Timescales in Models for Bursty Traffic

Kathleen Spaey\textsuperscript{1}, Danny De Vleeschauwer\textsuperscript{2}, Tom Hofkens\textsuperscript{1}, Chris Blondia\textsuperscript{1}

\textsuperscript{1}University of Antwerp, Department of Mathematics and Computer Science
Performance Analysis of Telecommunication Systems Research Group
Middelheimlaan 1, B-2020 Antwerpen, Belgium
\{kathleen.spaey,tom.hofkens,chris.blondia\}@ua.ac.be

\textsuperscript{2}Alcatel Bell, Network Strategy Group
Francis Wellesplein 1, B-2018 Antwerpen, Belgium
danny.de_vleeschauwer@alcatel.be

Keywords: Timescales, analytic modeling, queueing results, unfinished work, video streaming application.

Abstract
This paper investigates what the impact is of approximating packetized flows by fluid flows on the complementary cumulative distributions (CCDs) of the unfinished work. For that purpose two queueing models are compared. In the first model (referred to as the packet model) two timescales are present in the input traffic, i.e., a burst and a packet timescale. In the second one (referred to as the fluid model) only one (identical) burst timescale is present. The comparison shows that the slope of the CCDs obtained with both models is the same, i.e., that fluctuations in the arrival pattern at the finest timescale do not influence this slope, but that the CCD of the packet model intersects the ordinate at a higher value than the one of the fluid model. First, the influence of all relevant parameters on the difference between the two CCDs is explored. Because these CCDs of unfinished work are usually utilized to dimension the buffer, it follows that the fluid model underestimates the value of the required buffer, which can amount to a considerable value for the cases where the slope turns out to be small. Then, this extra amount is calculated in a case study dealing with a video streaming application.

INTRODUCTION
Following the standardization of the MPEG algorithm and its wide acceptance, researchers started studying the statistical characteristics of MPEG coded video traffic and developing statistical source models specific to this type of traffic. The result is a wide variety of models (see for example [1] and the references therein). An important characteristic of MPEG video traffic is that it exhibits multiple timescales (scenes, groups of pictures, frames, sub-frame levels, packets). Some studies claim that it is important to develop source models at the MPEG sub-frame level [2], because fine timescale fluctuations can have substantial queueing performance implications. Many other models neglect the sub-frame [3, 4] or even the frame level [5, 6], because these models are based on publicly available traces [7, 8], which are all frame size traces, and because fine timescales are less robust to buffering or shaping.

In this paper a simple bursty traffic model is defined in which two timescales are present. The goal of the model is to analyze how important fine timescale fluctuations in the traffic profile are. This is done by comparing queueing results obtained with this model to results obtained with a model where the fine time scale fluctuations are smoothed out over the larger timescale using a fluid-flow process. The performance measure of interest is the distribution of the amount of unfinished work in the queueing system. Although the two timescales in the model can represent any two timescales, the largest timescale will be referred to as the frame timescale, while the smallest one will be called the packet timescale. Remark that the considered traffic model does not necessarily model a video source, but is intended to investigate how large the influence of neglecting the fine timescale fluctuations is.

This paper is organized as follows: first a description of the traffic and queueing models is given. Next, the computation of the unfinished work distributions is discussed. Then numerical results are considered. By means of these results, the influence of the different parameters on the unfinished work distributions and on the differences between these distributions for both models, is investigated. Then the theory developed in the paper is applied to a case study of a video streaming service. Conclusions are drawn in the last section.

MODEL DESCRIPTION
Model for Bursty Traffic

Suppose that time is divided in units of constant length, called frame times, and that every frame time is further divided into \( x \) constant length time units, called packet times, where \( x \in \mathbb{N} \). As a typical model for a bursty traffic source, consider a two-state discrete-time Markov source. Assume that the Markov source has a frame time as underlying time unit, i.e., the source can only change state on frame time.
borders. Let $\alpha$, respectively $\beta$, be the probability that when
the source is in the first, respectively second, state, at the next
frame time border it makes a transition to the second,
respectively first, state. During a frame time in which the
source is in the first, respectively second state, it generates
$L_i$ bytes, respectively $L_2L_i$ bytes, with $L_2 < L_1 < x$. In a first
source model, which will be referred to as the packet source, these
$L_i$ bytes are divided into $L_i$ packets with a fixed length of $L$
bytes, and it is assumed that the source sends these packets
during the first $L_i$ packet times of the $x$ packet times in the
frame time (see also Figure 1). In a second source model,
the $L_i$ bytes are generated as ‘fluid’ at a constant rate of $L_i$
bytes per frame time, or thus $L_i$/x bytes per packet time.
This second source model is referred to further on as the fluid
source. Remark that in both cases, a source generates on
average $\lambda = (L_1\beta + L_2\alpha)L/\alpha + \beta$ bytes per frame time.

