Packet Loss Analysis of Load-Balancing Switch with ON/OFF Input Processes

Yury Audzevich[1], Levente Bodrog[2], Yoram Ofek[1], and Miklós Telek[2]

[1] Department of Information Engineering and Computer Science,
University of Trento, Italy,
{audzevi,ofek}@disi.unitn.it
[2] Department of Telecommunications,
Technical University of Budapest, Hungary,
{bodrog,telek}@hit.bme.hu

Abstract. Lately, the number of Internet users and, correspondingly, the amount of traversing traffic is growing extremely fast. In spite of the fact that transmission links – mostly optical fibers – have high capacity, the internet routers still remain a point of traffic bottleneck. The construction of highly scalable switches for high-speed transmission still remains a real challenge for designers. In this paper we focus our efforts on the analysis of Load-Balancing Birkhof-von Neumann switch which is lately considered to be a highly efficient distributed switch with simple control and high scalability. Due to the fact that Internet traffic represents an asynchronous traffic which supports a variety of applications, we have introduced the analysis of possible loss inside the load-balanced switch under consideration of *variable size packets* and *finite central stage buffers* previously in [1]. Although the analysis has showed some interesting features of the switch, it has exponential complexity of $O\left(N^N\right)$ which makes that model inapplicable for the switches with large number of ports, N. The main goal of this paper is to approximate the switch analysis with lower complexity, i.e., $O\left(2^N\right)$ which can be useful for evaluation of packet loss in the larger load-balanced switches.

1 Introduction

The traditional ways of packet switching are designed to connect multiple area networks (LANs, WANs, etc.) and forward asynchronous traffic between the communication links. Usually packet switches are implementing centralized control, in order to find the best possible link to forward data traffic from the source to the specific destination. Although in most of the cases these architectures are capable to provide high throughput, they have poor scalability for switches of large size. In this context, the switches with distributed control are more attractive with the advantage of their scalability due to the fact that each stage is making its own calculations for packet forwarding.

In this paper we examine the Load-Balanced switch (LB switch) [2,3], which is considered to be a particular case of two-stage switch. The first stage of the

switch is balancing the arriving traffic to the intermediate inputs of the second switch, which is in fact an input buffer switch with deterministic control (see Figure 1). Since all the interconnections inside are deterministic and periodic, the switch has a simple distributed control and can be highly scalable. Among the first significant results shown in [2] and [3] was the fact that under certain assumptions the switch can achieve high throughput (up to 100%) and low packet traversing delay. However these results were obtained under consideration that all the packets have equal length, traffic is admissible and central stage buffers are infinite. Even under these strong assumptions some important issues of packets mis-sequencing were investigated in detail in [4–8]. It is important to mention that some of the architectures to resolve packets mis-sequencing require extra control, introducing different overheads (communication and computational), that basically increases the control complexity of the LB switch. However, keeping correct sequence of packets through the system avoids unnecessary retransmissions of packets in the network protocol layer.

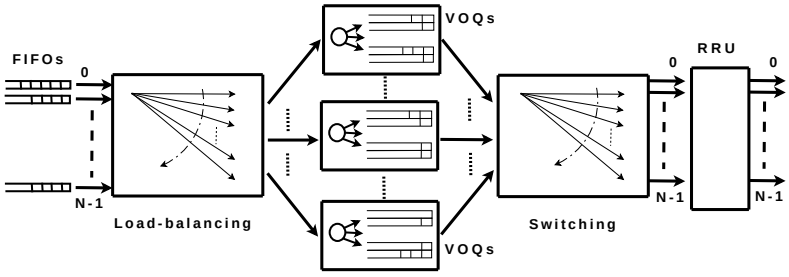


Fig. 1. The load-balanced switch considered for the analysis

Taking into account the fact that some of the assumptions mentioned in [2, 3, 8] are not practical, in [9] and [1] we examined the behavior of the LB switch with finite size central stage buffers. Under these considerations, the LB switch can experience a packet loss due to congestion. The first simulation results on this issue were presented in [10] and detailed mathematical analysis in [9]. However, the analysis in [9] was done only for fixed size packets (cells), and there were not taken into account variable size packets (multiple number of cells going to the same destination). It is considered that most of the internet switches are operating on the cell-based level (to increase buffer utilization), that means that arriving variable size packets are segmented at the inputs and reassembled at the outputs. The issue of possible cell and correspondingly a packet loss inside the switch, can introduce some significant posterior problems to the LB switch reassembly part [11]. That is why in [1] we presented the analysis of a packet loss experienced by the switch operating with variable size packets and finite central stage buffers.

We assumed Markovian behavior to be able to use numerically efficient algorithms to solve the model. This means geometrically distributed packet lengths

and interarrival times, which allows us to capture the mean of these distributions. Real internet traffic shows different packet size distributions [12] and one can fit more parameters using other, more complex Markovian structures like discrete Phase Type (DPH) distributions or discrete Markovian arrival processes (DMAPs). The number of fitted parameters can be increased at an arbitrary level, but it would greatly increase the complexity of the model as well and that would also hide the main contribution of our approach.

In spite of the same assumption in [1] the complexity of that model resulted unresolvable high Markov chains even in case of very small switches ($N \geq 4$). The main goal of this paper is to introduce the approximate model of the initial analysis – with complexity $O\left(2^N\right)$ – in order to make the evaluation of packet loss probabilities feasible for larger number of ports – at least in terms of the exact analysis provided in [1].

