Available online at www.sciencedirect.com
ScienceDirect
Future Generation Computer Systems 24 (2008) 549–560
FGCS FUTURE GENERATION COMPUTER SYSTEMS
www.elsevier.com/locate/fgcs

Scalable dimensioning of resilient Lambda Grids

Pieter Thysebaert*,[1], Marc De Leenheer, Bruno Volckaert, Filip De Turck[2], Bart Dhoedt, Piet Demeester

Ghent University - IBBT - IMEC, Department of Information Technology, Gaston Crommenlaan 8 bus 201, 9050 Gent, Belgium

Received 20 October 2006; received in revised form 13 August 2007; accepted 13 August 2007
Available online 31 August 2007

Abstract

Grids consist of the aggregation of numerous dispersed computational, storage and network resources, able to satisfy even the most demanding computing jobs. Due to the data-intensive nature of Grid jobs, there is an increasing interest in Grids using optical transport networks as this technology allows for the timely delivery of large amounts of data. Such Grids are commonly referred to as *Lambda Grids*.

An important aspect of Grid deployment is the allocation and activation of installed network capacity, needed to transfer data and jobs to and from remote resources. However, the exact nature of a Grid's network traffic depends on the way arriving workload is scheduled over the various Grid sites. As Grids possibly feature high numbers of resources, jobs and users, solving the combined Grid network dimensioning and workload scheduling problem requires the use of scalable mathematical methods such as Divisible Load Theory (DLT). Lambda Grids feature additional complexity such as wavelength granularity and continuity or conversion constraints must be enforced. Additionally, Grid resources cannot be expected to be available at all times. Therefore, the extra complexity of resilience against possible resource failures must be taken into account when modelling the combined Grid network dimensioning and workload scheduling problem, enforcing the need for scalable solution methods. In this work, we tackle the Lambda Grid combined dimensioning and workload scheduling problem and incorporate single-resource failure or unavailability scenarios. We use Divisible Load Theory to tackle the scalability problem and compare non-resilient lambda Grid dimensioning to the dimensions needed to survive single-resource failures. We distinguish three failure scenarios relevant to lambda Grid deployment: computational element, network link and optical cross-connect failure. Using regular network topologies, we derive analytical bounds on the dimensioning cost. To validate these bounds, we present comparisons for the resulting Grid dimensions assuming a 2-tier Grid operation as a function of varying wavelength granularity, fiber/wavelength cost models, traffic demand asymmetry and Grid scheduling strategy for a specific set of optical transport networks.

1. Introduction

By coupling numerous heterogeneous computational and storage resources distributed over various locations, Grids are able to satisfy the ever increasing demand for both processing and storage power, surpassing the capabilities of each of its individual resources. This allows a Grid to accommodate even the largest and most resource-demanding applications. These Grid applications typically need access to multiple resources simultaneously (so-called *co-allocation*); the most common types of resources include computational resources, data storage resources and the transport network interconnecting the various Grid sites. As the computational requirements for typical Grid applications originate from the large amounts of data they need to process, the transportation of this data between the involved Grid resources is an important factor when it comes to cost and time efficient scheduling of the Grid's workload. Optical circuit-switched transport networks allow for high-bandwidth end-to-end transfers capable of low latency delivery of these large amounts of data, and thus are well-suited to interconnect the various Grid resources. The relevance of optical networks in Grids is illustrated by the recent increase in research activities into these "supernetworks" [1–3]. Grids making use of optical circuit-switched transport networks are usually denoted as *Lambda Grids*.

If a Grid's aggregate power is to be successfully exploited, an important problem to be solved is to determine how

* Corresponding author.
E-mail address: pieter.thysebaert@intec.ugent.be (P. Thysebaert).
[1] Research Assistant with the Fund for Scientific Research—Flanders (FWO-V).
[2] Post-doctoral Fellow with the Fund for Scientific Research—Flanders (FWO-V).

0167-739X/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2007.08.003

the expected workload generated at each location should be distributed over the Grid's resources in case it cannot be handled locally. This is the Grid workload *scheduling* problem. Moreover, given a set of possible locations where Grid resources can be deployed, the question arises how operators interested in installing Grid infrastructure should decide on the capacities of the resources (influenced by the scheduling strategy) to be allocated at each site. This is known as a *dimensioning* problem.

An immediate observation is that the required network capacity directly depends on the expected traffic pattern, which in turn depends on the scheduling policy regulating workload distribution in the Grid. A study of the lambda Grid dimensioning problem thus necessarily needs to simultaneously take into account this scheduling policy. Thus, this combined lambda Grid dimensioning and scheduling problem differs from the isolated problems of optical transport network dimensioning (from static demand matrices) and parallel machine scheduling (in which computational resources are the major focal point). In addition, while these isolated problems can be modelled using Integer Linear Programs, naively combining both problems into a single linear program quickly yields intractable program sizes. This issue is complicated further by the unpredictable resource unavailability (due to different autonomous management policies, possible failures etc.) common to Grids, as the combined workload scheduling and network dimensioning problem should also model the resilience constraints protecting against these unavailabilities.

In this paper, we focus on the combined workload scheduling and optical transport network dimensioning problem in lambda Grids where processing and storage capacity has already been installed. This scenario is of importance to providers looking to deal with temporary spikes in local processing demand, as in such a scenario the connected remote sites can help us to address the excess load. We explicitly take into account possible failure scenarios featuring a single-resource failure (either a Computational Resource failure, network link failure or optical cross-connect failure).

We solve this combined workload scheduling and network dimensioning problem for these scenarios using the technique of Divisible Load Theory, which yields a more scalable approximation to the problem than yielded by combining classical Integer Linear Programming formulations for the workload scheduling and optical transport network dimensioning problems. We present heuristics which further reduce the combined problem's complexity and enable us to explore multiple operational scenarios simultaneously.

For regular optical transport network topologies, we derive analytical bounds on the additional cost incurred by taking into account possible resource failures in the combined workload scheduling and network dimensioning problem. To validate these bounds, we perform an extensive set of experiments using the heuristics described above to obtain these additional costs for a wide range of irregular network topologies. We present this additional cost as a function of varying network topology, wavelength granularity, fiber/wavelength cost models, traffic demand assymmetry and Grid scheduling strategy.

The remainder of the paper is structured as follows: Section 2 gives an overview of existing mathematical models and solution techniques to similar or related workload scheduling and dimensioning problems. Section 3 formally defines the problem under study. In Section 4, an accurate mathematical model extending one of the techniques from Section 2 for the problem at hand (optical network dimensioning in a Grid context) is provided. Several techniques are shown to reduce the model's computational complexity, one of which is our proposed DLT-based formulation. Results and discussions are presented in Section 6, and the main conclusions are presented in Section 7.

2. Related work

Optical network dimensioning and workload scheduling in parallel computer environments have been widely studied in the literature. Static network dimensioning starts from a given demand matrix, i.e. a matrix representation of the traffic demands between each pair of nodes in the network. In the case of optical circuit-switched networks (the type of networks considered in this paper), the required network dimensions follow from the solution to a so-called Routing and Fiber and Wavelength Assignment (RFWA) problem [4,5]. This type of problem can be modelled as a multi-commodity network flow problem [6], where every commodity maps to a single source–destination pair of nodes in the network. When the problem's objective does not depend on e.g. the number of wavelength conversions used (if any), a simpler formulation called *source formulation* is possible [7]. Restoration from failures and installation of spare capacity in such networks have been studied extensively in the literature [8,9,11]. The main difference with our work is that in our work, a vertex or node in the transport network may have a Computational Resource attached to it. Due to the operational Grid scenarios studied, failure of this node also alters traffic demands between other node pairs in the network as an alternative workload scheduling strategy must be used. This new scheduling strategy again forms an optimization problem coupled with the network dimensions needed to handle this failure scenario.

