QoS-aware bandwidth provisioning for IP network links

van den Berg, H.; Mandjes, M.R.H.; van de Meent, R.; Pras, A.; Roijers, F.; Venemans, P.H.A.

DOI
10.1016/j.comnet.2005.05.028

Publication date
2006

Published in
Computer Networks

Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

QoS-aware bandwidth provisioning for IP network links

H. van den Berg, M.R.H. Mandjes, R. van de Meent,
A. Pras, F. Roijers, P.Venemans

REPORT PNA-E0406A JUNE 2004

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

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

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

QoS-aware bandwidth provisioning for IP network links

ABSTRACT
Bandwidth provisioning is generally envisaged as a viable way to support QoS in IP networks. To guarantee at the same time cost-efficient use of resources, the crucial question is: what is the minimal bandwidth provisioning required to ensure the QoS level agreed upon (for instance: the probability that the traffic supply exceeds the available bandwidth, over some predefined interval T, is below some small fixed number epsilon. This paper deals with this dimensioning problem, with as crucial novelty that the resulting guidelines are based on coarse traffic measurements. Our approach relies on a powerful `interpolation' formula that predicts QoS on relatively short time scales (say the order of 1 s), by just using large time-scale measurements (say in the order of 5 m, as in the case of the standard MRTG measurements). As a result, we find that, measuring a load rho (in Mbit/s), the required bandwidth (to meet the QoS criterion) has the form rho +alpha sqrt{rho}, where alpha depends on T and epsilon -- this expression is derived under minimal model assumptions. Apart from its simplicity, the dimensioning formula has a number of attractive features, viz. its insensitivity and robustness (as just the load rho is needed), and its transparency (the impact of changing the `QoS-parameters', i.e., T and epsilon, on alpha is explicitly given). The dimensioning rule is validated through extensive measurements obtained in several operational network environments.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: provisioning, IP networks, Gaussian traffic
Note: This work was carried out under the SENTER-project EQUANET

QoS-aware bandwidth provisioning for IP network links

Hans van den Berg [a,b] Michel Mandjes [c,b] Remco van de Meent [b]
Aiko Pras [b] Frank Roijers [a,c,*] Pieter Venemans [a]

[a] *TNO Telecom, Delft, The Netherlands*
[b] *University of Twente, Enschede, The Netherlands*
[c] *CWI, Amsterdam, The Netherlands*

Abstract

Current bandwidth provisioning procedures for IP network links are mostly based on simple rules of thumb, using coarse traffic measurements made on a time scale of e.g, 5 or 15 minutes. A crucial question, however, is whether such coarse measurements give any useful insight into the capacity actually needed: QoS degradation experienced by the users is strongly affected by traffic rate fluctuations on a much smaller time scale. The present paper addresses this question. The goal is to develop provisioning procedures that require a minimal measurement effort.

The bandwidth provisioning formula that we propose (and which we justify under minimal model assumptions) is of the form $\rho + \alpha\sqrt{\rho}$. Here ρ (in Mbit/s) is the load of the system, which can evidently be estimated by coarse traffic measurements (e.g., 5 or 15 minutes measurements). The α depends on the characteristics of the individual flows and the QoS requirements. The QoS measure used is the probability that the traffic supply exceeds the available bandwidth, over some predefined (small) interval T, is below some small fixed number ε. The impact of changing the 'QoS parameters', i.e., T and ε, on the coefficient α is explicitly given. The validity of the bandwidth provisioning rule is assessed through extensive measurements performed in several operational network environments.

Key words: bandwidth provisioning • Quality of Service • traffic measurements • Gaussian traffic • M/G/∞ input

* Corresponding author: TNO Telecom, P.O.Box 5050, 2600 GB Delft, The Netherlands, e-mail: F.Roijers@telecom.tno.nl, phone: +31 15 285 7244, fax: +31 15 285 7057.

1 Introduction

Bandwidth provisioning procedures require a thorough understanding of the relation between the characteristics of the offered traffic, the link speed, and the resulting Quality of Service (QoS). The availablity of such a relation enables the selection of a link capacity that guarantees that the aggregate rate of the offered traffic exceeds the link capacity less than some predefined (small) fraction of time.
In such a 'QoS by provisioning' approach (i.e., allocating enough bandwidth to meet the QoS requirements of all applications present), all traffic streams are treated in the same way. An alternative is to use traffic differentiation mechanisms, such as those of the *IntServ* approach developed within the Internet Engineering Task Force (IETF). IntServ enables stringent QoS guarantees, by relying on per flow admission control, but suffers from the inherent scalability problems. Therefore it may be applied in the edge of the network (where the number of flows is relatively low), but not likely in the core.
The *DiffServ* architecture can be seen as a hybrid approach between pure provisioning and IntServ: agreements are made for *aggregates* of flows rather than micro-flows, thus solving the scalability problems. However, then the lack of admission control demands adequate bandwidth provisioning, in order to actually realize the QoS requirements. Thus both in the pure provisioning approach and in DiffServ, i.e., the most promising QoS-enabling mechanisms, a prominent role is played by bandwidth provisioning procedures.

Bandwidth provisioning has several other advantages over traffic differentiation mechanisms, see also, e.g., [9]. In the first place, the complexity of the network routers can be kept relatively low, as no advanced scheduling and prioritization capabilities are needed. Secondly, traffic differentiation mechanisms require that the parameters involved are 'tuned well', in order to meet the QoS needs of the different classes – this usually requires the selection of various parameters (for instance: weights in weighted fair queueing algorithms, token bucket parameters, etc.).
Bandwidth provisioning has drawbacks as well: the lack of any QoS differentiation mechanism dictates that all flows should be given the most stringent QoS requirement, thus reducing the efficiency of the network (in terms of maximum achievable utilization). However, it is expected that this effect is mitigated if there is a high degree of aggregation, even in the presence of heterogeneous QoS requirements across users, as argued in, e.g., the introduction of [9] and in [12].
Clearly, the challenge for a network operator is to provision bandwidth such that an appropriate trade-off between efficiency and QoS is achieved: Without sufficient bandwidth provisioning, the performance of the network will drop below tolerable levels, whereas by provisioning too much the performance hardly improves and is potentially already better than needed to meet the

users' QoS requirements, thus leading to inefficient use of resources.

The bandwidth provisioning procedures currently used in practice are usually very crude. A common procedure is to (i) use MRTG [17] to get coarse measurement data (e.g., 5 min. intervals), (ii) determine the average traffic rate during these 5 min. intervals, and (iii) estimate the required capacity by some quantile of the 5 min. measurement data – a commonly used value is the 95% quantile. This procedure is sometimes 'refined' by focusing on certain parts of the day (for instance office hours, in the case of business customers), or by adding safety and growth margins. The main shortcoming of this approach is that it is not clear how the coarse measurement data relates to the traffic behavior at time scales relevant for QoS. More precisely, a crucial question is whether the coarse measurements give any useful information on the capacity needed: QoS degradation experienced by the users may be caused by fluctuations of the offered traffic on a much smaller time scale, e.g., seconds (file transfers, web browsing) or even less (interactive, real-time applications).

Contribution. The goal of our work is to develop accurate and reliable provisioning procedures that require a minimal measurement effort. In particular, we derive an 'interpolation' formula that predicts the bandwidth requirement on relatively short time scales (say the order of 1 sec.), *by using large time scale measurements* (e.g., in the order of 5 min.). In our approach we express QoS in terms of the probability (to be interpreted as fraction of time) that, on a predefined time scale T, the traffic supply exceeds the available bandwidth. The bandwidth C should be chosen such that this probability does not exceed some given bound ε. The time scale T and performance target ε are case-specific: they are parameters of our model, and can be chosen on the basis of the specific needs of the most demanding application involved. We remark that in this setting buffers are not explicitly taken into account; evidently, there is a relation between the time scale T in which the traffic rate exceeds the link rate and the buffer size needed to absorb the excess traffic.

