Comparison of Link-by-Link Admission Control and Capacity Overprovisioning

Ruediger Martin [1], Michael Menth [1], and Joachim Charzinski [2]

[1] Dept. of Distributed Systems, Inst. of Computer Science, University of Würzburg
Am Hubland, D-97074 Würzburg, Germany
{martin,menth}@informatik.uni-wuerzburg.de
[2] Siemens AG, Munich, Germany
joachim.charzinski@siemens.com

Abstract. There are two basic approaches to achieve Quality of Service (QoS) for communication networks: admission control (AC) and capacity overprovisioning (CO). AC requires less capacity than CO because it can block excess traffic, preventing it from disturbing the QoS of admitted flows. CO on the other hand is simple and cheaper to implement than AC, which makes it an attractive option for Internet service providers that do not have an AC infrastructure yet, at the risk of the network being overwhelmed by excess traffic in rare cases. In this paper we contribute further insights to this discussion by quantifying the ratio of required capacity for CO:AC in certain networking scenarios including rare overload and hot spot scenarios.

Keywords: QoS, admission control, capacity overprovisioning

1 Introduction

Today's Internet offers (almost) global reachability at low cost. On the one hand, the "Best-Effort" delivery of packets does not guarantee any Quality of Service (QoS) level, but it is often sufficient even for high bit rate transfers. On the other hand, there is an ever increasing tendency to move value added services like telephony or video conferencing onto the Internet, which require bounded packet delay and predictable throughput. High precision applications like tele-surgery, tele-robotics or tele-music additionally require extremely low packet loss rates. Therefore, QoS in terms of short packet delay and low packet loss will be required in future versions of the Internet, so called "next generation networks" (NGNs), to support these services.

QoS can be achieved by introducing an admission control (AC) infrastructure into the network. Demanding applications reserve network resources before transmitting traffic over the network, at least in high-QoS classes. AC blocks out reservations when the capacity does not suffice to guarantee the required QoS level both for the new reservation

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.

and for the already established reservations. Of course, the blocking probability has to be small in order not to annoy customers such that AC admits all reservations most of the time. This fact is exploited by the capacity overprovisioning (CO) approach that simply trades in complexity cost for bandwidth cost. Instead of blocking flows in overload situations, the CO approach tries to provide enough bandwidth so that the resulting QoS violations are low enough to be tolerated by the relevant applications.

Many investigations compare blocking probabilities and required capacity for different AC schemes, for which several signalling protocols exist but none has been introduced on a large scale on the Internet. Introducing an AC architecture for the Internet would require significant investment into control plane elements and operation. Thus, currently all high-quality providers apply CO, leading to a low utilization of core networks [1].

In this paper, we quantify the capacity requirements for networks relying on AC or CO for QoS provisioning to economically assess both approaches. We assume the traditional AC approach of deciding link by link if a reservation can be admitted. We concentrate on the high priority traffic with given bandwidth requirements per flow and compute the network capacity required for it. To explain the fundamental relations between flow bandwidth and link capacity requirements, we evaluate first the single link scenario. Then, we extend the study to an entire network domain to reflect realistic scenarios.

The paper is structured as follows. In Sec. 2, we give a short introduction into AC and CO, discussing related work and our assumptions. Sec. 3 states the traffic models used for multi-rate traffic and hot spot scenarios. In Sec. 4, the capacity requirements for AC and CO investigated under various networking conditions are presented. Sec. 5 discusses the results and our conclusions. The capacity dimensioning methods used for AC and CO both for a single link and for entire networks are documented in the appendix.

2 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. Capacity dimensioning algorithms for both schemes are documented in the appendix.

2.1 Admission Control

QoS can be defined by a loss and a delay parameter. For example, the packet loss probability should be smaller than 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 to be transported over certain resources which can be granted or denied. In principle, there are numerous AC methods for operating a network, which are discussed in detail in [2]. Here, we focus on link-by-link AC where separate AC decisions are performed for all links in the path of a flow.