As the present model is still exponentially complex with regard to the number of ports – $O\left(2^N\right)$ – we have kept on with the research and introduced the model of complexity $O\left(N\right)$ in [13]. However, in the least complex model [13], we assumed stochastically identical input processes. As a consequence the reader should take into mind that the present model is less complex than that of [1] and more complex than that of [13], but it takes into consideration inhomogeneous input processes. These features of the three models are summarized in Table 1

citation	[1]	this paper	[13]
submission order	1st	2nd	3rd
complexity	$O\left(N^N\right)$	$O\left(2^N\right)$	$O\left(N\right)$
homogeneous inputs	✗	✗	✓

Table 1. The authors' recent work on the LB switch topic

The rest of the paper is organized as follows. We summarize the LB switch's operation principles and main assumptions in Section 2. Next, in Section 3 we introduce the "ON/OFF" model of the system. In Section 4 we verify the result by comparing it with initial analytical model as well as with simulation results. Finally, Section 5 concludes the paper.

2 The main assumptions and operating principles

Let denote $N \times N$ the LB switch with both N input and output ports. The single-stage buffering LB switch is equipped with First-In-First-Out (FIFO) buffers in the inputs, N sets of N Virtual Output Queues (VOQs) in the central stage and re-sequencing and reassembly units (RRU) in the output (see the illustration in Figure 1). In the kth set of VOQs there is one VOQ (VOQ_{kj}) dedicated to store cells directed to output j. Hereinafter the term VOQ_k with a single index denotes the kth set of VOQs and the term VOQ_{kj} with the pair of indices denotes the specific VOQ stores cells directed to output $j \quad j, k \in [0, N-1]$. As

it is out of the scope of this paper and it does not affect the modeled parts of the switch, the implementation of the RRU is not discussed in this paper, but it can be taken from the ones proposed in research, e.g., in [11]. In this analysis there is no feedback link between the switch stages and each stage is operating independently. After segmentation of an incoming packet, the cells are load-balanced between the central stage VOQs according to the final destination [2]. The interconnections between stages are made by means of crossbar switches without buffers inside (contrary to [14]). The crossbar switches are indicated as "Load-balancing" and "Switching" in Figure 1. In the tth time slot – the transmission time of one cell – the interconnection pattern is the periodic round-robin sequence according to the rules

$$\begin{aligned} k &= (i+t) \mod N \\ j &= (k+t) \mod N, \end{aligned} \tag{1}$$

where i denotes the ordinal number of the input port, j the output port and k the set of VOQs, $i, j, k \in [0, N-1]$ which implies the periodic behavior of the system. This N cell transmission time long period – hereinafter referred to as time period – will be the time unit of the discrete time Markov chain (DTMC) modeling the VOQ. As all the stages are synchronized, the transmission of cells is possible from all inputs simultaneously during a time slot [8].

If a single cell is lost in the central stage, there is no possibility to drop all the remaining cells of this "broken" packet from VOQs without sophisticated centralized controller (which is not the case in this paper). Such packets will waste the capacity of the central stage buffers, will increase the possibility of further packet loss and definitely will make impossible packets reassembling operation [11].

In a time slot, first, the VOQs are connected to the outputs and then the inputs to the VOQs. This order of interconnections inhibits a cell from traverse the switch in a single time slot. The transmission rate inside the switch is fixed and it is the service time of a cell. The mean service rate of the switch assumed to be greater than the mean arrival rate of the variable size packets – the switch is not overloaded.

The arrival pattern consists of packets with random distributed number of cells idle periods inbetween in time slots. The details of these distributions are

packet length geometric distributed with probability mass function (PMF) $\Pr(X = i) = p(1-p)^{i-1} \quad \forall i = 1, 2, \ldots$ and

idle period length geometric distributed with PMF $\Pr(Y = i) = q(1-q)^{i} \quad \forall i = 0, 1, \ldots$

The geometric distribution of the packets arrive from input i to output j have the parameter p_{ij} and the idle periods between packet arrivals at input i have the parameter q_i.

The destinations of the packets can be set via matrix $\mathbf{T}$ whose ijth element (t_{ij}) gives the probability that if a packet arrives to input i is directed to output j. The rowsum of $\mathbf{T}$ thus equals to $\mathbf{h}$ an appropriate size column vector of ones.

Moreover, as shown in our analytical results, the packet loss probability of the specific VOQ strongly depends on the specific traversing path of the traffic inside the switch (i.e., input, VOQ and output), which is an interesting phenomenon described in Section 2.1 for the interconnection pattern applied.

2.1 Properties of the different paths

An important finding of our analysis of the LB switch is that there are differences between the loss probabilities of paths traversing the switch. Here path means the triple, $\{i, j, k\}$, containing the ordinal number of the input, the output and the VOQ respectively.

Using the interconnection pattern policy given in (1) the time difference between the service of the VOQ and the arrival to it can be expressed as

$$d = (2k - i - j) \mod N. \tag{2}$$

d also expresses the number of inputs that have the right to send a packet to VOQ_{kj} before input i in the same time period.

A particular VOQ is served once a time period. It is also true that in a time period all the inputs have the right – in a particular order determined by (2) – to send cells to the VOQ. In case of "almost full" buffer the higher the d value is the higher the probability that there are enough inputs that can fill up the buffer, i.e., make the cell of the observed input to be lost. According to this observation we introduce the notation type-d for paths with value d.

For example, using the above introduced notation, we can say that the type-0 paths cannot have cell loss. Its short explanation is that even if the buffer is full the cell in the head of the queue is served and thus there is always a free position in the tail accordingly which is used by the type-0 path to push its cell into it. This makes impossible the cell loss at a type-0 path.

3 The ON/OFF model of the 3×3 switch

In this section we give the approximate model of a VOQ of the 3×3 LB switch. Compared to the exact analysis in [1] the approximation is that we model the input process, i.e., the arrival process of the VOQ, with a two state – ON/OFF – model. By this the state space of the model of the same VOQ can be reduced compared to the exact model of [1] where a size (N) dependent full characterization of the input process is given. Once we have the model of an input the complete model of the chosen VOQ is given in the same way as in case of the full characterization in [1]. Indeed the ON/OFF based model of the LB switch differs from the complete characterization in the DTMCs describing the input processes.