**Queueing Model**

Consider a queueing system with an infinite buffer, in
which the traffic of either $M$ identical packet sources or
$M$ identical fluid sources with the same parameters is mul-
tiplexed. It is assumed that the frame time borders of all
the sources are synchronized. So for the packet sources, if
$i$ $(0 \leq i \leq M)$ of the sources are in the first state during a
frame time (and thus $M - i$ sources are in the second state),
the packet arrival profile at the queue can be described as follows:

- $M$ arrivals during each of the first $L_1$ packet times of the
  frame time,
- $i$ arrivals during each of the next $L_2 - L_1$ packet times,
  and
- no arrivals during the remaining $x - L_1$ packet times.

For the fluid sources, in case that $i$ sources are in the first
state during a frame time, fluid arrives in that frame time at a
constant rate of $(iL_1 + (M - i)L_2)\lambda$ bytes per frame time. When
the input consists of packet sources, one packet can leave the
system (if the system is not empty) every packet time,
resulting in a service rate of $x\lambda$ bytes per frame time. When
the input sources are fluid sources, traffic (if present) leaves
the system at a constant rate of $x\lambda$ bytes per frame time.
Only scenarios with $M(L_1\beta + L_2\alpha)/(\alpha + \beta) < x < M$ will
be considered, since otherwise either the load of the system
is larger than one, leading to instability, or the queueing system
is always empty at frame time boundaries.

**MATHEMATICAL ANALYSIS**

The performance measure of interest is the distribution of
the amount of unfinished work in both queueing systems. To
calculate these distributions, first the stationary joint proba-
bility distribution of the amount of work in the system and
the number of sources that are in the first state at frame
time boundaries is calculated. Let $q_k$ be the unfinished work
(expressed in packets for the system with packet sources as input,
and expressed in whole number multiples of $L$ bytes
for the system with fluid sources as input) and let $f_k$ be the
number of sources that are in the first state at the beginning
of the $k$-th frame time. Then \{$(q_k, f_k)$, $k \geq 0$\} is a stationary
Markov chain with transition probability matrix

\[
Q = \begin{pmatrix}
S_0 + \cdots + S_n & S_{n+1} & S_{n+2} & \cdots \\
S_0 + \cdots + S_{n-1} & S_n & \cdots & \\
S_0 & \cdots & S_n & \\
0 & S_0 & S_1 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix},
\]

where the matrices $S_k$ govern transitions that correspond to
the arrival of $kL$ bytes of traffic in a frame time. Then

\[
(q_k)_{ij} = \begin{cases}
S_{ij}, & \text{if } k = iL_1 + (M - i)L_2, \\
0, & \text{otherwise},
\end{cases}
\]

and $S_{ij}$ is the $(i,j)$-th element of a matrix $S$ describing the
probability that when $i$ sources are in the first state at a frame
time border, at the next frame time border $j$ sources will be
in the first state:

\[
S_{ij} = \frac{\min\{i,j\}}{\max\{0,i+j-M\}} \frac{i}{j} (1 - \alpha)^{i-1} \beta^{j-i} \left( M - i \right)^{j-i} (1 - \beta)^{M-i-j+i}.
\]

The stationary probability vector of $Q$ will be denoted by
$q = (q_0, q_1, q_2, \ldots)$, where $q_k$ is a row vector of length $M + 1$, whose $i$-th element $(q_k)_i$ is the stationary joint probability
that at frame time boundaries there are $i$ sources in the first state
and the unfinished work equals $n$ packets (or $nL$ bytes).
The vector $q$ is calculated using the algorithms described in
[9] and [10]. Remark that $q$ is a vector of infinite length.
Components $q_n$ of $q$ are calculated until the probability mass
in the remaining components is negligible. In the following, a
frame time is said to be of type $(n,i)$, when at the beginning
of the frame time $n$ packets (or $nL$ bytes) are present in
the system, and $i$ of the $M$ sources are in the first state.

**Unfinished Work Distribution: System with Fluid Sources as Input**

Let the random variable $W_f$ represent the amount of unfin-
ished work in the system with fluid sources as input at an
arbitrary time instant, expressed in multiples of $L$ bytes. Re-
mark that $W_f$ is a continuous random variable. Denote by
$A(n,i)$ the event that an arbitrary time instant falls in a frame
time of type $(n,i)$. The cumulative distribution function of
$W_f$ is given by

\[
F_{W_f}(w) = P\{W_f \leq w\} = \sum_{n=0}^{\infty} \sum_{i=0}^{M} P\{W_f \leq w | A(n,i)\} P\{A(n,i)\},
\]

where $P\{A(n,i)\} = P\{\lim_{k \to \infty} (q_k, f_k) = (n,i)\} = (q_n)$,
since the probability that an arbitrary time instant falls in
a frame time of a certain type equals the relative occurrence of such frame types.

