Stochastic Bounds for Switched Bernoulli Batch arrivals observed through measurements

F. Aït-Salaht[1], H. Castel-Taleb[2], J-.M. Fourneau[3], and N. Pekergin[4]

[1] LIP6, Ensai, Rennes, France
[2] SAMOVAR, UMR 5157, Télécom Sud Paris, Evry, France
[3] DAVID, UVSQ, Univ. Paris Saclay,Versailles France
[4] LACL, Univ. Paris Est- Créteil, France

Abstract. We generalise to non stationary traffics an approach we have previously proposed to derive performance bounds of a queue under histogram-based input traffics. We use strong stochastic ordering to derive stochastic bounds on the queue length and the output traffic. These bounds provide probability inequalities on transient behaviours and on the steady-state when it exists. We provide some numerical techniques for SBBP traffic. Unlike approximate methods, these bounds can be used to check if the Quality of Service constraints are satisfied. Our approach provides a tradeoff between the accuracy of results and the computational complexity and it is much faster than the histogram-based simulation proposed in the literature.

1 Introduction

Measurements and traces are now much more frequent and we advocate that we can use them to make the performance analysis of networking elements more precise and more realistic. Typically, the traces are used as an input for a fitting algorithm which finds the best approximation inside a class of well-known stochastic processes (see, for instance, [12]). When this process can be associated to a Markov process or chain, the whole system can be modelled by a so-called structured Markov chain (see [13] for an example) and many algorithms have been derived to solve the steady-state distribution for this type of models.

In [2, 3], we have proposed a different approach for stationary arrivals: we model the system in discrete time and we use directly the measurements to obtain a discrete distribution of arrivals during a time slot. Thus, we avoid the fitting procedure and the approximations it may add in the model. Such an approximation due to the fitting of the processes may lead to incorrect results (see [6] for such a problem for service time distributions).

Such an idea has already been proposed and is known as the histogram based models for more than 20 years (see for instance, the work by Skelly et al. [16] in the area of network calculus to model the video sources and to predict buffer occupancy distributions). Recently, Hernàndez et al. [9–11] have introduced an approach called HBSP (Histogram Based Stochastic Process) to obtain histograms of buffer occupancy. Their use histograms as inputs and some specific operators

in discrete time to represent a finite capacity buffer with a constant service under the First Come First Served (FCFS) discipline. The model is solved numerically and as usual, the curse of dimensionality appears. When the number of bins in the histograms is too large, the computation times become extremely high and the authors present an approximation of the histograms of traffic which leads to a smaller complexity and a faster resolution. Unfortunately their accuracy of the approximation cannot be checked.

We propose a more accurate method to deal with histograms having a large number of bins. First in [4] we prove that the system is stochastically monotone. This allows to obtain bounds on the queue size and the output process when we consider bounds on the input process. Second, in [2] we provide several algorithms to derive stochastic bounds of the arrival process with a smaller complexity. As we build lower and upper bounds, our approach provides an estimation of the approximations. The complexity in the numerical computations in basically dependent of the number of bins in the histogram or the number of atoms in the discrete distribution. The main assumption of the approach is the stationarity of the input process.

Here, we do not assume that the traffic is stationary. Typical Internet services such as web surfing and high speed streaming services (Video On Demand (VOD) and video conferencing), tend to generate sporadic traffic, and hence it would be realistic to consider bursty packet arrivals for today's telecommunication traffic. There are some interesting queueing models and analytical results considering bursty sources and discrete time queueing systems.

In [20], they consider finite capacity queue in discrete time with constant service time of arbitrary length, and bursty on/off source with geometric distributed lengths of the phase. Closed form are derived for the loss ratio of cells. In [21] an infinite capacity discrete-time queue with Bernoulli bursty source and batch arrivals is analysed using the generating function technique. A closed form expressions of some performance measures as average buffer length, and average delays are obtained. Markov modulated arrivals have been quite often considered in the literature to represent traffic arrivals [15, 5]. In [15], they define an MMPP (Markov Modulated Poisson Process) traffic model that accurately approximates the characteristics of Internet traffic traces. Results prove that the queuing behaviour of the traffic generated by the MMPP model is coherent with the one produced by real traces. Some important results on MMPP traffic and queues with MMPP input are described in [5].

In this paper, we propose to apply stochastic bounds on the input traffic to derive stochastic bounds on the queue length and the departure flow. We propose a numerical technique to compute the bounds in an efficient way. We show how our approach which have been developed for stationary arrivals can be generalised to Switched Bernoulli Batch Process (SBBP in the following).

The technical part of the paper is as follows; We introduce briefly bounds for the $\leq_{st}$ ordering in the next section for the sake of completeness. We advocate that monotonicity of the evolution equation as well as stochastic bounds may help to solve such a queueing model when the arrival process is not stationary.

We first considered the stationarity assumption to derive some results, theorems and algorithms in section 3 which will be then generalised for non stationary arrival processes in Section 4.

2 A brief presentation of stochastic comparison

We refer to [14] for theoretical issues of the stochastic comparison method. We consider state space $\mathcal{G} = \{1, 2, \ldots, n\}$ endowed with a total order denoted as $\leq$. Let X and Y be two discrete random variables taking values on $\mathcal{G}$, with cumulative probability distributions F_X and F_Y, and probability mass functions (pmf) $\boldsymbol{d}2$ and $\boldsymbol{d}1$. The *ith* index of pmf vectors denotes the probability that the underlying random value takes value i: $\boldsymbol{d}2(i) = \text{Prob}(X = i)$, and $\boldsymbol{d}1(i) = \text{Prob}(Y = i)$, for $i = 1, 2, \ldots, n$. The stochastic comparison of two random variables in the sense of the strong stochastic order, $\leq_{st}$ can be defined as follows.

