Experience-Based Admission Control (EBAC)

Michael Menth, Jens Milbrandt, Simon Oechsner
Dept. of Distributed Systems, Inst. of Computer Science, University of Würzburg, Germany
Email: {menth,milbrandt,oechsner}@informatik.uni-wuerzburg.de

***Abstract*— Classical admission control approaches take either descriptor or measurement based information about the traffic into account without relating them to each other. We propose a experience-based AC (EBAC) which uses an empirical percentile of the effective reservation utilization to determine a suitable overbooking factor. In this paper, we show the impact of different measurement time scale resolutions and different quantiles on the performance of the system. We propose aging mechanisms for statistic collection to make the system adaptive to traffic mixes that change over time. We illustrate their effectiveness by simulation results.**

I. Introduction

Next Generation Networks (NGN) are expected to provide Quality of Service (QoS) to the customers. This can be achieved by bandwidth overprovisioning or by Admission Control (AC), which is the focus of this paper. AC can be subdivided into parameter-based AC (PBAC) methods and measurement-based AC (MBAC) methods. PBAC methods limit the acceptable traffic by accounting effective bandwidths of admitted flows. The effective bandwidth of a flow is calculated based on its traffic descriptor which is provided by the traffic source, and an underlying stochastic model [1]. PBAC is often inefficient because the AC decisions are rather pessimistic and the traffic descriptors usually overestimate the actually sent rate to avoid packet loss and delay due to spacing or policing. In addition, AC calculations for heterogeneous traffic mixes are very complex or even unfeasible. In contrast, MBAC approaches measure the current network load and take an estimate of the current characteristics of the new flow and the admitted aggregate to perform the AC decision [2], [3], [4], [5]. Other approaches [6], [7] work on end-to-end (e2e) measurements. MBAC methods take advantage of network measurements and admit traffic as long as enough network capacity is still available. The downside of these approaches is their susceptibility to measurement accuracy and QoS attacks, e.g., a set of streams can be silent for a while and congest the network later by sending simultaneously at high bitrate.

In this paper, we propose the experience-based AC (EBAC) which is a new type of MBAC. It pertains on a single link but it can be easily extended to a network-wide scope. It relies on peak rate traffic descriptors, which may be significantly overestimated. The utilization of the reserved capacity gives an estimate for the peak-to-mean rate ratio and allows for the calculation of an overbooking factor. The idea is simple but safety margins are required to provide sufficient QoS and questions arise regarding its robustness against variable traffic streams. We elaborate a feasible EBAC concept and illustrate the impact of the system parameters. After all, EBAC reveals to be a robust AC algorithm for flows with large peak-to-mean rate ratios.

The paper is organized as follows. Section 2 presents the EBAC and Section 3 explains our simulation and performance evaluation approach. The numerical results in Section 4 propose appropriate system parameters to achieve satisfying QoS and illustrate the robustness against highly variable traffic flows. The conclusion in Section 5 summarizes this work and gives an outlook on further research.

This work was funded by the Bundesministerium für Bildung und Forschung of the Federal Republic of Germany (Förderkennzeichen 01AK045) and Siemens AG, Munich. The authors alone are responsible for the content of the paper.

II. Experience Based Admission Control

This section describes EBAC in detail. We explain the AC decision for a single link together with its parameters and present the calculation of the overbooking factor $\varphi(t)$.

A. Admission Control Decision and Parameters

An AC entity for a link l limits the access to its capacity $c(l)$ and records the admitted flows $\mathcal{F}(t)$ at any time t together with their requested peak rates $\{r(f) : f \in \mathcal{F}(t)\}$. When a new flow f_{new} arrives, it requests for a peak rate $r(f_{new})$. If

$$r(f_{new}) + \sum_{f \in \mathcal{F}(t)} r(f) \quad \leq \quad c(l) \cdot \varphi(t) \cdot \rho_{max} \cdot \chi \qquad (1)$$

holds, admission is granted and f_{new} joins $\mathcal{F}(t)$. Flows are removed from $\mathcal{F}(t)$ on termination. The experience-based overbooking factor $\varphi(t)$ is calculated by statistical analysis and indicates how much more bandwidth than $c(l)$ can be safely allocated for reservations. The safety threshold $\chi \leq 1$ is a calibration parameter. The maximum link utilization threshold ρ_{max} limits the traffic admission such that the expected expected packet delay W exceeds an upper delay threshold W_{max} only with probability p_W.

B. Calculation of the Maximum Link Utilization Threshold ρ_{max}

The value for ρ_{max} depends significantly on the traffic characteristics and the link capacity and most prominent solutions are based on the $M/M/1-\infty$ and the $N \cdot /D/1-\infty$ queuing system. Real-time traffic like voice or video applications has a rather constant output rate and it can be rate-controlled by a spacer such that a maximum flow rate r_f is enforced. Therefore, we work with a ρ_{max} calculation based on the

$N \cdot D/D/1-\infty$ approach, which assumes N homogeneous flows, each sending packets of constant size and constant packet inter-arrival times. The flow rates are $C_f = \frac{E[B]}{E[A]}$ and the packet delay distribution of this periodic system is given by

$$\begin{aligned} P(W \leq t) &= 1 - e^{-2 \cdot x(\frac{x}{N}+1-\rho)} \\ \text{with } x &= \frac{t \cdot c(l)}{E[B]} \text{ and } \rho = \frac{N \cdot E[B]}{C_f}. \end{aligned} \quad (2)$$

provided that $\rho \leq 1$ holds (cf. [1] (15.2.4)). The maximum link utilization ρ_{max} is computed by

$$\rho_{max} = \max_{\rho} \{\rho : P(W > W_{max}) \leq p_W\}. \quad (3)$$

Due to Equation (2) the maximum link utilization ρ_{max} increases with increasing bandwidth $c(l)$ and decreases with increasing packet size $E[B]$.

