Packet Loss Minimization in Load-Balancing Switch

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}@disi.unitn.it
[2] Department of Telecommunications,
Technical University of Budapest, Hungary,
{bodrog,telek}@hit.bme.hu

Abstract. Due to the overall growing demand on the network resources and tight restrictions on the power consumption, the requirements to the long-term scalability, cost and performance capabilities appear together with the deployment of novel switching architectures. The load-balancing switch proposed in [1,2] satisfies to the above requirements due to a simple distributed control and good performance characteristics. However, as it was proven in [3, 4] a set of specific assumptions applied to the load-balancing switch restrains the above advantages. In particular, due to the limited information availability, cental stage buffers can overflow and, correspondingly, a packet loss can occur.
In this paper we present a novel load-balancing service protocol which uses the congestion allocation technique to allow the drop of the arriving packets also at the input stage. If congestion indication is detected, the input stage will drop the whole packet upon its arrival while reducing the probability of congestion at the output reassembly unit. In the following the mathematical model for joint input/central stage packet loss evaluation will be presented. Based on the presented analysis we account the ways to minimize the mentioned packet loss.

1 Introduction

Most of the packet switching technologies are forwarding packets from the ingress to the egress port while using substantial computation resources for decision making, header processing and packet storage. In this paper we focus our attention on the Load-Balancing (LB) switching architecture, which promises a simple distributed control with almost no communication and computation overheads and large set of performance benefits [1,2]. In particular, the architecture guarantees high throughput and small packet delay while applying certain assumptions. Some of them like usage of infinite buffers at switch stages, consideration of only admissible input traffic matrices and management of fixed size packets are simply impractical. [3–6] have considered the traditional load-balancing switching architecture under a practical set of assumptions. The finite buffers inside the system implies congestion and correspondingly non-zero packet loss of the system. In [3] performance characteristics of the switch were evaluated for both admissible and inadmissible traffic matrices for fixed size data cells. In contrast, papers [4–6] proposed and analyzed more realistic behavior of the switch with variable size packets arriving to the inputs. Although the variable size packets were considered, the data transmission was performed on the cell-by-cell basis (by means of segmentation and reassembly). The derived results have characterized the performance capabilities of the switch under various buffer and switch sizes. Apart from these, [6] has depicted a set of reassembly problems when incomplete packets arriving to the output for reconstruction. Since cells are transmitted inside the switch without any respect of the possible congestion in the next stage, cell drops at the central stage buffers will result in continuous arrival of incomplete packets to the output stage and also out-of-order delivery of cells. This aspect will provoke congestion at the output stage, wastage of considerable amount of buffering as well as implementation of some sophisticated algorithms for identification and removal of incomplete packets. Finally, as it was shown in [4–6] the internal packet loss probability strictly depends on the crossbar interconnection pattern and the evaluated path (input - central stage - output sequence).

In this paper we introduce a new packet acceptance policy which tries to minimize the packet loss by allowing the packets to be dropped at the input stage of the LB switch. The packet drop at the input helps to keep the number of waste cells (pointlessly processed cells of an already dropped packet), inside the switch, at the optimal level. Inhibiting packets to enter the almost saturated central stage (CS) avoids buffer saturation and correspondingly dropping the cells of the already accepted packets. As a consequence the accepted packets are dropped with lower probability and does not waste system capacity.

In particular, central stage buffer occupancy is controlled by means of an artificial buffering threshold. New packets arriving at an input are rejected to enter the switch if a certain CS buffer occupancy is above the threshold. Moreover, the analysis depicts the ways to control the input and central stage packet loss in such a way that the joint input-central stage (I-CS) packet loss is minimal.

In the following we present the overall description of the considered load-balancing (LB) switching architecture, paying particular attention to the practical integration of the service protocol (by means of a centralized controller – Section 2) into the traditional LB switch. Next, in Section 3 we present mathematical analysis which allows to evaluate the joint I-CS packet loss inside the switch. Section 4 presents computation study related to the protocol performance for various switch, central stage buffer and packet sizes. This part will also verify the mathematical analysis with developed simulation model. Finally, Section 5 concludes the paper.

2 Description of the considered architecture

In comparison to the traditional two-stage LB switch evaluated, e.g., in [6], the examined architecture includes also a centralized controller (Figure 1). It

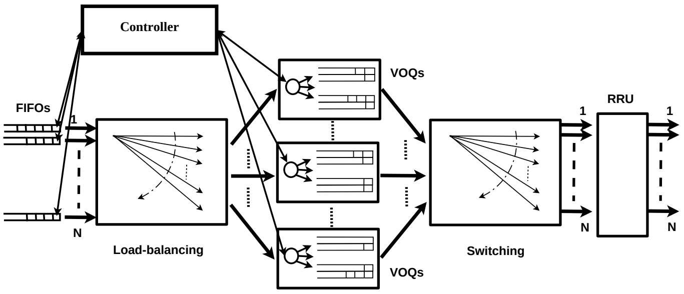


Fig. 1. The load-balanced switch considered for the analysis

is assumed that the variable length packets arriving to inputs are stored and segmented into fixed-size data cells in First-In-First-Out (FIFO) queues. The transmission process of cells through the switch is also well reported in [1]. Packets are reassembled at re-sequencing and reassembly unit (RRU) back to data cells arriving to outputs. The traditional load-balanced switch with finite central stage buffers can have a cell/packet loss due to buffer overflow [3, 4]. As a result, the arrival of incomplete packets to the output RRU can cause large buffering space wastage and enormous delays. To handle these issues the possibility of packet drop at the input is introduced in this paper. In our case the drop-tail discipline is used to keep particular attention on the theory of input packet dropping but one can improve the fairness of the packet acceptance by applying more fair active queue management (AQM) like random early detection (RED) [7] or adaptive RED [8].

Since the main point of congestion in the traditional LB switch can be found in central stage buffers, the following two values should be observed/controlled in order to minimize/avoid packet loss. In particular, it is necessary to carry the information: 1) about the input packet arrivals during a time slot (the basic time unit of the system) and 2) information about the occupancy of the central stage buffers. Since each stage of a basic two-stage switch is independent from all the other stages of the switch, the most appropriate way of the data collecting is by means of centralized unit. The controller is using detached links for information exchange and is interconnected with both inputs and central stage buffer sets (CSSs). Please note, that in order to maintain distributed control in the system with such a service protocol the switch might have considerably greater communication and computation overheads than that with centralized control (this issue is discussed in Section 2.2). The service protocol which is implemented in the centralized controller, can set the artificial buffering threshold at the central stage buffers (either statically or dynamically) in order to distribute a packet loss between input and central stage. One of the important considerations is that the service protocol allows to drop packets (of variable size and directed to some specific output) at an input stage in case if the occupancy of at least one virtual output queue (VOQ), where the packet is supposed to be distributed, is greater than the defined threshold. With this drop policy theoretically it is possible to drop a "very short" packet which would not use the specific VOQ, with queue length above the threshold, but in our approach it is neglected to reduce the computational need of the overheads to the minimal value.