Link Admission Control A single link has limited bandwidth and buffer space and a number of flows competes for these resources. The packet level properties of the admitted flows determine the achieved QoS. Thus, traffic descriptors are usually employed to inform a policer about a maximum peak rate and inter-packet distance, they characterize the packet streams by token bucket or dual token bucket parameters to capture the variability of the traffic on two different time scales. This information is used together with other assumptions to calculate 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 correlates these traffic descriptors and other parameters such as the link capacity and assigns a so-called *effective bandwidth* to each 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. The probability that a flow is rejected is called the flow blocking probability p_b, which is an additional "Grade of Service" (GoS) measure for performance.

Admission Control in a Network Protocols such as the Resource Reservation Protocol (RSVP) in the Integrated Services context implement a link-by-link AC scheme where for each link a flow traverses the corresponding resource reservation request is handled by the link AC procedure outlined above. If all AC instances on a flow's path admit the request, it is admitted; otherwise, it is blocked. To implement the scheme, information about all flows must be managed for all resources traversed by the flows. This can be implemented as an AC instance in each router, or by a bandwidth broker that manages reservation state and routing information for each link in a network domain in a central entity.

2.2 Capacity Overprovisioning

For GoS reasons, enough network capacity has to be provided in a network with AC to keep blocking probabilities at an acceptable level, which usually is considered to be $p_b \leq 10^{-3}$. Only if the network blocks traffic, the same network without an AC instance would cause problems. Thus, with a little bit of additional network capacity, the probability of these overload states can be further reduced, so that the overload effects can be neglected because they reach the same level of QoS distortions as defined on the packet level for an AC system. Capacity overprovisioning (CO) attempts to exploit that fact.

The AC-based QoS definition 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 the predicted traffic causes overload very rarely. The fundamental issue with the CO approach is that the corresponding QoS violation probability p_v is tightly coupled to the given input traffic model. That means, in the case of a sudden increase of traffic or a shift of traffic patterns, the QoS violation probability is increased in a network with CO whereas the additional traffic is blocked in a network with AC. This increases the blocking probability but keeps the originally assured QoS level for accepted reservations.

Note that the CO principle does not contradict the usage of multiple packet scheduling classes in a network, as it is considered simple to provide a higher priority class with access only for users that pay an increased flat rate or even combined with volume based charging. The main difference between CO and AC is the necessity of the AC infrastructure and the corresponding signalling protocols required to convey traffic characteristics in reservation requests from users to AC instances and the reply from the AC instances indicating if the reservation request was admitted or rejected.

2.3 Related Work

Bandwidth provisioning procedures differ fundamentally from access to core networks due to the degree of aggregation. Empirical evidence can be found in [4, 5] that core network traffic on the packet level, i.e. the packet arrival process, is modeled well by the Gaussian distribution due to the high level of aggregation. This is less the case in the access region where the aggregation level is inherently low due to the limited number of users.

A comparison of AC and CO in access network dimensioning is the topic of [6]. 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 with AC. However, we focus on core networks.

In [4] a core 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, $\mathrm{P}\{Q > x\}$. To satisfy end-to-end delay requirements as low as 3ms, the network requires only 15% extra bandwidth above the average data rate of the traffic in the highly aggregated Sprint network. Another approach [7] 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 time 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 *s* 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 [8]. The authors of [9] 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 4 *s* and 2 *min*. The Maximum Variance Asymptotic (MVA) approach for the tail probability of a buffer fed by an Gaussian input process makes the QoS requirement $\mathrm{P}\{delay > D\} < \epsilon$ explicit.

In an earlier paper [10], we compared the required capacity for CO and a *Network* Admission Control (NAC) scheme working on border-to-border bandwidth budgets developed within the KING project [11]. The fundamentally different architectural approaches of the link-by-link AC investigated here and the border-to-border NAC scheme investigated in [10] leads to significant changes in the CO:AC capacity tradeoff in network domains.