As we described in Section 2.1 it is relevant which type of path is considered. Here we describe a type-2 path lead through the 3×3 switch – as also done in [1] in case of the full characterization. For example it is path $\{1, 0, 0\}$ but we will also investigate all types of path later in Section 4.

3.1 Model of an input

In this section we will introduce the approximate – ON/OFF – input model of path $\{1, 0, 0\}$ of the 3×3 switch.

The ON/OFF model of the first input is derived from its complete characterization depicted in Figure 2 using the notations introduced for the input processes in Section 2. According to the geometric assumptions for the packet length and idle period length this is a DTMC having four states, $1\,id$ corresponds to the idle period, and the other three states corresponds to packet arrival from input 1 to either output 0 (state 10) or output 1 (state 11) or output 2 (state 12). The exact state transition probability matrix describing the behavior of input 1 is

$$\mathbf{P}_1^{\mathcal{C}} = \begin{pmatrix} (1-p_{10})+p_{10}q_1t_{10} & p_{10}q_1t_{11} & p_{10}q_1t_{12} & p_{10}(1-q_1) \\ p_{11}q_1t_{10} & (1-p_{11})+p_{11}q_1t_{11} & p_{11}q_1t_{12} & p_{11}(1-q_1) \\ p_{12}q_1t_{10} & p_{12}q_1t_{11} & (1-p_{12})+p_{12}q_1t_{12} & p_{12}(1-q_1) \\ q_1t_{10} & q_1t_{11} & q_1t_{12} & 1-q_1 \end{pmatrix}. \tag{3}$$

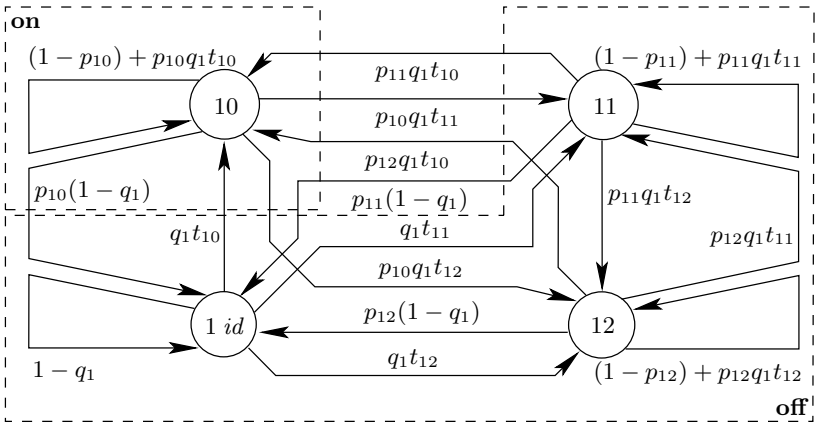


Fig. 2. The graph of the DTMC fully characterizing the first input of the 3×3 switch

In terms of path $\{1, 0, 0\}$ the states of the DTMC modeling input 1 can be divided into two subsets

on this is a one-element subset containing state 10 in which there are cell arrivals from input 1 to output 0 and

off the other states in which there is no arrival from input 1 to output 0

which is also indicated in Figure 2. Using this division we create the two state ON/OFF model of the input processes. Hereinafter lowercase bold **on** and **off**

denotes these two subsets and uppercase ON and OFF the two states of the newly derived DTMC model of the inputs.

In the following sections the detailed description of the ON and OFF states are given based on the aforementioned division.

OFF properties The OFF state is used to approximate the set of **off** states. Its properties are determined based on the absorbing time of a discrete phase type (DPH) distribution given in Figure 3 with transient states identical to the **off** states and absorbing state given as the **on** state. Its initial distribution then given as the renormalization of the zeroth row of $\mathbf{P}_1^{\mathcal{C}}$ in (3) without its zeroth element

$$\boldsymbol{\beta}_1 = \left(\frac{q_1 t_{11}}{q_1 t_{11} + q_1 t_{12} + (1-q_1)} \;\; \frac{q_1 t_{12}}{q_1 t_{11} + q_1 t_{12} + (1-q_1)} \;\; \frac{1-q_1}{q_1 t_{11} + q_1 t_{12} + (1-q_1)} \right). \tag{4}$$

$\mathbf{B}_1$, the transition probability matrix of the transient states, is the $N \times N$ matrix given as $\mathbf{P}_1^{\mathcal{C}}$ without its zeroth row and zeroth column

$$\mathbf{B}_1 = \begin{pmatrix} (1-p_{11}) + p_{11} q_1 t_{11} & p_{11} q_1 t_{12} & p_{11}\,(1-q_1) \\ p_{12} q_1 t_{11} & (1-p_{12}) + p_{12} q_1 t_{12} & p_{12}\,(1-q_1) \\ q_1 t_{11} & q_1 t_{12} & 1-q_1 \end{pmatrix}. \tag{5}$$

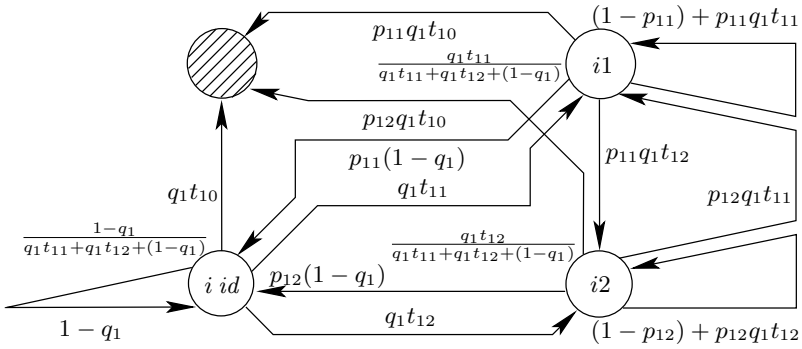


Fig. 3. The graph of the DPH substitution of the **off** states in terms of the pair input 1 - output 0

The mean absorbing time of this DPH is

$$\mu_1 = \boldsymbol{\beta}_1 \left(\mathbf{I} - \mathbf{B}_1 \right)^{-1} \mathbf{h}, \tag{6}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{h}$ is the column vector of ones of appropriate size.

