Abstract— In this paper we first construct a common classification guideline for adaptive provisioning schemes by introducing the feature tree concept. Afterward, we propose QoS-aware provisioning schemes that combine the use of the Gaussian traffic model, traffic measurements, and traffic predictions. In parallel, we give an overview on the recent adaptive provisioning schemes. We perform extensive numerical and simulative investigations to show that our proposed QoS-aware schemes outperform some previously available ones from the aspect of efficiency. Moreover, we point out the practical tradeoff between using buffered and bufferless constraints in QoS-aware schemes.

I. INTRODUCTION

Adaptive bandwidth provisioning currently emerges as a promising solution in practice for providing QoS. The term “adaptive” means that the bandwidth is dynamically allocated with regard to the live traffic dynamics. To this end, most of the work simply uses the link utilization as a basic factor for provisioning without specifying the role of the utilization level on QoS perception. Making one step further, we address in this paper the issue of QoS-aware adaptive provisioning by elaborating novel provisioning schemes with explicit respect to statistical, packet level QoS requirements.

In order to exploit the most retrievable benefits, we should also have a clear view on the specific features and performability of the currently available provisioning schemes. This is the motivation of the second main issue we tackle in this paper. Towards a better understanding of provisioning schemes, we discuss the superiority of QoS-aware provisioning schemes over the rest and we provide insights into the operation of QoS-aware schemes.

With the proposed novel provisioning schemes and the presented analysis guideline in this paper, we envision a wide use of QoS-aware provisioning towards Internet QoS. Some typical applicabilities are, for example, i) adaptive resizing of high-speed LSPs (Label Switched Paths) in MPLS networks; ii) adaptive resizing of customer-pipes in VPNs (Virtual Private Networks); and iii) adaptive bandwidth allocation of logical links in the Service Overlay Networks architecture [1] for providing end-to-end QoS over inter domains.

The paper is organised as follows. In Section II, we present the concept of the feature tree for the classification of provisioning schemes. Afterwards, in Section III, we propose novel QoS-aware provisioning schemes that are based on the Gaussian traffic model, periodical traffic measurements and traffic prediction. In Section IV, we review other provisioning schemes currently available in practice. In Section V, we provide insights into the operation of the QoS-aware schemes by examining in detail the issue of using bufferless or buffered constraints. Finally, we conclude the paper in Section VI.

II. CLASSIFICATION OF ADAPTIVE PROVISIONING SCHEMES

To make a classification of a given adaptive provisioning scheme, we can rely on a feature tree presented in Fig. 1. In this feature tree, the main features of provisioning schemes are sorted in levels. A common feature at a higher level can be followed by different features in a successive lower level, resulting in different provisioning schemes.

- Level 1: Tracing traffic dynamics. This is a common point every adaptive scheme should be able to achieve. Consequently, it can be considered as a root of the feature tree. The reason behind this is that bandwidth adjustments only have a tangible sense if they reflect truly the effect of the traffic dynamics that should, therefore, be traced. In most cases, the traffic dynamics are captured by using the monitoring capability of routers. The most usual and simplest way is to monitor the average traffic load (either per-class, or aggregate load) in successive measurement intervals. The bandwidth is updated periodically after a certain number of measurement intervals.

- Level 2: QoS respectability. Regarding the issue of QoS, a given provisioning scheme can be either QoS-aware or QoS-unaware. The schemes in the former group consider an explicit target QoS (e.g. packet loss and packet delay) as a basic criteria for provisioning. Very often, these schemes imply some analytical models to derive the needed bandwidth amount from the desired target QoS.
The schemes in the QoS-unaware group do provisioning regardless of QoS. This is understood in the sense that no explicit target QoS is given and the update rule in general is very simple requiring no sophisticated add-on considerations like traffic model, QoS parameter derivation.

- Level 3: Traffic Prediction. A given provisioning scheme can either imply traffic prediction or not. Traffic predictions are usually believed to improve the performance of the provisioning schemes. Of course, the benefit comes at the expense of more complicated analysis which is a tradeoff.

In the rest of the paper, we consider a provisioning task for a link that conveys traffic with high aggregation level. We adopt a common time division principle for all the schemes as follows. The traffic load is measured in consecutive basic time slots called measurement slots. A certain number of measurement slots constitutes a resizing window. Decision and execution, if necessary, concerning bandwidth updates are initiated at the end of each resizing window.

