Scalable model for packet loss analysis of load-balancing switches with identical 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. In this paper we present a scalable approximate model for packet loss analysis in load-balancing Birkhof-von Neumann switch with finite buffers and variable length packets assumption. We also present a numerical method to solve the model for large switches (up to the size ~ 30) equipped with large buffers (up to the buffer size ~ 1000). With regards to previously introduced models the main contribution of our model is its scalability in terms of the switch size as its computational complexity is linear with the number of ports. Contrary to previous models we assumed homogeneous input processes in this paper.

1 Introduction

Internet is a huge asynchronous mesh network which is composed of several sub-networks connected to each other through switches. As the traffic over the network and the number of links grow exponentially the transmitting media can be easily adopted using optical fibre. Although the links could provide high throughput, the switches are not always capable to fulfill both the throughput growth and increasing number of connections. Some solutions with high throughput and centralized control exist but they are poorly scalable.

Recently in [1, 2] the authors introduced a promising and highly scalable solution, a two-stage switching architecture called load-balancing (LB) Birkhof-von Neumann switch.

[1, 2] shows initial investigations on the switch under some strong assumptions (infinite buffers, traffic admissibility, equal size packets in the system). On the contrary [3] used realistic scenarios and carried out a simulation based throughput analysis of the LB switch with finite buffers. [4] pointed out that in cell-based (packets of the same size) LB switch a loss can occur because of buffer overflow. This latter paper also presents mathematical analysis for cell loss probability evaluation. Besides, going further in this approach, [5, 6] give analytical results for both finite buffers and variable size packets.

Whereas in [5] the authors present the full characterization of a realistic scenario, with finite buffers and variable size packets, a less complex approximating model is given in [6]. In spite of the complexity $O\left(2^N\right)$ we still need a fast procedure to solve the model. The aim of this paper is to present an analysis with fast solution procedure. However a restrictive assumption is applied, i.e., the model assumes identical stochastic processes on all the inputs.

We will demonstrate that, besides this assumption, the newly introduced model captures the two most important performance measures. We analyzed the packet loss – as the switch is equipped with finite buffers – and gave an estimate of the mean packet waiting time. The first parameter affects the Quality of Service (QoS) characteristics of data transfers (using TCP). The second parameter has high influence on real time traffic, e.g., speech (using UDP) over the network [7, 8].

We also introduced a folding algorithm-based numerical method to solve the model of switches with large buffers.

The rest of the paper is organized as follows. In Section 2 we give the modeling assumptions and the basic principles of the switch. Section 3 presents the model into detail. The numerical solution method is introduced in Section 4 and Section 5 verifies the model. Finally Section 6 concludes the paper.

2 Basic principles and main assumptions of the switching mechanism

2.1 Basic principles

The LB switch is considered to be a two-stage switching architecture. The first stage uniformly distributes the arriving traffic to the central stage, which is an input buffer of the second switch (see Figure 1). Its scalability lies in the distributed, distinct and deterministic control between different switching stages.

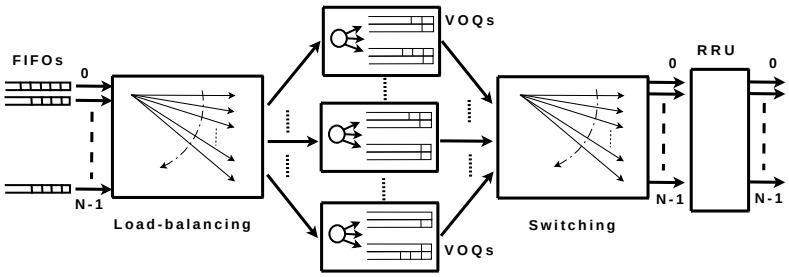


Fig. 1. The load-balancing switch considered for the analysis

To improve the buffer utilization the arriving packets are segmented into cells of equal size. The basic operating time unit is the service time of a cell – hereinafter referred to as *time slot.*

The arrival rate of the cells at the input ports are assumed to be identical to the service rate of the inputs, i.e., there is no cell loss here. The service rate of each output is assumed to be greater than the arrival rate of cell, i.e., the switch is not overloaded.

In the followings N denotes the size of the switch, i.e., the number of input and output ports. The central stage consists of N sets of N virtual output queues (VOQs). In each set there is one buffer dedicated to every output. Hereinafter VOQ_k denotes the kth set of VOQs. The cells directed to output j are put in the jth VOQ out of the kth set – hereinafter denoted as VOQ_{kj}.

During the t_1st time slot input i is connected to VOQ_k according to round-robin (RR) interconnection policy

$$k = i + t_1 \mod N \quad i, k \in [0, N-1] \tag{1}$$

by means of crossbar switches without buffers inside (contrary to [9]). The actual cell arriving from input i and directed to output j is put into VOQ_{kj} if a free position is available and it is dropped otherwise. In our assumption a cell can only be lost due to buffer overflow as the VOQs are finite. The VOQs are served according to the FIFO policy.

As the packets are segmented into cells we consider a packet to be lost, when at least one of its cells is lost, i.e., packet loss can occur also according to the finite VOQs.

VOQ_{kj} is served in the t_2nd time slot, when VOQ_k is connected to output j by means of crossbar switches operating according to RR policy

$$j = k + t_2 \mod N \quad j, k \in [0, N-1]\,. \tag{2}$$

As both crossbars applies RR interconnection policy with the same modulus (N), the LB switch itself has periodic behavior of period N time slots – hereinafter referred to as *time period.*

Finally a packet is reassembled in the re-sequencing and reassembly unit (RRU) at the output (see Figure 1), like in [10], and sent to the external link.