Example of the protocol function Observing a switch with N input and output ports ($N \times N$ switch) there are N set of virtual output queues (VOQ). In the kth set there is VOQ_{kj} dedicated to store cells of packets directed to output j. Supposing a packet directed to output j is distributed (depending on the current crossbar interconnection and packet size) cell-by-cell between $\mathrm{VOQ}_{0j}, \ldots,$ VOQ_{N-1j} at CSSs. However, in the current implementation of the switch, the controller performs congestion detection based on the value of an artificial buffer threshold T. The packet is allowed to be forwarded in case if the occupancy of $\mathrm{VOQ}_{0j}, \ldots, \mathrm{VOQ}_{N-1j}$ is less than or equal to T. Otherwise, if at least one of these queues has the occupancy greater than T, then the packet at the input is dropped. Such kind of comparison is performed for all the packets arriving to inputs during a time slot.

On the definition of the different loss probabilities The natural definition of the *cell loss probability* is the ratio between the cells dropped and the total number of cells entered the switch. The definition of the *packet loss probability* is similarly the ratio between the packets dropped and the total number of packets entered. A packet is considered to be dropped when one of its cells is dropped.

Cell loss probability inside the system is proved to be different for different transmission *path*s (input - central stage buffering set - output sequence), and the packet loss probability comes from the cell loss at the different paths. [4–6]

We distinguish input and CS packet loss. The *input packet loss* occurs based on a controller decision to drop an arrived packet. In this case input packets are removed entirely before the actual transmission is made. An arriving packet is dropped if there is at least one VOQ with occupancy greater than T, for the VOQs, potentially, used for the transmission of that particular packet. For example, packet arrived to input i and directed to output j is dropped if there is a k for which the occupancy of VOQ_{kj} is greater than T.

Throughout the paper we performed evaluation of the steady state joint input-central stage (I-CS) packet loss probability.

2.1 Information exchange in the controller

In this section we present the controller design and main operation principles used for the implementation of the service protocol considering an $N \times N$ switch. The controller which is interconnected with all inputs and CSSs is using N^2 bidirectional links for information exchange. The management unit is synchronized with the rest of the system, so the information exchange and all the computations are performed within a time slot basis.

The protocol can operate in two different modes. The first mode presumes statical initialization of a buffering threshold at the central stage, so the threshold

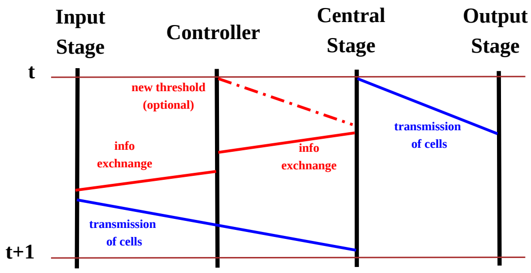


Fig. 2. Time diagram for considered load-balancing switch

is initialized before switch operation. In this case the threshold remains the same for the entire operation interval of the switch. The second approach assumes the threshold to be dynamically changing in time, e.g. can be modified at the beginning of any time slot. In this case the joint I-CS packet loss probability will be dynamically changing in correspondence to the actual threshold set. Section 4 presents results for the statically configured threshold operation.

The timing diagram in Figure 2 represents a set of consecutive operations performed in the switch during a time slot. At the beginning of a time slot transmission of cells from central stage buffers to the currently interconnected outputs is done. During this time the controller is capable to set a new threshold value. It is done by transmission of $\log b$ bits of information, where b is the physical length in cells of a VOQ (all VOQ buffers have the same size). As soon as forwarding from the CSSs to the outputs is done, CSSs check current occupancy of the virtual output queues and compare it with the prescribed threshold. Based on the comparison, each CSS creates a vector of N elements, each element of which is keeping one bit of information representing occupancy (congestion) status. If the current VOQ occupancy is greater than the threshold value, the bit is set to 1, otherwise it is set to 0. When occupancy vectors are formed, they are immediately transmitted from CCSs to the controller. The controller forms a decision matrix in a way that each arriving vector of N elements is placed as a column of the matrix. The decision is made on the destination basis in such a way that each row of the matrix is processed. If in a row of N elements at least one 1-bit exists (logical OR is applied), than all packets destined to that output are considered to be blocked for transmission. Otherwise if all elements of a row are 0-bits, transmission to this output is possible. Based on a simple logical OR operation, the final decision vector is created as it is shown in Figure 3. Finally N copies of the decision vector (N bits) are distributed to inputs. Based on the arrived final decision vector and availability of arrived packets, inputs either transmit packets to the next stage or drop them immediately (if a packet is already in transmission, no action is performed).

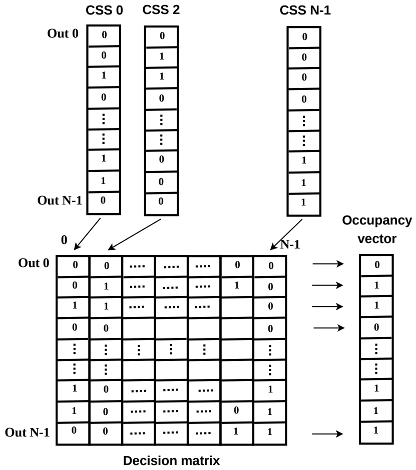


Fig. 3. Information processing and exchange at the centralized controller

2.2 The switch overheads and scalability

The traditional LB switch presented in [1, 2] is considered to be a highly scalable solution in comparison to a crossbar switch, driven by a stable matching algorithm, if relaxed assumptions (central stage buffers are infinite, packets are of the same size and admissible traffic arrivals) are applied. As a result the LB switch was able to provide high throughput and have zero information exchange implemented (each stage made self decisions). In contrast, it was shown in [3, 6] that if more realistic scenario is applied, the traditional switch is not able to provide high throughput due to significant internal packet loss.

