Dimensioning resilient optical grid/cloud networks

Chris Develder
Ghent University – iMinds, Ghent, Belgium
Massimo Tornatore
Politecnico di Milano, Milan, Italy
M. Farhan Habib
University of California, Davis, CA, USA
Brigitte Jaumard
Concordia University, Montreal, Quebec, Canada

ABSTRACT
Optical networks play a crucial role in the provisioning of grid and cloud computing services. Their high bandwidth and low latency characteristics effectively enable universal users' access to computational and storage resources that thus can be fully exploited without limiting performance penalties. Given the rising importance of such cloud/grid services hosted in (remote) data centers, the various users (ranging from academics, over enterprises, to non-professional consumers) are increasingly dependent on the network connecting these data centers, that must be designed to ensure maximal service availability, i.e., minimizing interruptions. In this chapter we will outline the challenges encompassing the design, i.e., dimensioning, of large-scale backbone (optical) networks interconnecting data centers. This amounts to extensions of the classical routing and wavelength assignment algorithms (RWA) to so-called anycast RWA, but also pertains to jointly dimensioning not just the network but also the data center resources (i.e., servers). We specifically focus on resiliency, given the criticality of the grid/cloud infrastructure in today's businesses, and, for highly critical services, we also include specific design approaches to achieve disaster resiliency.

INTRODUCTION
Back in the 1960s, John McCarthy envisioned the concept of "computation as a public utility", making computing power equally easily accessible as the classical utilities that provide users with water, gas, and electricity. That seminal idea reappeared in the 1990s under the form of grid computing, borrowing its name from the power grid, where "the grid" was aimed to be a highly powerful computing resource that scientists could easily tap into for performing challenging tasks. Similarly, today's cloud computing paradigm is built on the idea of relieving the user from worrying about the resources required to run applications and to store data, as well as on the idea of enabling access to such applications and data from basically any device. Clearly, such concept can be made possible only through a high capacity and low latency network that connects the user to "the cloud", i.e., the distributed computing/storage resources. Undeniably, development of optical network technology has been a major driver that enabled the realization of such grids/clouds.
The rise of broadband access networks, and high speed optical networking in wide area networks (WAN) has increased the geographical scale of distributed computing paradigms, extending their range from on-site computing facilities to the cost-efficient aggregation of IT resources for both processing and storage in large scale data centers. These now can supply a broad spectrum of applications, serving a wide audience ranging from end consumers, over business users, to scientists requiring high performance computing (HPC) facilities. Basic concepts underlying so-called grid technology, originating in the e-Science domain (e.g., to process massive data flows from the large hadron collider (LHC) at CERN, in Switzerland, used for the Higgs boson discovery), meanwhile evolved to today's cloud applications. For a more elaborate discussion of these applications, as well as relevant optical technology that can help to meet their challenging requirements, we refer to

(Develder, De Leenheer, et al., 2012). The resulting optical grid/cloud constituents are summarized in Figure 1.

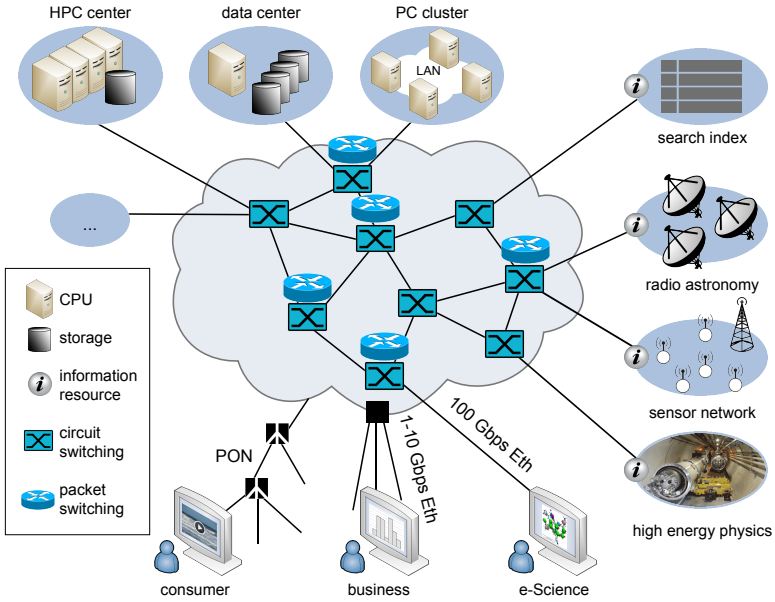


Figure 1. An optical grid/cloud interconnects various data sources (experimental facilities, sensors, etc.) to infrastructure for data storage and processing (data centers, high performance computing, etc.) to deliver services to various types of users. Such a distributed architecture owes its success to optical networking infrastructure, both in backbone and access networks. (Adapted from (Develder, De Leenheer, et al., 2012)).

Given that virtually all types of today's applications heavily rely on network connectivity, as well as the IT resources that constitute the work horses of the grid/cloud, it is crucial that this infrastructure is able to provide the services *resiliently*. Protection of cloud service and traditional traffic protection vary in nature. In the optical layer (which is the focus area of this chapter), protection of traffic between two nodes is generally provided by provisioning a backup path between the nodes. In an optical cloud, a specific service/content is generally available from multiple locations (such as data centers or servers). Thus, we no longer need to provide backup path between the requesting node and the server node as the service can be continued/restored from another location after a failure. Cloud service protection also includes protection of content that is an integral part of the service. Routing and protection of connections and services largely depend on the placement of content, which itself is another important problem in a cloud. Thus, protection of services in a cloud has different requirements than traditional traffic and can benefit from distinct protection methods. Moreover, large-scale network failures due to natural disasters and intentional attacks pose a major problem. Although upper layer schemes (such as TCP retransmission, IP layer re-routing, etc.) are in place to recover from a network failure, they are incapable of dealing with disaster failures, mostly since they are spatially correlated and may require cross-layer signaling between the optical backbone and the upper layers.

The remainder of this chapter is structured as follows: we start by introducing the general problem of resilient optical grid/cloud dimensioning, highlighting the fundamental principles of anycast routing and relocation. After providing a short literature overview of the resulting anycast routing and wavelength allocation (ARWA) problem, we will discuss two particular problems. The first is to dimension jointly the server and network capacities for an optical cloud, for which we propose an Integer Linear Programming (ILP) formulation and a scalable solving method based on column generation. Results on a case study on a European network topology are then presented. This first problem is rather generic, and will decide to route traffic to possibly 2 different locations (one for failure-free conditions, the other in case of resource failures), using a generic failure model (i.e., the

shared risk group concept). We then move to a second problem, presenting a specific approach to provide protection against disasters, in a scenario where we need to provide content placement (replication) in geographically scattered data centers, and network connectivity towards them. Here, data will thus be served at multiple (esp. more than 2) locations. Also in this case, we provide various solution methods (based on ILP, LP relaxations and heuristics) for that case of disaster resiliency. Finally, we conclude the chapter with a summary and possible future work.

We would like to point out that the problem of coordinated protection in cloud/network infrastructure is very challenging and presents a very high computational complexity. Computationally effective approaches must be devised for the solution of these classes of problems on realistic network instances. In this chapter we therefore describe possible solution methods based on LP relaxations, heuristics, and column generation, without claiming completeness, but with the intent of providing guidelines to devise scalable solution methods for similar problems arising in the context of resilient optical grid/cloud networks.

THE PROBLEM OF RESILIENT GRID/CLOUD DIMENSIONING

In this chapter, we study how to dimension optical grids/clouds, answering the question: What amount of resources, i.e., both network and server capacity, do we need to cater for a given demand for grid/cloud services? Our case studies will assume a particular user request can be served by a single data center, i.e., at a particular location (which however can be chosen out of a set of candidate sites). This assumption is representative of so-called bag-of-tasks applications in the e-science domain, as well as requests for the provisioning of virtual machines (VMs) in case of infrastructure-as-a-service (IaaS) clouds. The applications that we consider in our dimensioning studies can be seen as abstractions of any of these services, as our modeled applications imply non-negligible bandwidth and server (storage and/or computation) requirements.

In light of the high bandwidth requirements associated with such grid/cloud applications (Develder, De Leenheer, et al., 2012), we focus on optical circuit-switched networks exploiting wavelength division multiplexing (WDM). We particularly focus on backbone networks interconnecting various geographically spread data centers: intra-data center networks connecting the various server racks within a single data center will not be further discussed here (for a recent discussion of optics within the data center, see (Glick, Krishanmoorthy, & Schow, 2011). In the domain of backbone WDM networks, a substantial body of research literature already has widely addressed the offline dimensioning network problem. Yet, in the particular grid/cloud context addressed here, those works are not directly applicable. First of all, we need also to consider (and optimize the dimensions, hence cost, of) the *server resources and their location*, in addition to the network resources (i.e., wavelengths on each of the network links in WDM terms). Secondly, in classical WDM network design a so-called traffic matrix is assumed, specifying the amount of requests between any pair of optical network nodes. In a grid/cloud context we however do not a priori know the end points of such requests: grid/cloud users typically do not care where exactly their workload is processed ("in the cloud"), and therefore freedom arises to choose the most appropriate location for the data center to serve their requests. This concept of routing where the destination is not fully specified a priori, but rather can be freely chosen among a set of candidate locations is generally known as *anycast routing* (Partridge, Mendez, & Milliken, 1993). Therefore, the classical routing and wavelength assignment (RWA) needs to be rephrased as anycast routing and wavelength assignment (ARWA). We provide an overview of initial work in this arena in the following section.

Also resiliency, even before the advent of grids/clouds, has always been a major concern in optical networks given the traffic volumes affected by, e.g., a single failing optical channel or fiber. A popular basic principle to offer protection against failures of optical equipment is that of path protection: a primary path, running between source and destination of the request, is protected by a disjoint alternate (backup) path that does not share any possibly failing network resource with the primary. In light of the aforementioned anycast routing principle, we have proposed the idea of exploiting relocation (Buysse, De Leenheer, Dhoedt, & Develder, 2009), illustrated in Figure 2: the

backup path may end in a destination that possibly differs from that under failure-free conditions. Thus, we can save network resources (since the path to an alternate data center may be shorter than any other alternate path to the original one that is disjoint from the primary), as well as be protected against failure of server resources at the primary destination. Such relocation is crucial in both studies that we will present below. But first, let us briefly sketch the existing body of research work in the context of ARWA.

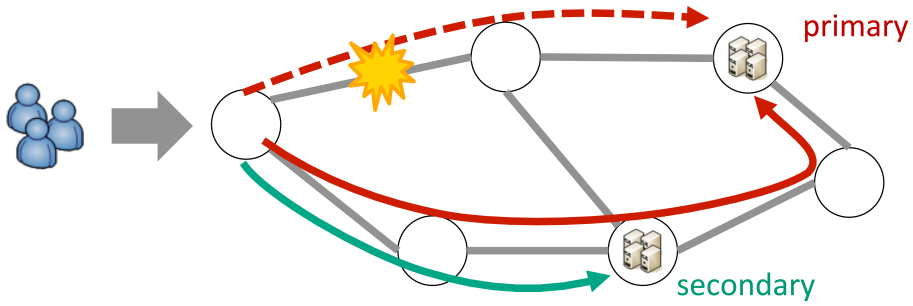


Figure 2. The relocation principle: since grid/cloud users are not very much concerned about the exact location where requests are being served, we can choose an alternate location to save network resources when failures affect the primary destination.

ANYCAST ROUTING AND WAVELENGTH ASSIGNMENT (ARWA)

In essence, the anycast routing problem amounts to finding a path from a source to a destination to be chosen among a given set of candidate destinations, while minimizing a certain cost (bandwidth used, delay, etc.). This has been considered in packet switched IP networks (Tim Stevens et al., 2007), or even optical burst switched (OBS) networks (Bathula & Elmirghani, 2009; De Leenheer et al., 2006). As indicated above, we consider anycast routing in optical circuit-switched (OCS) networks, where it amounts to the so-called anycast routing and wavelength assignment (ARWA) problem. We thus need to find so-called wavelength paths and minimize, e.g., the total number of wavelengths used summed over all network links, and/or the load on the links (Tang, Jia, Wang, & Wang, 2003).

The *offline* problem, where all requests are given at once and are considered to be static, has been proposed to be solved in three sequential phases (Din, 2005): (1) decide on the destination for each request, (2) route the paths for each (source, destination)-pair, (3) do the wavelength assignment for each path. A heuristic algorithm based on simulated annealing and genetic algorithms has been shown to outperform the former phased strategy (Din, 2007). Also applying on heuristics, (Hyytiä, 2004) and (Walkowiak, 2010) address generalized offline RWA where the requests are not solely anycast, but also include unicast (and in the case of Hyytiä et al. also multicast) requests. An *online* routing problem (while also briefly raised by Walkowiak, 2010) is studied specifically by Bhaskaran et al., who assume the number of anycast sites varies over time, according to a time-varying load (Bhaskaran, Triay, & Vokkarane, 2011). In (Bathula & Elmirghani, 2009), the authors propose anycast routing methods to improve the performance of reconfigurable WDM networks under the variations in the IP traffic. In (Tim Stevens, De Leenheer, De Turck, Dhoedt, & Demeester, 2006), authors show that the anycast routing problem can be reduced to unicast routing. Consequently, unicast routing algorithms (complemented by some specific constraints) can be applied to compute an optimal path based on several server selection criteria and achieve an effective distribution of the job scheduling requests.

