Stochastic modelling and control of communication networks

Zuraniewski, P.W.

Citation for published version (APA):
Chapter 6

Resource allocation issues of a two-level network resource

In this chapter we consider a network element, shared by two priority classes; while our model directly relates to various existing technologies, a timely application is Carrier Ethernet. In our model, two priority classes share the resource. The high-priority queue — intended for traffic generated by delay-sensitive applications — typically being small, the low-priority queue can be modelled as a queue with a time-varying service-rate; this service rate behaves independently of the input of the low-priority queue. The high-priority queue is standard, and we therefore provide an in-depth analysis of the performance of the low-priority queue, assuming that all traffic offered is subjected to a leaky-bucket type of policer. One of the approaches that we analyze borrows elements from the setup of [78], in which the notion of effective bandwidth plays a crucial role. It is concluded that this procedure has computationally attractive properties, but tends to be pessimistic. Through an alternative method we find, for a given set of users with a given traffic profile, more accurate estimates of the packet loss probability; these insights can be used when setting up resource allocation algorithms. The use of this type of approximation is supported by an extensive set of numerical experiments.

1 Introduction

In modern networks (with Carrier Ethernet as a leading example), in order to make efficient use of the network resources, traffic is split into multiple categories. In the
most basic form, one distinguishes between just two classes: delay-sensitive and delay-tolerant traffic, which are dealt with differently: the delay-sensitive traffic enjoys service priority over the delay-tolerant traffic. The idea behind this type of service differentiation is that, if one would have just one class, all traffic should be offered strict performance guarantees; the introduction of multiple classes enables a more efficient use of resources (in terms of the achievable utilization). The idea of traffic differentiation has been exploited several times before. For instance, in the 1990s this type of differentiation was pursued for ATM [79], later similar ideas were developed for IP (see for instance [80] for a recent account of the state of affairs). More detailed background on Carrier Ethernet is found in e.g. [81].

In an abstract form under the paradigm described above a network element can be modeled as a two-level queueing system. In the first place there is a high-priority queue, particularly intended to take care of the delay-sensitive traffic, getting strict priority, and being equipped with a small buffer (basically just to take care of packet-level queueing effects). Then there is a low-priority queue, meant to serve delay-tolerant traffic, which uses the residual service capacity, and which may be equipped with a sizeable buffer. Noticing that the high-priority queue essentially does not ‘see’ the low-priority traffic, the performance of this queue reduces to that of a single queue in isolation, which has been studied extensively in the literature. The performance of the low-priority queue is considerably more involved, though, as it is a queue that is served at a randomly fluctuating rate. It is the performance of this queue that we address in this chapter.

In the literature several studies have been devoted to performance issues of priority queues. In this respect we mention [82], which considers the system assuming that the input streams behave as Markov fluid sources, and [83], in a fairly general light-tailed framework. These studies do not take into account, though, that the traffic streams are policed: incoming traffic is checked for compliance with a traffic contract, and if not excess traffic is marked. In view of this, attention has been paid to identifying the ‘worst case traffic’ that complies with a given traffic contract; here ‘worst case’ refers to yielding the worst possible performance, according to a given metric [84–86].

In [78] it was argued that the worst case profile in case of a leaky bucket regulator (defined through a peak rate, a sustainable rate, and a maximum burst size) is a deterministic on-off stream. This stream alternates between transmitting at the peak rate and being silent, with the on- and off times chosen such that the burst size is exactly the maximum burst size, and the average rate equals the sustainable rate. When multiplexing several of these ‘worst case sources’, the only randomness in the model is the ‘phasing’ of the individual sources. For a single queue, [78] develops a technique
to (conservatively) estimate the loss probability (for a given mix of sources, each with their specific leaky bucket triplet), which also leads to a computationally attractive call admission procedure.

This chapter focuses on the two-level queueing system described above. We address issues related to the delivery of performance guarantees and resource allocation, and the techniques developed can in principle also be used for admission control purposes. In more detail, the contributions are the following:

- In Section 2 we consider two versatile models for the service rate consumed by the high-priority queue, one that is based on a multi-rate model [87], and another one on a Normal approximation. For both we give expressions for the loss-probability in the low-priority queue. We do so, both for the case of bufferless multiplexing in the low-priority queue (known as rate envelope multiplexing, or shortly REM, [88]), as well as the case of the worst case traffic (transmitted into a moderately sized or large buffer) that was discussed above.

- These expressions being computationally demanding, we consider in Section 3 their asymptotics. After a rescaling of the resources (buffer and link rate), and focusing on the so-called many-flows scaling, manageable expressions are derived, in terms of exact asymptotics for REM and in terms of logarithmic asymptotics for worst case traffic. The latter expressions generalize the setup of [78] to our two-level setting.

- Then we perform in Section 4 numerical experiments, underscoring the accuracy of our approximations. They also indicate that the effective bandwidth concept developed in [78] tends to be rather conservative, in line with earlier findings reported in [89] for the single-queue case.

2 Model

We consider a model with two classes of traffic: a high-priority (hp) class and a low-priority (lp) class. The hp traffic is fed into a buffered resource that is emptied at rate $C$. As this traffic is likely to be delay-sensitive, the buffer of this class is typically relatively small. For ease we assume the buffer has size 0, but in practice it is a small positive size (sufficiently large to absorb packet-level queueing). The remaining service capacity is used by the lp traffic. If temporarily the input rate of this lp traffic exceeds the available service rate, traffic can be stored in a buffer of size $B$. 
The hp queue does not ‘see’ the lp traffic, and therefore the hp queue can be treated as a queue in isolation. These single-queues have been thoroughly studied in the literature, and accurate performance models are available. The analysis of the lp queue, though, is more involved; it can be regarded as a queue with time-varying service rate, where the stochastic characteristics of this service rate behave independently of the input of the lp queue. To model the lp queue, we first propose two models that describe the amount of bandwidth (that is, service capacity) taken away by the hp class.

2.1 Bandwidth used by hp queue

In this subsection, we introduce two models that describe the amount of bandwidth used by hp traffic. We chose these models because of their flexibility: perhaps the hp class does not exactly obey the characteristics of the models we propose, but by tuning the models’ parameters an accurate fit can be realized.

Model 1: Multi-rate model. Let there be $I$ types of hp flows. Flows of type $i$ arrive according to a Poisson process of rate $\lambda_i$, and their mean duration is $\mathbb{E}D_i$; define the ‘load’ imposed by type $i$ by $\nu_i := \lambda_i \mathbb{E}D_i$ (in terms of number of users). Let $\alpha_i$ be the traffic rate of a single type-$i$ flow — we assume this rate to be more or less fixed during its duration (but it can also be thought of as some sort of effective bandwidth).