However, the $N \cdot D/D/1$ model is not a suitable for traffic with varying inter-arrival times and packet sizes. But we will see in Section IV that the adaptive overbooking factor $\varphi(t)$ and the safety threshold χ can compensate the effects of the traffic deviations from the exact model to some degree.

C. Computation of the Overbooking Factor $\varphi(t)$

The admitted flow aggregate $\mathcal{F}(t)$ and all its properties depend on time because new flows start and existing flows terminate. The mean rate $c(f)$ of a flow f is usually not known a priori and its declared peak rate $r(f)$ can be controlled by a shaper. Analogously, $R(t) = \sum_{f \in \mathcal{F}(t)} r(f)$ is the reserved bandwidth of all flows and $C(t)$ denotes their unknown aggregate mean rate.

EBAC measures the utilized bandwidth $M(t)$ of the aggregate reservation $R(t)$ and a time statistic for the reservation utilization $U(t) = \frac{M(t)}{R(t)}$ is collected. $U_p(t)$ is the p_u percentile $U_p(t)$ of this empirical utilization distribution and its reciprocal is the overbooking factor $\varphi(t) = \frac{1}{U_p(t)}$.

1) Measurement Process for $M(t)$: We use interval measurements to obtain $M(t)$. The time axis is divided into disjoint (not necessarily equidistant) intervals $I_i = [t_i, t_{i+1})$ of length $\Delta_i = t_{i+1} - t_i$. In our performance evaluation, we have a fixed Δ for all intervals. The corresponding rates $M_i = \frac{\Gamma_i}{\Delta_i}$ are determined by metering the traffic volume Γ_i sent during I_i. This method is simple to implement since Γ_i can be obtained via SNMP by standard routers. The desired measured rate is $M(t) = M_i$ for $t \in [t_{i+1}, t_{i+2}]$, i.e. the next interval length Δ_{i+1} is a measurement delay.

2) Statistic Collection $P(t,U)$: The aggregate reservation $R(t)$ is known from the AC process and the utilization values $U(t)$ are sampled every $10ms$. They are stored as hits in bins for a time-dependent histogram $P(t,U)$. The time-dependent utilization quantile $U_p(t)$ can be derived from $P(t,U)$ by

$$U_p(t) = \min_{u} \{u : P(t, U \leq u) \geq p_u\}. \quad (4)$$

To avoid an underestimation of $U_p(t)$ and an overestimation of $\varphi(t)$, enough statistical data must be collected before $\varphi(t)$ is calculated.

3) Statistics Aging: If traffic characteristics change over time, EBAC's utilization statistic must forget obsolete data to reflect the properties of the new traffic mix. Therefore, we devaluate the contents of the histogram bins every minute by a devaluation factor d. Typical values are between 0.9 and 1, whereby a system with $d = 1$ is a simple statistic collection that does not forget old events.

III. Performance Evaluation Method

EBAC is a LAC method taking advantage of the observed reservation utilization. Therefore, we focus on the simulation of a single link carrying traffic of a single traffic class.

A. Admission Control

We simulate the EBAC of Section II-A on a single link. Upon flow arrival, admission is granted if Equation (1) holds and f_{new} joins $\mathcal{F}(t)$. The flows in our simulation have different request sizes which leads to request size dependent blocking probabilities on a heavily loaded link. To avoid that, we apply trunk reservation in our simulations [8], i.e., a flow is only admitted if a maximum request size flow could also be accepted. As a consequence, all request types have the same blocking probability and contribute to the traffic aggregate according to their type specific offered load.

B. Traffic Model

We use the following packet and flow level model in our simulation.

1) Packet Level Model: Voice traffic consists typically of 50 equidistantly spaced packets per second, each of them carrying about 200 bytes (160 bytes uncompressed payload and 40 bytes RTP/UDP/IP header information), leading to 80 Kbit/s. For mobile traffic, compression techniques such as the Adaptive Multi-Rate (AMR) vocoder are used to generate traffic at variable rate and to reduce the mean payload size to 20 bytes with a range between 4 and 36 bytes. In addition, consecutive packet sizes are strongly correlated. This example shows the complexity of such traffic structures. However, in this work we are rather interested in a basic understanding of the EBAC under different conditions and not in its behavior for particular applications. Therefore, we abstract from real traffic patterns.

If not mentioned differently, we take the following standard parameters. Flows have a flow mean rate of $E[C_f] = 256\,Kbit/s$ and are shaped to a flow peak rate of $E[R_f] = 768\,Kbit/s$. Both packet size and packet inter-arrival time distributions contribute to the rate variability within a flow. To keep things simple, we assume fixed packet sizes of 512 bytes and use an exponential distribution for the packet inter-arrival time with rate λ_p.

