On the Performance of Bursty Traffic Dispersion over Multiple Server Clusters

Itamar Elhanany
Dept. of Electrical & Computer Engineering
University of Tennessee
Knoxville, TN 37919
Email: itamar@ieee.org

Michael Kahane
Dept. of Electrical & Computer Engineering
Ben-Gurion University,
Beer-Sheva, Israel
Email: michaelk@bgumail.bgu.ac.il

Abstract— **This paper presents a performance analysis of multiple subnets, each representing a cluster of computing systems, which are introduced with non-uniformly distributed bursty packet arrivals. In particular, we study the case of a multi-state Markov-modulated arrival process, heterogeneously dispersed among designated queues. Cluster processing is modeled by employing a batch memoryless service discipline. The probability generating functions of the interarrival times distributions are utilized to derive closed-form expressions for each of the queue size distributions.**

Index Terms— Markov-modulated arrivals, batch processing, traffic modeling, performance analysis**.**

I. Introduction

In recent years, extensive research has been conducted on the topic of multiple-queued systems, particularly in the context of packet switching architectures [1][2][3]. Much of the work focused on obtaining performance metrics, such as delay and jitter, under diverse traffic scenarios. In this context, the work appearing in the literature pertains to a single system, albeit a large one, to which all traffic arrives and from which it departs.

An interesting scenario is one in which traffic arrives through a high-speed link (e.g. 10 Gbps) to a site which distributes this traffic among a set of queues, each forwarding packets to a cluster of computing machines. A classic application of such topologies is high-performance parallel computation, such as massively complex visualization tasks [4]. Moreover, in the context of high-speed networks, wide area networks (WAN) often receive long-haul high-speed data links from which data is demultiplexed onto several, lower speed subnets. The majority of the studies performed on such topologies consider traffic that obeys a Bernoulli (uncorrelated) process and in most cases uniformly distributed such that all subnets consume the same load intensity.

This paper presents an analysis of a networking system comprising multiple subnet queues, each reflecting on a cluster of computing systems. The traffic arriving at the queues is assumed to be non-uniformly distributed and bursty, generated using an extended Markov-modulated arrival processes. Based on the per-queue probability generating function (p.g.f.) of the interarrival times distribution it is shown that precise depiction of the queues' behavior can be obtained.

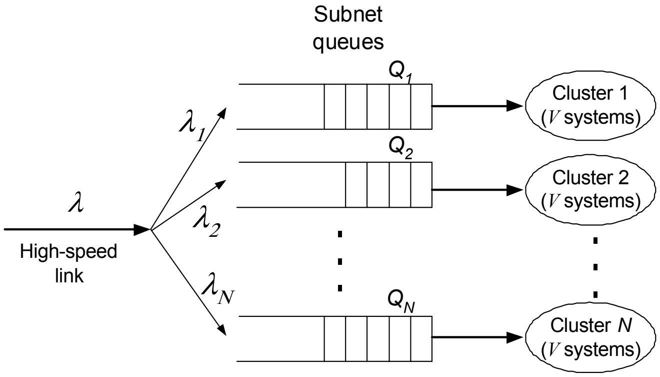

Fig. 1. The network topology model consisting of multiple subnets, each associated with a different queue, forwarding bursty traffic to designated clusters of computing servers.

The paper is organized as follows. Section II provides a description of the network model along with basic assumptions pertaining to the traffic applied. Section III describes the queueing notation and formulation used throughout the paper. In Section IV a detailed derivation of heterogeneous bursty traffic with bursty service is outlined. The simulation results are presented in section V while conclusions are drawn in Section VI.

II. Network Architecture

The network model is illustrated in figure 1. Traffic arriving from a high-speed link is assumed to be bursty and non-uniformly distributed among the N subnets. A unique queue is maintained for each of the subnets, aggregating traffic to be forwarded to a dedicated cluster of V computing servers. In our discussion, a burst is characterized by a sequence of packets destined to the same cluster (queue).

Letting λ_k denote the mean offered load traversing towards cluster k, the aggregate load is $\lambda = \sum_{k=1}^{N} \lambda_k$. Typical network platforms, particularly at the Internet backbone where ATM is commonly deployed, partition variable size packets (such as IP) into fixed sized datagrams. Processing fixed-size data units has proven both practical and easier to study. To that end, in our model all packets are assumed to be of fixed size.