III. THE PROPOSED, MVA BOUND BASED PROVISIONING SCHEMES

Suppose that we have to guarantee the packet level constraint \( P(\text{delay} > D) < \epsilon \), where \( D \) and \( \epsilon \) are the given delay bound (excluding the propagation delay) and violation probability, respectively. In case of high link capacity \( c \), the packet transmission delay is negligible, therefore this delay constraint is equivalent with the constraint on the queue tail probability \( P(q > Dc) < \epsilon \), where \( q \) stands for the queue length. The basic idea now is to use a Gaussian process as the model for the aggregate traffic. Two reasons justify this choice. First, this is a reasonable model for traffic with a high degree of aggregation. Second, using this traffic model, an asymptotic bound called maximum variance asymptotic (MVA) bound [2] on the queue tail probability is provided.

Let us consider a single discrete time queue with the aggregate input rate \( \lambda_n \) at time \( n \) and service rate \( c \). Define a stochastic process \( X_n \) as \( X_n = \sum_{i=1}^{n} \lambda_i - cn \). Let \( C_\lambda(l) \) be the autocovariance functions of \( \lambda_n \) and define \( \bar{\lambda} = \mathbb{E}\{\lambda_n\} \), \( \sigma^2 = \mathbb{V}ar\{\lambda_n\} = C_\lambda(0) \). The mean and variance of \( X_n \) can be computed from the mean and the autocovariance function of the input rate as

\[
\mathbb{E}\{X_n\} = n(\bar{\lambda} - c) \tag{1}
\]

\[
\mathbb{V}ar\{X_n\} = nC_\lambda(0) + 2 \sum_{l=1}^{n-1} (n - l)C_\lambda(l) \tag{2}
\]

Assuming that \( \lambda_n \) is a Gaussian process, then the MVA bound on the queue tail probability is given as

\[
Pr(q > x) = e^{-\mu_x/2}, \quad \text{where} \quad \mu_x := \frac{1}{\max_{n \geq 1} \sigma^2 \bar{x}^2}, \quad \text{and} \quad \sigma^2 \bar{x}^2 := \frac{\mathbb{V}ar\{X_n\}}{(x - \mathbb{E}\{X_n\})^2}. \tag{3}
\]

Denote the traffic rate measured in slot \( i \) of window \( j \) by \( y_{ij}^{(j)} \). When one resizing window consists of \( N \) measurement time slots, we compute the mean and covariance functions of traffic over a given window \( j \) as \( m_j = \frac{\sum_{i=1}^{N} y_{ij}^{(j)}}{N} \), and \( C_j(k) = \frac{1}{k} \sum_{i=1}^{N-k} (y_{ij}^{(j)})(y_{i+k}^{(j)} - m_j) \) for \( k = 0, 1, \ldots, N - 1 \). The computed quantities enable us to capture the mean and variance of the accumulated traffic process \( X_n \) by using equations (1) and (2) and in turn to calculate the MVA bound using (3).

Laying on the common above principle, we specify the following provisioning schemes that are distinguished from each other by the fact whether traffic prediction is involved and if yes what is the traffic prediction rule.

A. PS1: MVA bound based scheme with prediction.

In this scheme we propose to perform traffic prediction at the end of each resizing window. We predict both the mean rate of the aggregate traffic and the variance of the cumulative process \( X_n \). We opt the exponential smoothing (ES) technique for the prediction due to its proven stability and suitability on trend prediction [3]. Formally, for the resizing window \( j + 1 \) we predict

\[
m^*_j + 1 = m_j + (1 - w)m^*_j, \tag{4}
\]

where \( w \) is the weighting parameter (\( 0 \leq w \leq 1 \)), \( m^*_j \) and \( m_j \) are the predicted and measured value of the mean rate for the resizing window \( j \), respectively. Similarly, for the variance of the accumulated traffic, we predict

\[
V ar^*(X_{n,j+1}) = w \cdot V ar\{X_{n,j}\} + (1 - w) \cdot V ar^*\{X_{n,j}\} \tag{5}
\]

where \( V ar^*\{X_{n,j}\} \) and \( V ar\{X_{n,j}\} \) \((n = 0, 1, \ldots, N - 1)\) are the predicted and measured value of the corresponding accumulated variance for the resizing window \( j \), respectively. We then use the predicted \( m^*_j + 1 \) and \( V ar^*\{X_{n,j+1}\} \) values as the inputs for a binary search to define the needed bandwidth for the window \( j + 1 \). The output of the search is the bandwidth amount assuring that the achievable delay violation probability \( \epsilon_{ach} \) is sufficiently close (expressed via the parameter \( \epsilon^* \)) to the target \( \epsilon \), i.e., \( \log \epsilon_{ach} - \log \epsilon < \epsilon^* \).

