Transient analysis of the D-BMAP/G/1 queue with an application to the dimensioning of a playout buffer for VBR video[⋆]

T. Hofkens, K. Spaey, C. Blondia

University of Antwerp, Department of Mathematics and Computer Science
Performance Analysis of Telecommunication Systems Research Group
Middelheimlaan 1, BE-2020 Antwerpen - Belgium
{tom.hofkens, kathleen.spaey, chris.blondia}@ua.ac.be

Abstract. In this paper the D-BMAP/G/1 queue is considered. The goal is to derive an explicit expression for the transform of the queueing delay of the nth arriving customer, based on a transient analysis. While deriving this transform, intermediate results such as an explicit expression for the transform of the probability of having an empty system at the nth departure, are also obtained. These results are then applied to the dimensioning of a playout buffer for variable bit rate video traffic.

1 Introduction

In this paper the D-BMAP/G/1 queue is considered. This is a discrete-time single-server queue of infinite capacity with general service times. The arrival process is a discrete-time batch Markovian arrival process (D-BMAP), a quite general traffic model for discrete-time Markov sources [1,2]. In [1] and [2], a steady state analysis of queueing systems with a D-BMAP as input is performed. The goal of this paper is to derive an explicit expression for the transform of the queueing delay of the nth arriving customer of a D-MAP, based on a transient analysis. The paper is based on results presented in [3] about the transient analysis of the continuous-time BMAP/G/1 queue. While deriving the transform of the queueing delay of the nth arrival, intermediate results such as an explicit expression for the transform of the probability of having an empty system at the nth departure, are also obtained. The transform of the queueing delay of the nth arrival is used to dimension a playout buffer for a video application. The time the video application needs to keep the first packet of a video stream in the buffer before starting to playout is determined such that underflow is avoided.

The structure of the paper is as follows. Section 2 introduces the D-BMAP arrival process as well as the queueing model considered in this paper. It also summarizes the transient analysis of the queueing system and presents an expression for the transform of the queueing delay of the nth arrival in the D-MAP/G/1

[⋆] This work was carried out within the framework of the project CoDiNet sponsored by the Flemish Institute for the promotion of Scientific and Technological Research in the Industry (IWT).

queueing system. The obtained results are then applied in Section 3 to dimension a playout buffer for a video application. Finally, Section 4 concludes the paper.

2 The D-BMAP/G/1 Queue

2.1 The Discrete-Time Batch Markovian Arrival Process

A discrete-time batch Markovian arrival process (D-BMAP) is a general traffic model for discrete-time Markov sources. Consider a two-dimensional discrete-time Markov chain $\{N(k), J(k) | k \in \mathbb{N}\}$ on the state space $\mathbb{N} \times \{1, \ldots, M\}$. $N(k)$ is a counting variable representing the number of arrivals that have occurred since time 0 until time k (not including the possible arrivals at time k), and $J(k)$ is the phase of the arrival process immediately before the possible arrivals of time k occur. The transition matrix of the process has the following structure:

$$\mathbf{T} = \begin{pmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \ldots \\ 0 & \mathbf{D}_0 & \mathbf{D}_1 & \ldots \\ 0 & 0 & \mathbf{D}_0 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where the $\mathbf{D}_n$, $n \geq 0$, are $M \times M$ matrices. The matrices $\mathbf{D}_n$ govern the phase transitions of the arrival process for a batch arrival of size n. The matrix $\mathbf{D} = \sum_{n=0}^{\infty} \mathbf{D}_n$ is the transition matrix of the underlying Markov chain. Define the matrix generating function of the D-BMAP as $\mathbf{D}(z) = \sum_{n=0}^{\infty} \mathbf{D}_n z^n$, $|z| \leq 1$. Let $\boldsymbol{\pi}$ be the stationary probability vector of this Markov chain, i.e., $\boldsymbol{\pi}\mathbf{D} = \boldsymbol{\pi}, \quad \boldsymbol{\pi}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of 1's. The fundamental arrival rate λ of this process is then given by $\lambda = \boldsymbol{\pi} \left(\sum_{n=1}^{\infty} n \, \mathbf{D}_n \right) \mathbf{e}$.