Our focus in this paper is different from the literature presented above. We do not develop an AC or CO scheme to adjust network capacity to a specific application or scenario. In contrast, we develop a model to quantitatively compare the required capacity for AC and CO. In particular, we study the resource requirements for different kinds of hot spot scenarios, i.e. for varying traffic matrices. This approach provides additional insight into one of the major ideological disputes between AC and CO proponents, the additional level of security offered by AC in case the traffic deviates from the planned values.

The Basis for a Comparison between CO and AC Most CO studies use a flow and a packet level model. The former models the number of flows in the network whereas the latter models the required extra bandwidth above the mean data rate of the traffic.

When comparing AC to CO, the absolute amount of extra bandwidth above the *average* data rate of the traffic is not the characteristic of primary interest. This is essentially the purpose of the packet level model. An inadequate packet model leads to QoS degradation in both systems. Thus, we treat the packet level as comparable in both cases and use effective bandwidths in both cases as a common basis for a fair comparison.

We use the following QoS and GoS limits for our comparison: in the AC case, the network capacity is dimensioned for a target end-to-end blocking probability of $p_b = 10^{-3}$. For the CO case, a flow-level QoS measure is introduced as follows: In the overload case, the QoS degradation pertains to all flows. Thus, the QoS measure of interest is the QoS violation probability p_v, which is the time fraction with violated QoS. As argued in Sec. 2.2, p_v should be in the same order of magnitude as the packet loss or delay quantile values deemed to be acceptable on the packet level. Thus, we use a target value of $p_v = 10^{-6}$ to dimension the network capacity in the CO case. Details on the algorithms used for capacity dimensioning for AC and CO are given in the appendix.

3 Traffic Models for Overload Scenarios

We describe the capacity dimensioning methods both for single links and entire networks.

3.1 Traffic Model

Real-time flows are mostly triggered by human beings. Thus, their inter-arrival time is exponentially distributed [12]. The Poisson model for flow arrivals is also advocated by [13] and current evidence of Poisson inter-arrivals for VoIP call arrivals is given in [14]. 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 a multimedia world.

Multi-Rate Traffic We use a simplified multi-rate model from [2] as profile for the requested rate. Since we consider a future multi-rate Internet, we parameterize the request size C_t with $t = 1$, which leads to the following characteristics. We have two different request types with request sizes $c(r_0)=64$ kbit/s and $c(r_2)=2048$ kbit/s. They occur with a probability of $p_r(r_0)=\frac{28}{31}$ and $p_r(r_2)=\frac{3}{31}$ such that the mean rate is $E(C_1)=256$ kbit/s and the coefficient of variation is $c_{var}(C_1)=2.291$.

Traffic Matrix The network experiments in this paper are based on the KING [11] reference network given in Fig. 1. All network nodes are both ingress and egress routers. We use shortest path routing in our experiments because it is the basis for most Interior Gateway Protocols (IGPs).

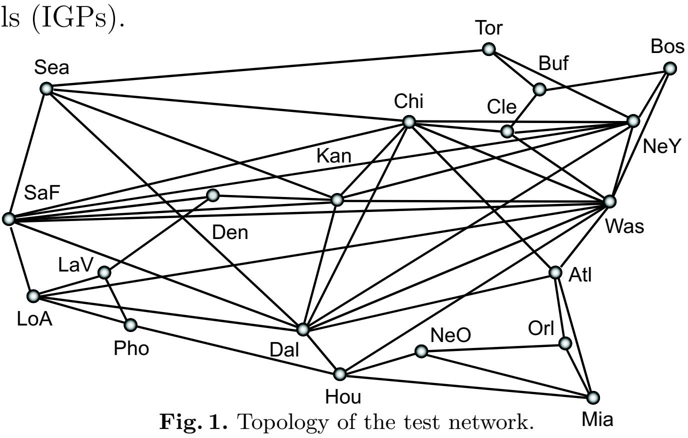

Fig. 1. Topology of the test network.

We scale the traffic matrix for the test network with the overall offered load a_{tot}. The generation of the 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 amounts to $a(v,w) = \frac{a_{tot} \cdot \pi(v) \cdot \pi(w)}{\sum_{x,y \in \mathcal{V}, x \neq y} \pi(x) \cdot \pi(y)}$ for $v \neq w$ and to zero for $v = w$. 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.