Although the traffic prediction used in PS1 scheme is quite simple and in fact the scheme works fairly well, we have learned from our experiments [4] that there is sometimes the problem of under provisioning. More precisely, the predicted value (either the mean traffic rate or the variance of the cumulated traffic) underestimates the real value in certain cases. Since the predicted values are used as an input for the binary search finding the bandwidth amount to provision, we suffer from under-provisioning, or equivalently QoS degradations, which should be highly avoided. We have revealed that the under estimation problem mainly occurs when the traffic trend changes from decreasing to increasing. Moreover, we can state the following.

**Proposition 3.1:** With the ES technique, if we have a predicted load smaller than the current measured load, then this under estimation remains as long as the aggregate load exhibits increasing trend.

**Proof:** see [4].
Having identified the above causes of under-provisioning, we work out the subsequent two provisioning schemes PS1* and PS1** with enhanced prediction rules, remedying the effect of traffic under-estimations.

B. PS1*: Modifying Prediction Rules with Linear Extrapolation

We change the prediction rule whenever we experience first (let’s say at the resizing window $j$) two facts together: increasing traffic trend and under prediction, i.e. $m_{j} > m_{j-1}$ and $m_{j} > m_{j}^{*}$. Instead of (4), we predict $m_{j+1}^{*} = m_{j} + (m_{j} - m_{j-1})$, assuming a local linear increasing trend of traffic. This modification is again applied in the window $j + 1$ if we still have $m_{j+1}^{*} > m_{j}$ and $m_{j+2}^{*} > m_{j+1}$, otherwise we switch back to the normal ES rule according to (4), and so on.

The same consideration is employed for the prediction of the measured link utilization in the last resizing window. This is Duan et al.’s proposal [1], to which we refer under the name PS4 throughout this paper.

C. PS1**: Modifying Prediction Rules with Predicted Increments

In this approach we use predicted increments between the loads from two successive resizing windows to give a load forecast as follows.

\[ m_{k}^{*} = m_{k} + (m_{k}^{*} - m_{k-1}). \]

By doing this we ensure that the original under-prediction in window $j$ becomes over-estimation with the degree observed earlier in window $j - 1$. For the following windows $k, k \geq j + 1$, as long as $m_{k} > m_{k-1}$ (i.e. the increasing trend is still valid), we predict $m_{k+1}^{*} = m_{k}^{*} + \Delta_{k+1}^{*}$ instead of using (4). The term $\Delta_{k+1}^{*}$ refers to the predicted increment (hence the name of the provisioning scheme) which is forecasted based on the measured load increments between two consecutive windows, and by using the ES technique as follows. By definition, for $k \geq j + 1$ the measured increment between window $k$ and $k - 1$ is $\Delta_{k} = m_{k} - m_{k-1}$. The estimation of $\Delta_{k+1}^{*}$ is that $\Delta_{k+1}^{*} = w \Delta_{k} + (1 - w) \Delta_{k}^{*}$. We initially set $\Delta_{j}^{*} = 0$. When the increasing trend stops, we switch back to the original ES rule (4).

The same consideration is employed for the prediction of the variance function $\text{Var}\{X_{n,j}\} (n = 0, 1, \ldots, N - 1)$.

D. PS2: MVA Bound based Scheme, without Prediction

In contrast with the series of PS1 scheme, in the PS2 scheme, we simply use the computed $m_{j}$ and $\text{Var}\{X_{n,j}\}$ $(n = 0, 1, \ldots, N - 1)$ as the input parameters of the binary search for the needed bandwidth of the window $j + 1$.

