absorb the stochastic fluctuations. We will see this numerous times below where we find it optimal to back off from the 100% utilization that optimizes pure deterministic systems and accept lower utilization of bottleneck resources to ameliorate the effects of stochastic traffic, while at the same time accepting some additional response time.

6.3. The M/G/1 queueing system

We now extend our analysis to the more general $M/G/1$ queueing system[12]. As in Sections 2 and 6 we set $G = \rho$ and $B(G) = \mu T(\rho)$. We will now apply the results of Section 5.1 to $M/G/1$. We know from Eq. (5.71) of [12] that $\mu T(\rho) = 1 + \frac{\rho(1+C_b^2)}{2(1-\rho)}$ where C_b is the coefficient of variation for the service time (i.e., the service time standard deviation divided by its mean). Preparing to apply Theorem 5.1, we observe that $d\mu T(\rho)/d\rho = \frac{1+C_b^2}{2(1-\rho)^2}$ and that $\mu T(\rho)/\rho = 1/\rho + \frac{1+C_b^2}{2(1-\rho)}$. Equating these last two as the condition for optimality, we see that maximum Power occurs at that ρ, namely ρ^*, which satisfies $1 = \frac{\rho^{*2}(1+C_b^2)}{2(1-\rho^*)^2}$. Now recall that $\overline{N} = \rho\mu T(\rho)$ and using ρ^* in this expression for $\overline{N}$ produces $\overline{N^*} = 1$ as the condition for optimal Power *for all M/G/1 queueing systems*! This interesting result for $M/G/1$ was first shown by the author in [3]. Once again, we see that our deterministic reasoning of *"Keep the pipe just full, and no fuller"*, leads us to obtaining optimal Power



(a) Optimality at $\overline{N^*} = 1$



(b) The Universal Power Profile for $M/G/1$.

Fig. 11. The queueing system $M/G/1$.

by running the system at a level such that the optimal average number in system, $\overline{N^*}$, is *exactly* equal to 1, i.e.,

$$\overline{N^*} = 1 \qquad for\ M/G/1 \tag{6.5}$$

Just as for $D/D/1$ and $M/M/1$, the *BBandwidth* is clearly μ and the *NLDelay* is $T(0) = 1/mu$, hence $BDP = 1 = \overline{N^*}$.

Moreover, as shown in [3], the optimal load, ρ^*, is

$$\rho^* = \frac{1}{1+\sqrt{(1+C_b^2)/2}} \quad for\, M/G/1 \tag{6.6}$$

As noted earlier, with stochastic systems, at optimality, we must allocate some residual capacity, $1-\rho^*$, to absorb the stochastic fluctuations, and for $M/G/1$ we see that this allocation of $1-\rho^*$ ranges from $\sqrt{2}-1 = 0.414$ (when $C_b^2 = 0$, i.e., $M/D/1$) to 0.5 (when $C_b^2 = 1$, i.e., $M/M/1$), to 1 (when $C_b^2 = \infty$). This basic results in this paragraph are generalized in Appendix B.

Let us now examine the performance of the $M/G/1$ system by filling in its behavior on the plot we showed in Fig. 10(a) (we choose not to clutter this figure with the full set of curves from Fig. 10(a) - specifically, we only need $\overline{N^*} = 1$); this gives us Fig. 11(a) in which we show $\mu T(\rho)$ *vs* ρ for a number of $M/G/1$ cases (i.e., $C_b^2 = 0$ which is $M/D/1$, $C_b^2 = 1$ which is $M/M/1$, and others up to $C_b^2 = 32$). We show the tangent out of the origin which locates the optimum operating point[14] for each of these curves and the locus of these optimal points is exactly at $\overline{N^*} = 1$ as just proven. Note, as with $M/M/1$, that the optimum has moved from the deterministic optimum at point β to the set of points $\{\beta_{C_b^2}\}$ in the interior of the diagram at various values of ρ and $\mu T(\rho)$, but still maintaining the value of $\overline{N^*} = 1$. This is interesting and elaborated upon in the next paragraph.

We now examine the performance of $M/G/1$ on the Universal Power Profile of Fig. 10(b) giving us Fig. 11(b) in which we show $1/\mu T(\rho)$ as curved solid lines and $P(\rho)$ as thin concave solid curves. Once again, in order to reduce any possible clutter, we show only $\overline{N^*} = 1$ and a smaller number of power curves than we did in Fig. 10(b). Note that maximum Power occurs for a set of points that lie on the line $f(\rho) = \rho$ shown as a linear heavy solid line at unit slope. This follows since, as we noted above, $\overline{N} = \rho\mu T(\rho)$ and if we set $\overline{N} = 1$ in this last equation, we see that the intersection of $1/\mu T(\rho) = \rho$ occurs at $\overline{N} = 1$. That is, once again we see that the optimum occurs at $\overline{N^*} = 1$. Observe that β is the optimal operating point for the deterministic case of $D/D/1$, but that for $M/G/1$ we find $\{\beta_{C_b^2}\}$, the set of optimal operating points, moving down the line $f(y) = \rho$ as C_b^2 grows. Note well that all of the optimal operating points lie on the line $\overline{N^*} = 1$ and so we may refer to this line, $\overline{N^*} = 1$, as the "Optimal Power Trajectory". As we have remarked, the best one can hope for is to operate at the deterministic point $\mu T(\rho) = 1$ and $\rho = 1$, but as the stochastic component increases (in the case of $M/G/1$ as C_b^2 grows), we must leave more and more capacity (i.e., lower utilization ρ while incurring more delay $\mu T(\rho)$) to allow the system to absorb the fluctuations. The point to be made is that, wherever we are on the Optimal Power Trajectory, we always maintain $\overline{N^*} = 1$ (*"Keep the pipe just full, and no fuller"*). And, this intuition comes right out of our deterministic reasoning supported by the *BDP*.

Let's examine this $M/G/1$ Universal Power Profile plot a bit further. We define $y(\rho) = 1/\mu T(\rho)$. First we show that $y(\rho)$ is symmetrical around the line $f(\rho) = \rho$. This requires that $\rho = y(y(\rho))$ and this is easily established from the expression for $y(\rho) = \frac{2(1-\rho)}{2(1-\rho)+\rho(1+C_b^2)}$. Further, we recall from Section 6.1 that Power on this plot is a set of hyperbolas (shown as dashed lines), each for a constant value of Power (i.e., $1/\mu T(\rho) = P_0/\rho$). By definition, these hyperbolas are clearly symmetrical about the line $f(\rho) = \rho$. For a given $y(\rho)$, one seeks that constant Power curve (dashed hyperbola) of maximum value with which $y(\rho)$ intersects. Since both functions are symmetric about the line $f(\rho) = \rho$ this will be a point of tangency (at a slope of -1) and will provide maximum Power, which, as was stated above, will lie on the line $f(\rho) = \rho$ which we have shown is the line $\overline{N^*} = 1$.

6.4. The G/M/1 queueing system

The queueing system $G/M/1$ does not enjoy the canonic properties of the $M/G/1$ system. That is, we no longer find that the optimal Power point occurs when $\overline{N} = 1$ as we did for all $M/G/1$ systems. However, we do find intuitive results similar to our earlier intuition which warns us about pumping too much traffic into the pipe's bottleneck, i.e., we find for a large class of $G/M/1$ systems that $\overline{N^*} \leq 1$.

