On Resource Sharing and Careful Overbooking for Network Virtualization

Markus Fiedler, *Member, IEEE*

***Abstract*—Network Virtualization implies the need for sharing network resources. However, exclusive resource allocation to users drives blocking probabilities and cost while potentially leaving precious resources under-utilized in constrained environments. Motivated by these observations, this paper analyzes careful overbooking according to Service Level Agreements that specify desired degrees of availability. Besides of full availability of the requested resources, a second level of limited availability, implying a well-defined reduction of the allocated resource, is taken into account. Particular attention is paid to the gain borderline, representing the possibility of accommodating one extra user beyond exclusive allocation without violating the SLAs. Simple but telling formulae provide insights into requirements for careful overbooking, worst-case capacity reduction factors from full to limited availability, and the conditions under which it is sensible to integrate users with different activity levels.**

***Index Terms*—Network virtualization, resource virtualization, resource sharing, overbooking, gain, availability, Service Level Agreements**

I. Introduction

Network Virtualization (NV), addressing “the simultaneous operation of multiple logical networks or overlays on a single physical platform” [1], has become a hot topic during the recent years. Some providers are not necessarily able or willing to invest in own hardware. Instead, they lease shares of existing resources in order to offer services. Here, the notion *resource* refers to some processing or transportation capacity. Resources are utilized by *users*, which can be providers, operators, customers or end users. We assume that each user (process or transmission) needs a well-known amount of a resource (e.g. 10 Mbps on a link).

From the user’s point of view, virtual resources should appear real in the sense that they should be usable as if they were *exclusively allocated* to that particular user. Using a resource, there should be no side-effect on performance from other concurrent users on the same system, i.e. the *full* amount of subscribed capacity should be *available* upon request as if the user had it at its full disposal.

Virtual resources, however, imply the need for resource *sharing*, as several users use the same resource. Typically, such resource sharing happens in a time-division manner, e.g. on a first-come-first-serve basis or through allocated and scheduled *time slots*. During a dedicated time slot, a user can use the allocated resource in an exclusive manner. Then, it has to wait until the next time slot is scheduled. For example, consider ten users using a 100 Mbps-link in a round-robin fashion with time slot duration of 10 ms. On time scales of multiples of the *cycle time* of 100 ms, the user perceives 10 Mbps as if they were allocated in an exclusive manner, while on the 10 ms-scale, the allocation process introduces jitter. However, this paper confines itself to principles of resource sharing and leaves potential jitter issues for further study.

In case the population of potential users is larger than the number of allocable time slots, it is important to keep the risk of resource *blocking* as low as possible. As pointed out before, a user that has subscribed a virtual resource hardly accepts a denial of access. On the other hand, that user might not use exclusively allocated resources permanently and not either be ready to pay for unused time. However, providers of virtual environments need to maximize the use of a system in order to optimize their gain-to-investment ratio. Thus, the temptation to *overbook* a resource, i.e. to admit more users than the resource could serve in an exclusive manner, is real.

Overbooking *may* lead to a temporary decrease of the amount of resources allocated to affected users, a situation which we denote as *limited availability* or *non-availability* depending on whether or not such a decrease is acceptable to the user. In the above example, admitting an eleventh user would imply a decrease of the time slot to 9.1 ms or a growth of the cycle time to 110 ms. Both yield a throughput of 9.1 Mbps per cycle time instead of 10 Mbps in case of full availability. But having 20 users in the system would decrease the throughput to 5 Mbps, which might impact service performance beyond what is acceptable.

A key parameter that is limiting the degree of overbooking is the risk to affect the user. Users might be willing to accept a certain percentage of limited availability in exchange for other benefits in form of reduced blocking and/or reduced cost, as exclusive resource allocation may entail unnecessarily high blocking ratios, underutilized resources and overpriced service. Therefore, *careful overbooking* that keeps the probabilities for full respective limited availability above predefined levels and at the same time the consequences of overbooking – once it strikes – reasonably small, is a worthwhile alternative to exclusive resource allocation. For instance, the possibility to accommodate just one extra user under controlled conditions might reduce the blocking probability significantly. Or it might relieve an infrastructure provider from the need to immediately upgrade the offered resources once they get exhausted.