IV. OTHER PROVISIONING SCHEMES

A. Link Utilization based Scheme

In this scheme, link bandwidth adjustments are based on the relation between the link utilization threshold and the measured link utilization in the last resizing window. This is Duan et al.’s proposal [1], to which we refer under the name PS3 throughout this paper. In order to avoid too frequent bandwidth updates, the authors also suggest that the bandwidth should be added or released in quotas. One quota is a chunk of bandwidth which is several times larger than one unit of bandwidth and can be set to e.g. $\beta / \sqrt{\gamma}$, where $\nu$ is the variance of the measured traffic rate, $\beta$ is a scaling factor. This scheme can be considered to belong to the QoS-unaware group due to the lack of specifying the explicit role of link utilization on the target QoS.

B. Rate Overload Probability (ROP) based Scheme

The needed bandwidth in resizing window $j + 1$ is set to $c_{j+1} = m_{j} + v_{j} \sqrt{\nu}$, where $m_{j}$ and $v_{j}$ are the mean and variance of the measured traffic load in the current window $j$. This is the scheme by Duffield et al. [5], to which we refer under the name PS4 throughout this paper.

The rationale behind the setting of parameter $\alpha$ is as follows. If the aggregate traffic load is considered to be Gaussian distributed, then the probability that the allocated bandwidth is exceeded is given by the standard $Q$-function, $Q(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\infty} e^{-\frac{z^2}{2}} dz$.

The basic idea of this scheme can be thought of as if we have a bufferless model and we have to keep the rate overload probability under a given threshold $\epsilon$. The bandwidth amount to be assigned is calculated explicitly as $c_{j+1} = m_{j} + Q^{-1}(\epsilon / \sqrt{\nu})$. This scheme belongs to the QoS-aware group without traffic prediction.

C. Cisco Scheme

One of commercially available solutions for dynamic bandwidth provisioning is the Cisco MPLS AutoBandwidth allocator [6], where the local maximum approach is used. The average traffic rate is sampled in each measurement time slots. By means of RSVP-TE signalling, the bandwidth of a given MPLS tunnel is automatically adjusted after each resizing window. Both the measurement slot length and the window length are configurable parameters. The bandwidth for the next resizing window is set to the largest value from the set of the average rates measured in the last resizing window. We refer to this scheme under the name PS5.

For clarity, in Fig. 1, the assignment of the schemes to the corresponding path of the feature-tree is symbolically shown by the arrows originating from the schemes to the proper leaf.

V. ANALYSIS RESULTS AND DISCUSSIONS

Our investigations are based on both analytical results and trace-driven simulation results. For the input traffic, we use the traffic aggregated from real MPEG sources (MPEG traffic), Ethernet sources (Ethernet traffic), WAN sources (WAN traffic), and the mixture of the before-mentioned sources (Mix traffic). The measurement slots and resizing windows have length of 40ms and 4000ms, respectively. The details of real traffic sources and the scenario settings are given in [4]. Due to space limitation, we focus hereafter only on the insights of the QoS-aware schemes, skipping the investigations on the superiority of QoS-aware provisioning schemes over the QoS-unaware ones (we refer to [4] for the complete investigation).

---

1. In [4] we provide the detailed discussion and strict theoretical proof on the superiority of this proposed scheme (PS1**) over the local linear extrapolation based scheme (PS1*).
Recall that in the MVA bound based schemes, we accomplish bandwidth provisioning with explicit respect to the packet delay violation probability or equivalently, to the queue tail probability \( Pr(q > x) < \epsilon \) (we refer to this as a "buffered constraint"). In the ROP based scheme PS4, we calculate the bandwidth for the next resizing window such that the rate overload probability is below the threshold \( \epsilon \). This PS4 scheme can also be interpreted as if we have a buffered system and the target QoS is that the "queue tail" probability \( Pr(q > 0) \) should be kept below \( \epsilon \) (we refer to this as a "bufferless constraint").

