Heuristic performance model of optical buffers for variable length packets

W. Rogiest[1], K. Laevens, S. Wittevrongel, H. Bruneel

Department of Telecommunications and Information Processing, Ghent University (UGent)
St.-Pietersnieuwstraat 41, B-9000 Ghent, Belgium

.

Postprint – The final publication is available at
http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s11107-013-0409-z

Abstract

Optical switching (Optical Packet Switching, Optical Burst Switching, and others) provides alternatives to the current switching in backbone networks. To switch optically, also packet buffering is to be done optically, by means of Fiber Delay Lines (FDLs). Characteristic of the resulting optical buffer is the quantization of possible delays: only delays equal to the length of one of the FDLs can be realized.

An important design challenge is the optimization of the delay line lengths for minimal packet loss. To this end, we propose a heuristic based on two existing queueing models: one with quantization and one with impatience. Combined, these models yield an accurate performance modeling heuristic. A key advantage of this heuristic is that it translates the optical buffer problem into two well-known queueing problems, with accurate performance expressions available in the literature. This paper presents the heuristic in detail, together with several figures, comparing the heuristic's output to existing approaches, validating its high accuracy.

Keywords: stochastic analysis, performance evaluation, queueing theory, optical buffers, FDL buffers
2000 MSC: 68M20, 90B22

1. Introduction

Offering a network of ubiquitous broadband connectivity, operators are preparing for the impact of the ever-growing bandwidth hunger of their customers. The widespread interest in cloud computing and big data applications is rapidly pushing the network to its capacity limit.

For the current backbone network, the true capacity limit is not fiber throughput, but rather switching speed. Current fiber connections provide throughput of well beyond 10 Tbit/s per fiber (or even 1 Pbit/s [1]), but this capacity is only available for transmission from node to node. Within the nodes, switching is done through Optical Circuit Switching (OCS), with circuits configured beforehand. As an alternative, packet switching would provide much better utilization of the (enormous) fiber capacity, through the statistical sharing it achieves (that is, multiplexing gain). This difference is spectacular especially when the packet switching concerns small packets (like IP packets), as is envisaged by "pure" Optical Packet Switching (OPS) [2], but also occurs for Optical Burst Switching (OBS) [3].

Regardless of the exact nature of the switching technology (OPS or OBS), a crucial point of debate is the buffering. To date, Fiber Delay Line (FDL) remains the medium of choice for optical buffering, and this, mainly because FDL buffers are the only optical buffers that are cheap and reliable at the same time. Currently, research for "pure" OPS with FDLs steadily gains momentum, see [4, 5]. Although both FDL buffers and classic RAM (Random Access Memory) buffers serve as a means of buffering, they differ fundamentally.

First difference with RAM buffers is the footprint: a microsecond of delay requires about 200

[1]Corresponding author. wrogiest@telin.ugent.be

m of fiber. Large size being a drawback for optical switching, the number of FDLs is always kept low in practical implementations. This results in buffers with capacity in the order of kB to MB, as opposed to RAM buffers, with capacities in the order of GB. (However, as argued in [6], RAM buffer capacity is typically overprovisioned.)

Second, rather than being able to let a packet or burst wait for an arbitrary time period, an optical buffer can only provide delays that fall within the quantized set of delay line values. Assignable delays can thus be brought back to the physical lengths of the fiber delay lines. This is opposed to the situation in a RAM buffer, where data can reside in the buffer for any period of time, resulting in completely different queueing behavior.

Third, rather than being able to schedule a given amount of bits (in space, like RAM), FDL buffers are able to guarantee a certain period of delay (in time). The upper bound to this delay is called the impatience of the buffer (denoted τ, see further).

Given the specific nature of the optical buffer, an important aspect in design is the choice of the delay line lengths. These can be used in a feed-forward setup (as opposed to a feed-back setup, see, e.g., [7]), where each line provides the delay that corresponds to its length. The lengths in a feed-forward setup are usually assumed multiples of a basic delay unit called granularity [8]. In a network with fixed-length packets, a natural choice is to let the granularity match the packet length, although this is not necessarily the optimal choice [9]. Of particular interest to this work is the case of asynchronous variable length packets, where such a "natural choice" is not readily available, and the line length optimization problem involves both the buffer size and traffic load. The often-cited letter [8] accounts for the prime contribution treating optical buffering, relying on an iterative procedure defined on a classic queueing model, allowing for intuitive reasoning on the trade-off between quantization and impatience. However, the solution of [8] (and its extended versions, [10] and [11]) provided an approximate solution that is inaccurate in some cases, as discussed below. An enhanced approach with Markov chains enabled accurate numerical results [12] as well as a solution in closed form [13], but does not allow for intuitive insight in the queueing behavior of the system.