2.2 Modeling assumptions

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 passing through the switch in a single time slot.

During our work we assumed Markovian behavior of system, more precisely geometric distributed random variables. On the one hand we can fit one parameter of the observed distributions, but on the other hand we can use the sophisticated and numerically efficient algorithms to solve discrete time Markov chains (DTMCs). In order to increase the precision of the analysis, one can expand the number of fitted parameters to an arbitrary level by using more complex Markovian structures like discrete time Phase-Type (DPH) distributions or discrete time Markovian Arrival Processes (DMAPs). Yet such a choice

would increase the complexity of the model and, to a certain extent, shift the focus of the paper.

According to the Markovian assumption the packet length (X) distribution (in cells) of the arrival process is geometric distributed with probability mass function (PMF)

$$\Pr(X = i) = \hat{p}\,(1 - \hat{p})^{i-1} \quad i = 1, 2, \ldots \tag{3}$$

The length of the idle periods between packets (Y) are also geometric distributed (in time slots) with PMF

$$\Pr(Y = i) = \hat{q}\,(1 - \hat{q})^{i} \quad i = 0, 1, \ldots \tag{4}$$

The parameters are the same for all inputs according to the identical input process assumption, which makes us possible to introduce a compact approximate model of the LB switch. The packets arriving at an arbitrary input are spread uniformly between the outputs, i.e., the probability of sending a packet to a particular output is

$$\hat{t} = \frac{1}{N}. \tag{5}$$

Previous works [5,6] introduced the differences between traffic paths traversing the switch. This phenomenon is recalled in the next section.

2.3 On the different paths

It is shown in [5] and [6] that the cell loss probability and accordingly the packet loss probability depend on the path through which it traverses the switch. Where path means a triple, denoted as $\{i, j, k\}$, containing the ordinal number of the input, the output and the VOQ respectively.

Mainly the difference of the paths comes from the time difference between the service of a VOQ and the arrival to it. Using (1) and (2) the time difference between the service of a VOQ and the arrival to it is expressed as

$$t_2 - t_1 = d = 2k - i - j \mod N, \tag{6}$$

which also gives the number of inputs that have the right to send cells to VOQ_{kj} before input i in the same time period. d is then directly proportional to the loss probability of a path, i.e., the higher the d value is the higher the loss probability of that path is. Here we use the term *loss probability of a path* to emphasize the difference between the cell loss probabilities depending on the triple $\{i, j, k\}$ or equivalently depending on d.

Based on (6) we recall the term type-d path introduced in [6] for a given path with characteristic value d.

2.4 On the definition of the different loss probabilities

Cell loss probability is simply the ratio between the number of cells which were dropped from the observed VOQ versus the total number of cells sent through the VOQ.

The calculation of a *packet loss* inside the VOQ is not that trivial since cells belonging to the same packet can be spread to different VOQs. Accordingly we consider the packet loss in terms of a specific VOQ, i.e., the packet is considered to be lost, if at least one of its cells is lost in the observed VOQ.

The cell loss probability and accordingly the packet loss probability depends on the path as it is described in the previous section, it is referred as *loss probability of a path.*

Since our main interest is the packet loss probability, the precise way how it is calculated is given in Section 3.4 and the different loss probabilities for all the paths are considered in the numerical study in Section 5.

3 The model

In this section we give the detailed model of VOQ_{00} as part of path $\{1, 0, 0\}$ of the 3×3 switch. This is a type-2 path of that particular switch, but the detailed analysis of all 3 types of paths will be given in Section 3.4.

3.1 The model of the input processes

The parameters of the identical input process are

$\hat{p}$ the parameter of the geometric distributed packet length (3) in cells,
$\hat{q}$ the parameter of the geometric distributed idle period length (4) in time slots and
$\hat{t} = \frac{1}{N}$ the probability of choosing a specific output for a given packet (5).

Based on the geometric assumption we can build the DTMC model, fully characterizing any of the identical inputs, with state transition probability matrix

$$\mathbf{P}^{\mathcal{C}} = \begin{pmatrix} (1-\hat{p}) + \hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & \hat{p}\,(1-\hat{q}) \\ \hat{p}\hat{q}\hat{t} & (1-\hat{p}) + \hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & \hat{p}\,(1-\hat{q}) \\ \hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & (1-\hat{p}) + \hat{p}\hat{q}\hat{t} & \hat{p}\,(1-\hat{q}) \\ \hat{q}\hat{t} & \hat{q}\hat{t} & \hat{q}\hat{t} & 1-\hat{q} \end{pmatrix} \tag{7}$$

and graph given in Figure 2, where the state identifiers are the following

j corresponds to cell arrival from the input to output $j \quad j = 0, 1, 2$
id corresponds to the idle period of the input.