Definition 1. *The following definitions are equivalent.*

- **generic definition:**

$$X \leq_{st} Y \Longleftrightarrow \mathbb{E}f(X) \leq \mathbb{E}f(Y),$$

  for all increasing (non decreasing) functions $f : \mathcal{G} \to \mathbb{R}^+$ *whenever expectations exist.*
- **cumulative probability distributions:**

$$X \leq_{st} Y \Leftrightarrow F_X(a) \geq F_Y(a),\ \forall a \in \mathcal{G}.$$

- **probability mass functions:**

$$X \leq_{st} Y \Leftrightarrow \forall i,\ 1 \leq i \leq n,\ \sum_{k=i}^{n} \boldsymbol{d}2(k) \leq \sum_{k=i}^{n} \boldsymbol{d}1(k) \tag{1}$$

  Notice that we use interchangeably $X \leq_{st} Y$ *and* $\boldsymbol{d}2 \leq_{st} \boldsymbol{d}1$.

Property 1. If $X \leq_{st} Y$, then for any increasing function f,

$$f(X) \leq_{st} f(Y)$$

Example 1. We consider two discrete random variables with $\boldsymbol{d}2 = [0.1, 0.2, 0.1, 0.2, 0.05, 0.1, 0.25]$, and $\boldsymbol{d}1 = [0.3, 0.05, 0.1, 0.15, 0.1, 0.3]$ defined respectively on support $\{1, \ldots, 7\}$ and $\{2, \ldots, 7\}$. The set $\mathcal{G}$ is the union of support of the two distributions $\boldsymbol{d}1$ and $\boldsymbol{d}2$ with null probabilities if an element does not belong to one of them. We can easily verify that $\boldsymbol{d}2 \leq_{st} \boldsymbol{d}1$: the probability mass of $\boldsymbol{d}1$ is concentrated to higher states, and the probability cumulative distribution of $\boldsymbol{d}1$ is always below the cumulative distribution of $\boldsymbol{d}2$ (see Figure 1).

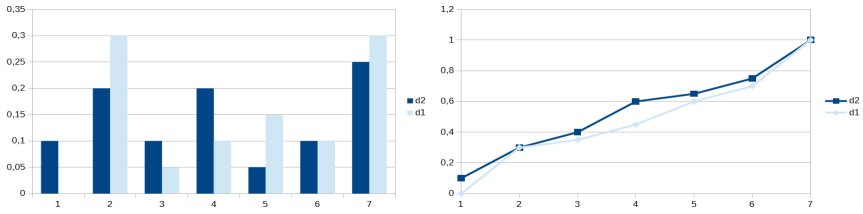


Fig. 1. $d2 \leq_{st} d1$: Probability mass functions (left) and cumulative distribution functions (right).

The $\leq_{st}$ ordering is closed under mixture (Theorem 1.2.15 in page 6 of [14]):

Theorem 1. *If X, Y and Θ are random variables such that $[X \mid \Theta = \theta] \leq_{st} [Y \mid \Theta = \theta]$ for all θ in the support of Θ, then $X \leq_{st} Y$.*

The following definition is used to compare Markov chains.

Definition 2. *Let $\{X(n),\, n \geq 0\}$ (resp. $\{Y(n),\, n \geq 0\}$) be a DTMC. We say $\{X(n),\, n \geq 0\} \leq_{st} \{Y(n),\, n \geq 0\}$, if $X(n) \leq_{st} Y(n),\ \forall n \geq 0$.*

Let $\mathbf{P}$ and $\mathbf{Q}$ be the probability transition matrix of $\{X(n),\, n \geq 0\}$ and $\{Y(n),\, n \geq 0\}$ respectively. If the chains are ergodic, let $\boldsymbol{\pi}_{\mathbf{P}}$ and $\boldsymbol{\pi}_{\mathbf{Q}}$ denote the corresponding steady state distributions, then $\boldsymbol{\pi}_{\mathbf{P}} \leq_{st} \boldsymbol{\pi}_{\mathbf{Q}}$.

The following theorem provides sufficient conditions to establish the comparison of DTMCs.

Theorem 2. *Let $\mathbf{P}$ (resp. $\mathbf{Q}$) be the probability transition matrix of the time-homogeneous Markov chain $\{X(n),\, n \geq 0\}$ (resp. $\{Y(n),\, n \geq 0\}$). The comparison of Markov chains is established $\{X(n),\, n \geq 0\} \leq_{st} \{Y(n),\, n \geq 0\}$, if the following conditions are satisfied*

- *$X(0) \leq_{st} Y(0)$,*
- *at least one of the probability transition matrices is monotone, that is, either $\mathbf{P}$ or $\mathbf{Q}$ (say $\mathbf{P}$) is $\leq_{st}$ monotone, if for all probability vectors $\mathbf{p}$ and $\mathbf{q}$,*

$$\mathbf{p} \leq_{st} \mathbf{q} \Longrightarrow \mathbf{p}\,\mathbf{P} \leq_{st} \mathbf{q}\,\mathbf{P}$$

  which is equivalent to

$$1 \leq i \leq n-1, \quad \boldsymbol{P}[i,*] \leq_{st} \boldsymbol{P}[i+1,*]$$

  where $\boldsymbol{P}[i,*]$ denotes the row of matrix $\mathbf{P}$ for state i.
- *the transition matrices are comparable in the sense of the $\leq_{st}$ order :*

$$\boldsymbol{P} \leq_{st} \boldsymbol{Q} \Leftrightarrow 1 \leq i \leq n,\ \boldsymbol{P}[i,*] \leq_{st} \boldsymbol{Q}[i,*]$$

3 Bounding performance measures under stationary traffic

We present in this section the method we have developed in various publications [4, 2, 3].