2) Flow Level Model: Both the flow inter-arrival time and the call holding time follow a Poisson model, i.e., they are exponentially distributed with rate λ_f and $\mu_f = \frac{1}{90\,s}$, respectively. The absolute offered load $a_f = \frac{\lambda_f}{\mu_f}$ is the mean number of simultaneous connections in Erlang if no blocking occurred. We define the offered load relative to the link size

by $a_r = \frac{E[R_f] \cdot a_f}{c(l)}$. If not mentioned differently, we have $c(l) = 10\,Mbit/s$ and take $a_r = 1.2$ as default to study the EBAC performance under heavy traffic load.

C. Peak-to-Mean Rate Ratio

The peak rate $r(f)$ of a flow f is shaped by a spacer and unlike in reality, in our simulation its mean rate $c(f)$ is known a priori. We define the peak-to-mean rate ratio of a flow by $k(f) = \frac{r(f)}{c(f)}$, i.e., as the ratio of its peak and mean rate. Analogously, $K(t) = \frac{R(t)}{C(t)}$ is the peak-to-mean rate ratio of the traffic aggregate. It is a natural upper limit on the achievable overbooking factor.

D. Problem Description

EBAC chooses automatically an overbooking factor $\varphi \in [1; \infty)$, however, this is influenced by the system parameters. We consider two extreme settings for $\varphi(t)$.

- If φ is very small, no multiplexing gain is achieved and only a low percentage of the existing bandwidth is used for the transport of premium traffic. This leads to increased bandwidth requirements compared to the offered – but unknown – aggregate mean rate $C(t)$ or to an increased flow blocking probability by AC.
- If φ is very large, i.e. in the range of the peak-to-mean rate ratio $K(t)$, only about $C(t)$ capacity is allocated for the aggregate reservations $R(t)$. If the link is fully loaded, i.e. if EBAC blocks new requests, $C(t)$ approximates $c(l)$ and QoS is jeopardized. Depending on the burstiness of the aggregate traffic, packet loss and delay can become quite high.

Both extremes are not desired. The goal of the EBAC is the determination of a suitable φ resulting in a tradeoff between improved bandwidth economy and probabilistic QoS guarantees.

E. Performance Measures

The objective of AC is to limit packet delay due to queueing and to avoid packet loss due to buffer overflow. If packet loss can be eliminated by sufficiently large buffers, packet delay is the natural performance measure for the assessment of AC mechanisms. For a small traffic load and $C(t) << c(l)$, the really experienced delay can be very small even for too large overbooking factors like $\varphi(t) >> K(t)$. As the overbooking factor must still be safe if the system works at its upper limit, i.e. if the link is highly utilized, the really experienced delay is not suitable for the validation of the EBAC parameters. As it is impossible to scale up the given traffic realistically to simulate a link under heavy load, we scale down the link capacity such that it should be just enough to meet the QoS requirements of the traffic. This equals a virtual server with capacity $C_v(t)$ and the observed waiting time is the virtual server delay W_v.

We can estimate the aggregate traffic mean rate by $\frac{R(t)}{\varphi(t)}$ and we want to guarantee a maximum delay W_{max} with a probability p_W, i.e. $P(W \leq W_{max}) > p_W$. Therefore, we compute the required virtual server rate $C_v(t)$ based on a $N \cdot D/D/1$ queuing system with a mean arrival rate $\frac{R(t)}{\varphi(t)}$. Homogeneous connections with constant packet inter-arrival times A, packet sizes B, and request rates C_f are assumed, which have better multiplexing properties than more variable traffic. Hence, it is a conservative approach because it rather underestimates the needed resources.

The required capacity can be calculated similarly to Equation (2):

$$C_v(t) = \frac{1}{\chi} \cdot \min\left(\frac{R(t)}{\varphi(t)}, \frac{N \cdot E[B]}{2 \cdot W_{max}} \left(-1 + \sqrt{1 + \frac{4 \cdot D \cdot E[C_f]}{E[B]} - \frac{2 \cdot \ln(1 - p_W)}{N}} \right) \right) \quad (5)$$

where $N = \frac{R(t)}{\varphi(t)} / E[C_f]$. Note that the virtual capacity takes the safety threshold χ into account. Finally, we take the mean $E[W_v]$ of the virtual spacer delay W_v and primarily its 99%-percentile as performance measures.

F. Simulation Setup

To obtain the desired EBAC performance measures from simulations, we use the setup in Figure 1. Flows are generated according to our flow level model (cf. III-B.2) and pass the AC process (cf. III-A). An admitted flow f sends packets at a mean rate $c(f)$ (cf. III-B.1) that are spaced according to their indicated peak rate $r(f)$. The spaced packet streams are cloned and forwarded to two different queuing systems. The first one with server capacity $c(l)$ simulates the link l itself and is required to obtain the rate measurements $M(t)$ (cf. II-C.1) from its output process. The second one with server capacity $C_v(t)$ calculates the virtual server delay $W_v(t)$ to validate the overbooking factor $\varphi(t)$ (cf. III-E).