In the literature, overbooking has so far not been discussed in the context of NV, as most processing and networking resources are believed to be abundant. However, there are scenarios in which capacity sharing is a must, e.g. on non-optical access links or on mobile equipment. In contrary to architecture [2], control [3, 4] and roadmap [1] issues, performance issues are hardly addressed in NV literature, although recognized [5]. Some few examples are found in [6, 7] investigating the performance impact of overlay networks used for implementing NV and mapping resources; in [8] dealing with transport system virtualization employing concurrent multipath transmission, and in [9] addressing performance optimization issues.

On this background, we investigate the regime of careful overbooking for NV, obeying *Service Level Agreements* (SLA) for full and limited availability. To our knowledge, the explicit use of two service levels (full and limited) is new in the context of overbooking. Thus, we will pay special attention to the amount of reduction of the resource allocation in case of limited availability. Furthermore, we particularly focus on systems that allow for one additional user as compared to exclusive allocation without violating the corresponding SLA(s).

The remainder of the paper is structured as follows. Section II introduces the resource allocation models, with definitions and notions to be used in the sequel. For – in terms of their activity levels – homogeneous users, Section III considers conditions for full availability, while Section IV addresses relationships between full and limited availability. Section V addresses the case of users with different activity levels and investigates whether and when to logically integrate or to separate these groups. Finally, Section VI wraps up the findings and provides an outlook to future work.

II. Resource Allocation Modeling

A. Definitions

We measure time in multiples of *allocation intervals* of duration ΔT , where interval i covers $t \in](i-1)\Delta T, i\Delta T]$. An *observation interval* W comprises of (the indices of) w allocation intervals. We denote the *resource request* by customer k during interval i by $Q_i^{(k)}$ and the *resource allocation* by $A_i^{(k)}$, respectively. Finally, we define an indicator function as follows:

$$I(x) = \begin{cases} 1 & \text{if expression } (x) \text{ holds} \\ 0 & \text{else} \end{cases}$$

B. Levels of Resource Allocation

From the viewpoint of the end-user, we distinguish different levels of resource allocation:

1. *Full availability*: The resource is fully or even over-allocated, i.e.
$$Q_i^{(k)} \le A_i^{(k)} .$$
The service level is optimal.
2. *Limited availability*: The resource is partly allocated to an extent described by the *reduction factor* $\gamma^{(k)}$, i.e.
$$\gamma^{(k)} Q_i^{(k)} \le A_i^{(k)} < Q_i^{(k)} .$$
The customer can observe a degradation of performance (e.g. longer response times). The service level is not optimal, but still acceptable.
3. *Non-availability*: The resource allocation is insufficient to grant an acceptable service level, i.e.
$$A_i^{(k)} < \gamma^{(k)} Q_i^{(k)} .$$

In case of *exclusive availability*, all accepted resource requests are granted full availability. Otherwise, dependent on the sharing situation, the user may perceive full, limited or non-availability.

C. Availability Degrees

During an observation interval, we obtain the following availability measures:

1. *Degree of full availability*:
$$h_{\text{full}}^{(k)} = \frac{1}{w} \sum_{i \in W} I(Q_i^{(k)} \le A_i^{(k)}) .$$
2. *Degree of limited availability*:
$$h_{\text{lim}}^{(k)} = \frac{1}{w} \sum_{i \in W} I(\gamma^{(k)} Q_i^{(k)} \le A_i^{(k)} < Q_i^{(k)}) .$$
3. *Degree of non-availability*:
$$h_{\text{non}}^{(k)} = \frac{1}{w} \sum_{i \in W} I(A_i^{(k)} < \gamma^{(k)} Q_i^{(k)}) = 1 - h_{\text{full}}^{(k)} - h_{\text{lim}}^{(k)}$$

In case of exclusive availability, $h_{\text{full}}^{(k)} = 1, h_{\text{lim}}^{(k)} = h_{\text{non}}^{(k)} = 0$.