We begin by looking at a class of $G/M/1$ systems in which $C_a{}^2$, the coefficient of variation of the interarrival time, satisfies $C_a{}^2 \leq 1$. This is the class of systems where the interarrival time distribution is a k-stage Erlangian distribution [12]. In particular, it is shown in [16] for all k-stage Erlangian distributions, that $0.796 \leq \overline{N^*} \leq 1.0$ with $\overline{N^*} = 1$ for $k = 1$ (which is equivalent to $M/M/1$) and decreasing monotonically to $\overline{N^*} = 0.796$ as $k \to \infty$ (which is equivalent to $D/M/1$). Thus we see that $\overline{N^*}$ hovers near $\overline{N^*} = 1$; apparently the Power metric is more sensitive to the stochastic behavior of the arrivals than it is to the stochastic behavior of the service times, but

[14] We denote these optimal operating points as $\beta_{C_b^2}$.

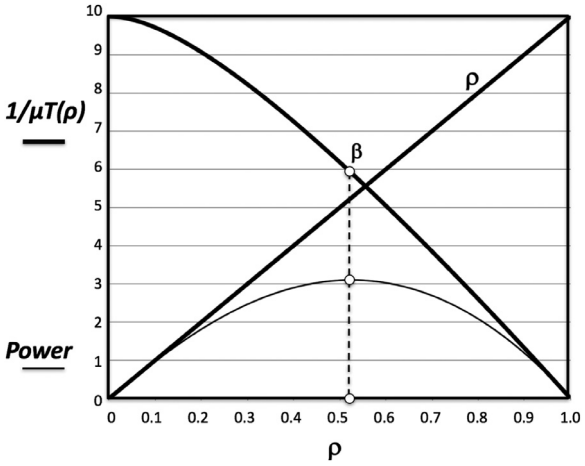


Fig. 12. Optimality for the queueing system $E_2/M/1$.

similarly drops the load (reducing the system efficiency) to avoid potential queue buildups. By way of illustration, we show an example of a $G/M/1$ system that behaves approximately as does $M/G/1$. Specifically, our example is the $E_2/M/1$ system described in Problem 6.2 of [12]. We find $\overline{N^*} = 0.890$ as the condition for optimal Power. Of special note is how close to our earlier $M/G/1$ optimal value of $\overline{N^*} = 1$ is this case. In Fig. 12, we show the usual Power Profile for this $E_2/M/1$ system. Note that the optimum, denoted by the label β, is close, but not (as earlier) at, the intersection of ρ and $1/\mu T(\rho)$.

Let us now look at $G/M/1$ systems in which ${C_a}^2 \geq 1$. Such a class includes the Hyperexponential interarrival time distribution [12]. In [16], it is shown for a class of Hyperexponential distributions, that as ${C_a}^2 \to \infty$, then $\overline{N^*} \to 0$. Again we suspect this is the effect of the Power metric responding to the potential queue buildups as C_a grows. It is worthwhile to note that $\overline{N^*} \leq 1$ for these $G/M/1$ systems which supports the *"... but no fuller"* portion of our intuitive conclusions.

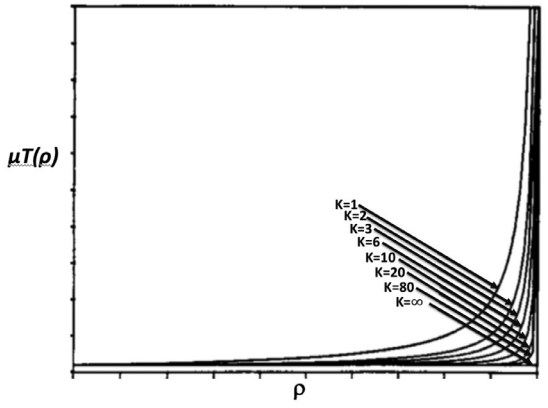
6.5. The M/M/K queueing system

As a further extension, let us extend this concept of *"Keep the pipe just full, and no fuller"* by looking at the multiple server system $M/M/K$ [12]. As usual, we set $G = \rho$ and $B(G) = \mu T(\rho)$.

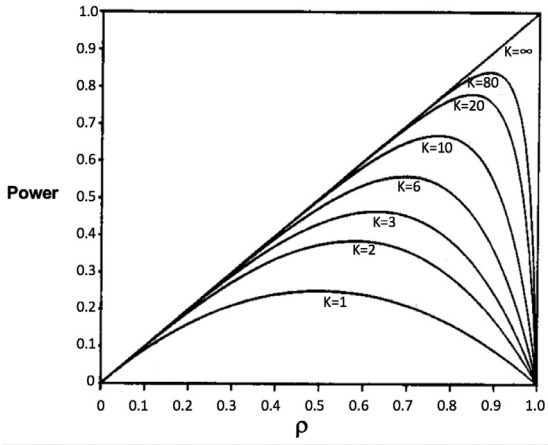
The limiting behavior of $\mu T(\rho)$ vs ρ for $M/M/K$ as $K \to \infty$ is the same as the behavior of D/D/1 as was shown in [3]. This behavior is shown in Fig. 13(a). Moreover, we see from Fig. 13(b) that as K increases, the optimum Power occurs at an increasing value of ρ which suggests that the optimum $\overline{N^*}$ is also increasing with K. Specifically, we see from [3] as shown in Fig. 14 below, that at optimum Power, there are, on average, *approximately* K customers in the system (one for each server), i.e., $\overline{N^*} \approx K$, but also $\overline{N^*} \leq K$ once again supporting *"Keep the pipe just full, but no fuller"* where the pipe consists of K servers, each of which is busy serving approximately one customer on average (and no "extra" customers are wasting their time waiting in the queue).

6.6. Summary for the power metric for queueing systems

The overwhelming intuition we extract from this Section 6 is that optimizing Power leads to the same deterministic intuition as earlier, namely that the optimal load on the system drives it to *"Keep the pipe just full, but no fuller"* by choosing $\overline{N^*}$ to be the *BDP* (which results in $\overline{N^*}$ typically being less than or equal to the number of resources in the pipe). In addition, we introduced the Universal Power Profile and the Optimal Power Trajectory as tools of great generality in the study and evaluation of stochastic systems of flow.



(a) Limiting Behavior $\to D/D/1$.



(b) Power.

Fig. 13. The queueing system $M/M/K$.

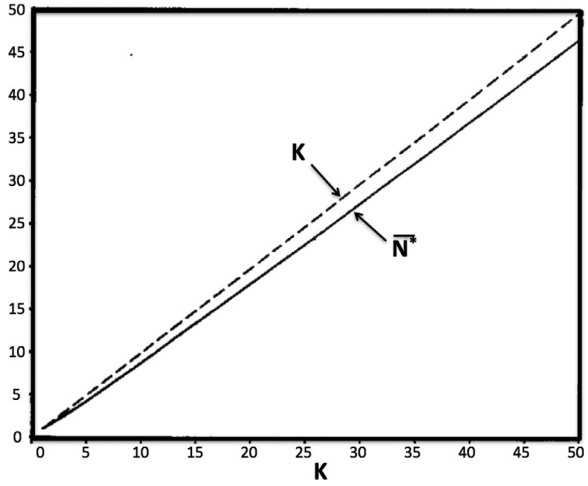


Fig. 14. The optimum number in system is approximately K for $M/M/K$.

