Centrum voor Wiskunde en Informatica

Probability, Networks and Algorithms

Analysis of jitter due to call-level fluctuations

M.R.H. Mandjes

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Analysis of jitter due to call-level fluctuations

ABSTRACT

In communication networks used by constant bit rate applications, call-level dynamics (i.e., entering and leaving calls) lead to fluctuations in the load, and therefore also fluctuations in the delay ("jitter"). By intentionally delaying the packets at the destination, one can transform the "perturbed" packet stream back into the original periodic stream; in other words: there is a trade off between jitter and delay, in that jitter can be removed at the expense of delay. As a consequence, for streaming applications for which the packet delay should remain below some predefined threshold, it is desirable that the jitter remains small. This paper presents a set of procedures to compute the jitter due to call-level variations. We onsider a network resource shared by a fluctuating set of constant bit rate applications (modelled as periodic sources). As a first step we study the call-level dynamics: supposing that a tagged call sees n_0 calls when entering the system, then we compute the probability that at the end of its duration (consisting of, say, i packets) n_i calls are present, of which $n_{0,i}$ stem from the original n_0 calls. As a second step, we show how to compute the jitter, for given n_0, n_i, and $n_{0,i}$; in this analysis generalized Ballot-problems have to be solved. We find an iterative exact solution to these, and explicit approximations and bounds. Then, as a final step, the (packet-level) results of the second step are weighed with the (call-level) probabilities of the first step, thus resulting in the probability distribution of the jitter experienced within the call duration. An explicit Gaussian approximation is proposed. Extensive numerical experiments validate the accuracy of the approximations and bounds.

Analysis of jitter due to call-level fluctuations

Michel Mandjes [*]

Abstract

In communication networks used by constant bit rate applications, call-level dynamics (i.e., entering and leaving calls) lead to fluctuations in the load, and therefore also fluctuations in the delay ('jitter'). By intentionally delaying the packets at the destination, one can transform the 'perturbed' packet stream back into the original periodic stream; in other words: there is a trade off between jitter and delay, in that jitter can be removed at the expense of delay. As a consequence, for streaming applications for which the packet delay should remain below some predefined threshold, it is desirable that the jitter remains small.

This paper presents a set of procedures to compute the jitter due to call-level variations. We consider a network resource shared by a fluctuating set of constant bit rate applications (modelled as periodic sources). As a first step we study the call-level dynamics: supposing that a tagged call sees n_0 calls when entering the system, then we compute the probability that at the end of its duration (consisting of, say, i packets) n_i calls are present, of which $n_{0,i}$ stem from the original n_0 calls. As a second step, we show how to compute the jitter, for given n_0, n_i, and $n_{0,i}$; in this analysis generalized Ballot-problems have to be solved. We find an iterative exact solution to these, and explicit approximations and bounds. Then, as a final step, the (packet-level) results of the second step are weighed with the (call-level) probabilities of the first step, thus resulting in the probability distribution of the jitter experienced within the call duration. An explicit Gaussian approximation is proposed. Extensive numerical experiments validate the accuracy of the approximations and bounds.

Key words: Packet-level models – Ballot theorems – Beneš approach – constant-bit rate applications – jitter

1 Introduction

Packet-level models are used to describe congestion phenomena in communication networks, particularly those related to relatively short time-scales. On this time-scale (typically in the ms-range), the network element (switch, router) has a (nearly) constant number of users, each transmitting packets at a constant rate. This resulted in the canonical *periodic* traffic model: a network resource that is used by N (independent) users, and each of these users generates a 'purely periodic packet stream with random phase': packets are generated in a periodic fashion, every D units of time, but the packet's position *within* these intervals of length D is, for each user, uniformly distributed.

For the above model of periodic streams with random phase, it is clear that, despite the regularity of the traffic streams, queueing can occur (namely, if multiple streams generate a packet nearly simultaneously within the period of length D). However, if the population of users N remains constant, and if the packet streams are served at a constant rate, then each of the users will experience the same delay for every packet he transmits. We say that there is no *packet delay variation* (or: *jitter*).

In practice, however, there *will* be fluctuations at the call level: supposing a tagged call that consists of i packets (and remains, as a consequence, in the system for iD units of time), the number of calls present at the moment it enters, say n_0, may differ considerably from the number it leaves behind after departing, say n_i. A complicating factor is that these n_0 and n_i calls may have a number of calls in common, say $n_{0,i}$; in other words, $n_{0,i} \in \{0, \ldots, \min\{n_0, n_i\}\}$ are present in both intervals (of length D). We conclude that the fact that the tagged call shares the network resources with a varying set of other users, i.e., the fluctuations at the call level, will induce packet delay variation. The goal of this paper is to analyze this jitter due to call-level fluctuations.

For several streaming applications, to meet Quality-of-Service requirements, the end-to-end delay should remain below some tolerable bound. Jitter can be removed by intentionally delaying packets (to obtain synchronous playout of the packets), such that the resulting signal is again periodic. It is clear that this jitter compensation is at the expense of extra delay, and this explains why the jitter should be controlled.

In this paper we focus on jitter due to call-level fluctuations. We remark that in practice this is not the only factor contributing to the jitter. Focusing on the example of voice over the Internet, many different speech codings are used, which are usually adaptive in the sense that their transmission rate depends on the network conditions. Also the assumption of perfectly periodic packet streams is not realistic, as in reality there is some 'initial jitter'. In addition, often silence suppression is applied.

Literature. The model with a constant number of periodic streams, say N, transmitting a packet every D units of time (at an epoch that is uniformly distributed on the interval $[0, D)$), served at constant rate, is called the $N\cdot$D/D/1 queue. The solution of the workload distribution in this queue goes back to Dempster [6], Pyke [20], and Takács [23], who independently found elegant explicit expressions. Takács' approach is based on combinatorial arguments, e.g., *Ballot theorems*; the translation of this result into queueing terms is due to Humblet *et al.* [10]. Eckberg [7], apparently not aware of the explicit results, found a recursive algorithm for computing the distribution function; see for another exact derivation also [8]. Independently of [10] (and nearly simultaneously), Virtamo and Roberts [19, 26] rederived Takács' closed-form solution; Norros *et al.* [16] noted that the approach followed in [19, 26] could be regarded as an application of the so-called *Beneš method* [1]. In [9, 16, 21] Brownian-bridge approximations are proposed, which are particularly accurate in a heavy-traffic environment. A concise survey on the single $N\cdot$D/D/1 queue is found in [18,

Section 15.2]. Extensions to tandem and priority systems can be found in [4, 11, 14, 25].

Remarkably, however, despite the explicitness of the solutions for the $N \cdot$D/D/1 queue, there are hardly any results on the impact of call-level fluctuations on the delay variation. To our best knowledge, the only result known so far is found in Humblet *et al.* [10, Section IV]. The situation considered there is more specific than ours: the number of calls remains constant, and in any period of length D one 'old' call is replaced by a 'fresh' call (an 'exchange'). Under the assumption that Q_i (with Q_i denoting the queue length in the ith period of length D) forms a Markov chain (which is obviously not correct, but it could of course serve as an approximation), [10] obtains the expected time until the delay is increased (or decreased) by some amount x. In this paper we look at more general call-level fluctuations than just exchanges, and in addition we do not need the approximation of the Q_i forming a Markov chain.

Jitter is considered as one of the main Quality-of-Service metrics for real-time streaming applications. This explains why there is a vast body of literature on the analysis of jitter. It is noted, however, that one usually considers other aspects than the impact of call-level fluctuations; we mention a few of those aspects here. Several papers address the propagation of some initial jitter (i.e., the original signal is a 'perturbed periodic stream') when traversing network nodes, see for instance the results in the context of IP's 'expedited forwarding' [2], and remarkably powerful 'low-jitter conservation laws' [5]. Other papers focus on the jitter that is induced in best-effort networks, and the playout adaptation policies that it requires, see, e.g., the recent paper [13].