3.2 Overload Scenarios

The question whether QoS can be achieved with pure CO depends on the predictability of the traffic and its variation. We first propose a simple overload model for a single link scenario. As this is not realistic for highly aggregated core networks, we suggest then a more complex overload model for a network that keeps the overall network load constant.

Rare Overload Scenarios on a Single Link 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 $\frac{1}{365}$. 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 $f_l = 3$.

Hot Spots in Networks We model overload in entire networks quite conservatively by 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 the cities in the set $\mathcal{H}$ by a hot spot factor f_h, which is expressed by a modified population function $\pi^{\mathcal{H}}_{overload}(w) = \begin{cases} \pi(w) & \text{if } w \notin \mathcal{H} \\ f_h \cdot \pi(w) & \text{if } w \in \mathcal{H} \end{cases}$. Then, the corresponding traffic matrix is generated proportionally to the population function $\pi^{\mathcal{H}}_{overload}$.

4 Capacity Requirements for AC and CO

First, we review our results for a single link to understand the basic tradeoffs [10] and then we extend our study to entire networks.

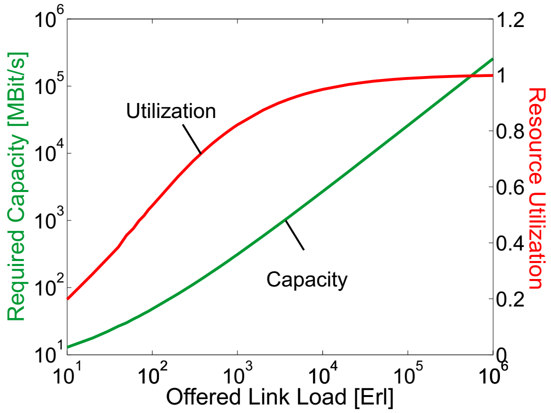

Fig. 2. Economy of Scale on a single link for AC.

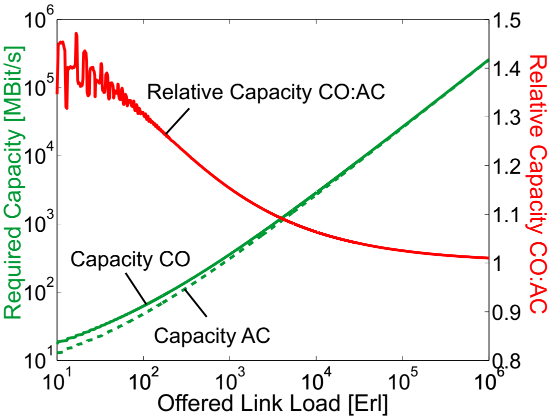

Fig. 3. Impact of offered load on capacity requirements for AC and CO.

4.1 Single Link with Constant Load

We explain economy of scale as the key to understand the phenomena in our study. Then, we compare the capacity requirements for AC and CO for constant load on a single link and, finally, we enhance the constant offered load scenarios by rare overload situations.

Comparison for Constant Offered Load 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. 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.

The ratio of both capacity curves in Fig. 3 also shows that they differ significantly only at low offered load. The oscillations here are due to the granularity limitation of the bandwidth and request size quantities. In particular, CO requires less than 5% additional capacity at 10^4 Erlang or more. 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.

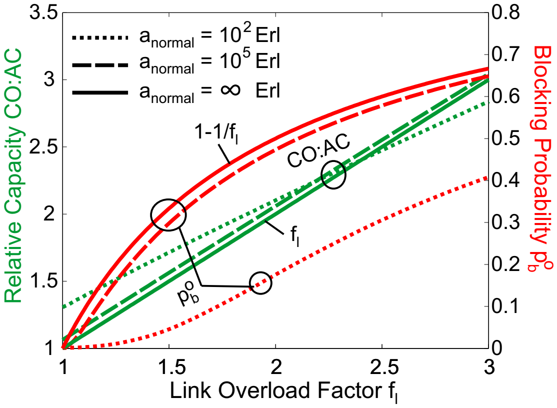

Fig. 4. Impact of rare overload on capacity requirements and blocking.