Assume that we use one of the MVA bound based schemes, but with the target QoS put on the bufferless constraint, i.e. \( Pr(q > 0) < \epsilon \). To find the needed bandwidth we have to evaluate the equality \( e^{-\mu_0/2} = \epsilon \) (see expression (3)). It has been shown in [7] that \( \mu_0 = (\epsilon - \lambda)/C_\lambda(0) \). Remember that \( C_\lambda(0) = \sigma^2 = \text{Var}[\lambda] \), i.e. \( C_\lambda(0) \) is the variance of the marginal distribution of the aggregate traffic rate. Thus, for a bufferless constraint the MVA bound \( e^{-\mu_0/2} \) depends only on the mean \( \lambda \) and the variance \( \sigma \) of the marginal distribution of the aggregate load, and not on the correlation structure of the input traffic expressed through the auto-covariance functions \( C_\lambda(l), l > 0 \). As a consequence, we can do QoS-aware provisioning regardless of the traffic correlation structure. This is a big advantage because we do not have to know and analyze the traffic correlation structure (that sometimes is quite complicated, e.g. the long-range dependent property) in order to decide the proper bandwidths. Moreover, the second advantage of using the bufferless constraint is that we can exactly compute the bandwidth amount as \( c = \lambda + 2C_\lambda(0) \ln \epsilon \), and we do not need to perform a binary search giving an approximate bandwidth value.

On the other hand, it is clear that for any given system, we have \( Pr(q > x) < Pr(q > 0) \) for \( x > 0 \). Therefore, if a specific provisioning scheme considers a given violation probability to be concerned with the bufferless constraint, rather than the buffered constraint, then it will advise us to use more bandwidth. A natural question is now raising. To what extent we need more bandwidth when provisioning is done with regard to the bufferless constraint, rather than to the buffered constraint? Does this extra-bandwidth amount pay off by the facts that the correlation structure is not required to be known and we can exactly define the bandwidth amount?

To answer the above question, we compare the schemes utilizing buffered and bufferless constraints. We first introduce the notion of provisioning factor, which stands for the average of normalized required bandwidth over the resizing windows. The base for the normalization is the bandwidth required by the chosen, reference scheme. Expressing in a mathematical way, given that the bandwidth provisioned by a specific scheme \( L \) over a window \( i \) is \( bw_{i,L} \), the bandwidth provisioned by the reference scheme is \( bw_{ref,i} \), the provisioning factor of the scheme \( L \) computed over \( M \) resizing windows is \( \frac{1}{M} \left( \sum_{i=1}^{M} \frac{bw_{i,L}}{bw_{i,ref}} \right) \). The provisioning factor of the reference scheme itself is 1.

In our first series of experiments, we fix the target QoS at \( Pr(delay > 10ms) < 10^{-4} \). We report the comparative results in Table II showing the provisioning factor of different schemes. In the tables, the "prediction" term refers to whether we use traffic prediction. If the answer is YES, the prediction rule of PS1 is used. The answers “enhanced 1” and “enhanced 2” mean that the prediction rule of PS1* and PS1** are used, respectively. Concerning the models, we always add the term "bufferless" to the model name if the bufferless constraint is employed. For clarity, in Table I we give the mapping between the scheme names we have so far and the combination of prediction rule and applied model. The blank element in Table I means that for a given combination we have not assigned a specific name yet. Also note that for each resizing window, the Cisco scheme (PS5) assigns a bandwidth amount that is independent of whether we use traffic prediction.

Table II clearly confirms that a given bufferless constraint based provisioning scheme always requires more bandwidth than its buffered counterparts using the same prediction rule. For our specific target QoS \( Pr(delay > 10ms) < 10^{-4} \), the degree of additional bandwidth amount captured by the provisioning factor is around 10-24%. Also observe that the ROP bufferless scheme always allocates a smaller bandwidth amount (around 4-6%) than the MVA bufferless scheme.

As can be read out from Table III, the two schemes PS1* and PS1**, where more emphasize was put on minimizing potential QoS degradation by using the modified prediction rules, already imply over provisioning compared to PS1. Consequently, the degree of additional bandwidth amount needed in the bufferless scenarios compared to the bandwidth of PS1 is about 17-24% for MPEG traffic, but is only 12-19% compared to PS1* and 5-12% compared to PS1**. This degree is even smaller for other (Ethernet, WAN, Mix) traffic as can be inferred from Table III.

Also from Table III, although the Cisco scheme is totally unaware of QoS parameters, this local maximum based provisioning provides a bandwidth amount in a surprisingly close range (around 1-5% bias) of that obtained with the PS1 scheme. Of course, this is only true for the case of specific delay constraint \( Pr(delay > 10ms) < 10^{-4} \). Changing either the delay bound or the violation probability will result in another quantitative bias of the Cisco scheme.

