Protected Connection Provisioning with Low Availability Overfulfillment in Meshed Core Networks

Tobias Enderle
Institute of Communication Networks and Computer Engineering (IKR)
University of Stuttgart
Stuttgart, Germany
tobias.enderle@ikr.uni-stuttgart.de

***Abstract*—Network operators sell connection services to their customers. The corresponding service contracts typically include a service level agreement (SLA) which defines a guaranteed availability level per month. An availability exceeding the SLA level means that the network operator complies with the SLA. A lower availability results in contractual penalties. Therefore, the probability of SLA compliance is a key figure for the operator. To adjust the compliance probability, many operators apply protection mechanisms, which tie up precious network resources to provide backup capacity. Therefore, it is desirable to dedicate only as many network resources to protection as necessary to reach a sufficiently high probability of SLA compliance—the operator's self-defined compliance target. However, in practice, this is difficult to realize because using protection, the amount of resources and, consequently, the compliance probability cannot be selected continuously. Adding protection to a connection service makes its compliance probability jump up, possibly to a level far above the operator's compliance target. The result is overfulfillment at the cost of precious network resources. In this paper, we propose an admission and routing approach that reduces said overfulfillment, frees network resources and by that allows more services to be accommodated in the network. We use a stochastic approach to estimate a service's probability of SLA compliance. Probability that exceeds the operator's compliance target is accumulated as surplus and allows other services to be accepted with a compliance probability below the operator's original compliance target. With this approach, the resulting SLA compliance ratio over all services matches the compliance target closely, i.e., the overfulfillment is reduced. We evaluate our mechanism in a simulation study covering several core network topologies. It is shown that the availability overfulfillment can be reduced or even eliminated and that the service blocking ratio can be decreased significantly.**

***Index Terms*—Availability management, Compliance probability, Interval availability, Overfulfillment, Routing, Service level agreement**

I. Introduction

The goal of every network provider is to achieve sustainable profit. In the past, the sale of connection services generated enough revenue. However, traffic demand is ever increasing and network operators have to expand their networks constantly to provide enough capacity. In recent years, though, network expansion has become increasingly costly, due to disparities between the growth rate of traffic demand and that of transmission technologies [1]. Additionally, network services are more and more considered a commodity and prices are falling. Therefore, instead of maintaining their course of generously overdimensioning their networks, operators have to drive their networks closer to the limits while trying to delay network expansion. This can be achieved, e.g., by replacing fixed-grid wavelength-division multiplexing (WDM) with a flexible-grid configuration or by reducing system margins during the network planning phase [2]. While these measures are purely technology-related, research proposes to take service level agreement (SLA)-related aspects into account as well. For example, the authors of [3] develop a model to estimate the availability of network services more precisely than with standard methods. In that way, the availability "safety margin" a network operator typically includes during provisioning can be selected appropriately. The authors of [4] consider the tradeoff between SLA penalties and protection costs with the goal of postponing costly protection investments.

In this work, we focus on the accurate fulfillment of a connection service's availability according to its SLA specifications. We propose a service admission and routing approach that allows different network services to share excess availability. As a result, the amount of network resources that has to be reserved for protection can be reduced. In that way, the availability specified in the SLAs is provided with less overfulfillment than with many other routing approaches. Overall, this leads to more efficient network operation and enables the operator to accommodate more services in the network. Our mechanism is applicable to dynamically arriving services in connection-oriented networking technologies, e.g. wavelength services, optical transport network (OTN) paths, or multi-protocol label switching (MPLS) tunnels.

In the following, we will first discuss some important aspects about availability in SLAs. Afterwards, we present related work in Section III. Section IV introduces our admission and routing approach consisting of a mathematical model to estimate the SLA compliance probability and an admission and

This work has been performed in the framework of the CELTIC-NEXT EUREKA project AI-NET-ANTILLAS (Project ID C2019/3-3), and it is partly funded by the German BMBF (Project ID 16KIS1312). The authors alone are responsible for the content of the paper.

routing algorithm. Section V shows some illustrative numerical results of the proposed approach for different meshed core networks. Finally, we conclude the paper in Section VI.

II. Availability in SLAs

In this section we will discuss two important aspects of availability in SLAs. First, the difference between the steady-state availability and the service availability (or interval availability, respectively), the latter being more relevant for SLAs. Second, the overfulfillment of availability due to topological or protection-related constraints.

Contracts for network services between the network operator and its customers typically include an SLA. Among other things, the SLA defines performance targets like latency, availability, and protection for the network service. In case the network operator cannot meet those performance targets the customer can claim penalty payments or refunds. Typically, availability and protection are two very important SLA aspects. While some SLAs guarantee the use of protection explicitly (e.g. [5]), others only stipulate an availability level α_{SLA} (e.g. [6]). In the latter case, the network operator is free to select appropriate protection mechanisms to provide the specified availability. In the following, we will focus on this particular scenario.

A common challenge in network operation is the admission and routing of dynamically arriving connection service requests. A simple but established strategy for network operators to select a suitable route and appropriate protection mechanisms is based on the evaluation of the steady-state availability a of a network service (e.g. [7], [8]). The steady-state availability is the probability to find a system working at an arbitrary point in time. A common formula to compute the steady-state availability is

$$a = \frac{MTTF}{MTTF + MTTR} \tag{1}$$

where *MTTR* is the mean time to repair and *MTTF* is the mean time to failure, i.e., the mean time between repair and next failure. If a combination of route and protection can be found that fulfills the condition $a \geq \alpha_{\mathrm{SLA}}$, the service is accepted and routed. However, it has been shown by several authors, e.g., in [3], [9], [10], that the use of the steady-state availability can lead to unexpectedly high SLA violation ratios.