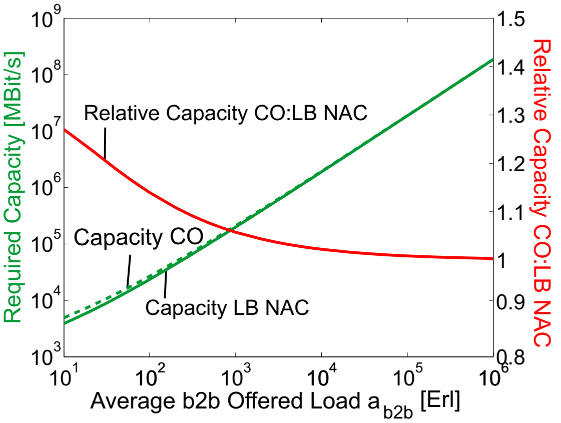

Fig. 5. Impact of offered load on the capacity requirements for CO and LB NAC and their ratio in the KING testbed.

Comparison for Rare Overload As mentioned above, we model rare overload by the link overload factor f_h. 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 4 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 the overload factor f_l quite well 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 impact of multiplexing gain when the offered link load is low [10].

Figure 4 also shows that the blocking probabilities p_b 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.

4.2 Networks with Constant Load

We have studied the single link to understand the basic tradeoffs and now proceed to entire networks. Figure 5 shows the capacity requirements for CO and LB NAC and their ratio depending on the average offered b2b load a_{b2b}. CO requires about 30% more capacity than AC for an offered load of $a_{b2b} = 10$ Erl, and the capacity requirements are about the same for CO and AC for an offered load of 10^4 Erl or more. This conforms with the single link experiment (cf. Fig. 3) but due to the aggregation of many b2b aggregates on a single link, the relative difference between the capacity requirements is smaller.

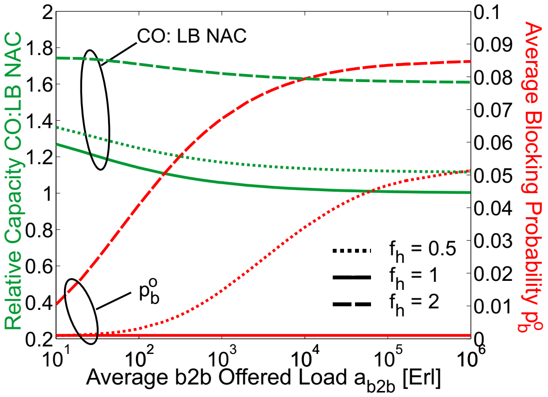

Fig. 6. Impact of offered load and single hot spots on the ratio of capacity requirements for CO and LB NAC, and blocking probabilities under overload conditions in the KING testbed.

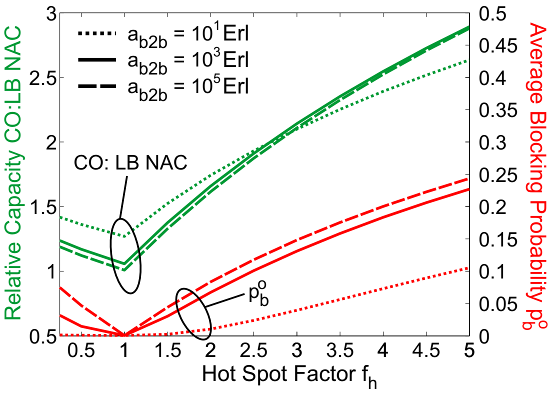

Fig. 7. Impact of the hot spot factor f_h in single hot spot scenarios on capacity requirements and blocking.