Our approach relies on minimal modeling assumptions. Notably, we assume that the underlying traffic model is Gaussian – empirical evidence for this assumption can be found in e.g., [9,13]. For the special case of peak-rate constrained traffic (peak rate r), we can use (the Gaussian counterpart of) M/G/∞ type of input processes [1], leading to an elegant, explicit formula for the required bandwidth; M/G/∞ corresponds to a flow arrival process that is Poisson with rate λ and flow durations that are i.i.d. as some random variable D (with $\delta := \mathrm{E}D$). We find that, measuring a load $\rho \equiv \lambda\delta r$ (in Mbit/s), the required bandwidth (to meet the QoS criterion) has the form $\rho + \alpha\sqrt{\rho}$. It is clear that the ρ can be estimated by coarse traffic measurements (e.g., 5 or 15 minutes measurements). The α depends on the characteristics of the individual flows, and its estimation requires detailed (i.e., on time scale T) measurements. In many situations, however, there are reasons to believe that

the α is fairly constant in time; the estimate needs to be updated only when one expects that the flow characteristics have changed (for instance due to the introduction of new applications).
Apart from its simplicity, our bandwidth provisioning formula $\rho + \alpha\sqrt{\rho}$ has a number of attractive features. In the first place it is *transparent*, in that the impact of changing the 'QoS parameters' (that is, T and ε) on α is explicitly given. Secondly, the provisioning rule is to some extent *insensitive*: α does not depend on λ, but just on characteristics of the individual flows, i.e., the flow duration D and the peak rate r. This property enables a simple estimate of the additionally required bandwidth if in a future scenario traffic growth is mainly due to a change in λ (e.g., due to growth of the number of subscribers). Furthermore, the analytical expression for α provides valuable insight into the impact of changes in D and r. Our bandwidth provisioning rule has been empirically investigated through the analysis of extensive traffic measurements in various network environments with different aggregation levels, user populations, etc.

Literature. There is a vast body of literature on bandwidth provisioning, see for instance [20]. With respect to traffic modeling, it was empirically shown that Poisson packet arrivals do not accurately capture the dependencies present in network traffic [19]. Gaussian approximations do incorporate these dependencies; their use was advocated in several papers, e.g., [1,9,13,16]. The use of flow level traffic models, like the M/G/∞ model (in which flows arrive according to a Poisson process), is justified in, e.g., [3,4,18]. In [18] it is pointed out that the M/G/∞ traffic model is extremely flexible, in that it allows all types of dependence structures: by choosing the flow durations Pareto-type one can construct long-range dependent traffic, whereas exponential-type flows lead to short-range dependent traffic. The use of M/G/∞ input is also investigated extensively in [2]; this paper also includes the analysis of a number of dimensioning rules.
The study by Fraleigh *et al.* [9] is related to ours, in that it uses bandwidth provisioning based on traffic measurements to deliver QoS. An important difference, however, is that in their case the performance metric is packet delay (rather than our link rate exceedance criterion). Also, in [9] measurements are used to fit the Gaussian model, and subsequently this model is used to estimate the bandwidth needed; this is an essential difference with our work, where our objective is to minimize the required measurement input/effort, and bandwidth provisioning is done on the basis of only coarse measurements. Another closely related paper is [8], where several bandwidth provisioning rules are empirically validated.

Organization. The remainder of this paper is organized as follows. In Section 2 we describe in detail the objectives of this study and the proposed modeling approach; next, we provide the analysis leading to our bandwidth provisioning rule. Numerical results of our modeling and analysis are presented and dis-

cussed in Section 3. In Section 4, the bandwidth provisioning rule is assessed through extensive measurements performed in several operational network environments. Finally, conclusions and topics for further research are given in Section 5.

2 Objectives, modeling and analysis

The typical network environment we focus on in this paper corresponds to an IP network with a considerable number of users generating mostly TCP traffic (from, e.g., web browsing, downloading music and video, etc.). Then the main objective of bandwidth provisioning is to take care that the links are more or less 'transparent' to the users, in that the users should not (or almost never) perceive any degradation of their QoS due to a lack of bandwidth. Clearly, this objective will be achieved when the link rate is chosen such that only during a small fraction of time ε the aggregate rate of the offered traffic (measured on a sufficiently small time scale T) exceeds the link rate. The values to be chosen for the QoS parameters T and ε typically depend on the specific needs of the application(s) involved. Clearly, the more interactive the application, the smaller T and ε should be chosen.

In more formal terms our objective can be stated as follows: The fraction ('probability') of sample intervals of length T in which the aggregate offered traffic exceeds the available link capacity C should be below ε, for prespecified values of T and ε. In other words, with $A(t)$ denoting the amount of traffic offered in $[0, t]$,

$$\Pr(A(T) \geq CT) \leq \varepsilon, \tag{1}$$

For provisioning purposes, the crucial question is: for given T and ε, find the *minimally* required bandwidth $C(T, \varepsilon)$ to meet the target.

In the remainder of this section we derive explicit, tractable expressions for our target probability $\Pr(A(T) \geq CT)$, see (1). We do this for a given traffic input process $\{A(t), t \geq 0\}$; the only explicit assumption imposed is that $\{A(t), t \geq 0\}$ has *stationary increments*, i.e., for any s, t and $u > 0$ we require that the amount of traffic $A(s + u) - A(s)$ arrived in $[s, s + u]$ has the same distribution as the amount of traffic $A(t + u) - A(t)$ arrived in $[t, t + u]$. In other words: the amount of traffic offered in a certain window depends on the window length only, and does *not* depend on the 'position' of the window. This stationarity will likely hold on time-scales that are not too long (up to, say, hours); on longer time-scales there is no stationarity due to day-patterns, and growth (or decline) of the number of subscriptions (time-scale of weeks, months, ...).

Once we have an expression for (an upper bound to) $\Pr(A(T) \geq CT)$, we

can find the minimal C required to make sure that this probability is kept below ε. We thus find the required bandwidth $C(T, \varepsilon)$ – it is expected that this function decreases in both T and ε (as increasing T or ε makes the service requirement less stringent).

2.1 General traffic

Based on the classical Markov inequality $\Pr(X \geq a) \leq (\mathrm{E}X)/a$ for non-negative random variables X, we have, by putting $X = \exp(\theta A(T))$, for $\theta \geq 0$, the upper bound

$$\Pr(A(T) \geq CT) = \Pr\left(e^{\theta A(T)} \geq e^{\theta CT}\right) \leq \mathrm{E}e^{\theta A(T) - \theta CT}.$$

Because this holds for *all* non-negative θ, we can choose the *tightest* upper bound:

$$\Pr(A(T) \geq CT) \leq \min_{\theta \geq 0} \left(\mathrm{E}e^{\theta A(T) - \theta CT}\right). \tag{2}$$

This bound, also known as the *Chernoff bound*, is unfortunately rather implicit, as it involves both the computation of the moment generating function $\mathrm{E}\exp(\theta A(T))$ and an optimization over θ.