3.1 Queue model and Evolution equations

Let us begin with some notation. The number of transmission units produced by the traffic source during the k^{th} slot is denoted by $A(k)$, and $Q(k)$ and $D(k)$ are respectively the buffer length and the output (departure) traffic (flow) during the k^{th} slot. The buffer size is noted by B and the service capacity during a slot by S. The input parameter $A(k)$ is specified by a discrete distribution (histogram), and the output parameters are also derived as histograms.

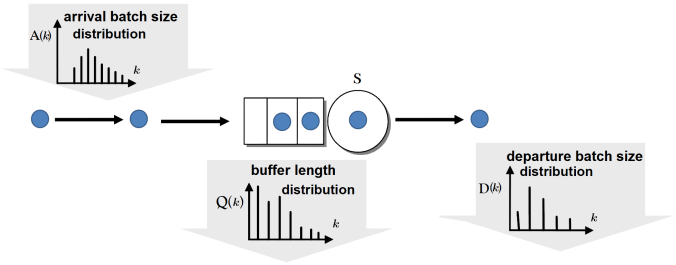


Fig. 2. Input and output parameters of a queueing model

The admission per packet is done with Tail Drop policy. Thus an arrival packet is accepted if there is a place in the buffer, otherwise it is rejected. The timing of events during a slot is as follows: arrivals occur first and they are followed immediately by services. The evolution equations for the buffer length ($Q(k)$) and the departure traffic ($D(k)$) can be given as follows:

$$Q(k) = \texttt{min}(B,\, (Q(k-1) + A(k) - S)^+), \quad k \geq 1, \tag{2}$$

where operator $(X)^+ = \texttt{max}(X,\, 0)$.

$$D(k) = \texttt{min}(S,\, Q(k-1) + A(k)), \quad k \geq 1. \tag{3}$$

The model of the queue is a time-inhomogeneous Discrete Time Markov Chains (DTMC), if the input arrivals are independent of the current queue state and the past of the arrival process. Under the stationary arrival assumptions, the underlying DTMC is time-homogenous.

The monotonicity of these equations under the $\leq_{st}$ order has been proved in [2, 3]. Intuitively speaking, the monotonicity property states that if we consider two models under different arrival processes but comparable in the sense of the $\leq_{st}$ order, then the corresponding output parameters are also comparable in the sense of the $\leq_{st}$ order.

Let consider two queues. The first one is under arrival process $A(k),\ \ k \geq 0$, and the output parameters (queue length, and departure traffic) noted by $Q(k)$ and $D(k), k \geq 0$. The second one is under arrival process $\tilde{A}(k)$, with output parameters: $\tilde{Q}(k)$, $\tilde{D}(k)$). At the beginning, $Q(0) \leq_{st} \tilde{Q}(0)$ and $D(0) \leq_{st} \tilde{D}(0)$. Without loss of generality, we assume that the queues are idle at $k = 0$, thus the

queue lengths and the departure processes are empty, thus $Q(0) =_{st} \tilde{Q}(0)$ and $D(0) =_{st} \tilde{D}(0)$.

Theorem 3. *If* $A(k) \leq_{st} \tilde{A}(k),\ \forall k > 0$, *then*

$$Q(k) \leq_{st} \tilde{Q}(k), \quad \textit{and} \quad D(k) \leq_{st} \tilde{D}(k),\ \forall k > 0.$$

The monotonicity results follow from the fact that the $\leq_{st}$ order is associated to increasing functions and the underlying measures are defined by increasing functions of input parameters.

This theorem lets us to construct bounding systems. For instance, for a given system, let say the one under the arrival process $A(k)$, it is possible to construct bounding performance measures, $\tilde{Q}(k)$, $\tilde{D}(k)$ by considering the bounding arrival process $\tilde{A}(k)$. Obviously, this approach is meaningful if the analysis under arrival $\tilde{A}(k)$ is more efficient to do. Notice that these are transient bounds thus the comparisons are satisfied at each instant k, and also for the steady state if it exists.

If a stationary bounding process $\tilde{\mathcal{A}}$ exists such that $A(k) \leq_{st} \tilde{A},\ \forall k > 0$, it has been proved that the stationary bounding performance measures can be derived by considering the system under the stationary bounding process $\tilde{\mathcal{A}}$ [3]. Clearly if both the real traffic ($A(k)$) and the (upper) bounding traffic ($\tilde{\mathcal{A}}(k)$), are stationary, we have the following corollary:

Corollary 1. *Let* $\mathcal{A}$ *(resp.* $\tilde{\mathcal{A}}$*) be the stationary exact (resp. upper bounding) input histogram (distribution) such that* $\mathcal{A} \leq_{st} \tilde{\mathcal{A}}$*, and* $\mathcal{Q}$, $\mathcal{D}$ *(resp.* $\tilde{\mathcal{Q}}$, $\tilde{\mathcal{D}}$*) be the stationary buffer length, departure flow under the exact* $\mathcal{A}$*, (resp. upper bounding* $\tilde{\mathcal{A}}$*) input arrival. If* $Q(0) \leq_{st} \tilde{Q}(0)$*, and* $D(0) \leq_{st} \tilde{D}(0)$*, then we have:*

$$\mathcal{Q} \leq_{st} \tilde{\mathcal{Q}} \ \textit{and}\ \mathcal{D} \leq_{st} \tilde{\mathcal{D}}.$$

The lower bounding case can be similarly derived.

3.2 Bounding histogram construction

The complexity of the numerical analysis of performance measures (Eq. 2-3) depends on the arrival distributions whatever the used method is. We advocate that, as the queue we model is stochastically monotone, it is possible to aggregate the input distribution (to reduce the number of atoms) for deriving in an easier way stochastic bounds on the performance measures. For a discrete distribution of probability, the complexity parameter is the number of atoms. Therefore we propose to apply the bounding approach to make the number of atoms smaller. The main advantage of this approach is the computation of bounds rather than approximations. Unlike approximations, the bounds allow us to have guarantees and check if QoS are satisfied or not.

