Photon Netw Commun (2006) 12:117–132
DOI 10.1007/s11107-006-0026-1

ORIGINAL ARTICLE

Scalable dimensioning of optical transport networks for grid excess load handling

Pieter Thysebaert · Marc De Leenheer · Bruno Volckaert · Filip De Turck · Bart Dhoedt · Piet Demeester

Received: 10 May 2005 / Revised: 8 February 2006 / Accepted: 8 February 2006 / Published online: 9 September 2006

Abstract Grids consist of the aggregation of numerous dispersed computational and storage resources, able to satisfy even the most demanding computing jobs. 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. Due to the data-intensive nature of Grid jobs, it is expected that optical transport networks will play an important role in Grid deployment. As Grids possibly consist of high numbers of resources, and users, solving the network dimensioning problem (i.e. determining the number of wavelength channels per fiber and wavelength granularity required) using straightforward Integer Linear Programs (ILP) does not scale well with increasing number of jobs. Therefore, we propose the use of Divisible Load Theory (DLT) when modeling the OCS (with wavelength translation) dimensioning problem in this context. We compare this approach to both an exact ILP and heuristic (derived from the exact ILP) approach as a function of the job arrival process, network related parameters and the Grid job scheduling strategy on the Grid. Results show the convergence of the DLT-based and the exact ILP approach, which indicates that the DLT-based approach is of practical use in cases where the exact ILP-based problem becomes intractable. We study an excess load scenario and evaluate the network cost for varying wavelength granularity, fiber/wavelength cost models, network topology and traffic demand asymmetry under multiple Grid scheduling strategies. Results indicate the suitability of our DLT-based approach as an Optical Transport Network dimensioning tool to be used by network operators.

Keywords Multifiber optical networks · Dimensioning · Grid computing · Divisible load theory · Integer linear programming

P. Thysebaert (✉) · M. De Leenheer · B. Vockaert · F. De Turck · B. Dhoedt · P. Demeester
Department of Information Technology,
Ghent University – IBBT – IMEC,
Gaston Crommenlaan 8, 9050 Gent, Belgium
e-mail: pieter.thysebaert@intec.UGent.be

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

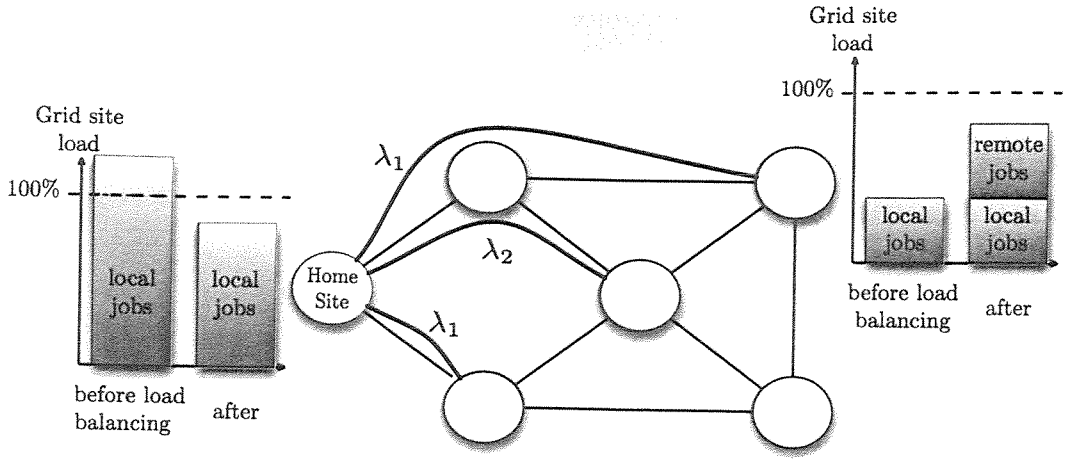


Fig. 1 Example load balancing scenario

by the recent increase in research activities into these "supernetworks" [1, 2].

If a Grid's aggregate power is to be successfully exploited, an important problem to be solved is to determine how the expected workload generated at each location should be distributed over the Grid's resources if it cannot be handled locally (Fig. 1). This is usually denoted as a *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.

In this paper, we address the dimensioning of an optical transport network connecting sites with already installed processing and storage capacity. 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 to address the excess load.

We propose to solve this dimensioning problem using the technique of Divisible Load Theory (DLT), which in fact yields an approximation. We compare the results from this solution method to the results obtained by modeling the problem as a classical Integer Linear Program (ILP), and to a heuristic derived from this ILP. These results reveal that our DLT-based method is an efficient and scalable technique that is able to find near-optimal solutions in cases where both the ILP and the derived heuristic become computationally intractable.

The remainder of the paper is structured as follows: Section Related work gives an overview of existing mathematical models and solution techniques to similar or related workload scheduling and dimensioning problems. Section Grid model and operational scenario formally defines the problem under study. In Section Dimensioning algorithms, an accurate mathematical model extending one of the techniques from Section Related work for the problem (optical network dimensioning in a Grid context) at hand is provided. Several alternatives are shown to reduce the model's computational complexity, one of which is our proposed DLT-based formulation. Results and discussions are presented in fifth section, and, after listing possible future extensions and applications in sixth section, the main conclusions are presented in the final section.

Related work

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 [3, 4].

This type of problem can be modeled as a multi-commodity network flow problem [5], 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 [6].

In a distributed computing environment (e.g. Grids, clusters, ...), 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*.

A common mathematical framework used to describe the scheduling problem of a given set of jobs on computational elements is the ILP formulation. This way, most scheduling problems can be seen as special instances of project scheduling problems [7–10].

Alternatively, the distributed computing platform can be treated as a queueing network. In [11, 12], 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.

When these complications need to be taken into account, or when large problems need to be tackled (which is not unlikely, given the nature of Grids) the use of the cited ILP or queueing theory can become cumbersome.

Due to 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 [13–17].

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 DLT [18, 19]. The use of DLT in a Grid environment, taking into account not only Computational Resources but also network parameters, has been demonstrated in [20]. 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 DLT and a load-balancing scheduling algorithm to derive traffic demands between network nodes, applicable to problems with large numbers of data-processing jobs. We readily map the network constraints laid out in [20] to physical constraints in an optical circuit-switched network supporting wavelength conversion. Once demands are known, the dimensioning problem can then be modelled by a *source routing* formulation. As real workloads are not arbitrarily divisible, we compare the DLT-derived network dimensions with an ILP approach based on an exact description of the job set where this latter approach is computationally tractable. Where it is not, the DLT results are compared to computationally tractable heuristics based on the ILP approach. The suitability of the DLT-based approach as a network dimensioning tool has been touched briefly in [21].

Grid model and operational scenario

Resources

We treat a Grid as a collection of different sites $\mathcal{R}$, connected through a transport network. 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 two time-shared resources—a Computational Resource and a Data Storage Resource. 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 Storage Resource holds input and output data for jobs; it is assumed they provide sufficient storage space for the jobs submitted at the resource's site.