Note that $C(T, \varepsilon)$ could be chosen as the smallest number C such that the right hand side of (2) is smaller than ε:

$$C(T, \epsilon) := \min\left\{C : \min_{\theta \geq 0} \left(\mathrm{E}e^{\theta A(T) - \theta CT}\right) \leq \varepsilon\right\}.$$

Rearranging terms, we find that equivalently we are looking for the smallest C such that there is a $\theta \geq 0$ such that

$$C \geq \frac{\log \mathrm{E}e^{\theta A(T)} - \log \varepsilon}{\theta T}.$$

This C is obviously equal to the infimum of the right-hand side over $\theta \geq 0$:

$$C(T, \varepsilon) = \min_{\theta \geq 0} \frac{\log \mathrm{E}e^{\theta A(T)} - \log \varepsilon}{\theta T}. \tag{3}$$

2.2 Explicit formula for Gaussian traffic

Assuming that $A(T)$ contains the contributions of many individual users, it is justified (based on the Central Limit Theorem) to assume that $A(T)$ is *Gaussian* if T is not too small, see e.g., [9,13]. In other words $A(T) \sim \mathrm{Norm}(\rho T, v(T))$, for some load ρ (in Mbit/s), and variance $v(T)$ (in Mbit^2).

For this Gaussian case we now show that we can determine the right hand side of (3) explicitly.

The first step is to compute the moment generating function involved (this is done by isolating the square):

$$\mathrm{E}e^{\theta A(T)} = \exp\left(\theta\rho T + \frac{1}{2}\theta^2 v(T)\right).$$

The calculation of the minimum in (3) is now straightforward:

$$C(T,\varepsilon) = \rho + \min_{\theta\geq 0}\left(\frac{\frac{1}{2}\theta v(T)}{T} - \frac{\log\varepsilon}{\theta T}\right) = \rho + \frac{1}{T}\sqrt{(-2\log\varepsilon)\cdot v(T)}; \qquad (4)$$

the minimum is attained at $\theta = \sqrt{(-2\log\varepsilon)/v(T)}$.
Evidently, $C(T,\varepsilon)$ can also be found by first computing the Chernoff bound for Gaussian traffic

$$\Pr(A(T)\geq CT) \leq \exp\left(-\frac{1}{2}\frac{(C-\rho)^2T^2}{v(T)}\right); \qquad (5)$$

then it is easily checked that (4) is the smallest C such that (5) is below ε.
As for any input process with stationary increments $v(\cdot)$ cannot increase faster than quadratically (in fact, a quadratic function $v(\cdot)$ corresponds to perfect positive correlation), $\sqrt{v(T)}/T$ is decreasing in T, and hence also the function $C(T,\varepsilon)$ – the longer T, the easier it is to meet the QoS requirement. Also, the higher ε, the easier it is to meet the requirement, which is reflected by the fact that the function decreases in ε.

Remark 1: effective bandwidth. There is some reminiscence between formula (4) and the *effective bandwidth* concept proposed earlier in the literature, see, e.g. [6], [7], [11], but there are major differences as well. One of the key attractive properties of effective bandwidths is their 'additivity': if there are two sources, both are assigned a bandwidth, parametrized by the QoS-criterion, such that their *sum* represents the bandwidth needed by their superposition:

$$C_{1+2}(\varepsilon) = C_1(\varepsilon) + C_2(\varepsilon).$$

Importantly, it can be argued that interpreting (4) as an effective bandwidth would lead to a bandwidth allocation *too pessimistic*: noting that

$$\sqrt{v_1(T)+v_2(T)} \leq \sqrt{v_1(T)} + \sqrt{v_2(T)},$$

the amount of bandwidth to be provisioned for the aggregate input could be substantially less than the sum of the individually required bandwidths. $\diamond$

Remark 2: equivalent capacity from Guérin et al. [10]. We remark

that expression (4) is of the same spirit as the 'Gaussian' equivalent capacity formula advocated in [10], but some remarks need to be made.

- *Time scale.* The (classical) formula proposed in [10] is of the form

$$C(\varepsilon) = \rho + \sqrt{-2\log\varepsilon - \log(2\pi)} \cdot \sigma, \tag{6}$$

where σ^2 is the variance of the 'instantaneous traffic rate' R:

$$\sigma^2 := \operatorname{Var} R = \lim_{T\downarrow 0} \operatorname{Var}\left(\frac{A(T)}{T}\right) = \lim_{T\downarrow 0} \frac{v(T)}{T^2}.$$

Hence, $C(\varepsilon)$ as derived in [10] relates to the time scale $T = 0$, and is in this sense less general than our $C(T, \varepsilon)$. We remark that for many Gaussian processes σ does not exist; think of fractional Brownian motion with $H < 1$.

- *Exceedance probability:* [10]*'s approximation vs. Chernoff bound.* It is easily verified that (6) essentially relies on the approximation

$$\Pr(R > C) \approx \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(C-\rho)^2}{\sigma^2}\right); \tag{7}$$

the actual value of $\Pr(R > C)$ is the (complementary) Gaussian distribution function

$$\Pr(R > C) = \int_C^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\rho)^2}{\sigma^2}\right) \mathrm{d}x.$$

In [10] it is reported that $C(\varepsilon)$ – based on the approximation (7) of the exceedance probability $\Pr(R > C)$ – is 'a good approximation' of the equivalent capacity, but no (mathematical) motivation was given. Relying on the approximation $\Pr(N > x) \sim x^{-1}(2\pi)^{-1/2}\exp(-x^2/2)$, where $N \sim \mathrm{Norm}(0,1)$, we find that $\Pr(R > C)$ reads

$$\int_C^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\rho)^2}{\sigma^2}\right) \mathrm{d}x \approx \left(\frac{\sigma}{C-\rho}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(C-\rho)^2}{\sigma^2}\right).$$

Hence, if $C - \rho \approx \sigma$, then we indeed find (7). However, we could not find a rationale for $C - \rho$ being of the same order as σ. In fact, we could construct cases in which (7) is extremely optimistic, in that it substantially *underestimates* $\Pr(R > C)$. Hence, its use is not appropriate for provisioning purposes.

This motivates why our approach above uses the (provably conservative) Chernoff bound (5) rather than approximations of the type of (7). Also, numerical experiments indicated that there is usually just a modest difference between the capacities based on the Chernoff bound and capacities based on inversion of the (complementary) Gaussian distribution function, where, evidently, the former are always conservative. ◇

The formula for $C(T,\varepsilon)$ indicates that, given that we are able to estimate the load ρ and the variance $v(T)$ on the 'advertised' time scale T, we have found a straightforward provisioning rule. In the next subsection, we focus on the special case of (the Gaussian counterpart of) M/G/∞ input; in that (still quite general) case the expressions simplify further.

2.3 M/G/∞ *traffic*

Whereas the above provisioning formula holds for general Gaussian traffic, we now focus on an important sub-class: Gaussian traffic that has the variance function of the M/G/∞ input process, also called the *Gaussian counterpart* of the M/G/∞ input process [1]. In the M/G/∞ input model, jobs arrive according to a Poisson process with rate λ, and stay in the system during a period that is distributed as the random variable D (i.i.d.). While in the system they generate traffic at rate r. Hence, $\rho = \lambda\delta r$, with $\delta = \mathrm{E}D$. Notice that the M/G/∞ traffic model is particularly appropriate in scenarios in which a peak-rate limitation is imposed, see also [1,18]. As we will see later, by choosing D appropriately, it covers a broad range of correlation structures.

