Dimensioning resilient optical Grids*

Jens Buysse†, Marc De Leenheer, Bart Dhoedt, *Member, IEEE*, Chris Develder‡, *Member, IEEE*
Ghent University – IBBT, Dept. of Information Technology - IBCN,
G. Crommenlaan 8 bus 201, BE-9050 Gent, Belgium
Tel: +32 9 3314961, Fax: +32 9 3314899, e-mail: chris.develder@intec.ugent.be

ABSTRACT
An important problem in optical networking is to dimension the network: given the amount of traffic to carry, determine the required amount of network resources (esp. wavelengths). In traditional scenarios, the traffic is specified in terms of a (source,destination)-based traffic matrix. In an optical Grid scenario however, the anycast principle applies: users submit jobs, and generally do not care where exactly they end up being executed. Thus, the destination of traffic is not known beforehand and traditional dimensioning algorithms are not directly applicable. On the other hand, this flexibility in choosing a destination opens opportunities to save on backup network resources: to protect against failures, we can opt to redirect jobs to another location (i.e. exploit relocation). In this paper we (i) outline how to derive a traffic matrix in a step-wise grid dimensioning approach, and (ii) present an assessment of potential network resource savings in resilient network dimensioning by exploiting relocation.
Keywords: optical Grids, dimensioning, resilience, shared path protection, relocation, integer linear programming

1. INTRODUCTION

In several research fields, the need arose to build powerful computer systems to face computational and data storage challenges (e.g. particle physics, astrophysics, etc.). To meet the demand for a huge common resource pool to process the tasks (jobs) at hand, networks interconnecting cluster centres were deployed. This led to the creation of so-called Grids. More recently, the potential of Grid infrastructure for more consumer/business oriented applications was acknowledged by industry, and referred to as cloud computing [1]. (In this paper, we will stick to the term Grids to also include cloud computing.) To realize the interconnecting Grid network, optical technology is the solution of choice, able to meet both the high data rates typical of many eScience applications and the low latency requirements associated with most business/consumer solutions.

In this paper we address the Grid dimensioning problem. The input is (i) the network topology comprising the locations of the sites where jobs originate (which could be aggregation points, e.g., points-of-presence (PoP) nodes of Grid service providers) and the (backbone) network interconnecting them, and (ii) the amount of jobs generated at each of the sites. Dimensioning amounts to figuring out the network resources required to process the submitted jobs. The major difference with classical (optical) network dimensioning arises from the anycast principle: only the source of the jobs is given, not the destination (which can be freely chosen by some job scheduling algorithm) and hence we are not given the complete so-called traffic matrix. Many dimensioning algorithms are available, either based on heuristics or exact solution methods using for example Integer Linear Programming (ILP). The algorithms vary depending on the network technologies and topologies, design criteria (such as survivability [2], availability), single or multi-period planning [3] (where the network evolves over time, usually spanning multiple years), etc. To apply any of these approaches for dimensioning grids, the problem arises of accurately estimating the traffic matrix (cf. anycast principle).

Section 2 outlines an iterative Grid dimensioning approach, translating job arrival rates to a traditional traffic matrix, by determining the locations and server capacity of Grid server sites. Given a particular scheduling algorithm, the site-to-site job rates are derived. The core of the paper then focuses on dimensioning of a resilient optical network to carry the considered jobs. For this, in Section 3 we propose to exploit the anycast principle and apply a relocation scheme to provide backup paths to alternate destinations (compared to the primary paths). A sample case study is discussed in the subsequent Section 4 and conclusions are summarized in Section 5.

2. AN ITERATIVE GRID DIMENSIONING APPROACH

To solve the general problem statement of Grid dimensioning (summarized in Fig. 1), we propose an iterative dimensioning approach. We start with an algorithm for choosing appropriate server site locations (cf. not every site will necessarily have servers). Next we calculate the required amount of servers (and distribute them over

* 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).
† Jens Buysse is supported by the research institute IWT.
‡ C. Develder is a post-doctoral fellow of the FWO.

the chosen server site locations). Lastly the inter-site job rates are determined, which at that point can be stated as a traditional (source,destination)-based traffic matrix. For a detailed discussion, and use of this methodology to assess the impact of server distributions and scheduling strategies on resulting network traffic, we refer to [4].