In steady state ($w \to \infty$), we refer to *stochastic degrees,* i.e.

$$h_{\text{full}}^{(k)} \xrightarrow{w \to \infty} \Pr\left\{ Q^{(k)} \le A^{(k)} \right\}$$

$$h_{\text{lim}}^{(k)} \xrightarrow{w \to \infty} \Pr\left\{ \gamma^{(k)} Q^{(k)} \le A^{(k)} < Q^{(k)} \right\}$$

Such probabilities are typically used for dimensioning issues and studies of mechanisms. Measured degrees $\hat{h}^{(k)}$ may deviate from the stochastic degrees, but should converge under steady-state conditions.

D. Availability Thresholds and Service Level Agreements

The desired degrees of availability and thus the SLA are reflected by the following parameters:

1. Parameter $\delta^{(k)}$ reflects the *desired degree of full availability*:
$$h_{\text{full}}^{(k)} \ge 1 - \delta^{(k)} \text{ alt. } \Pr\left\{ Q^{(k)} \le A^{(k)} \right\} \ge 1 - \delta^{(k)} .$$
2. Parameter $\varepsilon^{(k)}$ captures the *desired degree of limited availability*, which in the sequel also includes full availability:
$$h_{\text{full}}^{(k)} + h_{\text{lim}}^{(k)} \ge 1 - \varepsilon^{(k)} \text{ alt. } \Pr\{ \gamma^{(k)} Q^{(k)} \le A^{(k)} \} \ge 1 - \varepsilon^{(k)} .$$

It is common to specify $1-\delta^{(k)}$ (or $1-\varepsilon^{(k)}$) instead of $\delta^{(k)}$ (or $\varepsilon^{(k)}$) in the context of SLAs. They typically assume values such as 99 %, 99.9 %, etc. A sloppy but well-known notion is to refer to the "number of nines" (two nines, three nines, etc.). In the sequel, we will use any of these notations depending on the message to be conveyed.

E. User Model

Users are modeled by on-off processes with geometrically distributed on/off times in order to model a memory-less behaviour, i.e.

$$\Pr\left\{Q^{(k)}=r^{(k)}\right\}=1-\Pr\left\{Q^{(k)}=0\right\}=\alpha^{(k)}.$$

Here, we denote $\alpha^{(k)}$ as *activity level*, while $r^{(k)}$ characterises the *peak resource request* when the user is active. Consequently, the *average resource request* becomes $m^{(k)}=\alpha^{(k)}r^{(k)}$. If homogeneous users with similar parameters are considered, the index $^{(k)}$ is omitted.

III. Full Availability for Homogeneous Users

A. Maximal Numbers of Users and Gain

Exclusive use implies the need for peak allocation. Given a capacity of *C*, we can accommodate for a maximal number

$$N_{\max}^{\text{excl}}(C)=\left\lfloor\frac{C}{r}\right\rfloor$$

of users. However, we might ask ourselves under which circumstance(s) we can admit extra customers into the system without violating the corresponding SLA(s).

In case of *shared allocation* and full availability, we obtain the maximal number of users from

$$N_{\max}^{\text{full}}(C,\delta)=\max\left\{N:\Pr\left\{\sum\nolimits_k Q^{(k)}\le C\right\}\ge 1-\delta\right\}$$

with

$$\Pr\left\{\sum\nolimits_k Q^{(k)}\le C\right\}=\sum_{n=0}^{N_{\max}^{\text{excl}}}\frac{N!}{n!\left(N-n\right)!}\alpha^n(1-\alpha)^{N-n}.$$

The minimal allocation, on the other hand, is obtained from the average resource request, which leads to the maximal number

$$N_{\max}^{\text{avg}}(C)=\left\lfloor\frac{C}{\alpha r}\right\rfloor$$

of users.

As a first example, we consider the maximal number of users with $\alpha=50\,\%$ activity level as a function of the normalised capacity *C*/*r*, expressed in multiples of the resource requirement by one user. In addition to the shared allocations targeting full availability of two nines ($1-\delta=99$ %) and four nines ($1-\delta=99.99$ %), exclusive allocation of the peak resource requirement and allocation of the average resource requirement are taken into account.