III. Queueing Model and Formulation

A. Queueing Notation

We consider a discrete-time queueing system with N queues and N servers of infinite buffer capacity, in which all events occur at fixed time slot intervals. Within each time slot, at most a single arrival may occur, originating from the high-speed link. Since packets are stored at dedicated queues, at most N departures may occur within the same time slot. Let $Q_k(n)$ denote the occupancy of queue k at time slot n, such that the evolution of the queue may be described as

$$Q_k(n+1) = Q_k(n) - A_k(n) - D_k(n), \tag{1}$$

where $A_k(n)\in\{0,1\}$ and $D_k(n)\in\{0,1\}$ are the number of arrivals and departures to queue k during time slot n, respectively. In a stable queueing system, the arrival rate must converge to the departure rate. If the latter does not hold, the queue occupancy either grows to infinity or, alternatively, converges to zero. Interpreting the above balance equation for the generic case, we may equate the mean probability of arrival to the mean probability of departure by writing

$$\begin{aligned}\Pr[arrival] &= \lambda = \Pr[departure] \\ &= \Pr[Service \cap (Q>0)] = (1-\gamma_o)\mu\end{aligned} \tag{2}$$

where μ is the rate of service. From (2) we may isolate the expected stationary probability of the queue being empty, $\gamma_o = 1-\lambda/\mu$.

B. ON/OFF Arrivals with Geometric Service Times

It has been shown in the literature [5] that in a GI/Geo/1 discrete-time queueing system (general independent arrivals times and geometrically distributed service times), if f_n ($n\geq1$) is the interarrival time distribution, with a p.g.f., $F(z)$, and the service times are geometrically distributed with parameter μ, then the stationary queue size distribution as viewed by an arriving cell, π_m, will always be in the form $\pi_m = (1-\rho)\rho^m \quad m\geq 0$ where ρ is a unique root of the equation $z = F(\mu z + (1-\mu))$ that lies in the region (0,1). It has further been shown that the queue size distribution, as viewed by an outside observer, is [6]

$$\gamma_m = \begin{cases} 1-\xi & m=0 \\ \xi(1-\rho)\rho^{m-1} & m\geq 1 \end{cases} \tag{3}$$

The latter is, by definition, independent of packet arrivals. Hence utilizing (2) to derive ξ yields the first moment

$$E[Q] = \sum_{m=1}^{\infty} m\gamma_m = \frac{\xi}{(1-\rho)} = \frac{\lambda}{\mu(1-\rho)}, \tag{4}$$

which provides us with the mean queue occupancy. Employing Little's results [5], the mean waiting time is given by

$$E[W] = \frac{1}{\mu(1-\rho)}. \tag{5}$$

A late arrival model is considered, for reasons of convenience, such that within a time slot boundary a departure will always precede an arrival event. We observe the queue size at instances following the arrival phase, hence time slot boundaries are delimited by the observation instances.

Consider a discrete-time, two-state Markov chain generating arrivals modeled by an ON/OFF source which alternates between the ON and OFF states. Let the parameters p and q denote the probabilities that the Markov chain remains in states ON and OFF, respectively. An arrival is generated for each time slot that the Markov chain spends in the ON state. The result is a stream of correlated arrivals and silent periods, both of which are geometrically distributed in duration.

It can easily be shown that the parameters p and q are interchangeable with the mean arrival rate, $\lambda=(1-q)/(2-q-p)$, and mean burst size, $B=1/(1-p)$. Consequently, the offered load is identical to the steady-state portion of the time the chain spends in state ON. Recalling the notation f_n for the interarrival times distribution, the probability of two consecutive arrivals occurring is identical to the probability that following an arrival the Markov chain remains in state ON, i.e. $f_1 = p$. Similarly, f_2 is the probability that following an arrival, the chain transitions to the OFF state and then returns to the ON state. For $n>2$, it is apparent that following a transition from the ON state to the OFF state, there are $n-2$ time slots during which the chain remains in the OFF state before returning to the ON state. Accordingly, we obtain the following general expression for f_n:

$$f_n = \begin{cases} p & n=1 \\ (1-p)q^{n-2}(1-q) & n>1 \end{cases} \tag{6}$$

The corresponding p.g.f. is

$$F(z) = pz + (1-p)(1-q)\frac{z^2}{1-qz}. \tag{7}$$

Next we solve the equation $z = F(z\mu+(1-\mu))$ to find that the root in the region (0,1) is