The reason is that the steady-state availability considers an infinite amount of time, but an SLA considers a *billing cycle* of finite length T, which is usually one month [5], [6]. At the end of each billing cycle, the availability during that cycle is evaluated according to

$$A = \frac{T - X}{T} \tag{2}$$

where X is the accumulated downtime of the service during the billing cycle. A is known as *interval availability* and in the context of this work we refer to it as *service availability*. Due to the different time horizons, a and A are not necessarily equal in value. Furthermore, since network failures occur randomly and also the time it takes to put the failed network segment back into operation is fraught with uncertainty, the accumulated downtime X and, consequently, also the service availability A are random variables (RVs). Therefore, it is practically impossible to provide the stipulated availability to each and every service deterministically ($A \geq \alpha_{\mathrm{SLA}}$ almost surely). Instead, the service availability fulfills the stipulated availability only with a certain probability $P(A \geq \alpha_{\mathrm{SLA}})$ which we will call *compliance probability* or just *compliance* in the following (other works refer to the complementary probability as *SLA violation risk*, e.g. [9], [11]). Costly operator efforts, like protection, high-quality components, or fast repair, can increase this probability. However, even the most costly efforts cannot ensure $P(A \geq \alpha_{\mathrm{SLA}}) = 1$, and hence, there will still be SLA violations. Therefore, from an economical perspective, increasing the compliance probability at any price is not necessarily the optimal choice. Instead, a network operator has to balance costs for the provision of availability against expected SLA penalties and by that select a suitable target level f_{t} for the compliance probability.

Regardless of whether the operator selects a service's route and protection based on the steady-state availability or the compliance probability, a fundamental problem that arises is that a service does usually not meet the SLA level or the operator's compliance target with equality ($a = \alpha_{\mathrm{SLA}}$ or $P(A \geq \alpha_{\mathrm{SLA}}) = f_{\mathrm{t}}$) but only with *overfulfillment* ($a > \alpha_{\mathrm{SLA}}$ or $P(A \geq \alpha_{\mathrm{SLA}}) > f_{\mathrm{t}}$). The reason is that different candidates for route selection do not offer an arbitrarily fine granularity of availability and also protection cannot increase availability in a continuous way but only in discrete steps. As a consequence, the network operator permanently provides more availability than has been specified in the SLAs. The routing methodology we propose reduces this overfulfillment and thus allows a more efficient network operation.

III. Related Work

Our admission and routing approach is not the only one to employ availability sharing or to make use of availability margins.

The authors of [12] propose a protection mechanism in which a broken wavelength service that is close to an SLA violation can preempt a wavelength service that still has downtime budget left. In that way, excess availability of services with little or no downtime is used to reduce SLA violations. A similar approach is proposed in [13] where a network service with a high SLA violation risk can preempt backup resources of services with lower risk. The violation risk is represented by the so-called urgency level (UL), a metric that takes the remaining holding time, the remaining affordable downtime, and the penalty costs of the service into account. The UL is also used in [14] where protection schemes (dedicated link protection, 1+N protection, and shared path protection) are changed dynamically to provide a suitable availability level on the one hand and to delay network upgrades on the other hand. A very similar approach using steady-state availability instead of UL is presented in [15]. In [16], the goal is the maximization of the overall profitability, i.e., the tradeoff between expected SLA penalties and service returns plays a key role during the service admission process. Similar to our approach, the authors employ the SLA violation probability instead of the steady-state

availability to find a suitable route. Also, the selected route is not required to be the one with the lowest SLA violation probability if this results in a higher total profitability. In that way, availability margins are exploited to improve profitability. The authors of [17] present a rather radical algorithm in which the network resources of an existing service are released for new services when the remaining holding time of the existing service is less than its remaining allowed downtime, i.e., when an SLA violation is no longer possible. The approach in [18] implements the idea of sharing excess availability by forming clusters of network services with heterogeneous SLAs. Services inside a cluster that have experienced only little downtime act as protection for services that are close to violating their SLA. In that way, no explicit backup resources are required. Finally, in [19]–[21], the authors propose an admission and routing approach for shared-path protection. Whenever a new service request arrives, the target availability of each existing service is recomputed, taking its remaining holding time and outage history into account. In that way, the target availability of a service with only little or no downtime is relaxed, and as a consequence, its potential sharing degree for shared-path protection is increased. For a service that already experienced much downtime, the target availability is increased and the potential sharing degree is decreased. Overall, it is shown that the adaptation of the target availabilities leads to less availability overfulfillment and a lower blocking ratio. [21] also discusses a possible way to employ the SLA violation risk instead of the steady-state availability but no results are shown.

All discussed publications contain aspects of our work but to the best of our knowledge there is no publication that targets the accurate fulfillment of the service availability according to the SLA specifications and employs the sharing of SLA compliance probability or violation risk to achieve this.

IV. Admission and Routing Approach

A. Overview

Our approach is responsible for the admission and routing or the rejection of randomly arriving connection service requests. As mentioned in the previous section, the goal of the approach is the reduction of availability overfulfillment. For the reasons described above, we employ the compliance probability and not the steady-state availability as a decision criterion in our approach. We assume that the network operator has set an internal compliance target level f_t. Therefore, the probability of SLA compliance must be larger than or equal to f_t, however, equality is intended. More formally, the operator wants to achieve

$$P(A \geq \alpha_\mathrm{SLA}) \geq f_\mathrm{t} \tag{3}$$

where A is the RV defined in (2) describing the monthly service availability of the provisioned services. All routing decisions of the algorithm are, in principle, based on (3).