With $k \in \mathbb{N}$ denoting the ($I$-dimensional) vector representing the numbers of jobs of all types, such a vector is admissible if $\sum_i \alpha_i k_i \leq C$, where we recall that $C$ denotes the link rate; we call this admissible set $\mathcal{S}(C)$. It is well-known from the theory of loss networks [87] that, with $X_i$ denoting the number of type-$i$ calls in steady state, for $k \in \mathcal{S}(C)$

$$
P(X = k) = \frac{1}{\mathbb{B}^{\mathcal{M}}(C)} \prod_{i=1}^{I} \frac{\nu_i^{k_i}}{k_i!},$$

where $1/\mathbb{B}^{\mathcal{M}}(C)$ serves as normalizing constant:

$$
\mathbb{B}^{\mathcal{M}}(C) := \sum_{k \in \mathcal{S}(C)} \prod_{i=1}^{I} \frac{\nu_i^{k_i}}{k_i!}.
$$

We assume that $\sum_i \alpha_i \nu_i \leq C$; this is a highly natural assumption, as it entails that the ‘point of operation’ is admissible. The rate used by the hp class is $C_{\text{hp}} = \sum_i \alpha_i X_i$. Observe that $X$ has a truncated Poisson distribution: with $\bar{X}_i$ having a Poisson distribution with mean $\nu_i$, it holds that

$$
P(X = k) = P(\bar{X} = k \mid \bar{X} \in \mathcal{S}).$$
In this multi-rate model based approximation, one could try to choose the $\nu$ and $A$ such that models matches the measurement data; in case $I = 1$ this gives two parameters to tune (and in general $2I$).

**Model 2: Normal approximation.** A simpler model for the rate consumed by the high priority class could be the following. With $\mathcal{N}(\mu, \sigma^2)$ denoting a normal random variable with mean $\mu$ and variance $\sigma^2$, one can use the model that the rate used by the high-priority class reads, for some $\mu \in (0, C)$ and $\sigma^2$,

$$C_{hp}^N := \max\{0, \min\{C, \mathcal{N}(\mu, \sigma^2)\}\}.$$  

In this normal approximation, one should pick the $\mu$ and $\sigma$ such that the fit with the measurement data is sufficiently good.

The applicability of Normal approximations has been extensively validated in [5, 32, 90, 91]. Their use is justified as long as there is a sufficient level of aggregation. This refers in the first place to aggregation in terms of the number of end-users: this should be high enough to make sure that the central limit theorem has kicked in, and in addition these users should be ‘sufficiently homogeneous’. At the same time, there should be enough aggregation in time: at very short time scale, the rate process is so erratic that the normal approximation does not work properly.

### 2.2 Performance of the lp queue

Suppose now that the lp customers can use the remaining capacity, that is, $C - C_{hp}$. Let there be $J$ types of lp flows, and let $\varepsilon$ be maximally allowable loss probability. We now consider two ways of dealing with the lp queue. The first approach does not exploit the buffering capability of the lp queue (traffic is assumed to be lost as soon as the input rate of the lp queue exceeds the available bandwidth); traffic is modeled as on-off (where no assumptions are imposed on the distributions of the on- and off times). The other approach explicitly takes into account the buffering capability, and in addition the lp traffic is assumed to be compliant with a leaky-bucket regulator. Due to the fact that the latter approach relies on the ideas of [78], on the computational level the procedures are very similar.

A more detailed description of these two options is as follows:

- In the first approach we abstract from the lp buffer $B$, and we do as if packets are lost as soon as the input rate of lp traffic exceeds $C - C_{hp}$. 
Suppose a flow of type \( j \) transmits traffic at rate \( \beta_j \) during random times \( T_{\text{on},j} \), and is silent during random times \( T_{\text{off},j} \); let the sequence of on- and off-times be i.i.d. random variables, and let both sequences be mutually independent. We define the probability of a source of type \( j \) transmitting traffic by 
\[
\pi_j := \frac{T_{\text{on},j}}{T_{\text{on},j} + T_{\text{off},j}}.
\]
Supposing there are \( \ell_j \) jobs of type \( j \), then the aggregate rate required by type \( j \) is 
\[
\beta_j Y_j,
\]
where \( Y_j \) has a binomial distribution with parameters \( \ell_j \) and \( \pi_j \). It is concluded that, in the multi-rate setting, a vector \( \ell \) can be admitted if
\[
q_1^\ell := \sum_{k \in \mathcal{N}(C)} \mathbb{P}(X = k) \mathbb{P} \left( \sum_{j=1}^{J} \beta_j Y_j > C - \sum_{i=1}^{I} \alpha_i k_i \right) \leq \varepsilon.
\]
Using the normal approximation, this criterion becomes
\[
q_1^\nu := \frac{1}{\mathbb{E}^\nu(C)} \int_0^C \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right) \mathbb{P} \left( \sum_{j=1}^{J} \beta_j Y_j > C - x \right) dx \leq \varepsilon,
\]
where
\[
\mathbb{E}^\nu(C) := \int_0^C \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right) dx.
\]
This admission policy for the \( l_p \) class is often referred to as rate envelope multiplexing (REM) [88]; the non-standard feature here is that the service rate is random.

- There are computationally attractive ways to take into account the buffering capability of the low-priority queue, though. One of these possibilities relies on the framework presented in [78]. In this framework the focus is on users transmitting traffic in the ‘worst-case-manner’ that still fits in the traffic profile characterized by a regulator. For customers of type \( j \), this regulator is characterized by a sustainable rate \( r_j \) (or ‘token rate’), a peak rate \( p_j \) (larger than or equal to \( r_j \)), and a token buffer size \( \tau_j \). In [78] it is argued that the worst-case profile corresponding to this regulator is a deterministic on-off pattern: it is alternatingly transmitting at rate \( p_j \) during \( \tau_j / (p_j - r_j) \) time units, and then silent during \( \tau_j / r_j \), thus generating an average traffic rate \( r_j \).

Then [78] shows, for a model with just one priority class, that it can be guaranteed that there is zero loss if one uses the following algorithm. Let \( (r_j, p_j, \tau_j) \) be the regulator profile of a source of type \( j \). Let \( K \) be the amount of bandwidth available, and \( B \) the amount of buffering capacity, and define
\[
c_j \equiv c_j(K) := \max \left\{ \frac{p_j}{1 + B(p_j - r_j)/(\tau_j K) r_j} \right\}.
\]
Let there be \( \ell_j \) users of type \( j \). Then the vector \( \ell \) can be admitted with zero loss if \( \sum_j c_j \ell_j \leq K \). This algorithm can be adapted to allow a fraction \( \varepsilon > 0 \) loss. The key observation is that the effective bandwidth \( c_j \) is needed just a fraction \( r_j/c_j \) of the time, and 0 in the remaining fraction. This means that we can use an algorithm similar to the one defined above: for the multi-rate model, we can admit a vector \( \ell \) if

\[
q_2'' := \sum_{k \in \mathcal{S}(C)} \mathbb{P}(X = k) \mathbb{P} \left( \sum_{j=1}^{J} c_j Y_j > C - \sum_{i=1}^{I} \alpha_i k_i \right) \leq \varepsilon,
\]

where \( Y_j \equiv Y_j(\ell_j, \pi_j) \) has a binomial distribution with parameters \( \ell_j \) and \( \pi_j := r_j/c_j \). One needs to be careful with the precise definition of the effective bandwidth: instead of the \( K \) used in the case of a constant capacity, one now needs to take the available service capacity:

\[
c_j \equiv c_j \left( C - \sum_{i=1}^{I} \alpha_i k_i \right).
\]

One computes \( q_2'' \) analogously.

