Heterogeneous Bursty Traffic Dispersion over Multiple Server Clusters

Itamar Elhanany, *Senior Member, IEEE,* and Michael Kahane, *Student Member, IEEE*

***Abstract*— This letter presents performance analysis of multiple subnets, each representing a cluster of computing server nodes, that 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 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 directed towards the area of multiple-queued systems, particularly in the context of packet switching architectures [1], [2]. Much of the work focuses 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. 40 Gb/s) to a site which distributes this traffic among a set of subsystems (queues), each forwarding packets to a cluster of computing machines (nodes). A classic application of such topologies is high-performance parallel computation, such as massively complex visualization tasks [3]. Moreover, in the context of high-speed networks, wide area networks (WAN) often receive long-haul high-speed data links from which packets are 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.

In this letter we present analysis of a networking system comprising multiple subnet queues, each reflecting on a cluster of computing server 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 accurate depiction of the queues' behavior can be obtained.

Manuscript received June 22, 2004. The associate editor coordinating the review of this letter and approving it for publication was Jinwoo Choe. This work has been partially supported by the Department of Energy (DOE) under research grant DE-FG02-04ER25607.

I. Elhanany is with the Department of Electrical & Computer Engineering, University of Tennessee (e-mail: itamar@ieee.org).

M. Kahane is with the Department of Electrical & Computer Engineering, Ben-Gurion University, Beer-Sheva, Israel (e-mail: xmk@ieee.org).

Digital Object Identifier 10.1109/LCOMM.2005.03026.

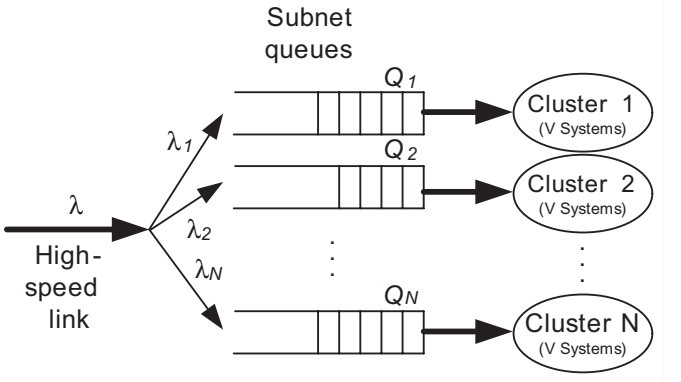

Fig. 1. The network topology model consisting of multiple subnets, each associated with a different queue, forwarding bursty traffic to designated clusters of computing machines.

II. Network Architecture

The network model is illustrated in Fig. 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 computing nodes. Each server processes packets at an independent rate of μ. 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. 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. It has been shown in the literature [4] that in a GI/Geo/1 discrete-time queueing system (general independent arrival times and geometrically distributed service times), if f_n $(n \geq 1)$ 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 $\pi_m = (1-\rho)\rho^m \ m \geq 0$ where ρ is a unique root of the equation $z = F(z\mu + (1-\mu))$ that lies in the region $(0,1)$.

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. 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:

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

The corresponding p.g.f. is

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

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

$$\rho = \frac{1-\mu}{\mu}\left[\frac{1}{\mu(1-p-q)+1} - 1\right]. \tag{3}$$

B. ON/OFF Arrivals with Geometric Batch Service

Extending the above model to address the case of batch service, we next assume that V computing nodes are extracting packets from each queue. The service times for each of the computing nodes is independent and identically distributed with parameter μ. To that end, we utilize the *GI/Geo$^{(V)}$/1* model in which the V nodes 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,

$$\mu_m = (1-\rho)\rho^m \qquad m \geq 0, \tag{4}$$

where $F(z)$ is the p.g.f. of the interarrival times distribution. Utilizing Little's theorem, we directly obtain the mean delay. We are thus left with finding the p.g.f. $D(z)$ for a set of independent memoryless servers (computing nodes). An aggregation of independent service events forms a binomial process in which 1 to V nodes may service a queue at once. The p.g.f. of the binomial process discussed is

$$D(z) = \left[(1-\mu) + \mu z\right]^V, \tag{5}$$

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.