In this paper, we present an approximative heuristic for the calculation of the loss probability of an M/M/1 optical FDL buffer system with finite size, both in continuous time and in discrete time. In continuous time, this corresponds to a Poisson arrival process and exponential packet length. In discrete time, this corresponds to a Bernoulli arrival process and geometric packet length. The heuristic aims at providing a more accurate alternative to the approach presented in [8, 10]. At the same time, by combining classic queueing models in a very straightforward manner, the heuristic is easy to understand, providing a simple engineering tool in the design of complicated optical packet/burst switches.

The remainder of this paper is structured as follows. In Section 2, we consider the prerequisites for the analysis: a model for quantization, combined with a model for impatience. In Section 3, the analysis is presented. In Section 4, the heuristic's output is compared to an existing approximate approach as well as exact numerical results, allowing to assess the validity and accuracy of our approximation. Conclusions are drawn in Section 5. Finally, the Appendix provides the analysis in case of a discrete-time setting.

In the context of this work, the words "packet" and "burst" are interchangeable, since results are generically applicable to both OPS and OBS. Below, the term "packet" is used.

2. Method

In this section, we set out the queueing models needed as input to the heuristic. We subsequently look at the modeling of quantization and impatience, and then to the heuristic itself.

2.1. Quantization Model

Regardless of time setting, a model with quantization of the delay is characterized by the fact that not any delay is available. In most cases, the FDL buffer is studied for a *degenerate* buffer setting. In a degenerate setting, FDL lengths equal to multiples of some basic value called *granularity*, denoted D [8]. Buffers of this type contain $(N + 1)$ FDLs with lengths $i \cdot D$, for $i = 0, \ldots, N$. Since this setting is known to be optimal in many (but not all) cases (for discussion, see [9]), it is also assumed in this work.

The previously mentioned [8] presents the first approximate model for evaluation of FDL buffer performance. In particular, it focuses on the effect of quantization (referred to as granularity), so

allowing to evaluate the performance of an M/M/1 optical buffer, by means of an iterative procedure. Here, M/M/1 is the well-known Kendall notation for a queueing system with a Poisson arrival process, exponential service times (here, packet sizes), and a single output for service (here, a single wavelength). The approach in [8], and in its extended version [10], is to introduce an excess packet length (there, denoted θ_e), that is in general larger than the regular packet length, and so incorporates the effect of quantization. On the other hand, the approach takes into account the limited size of the buffer by means of a bound on the sum of accumulated packet lengths. As such, the effects of quantization (excess packet length) and impatience (bound on the accumulated sum of packet lengths) are treated separately, as in the current contribution. The main difference is that we do not use the concept of an excess packet length, as this leads to an inaccurate definition of the equivalent load, as discussed below [see (6)].

In [14] (discrete time) and [15] (continuous time), the effect of quantization was traced in an exact manner, for a degenerate buffer setting with infinite buffer size. The effects of impatience were also taken into account, but only by means of an indirect approximate approach, which was hardly open for intuition, and inaccurate if the number of lines is few (say, smaller than 5). More precisely, the approach was based on the asymptotic of the tail probabilities of the waiting time distribution. Such approximation indeed (partially) reflects impatience, but loses accuracy if the tail probabilities are evaluated for small delays (and, related, small buffer sizes).

Opposed to this, in [12], the combined effect of quantization and impatience was modeled in an exact manner, and a numerical procedure was provided for exact performance evaluation. Further, in [13], this numerical framework provided the starting point to extract exact closed-form performance expressions, for optical buffers in an M/D/1 and an M/M/1 setting, respectively. While these contributions capture the system by a single Markov chain model, they do not allow for any distinction between the separate characteristics of impatience and quantization. Rather, both effects are implied by the system description in an entwined manner, and the interplay among these effects remains hidden.

2.2. Impatience Model

Classic impatience models were developed in a continuous-time setting, mostly in a setting with a single server (in the context of optical buffers: a single wavelength). An elegant approach to impatience is provided by a paper by Barrer [16]. There, the delay or waiting time (and the associated queue length) can in principle take on every value between 0 and some upper bound τ, which is the impatience associated with the system. This can be a random variable, but it can also have a simpler nature, a fixed parameter. Regardless of the nature of τ, the number of available waiting times in an impatient system falls within the interval $[0,\tau]$. In case of a continuous-time setting, this interval is a continuum, with a non-denumerable (infinite) number of possible delays. When considered in a discrete-time setting [17], this interval translates into a finite set of possible waiting times $\{0, 1, \ldots, \tau\}$. The continuous-time case is treated below; for the discrete-time case, we refer the reader to the Appendix.

2.3. Heuristic

The proposed heuristic unifies two seemingly disparate elements: recent results for systems with quantization of delay [14] on the one hand, and a well-known result for systems with impatience [16] on the other hand. Main advantage is that this approach is "simple", in the sense that it is complex only to the degree needed to model the most characteristic features of the original system, and nothing more.

In the remainder of this paper, the approach will remain complementary: we first trace the effects of quantization and impatience separately, without any entwinement. For quantization, we will rely on a measure from the infinite-size system, that proves a useful and accurate approximation for the finite system. For impatience, we will consider an exact model. Both elements are combined in a third step, to yield an encompassing model of impatience, with the quantization embedded within, as a single parameter, called the equivalent load.