Jobs

At each site, users can submit jobs from a job pool $\mathcal{J}$. 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. 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.

Excess load scenario

In our approach, 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.

We consider the set (parameterized by some integer k) of load-balancing scheduling strategies where the excess load is evenly 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 mode of operation (periodic with period T). 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.

Dimensioning algorithms

In the following, we present several approaches to the optical grid dimensioning problem. More specifically, Section Exact work load ILP introduces an ILP formulation for the exact workload, which is reformulated in a parallelized form in the following section. The complexity of the exact workload ILP model is reduced for large job counts in Section Large job count heuristic. Finally, the DLT-based approach is given in Section DLT. Each approach results in a Linear Program, which is solved as described in Section ILP solver. The common objective of all linear programs is the minimization of the network cost, which is defined as the weighted sum of the number of activated fibers and the number of used wavelength channels.

Exact workload ILP

Single scenario formulation

The following ILP formulation allows us to optimally dimension the optical network for a single overloaded resource S. Suppose the excess load of this resource is given explicitly by a set of jobs $\mathcal{J}^S$. We introduce the following integer variables (Fig. 2):

- f_e^S = number of fibers on edge e,
- c_{er}^S = number of wavelengths originating from resource r carried by edge e, with $0 \le c_{er}^S \le \sum_{j\in\mathcal{J}^S} \frac{d_j^{\mathrm{I}}}{B\cdot T}$ for $r = S$ and $0 \le c_{er}^S \le \sum_{j\in\mathcal{J}^S} \frac{d_j^{\mathrm{O}}}{B\cdot T}$ for $r \in \mathcal{R}\backslash\{S\}$,
- $y_{jr}^S = 1$ iff job j is executed on resource r, 0 otherwise,
- d_{uv}^S = demand (number of required wavelengths) between resources u and v, with $0 \le d_{uv}^S \le \sum_{j\in\mathcal{J}^S} \frac{d_j^{\mathrm{I}}}{B\cdot T}$ for $u = S$ and $v \in \mathcal{R}\backslash\{S\}$, $0 \le d_{uv}^S \le \sum_{j\in\mathcal{J}^S} \frac{d_j^{\mathrm{O}}}{B\cdot T}$ for $u \in \mathcal{R}\backslash\{S\}$ and $v = S$, $d_{uv}^S = 0$ otherwise.

The first set of constraints ensures that all excess jobs are remotely executed, while protecting each resource from being overloaded:

$$\forall j \in \mathcal{J}^S \cdot \sum_{r\in\mathcal{R}} y_{jr}^S = 1 \tag{1}$$

$$\forall j \in \mathcal{J}^S \cdot y_{jS}^S = 0 \tag{2}$$

$$\forall r \in \mathcal{R} \cdot \frac{\sum_{j\in\mathcal{J}^S} t_j \cdot y_{jr}^S}{T} \le P_r \tag{3}$$

Next, the demand variables are bound by the CBR traffic generated by each job:

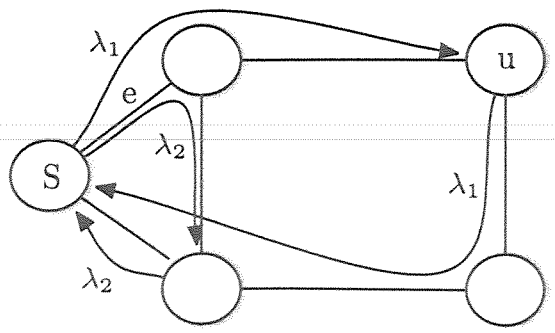


Fig. 2 Example 5-node network with $d_{Su}^S = d_{uS}^S = 1$ and $c_{eS}^S = 2$

$$\forall r \in \mathcal{R}\backslash\{S\} \cdot d^S_{Sr} \geq \frac{\sum_{j\in\mathcal{J}^S} d^{\mathrm{I}}_j \cdot y^S_{jr}}{B \cdot T} \tag{4}$$

$$\forall r \in \mathcal{R}\backslash\{S\} \cdot d^S_{rS} \geq \frac{\sum_{j\in\mathcal{J}^S} d^{\mathrm{O}}_j \cdot y^S_{jr}}{B \cdot T} \tag{5}$$

The following constraints express the network flow conservation ($\mathcal{E}^+_v$ is the set of outgoing directed links from resource v, $\mathcal{E}^-_v$ the set of incoming links):

$$\forall u \in \mathcal{R}, \quad \forall v \in \mathcal{R}\backslash\{u\} \cdot \sum_{e\in\mathcal{E}^+_v} c^S_{eu} + d^S_{uv} = \sum_{e\in\mathcal{E}^-_v} c^S_{eu} \tag{6}$$

$$\forall r \in \mathcal{R} \cdot \sum_{u\in\mathcal{R}} d^S_{ru} = \sum_{e\in\mathcal{E}^+_r} c^S_{er} \tag{7}$$

Finally, connections carried on an edge force the activation of fibers on that particular edge:

$$\forall e \in \mathcal{E} \cdot \sum_{r\in\mathcal{R}} c^S_{er} \leq W \cdot f^S_e \tag{8}$$

Our goal is to minimize the network cost, which is given by

$$\sum_{e\in\mathcal{E}} \left(\alpha \cdot f^S_e + \beta \cdot \sum_{r\in\mathcal{R}} c^S_{er} \right).$$

However, to model the fair distribution of workload over all sites, the actual objective function to be minimized is:

$$\sum_{e\in\mathcal{E}} \left(\alpha \cdot f^S_e + \beta \cdot \sum_{r\in\mathcal{R}} c^S_{er} \right) + M \cdot \max_{r\in\mathcal{R}} \sum_{j\in\mathcal{J}^S} t_j \cdot y_{jr}$$

In this last expression, M is a penalty factor, large enough to force the fair workload distribution in the solution without interfering with the network cost. For $M \gg \sum_{e\in\mathcal{E}}(\alpha \cdot f^S_e + \beta \cdot \sum_{r\in\mathcal{R}} c^S_{er})$, the resulting network cost can easily be found from the solution's objective value. Unless stated otherwise, the costs presented in this paper were obtained for $\alpha = \beta = 1$.