Contribution & organization. The model and some preliminaries are provided in Section 2. Section 3 studies – separately – the call level and the packet level. First the call-level computations are done (both for the situation with and without admission control): suppose a tagged call sees n_0 calls when entering the system, then we compute the probability that at the end of its duration (consisting of, say, i packets) n_i calls are present, of which $n_{0,i}$ stem from the original n_0 calls. The packet-level computations are more involved. To find the jitter, for given n_0, n_i, and $n_{0,i}$, generalized Ballot-problems have to be solved. We find an iterative exact solution to these, but for problems of realistic size, this procedure tends to be slow, particularly for the situation with admission control. Therefore we also propose explicit approximations and bounds. In Section 4 the call-level and packet-level results are combined into a multi-level model for analyzing jitter (i.e., the part of the jitter that is the consequence of call-level fluctuations). An explicit Gaussian approximation is proposed. Extensive numerical experiments validate the accuracy of the approximations and bounds.

2 Model

In this paper we analyze a two-layer model, i.e., a model with both a call level and a packet level. Time is slotted, where slots correspond to packet service times – we normalize this packet service time to 1 (in any time slot exactly one packet can be served). Let interval I_i correspond to the slots $\{iD + 1, \ldots, (i+1)D\}$, for $i \in \mathbb{Z}$.

Call-level traffic characteristics. We assume that calls arrive and depart at the beginning of the intervals I_i, for $i \in \mathbb{Z}$.

- *Arrivals:* arrivals occur according to (a discrete-time version of) a Poisson process. More precisely, with A_i (for $i \in \mathbb{Z}$) denoting the number of arrivals at the beginning of interval I_i, we assume that

$(A_i)_{i \in \mathbb{Z}}$ constitute an i.i.d. sequence of random variables, each of them distributed Poisson with mean $\lambda > 0$.

- *Departures:* Each call leaves the system after a random number of intervals; these durations also constitute a sequence of i.i.d. random variables. We assume that they are distributed on $\mathbb{N}$ according to some random variable B with mean $\mathbb{E}B < \infty$.

In the sequel, we consider two generic situations: (i) no *admission control* is imposed, i.e., all jobs can enter, irrespective of the number of flows present in the system; (ii) admission control is imposed, i.e., the number of simultaneous calls is truncated at K. If the blocking probability in the model with admission control is low, then it follows that the difference between both situations will be small.

Let N_i denote the number of calls present at the beginning of slot i. If there is no admission control, the classical Erlang loss formula yields that the steady-state number of jobs in the system, say N, has a Poisson distribution with mean $\lambda \mathbb{E}B$; importantly, this distribution is insensitive in that it depends on the call-duration distribution only through its mean $\mathbb{E}B$. If there is admission control we have also have insensitivity: for $n \in \{0, \ldots, K\}$ the truncated Poisson distribution applies:

$$\mathbb{P}(N = n) = \frac{(\lambda \mathbb{E}B)^n}{n!} \Big/ \sum_{k=0}^{K} \frac{(\lambda \mathbb{E}B)^k}{k!}.$$

Packet-level traffic characteristics. During their stay in the system, calls generate a periodic stream of packets. More precisely: if a call enters at the beginning of I_i, it generates a packet at an epoch that is uniformly distributed on the slots $\{iD + 1, \ldots, (i+1)D\}$. Suppose this turns out to be slot $iD + k$, and suppose the call leaves at the beginning of interval I_j, then also packets are generated at epochs $\ell D + k$, for $\ell = i + 1, \ldots, j - 1$. In other words: the per-call generated traffic stream is purely periodic.

Let $A_i(k)$ be the number of packets generated in $\{iD + 1, \ldots, iD + k\}$, for $k \in \{1, \ldots, D\}$. If there are n_i calls present at the start of interval I_i, then clearly $A_i(D) = n_i$. Also, $A_i(k)$ has a binomial distribution with parameters n_i and k/D. With $k \leq \ell$, we denote $A_i(k, \ell) := A_i(\ell) - A_i(k)$.

Definitions. The following definitions are used frequently in this paper: we say that $X \sim \mathbb{B}\text{in}(N, p)$ if

$$\mathbb{P}(X = n) = \mathbb{B}\text{in}(n \mid N, p) := \binom{N}{n} p^n (1 - p)^{N-n},$$

for $N \in \mathbb{N}_0$, $p \in (0, 1)$ and $n \in \{0, \ldots, N\}$. We say that $X \sim \mathbb{P}\text{ois}(\lambda)$ if

$$\mathbb{P}(X = n) = \mathbb{P}\text{ois}(n \mid \lambda) := e^{-\lambda} \frac{\lambda^n}{n!},$$

for $\lambda > 0$ and $n \in \mathbb{N}_0$. We say that $X \sim \mathbb{H}\text{g}(M, N, n)$ if

$$\mathbb{H}\text{g}(m \mid M, N, n) := \binom{M}{m} \binom{N}{n - m} \Big/ \binom{M + N}{n},$$

for $M, N \in \mathbb{N}_0$, $m \in \{0, \ldots, M\}$ and $n - m \in \{0, \ldots, N\}$.

3 Analysis of call level and packet level

In this section we present detailed analyses of both the call level and the packet level. In Section 4, these are combined into a two-level model that can be used for jitter computations.

3.1 Call level

Suppose we observe the number of calls present at the time the tagged call arrives. Without loss of generality, we let this happen at the start of interval I_0, and hence we have the information $N_0 = n_0$ at our disposal. To compute the jitter resulting from the intervals I_0 and some I_i (for $i \in \mathbb{N}$), we need to know (i) how many calls are present at the beginning of interval I_i (i.e., the value of the random variable N_i), and (ii) how many calls are present in *both* intervals (which we denote by $N_{0,i}$). Therefore, we concentrate in this subsection on the computation of

$$\pi_i(n_i, n_{0,i} \mid n_0) := \mathbb{P}(N_i = n_i, N_{0,i} = n_{0,i} \mid N_0 = n_0).$$

Evidently, $n_{0,i} \in \{0, \ldots, \min\{n_0, n_i\}\}$.

Call level without admission control

We first focus on the situation without admission control. Let B denote the generic random variable of the call duration. Each of the n_0 calls has a redidual lifetime with distribution, for $i \in \mathbb{N}_0$,

$$\mathbb{P}(B^{\mathrm{res}} = i) = \frac{1}{\mathbb{E}B} \cdot \mathbb{P}(B \geq i).$$

We first apply the following decomposition:

$$\pi_i(n_i, n_{0,i} \mid n_0) := \mathbb{P}(N_i - N_{0,i} = n_i - n_{0,i}) \cdot \mathbb{P}(N_{0,i} = n_{0,i} \mid N_0 = n_0).$$

It is clear that

$$\mathbb{P}(N_{0,i} = n_{0,i} \mid N_0 = n_0) = \mathbb{B}\mathrm{in}\left(n_{0,i} \mid n_0, \mathbb{P}(B^{\mathrm{res}} \geq i)\right).$$

Now realize that $N_i - N_{0,i}$ corresponds to the calls (out of the $\sum_{n=1}^{i} A_n$ arrived calls) that are still present at the beginning of I_i. In other words: we can rewrite $N_i - N_{0,i}$ as $\sum_{n=1}^{i} A_{n,i}$, where $A_{n,i}$ corresponds to the calls that arrived at the beginning of I_n, which are still present at the beginning of I_i. It is a property of the Poisson process that

$$\mathbb{P}(A_{n,i} = j) = \mathbb{P}\mathrm{ois}(j \mid \lambda \mathbb{P}(B \geq i - n)),$$

and consequently

$$\mathbb{P}(N_i - N_{0,i} = n_i - n_{0,i}) = \mathbb{P}\mathrm{ois}\left(n_i - n_{0,i} \mid \lambda \cdot \sum_{n=0}^{i-1} \mathbb{P}(B \geq n)\right).$$

Similarly, we find that

$$N_{0,i} \sim \mathbb{P}\mathrm{ois}(\mu_{0,i}), \quad \text{with} \quad \mu_{0,i} := \lambda \mathbb{E}B \cdot \sum_{n=i}^{\infty} \mathbb{P}(B^{\mathrm{res}} = n) = \lambda \cdot \sum_{n=i}^{\infty} \mathbb{P}(B \geq n);$$

$$N_0 - N_{0,i} \sim \mathbb{P}\mathrm{ois}(\mu_i), \quad \text{with} \quad \mu_i := \lambda \cdot \sum_{n=0}^{i-1} \mathbb{P}(B \geq n),$$

where $N_0 - N_{0,i}$, $N_{0,i}$, and $N_i - N_{0,i}$ are mutually independent; remark that, by observing that $\mu_0 + \mu_{0,i} = \mu_i + \mu_{0,i} = \lambda \mathbb{E}B$, we indeed find that both N_0 and N_i have a Poisson distribution with mean $\lambda \mathbb{E}B$.