When a service request arrives, the compliance probability of potential routes can be computed based on link properties (see Section IV-B). As argued above, it is unlikely that a route can be found for a service request i such that its compliance probability f_i equals the operator target precisely. Instead, there will be some overfulfillment Δf_i. In our routing algorithm, we accumulate the overfulfillment of individual services stepwise as *surplus* and share it with other services to reduce the overall overfulfillment. The sharing is realized by dynamically relaxing the compliance target for future service requests based on the stepwise accumulated surplus, i.e., instead of requiring a compliance probability of at least f_t for each and every service request, the i-th service request can be routed with a lower required compliance probability of

$$f_{\mathrm{req},i} = f_\mathrm{t} - \Delta f_{i-1} \tag{4}$$

where

$$\Delta f_{i-1} = f_{i-1} - f_{\mathrm{req},i-1} \tag{5}$$

is the surplus accumulated until the previous service $i-1$, and f_{i-1} is its compliance probability (or more precisely, that of the underlying route).

TABLE I
Exemplary sequence of service requests and surplus sharing. The targeted compliance level is $f_\mathrm{t} = 0.99$ and the initial surplus is $\Delta f_0 = 0$.

Request number i	Required compliance $f_{\mathrm{req},i} = f_\mathrm{t} - \Delta f_{i-1}$	Route compliance f_i	Resulting surplus $\Delta f_i = f_i - f_{\mathrm{req},i}$
1	0.990	0.995	0.005
2	0.985	0.986	0.001
3	0.989	0.989	0.000

Table I shows an example with three consecutive service requests. The operator's target level is set to $f_\mathrm{t} = 0.99$ and the initial surplus Δf_0 is zero. For the first request, potential routes must have a compliance probability of at least $f_{\mathrm{req},1} = 0.99$. Next, our routing algorithm tries to find a suitable route. Assume that the route the algorithm selects has a compliance probability of $f_1 = 0.995$. The resulting difference is the surplus, namely $\Delta f_1 = 0.995 - 0.99 = 0.005$. This surplus is used to relax the compliance requirement of the second request. A route with a compliance probability of $f_{\mathrm{req},2} = 0.985$ is sufficient for the second request. The operator can benefit from this by selecting a less reliable route or by provisioning less protection.

As already mentioned, the algorithm needs to estimate the probability of SLA compliance for routes through the network. The mathematical model for this estimation is introduced in the next section. Afterwards, the details of the algorithm are presented.

B. Estimation of Compliance Probability

In this section, the estimation of the compliance probability for repairable services is introduced. A service can be in a working or a failure state. If a service fails, it will be repaired and is then working again. As in many other works, we assume that the time it takes to repair a service and the time until the next failure occurs both follow exponential distributions with means *MTTR* and *MTTF*, respectively. Based on (2), the availability A

is a function of the cumulated downtime X during a billing cycle T. Therefore, the compliance probability can be expressed as

$$P(A \geq \alpha_{\text{SLA}}) = P\big(X \leq T \cdot (1 - \alpha_{\text{SLA}})\big) \tag{6}$$
$$= P(X \leq x_{\text{SLA}}) \tag{7}$$

where x_{SLA} is the maximum allowed cumulated service downtime per billing cycle T. The cumulative distribution function (CDF) of X has been derived in [22] and is given by

$$F(x) = a \cdot \Omega_{\lambda,\mu}(x) + (1-a)\big(1 - \Omega_{\mu,\lambda}(T-x)\big) \tag{8}$$

where $a = MTTF/(MTTF + MTTR)$ is the steady-state availability of the service, $\lambda = 1/MTTF$, and $\mu = 1/MTTR$. The function $\Omega_{\gamma,\delta}(z)$ is the CDF of the RV Z. Like X, Z describes the cumulated downtime during a time interval. However, in contrast to X, the service must be working at the beginning of the time interval. According to [23], we have

$$\Omega_{\gamma,\delta}(z) = e^{-\gamma(T-z)}\left(1 + \sqrt{\gamma\delta(T-z)} \cdot \int_0^z e^{-\delta y} y^{-\frac{1}{2}} \mathrm{I}_1\left(2\sqrt{\gamma\delta(T-z)y}\right) \mathrm{d}y\right) \tag{9}$$

where $\mathrm{I}_1(x)$ is the modified Bessel function of the first kind of order 1, and γ and δ correspond to the failure and repair rates of the process underlying Z. Finally, we obtain for the compliance probability

$$P(A \geq \alpha_{\text{SLA}}) = F\big(T \cdot (1 - \alpha_{\text{SLA}})\big). \tag{10}$$

In this work, we consider unprotected services and services with dedicated path protection, both of which consist of several network components. Therefore, in order to use (10), we have to aggregate the failure rates λ_j and repair rates μ_j of the underlying components.

An unprotected connection service consists of a series of N network components. The service is working only if all components are working. We assume that component failures and repairs are independent of each other. Therefore, the aggregated failure rate λ_{U} is the sum of the individual failure rates [24, Ch. 6.3]

$$\lambda_{\text{U}} = \sum_{j=1}^{N} \lambda_j. \tag{11}$$

The steady-state availability of a series is given by the product of the individual availabilities, i.e.,

$$a_{\text{U}} = \prod_{j=1}^{N} a_j \tag{12}$$

and with (1) the repair rate of the service is

$$\mu_{\text{U}} = \lambda_{\text{U}} \frac{a_{\text{U}}}{1 - a_{\text{U}}}. \tag{13}$$

Dedicated path protection employs two parallel, disjoint routes: a working route and a backup route. Aggregating the two routes individually using the serial substitution above yields λ_1 and μ_1 for the working route and λ_2 and μ_2 for the backup route. Then, according to [24, Ch. 6.4], the failure rate of such a protected service is

$$\lambda_{\text{P}} = \frac{\lambda_1\lambda_2(\lambda_1 + \lambda_2 + \mu_1 + \mu_2)}{(\lambda_1 + \mu_2)(\lambda_2 + \mu_1) + \lambda_1(\lambda_1 + \mu_2) + \lambda_2(\lambda_2 + \mu_1)}. \tag{14}$$