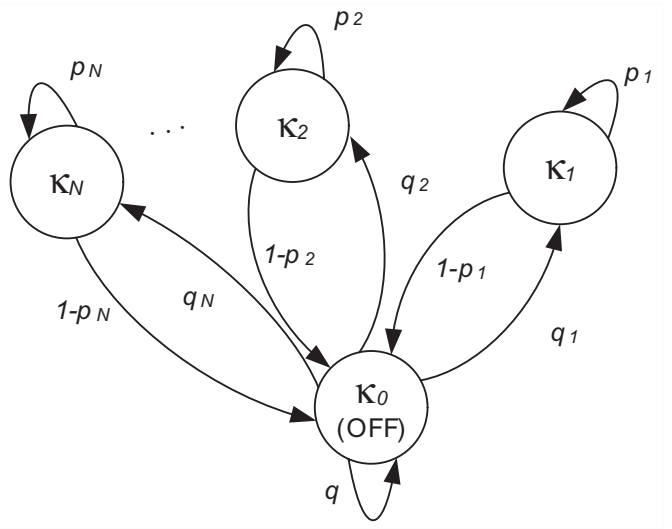

Fig. 2. Markov chain governing the generation of bursty traffic to a set of N queues. Each cluster receives an offered load of λ_i.

IV. Heterogeneous Distribution of Bursty 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. By doing so, we allow for a diverse range of non-uniformly distributed bursty arrival patterns to be generated. Such patterns better emulate real-life traffic scenarios, which tend to be bursty on different levels. 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 $(k = 1, 2, \ldots, N)$, and a mean burst size, B_k.

We construct a Markov chain corresponding to the behavior of the investigated bursty arrival process, as shown in Fig. 2. The chain consists of $N+1$ states, N of which represent arrivals going to the N queues, while the remaining 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 the probability of a transition from the OFF state to any of the ON states equals q_i. Thus, we can represent the Markov chain as a $(N+1)\times(N+1)$ transition probability matrix P where each element, p_{ij}, denotes the probability of

transitioning from the i^{th} state to the j^{th} state. For the Markov chain to be stable we observe that any pair (κ_0, κ_i) must satisfy $\lambda_i(1-p_i) = (1-\lambda)q_i$, yielding the relationships $p_i = 1 - \frac{1}{B_i}$ and $q_i = \frac{\lambda_i}{B_i(1-\lambda)}$.

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 [4] we have $P^{(k)} = P^k$ ($k \geq 1$), for which the p.g.f. is

$$P(z) = \sum_{n=0}^{\infty}(zP)^n = [I - zP]^{-1}, \tag{6}$$

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. 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 for the i^{th} queue. It has been shown that the following relationship exists between $p_{ii}(z)$, the p.g.f. diagonal elements of the transition probability matrix, and $f_{ii}(z)$, the p.g.f. diagonal elements of the first-passage times distribution, [4]:

$$f_{ii}(z) = 1 - \frac{1}{p_{ii}(z)}. \tag{7}$$

To obtain $f_{ii}(z)$, we first need to attain $P(z) = [I - zP]^{-1}$. Algebraic exploration of the latter yields the following result,

$$p_{ii}(z)|_{i>1} = \frac{1 - zp_{11} - \sum_{j=2, j\neq i}^{N+1}\varphi_j(z)}{\left[1 - zp_{11} - \sum_{j=2}^{N+1}\varphi_j(z)\right](1 - zp_{ii})}, \tag{8}$$

where

$$\varphi_j(z) = \frac{z^2 p_{j1}p_{1j}}{1 - zp_{jj}}, \tag{9}$$

from which we find $f_{ii}(z)$ using (7). The latter offers the desired interarrival time distribution p.g.f., for each of the N queues. To facilitate the completion of the analysis, we solve the equation $z = f_{ii}(D_i(z))$ 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 (7), (8) and (9), we obtain the following set of equations

$$\phi\left(D_i(z)\right)\left[z + D_i(z)p_{ii}\right] + \varphi_i\left(D_i(z)\right)(1-z) = 0, \tag{10}$$

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

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 to the clusters.

V. Conclusion

In this letter we present an analytical framework for evaluating the queueing behavior of multiple computation clusters introduced with heterogeneous bursty traffic. We utilize the p.g.f. 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. The methodology presented here may be broadened to address additional traffic scenarios and network topologies.

References

[1] I. Elhanany and D. Sadot, "DISA: A robust scheduling algorithm for scalable crosspoint-based switch fabrics," *IEEE J. Sel. Areas Commun.*, vol. 21, pp. 535-545, May 2003.

[2] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches," *IEEE/ACM Trans. Networking*, vol. 7, pp. 188-201, Apr. 1999.

[3] C. M. Wittenbrick, "Survey of parallel volume rendering algorithms," in *Proc. Parallel and Distributed Processing Techniques and Applications (PDPTA '98)*, July 1998, pp. 1329-1336.

[4] J. J. Hunter, *Mathematical Techniques of Applied Probability*. New York: Academic Press, 1983.

[5] M. L. Chaudhry, U. C. Gupta, and J. G. C. Templeton, "On the relations among the distributions at different epochs for discrete-time GI/Geom/1 queues," *Operations Research Lett.*, vol. 18, pp. 247-255, Mar. 1996.