Remark. It can be verified that for $n_{0,i} = 0$ and any n_0, we have that for i large that $\pi_i(n_i, n_{0,i} \mid n_0)$ tends to $\mathbb{P}\text{ois}(n_i \mid \lambda \mathbb{E}B)$, as could be expected from the Erlang loss formula (for any other value of $n_{0,i}$, it follows that this limit equals 0); here it is used that

$$\mathbb{E}B = \sum_{n=0}^{\infty} \mathbb{P}(B \geq n).$$

In other words, the complete initial population n_0 has left. ◇

Call level with admission control

The computation with admission control of $\pi_i(n_i, n_{0,i} \mid n_0)$ for the case *with* admission control is considerably more involved, and requires substantial administration. We therefore restrict ourselves to the special case of geometric durations; the following elementary recursive scheme follows from the memoryless property. Let $p \in (0,1)$ be the probability that an existing call 'survives' the next interval.

- First concentrate on the one-step transition probability, i.e., consider $i = 1$. Suppose first $n_1 < K$. Then our probability corresponds to the event that $n_{0,1}$ out of the original n_0 calls are still present at $i = 1$, whereas $n_1 - n_{0,1}$ new calls entered. This entails

  $$\pi_1(n_1, n_{0,1} \mid n_0) = \mathbb{B}\text{in}\,(n_1^\star \mid n_0, p)\,\mathbb{P}\text{ois}(n_1 - n_{0,1} \mid \lambda).$$

  Because of the truncation at K calls, we also have

  $$\pi_1(K, n_{0,1} \mid n_0) = \mathbb{B}\text{in}\,(n_{0,1} \mid n_0, p) \sum_{k=K}^{\infty} \mathbb{P}\text{ois}(k - n_{0,1} \mid \lambda).$$

- Now consider $i \in \{2, 3, \ldots\}$, and suppose we know $\pi_{i-1}(n_{i-1}, n_{0,i-1} \mid n_0)$. To make sure that $N_i = n_i$, among which $n_{0,i}$ of the original n_0 calls, suppose $N_{i-1} = n_{i-1}$, of which $n_{0,i-1}$ belong to the original n_0 calls; we have that $n_{0,i-1} \in \{n_{0,i}, \ldots, n_{i-1}\}$ and $n_{i-1} \in \{n_{0,i-1}, \ldots, K\}$. This happens with probability $\pi_{i-1}(n_{i-1}, n_{0,i-1} \mid n_0)$.

  Again, first focus on $n_i < K$. From the $n_{0,i-1}$ original calls, $n_{0,i}$ have to persist to the next interval, which happens with probability $\mathbb{B}\text{in}(n_{0,i} \mid n_{0,i-1}, p)$. Then, the net number of calls joining should be $n_i - n_{i-1}$, where it is already known that $n_{0,i-1} - n_{0,i}$ of the original calls left. This means that if j calls enter, then $n_i - n_{i-1} - (n_{0,i} - n_{0,i-1}) - j$ of the $n_{i-1} - n_{0,i-1}$ non-original calls leave (or, equivalently, $n_i - n_{0,i} - j$ of the non-original calls stay). This leads to the probability

  $$\sum_{j=0}^{n_i - n_{0,i}} \mathbb{B}\text{in}(n_i - n_{0,i} - j \mid n_{i-1} - n_{0,i-1}, p)\mathbb{P}\text{ois}(j \mid \lambda) =: \rho(n_{i-1}, n_i, n_{0,i-1}, n_{0,i}).$$

  Summarizing, $\pi_i(n_i, n_{0,i} \mid n_0)$ equals

  $$\sum_{n_{0,i-1}=n_{0,i}}^{n_i} \sum_{n_{i-1}=n_{0,i-1}}^{K} \pi_{i-1}(n_{i-1}, n_{0,i-1} \mid n_0)\mathbb{B}\text{in}(n_{0,i} \mid n_{0,i-1}, p)\rho(n_{i-1}, n_i, n_{0,i-1}, n_{0,i}).$$

  For $n_i = K$, we have an analogous formula, but now with $\rho(n_{i-1}, K, n_{0,i-1}, n_{0,i})$ equal to

  $$\sum_{j=0}^{K-n_{0,i}} \sum_{j'=j}^{\infty} \mathbb{B}\text{in}(K - n_{0,i} - j \mid n_{i-1} - n_{0,i-1}, p)\mathbb{P}\text{ois}(j' \mid \lambda).$$

3.2 Packet level

In this section we analyze the packet level. We consider two intervals, I_0 and I_i, and fix the number of calls present in each (n_0 and n_i), and the number of calls present in both $n_{0,i}$.

Exact analysis

We compare two (disjoint) intervals I_0 and I_i, for $i \in \mathbb{N}$. We condition on the situation in which in the first interval, there are n_0 calls present, in the second interval there are n_i, whereas $n_{0,i}$ calls are active in both intervals.

In this subsection, our goal is to compute, for $x_0 \in \{0, \ldots, n_0\}$, and $x_i \in \{0, \ldots, n_i\}$, the following 'generalized Ballot theorem', cf. [10],

$$\mathbb{P}\{\underline{x} \mid \underline{n}, n_{0,i}, D\} := \mathbb{P}(\forall \tau \in \{1, \ldots, D\} : \{A_0(\tau) \leq x_0 + \tau\} \cap \{A_i(\tau) \leq x_i + \tau\}).$$

i.e., the probability that in a first interval I_0 the queue length, say Q_0, does not exceed level x_0, and in a second interval I_i the queue lenth Q_i does not exceed x_i.

We first write our target probability $\mathbb{P}\{\underline{x} \mid \underline{n}, n_{0,i}, D\}$ as the difference of $\mathbb{P}\{x_i \mid n_i, D\}$ and

$$\mathbb{P}(\{\exists \tau \in \{1, \ldots, D\} : A_0(\tau) \geq x_0 + \tau\} \cap \{\forall \tau \in \{1, \ldots, D\} : A_i(\tau) \leq x_i + \tau\}); \tag{1}$$

here $\mathbb{P}\{x_i \mid n_i, D\} := \mathbb{P}(\{\forall \tau \in \{1, \ldots, D\} : A_i(\tau) \leq x_i + \tau\})$. Probability $\mathbb{P}\{x_i \mid n_i, D\}$ is standard, and follows from the explicit solution of the $N{\cdot}\mathrm{D/D/1}$ queue (see for instance [10, 18]); probability (1) is more involved and will be decomposed now.

- First condition on the *last epoch* that $A_0(\tau) \geq x_0 + \tau$; call this epoch m, cf. the Beneš method [1, 16, 17]. Hence, $A_0(m) = x_0 + m$. Then

$$\mathbb{P}(A_0(m) = x_0 + m) = \mathbb{B}\mathrm{in}\left(x_0 + m \mid n_0, \frac{n_0}{D}\right).$$

  Given $A_0(m) = x_0 + m$, clearly $A_0(D) - A_0(m) = n_0 - x_0 - m$.