4.3 Networks with Rare Overload

The comparison of the capacity requirements is now enhanced by rare overload situations caused by local hot spots while keeping the total offered load constant. AC is designed for providing QoS in exactly such scenarios and a temporary increase of the blocking can be accepted. Therefore, the network capacity is dimensioned for the normal traffic matrix. In contrast, CO requires sufficient capacity to meet $p_v = 10^{-6}$ in all overload conditions. Therefore, the network capacity must be provisioned for CO such that the maximum expected load can be carried in any overload scenario. We first consider single hot spots only to better isolate the effects of this extended traffic model.

Capacity Requirements for Single Hot Spots Figure 6 shows the relative network capacity CO:AC and the blocking probability p_b^o depending on the average offered b2b load a_{b2b} for single hot spots and $f_h \in \{0.5, 1, 2\}$. For a hot spot factor of $f_h = 2$, CO requires 74% more capacity than AC for an offered b2b load $a_{b2b} = 10$ and it still needs additional 61% for high offered load of 10^6 Erl. During overload, the blocking probability is 8.5% – averaged over all b2b relationships.

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 for AC only slightly in contrast to the previous experiment.

Figure 7 shows the impact of the hot spot factor f_h on the ratio of the capacity requirements for CO and LB NAC as well as the corresponding average blocking probability p_b^o for an offered b2b load $a_{b2b} \in 10^{\{1,3,5\}}$ Erl. In a situation with an increased attractiveness of node v, the links leaving from and leading to v carry traffic aggregates whose offered load rate scales with f_h. Since the total offered load remains constant, they carry also transit traffic whose rate is rather slightly decreased. 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. From this analysis we can predict that the required relative overprovisioning depends on the network topology, which has an impact on the routing.

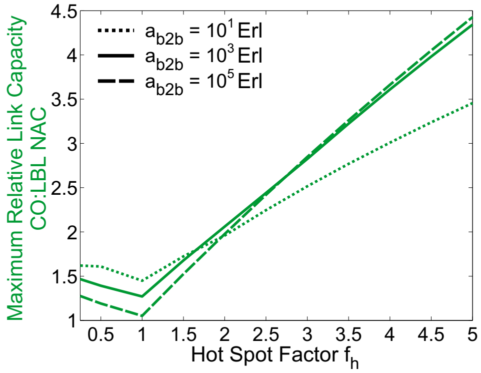

Fig. 8. Impact of the overload in single hot spot scenarios on the maximum additional relative link capacity requirements.

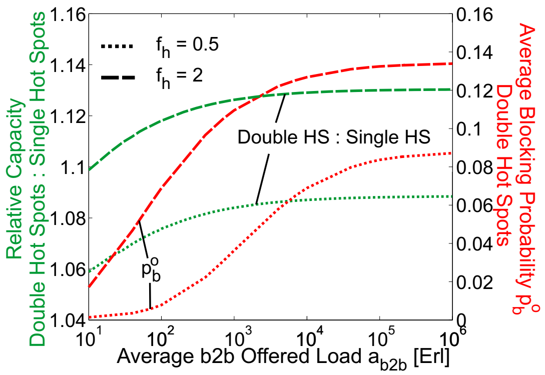

Fig. 9. Impact of double hot spots on the capacity requirements for CO.

Figure 8 shows the maximum values for the relative capacity requirements of CO compared to LB NAC on single links within the network. These maximum relative capacities per link are significantly larger than for the entire network. They result from links with only little transit traffic. For example, the network requires only 190% more capacity for CO in case of a hot spot factor of 400%, but individual links require 340% more capacity. The reciprocals of these values yield practically an upper bound to which utilization overprovisioned networks should be loaded in normal situations. This observation is crucial for the application of CO in the Differentiated Services architecture [15].