Consider now a frame time of type \((n, i)\) and choose a random instant \(t\) in the frame time. Denote by \(T_J\) the length of the frame time. Then, \(T\), the time between \(t\) and the beginning of the frame time, is uniformly distributed over \([0, T_J]\), and the relation between \(W_J\) and \(T\) is given by:

\[
W_J = \left( n + \frac{d_1 + (M - i)L_2 - x}{T_J} \right)^+,
\]

where \((a)^+ = \max\{a, 0\}\). So \(P\{W_J \leq w | A(n, i)\}\), denoted below by \(P\), depends on the relation between \(x\) and \(\chi := d_1 + (M - i)L_2\):

- \(\chi > x\), i.e., during the frame time, more work arrives in the system than can be left. Then

\[
P = \begin{cases} 
0, & \text{if } w < n, \\
\frac{w - n}{\chi - x}, & \text{if } n \leq w \leq n + \chi - x, \\
1, & \text{if } w > n + \chi - x.
\end{cases}
\]

- \(\chi < x\), i.e., during the frame time, less work arrives in the system than can be left. Consider the following cases:

  - if \(n + \chi - x > 0\), i.e., during the frame time the system does not become empty, then

\[
P = \begin{cases} 
0, & \text{if } w < n + \chi - x, \\
1 - \frac{w - n}{\chi - x}, & \text{if } n + \chi - x \leq w \leq n, \\
1, & \text{if } w > n.
\end{cases}
\]

  - if \(n + \chi - x \leq 0\), i.e., during the frame time the system becomes empty, namely from \(T = -nT_J / (\chi - x)\) on, then

\[
P = \begin{cases} 
0, & \text{if } w < 0, \\
1 - \frac{w - n}{\chi - x}, & \text{if } 0 \leq w \leq n, \\
1, & \text{if } w > n.
\end{cases}
\]

- \(\chi = x\), i.e., during the frame time, the amount of work that arrives in the system equals the amount that can leave it. Then

\[
P = \begin{cases} 
0, & \text{if } w < n, \\
1, & \text{if } w \geq n.
\end{cases}
\]

**Unfinished Work Distribution: System with Packet Sources as Input**

Let the random variable \(W_p\) represent the amount of unfinished work in the system with packet sources as input at an arbitrary packet time boundary, expressed in number of packets. Since at packet time boundaries there are always an integer number of packets present in the system, \(W_p\) is a discrete random variable. Denote by \(B(n, i)\) the event that an arbitrary packet time boundary falls in a frame time of type \((n, i)\). The cumulative distribution function of \(W_p\) is given by

\[
F_{W_p}(w) = P\{W_p \leq w | B(n, i)\} = \sum_{n=0}^{\infty} \sum_{i=0}^{M} P\{W_p \leq w | B(n, i)\} P\{B(n, i)\},
\]

where \(P\{B(n, i)\} = P\{\lim_{k \to \infty} (q_k, f_k) = (n, i)\} = (q_k)\). Consider now an arbitrary packet time boundary in a frame time of type \((n, i)\). Let \(B(m)\) represent the event that the arbitrary packet time boundary corresponds to the \(m\)-th of the \(x\) packet times in a frame time. Then

\[
P\{W_p \leq w | B(n, i)\} = \sum_{m=1}^{x} P\{W_p \leq w | B(n, i), B(m)\}
\]

with \(P\{B(m) | B(n, i)\} = 1/x\), since the packet time to which an arbitrary packet time boundary corresponds is independent of the queue length or state of the sources at the beginning of the frame time. Then is obtained as follows: if

\[
P^* = \begin{cases} 
1, & \text{if } w \geq n + m(M - 1), \\
0, & \text{otherwise},
\end{cases}
\]

\[
m \in \{1, 2, \ldots, l_2\}, \text{i.e., during the packet time considered, } M \text{ packets arrive and one packet leaves the system,}
\]

\[
P^* = \begin{cases} 
1, & \text{if } w \geq (n + l_2 M + (m - l_2) i - m)^+, \\
0, & \text{otherwise},
\end{cases}
\]

\[
m \in \{l_2 + 1, \ldots, l_1\}, \text{i.e., during the packet time considered, } i \text{ packets arrive and one packet leaves the system,}
\]

\[
P^* = \begin{cases} 
1, & \text{if } w \geq (n + l_2 M + (l_1 - l_2) i - m)^+, \\
0, & \text{otherwise},
\end{cases}
\]