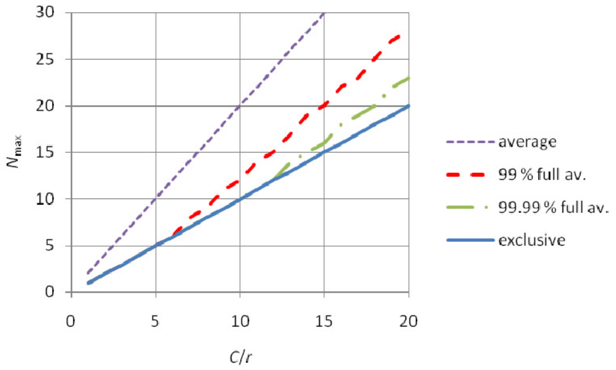

Fig. 1. Maximal number of users as a function of normalized capacity for different allocation strategies, activity level of 50 %.

Fig. 1 reveals that for quite small systems (with capacity for six exclusive users in our example), sharing cannot be exploited; there is practically no difference to exclusive allocation. The larger the desired degree of full availability becomes, the larger the system has to become (*C*/*r* = 13 for four nines as opposed to *C*/*r* = 7 for two nines) to allow for additional users. The special case of one additional admissible user will be investigated in detail in Section III.B.

As the capacity grows, the maximal admissible number of users grows significantly beyond the number of exclusive users. The ratio between those numbers determines the *gain*

$$G_{\max}^{\text{full}}(C,\delta)=\frac{N_{\max}^{\text{full}}(C,\delta)}{N_{\max}^{\text{excl}}(C)}-1.$$

Obviously, the gain grows with rising capacity *C*, while a rising desired degree of full availability $1-\delta$ (more nines) decreases the gain.

In the second example, we reduce the activity level to $\alpha=10\,\%$. Fig. 2 shows the corresponding maximal numbers of users.

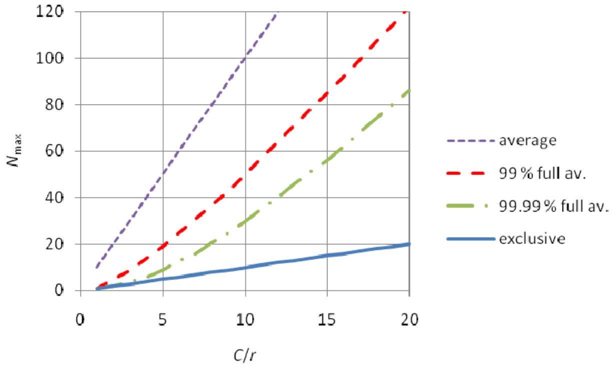

Fig. 2. Maximal number of users as a function of normalized capacity for different allocation strategies, activity level of 10 %.

Comparing Fig. 2 to Fig. 1, we note much higher gains that are exploitable for quite small systems. The smaller the activity level, the larger becomes the gain, which is also communicated by the formula for maximally achievable gain

$$G_{\max}\le\frac{1}{\alpha}-1.$$

B. *Gain Borderline: One Extra User*

We now take a closer look at cases where one extra user can be accommodated while keeping the desired degree of full availability. For 50 % activity (Fig. 1), this was the case from a capacity for seven (for two nines) respective 13 (for four nines) users. For 10 % activity (Fig. 2), those numbers sink to one (for two nines) respective three (for four nines) users. The number of users is thus given by $N_{\max}^{\text{excl}}+1$, and the corresponding gain amounts to

$$G_{\max}^{\text{full}}(C)=\left(N_{\max}^{\text{excl}}(C)\right)^{-1}.$$

At the *gain borderline*, we observe that the probability for resource shortage has only one component, namely

$$\Pr\{\text{all } N_{\max}^{\text{excl}}+1 \text{ users active}\}=\alpha^{N_{\max}^{\text{excl}}+1}\le\delta\,.$$

Denoting N^{+} as the real value close to $N_{\max}^{\text{excl}}$ such that $\alpha^{N^{+}}=\delta$, we obtain

$$N_{\max}^{\text{excl}}(\alpha,\delta)\le N^{+}(\alpha,\delta)=\log_{\alpha}(\delta)\,. \qquad (1)$$

The value N^{+} will be used to quantify the gain borderline, i.e. the boundary for the system size to integrate one more customer with an activity level of α and a degree of full availability of $1-\delta$, cf. Fig. 3.