For proper resource allocation, there is a need for fast and accurate techniques to determine the probabilities \( q_1 \) and \( q_2 \). These are developed in the next section.

### 3 Asymptotic analysis

In the previous section we have distinguished essentially two cases: the lp class performing either rate envelope multiplexing (REM) or following the ‘lossy’ EMW algorithm; on top of that we distinguished between two models describing the bandwidth left over to the lp queue. In all these cases one still needs to evaluate the loss probability experienced by the lp queue, which boils down to evaluating sums over the state space \( \mathcal{S}(C) \) in the case of the multi-rate model, and an integral in the case of the normal approximation. For both models we now develop accurate approximations of the loss probability.

#### 3.1 Multi-rate model

In the multi-rate model, the main difficulty lies in the fact that evaluation of \( q_1'' \) and \( q_2'' \) requires summation over the state space \( \mathcal{S}(C) \), which tends to be large. The idea
Chapter 6. Resource allocation issues of a two-level network resource

is now to circumvent this problem by scaling by \( n \): we replace \( \nu_i \mapsto n\nu_i \), \( C \mapsto nC \), \( B \mapsto nB \), and \( \ell_j \mapsto n\ell_j \); under this scaling the decay rate of the above probabilities can be determined, as we will show in this section.

We first focus on the setup of rate envelope multiplexing. \( X \) has now a truncated Poisson distribution with parameters \( n\nu_1 \) up to \( n\nu_I \), and we therefore write \( X^{(n)} \). The state space is now \( \mathcal{S}(nC) \). The random variables \( Y_j \) have now binomial distributions with parameters \( n\ell_j \) and \( \pi_j \), and to express the dependence on \( n \) we write \( Y^{(n)}_j \). The goal is to find the exact asymptotics of

\[
q_{M1}^{(n)} := \sum_{k \in \mathcal{S}(nC)} P(X^{(n)} = k) \left( \sum_{j=1}^J \beta_j Y^{(n)}_j > nC - \sum_{i=1}^I \alpha_i k_i \right),
\]

that is, an explicit function \( \phi(\cdot) \) such that \( \phi(n)q_{M1}^{(n)}(n) \to 1 \) as \( n \to \infty \). From the interpretation of the distribution of \( X \) being truncated Poisson, we can alternatively write, with \( \bar{X}_i^{(n)} \) having a Poisson distribution with mean \( n\nu_i \),

\[
q_{M1}^{(n)} = \mathbb{P}\left( \sum_{i=1}^I \alpha_i \bar{X}_i^{(n)} + \sum_{j=1}^J \beta_j Y_j^{(n)} > nC \mid \sum_{i=1}^I \alpha_i \bar{X}_i^{(n)} \leq nC \right).
\]

With \( V_m, m = 1, \ldots, n \), i.i.d. random variables distributed as \( \sum_{i=1}^I \alpha_i \bar{X}_i^{(1)} \), and \( W_m, m = 1, \ldots, n \), i.i.d. random variables distributed as \( \sum_{j=1}^J \beta_j Y_j^{(1)} \), we have

\[
q_{M1}^{(n)} = \mathbb{P}\left( \frac{1}{n} \sum_{m=1}^n (V_m + W_m) > C \mid \frac{1}{n} \sum_{m=1}^n V_m \leq C \right).
\]

Now consider the following lemma.

**Lemma 3.1.** Let \( V_m, m = 1, \ldots, n \), and \( W_m, m = 1, \ldots, n \), be mutually independent sequences of i.i.d. random variables. Let \( \mathbb{E}V_i < C \). Then

\[
\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left( \frac{1}{n} \sum_{m=1}^n (V_m + W_m) > C \mid \frac{1}{n} \sum_{m=1}^n V_m \leq C \right) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left( \frac{1}{n} \sum_{m=1}^n (V_m + W_m) > C \right).
\]

**Proof:** It is immediate that the decay rate (6.2) equals the difference of two decay rates:

\[
\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left( \frac{1}{n} \sum_{m=1}^n (V_m + W_m) > C, \frac{1}{n} \sum_{m=1}^n V_m \leq C \right) - \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left( \frac{1}{n} \sum_{m=1}^n V_m \leq C \right),
\]
The latter decay rate equals 0 due to the law of the large numbers. Noticing that
\[
\mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} (V_m + W_m) > C\right) \geq \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} (V_m + W_m) > C, \frac{1}{n} \sum_{m=1}^{n} V_m \leq C\right)
\]
\[
= \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} (V_m + W_m) > C\right) - \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} (V_m + W_m) > C, \frac{1}{n} \sum_{m=1}^{n} V_m > C\right),
\]
in conjunction with
\[
\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} (V_m + W_m) > C, \frac{1}{n} \sum_{m=1}^{n} V_m > C\right)
\]
\[
\leq \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} V_m > C\right) < \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} (V_m + W_m) > C\right),
\]
the stated follows.

As we assumed that \(\sum_i \alpha_i \nu_i < C\), Lemma 3.1 implies that the exact asymptotics of \(q_1^{\#}(n)\) and those of
\[
\bar{q}_1^{\#}(n) := \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} (V_m + W_m) > C\right)
\]
coincide. Now we are in a position to apply the celebrated Bahadur-Rao result to establish the exact asymptotics of \(q_1^{\#}(n)\). Let, with \(Z_m := V_m + W_m\),
\[
\Lambda(s) := \log \mathbb{E} e^{sZ_1} = \sum_{i=1}^{I} \nu_i (e^{s\alpha_i} - 1) + \sum_{j=1}^{J} \ell_j \log (1 - \pi_j + \pi_j e^{s\beta_j})
\]
denote the log-moment generating function of \(Z_1\), and
\[
\Phi(C) := \sup_{s > 0} (sC - \Lambda(s))
\]
the corresponding Legendre transform. In addition, let \(s^*\) be the optimizing \(s\) in the definition of \(\Phi(C)\). Let \(f(n) \sim g(n)\) denote \(f(n)/g(n) \to 1\) as \(n \to \infty\).

**Theorem 3.2.** Distinguish between \(Z_m\) being lattice and non-lattice.

1. Assume \(Z_1\) is a lattice random variable, i.e., for certain numbers \(z_0\) and \(d\) the random variable \((Z_1 - z_0)/d\) is an integer number, with \(d\) being the largest one with this property. Then, as \(n \to \infty\)
\[
q_1^{\#}(n) \sim \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} Z_m > C\right) \sim \frac{s^* d}{1 - e^{-s^* d}} \frac{e^{-n\Phi(C)}}{s^*/\sqrt{2\pi n \Lambda''(s^*)}}.
\]
2. If $Z_1$ is non-lattice, then, as $n \to \infty$,

$$q_1^{ll}(n) \sim \mathbb{P}\left(\frac{1}{n} \sum_{m=1}^{n} Z_m > C\right) \sim \frac{e^{-n\Phi(C)}}{s^* \sqrt{2\pi n \Lambda''(s^*)}}.$$ 

There are alternative ways to compute the exponent $\Phi(C)$ in the expansion of $q_1^{ll}(n)$. To this end, let $\mathcal{F}(C) := \{x \in \mathbb{R}^I : \alpha_i x_i \leq C\}$, and define

$$\xi(x) := \sup_{s \geq 0} \left( sx - \sum_{j=1}^{J} \ell_j \log(1 - \pi_j + \pi_j e^{s\beta_j}) \right);$$

$$\omega(x) := \sum_{i=1}^{I} \left( x_i \log \frac{\nu_i}{x_i} + x_i - \nu_i \right) - \xi\left( C - \sum_{i=1}^{I} \alpha_i x_i \right).$$

This following proposition is proven in the appendix.

