Exploiting relocation to reduce network dimensions of resilient optical Grids.

Jens Buysse, Marc De Leenheer, Chris Develder, Bart Dhoedt

Department of Information Technology
Ghent University - IBBT, Ghent, Belgium
Email: jens.buysse@intec.ugent.be

***Abstract*—Optical Grids are widely deployed to solve complex problems we are facing today such as climate modeling and multimedia editing. An important aspect of the network supporting the Grid is resilience i.e. the ability to overcome network failures. In contrast to classical network protection schemes, we will not necessarily provide a back-up path between the source and the original destination. Instead, we will try to relocate the job to another server location if this means that we can provide a backup path which comprises less wavelengths than the one the traditional scheme would suggest. This relocation can be backed up by the Grid specific anycast principle: a user generally does not care where his job is executed and is only interested in its results. We present ILP formulations for both resilience schemes and we evaluate them by applying them on a European network topology. Results show us that this relocation strategy has a significant impact on the network load, with an acceptable penalty in extra resource load needed in order to execute the relocated jobs.**

I. Introduction

A. Optical Grids

Although a significant effort has been put into the development of the performance of computing systems, there are still a large number of problems which cannot be solved in an acceptable time frame by traditional hardware. These problems range from several research domains such as astrophysics[2], climate modeling[1] and fluid dynamics [3]. To address these complex problems, Grid computing has been devised. A Grid consists of different heterogeneous resources (computational, storage and networking) which are geographically spread within various control domains, implying that resource coordination is not subject to centralized control. The underlying network should be able to support high bandwidth traffic with low latency in a reliable way. The introduction of wave division multiplexing (WDM) in optical networks, has made the optical network an ideal candidate. In circuit-switched optical networks, bandwidth granularity is at the wavelength level since one or more wavelengths are allocated to a connection, while connectivity between source and destination is established using a two-way reservation. A Grid, supported by an optical network is called an Optical Grid. Typical for a Grid is the anycast principle. In general, multiple processing locations exist in a Grid network, so the exact location of execution (the destination), is of less importance to the end user. The main interest of the user lies in the successful execution of the job subject to predetermined requirements such as a deadline or some quality guarantee.

In this study we will consider a new way to protect a path from one point to another. There are two basic strategies to protect an optical network, namely *restoration* and *protection*. The former is a reactive procedure where the spare capacity is only used after the fault, the latter deals with techniques where the spare capacity is reserved in advance. This paper discusses two techniques which fall in the protection category. A traditional resiliency scheme is *shared protection* where each primary path has a backup path where wavelengths can be shared between several back-up paths, as long as their corresponding primary light paths are unlikely to fail simultaneously (e.g. primary paths have no links in common). Its counterpart, *dedicated protection*, does not allow this kind of sharing. In this paper, we have extended the shared protection algorithm by incorporating the anycast principle. Instead of reserving a back-up path to the original destination determined by the Grid scheduler, it could be better to relocate the job to another resource which may lead to a reduction in the number of backup wavelengths required. This could mean an overall network optimization with a relatively small computing penalty, due to the extra load for the server which receives the relocated jobs.

The remainder of this paper is structured as follows. First, in section II, we briefly discuss the possible failures which can occur in optical Grids. We continue in section III by introducing integer linear programming (ILP) formulations for the algorithms we have described. We present an evaluation of these models by a case study in section IV.

B. Related work

Several studies have been conducted to model different network planning strategies. In [6] the problem is investigated of fault management in a wavelength-division multiplexing (WDM)-based optical mesh network in which failures occur due to fiber cuts. Several off line-algorithms and heuristics are examined and their performance is compared through numerical examples. Our ILP formulation for shared protection is loosely based on the formulations described in this paper. In [5] the authors focus their attention on the computational efficiency of the ILP model in order to provide a more effective tool for planning. The formulation exploits flow aggregation and consists in a new ILP formulation that allows to reach optimal solutions with less computational effort compared