The central stage packet loss of the system cannot be avoided, but it can be minimized knowing the arrivals to the inputs and the actual VOQ occupancies. The packet loss avoidance would imply the prediction of future arrivals which is impossible, i.e., it is impossible to evaluate the potential central stage packet loss based only on the existing information and current switch configuration. In order to improve the overall throughput of the switch, the additional information exchange can be implemented between the stages (giving non-zero communication and computation overheads). These modifications, in their turn will make an impact on the scalability properties of the system. Therefore the *tradeoff between the system scalability and throughput characteristics* exists.

The protocol used for minimization of packet loss described in this paper can be implemented either using *distributed information exchange* or *centralized information exchange.* In the following, all possible overheads of these solutions are compared in order to motivate the choice of centralized controller.

The distributed scheme Lets assume that all elements of the LB switch perform independent transmission decisions. In order to set a buffering threshold at all CSSs it is enough to send a request from a single input to all N CSSs. As soon

Table 1. Total system overheads for various implementation schemes

Management	Distributed	Centralized
Communication overhead	N^2 vectors of N bits	$2N$ vectors of N bits
Computation overhead	$N * N^2$ bits to compare	N^2 bits to compare
Additional wiring	0 links	N^2 links

as the threshold set, each CSS determines its occupancy vector, composed of N bits, for the current time slot and distribute this information between all N inputs. In total, during a time slot, each CSS sends N vectors of N bits, resulting in the transmission of N^2 vectors of N bits for the whole system. Based on the information arrived from all CSSs each input builds a decision matrix and performs logical comparison of bits (in total, the system will have N decision matrices). As a result the decisions for the currently arrived packets are performed. It is summarized in Table 1.

The centralized management The protocol realized in this paper is utilizing a centralized management which introduces extra wiring costs to the system since it utilizes N^2 detached links for information exchange. Considering the fact that only a centralized controller is performing the information exchange, we get $2N$ vectors of N bits as a total communication overhead in the system. Moreover, due to a simple bit-by-bit comparison (logical OR) performed in decision matrix of the controller the computation overhead is negligible and is constant in time. As a result, in terms of total overheads, the realization of the controller with centralized management is less complicated. The comparison is made in Table 1.

3 The mathematical analysis

The Markov model of the LB switch with identical input processes, without packet rejection, is given in [6] from which the present model, with packet rejection, is deduced. In this section we give the model with packet rejection in such a way that it can be understood on its own, but for more on all the detailed considerations of the original model, without packet rejection, we refer to [6].

First we summarize the model, without packet rejection, in Section 3.1 and than we give the differences of the model due to the introduction of the new packet acceptance protocol introduced in Section 2. The detailed model of the 3×3 switch (i.e. $N = 3$) is given in Section 3.2 which gives the steady state loss probability of the switch caused by finite central stage buffers.

3.1 Model of the LB switch without packet rejection

The LB switch, without packet rejection, can have packet loss due to cell loss in the finite central stage buffers. This is observed via the life cycle of a tagged packet which can either be transmitted successfully or be dropped due to the fact that one of its cells is dropped.

In Figure 4 there is a two dimensional, transient, discrete time Markov chain (DTMC) describing the life cycle of the tagged packet. Its level process (horizontal dimension) is the length of the tagged VOQ and its phase process (vertical dimension) is the state of the aggregated input process. The transient part has a quasi birth-like (QB-like) structure with possibly more than one (more precisely $N-1$) forward level transitions. The two absorbing states of the transient DTMC are the one representing the successful packet transmission (ST) and the packet loss (PL).

The main steps of the analysis of the original model are summarized in Algorithm 2 using Algorithm 1, but we refer to [6] for the details of the model. We used the following notations of [6]

N is the size of the switch, i.e., the number of the input and output ports,
b is the buffer size of the VOQs,
$\hat{p}$ is the parameter of the geometric distributed packet length in cells,
$\hat{q}$ is the parameter of the geometric distributed idle period length in time slots,
$\hat{t} = \frac{1}{N}$ is the probability of choosing a specific output for a given packet and
$\{i, j, k\}$ is a path, i.e., the ordinal number of the input, output and the VOQ respectively, $i, j, k \in [0, N-1]$.

3.2 Model of the LB switch with packet rejection

Here we give the model of the load-balancing switch with the possibility of packet rejection at the input stage, it is summarized in Algorithm 3. Like in [6] in case of the LB switch with packet rejection we also give the detailed model of the switch of size $N = 3$.

The packet acceptance threshold (T) is defined as the queue length, in cells, counted from the beginning of the observed VOQ in the central stage buffers.

If the queue length of the observed VOQ is above T then the arrival processes of the inputs are forced to be OFF, $\hat{q} = 0$. From modeling point of view $\hat{q} = 0$ represents the drop of the packets at the inputs. Setting $\hat{q} = 0$, when the queue length is greater than T, results in the model of the LB switch with packet rejection.

Algorithm 3 gives the joint input - central stage (I-CS) loss probability by determining the loss probability at the inputs and at the central stage by modeling the life cycle of a tagged packet. The transmission of the tagged packet is modeled by a similar transient DTMC, given in Figure 5, to the one modeling the LB switch without packet rejection, given in Figure 4.

During Algorithm 3 we use the results of [6], summarized in Algorithm 2. To emphasize the differences between the switch with and without packet rejection the notation $^{(\mathrm{th})}$ in the superscript is used to distinguish variables corresponding to the introduction of the packet rejection in the model for the packet loss minimization. Furthermore in the DTMCs, and accordingly in the state transition probability matrices, the differences caused by the introduction of the buffering threshold can be well distinguished. Both the graph of the packet level model in Figure 5 and the state transition probability matrices can be strictly divided into

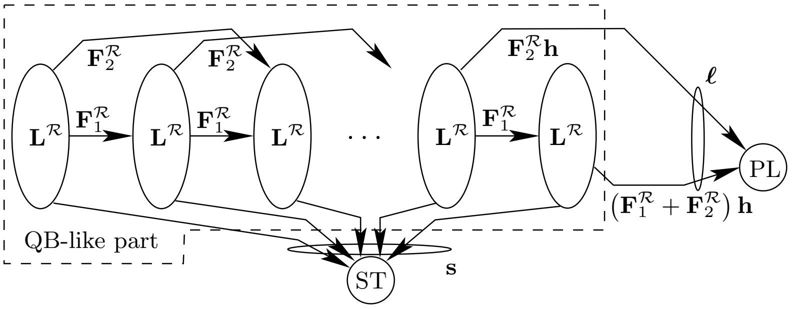

$\mathbf{F}_2^{\mathcal{R}}$
$\mathbf{F}_2^{\mathcal{R}}$
$\mathbf{F}_2^{\mathcal{R}}\mathbf{h}$
$\boldsymbol{\ell}$
$\mathbf{L}^{\mathcal{R}}$
$\mathbf{F}_1^{\mathcal{R}}$
$\mathbf{L}^{\mathcal{R}}$
$\mathbf{F}_1^{\mathcal{R}}$
$\mathbf{L}^{\mathcal{R}}$
$\cdots$
$\mathbf{L}^{\mathcal{R}}$
$\mathbf{F}_1^{\mathcal{R}}$
$\mathbf{L}^{\mathcal{R}}$
PL
$\left(\mathbf{F}_1^{\mathcal{R}}+\mathbf{F}_2^{\mathcal{R}}\right)\mathbf{h}$
QB-like part
ST
$\mathbf{s}$