Global scenario

In order to dimension the network so that it is capable of handling all individual scenarios, we must ensure that there is enough network capacity to handle each overload scenario (cfr. Section Excess load scenario). We require, therefore, all constraints from the previous section for each possible source node $S \in \mathcal{R}$. Additionally, we introduce the following variables:

- F_e = number of fibers on edge e for all scenarios,
- C_{er} = average number of wavelengths departing from resource r carried by edge e over all individual scenarios, with $0 \leq C_{er} \leq \sum_{S\in\mathcal{R}} \sum_{j\in\mathcal{J}^S} \frac{(d^{\mathrm{I}}_j/B\cdot T)+(|\mathcal{R}|-1)\cdot(d^{\mathrm{O}}_j/B\cdot T)}{|\mathcal{R}|}$.

$$\forall e \in \mathcal{E}, \quad \forall S \in \mathcal{R} \cdot F_e \geq f^S_e \tag{9}$$

$$\forall e \in \mathcal{E}, \quad \forall r \in \mathcal{R} \cdot C_{er} = \frac{\sum_{S\in\mathcal{R}} c^S_{er}}{|\mathcal{R}|} \tag{10}$$

The former constraint ensures sufficient fibers are activated to carry traffic for all scenarios, while the latter fixes the number of connections to the average over all scenarios. The network cost becomes in this case:

$$\sum_{e\in\mathcal{E}} \left(\alpha \cdot F_e + \beta \cdot \sum_{r\in\mathcal{R}} C_{er} \right) \tag{11}$$

Parallellizing heuristic

The previous section showed how to combine the individual scenarios into a model that is able to satisfy all individual scenarios at once. However, this approach becomes intractable very quickly for an increasing number of variables, in particular because of the number of jobs. We therefore propose an alternative technique, which is able to return solutions within reasonable calculation time and resource limits, however, at increased network cost.

As illustrated in Fig. 3, 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^S_e). These values are used to initialize the parameters G_e:

$$\forall e \in \mathcal{E} \cdot G_e = \max_{S\in\mathcal{R}} f^S_e,$$

and then we proceed by solving the problem as defined in Section Global scenario, but replace constraints (9) by

$$\forall e \in \mathcal{E} \cdot F_e = G_e. \tag{12}$$

Large job count heuristic

Given workload L is generated by a large number of jobs n. In this case, the actual probability of a certain number of random jobs totaling a given workload depends on that number and the distributions associated with the arrival process and the job lengths. Therefore, we can approximate this set of jobs by substituting them by n jobs of equal length L/n. Assume that $L_{\max}$ is the

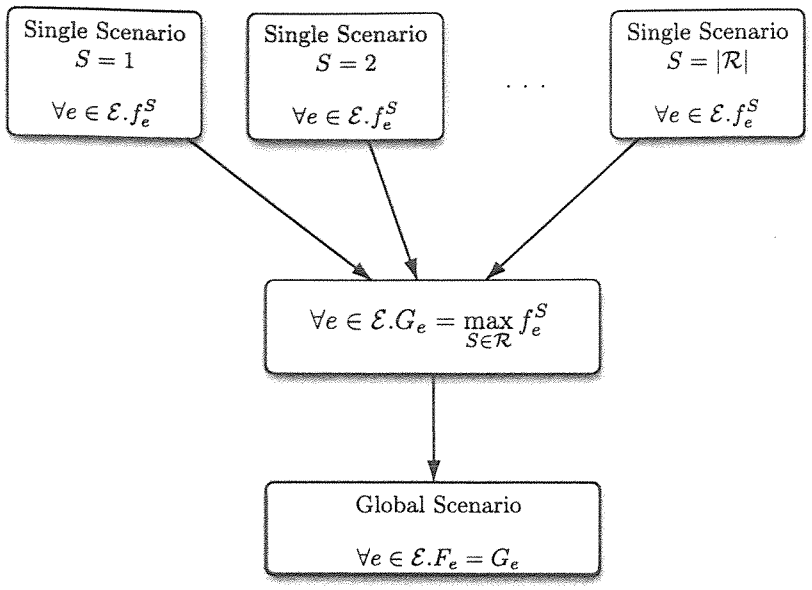


Fig. 3 Overview of heuristic method

a-priori maximum length of a single random job. Then, typically, a large set of jobs totalling workload $L \ll nL_{\max}$ will contain a relatively large number of "small" jobs (i.e., a length around L/n). This means that the total workload can be divided into $|\mathcal{R}| - 1$ parts of equal size plus some excess jobs of size L/n. Obviously, the case of n equal-sized jobs is a special instance of this. In this approximation, jobs are perfectly interchangeable.

Since we are handling a load balancing scenario, this implies that each resource must execute at least $\left\lfloor \frac{|\mathcal{J}^S|}{|\mathcal{R}|-1} \right\rfloor$ jobs. The assignment of the remaining jobs (at most $|\mathcal{R}| - 2$) is then limited by the fact that each resource may not receive more than one job (because of the load balancing constraint).

Clearly, it is not necessary to hold on to the binary variables y_{jr}^S for each job. Instead, we introduce new binary variables δ_r^S, which equal 1 iff one of the remaining jobs is executed on resource r, and 0 otherwise. Constraints (1) and (3) are replaced by

$$\sum_{r \in \mathcal{R}} \delta_r^S = |\mathcal{J}^S| \bmod (|\mathcal{R}| - 1), \tag{13}$$

while constraint (2) becomes

$$\delta_S^S = 0. \tag{14}$$

This effectively reduces the influence of the amount of excess jobs in the single scenario model, by elimination of $|\mathcal{J}^S| \cdot |\mathcal{R}|$ job decision variables to $|\mathcal{R}|$, and $2 \cdot |\mathcal{J}^S| + |\mathcal{R}|$ job-related constraints to only 2.

DLT

In the previous sections, Integer Linear Programming formulations for the combined load distribution and the optical transport network dimensioning problem were presented. The central concept in these formulations (regarding the load distribution and thus traffic demand generation) is the use of per-job (integer) decision variables. These variables ensure that the workload distribution and network dimensioning (obtained by solving the ILP) is feasible for a given set of jobs. However, as the number of jobs increases, the ILP's resulting complexity makes it difficult to obtain an optimal solution in reasonable time (Fig. 4).

For steady-state analysis of Grid systems processing large amounts of tasks, it has proven useful to treat massively parallel applications as arbitrarily divisible. By extension, one can imagine treating the workload generated at a single Grid site as arbitrarily divisible. That is, one does not consider the individual jobs (each of which is, in reality, not divisible at all) but only takes into account the aggregate workload (i.e., sum-of-jobs) generated at each site during some interval T. Using the divisible load approach, the network dimensioning problem (and the related workload scheduling problem) can be restated as a Linear Programming problem without the per-job variables. The load distribution variables in this problem are now real-valued instead of integer.

As it is common for a workload to be described in terms of stochastic variables (e.g., interarrival time of the jobs, job length, etc.), we derived appropriate DLT parameters as follows. First, we set the local processing

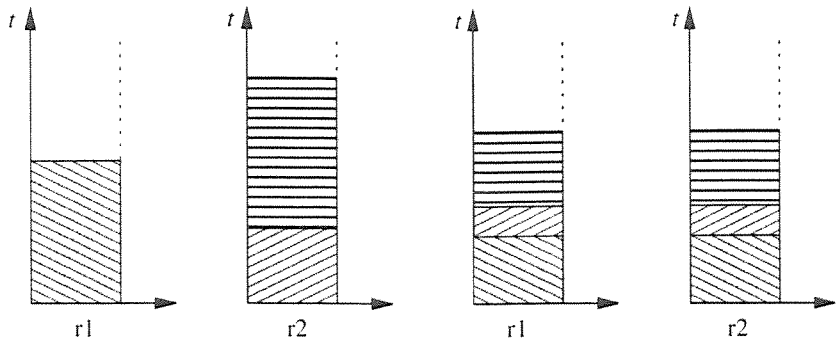


Fig. 4 Sample schedule (three jobs, two resources) when using the ILP method (left) and the DLT method (right)

capacity of each Computational Resource to the x-percentile of the stochastic workload arriving per unit of time at that Computational Resource. This value can be computed from the known distributions of job inter-arrival time and length and the size of the steady-state period T. Next, the actual workload arriving at such a CR per time unit is projected to be the y-percentile of the same stochastic workload.