Given:	***Find:***
- Graph representing the network topology (nodes representing Grid sites and switches, links the optical fibers interconnecting them), - Arrival process of jobs originating at each site, - Job processing capacity of a single server CPU (an average of λ jobs/s), and - Target maximum job loss rate,	- Locations of the server sites, - Amount of Grid server CPUs at each site, and - Amount of link bandwidth to install, - While meeting the maximum job loss rate criterion and minimizing network capacity

Fig. 1. The Grid dimensioning problem statement

The first step in solving our Grid dimensioning problem is to figure out which locations are best suited for placing the servers. The cost criterion to measure by will be the total expected link bandwidth. The major difficulty in evaluating that cost for a given choice of *K* locations, is that the required bandwidth depends also on the amount of server capacity installed at each of the server sites and possibly the Grid scheduling and routing algorithm. Therefore, we make some simplifying assumptions: (i) each Grid site *i* will send all its jobs to a single destination D_i, and (ii) shortest path routing is used. Hence, given a choice of *K* locations, a site *i* will send its jobs to server site *j* if the routing distance H_{ij} is the minimum over all H_{ik} values for $k = 1..K$. This gives rise to the ILP formulation in Fig. 2.

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

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

$$\min \sum_i \sum_j \lambda_i \cdot H_{ij} \cdot S_{ij} \quad \text{with} \quad \begin{cases} \sum_j T_j = K & \text{(only K server locations)} \\ \sum_j S_{ij} = 1 \quad \forall i & \text{(simplifying assumption: all traffic is sent to exactly one server, i.e. the closest one)} \\ S_{ij} \le T_j \quad \forall i, j & \text{(only send traffic to server sites)} \end{cases}$$

Fig. 2. ILP for choosing K server site locations

Backed by real world Grid measurements [5], we assume Poisson job arrivals (mean arrival rate λ_i at site *i*). This implies that, given the lack of buffers, we can use the ErlangB formula (1) to calculate the total number of server CPUs *n* required to achieve a maximal loss rate *L*. We then distribute this amount of servers over the *K* chosen locations proportionally (*prop*) to the to the (cluster) arrival rate at each server site *k*: $n_k = \lambda_k^* / (K \cdot \lambda)$, with $\lambda = \Sigma \lambda_i$ and $\lambda_k^* = \Sigma \lambda_i \cdot S_{ik}$, where S_{ik} is 1 if and only if *k* is the server site closest to *i* (as defined in the ILP of Fig. 2; hence λ_k^* equals the total job arrival rate summed over all Grid sites in cluster *k*). This proportional distribution was shown to be most beneficial to reduce bandwidth requirements [4].

$$L = \mathrm{ErlangB}(n, \lambda, \mu) = \frac{(\lambda/\mu)^n / n!}{\sum_{k=0}^{n} (\lambda/\mu)^k / k!} \tag{1}$$

The final step then involves simulation, taking into account the Grid scheduling algorithm, to determine the amount of jobs actually exchanged between each (source, destination)-pair (a destination being one of the *K* server site locations). The resulting traffic matrix is used as an input for the following resilient Grid dimensioning algorithms in the next section.

3. RESILIENT GRID DIMENSIONING

In the following, we propose a resilience scheme to protect an optical circuit-switched (OCS) network against unplanned single link failures. We consider a *path protection* scheme, i.e. we reserve backup wavelengths in advance (as opposed to restoration, where a recovery path is sought only after the failure occurred) to protect failures along a primary path. We focus on *shared* protection, i.e. the backup wavelengths can be shared among backup paths that protect against non-simultaneously occurring failures (cf. single link failure assumption). Obviously, this sharing allows for reserving a lower total number of wavelengths. In a Grid scenario, further reduction of necessary backup wavelength capacity can be achieved by exploiting the aforementioned anycast

principle. Indeed, since in general users do not care where exactly their jobs end up being executed, we can choose to set-up backup wavelength path to a different end point than the one of the primary path. This amounts to *relocation*, where jobs are relocated to another resource.

In Fig. 3, we present ILP formulations for dimensioning an OCS network both the classical shared path protection scheme and the relocation scheme. Given the network topology and a (source,destination)-based traffic matrix, along with the set of Grid server sites (e.g. the *K* locations as determined with the Grid site dimensioning scheme of Section 2), we try to minimize the total number of wavelengths needed, summed over all links, to carry the given traffic. The topology is modelled as a graph *G=(L,V)*, with *L* the links and *V* the nodes (representing OXCs). The traffic matrix is reformulated as a set *C* of connections φ_c. The ILP formulation in Fig. 3(a) is loosely based on those in [6]: we also assume wavelength conversion, and use a so-called flow formulation. (In future work, we plan to adopt flow aggregation to end up with a so-called source-formulation, to improve scalability of the ILP, as originally proposed in [7].)