$$\rho = \frac{(1-\mu)}{\mu}\left[\frac{1}{\mu(1-p-q)+q} - 1\right]. \tag{8}$$

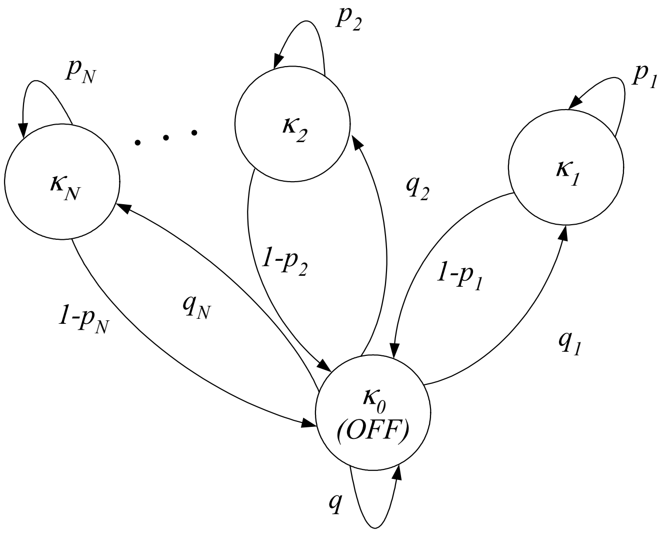

Fig. 2. Markov chain governing the generation of bursty traffic to a set of N queues. Each cluster/queue receives an offered load of λ_i.

Examining the condition $\rho<1$, which must be satisfied for stability, yields the anticipated inequality

$$\mu > (1-q)/(2-p-q) = \lambda \tag{9}$$

C. OFF Arrivals with Geometric Batch Service

Extending the above model to address the case of batch service, we next assume that V computing systems are extracting packets from each queue. The service times for each of the computing systems is independent and identically distributed with parameter μ. To that end, we utilize the $GI/Geo^{(V)}/1$ model in which the V systems may be reflected. It can be shown [5] that if $D(z)$ denotes the p.g.f. of the number of packets served in each time slot then ρ, the unique root of the equation $z=F(D(z))$ in the range (0,1), is the parameter of the stationary queue size distribution,

$$\gamma_m = (1-\rho)\rho^m \qquad m \geq 0 \tag{10}$$

where $F(z)$ is the p.g.f. of the interarrival times distribution. From (10) and Little's theorem, we may directly obtain the mean delay.

Next, we are faced with finding the p.g.f. $D(z)$ for a set of independent memoryless servers (computing systems). An aggregation of independent service events forms a binomial random variable in which 1 to V systems may service a queue at once. The p.g.f. of the binomial random variable discussed is

$$D(z) = \left[(1-\mu) + \mu z\right]^V \tag{11}$$

where μ is the independent service rate of each computing system. Accordingly, we are left with solving $z=F(D(z))$ for which the root, ρ, is the parameter of the queue size distribution. Note that under heterogeneous traffic conditions, each queue will be associated with a different arrival process and thus will result in a different queue size distribution.

IV. Heterogeneous Distribution of Busty Arrivals with Batch Service

We extend the foundations presented in section III to investigate the case of bursty arrivals that are heterogeneously distributed over several clusters. Letting N denote the number of queues, a burst is defined as a sequence of consecutive arrivals destined to the same queue. We further characterize the traffic for each queue by the portion of the offered load it receives, λ_k ($k=1,2,..N$), and a mean burst size, B_i. We construct a Markov chain corresponding to the behavior of the investigated bursty arrival process, as shown in figure 2. The chain consists of N+1 states, N of which represent arrivals going to the N queues, while the last state is the OFF state. We label the ON states as $\kappa_1,\kappa_2,\ldots\kappa_N$, and the OFF state as κ_0. The probability of remaining in the OFF state is q while the probability of remaining in each of the ON states is p_i, respectively.

To complement the latter, the probability of returning from any ON state to the OFF state is (1-p_i) while a transition from the OFF state to any of the ON states equals q_i. Thus, we can represent the Markov chain as an $(N+1)\times(N+1)$ transition probability matrix P where each element, p_{ij}, denotes the probability of transitioning from the i^{th} state to the j^{th} state among the states.