The relevant portions of the scenario of a single overloaded Computational Resource r can then be reformulated (using the DLT-derived parameters) as

$$\sum_{r\in\mathcal{R}\setminus\{S\}} \alpha_r = \alpha_S^y - \alpha_S^x \tag{15}$$

$$\forall r \in \mathcal{R}\setminus\{S\} \cdot d_{Sr} \geq \frac{\alpha_r \cdot D_\mathrm{I}}{B} \tag{16}$$

$$\forall r \in \mathcal{R}\setminus\{S\} \cdot d_{rS} \geq \frac{\alpha_r \cdot D_\mathrm{O}}{B} \tag{17}$$

$$\forall r \in \mathcal{R}\setminus\{S\} \cdot \alpha_r + \alpha_r^y \leq \alpha_r^x \tag{18}$$

In these equations, α_r^y and α_r^x represent the y- and x-precentiles for site r, calculated as described above. The amount of excess load (generated at resource S) that is scheduled for remote execution at resource r is dubbed α_r (real-valued). The value of D_I (D_O), which represents the average amount of input (output) data per processing unit, can be calculated from the job interarrival time, job length and job input (output) data size stochastic variables.

For simple network topologies (full mesh, star, ring) and shortest-path routing, the resulting network cost for a single excess load scenario (overloaded resource $S = 0$) can be expressed analytical in terms of $\alpha_r, \alpha_S^y, \alpha_S^x, B, D_\mathrm{I}$ and D_O (substituting $\left\lceil \frac{D\cdot(\alpha_S^y-\alpha_S^x)}{B\cdot(|\mathcal{R}|-1)} \right\rceil$ for d_{Sr}—assuming that $D_\mathrm{I} = D_\mathrm{O} = D$ and excess load is distributed evenly across all remote sites), which gives, respectively:

$$\mathrm{Cost}_\mathrm{Mesh} = 2 \cdot \sum_{r\neq 0} \left(d_{0r} + C \cdot \left\lceil \frac{d_{0r}}{W} \right\rceil \right), \tag{19}$$

$$\mathrm{Cost}_\mathrm{Star} = 2 \cdot \sum_{r\neq 0} \left(d_{0r} + C \cdot \left\lceil \frac{d_{0r}}{W} \right\rceil \right) + 2 \cdot \sum_{r\neq 0} d_{0r} + 2 \cdot C \cdot \left\lceil \frac{\sum_{r\neq 0} d_{0r}}{W} \right\rceil, \tag{20}$$

$$\mathrm{Cost}_\mathrm{Ring} = \sum_{k=1}^{\left\lfloor \frac{|\mathcal{R}|}{2} \right\rfloor} \left(\sum_{r=k}^{\left\lfloor \frac{|\mathcal{R}|}{2} \right\rfloor} d_{0r} + C \cdot \left\lceil \frac{\sum_{r=k}^{\left\lfloor \frac{|\mathcal{R}|}{2} \right\rfloor} d_{0r}}{W} \right\rceil \right) + \sum_{k=l}^{|\mathcal{R}|-1} \left(\sum_{r=l}^{k} d_{0r} + C \cdot \left\lceil \frac{\sum_{r=l}^{k} d_{0r}}{W} \right\rceil \right), \tag{21}$$

where $l = \left\lfloor \frac{|\mathcal{R}|}{2} + 1 \right\rfloor$. Here, $\lceil x \rceil$ denotes the smallest integer larger than or equal to x, and $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x.

Results and discussion

ILP solver

All ILP needed to evaluate the different solution methods have been solved using ILOG CPLEX 8.0, running on an AMD Athlon XP1700+ based OpenMosix cluster (20 Debian GNU/Linux nodes) with 1 GB RAM per node.

Reference topology

We used the European core network depicted in Fig. 5, which is composed of 13 nodes and 17 bidirectional links. These links constitute fiber ducts; the exact number of fibers needed on each link follows from the solution to the dimensioning problem. 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

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

site (thus, we make abstraction of any access networks in place), so that our Grid has as many cross connects as Grid sites. Unless stated otherwise, we have solved the dimensioning problem on this topology assuming each wavelength provides a data transfer rate of 2.5 Gbps, each fiber carries at most four wavelengths and the workload consisted of 2,000 jobs instantiated as described in the following section.

Job parameters

For the evaluation of the global optimization of the single overloaded source scenarios using an exact ILP, a heuristic decomposition of the exact program and the divisible load technique, we chose the following synthetic workload:

- job interarrival times are assumed to be independent and identically distributed, following an exponential distribution (thus, the number of jobs arriving over some interval follows a Poisson distribution),
- job lengths are also chosen to be independent and identically distributed, following a uniform distribution over $[0, L_{\max}]$, with $L_{\max} \ll T$,
- the size of the input data processed by a job is proportional to that job's length (factor D_{I}, resulting in size d_j^{I}),
- in analogy, the size of the output data generated by a job is proportional to that job's length (factor D_{O}, resulting in size d_j^{O}).

Excess load

As previously explained, we assume a single Computational Resource is experiencing excessive load. We set the amount of load that can be handled locally (i.e., the Computational Resource's capacity) to be the 60%-percentile of the load as described by the arrival process and the job length distribution. We assumed that an excessive load is made up by a job set instance whose aggregate load equals the 90%-percentile of the arriving workload. For the job sets used in the ILP model, we generated sets with total load equal to the difference of these percentiles. Except for the results presented in section Scheduling strategies, it is always assumed that excess load is distributed evenly among all remote sites (i.e. $k = |\mathcal{R}| - 1$).

Computational complexity

We have compared the computational complexity of the (Mixed) Integer Programs resulting from the application of the ILP, Heuristic and DLT methods in Table 1, which lists the (order of magnitude of the) number of variables and constraints in each program as functions of network dimensions, number of resources, and number of jobs.

As explained, the main simplification introduced by the DLT method is the absence of per-job variables, while the computational advantage of the heuristic over the ILP method lies in the reduced size of the individual subproblems. Note that, for example, the calculation of distribution fractiles for job length and data sizes is not contained in any of the complexity metrics in Table 1.