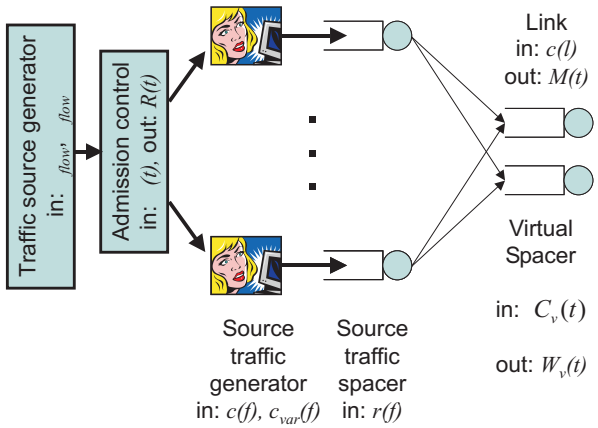

Fig. 1. The virtual server delay W_v is used for the assessment of the EBAC parameters.

IV. EBAC Performance

The purpose of our performance study is threefold. First, we intend to give parameter recommendations for EBAC, second, we want a proof of concept and show that the peak-to-mean

rate ratio can be exploited for overbooking, and third, we want to illustrate the robustness of EBAC against variable traffic sources.

A. Recommendation for EBAC Parameters

The measurement interval length Δ influences the smoothness of the measured traffic time series $M(t)$ which also affects the utilization distribution $P(t,u)$. As the overbooking factor $\varphi(t) = \frac{1}{U_p(t)}$ is based on the p_u utilization percentile (cf. Equation (4)), p_u is obviously also a critical EBAC parameter. We vary Δ from $10ms$ to $10s$ and p_u from 70% to 99.9%.

Figure 2 illustrates the mean overbooking factor $E[\varphi]$ depending on Δ and p_u on a $10Mbit/s$ link. We have set $\chi = 1$ and $d = 1$. The overbooking factor is of particular interest because it indicates the achievable gain by the EBAC method compared to peak rate allocation.

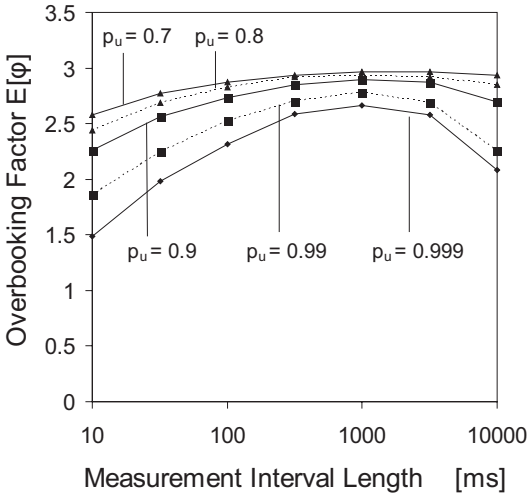

Fig. 2. Impact of the measurement interval Δ and the utilization percentile p_u on the mean overbooking factor $E[\varphi]$.

An increasing measurement window size Δ reduces the fluctuation of the measured rate $M(t)$ and the one of the utilization $U(t) = \frac{M(t)}{R(t)}$. This reduces $U_p(t)$ leading to a larger overbooking factor $\varphi(t)$. A decreasing percentile value p_u decreases $U_p(t)$ by definition and it increases thereby $\varphi(t)$. However, a stronger traffic concentration causes an increment in virtual server delay W_v. This can be well observed by the 99% quantile in Figure 3. The quantile of the virtual server delay is well limited up to a measurement interval of $\Delta = 1s$ and it can be compensated by a more conservative p_u.

For longer Δ, the measured $U(t)$ becomes too smooth, $U_p(t)$ too small, $\varphi(t)$ too large, and the virtual server delay too excessive. For very long $\Delta \approx 10s$ $\varphi(t)$ shrinks again, however, the delay values continue to increase. A closer look on the simulation data reveals that $c_{var}[\varphi]$ is about 20 times larger compared to the value for $\Delta = 1s$, i.e. $\varphi(t)$ is not a constant value but it fluctuates significantly. The value for the measured rate $M(t)$ is delayed and constant for time Δ.

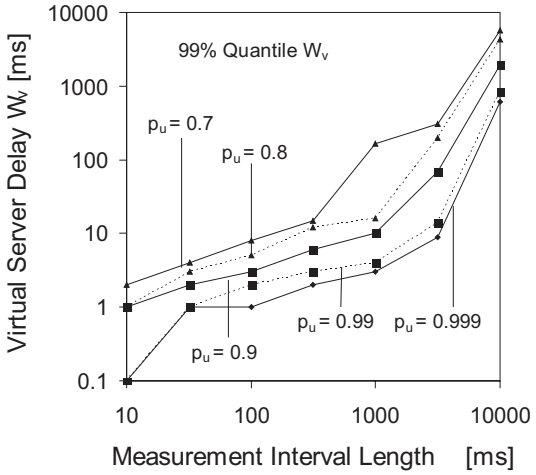

Fig. 3. The virtual server delay W_v for $\chi = 1.0$.