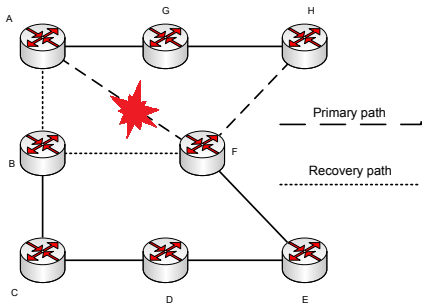
to other ILP approaches. Yet, the solution of the so-called source formulation ILP in [5] requires a post-processing step to find the actual routing and wavelength assignment (RWA). In this paper, we stick to traditional source-destination based flow formulation. We do however plan follow-up work to improve ILP scalability by also adopting source formulation (to increase the execution speed of the ILP solution programs and to make them more memory-efficient).

II. Failures in Optical Grids

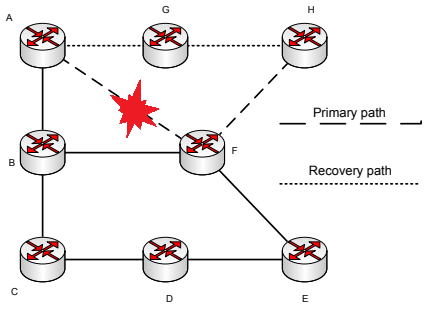
The causes of a network failure in optical networks can be divided into two categories, planned and unplanned failures. The former is caused by operational actions intentionally performed by the operator so they are known ahead and specific preventive techniques can be taken to overcome them. The latter is concerned with failures which are not preplanned and therefore general protection measures must be taken in advance to overcome these failures, e.g. fibreoptic cable cuts due to digging activities or natural disasters. It is almost impossible to provide measures against all possible failures in a communication network, therefore operators classify the most frequently occurring failures into a limited set of failure scenarios and provide protecting mechanisms to overcome them in a gracious manner. When considering the physical network layer where cable cuts and equipment failures represent the most common failures, two failure scenarios are considered.

1) *Single link failure*. In this failure scenario a link between two adjacent network nodes fails and as a consequence no information can be sent between them. Protection schemes protecting against these kind of failures can reroute around the end nodes of the failed link (figure 1(a)) or find a new path from the source to destination (figure 1(b)).
2) *Single node failure*. This is a situation in which a network element fails and hence all incident links are out of service (figure 1(c)).

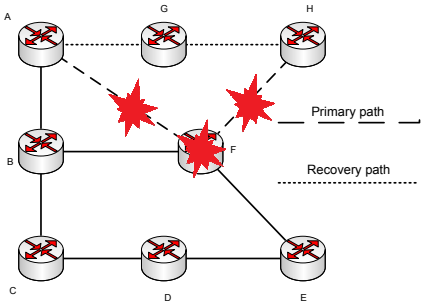
In this paper we are concerned with the first failure class. We will introduce two protection schemes which help to protect against link failures, by selecting an alternate back up path. A light path which carries traffic in failure-free conditions is known as a primary light path. We will reserve back-up capacity to make sure that in case of a failure of a link, all primary paths which were using that link can be rerouted via a back-up path. Before going into detail, we first have made a distinction regarding the allocation of this back up capacity. Two major options exist: dedicated or shared backup capacity. In the case of dedicated back-up capacity, a particular backup wavelength is exclusively reserved to protect a primary wavelength affected by the considered failure scenario. In the other case, one is able to share a backup wavelength between several primary wavelengths. The protection schemes we will introduce all fall in the latter category.



(a) Single link failure, with link protection. When the link A-F fails, this link is bypassed by the links A-B and B-F after which the original path is reused.



(b) Single link failure with path protection. When a link on the path from A to H fails, the backup path A-G-H is taken.