The steady-state availability of the two routes is given by

$$a_{\text{P}} = 1 - (1 - a_1)(1 - a_2) \tag{15}$$

and similar to (13) the repair rate is

$$\mu_{\text{P}} = \lambda_{\text{P}} \frac{a_{\text{P}}}{1 - a_{\text{P}}}. \tag{16}$$

C. Admission and Routing Algorithm

The routing algorithm handles a service request by accepting and routing it if enough network capacity is available and a route with sufficient compliance can be found. If this is not the case, the service request is blocked. The algorithm has a global view on the network, i.e., it knows all link capacities and the current traffic on the links as well as the link failure and repair rates. The network is represented by a graph G comprising the nodes and links. A service request R consists of a source and a destination node, a data rate, and an SLA availability. We assume that a service cannot be split but must be served by a single route.

Algorithm 1 shows how an arriving service request is handled. First, a subgraph of G is created which only contains links that are currently working and have enough free capacity to carry the new service. We assume that network nodes are failure-free. In line 3, the required compliance probability for the service route is determined. As explained above, our algorithm uses surplus sharing, i.e., the required compliance is the operator's target compliance f_{t} relaxed by the current surplus Δf. The case without surplus sharing serves as a reference in the evaluation below. Next, the algorithm searches for a suitable route. If the search succeeds, the service is routed accordingly and the accumulated surplus is set to the current route's surplus. Otherwise, the service is blocked.

Algorithm 2 shows how a suitable route is found. Essentially, the algorithm first searches for an unprotected primary route with enough compliance probability using the shortest path in the subgraph G_{s}. If the unprotected route is not reliable enough, a link-disjoint backup route is added and the compliance is evaluated again (lines 4–16). The whole procedure is repeated up to $k_{\max}$ times with alternative primary paths (line 3). This is important for two reasons. First, depending on the shortest path metric, the shortest path is not necessarily the path with the highest compliance probability. As a result, a k-th shortest path with $k > 1$ could provide sufficient compliance while the first shortest path does not. Second, if G_{s} is rather sparse, the bad choice of a primary path can lead to the situation that no additional backup path is available. The compliance probability in lines 8 and 14 is calculated using the procedure presented in Section IV-B.

Algorithm 1 Admission and routing.

Global state
- Network graph G
- Operator compliance target $f_\mathrm{t} \in [0,1]$
- Cumulated compliance surplus $\Delta f \geq 0$

Input
Service request $R = (s, d, h, \alpha_\mathrm{SLA})$, with source node s, destination node d, data rate h, and SLA availability α_SLA

procedure HANDLESERVICEREQUEST(R)
  $G_\mathrm{s} \leftarrow$ subgraph of G including only links that are working and have free capacity $\geq h$
  $f_\mathrm{req} \leftarrow f_\mathrm{t} - \Delta f$ **if** surplus sharing enabled **else** f_t
  route, $f \leftarrow$ FINDROUTE(G_s, R, f_req)
  if *route* $\neq$ None **then**
    Route service request on *route*
    $\Delta f \leftarrow f - f_\mathrm{req}$
  else
    Block service request

Algorithm 2 Route selection.

Global state
Maximum primary path trials $k_\mathrm{max} \in \mathbb{N}^+$

Input
- Network subgraph G_s
- Service request $R = (s, d, h, \alpha_\mathrm{SLA})$
- Required compliance $f_\mathrm{req} \in \mathbb{R}$

Output
Feasible route and its compliance probability or None

function FINDROUTE(G_s, R, f_req)
  $k \leftarrow 1$
  while $k \leq k_\mathrm{max}$ **do**
    $p \leftarrow$ KTHSHORTESTPATH(G_s, s, d, k)
    if $p =$ None **then**
      return None
    else
      $f \leftarrow$ COMPLIANCEPROB(p, α_SLA)
      if $f \geq f_\mathrm{req}$ **then**
        return p, f ▷ *Unprotected route*
    $G_\mathrm{b} \leftarrow$ subgraph of G_s excluding links in p
    $b \leftarrow$ SHORTESTPATH(G_b, s, d)
    if $b \neq$ None **then**
      $f \leftarrow$ COMPLIANCEPROB((p, b), α_SLA)
      if $f \geq f_\mathrm{req}$ **then**
        return $(p, b), f$ ▷ *Protected route*
    $k \leftarrow k + 1$
  return None

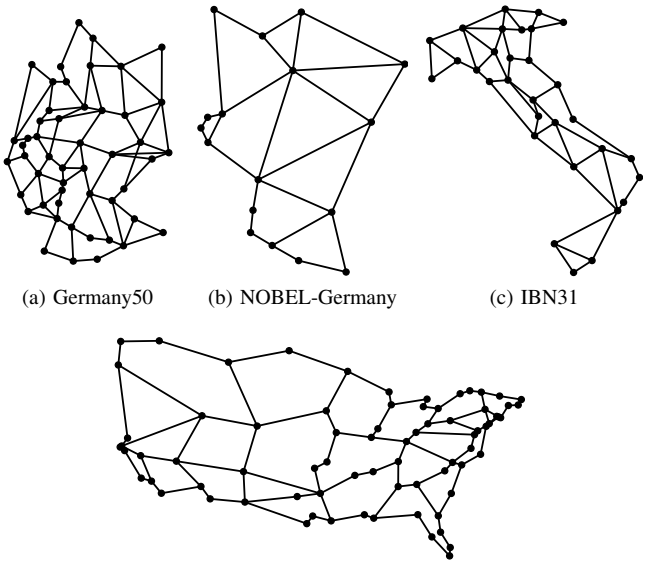
V. ILLUSTRATIVE NUMERICAL EXAMPLE

A. Simulation Setup

We evaluate the performance of our routing approach in four different core networks. The networks are taken from [25]–[27] and are depicted in Figure 1. Table II provides additional information. Based on realistic values provided in [28], we assume an MTTR of 9 hours for all links, and an MTTF of