We set the sojourn probability of the state OFF to $1 - \frac{1}{\mu_1}$ which sets the mean sojourn time to μ_1. Then the state transition probability from OFF to ON is $\frac{1}{\mu_1}$.

ON properties In case of ON the sojourn probability remain the same as in the complete characterization, i.e. in case of output 0 the upper left element of $\mathbf{P}_1^{\mathcal{C}}$ in (3). The state transition probability from ON to OFF is the summation of the remaining elements of the zeroth row of $\mathbf{P}_1^{\mathcal{C}}$ which is 1 minus the sojourn probability.

Summation of the ON/OFF DTMC Here we summarize all the properties of the ON/OFF DTMC by giving its graph for the general path $\{i, j, k\}$ in Figure 4 together with its state transition probability matrix

$$\mathbf{P}_i = \begin{pmatrix} \left(\mathbf{P}_i^{\mathcal{C}}\right)_{jj} & 1 - \left(\mathbf{P}_i^{\mathcal{C}}\right)_{jj} \\ \frac{1}{\mu_i} & 1 - \frac{1}{\mu_i} \end{pmatrix} = \begin{pmatrix} (1 - p_{ij}) + p_{ij}q_i t_{ij} & p_{ij} - p_{ij}q_i t_{ij} \\ \frac{1}{\mu_i} & 1 - \frac{1}{\mu_i} \end{pmatrix}, \quad (7)$$

where $(*)_{ij}$ denotes the ijth element of a matrix.

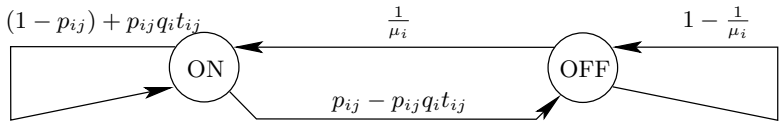


Fig. 4. The graph of the ON/OFF DTMC describing the pair input i - output j

3.2 The cell level model

Up to now we have introduced the differences between the full model of [1] and the ON/OFF model of the input processes. From now on we recall the remaining part of building the model of the VOQ using the ON/OFF model of each input. Here we keep on with building the model of the VOQ of path $\{1, 0, 0\}$.

First of all we give the cell level model of VOQ_{00} which is a quasi birth-deathlike (QBD-like) DTMC where the level represents the queue length and the phase is the combined state $\left(0, 1, \ldots, 2^N - 1\right)$ of the inputs.

According to the periodic operation of the switch mentioned in Section 2 the time unit of the QBD-like model is N time slots – the time period of the operation of the switch.

Since the DTMC given in Figure 4 and in (7) gives the behavior of the input process in a single time slot we raise all of them to the Nth = 3rd power to have the model of the input processes in a time period.

Then the joint behavior of the input processes – for all inputs ($i = 0, 1, 2$) – gives the phase process of the QBD-like model which is the Kronecker product of their 3rd power as

$$\boldsymbol{\mathcal{P}} = \mathbf{P}_0^3 \otimes \mathbf{P}_1^3 \otimes \mathbf{P}_2^3. \quad (8)$$

The number of arrivals to the observed VOQ is determined as the sum of the arrivals from each input, but we cannot forget that each input can transmit a

cell into the VOQ in its dedicated time slot. This is determined by the interconnection pattern given in (1), i.e. input 0 sends cell to VOQ_{00} in the 1st time slot of a time period, input 1 sends in the 3rd time slot of a period and input 2 sends in the 2nd time slot of a time period. Here we note that the ordinal number of the dedicated time slot equals to $d+1$ for each input i in any path. According to this we replace the 1st, the 3rd and the 2nd factor of the powers of $\mathbf{P}_0^3$, $\mathbf{P}_1^3$ and $\mathbf{P}_2^3$ respectively in (8) to

$$\mathbf{P}_i = \mathbf{A}_i + \mathbf{K}_i \quad \forall i \in [0, N-1]\,, \tag{9}$$

in which the first term corresponds to arrival from input i and the second term corresponds to the case when there is no arrival from input i. The substitution is then

$$\boldsymbol{\mathcal{P}} = \mathbf{P}_0^3 \otimes \mathbf{P}_1^3 \otimes \mathbf{P}_2^3 = (\mathbf{A}_0 + \mathbf{K}_0)\,\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\,(\mathbf{A}_1 + \mathbf{K}_1) \otimes \mathbf{P}_2\,(\mathbf{A}_2 + \mathbf{K}_2)\,\mathbf{P}_2 \tag{10}$$

based on the d values of the inputs calculated as given in (2). Expanding this expression and collecting the terms according to 0, 1, 2 and 3 arrivals we get

$$\begin{aligned}
\boldsymbol{\mathcal{P}} = {} & \underbrace{\mathbf{K}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\mathbf{K}_1 \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2}_{\text{no arrivals} - \mathbf{B}} + \underbrace{\mathbf{A}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\mathbf{K}_1 \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2}_{\text{1 arrival} - \mathbf{L}} + \\
& + \underbrace{\mathbf{K}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\mathbf{A}_1 \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2 + \mathbf{K}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\mathbf{K}_1 \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2}_{\text{1 arrival} - \mathbf{L}} + \\
& + \underbrace{\mathbf{K}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\mathbf{A}_1 \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2 + \mathbf{A}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\mathbf{K}_1 \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2}_{\text{2 arrivals} - \mathbf{F}_1} + \\
& + \underbrace{\mathbf{A}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\mathbf{A}_1 \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2}_{\text{2 arrivals} - \mathbf{F}_1} + \underbrace{\mathbf{A}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^2\mathbf{A}_1 \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2}_{\text{3 arrivals} - \mathbf{F}_2} = \\
& = \mathbf{B} + \mathbf{L} + \mathbf{F}_1 + \mathbf{F}_2,
\end{aligned} \tag{11}$$

where we have also indicated the level transition based decomposition, $\boldsymbol{\mathcal{P}} = \mathbf{B} + \mathbf{L} + \mathbf{F}_1 + \mathbf{F}_2$, of such a QBD-like model.