More details and properties about D-BMAPs can be found in [1,2].

2.2 The Queueing Model

Consider a discrete-time single-server queue of infinite capacity with a D-BMAP $(\mathbf{D}_n)_{n \in \mathbb{N}}$ as arrival process. Call the underlying time unit of the D-BMAP a slot, where slot l is the time unit between time instants $l-1$ and l. Let the service time have an arbitrary distribution H with z-transform $h(z) = \sum_{k=1}^{\infty} H(k) \; z^k$, where $H(k)$ is the probability that the service time equals k slots.

2.3 The Embedded Process at Departures

Define $\left[\hat{\mathbf{A}}_n(m)\right]_{i,j}$ as the probability that, given a departure at time 0 leaving at least one customer in the system and the phase of the arrival process is i, the next departure occurs at time m, at that time the phase of the arrival process is j, and there have been n arrivals since time 0. Define $\left[\hat{\mathbf{B}}_n(m)\right]_{i,j}$ as the probability that, given a departure at time 0 leaving the system empty and the phase of the arrival process is i, the next departure occurs at time m, at that time the phase of the arrival process is j, and there have been $n+1$ arrivals since time 0.

Consider the queueing system at departure instants $t_0, t_1, t_2, \ldots$. Let $L(t_k)$ be the number of customers in the system at instant t_k (after the departure), and let $J(t_k)$ be the phase of the arrival process at time t_k. Then the process $\{(L(t_k), J(t_k), t_{k+1} - t_k) | k \geq 0\}$ is a semi-Markov chain with state space $\mathbb{N} \times \{1, \ldots, M\}$. The transition matrix of the semi-Markov chain is given by

$$\mathbf{Q}(k) = \begin{pmatrix} \hat{\mathbf{B}}_0(k) & \hat{\mathbf{B}}_1(k) & \hat{\mathbf{B}}_2(k) & \ldots \\ \hat{\mathbf{A}}_0(k) & \hat{\mathbf{A}}_1(k) & \hat{\mathbf{A}}_2(k) & \ldots \\ 0 & \hat{\mathbf{A}}_0(k) & \hat{\mathbf{A}}_1(k) & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad k \geq 0,$$

which shows that the system has an embedded Markov chain of M/G/1-type.

2.4 The Delay of the *n*th Arrival in the D-MAP/G/1 Queue

Let $\left[\hat{\mathbf{G}}^{(r)}(k,m)\right]_{i,j}$ be the probability that the first passage from state $(l+r, i)$ to state (l, j), with $1 \leq i, j \leq M$, $l \geq 0$ and $r \geq 1$, occurs in k transitions during m slots and (l, j) is the first state visited in the set $\{(l, l') | 1 \leq l' \leq M\}$. Define $\mathbf{G}(z) = \sum_{k=1}^{\infty} \sum_{m=k}^{\infty} \hat{\mathbf{G}}^{(1)}(k,m) y^m$ with $|y| \leq 1$. Then from [4], $\mathbf{G}(z)$ satisfies

Property 1. $\mathbf{G}(z) = z\, h(\mathbf{D}(\mathbf{G}(z)))$.

Define the n-step transition probability matrices $\mathbf{P}_{i,j}^{(n)}$ as

$$\left[\mathbf{P}_{i,j}^{(n)}\right]_{k,l} = \mathrm{P}\left[L(t_n) = j, J(t_n) = l | L(t_0) = i, J(t_0) = k\right],$$

and the transform matrix $\tilde{\mathbf{P}}_{i,j}(w) = \sum_{n=0}^{\infty} \mathbf{P}_{i,j}^{(n)} w^n$, $|w| \leq 1$. Then it is proven in [4] that

Property 2. $\tilde{\mathbf{P}}_{i,0}(w) = [\mathbf{G}(w)]^i \left[\mathbf{I} - (\mathbf{I} - \mathbf{D}_0)^{-1} [\mathbf{D}[\mathbf{G}(w)] - \mathbf{D}_0]\right]^{-1}$.