**NUMERICAL RESULTS**

**Reference Example**

All results presented in this section will be compared to results obtained with a reference example. In this example the traffic is generated by \(M = 10\) sources with parameters \(\alpha = 0.4\), \(\beta = 0.1\), \(l_1 = 40\) and \(l_2 = 4\). The number of packet times in a frame time is chosen equal to \(x = 160\). Consequently, the load of the system equals 70%, a source is in the first state with probability 0.2, and thus in the second state with probability 0.8, and stays on average in the first state,
Table 1. Parameters and characteristics of the different scenarios considered.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Parameters of the scenario</th>
<th>Characteristics of the scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference example</td>
<td>M</td>
<td>α</td>
</tr>
<tr>
<td>Example 1</td>
<td>10</td>
<td>0.4</td>
</tr>
<tr>
<td>Example 2</td>
<td>10</td>
<td>0.2</td>
</tr>
<tr>
<td>Example 3</td>
<td>20</td>
<td>0.4</td>
</tr>
<tr>
<td>Example 4</td>
<td>5</td>
<td>0.4</td>
</tr>
<tr>
<td>Example 5</td>
<td>10</td>
<td>0.8</td>
</tr>
<tr>
<td>Example 6</td>
<td>10</td>
<td>0.2</td>
</tr>
<tr>
<td>Example 7</td>
<td>10</td>
<td>0.4</td>
</tr>
<tr>
<td>Example 8</td>
<td>10</td>
<td>0.4</td>
</tr>
<tr>
<td>Example 9</td>
<td>6</td>
<td>0.4</td>
</tr>
<tr>
<td>Example 10</td>
<td>8</td>
<td>0.4</td>
</tr>
<tr>
<td>Example 11</td>
<td>10</td>
<td>0.4</td>
</tr>
<tr>
<td>Example 12</td>
<td>10</td>
<td>0.4</td>
</tr>
</tbody>
</table>

(1): load
(2): probability that a source is in state 1, state 2
(3): average time (in packet times) a source is in state 1, state 2
(4): split-up of the load a source generates when it is in state 1, state 2

Figure 2. Results obtained with the reference example.

respectively the second state, during respectively 2.5 and 10 frame times. Figure 2 shows for the example the complementary cumulative distributions of $X$ ($P\{X > k\}$), where $X$ represents respectively the unfinished work for the fluid sources ($W_f$) and the unfinished work for the packet sources ($W_p$). The corresponding units in which $k$ is expressed are respectively multiples of $L$ bytes and packets. Complementary cumulative distributions are commonly used to determine the buffer size that would be sufficient to guarantee a loss or a delay quantile of at most a certain value. As can be seen from the figure, the two curves decrease almost linearly with increasing $k$ (which means that $P\{X > k\}$ decays exponentially with increasing $k$), except in the region of lesser importance where $k$ is small. Moreover, the slope of both curves is the same (~0.0049). The curve for the packet sources lies above that for the fluid sources. These trends are general, and will be observed in all following examples. The reason that the curve for the packet sources lies above that for the fluid sources is that although the traffic profile of both kind of sources is the same over the frame times, within each frame time the traffic of the fluid sources arrives in a more smooth way than that of the packet sources. The same reason explains why the probability of having an empty system (given by $1 - P\{X > 0\}$) is larger for the fluid than for the packet sources: assume that at the beginning of a frame time the buffer is empty, and there are $i$ sources in the first state, where $i$ is such that during the frame time less work will arrive than can be served. In the case of the fluid sources, the buffer will stay empty for the whole duration of the frame time, while in the case of the packet sources the buffer content first grows, then decreases, and the buffer only becomes empty again after $li_1 + (M - i)l_2$ packet times. So it is only empty for a fraction $(x - li_1 - (M - i)l_2)/x$ of the frame time.

For a fixed probability $p = P(W_f > w_f) = P(W_p > w_p)$ in the region where the curves decrease linearly (say $p < 10^{-7}$), the absolute difference between $w_p$ and $w_f$ is

\[w_p - w_f < 2 \times 10^{-7}
\]

Figure 3. Influence of the length of a frame time.
Table 2. Influence of the length of a frame time (examples 1-2), the number of sources (examples 3-4) and the state sojourn time (examples 5-6).