In a distributed computing environment (e.g. Grids, clusters etc.), traffic demand between two nodes arises when computational load and any data processed by this load are transferred between the particular nodes. How this computational load is distributed among the participating nodes, is determined by the *scheduling strategy*. In Lambda Grids, this scheduling is concerned with at least two types of resources: computational resources and the interconnecting Lambda network. Combined resource allocation and scheduling for these two resource types has been demonstrated in G-Lambda [12], a Web services-based resource reservation architecture on top of a GMPLS network resource management system. One step further down the road a transparent, single system image of the Lambda Grid can be created using a Service Address Routing architecture in which the different scheduling and allocation mechanisms are made available as services offered in the network [13].

Our work focusses on analyzing the Lambda Grid's resource requirements from an off-line combined dimensioning and scheduling approach, and is independent of the actual architectural concepts providing the resource scheduling and allocation services.

An additional complexity that needs to be taken into account when dealing with the Grid scheduling problem is the occurrence of resource failures. While fault-tolerant scheduling algorithms have been extensively studied and benchmarked [14,15], in this work we focus on the incorporation of possible resource failures in an off-line dimensioning problem, rather than on the behaviour of on-line scheduling algorithms in the presence of resource failures. We focus on simultaneously solving the dimensioning and scheduling problem for the failure-free and resource failure Grid scenarios rather than focussing on protecting individual work units by replicating computation and communication.

A common mathematical framework used to describe the scheduling problem of a given set of jobs on computational elements is the Integer Linear Program formulation. This way, most scheduling problems can be seen as special instances of project scheduling problems [16–19].

Alternatively, the distributed computing platform can be treated as a queueing network. In [20,21], for instance, average values for metrics such as idle time and resource utilization are derived for a pure space-shared system by analyzing the system as a steady-state queueing system.

In our Grid model, resources are time-shared; moreover, resource co-allocations made by a single job are, in general, not independent: a data-processing job which gets only small CPU shares will output less data per time unit (thus use less network resources too) when compared to the same job running exclusively on the same CPU.

The cited Integer Linear Programs or queueing theory can become cumbersome when taking into account these complications or when dealing with large problem instances (which is not unlikely, given the nature of Grids).

Because of the complexities involved in obtaining analytical results for realistic Grids, a lot of authors have used simulations on Grid models to obtain quantitative results with regard to schedule quality and workload distribution [22–28].

However, a formal and scalable mathematical approach is possible, if it can be assumed that the total load carried by the jobs behaves like an arbitrarily divisible workload. This approach is central to the Divisible Load Theory (DLT) [29,30]. The use of DLT in a Grid environment, taking into account not only Computational Resources but also network parameters has been demonstrated in [31]. In that work, the network constraints enforced include a limited number of (TCP) connections per link and a fixed bandwidth per connection. Traffic is entirely composed of the workload itself, and does not include "external" data processed by the load.

This contrasts with the approach used in this paper, as we use Divisible Load Theory and a load-balancing scheduling algorithm to derive traffic demands between network nodes, applicable to problems with large numbers of data-processing jobs. We translate the network constraints laid out in [31] to physical constraints in an optical circuit-switched network supporting wavelength conversion, expressing the discrete wavelength granularity and wavelength continuity and conversion constraints. Once demands are known, the dimensioning problem can then be modelled by a source routing formulation.

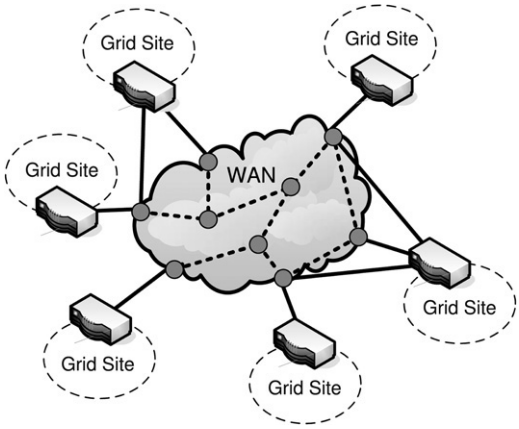

Fig. 1. Grid model: Sites and core transport network.

The suitability of the DLT-based approach as a Lambda Grid network dimensioning tool (without taking into account possible resource failures) has already been demonstrated in [32,33], and in this paper we show how this approach can be extended to include these possible resource failures.

3. Grid model and operational scenario

3.1. Resources

In this paper, we treat a Grid as a collection of different sites $\mathcal{R}$, connected through a transport network (Fig. 1).

The core network (which is to be dimensioned) is an optical circuit-switched transport network. It consists of core and access optical cross-connects (OXC) connected through directed links from the set $\mathcal{E}$. Each link $e \in \mathcal{E}$ contains optical fibers; each fiber can carry a (technology-dependent) number of wavelengths W, and each wavelength supports a (also technology-dependent) data rate B. All cross-connects have unlimited wavelength conversion capabilities.

Each Grid site $r \in \mathcal{R}$ connects to an access router of the optical network and offers several time-shared resources—a computational resource, a data storage resource and a data replica resource as illustrated in Fig. 2. The computational resource can process locally submitted as well as "foreign" jobs, and has a maximum computational capacity of P_r. It will only send locally generated jobs to a remote site if it cannot process or store that job locally. The data replica resource holds input datasets read by the jobs; the same datasets can be replicated among multiple of these resources in different Grid sites. The data storage resources hold output data generated by the jobs; it is assumed that they provide sufficient storage space for the jobs submitted at the resource's site.

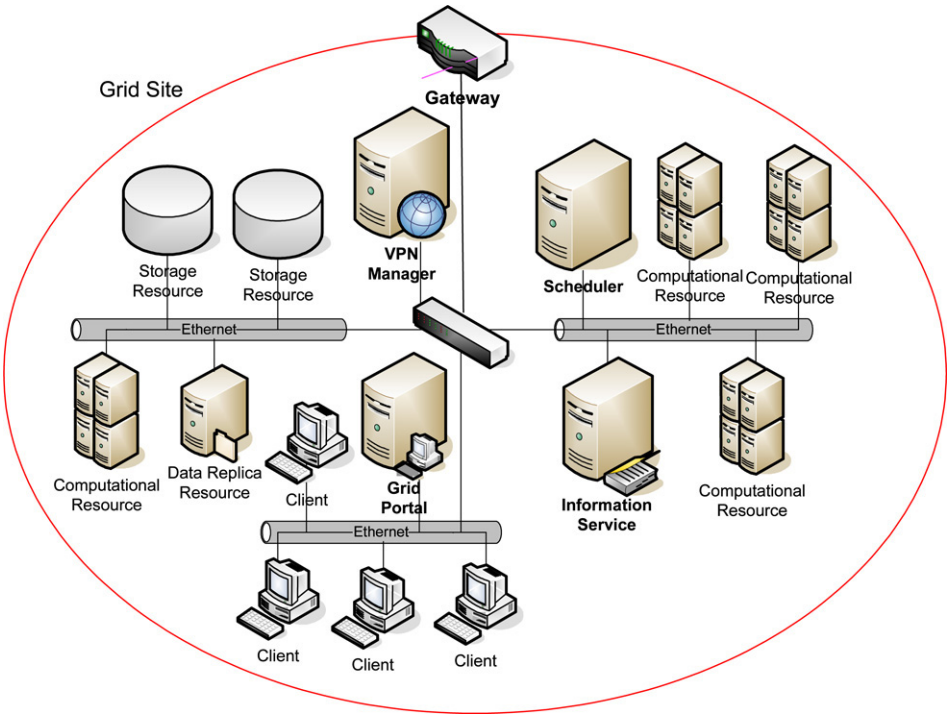

Fig. 2. Grid site model: Resources and gateway.