The studies mentioned so far do not address resiliency: they only find a single working path. The *online* problem of finding not only a primary, but also a backup path, is addressed in (She, Huang, Zhang, Zhu, & Jue, 2007): they propose an algorithm based on an auxiliary graph, modeling also grooming, to find a working and backup route for a single anycast request. The *offline* problem, which is the focus of this chapter, has been addressed in (Walkowiak & Rak, 2011) by considering joint optimization of anycast and unicast requests that are protected against single link failures with shared backup paths.

All works discussed above only address dimensioning the capacity of the optical network. However, here we are interested not only in the network, but also in the data center resources (i.e., the amount of servers for storage and/or computation). The *online* routing problem, taking both network and server constraints into account, has been discussed in, e.g., (Demeyer, De Leenheer, Baert, Pickavet, & Demeester, 2008; T. Stevens et al., 2009). Here, we are focusing on the *offline* dimensioning of a grid/cloud for a given set of multiple requests. Another approach related to resilient design methods for virtual networks minimizing the cost or the latency of the virtual network in presence of link and data center failures is reported in (Barla, Schupke, & Carle, 2012).In our previous study, we considered the basic case without offering resiliency, and propose a phased approach to both decide on the location of the data center sites, their capacity, and subsequently the network capacity (Develder, Mukherjee, Dhoedt, & Demeester, 2009). Here, we will further integrate the dimensioning process and offer resiliency, subsequently also explore the specific case of disaster resiliency, with the additional complexity of placement of contents in a selected subset of the data centers.
In the approaches discussed below, we present solutions to resilient dimensioning, jointly determining server and network capacities. Note that in our case studies, a single request will eventually be served at a single location (which however might depend on the failure condition): accommodating requests that comprise a set of interdependent tasks is a largely unaddressed problem, although some have proposed solutions to the non-resilient *online* routing and scheduling problem (X. Liu et al., 2009; X. Liu, Qiao, Yu, & Jiang, 2010).

The following section will discuss a first, generic problem of dimensioning jointly the server and the network resources to resiliently provide cloud services, exploiting so-called relocation. After this, we will present a second study on providing disaster resiliency.

JOINT DIMENSIONING OF RESILIENT SERVER AND NETWORK INFRASTRUCTURE

This first case study solves the offline dimensioning problem for optical grids/clouds comprising a backbone network of interconnected optical cross connects (OXCs), where some of these are collocated with data centers. A formal problem statement, as depicted in Figure 3, is:

Given

- The *topology*, comprising the sites where the grid/cloud requests originate, as well as the optical network of OXCs interconnecting them;
- The *demand* stating the amount of requests originating each single site; and
- The *survivability* requirements, detailing the failure scenarios that should be protected against,
- Candidate data center site locations, of out which up to? *K* should be chosen,

Find

- The *K data center locations* where server infrastructure should be provided;
- The *destination site(s) and routes* thereto for each of the requests in the demand, under each possible failure scenario; and
- The *network and server capacity* to provide on each of the network links (i.e., number of wavelengths) and in each of the K data centers (i.e., number of servers).

Such that the total network and data center resource capacity is minimized.

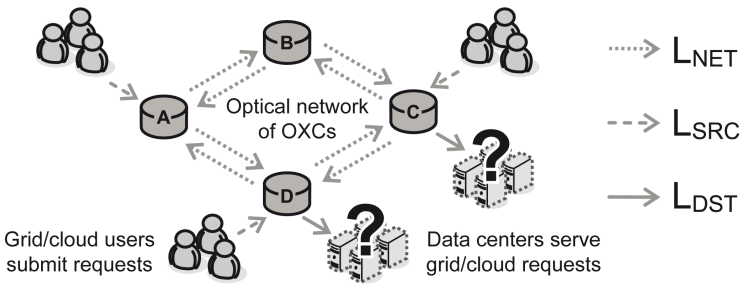


Figure 3. The problem we address is to dimension the optical backbone network comprising optical cross connects (OXCs) and data centers, required to meet a given demand of grid/cloud requests with

known sources. We will decide on the location of the data centers, as well as the amount of bandwidth and server capacity. (Adapted from (Develder et al., 2011).)

The 'requests' in this problem statement will be expressed in abstract unit capacities, and can be interpreted as a certain volume of "jobs" from grid applications (see above, especially bag-of-task types), but in the cloud case, one can use the same model and interpret "requests" to refer to virtual machines (VMs) to be provisioned in IaaS clouds. The demand will be expressed as request arrival intensity, and a unit capacity will be associated with a certain amount of server capacity (to be interpreted as storage and/or processing power, e.g., a single CPU) as well as a certain network bandwidth (a wavelength in the assumed optical circuit-switched network context). Our model is generic and can be used both for data- and computing-intensive grid/cloud scenarios.

The 'survivability' addressed will involve protection against single failures of either a bi-directional network link, or a server at one of the data center locations. Yet, note that the mathematical model we use to solve the problem is generic and can cater for any failure that can be represented as a so-called shared risk link group (SRLG): a set of links that can simultaneously fail, because they share a common dependency. Thus, a failing server will be modeled by a failure of a link connecting it to the rest of the network.
We will cater for failures with minimal resource requirements by sharing/reusing capacity: the same wavelength (and similarly, the same server in a data center) may be used as backup capacity under disjoint failure scenarios, i.e., to protect against failures of different resources. Also, we will exploit relocation as explained before, if this allows for a reduction in the overall network and server capacity.

Solution approach and network model

Our solution to the above resilient optical grid/cloud dimensioning problem comprises two steps:

- *Step 1:* Find the K best locations where it is most beneficial (as to minimize both network and server resources) to install data centers.
- *Step 2*: Determine the amount of network capacity (i.e., number of wavelengths on each of the link) and data center capacity (i.e., number of servers) to install, based on appropriate choices of so-called working paths (under failure free conditions) and backup paths for all grid/cloud traffic.

Clearly, the resource dimensions found in step 2 will depend on the choice of K in step 1 (see further, e.g., Figure 4). Solving the two steps jointly however is quite complex (see e.g., first attempts in (Jaumard et al., 2012). Here we will limit the discussion to the 2-stage approach.

Step 1 can be rephrased as a well-known k-medoids clustering problem (Develder et al., 2009), for which, e.g., heuristic approaches exist (Park & Jun, 2009). Yet, in our context we can quickly solve the problem using an integer linear programming (ILP) approach, as explained in detail below. Step 2 will be solved using linear programming techniques as well. Whereas it is possible to elegantly formulate the problem as a conventional ILP, solving the Step 2 problem for large problem instances (in terms of topology size and number of unit requests) is not scalable, and hence we successfully use a column generation (CG) approach. This generic technique to solve large problems limits the number of so-called configurations that are explicitly considered in the model, where typically a combinatorial explosion of the number of possible configurations occurs.
In CG, the linear program (LP) only considers a subset of all possible configurations. (In our case, a configuration comprises a combination of a working and a backup path.) Starting from an initial (limited) set of configurations C, a so-called restricted master problem (RMP) determines the optimal combination of configurations (restricted to those in C). In a next step, a so-called pricing problem finds a new configuration c that could improve the value of the objective function of the RMP. If such a c can be found, it is added to C, whereupon the master problem (with the extended C) is solved again. Using that as an input, the pricing problem (PP) looks for a new configuration, etc.: this process of adding new configurations and re-solving the RMP is repeated until the PP no longer finds new configurations. For more details on column generation, we refer to, e.g., (Desrosiers & Lübbecke,

2005; Vanderbeck & Wolsey, 1996) and specifically for its application to the RWA problem to (Jaumard, Meyer, & Thiongane, 2009).

Before presenting the ILPs, we first introduce the network model as illustrated in Figure 3 and its variables:

- $G = (V, L)$ is the directed graph, with nodes V and directed edges L, constituting the optical grid/cloud.
- $V = V_{SRC} \cup V_{NET} \cup V_{DST}$ is the set of all nodes, which is partitioned in the set of optical switches (the OXCs) V_{NET}, the data center sites V_{DST} (with $|V_{DST}| = K$), and the explicitly modeled sources of the requests V_{SRC}.
- $L = L_{SRC} \cup L_{NET} \cup L_{DST}$ includes the backbone network links L_{NET} interconnecting the OXCs, the links L_{SRC} that originate at the sources and hence can be interpreted as representations of the access network links (where possibly an appropriate cost factor can be added), and the L_{DST} links towards a data center. Note that L_{SRC} and L_{DST} usually have no direct correspondence to a single real-life link; e.g., the data source can be accessing the core through a particular, multi-hop, access network technology. Yet, in particular L_{DST} are modeled to be able to represent failing servers as link failures.
- Δ_v indicates the number of unit requests originating at node $v \in V_{SRC}$. Here, a unit request is interpreted as requiring a single wavelength as well as a single server. (The model can be extended by only using $\Delta_v$for network capacity requests, and decouple the server resource requirements in a separate server demand Γ_v; in the following we however stick to corresponding network and server requirements in a single demand vector Δ_v.)
- The set of SRLGs S, with elements $s \in S$ representing a set of links that could be jointly affected by the same root failure cause and hence simultaneously fail. Note that this modeling as SLRGs is very generic, and in our numerical case studies we will particularly focus on single failures of bidirectional links or servers.

Note that the network links that will be used to derive the total network capacity comprise only L_{NET} (where wavelengths will be the units of capacity). In fact, links L_{DST} will rather be used to determine the total server capacity: the number of capacity units on link $l \in L_{DST}$ will represent the number of servers to install in the data center represented by the node $v \in V_{DST}$ that l connects to.

Step 1: Choosing the *K* data center locations

First of all, we find the best locations to install data centers: we decide on V_{DST}. Thus, we assume only $V_{SRC} \cup V_{NET}$ is given, and we need to pick and choose K locations out of V_{NET} (or a subset thereof) where to attach data center nodes (thus adding $v \in V_{DST}$ to the topology to be further used in Step 2). We define and solve an ILP, under the following simplifying assumptions: (a) all requests originating at the same source $v \in V_{SRC}$ will be sent to the same data center location, (b) the choice for a particular data center location will be based on shortest path routing. The latter implies that a source v will send its requests to a data center attached to v' only if for some routing metric $h_{vv'}$ expressing the 'distance' between v and v' this is the minimum over all values $h_{vv''}$ (i.e., over all K possible destinations v'').

The ILP for choosing K locations, as we originally proposed without considering resilience (Develder et al., 2009), uses the following decision variables and given parameters:

- t_v is a binary decision variable that will be 1 if the site v is chosen as one of the K data center locations,
- $f_{vv'}$ is a binary decision variable that will be 1 if and only if requests from source node v are sent to a data center attached to v',
- $h_{vv'}$ is a given parameter that accounts for the cost of having a unit request being sent from v to v'.

Since we aim to minimize resource capacity, which for the network amounts to the number of used wavelengths summed over all links, and a unit request represents the use of a single wavelength, $h_{vv'}$ will be measured as hop count of the shortest path between v and v'. Our ILP thus becomes:

$$\min \sum_{v \in V_{NET}} \sum_{v'' \in V_{NET}} \Delta_v \cdot h_{vv'} \cdot f_{vv'} \quad (1)$$

Subject to:

$$\sum_{v \in V_{NET}} t_v = K \quad (2)$$

$$\sum_{v' \in V_{NET}} f_{vv'} = 1 \qquad \forall v \in V_{NET} \quad (3)$$

$$f_{vv'} \leq t_{v'} \qquad \forall v, v' \in V_{NET} \quad (4)$$

To account also for capacity that will be required for backup against failures, we define[1] $h_{vv'}$ as the length of the shortest combination of two link-disjoint paths between v and v' (as found by, e.g., (Suurballe & Tarjan, 1984)).

Note that the problem of data-center location in a network is a special case of the generic problem of network hub location that applies to any kind of facilities that serve as switching, shipment and sorting points in a distribution system under different assumption and configurations. For an overview of such problems, we refer to (Alumur & Kara, 2008).

Step 2: Dimensioning the network and data centers

Once we fixed the K locations of the data centers in Step 1, we need to find the actual routes for each individual request, as to ensure minimal network and server capacity requirements. Note that this routing can differ from the simplifying assumptions taken in Step 1: multiple unit requests originating from the same source could in principle take different routes to different destinations, and those routes don't necessarily coincide with a shortest path as found to calculate the h_* metrics.