3. Analysis

In this section, we subsequently analyze quantization and impatience, for the case of an optical buffer in an M/M/1 setting, in continuous time (here), and in discrete time (see Appendix). For the

former time setting, closed-form performance measures have been obtained recently [13], so providing an "ideal" reference when we assess the heuristic's accuracy.

3.1. Assumptions

We study a single outgoing channel, where contention is resolved by means of a degenerate FDL buffer, situated at the output port of the optical switch. Each incoming packet is routed to the shortest of these FDLs such that the packet will not overlap on departure with packets from the other FDLs. If such an FDL cannot be found, the packet is lost. The chance for this to occur is the loss probability (LP), a key performance measure, which we study in detail in this paper. While each packet travels through its assigned delay line only once, several packets may be traveling through the same delay line at the same time (without overlapping, however). Further, note that there is no formal limitation on the length of packets that can be accepted.

In the continuous-time setting we assume, all events take place in an asynchronous fashion, and time-related variables like inter-arrival times and packet sizes can take on any positive real value. Packets are assumed to arrive one by one; upon arrival, a packet is either accepted or lost. Numbering the packets in the order at which they arrive with index k, we associate with packet k an inter-arrival time $T_k \in \mathbb{R}^+$, that captures the time between the kth arrival and the next. The inter-arrival times form a sequence of independent and identically distributed (iid) random variables with common cumulative distribution function $T(x)$,

$$T(x) = \Pr[T_k \leq x] = 1 - e^{-\lambda x}\,,\; x \in \mathbb{R}^+\,, \qquad (1)$$

where $\lambda \in \mathbb{R}^+$ denotes the arrival intensity such that $\mathrm{E}[T_k] = 1/\lambda$.

With each packet, we associate a packet size B_k. The packet sizes also form a sequence of iid random variables with a common cumulative distribution function $B(x) = \Pr[B_k \leq x]$, $x \in \mathbb{R}^+$. The nature of this distribution is also exponential, but now with parameter μ, such that $\mathrm{E}[B_k] = 1/\mu$.

3.2. Quantization

To capture the effect of quantization, we can limit ourselves to coming up with a good definition of an equivalent load, denoted ρ_{eq}, that incorporates the effects of quantization into an increased traffic load for the system in general. Still using the same numbering as above, we assume for now that the quantized buffer has an infinite number of delay lines, and therefore no impatience. This assumption allows to derive a simple characterization of the equivalent load that is independent of the actual number of delay lines.

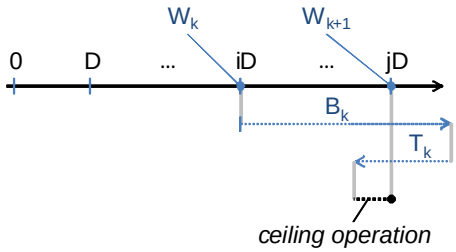

Figure 1: Illustration of the waiting time evolution. In this example, $W_k = iD$, $W_{k+1} = jD$, and $j > i > 0$.

As discussed in [9], the most efficient way to tackle the analysis is by focusing on the evolution of assigned waiting times. In the infinite FDL buffer system, we associate waiting time W_k with the kth packet, and define it as the time between the acceptance of packet k, and the start of its transmission. We focus on the evolution of the waiting time, as illustrated in Fig. 1 for a specific example with $W_k = i$, $W_k = j$, and $j > i > 0$. Packet k has packet size B_k and assigned waiting time W_k (equal to iD in the figure). A period T_k later, packet $k+1$ requests for waiting time W_{k+1} (jD in the figure). In order not to collide with packet k, buffer control assigns a waiting time to packet $k+1$ that is larger than or equal to $W_k + B_k - T_k$, chosen from the FDL set $l \cdot D$ ($l = 0, 1, \ldots$). Inferring the waiting time of packet $k+1$ from this, irrespective of time setting, we obtain

$$W_{k+1} = \left[W_k + D \cdot \left\lceil \frac{B_k - T_k}{D} \right\rceil \right]^+ . \qquad (2)$$

Here, $\lceil x \rceil$ denotes the well-known ceiling operation, and $[x]^+$ denotes the operation $\max(0,x)$.

As a system equation, (2) allows to define the maximum load, on the one hand, and a simple definition for an equivalent load, on the other hand.

To define a maximum load, one can easily see that the drift of the random variable W_k in (2) is brought about by the term

$$D \cdot \left\lceil \frac{B_k - T_k}{D} \right\rceil .$$

It can be understood intuitively that the stability of this infinite-size queueing system depends critically on the drift of this component. For degenerate settings, stability is studied in detail in [14], and yields following stability condition,

$$\mathrm{E}\left[\left\lceil \frac{B_k - T_k}{D} \right\rceil\right] < 0\,. \tag{3}$$

Notice that the negative drift condition (3) is valid only for degenerate buffer structures. Its counterpart for non-degenerate buffer structures is more complicated, as discussed in [18].