3.2. Jobs

At each site, users can submit jobs from a jobpool $\mathcal{J}$. The *home site* of a job $j \in \mathcal{J}$ is the site where it has been submitted. Jobs are indivisible work packets, characterized by their length t_j (i.e. processing time on a reference processor), the size of the input data d_j^I they process and the size of the output data d_j^O they generate. When looking at e.g. a periodic job load, the equivalent divisible workload is characterized by the average amount of workload arriving per unit of time α over this period and by the average data sizes processed and generated per unit of workloads, D_I and D_O, respectively.

It is assumed that all jobs read their input data from their home site and that they submit any output data to their home site as well, that is, only remotely processed jobs produce network traffic (between their processing site and their home site). Furthermore, jobs are assumed to process data at a constant rate throughout their lifetime; this way, remotely executed jobs can be treated as Constant Bit Rate (CBR) sources from a network point of view.

3.3. Grid excess load scenario

In our approach, the Grid's computational resources are first dimensioned to be able to deal with a specified steady-state load. Next, we assume that a single computational resource suffers from excessive (locally generated) load and that it needs to invoke remote computational resources. Under this assumption, load (which, in our model, is assumed to be arbitrarily divisible) is distributed from the top-tier site to lower-tier sites.

For our simulations, excess jobs have been generated with an average interarrival time of 0.5. The average computational load of this set of jobs is 30 units of work per second. We have chosen $D_I = D_O$ in such a way that the resulting average unidirectional excess load network demand is 220 GBps, which is roughly the order of magnitude of data that will be processed and transported in the EGEE Grid [34].

We consider the set (parameterized by some integer k) of load-balancing scheduling strategies where the excess load is uniformly distributed across k remote computational resources, a scenario not unlikely given that these remote resources may also be processing or storing local load. Again, we assume the Grid to converge into a steady-state (periodic with period T) mode of operation. For a given excess load instance per time period, we can decide which jobs are to be processed where (under the constraint of fair distribution across all remote resources), which determines the amounts of input and output data transferred per period between Grid sites.

Once traffic demands between each pair of Grid sites have been determined, solving the optical network dimensioning problem (for this single overloaded Grid site scenario) means deciding how lightpaths should be set up and routed in order to accommodate these demands with minimal cost. Here, only activation costs (fiber and wavelength) are taken into account.

The final network dimensions (i.e. number of installed fibers on each link and number of wavelengths activated on each fiber) are determined by the global optimum over all single-site overload problems.

4. Divisible load theory model

Using the concept of divisible load, the Grid dimensioning problem can be written down as an integer linear program (ILP) which does not suffer from scalability issues with increasing

Table 1
DLT Model: Overview of symbols used and their description

Symbol	Type	Description
D_I	Input Constant	Average amount of input data per unit of excess load
D_O	Input Constant	Average amount of output data per unit of excess load
B	Input Constant	Bandwidth of single wavelength
W	Input Constant	Maximum number of wavelengths available per fiber
α	Input Constant	Total excess load arriving per time unit
α_r^s	Cont. Var.	Excess load per time unit processed at r, top-tier site s
k	Input Constant	Number of remote sites processing load from top-tier site
δ_r^s	Boolean Var.	Equals 1 iff. r processes excess load from top-tier site s
d_{uv}^s	Integer Var.	Wavelength path demand between u and v (top-tier site s)
c_{eu}^s	Integer Var.	Number of wavelengths originating at u carried over link e (top-tier site s)
f_e^s	Integer Var.	Number of fibers needed on link e (top-tier site s)
f_e	Integer Var.	Ultimate number of fibers needed on link e

number of users and Grid jobs. The resulting ILP contains roughly two parts: one part describes how excess load generated at a single site is distributed among the remote sites; the other part describes how virtual wavelength paths need to be setup between Grid site pairs to support the transfer of excess load in this scenario.

In the absence of resilience considerations, the Grid dimensions follow from the following ILP (all symbols used have been conveniently summarized in Table 1).

Let α_r^s be the computational load per time unit transferred from the single top-tier site s to remote site r. Furthermore, the average amount of input data processed per unit of workload is given by D_I and the average amount of output data generated per unit of workload by jobs is D_O (see Section 3.3). Furthermore, let δ_r^s be a binary variable equalling 1 if and only if remote site r processes excess load from top-tier site s.

If excess load α is to be uniformly distributed among k remote sites, we have that

$$\sum_{r \in \mathcal{R} \setminus \{s\}} \delta_r^s = k \tag{1}$$

and

$$\forall r \in \mathcal{R} \setminus \{s\} \cdot \alpha_r^s = \frac{\delta_r^s \alpha}{k}. \tag{2}$$

Steady-state network demands (i.e. required number of wavelength paths for input and output data) between top-tier site s and remote site r then follow from

$$\forall r \in \mathcal{R} \setminus \{s\} \cdot d_{sr}^s \geq \frac{\alpha_r^s \cdot D_I}{B} \tag{3}$$

$$\forall r \in \mathcal{R} \setminus \{s\} \cdot d_{rs}^s \geq \frac{\alpha_r^s \cdot D_O}{B} \tag{4}$$

where B denotes the data rate supported by a single wavelength. If wavelength routing is written down using source formulation, we can use variables c_{er}^s to denote the number of wavelength paths originating at node r carried on link e in case the top-tier site is site s. Using these variables, network flow constraints can be expressed as

$$\forall u \in \mathcal{R}, \forall v \in \mathcal{R} \setminus \{u\} \cdot \sum_{e \in \mathcal{E}_v^+} c_{eu}^s + d_{uv}^s = \sum_{e \in \mathcal{E}_v^-} c_{eu}^s \tag{5}$$

$$\forall r \in \mathcal{R} \cdot \sum_{u \in \mathcal{R}} d_{ru}^s = \sum_{e \in \mathcal{E}_r^+} c_{er}^s. \tag{6}$$

In these equations, $\mathcal{E}_v^+$ denotes those links incident from node v and $\mathcal{E}_v^-$ denotes the links incident to node v.

The resulting global optical transport network cost is determined by the ultimate number of activated fibers and wavelengths in each scenario. In a single scenario featuring top-tier site s, the number of fibers needed on link e can be described by f_e^s which obeys

$$W \cdot f_e^s \geq \sum_{r \in \mathcal{R}} c_{er}^s \tag{7}$$

where W equals the maximal number of wavelengths that can be carried on a single fiber.