**Proposition 3.3.** $\Phi(C) = \sup_{x \in \mathcal{F}(C)} \omega(x)$.

**Example 1.** Let us consider the most simple example: one traffic type within both classes, and $\alpha = \beta = 1$. The stability condition is then $\nu + \ell \pi < C$. It is easily verified that

$$\xi(x) = (C - x) \log \frac{C - x}{\pi} + (\ell - C + x) \log \frac{\ell - C + x}{1 - \pi} - \ell \log \ell$$

so that $\Phi(C)$ equals

$$\max_{x \leq C} \left( x \log \frac{\nu}{x} + x - \nu - (C - x) \log \frac{C - x}{\pi} - (\ell - C + x) \log \frac{\ell - C + x}{1 - \pi} + \ell \log \ell \right).$$

After tedious calculations, this eventually leads to the following optimizing $x$:

$$x^* = -\frac{1}{2}(\ell - C) - \frac{1 - \pi}{2\pi} \nu + \sqrt{\left( \frac{\ell - C}{2} \right)^2 + \frac{1}{2} \frac{1 - \pi}{\pi} \nu(\ell + C) + \left( \frac{(1 - \pi)\nu}{2\pi} \right)^2}.$$ 

It is easy to show that $x^* \leq C$ with equality being possible only if $\ell = 0$ or, trivially, $C = 0$.

If we wish to find the exact asymptotics, we have to work with the setup of Thm. 3.2. It can be calculated that

$$s^* = \log \left( \frac{1}{2} \frac{z - \ell}{\nu} - \frac{1}{2} \frac{1 - \pi}{\pi} + \sqrt{\left( \frac{\ell - z}{2\nu} + \frac{1 - \pi}{2\pi} \right)^2 + \left( \frac{(1 - \pi)z}{\pi\nu} \right)^2} \right)$$

and

$$\Lambda''(s) = e^s \left( \nu + \frac{\ell \pi (1 - \pi)}{(1 - \pi + \pi e^s)^2} \right).$$
This enables the computation of the exact asymptotics.

We now continue with the evaluation of \( q_{2n}(n) \), i.e., the setting in which the lp traffic streams are allocated their effective bandwidth. As before, we use the scaling \( \nu_i \mapsto n
u_i \), \( C \mapsto nC \), \( B \mapsto nB \), and \( \ell_j \mapsto n\ell_j \). The probability \( q_{2n}(n) \) can alternatively be written as

\[
P \left( \sum_{i=1}^{I} \alpha_i \bar{X}_i^{(n)} + \sum_{j=1}^{J} c_j \left( nC - \sum_{i=1}^{I} \alpha_i \bar{X}_i^{(n)} \right) \cdot Y_j^{(n)} > nC \middle| \sum_{i=1}^{I} \alpha_i \bar{X}_i^{(n)} \leq nC \right).
\]

As before, it can be argued that, when looking at asymptotics, the impact of the condition is negligible; we therefore leave it out in the sequel. Let \( V_{m,i} \), \( m = 1, \ldots, n \), be i.i.d. samples from a Poisson distribution with mean \( \nu_i \), and \( W_{m,j} \) i.i.d. samples from a binomial distribution with parameters \( \ell_j \) and \( \pi_j \). We thus arrive at

\[
P \left( \sum_{i=1}^{I} \alpha_i \left( \frac{1}{n} \sum_{m=1}^{n} V_{m,i} \right) + \sum_{j=1}^{J} \gamma_j \cdot \left( \frac{1}{n} \sum_{m=1}^{n} W_{m,j} \right) > C \right),
\]

where

\[
\gamma_j := \max \left\{ \frac{p_j}{1 + nB(p_j - r_j) / (\tau_j (nC - \sum_{i=1}^{I} \alpha_i \sum_{m=1}^{n} V_{m,i}))}, r_j \right\}.
\]

Note, however that \( (\alpha_i / n) \sum_{m=1}^{n} V_{m,i} \) and \( (\gamma_j / n) \sum_{m=1}^{n} W_{m,j} \) are not independent: the value of \( n^{-1} \sum_{m=1}^{n} V_{m,i} \) affects \( \gamma_j \). As a result, we cannot write

\[
\sum_{i=1}^{I} \alpha_i \left( \frac{1}{n} \sum_{m=1}^{n} V_{m,i} \right) + \sum_{j=1}^{J} \gamma_j \cdot \left( \frac{1}{n} \sum_{m=1}^{n} W_{m,j} \right)
\]

as the sum of \( n \) i.i.d. random variables. This also means that Bahadur-Rao cannot be applied here, and therefore we cannot identify the exact asymptotics of \( q_{2n}(n) \) in this way. Logarithmic asymptotics can be derived, though, as follows.

The idea is to condition on \( n^{-1} \sum_{m=1}^{n} V_{m,i} \approx x_i \), for \( i = 1, \ldots, I \), and then to find the most likely value of \( x \) corresponding to the event under consideration. To this end, first recall that

\[
\lim_{n \to \infty} \frac{1}{n} \log P \left( \frac{1}{n} \sum_{m=1}^{n} V_{m,i} \approx x_i \right) = -H_i(x_i), \quad \text{where} \quad H_i(x) := x \log \frac{x}{\nu_i} + x - \nu_i.
\]
Also, conditioning on $n^{-1} \sum_{m=1}^{n} V_{m,i} \approx x_i$, we have that, informally,

$$
\lim_{n \to \infty} \frac{1}{n} \log P \left( \sum_{j=1}^{J} \gamma_j \cdot \left( \frac{1}{n} \sum_{m=1}^{n} W_{m,j} \right) > C \right) = - \sup_{s \geq 0} \left( sC - \sum_{j=1}^{J} \ell_j \log \left( 1 - \pi_j(x) + \pi_j(x)e^{s\gamma_j(x)} \right) \right),
$$

where

$$
\gamma_j(x) := \max \left\{ \frac{p_j}{1 + B(p_j - r_j)/\left( \tau_j(\sum_{i=1}^{I} \alpha_i x_i) \right)}, r_j \right\}, \quad \pi_j(x) := \frac{r_j}{\gamma_j(x)}.
$$

Upon combining these decay rates, we find

$$
\lim_{n \to \infty} \frac{1}{n} \log q_{2^*}^{\mathcal{N}}(n) = - \inf_{x} \left( \sum_{i=1}^{I} H_i(x_i) + \sup_{s \geq 0} \left( s \left( C - \sum_{i=1}^{I} \alpha_i x_i \right) - \sum_{j=1}^{J} \ell_j \log \left( 1 - \pi_j(x) + \pi_j(x)e^{s\gamma_j(x)} \right) \right) \right).
$$

Here the optimization over $x$ essentially finds the most likely value of $n^{-1} \sum_{m=1}^{n} V_{m,i}$, for $i = 1, \ldots, I$ such that (6.3) exceeds $C$. This result can be rigorously proven applying the same line of arguments as in the proof of Prop. 3.3. This proof is tedious, though, and does not add any new insights. It is stressed that the optimization over $s$ can be performed explicitly, and therefore the computation of this decay rate just requires the evaluation of a $I$-dimensional minimization in order to find the optimal $x$; the objective function is rather ‘ill-behaved’ though, particularly due to the non-smooth behavior of the $\gamma_j(x)$.