This result can then be used to derive an expression for the transform of the delay for the D-MAP/G/1 queue. A D-MAP is a D-BMAP in which no batch arrivals occur, i.e., $\mathbf{D}_n = \mathbf{0}$ for $n \geq 2$.

Let $[\mathbf{W}_n(k)]_{i,j}$ be the probability that, given a departure at time 0 and the phase of the arrival process is i, the queueing delay of the nth arrival is k slots and the phase of the arrival process immediately after the nth arrival is j. Denote its z-transform as $\mathbf{w}_n(z) = \sum_{k=0}^{\infty} \mathbf{W}_n(k) z^k$, and let $\mathbf{w}(y,z) = \sum_{n=1}^{\infty} \mathbf{w}_n(z) y^n$, $|y| \leq 1, |z| \leq 1$. Define the matrix $\mathbf{U}$ as $\mathbf{U} = (\mathbf{I} - \mathbf{D}_0)^{-1} \mathbf{D}_1$. Then the following theorem holds [4]:

Theorem 1. $\mathbf{w}(y,z) = y(\mathbf{w}_1(z) + \sum_{l=0}^{\infty} \mathbf{W}_1(l) \mathbf{D}(\mathbf{G}(y))^l \mathbf{G}(y)(\mathbf{I} - \mathbf{U}\mathbf{G}(y))^{-1}$

$$(\mathbf{U} - (z\mathbf{I} - \mathbf{D}_0)^{-1}\mathbf{D}_1)) \left(\mathbf{I} - h(z) y (z\mathbf{I} - \mathbf{D}_0)^{-1} \mathbf{D}_1\right)^{-1}.$$

$\mathbf{w}_1(z)$ gives the transform of the queueing delay of the first arrival and is set to the initial conditions of the system when the first customer arrives.

More details about the transient analysis can be found in [4].

3 Application to the dimensioning of a playout buffer

In this section the results are applied to the dimensioning of a playout buffer for a video application. This is achieved by numerically inverting the two-dimensional transform of the queueing delay of the nth arrival in Theorem 1 using [5].

Consider a scenario in which the traffic of a variable bit rate video source is sent towards a video player. Because of varying delays within the network caused by the random queueing delays in the routers in the network, the end-to-end delay between source and receiver can fluctuate from packet to packet. This phenomenon is called jitter. To compensate for the jitter, the video player uses a playout buffer. The player waits a fixed amount of time $\varDelta$ after the first packet has arrived before starting the video playout. In order to avoid underflow, it is important to carefully choose the initial delay $\varDelta$. In [4] it is derived that this is achieved if the delay of the jth packet d_j satisfies $d_j \leq d_1 + \varDelta$ $(j \geq 1)$.

Recent measurements [6] revealed that video streamers generate video traffic in bursts of multiple video frames. The duration of these bursts can vary from 1.5-2 ms for short bursts and 5-7 ms for long bursts. The silences between the bursts may be much longer than the bursts themselves, implying that the instantaneous bit rate during a burst is much higher than the average bit rate of the stream.

D-MAPs and D-BMAPs are good models for bursty traffic sources (e.g., VBR video) [1,7]. In this paper the traffic generated by a video streamer is modelled by a 4-state D-MAP which incorporates the typical characteristics of the video traffic as described above. The transition matrix $\mathbf{D}$ of the D-MAP is given by