(a)

(2) $min\left(\sum_{(i,j)\in L}\pi_{(i,j)}+\sum_{(i,j)\in L}\sum_{\varphi_c\in C}P_{(i,j)}^{\varphi_c}\right)$

Primary path flow constraints:

(4) $\sum_{i:(i,j)\in L}P_{(i,j)}^{\varphi_c}-\sum_{k:(j,k)\in L}P_{(j,k)}^{\varphi_c}=\begin{cases}-1 & j=s\\ +1 & j=d\\ 0 & else\end{cases},$
$\forall\varphi_c\in C$

Backup path flow constraints:

(6) $P_{(i,j)}^{\varphi_c}=\sum_{e:(s,e)\in L}R_{(s,e)(i,j)}^{\varphi_c}\quad\forall(i,j)\in L,\forall\varphi_c\in C$

(7) $P_{(i,j)}^{\varphi_c}=\sum_{e:(e,d)\in L}R_{(e,d)(i,j)}^{\varphi_c}\quad\forall(i,j)\in L,\forall\varphi_c\in C$

(8) $\sum_{k:(k,l)\in L}R_{(k,l)(i,j)}^{\varphi_c}-\sum_{m:(l,m)\in L}R_{(l,m)(i,j)}^{\varphi_c}=0,$
$\forall l\in V\setminus\{s,d\},\forall(i,j)\in L,\forall\varphi_c\in C$

Non-overlapping primary and backup paths:

(12) $R_{(i,j)(k,l)}^{\varphi_c}+P_{(i,j)}^{\varphi_c}\le 1,$
$\forall(i,j)\in L,\forall(k,l)\in L,\forall\varphi_c\in C$

Counting the number of backup wavelengths:

(14) $\pi_{(i,j)}\ge\sum_{\varphi_c\in C}R_{(i,j)(k,l)}^{\varphi_c},$
$\forall(i,j)\in L,\forall(k,l)\in L\setminus\{(i,j)\}$

(b)

(3) $min\left(\sum_{(i,j)\in L}\pi_{(i,j)}+\sum_{(i,j)\in L}\sum_{\varphi_c\in C}P_{(i,j)}^{\varphi_c}\right)$

Primary path flow constraints:

(5) $\sum_{i:(i,j)\in L}P_{(i,j)}^{\varphi_c}-\sum_{k:(j,k)\in L}P_{(j,k)}^{\varphi_c}=\begin{cases}-1 & j=s\\ +1 & j=d\\ 0 & else\end{cases},$
$\forall\varphi_c\in C$

Backup path flow constraints:

(9) $P_{(i,j)}^{\varphi_c}=\sum_{\delta\in D}\sum_{e:(s,e)\in L}R_{(s,e)(i,j)}^{\varphi_c,\delta},$
$\forall(i,j)\in L,\forall\varphi_c\in C$

(10) $P_{(i,j)}^{\varphi_c}=\sum_{\delta\in D}\sum_{e:(e,d)\in L}R_{(e,d)(i,j)}^{\varphi_c,\delta},$
$\forall(i,j)\in L,\forall\varphi_c\in C$

(11) $\sum_{k:(k,l)\in L}R_{(k,l)(i,j)}^{\varphi_c,\delta}-\sum_{m:(l,m)\in L}R_{(l,m)(i,j)}^{\varphi_c,\delta}=0,$
$\forall l\in V\setminus\{s,d\},\forall(i,j)\in L,\forall\delta\in D,\forall\varphi_c\in C$

Non-overlapping primary and backup paths:

(13) $\sum_{\delta\in D}R_{(i,j)(k,l)}^{\varphi_c,\delta}+P_{(i,j)}^{\varphi_c}\le 1,$
$\forall(i,j)\in L,\forall(k,l)\in L,\forall\varphi_c\in C$

Counting the number of backup wavelengths:

(15) $\pi_{(i,j)}\ge\sum_{\delta\in D}\sum_{\varphi_c\in C}R_{(i,j)(k,l)}^{\varphi_c,\delta},$
$\forall(i,j)\in L,\forall(k,l)\in L\setminus\{(i,j)\}$

Fig. 3. ILP for resilient dimensioning: (a) shared path protection, (b) shared path protection with relocation.

The decision variables (the unknowns) in this formulation are:

- $P_{(i,j)}^{\varphi_c}\in\{0,1\}$ equals 1 if and only if link (*i*,*j*) is used for the primary path of connection φ_c.
- $R_{(i,j),(k,l)}^{\varphi_c}\in\{0,1\}$ equals 1 if and only if link (*i*,*j*) is used in the backup path of connection φ_c to protect against failure of link (*k*,*l*).
- $\pi_{(i,j)}\in\aleph$ is an auxiliary variable identifying the total number of protection wavelengths on link (*i*,*j*).

For the formulation with relocation, we also assume the possible destinations, i.e. the Grid server sites are given as a set *D*. In this case, we define the decision variables *R* as follows:

- $R^{\varphi_c,\delta}_{(i,j),(k,l)} \in \{0,1\}$ equals 1 if and only if link (*i*,*j*) is used in the backup path towards destination site $\delta \in D$ of connection φ_c to protect against failure of link (*k*,*l*).

