University of Würzburg
Institute of Computer Science
Research Report Series

Comparison of Border-to-Border Budget Based Network Admission Control and Capacity Overprovisioning

Rüdiger Martin[1], Michael Menth[1], Joachim Charzinski[2]

Report No. 359 Mai 2005

[1] University of Würzburg
Institute of Computer Science
Department of Distributed Systems
Am Hubland, 97074 Würzburg, Germany
{martin,menth}@informatik.uni-wuerzburg.de

[2] Siemens AG
Munich
Germany
joachim.charzinski@siemens.com

Comparison of Border-to-Border Budget Based Network Admission Control and Capacity Overprovisioning

Rüdiger Martin, Michael Menth
University of Würzburg
Institute of Computer Science
Department of Distributed Systems
Am Hubland, 97074 Würzburg, Germany
{martin,menth}@informatik.uni-wuerzburg.de

Joachim Charzinsky
Siemens AG
Munich
Germany
joachim.charzinski@siemens.com

1 Introduction

The traditional Internet offers a best effort service and almost global reachability at low cost. Bulk transfers require high throughput, value added services like telephony or video conference depend on short packet delays and predictable throughput, and precision applications like tele-medicine or tele-robotics cannot even afford packet loss. Therefore, Quality of Service (QoS) in terms of short packet delay and low packet loss is required for next generation networks (NGNs) to support these services.

QoS can be achieved by limiting the traffic volume in the network by admission control (AC) and thereby preventing overload situations. As an alternative, sufficient capacity can be provisioned such that no congestion occurs. This is called capacity overprovisioning (CO). On the one side, many investigations compare blocking probabilities of different AC schemes for which several signalling protocols exist. On the other side, practical experience shows that CO is already applied since the utilization of core networks today is very low [1].

In this paper we quantify the capacity requirements for networks that rely on AC and CO for QoS provisioning, which is crucial for an economic assessment of both approaches. We focus on the simplest method for network AC (NAC), namely on border-to-border (b2b) budget (BBB) based NAC. We consider networks with only high priority real-time traffic which is not elastic and which has a connection structure. This is a typical scenario in, e.g., core networks for cellular systems. We consider first an unrealistically simple static traffic model on a single link. We add overload situations and extend the study to entire networks such that our final results reflect realistic scenarios.

The paper is structured as follows. In Sec. 2, we give a short introduction to AC and CO, discussing related work and our assumptions. Section 3 develops a traffic model and suggests capacity dimensioning methods for AC and CO both for a single link and for entire networks. In Sec. 4, we present the capacity requirements for AC and CO under various networking conditions. Sec. 5 discusses the results and our conclusions.

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.

2 Overview on Admission Control and Capacity Overprovisioning

We give an overview on various aspects of AC by focusing first on the packet level and then on the flow level. Then we consider related work regarding CO and find a suitable level on which we can compare AC and CO.

2.1 Admission Control

QoS can be defined by a loss and a delay parameter. The packet loss probability should be smaller than, e.g., 10^{-6} and the 99.99%-percentile of the waiting time should not exceed a given delay budget DB, i.e., the probability for a packet to wait longer than DB must be smaller than 0.01%. This is achieved by limiting the traffic per transmission resource to avoid overload, i.e., flows request admission for their transportation over certain resources which can be granted or denied. We identify AC methods for a single resource that we call link AC (LAC) and methods that coordinate several resources which we call network AC (NAC). An extensive overview on AC can be found in [2].

2.1.1 Link Admission Control

Link AC (LAC) methods concentrate primarily on the packet level and on a single resource. Thus, traffic descriptors characterize the packet streams by token bucket or dual token bucket parameters to capture the variability of the traffic on two different time scales. They inform the AC entity and the policer about a maximum peak rate and inter-packet distance. This information is used to calculate together with other assumptions the packet loss probability and the expected delay distribution on the link. A generalization and simplification of that approach is the concept of effective bandwidth [3]. It depends on such traffic descriptors and other parameters like the link capacity and assigns a so-called effective bandwidth to any flow request. If the effective bandwidth sum of admitted flows plus the effective bandwidth of a new request exceeds a certain capacity budget, e.g. the link capacity, then the flow is rejected; otherwise it is accepted. Thus, we get the flow blocking probability p_b as additional measure for GoS.

2.1.2 Network Admission Control

Network AC (NAC) methods concentrate on AC for several resources, e.g., for a path consisting of several single links within a network. The link budget (LB) based NAC performs the above described AC successively on each link of the flow's path and it is successful if all AC decisions are positive. This is the most intuitive NAC approach and it has been implemented by many signalling protocols, e.g. by RSVP. The drawback of this method is that information about flows must be kept not only by the ingress router but also by all routers along the path. This increases the administration overhead and it complicates a resilient architecture. The border-to-border (b2b) budget (BBB) based NAC defines capacity budgets for each b2b relationship (v, w) within the network and assigns them a capacity portion. A new flow originating at ingress border router v and destined for egress border router w asks for admission only at its ingress router

v. This ingress router performs AC based on $BBB(v, w)$ like on a single resource. This AC type has been enhanced by resilience mechanisms and it has been successfully implemented within the KING project [4]. Another example for BBB NAC is AC based on label switched paths (LSPs) with fixed capacity. In this work, we compare the required capacity for BBB NAC and CO.