### 3.2 Normal approximation

Also in the case of the normal approximation, we can scale by $n$. It effectively means that we have to replace $\mu \mapsto n\mu$ and $\sigma^2 \mapsto n\sigma^2$. Let

$$
U^{(n)} \overset{d}{=} \mathcal{N}(n\mu, n\sigma^2),
$$

and let $U_m, m = 1, \ldots, n$, be a sequence of i.i.d. $\mathcal{N}(\mu, \sigma^2)$ random variables. Then we have that

$$
q_{1^*}^{\mathcal{N}}(n) = \mathbb{P} \left( \frac{1}{n} \left( U^{(n)} + \sum_{m=1}^{n} W_m \right) > C \ \bigg| \ \ 0 \leq U^{(n)} \leq C \right).
$$
As in the multi-rate model, under the assumption of \( \mu < C \), one can show that

\[
q_1^N(n) \sim q_1^N(n) := \mathbb{P}\left( \frac{1}{n} \sum_{m=1}^{n} (U_m + W_m) > C \right).
\]

Then using the Bahadur-Rao result, we conclude that

\[
q_1^N(n) \sim e^{-n \Psi(C)} / s^* \sqrt{2\pi n M''(s^*)},
\]

where

\[
M(s) := (\mu s + 1/2 \sigma^2 s^2) + \sum_{j=1}^{J} \ell_j \log(1 - \pi_j + \pi_j e^{s \beta_j})
\]

and \( \Psi(C) := \sup_{s \geq 0} (sC - M(s)) \), and \( s^* \) is the optimizing argument in the definition of \( \Psi(C) \); the distribution of \( U_1 + W_1 \) is obviously non-lattice.

We conclude this section by presenting the decay rate of \( q_2^N(n) \). This decay rate can be identified analogously to the decay rate of \( q_2^N(n) \). Without proof we state that

\[
\lim_{n \to \infty} \frac{1}{n} \log q_2^N(n) = -\inf_{x} \left( \frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 + \sup_{s \geq 0} \left( s(C - x) - \sum_{j=1}^{J} \ell_j \log \left( 1 - \pi_j(x) + \pi_j(x) e^{s \gamma_j(x)} \right) \right) \right),
\]

where

\[
\gamma_j(x) := \max \left\{ \frac{p_j}{1 + B(p_j - r_j)/(r_j(C - x))}, r_j \right\}, \quad \pi_j(x) := \frac{r_j}{\gamma_j(x)}.
\]

### 3.3 Some remarks on the efficiency of the algorithm of EMW

In this section we briefly address efficiency issues of the effective bandwidth approach from [78], both in the lossless case (i.e., based on (6.1)) and in the case were loss is allowed.

We first remark that it is trivial to see that, when there are \( \ell_j \) sources of class \( j \), there is indeed no loss as long as \( \sum_j \ell_j c_j \leq C \); it is noticed that this reasoning was not given in [78]. With \( T_i := \tau_i/(p_i - r_i) \) denoting the length (in time) of the burst in the worst case
profile,

\[ \sum_{j=1}^{J} \ell_j (p_j - c_j) T_j \leq \sum_{j=1}^{J} \ell_j T_j \left( p_j - \max \left\{ r_j, \frac{p_j T_j}{B/C + T_j} \right\} \right) \]
\[ = \sum_{j=1}^{J} \ell_j \min \left\{ T_j (p_j - r_j), \frac{B p_j T_j}{B/C + T_j} \right\} \]
\[ \leq \sum_{j=1}^{J} \ell_j \frac{B}{C} \max \left\{ r_j, \frac{p_j T_j}{B/C + T_j} \right\} = \sum_{j=1}^{J} \ell_j c_j B \leq B. \]

In [89], however, it was argued that the ‘lossless admissible region’ was substantially larger. With the source types ordered such that \( T_1 \geq T_2 \geq \ldots \geq T_J \), there is no loss if the combination is ‘stable’ (i.e., \( \sum_{j} \ell_j r_j \leq C \)), and for all \( j = 1, \ldots, J \),

\[ \left( \sum_{k=1}^{j-1} \ell_k p_k T_j \right) + \ell_j p_j T_j + \left( \sum_{k=j+1}^{J} \ell_k (r_k (T_j - T_k) + p_k T_k) \right) \leq T_j C + B; \quad (6.4) \]

this set can be significantly larger than the set resulting from [78].

Example 2. Consider the easy case of \( J = 2 \) and \( \ell_1 = \ell_2 = 1 \). We assume \( p_1 = p_2 = 1 \), \( r_1 = r_2 = \frac{1}{2} \), and \( T_2 = 2T_1 = 2 \); we take \( C = 1 \). It is seen easily that the effective bandwidths of both sources, as functions of \( B \), are given through

\[ c_1 = \max \left\{ \frac{1}{2}, \frac{1}{B + 1} \right\}, \quad c_2 = \max \left\{ \frac{1}{2}, \frac{2}{B + 2} \right\}, \]

so that a buffer of size \( B = 2 \) would be needed to avoid any loss.

Constraints (6.4) result from solving an optimization program of the type (in our situation with \( J = 2 \) and \( \ell_1 = \ell_2 = 1 \))

\[ \min(p_1 - c_1) T_1 + (p_2 - c_2) T_2 \quad \text{under} \quad c_1 + c_2 \leq C, \quad c_i \in [r_i, p_i]; \]

it is easily seen that this program yields an upper bound to the buffer space needed. In our example, the optimum is achieved for \( c_1 = c_2 = \frac{1}{2} \), such that just \( \frac{3}{2} \) buffer space is needed, i.e., less than what would follow from [78].

If one draws the graph of the buffer content of time, however, one observes that even less buffer is needed. In the worst-case situation of both sources switching on at time 0, we see that the buffer level reaches 1 at time 1, stays on that level till time 3, and drops to 0 at time 4 (which is then repeated periodically). In other words: a buffer space of size 1 suffices in this case.
Chapter 6. Resource allocation issues of a two-level network resource

The above example shows that the lossless algorithm can be quite conservative; a more realistic example is given in [89, p. 26]. This effect becomes even more pronounced if we consider the variant of [78] that does allow loss, due to the fact that the algorithm proposed in [78] does not fully exploit statistical multiplexing. This issue was studied in detail in [89], relying on large-deviations-based asymptotics that capture the effect of statistical multiplexing more accurately, as developed in e.g. [92]. With $A_j(t)$ denoting the amount of traffic generated by a periodic on-off stream of type $j$ in an arbitrary window of length $t$, the loss probability in a single queue equipped with a (non fluctuating) service rate $nC$ and a buffer size $nB$ is, due to the findings of [92], well approximated by

$$
\exp \left( -n \inf_{t > 0} \sup_{s > 0} \left( (B + Ct)s - \sum_{j=1}^{J} \ell_j \log \mathbb{E}e^{sA_j(t)} \right) \right),
$$

in the situation that there are $n\ell_j$ streams of type $j$.