- Let $\bar{A}(m)$ be the part of $A_0(m)$ that stems from the common 'pool' of size $n_{0,i}$. It can be seen that

$$\mathbb{P}(\bar{A}(m) = k, A_i(m) = \ell \mid A_0(m) = x_0 + m) =$$

$$\mathbb{H}\mathrm{g}(k \mid n_0, n_{0,i}, x_0 + m)\,\mathbb{B}\mathrm{in}\left(\ell - k \mid n_i - k, \frac{n_0}{D}\right).$$

  Define $E(k, \ell, m) := \{A_0(m) = x_0 + m\} \cap \{\bar{A}(m) = k\} \cap \{A_i(m) = \ell\}$; let $p(k, \ell, m)$ denote $\mathbb{P}(E(k, \ell, m))$.

- Now condition on $E(k, \ell, m)$. Before m (i.e., for $\tau \in \{1, \ldots, m\}$), (i) the process $A_i(\tau)$ has to remain below $x_i + \tau$; this happens with probability $\mathbb{P}\{x_i \mid \ell, m\}$. Then, after epoch m (i.e., for $\tau \in \{m+1, \ldots, D\}$), both (ii) $A_0(\tau) - A_0(m)$ should stay below $\tau - m$, and (iii) $A_i(\tau) - A_i(m)$ below $x_i + \tau - \ell$; the intersection of these two events has probability

$$\mathbb{P}\{(0, x_i + m - \ell) \mid (n_0 - x_0 - m, n_i - \ell), n_{0,i} - k, D - m\}. \tag{2}$$

  Let $q(k, \ell, m)$ denote the probability, conditional on $E(k, \ell, m)$, of the intersection of the events (i), (ii), and (iii).

We get that our target probability (1) equals

$$\sum_{k,\ell,m} p(k,\ell,m) q(k,\ell,m). \tag{3}$$

Here m runs from 0 to $n_0 - x_0$; k from 0 to $\min\{n_{0,i}, n_0 + x_0\}$; ℓ from k to $k + n_i - n_{0,i}$.

In the above scheme we have expressed, through Eq. (3), our target probability $\mathbb{P}\{\underline{x} \mid \underline{n}, n_{0,i}, D\}$ in terms of probabilities $p(k,\ell,m)$. Importantly, *these $p(k,\ell,m)$ are of the same type as* $\mathbb{P}\{\underline{x} \mid \underline{n}, n_{0,i}, D\}$, but with other ('lower') arguments, see (2). Hence we found a *recursion* for our target probability. Together with the following start conditions, it yields a procedure for evaluating $\mathbb{P}\{\underline{x} \mid \underline{n}, n_{0,i}, D\}$. These start conditions follow from the fact that, for any $\underline{x}, \underline{n}, D$,

$$\mathbb{P}\{\underline{x} \mid \underline{n}, 0, D\} = \mathbb{P}\{x_0 \mid n_0, 0, D\} \times \mathbb{P}\{x_i \mid n_i, 0, D\};$$

in words: if both intervals do not have any calls in common, our probability factorizes.

Explicit upper bounds and approximations

The recursive scheme of Section 3.2 gives an explicit procedure for computing $\mathbb{P}\{\underline{x} \mid \underline{n}, n_{0,i}, D\}$, but this could be rather slow when the number of calls involved is large. In this subsection we study more efficient schemes that provide us with an upper bound on the probability that the jitter exceeds some predefined value.

We settle for a probability that is slightly less detailed than $\mathbb{P}\{\underline{x} \mid \underline{n}, n_{0,i}, D\}$, namely

$$
\begin{aligned}
\bar{\mathbb{P}}\{x \mid \underline{n}, n_{0,i}, D\} \quad &:= \quad \mathbb{P}(Q_i - Q_0 \geq x) \\
&= \quad \mathbb{P}\left(\left(\sup_{\tau \in \{1,\dots,D\}} A_i(\tau) - \tau\right) - \left(\sup_{\sigma \in \{1,\dots,D\}} A_0(\sigma) - \sigma\right) \geq x\right) \\
&= \quad \mathbb{P}(\exists \tau \in \{1,\dots,D\} : \forall \sigma \in \{1,\dots,D\} : A_i(\tau) - A_0(\sigma) - \tau + \sigma \geq x).
\end{aligned}
\tag{4}
$$

- *Crude upper bound; approximation.* A first crude upper bound is

$$\sum_{\tau=1}^{D} \mathbb{P}(\forall \sigma \in \{1,\dots,D\} : A_i(\tau) - A_0(\sigma) - \tau + \sigma \geq x).$$

  As the probability of an intersection is smaller than those corresponding to any of the individual events, we obtain

$$\bar{\mathbb{P}}\{x \mid \underline{n}, n_{0,i}, D\} \leq \sum_{\tau=1}^{D} \inf_{\sigma \in \{1,\dots,D\}} \mathbb{P}(A_i(\tau) - A_0(\sigma) - \tau + \sigma \geq x). \tag{5}$$

  From the fact that both intervals have $n_{0,i}$ calls in common, observe that $A_i(\tau) - A_0(\sigma)$ can be decomposed into $T_i(\tau) - T_0(\sigma) + T_{0,i}(\tau,\sigma)$, where $T_i(\tau), T_0(\sigma)$, and $T_{0,i}(\tau,\sigma)$ are independent; here

$$T_i(\tau) \sim \mathbb{Bin}\left(n_i - n_{0,i}, \frac{\tau}{D}\right); \quad T_0(\sigma) \sim \mathbb{Bin}\left(n_0 - n_{0,i}, \frac{\sigma}{D}\right),$$

  and

$$T_{0,i}(\tau,\sigma) \sim \mathbb{Bin}\left(n_{0,i}, \frac{\tau-\sigma}{D}\right) \text{ if } \tau \geq \sigma; \quad T_{0,i}(\tau,\sigma) \sim -\mathbb{Bin}\left(n_{0,i}, \frac{\sigma-\tau}{D}\right) \text{ if } \tau < \sigma.$$
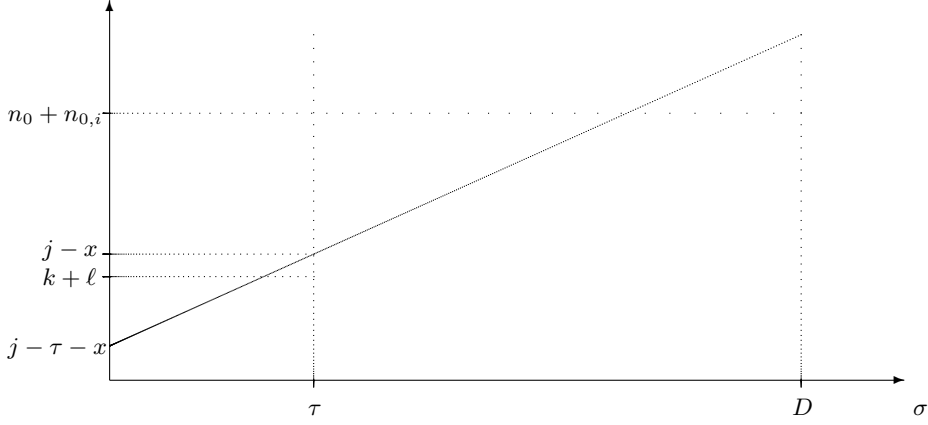
Figure 1: Illustration supporting the derivation of $p_\tau(j, k, \ell)$. The depicted straight line is $\sigma + j - \tau - x$; we are considering the probability that $A_0(\sigma)$ hits (but does not exceed) this line. At time τ the number of packets generated is $k + \ell$.

Replacing a union by a sum, as has been done above, is often quite crude. In many situations – justified by large deviations heuristics – it is better to replace the union by a maximum:

$$\bar{\mathbb{P}}\{x \mid \underline{n}, n_{0,i}, D\} \approx \sup_{\tau \in \{1, \ldots, D\}} \inf_{\sigma \in \{1, \ldots, D\}} \mathbb{P}(A_i(\tau) - A_0(\sigma) - \tau + \sigma \geq x). \tag{6}$$

It is clear that the right hand side is not necessarily an upper bound anymore; it can however be expected that it is a relatively good approximation. We assess the quality of both (5) and (6) later in this section.