Fig. 4. The transient DTMC modeling the VOQ during the life cycle of a packet in a 3×3 switch

Algorithm 1 `Level Transitions`$(N, M, \hat{p}, \hat{q}, \hat{t})$, the arrival based decomposition of the input process [6]

```
INPUT: $N, M, \hat{p}, \hat{q}, \hat{t}$
OUTPUT: $\mathbf{B}, \mathbf{L}, \mathbf{F}_1, \ldots \mathbf{F}_{N-1}, p, q$ //the arrival based decomposition and the ON/OFF properties
$\mathbf{P}^{\mathcal{C}} = \begin{pmatrix} (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & \hat{p}(1-\hat{q}) \\ \hat{p}\hat{q}\hat{t} & (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & \hat{p}(1-\hat{q}) \\ \hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}(1-\hat{q}) \\ \hat{q}\hat{t} & \hat{q}\hat{t} & \hat{q}\hat{t} & 1-\hat{q} \end{pmatrix}$ //the complete input model
$\boldsymbol{\beta} = \left( \frac{\hat{q}\hat{t}}{2\hat{q}\hat{t}+(1-\hat{q})} \; \frac{\hat{q}\hat{t}}{2\hat{q}\hat{t}+(1-\hat{q})} \; \frac{1-\hat{q}}{2\hat{q}\hat{t}+(1-\hat{q})} \right), \mathbf{B} = \begin{pmatrix} (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}\hat{q}\hat{t} & \hat{p}(1-\hat{q}) \\ \hat{p}\hat{q}\hat{t} & (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}(1-\hat{q}) \\ \hat{q}\hat{t} & \hat{q}\hat{t} & 1-\hat{q} \end{pmatrix}$ //the initial vector and the state transition probability matrix of the DPH substitution of the off states
$\mathbf{h} = \begin{pmatrix} 1 \\ \ldots \\ 1 \end{pmatrix}$ //an appropriate size column vector of ones
$\mu = \boldsymbol{\beta}\left(\mathbf{I} - \mathbf{B}\right)^{-1}\mathbf{h}$ //the solution of the DPH
$1 - \frac{1}{\mu}$ //the sojourn probability of the substituting OFF state
$(1 - \hat{p}) + \hat{p}\hat{q}\hat{t}$ //the sojourn probability of the ON state
$\mathbf{P} = \begin{pmatrix} (1-\hat{p})+\hat{p}\hat{q}\hat{t} & \hat{p}-\hat{p}\hat{q}\hat{t} \\ \frac{1}{\mu} & 1-\frac{1}{\mu} \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ //the ON/OFF input model
$\left(\boldsymbol{\mathcal{P}}_{N,1}(p,q)\right)_{ij} = \sum_{k=\max(0,j-i)}^{\min(i,N-j)} \binom{i}{k} p^k (1-p)^{i-k} \binom{N-i}{j-i+k} q^{j-i+k} (1-q)^{N-j-k}$ //the aggregate input model during one time slot
$\boldsymbol{\mathcal{P}}_{N,1}^{M}(p,q)_{(N+1)\times(N+1)} = \boldsymbol{\mathcal{P}}_{N,M}(p,q)_{(N+1)\times(N+1)} = \begin{pmatrix} \mathbf{p}_{1\times(N+1)}^{0} \\ \mathbf{p}_{1\times(N+1)}^{1} \\ \ldots \\ \mathbf{p}_{1\times(N+1)}^{N} \end{pmatrix}$ //aggregate input model during $M$ time slots and its row based decomposition with their sizes
$\underbrace{\mathbf{B} = \begin{pmatrix} \mathbf{p}^0 \\ 0 \\ \ldots \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{\text{0 arrivals}}, \quad \underbrace{\mathbf{L} = \begin{pmatrix} 0 \\ \mathbf{p}^1 \\ 0 \\ \ldots \\ 0 \\ 0 \end{pmatrix}}_{\text{1 arrival}}, \quad \underbrace{\mathbf{F}_1 = \begin{pmatrix} 0 \\ 0 \\ \mathbf{p}^2 \\ 0 \\ \ldots \\ 0 \end{pmatrix}}_{\text{2 arrivals}}, \ldots, \quad \underbrace{\mathbf{F}_{N-1} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \ldots \\ 0 \\ \mathbf{p}^N \end{pmatrix}}_{\text{N arrivals}}$ //the arrival based decomposition of the aggregate input model
return $(\mathbf{B}, \mathbf{L}, \mathbf{F}_1, \ldots \mathbf{F}_{N-1}, p, q)$
```