7. Applications to optimization of networks

Let us now extend our use of Power to find optimum operating points for *networks* with stochastic traffic. By networks, we mean networks of queues, i.e., systems of more than one service station,

be it in a parallel network[15], a finite population network, a series network, or a more general network of arbitrary topology. The series networks discussed in Section 7.2 below are of special interest to our later discussion in Section 7.4 on Internet congestion control since an Internet TCP connection can be modeled as a path of links in series between the source and destination nodes of that Internet connection.

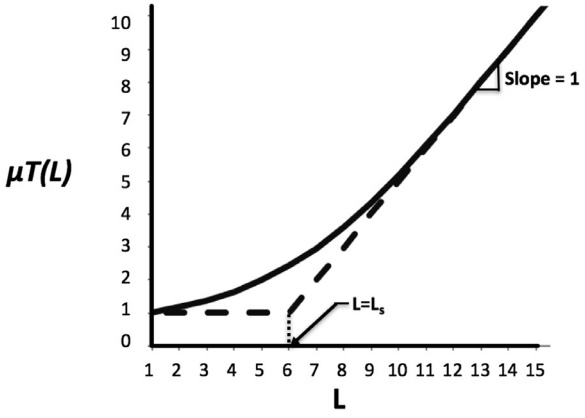
As usual in these systems of flow, we set $G = \rho$ and $B(G) = T(\rho)/T(0)$, the normalized average response time for data to traverse the network. In these networks, the normalization constant we use is $T(0)$ which is the average time to traverse the network when no other traffic is in the network (i.e., the *"no-load"* response time); for each of the networks considered below, we will give explicit expressions for $T(0)$.

In the case of networks below, we find we occasionally need to distinguish between maximizing global network power and maximizing the power of the individual flows. In addition we will discuss the issue of whether we can control all the flows in the network or if the flows act on their own. These issues add considerable complexity to the discussion.

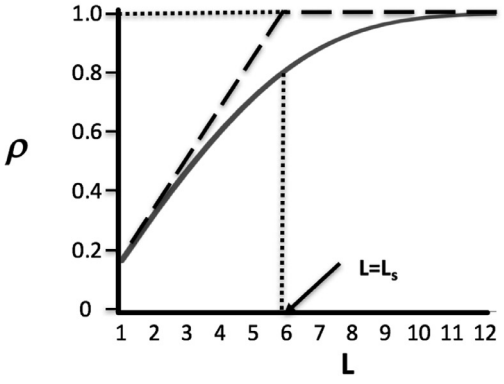
7.1. The stochastic finite population model

This discussion of finite population networks is of limited importance for us, but we include it to expose the way in which these networks reinforce our continuing theme of the value of deterministic reasoning and its affirmation of the rule of thumb *"Keep the Pipe Just Full, But No Fuller"*.

We now return to the single resource finite population model of Section 3.3.3 shown earlier in Fig. 4(a), but this time we consider a stochastic system in which the service times are exponentially distributed with the same mean as earlier, namely $1/\mu$ seconds, and the thinking time is exponentially distributed with the same mean as earlier, namely, $1/\lambda$ seconds. The mean response time, $T(L)$, is defined as the mean time spent by a user in the cloud waiting for and using the cloud server, R_1, in this L-user system after that user has finished thinking and has requested service from the cloud shown. The deterministic system model of Section 3.3.3 gives us a lower bound for $\mu T(L)$ in this stochastic system, and that is shown in Fig. 15(a) as the dashed line whereas the true mean response time for the stochastic system is shown as the solid line in Fig. 15(a) (this curve was calculated using Eq. (4.65) from [13] for which the parameters were chosen as $\lambda = 0.2$, $\mu = 1.0$ and thus $L_s = 6$). We also plot the efficiency, ρ vs L in Fig. 15(b) for these same parameters (where the dashed line is the deterministic ideal upper bound case from Section 3.3.3 and the true efficiency is the solid line). Our *"Keep the pipe just full, but no fuller"* intuition suggests that we drive the system with the optimum value L^* in the range of L_s customers (giving an almost busy server and an almost empty queue), but since the system is actually stochastic we expect to load it below it's saturation point (as discussed in Section 4), that is, we expect $L^* < L_s$. As we did in Section 3.3.3, we can cross-plot the two graphs of Fig. 15 and create a single plot eliminating L and mapping $\mu T(\rho)$ directly vs ρ; this is shown in Fig. 16.



(a) Mean Response Time.



(b) Efficiency vs ρ.

Fig. 15. Performance of the finite resource stochastic model.

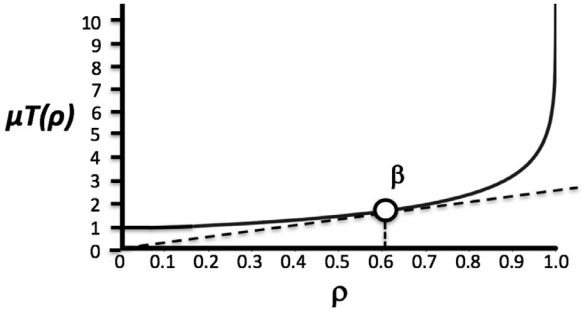


Fig. 16. The finite resource queueing system.

Unsurprisingly, it turns out that when we calculate the Power for this stochastic system we find that the optimum Power point does actually result in $L^* < L_s$; indeed, for the example shown in Fig. 15, we find that the optimum $L^* \approx 4$ and this corresponds to the tangent line out of the origin of Fig. 16 which occurs at $\rho \approx 0.6$ and identified by the optimal operating point β as usual. Most importantly, we find for this example our earlier intuition that the deterministic optimum $\overline{N^*} = 1$ holds very well in this example for which we find the stochastic optimum $\overline{N^*} \approx 1$. As in the deterministic case in Section 3.3.3, $BDP = 1$.

[15] We do not pursue parallel networks in this paper, but point to some of the results in [18] and [19] which include the following. Consider a Poisson arrival stream at rate λ which splits into K streams, where the kth stream has rate $\lambda_k = p_k\lambda$ according to a given set of probabilities, p_k. Each stream is served by its own parallel server with mean service time $\overline{x_k}$. If the service time for each is exponentially distributed and if we scale λ to maximize Power for the system, we find that $\sum_{k=1}^{K} \mu_s/\mu_k \le \overline{N^*} \le K$ where, as usual, μ_k is the service rate of the kth server and μ_s is the slowest of the exponential servers. If the λ_k can be selected independently to maximize Power for the system, then $\overline{N^*} = K$. On the other hand, if the service time for each is of its own General type (G), and if $\rho_k = \rho\ \forall k$, then optimum Power gives $\overline{N^*} = K$.