$$\begin{pmatrix} 1-\alpha & \alpha & 0 & 0 \\ \beta & 1-\beta-\gamma & \gamma & 0 \\ 0 & 0 & 1-\delta & \delta \\ \phi & 0 & \epsilon & 1-\epsilon-\phi \end{pmatrix}, \quad \mathbf{D}_1 = \begin{pmatrix} \lambda_1(1-\alpha) & \lambda_1\alpha & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_2(1-\delta) & \lambda_2\delta \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and $\mathbf{D}_0 = \mathbf{D} - \mathbf{D}_1$. Note that this D-MAP is a kind of on/off source, with two on states (state 1 and state 3) during which packets are generated in a slot with probabilities λ_1 and λ_2 respectively, and two off states during which no packets are generated. A period in the first, respectively second on state is always followed by a period in the first, respectively second off state, while a period in an off state is always followed by a period in an on state. So this D-MAP mimics the bursty character of a video source. By carefully choosing the values of the parameters, properties such as the mean burst and silence durations, the average arrival rate and the instantaneous arrival rates during the bursts can be tuned.

In this example, the parameters are set as follows: $\alpha = 1/50$, $\beta = (1/950) - \gamma$, $\gamma = 10^{-3}$, $\delta = 3/50$, $\phi = 19\gamma/59$, $\epsilon = (3/2950) - \phi$, and $\lambda_1 = \lambda_2 = 0.6$. Assuming that the video traffic enters the network over a link of 100 Mbit/s in packets of 1500 bytes, this means that the average bit rate of the source is 1.5 Mbit/s. Packets are generated during bursts which have an average duration of respectively 6 ms and 2 ms and a standard deviation of respectively 5.94 ms and 1.94 ms, and these bursts are followed by silence periods of on average 114 ms or 118 ms respectively and a standard deviation of respectively 113.94 ms and

117.94 ms. 75% of the bursts are ‘long’ bursts, the remaining bursts are ‘short’ bursts. During a burst, packets are generated at a rate of 60 Mbit/s.

The transport of the video stream through the network and the introduction of delay and delay jitter by the network is modelled by the D-MAP/G/1 queue. The general service time distribution H follows a shifted binomial distribution $B(9, 1/3)$, i.e., if $H_B \sim B(9, 1/3)$, then $H(k) = H_B(k-1)$ for $k > 0$, where $H(k)$ is the probability that the service time of a packet equals k slots. This distribution has a mean of 4 slots and a standard deviation of $\sqrt{2}$ slots.

Using the theory developed before, values Δ_n are determined such that with probability 1-p all of the first n generated packets arrive before their scheduled playout time. It is assumed that at time 0 the D-MAP/G/1 queueing system is empty, and the phase of the arrival process is i with probability $\boldsymbol{\pi}_i$, where $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_4)$ is the stationary probability vector of the D-MAP. Figure 1 shows the complementary cumulative distributions of the queueing delay of the nth packet, for $n \in \{40, 60, 80, 100, 120, 142\}$ with $p = 0.05$ and $p = 10^{-8}$. The values for Δ_n are denoted by $\Delta_n^{(1)}$ for $p = 0.05$ and by $\Delta_n^{(2)}$ for $p = 10^{-8}$. Figure 2 shows the different values of $\Delta_n^{(1)}$ for increasing n. Both the queueing delay and Δ_n are measured in slots. The horizontal lines on Figure 1 are positioned at probability p. It is the intersection of these lines and the delay curve for packet n that gives the corresponding values for $\Delta_n^{(1)}$ and $\Delta_n^{(2)}$. Because of the bursty nature of the traffic, subsequent packets in a burst have larger delays with a higher probability. Thus the values for Δ_n need to increase with increasing n since the probability that a packet needs more time to arrive at the playout buffer than a previous packet also increases. Note the influence of the high variance of the burst length on the delay curves. For a long burst the average burst length is 6 ms, which corresponds to sending on average 30 packets. Because of the high variance of the burst length however, the actual number of packets that are sent in a burst can be much higher, hence the values for Δ_n increase with

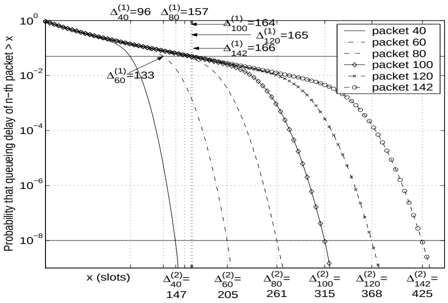

Fig. 1. The complementary cumulative distribution of the queueing delay.