$$MTTF_e = \frac{628\,\mathrm{km} \cdot 360\,\mathrm{days} \cdot 24\,\mathrm{h/day}}{\ell_e} \tag{17}$$

for link e, where ℓ_e is the link length. We assume that nodes do not fail. For the sake of simplicity, a year in the simulation has 360 days with 30 days per month. Each link provides a capacity of 16 Tbit/s. Service requests arrive randomly with an exponentially distributed inter-arrival time (IAT). The mean IAT is varied over different simulation runs to evaluate different network loads. The source-destination node pair of a service request is chosen uniformly from all node pairs in the network. The requested data rate and holding time (contract period) are selected uniformly from 40 or 100 Gbit/s and 3, 6, 12, or 24 months, respectively. The parameter k_max is set to 5. The weight of a link e, used by the shortest path algorithm, is $-\log(a_e)$, where a_e is the steady-state availability of the link. This link weight allows a standard shortest path algorithm to find the path with the highest availability—the so-called *most reliable path* [29]. In general, the path with the highest availability is not necessarily the same as the path with the highest compliance probability[1]. However,



(a) Germany50 (b) NOBEL-Germany (c) IBN31

(d) CORONET

Fig. 1. Network topologies.

TABLE II
NETWORK PROPERTIES.

Network	Nodes	Links	Link length in km		
			Min.	Avg.	Max.
Germany50	50	88	26	101	252
NOBEL-Germany	17	26	29	143	294
IBN31	31	52	34	128	283
CORONET	75	99	20	330	1017

[1]For example, with $T = 1\,\mathrm{month}$, $\alpha_\mathrm{SLA} = 0.99999$, $MTTR_1 = 7\,\mathrm{h}$, and $a_1 = 0.9999$, the compliance probability is $f_1 = 0.9897$, while with $MTTR_2 = 15\,\mathrm{h}$, and $a_2 = 0.9998 < a_1$ it is $f_2 = 0.9903 > f_1$, due to the change in the MTTR.

in practice, they often coincide. We do not assume a specific networking technology in this evaluation, therefore, typical challenges, especially from the optical domain, e.g. spectrum fragmentation, contiguity, and continuity, are not considered.

We study the behavior of our algorithm for compliance targets f_t of 0.99, 0.995, and 0.999. As an example, for $f_\mathrm{t} = 0.99$, the operator only allows $1\,\%$ of all billing cycles to violate the SLA availability. For the networks Germany50, NOBEL-Germany, and IBN31, the SLA availability is set to $\alpha_\mathrm{SLA} = 0.99999$. For the network CORONET, $\alpha_\mathrm{SLA} = 0.9828$. With those availability levels, the feasibility of a service is guaranteed for all nodes pairs, even with the most strict compliance target of 0.999.

For each set of parameters, 500 consecutive periods (batches) of 5 years each have been simulated in an event-based batch simulation using the IKR SimLib library [30]. The startup phase in each simulation run has been set to 100 years to eliminate transient effects.

B. Results

We first discuss detailed results for the Germany50 network with a compliance target of $f_\mathrm{t} = 0.99$. Figure 2 shows plots for the protection overbuild, the compliance ratio, and the blocking ratio. The protection overbuild is the number of links in the protection path divided by the number of links in the primary path. In the case of an unprotected route, the overbuild is zero. For each plot, the *load factor* is varied to evaluate the algorithm under different network loads. The load factor is varied by changing the mean IAT of service requests. The mean IAT has been calibrated per parameter combination in a preparatory simulation such that the blocking ratio for a load factor of 1 is at around 10^{-2} when surplus sharing is disabled. The obtained mean IATs range from 7 h to 16 h. The simulation with surplus sharing enabled uses the same calibration. The error bars in Figure 2c depict $95\,\%$-confidence intervals for the mean. In Figures 2a and 2b, the confidence intervals are omitted because they are too small to be visible.

Figure 2a shows the average protection overbuild. Using surplus sharing, the protection overbuild is reduced by around $38\,\%$ because some services can be routed without protection due to the accumulated surplus. With increasing network load, the overbuild increases as well because the shortest paths to connect a node pair are more and more occupied and alternate paths have to be taken.

Figure 2b shows the compliance ratio. The compliance ratio in a simulation batch describes the share of all billing cycles in which the service availability fulfilled the SLA. It is calculated based on all provisioned services and their billing cycles as

$$\frac{1}{\sum_{i=1}^{N} M_i} \sum_{i=1}^{N} \sum_{m=1}^{M_i} \mathbf{1}_{a_{i,m} \geq \alpha_\mathrm{SLA}} \tag{18}$$