In [93] it is explained how to compute the moment generating function $\mathbb{E}e^{sA_j(t)}$ for our worst case sources. It is noted that this approach requires many cases to be distinguished, and therefore one may look for simpler methods. An easier alternative is the following. The loss probability can be conservatively approximated by replacing the worst case sources by their ‘instantaneous counterpart’: the full burst arrives instantaneously, rather than gradually (that is, spread over $T_{on}$). In more detail, first note that the period of a worst case source with profile $(r, p, \tau)$ is given by

$$
\bar{T} := \frac{\tau}{p - r} + \frac{\tau}{r},
$$

and that the maximum burst size is

$$
m := pT_{on} = \frac{pr}{p - r};
$$

it can be easily seen that the average traffic rate generated by the source equals $m/\bar{T} = r$, as desired. It is now easily checked that, with $[t]_T$ defined as $[t/\bar{T}]$, the moment generating function of the instantaneous-input variant equals

$$
\mathbb{E}e^{s\tilde{A}(t)} = e^{sm[t]_T} \left( \frac{t - \bar{T} \cdot [t]_T}{\bar{T}} e^{sm} + \left( 1 - \frac{t - \bar{T} \cdot [t]_T}{\bar{T}} \right) \right)
$$

Finally, one may look at an approximation in which it is assumed that for all $t$ we have that $A(t) = rt$, i.e., that every source transfers at constant rate of $r$ in a continuous fashion. This may be regarded as an ‘optimistic approach’. On the other hand, it greatly reduces the numerical cost, and could therefore at least be seen as some sort of ‘sanity check’.
4 Numerical results

We now consider a number of experiments that assess the quality of the approximations made in our analysis. To shed light on resource allocation issues, our goal is to assess the impact of the various parameters on the performance. To come up with realistic scenarios, we took the ratio of the mean and standard deviation of the traffic rate (at a given time scale) roughly in line with the findings of [32]. For the buffer parameters we based our choice on the values given by leading equipment vendors in their 10GE line cards specifications.

Regarding the scaling parameter $n$, the following remark is in place. The reason for introducing this parameter is that under the proportional scaling of resources as well as input, we could prove results on the various decay rates. In the present section, our objective is to perform numerical experiments with representative parameter values. From the formulas it is readily verified that we can then pick $n = 1$, such that we can take for $B$ the ‘real’ (that is, non-scaled) buffer size, for $C$ the real link rate, for the $\nu$ the real load generated by the hp traffic, and for the $\ell$ the real number of lp sources. It is left to the reader to verify that these approximations do not change when taking $n = K$ and buffer $B/K$, link rate $C/K$, load $\nu/K$ and number of lp sources $\ell/K$; in other words, to evaluate the approximations the value of the scaling parameter is irrelevant; the only thing that matters is the product of the scaling parameters and the buffer size, link rate, load, and number of lp sources. A similar line of reasoning is followed when the hp traffic is modeled by a truncated Normal distribution.

4.1 Experiments with a non-buffered resource and rate envelope multiplexing

We first consider the setting of rate envelop multiplexing, where the rate consumed by hp traffic is described by a multi-rate model. In the top-left panel of Fig. 6.1 we take $I = J = 1$ along with $\alpha = \beta = 1$, so that we are in the situation of Example 1. Picking as indicated above without loss of generality $n = 1$, the link speed of 1 Gbps results in $C = 1000$ (measured in Mbps). We choose $\nu = 420$, $\pi = 0.6$. Then we vary $\ell$ to see how the loss probability $q_1(1)$ is affected. The stability region is $420 + 0.6\ell \leq 1000$.

The graph shows the $10\log$ of the loss probability $q_1(1)$ as a function of $\ell$. There are three lines: (i) a curve based on the exact asymptotics of Thm. 3.2, relying on the Bahadur-Rao result, (ii) a curve based on the logarithmic approximation (only containing the exponential part of the Bahadur-Rao approximation), and (iii) a curve that
estimates the loss probability $q^{M_1}(1)$ based on simulations. Regarding the simulated values, it is mentioned that per data point we used $10^6$ independent samples.

The graphs show that the exact asymptotics match excellently with the simulated values. The fit is less good in the region close to saturation: there in reality the loss probabilities are still well below 1, but the exact asymptotics predict that $q^{M_1}(1) \approx 1$. Notice, however, that this region close to saturation is practically irrelevant (as the target values of the loss probability are considerably lower than $10^{-1}$). It is further observed that the curve that is based on just the exponent (that is, the logarithmic asymptotics) is too pessimistic: in reality $q^{M_1}(n)$ is roughly one order of magnitude smaller; however, the curve captures the right trend excellently, as it is nearly parallel to the simulated curve.

The other three panels show admissible regions for different values of $10^{\log q^{M_1}(1)}$. The top right panel is based on simulation (with $10^6$ runs per data point), the bottom-left panel is based on Thm. 3.2, and the bottom-right panel on just the decay rate (logarithmic asymptotics). Again we observe that the results based on Thm. 3.2 match the simulated regions excellently. The regions based on the logarithmic asymptotics are again conservative. Also observe that the boundaries of the admissible regions (for various values of the loss probability) are (nearly) straight lines.

### 4.2 Experiments with buffered resource and worst case traffic

In this situation the lp traffic is modeled by worst sources. Whereas in the previous example we assumed that the rate consumed by the hp traffic was described by a multi-rate model, we now assume it to have a truncated normal approximation. In these numerical experiments we also want to examine whether the conservative nature of the setup of [78] in the single queue, as reported on in Section 3.3, carries over to our two-level queueing model. To this end, we can compare the loss probabilities (and admissible regions), as obtained by the asymptotics we found in Section 3, with the counterpart of (6.5) for our priority queue (specified in more detail below). We model the hp traffic by fractional Brownian motion (fBm): the mean and the variance of the amount of hp traffic transmitted in an interval of length $t$ are $\mu_{fBm} t$ and $\sigma^2_{fBm} t^{2H}$, for parameters $\mu_{fBm}$, $\sigma^2_{fBm}$ and $H \in (0, 1)$. In real network traffic, we have that $H \in (\frac{1}{2}, 1)$: there is positive correlation.