We will indeed look for routes such that backup capacity can be maximally shared among backup paths whose corresponding primary paths (i.e., those under failure-free conditions) are disjoint. Failures that we consider are single bidirectional network link failures, and single server failures. For the latter, we will assume $1{:}N$ server protection, implying that per N servers, we will add 1 extra for protection purposes. In terms of modeling, every failure scenario will be represented as a set of jointly failing modeled links, aka an SRLG: bidirectional link failures will be modeled as a set $s = \{l, l'\}$ comprising the two opposite directed links $l, l' \in L_{NET}$; server failures as a failing link $l \in L_{DST}$.

To protect against all possible failure scenarios, there are basically two fundamental approaches. The first is generally known as *failure-independent (FID)* rerouting (Y. Liu, Tipper, & Siripongwutikorn, 2005; Zang, Ou, & Mukherjee, 2003) or state-independent restoration (Xiong & Mason, 2002, 1999): the alternate route under any failure condition that impacts the original primary path under failure-free circumstances is always the same. This approach is detailed below in its column generation model. The second approach, for which we refer to the model in (Develder, Buysse, De Leenheer, Jaumard, & Dhoedt, 2012), is named *failure-dependent (FD)*, or state-dependent, restoration: the backup path to cater for a specific failure affecting the primary can be tailored to that particular failure.

For the FID case we resort to a column generation model. The so-called restricted master problem (RMP) will use a set of configurations as input, deciding how many times to use each configuration to meet the given requests while jointly minimizing the server and network capacity. A configuration c will be associated with a given source node v and comprise a combination of (i) a so-called working path, to use for a request originating at v under failure-free conditions, and a (ii) backup path to use under any failure scenario that affects the working path. Such configurations will be found by subsequently and iteratively solving the pricing problem (PP) as indicated before. To kick-start the problem, we clearly have to compose an initial set of configurations that can meet the requests. A procedure to find such initial configurations can be found in (Develder et al., 2011). Now, let's specify the RMP and PP formulations.

[1] Note that this choice is different from the definition of $h_{vv'}$ in (Develder, Mukherjee, Dhoedt, & Demeester, 2009), since there we did not consider resiliency yet, i.e., there was no backup path.

The parameters and decision variables of the column generation model are listed in Table 1.

Table 1. ILP model parameters/variables for the column generation model for Step 2.

Case	ILP model parameters/variables
Common	• c denotes a configuration, which is defined for a particular source node $v \in V_{SRC}$ • C_v is the set of all such configurations c associated with a given source $v \in V_{SRC}$ • $C = \bigcup_{v \in V_{SRC}} C_v$ is the set of all considered configurations for any source • S is the set of SRLGs, each $s \in S$ representing a given failure scenario we want to protect against • z_c is an integer decision variable that counts how many times configuration c is used • p_{cl}^W is a binary parameter, equaling 1 if and only if link l is used in the working (hence superscript W) path in configuration c • p_{cl}^B similarly is a binary parameter that equals 1 if link l is used in the backup path of c • w_l is an auxiliary variable counting the number of capacity units on link l. Note that 'capacity' will denote wavelengths for network links $l \in L_{NET}$, and number of servers for data center links $l \in L_{DST}$.
Restricted Master Problem (RMP)	• ρ_l is a parameter that will be used in the NR case, for our model's server links $l \in L_{DST}$ to model $1{:}N$ server protection, where $\rho_l = 1 + 1/N$. In any other case (for links $l \in L_{NET}$, or in cases where we do consider relocation even for $l \in L_{DST}$) we will have $\rho_l = 1$. • π_{cls}^W is a binary parameter that will be 1 if and only if the working path of configuration c traverses link l, and this link l remains unaffected by failure of s • π_{cls}^W is a binary parameter that will be 1 if and only if link l is crossed by the backup path of configuration c, while c is affected by failure of s.
Pricing Problem (PP)	• u_v^1 is the dual variable corresponding to the RMP demand constraint (6), • u_l^2 is the dual variable for the RMP constraint (7) determining working capacity, • u_{ls}^3 is the dual variable for the RMP constraint (8) to satisfy failure scenarios. • a_s^W is a binary variable that is set to 1 if the working path of the configuration is affected by s, thus, if any of the links $l' \in s$ is used on the working path (formally: $\exists l' \in s{:}\ p_{l'}^W = 1$). • The π_{ls}^W and π_{ls}^B now will be variables (rather than given parameters), but with the same semantics as for the RMP.

The **restricted master problem (RMP)** will fix the values of decision variables z_c. Its objective function (5) constitutes two terms, the first being the network capacity and the second the server capacity.

$$\min\left(\sum_{l \in L_{NET}} w_l + \alpha \cdot \sum_{l \in L_{DST}} w_l\right) \quad (5)$$

Note that our formulation assumes a linear relation between the server and network requirements associated with a request (represented by the demand Δ_v associated with source node v). While the formulation can straightforwardly be extended to account for arbitrary server requirements[2], we stick

[2] E.g., via an extra parameter Γ_v, then adding a factor Γ_v/Δ_v to the z_v occurrences in (7)-(8) for links $l \in L_{DST}$.

to the linear relation assumption to not overload the model at this point. The demand requirement constraints thus amount to:

$$\sum_{c \in C_v} z_c \geq \Delta_v \quad \forall v \in V_{SRC} \tag{6}$$

To determine server capacity in the case where we do not relocate (noted as *NR*), we can account for the extra server resources by quite simply using an overprovisioning factor (rather than modeling server failures explicitly via failing links $l \in L_{DST}$). Therefore, we introduce a given parameter:

- ρ_l will be used in the NR case, for our model's server links $l \in L_{DST}$ to model $1{:}N$ server protection, where $\rho_l = 1 + 1/N$. In any other case (for links $l \in L_{NET}$, or in cases where we do consider relocation even for $l \in L_{DST}$) we will have $\rho_l = 1$.

We use this factor in the constraints counting the capacity (wavelengths for $l \in L_{NET}$, servers for $l \in L_{DST}$). Constraint (7) determines that capacity for failure-free conditions. For each of the failure cases, represented as an SRLG $s \in S$, we have capacity lower bounds (8). Herein, we have two summations: the first considers all unaffected configurations, the second those that are impacted by the failure of s. The auxiliary parameters (which in this RMP are given, but will become decision variables in the configuration finding pricing problem, PP) that we use to determine whether or not a configuration is impacted are:

- π^W_{cls} is a binary that will be 1 if and only if the working path of configuration c traverses link l, and this link l remains unaffected by failure of s
- π^W_{cls} is a binary that will be 1 if and only if link l is crossed by the backup path of configuration c, while c is affected by failure of s.

Using these parameters, we have the link dimension constraints[3]:

$$w_l \geq \rho_l \cdot \sum_{c \in C} p^W_{cl} \cdot z_c \qquad \forall l \in L \tag{7}$$

$$w_l \geq \rho_l \cdot \left(\sum_{c \in C} \pi^W_{cls} \cdot z_c + \sum_{c \in C} \pi^B_{cls} \cdot z_c\right) \qquad \forall s \in S, \forall l \in L \backslash s \tag{8}$$

As explained before, solving the above RMP for all possible configurations is not feasible. Starting from the latest RMP solution, the **pricing problem (PP)** will try to find new configurations to consider, which can lower the RMP objective function (5). A PP will be associated with a particular source node $v \in V_{SRC}$, and will use the values of dual variables corresponding to the RMP constraints:

- u^1_v is the dual variable corresponding to the RMP demand constraint (6),
- u^2_l is the dual variable for the RMP constraint (7) determining working capacity,
- u^3_{ls} is the dual variable for the RMP constraint (8) to satisfy failure scenarios.

(Considering the respective position of z_c in regard to the inequality sign in (6)–(8), u^1_v will be positive, while u^2_l and u^3_{ls} will be negative.)

The PP objective for a given source node $v \in V_{SRC}$ is to minimize the (negative) reduced cost, as stated in (9) (where the first explicit zero term is the coefficient of z_c in the RMP objective (5)). The decision variables of the PP are the configuration routing variables for working and backup path, p^W_l resp. p^B_l, with the same meaning as before but dropping the c index. Also the auxiliary π variables keep their original meaning, but now are variables that will follow out of the ILP solution (rather than a priori given as in the RMP).

$$\min \overline{cost}(p, \pi) = 0 - u^1_{v_{SRC}} + \sum_{l \in L} u^2_l \cdot \rho_l \cdot p^W_l + \sum_{s \in S} \sum_{l \in L} u^2_l \cdot \rho_l \cdot \left(\pi^W_{ls} + \pi^B_{ls}\right) \tag{9}$$

The first constraints are the traditional flow constraints, enforcing that the net flow into a node is either -1 (for the source node), $+1$ (for the chosen destination), or 0 otherwise:

$$\sum_{l \in in(v)} p^\star_l - \sum_{l \in (v)} p^\star_l = \begin{cases} -1 & \text{if } v = v_{SRC} \\ \sum_{l \in in(v)} p^\star_l & \text{if } v \in V_{DST} \\ 0 & \text{else} \end{cases} \tag{10}$$

[3] Note that we can in principle omit (7), since the failure-free case could be represented as $s = \emptyset$. Hence constraint (7) can be subsumed in (8) by adding this $s = \emptyset$ to S and noting that for $s = \emptyset$ we have $\pi^W_{cls} = p^W_{cl}$ and $\pi^B_{cls} = 0$. Then the corresponding u^2_l summation can be removed from the PP objective (9).

Subsequently, we assure that no loops occur (11) and that exactly one working and one backup destination is picked (12). The working and backup paths trivially need to be disjoint in terms of possible failure scenarios (13), since we protect against a single SRLG failure.

$$\sum_{l\in in(v)} p_l^\star \le 1,\ \sum_{l\in out(v)} p_l^\star \le 1 \qquad \forall v \in V,\ \star \in \{W,B\} \tag{11}$$

$$\sum_{l\in L_{DST}} p_l^\star = 1 \qquad \text{for } \star \in \{W,B\} \tag{12}$$

$$p_l^W + p_{l'}^B \le 1 \qquad \forall s \in S,\ \forall l,l' \in s \tag{13}$$

Now we still need to enforce that the π_*^W and π_*^B variables adhere to their definitions as given above. Therefore, we introduce additional auxiliaries a_s^W associated with an SRLG $s \in S$:

- a_s^W is a binary variable that is set to 1 if the working path of the configuration is affected by s, thus, if any of the links $l' \in s$ is used on the working path (formally: $\exists l' \in s$: $p_{l'}^W = 1$).

The definition of π_{ls}^W amounts to the logical equivalence relation $\pi_{ls}^W \equiv p_l^W \wedge \neg a_s^W$, which can be translated as linear constraints (14). Similarly, $\pi_{ls}^B \equiv p_l^B \wedge a_s^W$, or (15). Finally, logically a_s^W can be expressed as $a_s^W \equiv \bigvee_{l'\in s} p_{l'}^W$, as enforced by (16).

$$\left.\begin{array}{c} \pi_{ls}^W \ge p_l^W - a_s^W \\ \pi_{ls}^W \le p_l^W \\ \pi_{ls}^W \le 1 - a_s^W \end{array}\right\} \forall s \in S,\ \forall l \notin s \tag{14}$$

$$\left.\begin{array}{c} \pi_{ls}^B \ge p_l^B + a_s^W - 1 \\ \pi_{ls}^B \le p_l^B \\ \pi_{ls}^B \le a_s^W \end{array}\right\} \forall s \in S,\ \forall l \notin s \tag{15}$$

$$\left.\begin{array}{c} M \cdot a_s^W \ge \sum_{l'\in s} p_{l'}^W \\ a_s^W \le \sum_{l'\in s} p_{l'}^W \end{array}\right\} \forall s \in S \text{ with } M = |s| \tag{16}$$

Thus, the complete PP amounts to (9)–(16), assuming that we allow for relocation. Our case studies discussed next will compare that against the case where we do not relocate. In the latter case, we need an additional constraint forcing the working and backup path of the configuration to end at the same destination:

$$\sum_{l\in in(v)} p_l^{\mathrm{W}} = \sum_{l\in in(v)} p_l^{\mathrm{B}},\ \forall v \in V_{DST} \tag{17}$$

As a final remark, we note that our model can easily accommodate unicast requests simultaneously with anycast requests, as in (Walkowiak & Rak, 2011). For unicast requests, one simply can enforce $p_l^W = p_l^B = 1$ for the particular unicast request's destination link $l \in L_{DST}$, and setting $p_{l'}^W = p_{l'}^B = 0$ for all other server links $l' \in L_{DST} \setminus \{l\}$. In the following we will however consider anycast requests only.

Case study set-up

The basic questions we want to answer are the following:

- What is the best of our proposed data center location chooser strategies?
- What is the benefit of exploiting relocation, in terms of cost reduction, where cost is expressed as amount of network (wavelength) and server resources?
- What is the additional benefit of adopting a failure-dependent (FD) rerouting strategy versus a failure-independent (FID) rerouting approach?