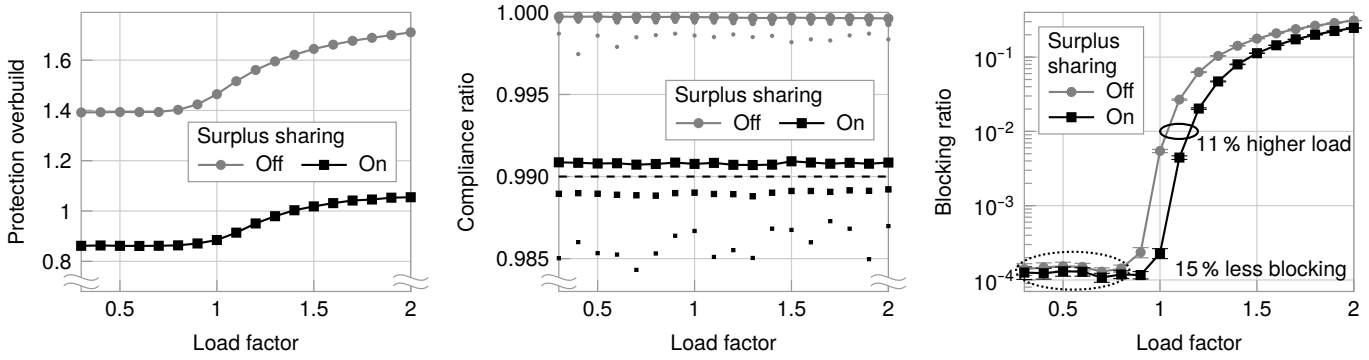
where N is the number of provisioned services, M_i is the number of billed months of the i-th service (3, 6, 12, or 24) and $a_{i,m}$ is its actual service availability in its m-th billed month, i.e., the realization of the RV A in (2). $\mathbf{1}_{a_{i,m} \geq \alpha_\mathrm{SLA}}$ is the indicator function which equals 1 if $a_{i,m} \geq \alpha_\mathrm{SLA}$ (availability high enough) and 0 otherwise. The connected marks in the figure show the average compliance ratio over all 500 batches. The medium-sized marks correspond to the first decile and the small marks represent the minimum 5-year compliance ratio of the 500 batches. Without surplus sharing, the average compliance ratio is close to 1 which means that the operator target of $f_\mathrm{t} = 0.99$ is overfulfilled. With surplus sharing enabled, the average compliance ratio is close to 0.99. Consequently, the overfulfillment has been reduced significantly. The average ratios are independent of the network load. Since the SLA compliance is the result of a random failure and repair process, there are periods in which the compliance ratio is below the targeted level. As can be seen, of the 500 5-year periods, the minimum compliance ratios with surplus sharing are around 0.986 while the first deciles are around 0.989. Of course, these variations pose a risk for the network operator which should be minimized. The amount of variation depends on the failure characteristics of the network components and on the service arrival and holding time behavior. Furthermore, the considered time period plays an important role, i.e., time horizons longer than the currently considered 5 years will result in less variation. However, the detailed relations will be the subject of future work.

Figure 2c shows the average blocking ratio. It can be seen that surplus sharing leads to consistently lower blocking ratios. Considering the network load at a blocking ratio of 10^{-2}, surplus sharing allows a significant increase of about $11\,\%$ (solid ellipse in Figure 2c). For low-load situations (load factor ≤ 0.8), the blocking ratio is reduced by around $15\,\%$ (dotted ellipse in Figure 2c).

To summarize, the proposed admission and routing approach is able to almost eliminate the availability overfulfillment, it reduces the resource overbuild required for protection considerably, and it allows more services to be accommodated in the Germany50 network. We will now include results for the other networks and stricter compliance targets of 0.995 and 0.999. For the protection overbuild and the compliance ratio, the relative change when enabling surplus sharing is almost independent from the load factor. Therefore, we show values averaged over all load factors in the following. Furthermore, we show the attainable load increase at a blocking ratio of 10^{-2} but do not consider the change in the blocking ratio for low network loads as we did above. Notice that in each of the following figures the bar for Germany50 and $f_\mathrm{t} = 0.99$ relates to the values presented in Figure 2.

Figure 3 shows the reduction in protection overbuild averaged over all load factors. The behavior is very similar for all networks. The highest reduction of more than $36\,\%$ is possible for a compliance target of $f_\mathrm{t} = 0.99$. For stricter targets the reductions decrease. Nevertheless, a significant reduction of around $10\,\%$ is possible in all networks even for a compliance target of $f_\mathrm{t} = 0.999$.

Figure 4 compares the compliance ratio averaged over all load factors with and without surplus sharing. Figure 4a depicts the case without surplus sharing. It is visible that the resulting compliance ratio is above 0.999 for all networks



(a) Average protection overbuild.

(b) Compliance ratio (average: large marks, decile: medium-sized marks, minimum: small marks). The dashed line indicates the operator target $f_t = 0.99$.

(c) Average blocking ratio. The y-axis is in log scale. The error bars depict the 95 %-confidence intervals for the mean.

Fig. 2. Results for the Germany50 network with a compliance target of $f_t = 0.99$ for various network loads.

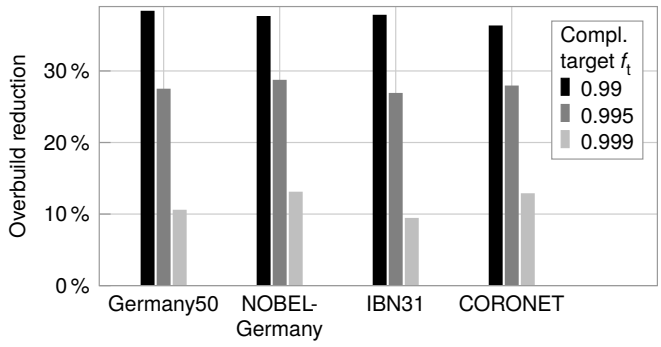


Fig. 3. Reduction in protection overbuild.

almost independently of the actual compliance target in use. Consequently, among the scenarios in this study, the availability overfulfillment is the highest for a compliance target of $f_t = 0.99$. As mentioned earlier, the high overfulfillment mainly originates from the coarse availability granularity of the protection—a backup path raises the compliance probability considerably even though a small increase might have been sufficient. Figure 4b shows the compliance ratio with surplus sharing enabled. It can be seen that the overfulfillment is reduced significantly in all networks. For the CORONET network, the compliance ratio matches the compliance target tightly. For the other networks, a certain margin persists.

