Queueing Analysis of Markov Modulated ON/OFF Arrivals with Geometric Service Times

I. Elhanany, D. Sadot
Dept. Electrical & Computers Engineering, Ben-Gurion University, Beer-Sheva, Israel

1. INTRODUCTION

There has been extensive work focused on modeling discrete-time queueing systems with correlated arrivals [1], [2], [3], [4]. Specific attention has been drawn to queueing models with application to data switching, where bursty arrival streams represent more accurately real-life network traffic. To that end, the Markov modulated ON-OFF model has been frequently incorporated as a building block for constructing more complex traffic scenarios. Multimedia traffic, which by nature tends to be correlated on several levels, is commonly modeled by a superposition of several ON-OFF sources [5]. Contemporary traffic models include storage area networks (SAN), which are characterized by highly bursty arrival patterns due to long data blocks in conjunction with short control messages.

The majority of the work presented in the literature, however, addresses various cases of correlated arrivals with deterministic (constant) inter-service times [6]. Assuming constant interservice times provides solid framework for many systems including those deploying leaky-bucket-type service disciplines. For many other applications, such as several high-end packet scheduling architectures, a more accurate model for the interservice times is that of geometrical distribution, corresponding to a scheme in which within each time slot there is a given and independent probability of service rendered for each queue or set of queues. This probability of service, μ, is directly derived from the scheduling algorithm employed and, in general, a higher service rate is associated with more efficient scheduling.

In order to better evaluate the performance of a given switching system under bursty traffic, many packet-scheduling algorithms are examined under traffic obeying Markov modulated arrival processes [8], [9]. In some cases, significant degradation in performance metrics, such as the mean queueing latency, is observed under such bursty scenarios. Typically, performance metrics for switching systems under bursty traffic loads are attained by means of simulations. In this paper, we present an analytical tool that exploits the probability generating function of the interarrival times for obtaining steady-state queueing information from which performance metrics are analytically derived.

2. BACKGROUND: GI/Geo/1 QUEUEING SYSTEMS

We assume a discrete queueing system with a single-server and infinite buffer capacity, in which all events occur at fixed time slot intervals. Within each interval, at most a single arrival and a single service event may occur. An early arrival model is considered, for reasons of convenience, such that during a time slot an arrival event will always precede a service event. The service discipline is governed by an i.i.d. Bernoulli process, resulting in geometrically distributed interservice times. Let μ denote the homogeneous probability of service at any given time slot. The basic stability condition dictates that $\lambda/\mu < 1$ [10], where λ is the probability of arrival resulting from Bernoulli trials.
Let $Q(n)$ denote the queue occupancy at time slot n, such that

$$Q(n) = \max\big(Q(n-1) + A(n) - D(n), 0\big) \tag{1}$$

where $A(n) \in \{0,1\}$ and $D(n) \in \{0,1\}$ are the number of arrivals and departures during time slot n, respectively. For stability, the arrival rate should converge to the departure rate, such that

$$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{n=1}^{\infty} A(n)\right) = \lambda = \lim_{n\to\infty}\left(\frac{1}{n}\sum_{n=1}^{\infty} D(n)\right) \quad . \tag{2}$$

It has been shown in the literature [7] that for the *GI/Geo/1* queueing model (general inter-arrival process and geometrically distributed interservice times), if f_n ($n \geq 1$) is the inter-arrival time distribution, with a probability generating function (p.g.f.),

$$F(z) \equiv \sum_{n=0}^{\infty} f_n z^n = \sum_{n=1}^{\infty} f_n z^n \quad , \tag{3}$$

then the stationary queue size distribution, $\pi_m = \Pr\{Q = m\}$, will be in the form

$$\pi_m = \begin{cases} 1-\xi & m = 0 \\ \xi(1-\rho)\rho^{m-1} & m \geq 1 \end{cases} \tag{4}$$

where ρ ($0 < \rho < 1$) is a unique root of the equation

$$z = F\big(\mu z + (1-\mu)\big) \tag{5}$$

that lies in the region (0,1) and ξ is a constant. The queue state is examined subsequent to the arrival phase and before the service phase. One way of interpreting f_n is the steady-state probability of precisely n time slots separating two

consecutive arrivals, hence using the definition of conditional probability we have

$$f_n = \Pr\{(A(n+m)=1, A(m+j)=0, 0<j<m, m>0) \mid A(n)=1\} = \\ = \frac{\Pr\{(A(n+m)=1, A(m+j)=0, 0<j<m, m>0) \cap A(n)=1\}}{\Pr\{A(n)=1\}} \quad . \tag{6}$$