- *Refined upper bound.* Recall that Q_0 denotes the queue length in I_0, i.e., $Q_0 := \sup_{\sigma \in \{1, \ldots, D\}} A_0(\sigma) - \sigma$. The Beneš approach (i.e., partition the overflow event with respect to the *last* epoch of exceeding x) tells us that this probability equals

$$\sum_{\tau=1}^{D} \mathbb{P}(A_i(\tau) - \tau = x + Q_0, \forall \bar{\tau} \in \{\tau + 1, \ldots, D\} : A_i(\bar{\tau}) - \bar{\tau} < x - Q_0).$$

A standard upper bound to this probability is – see also [18, p. 375] – given by

$$\sum_{\tau=1}^{D} \mathbb{P}(A_i(\tau) - \tau = x + Q_0).$$

Now let us try to explicitly compute $\mathbb{P}(A_i(\tau) - \tau = x + Q_0)$. Suppose that $A_i(\tau) = j$, which happens with probability $\mathbb{Bin}(j \mid n_i + n_{0,i}, \tau/D)$. A part of these j packets belong to the $n_{0,i}$ packets that are present in both intervals I_0 and I_i; this number, say k, is $\mathbb{Hg}(n_i, n_{0,i}, j)$ distributed. The part of the n_0 'non-common' packets that arrives in $\{1, \ldots, \tau\}$, say ℓ, has a $\mathbb{Bin}(n_0, \tau/D)$ distribution.

Now condition on these three events: (i) $\{A_i(\tau) = j\}$, (ii) $T_{0,i}(\tau, 0) = k$, and (iii) $T_0(\tau) = \ell$; we call the composite event $F(j, k, \ell)$. From the above we have that

$$q_\tau(j, k, l) := \mathbb{P}(F(j, k, \ell)) = \mathbb{Bin}\left(j \mid n_i + n_{0,i}, \frac{\tau}{D}\right) \cdot \mathbb{Hg}(k \mid n_i, n_{0,i}, j) \cdot \mathbb{Bin}\left(k \mid n_0, \frac{\tau}{D}\right).$$
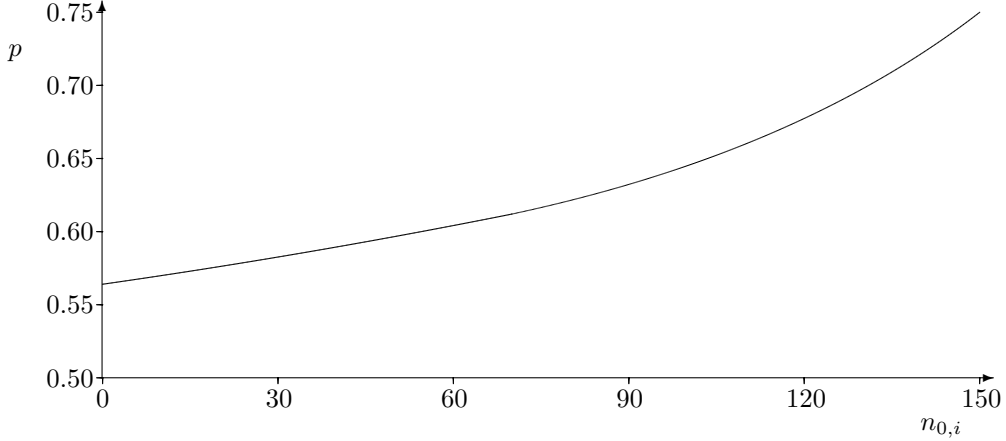
9

Figure 2: The probability p of a non-empty buffer in both intervals, as a function of $n_{0,i}$; $D = 200$ and $N := n_0 = n_i = 150$.

Hence in I_0, we have that $k + \ell$ packets arrive before (or at) τ (each of them at a position that is uniformly distributed on $\{1, \ldots, \tau\}$), and the remaining $n_0 + n_{0,i} - k - \ell$ after τ (each of them on a position that is uniformly distributed on $\{\tau + 1, \ldots, D\}$). Now define $\mathbb{P}_e\{x \mid N, D\}$ the probability that the queue of an $N \cdot D/D/1$ system reaches *exactly* level x, or, in other words,

$$\mathbb{P}_e\{x \mid N, D\} = \mathbb{P}\{x \mid N, D\} - \mathbb{P}\{x - 1 \mid N, D\}.$$

Hence we obtain

$$
\begin{aligned}
p_\tau(j, k, \ell) &:= \mathbb{P}(Q_0 = A_i(\tau) - \tau - x \mid E(j, k, \ell)) \\
&= \mathbb{P}_e\{j - \tau - x \mid k + \ell, \tau\} \cdot \mathbb{P}\{j - x - k - \ell \mid n_0 + n_{0,i} - k - \ell, D - \tau\} \\
&\quad + \mathbb{P}\{j - \tau - x \mid k + \ell, \tau\} \cdot \mathbb{P}_e\{j - x - k - \ell \mid n_0 + n_{0,i} - k - \ell, D - \tau\};
\end{aligned}
$$

i.e., the process $A_0(\sigma)$ hits $\sigma + j - \tau - x$ before τ (without exceeding it), and stays below (or at) the line after τ, or vice versa; see Fig. 1. We arrive at the upper bound

$$\sum_{\tau=1}^{D} \sum_{j,k,\ell} p_\tau(j, k, \ell) q_\tau(j, k, \ell). \tag{7}$$

Numerical study

We now assess the algorithms proposed, by means of two numerical experiments. The first illustrates what type of questions could be solved by applying the recursive algorithm. In the second experiment, the performance of the proposed bounds and approximation is evaluated.

Experiment 1 We first consider the situation $D = 200$, and $N := n_0 = n_i = 150$. We let $n_{0,i}$ vary from 0 to 150, and we consider the probability p that both Q_0 and Q_i are empty. Evidently, when $n_{0,i} = 0$ this is $(N/D)^2 = 0.5625$ (two independent experiments), whereas when $n_{0,i} = N$ we get $N/D = 0.75$ (essentially a single experiment). Fig. 2 shows how the probability of our interest depends on $n_{0,i}$. The numbers are computed by using the (exact) recursive scheme. These computations are very time-consuming, due to the fact that the number of sources is relatively high. From the graph we conclude that for low $n_{0,i}$ the probability p is still close to $(N/D)^2$; the graph sharply increases for $n_{0,i}$ close to N. $\diamond$

10

x	exact	bound (5)	appr. (6)	bound (7)
2	$2.07 \cdot 10^{-1}$	$3.12 \cdot 10^{0}$	$1.33 \cdot 10^{-1}$	$8.16 \cdot 10^{-1}$
4	$4.53 \cdot 10^{-2}$	$8.47 \cdot 10^{-1}$	$3.52 \cdot 10^{-2}$	$2.01 \cdot 10^{-1}$
6	$1.17 \cdot 10^{-2}$	$1.98 \cdot 10^{-1}$	$9.46 \cdot 10^{-3}$	$4.67 \cdot 10^{-2}$
8	$3.34 \cdot 10^{-3}$	$3.82 \cdot 10^{-2}$	$2.10 \cdot 10^{-3}$	$9.07 \cdot 10^{-3}$
10	$4.25 \cdot 10^{-4}$	$4.40 \cdot 10^{-3}$	$4.10 \cdot 10^{-4}$	$1.47 \cdot 10^{-3}$

x	exact	bound (5)	appr. (6)	bound (7)
2	$1.03 \cdot 10^{-1}$	$2.72 \cdot 10^{0}$	$1.13 \cdot 10^{-1}$	$3.63 \cdot 10^{-1}$
4	$1.42 \cdot 10^{-2}$	$7.11 \cdot 10^{-1}$	$2.96 \cdot 10^{-2}$	$4.32 \cdot 10^{-2}$
6	$1.59 \cdot 10^{-3}$	$1.59 \cdot 10^{-1}$	$7.20 \cdot 10^{-3}$	$5.12 \cdot 10^{-3}$
8	$1.12 \cdot 10^{-4}$	$3.10 \cdot 10^{-2}$	$1.56 \cdot 10^{-3}$	$2.70 \cdot 10^{-4}$
10	$4.91 \cdot 10^{-6}$	$4.54 \cdot 10^{-3}$	$3.20 \cdot 10^{-4}$	$2.80 \cdot 10^{-5}$

Table 1: Accuracy of the upper bounds and approximation; $D = 100$ and $N := n_0 = n_i = 80$. In the upper panel $n_{0,i} = 20$, in the lower panel $n_{0,i} = 60$.