ILP vs DLT

In Fig. 6, we have depicted the resulting cost for the dimensioning problem, applied to the topology from Fig. 5. These results show the cost for increasing number of jobs under the constraint that the total job load (per period) remains constant. Each value obtained for the parallellizing heuristic is the result of averaging the cost over ten different job instances. For low values of the job count ($\leq$2000), we used the same approach for the costs obtained with the exact ILP formulation. The granularity of the job requirements causes the ILP method to perform better than the DLT approach in some cases, and worse in others. For higher number of jobs however, the computational intractability forced us to resort to the approximation described in Section Large job count heuristic. In this case, each measurement is the average cost obtained from $|\mathcal{R}| - 2$ evaluations of this approximation for successive numbers of jobs around

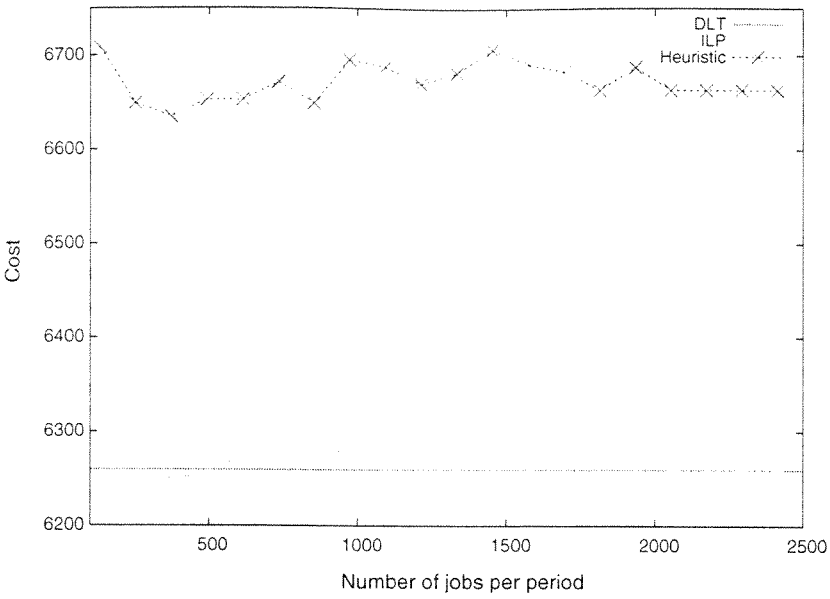


Fig. 6 Cost vs. number of jobs per period for European network

Table 1 Algorithm comparison: computational complexity (reduction by factor $|\mathcal{R}|$ from ILP to parallellizing heuristic, and term $|\mathcal{J}|$ from ILP to DLT)

Algorithm	Integer Vars	Float Vars	Constraints
ILP	$\|\mathcal{R}\|^2 \cdot (\|\mathcal{R}\| + \|\mathcal{E}\| + \|\mathcal{J}\|)$	0	$\|\mathcal{R}\| \cdot (\|\mathcal{R}\|^2 + \|\mathcal{E}\| + \|\mathcal{J}\|)$
Heuristic	$\|\mathcal{R}\| \cdot (\|\mathcal{R}\| + \|\mathcal{E}\| + \|\mathcal{J}\|)$	0	$\|\mathcal{R}\|^2 + \|\mathcal{E}\| + \|\mathcal{J}\|$
DLT	$\|\mathcal{R}\|^2 \cdot (\|\mathcal{R}\| + \|\mathcal{E}\|)$	$\|\mathcal{R}\| \cdot \|\mathcal{E}\|$	$\|\mathcal{R}\| \cdot (\|\mathcal{R}\|^2 + \|\mathcal{E}\|)$

Table 2 Network cost for different wavelength/fiber cost models and wavelength granularity

B (Gbps)	Algorithm	Constant	Geo (1.5)	Geo (2.5)	Geo (3.5)	Linear
0.155	DLT	58361.1	27602.13	11815.0	7384.46	6241.2
	ILP	58241.08	27548.26	11795.06	7373.76	6254.76
	Heuristic	58485.23	27755.44	12072.47	7739.88	6648.85
0.622	DLT	16780.2	12147.2	8407.4	6772.856	6241.2
	ILP	16760.0	12134.00	8398.68	6766.76	6254.96
	Heuristic	16992.92	12395.82	8723.89	7136.62	6650.31
2.5	DLT	6259.54	6259.54	6259.54	6259.54	6259.54
	ILP	6259.54	6259.54	6259.54	6259.54	6259.54
	Heuristic	6660.23	6660.23	6660.23	6660.23	6665.23
10	DLT	3605.77	4079.58	4986.93	5909.23	6364.6
	ILP	3605.77	4066.15	4986.92	5906.54	6364.62
	Heuristic	4116.92	4555.38	5432.30	6309.23	6762.69

the measurement's corresponding x-value. Obviously, the cost of the DLT-based method remains constant as by its very nature, only the aggregate load is of importance. From the figure, it is clear that for high number of jobs (totalling a constant load), the cost for the ILP method converges to the DLT cost. Furthermore, the DLT-based approximation performs consistently better than the heuristic, which reduces computational complexity by parallellizing the dimensioning problem into independent subproblems (maximal deviation from ILP method is about 0.5% in this case, vs. 5% deviation for the heuristic).

Connectivity

In the previous section, results were obtained for a single network topology. Figure 7 shows the resulting cost for the dimensioning problem when solved for a wider range of network topologies, for 2,000 excess jobs per period. We randomly generated connected networks (with number of nodes equal to the number of nodes in the reference network) for varying random-link probabilities 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

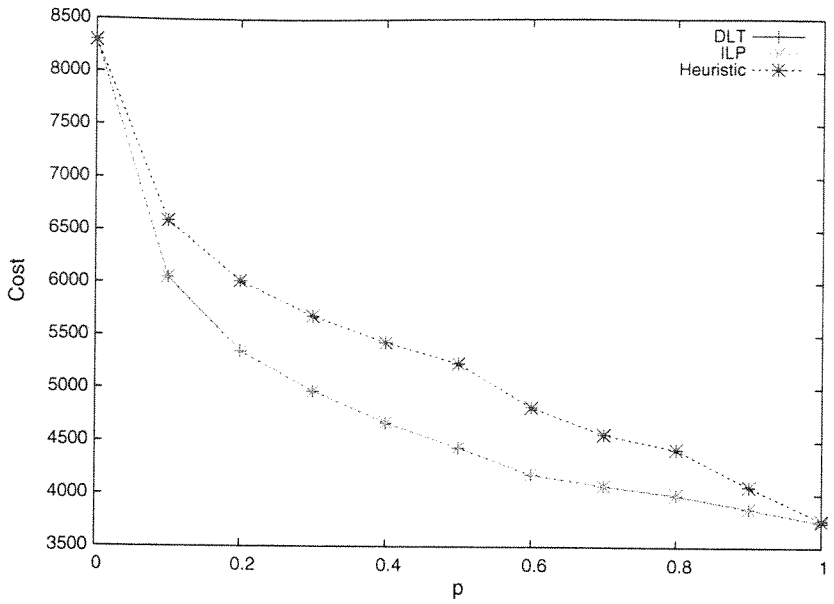


Fig. 7 Cost vs. average connectivity for random networks with 13 nodes each

network), 10 topologies were generated. For all topologies, the DLT-based solution is very close to the solution obtained from the ILP, the difference staying below 1% for all values of p. Only in trivial cases ($p \to 0, p \to 1$) the heuristic approach obtains the quality of the DLT-based solution.