To quantitatively answer these questions, we set up experiments on European backbone network topologies from (Maesschalck et al., 2003), which all comprise 28 nodes but have varying densities: the *basic* topology has 41 bidirectional links, while the *sparse* variant only has 34 and the *dense* one 60. We consider demands that contain a varying number of unit requests (i.e., asking for a single wavelength capacity towards a single server to be instantiated at a data center location of choice). Clearly demonstrating the scalability of our solution, the number of unit demands that constitutes a particular demand will vary in [10, 350]. Reported results for a particular demand size $x \in [10, 350]$ will be averages taken over 10 randomly generated instances of that particular size, where each unit request is equally likely to originate from any of the 28 nodes of the EU network.

The assessment of the benefit of relocation will be made for two scenarios: (i) single bidirectional network link failures only (denoted as 1L), or (ii) single failures that are either one failing bidirectional network link, or a single data center failure (1LS). For those data center failures, we will consider that up to $1/N^{\text{th}}$ of the server capacity may be impacted. In other words, we will adopt $1:N$ server protection. Using either relocation (RO), or not (NR), we thus consider effectively four scenarios:

- **NR**: No relocation, implying that to serve a request, the destination data center's location will be identical under a failure or in failure-free conditions
  - *1L*: A single bidirectional network link failure implies an SRLG comprising the two opposite directed links in the model.
  - *1LSN*: Single network link failures will be modeled as for 1L. To cater for single server failures, there is however no need to model failures. We can simply calculate the extra amount of servers that we need for $1:N$ server protection by adopting the overprovisioning factor ρ_l as explained before.
- **RO**: When relocation is optional, primary and corresponding backup paths can (but not necessarily need to) end at different data centers. The solution of the ILP will determine whether or not relocation is beneficial (in terms of cost, i.e., overall network and server capacity) for each individual unit request.
  - *1L:* Single link failures are modeled as in the NR case.
  - *1LS:* Single link failures will be still modeled as SRLGs, and similarly, data center failures will be modeled as failures of corresponding server links. Given that we adopt $1:N$ server protection, we will construct the model to have $1+N$ parallel links to each of the data center nodes. Out of these parallel links to a particular data center, at most one will fail, thus the singleton SRLGs corresponding to each modeled link $l \in L_{DST}$ will be considered.

The resulting settings for the ILP models of Step 2 are summarized below in Table 2. For the RO case, it is important to understand that this model indeed amounts to optional relocation, even under server failures (i.e., the 1LS case): we offer the choice either to add extra server capacity on a parallel link $l' \in L_{DST}$ to the same destination $v \in V_{DST}$, or to relocate to another server site $v' \in V_{DST} \setminus \{v\}$ (possibly implying extra network capacity on the path towards it).

Table 2. Model settings for the various resiliency strategies in Step 2.

Case	ILP model settings
1L	$S = \{\{l, l'\}: l, l' \in L_{NET}, l \text{ and } l' \text{ are each other's reverse}\} \triangleq S_{1L}$ $\rho_l = 1, \forall l \in L$ $\lvert L_{DST} \rvert = K$ (single server link per data center site)
1LS	$S = S_{1L} \cup \{\{l\}: l \in L_{DST}\}$ $\rho_l = 1, \forall l \in L$ $\lvert L_{DST} \rvert = (1+N) \cdot K$ (parallel server links at each data center site)
1LSN	$S = S_{1L}$ $\rho_l = 1 + \frac{1}{N}$ if $l \in L_{DST}$, else 1 $\lvert L_{DST} \rvert = K$ (single server link per data center site)

Quantitative benefits of exploiting relocation and failure-dependent rerouting

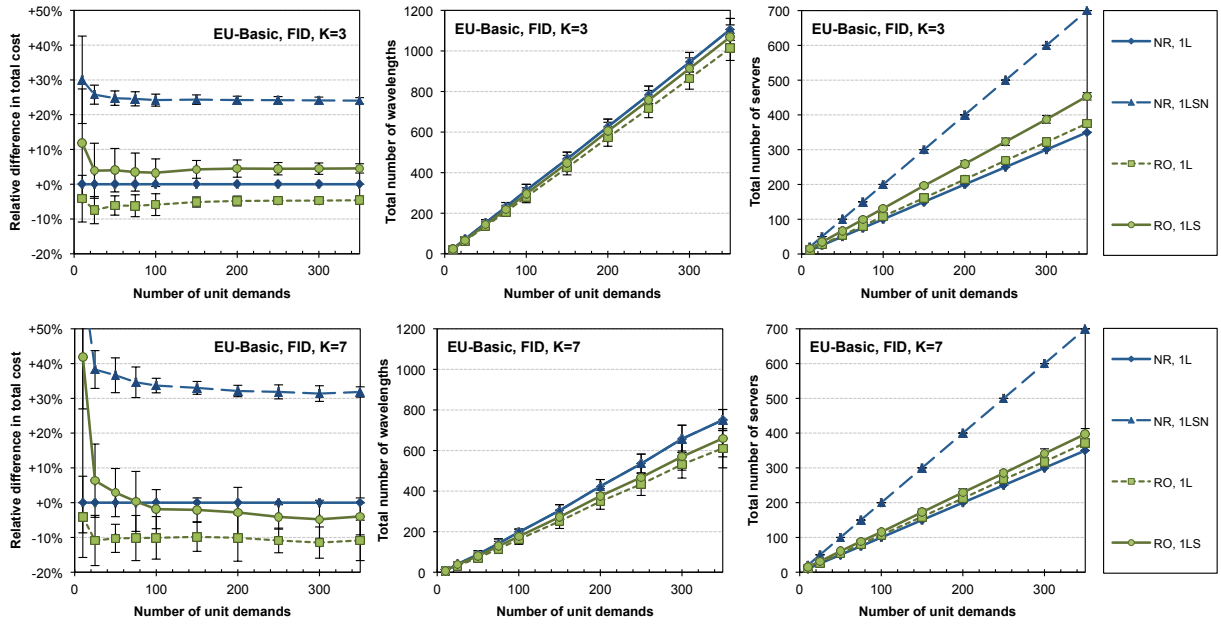
We first address the question which one out of proposed choosers is the best to determine the K locations for data centers. To decide what is "best", we consider the total cost in terms of joint network and server capacity, as expressed by the RMP objective (5). It can be expected that the difference will mainly pertain to the network capacity (number of wavelengths summed over all links), as the total number of servers (summed over the K data centers) required to meet the given demand is not expected to depend very much on the location of those servers.

Let's now focus on the cost reduction that the exploitation of relocation can bring. When protecting against link failures only (the 1L case), we observe a clear advantage of adopting relocation (RO): for $K = 3$ the total network capacity decreases with around 8.9% (averaged over the larger demand instances, $\Delta_v \in [100, 350]$). The price paid for this network advantage is a slight increase in total number of servers: extra capacity needs to be provided at the relocation locations. The net cost balance however is still advantageous for RO. When we offer protection also against server failures (1LS), the benefit of relocation (see RO-1LS vs NR-1LSN) is substantially more pronounced. Indeed, protection against server failures implies backup server capacity (increase with a factor $1 + 1/N$ for the assumed $1{:}N$ server protection; results are plotted for $N = 1$) in the NR case as well. When allowing relocation, i.e., in case of RO, we can however maximally share any backup capacity among all failure scenarios (amounting to a factor in the order of $1 + 1/K$ for K data center locations). When we increase the number of data center locations K, we learn that the relative benefit of relocation in terms of network capacity reduction becomes more pronounced (since paths to alternate destinations become on average shorter when increasing K), whereas the penalty of increased server capacity dissolves. For larger K, we see in Table 3 and Figure 4(a) that relocation even allows protection against both network and server failures (see RO-1LS) at a lower total cost than we can protect against network failures only without relocation (NR-1L). When comparing the results for different topologies (results omitted to save space) we find that, as intuitively expected, relocation is especially beneficial in sparser topologies: in a sparse network, a backup path towards an alternate destination has more chance to be substantially shorter than one towards the original that is disjoint from the working path (e.g., think of a ring network).

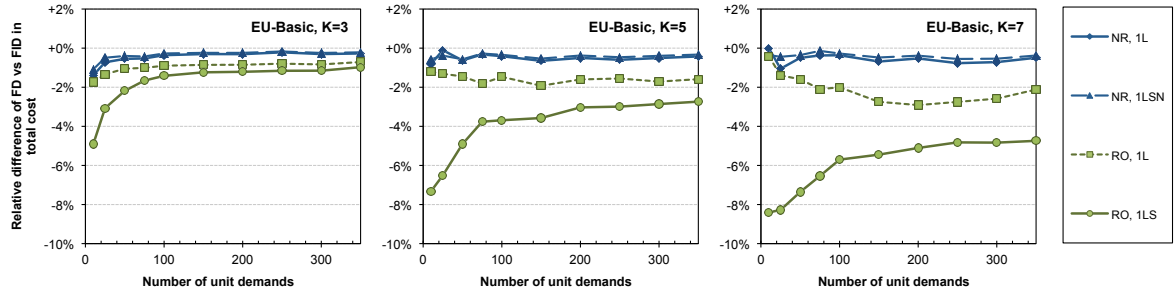
Table 3. The value of relocation (RO) is that it allows substantial reduction of the required network capacity (measured as wavelengths summed over all links), at the cost of extra server capacity. We list the relative values compared to the baseline of using no relocation (NR) to protect against single link failures (1L) only. Reported values are averages over the demand cases $\Delta_v \in [100, 350]$*, for EU-Basic topology, failure-independent (FID) rerouting and a server cost factor* $\alpha = 1$.

	K	Total wavelengths	Total servers	Total cost
1L, RO	3	–8.9%	+7.5%	–5.0%
	5	–14.3%	+6.1%	–8.6%
	7	–18.3%	+6.4%	–10.5%
1LS, RO	3	–3.9%	+29.9%	+4.3%
	5	–8.9%	+20.5%	–0.5%
	7	–11.8%	+14.3%	–3.2%

The last question concerned the difference between failure-dependent (FD) rerouting and failure-independent (FID) rerouting. Obviously, the qualitative advantage of relocation is expected to continue to hold when applying FD rather than FID. Our results (not included here in detail to save space) confirm that expectation. Here we particularly focus on quantifying the expected advantage of applying FD over FID. Comparison of FID vs FD in other contexts, i.e., in unicast routing problems (thus without an option to resort to relocation), concluded that the potential advantage of adopting FD rerouting that can be tailored to each specific failure instance is very limited in terms of total network capacity (Xiong & Mason, 2002; Zang et al., 2003).
Figure 4(b) plots our comparison of FD vs FID in terms of total cost. We note that for the case without relocation (NR), the advantage of FD indeed seems limited. However, when we do exploit relocation (RO), and then especially to protect against both link and server failures (1LS), the relative advantage of FD over FID seems significant (and more so for larger number of data center locations K). For instance, for $K = 7$ we note a cost reduction of applying FD compared to FID that amounts to around 6%. This could very well outweigh the higher operational complexity that FD incurs: FD needs to maintain more state (e.g., multiple pre-computed routes to be signaled and stored as routing state) and conditional rerouting is required, which implies that the exact failure affecting the working path needs to be properly identified (vs unconditional rerouting to the single backup path for FID).



(a)



(b)

Figure 4. (a) The advantage of relocation increases for larger number of data center locations: the network savings increase, while the penalty of higher server capacity diminishes. The net cost of protecting against both link and server failures with relocation (RO, 1LS) even is lower than protecting against single link failures only without relocating (NR, 1L).
(b) The advantage of adopting failure-dependent (FD) rerouting, compared to using the same backup configuration regardless of failure scenario (FID), is limited when we do not exploit relocation (NR). However, when exploiting relocation (RO), the total cost reduction (in terms of combined network and server capacity) is more pronounced. The advantage of FD increases for higher number of chosen data center locations (increasing K).