In the second series of experiments, we check the effect of the delay bound and the violation probability on the provisioning factor of different schemes. The results are shown in Figs. 2-5, where we plot the provisioning factor versus the delay bound (while the violation probability is fixed at \( 10^{-4} \)) and versus the violation probability (while the delay bound is...
Firstly, we see in the figures that the degree of additional bandwidth for PS1* and PS1** schemes compared to PS1 scheme is independent of the applied delay bound and violation probability, i.e. it remains nearly constant for any setting of these two parameters. Secondly, Figs. 2, 3 indicate that except for the PS5 scheme, if the same delay bound ($D = 10\text{ms}$) is applied to buffered constraint based schemes, the larger the violation probability, i.e. the looser the target QoS, the smaller the degree of the additional bandwidth needed by the bufferless constraint based schemes. The Cisco scheme exhibits the converse trend, because for each resizing window it allocates a bandwidth amount independently of the violation probability, while the PS1 scheme (the reference scheme in this case) allocates less and less bandwidth with the increase of the violation probability.

**TABLE II**

**Provisioning factor of different schemes (I), the target QoS $Pr(\text{delay} > 10\text{ms}) < 10^{-4}$**

<table>
<thead>
<tr>
<th>Model \ Prediction</th>
<th>MPEG traffic</th>
<th>Ethernet traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>MVA</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MVA, bufferless</td>
<td>1.245</td>
<td>1.245</td>
</tr>
<tr>
<td>ROP, bufferless</td>
<td>1.182</td>
<td>1.181</td>
</tr>
<tr>
<td>Cisco rule</td>
<td>1.041</td>
<td>1.046</td>
</tr>
</tbody>
</table>

**TABLE III**

**Provisioning factor of different schemes (II), the target QoS $Pr(\text{delay} > 10\text{ms}) < 10^{-4}$**

<table>
<thead>
<tr>
<th>Scheme \ Traffic trace</th>
<th>MPEG</th>
<th>WAN</th>
<th>Ethernet</th>
<th>Mix</th>
</tr>
</thead>
<tbody>
<tr>
<td>buffered MVA with prediction (PS1), reference scheme</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>buffered MVA without prediction (PS2)</td>
<td>0.994</td>
<td>0.999</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>buffered MVA with enhanced prediction (PS1*)</td>
<td>1.093</td>
<td>1.093</td>
<td>1.067</td>
<td>1.054</td>
</tr>
<tr>
<td>buffered MVA with enhanced prediction (PS1**)</td>
<td>1.128</td>
<td>1.090</td>
<td>1.143</td>
<td>1.130</td>
</tr>
<tr>
<td>bufferless MVA with prediction rule of PS1</td>
<td>1.245</td>
<td>1.145</td>
<td>1.150</td>
<td>1.161</td>
</tr>
<tr>
<td>bufferless MVA without prediction</td>
<td>1.237</td>
<td>1.144</td>
<td>1.147</td>
<td>1.158</td>
</tr>
<tr>
<td>bufferless ROP, with prediction rule of PS1</td>
<td>1.182</td>
<td>1.112</td>
<td>1.104</td>
<td>1.119</td>
</tr>
<tr>
<td>bufferless ROP, without prediction</td>
<td>1.174</td>
<td>1.110</td>
<td>1.101</td>
<td>1.116</td>
</tr>
<tr>
<td>Cisco scheme</td>
<td>1.041</td>
<td>1.050</td>
<td>1.014</td>
<td>1.041</td>
</tr>
</tbody>
</table>

Fig. 2. Effect of violation probability on the provisioning factor, Ethernet traffic

Fig. 3. Effect of violation probability on the provisioning factor, MPEG traffic

Fig. 4. Effect of delay bound on the provisioning factor, Ethernet traffic

Fig. 5. Effect of delay bound on the provisioning factor, MPEG traffic
On the other hand, when the violation probability is fixed ($\epsilon = 10^{-4}$), the degree of the additional bandwidth increases with the delay bound (see Figs. 4, 5). This is because increasing the delay bound in the buffered constraint based schemes implies looser and looser QoS requirement compared to the QoS of the bufferless counterpart scenario. For example, the bandwidth difference between bufferless and buffered scenarios when the MVA bound based dimensioning rule is applied can be up to 40% in case of MPEG and more than 20% in case of Ethernet traffic, when $D = 40\text{ms}$. It is again observable that when the bufferless constraint is considered, the MVA bound based scheme requires more bandwidth (around 4-6%) than the ROP based, PS4 scheme.