According to the observed output, i.e., output 0, the states of the DTMC in Figure 2 are divided into two subsets, a one element subset and all the others, hereinafter denoted as **on** and **off**, respectively. Their meaning are

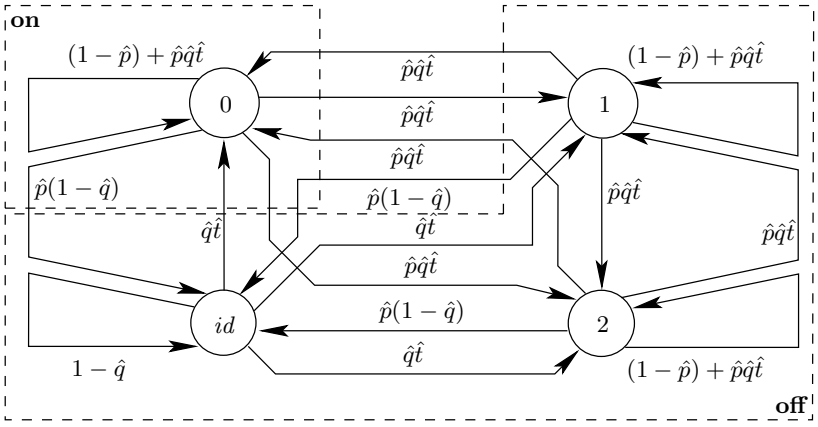


Fig. 2. The graph of the DTMC fully characterizing any input of the 3×3 switch

on the state represents cell arrival from the observed input to output 0 and **off** the states represent no cell arrival from the observed input to output 0.

In the following we introduce the approximating two state ON/OFF model of the general input mainly as replacing the set **off** with a single state OFF. Hereinafter uppercase ON and OFF denote the states of the approximating two state description of the input process.

The ON properties State ON replaces the one element subset **on** with the same sojourn probability $(1-\hat{p})+\hat{p}\hat{q}\hat{t}$. Accordingly the state transition probability from ON to OFF is 1 minus the sojourn probability $\hat{p}-\hat{p}\hat{q}\hat{t}$.

The OFF properties The OFF state replaces the set of **off** states by approximating their sojourn time with the absorbing time of a DPH distribution described in the followings.

For output 0 the transient states of the DPH are the **off** states and the absorbing state is the **on** state as depicted in Figure 3.

Based on $\mathbf{P}^{\mathcal{C}}$, given in (7), we give the initial distribution ($\boldsymbol{\beta}$) and the state transition probability matrix ($\mathbf{B}$) of the DPH. The initial distribution is the state probability right after entering **off** from **on**. It is obtained as the renormalization of the zeroth row of $\mathbf{P}^{\mathcal{C}}$ without its zeroth element

$$\boldsymbol{\beta} = \left(\frac{\hat{q}\hat{t}}{2\hat{q}\hat{t}+(1-\hat{q})} \;\; \frac{\hat{q}\hat{t}}{2\hat{q}\hat{t}+(1-\hat{q})} \;\; \frac{1-\hat{q}}{2\hat{q}\hat{t}+(1-\hat{q})} \right),$$

which is also indicated in Figure 3. The 3×3 sized state transition probability matrix of the **off** states is obtained from $\mathbf{P}^{\mathcal{C}}$ by cutting its zeroth row and zeroth column

$$\mathbf{B} = \begin{pmatrix} (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & \hat{p}\,(1-\hat{q}) \\ \hat{p}\hat{q}\hat{t} & (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}\,(1-\hat{q}) \\ \hat{q}\hat{t} & \hat{q}\hat{t} & 1-\hat{q} \end{pmatrix}.$$

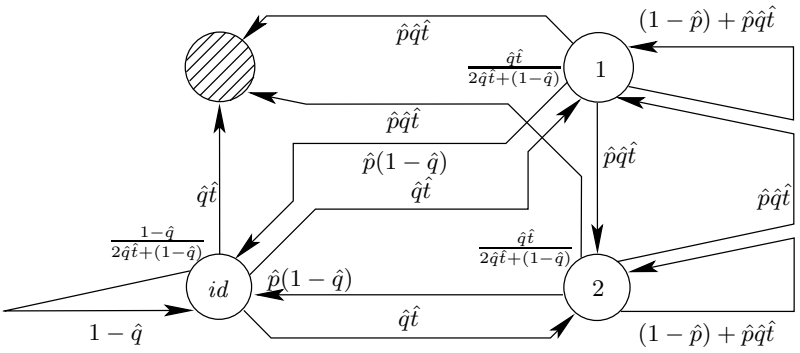


Fig. 3. The graph of the DPH substitution of the **off** states in terms of output 0

The mean absorbing time of this DPH is then

$$\mu = \boldsymbol{\beta}\left(\mathbf{I}-\mathbf{B}\right)^{-1}\mathbf{h}, \tag{8}$$

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

Here we note that according to the structure of (7) μ is the same for any output and any input – indeed the input processes are identical.

Consequently the sojourn probability of state OFF is $1-\frac{1}{\mu}$. The state transition probability from OFF to ON is $\frac{1}{\mu}$ which sets the mean sojourn time in state OFF equal to μ.

The state transition probability matrix of the two state DTMC describing the ON/OFF input process for the general path is

$$\mathbf{P} = \begin{pmatrix} (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}-\hat{p}\hat{q}\hat{t} \\ \frac{1}{\mu} & 1-\frac{1}{\mu} \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}, \tag{9}$$

where we also introduced a simplified notation with p and q. The graph of the ON/OFF DTMC using the simplified notation is given in Figure 4 which is the same for all the inputs according to the identical input process assumption.