DISASTER-RESILIENT OPTICAL CLOUDS

Above, we dealt with the generic problem of resilient cloud network design. Now we discuss protection of optical clouds specifically against large-scale disaster failures.
Whereas traditional studies were more focused on small-scale (e.g., single link/node) failures, protection against disaster failures has become a major concern, given their risk of affecting communication networks (Neumayer, Zussman, Cohen, & Modiano, 2009; Reuters, 2005). Disasters can have natural (earthquakes, tornados, tsunamis, hurricanes, etc.) or human causes (e.g., weapons of mass destruction (WMD) and electro-magnetic pulse (EMP)) (Neumayer et al., 2009). In this chapter, the main focus of our analysis is optical backbone networks. These networks act as a substrate for upper layer networks (e.g., SONET/SDH, Ethernet, IP/MPLS) providing end-to-end connectivity and network services (such as cloud computing). Due that dependency on optical networks, a disaster that severely disrupts optical backbone networks (implying huge data loss and

disruption of high-bandwidth optical channels) thus might cause disruption of essential services for weeks and further complicates rescue operation. A few examples: the 2008 China Sichuan earthquake caused damage to around 30,000 kilometers of fiber optic cables and 4,000 telecom offices (Ran, 2011); the 2005 Hurricane Katrina and the flood and power outage following the hurricane caused reduction of communication network availability from approximately 99.99% to 85% (Kwasinski, Weaver, Chapman, & Krein, 2009); the 2006 Taiwan earthquake caused a fiber-cut and reduced Hong-Kong and China's Internet access capacity by 100% and 74% respectively (Sterbenz, Cetinkaya, Hameed, Jabbar, & Rohrer, 2011). Although upper layer schemes (e.g, TCP-layer retransmission, IP-layer re-routing) can deal to some extent with network failures, they are not capable of dealing with disaster failures as efficiently as physical (e.g., optical) layer schemes: firstly, since disaster failures are geographically collocated, physical topology information (more easily available at the physical layer) is required to efficiently combat disaster failures, and end-to-end mechanisms such as those applied in upper network layers might be ineffective or extremely expensive; secondly, a single failure in the optical layer might cause disruption of thousands of upper layer connections and services, thus requiring more resources and time to recover in the upper layer.

Optical networks, as previously pointed out, are ideally suited to meet the rising traffic demand from clouds. Cloud service providers (e.g., Google, Comcast) have built geo-distributed networks of data centers to provide lower latency and reliability through redundancy in case of a failure, and have deployed optical backbone networks connecting data centers and the customer networks. Here, we consider an optical cloud comprising an optical WDM backbone network providing circuit-switched paths for high-bandwidth connections between users and data centers. Cloud services yield new opportunities to provide protection against disasters exploiting the anycast principle. Note that, in addition to path protection, data center networks also require protection of content, i.e., failure of data centers should not cause the disappearance of a specific content/service from the whole network. Below, we present a model to design data center networks while providing disaster survivability to both paths and content. This model allows effective analysis of the effect of different parameters (e.g., protection schemes, number and locations of data centers, number of replicas per content item) on disaster survivability as well as resource usage in a data center network.

DESIGN OF DISASTER-RESILIENT OPTICAL DATA CENTER NETWORKS

Most research on protection in optical networks focused on single-link failures (Ou & Mukherjee, 2005). Recently, a few studies on multiple failure scenarios have been conducted (Johnston, Lee, & Modiano, 2011; Lee, Lee, & Modiano, 2011; Sen, Murthy, & Banerjee, 2009). Traditional backup path based protection schemes cannot provide protection against destination (data center) node failures. As explained before, introducing backup data centers following the anycast principle can reduce bandwidth consumption (Buysse et al., 2009). This scheme does not protect against disasters affecting both primary and backup resources (network links and/or data centers). Also, content/service protection is a fundamental problem in data center networks. Moreover, location of contents directly affects routing performance. Thus, three problems, namely (i) content/service placement, (ii) routing, and (iii) protection of paths and content/service should be addressed simultaneously. In this case study, we use shared risk groups to define potential disaster zones. A Shared Risk Group (SRG) or Disaster Zone (DZ) is defined as a set of nodes and links that might be affected simultaneously by a single disaster event. In this case study, we consider placement of a content in DZ-disjoint data centers and providing a pair of DZ-disjoint paths (primary and backup) for a mission-critical connection requiring high bandwidth and/or low latency; the proposed solutions can easily be extended for connections with arbitrary bandwidth requirements by aggregating multiple connections using grooming techniques (Zhu & Mukherjee, 2002). (Note that it may be unnecessary to provide disaster protection for all services: our study applies to high-priority content and high-bandwidth requests that do require such protection.)

Solution approach and problem statement

We first propose an integrated integer linear program (ILP) to design data center networks and provide disaster survivability (i.e., protection against a single disaster event). Our formulation solves

the following problems simultaneously: content placement (i.e., replication), as well as routing and disaster protection for both paths and content. We consider a circuit-switched optical data center mesh network and assume that a single wavelength path is required for each request (extension to arbitrary capacity requirements per request is straightforward). To simplify the model, we assume that data centers have no constraints on storage and computing capacity. Since ILPs are computationally intractable and thus do not scale well, our design strategy is based on relaxations of the ILP and heuristics to solve for large problem instances (presented in our previous work (Habib, Tornatore, De Leenheer, Dikbiyik, & Mukherjee, 2012)). We formally state the problem as follows:

Given

- $G(V, E)$: Physical topology comprising the physical node set V and the set of directed links E
- V': Set of data center locations, $V' \subset V$
- Z: Set of disaster zones (DZs)
- C: Set of contents
- $R = \{(s, c)\}$: R is the set of anycast requests, $s \in V, c \in C$
- K: Maximum number of replicas per content
- B: Capacity of a directed link (every link is assumed to have the same capacity)

Objective

- Minimize total wavelength usage

Output

- Content placement
- Disaster-zone-disjoint primary and backup paths for each request

Integrated ILP model for disaster-resilient content placement and routing

We formulate the problem of assigning paths to high bandwidth connections, determining content replica placement, and providing shared protection against a single disaster failure (i.e., multiple failures caused by a single disaster) for both paths and contents, using an ILP as shown below.

Variables

- $p_{scij}^{W} \in \{0,1\}$: Primary (working) path of request (s, c) (hence superscript W) goes through link (i, j); $s \in V, c \in C$
- $p_{scij}^{B} \in \{0,1\}$: Backup path of request (s, c) goes through link (i, j) ; $s \in V, c \in C$
- $\tau_{ij} \in \{0,1,2, \dots\}$: Number of shared wavelengths used in link (i, j) to support backup paths
- $r_d^c \in \{0,1\}$: data center $d \in V'$ hosts a replica of content item $c \in C'$
- $f_{scd}^{W} \in \{0,1\}$: $d \in V'$ serves as the primary data center for request (s, c), i.e., primary path for request (s, c) is routed to d
- $f_{scd}^{B} \in \{0,1\}$: $d \in V'$ serves as the backup data center for request (s, c), i.e., backup path for request (s, c) is routed to d
- $b_{ijx}^{sc} \in \{0,1\}$: Backup path of request (s, c) through link (i, j) is used when the primary path is down due to a disaster at $x \in Z$
- $\alpha_x^{sc} \in \{0,1\}$: Primary path of request (s, c) is down due to a disaster at $x \in Z$
- $\beta_x^{sc} \in \{0,1\}$: Backup path of request (s, c) is down due to a disaster at $x \in Z$

Using the variables, the objective becomes:

$$min\left(\sum_{i,j}\sum_{s,c} p_{scij}^{W} + \sum_{i,j} \tau_{ij}\right)$$

Here, the first term minimizes resources used to provision primary paths and the second term minimizes resources used to provision shared backup paths.

Flow-Conservation Constraints

$$\sum_{j:(i,j)\in E} p^W_{scij} - \sum_{j:(j,i)\in E} p^W_{scji} = \begin{cases} 1 & \text{if i == s} \\ -f^W_{sci} & \text{if } i \in V' \\ 0 & \text{otherwise} \end{cases} \quad \forall (s,c) \in R, \forall i \in V \tag{18}$$

$$\sum_{j:(i,j)\in E} p^B_{scij} - \sum_{j:(j,i)\in E} p^B_{scji} = \begin{cases} 1 & \text{if i == s} \\ -f^B_{sci} & \text{if } i \in V' \\ 0 & \text{otherwise} \end{cases} \quad \forall (s,c) \in R, \forall i \in V \tag{19}$$

$$\sum_{d\in V'} f^W_{scd} = 1 \qquad \forall (s,c) \in R \tag{20}$$

$$\sum_{d\in V'} f^B_{scd} = 1 \qquad \forall (s,c) \in R \tag{21}$$

Equations (18) and (19) enforce flow conservation on primary and backup paths, respectively. Following the anycast principle, the primary and backup data centers are not fixed, as represented by variables f^W_{sci} and f^B_{sci} in eq. (18) resp. (19). Equations (20) and (21) constrain the number of primary and backup data centers to one.

Data center assignment and content placement

$$r^c_d \geq f^W_{scd} + f^B_{scd} \qquad \forall c \in C, \forall d \in V', \forall (s,c) \in R \tag{22}$$

$$\textstyle\sum_{d\in V'} r^c_d \leq K \qquad \forall c \in C \tag{23}$$

Equation (22) ensures that (i) a data center d is not used to serve a request (*s,c*) if it does not host content c and (ii) primary and backup data centers of a request are different. Equation (23) constrains the number of replicas per content.

Capacity Constraint

$$\textstyle\sum_{(s,c)} p^W_{scij} + \tau_{ij} \leq B \qquad \forall (i,j) \in E \tag{24}$$

Equation (24) constrains link capacity. The computation of τ_{ij} will be explained later.

Disaster-zone-disjoint path constraint

$$\frac{\sum_{(i,j)\in z} p^W_{scij}}{M} \leq \alpha^{sc}_z \leq \sum_{(i,j)\in z} p^W_{scij} \qquad \forall (s,c) \in R, \forall z \in Z \tag{25}$$

$$\frac{\sum_{(i,j)\in z} p^B_{scij}}{M} \leq \beta^{sc}_z \leq \sum_{(i,j)\in z} p^B_{scij} \qquad \forall (s,c) \in R, \forall z \in Z \tag{26}$$

$$\alpha^{sc}_z + \beta^{sc}_z \leq 1 \qquad \forall (s,c) \in R, \forall z \in Z \tag{27}$$

Equations (8) and (9) set the value of α^{sc}_z and β^{sc}_z, respectively. Here, *M* is a large integer constant (i.e., greater than the maximum possible numerator ($\sum_{(i,j)\in z} p^W_{scij}$ and $\sum_{(i,j)\in z} b^W_{scij}$)). By definition, α^{sc}_z (β^{sc}_z) is 1 if at least one link on the primary (backup) path for request (s,c) goes through DZ z. Equation (10) ensures that primary and backup paths are DZ-disjoint.

Shared Protection Constraint

$$\tau_{ij} \geq \textstyle\sum_{(s,c)} b^{sc}_{ijz} \qquad \forall (i,\mathrm{j}) \in E, \forall z \in Z \tag{28}$$

$$b^{sc}_{ijz} \leq \alpha^{sc}_z \qquad \forall (s,c) \in R, \forall (i,j) \in E, \forall z \in Z \tag{29}$$

$$b^{sc}_{ijz} \leq p^B_{scij} \qquad \forall (s,c) \in R, \forall (i,j) \in E, \forall z \in Z \tag{30}$$

$$b^{sc}_{ijz} \geq \alpha^{sc}_z + p^B_{scij} - 1 \qquad \forall (s,c) \in R, \forall (i,j) \in E, \forall z \in Z \tag{31}$$

Equations (28)-(31) bound the number of shared wavelengths used in a link for protection. Combining the objective and eq. (28), we get $\tau_{ij} = \max_z \sum_{(s,c)} b^{sc}_{ijz}$. The example shown in Figure 5 explains the computation of τ_{ij}. Here, we have three requests with primary paths: 1-2-3-4, 8-3-4, and 8-7-6; and

corresponding backup paths: 1-9-5, 8-9-5, and 8-9-5. The backup data centers are different from the primary ones. Circles A and B represent two DZs. Link (9, 5) is shared by all three backup paths. Failure of DZ A affects primary paths 1-2-3-4 and 8-3-4; and failure of DZ B affects primary path 8-7-6. Thus, $\sum_{(s,c)} b_{95A}^{sc} = 2$ and $\sum_{(s,c)} b_{95B}^{sc} = 1$. We get, $\tau_{95} = max_{z\in\{A,B\}} \sum_{(s,c)} b_{ijz}^{sc} = 2$. So, only two shared wavelengths on link 9-5 are needed to provide protection for all three requests.

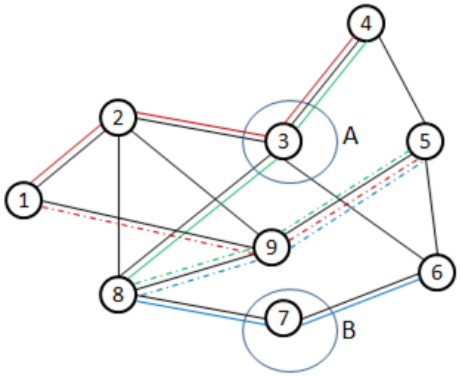


Figure 5. Computation of τ_{ij}, which denotes the number of shared wavelengths used in link (i, j) to provide backup paths for connections disrupted by a disaster failure.

This ILP does not explicitly include constraints to provide protection for content. Assuming that each content is requested by at least one request, and also noting that primary and backup paths for a request are DZ-disjoint and connect to two different data centers, it is ensured that content is replicated at disaster-disjoint data centers.