Eq. (6) states that f_n is the joint probability of the next arrival occurring after n time slots and having an arrival at the current time slot, divided by the probability of having an arrival during the current time slot. For the *Geo/Geo/1* model (with parameter λ) the probability of two successive arrivals is λ and similarly we have $f_n = \lambda(1-\lambda)^{n-1}$ which may be interpreted as the probability of having *n-1* non-arrival time slots separating two arrivals. The corresponding p.g.f. is $F(z) = \lambda\,(1-(1-\lambda))z^{-1}$. When solving for $z=F(z\mu+(1-\mu))$ we find the root

$$\rho = \frac{\lambda(1-\mu)}{\mu(1-\lambda)} \tag{7}$$

which is consistent with the well know result for *Geo*/*Geo*/*1* [11], where $\rho = \xi = 1-\pi_0$.

3. Markov Modulated ON/OFF Arrival Process

Consider a discrete-time, two-state Markov chain generating arrivals modeled by an ON-OFF source which alternates between the ON and OFF states. The parameters α and β 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 is in state ON. The outcome is a stream of correlated bursts and silent periods both of which are geometrically distributed in length. It has been shown [7] that α and β are interchangeable with the mean arrival rate, $\lambda = (1-\beta)/(2-\alpha-\beta)$ and mean burst length, $B = 1/(1-\alpha)$. Consequently, the offered load is identical to the steady-state portion of the time the chain spends in state ON. Given this arrival model and a Bernoulli service process with parameter μ, we would like to derive closed-form expressions for the steady-state queue size probabilities from which we can obtain the mean queue size and mean waiting time.

In the ON-OFF models investigate in this paper, we assume that homogeneously each state either generates an arrival or does not. To that end, $\Pr\{A(n)=1\}$ denotes the sum of probabilities of all states which generate arrivals. In the discussed model, arrivals are generated only in state ON, hence $\Pr\{A(n)=1\} = P_{ON} = 1-P_{OFF} = \lambda$. Accordingly, we can conclude from (6) that $f_1 = (P_{ON}\times\alpha/P_{ON}) = \alpha$, thereby expressing the probability that following an arrival the Markov chain will remain in state ON. 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 were $n-2$ time slots during which the chain remained in the OFF state before returning to state ON. Consequently, we obtain the following general expression for f_n:

$$f_n = \begin{cases} \alpha & n=1 \\ (1-\alpha)\beta^{n-2}(1-\beta) & n>1 \end{cases} \quad . \tag{8}$$

The corresponding probability generating function, $F(z)$, is thus

$$F(z) = \sum_{n=0}^{\infty} f_n z^n = \sum_{n=1}^{\infty} f_n z^n \\ = \alpha z + \sum_{n=2}^{\infty} \beta^{n-2}(1-\alpha)(1-\beta)z^n \\ = \alpha z + (1-\alpha)(1-\beta)\frac{z^2}{1-\beta z} \tag{9}$$

The mean inter-arrival time can be found by differentiating $F(z)$ with respect to z and letting $z=1$,

$$\left.\frac{dF(z)}{dz}\right|_{z=1} = \sum_{n=1}^{\infty} n f_n = \frac{2-\alpha-\beta}{1-\beta} \tag{10}$$

which, as expected, equals to $1/\lambda$. Next we replace z with $z\mu+(1-\mu)$ and solve the equation $z = F(z\mu+(1-\mu))$, to find that the root in the region (0,1) is

$$\rho = \frac{(1-\mu)}{\mu}\left[\frac{1}{\mu(1-\alpha-\beta)+\beta} - 1\right] \quad . \tag{11}$$

As with the *Geo/Geo/1* model, the term ρ is a function of both the service probability, μ, and the parameters of the arrival process. Note that for $\mu\to 1$ and $\mu\to 0$ we attain values of ρ corresponding to an empty queue and an infinitely growing queue, respectively. Moreover, the required condition $\rho<1$ yields that $\mu < (1-\beta)(2-\alpha-\beta) = \lambda$. In order to

complete the analysis we need to find ξ or, alternatively, π_0 (given that $\xi = 1-\pi_0$). For stability, the mean probability of arrival must equal the mean probability of departure. A departure will occur if, at a given time slot, the queue is granted service and at the same time happens to be non-empty. With that in mind, equating the mean probability of arrival to the mean probability of departure, as reflected in (2), we find that

$$\lambda = \mu(1-\pi_0) \;\Rightarrow\; \pi_0 = 1-\frac{\lambda}{\mu} \tag{12}$$

where $\lambda = (1-\beta)/(2-\alpha-\beta)$. Therefore, according to (4), $\xi = \lambda/\mu$ and we obtain a close-form expression for the stationary queue size distribution, π_m. From the result for π_m and Little's theorem [11] we derive two important performance metrics: the mean queue size and mean waiting time

$$E[Q] = \sum_{n=1}^{\infty} n\pi_n = \frac{\xi}{1-\rho} = \frac{\lambda}{\mu(1-\rho)}$$
$$E[W] = E[L]/\lambda = \frac{1}{\mu(1-\rho)} \tag{13}$$