Experiment 2 We now assess the performance of the proposed bounds (5) and (7) as well as approximation (6), see Table 1. To this end, we have considered a situation with $D = 100$ and $N := n_0 = n_i = 80$, and both $n_{0,i} = 20$ (i.e., a relatively low number of calls is present in both intervals) and $n_{0,i} = 60$ (relatively many of the initial calls are still present). Besides these experiments, we have performed extensive additional simulations. Some general conclusions are: (i) The upper bound (5) is usually rather pessimistic; apparently it is very conservative to replace the union by a sum. (ii) If the number of calls in common $n_{0,i}$ is relatively low, the approximation (6) performs well; for high values of $n_{0,i}$ the approximation is not that good, but still considerably better than the upper bound (5). (iii) The approximation (6) yields neither a systematic underestimation nor a systematic overestimation. (iv) The upper bound (7) performs considerably better than the upper bound (5); for high values of $n_{0,i}$ the upper bound (7) is closer to the real value than the approximation (6). $\diamond$

4 Analysis of the integrated model

In this section we study the model with both a call level and a packet level.

4.1 Integrated model

One of the interesting performance measures related to jitter is the probability $\mathbb{P}(Q_B - Q_0 \geq x)$ for $x \geq 0$, i.e., the probability that, at the end of the tagged call the delay is at least x larger than its initial value. One could think of more precise performance measures.

- For instance, it should also be avoided that the delay becomes considerably *smaller* than its initial value, such that $\mathbb{P}(|Q_B - Q_0| \geq x)$ is perhaps more relevant. It is clear, however, that, by reasons of

symmetry,

$$\mathbb{P}(|Q_B - Q_0| \geq x) = 2\mathbb{P}(Q_B - Q_0 \geq x). \tag{8}$$

- Also, one could wonder what the probability is that *during the call* the jitter exceeds x:

$$\mathbb{P}(\exists i \in \{1, \ldots, B\} : Q_i - Q_0 \geq x).$$

  It can be expected, however, that the longer the call is, the more likely a delay variation of at least x. For that reason, the above probability is accurately approximated by $\mathbb{P}(Q_B - Q_0 \geq x)$.

The above considerations made us decide to concentrate in the sequel on the computation of $\mathbb{P}(Q_B - Q_0 \geq x)$.

The theory of the previous section can be used to explicitly compute $\mathbb{P}(Q_B - Q_0 \geq x)$, by using the decomposition into a call-level and a packet-level. In the first place, we can condition on the duration of the tagged call:

$$\mathbb{P}(Q_B - Q_0 \geq x) = \sum_{i=1}^{\infty} \mathbb{P}(B = i)\mathbb{P}(Q_i - Q_0 \geq x).$$

Note, however, that, unlike in Section 3.2 (for instance display (4)), the number of calls N_0, N_i, and $N_{0,i}$ are still random here. We can therefore proceed as follows:

$$\mathbb{P}(Q_i - Q_0 \geq x) = \sum_{\underline{n}, n_{0,i}} \pi_i(n_i, n_{0,i} \mid n_0)\mathbb{P}(N_0 = n_0)\bar{\mathbb{P}}\{x \mid \underline{n}, n_{0,i}, D\}.$$

This can be done for both the situation with and without call admission control. As the exact computation of $\bar{\mathbb{P}}\{x \mid \underline{n}, n_{0,i}, D\}$ (using the recursive scheme) can be rather slow (especially when the number of sources is large), one might resort to the proposed bounds and approximations.

4.2 Explicit expressions for the case without admission control

For the situation without admission control significant simplifications can be made. Recall from Section 3.1 that $N_{0,i} \sim \mathbb{P}\text{ois}(\mu_0)$, $N_{0,i} \sim \mathbb{P}\text{ois}(\mu_{0,i})$, and $N_i - N_{0,i} \sim \mathbb{P}\text{ois}(\mu_i)$ where $N_0 - N_{0,i}$, $N_{0,i}$, and $N_i - N_{0,i}$ are mutually independent. Now it follows that for $\mathbb{P}(Q_i - Q_0 \geq x)$ again upper bound (5) applies, but now with $A_i(\tau) - A_0(\sigma)$ decomposed into $U_i(\tau) - U_0(\sigma) + U_{0,i}(\tau, \sigma)$, with $U_i(\tau), U_0(\sigma)$, and $U_{0,i}(\tau, \sigma)$ independent, and

$$U_i(\tau) \sim \mathbb{P}\text{ois}\left(\mu_i \frac{\tau}{D}\right); \quad U_0(\sigma) \sim \mathbb{P}\text{ois}\left(\mu_0 \frac{\sigma}{D}\right), \text{ and } U_{0,i}(\tau, \sigma) \sim \text{sgn}(\tau - \sigma)\mathbb{P}\text{ois}\left(\mu_{0,i} \frac{|\tau - \sigma|}{D}\right);$$

$\text{sgn}(x) = 1$ for $x \geq 0$ and -1 for $x < 0$. Hence, for $\tau \geq \sigma$, the probability $\mathbb{P}(A_i(\tau) - A_0(\sigma) - \tau + \sigma \geq x)$ boils down to

$$\sum_{k=0}^{\infty} \sum_{j=k+\tau-\sigma+x}^{\infty} \mathbb{P}\text{ois}\left(k \mid \mu_0 \frac{\sigma}{D}\right) \mathbb{P}\text{ois}\left(j \mid \mu_i \frac{\tau}{D} + \mu_{0,i} \frac{\tau - \sigma}{D}\right);$$

a similar expression is found for the case $\tau < \sigma$. In the same fashion, we can find the (integrated packet/call level) analog of the approximation (6) (replace the sum over τ by a maximum). The numerical results in Section 3.2 have indicated that this 'sup inf method' is reasonably accurate, while still of low complexity, which motivates why we will use this analog of (6) in the numerical experiments of the next subsection.

i (slots)	$\mu_{0,i}$	$\mu_0 = \mu_i$	$x = 5$	$x = 10$	$x = 15$	$x = 20$
80	217.42	22.58	$6.33 \cdot 10^{-3}$	$6.52 \cdot 10^{-5}$	$8.12 \cdot 10^{-6}$	$2.71 \cdot 10^{-7}$
240	177.99	62.01	$3.02 \cdot 10^{-2}$	$9.84 \cdot 10^{-4}$	$4.98 \cdot 10^{-5}$	$2.43 \cdot 10^{-6}$
480	131.83	108.17	$4.20 \cdot 10^{-2}$	$2.99 \cdot 10^{-3}$	$2.00 \cdot 10^{-4}$	$1.43 \cdot 10^{-5}$
1200	53.57	186.43	$8.42 \cdot 10^{-2}$	$7.96 \cdot 10^{-3}$	$6.18 \cdot 10^{-4}$	$4.21 \cdot 10^{-5}$
2400	11.94	228.06	$1.08 \cdot 10^{-1}$	$9.09 \cdot 10^{-3}$	$6.38 \cdot 10^{-4}$	$4.40 \cdot 10^{-5}$
∞	0.00	240.00	$1.17 \cdot 10^{-1}$	$9.21 \cdot 10^{-3}$	$6.88 \cdot 10^{-4}$	$4.53 \cdot 10^{-5}$
$B \quad \sim \mathbb{G}\text{eom}(p)$			$5.19 \cdot 10^{-2}$	$3.70 \cdot 10^{-3}$	$3.00 \cdot 10^{-4}$	$1.62 \cdot 10^{-5}$

Table 2: $\mathbb{P}(Q_i - Q_0 \geq x)$ for different values of i. The last row corresponds to a *random* call duration B, i.e., $\mathbb{P}(Q_B - Q_0 \geq x)$, with B having a geometric distribution with mean $1/p = 800$ slots.