Let the input arrival process is specified by a probability mass function (discrete distribution) $\boldsymbol{d}$ defined on N atoms. In [4], we have proposed an algorithm to build an upper and a lower bounding distribution, $\boldsymbol{d}1$ and $\boldsymbol{d}2$ with $n << N$ atoms. Moreover, $\boldsymbol{d}1$ and $\boldsymbol{d}2$ are the optimal bounds with respect to a given

positive, increasing reward function, $\boldsymbol{r}$. Formally, for a given distribution $\boldsymbol{d}$ defined on $\mathcal{H}$ ($|\mathcal{H}| = N$), we compute bounding distributions $\boldsymbol{d}1$ and $\boldsymbol{d}2$ defined respectively on $\mathcal{H}^u$, $\mathcal{H}^l$ ($|\mathcal{H}^u| = n$, $|\mathcal{H}^l| = n$) such that:

1. $\boldsymbol{d}2 \leq_{st} \boldsymbol{d} \leq_{st} \boldsymbol{d}1$,
2. $\sum_{i\in\mathcal{H}} \boldsymbol{r}(i)\boldsymbol{d}(i) - \sum_{i\in\mathcal{H}^l} \boldsymbol{r}(i)\boldsymbol{d}2(i)$ is minimal among the set of distributions on n atoms that are stochastically lower than $\boldsymbol{d}$,
3. $\sum_{i\in\mathcal{H}^u} \boldsymbol{r}(i)\boldsymbol{d}1(i) - \sum_{i\in\mathcal{H}} \boldsymbol{r}(i)\boldsymbol{d}(i)$ is minimal among the set of distributions on n atoms that are stochastically upper than $\boldsymbol{d}$.

Notice that $\forall\, i \in \mathcal{H}$ and $i \notin \mathcal{H}^u$ (resp. $\forall\, i \in \mathcal{H}$ and $i \notin \mathcal{H}^l$), $\boldsymbol{d1}(i) = 0$ (resp. $\boldsymbol{d2}(i) = 0$) to establish the stochastic comparisons. Thus $\boldsymbol{d}1$ and $\boldsymbol{d}2$ denote the optimal bounding distributions on n atoms with respect to reward $\boldsymbol{r}$.

The proposed algorithm is based on dynamic programming and has a complexity of $O(N^2\, n)$. Some heuristics with a smaller complexity which let to construct stochastic bounds with the required number of atoms but which are not in general optimal can be found in the same reference.

The number of atoms provide a trade-off between the accuracy of the bounds and the computation time. It can be determined in an incremental manner: one begins with a reduced number of atoms, if the accuracy of bounds is not satisfactory, the number of atoms can be incremented. The iteration can be stopped, if the required accuracy is reached and/or the computation time of bounds exceeds a fixed threshold.

Example 2. Let $\boldsymbol{d} = [0.1, 0.4, 0.05, 0.15, 0.1, 0.2]$ be a discrete distribution defined on a support $\mathcal{H} = \{1, 2, 3, 4, 5, 6\}$ ($N = 6$). For reward function $\boldsymbol{r}$ sets to $\boldsymbol{r}(i) = a_i$, $\forall\, a_i \in \mathcal{H}$, the expected reward of $\boldsymbol{d}$ is $R[\boldsymbol{d}] = \sum_{a_i\in\mathcal{H}} \boldsymbol{r}(i)\ \boldsymbol{d}(i) = 3.35$.

The computation of the optimal stochastic upper bound $\boldsymbol{d}1$ (resp. lower bound $\boldsymbol{d}2$) of $\boldsymbol{d}$ with only 3 states (atoms) consists in exploring all 3 single hops paths from the largest (resp. smallest) atom and select the path for which $R[\boldsymbol{d}1] - R[\boldsymbol{d}]$ ($R[\boldsymbol{d}] - R[\boldsymbol{d}2]$) is the minimal.

We illustrate in the following figure the probability mass functions and the cumulative distribution functions of the exact and the computed optimal bounding distributions. The expected reward of the bounding distributions are: $R[\boldsymbol{d}2] = 3.1$ and $R[\boldsymbol{d}1] = 3.8$.

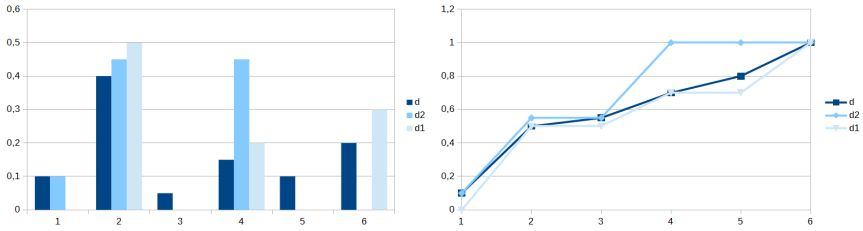


Fig. 3. $\boldsymbol{d}2 \leq_{st} \boldsymbol{d} \leq_{st} \boldsymbol{d}1$: Probability mass functions (left) and cumulative distribution functions (right).

3.3 Performance measure bounds under stationary arrivals

We are indeed interested in the performance analysis of the queue under real traffic traces. We present here an example given in [3] under stationary traffic assumption of real traces. We illustrate in Figure 4, a real traffic trace extracted from the MAWI traffic traces [17]. Precisely, it corresponds to an IP traffic trace during one hour for a 150 Mbps transpacific line (samplepoint-F) for the 9th of January 2007 between 12:00 and 13:00. This traffic trace has an average rate of 109 Mbps. Using a sampling interval of T = 40 ms (25 samples per second), the resulting traffic trace has 90,000 frames (periods), an average of 4.37 Mb per frame and 80511 distinct values (atoms).