From (11) we note that as the probability of service increases, ρ decreases, and hence the mean queue size and mean waiting time decrease, as anticipated. Fig. 1 illustrates a set of results derived from the proposed model. The x-axis denotes the mean arrival rate, λ, while the y-axis represents the mean queue waiting time (in time slots). For 64-byte data units over a 10 Gbps link, a time slot duration would be approximately 50 nsec. The results assume μ=0.9 and are shown for various mean burst sizes. We observe an increase in the waiting time that is proportional to the mean burst size. The simulation results clearly validate the presented analysis.

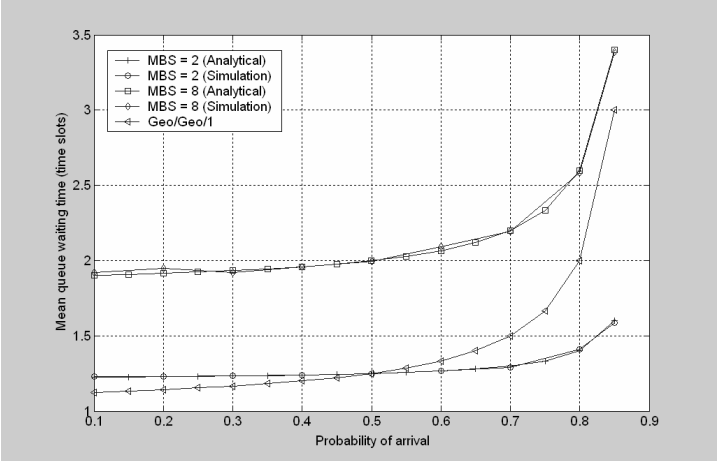

Fig. 1 – Mean queue waiting time as a function of the probability of arrival, λ, with μ = 0.9 for Geo/Geo/1 and mean burst size (MBS) of 2 and 8 time slots.

References

1. S. Wittevrongel, "Discrete-Time Buffers with Variable-Length Train Arrivals," *Electronics Letters*, Vol. 34, No. 18, pp. 1719-1721, Sep. 1998.
2. J. F. He, K. Sohraby, "A New Analysis Framework for Discrete Time Queueing Systems with General Stochastic Sources," *Proc. INFOCOM 2001*, pp. 1075-1084.
3. S. Q. Li, "Queue Response to Input Correlation Functions: Discrete Spectral Analysis," *IEEE/ACM Trans. on Networking*, Vol. 1, No. 5, pp. 552-533, May 1993.
4. A. Khamisy and M. Sidi, "Discrete-Time Priority Queueing Systems with Two-State Markov Modulated Arrival Processes," *Stochastic Models*, Vol. 8, No. 2, pp. 337-357, 1992.
5. A. La Corte, A. Lombardo, G. Schembra, "Modeling Superposition of ON-OFF Correlated Traffic Sources in Multimedia Applications," *Proc. INFOCOM 1995*, pp. 993-1000.
6. H. Bruneel, I. Wuyts, "Analysis of Discrete-Time Multiserver Queueing models with Constant Service Times," *Operations Research Letters*, Vol. 15, pp. 213-236, 1994.
7. J. J. Hunter, *Mathematical Techniques of Applied Probability: Discrete Time Models; Techniques and Applications*, Vol. 2, pp. 237-259. Academic Press, 1983.
8. N. McKeown, "The iSLIP Scheduling Algorithm for Input-Queued Switches," *IEEE Trans. on Communications*, Vol. 7, No. 2, pp. 188-201, April 1999.
9. I. Elhanany, D. Sadot, "A Contention-Free Tbit/sec Packet Switching Architecture for ATM over WDM Networks," *IEICE Trans. on Communications*, Vol. E83-B, No. 2, Feb. 2000.
10. S. M. Ross, *Introduction to Probability Models*, San Diego, Academic Press, 1997.
11. T. G. Robertazzi, *Computer Networks and Systems,* New York, Springer, 2000.