Capacity Requirements for Double Hot Spots Analogous to the single link overload model, we assume a hot spot probability for a single node of $p_h = \frac{1}{365}$. Under the further assumption of independence, any n simultaneous hot spots occur with the probability of $\binom{|V|}{n} \cdot (\frac{1}{365})^n \cdot (\frac{364}{365})^{|V|-n}$ where $|V| = 20$ in the KING network. Hence, a single hot spot occurs with the probability of $\approx 5.2 \cdot 10^{-2}$, a double hot spot with the probability of $\approx 1.4 \cdot 10^{-3}$. This is orders of magnitude larger than p_v. Therefore, double hot spot scenarios must be taken into account. We assume a common hot spot factor f_h for all hot spots. Figure 9 shows the ratio of the capacity requirements for single and double hot spot scenarios and the blocking probability p_b^o for double hot spot scenarios. Double hot spots require only slightly more capacity, i.e. up to 13% – even for the hot spot factor $f_h = 2$. The blocking probability during the increased overload by double hot spots remains significantly below 14% – averaged over all b2b aggregates. As the total load in the network is kept constant, the structure of the traffic matrix is further disturbed but both hot spots mutually decrease their impact relative to the single hot spot scenario. Thus, CO requires 84% more capacity than AC for high offered load and $f_h = 2$.

5 Discussion and Conclusion

In this paper we compared the capacity requirements for networks with admission control (AC) and capacity overprovisioning (CO) under various conditions. First, we illustrated the resource demands for a single link. As long as the offered load is constant for the link, the additionally required capacity for CO compared to AC is lower than 15% for an offered load larger than 10^3 Erl although we have considered highly variable flow request sizes. We enhanced our Poisson traffic model by rare overload scenarios. The overload factor f_l governs exactly the capacity requirements for CO while they remain constant for AC as traffic can be blocked to avoid congestion.

More interesting is the corresponding comparison for entire networks. We use the link-by-link AC for our experiments, which is the most popular approach for AC in networks [2]. As long as the offered load is constant for the network, the additionally required capacity for CO compared to AC is even lower than 6% for an average offered load larger than 10^3 Erl per border-to-border relationship. The relative savings of additional capacity (in contrast to 15% on a single link) are due to an increased economy of scale on network links compared to the single link experiment due to transit traffic.

A sudden load increase is quite unlikely in highly aggregated networks and overload on some links may be caused by link failures or local hot spots that show temporarily increased activity. In this paper we neglected the first issue and concentrated on the second one. Our traffic model was based on constant total load which is spread over the network according to the population in the catchment area of the routers. We modelled hot spots by increasing their virtual population by a hot spot factor f_h with a probability of $\frac{1}{365}$. Our results showed that the impact of the hot spot factor f_h on the additionally required capacity is clearly less in single hot spot scenarios than the impact of the overload factor f_l in the single link experiment. This is due to transit traffic on network links and the exact impact depends on the network topology, the routing, and the traffic matrix. The experiments extended to double hot spots showed that they require at most 13% more capacity than single hot spots when we keep the overall traffic in the network constant.

Future Internet architectures propose multi-service networks with high and low priority traffic classes together with suitable scheduling mechanisms [15]. If CO is used for high priority traffic, the excess capacity can be used for low priority traffic. In overload situations caused by high priority traffic, less capacity is available for low priority traffic. For a proper dimensioning of such networks, the above analysis should be applied to assess the bandwidth requirements of the links because their demands for extra capacity vary depending on their transit traffic proportion.

We have shown that AC leads to considerable bandwidth savings compared to CO when strong overload scenarios are likely. However, the additional capacity requirements for networks are significantly smaller than expected from single link experiments. If the excess capacity can be used for the transport of low priority traffic, CO is a viable alternative to AC. Our analysis is based on a traffic model with limited assumptions regarding overload; we did not answer the question how much overload can occur at all, which remains a further research issue, and we did not include disaster scenarios. Currently, we are integrating link failures in our analysis and expect a relative bandwidth reduction for CO because the backup capacity for network outages can be shared with excess capacity for overload situations.

Appendix

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 [16] computes the blocking for a given traffic mix and capacity. Our capacity dimensioning algorithm for AC performs an efficient computational inversion of these formulae [2].

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 $p(x) = \prod_{i=0}^{k-1} \frac{\rho_i^{x_i}}{x_i!} e^{-\rho_i}$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 $C = \min_{C'}\{1 - \sum_{c(x) \leq C'} p(x) \leq p_v\}$.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 [17]. Originally derived for the $M/M/\infty$ system, it is insensitive to the holding time and holds for $M/G/\infty$ systems, too.