So, the utilization $\frac{M(t)}{R(t)}$ is significantly influenced by $R(t)$ whose evolution is determined by random flow arrivals and terminations which are not reflected fast enough by $M(t)$. We conclude that EBAC is not feasible for long measurement intervals in the order of $\Delta = 10s$. This may be different for links that aggregate more traffic and for flows with different holding times but this is subject to further research.

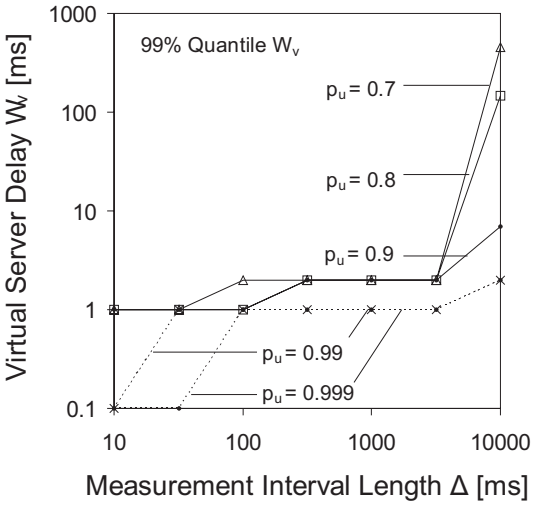

Fig. 4. The virtual server delay W_v for $\chi = 0.9$.

We change the safety threshold from $\chi = 1.0$ to $\chi = 0.9$. According to Figure 4, a measurement interval of $\Delta = 10s$ can still provide QoS. The plateaus in the figure result from our measurement accuracy for distributions.

However, the safety threshold χ is actually required if traffic characteristics deviate from their typical long time behavior, e.g., if the peak-to-mean rate ratio is slowly decreasing because it takes a while until the overbooking factor reflects the new

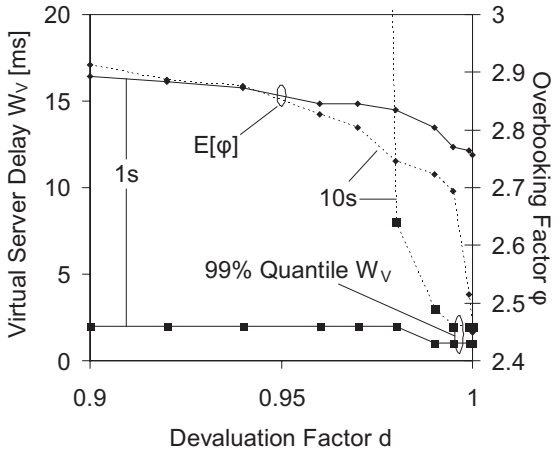

Fig. 5. Impact of the devaluation factor d for $\Delta = 1s$ and $\Delta = 10s$ with $\chi = 0.9$.

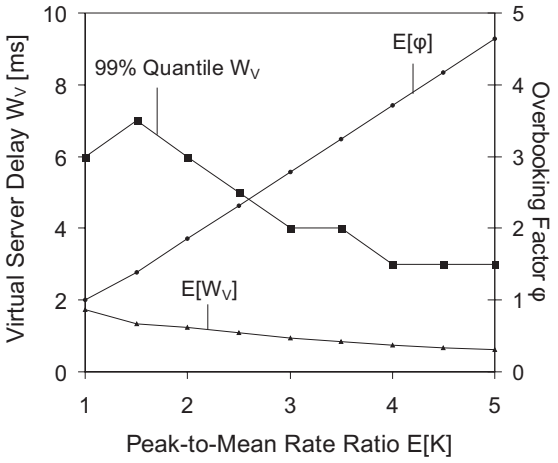

Fig. 6. The sensitivity of EBAC to the mean of traffic peak-to-mean rate ratios.

conditions due to statistic aging. Therefore, we consider the QoS performance with statistics aging. Figure 5 shows that $E[\varphi]$ is more stable for $\Delta = 1s$ than for $\Delta = 10s$ and that the QoS objective can only be met for $\Delta = 10s$ if $d \geq 0.99$ holds. With this devaluation factor, observed measurements still count more than 54% after one hour, i.e., the aging process is probably too slow. Therefore, we recommend the use of $\Delta = 1s$ and $p_u = 0.99$. Shorter measurement intervals are difficult to realize with existing hardware and more conservative percentiles are hard to compute with sufficient reliability. So, the proposed parameter set is a feasible solution.

As we do not consider changing traffic mixes, we work with $d = 1$ and $\chi = 1$ in the sequel.

B. Proof of Concept

The intrinsic idea is the exploitation of the peak-to-mean rate ratio and we show that the EBAC overbooking factor achieves this goal.

1) Impact of the Average Peak-to-Mean Rate Ratio: We performed simulations with different peak-to-mean rate ratios by using different peak rates for source traffic shaping. Figure 6 illustrates that EBAC reacts very well to traffic with different but constant peak-to-mean rate ratios. The average overbooking factor $E[\varphi]$ is approximately as large as the aggregate peak-to-mean rate ratio K. At the same time the virtual server delay W_v is well limited.

2) Impact of the Peak-to-Mean Rate Ratio Variability: We release the assumption of homogeneous traffic sources and use a traffic mix with flows having different peak-to-mean rate ratios. Table I allows for $c_{var}[K_f] \in [0; \sqrt{\frac{3}{2}}]$ while the average mean rate is $E[C_f] = 256\,Kbit/s$ and the average peak-to-mean rate ratio is $E[K_f] = \frac{E[R_f]}{E[C_f]} = 3$.