Asymmetric jobs

So far, symmetry between incoming and outgoing traffic for each job was assumed ($D_I = D_O$). Figure 8 shows the cost (for the reference network, again using 2,000 excess jobs per period) for varying ratios of D_I–D_O but constant $D_I + D_O$. Since the dimensioning problem we study a global optimization over individual similar scenarios, we expect these results to show symmetry around $s = D_I/D_O = 1$ as each pair of directed links (connecting two routers) considered for transporting input data will also be considered for transporting output data. The net result is that the chosen 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.

Wavelength granularity

Results presented so far were obtained using at most four wavelengths per fiber, each wavelength able to carry 2.5 Gbps. Below, we have evaluated the dimensioning problem for the reference network for different wavelength granularities. In all cases, fiber capacity was fixed at 10 Gbps. This value and the wavelength granularity determined the number of wavelengths that can be carried on each fiber. Additionally, we have compared different cost models of the wavelength per fiber parameter $C = \beta/\alpha$. Each model is represented by a non-decreasing function, which mirrors the economic reality of the higher cost for technologies with larger wavelength capacity (Fig. 9). However, since smaller wavelength capacity implies a larger number of activated wavelengths, and thus increasing number of line termination equipment, we introduce three different functions for the parameter C. First, the constant function is invariant to changes in wavelength granularity. The second function scales the cost of a wavelength over a fiber linearly with the wavelength's bandwidth. Finally, three different geometric functions (factors 1.5, 2.5, and 3.5) are bounded by the constant and linear functions.

Table 2 summarizes our results for different wavelength bandwidths and cost models. For all wavelength granularities presented (and thus, maximal number of wavelengths per fiber), our DLT-based approach outperforms the heuristic and follows the ILP approach closely. This remains the case over all different cost models.

Figure 10 shows the resulting network cost obtained with the DLT approach for different models of the wavelength per fiber cost.

Scheduling strategies

In the previous sections, all results were obtained for uniform excess load distribution over all remote sites (i.e., $k = 12$ for 13-node networks). As our model supports the combined dimensioning of the network and optimal selection of load-balancing sites, Figs. 11 and 12 show the results for the DLT, ILP and heuristic approaches for different scheduling strategies (k-values). These results

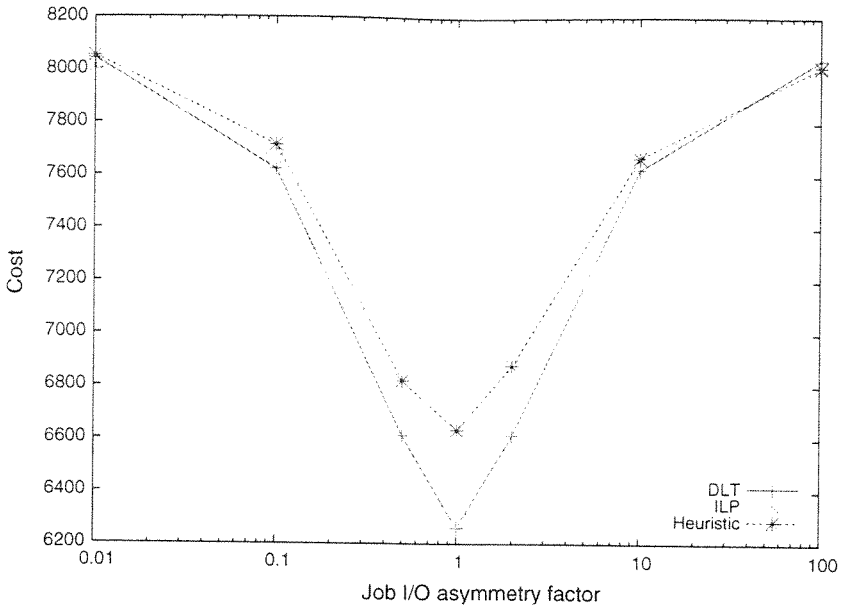


Fig. 8 Cost vs. traffic asymmetry for European network

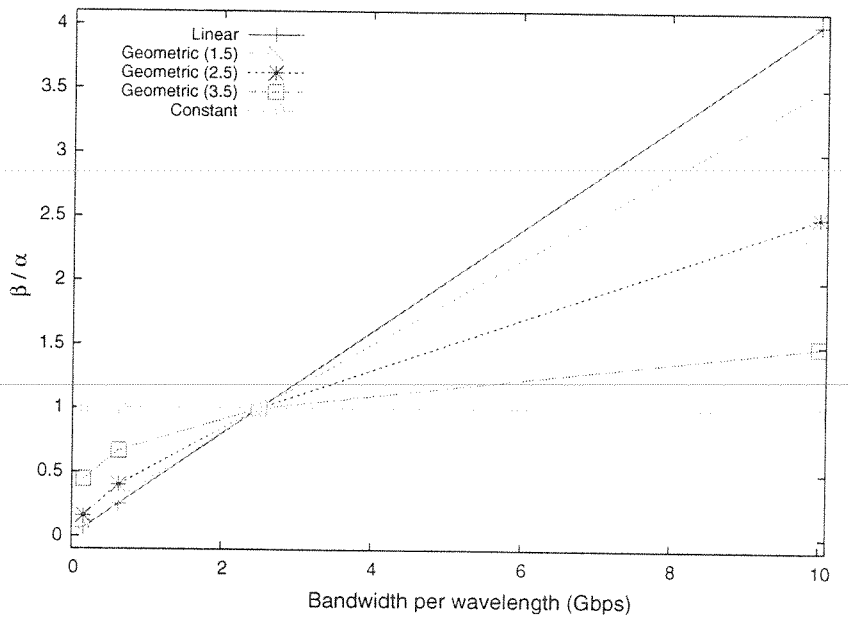


Fig. 9 Different wavelength/fiber cost models vs. wavelength granularity

were obtained on the previously used sets of randomly generated networks with average connectivity 0.1 and 0.9, respectively.

In all cases, the DLT and ILP approaches give similar network costs, outperforming the heuristic. Note though that in some cases, the DLT approach yields better solutions than the ILP approach, which can be attributed to the limited computational resources available to the ILP solver. The results for the heuristic approach (as a function of k) can be explained as follows: for each individual excess load scenario, the heuristic "optimizes" network cost by selecting the k closest (with respect to hop count) remote sites. For high average connectivity values, the amount of sets resulting in minimal cost is higher than for lower average connectivity values. This means that the heuristic approach (which does not correlate individual scenarios) is prone to selecting previously unused resources, resulting (after deciding on global capacities) in high network costs. This effect decreases for larger values of k, as the number of sets resulting in minimal network costs are lowered.