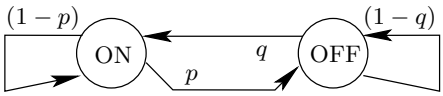


Fig. 4. The ON/OFF model of the input process with the simplified notation

3.2 Aggregate input model

We describe the combined behavior of the N inputs by a DTMC of $N+1$ states representing the number of inputs in ON. Using the considerations in Section 3.1

and especially (9) the ijth element of the state transition probability matrix of such a DTMC describing N inputs after 1 time slots is

$$\left(\boldsymbol{\mathcal{P}}_{N,1}(p,q)\right)_{ij} = \sum_{k=\max(0,j-i)}^{\min(i,N-j)} \binom{i}{k} p^k (1-p)^{i-k} \binom{N-i}{j-i+k} q^{j-i+k} (1-q)^{N-j-k} \tag{10}$$

where we also indicated that these probabilities depend on the parameters of (9) – p, q. The first binomial factor of (10) represents that out of i ON sources k moves to OFF and the second factor represents that out of $N-i$ OFF sources $j-i+k$ moves to ON, $i, j \in [0, N-1]$. (10) also introduces the notation $\boldsymbol{\mathcal{P}}_{N,M}(p,q)$ hereinafter denoting the state of N inputs during M time slots with each input modeled by an ON/OFF DTMC with parameters p and q given in (9). For example the state of N inputs after M time slots is

$$\boldsymbol{\mathcal{P}}_{N,M}(p,q) = \boldsymbol{\mathcal{P}}_{N,1}^{M}(p,q). \tag{11}$$

Using the above method there can be given behavior of any number of inputs in any number of time slots.

Based on $\boldsymbol{\mathcal{P}}_{N,M}(p,q)$ we give the arrival based decomposition of the arrival process as

$$\underbrace{\mathbf{B} = \begin{pmatrix} \mathbf{p}^0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{\text{0 arrivals}} \quad \underbrace{\mathbf{L} = \begin{pmatrix} 0 \\ \mathbf{p}^1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}}_{\text{1 arrival}} \quad \underbrace{\mathbf{F}_1 = \begin{pmatrix} 0 \\ 0 \\ \mathbf{p}^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\text{2 arrivals}} \quad \dots \quad \underbrace{\mathbf{F}_{N-1} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mathbf{p}^N \end{pmatrix}}_{\text{N arrivals}}, \tag{12}$$

where $\mathbf{p}^i$ denotes the ith row vector of $\boldsymbol{\mathcal{P}}_{N,M}(p,q)$.

The arrival based decomposition of the $N \times N$ switch in M time slots, is formalized in Algorithm 1.

Algorithm 1 Arrival based decomposition of the input process

INPUT: $N, M, \mathbf{P}$ from (9)
OUTPUT: $\mathbf{B}, \mathbf{L}, \mathbf{F}_1, \dots, \mathbf{F}_{N-1}$ the arrival based decomposition
1: determine $\boldsymbol{\mathcal{P}}_{N,M}(p,q)$ similar to (11) using $\mathbf{P}$
2: decompose $\boldsymbol{\mathcal{P}}_{N,M}(p,q)$ as in (12)
3: **return** $\mathbf{B}, \mathbf{L}, \mathbf{F}_1, \dots, \mathbf{F}_{N-1}$

3.3 The cell level model of the 3 × 3 switch

We show how the packet loss is calculated in path $\{1,0,0\}$ for which we give the cell level model of the corresponding VOQ – VOQ_{00}. It is a quasi birth-death

like (QBD-like) structure whose level represents the queue length and phase represents the state of the input process.

As the phase process of the QBD-like model is the combined state of the inputs their arrival based decomposition gives the level transition matrices used to build the QBD-like structure. $\mathbf{B}, \mathbf{L}, \mathbf{F}_1, \mathbf{F}_2$ are determined by Algorithm 1 with input parameters $N = 3$, according to the number of inputs, $M = 3$ the number of time slots in a time period and $\mathbf{P}$ (from (9)). Here $M = 3$ since the time period of the DTMC is 3 time slots long – as it is given in Section 2.1.

A level transition backward is according to $\mathbf{B}$ since there is one cell served during a time period and $\mathbf{B}$ represents 0 arrivals. Local state transition is according to $\mathbf{L}$ and there are 1(2) forward level transition(s) according to $\mathbf{F}_1(\mathbf{F}_2)$.

The state transition probability matrix of the QBD-like model is

$$\mathbb{P} = \begin{pmatrix} \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \dots \\ \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 \\ \dots & 0 & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}'_1 \\ \dots & 0 & 0 & 0 & \mathbf{B} & \mathbf{L}' \end{pmatrix}, \tag{13}$$

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

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

$$\boldsymbol{\pi}\mathbb{P} = \boldsymbol{\pi}, \qquad \boldsymbol{\pi}\mathbf{h} = 1. \tag{14}$$

3.4 Packet level model

With the geometric assumption for the packet length, given in Section 2.2, the life cycle of a packet in the observed path can be modeled by a transient DTMC in which there are two absorbing states corresponding to the two possible ending of a packet. The first absorbing state corresponds to the first cell loss, or equivalently the packet loss (PL) and the other one corresponds to the successful packet transmission (ST). The transient DTMC with two absorbing states is given in Figure 5. In this section we present this transient DTMC with its state transition probability matrix and initial distribution based representation.

The transient part of the DTMC Basically during the life cycle of a packet VOQ_{00} is modeled by a quasi birth like (QB-like) structure. Its level represents the queue length and its phase process is the combined state of the 3 inputs. In this case there is one important difference compared to the model given in the previous section. Input 1 is in ON for sure, since this is the model of the life cycle of a packet arrives from input 1, which also implies that there is no backward level transition.