7.2. Series networks

As we stated above, the series networks discussed herein are of special interest to our later discussion in Section 7.4 on Internet congestion control since a single Internet TCP connection can be modeled as a path of links in series between the source and destination nodes of that Internet connection. This discussion of series networks is of value in modeling and optimizing the performance of single flows over Internet connections. A summary of our findings for series networks, as well as other related results is given in Theorem 8.1 of Section 8.

7.2.1. The series network of K identical M/M/1 queueing systems

We first consider a series network consisting of K identical $M/M/1$ queueing systems in tandem, i.e, a stochastic version of the series network considered in Section 3.3.1. This system was considered in our previous paper [2] in which we assume each $M/M/1$ system is independent of the others (see the Independence Assumption of [12]). ρ is, as usual, the efficiency of each queueing system (and, due to them being identical, is also the efficiency of the entire tandem system). The results for this network are that optimal Power occurs at $\rho^* = 0.5$ for each member of the K-member chain and that $\overline{N^*}$, the average number of customers in the full chain, is

$$\overline{N^*} = K \tag{7.1}$$

and these K are uniformly distributed among the K members such that for each member, say the kth member of the chain, the Power optimal average number is $\overline{N_k^*} = 1$ (as in Section 3.3.1). Once again, we see that each node is an equivalent bottleneck, and so each node satisfies *"Keep the pipe just full, and no fuller"*. The *BBandwidth* is obviously μ and the *NLDelay* to pass through the chain is K/μ, hence, $BDP = K$. Once again we have $BDP = \overline{N^*} = K$.

7.2.2. The series network of K heterogeneous M/M/1 queueing systems

Next we consider a series network consisting of K heterogeneous M/M/1 queueing systems in tandem, i.e., the kth server has a mean service time of $1/\mu_k$ seconds; this is a stochastic version of the series network considered in Section 3.3.2. As shown in [16] and [20], we find that when Power is optimized, then $\overline{N^*} \le K$ and also $\overline{N^*} = \sum_{k=1}^{K} (\overline{N_k^*})^2$ where $\overline{N_k^*}$ is the Power optimized average number in the kth node of the tandem. Furthermore, in [16] it is shown that $\sum_{k=1}^{K} \mu_s/\mu_k \le \overline{N^*}$ where μ_s is the rate of the slowest server, i.e., $\mu_s \le \mu_k$ for all k. Thus, at optimal Power we see that $\overline{N^*}$ is bounded above and below by

$$\sum_{k=1}^{K} \frac{\mu_s}{\mu_k} \le \overline{N^*} \le K \tag{7.2}$$

The *BBandwidth* is simply μ_s and the *NLDelay* to pass through the chain is $\sum_{k=1}^{K} 1/\mu_k$, hence, $BDP = \sum_{k=1}^{K} \mu_s/\mu_k$. In this case we have $BDP \le \overline{N^*} \le K$.

7.2.3. The series network of K identical "M/D/1" queueing systems

Again we consider K servers in series, the first of which is fed with Poisson traffic, but now where the service time of each user is constant (and identical) at each server. The first node is an M/D/1 queue, but the subsequent nodes are more complicated; we abuse the notation and refer to this as a series of "$M/D/1$" systems. In [19] we show that

$$\overline{N^*} = K \tag{7.3}$$

This equation is true even though the average number in the first member of the chain is considerably larger than the number in each of the subsequent members of the chain; specifically, all queueing occurs in the first node, and no queues form at any nodes beyond the first. We also note that the (Power) optimal load for this system is $\rho^* = \frac{\sqrt{2K}}{1+\sqrt{2K}}$. Here, as in both series systems with identical servers we studied above (i.e., the K $D/D/1$ systems of Section 3.3.1 and the K $M/M/1$ systems of Section 7.2.1), we see the full meaning of *"Keep the pipe just full, and no fuller"* at optimal Power, i.e., on average, as many customers are allowed in the tandem as there are nodes in the tandem (i.e., K). The *BBandwidth* is μ and the *NLDelay* to pass through the chain is K/μ, hence, $BDP = K$. Once again we have $BDP = \overline{N^*} = K$.

7.2.4. The series network of K heterogeneous M/D/1 queueing systems

Here again we consider K servers in series, the first of which is fed with Poisson traffic, and where the service time of each user is constant at each server, but in this case, they need not be identical; hence we refer to this as a heterogeneous system. Again, the first node is an M/D/1 queue and the subsequent nodes are more complicated. As in Section 7.2.2, let us label the slowest server in the chain as the "saturated" server and denote it by the subscript s and whose average service time is $1/\mu_s$. It was shown in [21–23] that this series chain has a mean response time equal to the sum of $\sum_{k=1}^{K} 1/\mu_k$ for $k \ne s$ plus the response time of a *single* M/D/1 queue with a service time equal to the maximum of the service times of the chain (i.e., with a service time $= 1/\mu_s$); thus we see that

$$T(\rho) = \frac{\rho_s}{2\mu_s(1-\rho_s)} + \sum_{k=1}^{K} 1/\mu_k \tag{7.4}$$

For this system, we show in [19] that at maximum Power, we have

$$\overline{N^*} = \sum_{k=1}^{K} \frac{\mu_s}{\mu_k} \le K \tag{7.5}$$

The *BBandwidth* is the service rate of the slowest server, μ_s, and the *NLDelay* to pass through the chain is $\sum_{k=1}^{K} 1/\mu_k$, hence, $BDP = \sum_{k=1}^{K} \mu_s/\mu_k$. In this case we have $BDP = \overline{N^*} \le K$.

Note if we compare Eqs. (7.2) and (7.5), we see that for the heterogenous cases, we have

$$\overline{N^*}_{M/D/1} \le \overline{N^*}_{M/M/1} \tag{7.6}$$

whereas for identical cases (Eqs. (7.1) and (7.3)) we have that $\overline{N^*}_{M/D/1} = \overline{N^*}_{M/M/1} = K$, and, of course, for both identical cases we have $BDP = \overline{N^*} = K$.

It is also interesting to see that whereas $\overline{N^*}$ is independent of the order of the individual nodes, the individual values for $\overline{N_k^*}$ do depend on their order (and although it is tempting from Eq. (7.5) to think that $\overline{N_k^*} = \mu_s/\mu_k$, it is not true).

In this important case, we have the same guiding intuition, *"Keep the pipe just full, and no fuller"*. Note as well that whereas queueing systems in general can operate with $\overline{N}$ at very large numbers, our result in Eq. (7.5) shows that the Power optimal average number in system does not exceed the number of servers in the system! Furthermore, since the message length does not change as it travels along an Internet connection, this $M/D/1$ series network is often used to model a TCP connection in today's Internet which we discuss below in Section 7.4.