The global optimization problem decides on variable values for all values of s. The ultimate number of fibers f_e needed on link e is the maximal value of f_e^s over all scenarios (i.e. over all s), and is thus constrained by

$$\forall s \in \mathcal{R}, \forall e \in \mathcal{E}. f_e \geq f_e^s. \tag{8}$$

The average number of activated wavelengths (over all scenarios) is given by

$$\sum_{s \in \mathcal{R}} \sum_{e \in \mathcal{E}} \sum_{r \in \mathcal{R}} \frac{c_{er}^s}{|\mathcal{R}|}. \tag{9}$$

As a cost is associated with the number of activated fibers and wavelengths, the optimization goal is to minimize the cost as described in expression (10).

$$\min \sum_{s \in \mathcal{R}} \sum_{e \in \mathcal{E}} \sum_{r \in \mathcal{R}} \frac{c_{er}^s}{|\mathcal{R}|} + C \cdot \sum_{e \in \mathcal{E}} f_e. \tag{10}$$

In this last expression, C determines the relative weight assigned to the installed fibers over the average amount of activated wavelengths. In this paper, results presented have been obtained for reference values of $C = 1$, $B = 2.5$ Gbps and $W = 4$.

4.1. Single computational element failure

The initial set of base scenarios described in Section 4 only involved a single source node and assumed reliable Grid resources.

In this section, we extend our base scenarios by taking into account possible resource failures. In particular, a single scenario in this section is described by a (source, failure) pair, denoting the single source present in each scenario, and the single-resource failure (if any) occurring in this scenario.

We can discern 3 kinds of resource failures, corresponding to the most prominent resource types in our analysis:

- computational resource failures
- optical cross-connect failures
- network link failure, effectively cutting all fibers inside.

The effects of a computational resource failure are that the affected resource cannot handle any excess load from the overloaded source. The wavelength routing capabilities of the underlying optical cross-connect, however, are deemed to remain operational. In contrast, in case of an optical cross-connect failure, network traffic can no longer be routed over this cross-connect. This means the failure scenario can be treated as a base scenario where the affected cross-connect and all links incident to and from it have been left out of the original network. The last failure scenario only affects a single link.

Note that an optical cross-connect failure implies the unavailability of the connected computational resource and thus encompasses the computational resource failure scenario. In addition, the unavailability of an optical cross-connect implies that the links connected to it remain unused (it is assumed that no jobs are submitted to the OXC's corresponding computational resource, as there is effectively no way to route these jobs elsewhere).

Thus, the failure of an optical cross-connect can be regarded as the worst-case single-resource failure scenario. In our dimensioning, we provide shared spare capacity for all relevant worst-case scenarios. The relevant worst-case scenarios are those optical cross-connect failure scenarios that do not partition the interconnecting network. Thus, in case the interconnecting network features low connectivity, only a limited number of these failure scenarios are taken into account.

In this paper, when representing the cost of protecting the lambda Grid against different types of resource failures, we have limited ourselves to the additional network dimensioning cost related to the protection against the relevant (i.e. not partitioning the interconnecting optical network) optical cross-connect failures. We have compared this cost to the cost associated with the single source scenarios discussed previously.

In order to introduce the notion of a failing resource in the elementary scenario, the combined dimensioning and workload scheduling linear program needs to be adapted as follows.

4.2. Computational resource failure

The notion of a single failing computational resource at node n can easily be incorporated into the combined dimensioning and workload scheduling linear program by modifying Eq. (1) to read

$$\sum_{r \in \mathcal{R} \setminus \{s,n\}} \delta_r = k. \tag{11}$$

This effectively excludes any excess load generated at source node s from being scheduled on the now-defunct computational resource at node n.

4.3. Optical cross-connect failure

To model the failure of an optical cross-connect, note that such an optical cross-connect will be completely unused if all links incident to and from it are void of traffic. Therefore, we can model the failure of an optical cross-connect at node n by modifying network flow Eq. (5) to exclude these links from the flow conservation constraints as follows:

$$\forall u \in \mathcal{R} \setminus \{n\}, \forall v \in \mathcal{R} \setminus \{u, n\} \cdot \sum_{e \in \mathcal{E}_v^+ \setminus \mathcal{E}_n^-} c_{eu}^s + d_{uv}^s = \sum_{e \in \mathcal{E}_v^- \setminus \mathcal{E}_n^+} c_{eu}^s. \tag{12}$$

Indeed, as network cost increases when more fibers and wavelengths are activated and when we are dealing with a cost minimization problem, the net effect of the exclusion is that the affected cross-connect as well as any links to and from it will not be used in the observed scenario. After all, suppose that a solution to the network dimensioning problem is obtained in which an affected link is not void of traffic, we can immediately derive another valid solution which is cheaper.

Note that only those failure scenarios preserving the network's connectedness have been studied.

4.4. Link failure

In a similar way to the approach described in the previous section, we can model the failure of a single link l by modifying flow Eq. (5). Replacing that equation with

$$\forall u \in \mathcal{R} \setminus \{n\}, \forall v \in \mathcal{R} \setminus \{u, n\} \cdot \sum_{e \in \mathcal{E}_v^+ \setminus \{l\}} c_{eu}^s + d_{uv}^s = \sum_{e \in \mathcal{E}_v^- \setminus \{l\}} c_{eu}^s \tag{13}$$

ensures that traffic is routed over operational links only.

4.5. Parallelizing heuristic

Although our use of divisible workloads and source routing formulations have significantly reduced the complexity of the combined workload scheduling and dimensioning problem, our inclusion of resource failures has increased the number of simultaneous scenarios to be investigated (and thus, the number of variables and constraints) by a large number. Therefore, a solution method combining solutions to the individual elementary scenarios (rather than solving the global problem over all scenarios) is of interest. One such approach—which

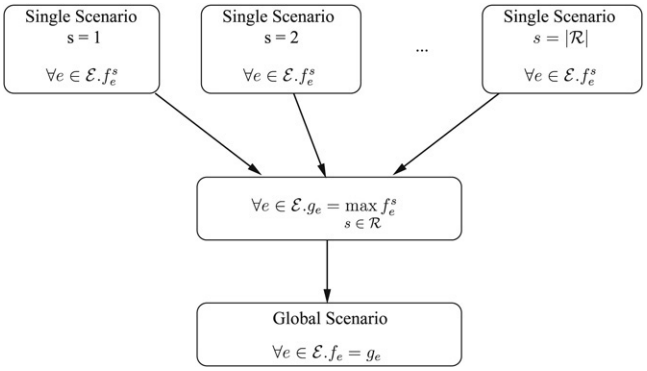

Fig. 3. Parallelizing heuristic: Overview.

we refer to as a parallelizing heuristic—is depicted in Fig. 3. This approach is able to return solutions within reasonable calculation time and resource limits, however at increased network dimensioning cost.

As illustrated in the figure, we start by solving all individual scenarios independently. This step can be performed in parallel, and results in a series of fiber counts on each edge (variables f_e^s). These values are used to initialize the parameters g_e:

$$\forall e \in \mathcal{E}.g_e = \max_{s \in \mathcal{R}} f_e^s,$$

and then we proceed by solving the problem as defined in Section 4, but replacing constraints (8) by:

$$\forall e \in \mathcal{E} \cdot f_e = g_e. \tag{14}$$

4.6. Incremental heuristic

The parallelizing heuristic presented in Section 4.5 performs a simple maximization over the solutions to a set of independently solved problems, instead of solving a single problem tackling all of these problems simultaneously.

Another heuristic method to solve this complex global problem is shown in Fig. 4. This *incremental* approach solves the dimensioning problem for a set of single excess load scenarios as follows. First, the elementary scenarios are ordered. The heuristic then solves the Grid network dimensioning problem for the first elemental scenario in the ordered list and saves the resulting network dimensions, in particular the number of installed fibers on each edge and the number of activated wavelengths on each fiber in the scenario at hand.