In its turn, condition (3) corresponds to some bound ρ_{max} on the traffic load ρ, by definition given by

$$\rho = \frac{\mathrm{E}[B_k]}{\mathrm{E}[T_k]}\,.$$

Here, ρ is thus given by λ/μ. The bound ρ_{max} is more restrictive than the classic bound on stability ($\rho < 1$), and, as such,

$$\rho < \rho_{max} \leq 1\,.$$

For the equivalent load, we consider an altered definition of the load, that incorporates the effects of quantization for any value of the traffic load (not only the maximum load). To this end, we extend the notion of *drift*, by proposing the following definition for the equivalent load,

$$\rho_{eq} = 1 + \frac{\mathrm{E}\left[D \cdot \left\lceil \frac{B_k - T_k}{D} \right\rceil\right]}{\mathrm{E}[T_k]}\,, \tag{4}$$

which is very similar to the expression of the classic load, especially when the latter is written as

$$\rho = 1 + \frac{\mathrm{E}[B_k - T_k]}{\mathrm{E}[T_k]}\,.$$

Note however, that the proposed equivalent load does not simply replace the classic load in general, and should not be applied in this manner. For example, while it is well-known for an M/G/1 classic system in a continuous-time setting that the probability of finding an empty system upon arrival equals $1-\rho$, this is not the case for an optical system: neither $(1-\rho)$ nor $(1-\rho_{eq})$ provides an answer, and the probability in question can only be determined after full queueing analysis, as available in [14]. In this regard, the definition of ρ_{eq} proposed here is prone to discussion: $\hat{\rho}_{eq} = 1 - \Pr[W_k = 0]$ can indeed provide an alternative definition for an alternative equivalent load $\hat{\rho}_{eq}$, but leads to impractical expressions, whose form is highly sensitive to the assumptions on the inter-arrival time and packet size distributions. As such, the main advantage of this definition of the equivalent load is that it allows for a simple, closed-form definition, that does not require analysis of the complete system. For the current case of M/M/1 in continuous time, also treated in [15], this expression is

$$\rho_{eq} = 1 + \frac{\lambda D}{\mu + \lambda}\left(\frac{\lambda}{1 - e^{-\mu D}} + \frac{\mu}{1 - e^{+\lambda D}}\right)\,. \tag{5}$$

For the case of a general GI/G/1 optical buffer, an efficient formulation can be done in terms of Laplace-Stieltjes transforms. A general expression can be readily obtained from the expression available in [14], valid for discrete time, by applying a limit procedure, similar as is done in [15] for an optical M/G/1 setting.

Importantly, note that (5) is not consistent with the definition of the equivalent load proposed in [8, 10], and, later on, [11],

$$\frac{\rho}{1 - \frac{\mu D}{2}\rho}\,. \tag{6}$$

In the infinite-size system, the maximum tolerable load (and system stability) is defined exactly by requiring $\rho_{eq} < 1$ in (5), but not with (6). This may be a cause for the inaccuracy of the related model [10] in some cases, such as the one displayed in Fig. 4 (see below).

3.3. Impatience

To model impatience, we rely on the result obtained by Barrer [16]. In [16], a classic continuous-time M/M/1 model with impatience is considered. In such a classic model, no effect of quantization comes about, and any delay can be realized. As such, it can be seen as a limit situation of an optical buffer, with the granularity D considered infinitely small. As such, we come to study an optical buffer that still is degenerate (FDLs have lengths iD), that still has a finite maximum waiting time ($\tau = ND$), but that has infinitely small granularity, $D \to 0$, and an infinitely large number of FDLs, $N \to \infty$. Clearly, this system cannot actually be implemented, but serves as an instructive point of reference, that can be analyzed with the simple formula presented in [16].

Barrer obtains a closed-form expression for the loss probability (LP) of an arbitrary customer (or

granularity value D	0	0.2	0.5	1	2
N for $\tau = 4$	∞	20	8	4	2
N for $\tau = 20$	∞	100	40	20	10

Table 1: Parameter setting used for the comparison presented in Fig. 2 and 3.

packet) in an M/M/1 non-optical system with impatience, namely

$$\mathrm{LP} = \begin{cases} \dfrac{(1-\rho)\rho}{e^{\mu\tau(1-\rho)} - \rho^2} & , \rho \neq 1\,, \\ \dfrac{1}{\mu\tau + 2} & , \rho = 1\,, \end{cases} \tag{7}$$

where we used the notations μ, ρ and λ as introduced above.

3.4. Heuristic

As mentioned, Barrer's result can be seen as a limit situation of an optical buffer, with D infinitely small. As such, it can be brought in tight relation to an optical buffer's performance with the same impatience (or maximum delay) $\tau = N \cdot D$. The heuristic consists in using the equivalent load ρ_{eq} to incorporate the effect of quantization and using this expression as load measure in the formula of Barrer. The heuristic thus corresponds to a translation: the original system parameters N and D are translated into one parameter for quantization (ρ_{eq}, replacing ρ) and one for impatience (τ, equal to $N \cdot D$).