Denote by f_X and F_X the density and the distribution function, respectively, of the random variable X. Let D^r be the residual distribution of D, i.e., $1 - F_D(x) = \delta f_{D^r}(x)$. Now the variance of $A(T) = r\int_0^T N(t)\mathrm{d}t$ (with $N(t)$ denoting the number of flows present at time t) can be explicitly calculated by distinguishing between the flows that were already present at time 0, and those entering in $(0,T]$. As was derived in [14,15], for M/G/∞ input,

$$\begin{aligned} v(T) = {} & \lambda r^2\delta\left(\int_0^T x^2 f_{D^r}(x)\mathrm{d}x + T^2(1-F_{D^r}(T))\right) \\ & + \lambda r^2\int_0^T\int_u^T (x-u)^2 f_D(x-u)\mathrm{d}x\mathrm{d}u + \lambda r^2\int_0^T (T-u)^2(1-F_D(T-u))\mathrm{d}u. \end{aligned}$$

This can be simplified further to

$$\lambda r^2\left(2T\int_0^T x(1-F_D(x))\mathrm{d}x - \delta\int_0^T x^2 f_{D^r}(x)\mathrm{d}x + \delta T^2(1-F_{D^r}(T))\right). \quad (8)$$

Hence, cf. (4), the required bandwidth $C(T,\varepsilon)$ can be expressed as:

$$C(T,\varepsilon) = \rho + \alpha\sqrt{\rho}. \quad (9)$$

Importantly, α depends exclusively on the characteristics of the individual flows, i.e., the distribution of the flow duration D and the peak rate r (and QoS requirements T and ϵ), but does *not* depend on the flow arrival rate λ – this will turn out to be a key property in our experimental investigations on

provisioning that are presented in Section 4. Now we evaluate (8) for different distributions D, covering both the long-range dependent and short-range dependent case.

2.3.1 Exponential flow durations

For exponentially distributed flow lengths D, the variance $v(T)$ reads

$$v(T) = 2\rho\delta^2 r(e^{-T/\delta} - 1 + T/\delta),$$

such that

$$\alpha = \left(\frac{T}{\delta}\right)^{-1} \sqrt{(-2\log\varepsilon)\cdot 2r(e^{-T/\delta} - 1 + T/\delta)}.$$

Observe that $v(T)$ is, for T large, linear, corresponding to short-range dependent input. Also observe, that α depends on T only through the ratio T/δ.

2.3.2 Pareto flow durations

For Pareto-distributed flow lengths D, i.e., obeying

$$F_D(x) = 1 - \left(\frac{b}{x+b}\right)^a, \quad x \geq 0, \tag{10}$$

and $\delta = b/(a-1)$ (where $a > 1$ and $b > 0$), substantial calculus gives (assume for ease $a \neq 2$, $a \neq 3$)

$$v(T) = \frac{2\rho r}{(3-a)(2-a)} \cdot (b^{a-1}\left(T+b\right)^{3-a} - (3-a)bT - b^2\Big);$$

$$\alpha = \frac{1}{T}\sqrt{(-2\log\varepsilon)\cdot\frac{2r}{(3-a)(2-a)}\cdot(b^{a-1}(T+b)^{3-a} - (3-a)bT - b^2)}.$$

(Notice that [1] uses $F_D(x) = 1 - (b/x)^a$ for $x \geq b$, but, as flow sizes do not obey some natural lower bound b, we have chosen to use the more natural 'shifted version' (10) instead.) If $a < 2$, $v(T)$ grows 'superlinearly' for large T (in fact, it grows as T^{3-a}), corresponding to long-range dependent input; for $a > 2$, we see that $v(T)$ is essentially linear, cf. [5].

2.3.3 Discussion on the M/G/∞ input model

1. If T is small (i.e., small compared to δ), then α becomes insensitive in the flow duration D. This can be seen as follows. From (8) it can be derived that $v(T)/T^2 \to \rho r$ if $T \downarrow 0$. Then (4) yields $C(T,\varepsilon) \approx \rho + \sqrt{(-2\log\varepsilon)\cdot\rho r}$, exclusively depending on ρ, for T small.

This result can be derived differently, by noting that for $T \downarrow 0$, the performance criterion boils down to requiring that the number of active users does not exceed C/r. It is well-known that the number of active users has a Poisson distribution with mean $\lambda\delta$; this explains the insensitivity.

2. The case of exponential flow lengths can be easily extended to, e.g., *hyper-exponentially* distributed flows; a random variable X is hyperexponentially distributed [21, p. 446] if with probability $p \in (0, 1)$ it is distributed exponentially with mean δ_1, and else exponentially with mean δ_2. Then the hyperexponential case is just the situation with two flow types feeding independently into the link (each type has its own exponential flow length distribution); note that the variance of the total traffic is equal to the sum of the variances of the traffic generated by each of the different exponential flow types.
3. The above approach assumes that traffic arrives as 'fluid': it is generated at a constant rate r. It is perhaps more realistic to assume that, during the flow's 'life time', traffic arrives as a Poisson stream of packets (of size s); the rate of the Poisson process is γ, where γs is equal to r. Denoting the above, fluid-based, variance function by $v_{\mathrm{f}}(T \mid r)$, and the packet-based variance function by $v_{\mathrm{p}}(T \mid \gamma, s)$, it can be verified that

$$v_{\mathrm{p}}(T \mid \gamma, s) = v_{\mathrm{f}}(T \mid r) + \rho s T, \tag{11}$$

irrespective of the flow duration distribution. Importantly, the provisioning formula $C(T, \varepsilon) = \rho + \alpha\sqrt{\rho}$ remains valid (for an α that does not depend on λ).

3 Numerical results

This section presents numerical results obtained by using the analytical model of Section 2. The goal is to illustrate a few key features of our bandwidth provisioning formula. We use the traffic parameters and QoS parameters displayed in Table 1, unless specified otherwise.

Experiment 1: Fluid model vs. packet-level model

Figure 1 shows the required capacity obtained by the packet-level and fluid model as a function of T, for various mean flow durations δ. It is seen that for large values of the time scale T, both models obtain the same required capacity. This can be understood by looking at the extra term $\rho s T$ of (11), which influence on $C(T, \varepsilon)$ becomes negligible for increasing T, cf. (4).
For $T \downarrow 0$ the required capacity obtained by the packet-level model behaves

Table 1
Default parameter settings for the numerical results.

TRAFFIC MODEL			QOS PARAMETERS		
δ	1	sec	T	1	sec
distr. of D	exp.		ε	0.01	-
ρ	10	Mbit/s			
model	fluid				
FLUID MODEL			PACKET-LEVEL MODEL		
r	1	Mbit/s	γ	83.3	Packet/s
			s	1500	Bytes

like

$$C(T,\varepsilon) \sim \rho + \sqrt{\rho s}\frac{1}{\sqrt{T}}\sqrt{-2\log\varepsilon}$$

and hence $C(T,\varepsilon) \uparrow \infty$ as $T \downarrow 0$, whereas the required capacity of the fluid model converges to $\rho + \sqrt{(-2\log\varepsilon)\cdot\rho r}$, as was already argued in Section 2.3.3.

The fast increase in the required capacity in the packet-level model for a decreasing time scale T was also observed in e.g., [9]. Note that, in fact, the required capacity is not influenced by the *absolute value* of T, but rather by the *ratio* of T/δ. The right graph of Figure 1 shows the same results as the left graph, but now on a linear axis and only for $T \in [0,1]$.

In the remainder of this section we restrict ourselves to the flow-level model, as we will focus on situations with values of $T/\delta > 0.1$ for which the required capacity is almost identical in both models.