<table>
<thead>
<tr>
<th>Reference example: $M = 10, \alpha = 0.4, \beta = 0.1, l_1 = 40, l_2 = 4, x = 160$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference example</td>
</tr>
<tr>
<td>Slope curves</td>
</tr>
<tr>
<td>Absolute diff</td>
</tr>
<tr>
<td>Relative diff</td>
</tr>
<tr>
<td>$p = 10^{-2}$</td>
</tr>
<tr>
<td>$p = 10^{-3}$</td>
</tr>
<tr>
<td>$p = 10^{-4}$</td>
</tr>
<tr>
<td>$p = 10^{-5}$</td>
</tr>
<tr>
<td>$p = 10^{-6}$</td>
</tr>
<tr>
<td>$p = 10^{-7}$</td>
</tr>
<tr>
<td>$p = 10^{-8}$</td>
</tr>
<tr>
<td>$p = 10^{-9}$</td>
</tr>
<tr>
<td>$p = 10^{-10}$</td>
</tr>
</tbody>
</table>

approximately 71. Because this absolute difference stays constant with decreasing $p$, the relative differences between $w_p$ and $w_f$ decrease with decreasing $p$ (see Table 2). To calculate the relative differences, the value of $w_f$ is always used as reference value (i.e., the relative differences are calculated as $(w_p - w_f)/w_f$), because it is expected that one might prefer a model that neglects the smallest timescale (and thus is simpler), and use the results obtained with this model to make predictions about the results that would be obtained when the smallest timescale is not neglected.

As illustrated with this reference example, using results obtained with the fluid source model that ignores the fine timescale fluctuations in the traffic arrival process, may lead to an underestimation of the required buffer space. The remaining part of this section investigates the influence of the different parameters of the model on the unfinished work distributions on one hand, and on their influence on this underestimation on the other hand. Results are always compared to results obtained with the reference example. An overview of the parameters and characteristics of the different examples considered is given in Table 1.

Influence of the Length of a Frame Time

Consider the reference example and two other examples that are obtained by doubling (Example 1), respectively halving (Example 2), the parameters $\alpha, \beta, l_1, l_2$ and $x$. By choosing the parameters in this way, the three examples have many common properties (see Table 1). Notwithstanding these common properties, it is seen from Figure 3 that the smaller $x$, the larger the slope of the curves (i.e., the more to the right the tail of the distributions lie). This is because although the sources in the three examples stay on average in the first state for the same amount of time, and also the probability that there are $i$ sources in the first state is the same $(\binom{M}{i}\binom{M-1}{i}/\binom{M}{i})$, the standard deviation of the amount of time a source stays in the first state is different, namely $\sqrt{\frac{i(M-1)}{x}}$ packet times. Because $x/\alpha$ is constant, a smaller $x$ corresponds to a larger standard deviation. So the smaller $x$, the higher the probability that a source stays for a long time in the first state, the state at which it generates the highest load, and thus the higher the probability that for a long time the sources generate traffic at a higher rate than the service rate, such that the buffer content grows.

Further it is observed from Figure 3 and Table 2 that the more packet times in a frame time, the larger the absolute and relative difference between the curves of $W_p$ and $W_f$. The reason is the following: when looking over a period in time corresponding to the largest frame time (320 packet times), then for $x = 320$ the packets arrive concentrated at the first 80 packet times of this period. In the case that $x = 160$, they arrive concentrated at packet times 1-40 and 161-200 of this period, while for $x = 80$ this is at packet times 1-20, 81-100, 161-180 and 241-260. So the larger $x$, the more the arrival pattern of the packets deviates from how the fluid traffic arrives.

Influence of the Number of Sources

Consider the reference example and two other examples in which the same load is generated by twice as much (Example 3) and half as much (Example 4) sources. In all examples, the parameters $x, \alpha$ and $\beta$ are chosen the same, such that the sources in all examples have the same probability of being in one of the two states, with the same mean sojourn time. The parameters $l_1$ and $l_2$ are chosen such that in all examples each state of a source contributes the same fraction to the to-
tal load the source generates. Remark however that since the same total load is generated by a different number of sources in each example, the load one source should generate is different in each example, resulting in different values for $I_1$ and $I_2$ in the different examples. Figure 4 and Table 2 show the results obtained with these examples.

From Figure 4 the well-known property is observed that increasing the number of sources while keeping the load fixed, results in a smaller slope of the curves, or thus a smaller probability corresponding to larger buffer sizes. This is because the more sources, the larger the statistical multiplexing gain. So the more sources, the smaller the probability that a long queue will build up.

When considering the absolute and relative differences between the curves for $W_p$ and $W_f$, it is seen from Figure 4 and Table 2 that the more sources, the larger the absolute and relative differences. Indeed, the packet arrivals are spread out more in the time when there are less sources, and thus the larger the number of sources, the more the traffic pattern within a frame time deviates from the fluid traffic pattern.