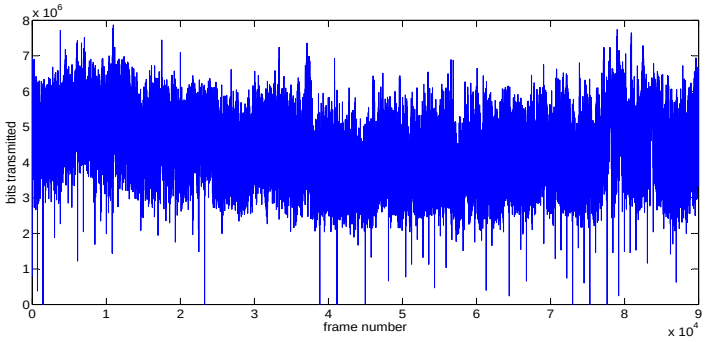


Fig. 4. MAWI traffic trace.

We present in Figure 5, the lower and upper bounding histograms with $n = 10$ atoms for this trace, and the exact histogram without size reduction. The reward considered in the histogram size reduction algorithm is the identity function in order to construct optimal $\leq_{st}$ bounds with respect to the expectation. The expectation of the original histogram (noted as exact) is 4.3757×10^6 bits while the expectation of the upper bound is 4.5843×10^6 bits, and that of the lower bound is 4.1644×10^6 bits.

In Figure 6, some performance measures under MAWI traffic using stochastic bounds are given. We present respectively the blocking probability and the mean buffer length for different values of reduction (atoms varying from 10 to 200). In each figure, we give the results computed under: 1) exact MAWI histogram (without reduction 80511 atoms), 2) Lower bound histogram and 3) Upper bound histogram.

Through this example, we can see that the computed performance measures are bounds on exact results. We observe also that when the size of the support increases, the bounds become tighter. So, we can see here that in an empirical way, we can reach the required accuracy with reduced complexity.

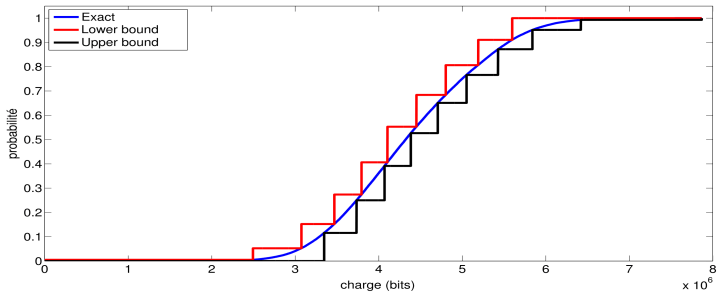
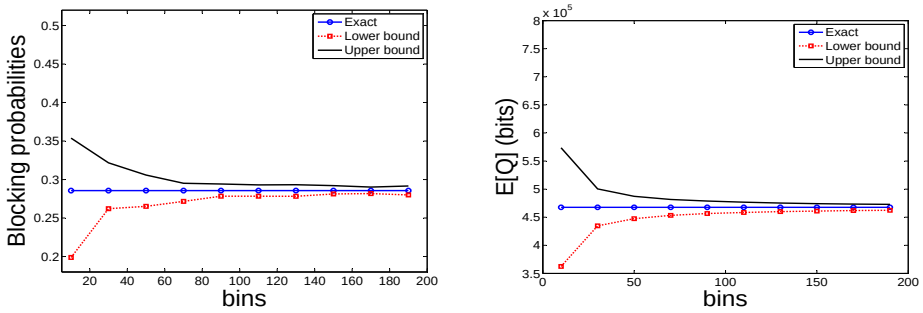


Fig. 5. Cumulative probability distributions (cdf) for the MAWI traffic.



(a) Blocking probability (b) Mean buffer length

Fig. 6. Accuracy versus the number of atoms: QoS parameters using the MAWI traffic

4 SBBP input traffic

We now consider that the traffic is modelled by a *Switched Batch Bernoulli Process* (SBBP)[8], and we show that our method can also be applied in this case. The SBBP process is an arrival process modulated by a Markov chain. This model is useful to characterise phase-dependent arrivals, i.e. the arrival processes during different phases have different characteristics. If there are p arrival phases, the phase evolution is controlled by a time-homogeneous DTMC defined on state space $\mathcal{P} = \{1, \ldots, p\}$, Let $\mathbf{F}$ be the probability transition matrix for phase changes, then $\mathbf{F}(i, j)$ is the probability of the transition from phase i to phase j.

The state of the system at time k can be denoted by $QP(k) = (Q(k), \phi(k))$. The first component $Q(k)$ is the number of entities in the buffer and $\phi(k)$ is the arrival phase during slot k. In each arrival phase $i \in \mathcal{P}$, the arrival process $\mathcal{A}^i$ is assumed to be stationary and independently, identically distributed. The underlying system $\{QP(k), k \geq 0\}$ is a time-homogeneous DTMC. During time k, the arrival phase is $\phi(k)$, and the evolution of $Q(k)$ is the same as in the stationary arrival case, but under arrival $\mathcal{A}^{\phi(k)}$ instead of $\mathcal{A}$. Thus, $Q(k)$ takes

values in the set $\mathcal{N} = \{0 \cdots B\}$, and evolves as follows:

$$Q(k+1) = \texttt{min}\left(B, (Q(k) + \mathcal{A}^{\phi(k)} - S)^+\right).$$

The evolution of the second component, $\phi(k)$ is controlled by a Markov chain. The state space of $\{QP(k), k \geq 0\}$ is the product space $\mathcal{S} = \mathcal{N} \times \mathcal{P}$.