Experiment 2: Impact of the flow duration distribution

Next we investigate the impact of the flow duration distribution on the re-

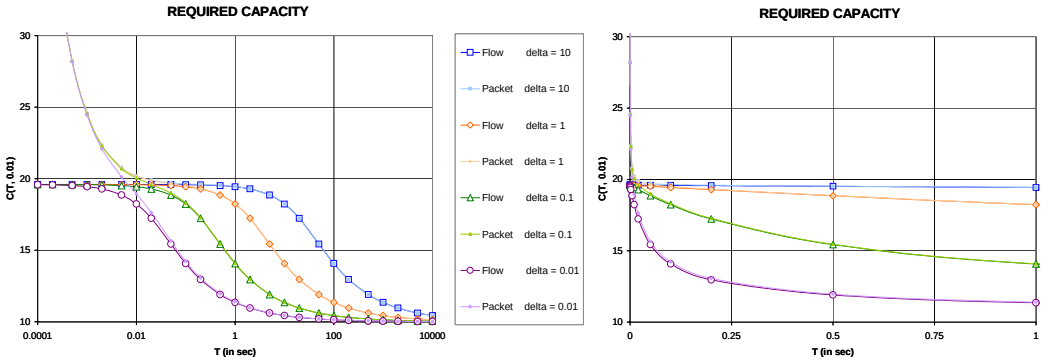

Fig. 1. Comparison of the required capacity for the flow-level and packet-level model as a function of the time scale T. Left: logarithmic axis. Right: linear axis.

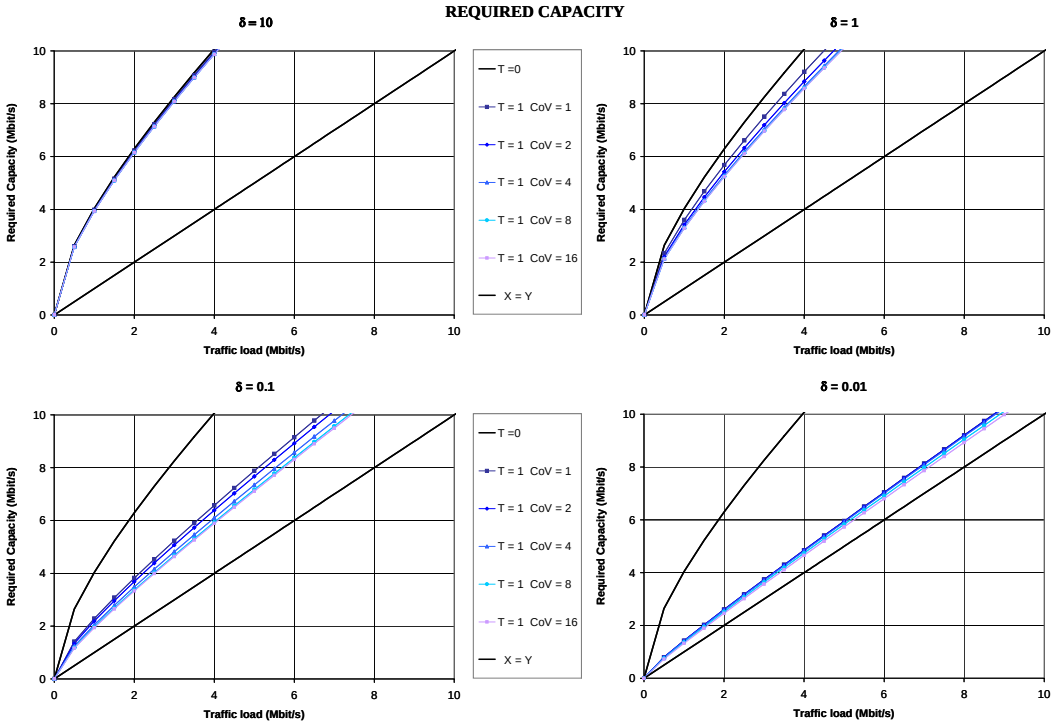

Fig. 2. Required capacity for hyper-exponential flow durations with different means and CoVs.

quired capacity. Figure 2 contains four graphs with results for hyperexponentially distributed flow durations D. Each graph shows, for a particular value of the mean flow size δ, the required capacity as a function of the offered load ρ, for different Coefficients of Variation (CoV) of D. These graphs show that the required capacity is almost insensitive to the CoV for the long ($\delta = 10$ sec.) and short flow durations ($\delta = 0.01$ sec.). For the other cases ($\delta \in \{0.1, 1\}$ sec.) the required capacity is somewhat more sensitive to the CoV. The graphs show that for hyperexponentially distributed flow durations *less* capacity is required if the CoV increases.

It should also be noticed that the required capacity for $T = 0$, also shown in Figure 2, corresponds to the often used M/G/∞ bandwidth provisioning approach, cf. the discussion in Section 2.3.3 and the discussion on Experiment 1. The numerical results show that particularly for short flow durations significantly less capacity is required than suggested by the classical M/G/∞ approach; for longer flows this effect is less pronounced.

Experiment 3: Impact of QoS parameter ε

Figure 3 shows the required capacity as a function of the QoS requirement ε, which specifies the fraction of intervals in which the offered traffic may exceed the link capacity. A larger value of ε means relaxing the QoS requirement, and hence less capacity is needed. Obviously, for $\varepsilon \to 1$ the required capacity

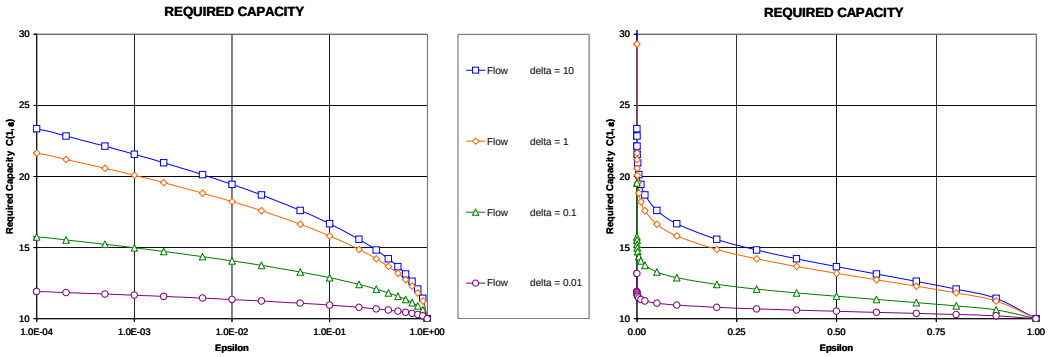

Fig. 3. Comparison of the required capacity for the flow-level and packet-level model as a function of the QoS requirement T. Left: logarithmic axis. Right: linear axis.

converges to the long term average load $\rho = 10$. For $\varepsilon \downarrow 0$ the required capacity increases rapidly to infinity (according to $\sqrt{-2\log\varepsilon}$).

Experiment 4: Impact of the CoV of the flow duration distribution

To investigate the impact of the flow duration characteristics, we computed the required capacity for exponential, hyperexponential, and Pareto distributed flow durations with different CoV values, see the left panel of Figure 4. The graph shows that the required capacity is almost insensitive to the flow duration distribution. The left graph also confirms the earlier observations that the capacity is almost insensitive to the CoV of the flow duration distribution. Note, that for hyperexponentially distributed flow durations the required capacity slightly decreases for increasing CoV, while for Pareto distributed flow durations the required capacity slightly increases for increasing CoV.