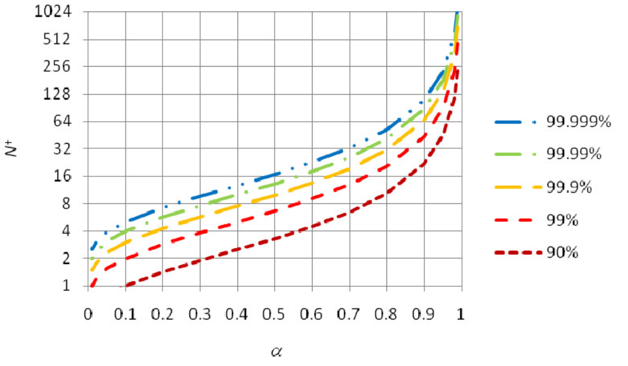

Fig. 3. Gain borderline as a function of the activity level for different desired degrees of full availability.

Fig. 3 illustrates how N^{+} is affected by the activity level and the desired degree of full availability, respectively. We observe:

- For large activity levels (above 70 %), systems need to be very large in size to enable inclusion of any extra user beyond the number given by exclusive allocation.
- For activity levels between 30 % and 70 %, the curves are almost linear in logarithmic N^{+}-coordinates. A growth of the activity by 20 % yields roughly a doubling of N^{+}.
- The – in terms of gain – most interesting area is that of small activity levels. Less than 30 % activity yields gain for as little as ten users even at the quite high demand for five nines.

From these observations, we can conclude that it might be quite difficult to accommodate yet another user when activity levels and demands for full availability are rather high. In such cases, exclusive allocation might be the method of choice. However, rather low demands for full availability (few nines) relax this situation significantly. Under such circumstances, it is worthwhile to consider a second criterion on limited availability, which will be investigated in detail in the next section.

IV. Full vs. Limited Availability for Homogeneous Users

A. *Maximal Numbers of Users and Gain*

We will now investigate the relationship between full and limited availability. In particular, we will look for quasi-optimal relationships between the parameters δ,ε and γ. The latter parameter scales the resource requests so that, in case of *limited availability*, we obtain the maximal number of users

$$N_{\max}^{\text{lim}}(C,\gamma,\varepsilon)=\max\left\{N:\Pr\left\{\gamma\sum\nolimits_{k}Q^{(k)}\le C\right\}\ge 1-\varepsilon\right\}$$

and the corresponding gain as

$$G_{\max}^{\text{lim}}(C,\gamma,\varepsilon)=\frac{N_{\max}^{\text{lim}}(C,\gamma,\varepsilon)}{N_{\max}^{\text{excl}}(C)}-1\,.$$

When both conditions for full and limited availability need to be fulfilled simultaneously, then the minimum of the number of users can be admitted without violating any of the degrees:

$$N_{\max}^{\text{both}}(C,\delta,\gamma,\varepsilon)=\min\left\{N_{\max}^{\text{full}}(C,\delta),N_{\max}^{\text{lim}}(C,\gamma,\varepsilon)\right\}.$$

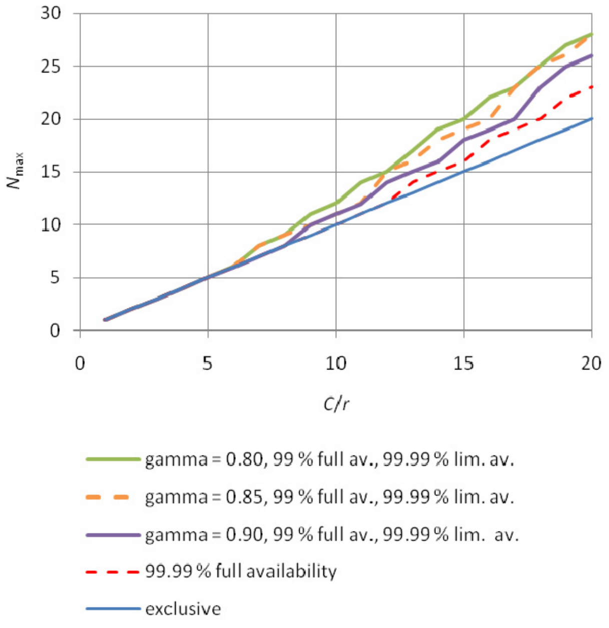

Fig. 4. Maximal number of users as a function of normalized capacity for different allocation strategies involving criteria for both full and limited availability and different reduction factors γ, activity level of 50 %.