**Influence of the State Sojourn Time**

In this section, the reference example is compared to scenarios that differ in the average number of subsequent frame times the sources are in the first, respectively the second state. It are the parameters $\alpha$ and $\beta$ that influence the average state sojourn times. In the reference example ($\alpha = 0.4, \beta = 0.1$), a source stays on average 2.5 frame times in the first state, and 10 frame times in the second state. In Examples 5 and 6, the values of the parameters $\alpha$ and $\beta$ are halved, respectively doubled, resulting in average state sojourn times that are half as long (Example 5), respectively double as long (Example 6). Figure 5 and Table 2 show the results obtained with these examples.

First of all, it is seen in Figure 5 that the larger the mean sojourn times of the sources in a scenario, the larger the slope of the curves is. This is again a well-known result, which is easy to explain: the longer in time sources stay in the first state at which they generate traffic at their peak rate, the longer the buffer grows. To keep the load a source generates the same in all examples, long periods in the first state are compensated by long periods in the second state.

Table 2 and Figure 5 also indicate that with increasing mean sojourn times, the absolute and relative difference between the curves $W_p$ and $W_f$ of the unfinished work decrease.

**Influence of the Peak Rate of the Sources**

In this section the reference example is considered together with two scenarios in which the values of the parameters $I_1$ and $I_2$ are changed. As a consequence, the split up of the load one source generates (7%) between the load a source generates while it is in the first state and while it is in the second state changes. The higher $I_1$, the higher the peak rate of a source, and the higher the load it generates when it is in the first state. Where in the reference example the split up is $5\% + 2\%$, it is chosen $4\% + 3\%$ in Example 7, and $6\% + 1\%$ in Example 8. Results are shown in Figure 6 and Table 3.

Figure 6 shows that the higher the peak rate a source generates, the larger the slope of the curves, which is as expected since the higher the peak rate, the more bursty the traffic a source generates. Table 3 and Figure 6 further show that with increasing peak rate of the sources, the absolute and relative differences between the curves of $W_p$ and $W_f$ decrease. The reason for this is that within a frame time, packets arrive over the first $I_1$ packet times of the frame time. The larger $I_1$, the more the arrival pattern of the packets within the frame time resembles that of the fluid sources.

**Influence of the Load**

In all previous sections, the load was kept constant in all scenarios. In this section, the reference example will be compared to scenarios with a different load. The load is changed in two different ways: on one hand by considering a different number of the same sources as in the reference example ($M = 6$ in Example 9, $M = 8$ in Example 10 and $M = 10$ in the reference example); on the other hand by keeping the
Table 3. Influence of the peak rate of the sources (examples 7-8), the load when the number of sources is changed (examples 9-10), and the load when the rate the sources generate is changed (examples 11-12).

<table>
<thead>
<tr>
<th>Slope curves</th>
<th>Reference example</th>
<th>Example 7</th>
<th>Example 8</th>
<th>Example 9</th>
<th>Example 10</th>
<th>Example 11</th>
<th>Example 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>w_p, w_f</td>
<td>w_p, w_f</td>
<td>w_p, w_f</td>
<td>w_p, w_f</td>
<td>w_p, w_f</td>
<td>w_p, w_f</td>
<td>w_p, w_f</td>
<td>w_p, w_f</td>
</tr>
<tr>
<td>Slope curves</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-2}</td>
<td>3.86</td>
<td>97.65</td>
<td>15.04</td>
<td>1306.67</td>
<td>119.40</td>
<td>593.75</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-3}</td>
<td>16.63</td>
<td>43.62</td>
<td>8.56</td>
<td>180.85</td>
<td>43.18</td>
<td>126.03</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-4}</td>
<td>11.09</td>
<td>27.09</td>
<td>5.97</td>
<td>85.29</td>
<td>25.51</td>
<td>59.59</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-5}</td>
<td>8.50</td>
<td>19.66</td>
<td>4.51</td>
<td>48.52</td>
<td>18.03</td>
<td>39.82</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-6}</td>
<td>6.83</td>
<td>15.43</td>
<td>3.72</td>
<td>35.90</td>
<td>13.91</td>
<td>29.33</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-7}</td>
<td>5.70</td>
<td>12.85</td>
<td>3.13</td>
<td>27.67</td>
<td>11.16</td>
<td>22.83</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-8}</td>
<td>4.35</td>
<td>10.77</td>
<td>2.70</td>
<td>22.28</td>
<td>9.54</td>
<td>18.82</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-9}</td>
<td>3.76</td>
<td>9.32</td>
<td>2.38</td>
<td>19.12</td>
<td>8.24</td>
<td>16.02</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-10}</td>
<td>3.39</td>
<td>8.38</td>
<td>2.12</td>
<td>16.37</td>
<td>7.25</td>
<td>13.76</td>
</tr>
<tr>
<td>Reference example</td>
<td>p = 10^{-11}</td>
<td>3.00</td>
<td>7.41</td>
<td>1.91</td>
<td>14.44</td>
<td>6.48</td>
<td>12.32</td>
</tr>
</tbody>
</table>

Figure 6. Influence of the peak rate of the sources.