7.3. The general network of K heterogeneous M/M/1 queueing systems

A general computer network with K nodes was modeled and analyzed by the author in [24] and used to evaluate its performance. The model used was a modification of Jackson networks [25]. The Power metric can be extended to this model as well, and it can be shown [26] that maximizing Power based on the mean

response time of the network derived in [24] leads to the consistent conclusion that, if the traffic can be so arranged, then the traffic at each node in the network should be chosen so that there should be an average of exactly one customer in each node, i.e., $\overline{N_k^*} = 1$; this also gives us that $\overline{N^*} = K$. Once again we see the deterministic rule of thumb *"Keep the pipe just full, but no fuller"* where each node may be a bottleneck. *However*, it is *not* generally true that this traffic pattern can be achieved for an arbitrary network. For the more realizable model where the traffic matrix is given (rather than designed as with [26]), then in [16], it is shown that if we scale all traffic levels so as to optimize Power for the total network, then $\overline{N^*} = \sum_{k=1}^{K} (\overline{N_k^*})^2$ and in particular, $\overline{N^*} \leq K$, where K is the number of links in the network, a result we have seen so many times.[16]

Selecting a feasible set of Power optimum flows in a general network is challenging. One approach to the problem is that presented in [17] in which is considered Pareto optimum allocations of flow using the metric of Power which balances the individual gains of a flow against the interference that flow may cause other users. We mentioned this approach earlier in Section 6.2 where we saw the need to leave sufficient server capacity to absorb the fluctuations in the traffic. Another approach for general networks as considered in [27] uses Nash Equilibrium as the greedy algorithm for flow control formulated as a multi-user noncooperative game and it is shown that there exists an equilibrium set of Power optimized (Nash) flows.

7.4. Internet congestion control

The concept of optimal Power (and thus optimal traffic in the pipe) is a natural metric for computer networks. Recently, the Google "make-TCP-fast" team [1] used the principle of optimum Power to control of the amount of in-flight data as articulated in [2] and [3] to dramatically improve congestion control in the Internet.[17] This is a TCP flow control algorithm from Google that they call *BBR* (Bottleneck Bandwidth Round-trip propagation time). They provide a fine elucidation of the behavior of a (full-duplex) TCP connection in a network by recognizing that the behavior of that connection is the same[18] as a single link with the same round-trip time and the same bottleneck bandwidth as has the connection itself. By using a deterministic model, they identify the bounds on performance in terms of *RTprop*, the minimum round-trip time to cycle the connection with no congestion, and *BtlBW*, the bottleneck bandwidth of the connection. They refer to the product *BtlBW*RTprop* as the "pipe's bandwidth-delay-product"; of course this is the same as our *BDP* (except we consider the one-way Bandwidth Delay Product, which is easily converted to theirs). They plot the round-trip time as well as the delivery rate, each versus the amount of data in flight (as shown using two coupled graphs in their Fig. 1). Their coupled plot is similar to the plot that we presented as two separate plots in Figs. 15(a) and (b). Here we choose to replot the information in their coupled graph onto a single graph of Round-Trip Time vs. Delivery Rate, (similar to what we did to create the graph in Fig. 16) as shown in Fig. 17; the straight

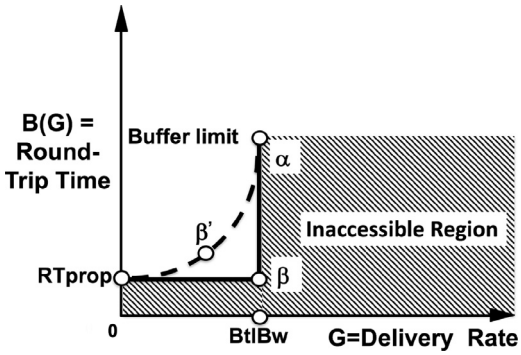


Fig. 17. Comparing *BBR* (β), *TCP* (α) and Power (β').

line behavior is a consequence of their deterministic model, but to show the qualitative performance of a stochastic connection, we have added the convex dashed line as the round-trip response time curve. Note that this is a $B(G)$ vs G plot where G=Delivery Rate (i.e., throughput) and $B(G)$ = Round-trip Time (i.e., response time). This straight line plot is very much like the plot of K $D/D/1$ systems in series shown above in Fig. 2. From our usual considerations, the optimal Power point is at the "knee" of the curve which, for the deterministic case is located at the intersection of $G = BtlBW$ and $B(G) = RTprop$, this point being denoted by β in Fig. 17, as we have used earlier to identify the location of the optimal deterministic Power point. At this maximal Power point, we get the minimal Round-trip Time with the maximum Delivery Rate. This point also satisfies *"Keep the pipe just full, and no fuller"* by sending exactly as many message units (packets in Internet terminology) as the pipe can hold without causing congestion. In [1], it is clearly stated that many of the current loss-based congestion control versions (e.g., Reno [29] and Cubic [30]) of the Internet's TCP protocol tend to put excessive flow into the pipe and cause queues to form at the bottleneck, thereby driving the flow away from the point β up to the point α which is an undesirable situation since it leads to buffer bloat and/or packet loss. *BBR*, on the other hand, recognized the value of the Power optimization approach taken in [2] and [3] which leads the system to operate at the point β. However, in reality, the flow has certain stochastic properties and so the point β may be unattainable since the performance profile may look like the dashed curve in Fig. 17 (this is an example of the performance profiles $\mu T(\rho)$ shown in Fig. 11(a)). To find the optimal operating point in this case, we can revert back to the discussion in Theorem 5.1 and seek to find the appropriate tangent to the dashed curve (or the line of minimum slope) to identify the optimal point as, for example in Fig. 17 at the point β' representing a point such as we saw in the Fig. 11(a) profiles (in which examples of β' were shown as the points $\{\beta_{C_b^2}\}$). This leaves us with the need to develop an algorithm to find this point dynamically in an operating network, an issue we discuss further below. The basic ideas of the *BBR* algorithm are to: (i) track the windowed maximum bandwidth and the minimum round-trip time on each ACK that gets returned to the source end of the link, to control the sending rate based on the model; (ii) to sequentially probe for the maximum bandwidth and minimum round-trip times to feed the model samples; (iii) to seek high throughput with small queues; (iv) to approach the maximum achievable throughput for random losses less that 15%; and (v) to maintain small bounded queues independent of the depth of the buffers.

Following the introduction of the *BBR* paper [1] in late 2016, there has followed a continual flurry of discussions, papers and active work in progress by the community on the *BBR* Development site [4] which addresses improvements to [1]. The issues revolve around improving the dynamics of the flow rate algorithm

[16] The problem of finding optimal flow to minimize response time alone (this was before the concept of Power was introduced) in these general networks was solved much earlier and led to the Flow Deviation algorithm [28].

[17] The reason that the early work of 40 years ago took so long to make its current impact is because in [31] it was shown that the mechanism presented in [2] and [3] could not be implemented in a decentralized algorithm. This delayed the application of Power until the recent work by the Google team in [1] demonstrated that the key elements of response time and bandwidth could indeed be estimated using a distributed control loop sliding window spanning approximately 10 round-trip times.