Figure 7 shows the EBAC performance depending on the coefficient of variation of the flow compressibility of the traffic mix. A mean overbooking factor $E[\varphi] < 3$ makes sense as the average peak-to-mean rate ratio is 3. It is decreased by an increasing peak-to-mean rate ratio variance. Obviously, EBAC adapts the overbooking factor such that W_v is well limited for $c_{var}[K_f] \leq 0.7$. Hence, EBAC works also well for heterogeneous traffic sources.

TABLE I

PEAK TO MEAN RATE RATIO DISTRIBUTION DEPENDING ON $c_{var}[K_f]$.

request type	0	1	2
$E[C_f]$	$256\,Kbit/s$	$256\,Kbit/s$	$256\,Kbit/s$
$E[R_f]$	$256\,Kbit/s$	$768\,Kbit/s$	$1536\,Kbit/s$
$E[K_f]$	1	3	6
probability	$\frac{1}{5} \cdot c_{var}[K_f]^2$	$1 - \frac{2}{3} \cdot c_{var}[K_f]^2$	$\frac{4}{15} \cdot c_{var}[K_f]^2$

3) Impact of Transmission Start Delays: The data transmission start is usually delayed regarding the admission time of a connection due to signaling and human reaction. On the one side, this affects the calculation of the reservation utilization $U(t) = \frac{M(t)}{R(t)}$ and leads to underestimation. On the other side, the aggregate peak-to-mean rate ratio $K(t)$ also increases.

Figure 8 shows the EBAC performance for delayed transmission starts. The transmission starts have an exponentially distributed latency L_{TS} which extends the reservation time by $E[L_{TS}]$ and increases the compressibility by a factor $1 + E[L_{TS}] \cdot \mu_f]$. As a consequence, we observe that $\varphi(t)$ increases with $E[L_{TS}]$ and that W_v is still under control for moderate transmission start delays, i.e. EBAC can take advantage of the start latency to increase the resource utilization.

C. Robustness against Traffic Variability

To meet a target delay bound W_{max}, both the required virtual server capacity $C_v(t)$ for a traffic aggregate and the maximum admission threshold for a link ρ_{max}, respectively, depend on the traffic characteristics of the admitted flows. Those are in particular packet size and inter-arrival time distributions as well as correlations thereof. We investigate

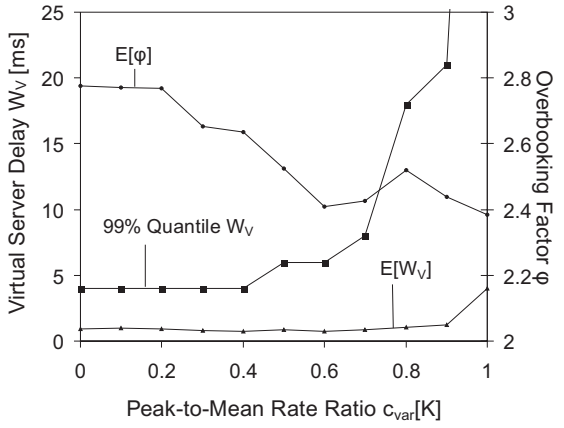

Fig. 7. The sensitivity of EBAC to the variability of traffic peak-to-mean rate ratios.

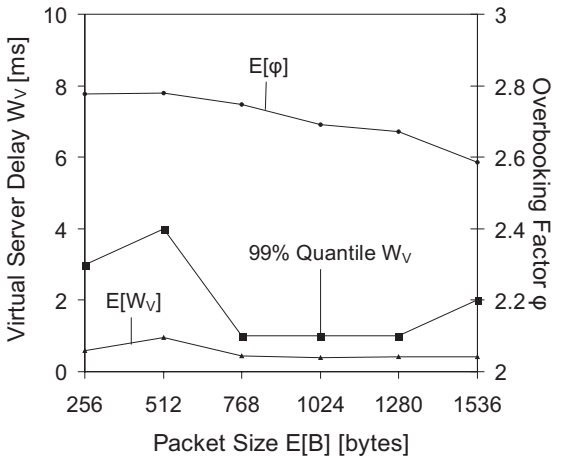

Fig. 9. Impact of the mean packet size $E[B]$.

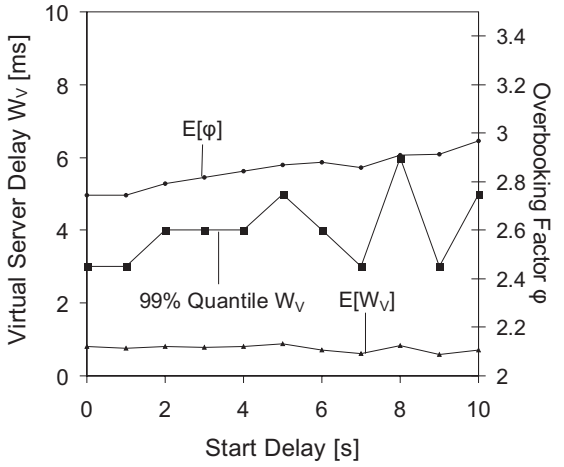

Fig. 8. Impact of the transmission start latency L_{TS}.