On the other hand, for low average connectivity values, the remote sites yielding minimal network cost are more likely to form a unique set. As such, higher

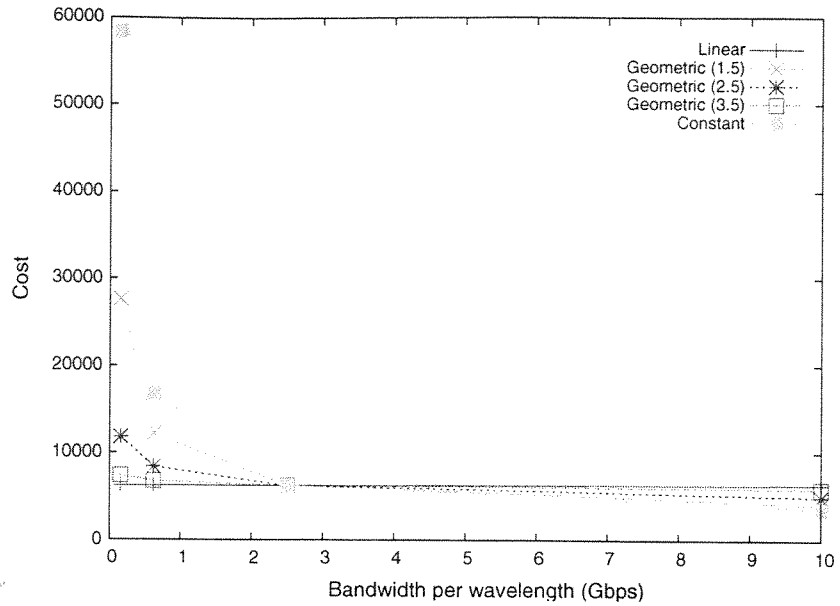


Fig. 10 DLT Cost vs. Wavelength Granularity for European network under different wavelength/fiber cost models

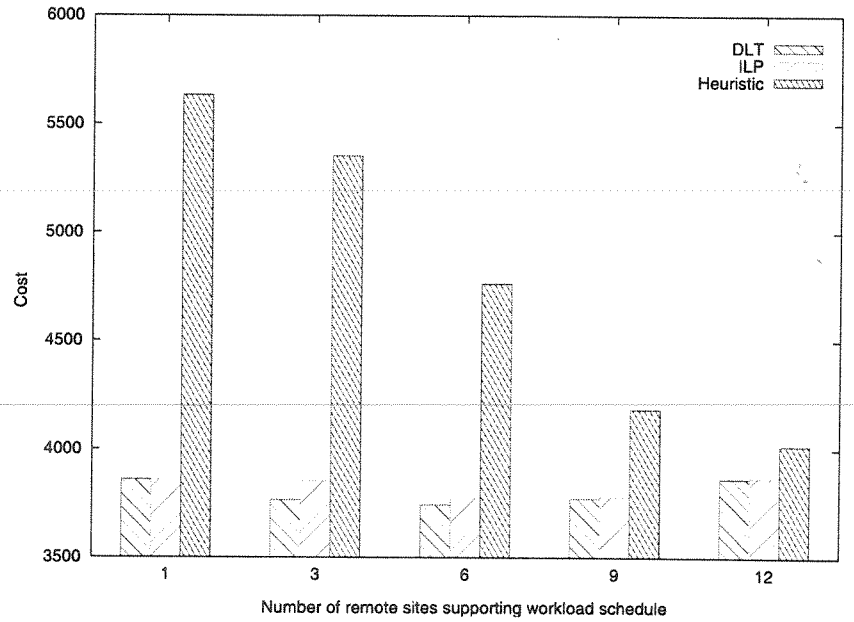


Fig. 11 Cost vs. scheduling strategy for random networks with 13 nodes, average connectivity 0.9

k-values imply higher network costs due to the use of remote sites located further away (with regard to hop-count). In addition, the difference between the network cost obtained through the heuristic and the network costs obtained by using the DLT or ILP approaches is smaller in the case of lower average connectivity because the heuristic is forced into using the same resource set as used by the DLT and ILP methods.

The same argument explains why the heuristic seems to perform badly for high connectivity and low k-values. Here, we compare to the heuristic operating in the low connectivity, low k-value scenarios. This is because of the possible existence of multiple optimal solutions to a single overload scenario in the former case, a simple maximization over all these individual scenarios (without correlation) may perform erratically, depending on which optimal solution was selected in each individual scenario. An interesting improvement to the heuristic therefore consists of determining the *set* of optimal solutions to each individual scenario, and selecting the best global combination of these solutions.

Applicability of DLT

Results show that the DLT approach to modeling and solving the optical transport network dimensioning problem (in the context of Grid excess load handling) approximates the ILP-based approach quite well in case this

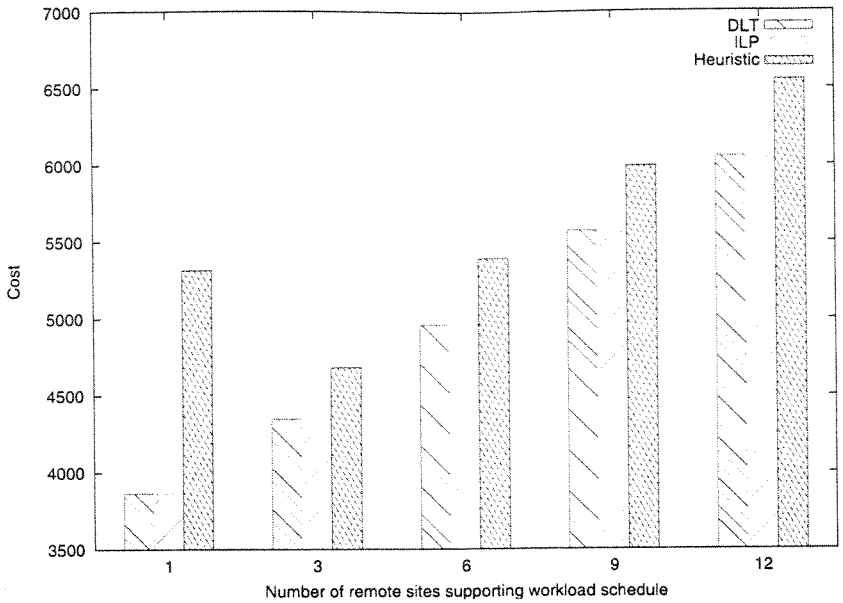


Fig. 12 Cost vs. scheduling strategy for random networks with 13 nodes, average connectivity 0.1