2.2 Capacity Overprovisioning

Capacity overprovisioning (CO) purely relies on provisioning enough bandwidth to meet a desired QoS. The QoS definition from above in terms of packet loss probability and delay budget still holds. As CO does not limit the traffic to avoid overload, all flows are admitted. The link capacities are chosen such that they are very rarely exceeded by the predicted traffic. Like AC, CO can also be combined with different traffic classes by implementing priority scheduling mechanisms. Low priority traffic can use the bandwidth provisioned for high priority traffic under non-overlad situations without additional mechanisms.

2.2.1 Related Work

Bandwidth provisioning procedures differ fundamentally from access to core networks due to the degree of aggregation. Empirical evidence can be found in [5, 6] that core network traffic on the packet level, i.e. the average traffic arrival rate, is modeled well by the Gaussian distribution due to the high level of aggregation. This is clearly not the case in the access due to the limited number of users where the aggregation level is inherently low.

A comparison of AC and CO in access network dimensioning is the topic of [7]. The authors find a clear benefit of AC. Depending on network parameters like blocking probability, packet loss probability and user activity, the number of subscribers for a given access network capacity is substantially higher when AC is used. However, we focus on core networks.

In [5] the network is dimensioned to support latency sensitive traffic. Accordingly, the QoS measure the network is dimensioned for is the probability that the queue length Q of a router exceeds a certain value x: $\mathrm{P}\{Q > x\}$. End-to-end delay requirements of 3ms require only 15% extra bandwidth above the average data rate of the traffic in the highly aggregated Sprint network. Another approach [8] focuses on the probability that the amount of traffic $A(T)$ generated on a link within a specified time interval T exceeds the capacity C of the link: $\mathrm{P}\{A(T) \geq C \cdot T\}$. The authors argue that applications can cope with lack of bandwidth within an application-dependent small interval T if this occurs sufficiently rarely. They develop an interpolation formula that predicts the bandwidth requirement on a relatively short time scale in the order of 1 second by relying on coarse traffic measurements.

Another closely related problem is forecasting of Internet traffic. A recent approach for long-term forecasting can be found in [9]. The authors of [10] combine both tasks to yield an adaptive bandwidth provisioning algorithm. Based on measurements, the

required capacity is predicted and adjusted on relatively small time scales between $4s$ and $2min$. The Maximum Variance Asymptotic (MVA) approach for the tail probability of a buffer fed by an Gaussian input process is used to make the QoS requirement $\mathrm{P}\{delay > D\} < \epsilon$ explicit.

Our work is different from the literature presented here. Our focus is not the development of an AC or CO scheme to adjust the network capacity to the needs of a specific application or specific scenario. We develop a model to qualitatively compare the required capacity for AC versus the required capacity for CO.

2.2.2 Our View on CO for Comparison with AC

Most CO studies use both a flow and a packet level model. The first models the number of flows in the network whereas the second causes the required extra bandwidth above the mean data rate of the traffic.

Both AC and CO can be combined with a packet level model to asses the relation of effective bandwidth to average and peak data rates, and an inadequate packet level model will lead to QoS degradations in both systems.

However, we are primarily interested in comparing AC to CO and not in specific statistical multiplexing schemes. Therefore, we eliminate the packet level by working on effective bandwidths for both systems. This is a prerequisite for a fair comparison.

If the requested rate of all flows on a link exceeds the link bandwidth, all flows are affected by QoS degradation. This view leads to the definition of a new QoS measure for CO: the QoS violation probability p_v which is the time fraction with violated QoS. As all flows are concerned, it should be low and we use an objective value of $p_v = 10^{-6}$ for bandwidth dimensioning in our study.

3 Capacity Dimensioning

Now we describe the capacity dimensioning methods used for AC and CO both for a single link and an entire network.

3.1 Traffic Model

Real-time flows are mostly triggered by human beings. Thus, their inter-arrival time is exponentially distributed [11]. The Poisson model for flow arrivals is also advocated by [12] and current evidence of Poisson inter-arrivals for VoIP call arrivals is given in [13]. Therefore, a flow level model that is characterized by exponentially distributed inter-arrival time and an independently and identically distributed call holding time is appropriate in an evolving multimedia world.