Finally, Figure 5 shows the increase in attainable network load for a blocking ratio of 10^{-2}. The highest increase is achieved for a target compliance of 0.99. Stricter compliance targets result in less load increase. To a certain degree, the load increase is related to the overfulfillment reduction. Stronger overfulfillment reductions (see Figure 4) yield stronger increases in network load. Except for the network IBN31, load increases of more than 11 % for a target compliance of 0.99 are achieved by our admission and routing approach. For a target compliance of 0.995, load increases between 3.6 % and 7.9 % are realized.

The results show that the approach we propose is able to substantially increase the amount of services a network operator can accommodate in its network.

VI. Conclusion

In order to improve network efficiency, operators have to reduce margins that are present in the network. In this work, we have identified availability overfulfillment as a type of margin that appears unintentionally due to topological or protection-related constraints. Since availability overfulfillment ties up valuable network resources, an overfulfillment reduction can improve the network efficiency.

We have proposed a service admission and routing approach that is able to reduce the availability overfulfillment by sharing surplus availability among services. More precisely, our algorithm considers the probability of SLA compliance of connection services with and without dedicated path protection. Excess compliance probability of one service is accumulated as surplus, which is used to relax availability requirements of other services.

The simulation results confirm that the availability overfulfillment is reduced or even eliminated, and that fewer resources have to be dedicated to protection. As a result, the blocking ratio of new connection services is reduced significantly and more services can be accommodated in the network. The results also show that, even though the average compliance ratio matches the operator target accurately, the compliance ratio can fall below the target level for certain periods of time due to statistical variations. Further work is required to incorporate this behavior into the overall model. Nevertheless, the presented approach is a powerful mechanism to exploit existing margins in the network and to improve network efficiency.

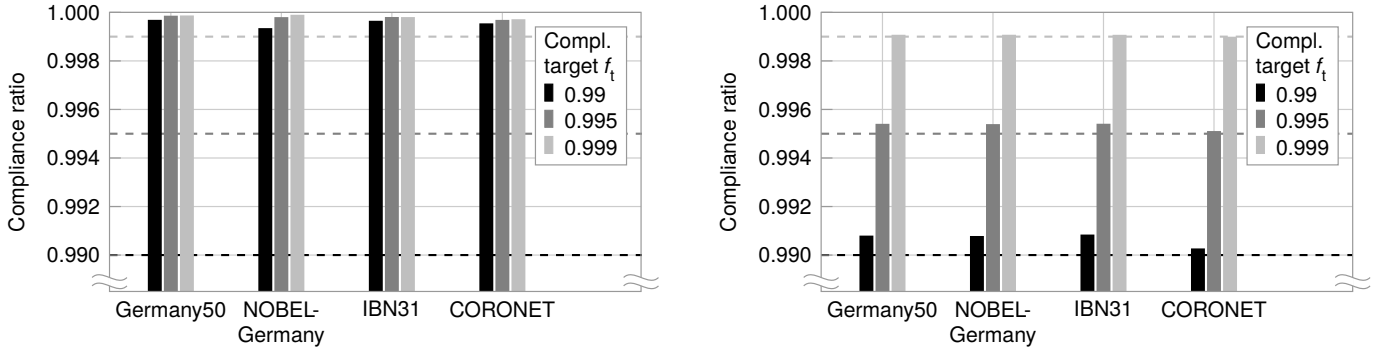
References

[1] P. J. Winzer and D. T. Neilson, "From Scaling Disparities to Integrated Parallelism: A Decathlon for a Decade," *Journal of Lightwave Technology*, vol. 35, no. 5, pp. 1099–1115, 2017.

[2] Y. Pointurier, "Design of Low-Margin Optical Networks," *Journal of Optical Communications and Networking*, vol. 9, no. 1, pp. A9–A17, 2017.

[3] L. Zhou and W. D. Grover, "A Theory for Setting the "Safety Margin" on Availability Guarantees in an SLA," in *5th International Workshop on Design of Reliable Communication Networks (DRCN)*. IEEE, 2005.

[4] C. Meusburger and D. A. Schupke, "Method to Estimate the Break-Even Point Between SLA Penalty Expenses and Protection Costs," in *2008 Optical Fiber Communication Conference (OFC)*. Optical Society of America, 2008.



(a) Without surplus sharing.

(b) With surplus sharing.

Fig. 4. Compliance ratio. The dashed lines correspond to the targeted compliance probabilities.

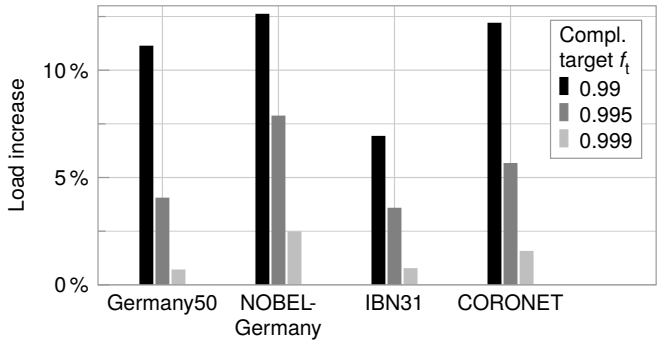


Fig. 5. Increase in attainable network load for a blocking ratio of 10^{-2}.

[5] Verizon Communications Inc., "WAVELENGTH SERVICES SOLUTION +; Part IV: Service Level Agreement," https://www.verizon.com/business/service_guide/reg/cp_wss_plus_sla.pdf, 2020, Online, accessed 31 March 2022.

[6] Windstream Enterprise, "COMPLETE DATA MPLS: SERVICE LEVEL AGREEMENT," https://www.windstreamenterprise.com/wp-content/uploads/legal/MPLS_SLA.pdf, 2019, Online, accessed 31 March 2022.