Relaxations of the integrated ILP

To solve the problem for large networks, we propose two relaxations of the integrated ILP:

- **Two-Step ILP**: We introduce separate but interlaced ILP formulations for content placement and routing.
- **LP Relaxations**: The relaxed two-step ILP, obtained by relaxing the integrality constraint of the variables, gives us a lower bound on the optimal. Since the LP-relaxed solution may not be a feasible one, we propose heuristics to find a feasible solution from the LP-relaxed solution

Two-Step ILP

Here, we consider content placement with content protection and routing with path protection as separate processes, and propose two separate ILPs for these two problems. This ILP allows us to handle larger networks with more connections as it has lower complexity than the integrated one.

Step 1: Content placement with content protection

This ILP uses a heuristic objective that minimizes the average distance from the requesting nodes to the data center nodes for all the requests. As a result, this objective tends to place content in data centers closer to its popular region, which reduces resource usage by primary and backup paths while routing connection requests. We pre-compute h_{sd}, which is the minimum number of hops to reach node *d* from node *s*.

Objective $\min (\sum_{d\in V'} \sum_{sc\in R} h_{sd} r_d^c)$

Constraints

$$\sum_{d\in V'} r_d^c \leq K \qquad \forall c \in C \tag{32}$$

$$\sum_{d\in z} r_d^c \leq 1 \qquad \forall c \in C, \forall z \in Z \tag{33}$$

Equation (33) bounds the number of replicas to *k*. Equation (32) ensures that no DZ can have more than one replica of a content item.

Step 2: Routing with path protection
We derive the ILP for routing by eliminating eq. (23) from the integrated formulation and using constant values for R_a^cs as found from Step 1.

LP Relaxation
One way to make an ILP more scalable is LP relaxation: remove the integrality constraints on variables and then solve the corresponding Linear Program (LP). The search space of the LP includes the search space of the ILP, so LP relaxation provides a lower bound on the optimal solution. As some constraints are relaxed in the LP formulation, we may obtain an infeasible solution, particularly since this solution may give fractional values for the variables (some of which should be integers in reality). Heuristics can be used to obtain a feasible solution for the original ILP problem from the infeasible relaxed solution.

As there is no constraint on the storage capacity of data centers in the ILP to solve the content placement problem (i.e., Step 1 of two-step ILP), the placement of a content item does not depend on the placement of other contents. Thus, each item can be replicated separately. This ILP uses a heuristic objective which tends to place an item in data centers closer to its popular region. Thus, instead of LP relaxation, we propose Algorithm 1, which solves the content placement problem (one item at a time) to achieve the same objective in polynomial time.

Basically, for a specific content item and a specific data center location, we consider the requests for the item and compute the average shortest distance of that data center from the requesting nodes. We sort the data centers in non-decreasing order of the computed distances and replicate content into data centers in succession following the order of the sorted list and maintaining constraint (38).

We apply LP relaxation on the routing problem (i.e., Step 2 of the two-step ILP). We relax all the integer variables. Note that, as the relaxed solution may give fractional values for variables, we may have multiple primary paths and multiple backup paths for each connection request.

We propose Algorithm 2 to derive a feasible solution for a request (s, c) from the infeasible relaxed one. Here, *SP* is the pre-computed set of *k*-shortest paths from node s to each of the data center nodes. Using any polynomial time search algorithm, we can find the set of primary paths, P, and backup paths, B, from the relaxed solution. Here, $\alpha_i^p(\beta_i^b)$ is equal to 1 if primary path p (backup path b) goes through DZ i. Then, we compute $max_i(\alpha_i^p \times \beta_i^b)$ for each possible pair (p, b), $p \in P, b \in B$. If $max_i(\alpha_i^p \times \beta_i^b) = 0$ for a pair (p, b), then p and b are DZ-disjoint paths. We compute set S which holds those pairs (p, b) having $max_i(\alpha_i^p \times \beta_i^b) = 0$. If S is not empty, we take the pair that consumes the least amount of resources (sum of primary and backup wavelengths to provide shared protection). To compute maximum sharing for a path, we use the procedure in Chapter 11 of (Mukherjee, 2006). If S is empty, we do not have any DZ-disjoint pair of paths from the sets P and B. We then take the least-cost primary path from P, and compute *candidateB* as the set of paths $b' \in SP$ such that b' is DZ-disjoint to p. If *candidateB* is not empty, we take a path b from it such that (p, b) pair consumes

Algorithm 1: Content Placement
Input:
T: Set of user requests (s, c) for content c
V′: Set of data center locations
h_{sd}: Minimum number of hops to reach node d from node s
Output:
L_c: Set of replica locations for content c

1. **for** each $d \in V'$ **do**
2. $cost_d = \sum_{(s,c)\in R} h_{sd}$
3. **end for**
4. Sort all data centers in non-decreasing order of $cost_d$
5. $i = 1$, n =number of data centers
6. **while** L_c does not have k members and $i \leq n$ **do**
7. Add i'th data center from the sorted list into L_c if it does not violate constraint (38)
8. $i = i + 1$
9. **end while**

Algorithm 3: Compute primary and backup paths for a request
Input:
(s, c): User request with s as requesting node and c as content
SP: Set of k shortest paths from node s to all data centers
PP: Set of already-provisioned primary paths
PB: Set of already-provisioned backup paths
Output:
bestPrimary: Primary path
bestBackup: Backup path
minCost: Cost of the pair (*bestPrimary*, *bestBackup*)

1. $minCost = \infty$
2. $bestPrimary = NULL$
3. $bestBackup = NULL$
4. **for** each path $p_1 \in SP$ **do**
5. **if** all links on p_1 have enough capacity, **then**
6. **for** each path $p_2 \in SP$ **do**
7. **if** all links on p_2 have enough capacity and p_1 and p_2 are DZ-disjoint **then**
8. **if** *minCost* > cost($p_1 + p_2$) **then**
9. *bestPrimary*=p_1, *bestBackup*=p_2.
10. *minCost* = cost($p_1 + p_2$)
11. **end if**
12. **end if**
13. **end for**
14. **end if**
15. **end for**

Algorithm 2: Routing: Compute primary and backup paths for the given request from relaxed LP solution
Input:
(s, c): User request with s as requesting node and c as content
P: Set of primary paths found for (s, c) from relaxed ILP
B: Set of backup paths found for (s, c) from relaxed ILP
SP: Set of k shortest paths from node s to all data centers
α_i^p: Primary path $p \in P$ for (s, c) passes DZ i
β_i^b: Backup path $b \in B$ for (s, c) passes DZ i
PP: Set of already-provisioned primary paths
PB: Set of already-provisioned backup paths
Output:
(*bestP, bestB*): primary path *bestP* and backup path *bestB*

1. Compute $S = \{(p,b): max_i(\alpha_i^p \times \beta_i^b) = 0; p \in P, b \in B\}$
2. **if** S is not empty and for at least one pair in S, all links on the two paths have enough capacity, **then**
3. Take lowest cost pair $(p', b') \in S$ such that all links on the two paths have enough capacity. Set $bestP = p'$ and $bestB = b'$.
4. **else**
5. **if** P is empty then
6. GOTO Step 20.
7. **end if**
8. Take lowest-cost primary path $minP \in P$ such that all links on $minP$ have enough capacity. Set $bestP = minP$.
9. Compute $candidateB = \{b': b' \in SP, b'$is disaster-zone-disjoint to $bestP$.
10. **if** *candidateB* is not empty and for at least one path of *candidateB*, all links on that path have enough capacity, **then**
11. Take least-cost pair $(bestP, b')$ where $b' \in candidateB$ and links on b' have enough capacity. Set best$B = b'$.
12. **else**
13. From the topology graph find (if possible) a shortest path b', disaster-zone-disjoint to *bestP*, from node *S* to any of the data center nodes that has content c. Set $bestB = b'$.
14. **if** *bestB* not found **then**
15. Delete minP from P. GOTO Step 5.
16. **end if**
17. **end if**
18. **end if**
19. **if** $(bestP, bestB)$ not found **then**
20. Use Algorithm 3 to compute *bestPrimary* and *bestBackup*. Set *bestP=bestPrimary* and *bestB=bestBackup*
21. **end if**

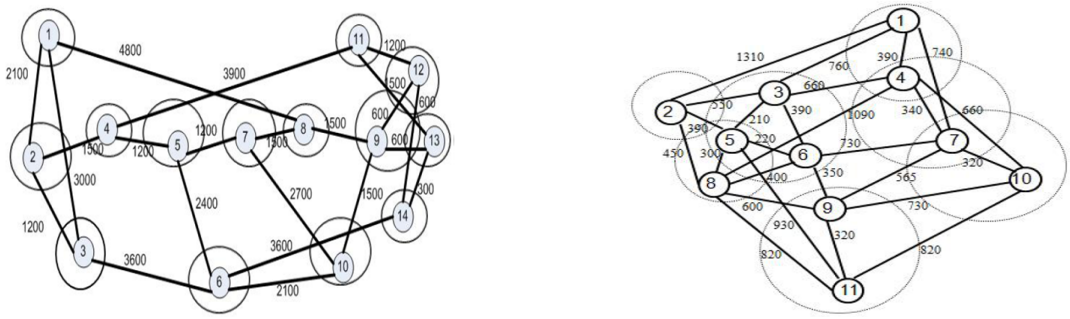
lowest cost. If *candidateB* is empty, we delete the links on p from the topology, and find a shortest path from the modified topology as backup path. If a backup path is not found, we delete the primary path from P, take the least-cost primary path from P, compute *candidateB* for the new primary path, and repeat the steps. If none of these works, we use Algorithm 3 to compute the paths, which we explain below.

Heuristic

Here, we explore non-mathematical heuristic approaches to solve the problem for large problem instances. We consider a static-traffic case, where all requests are known beforehand, yet, these heuristics can also be applied for dynamic traffic, where requests arrive and are processed one-by-one. For content placement with disaster protection, we propose Algorithm 1, as previously discussed. Algorithm 3 shows the heuristic to compute DZ-disjoint primary and backup paths for a given request (s, c). The k shortest paths from node s to all data center nodes are pre-computed. Thus, if we have r replica locations for content c, we have $k \times r$ paths to be considered for request (s, c). The heuristic considers all possible pairs of paths and selects the lowest-cost disaster-disjoint pair as solution. Cost of a pair of paths is the sum of the number of wavelengths used for the primary path and additional wavelengths used for shared backup path. For a comparison of the running times of the proposed methods, we refer to (Habib, Tornatore, De Leenheer, Dikbiyik, & Mukherjee, 2012).

Illustrative numerical examples

We present illustrative results by solving the ILP formulations and heuristics on NSFNet and COST239 networks shown in Figures 6(a) and 6(b). Existing studies show that a disaster zone (DZ) can span up to 160 km (Weems, 2003). Following this, we randomly specify 14 DZs for NSFNet and 7 DZs for COST239. The COST239 network is denser (i.e., shorter link lengths and higher connectivity) than NSFNet. The maximum number of wavelengths per link is 32. We compare three protection schemes: dedicated single-link failure (SLF) protection (i.e., dedicated path protection against a single link failure), shared single-link failure (SLF) protection (i.e., shared path protection against a single link failure), and the proposed shared disaster zone failure (DZF) protection (i.e., shared path protection against a single disaster-zone failure). In all three schemes, the primary and backup data centers for a request are always different. The formulations for dedicated and shared SLF protection can be easily derived from our disaster protection model with minor changes.



(a) NSF Net (Average link length = 1936 km, Average node degree = 3.14)

(b) Cost239 (Average link length = 578 km, Average node degree = 4.73)

Figure 6. Topologies used in the study (DZs in circles, link lengths in km).

Results from the integrated ILP model

We present the impacts of protection scheme, number of data centers, and number of content replicas on network resource (i.e., wavelength) usage. The traffic matrix used in the simulations is generated randomly and uniformly distributed among network nodes.

1) Protection scheme: We first compare the wavelength usage of the three schemes. For both NSFNet and COST239, we have 3 data centers, 10 content items, and unconstrained number of replicas per item. For NSFNet, the data centers are located at nodes 2, 6, and 9; and for COST239, the data centers are located at nodes 4, 5, and 9. Data center locations are chosen in a way such that at

least one of the data centers is at most two hops away from a node in the network. Figure 7(a-b) compares the wavelength usage for shared DZF protection with dedicated and shared SLF protection.