Experiment 5: Impact of the access rate

Finally, the right panel of Figure 4 studies the effect of the access rate r on the required capacity. Three values of the access rate r and the mean flow duration

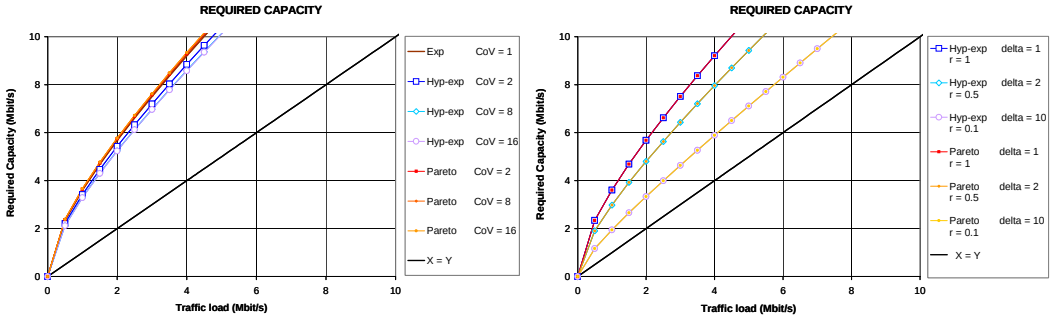

Fig. 4. Left: required capacity for different flow duration distribution and CoVs. Right: required capacity for different access rates.

δ are chosen such that the mean flow size $\delta \cdot r$ remains constant. As expected, the required capacity increases considerably when r becomes larger (i.e. the traffic burstiness grows). The results in this graph for hyperexponential and Pareto flow sizes confirm the conclusions from Experiments 2 and 4 that the required capacity is almost insensitive to the flow duration distribution.

4 Experimental verification & bandwidth provisioning

In this section we will analyze measurement results obtained in operational network environments in order to validate the modeling approach and bandwidth provisioning rule presented in Section 2. In particular, we will investigate the relation between measured traffic load values $\hat{\rho}$ during 5 min. periods (long enough to assume stationarity) and the traffic fluctuations at a 1 sec. time scale within these periods.
Clearly, if (A) our M/G/∞ traffic modeling assumptions of Section 2.3 apply *and* if (B) differences in the load ρ are caused by changes in the flow arrival rate λ (i.e., the flow size characteristics remain unchanged during the measurement period), then, as a function of ρ, for given (T, ε), the required bandwidth C_ρ should satisfy

$$C_\rho = \rho + \alpha\sqrt{\rho}, \tag{12}$$

for some fixed value of α. To assess the validity of this relation, we have carried out measurements in three different network environments: (i) a national IP network providing Internet access to residential ADSL users, (ii) a college network, and (iii) a campus network.
In the ADSL network environment the main assumptions made in Section 2.3 in order to justify use of the M/G/∞ traffic model seem to be satisfied, i.e., the flow peak rates are limited due to the ADSL access rates (which are relatively small compared to the network link rates), and the traffic flows behave more or less independently of each other (the IP network links are generously provisioned and, hence, there hardly is any interaction among the flows). The other network environments have essentially different characteristics. In particular, in the college and campus network the ratio of the access rate and link rate is relatively large, which, obviously, may lead to violation of our traffic modeling assumptions.
The measurement scenarios and results will be described and discussed in more detail in the following subsections.

4.1 ADSL *network environment*

We first focus on the ADSL network environment with residential users, see Figure 5. An ADSL connection consists of an ADSL modem on both sides of

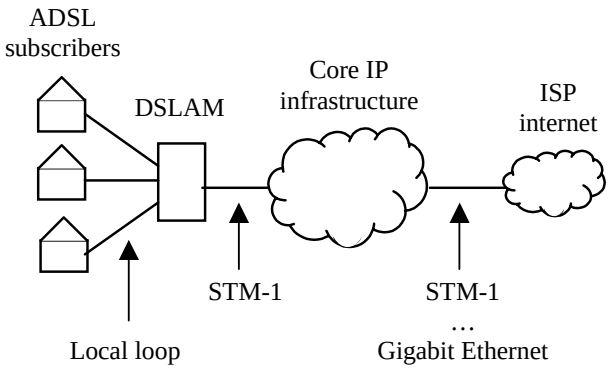

Fig. 5. Overview of ADSL infrastructure

the local loop between the subscriber and the local exchange. On the local exchange side, up to 500 modems are contained in Digital Subscriber Line Access Multiplexers (DSLAM).
The DSLAMs are connected to the core IP infrastructure by means of optical STM-1 (155 Mbit/s) links. The aggregated traffic of all the ADSL subscribers of a certain Internet Service Provider (ISP) is carried over a high-capacity link between the core infrastructure and the ISP. Depending on the size of the ISP, this can vary between a single STM-1 link and multiple Gigabit Ethernet links. At the time of the measurements, none of the network links were saturated, and hence the traffic was not affected by any shortage of capacity in the ADSL network.

We choose the sample size $T = 1$ sec., motivated by the fact that this can be assumed to be the time scale that is most relevant for the Quality of Service perception of end-users of typical applications like web browsing. Elementary transactions, such as retrieving single web pages, are normally completed in intervals roughly in the order of 1 sec. If the network performance is seriously degraded during one or several seconds, then this will affect the quality as perceived by the users.

The method that was chosen to measure the traffic at the 1 sec. time scale was to use the internal traffic counters (interface MIBs) of the DSLAMs. These counters keep track of the accumulated number of bytes that are transported on each port in each direction. In this experiment, the counters for the STM-1 ports in the downstream direction (toward the subscribers) were used. The counters were read-out using SNMP.
The measurements were done during several evenings (between 5 PM and 11 PM), as this is the busiest period of the day for ADSL traffic. The measurements were performed on a large number of DSLAMs, in locations ranging from small villages to major cities.
Time was split into 5 minute chunks, over which the load ρ is determined for each STM-1 link. In addition, for each 5 min. period, the 99% quantile of the 1 sec. measurements was determined. This quantile was assumed to

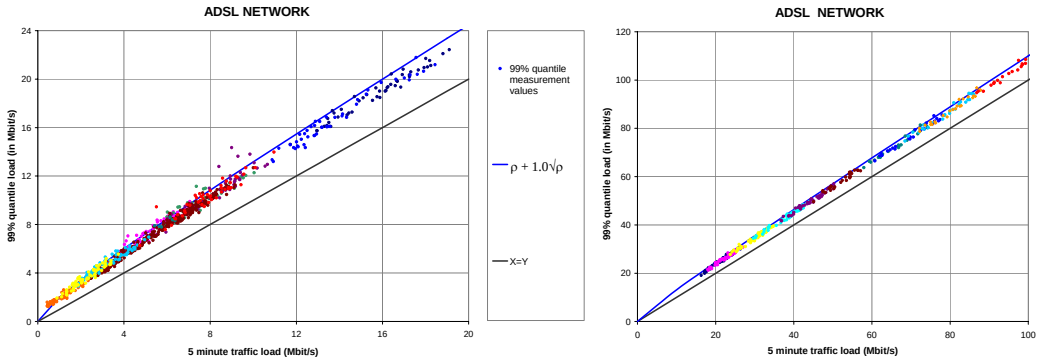

Fig. 6. Left: 99% maximum 1 sec. traffic as function of 5 min. traffic mean. Right: synthesized traffic measurements for higher traffic volumes.

indicate the minimum capacity C that is needed to fulfill the QoS requirement $\Pr(A(T) \geq CT) \leq \varepsilon$, with $\varepsilon = 1\%$ and $T = 1$ sec.