[7] M. Tornatore, G. Maier, and A. Pattavina, "Availability Design of Optical Transport Networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 8, pp. 1520–1532, 2005.

[8] L. Song, J. Zhang, and B. Mukherjee, "Dynamic Provisioning with Availability Guarantee for Differentiated Services in Survivable Mesh Networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 35–43, 2007.

[9] M. Xia, J. ho Choi, and T. Wang, "Risk Assessment in SLA-Based WDM Backbone Networks," in *2009 Optical Fiber Communication Conference (OFC)*. Optical Society of America, 2009.

[10] A. J. González and B. E. Helvik, "Guaranteeing Service Availability in SLAs; a Study of the Risk Associated with Contract Period and Failure Process," *IEEE Latin America Transactions*, vol. 8, no. 4, pp. 410–416, 2010.

[11] L. Mastroeni and M. Naldi, "Violation of Service Availability Targets in Service Level Agreements," in *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2011.

[12] O. Gerstel and G. Sasaki, "Meeting SLAs by Design: a Protection Scheme with Memory," in *2007 Optical Fiber Communication Conference (OFC)*. Optical Society of America, 2007.

[13] M. Xia, M. Batayneh, L. Song, C. U. Martel, and B. Mukherjee, "SLA-Aware Provisioning for Revenue Maximization in Telecom Mesh Networks," in *2008 IEEE Global Telecommunications Conference (GLOBECOM)*. IEEE, 2008.

[14] F. Dikbiyik, M. Tornatore, and B. Mukherjee, "Exploiting Excess Capacity, Part II: Differentiated Services Under Traffic Growth," *IEEE/ACM Transactions on Networking*, vol. 23, no. 5, pp. 1599–1609, 2015.

[15] X. Chen, M. Tornatore, S. Zhu, F. Ji, W. Zhou, C. Chen, D. Hu, L. Jiang, and Z. Zhu, "Flexible Availability-Aware Differentiated Protection in Software-Defined Elastic Optical Networks," *Journal of Lightwave Technology*, vol. 33, no. 18, pp. 3872–3882, 2015.

[16] A. Das, "Maximizing Profit Using SLA-Aware Provisioning," in *2012 IEEE Network Operations and Management Symposium (NOMS)*. IEEE, 2012.

[17] A. Nafarieh, S. C. Sivakumar, W. Phillips, and W. Robertson, "Memory-aware SLA-based Mechanism for Shared-Mesh WDM Networks," in *3rd International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT)*. IEEE, 2011.

[18] M. Xia, C. U. Martel, M. Tornatore, and B. Mukherjee, "Service Cluster: A New Framework for SLA-Oriented Provisioning in WDM Mesh Networks," in *2009 IEEE International Conference on Communications (ICC)*. IEEE, 2009.

[19] M. Tornatore, D. Lucerna, L. Song, B. Mukherjee, and A. Pattavina, "Dynamic SLA Redefinition for Shared-Path-Protected Connections with Known Duration," in *2008 Optical Fiber Communication Conference (OFC)*. Optical Society of America, 2008.

[20] D. Lucerna, M. Tornatore, B. Mukherjee, and A. Pattavina, "Availability Target Redefinition for Dynamic Connections in WDM Networks with Shared Path Protection," in *7th International Workshop on Design of Reliable Communication Networks (DRCN)*. IEEE, 2009.

[21] ——, "Trading availability among shared-protected dynamic connections in WDM networks," *Computer Networks*, vol. 56, no. 13, pp. 3150–3162, 2012.

[22] T. Enderle, "On the Impact of Billing Cycles on Compensations in Network SLAs," in *25th International Conference on Optical Network Design and Modeling (ONDM)*. IEEE, 2021.

[23] L. Takács, "ON CERTAIN SOJOURN TIME PROBLEMS IN THE THEORY OF STOCHASTIC PROCESSES," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 8, pp. 169–191, 1957.

[24] A. Birolini, *Reliability Engineering: Theory and Practice*, 8th ed. Berlin, Germany: Springer-Verlag, 2017.

[25] S. Orlowski, R. Wessäly, M. Pióro, and A. Tomaszewski, "SNDlib 1.0—Survivable Network Design Library," *Networks: An International Journal*, vol. 55, no. 3, pp. 276–286, 2010.

[26] A. Eira, J. Pedro, and J. Pires, "Optimized Design of Shared Restoration in Flexible-Grid Transparent Optical Networks," in *2012 Optical Fiber Communication Conference (OFC)*. Optical Society of America, 2012.

[27] A. Von Lehmen, R. Doverspike, G. Clapp, D. M. Freimuth, J. Gannett, A. Kolarov, H. Kobrinski, C. Makaya, E. Mavrogiorgis, J. Pastor, M. Rauch, K. K. Ramakrishnan, R. Skoog, B. Wilson, and S. L. Woodward, "CORONET: Testbeds, Demonstration, and Lessons Learned [Invited]," *Journal of Optical Communications and Networking*, vol. 7, no. 3, pp. A447–A458, 2015.

[28] S. Verbrugge, D. Colle, P. Demeester, R. Huelsermann, and M. Jaeger, "General Availability Model for Multilayer Transport Networks," in *5th International Workshop on Design of Reliable Communication Networks (DRCN)*. IEEE, 2005.

[29] J. Zhang, K. Zhu, H. Zang, and B. Mukherjee, "A New Provisioning Framework to Provide Availability-Guaranteed Service in WDM Mesh Networks," in *2003 IEEE International Conference on Communications (ICC)*. IEEE, 2003.

[30] J. Sommer and J. Scharf, "IKR Simulation Library," in *Modeling and Tools for Network Simulation*, K. Wehrle, M. Güneş, and J. Gross, Eds. Springer, Berlin, Heidelberg, 2010, pp. 61–68.