Next, the heuristic solves a modified single scenario Grid dimensioning problem. The single scenario is the second scenario in the ordered list, while the modifications encompass the inclusion of the solution to the first dimensioning problem as additional constraints. Thus, in this second phase, the problem solved is to dimension the lambda Grid with minimal network cost, given that two scenarios need to be supported and that workload distribution and network routing for the first scenario have already been decided upon. The modifications thus add constraints to a standard single scenario dimensioning problem, but do not add additional variables.

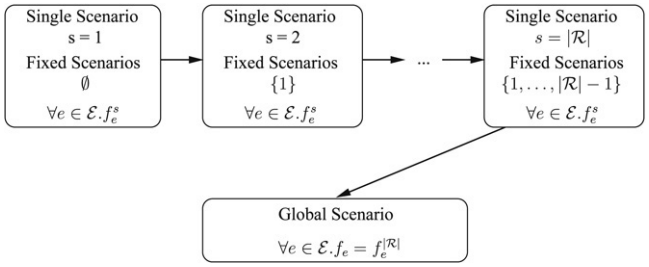

Fig. 4. Incremental heuristic: Overview.

This process is repeated; the number of iterations needed is the number of scenarios that needs to be supported. At iteration n, a modified single scenario dimensioning problem is solved for the nth scenario in the ordered list, given the solution to the previous $n-1$ scenarios.

When the last iteration is finished, we have obtained a solution to the global dimensioning problem based upon this particular ordering of the elemental scenarios. By repeating this whole process for different scenario orderings, we can select the ordering with the lowest resulting lambda Grid dimensioning cost for the collection of all single source scenarios.

5. Analytical bounds

In this section, we derive analytical bounds for the Lambda Grid dimensioning cost (defined in Eq. (10) and consisting of wavelength path and fiber components) in case a regular network topology with N nodes is used. Throughout this section, we assume that shortest-path routing is used, that excess workload (featuring perfect I/O symmetry) is distributed uniformly over all operational remote sites and that wavelength granularity is fixed at 2.5 Gbps.

For a regular topology like the bidirectional ring, it is possible to derive analytical results with regard to the expected additional costs components (both wavelength path and fiber costs) as follows.

Assume that, in the absence of resource failures, the wavelength demand from the single top-tier site to each remote site is given by λ_1, and that in this scenario none of these λ_1 wavelengths has residual capacity left. If the excess source node is node 0 and the OXC at node $N-1$ fails, the link carrying the bulk of the traffic is situated between nodes 0 and 1, as all working traffic for nodes $1, \ldots, \lfloor \frac{N}{2} \rfloor$ as well as traffic routed on the backup path for nodes $\lfloor \frac{N}{2} \rfloor + 1, \ldots, N-2$ is routed over this link.

In order to support all node failure scenarios in such a double ring, each of the $2N$ directed links needs to be provisioned with at least an amount of fibers equal to

$$\left\lceil \frac{\left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right)(N-2)}{W} \right\rceil \tag{15}$$

whereas in the single-site excess load scenario (without taking into account possible OXC failures) this number is only

$$\left\lceil \frac{\lambda_1 \left\lceil \frac{N-1}{2} \right\rceil}{W} \right\rceil. \tag{16}$$

In the envisioned double ring failure scenario, the number of wavelength paths (originating from node 0, with node F failing) carried on the network links equals

$$2\sum_{k=1}^{F-1} k \left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right) + 2\sum_{k=F+1}^{N-1} (N-k) \left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right). \tag{17}$$

Thus, averaged over all scenarios we obtain

$$\frac{\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil}{N-1} \sum_{F=1}^{N-1} \left(2\sum_{k=1}^{F-1} k + 2\sum_{k=F+1}^{N-1} (N-k)\right) \tag{18}$$

which eventually reduces to

$$\frac{2N(N-2)\left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right)}{3}. \tag{19}$$

For the base scenarios, this average number of wavelength paths carried on the links is given by

$$2\sum_{k=1}^{\left\lfloor \frac{N}{2} \right\rfloor} k\lambda_1 + 2\sum_{k=\left\lfloor \frac{N}{2} \right\rfloor + 1}^{N-1} \left(k - \left\lfloor \frac{N}{2} \right\rfloor\right)\lambda_1 \tag{20}$$

which in turn can be simplified to

$$2\lambda_1 \left(\left\lfloor \frac{N}{2} \right\rfloor\right)^2. \tag{21}$$

For a double ring topology, the cost increase for the OXC failure protection (compared to the base scenario) is therefore

$$\frac{\frac{2N(N-2)\left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right)}{3} + 2N \left\lceil \frac{\left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right)(N-2)}{W} \right\rceil}{2\lambda_1 \left(\left\lfloor \frac{N}{2} \right\rfloor\right)^2 + 2N \left\lceil \frac{\lambda_1 \left\lceil \frac{N-1}{2} \right\rceil}{W} \right\rceil}. \tag{22}$$

For a full-mesh topology in which shortest-path routing is enforced, the average number of wavelength paths on the links in over all OXC failure scenarios changes to

$$2(N-2)\left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right) \tag{23}$$

from

$$2(N-1)\lambda_1. \tag{24}$$

The total number of fibers needed over all scenarios is then given by

$$N(N-1) \left\lceil \frac{\left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right)}{W} \right\rceil \tag{25}$$

and is simply given by

$$N(N-1) \left\lceil \frac{\lambda_1}{W} \right\rceil \tag{26}$$

in the absence of OXC failures.

Under our assumptions (i.e. shortest-path routing, $D_I = D_O$, workload as in Section 3.3), the relative cost increase due to OXC failure protection for full-mesh interconnection networks is thus given by

$$\frac{2(N-2)\left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right) + N(N-1) \left\lceil \frac{\left(\lambda_1 + \left\lceil \frac{\lambda_1}{N-2} \right\rceil\right)}{W} \right\rceil}{2(N-1)\lambda_1 + N(N-1) \left\lceil \frac{\lambda_1}{W} \right\rceil}. \tag{27}$$

6. Evaluation

In this section, we study the increased dimension (defined in Eq. (10) and consisting of wavelength path and fiber components) incurred by considering the possible optical cross-connect failure scenarios and compare it to the dimensioning cost of the base scenario featuring a single source and no failures. We have performed the Lambda Grid dimensioning for the failure scenarios for different parameter sets including the Grid's scheduling strategy, the wavelength granularity and job I/O asymmetry.

6.1. Setup

The optical transport network topologies considered in this study have been inspired by the European network (similar to the "basic network" reference network from the European COST 266 project [10]) shown in Fig. 5. As each OXC is located in a major European city, it is conceivable that each such cross-connect has a Grid site attached to it. We therefore assume that each such OXC actually doubles as a Grid site (thus, we make an abstraction of any access network in place), so that our Grid has as many cross-connects as Grid sites.

For this paper, we have generated sets of connected networks (with number of nodes equal to the number of nodes in the reference network) for varying random-link probabilities p. These networks were obtained through repeated addition of node-link pairs and then by adding extra links with probability p. Using this method, the European reference network is similar to the networks obtained for $p = 0.1$. For each value of p (except for $p = 1$, denoting a full-mesh network), ten topologies have been generated.

The excess load generated in each scenario is the one described in Section 3.3. For our reference parameter settings, jobs feature equal-sized inputs and outputs and 2.5 Gbps wavelengths are used while the excess load is distributed over all operational remote sites.