Before applying this approach for approximation purposes, we first examine the relation between exact results for optical buffers with impatience (with results from [13]), and exact results for classic (non-optical) systems (with results from the formula of Barrer). A comparison is provided in Fig. 2 and 3. For $D = 0$ (formula of Barrer), the curve corresponds to a classic (non-optical) M/M/1 system with deterministic impatience fixed to τ. Fixing the maximum achievable amount of delay $\tau = N \cdot D$ to 4 μs (Fig. 2) and to 20 μs (Fig. 3), the granularity values considered lead either to a finite amount of FDLs, or to the limit of an infinite number of FDLs, as displayed in Table 1. As a reference, all curves carry a diamond ($\diamond$) to indicate the load level at which the equivalent load, ρ_{eq}, equals one.

In Fig. 2, with $\tau = 4$ μs, it comes as no surprise that decreasing granularity (more FDLs) leads to significant performance improvement. Further, it shows that, even with τ fixed, the granularity has a paramount impact on performance, that remains visible even for granularity values as small as 0.2. Also, the point where ρ_{eq} turns one ($\diamond$) comes about as a reference point: for $0 < \rho < \rho_{eq}$, the LP grows fast with the traffic load (as reflected in quasi-linear curves on the log-lin scale applied), while for $\rho > \rho_{eq}$, loss grows slowly, with an asymptote at $(\rho - 1)/\rho$ for $\rho \to \infty$, with the effect of granularity gradually fading. This role of reference point comes about even more distinctly when we consider the curves for a larger achievable delay, $\tau = 20$ μs, in Fig. 3. The latter curves further confirm the major role of the granularity in performance evaluation, since even the case of $D = 0.2$ significantly differs from the limiting case with $D = 0$. However, for D even smaller, ($D \ll \mu^{-1}$, here, $D < 0.1$), the curves nearly overlap, as should. As such, it is clear that the impatience model without quantization is indeed identical to the optical buffer model with $D \to 0$.

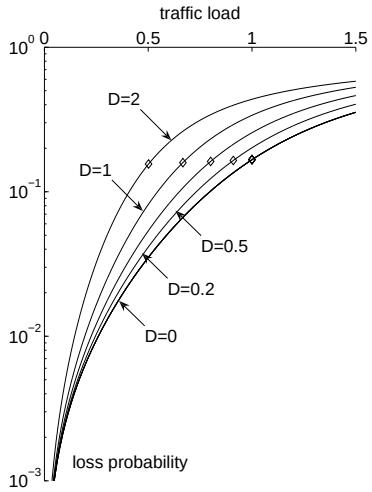

Figure 2: Loss probability vs. traffic load ρ. Comparison for impatience $\tau = 4$ μs and granularity $D \in \{0, 0.2, 0.5, 1, 2\}$ μs.

4. Numerical Comparison

To assess the accuracy of the proposed heuristic, we first compare its output to that of the existing approximate model proposed by Callegati [10], in Fig. 4, displaying the loss probability as function of the granularity D. Additionally, exact results obtained from [13] are displayed. The figure is ob-

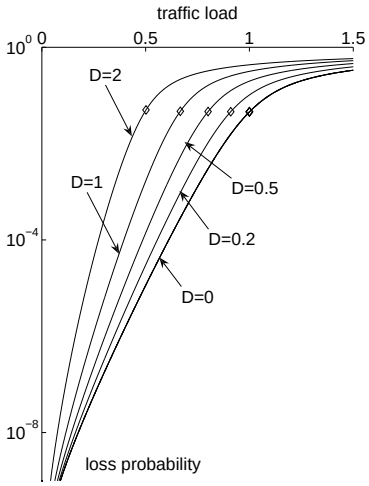

Figure 3: Same as Fig. 2, but now for impatience $\tau = 20$ μs.

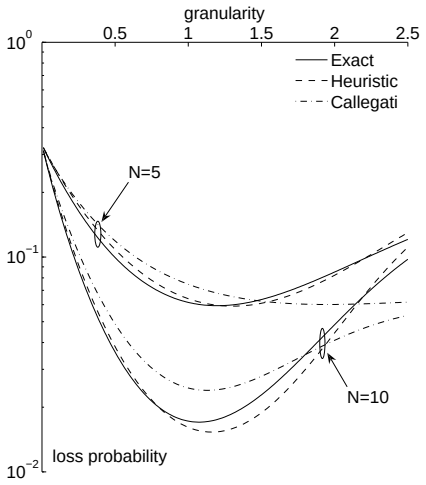

Figure 4: Loss probability vs. granurity D. Comparison of heuristic results to exact results, and the output of the approximate model by Callegati [10], for mean packet size $\mathrm{E}[B] = 1$ μs, load $\rho = 0.5$ and buffer size $N \in \{5, 10\}$.