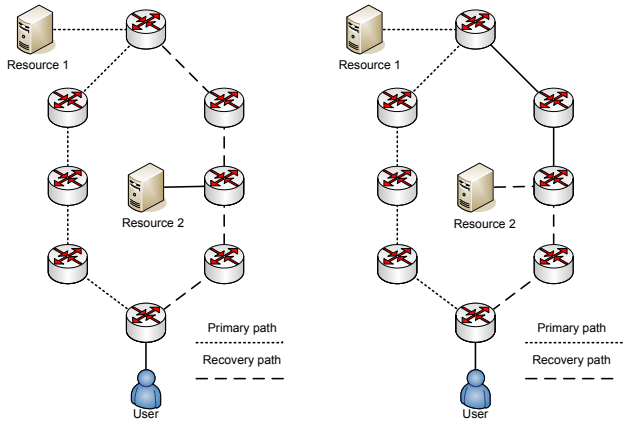


(c) Single node failure. When node F fails, the recovery path A-G-H is taken.

Fig. 1. The different failure scenario's in a communication network. The primary path is from A to H.

A. Shared path protection with relocation

In a standard fibreoptic path protection scheme where wavelengths can be shared, a light path is protected by a link-disjoint light path going from the source to the destination of that light path. These back-up paths can be shared, in case the corresponding primary paths are unlikely to fail at the same time. Now, we can extend this technique by incorporating the *anycast principle*. In anycast, there is a one-to-many association between network endpoints: when a user creates a job, there is a set of several resources which are able to execute it and only one of them is chosen, generally by the Grid scheduler. Now, instead of protecting the path by allocating bandwidth from the source to the indicated destination, we



(a) Traditional shared protection. (b) Shared protection with relocation possibility.

Fig. 2. Traditional protection compared to protection with relocation. In the traditional case the path is protected by a link disjoint backup path to the original destination. In the relocation case, we can decrease the length of the backup path with one link by relocating to resource 2, but it does increase the load at this resource.

can create a back-up path to another resource which could be closer to the destination in terms of back-up wavelengths, as outlined in Fig. 4(a) and Fig. 4(b). Of course, this means that there is a trade off between the protection against network-faults and resource capacity. Because in case of a single link fault, the jobs on a circuit using that link have to be relocated to another resource which will need more processing capacity.

III. Models

We aim to evaluate the aforementioned relocation scheme against traditional shared protection, from a network dimensioning perspective. Hence, we will use ILP formulations to derive the required amount of wavelengths to equip for a given demand. As explained before, in a Grid scenario the traffic is specified in number of jobs (and their characteristics) arriving at source sites only. The destination can be more or less freely chosen, hence there is no clearly defined (source, destination)-based traffic matrix. Yet, this is what is assumed as given for ILP dimensioning as outlined in section III-B. Hence, we first present the methodology used to derive a classical traffic matrix from the source job rates.

A. Job Demand Model

The (source, destination)-based traffic matrix will depend on (i) the location of the Grid resources (servers) capable of executing the jobs, (ii) the amount of servers at each of the chosen location, and (iii) the scheduling algorithm used to chose an available server. Thus, to obtain our traffic matrix, we first have to dimension the Grid and decide where to locate the server sites and the amount of server CPU's at each site (e.g. while meeting a maximum job loss rate criterion). For this, we resorted to iterative algorithm which is fully described in [4]. We start with a list of a job arrival rates for every site in the network.

1) Find the K best servers: The first thing we have to do is find the optimal choice for the server locations which actually is a k-medoid problem: we are looking for K clusters with the cluster centers representing the server sites and the members of the cluster corresponding to the Grid sites sending the jobs to the respective server. For this step we have written a small ILP formulation. The decision variables deciding on the server site locations are:

- $T_j = 1$ if and only if site j is chosen as a server site location, else 0.
- $S_{i,j} = 1$ if and only if site j is the target server for traffic from site i, else 0.