4.1 Bounds under SBBP input traffic

We construct the bounding models by fixing the arrival phase, and the comparisons are established arrival phase by arrival phase. The comparison of two states $x, y \in \mathcal{S}$ is defined by the partial order $\preceq$ on $\mathcal{S}$:

Definition 3. *Let $x = (x_q, x_p), y = (y_q, y_p) \in \mathcal{S}$, where the first components correspond to the buffer lengths (Q) and the second components correspond to the arrival phases (ϕ).*

$$x \preceq y \quad \textit{iff} \;\; x_q \leq y_q \;\; \textit{and} \;\; x_p = y_p$$

In the bounding system denoted by $\tilde{QP}(k)$, the arrival processes in each phase ($\tilde{\mathcal{A}}^i$) are the upper bounds of the real traffic ($\mathcal{A}^i$) and they are constructed as explained in subsection 3.2. Formally,

$$\forall i \in \mathcal{P}, \qquad \mathcal{A}^i \leq_{st} \tilde{\mathcal{A}}^i \tag{4}$$

The second component of both models are controlled by the same DTMC independently of the first component. We assume that at the beginning, $k = 0$,

$$(Q(0), \phi(0)) =_{st} (\tilde{Q}(0), \tilde{\phi}(0)).$$

Thus, if we start with the same initial states in both models, the evolution of the second component will be the same at each time k.

Corollary 2. *Let $\mathcal{Q}, \mathcal{D}$ be the steady-state marginal distributions of the buffer length and the departure flow under arrival distributions $\mathcal{A}^i$, while $\tilde{\mathcal{Q}}$ and $\tilde{\mathcal{D}}$ denote the corresponding distributions under the upper bounding arrival distributions $\tilde{\mathcal{A}}^i$.*

If $\mathcal{A}^i \leq_{st} \tilde{\mathcal{A}}^i, \quad \forall i \in \mathcal{P}$, then

$$\mathcal{Q} \leq_{st} \tilde{\mathcal{Q}} \;\; \textit{and} \;\; \mathcal{D} \leq_{st} \tilde{\mathcal{D}}.$$

Proof. By fixing the arrival phase, we derive bounds on conditional distributions. At each time k, for all arrival phases $i \in \mathcal{P}$, we have:

$$[Q(k) \mid \phi = i] \leq_{st} [\tilde{Q}(k) \mid \phi = i] \text{ and } [D(k) \mid \phi = i] \leq_{st} [\tilde{D}(k) \mid \phi = i].$$

As the $\leq_{st}$ ordering is closed under mixtures (Theorem 1 in Section 2), we have the comparison of the marginal distributions at each time k:

$$Q(k) \leq_{st} \tilde{Q}(k) \text{ and } D(k) \leq_{st} \tilde{D}(k).$$

By construction, the steady-states exist, then it follows from the convergence in distribution:

$$\mathcal{Q} \leq_{st} \tilde{\mathcal{Q}} \text{ and } \mathcal{D} \leq_{st} \tilde{\mathcal{D}}.$$

4.2 Numerical Analysis

Due to the SBBP arrivals, we have a block structured Markov chain.

$$\mathbf{P} = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1p} \\ P_{21} & P_{22} & \cdots & P_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ P_{p1} & P_{p2} & \cdots & P_{pp} \end{pmatrix}.$$

Let (x_1, x_2) and (y_1, y_2) two states of the DTMC, and let $\mathbf{R}_\phi$ be the transition matrix of the system when the arrivals are in phase ϕ. The transition matrix $\mathbf{P}$ of the Markov chain $(QP(k))$ is

$$\mathbf{P}((x_1, x_2), (y_1, y_2)) = \mathbf{F}(x_2, y_2)\mathbf{R}_{x_2}(x_1, y_1).$$

Such a structured matrix is denoted as a functional Kronecker product in the theory of Stochastic Automata Networks [19, 7]. It has many important properties which can be taken into account to obtain efficient numerical techniques.

Property 2. The Markov chain $(QP(k))$ of the model with SBBP arrival is lumpable according to the partition defined by the phase of the arrival process.

Let $\pi_{\mathbf{P}}$ (resp. $\pi_{\mathbf{F}}$) be the steady-state distribution for matrix $\mathbf{P}$ (resp. $\mathbf{F}$). We know that the lumpability implies that there exists p vectors ψ_j of size $B+1$, denoting the conditional queue length probabilities when the arrival phase is j. The stationary distribution $\pi_{\mathbf{P}}$ is then computed as follows:

$$\pi_{\mathbf{P}}(i) = \sum_{j=1}^{p} \pi_{\mathbf{F}}(j)\ \psi_j(i), \quad \forall\, i = 0 \cdots B.$$

To compute the steady-state solution of the model, we use the Iterative Aggregation Disaggregation (IAD) algorithm specialised for lumpable matrices published in [7] to obtain successive values of vectors ψ_i which are denoted $\psi_i^{(t)}$ at iteration t. This algorithm is based on the following steps.