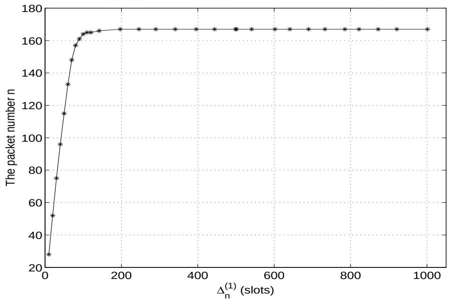

Fig. 2. The different values of $\Delta_n^{(1)}$ ($p = 0.05$).

increasing n for values of n much larger than 30. This increase will however not continue indefinitely because bursts are followed by silence periods during which the buffer of the D-MAP/G/1 system is able to empty again. Therefore, Δ_n will stabilize to a fixed Δ, as is shown in Figure 2 for $p = 0.05$.

All of the delay curves first follow a common straight line and then drop relatively fast. As n increases, the common portion of the curves becomes larger because of the increasing delays. Therefore Δ_n no longer increases when this common portion crosses the horizontal line indicating that the required condition of having a probability of $1 - p$ that packets arrive on time, is satisfied.

When $p = 10^{-8}$ a stronger demand is imposed on the system, i.e., a higher probability that packets arrive before their scheduled playout time is required. Where for $p = 0.05$, $\Delta_n^{(1)}$ stabilizes to $\Delta^{(1)} = 167$ slots $= 20.04$ ms, $\Delta^{(2)}$ will take a much larger value for $p = 10^{-8}$, as is confirmed by the different values of $\Delta_n^{(1)}$ and $\Delta_n^{(2)}$ in Figure 1.

4 Conclusion

In this paper the D-BMAP/G/1 queue was considered. For this queueing system a transient analysis was done in order to derive an explicit expression for the transform of the queueing delay of the nth arriving customer of a D-MAP. These results were then applied to dimension a playout buffer for a video application. A simple model was proposed to model the bursty nature of variable bit rate video and used as traffic source into a network. The transport of the video stream and the introduction of delay and delay jitter by the network was modelled by the D-MAP/G/1 queue. Using the developed theory, values for the time Δ_n the video application needs to keep the first packet of a video stream in the playout buffer were determined, such that with probability $1-p$ all of the first n packets arrive before their scheduled playout time in order to avoid buffer underflow.

References

1. Blondia, C., Casals, O.: Statistical multiplexing of VBR sources: A matrix-analytic approach. Performance Evaluation **1** (1992) 5–20
2. Blondia, C.: A discrete-time batch Markovian arrival process as B-ISDN traffic model. Belgian Journal of Operations Research, Statistics and Computer Science **32** (1993) 3–23 http://www.pats.ua.ac.be/chris-personal.html.
3. Lucantoni, D.: Further Transient Analysis of the BMAP/G/1 Queue. Stochastic Models **14** (1998) 461–478
4. Hofkens, T., Spaey, K., Blondia, C.: Transient analysis of the D-BMAP/G/1 queue with an application to the dimensioning of a playout buffer for VBR video (extended version). (http://www.pats.ua.ac.be/publications.html)
5. Choudhury, G.L., Lucantoni, D.M., Whitt, W.: Multidimensional transform inversion with applications to the transient M/G/1 queue. Ann. Appl. Prob. **4** (1994)
6. Balint, Z., Truyts, B.: Traffic characteristics: measurements. (Internal report of the CoDiNet project)
7. Spaey, K., Blondia, C.: Circulant matching method for multiplexing ATM traffic applied to video sources. In: Proceedings IFIP PICS'98, Lund, Sweden (1998)