The other two inputs behave in the "normal" manner, i.e., their corresponding level transition matrices are determined by Algorithm 1 with input parameters $N = 2$, $M = 3$ and $\mathbf{P}$ in (9). $M = 3$ since the time unit of the 3×3 switch

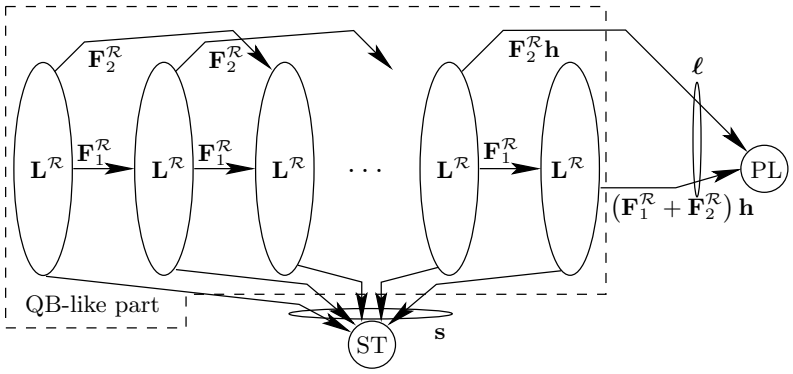


Fig. 5. The transient DTMC modelling the VOQ during the life cycle of a packet

is 3 time slots. The result of the algorithm is

$$\mathbf{B}, \mathbf{L} \text{ and } \mathbf{F}, \tag{15}$$

of size 3×3 as they describe 2 inputs (the possible states of this phase process are $0, 1$ and 2 – the number of inputs that are in ON).

According to these considerations the state transition probability matrix of the QB-like structure is built using the blocks

$$\mathbf{L}^{\mathcal{R}} = (1-p)^3\,\mathbf{B}, \qquad \mathbf{F}_1^{\mathcal{R}} = (1-p)^3\,\mathbf{L} \qquad \text{and} \qquad \mathbf{F}_2^{\mathcal{R}} = (1-p)^3\,\mathbf{F}. \tag{16}$$

Superscript $\mathcal{R}$ denotes quantities describing this transient DTMC of Figure 5. (16) describes the joint behavior of input 1 (given by $(1-p)^3$, the probability that input 1 remains in ON) and the other two inputs (given by matrices $\mathbf{B}, \mathbf{L}, \mathbf{F}$).

Finally using (16) the state transition probability matrix of the transient part and the state transition probability vector to state PL are

$$\mathbb{P}^{\mathcal{R}} = \begin{pmatrix} \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} \\ \dots & 0 & 0 & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} \\ \dots & 0 & 0 & 0 & \mathbf{L}^{\mathcal{R}} \end{pmatrix}, \qquad \boldsymbol{\ell} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{F}_2^{\mathcal{R}}\mathbf{h} \\ \left(\mathbf{F}_1^{\mathcal{R}}+\mathbf{F}_2^{\mathcal{R}}\right)\mathbf{h} \end{pmatrix}, \tag{17}$$

where $\boldsymbol{\ell}$ tells that if input 1 is in ON (which is the fundamental assumption here) then there is packet loss if at the beginning of the time period there is either

- one free position in the VOQ and there are three arrivals $\left(\mathbf{F}_2^{\mathcal{R}}\mathbf{h}\right)$ or
- no free positions in the buffer and there are either
  - two arrivals $\left(\mathbf{F}_1^{\mathcal{R}}\mathbf{h}\right)$ or
  - three arrivals $\left(\mathbf{F}_2^{\mathcal{R}}\mathbf{h}\right)$.

Using $\mathbb{P}^{\mathcal{R}}\mathbf{h} + \boldsymbol{\ell} + \mathbf{s} = \mathbf{h}$ the state transition probability vector to state ST is

$$\mathbf{s} = \mathbf{h} - \left(\mathbb{P}^{\mathcal{R}}\mathbf{h} + \boldsymbol{\ell}\right). \tag{18}$$

The initial distribution of the transient DTMC The initial distribution of $\mathbb{P}^{\mathcal{R}}$ in (17) is determined as the state of the system right after the arrival of an incoming packet. In this section we determine the probability distribution of the system at this time instance, right after a new packet arrival.

Here we give the joint probability of arriving a new packet at input 1 and the "normal" behavior of the other two inputs. Using the notations introduced in (9) the first probability is $1-(1-q)^3$ and latter one is determined as the output of Algorithm 1 with input parameters $N=2, M=3, \mathbf{P}$, the same as in (15). If $\tilde{q}=1-q$ then their joint behavior is described by the matrices

$$\hat{\mathbf{B}}^{\mathcal{N}} = \left(1-\tilde{q}^3\right)\mathbf{B}, \qquad \hat{\mathbf{L}}^{\mathcal{N}} = \left(1-\tilde{q}^3\right)\mathbf{L} \qquad \text{and} \qquad \hat{\mathbf{F}}^{\mathcal{N}} = \left(1-\tilde{q}^3\right)\mathbf{F}. \tag{19}$$

The block sizes of $\boldsymbol{\pi}$ in (14) are 4 since they describe all the 3 inputs. According to this there is a row of zeros appended to every level transition matrices in (19) as

$$\mathbf{B}^{\mathcal{N}} = \begin{pmatrix} \hat{\mathbf{B}}^{\mathcal{N}} \\ 0 \end{pmatrix} \qquad \mathbf{L}^{\mathcal{N}} = \begin{pmatrix} \hat{\mathbf{L}}^{\mathcal{N}} \\ 0 \end{pmatrix} \qquad \mathbf{F}^{\mathcal{N}} = \begin{pmatrix} \hat{\mathbf{F}}^{\mathcal{N}} \\ 0 \end{pmatrix}. \tag{20}$$

The last row expresses that in case of a new packet arrival there cannot be all the $N=3$ inputs in ON. Here we recall that in our model there is no corresponding cell arrival to state change from OFF to ON, i.e., in case of new packet arrival there is no cell arrival from the observed input.