Fig. 4 presents the results of different allocation strategies, focusing on the reduction factor γ in order to illustrate its impact on the outcome. To facilitate orientation and comparison with Fig. 1, the cases of two nines full availability and exclusive allocation have been included.

In the case of a reduction factor $\gamma=0.8$, the comparison with Fig. 1 shows that the allocation process is dominated by the two nines requirement for full availability. This means that for such a strong capacity reduction, the degree of limited availability is considerably higher than the required four nines.

The reverse holds for the reduction factor $\gamma = 0.9$. Now, the four nines for the limited availability determine the maximal numbers. The reason for this change is to be found in the significantly larger capacity share that is needed during phases of limited availability.

Thus, we might expect some in-between behavior for a reduction factor γ between 0.8 and 0.9, which is actually confirmed by the curve for $\gamma = 0.85$. While for small systems, the allocation is governed by the two nines full availability requirement, larger systems have to consider the four nines-limited availability requirement, providing them with at least 85 % of their resource needs.

B. Gain Borderline: One Extra User

We again turn our interest to the point when exactly one more user can be admitted to the system, the so-called gain borderline. Recalling (1), $N^{+}_{\text{full}}(\alpha,\delta)$ for full availability, we arrive at a similar equation for $N^{+}_{\text{lim}}(\alpha,\varepsilon)$ for limited availability. We turn our interest in the specific reduction factor γ^{*} that yields the gain *simultaneously* for *both* criteria.

At the gain borderline, the capacity is fully exploited in case all users are active at the same time. Compared to full availability, limited availability means γ^{*} the capacity per user and thus $1/\gamma^{*}$ times the number of users. We obtain

$$\gamma^{*} = \frac{N^{+}_{\text{full}}(\alpha,\delta)}{N^{+}_{\text{lim}}(\alpha,\varepsilon)} = \frac{\log_{\alpha}(\delta)}{\log_{\alpha}(\varepsilon)} = \frac{\lg\delta}{\lg\varepsilon}\,. \qquad (2)$$

The numbers of nines determine the capacity ratio between limited and full availability when just one extra customer is to be accepted beyond exclusive allocation. This relationship, which is actually independent of the activity level, is illustrated in Table I.

TABLE I
EXAMPLES OF RELATIONSHIPS BETWEEN DESIRED DEGREES FOR FULL AND LIMITED AVAILABILITY AND SPECIFIC REDUCTION FACTOR AT THE GAIN BORDERLINE

$1-\delta$	# of nines	$1-\varepsilon$	# of nines	γ^{*}
0.9	1	0.99	2	1/2 = 0.5
		0.999	3	1/3 ≈ 0.33
0.99	**2**	**0.999**	**3**	**2/3 ≈ 0.67**
		0.9999	4	2/4 = 0.5
		0.99999	5	2/5 = 0.4
0.999	**3**	**0.9999**	**4**	**3/4 = 0.75**
		0.99999	5	3/5 = 0.6

The values in Table I indicate that small numbers of nines and large differences in the number of nines between the desired degrees of full and limited availability entail a distinct change in service quality when moving from full to limited availability. Cases in which limited availability provides more than half of the originally desired resource to the user have been highlighted. Obviously, a rule-of-thumb for reasonable reduction factors is as follows: (a) at least two nines for full availability, and (b) one more nine for limited availability.

It has to be pointed out that the values shown in Table I are not necessarily significant for larger gains. Looking at the example in the subsection above, we observed roughly the same admissible numbers of users for two nines full availability and four nines limited availability at about double the gain borderline ($C/r \simeq 12$) for a reduction factor $\gamma \simeq 0.85$, which is significantly higher than the value $\gamma^{*} = 0.5$ shown in Table I. We may conclude that (2) represents a lower bound for the individual performance to be observed when actual resource sharing leads to limited availability.

V. FULL AVAILABILITY FOR HETEROGENEOUS USERS

In this section, we consider heterogeneous users in the sense that there are two groups of N_i users with different activity levels α_i. In particular, we investigate the conditions under which the gain borderline is reached (i.e. one extra user can be accommodated) and aim at insights to which extent resource sharing between different groups makes sense. The alternative is to treat and exploit gain within each group individually. We confine ourselves to full availability, whose desired degree is assumed to be δ, i.e. the same value for both types of users.