TABLE II
PACKET SIZE DISTRIBUTIONS FOR DIFFERENT PACKET SIZE VARIABILITIES.

packet size [bytes]	256	512	1536
probability	$4 \cdot \frac{c_{var}^2}{5}$	$1 - c_{var}^2$	$\frac{c_{var}^2}{5}$

parameter ranges and test whether EBAC is able to take the different queuing behavior into account by the calculation of the overbooking factor φ and to control the virtual server delay W_v.

1) Impact of the Average Packet Size: According to the $N \cdot D/D/1$ queuing formula (cf. Equation (2)), the average packet size is a key factor for the multiplexing property of a traffic mix. Figure 9 shows the average overbooking factor $E[\varphi]$, the mean, and the 99% quantile of the virtual server delay depending on the average packet size $E[B]$. The increasing packet size decreases $E[\varphi]$ such that the virtual server delay does not increase. Hence, EBAC can well cope with different mean packet sizes.

2) Impact of the Packet Size Variability: Instead of homogeneous traffic, we consider traffic mixes with flows of different but constant packet sizes. The parametrization of the packet size distribution in Table II allows for $c_{var}[B] \in [0; 1]$.

Figure 10 shows the EBAC performance for different packet size variabilities. The packet size variability $c_{var}[B]$ has no impact, neither on the overbooking factor $E[\varphi]$ nor on the virtual server delay, at least not in a visible way. Hence, EBAC is robust to different packet size distributions.

3) Impact of the Packet Inter-Arrival Time Variability: The average packet inter-arrival time $E[A]$ was implicitly investigated in IV-C.1 and has proven to be well treatable by EBAC. Now we study the impact of packet inter-arrival time variability by using ErlangK and hyper-exponential distributions for A within a single flow. According to Figure 11 $E[\varphi]$ is slightly reduced with increasing packet inter-arrival time variability. Therefore, both the mean and the 99% quantile of the virtual server delay increase to values 3 times larger than the target values. However, for a reasonable non-conservative assumption of $c_{var}[A] = 1$ the objective of 10 msec for the 99% quantile of the virtual server delay W_v is met. Hence, EBAC can well cope with flows with clearly different packet inter-arrival time distributions.

4) Impact of Traffic Correlations: Correlations within inter-arrival times of a traffic stream have a tremendous impact on its queuing behavior. We use a simple model for correlated traffic. The individual traffic sources have two discrete Markov states s_0 (off) and s_1 (on). In the on mode, the source sends packets as usual at the beginning of an inter-arrival time whereas this is suppressed in the off mode.

The state of a source changes from s_i to s_j with probability p_{ij} at the end of an inter-arrival time. The probability that the source is in state s_1 is $p_1 = \frac{p_{01}}{1-p_{11}+p_{01}}$. We use p_{11} to control

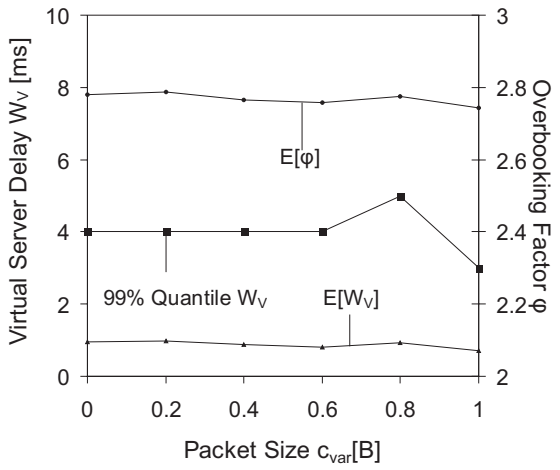

Fig. 10. Impact of the coefficient of variation $c_{var}[B]$ of the packet size.

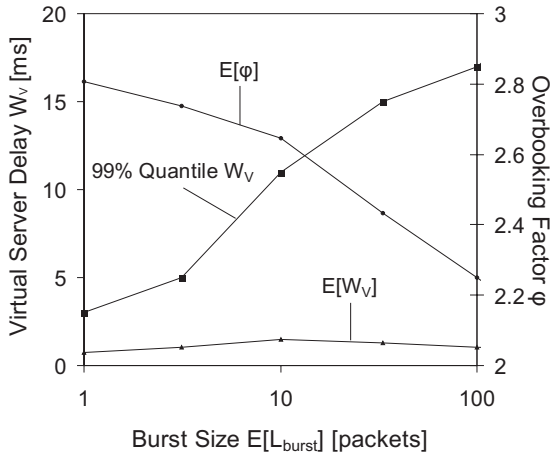

Fig. 12. Impact of correlated inter-arrival times.