Algorithm 2 `Scalable Model`$(N=3, b, \hat{p}, \hat{q}, \hat{t}, \{i,j,k\}=\{1,0,0\})$, the scalable model of the 3×3 LB switch in [6]

```
INPUT: $N=3, b, \hat{p}, \hat{q}, \hat{t}, \{i,j,k\}=\{1,0,0\}$
OUTPUT: $p_s, p_\ell$ //the probabilities of successful packet transmission and packet drop
 1: $(\mathbf{B}, \mathbf{L}, \mathbf{F}_1, \mathbf{F}_2, p, q) =$ Level Transitions$(N, M=N, \hat{p}, \hat{q}, \hat{t})$//the arrival based decomposition of the aggregate process of all inputs during 3 time slots using Algorithm 1
 2: $\mathbb{P} = \begin{pmatrix} \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \cdots \\ \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 \\ \cdots & 0 & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}_1' \\ \cdots & 0 & 0 & 0 & \mathbf{B} & \mathbf{L}' \end{pmatrix}$//the cell level model of the $3\times 3$ switch
 3: $\boldsymbol{\pi}\mathbb{P} = \boldsymbol{\pi}, \quad \boldsymbol{\pi}\mathbf{h} = 1.$ //the steady state solution of the cell level model
 4: $(\mathbf{B}, \mathbf{L}, \mathbf{F}, p, q) =$ Level Transitions$(N-1, M=N, \hat{p}, \hat{q}, \hat{t})$//the arrival based decomposition of the aggregate process of two inputs during 3 time slots using Algorithm 1
 5: $\mathbf{L}^{\mathcal{R}} = (1-p)^3\,\mathbf{B}, \quad \mathbf{F}_1^{\mathcal{R}} = (1-p)^3\,\mathbf{L}, \quad \mathbf{F}_2^{\mathcal{R}} = (1-p)^3\,\mathbf{F}$//the arrival based decomposition of the aggregate process of the two non-observed and the observed input during 3 time slots
 6: $\mathbb{P}^{\mathcal{R}} = \begin{pmatrix} \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} \\ \cdots & 0 & 0 & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} \\ \cdots & 0 & 0 & 0 & \mathbf{L}^{\mathcal{R}} \end{pmatrix}, \quad \boldsymbol{\ell} = \begin{pmatrix} 0 \\ \cdots \\ 0 \\ \mathbf{F}_2^{\mathcal{R}}\mathbf{h} \\ (\mathbf{F}_1^{\mathcal{R}}+\mathbf{F}_2^{\mathcal{R}})\mathbf{h} \end{pmatrix}$ //the state transition probability matrix of the QB-like part and the absorption vector to state PL
 7: $\mathbf{s} = \mathbf{h} - \left(\mathbb{P}^{\mathcal{R}}\mathbf{h} + \boldsymbol{\ell}\right)$ //the absorption vector to state ST
 8: $\tilde{q} = 1 - q$
 9: $\hat{\mathbf{B}}^{\mathcal{N}} = \left(1-\tilde{q}^3\right)\mathbf{B}, \quad \hat{\mathbf{L}}^{\mathcal{N}} = \left(1-\tilde{q}^3\right)\mathbf{L}, \quad \hat{\mathbf{F}}^{\mathcal{N}} = \left(1-\tilde{q}^3\right)\mathbf{F}$//the level transitions according to packet arrival during 3 time slots using the results of line 4
10: $\mathbf{B}^{\mathcal{N}} = \begin{pmatrix} \hat{\mathbf{B}}^{\mathcal{N}} \\ 0 \end{pmatrix}, \quad \mathbf{L}^{\mathcal{N}} = \begin{pmatrix} \hat{\mathbf{L}}^{\mathcal{N}} \\ 0 \end{pmatrix}, \quad \mathbf{F}^{\mathcal{N}} = \begin{pmatrix} \hat{\mathbf{F}}^{\mathcal{N}} \\ 0 \end{pmatrix}$ //the size-corrected level transitions according to packet arrival
11: $\boldsymbol{\pi}_u^{\mathcal{N}} = \boldsymbol{\pi} \begin{pmatrix} \mathbf{B}^{\mathcal{N}} & \mathbf{L}^{\mathcal{N}} & \mathbf{F}^{\mathcal{N}} & 0 & \cdots \\ \mathbf{B}^{\mathcal{N}} & \mathbf{L}^{\mathcal{N}} & \mathbf{F}^{\mathcal{N}} & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & \mathbf{B}^{\mathcal{N}} & \mathbf{L}^{\mathcal{N}} & \mathbf{F}^{\mathcal{N}} \\ \cdots & 0 & 0 & \mathbf{B}^{\mathcal{N}} & \mathbf{L}^{\mathcal{N}} \\ \cdots & 0 & 0 & 0 & \mathbf{B}^{\mathcal{N}} \end{pmatrix}$//the unnormalized initial distribution
12: $\boldsymbol{\pi}^{\mathcal{N}} = \frac{\hat{\boldsymbol{\pi}}^{\mathcal{N}}}{\hat{\boldsymbol{\pi}}^{\mathcal{N}}\mathbf{h}}$//the normalized initial distribution of the packet level model
13: $p_s = \boldsymbol{\pi}^{\mathcal{N}}\left(\mathbf{I} - \mathbb{P}^{\mathcal{R}}\right)^{-1}\mathbf{s}, \quad p_\ell = \boldsymbol{\pi}^{\mathcal{N}}\left(\mathbf{I} - \mathbb{P}^{\mathcal{R}}\right)^{-1}\boldsymbol{\ell}$ //the solution of the packet level model
14: return $p_s, p_\ell$
```

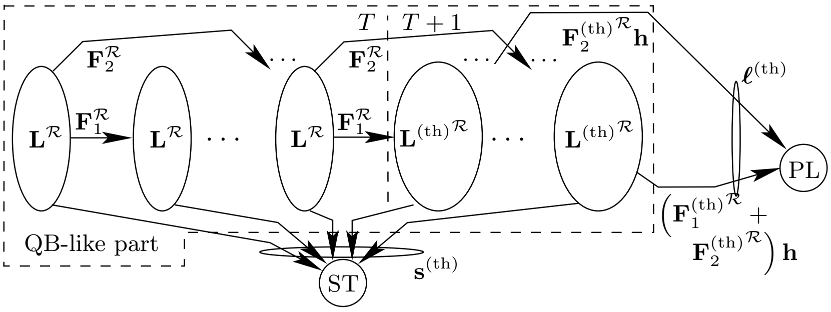

T
$T+1$
$\mathbf{F}_2^{\mathcal{R}}$
$\mathbf{F}_2^{\mathcal{R}}$
$\mathbf{F}_2^{(\mathrm{th})\mathcal{R}}\mathbf{h}$
$\boldsymbol{\ell}^{(\mathrm{th})}$
$\mathbf{L}^{\mathcal{R}}$
$\mathbf{F}_1^{\mathcal{R}}$
$\mathbf{L}^{\mathcal{R}}$
$\mathbf{L}^{\mathcal{R}}$
$\mathbf{F}_1^{\mathcal{R}}$
$\mathbf{L}^{(\mathrm{th})\mathcal{R}}$
$\mathbf{L}^{(\mathrm{th})\mathcal{R}}$
PL
$\left(\mathbf{F}_1^{(\mathrm{th})\mathcal{R}} + \mathbf{F}_2^{(\mathrm{th})\mathcal{R}}\right)\mathbf{h}$
QB-like part
$\mathbf{s}^{(\mathrm{th})}$
ST

Fig. 5. The transient DTMC modeling the VOQ with packet rejection during the life cycle of a packet

two parts along the threshold. It is marked in Figure 5 by a vertical line between the Tth and the $T+1$st level. There is a horizontal line mark the threshold in the state transition probability matrix of the cell and packet level models in lines 3 and 10 of Algorithm 3 respectively and in the block matrix in line 14 of Algorithm 3.