1. Initialise $\psi_i^{(0)}$, for all i
2. Compute $\pi_{\mathbf{F}}$, the steady state probability vector of $\mathbf{F}$
3. Compute vectors $\psi_i^{(t+1)}$ using a Block Gauss Seidel iteration for matrix $\mathbf{P}$ in block form:
   a) $Z_i^{(t+1)} = \pi_{\mathbf{F}}(i)\frac{\psi_i^{(t)}}{||\psi_i^{(t)}||_1},\ \forall\, i = 1 \cdots p$
   b) $\psi_i^{(t+1)} = \psi_i^{(t)}\,\mathbf{P}_{ii} + \sum_{j=i+1}^{p} Z_j^{(t+1)}\,\mathbf{P}_{ji} + \sum_{j=1}^{i-1} \psi_j^{(t+1)}\,\mathbf{P}_{ji},\ \forall\, i = 1 \cdots p$
4. Normalise vectors $\psi_i^{(t+1)}$ to be distributions of probability
5. If $\sum_i ||\psi_i^{(t+1)} - \psi_i^{(t)}||_\infty$ is smaller than a threshold, go to step 6. Otherwise set $t = t + 1$ and go to step 3.
6. Compute $\pi_{\mathbf{P}}$, such that: $\pi_{\mathbf{P}}(i) = \sum_{j=1}^{p} \pi_{\mathbf{F}}(j)\ \psi_j^{(t)}(i), \quad \forall\, i = 0 \cdots B.$

Theoretically, in the first step we can initialise vectors $\psi_i^{(0)}$ with any distribution of probability. Taking into account the properties of the arrivals during the phases as defined in the next paragraph, we have used three phases and the following guess: $\psi_1^{(0)} = \delta_0$, $\psi_3^{(0)} = \delta_B$, and $\psi_2^{(0)}$ equal to the steady state probability vector of matrix $\mathbf{R}_2$ (transition matrix of the system when the arrivals are in phase 2).

4.3 Numerical Results

In order to illustrate the results stated in this paper, we propose to compute the performance measures of a finite single queue under real traffic trace modelled as SBBP arrival process and constant service. We consider the MAWI trace [17] which corresponds to a little more than 10 hours of an IP traffic on transpacific line with link capacities of 128 Kbps, carried between the 6th of march 2007 at 18 : 00 and the 7th of march 2007 at 4 : 24 : 27. For a sampling period $T = 40$ ms, we obtain the trace shown in Figure 7 with 922873 frames and 4579 different atoms.

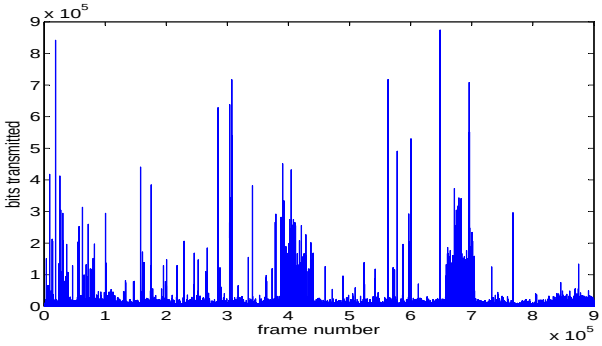


Fig. 7. MAWI traffic trace (more than 10 hours).

We distinguish three phases, phase 1 corresponds to low traffic, phase 2 to medium traffic, and phase 3 to heavy traffic. We assume that for each slot, the traffic trace is characterised by its volume per sampling period. If the traffic per sampling period is less or equal to the minimum threshold (10 Kbps), the arrival phase is 1, and if it is greater or equal to the maximum threshold (100 Kbps), the arrival phase is 3. When the traffic is between the thresholds, the arrival phase is 2. In each phase, the traffic is defined by a stationary arrival process associated to this phase. The probability transition matrix for phase modulation is defined as follows:

$$\mathbf{F}(i,\, j) = \frac{\text{number of transition between phase } \ i \ \text{ and phase } j}{\text{number of slots in phase } \ i}.$$

The resulting transition matrix of phases $\mathbf{F}$ is

$$\mathbf{F} = \begin{pmatrix} 0.9982 & 0.0018 & 0.0000 \\ 0.5563 & 0.4163 & 0.0274 \\ 0.2706 & 0.2615 & 0.4679 \end{pmatrix}$$

The histogram of each phase is defined respectively on 1228 atoms (phase 1), 2568 atoms (phase 2) and 783 atoms (phase 3). They are characterised by the following statistical descriptions:

	Expected value (bits)	Standard deviation (bits)	Coefficient of variation
Phase 1	433.56	1.0503×10^3	5.8684
Phase 2	28953	2.13×10^4	0.5413
Phase 3	2.1515×10^5	1.2844×10^5	0.3564

Table 1. Statistical descriptions of the considered MAWI traffic trace.

Let us emphasize here that our goal is not to study how to obtain an accurate SBBP model for a given trace. We just aim to construct such a model to apply our bounding algorithms and explain how our approach works and can be accurate if the input arrival is a SBBP process. The thresholds have been arbitrarily chosen. The statistical analysis of traces to derive fitting models is out of the scope of this paper.

We now apply our numerical bounding approach to this model to obtain two performance measures (expected buffer length and blocking probability) versus the buffer size (B) which varies from 100 Kb to 3 Mb. We consider a deterministic service capacity of 35 Kbps. The bounding histograms (noted by $L.b$ for the lower bound and by $U.b$ for the upper bound) are constructed on reduced state space with 100 atoms. The exact results (without reduction) and the bounds of these performance measures are given in Table 2. The computation times are presented in Table 3.

We observe that the computed bounds under SBBP arrivals are relevant, essentially for the upper bound and the accuracy of bounds is not degraded when the histogram sizes increase. In terms of complexity, we remark that the computation times of bounds are significantly less than the exact one (the computation time is divided approximately by three when $B = 10^6$, by four when $B = 2 \times 10^6$, and by five for $B = 5 \times 10^6$). In view of these results, we can conclude and say that in order to satisfy the required QoS constraints the use of stochastic bounds for SBBP arrivals observed through measurements offer to the user an interesting tradeoff between accuracy of the results and the computational complexity.