As we consider all relevant single-failure scenarios (leaving the Grid in a connected state), it follows that the number of possible scenarios greatly increases when compared to the number of scenarios in the absence of possible resource

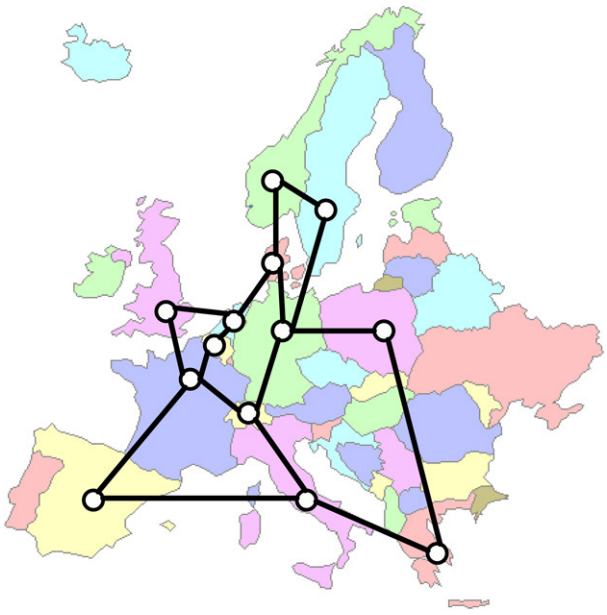

Fig. 5. Reference Grid Topology: European Core Network (13 nodes, 17 bidirectional links).

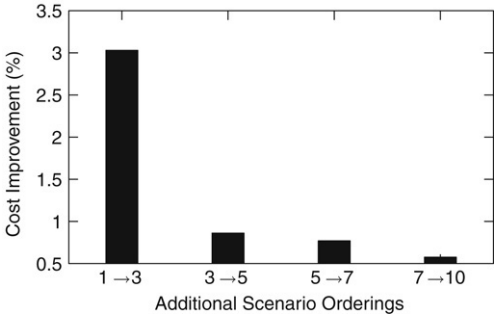

Fig. 6. Incremental heuristic: Sensitivity to number of investigated scenario orderings.

failures. Because this increased number of scenarios leads to significantly longer calculation times, from this point on, all dimensioning costs (which represent the dimensioning cost allowing the Grid to support every scenario taken into account) in this section have been obtained using the incremental heuristic described in Section 4.6.

As this heuristic evaluates the scenarios sequentially, we have repeated each heuristic run 10 times (for different scenario orderings) in order to reduce the resulting solution's sensitivity to scenario reordering. This property is demonstrated by the numbers shown in Fig. 6 which represent sequential improvements in network dimensioning cost when dimensioning our set of random networks (with $p = 0.1$) for the optical cross-connect failure scenarios using the incremental heuristic.

6.2. Results

The cost increase for the reference case on our set of random networks due to OXC failure protection has been plotted in Fig. 7.

Using the workload as described in Section 3.3 and the number of nodes in the networks we study ($N = 13$), we

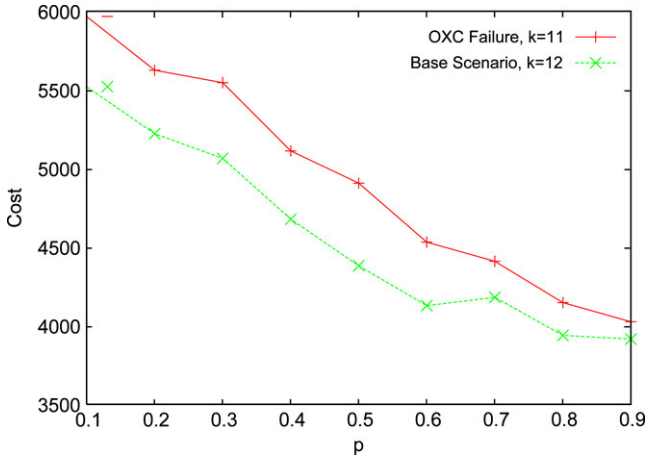

Fig. 7. OXC failure protection cost increase for random networks.

can now evaluate the analytical bounds derived in Section 5. Assuming shortest-path routing, 2.5 Gbps wavelengths and equal-sized inputs and outputs, the cost increase for a double ring topology as given by Eq. (22) then equals 1.60.

For the full-mesh topology, under the same assumptions, Eq. (27) yields 1.06. The relationship of these numbers to the numbers shown in Fig. 7 is as follows: from Fig. 7, it follows that in our reference case the cost increase incurred by providing OXC failure resilience to the base scenarios is no more than 10% for our set of random networks.

The figures obtained for the double ring topology (1.60) are much higher because that particular topology features a failure scenario in which all traffic is rerouted over a single network link (regardless of the excess load source under observation), and in the above formulas it has been assumed that network traffic in the base scenario fills exactly λ_1 wavelength paths.

For full-mesh topologies ($p = 1$), note how the value obtained analytically (1.06) corresponds closely to the cost difference shown in Fig. 7 for highly connected networks ($p = 0.9$).

6.2.1. Job I/O Asymmetry

While in the reference case jobs are assumed to produce as much output data as they need input data ($D_I = D_O$), Fig. 8 shows the average dimensioning costs for OXC failure resilient lambda Grids for I/O asymmetric jobs. Results are obtained on our set of random networks corresponding to $p = 0.1$.

Because of input/output symmetry in our global dimensioning problem featuring input and output datasets of equal size, we expect these results to show symmetry around $s = \frac{D_I}{D_O} = 1$. In addition, due to the optimization over all individual scenarios, the chosen network dimensions are actually determined by $\max(D_I, D_O)$. This is also an indicator that minimal cost is expected for $s = 1$. Clearly, the figure confirms these expectations. For all asymmetry factors studied, the additional dimensioning cost in case OXC failure protection is incorporated does not exceed 10%. The cost increase is maximal around the point where input and output data are of equal size (i.e. $D_I = D_O$). In this case, network dimensions are determined by $\max(D_I, D_O)$ which is minimal when $D_I = D_O$ (if $D_I + D_O$ is constant). As

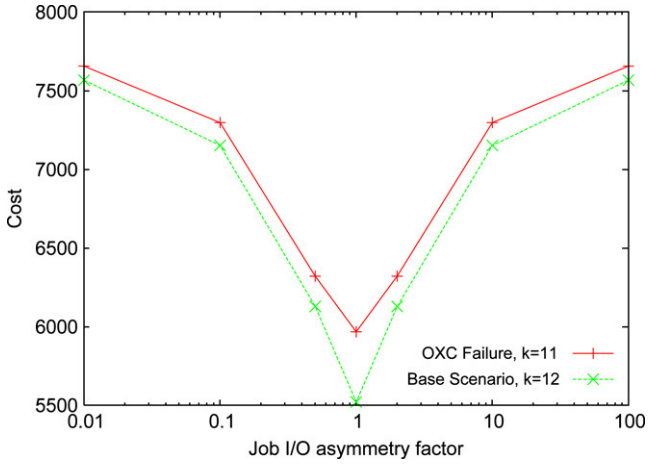

Fig. 8. Traffic asymmetry: OXC failure protection cost for random networks ($p = 0.1$).