The steady state solution of the cell level model is computed in line 3 of Algorithm 2 for the model without packet rejection and in line 4 of Algorithm 3 for the model with packet rejection. Here we note that contrary to [6] which proposes a Folding algorithm [9] based method for the steady state solution of the cell level model a more effective numerical solutions can be applied for such a Markov chain in both cases with/without packet rejection. Both $\mathbb{P}$ and $\mathbb{P}^{(\mathrm{th})}(T)$ are skip-free to the left (upper Hessenberg matrix) with regenerative structure (during backward level transition the phase process regenerates). An effective numerical solution method of this kind of QBD-like Markov chain can be found in [10, 11].

4 Computational study

In this section we study the joint I-CS packet loss probability of the switch as a function of the CSSs' buffering threshold (T) by the consecutive execution of Algorithm 3 for all $T \in [0, b]$. The analytical results are also verified by simulations using our own simulator.

The simulator is written in `c++` and acts exactly as the specifications of the load-balancing switch, i.e., it plays its operation during the predefined simulation runtime. The runtime is set up such that the observed parameters does not change within a confidence interval, i.e., the statistical error are kept within that value.

In correspondence with [6], from which the present model is deduced, there are identical input processes assumed. The computational studies, given here, are drawn using the parameters of Table 2. The joint I-CS loss probability results are determined for the input 1 - VOQ_{00} - output 0 traversing path.

Algorithm 3 `Loss Minimizing Model`$(N = 3, b, \hat{p}, \hat{q}, \hat{t}, \{i, j, k\} = \{1, 0, 0\}, T)$, the packet minimizing model of the 3×3 LB switch

```
INPUT: $N = 3, b, \hat{p}, \hat{q}, \hat{t}, \{i, j, k\} = \{1, 0, 0\}, T$
OUTPUT: $p_{\text{I-CS}}(T)$ //the buffering threshold dependent, joint input-central stage packet loss probability
 1: $(\mathbf{B}, \mathbf{L}, \mathbf{F}_1, \mathbf{F}_2, p, q) =$ Level Transitions$(N, M = N, \hat{p}, \hat{q}, \hat{t})$//using Algorithm 1
 2: $(\mathbf{B}^{(\mathrm{th})}, \mathbf{L}^{(\mathrm{th})}, \mathbf{F}_1^{(\mathrm{th})}, \mathbf{F}_2^{(\mathrm{th})}, p, q) =$ Level Transitions$(N, M = N, \hat{p}, \hat{q} = 0, \hat{t})$//using Algorithm 1
 3: $\mathbb{P}^{(\mathrm{th})}(T) = \left(\begin{array}{ccccccccc} \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \dots & 0 & \dots & 0 \\ \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & \mathbf{B} & \mathbf{L} & \mathbf{F}_1 & \mathbf{F}_2 & 0 & \dots & 0 \\ \hline \dots & 0 & 0 & \mathbf{B}^{(\mathrm{th})} & \mathbf{L}^{(\mathrm{th})} & \mathbf{F}_1^{(\mathrm{th})} & \mathbf{F}_2^{(\mathrm{th})} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & 0 & \mathbf{B}^{(\mathrm{th})} & \mathbf{L}^{(\mathrm{th})} & \mathbf{F}_1^{(\mathrm{th})} & \mathbf{F}_2^{(\mathrm{th})} \\ 0 & \dots & 0 & \dots & 0 & 0 & \mathbf{B}^{(\mathrm{th})} & \mathbf{L}^{(\mathrm{th})} & \mathbf{F}_1^{(\mathrm{th})\prime} \\ 0 & \dots & 0 & \dots & 0 & 0 & 0 & \mathbf{B}^{(\mathrm{th})} & \mathbf{L}^{(\mathrm{th})\prime} \end{array}\right)$ //the cell level model
 4: $\boldsymbol{\pi}^{(\mathrm{th})}(T)\mathbb{P}^{(\mathrm{th})}(T) = \boldsymbol{\pi}^{(\mathrm{th})}(T), \quad \boldsymbol{\pi}^{(\mathrm{th})}(T)\mathbf{h} = 1.$ //the solution of the cell level model
 5: $p_i(T) = \sum_{i=b-(N+1)T+1}^{b} \pi_i^{(\mathrm{th})}$//packet drop probability at the input
 6: $(\mathbf{B}, \mathbf{L}, \mathbf{F}, p, q) =$ Level Transitions$(N - 1, M = N, \hat{p}, \hat{q}, \hat{t})$//using Algorithm 1
 7: $\mathbf{L}^{\mathcal{R}} = (1-p)^3\,\mathbf{B}, \quad \mathbf{F}_1^{\mathcal{R}} = (1-p)^3\,\mathbf{L}, \quad \mathbf{F}_2^{\mathcal{R}} = (1-p)^3\,\mathbf{F}$//the arrival based decomposition of the aggregate process of the two non-observed and the observed input during 3 time slots
 8: $(\mathbf{B}^{(\mathrm{th})^{\mathcal{R}}}, \mathbf{L}^{(\mathrm{th})^{\mathcal{R}}}, \mathbf{F}^{(\mathrm{th})^{\mathcal{R}}}, p, q) =$ Level Transitions$(N - 1, M = N, \hat{p}, \hat{q} = 0, \hat{t})$
 9: $\mathbf{L}^{(\mathrm{th})^{\mathcal{R}}} = (1-p)^3\,\mathbf{B}^{(\mathrm{th})^{\mathcal{R}}}, \quad \mathbf{F}_1^{(\mathrm{th})^{\mathcal{R}}} = (1-p)^3\,\mathbf{L}^{(\mathrm{th})^{\mathcal{R}}}, \quad \mathbf{F}_2^{(\mathrm{th})^{\mathcal{R}}} = (1-p)^3\,\mathbf{F}^{(\mathrm{th})^{\mathcal{R}}}$//the same for packet rejection
10: $\mathbb{P}^{(\mathrm{th})^{\mathcal{R}}}(T) = \left(\begin{array}{cccccccc} \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & \mathbf{L}^{\mathcal{R}} & \mathbf{F}_1^{\mathcal{R}} & \mathbf{F}_2^{\mathcal{R}} & 0 & \dots & 0 \\ \hline \dots & 0 & 0 & \mathbf{L}^{(\mathrm{th})^{\mathcal{R}}} & \mathbf{F}_1^{(\mathrm{th})^{\mathcal{R}}} & \mathbf{F}_2^{(\mathrm{th})^{\mathcal{R}}} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & 0 & \mathbf{L}^{(\mathrm{th})^{\mathcal{R}}} & \mathbf{F}_1^{(\mathrm{th})^{\mathcal{R}}} & \mathbf{F}_2^{(\mathrm{th})^{\mathcal{R}}} \\ 0 & \dots & 0 & \dots & 0 & 0 & \mathbf{L}^{(\mathrm{th})^{\mathcal{R}}} & \mathbf{F}_1^{(\mathrm{th})^{\mathcal{R}}} \\ 0 & \dots & 0 & \dots & 0 & 0 & 0 & \mathbf{L}^{(\mathrm{th})^{\mathcal{R}}} \end{array}\right)$
    ${\boldsymbol{\ell}^{(\mathrm{th})}}^{\mathsf{T}}(T) = \left( 0 \dots 0 \,\middle|\, 0 \dots 0 \;\mathbf{F}_2^{(\mathrm{th})^{\mathcal{R}}}\mathbf{h} \;\left(\mathbf{F}_1^{(\mathrm{th})^{\mathcal{R}}} + \mathbf{F}_2^{(\mathrm{th})^{\mathcal{R}}}\right)\mathbf{h} \right)$//the state transition probability matrix of the QB-like part and the absorption vector to state PL
11: $\tilde{q} = 1 - q$
12: $\hat{\mathbf{B}}^{\mathcal{N}} = \left(1 - \tilde{q}^3\right)\mathbf{B}, \quad \hat{\mathbf{L}}^{\mathcal{N}} = \left(1 - \tilde{q}^3\right)\mathbf{L}, \quad \hat{\mathbf{F}}^{\mathcal{N}} = \left(1 - \tilde{q}^3\right)\mathbf{F}$//the level transitions according to packet arrival during 3 time slots using the results of line 6
13: $\mathbf{B}^{\mathcal{N}} = \left(\begin{smallmatrix} \hat{\mathbf{B}}^{\mathcal{N}} \\ 0 \end{smallmatrix}\right), \quad \mathbf{L}^{\mathcal{N}} = \left(\begin{smallmatrix} \hat{\mathbf{L}}^{\mathcal{N}} \\ 0 \end{smallmatrix}\right), \quad \mathbf{F}^{\mathcal{N}} = \left(\begin{smallmatrix} \hat{\mathbf{F}}^{\mathcal{N}} \\ 0 \end{smallmatrix}\right)$ //the size-corrected level transitions according to packet arrival
14: $\boldsymbol{\pi}_u^{(\mathrm{th})^{\mathcal{N}}}(T) = \boldsymbol{\pi}^{(\mathrm{th})}(T) \left(\begin{array}{ccccccc} \mathbf{L}^{\mathcal{N}} & \mathbf{F}_1^{\mathcal{N}} & \mathbf{F}_2^{\mathcal{N}} & 0 & 0 & \dots & 0 \\ \mathbf{L}^{\mathcal{N}} & \mathbf{F}_1^{\mathcal{N}} & \mathbf{F}_2^{\mathcal{N}} & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & \mathbf{L}^{\mathcal{N}} & \mathbf{F}_1^{\mathcal{N}} & \mathbf{F}_2^{\mathcal{N}} & 0 & \dots \\ \hline 0 & 0 & 0 & \dots & 0 & 0 & 0 \end{array}\right)$ //the unnormalized initial distribution
15: $\boldsymbol{\pi}^{(\mathrm{th})^{\mathcal{N}}}(T) = \frac{\boldsymbol{\pi}_u^{(\mathrm{th})^{\mathcal{N}}}(T)}{\boldsymbol{\pi}_u^{(\mathrm{th})^{\mathcal{N}}}(T)\mathbf{h}}$ //the initial distribution of the packet level model
16: $p_\ell(T) = \boldsymbol{\pi}^{(\mathrm{th})^{\mathcal{N}}}(T)\left(\mathbf{I} - \mathbb{P}^{(\mathrm{th})^{\mathcal{R}}}(T)\right)^{-1}\boldsymbol{\ell}^{(\mathrm{th})}(T).$//the CS packet loss probability
17: $p_{\text{I-CS}}(T) = p_i(T) + \big(1 - p_i(T)\big)p_\ell(T)$//the joint input - CS loss probability
18: return $p_{\text{I-CS}}(T)$
```