Regarding the difference between stationary input traffic and SBBP traffic, we note that the blocking probabilities and the expected buffer length are much greater for SBBP traffic except for the small buffer values. This phenomenon is

	SBBP input traffic						Stationary input traffic	
	Blocking probabilities (BP)			Expected buffer length (E[Q])			BP	E[Q]
B	Exact	L.b	U.b	Exact	L.b	U.b	Exact	
10^5	0.0032218	0.0031552	0.0035345	19297.7	17628.2	19448.6	0.00419	21651.7
2×10^5	0.0021574	0.0020811	0.0022456	46352.2	41679.7	46696.1	0.00238	51641.8
5×10^5	0.0012534	0.0011796	0.0013074	154693	137686	156254	0.00101	147084
10^6	0.0008447	0.0007416	0.0008902	401307	351574	405695	0.000295	260630
2×10^6	0.0005545	0.0004148	0.0005890	975858	813564	985124	1.75742e-05	304474
5×10^6	0.0003562	0.0001691	0.0003835	3046090	2205710	3077930	1.62373e-10	306020

Table 2. Blocking probabilities and expected buffer lengths versus buffer size.

	SBBP input process			Stationary input process
B	Exact	L.b	U.b	Exact
10^5	2.58	2.41	2.29	78.6
2×10^5	5.584	4.45	3.89	554.55
5×10^5	41.94	17.37	17.70	3710.7
10^6	203.61	74.08	79.57	7564.05
2×10^6	1180.62	359.61	422.9	13736.2
5×10^6	14085	3325	3695	44999.4

Table 3. Computation times in second.

due to the dependence of the variance of the arrival process. We note that the mean input stream for both traffic (stationary and SBBP) is identical, however we have more variance in the SBBP model which is reflected in its performance measures.

5 Conclusion

The stochastic performance bounds of a queue under stationary histogram-based input traffics is generalised to the Markov modulated arrivals. The traffic is assumed to be stationary during a phase and the traffic phase transition is controlled by a DTMC. We illustrate the applicability of this approach by giving some numerical results for a system with arrivals derived from a real traffic trace. We want to emphasize that despite the bivariate process we can use strong stochastic bounds rather than weak or weak* comparisons (see [14]). The techniques we develop here and the associated publications [2, 3] lead to an algorithmic analysis of queues based on measurements for the arrival process.

References

1. Caida, traces of oc48 link at ames internet exchange (aix) (april 24, 2003), accessed via datcat - internet data measurement catalog, http://imdc.datacat.org.

2. Aït-Salaht F, Castel Taleb H, Fourneau J.-M., and Pekergin N (2013) Stochastic bounds and histograms for network performance analysis. In *10th European Workshop on Performance Engineering (EPEW'13)*, volume 8168, pages 13–27. Lecture Notes in Computer Science.
3. Aït-Salaht F, Castel Taleb H, Fourneau J.-M., and Pekergin N (2016) In *Computer Journal*, 59(12): 1817-1830, 2016.
4. Aït-Salaht F, Cohen J, Castel Taleb H, Fourneau J. M, and Pekergin N (2012) Accuracy vs. complexity: the stochastic bound approach. In *11th International Workshop on Discrete Event Systems*, pp 343–348.
5. Fischer W.,Hellstern K.M, (1992) The Markov-modulated Poisson process (MMPP) cookbook, *Performance Evaluation*, 18 : 149-171.
6. Gupta V, Harchol-Balter M, Dai J. G, Zwart B (2010) On the inapproximability of $M/G/K$: why two moments of job size distribution are not enough. *Queueing Syst.*, 64(1):5–48.
7. Gusak O, Dayar T, Fourneau J.-M (2003) Iterative disaggregation for a class of lumpable discrete-time stochastic automata networks. *Perform. Eval.*, 53(1):43–69.
8. Hashida O, Takahashi Y, Shimogawa S (1991) Switched batch bernoulli process (SBBP) and the discrete-time sbbp/g/1 queue with application to statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 9(3):394–401.
9. Hernández-Orallo E, Vila-Carbó J (2007) Network performance analysis based on histogram workload models. In *MASCOTS*, pp 209–216.
10. Hernández-Orallo E, Vila-Carbó J (2009) Web server performance analysis using histogram workload models. *Computer Networks*, 53(15):2727–2739.
11. Hernández-Orallo E, Vila-Carbó J (2010) Network queue and loss analysis using histogram-based traffic models. *Computer Communications*, 33(2):190–201.
12. Horváth G, Telek M, Buchholz P (2005) A map fitting approach with independent approximation of the inter-arrival time distribution and the lag correlation. In *QEST*, pages 124–133. IEEE Computer Society.
13. Klemm A, Lindemann C, Lohmann, M (2003) Traffic modeling of IP networks using the batch markovian arrival process. *Perform. Eval.*, 54(25):149–173.
14. Muller A, Stoyan D (2002) Comparison Methods for Stochastic Models and Risks. Wiley, New York, NY.
15. Muscarielloa L, Melliaa M., Meoa M., Ajmone Marsan M., Lo Cignob R, 2005 Markov models of internet traffic and a new hierarchical MMPP model, *Computer Communications*, 28: 1835-1851
16. Skelly P, Schwartz M, Dixit S. S. (1993) A histogram-based model for video traffic behavior in an atm multiplexer. *IEEE/ACM Trans. Networking.*, 1(4):446–459.
17. Sony K. C, Cho K (2000) Traffic data repository at the wide project. In *In Proceedings of USENIX 2000 Annual Technical Conference: FREENIX Track*, pages 263–270.
18. Stathis C, Maglaris B. S (2000) Modeling the self-similar behavior of network traffic. *Computer Networks*, 34(1):37–47.
19. Stewart W (1995) Introduction to the numerical Solution of Markov Chains. *Princeton University Press, New Jersey.*
20. Wittevrongel S, Bruneel H, 1999 Discrete-time queues with correlated arrivals and constant service times *Computers and Operations Research*, 26 : 93-108.
21. Zhou W , Wang A, (2013) Discrete-time queue with Bernoulli bursty source arrival and generally distributed service times *Applied Mathematical Modelling* , 3,2223-2240.