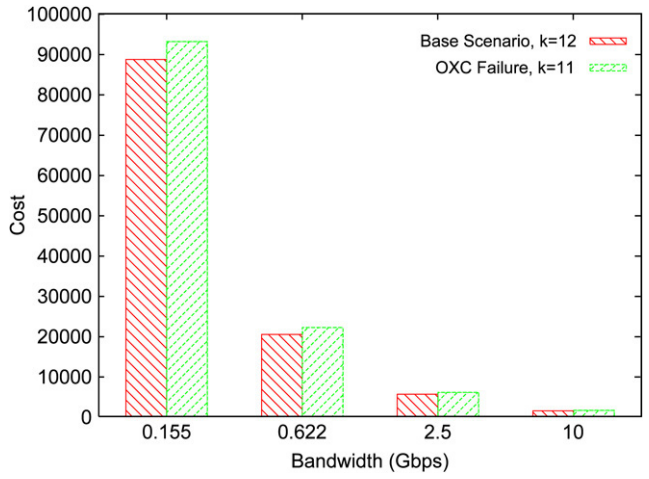

Fig. 9. Wavelength granularity: OXC failure protection cost for random networks ($p = 0.1$).

taking into account possible OXC failures needs extra network capacity when compared to the base scenario, the network capacity needed to support possible OXC failures will differ most from the base scenario cost in case $D_I = D_O$ as in this case there is minimal room for wavelength and fiber re-use (and thus, the most extra capacity needs to be installed when $D_I = D_O$).

6.2.2. *Wavelength granularity*

We have calculated the resulting network dimensioning costs for different wavelength granularities (155 Mbps, 622 Mbps, 2.5 Gbps and 10 Gbps) while limiting the number of wavelengths per fiber in such a way that total fiber capacity remains fixed at 10 Gbps. We have used a linear fiber/wavelength cost model. For all wavelength granularities examined, the dimensioning cost for the OXC failure resilient lambda Grid (see Fig. 9) does not exceed the base scenario dimensioning cost by more than 10%.

6.2.3. *Scheduling strategies*

In the previous sections, excess workload in each scenario was distributed among all remote Grid sites. Fig. 10 compares

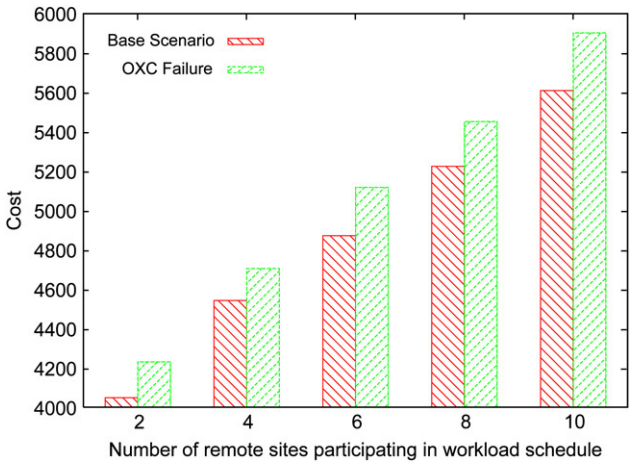

Fig. 10. Scheduling strategy: OXC failure protection cost for random networks ($p = 0.1$).

the resulting dimensioning cost for the base scenarios and the OXC failure resilient lambda Grid for varying numbers of remote sites participating in the excess workload absorption, and we see comparable increases in dimensioning cost for all numbers of active remote sites participating in the excess workload schedule.

7. Conclusions

In this paper, we have studied the dimensioning problem of an optical circuit-switched transport network for Grid applications. Our main operational scenario of concern is that when excess load is generated at a single site, remote sites are then needed to process this excess load. We discussed the need to solve the combined dimensioning and workload scheduling problem, and argued why the use of Divisible Load Theory can help us to model this problem in a scalable fashion. As the inclusion of possible resource failures greatly increases the number of scenarios to be studied, the need for scalable modelling techniques becomes even more pressing.

In order to cope with a global optimization problem dealing with all operational scenarios of interest, we have proposed additional simplifications to solve the problem. These simplifications consist of a parallelizing heuristic and an incremental heuristic, both attempting to solve the resulting linear programs in a more timely fashion.

For the topologies and scenarios studied in this paper, the additional Lambda Grid dimensioning cost incurred by explicitly incorporating possible optical cross-connect failures in the DLT-based dimensioning problem remained below 10% when compared to the dimensioning cost of our base problem.

We validated these conclusions for a wide range of parameter variations, most notably network topology (through variation in average link probability), wavelength granularity and cost model, changes in traffic demand (a) symmetry and Grid scheduling policy.

We can conclude that our approach is of practical use for selecting and dimensioning a suitable OCS Grid interconnection topology, including selection of optimal wavelength granularity in the presence of possible resource failures.

References

[1] L. Smarr, A. Chien, T. DeFanti, J. Leigh, P. Papadopoulos, The optiputer, Communications of the ACM 46 (2003) 58–67.

[2] N. Taesombut, X. Wu, A. Chien, A. Nuyak, B. Smith, D. Kilb, T. Im, D. Samilo, G. Kent, J. Orcutt, Collaborative data visualization for Earth Sciences with the OptIPuter, Future Generation Computer Systems 22 (2006) 955–963.

[3] T. DeFanti, C. de Laat, J. Mambretti, K. Neggers, B. Arnaud, Translight: A global-scale lambdagrid for e-science, Communications of the ACM 46 (2003) 34–41.

[4] N. Wauters, P. Demeester, Design of the optical path layer in multiwavelength cross-connected networks, IEEE Journal on Selected Areas in Communications 14 (1996) 881–892.

[5] D. Banerjee, B. Mukherjee, Wavelength-routed optical networks: Linear formulation, resource budgeting tradeoffs, and a reconfiguration study, IEEE/ACM Transactions on Networking 8 (2000) 598–607.

[6] D. Coudert, H. Rivano, Lightpath assignment for multifibers wdm networks with wavelength translators, in: Proceedings of IEEE Globecom'02, vol. 3, 2002, pp. 2686–90.

[7] M. Tornatore, G. Maier, A. Pattavina, Wdm network optimization by ilp based on source formulation, in: Proceedings of IEEE Infocom'02, vol. 3, 2002, pp. 1813–21.

[8] B. Van Caenegem, W. Van Parys, F. De Turck, P. Demeester, Dimensioning of survivable wdm networks, IEEE Journal on Selected Areas in Communications 16 (1998) 1146–1157.

[9] W. Grover, J.D. amd, M. Clouqueur, D. Leung, D. Stamatelakis, New options and insights for survivable transport networks, IEEE Communications Magazine 40 (2002) 34–41.

[10] Cost 266 - Main Page, http://www.ufe.cz/dpt240/cost266/.

[11] T.-H. Wu, Fiber Network Service Survivability, Artech House, 1994.

[12] A. Takefusa, M. Hayashi, N. Nagatsu, H. Nakada, T. Kudoh, T. Miyamoto, T. Otani, H. Tanaka, M. Suzuki, T. Sameshima, W. Imajuku, M. Jinno, Y. Takigawa, S. Okamoto, Y. Tanaka, S. Sekiguchi, G-Lambda: Coordination of a Grid scheduler and lambda path service over GMPLS, Future Generation Computer Systems 22 (2006) 868–875.

[13] I. Schersom, D. Valencia, E. Cauich, J. Duselis, R. Wang, Federated Grid clusters using service address routed optical networks, Future Generation Computer Systems 23 (2007) 957–967.

[14] B. Kalyanasundaram, K. Pruhs, Fault-tolerant scheduling, in: ACM Symposium on Theory of Computation, 1994, pp. 115–124.