Following the reasoning in Section III.B, we define the gain borderline through the joint probability that all users of both groups are active, i.e.

$$\Pr\{N_1 \text{ and } N_2 \text{ users active}\} = \alpha_1^{N_1}\alpha_2^{N_2} \le \delta\,.$$

Without loss of generality, the more active users are placed in group 1. Freezing the number N_1, we can express the boundary value N_2^{+} through a rather simple linear relationship:

$$N_2^{+}(N_1,\alpha_1,\alpha_2,\delta) = \underbrace{\frac{\log\delta}{\log\alpha_2}}_{a} - N_1\underbrace{\frac{\log\alpha_1}{\log\alpha_2}}_{b} \quad (\alpha_2 < 1)\,. \qquad (3)$$

For the logarithms, any convenient base can be applied.

While the decrease of N_2^{+} as function of growing N_1 (term *b*) is solely governed by the ratio between the logarithms of the activity levels, the vertical position of the curve (term *a*) is determined by the ratio of the logarithms of the desired availability threshold and the activity level of the users in question.

Permanently active users of type 1 ($\alpha_1 = 1$) imply *b* = 0, i.e. the impact of N_1 on the gain borderline for the group 2-users disappears. This means that we do not gain anything from integrating those two groups.

Table II shows Eqn. (3) in cases of non-permanently active users in both groups. We observe the linear impact of the number of nines in the desired degree of full availability on term *a* and a clear decrease of the gain borderline as the activity levels of users of type 2 decrease. In case of large activity levels of these users, the gain boundary decreases quite slowly (shown by a large ratio *a*/*b* > 10), which indicates small gains when integrating both groups. The situation looks different for small activity levels. Here, integration of both groups helps to significantly decrease the gain boundary.

TABLE II
GAIN BORDERLINES IN CASE OF NON-PERMANENTLY ACTIVE USERS

α_1	α_2	N_2^+ $(\delta = 0.01)$	N_2^+ $(\delta = 0.0001)$
0.75	0.5	$N_2^+ \simeq 6.6 - 0.42N_1$	$N_2^+ \simeq 13.3 - 0.42N_1$
0.75	0.25	$N_2^+ \simeq 3.3 - 0.21N_1$	$N_2^+ \simeq 6.6 - 0.21N_1$
0.75	0.1	$N_2^+ \simeq 2 - 0.12N_1$	$N_2^+ \simeq 4 - 0.12N_1$
0.5	0.25	$N_2^+ \simeq 3.3 - 0.5N_1$	$N_2^+ \simeq 6.6 - 0.5N_1$
0.5	0.1	$N_2^+ \simeq 2 - 0.3N_1$	$N_2^+ \simeq 4 - 0.3N_1$
0.25	0.1	$N_2^+ \simeq 2 - 0.6N_1$	$N_2^+ \simeq 4 - 0.6N_1$

For instance, consider $\alpha_1 = 0.25$ and $\alpha_2 = 0.1$ for $\delta = 0.0001$. Separation of both groups means $N_1^+ \simeq 6.6$ and $N_2^+ = 4$. If both groups were integrated and $N_1 = 6$ users existed, the gain borderline for group 2 was decreased to $N_2^+ \simeq 0.4$, which clearly denotes the extra gain obtained from the sharing of the resources when activity factors are low.

VI. CONCLUSIONS

We considered resource sharing issues in virtualized network environments. Going beyond exclusive assignment of resources to users, we investigated careful overbooking in the sense that (a) full availability, i.e. the full amount of required resources, and (b) limited availability of at least a certain share of required resources is (statistically) guaranteed at given degrees. In particular, we paid attention to cases in which one extra user can be accommodated without sacrificing the desired degrees of full and/or limited availability, typically specified through the number of nines in the decimals.

Assuming an on-off user model with geometrically distributed phases of activity and inactivity, we first considered – in terms of parameters – homogeneous users. We calculated quantitative results for the maximal number of users in order to keep the desired degree of full availability. We furthermore investigated the minimal number of users for which a gain can be realized, called the gain borderline. We derived a closed formula for that gain borderline, revealing that large activity levels and many nines provide hinders for gain. No advantage over exclusive allocation is obtained unless the system becomes very large.