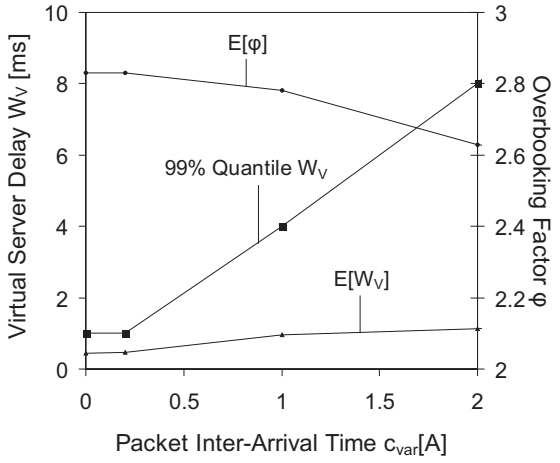

Fig. 11. Impact of the coefficient of variation $c_{var}[A]$ of the packet inter-arrival time.

the average burst length $E[L_{burst}] = \frac{1}{1-p_{11}}$ in packets. We set $p_1 = 0.5$, which leads to $p_{00} = p_{11} = \frac{(1-p_{11})\cdot p_1}{1-p_1}$ and to $p_{01} = p_{10} = 1 - p_{11}$. In addition, we set the mean packet inter-arrival time to $E[A] = \frac{E[B]}{2 \cdot E[C_f]}$ to achieve a flow rate of $E[C_f]$.

Figure 12 illustrates the EBAC performance for traffic with different inter-arrival time correlations indicated by the average burst length $E[L_{burst}]$ in packets. EBAC reduces the overbooking factor $E[\varphi]$ in the presence of correlated traffic sources which is sufficient to keep the 99% quantile of the virtual server delay limited to 10ms for moderately correlated traffic ($E[L_{burst}] = 10$ packets). Extremely correlated traffic reveals the double flow rate $2 \cdot E[C_f]$ over long time since a mean burst length of $E[L_{burst}] = 100$ packets takes $1.6s$. Even for this extreme case the mean virtual server delay is low ($E[W_v] = 1ms$) and only the quantile is large ($17ms$). Hence, EBAC is able to work with moderately correlated traffic streams.

V. Conclusion

We proposed the Experience-Based AC (EBAC) as a new AC method. Flows signal their possibly overestimated peak rate demands to request a reservation and EBAC performs its AC decision taking an overbooking factor $\varphi(t)$ into account which is based on the time series of observed reservation utilizations. The objective is efficient resource allocation in presence of overestimated peak rates and varying traffic rates. We presented the EBAC concept in detail and explained our performance evaluation methodology.

The calculation of the overbooking factor is influenced by many system parameters such as the applied utilization percentile p_u, the measurement interval length Δ, the devaluation factor d. In addition, the QoS of the admitted flows depends on the safety threshold χ and the maximum link utilization ρ_{max}. We introduced the notion of the virtual spacer delay W_v to estimate the suitability of the system parameters.

We have found that a utilization percentile $p_u = 99\%$ and a measurement interval length $\Delta = 1s$ works well on a 10 Mbit/s link. The devaluation factor can then be safely chosen from the range $[0.9, 1]$. Our simulation results showed that EBAC is adaptive such that QoS can be guaranteed even for traffic with high variability and burstiness. The objective could always be met, i.e. almost the peak-to-mean ratio could be used for overbooking while high QoS could be maintained.

In future work, we plan to investigate the EBAC behavior and to adapt the system parameters for traffic mixes that change over time significantly. We want to study the impact of different measurement and statistic aging methods as well as link sizes. We intend to use typical data traces from real-time applications for further validations.

We plan to extend the EBAC concept from a single link to a network-wide AC [9], to support different traffic classes and to optimize the system parameters for production environments.

REFERENCES

[1] James Roberts, Ugo Mocci, and Jorma Virtamo, *Broadband Network Teletraffic - Final Report of Action COST 242*, Springer, Berlin, Heidelberg, 1996.

[2] R. Gibbens and F. Kelly, "Measurement-Based Connection Admission Control," in 15^{th} *International Teletraffic Congress*, June 1997.

[3] Matthias Grossglauser and David N. C. Tse, "A Framework for Robust Measurement-Based Admission Control," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 293–309, 1999.

[4] Lee Breslau, Sugih Jamin, and Scott Shenker, "Comments on the Performance of Measurement-Based Admission Control Algorithms," in *INFOCOM (3)*, 2000, pp. 1233–1242.

[5] Hans van den Berg and Michel Mandjes, "Admission Control in Integrated Networks: Overview and Evaluation," in *Proceedings 8th International Conference on Telecommunication Systems*, Nashville, US, 2000, pp. 132 – 151, Guter Ueberblick.

[6] Coskun Cetinkaya and Edward W. Knightly, "Egress Admission Control," in *INFOCOM (3)*, 2000, pp. 1471–1480.

[7] Ignacio Más and Gunnar Karlsson, "PBAC: Probe-Based Admission Control," in 2^{nd} *International Workshop on Quality of future Internet Services (QofIS2001)*, September 2001.

[8] Phuoc Tran-Gia Frank Hübner, "An Analysis of Multi-Service Systems with Trunk Reservation Mechanisms," Technical Report, No. 40, University of Würzburg, Institute of Computer Science, April 1992.

[9] Michael Menth, Stefan Kopf, Jens Milbrandt, and Joachim Charzinski, "Introduction to Budget Based Network Admission Control Methods," in 28^{th} *Annual IEEE Conference on Local Computer Networs (LCN2003)*, Bonn, Germany, Oct. 2003.