3.1.1 Multi-Rate Traffic

As the request profile is multi-rate in a multi-service network like the Internet, we use a simplified multi-rate model (cf. [2]). We have $n_r = 3$ different request types r_i, $0 \leq i < n_r$ with request sizes $c(r_i) \in \{64, 256, 2048\}$ kbit/s. The mean of the request-type-specific

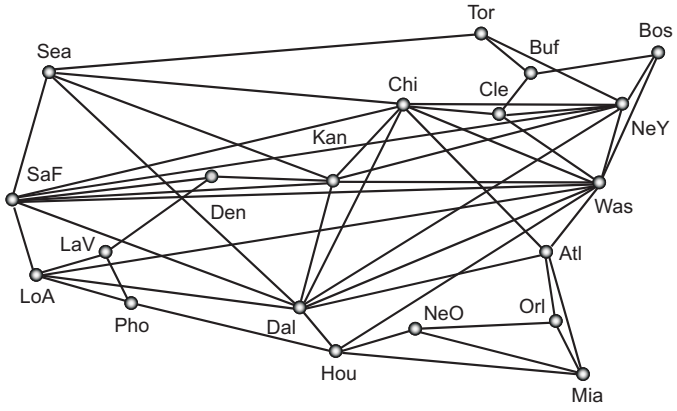

Figure 1: Topology of the test network.

inter-arrival and the mean of the call holding time determine the request-type-specific offered load $a(r_i)$. The overall load is $a = \sum_{0 \leq i < n_r} a(r_i)$. The random variable C_t indicates the requested rate in case of a flow arrival and the request size probability $P\{C_t = c(r_i)\}$ depends on the parameter $t \in [0, 1]$. The statistical properties of the request types are compiled in Table 3.1.1. They are chosen such that we get a constant mean of $E(C_t) = 256$ kbit/s and a coefficient of variation of $c_{var}(C_t) = 2.291 \cdot t$ that depends linearly on t.

request type r_i	$c(r_i)$	$P\{C_t = c(r_i)\}$
r_o	64 kbit/s	$\frac{28}{31} \cdot t^2$
r_1	256 kbit/s	$(1 - t^2)$
r_2	2048 kbit/s	$\frac{3}{31} \cdot t^2$

Table 1: Request type statistics.

3.1.2 Traffic Matrix

The network experiments in this paper are based on the KING [4] reference network given in Fig. 1. All network nodes are both ingress and egress routers. We scale the traffic matrix for the test network with the overall offered load a_{tot}. The generation of the traffic matrix is based on the population of the cities and their surroundings [2]). For two cities v and w with population sizes $\pi(v)$ and $\pi(w)$, the border-to-border (b2b) offered load $a(v, w)$ amounts to

$$a(v,w) = \begin{cases} \frac{a_{tot} \cdot \pi(v) \cdot \pi(w)}{\sum_{x,y \in \mathcal{V}, x \neq y} \pi(x) \cdot \pi(y)} & \text{for } v \neq w, \\ 0 & \text{for } v = w. \end{cases} \quad (1)$$

The average offered b2b load a_{b2b} specifies the overall offered load in the network $a_{tot} = \sum_{v,w \in \mathcal{V}, v \neq w} a(v,w) = |\mathcal{V}| \cdot (|\mathcal{V}| - 1) \cdot a_{b2b}$, where $\mathcal{V}$ is the set of all nodes in the network. We use shortest path routing in our experiments, which is the basis for most Interior Gateway Protocols (IGPs). Our reference populations are given in [2].

3.2 Capacity Dimensioning for AC: $M/G/n - 0$

Capcity dimensioning for AC on a single link with a multi-rate Poisson flow model and the usage of effective bandwidths is the task of finding the capacity n of a multi-rate $M/G/n - 0$ blocking system. The capacity n – the number of basic bandwidth units – must be chosen to accommodate sufficiently many flows in the network to fulfill the desired blocking probability $p_b = 10^{-3}$. The well-known Kaufman/Roberts algorithm presented in [14] computes the blocking probability for a given traffic mix and capacity. Our capacity dimensioning algorithm for AC performs a computational inversion of these formulae in an efficient way [2].

3.3 Capacity Dimensioning for CO: $M/G/\infty$

With CO, the number of flows in the system is not bounded. Therefore, dimensioning for CO on a single link with a multi-rate Poisson model can be done using a $M/G/\infty$ system. We calculate the equilibrium state probabilities of the system. The request types constitute the $k = n_r$ classes for which the k-dimensional state space is described by $X = \{x = (x_0, x_1, \ldots, x_{k-1}) \in \mathbb{N}_0^k\}$. With the class-specific arrival rate λ_i and the class-specific mean holding time $\frac{1}{\mu_i}$ the equilibrium state probabilities are

$$p(x) = \prod_{i=0}^{k-1} \frac{\rho_i^{x_i}}{x_i!} e^{-\rho_i} \tag{2}$$

with $\rho_i = \frac{\lambda_i}{\mu_i}$. The consideration of the request type rates $c(r_i)$ yields the required link capacity $c(x) = \sum_{i=0}^{k-1} c(r_i) \cdot x_i$ of state x. Thus, the required capacity C for the overprovisioned system is

$$C = \min_{C'} \{1 - \sum_{c(x) \leq C'} p(x) \leq p_v\}. \tag{3}$$

This is the smallest capacity such that the rates of the flows crossing the link exceed the link capacity at most with the desired QoS violation probability $p_v = 10^{-6}$. The calculation of the state probabilities is also known as the stochastic knapsack with infinite capacity [15]. Its solution was originally derived for the $M/M/\infty$ system but it is insensitive to the holding time distribution and holds for $M/G/\infty$ systems, too.