Using these level transition matrices the state transition probability matrix has the QBD-like structure

$$\mathbf{P} = \begin{pmatrix}
\mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \dots \\
\mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots \\
\dots & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 \\
\dots & 0 & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}_1' \\
\dots & 0 & 0 & 0 & \mathbf{B} & \mathbf{L}'
\end{pmatrix}, \tag{12}$$

where $\mathbf{F}_1' = \mathbf{F}_1 + \mathbf{F}_2$ and $\mathbf{L}' = \mathbf{L} + \mathbf{F}_1 + \mathbf{F}_2$.

The building of this kind of QBD-like DTMC for $N = 3$ is given in Algorithm 1.

The steady state solution of this QBD-like model is the solution of the linear equation system

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}, \qquad \boldsymbol{\pi}\mathbf{h} = 1. \tag{13}$$

Algorithm 1 Building the QBD-like model of a VOQ

INPUT: $\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2$ from (7)
OUTPUT: $\mathbf{P}$ the QBD-like model similar to (12)

1: **for** $i = 0$ to 2 **do**
2: compute $\mathbf{A}_i, \mathbf{K}_i$ as given in (9)
3: calculate d for the ith input as given in (2)
4: replace the $(d+1)$st factor of $\mathbf{P}_i^3$ in (8) with $\mathbf{A}_i + \mathbf{K}_i$ as given in (10)
5: **end for**
6: expand the resulting expression for $\boldsymbol{\mathcal{P}}$ and
7: identify the level transition matrices $\mathbf{B}, \mathbf{L}, \mathbf{F}_1, \mathbf{F}_2$ as given in (11)
8: build $\mathbf{P}$ as in (12)
9: **return** $\mathbf{P}$

3.3 The packet level model

With the geometric assumption for the packet length, given in Section 2, the life cycle of a packet in the observed path can be modeled by a transient DTMC in which the two absorbing states corresponds to the two possible ending of a packet transmission – the successful transmission (ST) of the packet or its lost (PL), as given in Figure 5. In this section we present this transient DTMC with its state transition probability matrix and initial distribution.

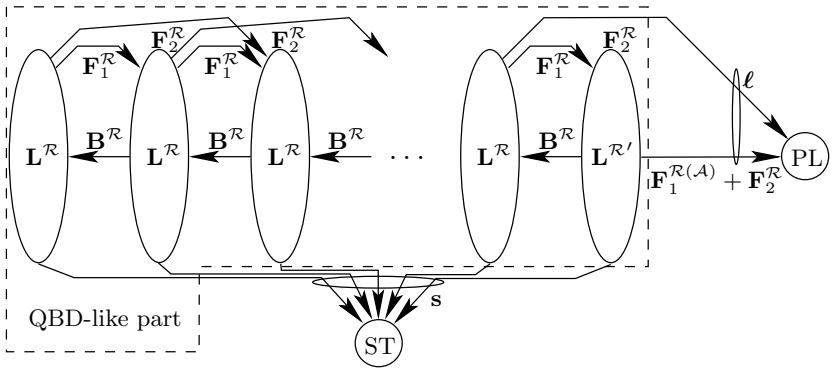


Fig. 5. The transient DTMC modelling the VOQ during the life cycle of a packet

The state transition probability matrix of the transient part The transient DTMC is mainly built in the same way as the QBD-like model of the VOQ on the cell level in Section 3.2. The exceptions are

- the state transitions responsible for packet completion in the observed path are removed (its DTMC is given in Figure 6(a)) and
- the cell losses in case of "nearly" full buffer are considered.

The removal of the state transitions is explained by the introduction of absorbing state ST. Indeed this transient DTMC move to state ST when the transmission of a packet is completed. Then according to these modifications the state transition probability matrix of the modified DTMC of input 1, with such state transitions removed (see Figure 6(a)), is

$$\mathbf{P}_1^{\mathcal{R}} = \begin{pmatrix} 1-p_{10} & 0 \\ 1-\frac{1}{\mu_i} & \frac{1}{\mu_i} \end{pmatrix}, \tag{14}$$

where superscript $\mathcal{R}$ refers to the DTMC with absorbing states PL and ST, in Figure 5. The DTMC of the other two inputs remain as in (7).

The state transition probability matrix of the QBD-like part of the DTMC in Figure 5 is $\mathbf{P}^{\mathcal{R}}$. It is determined by Algorithm 1 with input parameters $\mathbf{P}_0, \mathbf{P}_1^{\mathcal{R}}, \mathbf{P}_2$ with one exception in line 8.