The given input parameters to base these decisions on are:

- λ_i job arrival rate at site j $(i = 1 \dots N)$.
- $H_{i,j}$ routing distance (for instance hop count) from site i to site j $(i, j = 1 \dots N)$.
- K the number of server sites to choose

With objective function

$$min \left(\sum_i \sum_j \lambda_i \cdot H_{i,j} \cdot S_{i,j} \right) \tag{1}$$

and constraints

$$\sum_j T_j = K \tag{2}$$

$$\sum_j S_{i,j} = 1 \qquad \forall i \tag{3}$$

$$S_{i,j} \leq T_j \qquad \forall i, j \tag{4}$$

2) Determining the server capacities: The next step involves dimensioning the servers in terms of processing capacity expressed in the number of CPU's. For this, we assume Poisson arrivals and exponentially distributed service times, and use the well-known ErlangB formula to establish the total number of servers needed to meet a maximum job loss rate of $x\%$. Hence, we solve ErlangB in equation 5 for n. We subsequently distribute that amount of n CPU's among the server sites, proportionally to the cluster arrival rate at each server site hoping to have a high performance since we install the most CPU's where the most traffic is arriving. (Note that [4] indeed showed this choice results in lower network loads.)

$$erlangB(\lambda, \mu, n) = \frac{\frac{(\frac{\lambda}{\mu})^n}{n!}}{\sum_{k=0}^{n} \frac{(\frac{\lambda}{\mu})^k}{k!}} = x \tag{5}$$

3) Scheduling policy: We have adopted a *mostfree* scheduling policy: choose a free CPU at server site f , where f is the server site with the highest number of free server CPUs, in an attempt to avoid overloading sites and thus limiting non-local job execution. In this step we have resorted to simulations because of the anycast principle: it is hard to obtain accurate estimates for the inter-site traffic using analytical techniques.

After this step we know how many jobs are exchanged between every Grid node pair in the considered network. By

scaling and rounding these numbers, we finally end up with a demand matrix containing a number of connections between each Grid node pair.

B. Dimensioning Model

We investigate a network design model with a static traffic matrix in which a known set of connection requests is assigned to the network. Each connection represents a point-to-point light path (circuit) from a source to a destination, able to transport a given capacity. Furthermore, we assume that all optical cross-connects (OXC's) are able to perform wavelength conversion which we will refer to as the virtual wavelength path (VWP) network [6]. When OXC's do not support wavelength conversion, the wavelength continuity constraint arises and the resulting network is a plain wavelength path (WP) network. In this paper, we only consider the VWP case.

Our topology is modeled as a graph $G = (E, V)$ where the bidirectional links are represented by an edge $e \in E$ with $|E| = L$ while the vertices $v \in V$ with $|V| = N$ represent the OXC's. Our static traffic matrix is converted into a list of connection objects $\beta = \{\phi_1, \phi_2, \ldots \phi_n\}$ where a connection corresponds to a single wavelength path and is identified by its index. Two connections can have the same source and the same destination.

1) Shared Path Protection: The variables in the formulation are the following:

- $P^{\phi_c}_{(i,j)} \in \{0, 1\}$ is a binary variable which states that link (i, j) is used for a primary path for demand $\phi_c \in \beta$.
- $R^{\phi_c}_{(i,j),(k,l)} \in \{0, 1\}$ is a binary variable stating that link (i, j) is used in a backup path for a failure of link (k, l) for connection $\phi_c \in \beta$.
- $\pi_{(i,j)} \in \mathbb{N}$ identifies the total number of protection wavelengths.
- We denote s and d as the source and respectively the destination of the connection ϕ_c

The actual formulation of our ILP model can be detailed as follows. The cost function to be minimized is the total number of wavelengths used in the topology.

$$min \left(\sum_{(i,j)} \pi_{(i,j)} + \sum_{(i,j)} \sum_{\phi_c} P^{\phi_c}_{(i,j)} \right) \tag{6}$$

The first constraint deals with the demand constraints and the flow conservations in the topology for the primary paths.

$$\sum_{i:(i,j)\in L} P^{\phi_c}_{(i,j)} - \sum_{k:(j,k)\in L} P^{\phi_c}_{(j,k)} = \begin{cases} -1: & j = s \\ +1: & j = d \\ 0: & else \end{cases} \quad \forall \phi_c \in \beta \tag{7}$$