The first row of P, with the exception of its first element, consists of the probabilities of transitioning from the OFF state to each of the ON states, signifying a beginning of a burst. The first column, with the exception of its first element, contains the probability of returning from each of the ON states to the OFF state (i.e. terminating of a burst). The first element on the diagonal is the probability of remaining in the OFF state while the rest of the diagonal elements are the probabilities of remaining in the ON states. Accordingly, the examined transition probability matrix is

$$P = \begin{pmatrix} q & q_1 & q_2 & \cdots & q_N \\ 1-p_1 & p_1 & 0 & \cdots & 0 \\ 1-p_2 & 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1-p_N & 0 & 0 & \cdots & p_N \end{pmatrix} \tag{12}$$

For the Markov chain to be stable we observe that any pair (κ_0, κ_i) must satisfy $\lambda_i(1-p_i) = (1-\lambda)q_i$, hence using λ_i and B_i we can strictly obtain p_i and q_i, from which P is fully constructed.

As with the single queue case, we would like to find, for each queue, the p.g.f. of the interarrival times distribution. The latter is done by utilizing the k-step transition matrix, $P^{(k)}$, in which each element, $p_{ij}^{(k)}$, represents the probability of transitioning from the i^{th} state to the j^{th} state in precisely k-steps, with no restrictions made on passing through state j in any of the intermediate steps. In accordance with the Chapman-Kolmogorov equation [7] we have $P^{(k)} = P^k$ ($k\geq 1$), for which the p.g.f. is

$$P(z) = \sum_{n=0}^{\infty}(zP)^n = [I - zP]^{-1} \tag{13}$$

where $|z|<1$. We next define the k-step *first passage time* probability matrix [5], $F^{(k)}$, the elements of which, $f_{ij}^{(k)}$, are the probabilities of transitioning from state i to state j in *precisely* k-steps with the constraint that prior to the k^{th}-step the process has not visited state j. In other words, $f_{ij}^{(k)}$ denotes the probability of the first transition from state i to state j occurs in precisely k steps. It can be shown that $f_{ij}^{(k)}$ is

$$f_{ij}^{(k)} = \sum_{s_1\neq j}\sum_{s_2\neq j}\cdots\sum_{s_{k-1}\neq j} p_{is_1}p_{s_1s_2}\cdots p_{s_{k-2}s_{k-1}}p_{s_{k-1}j} \tag{14}$$

with $f_{ij}^{(1)} = p_{ij}$ and the s terms are the intermediate states between i and j. Since each diagonal element, $f_{ii}^{(k)}\big|_{i>1}$, is by definition the probability of k steps separating two consecutive arrivals to queue i, it is identical to the definition of the inter-arrival time distribution of the i^{th} queue. It has been shown that the following relationship exists between $p_{ii}(z)$ and $f_{ii}(z)$ [7]:

$$f_{ii}(z) = 1 - \frac{1}{p_{ii}(z)} \tag{15}$$

Accordingly, as a first step in finding F_{ii}(z), we need to find $P(z) = [I\text{-}zP]^{-1}$. Algebraic exploration of the latter yields the following generic result for $p(z)_{ii}$,

$$p(z)_{ii}\big|_{i>1} = \frac{1 - zp_{11} - \sum_{j=2,j\neq i}^{N+1}\varphi_j(z)}{\left[1 - zp_{11} - \sum_{j=2}^{N+1}\varphi_j(z)\right](1-zp_{ii})}, \tag{16}$$

where

$$\varphi_j(z) = \frac{z^2 p_{j1}p_{1j}}{1-zp_{jj}} \tag{17}$$

from which we find $F_{ii}(z)$ using (15). The latter offers the required interarrival time distribution p.g.f., for each of the N queues. To facilitate the completion of the analysis, we need to solve the equation $z = F_i(D_i(z))$ for each of the queues, where $D_i(z) = [(1-\mu_i) + \mu_i z]^V$, denoting the p.g.f. of the batch service distribution for each cluster. From (15), (16) and (17), we obtain the set of equations

$$\phi(D_i(z))[z + D_i(z)p_{ii}] + \varphi_i(D_i(z))(1-z) = 0, \tag{18}$$

where

$$\phi(z) = 1 - zp_{11} - \sum_{j=2}^{N+1}\varphi_j(z). \tag{19}$$

The roots, ρ_i, of the above equations allow us to obtain the stationary queue sizes distributions from which we establish the mean delay experienced by packets as they flow through the clusters.