Then starting from the steady state of the cell level model (14) and using the level transitions according to new packet arrival (20) the blocks of the initial distribution of the transient DTMC given in Figure 5 are

$$\begin{aligned}
\hat{\boldsymbol{\pi}}_0^{\mathcal{N}} &= \boldsymbol{\pi}_0\mathbf{B}^{\mathcal{N}} + \boldsymbol{\pi}_1\mathbf{B}^{\mathcal{N}} \\
\hat{\boldsymbol{\pi}}_1^{\mathcal{N}} &= \boldsymbol{\pi}_0\mathbf{L}^{\mathcal{N}} + \boldsymbol{\pi}_1\mathbf{L}^{\mathcal{N}} + \boldsymbol{\pi}_2\mathbf{B}^{\mathcal{N}} \\
\hat{\boldsymbol{\pi}}_2^{\mathcal{N}} &= \boldsymbol{\pi}_0\mathbf{F}^{\mathcal{N}} + \boldsymbol{\pi}_1\mathbf{F}^{\mathcal{N}} + \boldsymbol{\pi}_2\mathbf{L}^{\mathcal{N}} + \boldsymbol{\pi}_3\mathbf{B}^{\mathcal{N}} \\
\hat{\boldsymbol{\pi}}_i^{\mathcal{N}} &= \boldsymbol{\pi}_{i-1}\mathbf{F}^{\mathcal{N}} + \boldsymbol{\pi}_i\mathbf{L}^{\mathcal{N}} + \boldsymbol{\pi}_{i+1}\mathbf{B}^{\mathcal{N}} \quad 3 \le i \le b-1 \\
\hat{\boldsymbol{\pi}}_b^{\mathcal{N}} &= \boldsymbol{\pi}_{b-1}\mathbf{F}^{\mathcal{N}} + \boldsymbol{\pi}_b\left(\mathbf{L}^{\mathcal{N}} + \mathbf{F}^{\mathcal{N}}\right).
\end{aligned}$$

$\hat{\boldsymbol{\pi}}^{\mathcal{N}}$ is normalized as

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

resulting in the initial distribution of the packet level model in Figure 5.

The packet loss of the system Using (17), (18) and (21) the packet loss probability of the system and the probability of successful packet transmission on the given path are calculated as absorbing in state PL and ST, respectively, i.e.,

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

Estimation for the packet waiting time We estimate the mean packet waiting time with the mean cell waiting time. The mean cell waiting time equals to the mean system time of the cells entering the queue minus the cell service time. Since the service of the VOQ is deterministic the system time of a cell in the VOQ is $N = 3$ time slots times the queue length right after the cell arrival given that the cell is not dropped (denoted as $\tilde{\boldsymbol{\pi}}'$).

$\tilde{\boldsymbol{\pi}}'$ can be determined by the equation system

$$\begin{aligned}
\tilde{\boldsymbol{\pi}}'_1 &= \boldsymbol{\pi}_1 \left(\frac{1}{3}\mathbf{F}_2 + \frac{1}{2}\mathbf{F}_1 + \mathbf{L}\right) + \boldsymbol{\pi}_0 \left(\frac{2}{3}\mathbf{F}_2 + \mathbf{F}_1 + \mathbf{L}\right) \\
\tilde{\boldsymbol{\pi}}'_i &= \boldsymbol{\pi}_{i-2}\frac{1}{3}\mathbf{F}_2 + \boldsymbol{\pi}_{i-1}\left(\frac{1}{3}\mathbf{F}_2 + \frac{1}{2}\mathbf{F}_1\right) + \boldsymbol{\pi}_i \left(\frac{1}{3}\mathbf{F}_2 + \frac{1}{2}\mathbf{F}_1 + \mathbf{L}\right) \quad i \in [2, b-2] \\
\tilde{\boldsymbol{\pi}}'_{b-1} &= \boldsymbol{\pi}_{b-3}\frac{1}{3}\mathbf{F}_2 + \boldsymbol{\pi}_{b-2}\left(\frac{1}{3}\mathbf{F}_2 + \frac{1}{2}\mathbf{F}_1\right) + \boldsymbol{\pi}_{b-1}\left(\frac{1}{2}\mathbf{F}_2 + \frac{1}{2}\mathbf{F}_1 + \mathbf{L}\right) \\
\tilde{\boldsymbol{\pi}}'_b &= \boldsymbol{\pi}_{b-2}\frac{1}{3}\mathbf{F}_2 + \boldsymbol{\pi}_{b-1}\left(\frac{1}{2}\mathbf{F}_2 + \frac{1}{2}\mathbf{F}_1\right) + \boldsymbol{\pi}_b \left(\mathbf{F}_2 + \mathbf{F}_1 + \mathbf{L}\right).
\end{aligned}$$

$\tilde{\boldsymbol{\pi}}'$ is normalized as $\tilde{\boldsymbol{\pi}} = \frac{\tilde{\boldsymbol{\pi}}'}{\tilde{\boldsymbol{\pi}}'\mathbf{h}}$ resulting in the queue length distribution right after a cell arrival given that the cell enters the queue.

4 On the solution of large QBD-like DTMCs

Building the cell level DTMC as in (13) and solving it as in (14) results in the solution of a linear equation system of size $(b+1)(N+1)$ which can be easily led to inaccurate numerical results.