4. CASE STUDY

To assess the possible gain of adopting relocation, we performed a case study on a European network topology as depicted in Fig. 4. We translated job arrival traces (the same as in [5]) to a varying number of connections, and subsequently solved the above ILP. The results are summarized in Fig. 5. We observe that compared to classical shared path protection, we need about the same number of wavelengths for the primary paths (see Fig. 5b), but can save quite a lot on backup wavelengths by exploiting relocation: the total number of wavelengths required for both primary and backup paths amounts to only about 80%. The price paid for this is an increase in the amount of jobs to process at the sites jobs are relocated to. For this (limited) case study, we found it amounted to an average increase of maximal server site load with 25%.

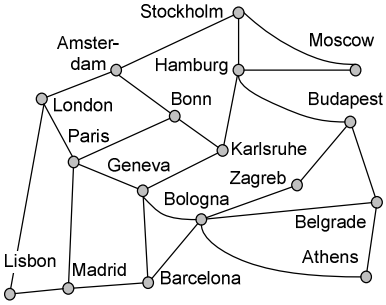
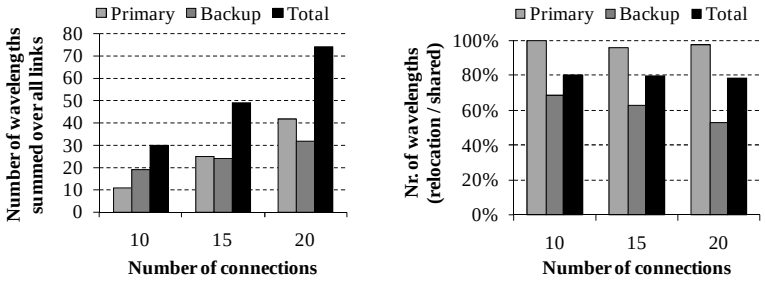


Fig. 4. A European network topology based on the EGEE sites and associated national research- and education networks (NRENs).



(a) (b)

Fig. 5. The number of wavelengths summed over all links, split up in primary wavelengths, backup and total wavelengths: (a) Absolute number for shared path protection, (b) For the relocation scheme, relative to classical shared protection.

5. CONCLUSIONS

The dimensioning problem in an optical Grid scenario, compared to classical optical networking, is complicated by the anycast principle: users generally do not care where exactly their jobs end up being executed. This extra degree of freedom implies the classical (source,destination)-based traffic matrix for network dimensioning is not given a priori, but (heavily) depends on the Grid site dimensioning and scheduling. However, as we outlined, it is possible to use a stepwise dimensioning scheme to derive the sought traffic matrix. To dimension a resilient optical network, the anycast principle can be exploited to choose for relocating traffic in case of failures (rather than providing backup capacity towards the original destination). We defined an ILP solution for the case of optical circuit-switched WDM networks with wavelength conversion. In a case study, we showed that exploiting relocation can achieve around 20% reduction of total wavelength capacity in the network.

REFERENCES

[1] G. Lawton, "Moving the OS to the web," *IEEE Computer*, vol. 41, no. 3, pp. 16–19, Mar. 2008.

[2] D. Colle, S. De Maesschalck, C. Develder, P. Van Heuven, A. Groebbens, J. Cheyns, I. Lievens, M. Pickavet, P. Lagasse, and P. Demeester, "Data-centric optical networks and their survivability," *IEEE J. Selected Areas in Communications*, vol. 20, no. 1, pp. 6–20, Jan. 2002.

[3] M. Pickavet and P. Demeester, "Long-term planning of WDM networks: A comparison between single-period and multi-period techniques," *Photonic Network Communications*, vol. 1, no. 4, pp. 331–346, Dec. 1999.

[4] C. Develder, B. Mukherjee, B. Dhoedt, and P. Demeester, "On dimensioning optical grids and the impact of scheduling," *Photonic Network Commun. (PNET)*, 2008.

[5] K. Christodoulopoulos, E. Varvarigos, C. Develder, M. De Leenheer, and B. Dhoedt, "Job demand models for optical grid research," in *Proc. 11th Int. IFIP TC6 Conf. on Optical Netw. Design and Modeling (ONDM2007)*, ser. Lecture Notes In Computer Science, vol. 4534, May 2007, pp. 127–136.

[6] H. Zang, C. Ou, and B. Mukherjee, "Path-protection routing and wavelength assignment (RWA) in WDM mesh networks under duct-layer constraints," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 248–258, Apr. 2003.

[7] M. Tornatore, G. Maier, and A. Pattavina, "WDM network design by ILP models based on flow aggregation," *IEEE/ACM Trans. Netw.*, vol. 15, no. 3, pp. 709–720, 2007.