3.4 Extension to Networks

The algorithms for link capacity dimensioning must be extended to entire networks.

3.4.1 BBB NAC

If a flow wants to pass the network from node v to w, the BBB NAC checks the single budget $BBB(v,w)$. We dimension the size of the budget with the single link AC algorithm in such a way that $p_b = 10^{-3}$ is achieved. The capacity of a link within the network is the sum of all budgets crossing this link.

3.4.2 CO

For CO, the QoS violation probability on the complete path from source to destination must be at most $p_v = 10^{-6}$. The corresponding probabilities $p_v(l)$ on the individual links are clearly rather positively correlated. An upper bound for p_v on the path is given by $p_v(path) = 1 - \prod_{l \in path}(1 - p_v(l))$ where $l \in path$ denotes the links on the path and $p_v(l)$ is the QoS violation probability on the link. Now we can compute $p_v(l) = 1 - \sqrt[len(path)]{1 - p_v}$ for a given link for every b2b relation and obtain the minimum of these values as the required QoS violation probability on this link. Based on $p_v(l)$ and the aggregate offered load $a(l)$ of all flows traversing link l we can dimension the capacity of each link l in the network.

4 Capacity Requirements for AC and CO

In this section, we compare the capacity requirements for AC and CO. We dimension the capacity for AC such that the blocking probability is $p_b = 10^{-3}$ under normal conditions. As CO cannot prevent overload situations, we dimension the capacity for CO in a very conservative manner such that the QoS violation probability is $p_v = 10^{-6}$. We concentrate first on a single link to understand the basic tradeoffs and then we extend our study to entire networks.

4.1 Single Link with Constant Load

First, we explain economy of scale as it is the key to understand the phenomena in our study. Then, we compare the capacity requirements for AC and CO for a constant load on a single link and consider the strength of the QoS violation by CO. Finally, we enhance the constant offered load scenarios by rare overload situations.

4.1.1 Economy of Scale

In Fig. 2 we dimensioned the required capacity on a single link for AC and a blocking probability of $p_b = 10^{-3}$. The required link capacity is almost proportional to the offered link load, at least for an offered load of 10^3 Erlang or more. The average resource utilization of that capacity by the offered traffic increases with the offered load and expresses the resource efficiency in a natural way. The fact that little offered load leads to low utilization and that large offered load leads to high utilization is a non-linear functional dependency and it is called economy of scale or multiplexing gain.

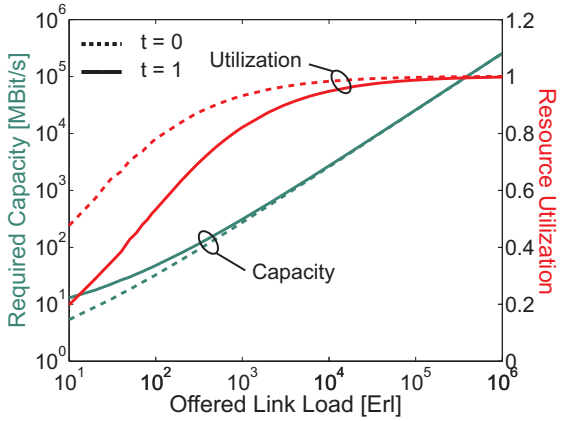

Figure 2: Economy of Scale on a single link for AC.

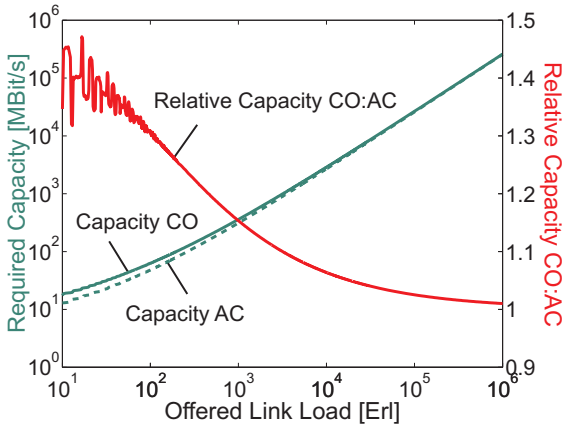

Figure 3: Impact of offered load on capacity requirements for AC and CO.

Figure 2 also shows that traffic with highly variable request sizes ($t=1$) requires more capacity. In the following, we use only highly variable traffic ($t=1$) due to the multi-rate nature of Internet traffic.