The left graph in Figure 6 results from measurements on 11 STM-1 links at various locations. Each location is represented by a distinct color. For orientation purposes, the dotted line shows the unity ($y = x$) relation. It is remarkable how the 99% quantiles almost form a solid curve. We fitted a function $\rho + \alpha\sqrt{\rho}$, such that roughly 95% of the 99% quantiles are lower or equal to this function. The reason for fitting an upper bound, instead of finding the function that gives the minimum least square deviation, is that eventually we intend to use this function for capacity planning: then it is better to *over*estimate the required bandwidth than to underestimate it. The graph shows an extremely nice fit for the function $C = \rho + 1.0\sqrt{\rho}$ (with C and ρ expressed in Mbit/s).

At the time of the measurements, the busiest STM-1's did not carry more traffic than 20 Mbit/s during the busiest hours, so we could not verify that the found upper bound also holds for higher traffic volumes. To overcome this problem at least partly, we synthesized artificial traffic measurements by taking the superposition of the traffic measured on several (unrelated) STM-1's. The right graph of Figure 6 shows the results of this experiment. As expected on theoretical grounds, the fitted function $C_\rho = \rho + 1.0\sqrt{\rho}$ remains valid.

4.2 College and campus network

We have performed similar experiments in two other network environments, viz. a college network and a campus network, with essentially different characteristics than the ADSL network. In particular, in these alternative network environments the ratio of the access rate and link rate is relatively small, and, hence, one would expect that the M/G/∞ modeling assumption underlying the analysis in Section 2 is not valid anymore. The question is whether (or up to what extent) the bandwidth requirement formula (9) still applies.

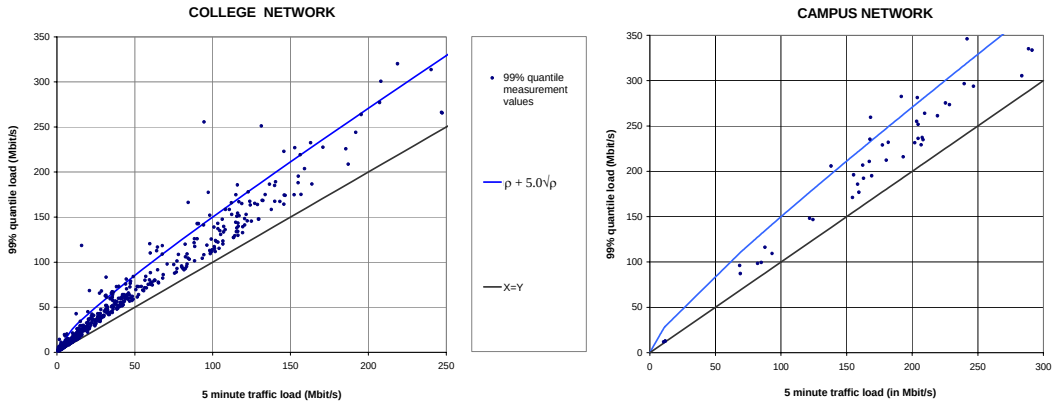

Fig. 7. 99% maximum 1 sec. traffic as function of 5 min. traffic mean. Left: college network. Right: campus residential network.

In the first scenario, we have measured a 1 Gbit/s link connecting a college network to the Internet. This link is shared by about 1000 students and teachers, each having a 100 Mbit/s FastEthernet connection (a ratio of 1:10). In the second scenario, we have measured a 300 Mbit/s (trunked) link connecting an university campus (residential) network to the Internet. This link is shared by some 2000 students, each of them having a 100 Mbit/s connection (a ratio of 1:3). Thus, theoretically, it takes only 10 or 3 users, respectively, to saturate the observed network links.

The left graph of Figure 7 shows the measurement results for the college network. As expected, the cloud of 99% quantiles of the 1 sec. traffic rate samples within 5 min. intervals does not form such a nice 'curve' as in the previous (ADSL) scenario, but the typical square-root behavior can still be recognized.

For the university campus network, the 'peak versus average load' is plotted in the right graph of Figure 7. Note that the traffic in both directions has been aggregated during the measurements, which explains that the link load as plotted in the graph is sometimes higher than the link capacity (which is one-way). Although the number of measurements available for the campus network is relatively low, we conclude from the graph that the relation between the average link loads and the 99% quantiles of 1 sec. samples shows a similar behavior as in the college network.

From the above results it is concluded that, as expected, for these alternative scenarios our model developed in Section 2 does clearly not apply as well as for the ADSL scenario. Indeed, it may be expected that this is caused by the high access link speed, which leads to a possibly high variability in the rate at which traffic is generated by the users in the alternative scenarios (while the M/G/∞ model assumes that sources generate traffic at a fixed rate). Apparently, under these highly variable traffic conditions the 5 min. average traffic rate does

not provide sufficient information to estimate the traffic behavior on much smaller time scales (i.e. more detailed information than just ρ is needed), and, consequently, other underlying traffic models should be applied.

4.3 Bandwidth provisioning procedure

Our formula (9) for the required bandwidth $C(T, \varepsilon)$ can be used to develop bandwidth provisioning procedures. Obviously, a first step in this procedure is to verify whether the main M/G/∞ modeling assumptions are satisfied in the network environment under consideration, such that formula (9) can indeed be applied. A next step is then to estimate ρ and α. Clearly, ρ can be estimated through coarse traffic measurements, as it is just the average load; α, however, contains (detailed) traffic characteristics on time scale T (viz., the variance $v(T)$). In particular, as noticed in Section 2.3, α depends on the flow peak rate r and on the parameters of the flow duration D, but, importantly, α does *not* depend on the flow arrival rate λ; α can be considered as a characteristic of the individual flows.
This 'dichotomy' between ρ and α gives rise to efficient provisioning procedures. Consider the following two typical situations:

- Situations in which there is a set of links, that differ (predominantly) in the *number* of connected users; across the links, the individual users have essentially the same type of behavior (in terms of the distribution D and the access rate r). Then the α can be estimated by performing detailed measurements at (a part of) the existing links. When a new link is connected, one could obtain an estimate $\hat{\rho}$ of the load by performing coarse measurements (e.g., every 5 min., by using the MRTG tool [17]). Then the provisioning rule $\hat{\rho} + \hat{\alpha}\sqrt{\hat{\rho}}$ can be used. An example is the ADSL scenario described above, in which one could use $\hat{\alpha} \approx 1.0$ to dimension a new link.
- Growth scenarios in which it is expected that the increase of traffic is (mainly) due to a growing number of subscribers (i.e., the λ), while the user behavior remains unchanged. Here it suffices to perform infrequent detailed measurements at time scale T, yielding an estimate $\hat{\alpha}$ of α. If a future load $\hat{\rho}$ is envisaged, the required bandwidth can be estimated by the provisioning rule $\hat{\rho} + \hat{\alpha}\sqrt{\hat{\rho}}$.

The estimate of α has to be updated after a certain period (perhaps in the order of months). This should correspond to the time at which it is expected that the 'nature' of the use of resources changes (due to, e.g., new applications, etc.).
The explicit formulas for α derived in Section 2 are also useful when examining the impact of changes in the user behavior or the QoS parameters. For instance, the impact of an upgrade of the access speed r can be evaluated. Also one

could assess the effect of imposing a stronger of weaker performance criterion ε: when replacing ε_1 by ε_2, the α needs to be multiplied by $\sqrt{\log \varepsilon_2 / \log \varepsilon_1}$.