Figure 7. Influence of the load when the number of sources is changed.

number of sources the same, but by decreasing (Example 11), respectively increasing (Example 12) the rate at which the sources generate traffic, by changing the parameters $l_1$ and $l_2$ of the sources. This is done in such a way that the proportion of the load a source generates when it is in the first state compared to when it is in the second state, stays the same. Also the parameters $\alpha$, $\beta$ and $x$ are not changed. Figures 7 and 8 and Table 3 show the results.

Figures 7 and 8 illustrate the well-known property that an increased load results in curves with a larger slope. About the absolute and relative differences between the curves for $W_p$ and $W_f$, it is noticed from Table 3 that they decrease with increasing load. The reason why the absolute difference between $W_p$ and $W_f$ decreases with increasing load when the rate at which the sources generate traffic is increased, is that the packet arrivals are more spread out over a frame time then, such that the arrival pattern of the packets resembles more how the fluid traffic arrives.

APPLICATION TO VIDEO STREAMING SERVICE

In this section the theory developed before is applied to the following case study. Consider $M$ users who get access to a video streaming service over an access network where the (sole) bottleneck is a link of capacity $C = 100$ Mb/s. All video footage that these $M$ users have access to, is stored on a local server. As it is an ‘on demand’ service, and hence, it is not known a priori when a user wants access to a particular
piece of a video, no multicasting or broadcasting technique can be used. There is one video flow per user. The bit rate associated with a flow will evolve over time, because users may tap from one video to another.

Although each video itself is encoded at Constant Bit Rate (CBR), not all videos are encoded at the same bit rate. There are two types of footage: high-motion video with highly textured areas (typically sport footage) and videos with a lot of still scenes which lack details (typically soaps, cartoons, …). The former type requires a higher bit rate than the latter one. Video bit rates $R_1$ and $R_2$ equal to 3 Mb/s and 2 Mb/s respectively, are considered in this case study.

The bit rate produced depends on the state the source (i.e., the user) is in, and state transitions can only occur at frame time boundaries. A ‘natural’ choice for the frame time $T_f$ (expressed in seconds) is the duration of a Group of Picture (GoP). With a GoP of 12 pictures, each of duration 40 ms, $T_f = 480$ ms.

A user watches half of the time content of the first type and half of the time content of the second type. The average time $T_v$ a user watches content of one particular type, which is in this case the same for both types of video, determines the transition probabilities $\alpha = \beta = \frac{T_f}{2}$. Two values for $T_v$ are considered: 3 minutes and 30 minutes.

A packet size $L$ of 1500 bytes is chosen. The packet time $T_p$ is then given by $\frac{L}{R} = 0.12$ ms. The number of packet times in a frame time equals $x = \frac{T_f}{T_p} = 4000$. The number of packets a flow of type $i$ ($i \in \{1, 2\}$) generates during a frame time $T_f$ is given by:

$$l_i = \frac{R_i T_f}{L} = \frac{R_i}{C} = \frac{R_i}{25 \text{ kb/s}},$$

which implies that $l_1 = 120$ and $l_2 = 80$.

\footnote{A picture is sometimes also referred to as a ‘video frame’. This terminology is avoided as in this paper a ‘frame time’ corresponds to the time between state transitions of a source, i.e., a GoP in this section.}

### Table 4. Buffer space required for the video streaming service, corresponding to an overflow probability of $10^{-5}$. The buffer space is expressed in multiples of $L$ bytes for the system with fluid sources as input (FS) and in packets for the system with packet sources as input (PS).

<table>
<thead>
<tr>
<th>$M$</th>
<th>$T_v = 30$ min</th>
<th>$T_v = 3$ min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FS</td>
<td>PS</td>
</tr>
<tr>
<td>34</td>
<td>0</td>
<td>3689</td>
</tr>
<tr>
<td>35</td>
<td>11719</td>
<td>13746</td>
</tr>
<tr>
<td>36</td>
<td>15419</td>
<td>156078</td>
</tr>
<tr>
<td>37</td>
<td>$&gt; 10^5$</td>
<td>$&gt; 10^5$</td>
</tr>
<tr>
<td>38</td>
<td>$&gt; 10^5$</td>
<td>$&gt; 10^5$</td>
</tr>
<tr>
<td>39</td>
<td>$&gt; 10^5$</td>
<td>$&gt; 10^5$</td>
</tr>
</tbody>
</table>

The interesting cases are the ones where the maximum of the offered traffic per frame time exceeds the capacity of the link (otherwise the system is always empty at frame time boundaries) and where the average of the offered traffic does not exceed the link capacity (otherwise the system is not stable), i.e.,

$$MR_1 > C > M \frac{\alpha R_1 + \beta R_2}{\alpha + \beta},$$

which leads to the following bounds on the number of users: $33 < M < 40$.