V. Simulation Results

The network topology modeled consisted of 12 subnets whereby each subnet is associated with a cluster of 8 servers. Traffic is assumed to obey the multiple ON/OFF process described in section IV. Each packet is 50nsec in duration, corresponding to a 10Gbps incoming link rate. Simulations were performed for three different traffic scenarios over the same network topology, aimed at evaluating the impact of cluster size as well as traffic and service rate distribution on the delay experienced by packets.

The first considered traffic scenario was identically distributed across subnets and servers, with a rate of service for each server set to 0.9/V. Different cluster sizes were considered. As illustrated in figure 3, an increase in the number of servers does not significantly impact the mean waiting time. The next simulation pertained to the scenario in which the mean burst size for each cluster was not identical. However, Traffic load was uniformly distributed between the subnet queues and servers in each cluster. As can be observed in figure 4, the mean waiting time increases proportionally with respect to the mean burst size.

The final simulation examined the impact of non-uniformity in the rate of service between clusters. The latter aims to shed light on the affect of the service rate on the overall waiting time experienced by packets. To that end, we choose to employ a non-uniform distribution function called Zipf's law, which was proposed by G. K. Zipf [8]. The Zipf law states that the frequency of occurrence of some events (P), as a function of the rank (k), where the rank is determined by the above frequency of occurrence, is a power-law function: $P_k \sim 1/k^r$, with the exponent r typically close to unity. It was shown that many natural and human phenomena such as Web access statistics, company size and biomolecular sequences obey the Zipf law with r close to 1. The Zipf distribution is given by

$$Zipf_r(k) = \lambda_k = \frac{k^{-r}}{\sum_{j=0}^{N} j^{-r}} \tag{20}$$

where r is the model order and N is the number of elements. While r=0 represents uniform distribution, as r increases the distribution becomes more biased towards preferred elements. We utilized the Zipf distribution with r=1, and examined the first 5 (highest service rate) clusters. Due to the topology of the network, each of the aggregate cluster service rates must be larger than 1/N. The results are shown in figure 5, where it becomes evident that service rate has a great influence on the mean waiting time. All things considered, the cases studied indicate an aggregate waiting time in the order of tens of microseconds.

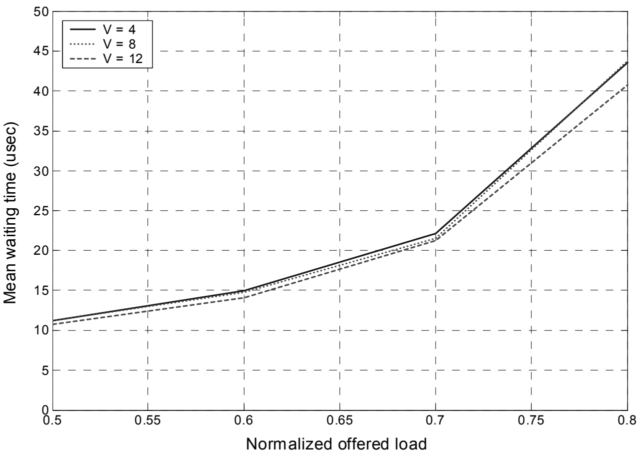

Fig. 3. Mean waiting time as a function of the offered load and the number of servers per cluster for a system with 12 subnets. Traffic is uniformly distributed across subnets and servers whereby arrivals have a mean burst size of 8 packets. The probability of service is $0.9/V$ across all clusters, where V is the number of servers in each cluster.