As a fast and numerically efficient solution of this we apply the folding algorithm, e.g., in [11], based solution of (14). The algorithm is prepared to block tri-diagonal matrices, hence we repartition (13) as

$$\mathbb{P} = \begin{pmatrix}
\mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & & & 0 & \\
\mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & & & & \\
0 & \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & & \\
0 & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
 & & & 0 & & & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}'_1 \\
 & & & & & & 0 & 0 & \mathbf{B} & \mathbf{L}'
\end{pmatrix} = \begin{pmatrix}
\mathbb{L}' & \mathbb{F}' & 0 & \\
\mathbb{B} & \mathbb{L} & \mathbb{F} & 0 \\
\cdots & \cdots & \cdots & \cdots \\
0 & & \mathbb{B} & \mathbb{L}''
\end{pmatrix}, \tag{23}$$

where we have enlarged the block size to $(N-1)(N+1)$. The inverse of the enlarged block is calculated in the folding algorithm, by which we have increased the complexity as well. We can increase the buffer size b to high values as the computational complexity of the folding algorithm is $O(\log_2 b)$.

In the followings we give the reduction of the matrix inversion of $\mathbf{I} - \mathbb{P}^{\mathcal{R}}$, in (22), to the inversion of its diagonal block – denoted as $\mathbf{V} = \mathbf{I} - \mathbf{L}^{\mathcal{R}}$. Considering the matrix equation

$$\mathbf{x}\left(\mathbf{I} - \mathbb{P}^{\mathcal{R}}\right) = \boldsymbol{\pi}^{\mathcal{N}} \tag{24}$$

where the coefficient matrix $\left(\mathbf{I}-\mathbb{P}^{\mathcal{R}}\right)$ has an upper triangular structure, on the block level, we can apply the following iterative solution of the matrix equation

$$\begin{aligned}\mathbf{x}_0\mathbf{V} = \boldsymbol{\pi}_0^{\mathcal{N}} \quad &\rightarrow \mathbf{x}_0 = \boldsymbol{\pi}_0^{\mathcal{N}}\mathbf{V}^{-1} \\ \mathbf{x}_0\mathbf{F}_1 + \mathbf{x}_1\mathbf{V} = \boldsymbol{\pi}_1^{\mathcal{N}} \quad &\rightarrow \mathbf{x}_1 = \left(\boldsymbol{\pi}_1^{\mathcal{N}} - \mathbf{x}_0\mathbf{F}_1\right)\mathbf{V}^{-1}\end{aligned}$$

and all the other blocks for $i = 2, \ldots, b$ are

$$\mathbf{x}_{i-2}\mathbf{F}_2 + \mathbf{x}_{i-1}\mathbf{F}_1 + \mathbf{x}_i\mathbf{V} = \boldsymbol{\pi}_i^{\mathcal{N}} \quad \rightarrow \mathbf{x}_i = \left(\boldsymbol{\pi}_i^{\mathcal{N}} - \mathbf{x}_{i-1}\mathbf{F}_1 - \mathbf{x}_{i-2}\mathbf{F}_2\right)\mathbf{V}^{-1}$$

Rearranging (24) results in $\mathbf{x} = \boldsymbol{\pi}^{\mathcal{N}}\left(\mathbf{I}-\mathbb{P}^{\mathcal{R}}\right)^{-1}$ which implies that from (22) the packet loss probability (p_ℓ) and the probability of successful packet transmission (p_s) of the observed VOQ can be calculated as

$$p_\ell = \mathbf{x}\boldsymbol{\ell} \qquad \text{and} \qquad p_s = \mathbf{x}\mathbf{s}. \tag{25}$$

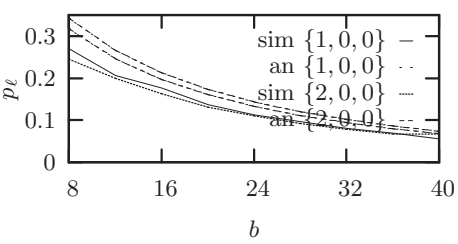
5 A numerical study

In contrast to [5, 6] where we described extended methodology of packet loss analysis in the LB switch, this paper presents optimized solution with linear complexity. The computational study has two parts. The first part shows the behavior of the packet loss and waiting time of the LB switch as a function of buffer length and switch size. The second part examines some extreme cases when central stage buffers are large to show the power of the folding algorithm based solution method presented in Section 4. For the results of this section we used the parameters given in Table 1. In order the comparative analysis, we made the specified measurements also with our LB switch simulation tool.

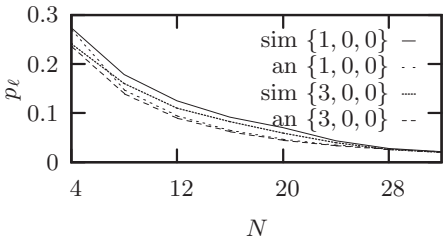
Figure	6(a)	6(b)	6(c)	6(d)	7(a)	7(b)
name	p_ℓ vs. b	p_ℓ vs. N	T vs. b	T vs. N	p_ℓ vs. b	T vs. b
	without folding algorithm				with folding	
N	4	$4, \ldots, 32$	4	$3, \ldots, 33$	3	
b	$8, \ldots, 40$	36	$8, \ldots, 40$	127	$9, \ldots, 999$	
$\hat{p}$	$\frac{1}{20}$	$\frac{1}{40}$	$\frac{1}{20}$	$\frac{1}{50}$		
$\hat{q}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{3}$			
$\hat{t}$	$\frac{1}{N}$					

Table 1. Parameters used for the numerical studies

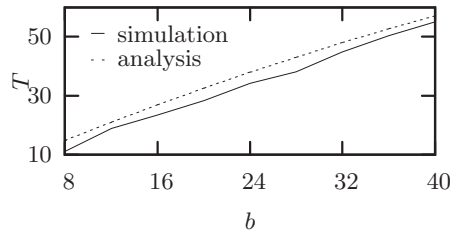
Part 1 In [5, 6] we examined the dependence of packet loss at the central stage buffers on the buffer size and switch size. It was found that the packet loss probability strongly depends on the chosen path ($\{i, j, k\}$). Figure 6(a) and 6(b) present similar results using the approximate model introduced in this paper.