4.1.2 Comparison for Constant Offered Load

Figure 3 indicates the required capacity for AC and CO depending on the offered load. The ratio of both curves shows that they differ significantly only at low offered load. The oscillations here and in the following figures are due to the granularity limitation of the bandwidth and request size quantities. In particular, CO requires less than 5% additional capacity at 20^4 Erlang or more. The capacity for CO with $p_v = 10^{-6}$ equals approximately the capacity for AC with $p_b = 10^{-6}$. Due to economy of scale, this requires only slightly more capacity than AC with $p_b = 10^{-3}$ for large offered load [2].

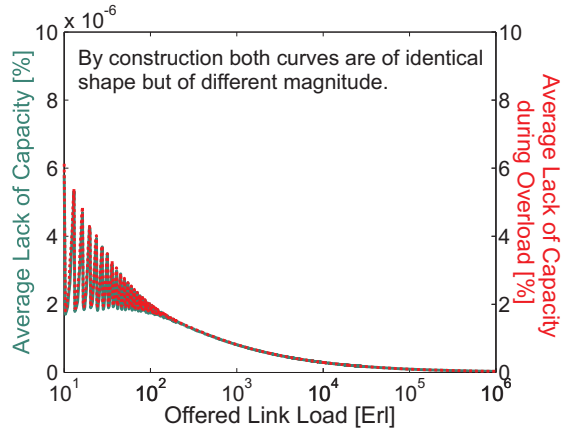

Figure 4: Lack of capacity for CO as time average and conditioned on overload situations.

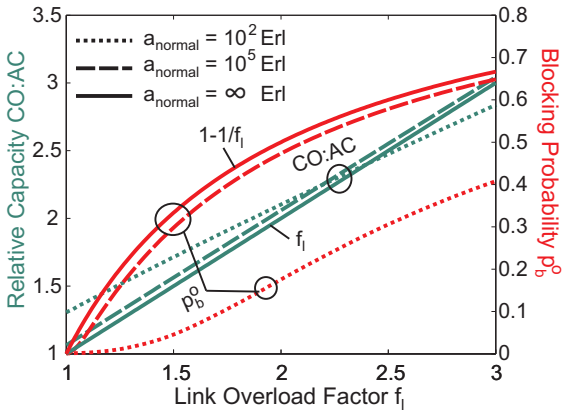

Figure 5: Impact of rare overload on capacity requirements and blocking.

Another concern is the degree to which QoS is violated. We can capture that by the lack of capacity in overload situations. The average lack of capacity over time is $E[L] = \sum_{c(x)>C, x\in X} (c(x) - C) \cdot p(x)$ where x is the state vector of flows in the system. Figure 4 illustrates this value in percent related to the provisioned capacity for CO. It is in the order of 10^{-6} for all considered scenarios because we have provisioned so much capacity that overload occurs very rarely with $p_v = 10^{-6}$. In case of overload situations, the lack of capacity is 6 orders of magnitude larger but it is not larger than 6%. The reason for that is the constant offered load in our experiment, which allows only small statistical oscillations but does not model occasional hot spots due to increased content attractiveness at certain locations.

4.1.3 Comparison for Rare Overload

Rare overload situations can occur due to hot spot scenarios caused by singular events. We capture this intuition by keeping the offered load at a normal level a_{normal} for a time fraction $\frac{364}{365}$ and increase it to an overload level of $a_{overload}$ for a short time fraction of $\frac{1}{365}$. In the following, we call the ratio $f_l = \frac{a_{overload}}{a_{normal}}$ the link overload factor. Note that the variability of the time series of offered load is still quite moderate with a coefficient of variation of 0.104 for an overload factor of $f_l = 3$. The capacity for AC is dimensioned based on a_{normal} since an increased blocking probability p_b can be tolerated for a short time interval whereas the capacity for CO is dimensioned based on $a_{overload}$ since CO cannot avoid congestion in severe overload situations. For very high offered load, a utilization of almost 100% can be achieved for AC. In this case, the ratio of the capacity requirements for CO and AC scales with the overload factor f_l and the blocking probability p_b during overload situations scales with $1 - \frac{1}{f_l}$ which are both analytical values. Figure 5 shows these performance metrics for an offered load $a_{normal} = 10^2$ Erl and $a_{normal} = 10^5$ Erl. Regardless of the offered load, the ratio of the capacity requirements for CO and AC follows quite well the overload factor f_l while the blocking probability p_b^o depends also significantly on the offered load. The fact that the CO:AC capacity requirement curves cross is due to the stronger impact of multiplexing gain when the offered link load is low.