Extension to Networks

The link capacity algorithms must be extended to entire networks. For that purpose we calculate the required link blocking (QoS violation) probability $p_b(l)$ ($p_v(l)$) for all links based on the desired border-to-border flow blocking (QoS violation) probability p_b (p_v) such that we can use the above algorithms to determine the required link capacities.

We discuss the mapping process for AC in detail and it works analogously for CO. Let $p_b(l)$ be the blocking probability of a specific single link l. If a flow wants to pass the network from node v to w, the link-by-link NAC performs AC for every link in the flow's path. The blocking probability on the individual links is rather positively correlated but it is impossible to assess this correlation. An upper bound for p_b on the path is given by $p_b(path) = 1 - \prod_{l \in path}(1 - p_b(l))$ where $l \in path$ denotes the links on the path. Now we can compute $p_v(l) = 1 - \sqrt[len(path)]{1 - p_b}$ for every aggregate traversing that link; the minimum of these values yields the required blocking probability for that link.

Acknowledgment

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

References

1. Odlyzko, A.: Data Networks are Lightly Utilized, and will Stay that Way. The Review of Network Economics **2** (2003)
2. Menth, M.: Efficient Admission Control and Routing in Resilient Communication Networks. PhD thesis, University of Würzburg, Faculty of Computer Science, Am Hubland (2004)
3. Kelly, F.P.: Notes on Effective Bandwidths. In: Stochastic Networks: Theory and Applications. Volume 4. Oxford University Press (1996) 141 – 168
4. Fraleigh, C., Tobagi, F., Diot, C.: Provisioning IP Backbone Networks to Support Latency Sensitive Traffic. IEEE Infocom 2003, San Francisco, USA (2003)
5. Kilkki, J., Norros, I.: Testing the Gaussian approximation of aggregate traffic. In: Internet Measurement Workshop, Marseille, France (2002)
6. Van Hoey, G., De Vleeschauwer, D.: Benefit of Admission Control in Aggregation Network Dimensioning for Video Services. Networking 2004, Athens, Greece (2004)
7. Van den Berg, H., et al.: QoS-aware bandwidth provisioning for IP network links. Technical Report PNA-E0406, Centrum voor Wiskunde en Informatica, The Netherlands (2004)
8. Papagiannaki, D., Taft, N., Zhang, Z., Diot, C.: Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models. In: IEEE Infocom, San Francisco, USA (2003)
9. Tuan Tran, H., Ziegler, T.: Adaptive Bandwidth Provisioning with Explicit Respect to QoS Requirements. Quality of Future Internet Services (QoFIS), Sweden (2003)
10. Martin, R., Menth, M., Charzinski, J.: Comparison of Border-to-Border Budget Based Network Admission Control and Capacity Overprovisioning. Networking 2005, Waterloo, OT, Canada (2005)
11. Hoogendoorn, C., Schrodi, K., Huber, M., Winkler, C., Charzinski, J.: Towards Carrier-Grade Next Generation Networks. In: ICCT, Beijing, China (2003)
12. Paxson, V., Floyd, S.: Wide-Are Traffic: The Failure of Poisson Modelling. IEEE/ACM Transactions on Networking (1995)
13. Roberts, J.W.: Traffic Theory and the Internet. IEEE Communications Magazine **1** (2001) 94–99
14. Dinh, T., Sonkoly, B., Molnár, S.: Fractal Analysis and Modeling of VoIP Traffic. In: Proceedings of Networks 2004, Vienna, Austria (2004) 123 – 130
15. Xiao, X., Ni, L.M.: Internet QoS: A Big Picture. IEEE Network Magazine **13** (1999) 8–18
16. Roberts, J., Mocci, U., Virtamo, J.: Broadband Network Teletraffic - Final Report of Action COST 242. Springer, Berlin, Heidelberg (1996)
17. Ross, K.W., Tsang, D.H.K.: The Stochastik Knapsack Problem. IEEE Transactions on Communications **37** (1989) 740–747