For each parting primary path, there should be a back-up path (Demand constraints for the back-up paths).

$$P^{\phi_c}_{(i,j)} = \sum_{(s,e):e\in V} R^{\phi_c}_{(s,e)(i,j)} \qquad \forall (i,j) \in L, \forall \phi_c \in \beta \tag{8}$$

$$P^{\phi_c}_{(i,j)} = \sum_{(e,d):d\in V} R^{\phi_c}_{(e,d)(i,j)} \qquad \forall (i,j) \in L, \forall \phi_c \in \beta \tag{9}$$

After this we have to create the flow conservations for the back up paths.

$$\sum_{i:(i,j)\in L} R^{\phi_c}_{(i,j)(k,l)} - \sum_{p:(j,p)\in L} R^{\phi_c}_{(j,p)(k,l)} = 0 \tag{10}$$
$$\forall j \neq s, j \neq d \text{ of } \phi_c, \forall \phi_c \in \beta, \forall (k,l) \in L$$

We continue with the constraint stating that a primary path and a back-up path protecting that primary path cannot overlap.

$$R^{\phi_c}_{(i,j)(k,l)} + P^{\phi_c}_{(i,j)} \leq 1 \tag{11}$$

$$\forall \phi_c \in \beta, \forall (i,j), (k,l) \in L$$

Now we have to create the variables which express the total number of wavelengths used for back-up paths on a specified link.

$$\pi_{(i,j)} \geq \sum_{\phi_c} R^{\phi_c}_{(i,j)(k,l)} \tag{12}$$

$$\forall (i,j) \neq (k,l) \in L, \forall (k,l) \in L$$

2) Shared Path Protection with Job Relocation: The ILP model for the shared protection scheme with relocation possibility differs from the normal model by the observation that the back-up path does not necessarily have to go to the resource which originally was proposed as destination for this connection. At first we declare $\Delta = \{\delta_1, \delta_2, \ldots, \delta_r\}$ as the set of all the resources which can be used to relocate a job to and secondly we introduce the variable $R^{\phi_c,\delta}_{(i,j)(k,l)} \in \{0, 1\}$ expressing that link (i, j) is protecting link (k, l) for connection ϕ_c by relocating to resource δ. Remark that δ can be the original resource as stated in the demand matrix.

The objective function looks just like the shared protection objective :

$$min \left(\sum_{(i,j)} \pi_{(i,j)} + \sum_{(i,j)} \sum_{\phi_c} P^{\phi_c}_{(i,j)} \right) \tag{13}$$

Just as in the normal shared protection scheme we begin with the demand constraints and flow conservations for the primary paths.

$$\sum_{i:(i,j)\in L} P^{\phi_c}_{(i,j)} - \sum_{k:(j,k)\in L} P^{\phi_c}_{(j,k)} = \begin{cases} -1: & j = s \\ +1: & j = d \\ 0: & else \end{cases} \quad \forall \phi_c \in \beta \tag{14}$$

ILP	Variables	Constraints
Shared	$L^2 \cdot \lvert\beta\rvert + L \cdot \beta$	$N \cdot \lvert\beta\rvert$ $+3 \cdot L \cdot \lvert\beta\rvert$ $+N \cdot \lvert\beta\rvert \cdot L + L$
Shared relocation	$L^2 \cdot \lvert\beta\rvert \cdot \lvert\Delta\rvert + L \cdot \beta$	$N \cdot \lvert\beta\rvert$ $+3 \cdot L \cdot \lvert\beta\rvert$ $+N \cdot \lvert\beta\rvert \cdot L \cdot \lvert\Delta\rvert + L$

TABLE I
THE NUMBER OF VARIABLES AND CONSTRAINTS FOR THE ILP FORMULATIONS

We continue with the demand constraints for the back up paths.