The measurement period of 5 min. mentioned above for estimating the load ρ is motivated by the fact that this is the time scale on which measurements in an operational network can be (and are) performed on a routine basis. A higher frequency would be desirable, but this would put a high load on the processing capacity of routers, the transport capacity of management links, etc., particularly if there are many routers and ports involved. On the other hand, measurements performed at lower frequencies (for instance 1 to several hours) are too coarse, as traffic is not likely to be stationary over such long periods. Therefore, 5 min. will usually be a suitable trade-off, as this is feasible to measure, and at the same time a reasonable period during which the traffic can still be assumed stationary.

5 Concluding remarks & further research

In this paper we have considered bandwidth provisioning for IP network links. Our goal was to develop accurate and reliable provisioning procedures that require a minimal measurement effort. First we derived a formula for the minimally required link bandwidth $C(T, \varepsilon)$, such that the aggregate traffic rate (measured on a time scale T) exceeds the link rate only during a small fraction ε of time. In particular, for the situation that the traffic is generated by peak rate constrained flows that arrive according to a Poisson process and remain active for some random time D (i.e., M/G/∞ input traffic) the resulting bandwidth provisioning rule is of the form $C(T, \varepsilon) = \rho + \alpha\sqrt{\rho}$; here ρ is the traffic load, which can be estimated easily from coarse traffic measurements (typically in the order of 5 min.). Importantly, the coefficient α is determined by characteristics of the individual flows, and does *not* depend on the flow arrival rate λ. We have shown that this property opens up the possibility of elementary (but adequate) provisioning procedures.
The explicit expression of α shows the impact of the flow size, peak rate and other traffic and system parameters on the required link bandwidth. In particular, α lies somewhere between 0 and $\sqrt{(-2\log\varepsilon)r}$; its exact value depends mainly on the ratio of the time scale of interest T and the mean flow duration. Extensive numerical results show that $C(T, \varepsilon)$ is quite insensitive to the flow size distribution (apart from its mean value).

The above provisioning rule has been empirically validated through the analysis of extensive traffic measurements in three practical scenarios: (i) an IP network connecting private and small business ADSL users to the Internet, (ii) a college network and (iii) a campus network. A particularly good correspondence with our theoretical results was found from the measurements in

the IP network scenario, where the flow rates are bounded by relatively small ADSL access rates. The measurement results for the other scenarios showed, as expected, less good correspondence: the M/G/∞ modeling assumptions are not really satisfied there; in particular the flow rates may be strongly variable due to the relatively high access rates (compared to the network link rates) in these scenarios. It remains for further research whether other underlying traffic models could be used to improve the results for network environments like the college and campus network. An attractive alternative traffic model is the fractional Brownian motion (fBm) model, as used in e.g., [9].

Another topic for further research is the investigation of the validity of the stationarity assumption in our modeling approach. In particular, up to which time scale $\hat{t}$ can the traffic arrival process be assumed stationary? It is clear that the estimates of the average rate ρ should be measured at a time scale smaller than $\hat{t}$. It should also be investigated in more detail under which conditions (and to what extent) the Gaussian traffic assumption is valid, cf. Section 2.2.

As a last topic for further research we mention the QoS criterion used in this paper, i.e., the fraction of time ε that the aggregate offered traffic rate (measured at time scale T) is restricted by the link rate. In particular, we could obtain more insight in the relation between this QoS criterion, which we used as link bandwidth provisioning objective, and the actual QoS that the users are offered. In other words: to what extent does this criterion actually determine the *duration* of a congestion period (this will depend on the traffic characteristics, in particular the flow-level dynamics)? What are appropriate choices of T and ε for different (TCP) application types (file downloading, interactive web browsing, etc.)?

Acknowledgements

This work was (partly) carried out within the EQUANET project, an 'ICT-doorbraakproject' which is supported by the Dutch Ministry of Economic Affairs via its agency Senter.

References

[1] R. ADDIE, P. MANNERSALO, and I. NORROS. Most probable paths and performance formulae for buffers with Gaussian input traffic. European Transactions on Telecommunications, Vol. 13, pp. 183–196, 2002.

[2] R. Addie, T. Neame, and M. Zukerman. Performance evaluation of a queue fed by a Poisson Pareto burst process. Computer Networks, Vol. 40, pp. 377–397, 2002.

[3] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, and J.W. Roberts. Statistical bandwidth sharing: a study of congestion at flow-level. *Proceedings* sigcomm *'01*, San Diego CA, USA, 2001.

[4] T. Bonald and J.W. Roberts. Congestion at flow level and the impact of user behaviour. Computer Networks, Vol. 42, pp. 521-536, 2003.

[5] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. ieee/acm Transactions on Networking, Vol. 5, pp. 835–846, 1997.

[6] A. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. IEEE/ACM Transactions on Networking, Vol. 1, pp. 329–343, 1993.

[7] A. Elwalid, D. Mitra, and R. Wentworth. A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM Node. IEEE Journal on Selected Areas in Communications, Vol. 13, pp. 1115–1127, 1995.

[8] M. Fiedler and A. Arvidsson. A Resource Allocation Law to Satisfy QoS Demands on ATM Burst and Connection Level. Technical Report 06, cost-257, 1999.

[9] C. Fraleigh, F. Tobagi, and C. Diot. Provisioning ip backbone networks to support latency sensitive traffic. *Proceedings* ieee infocom *2003*, San Francisco CA, USA, 2003.

[10] R. Guérin, H. Ahmadi, and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks. ieee Journal on Selected Areas in Communications, Vol. 9, pp. 968–981, 1991.

[11] F. Kelly. Notes on effective bandwidths. In: *Stochastic Networks: Theory and Applications*, eds. F. Kelly, S. Zachary, and I. Ziedins, 1996.

[12] F. Kelly. Mathematical modeling of the Internet. *Mathematics Unlimited - 2001 and Beyond* (editors B. Enquist and W. Schmidt), Springer-Verlag, Berlin, Germany, 2001.

[13] J. Kilpi and I. Norros. Testing the Gaussian approximation of aggregate traffic. *Proceedings Internet Measurement Workshop*, Marseille, France, 2002. Available at url `http://www.vtt.fi/tte/rd/traffic-theory/papers/`

[14] M. Mandjes, I. Saniee, and A. Stolyar. Load characterization, overload prediction, and load anomaly detection for voice over ip traffic. *Proceedings 38th Allerton Conference,* Urbana-Champaign, US, pp. 567–576, 2000.

[15] M. Mandjes and M. van Uitert. Sample-path large deviations for tandem and priority queues with Gaussian inputs. CWI report PNA-R0221. Available at URL `http://www.cwi.nl/ftp/CWIreports/PNA/PNA-R0221.pdf`, 2002. Accepted for publication in Annals of Applied Probability.
A short version appeared in *Proceedings* ITC *18*, pp. 521-530, Berlin, Germany, 2003.

[16] I. Norros. On the use of Fractional Brownian Motion in the theory of connectionless networks. IEEE Journal on Selected Areas in Communications, Vol. 13, pp. 953–962, 1995.

[17] T. Oetiker. *MRTG: Multi Router Traffic Grapher.* Available at URL `http://people.ee.ethz.ch/~oetiker/webtools/mrtg/`.

[18] M. Parulekar and A. Makowski. M/G/∞ input processes: a versatile class of models for network traffic. *Proceedings* IEEE Infocom, pp. 419–426, 1997.

[19] V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modeling. IEEE/ACM Transactions on Networking, Vol. 3, pp. 226–244, 1995.

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

[21] H.C. Tijms. *A first course in stochastic models.* Wiley, New York, 2003.