Also, the (integrated packet/call level) counterpart of upper bound (7) could be found; its computation is tedious and left out here.

Furthermore, it can be checked that

$$\mu\{\tau,\sigma\} := \mathbb{E}(A_i(\tau) - A_0(\sigma)) = (\mu_0 + \mu_{0,i})\frac{\tau - \sigma}{D};$$

$$v\{\tau,\sigma\} := \mathbb{V}\text{ar}\,(A_i(\tau) - A_0(\sigma)) = \mu_i\left(\frac{\tau}{D} + \frac{\sigma}{D}\right) + \mu_{0,i}\frac{|\tau - \sigma|}{D}.$$

Hence, one could alternatively use a Gaussian approximation (with mean $\mu\{\tau,\sigma\}$ and variance $v\{\tau,\sigma\}$) to estimate $\mathbb{P}(A_i(\tau) - A_0(\sigma) - \tau + \sigma \geq x)$.

4.3   Numerical experiments with the integrated model

We now perform a number of numerical experiments with the integrated packet/call-level model.

Experiment 3   In this experiment we assume the following parameter setting. Let D be equal to 312. Suppose that the number of packets generated in a call, i.e., B, has a geometric distibution with mean $1/p$, i.e.,

$$\mathbb{P}(B = n) = \mathbb{G}\text{eom}(n \mid p) := (1 - p)^{n-1}p.$$

Let p be $1.25 \cdot 10^{-3}$, such that the mean call duration amounts to 800 slots. In this example we take $\lambda = 0.3$, i.e., the number of new calls (arriving in each interval of D slots) has a Poisson distribution with mean λ. The mean number of calls in the system is $\lambda\,\mathbb{E}B = 240$, such that the load is 76.9%. It can be verified that the blocking probability in this model is small (in the order of 10^{-5}), which justifies our choice to use the (simpler) model without admission control.

In Table 2 we have used the results of Section 4.2 to get insight into the jitter due to call-level fluctuations. The first lines of the table show $\mathbb{P}(Q_i - Q_0 \geq x)$ for several (fixed) values of i. It is clear that, when i is small, there is a strong correlation between the delay experienced in I_0 and the delay in I_i. When i is getting larger, this dependence becomes weaker; for $i \to \infty$ the values of Q_i and Q_0 are essentially independent.

This can also be concluded from the numbers in the 3rd and 4th column: the larger i, the smaller $\mu_{0,i}$, which represents the mean number of calls present in *both* I_0 and I_i.

Evidently, a strong positive correlation between the arrival patterns in I_0 and I_i can be considered as 'benign': the jitter will be relatively low. This property is reflected in the numbers in Table 2: the smaller i, the lower the probability that the jitter exceeds threshold x. The limiting value (for $i \to \infty$) corresponds to, with Q_∞ being an independent copy of Q_0,

$$\mathbb{P}(Q_\infty - Q_0 \geq x) = \sum_{y=0}^\infty \mathbb{P}(Q_0 \geq x + y)\mathbb{P}(Q_0 = y). \tag{9}$$

The last row of Table 2 corresponds to the situation of a 'random lag', i.e., $\mathbb{P}(Q_B - Q_0 \geq x)$. Its value lies between the values of $\mathbb{P}(Q_i - Q_0 \geq x)$ for $i = 480$ and $i = 1200$, as could be expected from the fact that $\mathbb{E}B = 800$.

As a benchmark, we also computed the probability $P_{\mathrm{M/D/1}}(x \mid \varrho)$ that the delay in an M/D/1 queue (with load $\varrho := \lambda \mathbb{E}B/D = 0.769$) corresponds to at least x packets. It is not *a priori* clear whether $P_{\mathrm{M/D/1}}(x \mid \varrho)$ overestimates or underestimates $\mathbb{P}(Q_B - Q_0 \geq x)$. The following errors can be identified:

- In fact, the load is not constantly ϱ; it is more accurate to say that the load is k/D with probability $\mathrm{Pois}(k \mid \lambda\mathbb{E}B)$ (recall that the steady-state number of calls present is Poisson with mean $\lambda\mathbb{E}B$). But notice that

$$P_{\mathrm{M/D/1}}(x \mid \varrho) \neq \sum_k P_{\mathrm{M/D/1}}\left(x \mid \frac{k}{D}\right)\mathrm{Pois}(k \mid \lambda\,\mathbb{E}B).$$

  More precisely: the right-hand side of the previous display will majorize the left-hand side, due to the additional call-level variation that is incorporated. Due to this effect the approximation $P_{\mathrm{M/D/1}}(x \mid \varrho)$ may underestimate $\mathbb{P}(Q_B - Q_0 \geq x)$.

- On the other hand, the fact that the M/D/1 queue gives rise to higher delays than the $N\cdot/\mathrm{D/D/1}$ queue, may make the approximation $P_{\mathrm{M/D/1}}(x \mid \varrho)$ too pessimistic.

- Also, Q_0 is stochastically larger than $Q_B - Q_0$ (recall that Q_B has the same distribution as Q_0), which also may make $P_{\mathrm{M/D/1}}(x \mid \varrho)$ too pessimistic.

Hence, if the second and third effect dominate, then the approximation $P_{\mathrm{M/D/1}}(x \mid \varrho)$ is too pessimistic, whereas if the first effect dominates, then it is too optimistic. To gain more insight, we performed some numerical experiments. We find, for $x = 5, 10, 15, 20$, respectively, that $P_{\mathrm{M/D/1}}(x \mid \varrho)$ equals $6.80 \cdot 10^{-2}$, $5.88 \cdot 10^{-3}$, $4.63 \cdot 10^{-4}$, and $3.76 \cdot 10^{-5}$. From the table we see how how these numbers $P_{\mathrm{M/D/1}}(x \mid \varrho)$ compare to $\mathbb{P}(Q_B - Q_0 \geq x)$. It turns out that, in this specific scenario, the estimate is rather accurate (and slightly pessimistic), i.e., the 'positive' and 'negative effects' roughly cancel out. It is not *a priori* clear under what parameter settings this conclusion remains valid.

A counterpart of (9), in the same spirit as the estimate $P_{\mathrm{M/D/1}}(x \mid \varrho)$, is

$$\sum_{y=0}^\infty P_{\mathrm{M/D/1}}(x + y \mid \varrho)P_{\mathrm{M/D/1;e}}(y \mid \varrho),$$

with $P_{\mathrm{M/D/1;e}}(x \mid \varrho)$ being the probability that in the M/D/1 with load ϱ the buffer content is exactly x packets. We obtain, for $x = 5, 10, 15, 20$ the estimates $3.96 \cdot 10^{-2}$, $3.07 \cdot 10^{-3}$, $2.13 \cdot 10^{-4}$, and $1.33 \cdot 10^{-5}$;

from the table it turns out that these numbers are slightly too optimistic, also compared to the situation $i \to \infty$. Again, the M/D/1 assumption has a 'conservative effect' on the estimate, whereas the fact that call-level fluctuations are not taken into account works into the opposite direction; apparently the latter effect dominates.