Having the level transition matrices $\left(\mathbf{B}^{\mathcal{R}}, \mathbf{L}^{\mathcal{R}}, \mathbf{F}_1^{\mathcal{R}}, \mathbf{F}_2^{\mathcal{R}}\right)$ the construction of the QBD-like structure and the state transition vector to state PL are

$$\mathbf{P}^{\mathcal{R}} = \begin{pmatrix} \mathbf{B}^{\mathcal{R}} & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} & 0 & \dots \\ \mathbf{B}^{\mathcal{R}} & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & \mathbf{B}^{\mathcal{R}} & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} \\ \dots & 0 & 0 & \mathbf{B}^{\mathcal{R}} & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} \\ \dots & 0 & 0 & 0 & \mathbf{B}^{\mathcal{R}} & \mathbf{L}^{\mathcal{R}\prime} \end{pmatrix}, \qquad \boldsymbol{\ell} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{F}_2^{\mathcal{R}}\mathbf{h} \\ \left(\mathbf{F}_1^{\mathcal{R}(\mathcal{A})} + \mathbf{F}_2^{\mathcal{R}}\right)\mathbf{h} \end{pmatrix}, \tag{15}$$

where $\mathbf{L}^{\mathcal{R}\prime} = \mathbf{L}^{\mathcal{R}} + \mathbf{F}_1^{\mathcal{R}(\mathcal{K})}$. Here the forward level transition matrix $\left(\mathbf{F}_1^{\mathcal{R}}\right)$ is decomposed into two parts both of them corresponds to two cell arrivals. In the first case one of the cells arrives from input 1

$$\mathbf{F}_1^{\mathcal{R}(\mathcal{A})} = \mathbf{K}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^{\mathcal{R}\,2}\mathbf{A}_1^{\mathcal{R}} \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2 + \mathbf{A}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^{\mathcal{R}\,2}\mathbf{A}_1^{\mathcal{R}} \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2$$

and in the second case none of them arrive from input 1

$$\mathbf{F}_1^{\mathcal{R}(\mathcal{K})} = \mathbf{A}_0\mathbf{P}_0^2 \otimes \mathbf{P}_1^{\mathcal{R}\,2}\mathbf{K}_1^{\mathcal{R}} \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2.$$

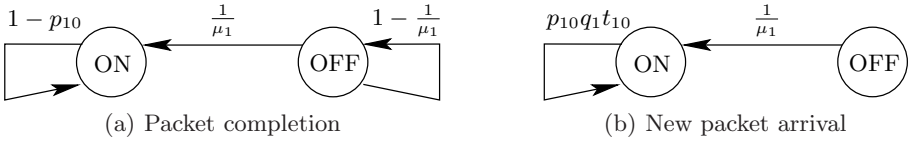
Due to this there is cell loss and accordingly packet loss in the observed path $\{1,0,0\}$ if at the beginning of the time period either

- there is one free position in VOQ_{00} and there are cell arrivals from all three inputs $\left(\mathbf{F}_2^{\mathcal{R}}\right)$ or
- the buffer is full and there are cell arrivals either
  - from all the three inputs $\left(\mathbf{F}_2^{\mathcal{R}}\right)$ or
  - there is two arrivals from which one arrives from input 1 $\left(\mathbf{F}_1^{\mathcal{R}(\mathcal{A})}\right)$.

Accordingly, if the buffer is full at the beginning of the time period and there is two arrival, but none of them from input 1 the DTMC stays in the last level $\left(\mathbf{F}_1^{\mathcal{R}(\mathcal{K})}\right)$.

Finally according to Figure 6(a) and (14) and using the notations of Figure 5 and (15) the state transition probability vector to the absorbing state ST is

$$\mathbf{s} = \mathbf{h} - \left(\mathbf{P}^{\mathcal{R}}\mathbf{h} - \boldsymbol{\ell}\right). \tag{16}$$



(a) Packet completion (b) New packet arrival

Fig. 6. The modified graphs of the ON/OFF DTMC describing input 1

The initial distribution of the transient DTMC The initial distribution of $\mathbf{P}^{\mathcal{R}}$ in (15) is determined as the state of the system right after the arrival of an incoming customer. In this section we determine the probability distribution of the system at this time instance, right after a new packet arrival.

A new packet arrives at input 1 according to the state transitions depicted in Figure 6(b). Its state transition probability matrix is

$$\mathbf{P}_1^{\mathcal{N}} = \begin{pmatrix} p_{10}q_1t_{10} & 0 \\ \frac{1}{\mu_1} & 0 \end{pmatrix}, \tag{17}$$

where superscript $\mathcal{N}$ refers to the DTMC according to new packet arrival.

Here we build a QBD-like model also using Algorithm 1 with input parameters $\mathbf{P}_0, \mathbf{P}_1^{\mathcal{N}}, \mathbf{P}_2$ with an exception in line 4 which also affects lines 6 and 7.

Instead of replacing the third factor of $\mathbf{P}_1^{\mathcal{N}3}$ (remind that $d = 2$ for $i = 1$) we give the state transition probability matrix of input 1 in a time period as

$$\mathbf{P}_1^3 - \left(\mathbf{P}_1 - \mathbf{P}_1^{\mathcal{N}}\right)^3. \tag{18}$$

It expresses the behavior of input 1 at new packet arrivals in a three time slots long time period. According to Algorithm 1 we expand (18), replace the third factors of its terms and simplify it we get

$$\mathbf{P}_1^3 - \left(\mathbf{P}_1 - \mathbf{P}_1^{\mathcal{N}}\right)^3 = \underbrace{\mathbf{P}_1^2\mathbf{A}_1^{\mathcal{N}} + \left(\mathbf{P}_1\mathbf{P}_1^{\mathcal{N}} + \mathbf{P}_1^{\mathcal{N}}\left(\mathbf{P}_1 - \mathbf{P}_1^{\mathcal{N}}\right)\right)\left(\mathbf{A}_1 - \mathbf{A}_1^{\mathcal{N}}\right)}_{\boldsymbol{\mathcal{A}}_1^{\mathcal{N}}} +$$

$$+ \underbrace{\mathbf{P}_1^2\mathbf{K}_1^{\mathcal{N}} + \left(\mathbf{P}_1\mathbf{P}_1^{\mathcal{N}} + \mathbf{P}_1^{\mathcal{N}}\left(\mathbf{P}_1 - \mathbf{P}_1^{\mathcal{N}}\right)\right)\left(\mathbf{K}_1 - \mathbf{K}_1^{\mathcal{N}}\right)}_{\boldsymbol{\mathcal{K}}_1^{\mathcal{N}}} = \boldsymbol{\mathcal{A}}_1^{\mathcal{N}} + \boldsymbol{\mathcal{K}}_1^{\mathcal{N}}, \tag{19}$$