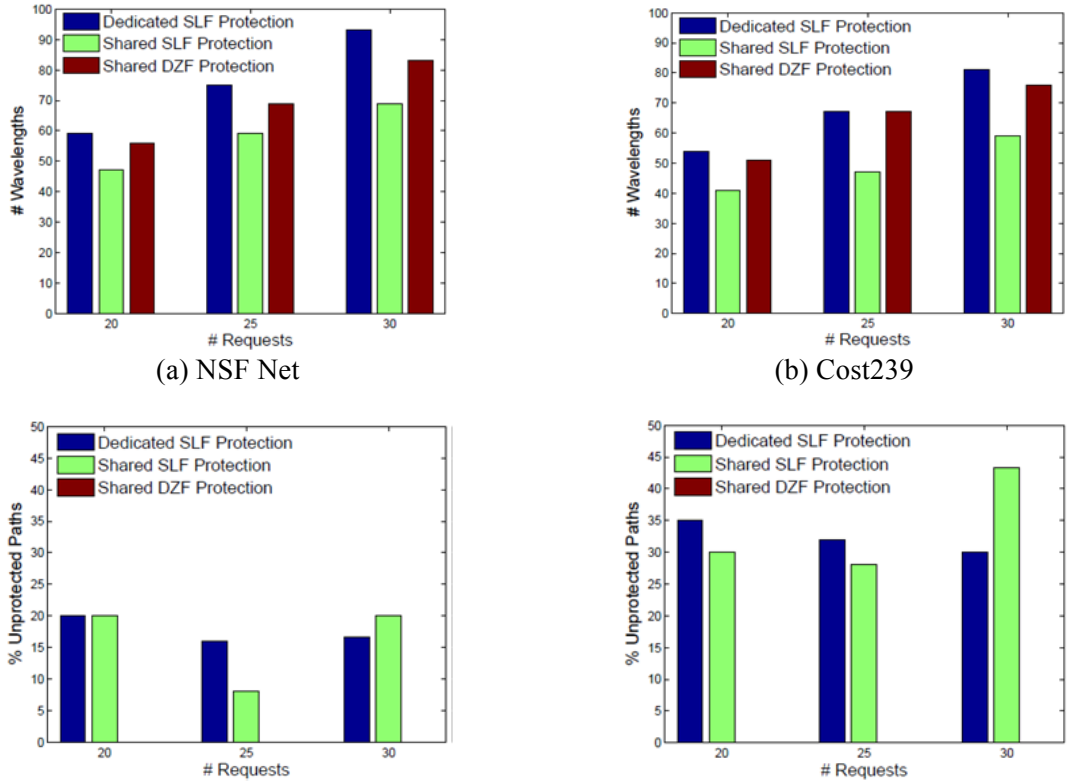


Figure 7. Total wavelength usage (a-b) and number of unprotected paths (c-d) for three schemes using integrated ILP.

We see that shared DZF protection uses more wavelengths than shared SLF protection but fewer than dedicated SLF protection. Dedicated SLF protection has more probability of being survivable in case of multiple random link failures than shared SLF protection. But, in reality, failures of multiple non-correlated links are quite unlikely. Rather, it is more likely that a set of correlated links/nodes are down simultaneously due to a disaster. Though disaster protection provides more protection, it is not a popular choice as it consumes significantly higher resources than protection against a single link failure. But we find that disaster protection exploiting anycast in a data center network consumes moderate resources while providing the required protection. As the average hop distance between two nodes is less in COST239 than in NSFNet, we find that the wavelength usage is less in COST239 than in NSFNet.

Figure 7(c-d) shows that, without DZF protection, a significant number of connections are vulnerable to disaster failures, even though data centers are distantly located in the network, and primary and backup data centers are different in all the three cases. With the same primary and backup data center, the connections are never protected against destination (data center) node failure. These results indicate that shared DZF protection, although it uses fewer resources, provides more protection against disasters than dedicated SLF protection. Figure 8(c-d) shows that the number of paths vulnerable to disasters is higher in COST239 than in NSFNet because COST239 is denser with shorter links than NSFNet. The denser a network is, the more vulnerable it is to a disaster failure.

2) Number of replicas: Table 4 shows the effect of number of content replicas on wavelength usage using shared DZF protection in NSFNet. To experiment with more replicas, we use four data centers at nodes 2, 5, 9, and 11. We compare the wavelength usage between 4 replicas per content item (i.e., every item is replicated in every data center) and 2 replicas per item. Note that with a small increase

in wavelength usage, the number of replicas can be decreased significantly. Based on user demands, replicas are distributed throughout the network, which allows flexibility to choose primary and backup data centers. More replicas do not always provide more flexibility to choose a shorter path; rather more replicas mean more usage of storage resources and more usage of bandwidth to perform replication and synchronization (i.e., consistency among replicas of a content).

	# Requests		
	20	25	30
4 replicas per content	49	64	75
2 replicas per content	55	73	83

Table 4. Wavelength usage in shared DZF protection for different numbers of replicas in NSFNet using Integrated ILP.

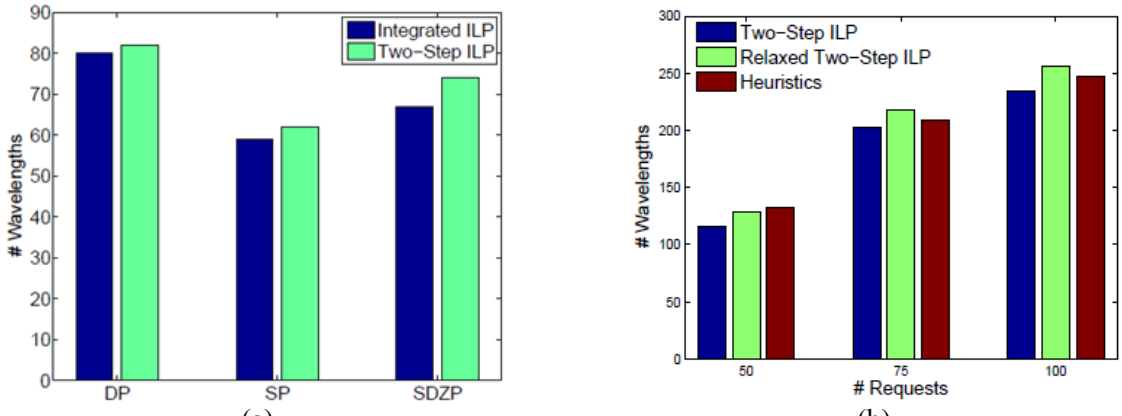
# Data centers	2	3	4	5
# Wavelengths	62	45	37	33

Table 5. Wavelength usage in shared DZF protection for different numbers of replicas in NSFNet using Integrated ILP.

3) Number of data centers: Table 5 shows the effect of the number of data centers on wavelength usage in shared protection with unconstrained number of replicas in NSFNet. For two data centers, we use locations 5 and 9; for three data centers, we use 2, 5, and 9; for four data centers, we use 2, 5, 9, and 11; and for five data centers, we use locations 2, 5, 6, 9, and 11. In this result, the number of wavelengths reduces significantly as the number of data centers increases, but after a certain value, increasing the number of data centers does not help much to reduce wavelength utilization. We conclude that a reasonable number of data centers with intelligent network design can provide survivability to disasters while supporting user demands.

Performance of relaxed formulations and heuristics

To check the performance of the two-step ILP, we compare its wavelength usage with that of the integrated ILP on NSFNet for 30 connections, 5 data centers (2, 6, 7, 11, and 14), and 3 replicas per content. Figure 8(a) shows that the performance of the two-step ILP is quite close to that of the integrated ILP. Here, DP, SP, and SDZP are short forms for dedicated SLF protection, shared SLF protection, and shared DZF protection, respectively. For different protection schemes, the wavelength usage in the two-step ILP is 2.5% to 10.4% more than that in the integrated ILP. Note that the result of the two-step ILP is closer to what can be achieved in a real-world scenario where future requests are not known beforehand and replication is done separately from routing. We find that the two-step ILP is more scalable in the number of connections than the integrated ILP. We compared the performance of two-step ILP, relaxed two-step ILP, and heuristics as shown in Figure 8(b) and found that their performance is quite close to each other. For more details of these results, we refer to (Habib et al., 2012).



(a) (b)

Figure 8. (a) Comparison of two-step ILP with integrated ILP for NSF Net. (b) Comparison of two-step ILP, relaxed two-step ILP, and heuristics for NSF Net.

CONCLUSION AND OUTLOOK

Cloud services are today integral part of our society. Considering the increasing dependency of social and financial activities on clouds, disruption of cloud services due to failure of data center network resources can be catastrophic. In this chapter, we proposed schemes to dimension optical clouds resilient to failures. We considered protection of both network connections and content and showed how anycasting and relocation can be exploited to provide protection in an optical cloud. From the two studies we presented in detail, we conclude two lessons learned: (1) The classical optical network dimensioning models (cf. routing and wavelength assignment, RWA) need to be reworked to deal with the fact that the end point of traffic flows is flexible and not a priori known (cf. anycast), and (2) The network infrastructure should not be designed independently of the server infrastructure (i.e., data centers). The proposed schemes confirm the necessity of further research on optical cloud resiliency and provide a solid base to investigate and build up new protection schemes for both traffic and content in cloud networks. We mention here some relevant issues that need to be addressed while designing a protection scheme for optical clouds.

In this study, we have considered static traffic to design and analyze the characteristic of a data center network. A dynamic scenario is more complex. First, connections have a finite lifetime and thus available network capacity and topology may be changing frequently. Routing should not only consider the instantaneously available resources, but rather also the (near) future availability to achieve better optimization. Secondly, content popularity may vary over time, requiring to make new replicas or remove old replicas for it. Opportunities and challenges deriving from the dynamicity of combined content placement and connection routing must be investigated. Even though the relaxations and heuristics we discussed can be applied for the dynamic traffic with some modifications, more in-depth research needs to be conducted.

We modeled potential disaster zones as shared risk groups (SRGs). Generally this model can be applied when the location and the span of a disaster is known beforehand. For example, seismic hazard maps (Petersen & others, 2008) provide information regarding potential earthquake zones and seismic hazard levels, which can be used to define SRGs in a network graph. In reality, the probability of the occurrence of a disaster at a region varies with time. Due to advanced scientific methods, occurrence of some disasters can now be predicted beforehand (e.g., an incoming hurricane can be reasonably forecast at least one day in advance, leaving to the network operator an chance to re-organize its network to withstand the impending damages). Path provisioning and content placement algorithms should be made able to dynamically capture this information to increase disaster resilience. Post-failure network management can also enable higher resource utilization in a cloud network by reprovisioning the already provisioned requests. Multipath routing enables a connection to be multiplexed onto multiple paths whereas traffic grooming enables aggregation of multiple small flows into a larger flow that can be switched along a single path. Both of these techniques have been extensively studied in optical WDM networks. In a data center network with anycast services, multipath routing and traffic grooming can play a crucial role to effectively provide disaster survivability with limited additional resource usage (overprovisioning).

ABBREVIATIONS

ARWA	Anycast Routing and Wavelength Assignment
CG	Column Generation
DP	Dedicated SLF Protection
DZ	Disaster Zone
DZF	Disaster Zone Failure
EMP	Electro-Magnetic Pulse
FD	Failure Dependent
FID	Failure InDependent
HPC	High Performance Computing
IaaS	Infrastructure-as-a-Service
ILP	Integer Linear Program

IP	Internet Protocol
IT	Information Technology
LHC	Large Hadron Collider
LP	Linear Program
MPLS	Multi Protocol Label Switch
NR	No Relocation
OBS	Optical Burst Switching
OCS	Optical Circuit Switching
OXC	Optical Cross-Connect
PP	Pricing Problem
RMP	Restricted Master Problem
RO	RelOcation
RWA	Routing and Wavelength Assignment
SC	Shortest Cycle
SDZP	Shared DZF Protection
SLF	Single-Link Failure
SRG	Shared Risk Group
SRLG	Shared Risk Link Group
SP	Shared SLF Protection
SW	Shortest Working
TCP	Transmission Congestion Protocol
VM	Virtual Machine

REFERENCES

Alumur, S., & Kara, B. Y. (2008). Network hub location problems: The state of the art. *Eur. J. Oper. Res.*, *190*(1), 1–21. doi:10.1016/j.ejor.2007.06.008

Barla, I. B., Schupke, D. A., & Carle, G. (2012). Resilient virtual network design for end-to-end cloud services. In *Proc. 11th Int. Conf. Networking (Networking 2012)* (pp. 161–174). Prague, Czech Republic: Springer-Verlag. doi:10.1007/978-3-642-30045-5_13

Bathula, B. G., & Elmirghani, J. M. (2009). Constraint-Based Anycasting Over Optical Burst Switched Networks. *IEEE/OSA J. Opt. Commun. Netw.*, *1*(2), A35–A43. doi:10.1364/JOCN.1.000A35

Bhaskaran, K., Triay, J., & Vokkarane, V. M. (2011). Dynamic Anycast Routing and Wavelength Assignment in WDM Networks Using Ant Colony Optimization. In *Proc. IEEE Int. Conf. Commun. (ICC 2011)*. Kyoto, Japan.