In principle it is not known at which time instants packets arrive within a frame time. Before, two systems were considered. In the first system, referred to as ‘the system with fluid sources as input’ (FS), the information generated by a source in a frame time arrives as a constant fluid flow within the frame time. In the second system, referred to as ‘the system with packet sources as input’ (PS), this information arrives as fast as possible in packets (i.e., with an interpacket time of $T_p$), which is some kind of worst case assumption.

For both models, the buffer occupancy level that is only exceeded with a probability of $10^{-5}$ is calculated. This occupancy level can be roughly interpreted as the buffer size that is required for this video service, i.e., the video service will be of bad quality (caused by buffer overflow) for a total of a mere 2.6 seconds a month if a buffer of this size is implemented.

Notice that a reasonable maximum value for the required buffer size is of the order of $10^4$ packets. This size corresponds to a maximum delay of $10^4 T_p = 1.2$ s, which is just about acceptable. A buffer of an order of magnitude larger than this would introduce too much jitter.

Table 4 gives the buffer space (expressed in multiples of $L$ bytes for the FS system, and expressed in packets for the PS system) required for the video streaming service for both considered values of $T_v$. To express it in bytes, the values in Table 4 have to be multiplied by 1500 bytes. It can be seen that the FS model underestimates the required buffer space. For low loads, i.e., $M = 34$, since the FS model does not capture the competition on the packet level, it predicts a buffer size of 0, where in fact it should be a few 1000 of packets. Notice that in this case the parameter $T_v$ does not play a dominant role. For loads corresponding to $M = 35$ for the case $T_v = 30$ min and corresponding to $M = 35$ and...
$M = 36$ for the case $T_v = 3$ min, respectively, the required buffer size is still acceptable. In these cases, the FS model underestimates the buffer size by a few 1000 packets. Higher loads than this would require a too large buffer space (that would introduce too much jitter). So, in the cases $T_v = 30$ min and $T_v = 3$ min, the number of sources needs to be restricted to 35 and 36, respectively.

**CONCLUSIONS**

In this paper a simple bursty traffic model was defined in which two timescales are present. This model is used to analyze how large the importance of fluctuations in the arrival pattern at the smallest timescale (called packet timescale) is. This is done by comparing queuing results obtained with this model to results obtained with a model which is the same at the larger timescale (called frame timescale), but which neglects the fluctuations in traffic pattern at the packet timescale by smoothing out the traffic over a frame time using a fluid-flow process.

It is observed that for the packet sources the complementary cumulative distribution of the unfinished work lies above that for the fluid sources. Both curves have the same slope (so the arrival process at the smallest timescale does not influence this slope), except in the region of lesser importance where the amount of unfinished work is small. As a consequence, in the region where the curves decrease linearly, the absolute distances between the two curves for a fixed probability $p$ stay constant independent of that probability, and the relative distances decrease with decreasing $p$.

The probability that the amount of unfinished work is larger than a certain value is the smallest for the fluid model that neglects traffic fluctuations at the finest timescale. Therefore, using the complementary cumulative distribution curve for the fluid sources to determine which buffer size would be sufficient to guarantee a packet loss or a delay quantile of at most a certain value when considering packet sources, may lead to too optimistic results. From the numerical results presented in this paper it is however learned that the absolute difference between the unfinished work curve for the fluid and packet sources becomes less important when the slope of the curves becomes larger, i.e., when the probability of having a large amount of unfinished work in the system becomes larger. There are several properties of traffic scenarios that result in a larger slope of the curves. The following properties were considered in this paper:

- a decrease of the duration of the largest time scale,
- less sources generate the same traffic load,
- an increase of the average time the sources stay in their different states,
- an increase of the peak rate the sources generate,
- an increase of the load by increasing the number of sources or the rate at which the sources generate traffic.

So this study shows that approximating packetized flows by fluid flows that ignore the fine timescale fluctuations in the traffic arrival pattern, may lead to an underestimation of the required buffer space to guarantee a packet loss or delay quantile of at most a certain value. This is further illustrated by applying the developed models to a case study of a video streaming service. For this service, buffer size requirements calculated using both the packet and the fluid source model, where a fluid source smooths out the traffic over the duration of a GoP, are compared.

**REFERENCES**