$$P^{\phi_c}_{(i,j)} = \sum_{\delta \in \Delta} \sum_{(s,e):s \in V} R^{\phi_c,\delta}_{(s,e)(i,j)} \qquad (15)$$
$$\forall (i,j) \in E, \qquad \forall \phi_c \in \beta$$
$$P^{\phi_c}_{(i,j)} = \sum_{\delta \in \Delta} \sum_{(e,d):d \in V} R^{\phi_c,\delta}_{(e,d)(i,j)} \qquad (16)$$
$$\forall (i,j) \in E, \qquad \forall \phi_c \in \beta$$

Again we have to formulate the flow conservation for the back-up paths.

$$\sum_{i:(i,j) \in L} R^{\phi_c,\delta}_{(i,j)(k,l)} - \sum_{p:(j,p) \in L} R^{\phi_c,\delta}_{(j,p)(k,l)} = 0 \qquad (17)$$
$$\forall j \neq s, j \neq d \text{ of } \phi_c \in \beta, \forall \delta \in \Delta, \forall (k,l) \in L$$

We follow with the constraint stating that a primary path and the recovery path protecting that primary path cannot share the same link.

$$\sum_{\delta \in \Delta} R^{\phi_c,\delta}_{(i,j)(k,l)} + P^{\phi_c}_{(i,j)} \leq 1 \qquad (18)$$
$$\forall \phi_c \in \beta, \forall (i,j) \in L, \forall (k,l) \in L$$

Finally we have to introduce the variables counting the total number of wavelengths on a link used for a back-up path.

$$\pi_{(i,j)} \geq \sum_{\delta \in \Delta} \sum_{\phi_c} R^{\phi_c,\delta}_{(i,j)(k,l)} \qquad (19)$$
$$\forall (i,j) \neq (k,l) \in L, \forall (k,l) \in L$$

C. Complexity

The execution time of an ILP program depends heavily on he number of variables the problem introduces and by a less important factor, the number of constraints which are introduced. We have put the number of variables and constraints in table I and as can be noted, the difference between the number of variables for the two ILP only differ by a factor $|\Delta|$. Both ILP's greatly depend on the number of links of the topology and the number of connections which should be established.

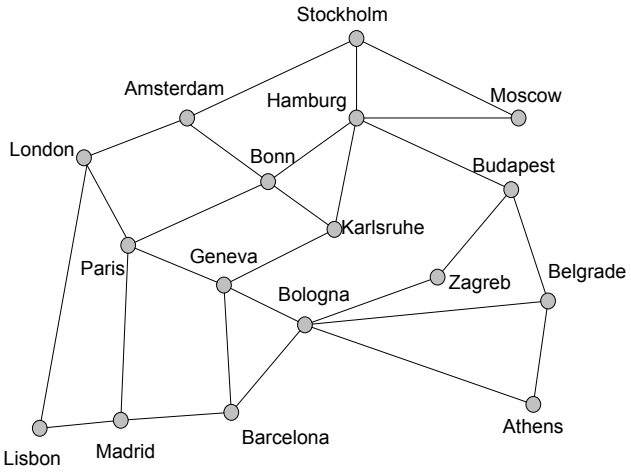


Fig. 3. Topology of the considered network

IV. CASE STUDY

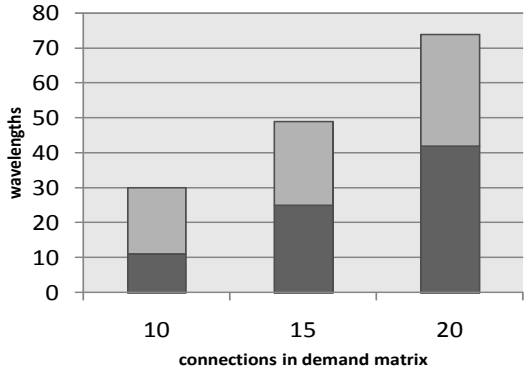
The topology we have considered is depicted in figure 3 and it is based on the Geant2 network topology and its associated various national research- and education networks (NRENs). It consists of 17 nodes and 54 links. We have made three traffic matrices, by applying the dimensioning strategy explained in III-A1,III-A2 and III-A3 resulting in traffic matrices with 10, 15 and 20 connection requests.