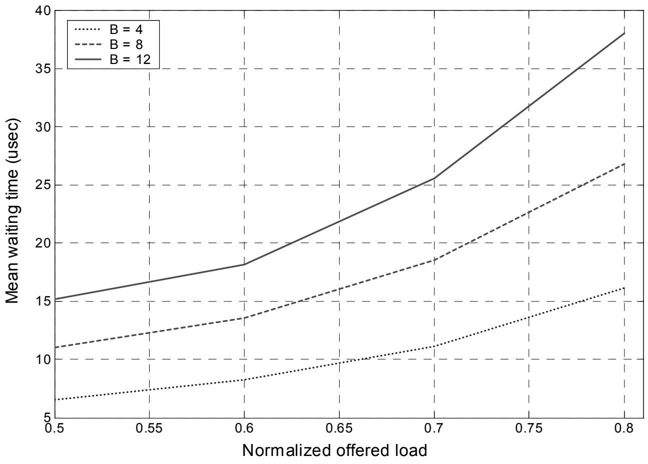

Fig. 4. Mean waiting time as a function of the offered load for a system with 12 subnets and 8 servers per cluster. Traffic is distributed uniformly across subnets and servers where arrivals have a mean burst size of 4, 8 and 12 packets. The probability of service for each server is $1/V$, where V is the number of servers in each cluster.

VI. CONCLUSION

In this paper we presented an analytical framework for evaluating the queueing behavior of multiple server clusters introduced with heterogeneous bursty traffic. We utilize the probability generating functions of the interarrival times distributions, in the context of GI/Geo$^{(x)}$/1 queueing models, to derive per-queue expressions for the queue size distribution and mean latency. Simulation results, which validate the analysis, outlined the impact of the heterogeneity of traffic on the waiting time experienced by packets traversing the network. The methodology presented in this paper may be broadened to address additional traffic scenarios and network topologies.

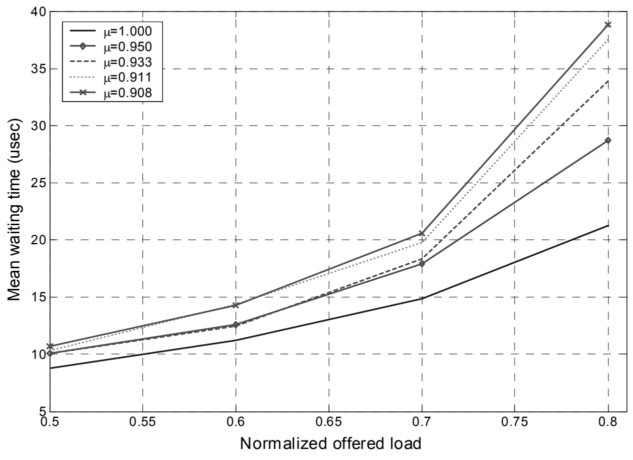

Fig. 5. Mean waiting time as a function of the offered load for a system with 12 subnets and 8 servers per cluster. Traffic is distributed uniformly across servers and subnets in which arrivals have a mean burst size of 8 packets. Each cluster is associated with a different service rate (μ) in accordance with the Zipf$_{(r=1)}$ distribution.

REFERENCES

[1] M. J. Karol, M. G. Hluchyj, and S. P. Morgan, “Input versus Output Queueing in a Space Division Switch,” *IEEE Trans. Communications*, Vol. COM-35, pp. 1347-1356, Dec. 1987

[2] I. Elhanany, D. Sadot, “DISA: A Robust Scheduling Algorithm for Scalable Crosspoint-Based Switch Fabrics,” *IEEE Journal on Selected Areas in Communications*, Vol. 21, No. 4, pp. 535-545, May 2003.

[3] N. McKeown, “The iSLIP Scheduling Algorithm for Input-Queued Switches,” *IEEE/ACM Trans. on Networking*, Vol. 7, No. 2, pp. 188-201, April 1999.

[4] C. M. Wittenbrick, “Survey of Parallel Volume Rendering Algorithms,” *Proc. Parallel and Distributed Processing Techniques and Applications (PDPTA ’98)*, July 1998, Las Vegas, NV.

[5] J. J. Hunter, *Mathematical Techniques of Applied Probability: Discrete Time Models*: *Techniques and Applications*, Vol. 2, Academic Press, 1983.

[6] M. L. Chaudhry, U. C. Gupta, J. G. C. Templeton, “On the Relations Among the Distributions at Different Epochs for Discrete-Time GI/Geom/1 Queues,” *Operations Research Letters*, Vol. 18, pp. 247-255, 1996.

[7] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "On the Implications of Zipf's Law for Web Caching", *Proc. of IEEE INFOCOM '99*, New York, March 1999.

[8] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "On the Implications of Zipf's Law for Web Caching", *Proc. of IEEE INFOCOM '99*, New York, March 1999.