Our last investigation is concerned with the comparison between the ROP based schemes and the MVA bound based scheme, when the bufferless constraint is used. We already know from experiments so far that the former scheme always requires less bandwidth than the latter one. Thus, the question now is whether the target QoS is really met with the ROP based scheme or it is safer from a QoS point of view to consume more bandwidth according to the MVA bound based scheme. We check this issue by performing trace driven simulation with the QoS target $Pr(q > 0) < 10^{-4}$ for provisioning. Table IV shows the actual value of this rate overload probability after the whole period of provisioning. Note that we again classify the provisioning schemes along two dimensions according to the model in use (MVA bound based or rate overload probability based) and whether (and what) prediction rule is applied.

### Table IV

<table>
<thead>
<tr>
<th>Pred. \ Sch.</th>
<th>MPEG traffic</th>
<th>Ethernet traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MVA based</td>
<td>ROP based</td>
</tr>
<tr>
<td>no</td>
<td>4.60e-05</td>
<td>1.02e-02</td>
</tr>
<tr>
<td>yes</td>
<td>5.80e-03</td>
<td>1.32e-02</td>
</tr>
<tr>
<td>enhance-1</td>
<td>2.30e-03</td>
<td>4.10e-03</td>
</tr>
<tr>
<td>enhance-2</td>
<td>6.00e-04</td>
<td>1.70e-03</td>
</tr>
</tbody>
</table>

Allocating more bandwidth than the ROP based counterparts, the MVA bound based schemes achieve definitely better QoS, as reported in Table IV. However, even they could not entirely meet the target QoS $Pr(q > 0) < 10^{-4}$, although the target QoS improves with the application of the enhanced traffic prediction rules of PS1* and PS1** scheme. In the best case, when prediction rules of PS1** scheme is used ("enhance-2" rule in the table), still 4-6 times larger rate overload probability is obtained compared to the target QoS. The reasons for this phenomena might be due to two facts, the non-perfectibility concerning the Gaussian approximation and the traffic under-prediction that we could not completely exclude. Remind that the original QoS we would like to have is the delay constraint $Pr(\text{delay} > D) < \epsilon$, or equivalently $Pr(q > D, c) < \epsilon$ which is surely looser than $Pr(q > 0) < \epsilon$. Therefore, despite the fact that we do not meet $Pr(q > 0) < \epsilon$, we still have a chance to meet the original packet delay QoS. Nevertheless, from the QoS point of view, when the bufferless constraint is employed, we advocate to use the MVA bound based provisioning scheme, because with only 4-6% more bandwidth allocation, it can tolerate better QoS degradation effects of the uncertainties lying on traffic model approximation and traffic under-prediction.

### VI. Conclusions

In this paper, we have dealt with adaptive bandwidth provisioning schemes. Besides revisiting some available schemes, we have proposed novel adaptive provisioning schemes that assign the link bandwidth with explicit respect to the target, packet level QoS. The operation of these novel schemes combines periodical measurements, the Gaussian traffic model and traffic prediction features. Our work [4] has confirmed that these proposed QoS-aware schemes achieve better provisioning efficiency than other existing schemes, i.e. bandwidth is properly regulated to ensure the target QoS, avoiding both extensive over-provisioning and severe under-provisioning.

Based on thorough analytical and simulative performance analysis with real traffic traces, the tradeoff of using bufferless or buffered constraints in QoS-aware provisioning schemes has been explored. By using the bufferless constraint, we do not need a knowledge on the correlation structure of the traffic, and we can exactly define the bandwidth to be allocated. But on the other hand, we need more bandwidth than using the buffered constraint. Our analysis has shown that the degree of additional bandwidth compared to the best, buffered constraint based scheme $PS1^{**}$ is acceptably small (at most 12%) if the delay bound is sufficiently small (in a range of 10ms or below) and the violation probability is not too tight (in a range of $10^{-4}$ or above). Thus, in such cases we propose to use the bufferless constraint based scheme, exploiting the before-mentioned advantages. Moreover, between two bufferless constraint based alternatives, the MVA bound based scheme and the rate overload probability (ROP) based scheme, we advocate to use the former one, because it can better cope with and alleviate potential QoS degradations with around 4-6% bandwidth excess.

### References