A. Network Load

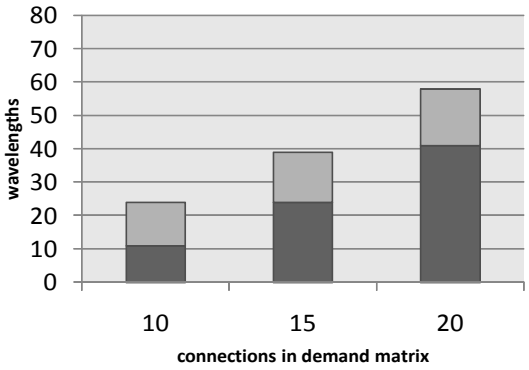
In figure 4 we can see the total number of wavelengths summed over all links which are being used for the primary paths and the back-up paths. We have run three tests, namely with a demand matrix with $|\beta| = 10$, 15, and 20 connection requests. We see that with an increasing load, the network occupation also increases as is expected. When we compare 4(a) with figure 4(b) we see that the number of primary wavelengths which are needed for both protection schemes is about the same, but the amount of wavelengths needed for the back-up paths differs a lot. As can be noted in figure 5, when applying the relocation protection scheme only about 80% of the number of wavelengths in the normal shared protection scheme is used, so this means a reduction of 20%.

B. Resources

As previously demonstrated, by relocating to another server site instead of the proposed server site by the Grid scheduler we can obtain an overall network optimization. But there is a trade off: by relocation, the relocation server receives more jobs than the scheduler intended and thus, needs to reserve some spare capacity in order to execute the relocated job. In figure 6 we have represented the maximum amount of extra connections a resource receives. These connections can be seen as the extra load due to a single link fault. For each server site there are three bars (10, 15 and 20 connection requests), with the lowest part representing the normal load (caused by the traffic matrix) and the upper part comprising



(a) Traditional shared protection.



(b) Shared protection with relocation.

Fig. 4. Traditional protection compared to protection with relocation, expressed in wavelengths. The dark part represents the number of wavelengths used for the primary paths, the lighter part stands for the number of wavelengths used for the back-up paths.

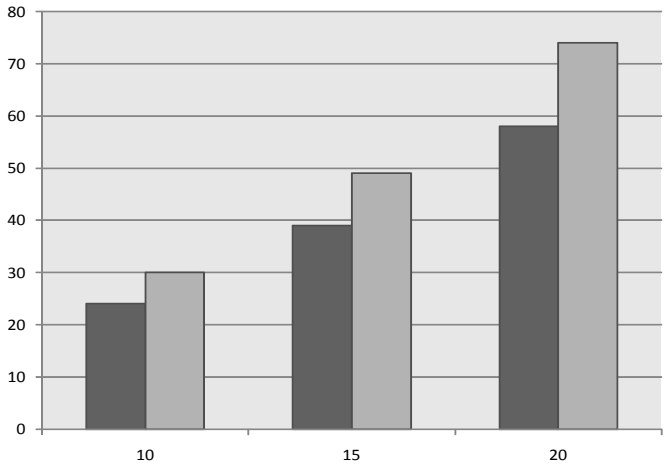


Fig. 5. The total number of wavelengths summed over all links. The darker bars represent the results when using the shared relocation scheme, the lighter bars stands for the traditional protection scheme.We see that the number of wavelengths needed when relocation is involved is significantly smaller.