where we have also indicated the two terms according to cell arrival $\left(\boldsymbol{\mathcal{A}}_1^{\mathcal{N}}\right)$ into VOQ_{00} and no cell arrival $\left(\boldsymbol{\mathcal{K}}_1^{\mathcal{N}}\right)$ in the time period. These two matrices are

used to replace the whole middle operand of (8) in line 6 of Algorithm 1 as

$$\begin{aligned}
\boldsymbol{\mathcal{P}}^{\mathcal{N}} &= \left(\mathbf{A}_0 + \mathbf{K}_0\right)\mathbf{P}_0^2 \otimes \left(\boldsymbol{\mathcal{A}}_1^{\mathcal{N}} + \boldsymbol{\mathcal{K}}_1^{\mathcal{N}}\right) \otimes \mathbf{P}_2\left(\mathbf{A}_2 + \mathbf{K}_2\right)\mathbf{P}_2 = \\
&= \underbrace{\mathbf{K}_0\mathbf{P}_0^2 \otimes \boldsymbol{\mathcal{K}}_1^{\mathcal{N}} \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2}_{\text{no arrivals} - \mathbf{B}^{\mathcal{N}}} + \underbrace{\mathbf{A}_0\mathbf{P}_0^2 \otimes \boldsymbol{\mathcal{K}}_1^{\mathcal{N}} \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2}_{\text{1 arrival} - \mathbf{L}^{\mathcal{N}}} + \\
&+ \underbrace{\mathbf{K}_0\mathbf{P}_0^2 \otimes \boldsymbol{\mathcal{A}}_1^{\mathcal{N}} \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2 + \mathbf{K}_0\mathbf{P}_0^2 \otimes \boldsymbol{\mathcal{K}}_1^{\mathcal{N}} \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2}_{\text{1 arrival} - \mathbf{L}^{\mathcal{N}}} + \\
&+ \underbrace{\mathbf{K}_0\mathbf{P}_0^2 \otimes \boldsymbol{\mathcal{A}}_1^{\mathcal{N}} \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2 + \mathbf{A}_0\mathbf{P}_0^2 \otimes \boldsymbol{\mathcal{K}}_1^{\mathcal{N}} \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2}_{\text{2 arrivals} - \mathbf{F}_1^{\mathcal{N}}} + \\
&+ \underbrace{\mathbf{A}_0\mathbf{P}_0^2 \otimes \boldsymbol{\mathcal{A}}_1^{\mathcal{N}} \otimes \mathbf{P}_2\mathbf{K}_2\mathbf{P}_2}_{\text{2 arrivals} - \mathbf{F}_1^{\mathcal{N}}} + \underbrace{\mathbf{A}_0\mathbf{P}_0^2 \otimes \boldsymbol{\mathcal{A}}_1^{\mathcal{N}} \otimes \mathbf{P}_2\mathbf{A}_2\mathbf{P}_2}_{\text{3 arrivals} - \mathbf{F}_2^{\mathcal{N}}} = \\
&= \mathbf{B}^{\mathcal{N}} + \mathbf{L}^{\mathcal{N}} + \mathbf{F}_1^{\mathcal{N}} + \mathbf{F}_2^{\mathcal{N}}.
\end{aligned} \tag{20}$$

Here we also indicated the level transition matrices used in line 8 of Algorithm 1 to build the state transition probability matrix $\left(\mathbf{P}^{\mathcal{N}}\right)$ in the same way as in (12).

Using (13) and $\mathbf{P}^{\mathcal{N}}$ the initial distribution of the DTMC in Figure 5 is

$$\boldsymbol{\pi}^{\mathcal{N}} = \frac{\boldsymbol{\pi}\mathbf{P}^{\mathcal{N}}}{\boldsymbol{\pi}\mathbf{P}^{\mathcal{N}}\mathbf{h}}. \tag{21}$$

The packet loss of the system Using (15), (16) and (21) the packet loss probability (p_ℓ) is given as the probability of absorbing in state PL and the probability of successful packet transmission (p_s) as absorbing in state ST

$$p_\ell = \boldsymbol{\pi}^{\mathcal{N}}\left(\mathbf{I} - \mathbf{P}^{\mathcal{R}}\right)^{-1}\boldsymbol{\ell} \qquad p_s = \boldsymbol{\pi}^{\mathcal{N}}\left(\mathbf{I} - \mathbf{P}^{\mathcal{R}}\right)^{-1}\mathbf{s} = 1 - p_\ell. \tag{22}$$

4 Computation study

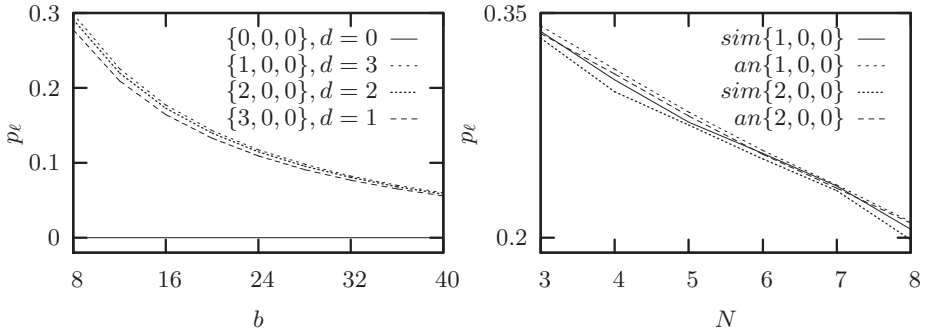
In this section we present the comparative study of the analysis with ON/OFF model and the simulation results using the memoryless (geometric) assumptions and the notations introduced in Section 2. We executed two studies with different sets of parameters given in Table 2 representing a set of considered parameters in detail, instead of just the ON and the OFF parameters (the model is derived from the detailed parameters). Although the independent variables are discrete we used continuous plots to improve visibility of Figure 7.