The table indicates that, in the scenario under consideration, a delay variation (due to call-level fluctuations) of 10 to 15 packets is quite likely, recall also (8). As argued in the intoduction, other factors contribute to the jitter as well. We also recall that jitter can be removed at the expense of additional delay. Other components of the end-to-end delay are queueing delay, packetization time, propagation delay, etc. Also the residual service time of packets of other traffic types (assumed that our streaming application has preemptive priority over these other types of traffic) has to be taken into account. A full procedure to determine the end-to-end delay is sketched in, e.g., [12, 15, 27]. $\diamond$

Remark. Interestingly, Table 2 indicates that $\mathbb{P}(Q_B - Q_0 \geq x)$ can be conservatively estimated by (9). This leads to the explicit expression

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \sum_{y=0}^{\infty} \mathbb{P}\text{ois}(n_1 \mid \lambda \, \mathbb{E}B) \, \mathbb{P}\text{ois}(n_2 \mid \lambda \, \mathbb{E}B) \, \mathbb{P}\{x + y \mid n_1, D\} \, \mathbb{P}_e\{y \mid n_2, D\}$$

The inner summation (i.e., the summation over y, for n_1, n_2 given) can be made more explicit when using the *Brownian bridge approximation*, see [18, Section 15.2.2] and also [22, Exercise 2.2.4]:

$$\mathbb{P}\{x \mid N, D\} \approx \mathbb{P}_{\text{BB}}\{x \mid N, D\} := \exp\left(-2x\left(\frac{x}{N} + 1 - \frac{N}{D}\right)\right).$$

With

$$p_{\text{BB}}\{x \mid N, D\} := -\frac{\partial}{\partial x}\mathbb{P}_{\text{BB}}\{x \mid N, D\},$$

the inner summation is approximated by

$$\int_0^{\infty} \mathbb{P}_{\text{BB}}\{x + y \mid n_1, D\} \, p_{\text{BB}}\{y \mid n_2, D\}\mathrm{d}y.$$

Defining $\rho_i' := 1 - \rho_i = 1 - n_i/D$, for $i = 1, 2$, the above integral reads

$$\int_0^{\infty} \left(4\frac{y}{n_2} + 2\rho_2'\right) \exp\left(-2y^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) - 2y\left(\frac{2x}{n_1} + \rho_1' + \rho_2'\right) - 2\frac{x^2}{n_1}\right) \mathrm{d}y.$$

We obtain after considerable calculus that this integral equals, with $\rho' := \rho_1' + \rho_2'$ and $n := n_1 + n_2$,

$$\frac{n_1}{n} \exp\left(-2\frac{x^2}{n_1}\right) + 2\left(\rho_2' - \frac{4x + 2\rho' n_1}{n}\right) \sqrt{\frac{\pi n_1 n_2}{2n}} \times$$

$$\bar{\Phi}\left(-\frac{2x\sqrt{n_2/n_1} + \rho'\sqrt{n_1 n_2}}{\sqrt{n}}\right) \exp\left(\frac{n_1 n_2}{8n}\left(-4\frac{x}{n_1} + \rho'\right)^2 - 2\frac{x^2}{n_1}\right),$$

where $\bar{\Phi}(\cdot)$ is the complementary standard normal distribution function. As an aside, we remark that for x large the $\bar{\Phi}(\cdot)$-term goes to 1; now it is readily derived that the second term is for large x negligible compared to the first term, such that the integral behaves asymptotically as $(n_1/n) \exp\left(-2x^2/n_1\right)$. $\diamond$

References

[1] V. Beneš. *General stochastic processes in the theory of queues.* Addison Wesley, Reading MA, 1963.

[2] J. Bennett, K. Benson, A. Charny, W. Cortney, and J.-Y. le Boudec. Delay jitter bounds and packet scale rate guarantee for expedited forwarding. *IEEE/ACM Transactions on Networking*, 10: 529 – 540, 2002.

[3] A. Bhargava, P. Humblet, and M. Hluchyj. Queueing analysis of continuous bit-stream transport in packet networks. *Proceedings IEEE Globecom*, 1989.

[4] P. Boyer, A. Dupuis, A. Gravey, and J.-M. Pitie. The output process of the single server queue with periodic arrival process and deterministic service time. In: *Modelling and performance evaluation methodology: proceedings of the international seminar, Paris, France,* F. Baccelli and G. Fayolle, Eds. Lecture Notes in Control and Information Sciences, Vol. 60. Springer, Berlin, 1984.

[5] F. Brichet, L. Massoulié, and J. Roberts. Stochastic ordering and the notion of negligible CDV. *Proceedings ITC 15,* 1433 – 1444, 1997.

[6] A. Dempster. Generalized D_n^+ statistics. *Annals of Mathematical Statistics,* 30: 593 – 597, 1959.

[7] A.E. Eckberg, Jr. The single server queue with periodic arrival process and deterministic service time. *IEEE Transactions on Communications,* 27: 556 – 562, 1979.

[8] A. Gravey. Temps d'attente et nombre de clients dans une file nD/D/1. *Annales d'institut H. Poincaré – Probabilités et Statistiques,* 20: 53 – 73, 1984.

[9] B. Hajek. A queue with periodic arrivals and constant service rate. In *Probability, Statistics and Optimisation—a Tribute to Peter Whittle.* (F.P. Kelly ed.; John Wiley and Sons), 147 – 158, 1994.

[10] P. Humblet, A. Bhargava, and M. Hluchyj. Ballot theorems applied to the transient analysis of nD/D/1 queues. *IEEE/ACM Transactions on Networking*, 1: 81 – 95, 1993.

[11] K. Iida, T. Takine, H. Sunahara, and Y. Oie. Delay analysis for CBR traffic under static priority scheduling. *IEEE/ACM Transactions on Networking*, 9: 177 – 185, 2001.

[12] M. Karam and F. Tobagi. Analysis of the delay and jitter of voice traffic over the Internet. *Proceedings IEEE Infocom*, 824 – 833, 2001.

[13] N. Laoutaris, B. Van Houdt, and S. Stravrakakis. Optimization of a packet video receiver under different levels of delay jitter: an analytical approach. *Performance Evaluation,* 55: 251 – 275, 2004.

[14] M. Mandjes. Packet models revisited: tandem and priority systems. *Queueing Systems*, 47: 363 – 377, 2004.

[15] M. Mandjes, K. van der Wal, R. Kooij, and H. Bastiaansen. End-to-end delay models for interactive services on a large-scale IP network. *Proceedings 7th IFIP workshop on modeling and evaluation of ATM/IP networks*, Antwerp, Belgium, 1999.

[16] I. Norros, J. Roberts, A. Simonian, and J. Virtamo. The superposition of variable bit rate sources in an ATM multiplexer. *IEEE Journal on Selected Areas in Communications*, 9: 378 – 387, 1991.

[17] I. Norros, A. Simonian, D. Veitch, and J. Virtamo. A Beneš formula for a buffer with fractional Brownian input. *Proceedings 9th ITC Specialists Seminar: Teletraffic Modelling and Measurement*, 1995.

[18] J. Roberts, U. Mocci, and J. Virtamo. *Broadband network teletraffic.* Final report of action COST 242. Springer, Berlin, 1996.

[19] J. Roberts and J. Virtamo. The superposition of periodic cell arrival streams in an ATM multiplexer. *IEEE Transactions on Communications*, 39: 298 – 303, 1991.

[20] R. Pyke. The supremum and infimum of the Poisson process. *Annals of Mathematical Statistics,* 30: 568 – 576, 1959.

[21] B. Sengupta. A queue with superposition of arrival streams with an application to packet voice technology. *Proceedings Performance 1990,* 53 – 60, 1990.

[22] G. Shorack and J. Wellner. *Empirical processes with applications to Statistics.* Wiley, New York, 1967.

[23] L. Takács. *Combinatorial methods in the theory of stochastic processes.* Wiley, New York NY, 1967.

[24] H. Takagi. *Queueing Analysis.* North-Holland, Amsterdam, 1991.

[25] J. Virtamo. Idle and busy period distributions of an infinite capacity $N*D/D/1$ queue. *Proceedings ITC 14*, 453 – 459, 1994.

[26] J. Virtamo and J. Roberts. Evaluating buffer requirements in an ATM multiplexer. *Proceedings IEEE Globecom*, 1989.

[27] K. van der Wal, M. Mandjes, and H. Bastiaansen. Delay performance analysis of the new internet services with guaranteed QoS. *Proceedings of the IEEE*, 85: 1947 – 1957, 1997.