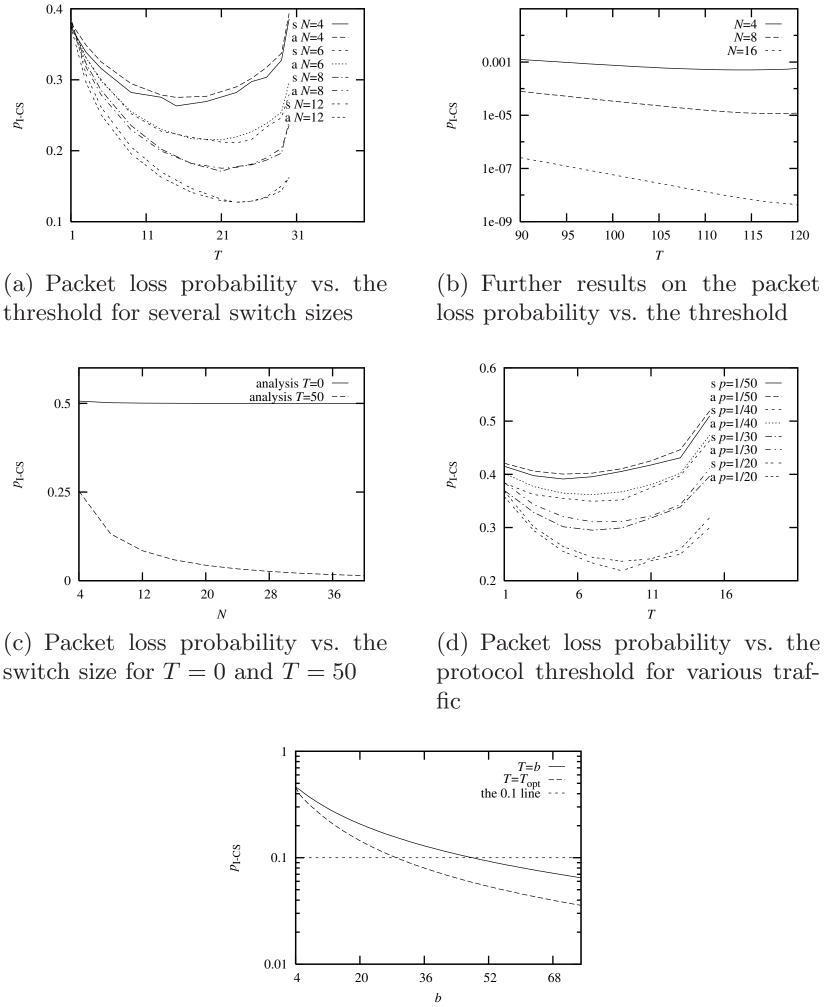
Table 2. Parameters used for the numerical studies