Study 1 Figure 7(a) plots the packet loss probability of different types of paths through VOQ_{00} versus the buffer size. The loss of a single queue is decreasing with increase of the buffer size, which is obvious with increase of system capacity. Here the dependence of packet loss on the chosen paths is also shown. The set of parameters of study 1 is given in the left hand side of Table 2. The experimental results proof the validity of our assumptions. In particular, in Figure 7(a) we

study 1		study 2	
variable	value	variable	value
N	4	N	3,. . . ,8
p_{ij}	$\frac{1}{20}$ (av. 20 cells)	p_{ij}	$\frac{1}{50}$ (av. 50 cells)
q_i	$\frac{1}{3}$ (av. 2 cells)	q_i	$\frac{1}{6}$ (av. 5 cells)
t_{ij}	$\frac{1}{N}$	t_{ij}	$\frac{1}{N}$
b	$8, \ldots, 40$	b	20

Table 2. The main parameters of the computation

show, that the queue does not experience any loss for the type-0 path $\{0, 0, 0\}$, and, as expected, the higher the d value is the higher the loss probability of the path is. It is also shown in Figure 7(a) that the higher the buffer size (b) is the less the difference between the loss values for types.



(a) The packet loss probability versus the buffer size (study 1)

(b) The packet loss probability versus the switch size (study 2)

Fig. 7. Numerical results for the packet loss analysis of LB switches

Study 2 Due to lower analysis complexity in comparison with [1], the packet loss of a single queue can be evaluated for larger switches – than those ones in [1]. Figure 7(b) plots the packet loss of the queue if the switch size is increasing – up to the solvable highest size of this model. The detailed set of parameters used in Study 2 is shown in the right hand side of Table 2. We present packet loss only for those two traffic path ($\{1, 0, 0\}$ and $\{2, 0, 0\}$) which exist for all considered switch sizes. As it is shown on the plot, with the increase of the switch size, the packet loss decreases. As the average packet size and idle period size keeps to be the same, the increase in number of ports increases the number of queues at the central stage and consequently the buffering capacity for the same set of parameters. Correspondingly, the higher is the LB switch buffering capacity the lower packet loss is experienced.

5 Conclusions

In this paper we have presented an approximate analytical model for evaluation of loss probabilities inside the load-balanced switch with finite buffers and variable length packets. In comparison to the analysis presented in [1], we reduced the complexity of the model from $O\left(N^N\right)$ to $O\left(2^N\right)$. Although the complexity has remained exponential, the new approach has extended the range of packet/cell loss probability evaluation for switches with $N \geq 4$ and large VOQ sizes. Since the load-balanced switch is the architecture of choice when N is large, our next step is the presentation of approximated analysis with linear complexity in [13]. This will enable us to remove restrictions on the port/buffer size of the switch in order to calculate the systems important characteristics (like different kinds of loss, delays, average buffers occupancy).

References

1. Audzevich, Y., Bodrog, L., Telek, M., Ofek, Y.: Variable Size Packets Analysis in Load-balanced Switch with Finite Buffers. Technical report, Technical University of Budapest (2009)
2. Chang, C., Lee, D., Jou, Y.: Load-Balanced Birkhoff-von Neumann switches, Part I: One-Stage Buffering. Computer Communications **25** (2002) 611–622
3. Chang, C., Lee, D., Lien, C.: Load-Balanced Birkhoff-von Neumann switches, Part II: Multi-Stage Buffering. Computer Communications **25** (2002) 623–634
4. Yu, C., Chang, C., Lee, D.: CR Switch: A Load-Balanced Switch with Contention and Reservation. In: IEEE INFOCOM'07, Anchorage, USA (May 2007) 1361–1369
5. Chang, C., Lee, D., Shih, Y.: Mailbox Switch: A Scalable Two-Stage Switch Architecture for Conflict Resolution of Ordered Packets. In: Proceedings of IEEE INFOCOM'04. Volume 3., Hong Kong (March 2004) 1995–2006
6. Shen, Y. andJiang, S.a.S., Chao, H.: Byte-Focal: A Practical Load-Balanced Switch. In: IEEE HPSR'05, Hong Kong (May 2005) 6–12
7. Lin, B., Keslassy, I.: The Concurrent Matching Switch Architecture. In: IEEE INFOCOM'06, Barcelona, Spain (April 2006) 1–12
8. Keslassy, I., Chuang, S., Yu, K., Miller, D., Horowitz, M., Solgaad, O., McKeown, N.: Scaling Internet Routers Using Optics. In: ACM SIGCOMM'03, Karlsruhe, Germany (2003)
9. Audzevich, Y., Ofek, Y., Telek, M., Yener, B.: Analysis of load-balanced switch with finite buffers. In: IEEE Globecom'08, New Orleans, LA, USA (2008) 1–6
10. Tu, C., Chang, C., Lee, D., Chiu, C.: Design a Simple and High Performance Switch Using a Two-Stage Architecture. In: IEEE GLOBECOM'05. Volume 2., St. Louis, MO, USA (November 2005) 6–11
11. Turner, J.: Resilient Cell Resequencing in Terabit Routers. Technical report, Washington University, Department of Computer Science (June 2003)
12. Thompson, K., Miller, G., Wilder, R.: Wide-area Internet traffic patterns and characteristics. IEEE Network **11** (Nov.-Dec. 1997) 10–23
13. Audzevich, Y., Bodrog, L., Telek, M., Ofek, Y.: Scalable model for packet loss analysis of load-balancing switches with identical input processes. In: ASMTA '09. LNCS, Madrid, Spain (June 2009) 1–15
14. Turner, J.: Strong Performance Guarantees for Asynchronous Crossbar Schedulers. In: IEEE INFOCOM '06, Barcelona, Spain (April 2006) 1–11