tained for mean packet size $\mathrm{E}[B] = \mu^{-1} = 1\,\mu$s, $\rho = 0.5$ (and thus, $\lambda = 0.5$) and buffer sizes $N = 5$ and $N = 10$. For this setting, the equivalent load ρ_{eq} (given by (5), and thus, independent of N) increases from 0.5 (for $D = 0$) to 1 (for $D = 2.0101$) to 1.1193 (for $D = 2.5$). Comparing the heuristic's output to exact results, obviously, accuracy is high over the whole range of D and for both values of N. Assessing Callegati's model, two observations can be made. Firstly, the accuracy typically is lower than that of our heuristic. Secondly, accuracy quickly decreases for higher values of D ($D > 2$), whereas our heuristic remains accurate over the whole range of D. Comparing the results of the heuristic to those of Callegati's model, the discrepancy in accuracy may be (partially) attributed to the difference in the definition of the equivalent load, see (5) and (6).

In Fig. 5–7, the heuristic is again put to test, by comparing its output to results from the exact analysis of [13], all for mean packet size $\mu^{-1} = 1\,\mu$s. In Fig. 5, the LP is plotted as function of the granularity, for four cases of the traffic load, and buffer size $N = 20$. Apparently, the approximation allows for accurate results, especially for high traffic load, and not too large values for the granularity ($D \leq \mu^{-1}$). For variable packet sizes, this is exactly the zone of practical relevance. More precisely, as argued in [13], several reasons make it imperative to set the granularity to about half the expected value of the packet size, $D \leq (2 \cdot \mu)^{-1}$, in actual implementations. Indeed, inspection of Fig. 5 shows that the accuracy of our approximation is always good for granularity values near $D \leq (2 \cdot \mu)^{-1}$. Further, considering the optimal (minimal) value of each curve separately, the heuristic apparently underestimates the loss probability for lower traffic load (e.g., $\rho = 0.5$), while the opposite holds true for high traffic load (e.g., $\rho = 0.8$), for which the heuristic (very) slightly overestimates the loss probability.

To further assess the impact of traffic load, we consider the plots of Fig. 6, where the loss probability is displayed as function of the load, for three different values of the granularity, and $N = 20$. This further confirms that the approximation works best for small granularity, since the curves for $D = 0.5\,\mu s$ nearly coincide with the exact ones. For large granularity, the approximation is most accurate for high traffic load.

Finally, Fig. 7 illustrates the impact of the buffer size N, with traffic load fixed to $\rho = 0.8$. As can be seen, the accuracy is not influenced much by the buffer size N, as the curves for $N \in \{20, 40, 60, 80\}$ all confirm the accuracy of the heuristic.

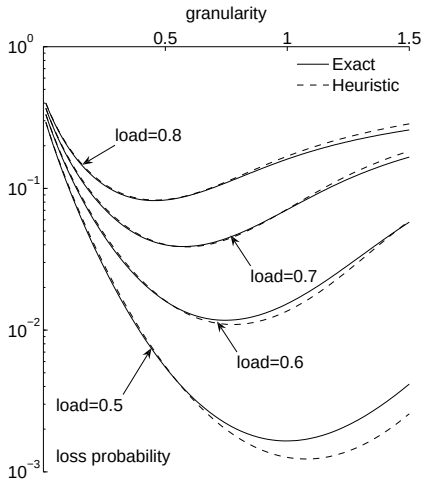

Figure 5: Loss probability vs. granularity D. Comparison of heuristic results to exact results, for buffer size $N = 20$, mean packet size $\mathrm{E}[B] = 1$ μs and load $\rho \in \{0.5, 0.6, 0.7, 0.8\}$.

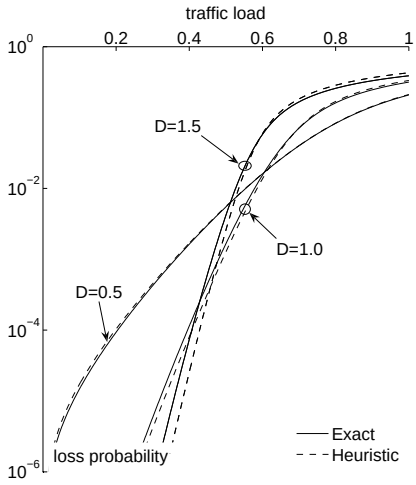

Figure 6: Same as Fig. 5, but here with varying traffic load instead of granularity, and with $D \in \{0.5, 1.0, 1.5\}$ μs.

5. Conclusions and Outlook

In this contribution, we presented a heuristic approach, allowing for simple performance analysis of optical buffers. The heuristic combines two existing queueing models: one with quantization, and one with impatience. As shown, for the considered M/M/1 FDL optical buffer, the heuristic yields accurate performance results. A key advantage of the heuristic is that it translates the optical buffer problem into two well-known queueing problems.