Buysse, J., De Leenheer, M., Dhoedt, B., & Develder, C. (2009). Exploiting relocation to reduce network dimensions of resilient optical Grids. In *Proc. 7th Int. Workshop Design of Reliable Commun. Netw. (DRCN 2009)* (pp. 100–106). Washington, D.C., USA. doi:10.1109/DRCN.2009.5340020

De Leenheer, M., Farahmand, F., Lu, K., Zhang, T., Thysebaert, P., De Turck, F., … Jue, J. P. (2006). Anycast Algorithms Supporting Optical Burst Switched Grid Networks. In *Proc. 2\textsuperscriptnd Int. Conf. Networking and Services (ICNS 2006)*. Santa Clara, CA, USA. doi:10.1109/ICNS.2006.27

Demeyer, S., De Leenheer, M., Baert, J., Pickavet, M., & Demeester, P. (2008). Ant colony optimization for the routing of jobs in optical grid networks. *J. Opt. Netw.*, *7*(2), 160–172. doi:10.1364/JON.7.000160

Desrosiers, J., & Lübbecke, M. E. (2005). A Primer in Column Generation. In G. Desaulniers, J. Desrosiers, & M. M. Solomon (Eds.), *Column Generation* (pp. 1–32). Springer US.

Develder, C., Buysse, J., De Leenheer, M., Jaumard, B., & Dhoedt, B. (2012). Resilient network dimensioning for optical grid/clouds using relocation (Invited Paper). In *Proc. Workshop on New Trends in Optical Networks Survivability, at IEEE Int. Conf. on Commun. (ICC 2012)*. Ottawa, Ontario, Canada. doi:10.1109/ICC.2012.6364981

Develder, C., Buysse, J., Shaikh, A., Jaumard, B., De Leenheer, M., & Dhoedt, B. (2011). Survivable optical grid dimensioning: anycast routing with server and network failure protection. In *Proc. IEEE Int. Conf. Commun. (ICC 2011)*. Kyoto, Japan. doi:10.1109/icc.2011.5963385

Develder, C., De Leenheer, M., Dhoedt, B., Pickavet, M., Colle, D., De Turck, F., & Demeester, P. (2012). Optical networks for grid and cloud computing applications. *Proc. IEEE*, *100*(5), 1149–1167. doi:10.1109/JPROC.2011.2179629

Develder, C., Mukherjee, B., Dhoedt, B., & Demeester, P. (2009). On dimensioning optical Grids and the impact of scheduling. *Photonic Netw. Commun.*, *17*(3), 255–265. doi:10.1007/s11107-008-0160-z

Din, D.-R. (2005). Anycast Routing and Wavelength Assignment Problem on WDM Network. *IEICE Trans. Commun.*, *EE88-B*(10), 3941–3951.

Din, D.-R. (2007). A hybrid method for solving ARWA problem on WDM networks. *Comput. Commun.*, *30*(2), 385–395. doi:10.1016/j.comcom.2006.09.003

Glick, M., Krishanmoorthy, A., & Schow, C. (2011). Optics in the Data Center: Introduction to the Feature Issue. *IEEE/OSA J. Optical Commun. Netw.*, *3*(8), OD1. doi:10.1364/JOCN.3.000OD1

Habib, M. F., Tornatore, M., De Leenheer, M., Dikbiyik, F., & Mukherjee, B. (2012). Design of Disaster-Resilient Optical Datacenter Networks. *IEEE/OSA J. Lightwave Techn.*, *30*(16), 2563 – 2573. doi:10.1109/JLT.2012.2201696

Hyytiä, E. (2004). Heuristic Algorithms for the Generalized Routing and Wavelength Assignment Problem. In *Proc. 17th Nordic Teletraffic Seminar (NTS-17)* (pp. 373–386). Fornebu, Norway.

Jaumard, B., Meyer, C., & Thiongane, B. (2009). On column generation formulations for the RWA problem. *Discrete Applied Mathematics*, *157*(6), 1291–1308. doi:10.1016/j.dam.2008.08.033

Johnston, M., Lee, H., & Modiano, E. (2011). A Robust Optimization Approach to Backup Network Design with Random Failures. In *Proc. 30th IEEE Conf. Computer Commun. (INFOCOM 2011)*. Shanghai, China. doi:10.1109/INFCOM.2011.5934940

Kwasinski, A., Weaver, W. W., Chapman, P. L., & Krein, P. T. (2009). Telecommunications Power Plant Damage Assessment for Hurricane Katrina-Site Survey and Follow-Up Results. *IEEE Syst. J.*, *3*(3), 277–287. doi:10.1109/JSYST.2009.2026783

Lee, K., Lee, H., & Modiano, E. (2011). Reliability in Layered Networks with Random Link Failures. *IEEE/ACM Trans. Netw.*, *19*(6), 1835 – 1848. doi:10.1109/TNET.2011.2143425

Liu, X., Qiao, C., Wei, W., Yu, X., Wang, T., Hu, W., … Wu, M.-Y. (2009). Task Scheduling and Lightpath Establishment in Optical Grids. *IEEE J. Lightwave Technol.*, *27*(12), 1796–1805. doi:10.1109/JLT.2009.2020999

Liu, X., Qiao, C., Yu, D., & Jiang, T. (2010). Application-specific resource provisioning for wide-area distributed computing. *IEEE Netw.*, *24*(4), 25–34. doi:10.1109/MNET.2010.5510915

Liu, Y., Tipper, D., & Siripongwutikorn, P. (2005). Approximating optimal spare capacity allocation by successive survivable routing. *IEEE/ACM Trans. Netw.*, *13*(1), 198–211. doi:10.1109/TNET.2004.842220

Maesschalck, S. D., Colle, D., Lievens, I., Pickavet, M., Demeester, P., Mauz, C., … Derkacz, J. (2003). Pan-European Optical Transport Networks: An Availability-based Comparison. *Photonic Netw. Commun.*, *5*(3), 203–225. doi:10.1023/A:1023088418684

Mukherjee, B. (2006). *Optical WDM Networks*. Springer.

Neumayer, S., Zussman, G., Cohen, R., & Modiano, E. (2009). Assessing the Vulnerability of the Fiber Infrastructure to Disasters. In *Proc. 28th IEEE Conf. Computer Commun. (INFOCOM 2009)* (pp. 1566–1574). Rio de Janeiro, Brazil. doi:10.1109/INFCOM.2009.5062074

Ou, C., & Mukherjee, B. (2005). *Survivable Optical WDM Networks*. Springer.

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2, Part 2), 3336–3341. doi:10.1016/j.eswa.2008.01.039

Partridge, C., Mendez, T., & Milliken, W. (1993). *Host Anycasting Service* (RFC No. 1546). USA: IETF.

Petersen, M. D., & others. (2008). *United States National Seismic Hazard Maps* (No. Fact Sheet 2008-3017) (pp. 1–4). U.S. Geological Survey. Retrieved from http://pubs.usgs.gov/fs/2008/3017/

Ran, Y. (2011). Considerations and Suggestions on Improvement of Communication Network Disaster Countermeasures after the Wenchuan Earthquake. *IEEE Commun. Mag.*, *49*(1), 44–47.

Reuters. (2005). *Experts warn of substantial risk of WMD attack*. Retrieved from http://research.lifeboat.com/lugar.htm

Sen, A., Murthy, S., & Banerjee, S. (2009). Region-based connectivity - a new paradigm for design of fault-tolerant networks. In *Proc. Int. Conf. High Performance Switching and Routing (HPSR 2009)* (pp. 1–7). Paris, France. doi:10.1109/HPSR.2009.5307417

She, Q., Huang, X., Zhang, Q., Zhu, Y., & Jue, J. P. (2007). Survivable Traffic Grooming for Anycasting in WDM Mesh Networks. In *Proc. IEEE Global Telecommun. Conf. (Globecom 2007)* (pp. 2253–2257). Washington D.C., USA. doi:10.1109/GLOCOM.2007.430

Sterbenz, J. P. G., Cetinkaya, E. K., Hameed, M. A., Jabbar, A., & Rohrer, J. P. (2011). Modelling and analysis of network resilience. In *Proc. 3rd Int. Conf. Commun. Sys. and Netw. (COMSNETS 2011)* (pp. 1 –10). Bangalore, India. doi:10.1109/COMSNETS.2011.5716502

Stevens, T., De Leenheer, M., Develder, C., Dhoedt, B., Christodoulopoulos, K., Kokkinos, P., & Varvarigos, E. (2009). Multi-cost job routing and scheduling in Grid networks. *Futur. Gener. Comp. Syst.*, *25*(8), 912–925. doi:10.1016/j.future.2008.08.004

Stevens, Tim, De Leenheer, M., De Turck, F., Dhoedt, B., & Demeester, P. (2006). Distributed Job Scheduling based on Multiple Constraints Anycast Routing. In *Proc. 3rd Int. Conf. Broadband Commun., Netw. and Sys. (Broadnets 2006)* (pp. 1–8). San Jose, CA, USA. doi:10.1109/BROADNETS.2006.4374374

Stevens, Tim, De Leenheer, M., Develder, C., De Turck, F., Dhoedt, B., & Demeester, P. (2007). ASTAS: Architecture for scalable and transparent anycast services. *J. Commun. Netw.*, *9*(4), 1229–2370. doi:1854/9884

Suurballe, J. W., & Tarjan, R. E. (1984). A quick method for finding shortest pairs of disjoint paths. *Networks*, *14*(2), 325–336. doi:10.1002/net.3230140209

Tang, M., Jia, W., Wang, H., & Wang, J. (2003). Routing and Wavelength Assignment for Anycast in WDM Networks. In *Proc. 3rd Int. Conf. Wireless and Optical Commun. (WOC 2003)* (pp. 301—306). Banff, Canada.

Vanderbeck, F., & Wolsey, L. A. (1996). An exact algorithm for IP column generation. *Operations Research Letters*, *19*(4), 151–159. doi:10.1016/0167-6377(96)00033-8

Walkowiak, K. (2010). Anycasting in connection-oriented computer networks: Models, algorithms and results. *J. Appl. Math. Comput. Sci.*, *20*(1), 207–220. doi:10.2478/v10006-010-0015-5

Walkowiak, K., & Rak, J. (2011). Shared Backup Path Protection for Anycast and Unicast Flows Using the Node-Link Notation. In *Proc. IEEE Int. Conf. Commun. (ICC 2011)*. Kyoto, Japan. doi:10.1109/icc.2011.5962478

Weems, T. L. (2003). How Far is Far Enough. *Disaster Recovery J.*, *16*(2).

Xiong, Y., & Mason, L. (2002). Comparison of two path restoration schemes in self-healing networks. *Comput. Netw.*, *38*(5), 663–674. doi:10.1016/S1389-1286(01)00279-1

Xiong, Y., & Mason, L. G. (1999). Restoration strategies and spare capacity requirements in self-healing ATM networks. *IEEE/ACM Trans. Netw.*, *7*(1), 98–110. doi:10.1109/90.759330

Zang, H., Ou, C., & Mukherjee, B. (2003). Path-protection routing and wavelength assignment (RWA) in WDM mesh networks under duct-layer constraints. *IEEE/ACM Trans. Netw.*, *11*(2), 248–258. doi:10.1109/TNET.2003.810313

Zhu, K., & Mukherjee, B. (2002). Traffic grooming in an optical WDM mesh network. *IEEE J. Selected Areas Commun.*, *20*(1), 122 –133. doi:10.1109/49.974667

KEY TERMS & DEFINITIONS

Anycast routing: Anycast routing specifically enables users to transmit information (for processing, storage or service delivery), without a priori assigning an explicit destination. This offers the service provider the flexibility to select "the most appropriate" destination, and thus this freedom can be exploited to realize e.g., load balancing. Yet, this also introduces more complex routing decisions.

Cloud Computing: Distributed computing paradigm, building on key concepts of grid computing, but manifesting itselves in more commercially oriented applications that often involve loosely coupled tasks, and are typically interactive. A key concept that clouds heavily build on is that of virtualization: physical infrastructure (e.g., servers) are logically partitioned so that multiple users/applications can share the capacity, thus providing extra scalability.

Grid Computing: Distributed computing infrastructure, providing coordination of resources that are not subject to centralized control, using standard, open, general-purpose protocols and interfaces, and delivers non-trivial qualities of service (QoS).

Routing and wavelength allocation (RWA): In optical wavelength division multiplexing (WDM) networks, traffic flows over so-called lightpaths: on each physical link, bits are transported over a given wavelength. For a given amount of traffic, each demand from a source s to a destination d has to follow a given route (i.e., sequence of links from s to d) and on each of these links the wavelength to be used has to be chosen (and has to be the same if switches cannot convert one wavelength to another, thus enforcing a so-called wavelength continuity constraint). This

Shared risk (link) group (SRG, SRLG): An SRG captures the concept that different network resources may be jointly impacted by a certain failure if they are sharing a common risk: an SRG is a group of resources that thus fail simultaneously when the common risk occurs. An SRLG is a particular example where the resources are network links: e.g., if fibers are grouped in ducts, an accident causing the duct to break will impact all fibers crossing it.