Figure 5 shows that the blocking probabilities p_b^o for $a = 10^{\{2,5\}}$ Erl are below the analytical value $1 - \frac{1}{f_l}$. In overload situations, 100% of the available bandwidth is used to transport traffic. If a high average utilization can be achieved under normal conditions, only relatively little extra capacity is available to accommodate extra traffic. Therefore, the blocking probability increases with offered load. Hence, QoS can be maintained with AC in overload situations and the effective blocking probability is significantly smaller than the simple analytical rule of thumb for a moderately aggregated traffic. However, the additional capacity for CO scales quite well with the assumed overload factor. These are the results from the analysis but there is another practical problem. The overload factor f_l is unknown and must be overestimated to guarantee QoS. This safety margin cannot be covered by our analysis but increases again the additional capacity requirements for CO.

4.2 Networks with Constant Load

We have studied the single link to understand the basic tradeoffs. If we proceed to entire networks, we have to take the impact of NAC into account which entails two different types of multiplexing gain.

A Traffic corresponding to a single b2b relationship (v, w) is carried within a single $BBB(v, w)$. Its offered load $a(v, w)$ determines the required capacity of that BBB and its utilization. Here, BBB NAC and CO can profit from economy of scale for budgets.

B Traffic corresponding to different b2b relationships is carried over a single link.

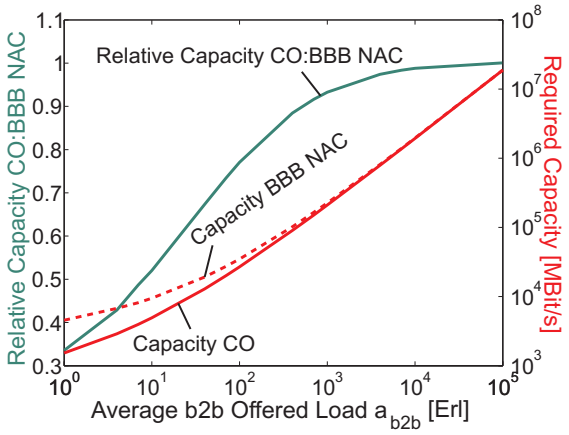

Figure 6: Impact of offered load on the capacity requirements for CO and BBB NAC and their ratio in the KING testbed.

The capacity requirement for the considered link is the sum of the capacities of the respective BBBs. With CO, in contrast, the required capacity of a link l is calculated on the basis of its overall offered traffic load $a(l)$ which is a larger aggregation level than the load pertaining only to a single budget. We call that economy of scale due to link sharing. It can be exploited by CO but not by BBB NAC.

Fig. 6 shows the capacity requirements for CO and AC and their ratio depending on the average offered b2b load a_{b2b}. CO requires significantly less capacity than AC for low offered load, and the capacity requirements are about the same for an offered load of 10^4 or more. We get this counterintuitive result because AC can exploit economy of scale only on the budget level (A) but not due to link sharing (B). The fact that p_v takes more stringent values than p_b plays obviously only a minor role in entire networks.

4.3 Networks with Rare Overload

The comparison of the capacity requirements is now enhanced by rare overload situations but it is still based on constant total offered load. AC can provide QoS also in such scenarios and a temporary increase in blocking can be accepted. Hence, capacity dimensioning is based on the normal traffic matrix. CO requires sufficient additional capacity for all overload conditions. Therefore, the link capacity must be provisioned for CO such that the maximum expected load can be carried for any overload scenario.

4.3.1 Overload Model for Networks

We model overload in entire networks quite conservatively by single hot spots whereby the overall offered load in the network does not increase, i.e., we change only the structure of the traffic matrix. We increase the traffic attraction of a single city v by a hot spot

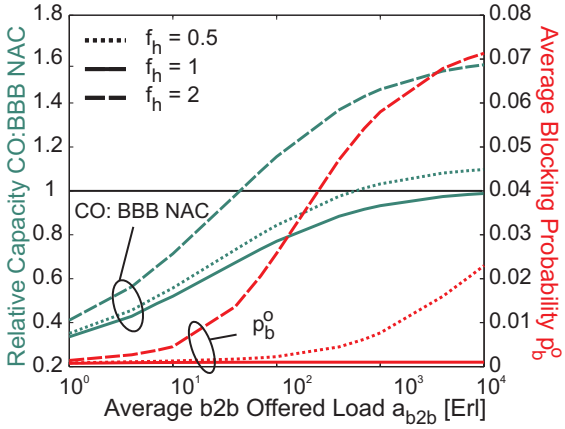

Figure 7: Impact of offered load and hot spot factor on the ratio of capacity requirements for CO and BBB NAC, and blocking probabilities under overload conditions in the KING testbed.

factor f_h, which is expressed by a modified population function

$$\pi^v_{overload}(w) = \begin{cases} \pi(w) & \text{if } w \neq v \\ f_h \cdot \pi(w) & \text{if } w = v \end{cases}. \tag{4}$$

For every potential single hot spot $v \in \mathcal{V}$, a traffic matrix is generated proportionally to $\pi^v_{overload}$ with the same overall offered load in the network.