[15] G. Wrzesinska, R.V. van Nieuwport, J. Maassen, T. Kielmann, H.E. Bal, Fault-tolerant scheduling of fine-grained tasks in grid environments, International Journal of High Performance Applications 20 (1) (2006) 103–114.

[16] R. Kolisch, R. Padman, An integrated survey of project scheduling, OMEGA International Journal of Management Science 29 (3) (2001) 249–272.

[17] J. Sgall, On-line scheduling - a survey, in: Lecture Notes in Computer Science, vol. 1442, 1998, pp. 196–231.

[18] L. Hall, A. Schulz, D. Shmoys, J. Wein, Scheduling to minimize average completion time: Off-line and on-line algorithms, in: SODA: ACM-SIAM Symposium on Discrete Algorithms (Conference on Theoretical and Experimental Analysis of Discrete Algorithms), 1996.

[19] D.G. Feitelson, L. Rudolph, U. Schwiegelshohn, K.C. Sevcik, P. Wong, Theory and practice in parallel job scheduling, in: D.G. Feitelson, L. Rudolph (Eds.), Job Scheduling Strategies for Parallel Processing, Springer Verlag, 1997, pp. 1–34.

[20] A. Bucur, D. Epema, An evaluation of processor co-allocation for different system configurations and job structures, in: Proceedings of SBAC-PAD, 2002.

[21] A. Bucur, D. Epema, The influence of the structure and sizes of jobs on the performance of co-allocation, in: Proceedings of JSSPP6, 2000.

[22] R. Buyya, M. Murshed, Gridsim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing, The Journal of Concurrency and Computation: Practice and Experience (CCPE) (2002).

[23] A. Legrand, L. Marchal, H. Casanova, Scheduling distributed applications: the simgrid simulation framework, in: CCGRID'03: Proceedings of the 3st International Symposium on Cluster Computing and the Grid, 2003.

[24] P. Thysebaert, B. Volckaert, F. De Turck, B. Dhoedt, P. Demeester, Evaluation of grid scheduling strategies through nsgrid: a network-aware grid simulator, in: Grid Computing, Neural, Parallel & Scientific Computations 12 (2004) 353–378 (special issue).

[25] I.F.K. Ranganathan, Simulation studies of computation and data scheduling algorithms for data grids, Journal of Grid Computing 1 (2003) 53–62.

[26] W.H. Bell, D.G. Cameron, L. Capozza, A.P. Millar, K. Stockinger, F. Zini, Simulation of dynamic grid replication strategies in optorsim, in: GRID '02: Proceedings of the Third International Workshop on Grid Computing, 2002, pp. 46–57.

[27] D.G. Cameron, R. Carvajal-Schiaffino, A.P. Millar, C. Nicholson, K. Stockinger, F. Zini, Evaluating scheduling and replica optimisation strategies in optorsim, in: 4th International Workshop on Grid Computing, Grid2003, 2003.

[28] E. Caron, V. Garonne, A. Tsaregorodtsev, Definition, modelling and simulation of a grid computing scheduling system for high throughput computing, Future Generation Computer Systems 23 (2007) 968–976.

[29] D. Yu, T. Robertazzi, Divisible load scheduling for grid computing, in: Proceedings of the IASTED 2003 International Conference on Parallel and Distributed Computing and Systems, PDCS, 2003.

[30] J. Hung, H. Kim, T. Robertazzi, Scalable scheduling in parallel processors, in: Proceedings of the 36th Annual Conference on Information Sciences and Systems, CISS'02, 2002.

[31] L. Marchal, Y. Yang, H. Casanova, Y. Robert, A realistic network/application model for scheduling divisible loads on large-scale platforms, Rapport de recherche de l'INRIA-Rhone-Alpes (RR-5197), 2004.

[32] P. Thysebaert, F. De Turck, B. Dhoedt, P. Demeester, Using divisible load theory to dimension optical transport networks for computational grids, in: Proceedings of OFC/NFOEC, 2005.

[33] P. Thysebaert, M. De Leenheer, B. Volckaert, F. De Turck, B. Dhoedt, P. Demeester, Scalable dimensioning of optical transport networks for grid excess load handling, published in Photonic Network Communications 12(2), pp. 117–132.

[34] Enabling Grids for E-SciencE, http://www.eu-egee.org/.

Pieter Thysebaert received his M.Sc. degree in Computer Science Engineering from Ghent University, Belgium, in June 2001. He is now a research assistant and Ph.D. student affiliated to the Department of Information Technology at Ghent University and has received a scholarship from the FWO (Fund for Scientific Research—Flanders). His main interests include Grid simulation and modelling of Grid scheduling problems.

Marc De Leenheer received his M.Sc. degree in Computer Science Engineering from Ghent University, Belgium, in June 2003. He is now a research assistant and Ph.D. student affiliated to the Department of Information Technology at Ghent University and has received a scholarship from the IWT (Institute for Innovation in Science and Technology—Flanders). His main interests include modelling and optimization of Grid management architectures, specifically in the context of photonic networks.

Bruno Volckaert received his M.Sc. degree in Computer Science from Ghent University, Belgium, in June 2001. He is now a research assistant and Ph.D. student affiliated to the Department of Information Technology at Ghent University and has received a scholarship from the IWT (Institute for Innovation in Science and Technology—Flanders). His main interests include Grid management architectural designs and Grid simulation.

Filip De Turck received his M.Sc. degree in Electrical Engineering from Ghent University, Belgium, in June 1997. In May 2002, he obtained the Ph.D. degree in Electrical Engineering from the same university. From October 1997 to September 2001, he was research assistant at the Fund for Scientific Research—Flanders, Belgium (FWO-V.). At the moment, he is a part-time professor and a post-doctoral fellow of the FWO-V., affiliated with the Department of Information Technology at Ghent University. Filip De Turck is the author or co-author of approximately 80 papers published in international journals or in the proceedings of international conferences. His main research interests include scalable software architectures for telecommunication networks, service management, performance evaluation and optimization of routing, admission control and traffic management in telecommunication systems.

Bart Dhoedt received a degree in Engineering from Ghent University, Belgium, in 1990. In September 1990, he joined the Department of Information Technology at the same university. His research, addressing the use of micro-optics to realize parallel free space optical interconnects, resulted in a Ph.D. degree in 1995. After 2 years of post-doctoral research in opto-electronics, he became professor at the Faculty of Engineering, Department of Information Technology. Since then, he has been responsible for several courses on algorithms, programming and software development. His research interests are software engineering and mobile and wireless communications. Bart Dhoedt is the author or co-author of approximately 100 papers published in international journals or in the proceedings of international conferences. His current research addresses software technologies for communication networks, peer-to-peer networks, mobile networks and active networks.

Piet Demeester received the M.Sc. degree in Electrical Engineering and the Ph.D. degree from Ghent University, Belgium, in 1984 and 1988, respectively. In 1992 he started a new research activity on broadband communication networks resulting in the IBCN group (INTEC Broadband Communications Network research group). He became professor at Ghent University in 1993 and is responsible for the research and education on communication networks. The research activities cover various communication networks (IP, ATM, SDH, WDM, access, active, mobile), including network planning, network and service management, telecom software, internetworking, network protocols for QoS support, etc. Piet Demeester is the author of more than 400 publications in the area of network design, optimization and management. He is a member of the editorial board of several international journals and has been a member of several technical program committees (ECOC, OFC, DRCN, ICCCN, IZS).