Figure	6(a)	6(b)	6(c)	6(d)	6(e)
N	$4,\dots,12$	$\{4,8,16\}$	$4,\dots,40$	4	4
b	30	120	50	15	$4\dots75$
T	$1,\dots,30$	$90,\dots,120$	$\{0,50\}$	$1,\dots,15$	various
$\hat{p}$	$\frac{1}{50}$	$\frac{1}{30}$	$\frac{1}{40}$	$\frac{1}{20},\dots,\frac{1}{50}$	$\frac{1}{15}$
$\hat{q}$	$\frac{9}{10}$	$\frac{1}{11}$	$\frac{9}{10}$		
$\hat{t}$	$\frac{1}{N}$				

The first experiment focuses on the threshold and switch size dependency of the optimal packet loss. In particular Figure 6(a) and 6(b) shows the dependency of the joint I-CS packet loss probability on the threshold value for several switch sizes and Figure 6(c) shows the dependency on the switch size for $T = \{0, 50\}$. The parameters used for packet loss evaluation are listed in Table 2. If the threshold is around 0, the input packet loss has the main impact on the joint packet loss. Basically the protocol is dropping most of the packets arriving to the inputs since none of the central stage buffers is allowed to be used for packets forwarding. Indeed, the loss value is almost independent of the switch size (see curve $T = 0$ in Figure 6(c)). Obviously when the threshold at the central stage is equal to the buffer size b the switch is operating in the traditional way (without protocol support) and the joint packet loss is composed only of the loss obtained due to the central stage buffers congestion. Finally, moving T in $[0, b]$ we can determine the threshold for which the joint packet loss probability is minimal.

Since the results were performed for different switch sizes it is also possible to see in Figures 6(a) and 6(b) how the optimal threshold (T_{opt}) of the minimal packet loss moves towards b as the switch size increases. The threshold aims to reduce the wasted capacity at the central stage. If the loss probability at the CS is high the introduction of $T < b$ reduces the amount of waste cells at VOQs. The higher the loss probability is the lower T results in the minimal joint I-CS loss. On the other hand the growth of the switch size results in larger system capacity and accordingly lower CS packet loss probability [4,6]. These two effects moves T_{opt} towards b with the increase of the switch size. From a given point on the CS packet loss probability decreases slowly with T, and from this point the packet loss at the input becomes dominant. In Figure 6(b) there is also demonstrated how the analysis performs in case of large buffers with low packet loss probability while in the other cases the phenomena are demonstrated with lower buffer capacity and accordingly higher loss probability values.

Figure 6(d) shows the joint I-CS packet loss probability, determined by the model and the simulations, versus the threshold. In this experiment we focus on the behavior of the system when various types of traffic matrices appear at the inputs. The set of parameters used for the experiment are given in Table 2. According to the obtained results, and also to our expectations, with the growth of the average packet size the joint packet loss of the system also increases. Figure 6(d) reflects to the fact that not only the system capacity plays significant role in the central stage loss probability but the average packet size too. If the



(a) Packet loss probability vs. the threshold for several switch sizes

(b) Further results on the packet loss probability vs. the threshold

(c) Packet loss probability vs. the switch size for $T = 0$ and $T = 50$

(d) Packet loss probability vs. the protocol threshold for various traffic

(e) An example of the gained buffer capacity using the threshold

Fig. 6. Numerical studies comparing the analytical results with simulations for the determination of the optimal threshold

average packet size is larger compared to the switch size the CS packet loss probability also increases. Similarly to the previous experiment the higher the CS packet loss probability is the lower T results in the minimal joint I-CS loss.

Our last study shows how the buffering threshold can be used to save buffer capacity. In Figure 6(e) there are two curves one of them plotting the packet loss values for the optimal threshold setting and the other of them the packet loss without the packet rejection policy. The intersection of the 10 % line for the tradition switch, without packet rejection, is at $b = 48$ while the same for the switch with packet rejection is at $b = 30$, i.e., by the use of the packet rejection protocol there is more than $\frac{1}{3}$ of the buffer capacity saved while the packet loss probability kept on the same level.

5 Conclusions

In this paper we presented a service protocol which allows to calculate and configure the LB switch in order to obtain the minimal joint packet loss probability of the input and central stage buffers. Using the protocol one can decrease the wasted capacity of load-balancing switch and accordingly the reassembly delay as well as the power equipment of the reassembly unit under some circumstances.

During the computational studies we have shown the experiments on the finding of threshold for the minimal packet loss probability. We have also given the explanations of three interesting phenomena, how the switch size and the load of the switch affects the threshold value at which the minimal joint I-CS loss probability is gained and how can the introduction of the packet rejection threshold reduce the buffer capacity needed to keep the packet loss probability on a predefined level.

Acknowledgements The authors would like to thank the constructive criticism of the reviewers which helped to improve the quality of the paper.

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. 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
4. Audzevich, Y., Bodrog, L., Telek, M., Ofek, Y.: Variable Size Packets Analysis in Load-balanced Switch with Finite Buffers. Technical report, TUB (2009)
5. Audzevich, Y., Bodrog, L., Telek, M., Ofek, Y.: Packet loss analysis of load-balancing switch with ON/OFF input processes. In: EPEW '09, London, UK (July 2009)
6. Audzevich, Y., Bodrog, L., Telek, M., Ofek, Y.: Scalable model for packet loss analysis of load-balancing switches with identical input processes. In: IEEE ASMTA '09, Madrid, Spain (June 2009)
7. Floyd, S., Jacobson, V.L.: Random early detection gateways for congestion avoidance. IEEE/ACM Transactions on Networking **1**(4) (1993) 397–413
8. Floyd, S., Gummadi, R., Shenker, S.: Adaptive RED: An algorithm for increasing the robustness of RED's active queue management (August 2001)
9. 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
10. Van Velthoven, J., Van Houdt, B., Blondia, C.: The impact of buffer finiteness on the loss rate in a priority queueing system. In: EPEW '06, Budapest, Hungary (June 2006)
11. Ishizaki, F.: Numerical method for discrete-time finite-buffer queues with some regenerative structure. Stochastic models **18**(1) (2002) 25–39