4.3.2 Comparison of Capacity Requirements for Rare Over- and Underload

Figure 7 shows the relative network capacity CO:BBB NAC and the blocking probability p_b^o depending on the average offered b2b load a_{b2b} for $f_h \in \{0.5, 1, 2\}$. For a hot spot factor of $f_h = 2$, CO already requires more capacity than BBB NAC for an offered b2b load $a_{b2b} = 70$ Erl and it needs 60% more capacity than BBB NAC for high offered load. During overload, the blocking probability is 7% – averaged over all b2b relationships – and it is at most 45.4% for a few b2b aggregates in some scenarios. Depending on the network operation policy, these values are well acceptable.

With $f_h = 0.5$, a single node looses global attractiveness, which means that the relative importance of all other cities is slightly increased. This causes a smooth shift of offered load in the traffic matrix from this city to other cities. It raises the capacity requirements for CO and the blocking probabilities for the BBB NAC only slightly in contrast to the previous experiment.

Figure 8 shows the impact of the hot spot factor f_h on the ratio of the capacity requirements for CO and BBB NAC as well as the corresponding average blocking probability p_b^o for an offered b2b load $a_{b2b} \in \{10^1, 10^3\}$ Erl. For the single link experiment, we could easily predict that the capacity requirements for CO scale with f_l. In case of a traffic shift due to increased attractiveness of a single node within an entire network, the effect

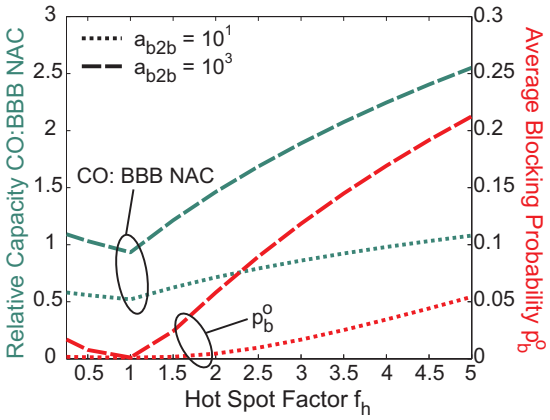

Figure 8: Impact of the hot spot factor f_h on capacity requirements and blocking.

of the hot spot factor f_h is significantly smaller. One reason is that BBB NAC can exploit economy of scale only within b2b budgets (cf. A above). This applies in particular for $a_{b2b} = 10$ and $a_{b2b} = 10^3$ Erl but not for clearly larger values of the offered b2b load. For $a_{b2b} = 10^3$ Erl and $f_h = 3$, the additional capacity requirements are about 200% for the single link experiment while they amount to only 150% for the network experiment. Another reason is the following. In a situation with an increased attractiveness of node v, the links leaving from and leading to v require most additional capacity. But they carry also transit traffic whose rate is rather slightly decreased by v's increased attractiveness. Hence, the increase of the capacity requirements of those links depends on a mixture of slightly decreased transit traffic rates and significantly increased rates for hot spot traffic. Therefore, their additional capacity requirements are smaller for a given hot spot factor f_h than the additional capacity requirement of a single link with a link overload factor f_l.

Figure 9 shows the maximum values for the relative capacity requirements of CO compared to BBB NAC on all single links within the network. These maximum relative capacities are significantly larger than for the entire network. For example, the network requires only 150% more capacity for CO than for AC in case of a hot spot factor of 400%, but some links require 300% more capacity. Hence, the additional capacity varies significantly among the links and the exact amount depends on the network topology, the traffic matrix, and the routing. Therefore, determining the appropriate degree of overdimensioning for individual links in a network is a non-trivial task for which our analysis can be useful. It may be applied, e.g., to provision Differentiated Services networks [16], the base architecture for the future Internet, which will be a multi-service network with high and low priority traffic and suitable scheduling mechanisms. If the fraction of high priority traffic is low, bandwidth overprovisioning can be done for high priority traffic and the required excess capacity may be used under normal conditions to transport low priority traffic. In case of overload in the high priority traffic class, low priority traffic is swamped out.

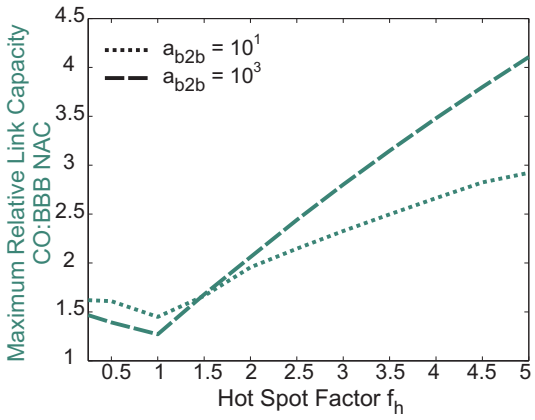

Figure 9: Impact of the overload on the maximum additional relative link capacity requirements.