[18] As we commented in Section 7.2.4, this equivalence derives from earlier work by [21]–[23].

so as to enhance fairness among multiple flows, prevent underutilization, reduce high queueing delays and avoid packet loss. Let us review some of these contributions/discussions. In May, 2017, Huston [11] was early to blog a lucid summary of the history of TCP flow control algorithms[19] including Reno, Cubic, Vegas, and *BBR* and then pointed out some issues with the first version of *BBR* including unfairness among multiple flows (especially with different TCP versions running). In July, 2017, the Google team provided a specification [9] of their *BBR* congestion control algorithm v1.0 including an overview of the design and details of the algorithm. Around the same time, Ma, et al. published some measurements that showed a fairness issue related to competing flows with different round-trip times [5]. In October, 2017, Hock, et al. [6], looked deeper into the issue of multiple flows competing for a share of the bottleneck link, confirming that *BBR* works well with a single flow but that the behavior of multiple flows at the bottleneck presents some challenges including unfairness among competing flows along with increased delays with large buffers as well as packet loss with small buffers; in addition, they summarize a number of approaches that have been made over the years to address congestion control. A subsequent paper [7] by the same group in October, 2017 offered their delay-based congestion control algorithm, TCP LoLa, as their approach to limit queueing delay while maintaining high utililization at the bottleneck link[20] as does *BBR*, but with the ability to provide flow rate fairness independent of round-trip times of competing flows using a technique they call "fair flow balancing". The group at Google described their version v2.0 of *BBR* in November, 2017 offering their efforts in the new version to address reducing loss rate in shallow buffers, reducing queueing delay, improving fairness, improving throughput on wifi, cellular, cable networks with widespread ACK aggregation, and reducing queueing and loss in data center networks with large numbers of flows; their slides and their presentation can be found at [10] and [8]. Active progress continues to be made as reported in [4].

8. Conclusion

In this paper, we studied congestion control in networks by generalizing our work in 1978 [2] and 1979 [3] and identified the optimal amount of data ($\overline{N^*}$) to pump into a network connection. By focusing on the performance metric *Power*, we identified the Power-optimal operating point β (or, more realistically, β'). Our approach began with developing *deterministic reasoning* as a rule of thumb which was confirmed in the stochastic flow case by considerations of *Power* both of which are supported by the *Bandwidth-Delay Product BDP*.

Theorem 5.1 describes how to find the optimal power point. When applied to queueing systems (which are models of Internet traffic flow), this informs us as to how much traffic to pump into the TCP connection to achieve optimality and drive us toward the operating point β. The general rule of thumb that emerges is *"Keep the pipe just full, and no fuller"*. We constructed a new diagram, the Universal Power Profile, which allows one to see the performance of any queueing system and, from that diagram, to define the Optimal Power Trajectory which identifies the location of the optimal operating point as the input process changes in its level of stochastic behavior (and for a large class of queueing systems, the trajectory travels along the line $\overline{N^*} = 1$).

In this paper, we showed a number of cases (e.g., the important case of a series chain of K links of identical $M/D/1$ queueing systems - as in Section 7.2.3) in which $\overline{N^*}$, the optimum number to place in a pipe of length K (i.e., how much traffic to keep in flight) is equal to the length of the pipe, i.e., $\overline{N^*} = K$. We also showed that $\overline{N^*} = BDP$ which further confirms our intuitive reasoning. In other cases (e.g., the important case of a series chain of K links of heterogeneous $M/D/1$ queueing systems - as in Section 7.2.4), the optimum number to place in a pipe of length K was given by the result in Eq. (7.5), namely, $\overline{N^*} = \sum_{k=1}^{K} \mu_s/\mu_k \leq K$. Once again, it turns out that $\overline{N^*} = BDP$. In this case, the reduction from K to $\overline{N^*}$ allows the system to absorb some of the stochastic fluctuations to which we referred earlier, and accounts for the convex dashed line behavior of the response time in Fig. 17 leading to the optimal operating point β'. In all these cases, we observe that $\overline{N^*}/K \leq 1$ which makes clear that the network connection should hardly ever be driven into congestion!

The relation between $\overline{N^*}$, BDP and the pipe length K is remarkably simple and links together three key variables for our systems. We summarize this relation in the following Theorem (proofs are in Sections 3, 6.3 and 7.2):

Theorem 8.1. *For all the systems considered below*

$$\overline{N^*} = BDP \tag{8.1}$$

- *For D/D/1 and for all M/G/1 systems*

$$\overline{N^*} = 1 \tag{8.2}$$

- *For D/D/K and any series network of K identical D/D/1 systems or of K identical M/M/1 systems or of K identical M/D/1 systems*

$$\overline{N^*} = K \tag{8.3}$$

- *For any series network of K heterogeneous D/D/1 systems or of K heterogeneous M/D/1 systems*

$$\overline{N^*} = \sum_{k=1}^{K} \frac{\mu_s}{\mu_k} \leq K \tag{8.4}$$

Note carefully, however, that our work focuses on the optimal *steady state* operating point and does not address the design of an algorithm that tracks the *dynamics* of traffic that interferes with our connection. In this case we must track and adapt the allocation of bandwidth and adjust the amount of data inflight to achieve optimal performance. It is this latter, more difficult problem, that [1,5–11] and its variations seek to solve. Based on the results of Theorem 5.1 we here suggest that one could build an algorithm that continually measures the tangent of $B(G)$ (i.e, $\mu T(\rho)$) at the current operating point and then adapt the operating point ($\overline{N^*}$) so that the tangent intersects the origin of the $[B(G), G]$ axes.

Appendix A. Generalizations of the Power function

Let us consider the following simple, but useful, generalization of the definition of Power in Eq. (5.1) which we denote by $P_r(G)$:

$$P_r(G) = \frac{G^r}{B(G)} \tag{A.1}$$

The reason for introducing this generalization of the basic Power function as given earlier in Eq. (5.1) is to account for the case where one perhaps values G more than one deplores $B(G)$ (i.e., $r > 1$), or vice-versa (i.e., $r < 1$). Assuming for the moment that $B(G)$ is differentiable and convex, and following the same procedure as in Section 5.1 above, we find the condition for maximum Power to be:

$$\frac{dB(G)}{dG} = \frac{rB(G)}{G} \tag{A.2}$$

This says that the optimal G, say G^*, occurs when the slope of $B(G)$ at G^* is r times the slope of a line out of the origin to the point $[G^*, B(G^*)]$. In Section 5.1, this was easy to visualize since all

[19] A detailed survey of the development of TCP published in 2010 can be found in [32].

[20] Note how this implies using Power as a useful metric.

we had to do was to find the tangent with minimum slope; in this generalization it is not that simple. However, we do note that if we plot $B(G)$ vs G^r, then, in these axes, the slope of a line out of the origin to the point $[G^r, B(G)]$ is $B(G)/G^r$ and this is exactly $1/P_r(G)$. As usual, we wish to maximize $P_r(G)$, and so we desire to find the optimum G^* for which this slope is a minimum. Thus we see that plotting $B(G)$ vs G^r allows us to proceed as in Section 5.1 to find the optimal operating point via a simple (minimum slope) tangent to $B(G)$ on these new axes. On the other hand, since $P_r(G) \geq 0$, then raising $P_r(G)$ to any power does not change the location of its maximum. This observation offers another way to find the optimum point, G^*, namely, to plot $B^{1/r}(G)$ vs G. On these axes, the slope of a line out of the origin to the point $[G, B^{1/r}(G)]$ is $B^{1/r}(G)/G$ and this is exactly $(1/P_r(G))^{1/r}$. In this case, if we find the point G^* for which this slope is a minimum, then we have found the point of maximum $P_r(G)$. In some cases, it might well be more convenient to consider this plot to find the optimum.