Realize that in the fBm model there is not such a notion as traffic rate [36]; to this end, observe that, as $t \downarrow 0$,

$$\text{Var} \left( \frac{A(t)}{t} \right) = \sigma^2_{fBm} t^{2H-2} \to \infty.$$
Note, however, that the normal approximation is based on the notion of traffic rate (recall that it is the traffic rate consumed by the hp traffic that has a truncated Normal distribution). This explains why we should define a ‘reference time interval’ Δ (roughly corresponding to the time scale of traffic fluctuations that can be absorbed by the hp queue), and we let μ and σ² the mean and variance, respectively, of the amount of service capacity used by the fBm in such an interval of length Δ. It is seen easily that μ = μ_{fBm}, whereas σ² follows from σ² = σ²_{fBm}Δ^{2H}.

In the scaled setting of Section 3, the counterpart of (6.5) for our two-level system is

\[ \exp \left( -n \inf_{t>0} \sup_{s \in [0,s^*_t]} \left( (B+Ct)s - \sum_{j=1}^J \ell_j \log \mathbb{E} e^{sA_j(t)} - \mu_{fBm}ts - \frac{1}{2} \sigma^2_{fBm}s^2 \Delta^{2H} \right) \right), \]

with \( s^*_t \) defined below, and under the assumption that \( \mu + \sum_j \ell_j r_j < C \). This approach relies on the fact that the total amount of service capacity used by the hp traffic is fractional Brownian motion (up to the truncations mentioned above), so that the moment
Chapter 6. Resource allocation issues of a two-level network resource

generating function of the capacity used in an interval of length \( t \) equals

\[
\Lambda_H(s, t) := \exp \left( \mu ts + \frac{1}{2} \sigma^2 s^2 t^{2H} \right).
\]

With \( s_t \) denoting the optimizer in the objective function appearing in our approximation (for given \( t \)), we have that

\[
\frac{\partial}{\partial s} \Lambda_H(s, t) \bigg|_{s=s_t}
\]

represents the amount of service capacity used by the hp traffic in the interval \([0, t] \). As this can never exceed \( Ct \), we fix \( s^*_t \) at

\[
s^*_t := \frac{(C - \mu)t}{\sigma^2 t^{2H}} = \frac{C - \mu}{\sigma^2} t^{1-2H}. \]

This idea stems from the ‘rough full link approximation’ in [94], and has shown to be highly accurate [95]. For ease we replace the moment generating functions by those of their instantaneous-input counterparts, as introduced in (6.6).

• In Experiment A, we consider a model with the following parameters at 1s reference time scale (as explained above): \( C = 10000 \) (it may be regarded as a 10 Gbps Ethernet link), \( B = 4096 \) (this would be expressed in bits as some vendors provide 10GE line cards with per-port buffer of 512 Mb), \( \mu = 1200 \), \( \sigma^2 = 1000^2 \). The parameters of the lp sources are \((r, p, \tau) = (250, 1000, 250)\). As we mentioned earlier, all the values are already scaled, in the sense that we set \( n = 1 \). We decided to consider \( \text{fBm} \) with \( H = 0.8 \) at the \( \Delta = 0.1 \)s time scale, so that \( \sigma^2_{\text{fBm}} = (158.48)^2 \), and \( C, \mu, r, p \) must be multiplied by \( \Delta \). Now the source’s period is \( \bar{T} = 13.333 \), and the maximal burst size is \( m = 333.3 \). It is readily verified that the stability condition is \( 120 + 25\ell < 1000 \), or equivalently, \( \ell \leq 35 \). With these parameters we performed the experiments.

As mentioned above, our objective is to assess the conservativeness of the method based on [78]. To this end we compared the decay rate of \( q^{(\nu)}_2 (n) \) with

\[
- \inf_{i \in \mathbb{N}} \sup_{s \in [0, s^*_t]} \left( (B + C i \Delta)s - \ell \log Ee^{s A(i \Delta)} - \mu_{\text{fBm}} i \Delta s - \frac{1}{2} \sigma^2_{\text{fBm}} s^2 (i \Delta)^{2H} \right).
\]

The results are shown in the left panel of Fig. 6.2. They once more indicate that the effective bandwidth-based approach from [78] tends to be conservative. Due to its attractive numerical properties, it still may be preferred over alternative methods.

We also compare our finding with the (easily computable) ‘mean’ case, i.e., the situation in which all the sources are assumed to transfer always at their mean rate \( r \) so \( A(t) = rt \). If we consider just a single class for the lp traffic, we get the following optimizing
arguments and objective function value respectively:

\[ s_t^* = \frac{t^{-2H}(B + Wt)}{\sigma_{Bm}^2}, \quad t^* = B \frac{H}{W} \frac{1}{1 - H}, \quad \Phi^* = -\frac{1}{2\sigma_{Bm}^2} \left( \frac{B}{1 - H} \right)^{2-2H} \left( \frac{W}{H} \right)^{2H}, \]

with \( W = C - \ell r - \mu_{Bm} \) The results of the experiment are shown in Fig. 6.2. Note that, with the chosen parameters, hp traffic tends to be highly persistent (\( H = 0.8 \)), while the lp sources may behave quite burstily, their peak rate being four times higher than the mean rate.

The lowest curves depict dynamic simulation results, that is, results based on simulation of the queueing system under consideration. In the situation at hand, all the methods turn out to be conservative, with EMW being quite far off, and REM (which does not take into account a presence of the buffer) being even worse (which is not a surprise as the buffer is very large relative to the bandwidth). All three methods that consider both buffer and self-similarity of the traffic, behave in the very similar way. It is noted, though, that for instance for a loss probability of \( 10^{-6} \), the simulated curve shows that in reality we could allow about 5 more clients than suggested by these three methods.

- In Experiment B, we again have \( C = 10000, \mu = 1200 \) and \( \sigma^2 = 10000^2 \), along with the triplet characterizing the worst case sources being \( (r, p, \tau) = (250, 1000, 250) \). We decrease, however, the ‘persistency level’ of the hp traffic so now we have \( H = 0.6 \) (i.e., hp traffic is ‘less long-range dependent’) and reduce the buffer to 1024, whereas we left the time scale of interest for fBm unchanged (\( \Delta = 0.1s \)). The numerical output is found in Fig. 6.3.

Note that the ‘abscissa range’ is considerably larger than in the previous experiment. The various fBm curves are now clearly separated, illustrating that different approaches to calculate \( A(t) \) may have significant impact.