Finally, we would like to point out that all these results were obtained by a mere traffic shift while the overall offered load in the network has been kept constant.

5 Discussion and Conclusion

Our results show that capacity overprovisioning (CO) requires only 5% more capacity on a single link than border-to-border (b2b) budget (BBB) based network admission control (AC) if we consider constant offered load. As this assumption is not realistic, we took temporary overload of up to 300% of the normal traffic into account. In this case, AC can guarantee QoS for admitted flows at constant capacity at the expense of higher blocking probability, whereas CO needs 200% more capacity. We extended this experiment to entire networks to better motivate the overload scenario. We assumed the traffic matrix to be proportional to the population of the catchment area of a router. To solicit overload, we kept the overall traffic in the network constant, increased the attractiveness of a single city by up to 400% by increasing the population by that factor, and performed that experiment for every city. This models realistic temporary hot spots without increasing the overall traffic volume. Here, CO requires 60% (150%) more capacity than AC for a hot spot factor of 100% (400%) while AC reacts with a blocking probability of 7% (25%).

Based on our temporary hot spot model, we could prove considerable bandwidth savings by AC compared to CO. AC, however, requires a substantial amount of signalling, coordination and interoperation that is not yet implemented in most networks. An economic assessment must take this into account.

We showed that the required degree of overdimensioning varies among the links within a network and, therefore, our analysis can be useful to determine the appropriate additional capacity for individual links for a given networking scenario with suitable assumptions regarding overload. This approach can be of particular interest for Differentiated Services networks [16].

Currently, we are working on a comparison between CO and link-by-link (LB) NAC. LB NAC is wider spread than BBB NAC and its resource utilization differs significantly. In addition, we intend to integrate resilience aspects into our work.

Acknowledgment

The authors would like to thank Prof. Tran-Gia for the stimulating environment which was a prerequisite for that work.

References

[1] A. Odlyzko, "Data Networks are Lightly Utilized, and will Stay that Way," *The Review of Network Economics*, vol. 2, September 2003.

[2] M. Menth, *Efficient Admission Control and Routing in Resilient Communication Networks*. PhD thesis, University of Würzburg, Faculty of Computer Science, Am Hubland, July 2004.

[3] F. P. Kelly, *Stochastic Networks: Theory and Applications*, vol. 4, ch. Notes on Effective Bandwidths, pp. 141 – 168. Oxford University Press, 1996.

[4] C. Hoogendoorn, K. Schrodi, M. Huber, C. Winkler, and J. Charzinski, "Towards Carrier-Grade Next Generation Networks," in *ICCT*, (Beijing, China), April 2003.

[5] C. Fraleigh, F. Tobagi, and C. Diot, "Provisioning IP Backbone Networks to Support Latency Sensitive Traffic," *IEEE Infocom 2003, San Francisco, USA*, April 2003.

[6] J. Kilpi and I. Norros, "Testing the Gaussian Approximation of Aggregate Traffic," in *Internet Measurement Workshop*, (Marseille, France), 2002.

[7] G. Van Hoey and D. De Vleeschauwer, "Benefit of Admission Control in Aggregation Network Dimensioning for Video Services," *Networking 2004, Athens, Greece*, May 2004.

[8] H. Van den Berg *et al.*, "QoS-aware Bandwidth Provisioning for IP Network Links," Tech. Rep. PNA-E0406, Centrum voor Wiskunde en Informatica, The Netherlands, June 2004.

[9] D. Papagiannaki, N. Taft, Z. Zhang, and C. Diot, "Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models," in *IEEE Infocom*, (San Francisco, USA), April 2003.

[10] H. Tuan Tran and T. Ziegler, "Adaptive Bandwidth Provisioning with Explicit Respect to QoS Requirements," *Quality of Future Internet Services (QoFIS), Sweden*, October 2003.

[11] V. Paxson and S. Floyd, "Wide-Are Traffic: The Failure of Poisson Modelling," *IEEE/ACM Transactions on Networking*, June 1995.

[12] J. W. Roberts, "Traffic Theory and the Internet," *IEEE Communications Magazine*, vol. 1, pp. 94–99, January 2001.

[13] T. Dinh, B. Sonkoly, and S. Molnár, "Fractal Analysis and Modeling of VoIP Traffic," in *Proceedings of Networks 2004*, (Vienna, Austria), pp. 123 – 130, June 2004.

[14] J. Roberts, U. Mocci, and J. Virtamo, *Broadband Network Teletraffic - Final Report of Action COST 242*. Berlin, Heidelberg: Springer, 1996.

[15] K. W. Ross and D. H. K. Tsang, "The Stochastik Knapsack Problem," *IEEE Transactions on Communications*, vol. 37, no. 7, pp. 740–747, 1989.

[16] X. Xiao and L. M. Ni, "Internet QoS: A Big Picture," *IEEE Network Magazine*, vol. 13, pp. 8–18, March 1999.