We further observe that if we do not require any condition on $B(G)$ beyond $B(G) > 0$ as earlier in Section 5.1 (e.g., it need be neither differentiable nor continuous nor convex), then this construct of locating the point G^* for which the slope of a line out of the origin to point G^* is a minimum, will still identify the point of maximum Power.

One could suggest that another generalized Power might be ${}_sP(G) = \frac{G}{[B(G)]^s}$ but this will lead to no more generality than given in Eq. (A.1) since we could set $s = 1/r$, raise the full expression to the rth power (and not affect where the maximum Power is obtained since, as above, ${}_sP(G) \geq 0$) and obtain the same expression as in Eq. (A.1). Similarly, were one to suggest ${}_sP_q(G) = \frac{G^q}{[B(G)]^s}$ we find by substituting $s = q/r$ and raising the full expression to the (r/q)th power, that once again we have Eq. (A.1) which shows that we have no more generality. Thus, the generalized Power in Eq. (A.1) is quite general.[21]

Generalized Power $P_r(G)$ given in Eq. (A.1) was first introduced years ago in [3] and it was applied to $M/M/1$ and $M/G/1$ queueing systems. For $M/M/1$, the following intriguing Theorem was proven:

Theorem 9.1. *For the M/M/1 queueing system, generalized Power (as defined in Eq. (A.1)) is maximized when*

$$\overline{N^*} = r \tag{A.3}$$

$$\rho^* = \frac{r}{r+1} \tag{A.4}$$

As compared to the case $r = 1$, when $r > 1$ the increase in $\overline{N^*}$ and ρ^* as r increases is consistent with our valuing efficiency more than deploring delay in that we are now willing to load the system more heavily (higher efficiency) at the expense of more delay; that is, we are willing to *"Keep the pipe fuller"* as r increases. The converse statements apply for $r < 1$. For $M/G/1$, we do not enjoy the same simple results as we do for $M/M/1$ in Eqs. (A.3) and (A.4), but in [3] explicit expressions for $\overline{N^*}$ and ρ^* were given in his Theorems 6.2 and 6.4 respectively.

Generalized Power $P_r(G)$ for a series chain of K $M/M/1$ queueing nodes was examined in [20] and again in [16]. It was shown that

$$\overline{N^*} = Kr \quad \textit{for identical nodes} \tag{A.5}$$

$$\overline{N^*} \leq Kr \quad \textit{for heterogeneous nodes} \tag{A.6}$$

once again showing the *"Keep the pipe fuller"* intuition as r increases.

[21] Certainly one could introduce a yet more general Power function such as ${}_{f(G)}P_{h(B(G)}(G) = \frac{f(G)}{h(B(G))}$ to gain more flexibility, but we choose not to address that in this paper.

Just as was found in [16] for the general network of K heterogeneous $M/M/1$ queueing systems discussed in Section 7.3 for $r = 1$ that $\overline{N^*} \leq K$, it was also found there that when using generalized Power (arbitrary $r > 0$) that the result is $\overline{N^*} \leq Kr$. Further, for the case of identical network nodes, each of the K nodes behaves individually as in Eqs., (A.3) and (A.4), i.e., we have $\overline{N_k^*} = r$ and $\rho_k^* = \frac{r}{r+1}$; in this case, once again we have $\overline{N^*} = Kr$. Gail [16] also considers a number of other network configurations for generalized Power.

One additional generalization of Power was introduced in [3] in which we included the negative effect of blocking in queueing and network systems that endure loss of arrivals when there is limited storage space in the queue. Let us define p_B as the blocking probability that an arriving message is rejected by the system due to buffer overflow. In this case we define Power, $P_{[p_B]}(G)$, as

$$P_{[p_B]}(G) = \frac{G(1-p_B)}{B(G)} \tag{A.7}$$

This metric, $P_{[p_B]}(G)$, was applied in [3] to a number of combined loss and delay systems. In addition, in that paper, cases of pure loss were also considered; for these, the metric was defined as in Eq. (A.7) but without the denominator $B(G)$. Of course, one could add the effect of loss to the generalized power given in Eq. (A.1) and define a mixed generalized power function, which we denote as $P_{[p_B],r}(G)$, to be

$$P_{[p_B],r}(G) = \frac{G^r(1-p_B)}{B(G)} \tag{A.8}$$

Appendix B. The *ZAP* Approximation - Beyond *M/G/1*

In [33], the *ZAP* approximation was introduced as a family of response time functions to represent the performance of various systems of flow. Here we follow that approach and consider the following three-parameter expression for $T(\rho)$,

$$T(\rho) = A\frac{Z-\rho}{P-\rho} \tag{B-1}$$

where Z, A, and P are constants to be selected with the following constraints: $A > 0$, $P > 0$ and $Z > P$ or $Z < 0$. Z represents a "zero" of $T(\rho)$ whereas P represents a "pole". Since we have been considering normalized response time functions, we note that $T(0) = AZ/P$ and then form the following:

$$\frac{T(\rho)}{T(0)} = \frac{P}{Z}\frac{Z-\rho}{P-\rho} \tag{B-2}$$

Note that A has dropped out of this expression. If we interpret $T(\rho)/T(0)$ as a normalized response time, then the range of interest is for ρ is $0 \leq \rho < P$. Looking at the $M/G/1$ expression for $\mu T(\rho)$ at the beginning of Section 6.3, we see that $M/G/1$ is a special case of *ZAP* with $P = 1$ and $Z = 2/(1 - C_b^2)$.

Let us optimize Power for the *ZAP* expression given in Eq. (B.2) in the range of interest. This is easily done by showing that its second derivative with respect to ρ in this range is non-negative and is therefore convex. Then we apply the result of Theorem 5.1 to find the optimal value of G^* which in our case is ρ^* and is given by

$$\rho^* = Z - \sqrt{Z}\sqrt{Z-P} \tag{B-3}$$

It is easy to prove that $0 \leq \rho^* < P$.

To find the Power-optimized number in system, $\overline{N^*}$, as earlier we use Little's Result (Eq. (2.1)), namely $\overline{N} = \rho T(\rho)/T(0)$, and plug in ρ^* from Eq. (B.3) to obtain the interesting result that

$$\overline{N^*} = P \tag{B-4}$$

Of course, for $P = 1$ we have our earlier result showing $\overline{N^*} = 1$ but for more general normalized response times.

References

[1] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, V. Jacobson, BBR: Congestion-Based Congestion Control in ACM Queue 1420–53. Sept-Oct 2016, and in Communications of the ACM, 60 (2), pp 58–66 (Feb 2017).

[2] L. Kleinrock, On flow control in computer networks, in conference record, in: International Conference on Communications, Toronto, Ontario, June, 1978, pp. 27.2.1–27.2.5.

[3] L. Kleinrock, Power and deterministic rules of thumb for probabilistic problems in computer communications, in: International Conference on Communications, Boston, Massachusetts, June, 1979, pp. 43.1.1–43.1.10.