While the focus of the current contribution was solely on the M/M/1 optical buffer system, both for continuous-time and discrete-time setting, other cases would certainly deserve further exploration. This approach may be the only alternative when an exact approach becomes infeasible, for example when multiple wavelengths are available for service. Since accurate models for multi-server queues with impatience are available in literature, the main feat is to come up with a good modeling of quantization in a multi-wavelength setting, including a characterization of the maximum and the equivalent load in that case. This problem is part of ongoing and future work.

Acknowledgement

The first author is a Postdoctoral Fellow with the Research Foundation Flanders (FWO-Vlaanderen), Belgium. This research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

Appendix: the discrete-time case

While the results of the body of this contribution apply only to a continuous-time setting, this Appendix shows that minor changes suffice to make it applicable to a discrete-time setting. As in the analysis for continuous time, we assume a degenerate buffer setting, with line lengths equal to multiples of the granularity D. We specifically focus on the case of an M/M/1 optical buffer, now in discrete time.

In a discrete-time setting, events take place synchronously, at the beginning of time slots. Therefore, all time-related variables and performance measures are expressed as multiples of the slot length, and for example inter-arrival times and packet sizes take on only strictly positive integer values, contained in $\mathbb{N}_0$. The slot length may be arbitrary, and is therefore not mentioned explicitly in this Appendix.

A first point of attention is that, in discrete time, the relation between queues with quantization

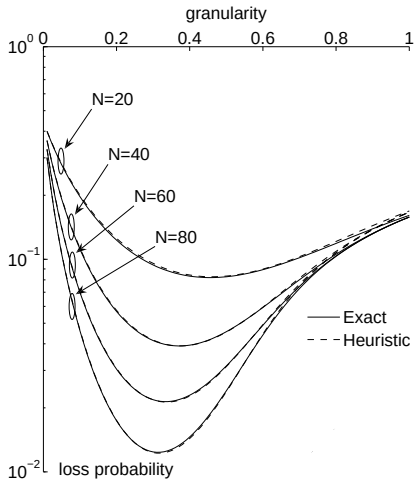

Figure 7: Same as Fig. 5, but here with load $\rho = 0.8$ and buffer size $N \in \{20, 40, 60, 80\}$.

and queues with impatience is tighter than in the continuous-time case. While in continuous time a queue with impatience was obtainable from the optical buffer by letting $D \to 0$, it suffices to set $D = 1$ in an optical buffer model to obtain a discrete-time model for queues with impatience. This has large impact on the way in which impatience is modeled. In the present context, it suffices to set $D = 1$ in all discrete-time results of [14], to obtain exact results for a general GI/G/1 system with deterministic impatience in discrete time. Contrasting this with results for continuous time, it is remarkable that even the less general M/G/1 system in continuous time of [19] requires numerical approximations in order to obtain results. Apparently, assuming a discrete-time setting somewhat simplifies the queueing problem with impatience.

Further, discrete-time is also the time setting studied in [17], where a general impatience distribution with upper bound r is assumed. This allows to handle a deterministic impatience distribution, by setting $r = \tau + 1$. (The term $+1$ is to be introduced to be compatible with the definition of r in [17]).

As for the traffic assumptions, we adopt the same indexing convention as for continuous time. For the arrival process, we assume a Bernoulli arrival process, which is the discrete-time counterpart of the Poisson arrival process. The inter-arrival times, a sequence of iid random variables, have a common geometric distribution with cumulative distribution function

$$T(n) = \Pr[T_k \leq n] = 1 - (1-p)^n \ , \ n \in \mathbb{N} , \quad (8)$$

with $p \in [0, 1]$. The latter probability is also the parameter of the geometric distribution, and gives the probability of having an arrival in an arbitrary slot, relating to the mean value, as $\mathrm{E}[T_k] = 1/p$. The packet sizes again form a sequence of iid random variables with geometric distribution, now with common probability mass function $b(n) = \Pr[B_k = n]$ and cumulative distribution function $B(n) = \Pr[B_k \leq n] = 1 - (1-f)^n$, $n \in \mathbb{N}$, with $f \in [0, 1]$, with mean value $\mathrm{E}[B_k] = 1/f$. As in continuous time, knowledge of the granularity D, $T(n)$ and $B(n)$ suffices as input for the analysis in discrete time.

For the modeling of impatience in discrete time, we rely on Section 3 of [17], reporting the probability that the age of the customer in service is zero (there denoted $\hat{\pi}_0$) equals

$$\hat{\pi}_0 = \frac{\bar{p}^r f(p-f)}{p^2 \bar{f}^r - \bar{p}^r f^2} ,$$

with $\bar{p} = 1 - p$, and $\bar{f} = 1 - f$. With now $\mathrm{LP} = 1 - (1 - \hat{\pi}_0)/\rho$ and $r = N + 1$, one easily obtains

$$\mathrm{LP} = \begin{cases} \dfrac{(1-\rho)\cdot\rho}{(\bar{p}/\bar{f})^{N+1} - \rho^2} & , \rho \neq 1 , \\ \dfrac{1}{(N+1)q/\bar{f} + 2} & , \rho = 1 , \end{cases} \quad (9)$$

with $\rho = p/f$. Note that this expression for the LP is tightly related to the one of continuous time (7). To see this, we rewrite the continuous-time expression for $\rho \neq 1$ as

$$\mathrm{LP} = \frac{(1-\rho)\rho}{e^{\mu\tau} e^{-\lambda\tau} - \rho^2} \ , \ \rho \neq 1 .$$