We also investigated the relationship between the degrees of full and limited availability, and the reduction of resource allocation related to limited availability. We found that at the gain borderline, the numbers of nines provide a hint on the capacity reduction factor related to limited availability, which also serves as worst-case estimation. In particular, the numbers of nines should neither be too small nor differ too much between full and limited availability.

Considering two groups of heterogeneous users with different activity levels, we investigated the conditions under which it makes sense to integrate both groups in order to yield gain. As in the homogeneous case, this makes sense for comparably small activity levels. As soon as at least one group is very active, it is advisable to keep the groups apart capacity-wise, as hardly any additional gain can be obtained.

Although the topic of overbooking has been well-researched in other areas such as the dimensioning and control of broadband networks, is has hardly been addressed in the NV context so far. Overbooking might be a problem in many kinds of systems, and this study could have been presented in a much more generic setting. Still, we perceive some value in tailoring known approaches to NV in order to highlight specific chances and risks when being tempted to applying overbooking, as it has implies the chance to reduce resource blocking at the expense of limited availability. In many cases – small systems, high activity factors, and high desired availability – overbooking has shown to be non-efficient. Still, with this paper, we have some simple formulae at hand to assess the circumstances under which overbooking is an option.

Thus, there are a couple of options for future work. Amongst others, we might consider refined user models matched to traces; validation of the findings of this paper through measurements in real NV systems, or through simulations; and some detailed investigation of potential gains beyond the gain borderline, or the dependency of gains from eventual correlations of the resource requests. As a generic topic even beyond the area of NV, a quantitative study of the coexistence of performance criteria for full and limited availability, in particular related to user-expressed Quality of Experience, is of interest.

ACKNOWLEDGEMENT

The author would like to thank Kurt Tutschku for interesting and fruitful discussions on the principles of NV.

REFERENCES

[1] K. Tutschku, T. Zinner, A. Nakao, and R. Tran-Gia, “Network virtualization: Implementation steps towards the Future Internet,” in *Proc. Workshop on Overlay and Network Virtualization at KIVS 2009*, 14 pp., Electronic Communications of the EASST, Volume X (2008).

[2] J.M.P. Cardoso, “On combining temporal partitioning and sharing of functional units in compilation for reconfigurable architectures,” *IEEE Trans. Computers*, vol. 52, no. 10, pp. 1362–1375, Oct. 2003.

[3] S. Graupner, V. Kotov, and H. Trinks, “Resource-sharing and service development in virtual data centers,” in *Proc. 22nd Int. Conf. on Distributed Computing Systems Workshops*, July 2002, pp. 666–671.

[4] P. Garbacki and V.K. Naik, “Efficient resource virtualization and sharing strategies for heterogeneous Grid environments,” in *Proc. 10th IFIP/IEEE Int. Symp. on Integrated Network Management (IM’07),* May 2007, pp. 40–49.

[5] S. Rixner, “Network virtualization – breaking the performance barrier,” *ACM QUEUE,* Jan./Feb 2008, pp. 36–43&52.

[6] M. Tsugawa and J.A.B. Fortes, “Characterizing user-level network virtualization: Performance, overheads and limits,” in *Proc. IEEE 4th Intl. Conf. on eScience (eScience’08),* Dec. 2008, pp. 206–213.

[7] I. Houidi, W. Louati, and D. Zeghlache, “A distributed and autonomic virtyal network mapping framework,” In *Proc. 4th Int. Conf. on Autonomic and Autonomous Systems (ICAS 2008),* March 2008, pp. 241–247.

[8] K. Tutschku, T. Zinner, A. Nakao, and P. Tran-Gia, “Re-sequencing buffer occupancy of a concurrent multipath transmission mechanism for transport system virtualization,” in *Proc. of Kommunikation in verteilten Systemen 2009 (KIVS 2009)*, Kassel, Germany, Mar. 2009.

[9] J. Zhang, X. Li, and H. Guan, “The optimization of Xen network virtualization,” in *Proc. 2008 Int. conf. on computer Science and Software Engineering*, vol. 3, Dec. 2008, pp. 431–436.