[4] https://groups.google.com/forum/#!forum/bbr-dev.

[5] S. Ma, J. Jiang, W. Wang, B. Li, Fairness of congestion-based congestion control: experimental evaluation and analysis in coRR abs/1706.09115, 2017.

[6] M. Hock, R. Bless, M. Zitterbart, Experimental Evaluation of BBR Congestion Control, in IEEE ICNP, October 2017, 2017.

[7] M. Hock, F. Neumeister, M. Zitterbart, R. Bless, TCP lola: Congestion Control for Low Latencies and High Throughput, in Local Computer Networks (LCN) IEEE 42nd Conference on Local Computer Networks, October, 2017.

[8] IETF100 ICCRG (November) 2017, https://www.youtube.com/watch?v=IGw5NVGBsDU&t=43m58s.

[9] N. Cardwell, Y. Cheng, S.H. Yeganeh, V. Jacobson, BBR congestion control internet congestion control research group, 2017. Internet Draft at https://tools.ietf.org/html/draft-cardwell-iccrg-bbr-congestion-control-00.

[10] N. Cardwell, Y. Cheng, C.S. Gunn, S.H. Yeganeh, I. Swett, J. Iyengar, V. Vasilev, V. Jacobson, BBR congestion control: IETF 100 update: BBR in shallow buffers, In IETF 100 ICCRG (November), 2017.

[11] G. Huston, BBR, the new kid on the TCP block, 2017. https://blog.apnic.net/2017/05/09/bbr-new-kid-tcp-block/.

[12] L. Kleinrock, Queueing Systems, wiley interscience, 1975. Vol. I: Theory

[13] L. Kleinrock, Queueing Systems, wiley interscience, 1976. Vol. II: Computer Applications.

[14] L. Kleinrock, Certain analytic results for time-shared processors, IFIP Congress Inf. Process. 68 (1968) 838–845. August

[15] A. Giessler, J. Hanle, A. Konig, E. Pade, Free buffer allocation - an investigation by simulation, Comput. Netw. 1 (3) (1978) 191–204. July

[16] H.R. Gail, On the Optimization of Computer Network Power. UCLA-CSD-830922, September 1983 (PhD Dissertation).

[17] Y. Yemini, A balance of power principle for decentralized resource sharing, Comput. Netw. J. 66 (June) (2014) 46–51.

[18] L. Kleinrock, R. Gail, An Analysis of Power for Simple Computer Network Configurations, Computer Science Department, University of California, Los Angeles, 1981. March

[19] R. Gail, L. Kleinrock, An invariant property of computer network power, Int. Conf. Commun. (1981) 63.1.1–63.1.5. June

[20] K. Bharath-Kumar, Optimum end-to-end control in computer networks, Internat. Conf. Commun. (1980) 23.3.1–23.3.6. June

[21] H.D. Friedman, Reduction methods for tandem queueing systems, Oper. Res. 13 (1965) 121–131. Jan-Feb 6

[22] B. Avi-Itzhak, A sequence of service stations with arbitrary input and regular service times, Manage. Sci. 11 (1965) 565–571. March

[23] I. Rubin, Communication networks: message path delays, IEEE Trans. Inf. Theory IT-20 (6) (1974) 738–745. November

[24] L. Kleinrock, Communication Nets; Stochastic Message Flow and Delay, McGraw-Hill Book Company, New York, 1964. (Out of Print.) Reprinted by Dover Publications, 1972 and in 2007

[25] J.R. Jackson, Networks of waiting lines, Oper. Res. 5 (1957) 518–521. August

[26] G. Rubino, On kleinrock's power metric for queueing systems, in: Proc. of the 5th International Workshop for Performance Modeling and Evaluation of Computer and Telecommunication Networks, 2011. August

[27] P. Chung, R.V. Slyke, 6, 2012, pp. 443–454.

[28] L. Fratta, M. Gerla, L. Kleinrock, The flow deviation method: an approach to store-and-forward communication network design, Networks 3 (2) (1973) 97–133. And updated in "Flow Deviation: 40 years of incremental flows for packets, waves, cars and tunnels," Computer Networks, vol. 66, pp. 18–31, (June 2014)

[29] V. Jacobson, M. Karels, Congestion avoidance and control in SIGCOMM 1988, Comput. Commun. Rev. 18 (4) (1988) 314–329.

[30] S. Ha, I. Rhee, L. Xu, CUBIC: A new TCP-friendly high-speed TCP variant, ACM SIGOPS Oper. Syst. Rev. 5 (July (5)) (2008) 64–74.

[31] J. Jaffe, Flow control power is nondecentralizable, in IEEE Transactions on Communications 29 (September(9)) (1981) 1301–1306.

[32] A. Afanasyev, N. Tilley, P. Reiher, L. Kleinrock, Host-to-host congestion control for TCP, IEEE Commun. Surv. Tutorials 12 (3) (2010) 304–342.

[33] L. Kleinrock, Performance of distributed multi-access computer-communication systems, in: Information Processing 77, Proceedings of IFIP Congress 77, Toronto, Canada, August, 1977, pp. 547–552.

Leonard Kleinrock is Distinguished Professor of Computer Science at UCLA. He is considered a father of the Internet, having developed the mathematical theory of packet networks, the technology underpinning the Internet as an MIT graduate student in 1962. His UCLA Host computer became the first node of the Internet in September 1969 from which he directed the transmission of the first Internet message. Kleinrock received the 2007 National Medal of Science, the highest honor for achievement in science bestowed by the President of the United States. Leonard Kleinrock received his Ph.D. from MIT in 1963. He has served as Professor of Computer Science at UCLA since then, serving as department Chairman from 1991 to 1995. He received a BEE degree from CCNY in 1957 and an MS degree from MIT in 1959. He has published over 250 papers and authored six books in areas including packet switching networks, packet radio networks, local area networks, broadband networks, nomadic computing, performance evaluation, intelligent agents, peer-to-peer networks and advanced network design. He has supervised the research for 48 Ph.D. students. Dr. Kleinrock is a member of the National Academy of Engineering, the American Academy of Arts and Sciences, is an IEEE fellow, an ACM fellow, an INFORMS fellow, an IEC fellow, an inaugural member of the Internet Hall of Fame, a Guggenheim fellow, and a founding member of the Computer Science and Telecommunications Board of the National Research Council. Among his many honors, he is the recipient of the National Medal of Science, the Ericsson Prize, the NAE Draper Prize, the Marconi Prize, the Dan David Prize, the Okawa Prize, the BBVA Foundation Frontiers of Knowledge Award, the IEEE Internet Millennium Award, the ORSA Lanchester Prize, the ACM SIGCOMM Award, the NEC Computer and Communications Award, the Sigma Xi Monie A. Ferst Award, the CCNY Townsend Harris Medal, the CCNY Electrical Engineering Award, the UCLA Outstanding Faculty Member Award, the UCLA Distinguished Teaching Award, the INFORMS President's Award, the ICC Prize Paper Award, the IEEE Leonard G. Abraham Prize Paper Award, the IEEE Alexander Graham Bell Medal, the SIGMOBILE 2014 Outstanding Contributions Award, and the IEEE Harry M. Goode Award.