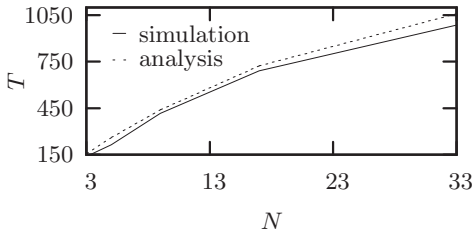
(a) Packet loss versus the buffer size



(b) Packet loss versus the switch size



(c) Packet waiting time versus the buffer size



(d) Packet waiting time versus the switch size

Fig. 6. Numerical results for the packet loss analysis of LB switches

Figures 6(c) and 6(d) indicate another performance characteristic, the packet waiting time estimator compared to simulation results. The packet waiting time is evaluated considering only the successfully transmitted packets. The packet waiting time is generally increases together with the buffer size (larger interval between cell arrivals and services), like in Figure 6(c) and switch size (cells are spread to more queues), like in Figure 6(d).

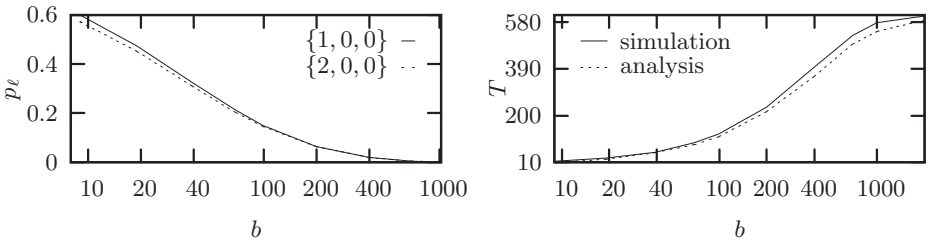
Part 2 Figure 7(a) and 7(b) shows the applicability of the analytical model for large buffer sizes. According to presented results, we admit that the ratio between the switch size and buffer length of the VOQs is a crucial issue for the expected packet loss and system performance. Unfortunately, the optimal set of parameters (e.g. switch size and buffer length) is not constant and should be chosen to the specific needs.

6 Conclusions

In this paper we present a scalable model for the packet loss and packet waiting time analysis in the load-balancing Birkhof-von Neumann switch.

This model also reflex the previously shown property of different loss probabilities on the chosen path traversing the switch [5, 6].

The computational complexity of the approximate model introduced in this paper is reduced to be linear with N, the number of ports of the switch. The



(a) Packet loss versus the buffer size

(b) Packet waiting time versus the buffer size

Fig. 7. Behavior of switches with large buffers

other contribution of the paper is the folding algorithm based, numerically stable and fast algorithm to solve the DTMCs for large buffer sizes (b). This allow us to solve switches of size up to ~ 30 equipped with buffer of size up to ~ 1000.

References

1. Chang, C., Lee, D., Jou, Y.: Load-Balanced Birkhoff-von Neumann switches, Part I: One-Stage Buffering. Computer Communications **25** (2002) 611–622
2. Keslassy, I., Chuang, S., Yu, K., Miller, D., Horowitz, M., Solgaad, O., McKeown, N.: Scaling Internet Routers Using Optics. In: SIGCOMM'03, Germany (2003)
3. 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
4. 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
5. Audzevich, Y., Bodrog, L., Telek, M., Ofek, Y., Yener, B.: Variable Size Packets Analysis in Load-balanced Switch with Finite Buffers. submitted for revision to IEEE HPSR'09 available at `http://webspn.hit.bme.hu/~bodrog/techrep/hpsr2009.pdf` (January 2009)
6. Audzevich, Y., Bodrog, L., Ofek, Y., Telek, M.: Packet Loss Analysis of Load-Balancing Switch with ON/OFF Input Processes. submitted for revision to EPEW2009 available at `http://webspn.hit.bme.hu/~bodrog/techrep/epew2009.pdf` (February 2009)
7. Thompson, K., Miller, G., Wilder, R.: Wide-area Internet Traffic Patterns and Characteristics. IEEE Network **11**(6) (November/December 1997) 10–23
8. Fomenkov, M., Keys, K., Moore, D., Claffy, K.: Longitudinal study of internet traffic in 1998-2003. In: Proceedings of WISICT, Mexico (5-8. January 2004)
9. Turner, J.: Strong Performance Guarantees for Asynchronous Crossbar Schedulers. In: IEEE INFOCOM '06, Barcelona, Spain (April 2006) 1–11
10. Turner, J.: Resilient Cell Resequencing in Terabit Routers. Technical report, Washington University, Department of Computer Science (June 2003)
11. Ye, J., Li, S.: Courier dover publication. Folding Algorithm: A Computational Method for Finite QBD Processes with Level-Dependent Transitions **42**(2/3/4) (February/March/April 1994) 652–639