Another comment should be made on the right panel, which shows the loss probability approximation of the type, where \( \Phi \) is the exponent from the left panel, obtained by the various methods. For some values of \( \ell \) we see that loss probability derived using the REM method is actually lower than its counterpart based on [78]. This may look suspicious at first glance, as REM, contrary to the framework of [78], does not exploit the buffer. The exponents themselves (see the left panel) are, however, in the expected order; an explanation is that it is the Bahadur-Rao-based refinement which makes the difference, see Section 3.2.
Chapter 6. Resource allocation issues of a two-level network resource

This chapter dealt with resource allocation issues in a multiple-priority system featuring Carrier Ethernet. The high-priority queue being standard, our attention focused on methods for the low-priority queue. We proposed various approaches, some of which led to highly accurate approximations, as witnessed by numerical experiments. Future research in this area may relate to (i) systems with more than two priority classes (for instance by introducing a third class, in between the lp and hp classes considered here, where only certain fraction of the traffic is treated in the best effort manner), (ii) the impact of active queue management mechanisms such as random early detection, (iii) a more extensive measurement-based backing of the procedures proposed (including the estimation of the parameters involved, such as the Hurst parameter $H$).
6 Appendix. Proof of Prop. 3.3

Upper bound. It is evident that the number of elements $\# \mathcal{S}(nC)$ of $\mathcal{S}(nC)$ is $O(n^l)$. It follows that $q_k^\mathcal{S}(n)$ is bounded from above by

$$
\# \mathcal{S}(nC) \times \max_{k \in \mathcal{S}(nC)} \mathbb{P}(X(n) = k) \mathbb{P}\left( \sum_{j=1}^l \beta_j Y_j(n) > nC - \sum_{i=1}^l \alpha_i k_i \right).
$$

As $n^{-1} \log \# \mathcal{S}(nC) \to 0$ as $n \to \infty$, we are left with determining an upper bound to

$$
\limsup_{n \to \infty} \frac{1}{n} \log \left( \max_{k \in \mathcal{S}(nC)} \mathbb{P}(X(n) = k) \mathbb{P}\left( \sum_{j=1}^l \beta_j Y_j(n) > nC - \sum_{i=1}^l \alpha_i k_i \right) \right). \tag{6.7}
$$

First consider the normalizing constant; to this end observe that for all $\eta > 0$ and $n$ large enough, with $Z_i(n)$ denoting a Poisson random variable with mean $n \nu_i$,

$$
\sum_{k \in \mathcal{S}(nC)} \prod_{i=1}^l \left( \frac{n \nu_i}{k_i} \right)^{k_i} = \left( e^{-n \sum_{i=1}^l \nu_i} \right) \times \left( \sum_{k \in \mathcal{S}(nC)} \prod_{i=1}^l \left( \frac{n \nu_i}{k_i} \right)^{k_i} e^{-n \nu_i} \right) = \mathbb{P}(Z(n) \in \mathcal{S}(nC)) \geq e^{-n \sum_{i=1}^l \nu_i (1 - \eta)},
$$

using that $Z_i(n)$ is distributed as the sum of $n$ independent Poisson($\nu_i$) random variables, and applying the law of large numbers (recalling that $\nu \in \mathcal{S}(C)$). Using the Stirling-type bound (uniformly in $k$)

$$
\log(k!) \leq k \log k - k + \kappa(k), \quad \kappa(k) := \frac{1}{2} \log(2\pi k) + \frac{1}{12 k},
$$

we thus arrive at

$$
\mathbb{P}(X(n) = k) \leq \prod_{i=1}^l \left( \frac{n \nu_i}{k_i} \right)^{k_i} e^{k_i - n \nu_i} e^{\kappa(k_i)} (1 - \eta),
$$

for all $k \in \mathcal{S}(nC)$. Due to the Chernoff bound, we also have

$$
\log \mathbb{P}\left( \sum_{j=1}^l \beta_j Y_j(n) > nC - \sum_{i=1}^l \alpha_i k_i \right) \leq -\xi_n \left( nC - \sum_{i=1}^l \alpha_i k_i \right),
$$

where $\xi(\cdot)$ is the associated Legendre transform:

$$
\xi_n(x) := \sup_{s \geq 0} \left( sx - n \sum_{j=1}^l \xi_j \log(1 - \pi_j + \pi_j e^{s \beta_j}) \right).
$$

Now realize that the maximum over $\mathcal{S}(nC)$ is smaller than the maximum over the (bigger) set $\mathcal{S}(nC)$. It follows that the decay rate (6.7) is bounded from above by

$$
\limsup_{n \to \infty} \frac{1}{n} \log \max_{k \in \mathcal{S}(nC)} \prod_{i=1}^l \left( \frac{n \nu_i}{k_i} \right)^{k_i} e^{k_i - n \nu_i} e^{\kappa(k_i)} (1 - \eta) \exp \left( -\xi_n \left( nC - \sum_{i=1}^l \alpha_i k_i \right) \right),
$$
which equals (note that the factor \((1 - \eta)\) trivially cancels)

\[
\limsup_{n \to \infty} \frac{1}{n} \log \max_{x \in \mathcal{F}(C)} \prod_{i=1}^{I} \left( \frac{\nu_i}{x_i} \right)^{n x_i} e^{\kappa (x_i - \nu_i)} e^{\kappa (n x_i)} \exp \left( - n \xi \left( C - \sum_{i=1}^{I} \alpha_i x_i \right) \right).
\]

This in turn can be majorized by

\[
\max_{x \in \mathcal{F}(C)} \omega(x) + \limsup_{n \to \infty} \frac{1}{n} \log \max_{x \in \mathcal{F}(C)} \prod_{i=1}^{I} e^{\kappa (n x_i)}.
\]

Now noting that \(\kappa (n x) \leq \frac{1}{2} \log (2 \pi n x) + 1\), and noting that there is a finite \(M\) such that \(x_i \leq M\) for all \(x \in \mathcal{F}(C)\) and all \(i = 1, \ldots, I\), we conclude that

\[
\limsup_{n \to \infty} \frac{1}{n} \log q_{1\#} (n) \leq \max_{x \in \mathcal{F}(C)} \omega(x). \tag{6.8}
\]

**Lower bound.** Call the optimizer in the right-hand side of (6.8) \(x^*\). It is evident that for all \(k \in \mathcal{I}(n)\),

\[
q_{1\#} (n) \geq \mathbb{P}(X(n) = k) \mathbb{P} \left( \sum_{j=1}^{J} \beta_j Y_j(n) > n C - \sum_{i=1}^{I} \alpha_i k_i \right).
\]

Now take \(k = \lfloor n x^* \rfloor \in \mathcal{I}(nC)\). First observe that, again using \(\nu \in \mathcal{I}(C)\) and ‘Stirling’

\[
\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(X(n) = \lfloor n x^* \rfloor) = \liminf_{n \to \infty} \frac{1}{n} \log \prod_{i=1}^{I} \frac{(n \nu_i) \lfloor n x^*_i \rfloor}{\lfloor n x^*_i \rfloor !} = \sum_{i=1}^{I} \left( x^*_i \log \frac{\nu_i}{x^*_i} + x^*_i - \nu_i \right)
\]

The other decay rate we need to study is

\[
\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P} \left( \frac{1}{n} \sum_{j=1}^{J} \beta_j Y_j(n) > C - \sum_{i=1}^{I} \frac{\lfloor n x^*_i \rfloor}{n} \alpha_i \right).
\]

Now choose \(n\) sufficiently large that \(\lfloor n x^*_i \rfloor / n > (1 - \delta)x^*_i\) for all \(i = 1, \ldots, I\). Then, due to ‘Cramér’, the above decay rate majorizes

\[
- \xi \left( C - \sum_{i=1}^{I} \alpha_i x^*_i (1 - \delta) \right).
\]

Using the continuity of the Legendre transform \(\xi(\cdot)\), it follows by letting \(\delta \downarrow 0\) that

\[
\limsup_{n \to \infty} \frac{1}{n} \log q_{1\#} (n) \geq \omega(x^*).
\]

Combining the above bounds, we have proven the stated.