Now, a substitution similar to the one in continuous time is needed to yield correspondence,

$$\bar{p} = e^{-\lambda} , \ p = 1 - \bar{p} \ ; \ \ \bar{f} = e^{-\mu} , \ f = 1 - \bar{f} ,$$

completed with $\tau = N + 1$, to indeed obtain (9). The expression for $\rho = p/f = 1$ follows by taking the limit for $p \to f$. The link between discrete time and continuous time is less intuitive at one point, since τ is "virtually expanded" to $N + 1$ in discrete time, rather than N. The latter however forms no stumbling block: it can be understood as (an indirect) result of the difference in the minimum of the

inter-arrival times in discrete time and continuous time, 1 and 0, respectively.

Together with the expression for ρ_{eq} in discrete time, whose derivation is straightforward, (9) can be used for an approximate modeling of optical buffers. This is not treated further here, since the accuracy and the obtained associated figures are very similar to the continuous-time case.

References

[1] World record one petabit per second fiber transmission over 50-km, NTT press release 20 September 2012, http://www.ntt.co.jp/news2012/1209e/120920a.html.

[2] G. I. Papadimitriou, C. Papazoglou, A. S. Pomportsis, Optical switching: switch fabrics, techniques, and architectures, Journal of Lightwave Technology 21 (2) (2003) 384–405.

[3] Y. Chen, C. Qiao, X. Yu, Optical burst switching: A new area in optical networking research, IEEE Network 18(3) (2004) 16–23.

[4] E. F. Burmeister, J. P. Mack et al., Photonic integrated circuit optical buffer for packet-switched networks, Optics Express 17(8) (2009) 6629–6635.

[5] T. Tanemura, I. Soganci, T. Oyama, T. Ohyama, S. Mino, K. Williams, N. Calabretta, H. J. S. Dorren, Y. Nakano, Large-capacity compact optical buffer based on InP integrated phased-array switch and coiled fiber delay lines, Journal of Lightwave Technology 29 (4) (2011) 396–402.

[6] G. Appenzeller, I. Keslassy, N. McKeown, Sizing router buffers, Proceedings of the 2004 ACM Conference of the Special Interest Group on Data Communication, SIGCOMM 2004 (New York) (2004) 281–292.

[7] K.-H. Chou, W. Lin, An analytical model for all-optical packet switching networks with finite FDL buffers, Photonic Network Communications 25 (2013) 144–155.

[8] F. Callegati, Optical buffers for variable length packets, IEEE Communications Letters 4 (9) (2000) 292–294.

[9] W. Rogiest, J. Lambert, D. Fiems, B. Van Houdt, H. Bruneel, C. Blondia, A unified model for synchronous and asynchronous FDL buffers allowing closed-form solution, Performance Evaluation 66 (7) (2009) 343–355.

[10] F. Callegati, Approximate modeling of optical buffers for variable length packets, Photonic Network Communications 3 (4) (2001) 383–390.

[11] M. Yasuji, An approximation for blocking probabilities and delays of optical buffer with general packet-length distributions, Journal of Lightwave Technology 30 (1) (2012) 54–60.

[12] R. C. Almeida, J. U. Pelegrini, H. Waldman, A generic-traffic optical buffer modeling for asynchronous optical switching networks, IEEE Communications Letters 9 (2) (2005) 175–177.

[13] W. Rogiest, H. Bruneel, Exact optimization method for an FDL buffer with variable packet length, IEEE Photonics Technology Letters 22 (4) (2010) 242–244.

[14] W. Rogiest, K. Laevens, J. Walraevens, H. Bruneel, Analyzing a degenerate buffer with general inter-arrival and service times in discrete time, Queueing Systems 56 (3-4) (2007) 203–212.

[15] W. Rogiest, K. Laevens, D. Fiems, H. Bruneel, A performance model for an asynchronous optical buffer, Performance Evaluation 62 (1-4) (2005) 313–330.

[16] D. Y. Barrer, Queuing with impatient customers and ordered service, Operations Research 5(5) (1957) 650–656.

[17] J. Van Velthoven, B. Van Houdt, C. Blondia, On the probability of abandonment in queues with limited sojourn and waiting times, Operations Research Letters 34(3) (2006) 333–338.

[18] W. Rogiest, E. Morozov, D. Fiems, K. Laevens, H. Bruneel, Stability of single-wavelength optical buffers, European Transactions on Telecommunications 3 (21) (2010) 202–212.

[19] W. Xiong, D. Jagerman, T. Altiok, M/G/1 queue with deterministic reneging times, Performance Evaluation 65 (2008) 308–316.