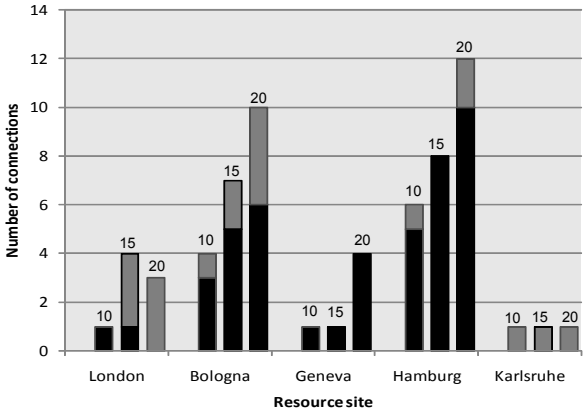


Fig. 6. Extra resource capacity needed for the resource, expressed in number of incoming connections. The black part of the bar is the number of incoming connections due to the demand matrix, the gray part of the bar, is the maximum number of extra connections caused by a link failure.

the maximal extra connections in case of single link failures. We see that this extra capacity can range from three times the normal capacity to no extra capacity. On average maximum spare, a resource needs about 25% extra connections compared to the original demand so he it able to execute all the relocated jobs in case of a single link fault.

V. Conclusion

In this work we have described an alternative method for path protection against single link failures. Whereas traditional protection resilience schemes try to reserve back-up capacity to the original destination of the primary path, we have integrated the Grid-specific any-cast principle into the scheme. This principle states that there are several destinations possible for a job to be executed. Therefore, in case of a network failure, we relocate the job to another possible resource, in order to minimize the bandwidth which needs to be allocated for the back-up path. We have described an ILP model for the traditional shared protection scheme and a ILP formulation for shared protection with relocation possibility. Our case study pointed out that we can have a reduction of the total number of necessary wavelengths of 20%, while on average each resource should only need 25% of extra capacity to handle the relocated jobs.

Acknowledgment

The work described in this paper was carried out with the support of the BONE-project (Building the Future Optical Network in Europe), a Network of Excellence funded by the European Commission through the 7th ICT-Framework Programme, as well as the IST Phosphorus-project. The Flemish government partly funded this work through the Research Foundation (FWO). C. Develder is a post-doctoral fellow of the FWO. Jens Buysse is supported by the research institute IWT.

REFERENCES

[1] Bill Allcock, Ian Foster, Veronika Nefedova, Ann Chervenak, Ewa Deelman, Carl Kesselman, Jason Lee, Alex Sim, Arie Shoshani, Bob Drach, and Dean Williams. High-performance remote access to climate simulation data: a challenge problem for data grid technologies. In *Supercomputing '01: Proceedings of the 2001 ACM/IEEE conference on Supercomputing (CDROM)*, pages 46–46, New York, NY, USA, 2001. ACM.

[2] G. Allen, G. Daues, J. Novotny, and J. Shalf. The astrophysics simulation collaboratory portal: A science portal enabling community software development. In *HPDC '01: Proceedings of the 10th IEEE International Symposium on High Performance Distributed Computing*, page 207, Washington, DC, USA, 2001. IEEE Computer Society.

[3] Stephen Barnard, Rupak Biswas, Subhash Saini, Robert Van der Wijngaart, Maurice Yarrow, Lou Zechtzer, Ian Foster, and Olle Larsson. Large-scale distributed computational fluid dynamics on the information power grid using globus. In *FRONTIERS '99: Proceedings of the The 7th Symposium on the Frontiers of Massively Parallel Computation*, page 60, Washington, DC, USA, 1999. IEEE Computer Society.

[4] Chris Develder, Biswanath Mukherjee, Bart Dhoedt, and Piet Demeester. On dimensioning optical grids and the impact of scheduling. *Photonic Network Communications (PNET)*, May 2008.

[5] Massimo Tornatore, Guido Maier, and Achille Pattavina. Wdm network design by ilp models based on flow aggregation. *IEEE/ACM Trans. Netw.*, 15(3):709–720, 2007.

[6] Hui Zang, Canhui Ou, and Biswanath Mukherjee. Path-protection routing and wavelength assignment (rwa) in wdm mesh networks under duct-layer constraints. *IEEE/ACM Trans. Netw.*, 11(2):248–258, 2003.