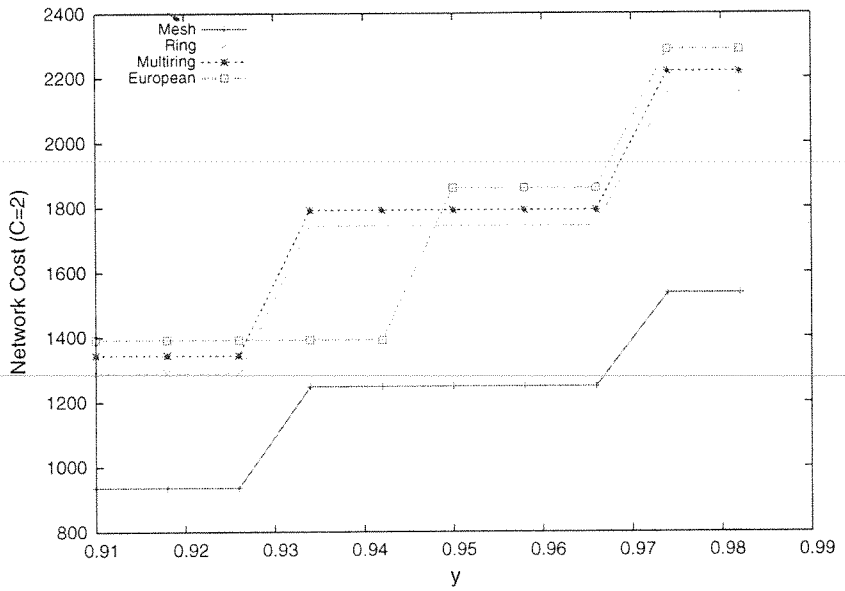


Fig. 13 DLT network cost vs. increasing excess load for different 13-node topologies

latter method becomes intractable due to large job numbers. This in turn means that the DLT approach can be used by network operators to obtain network costs for different interconnection topologies and compare them (e.g., in order to determine the most cost effective interconnection topology, given an excess load scenario that the resulting Grid must support). A sample comparison between the European network and the simple topologies mentioned in Section DLT is shown in Figs. 13 and 14. The Multiring network mentioned there consists of three rings, connected through a superring intersecting each ring exactly once. Costs in these figures were obtained for $C = \beta/\alpha = 2$.

Future work

Due to the proven applicability of the DLT-based approach, we will study how it can be extended into more complex, non-load balancing scheduling scenarios featuring multiple workload sources. As we are particularly interested in dimensioning optical networks, an interesting topic to investigate is the influence and/or necessity of limited wavelength conversion capabilities in the network, possibly introducing (the number of) wavelength conversions in the cost function.

Finally, we will extend our formulations to support the dimensioning of survivable networks.

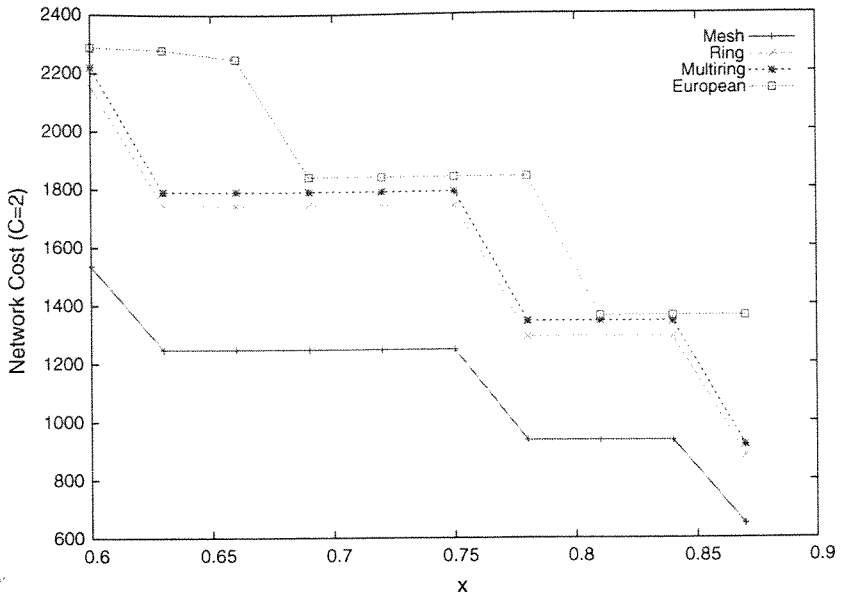


Fig. 14 DLT network cost vs. increasing resource capacity for different 13-node topologies

Conclusions

In this paper, we have studied the dimensioning problem of an optical transport network for Grid applications under single site excess load assumption. We presented and discussed a solution for this problem using a model based on DLT. We compared this model to an Integer Linear Programming formulation using an exact job-level workload description, and a parallelizing heuristic based on this ILP.

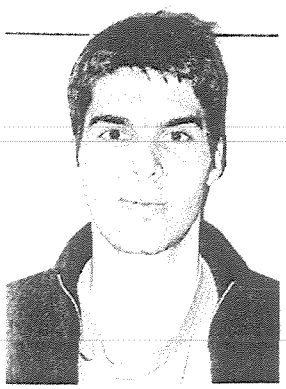
Results show that the global optimization of single overloaded source scenarios using the exact job-level ILP formulation is possible only for a low number of jobs. However, we have established the convergence of the DLT-based approach and this job-level ILP formulation for increasing number of jobs. This indicates that the DLT formulation is of practical use in cases where the exact ILP becomes computationally intractable. Additionally, we have presented a heuristic method based on the job-level ILP model, which shows good results and scales better for a high number of jobs, although it is consistently outperformed by the DLT-based method. We showed that these conclusions remain valid 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. This means that our DLT-based approach is of practical use to network operators interested in selecting and dimensioning a suitable OCS Grid interconnection topology, including selection of optimal wavelength granularity.

Acknowledgements P. Thysebaert is supported as a Research Assistant by the Fund for Scientific Research—Flanders (F.W.O.—Vlaanderen). M. De Leenheer and B. Volckaert are supported by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT—Vlaanderen). F. De Turck is a postdoctoral Fellow of the F.W.O.—Flanders.

References

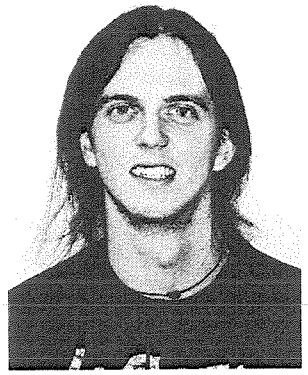
1. Smarr, L., Chien, A., DeFanti, T., Leigh, J., Papadopoulos, P.: The OptIPuter. Commun. ACM **46**(11), 58–67 (2003)
2. DeFanti, T., de Laat, C., Mambretti, J., Neggers, K., Arnaud, B.: TransLight: A global-scale LambdaGrid for e-Science. Commun. ACM **46**(11), 34–41 (2003)
3. Wauters, N., Demeester, P.: Design of the optical path layer in multiwavelength cross-connected networks. IEEE J. Selected Areas Commun. **14**(5), 881–892 (1996)
4. Banerjee, D., Mukherjee, B.: Wavelength-routed optical networks: Linear formulation, resource budgeting tradeoffs, and a reconfiguration study. IEEE/ACM Trans. Networking **8**(5), 598–607 (2000)
5. Coudert, D., Rivano, H.: Lightpath assignment for multifibers WDM networks with wavelength translators. Proc. IEEE Globecom'02, vol. 3, pp. 2686–2690. Taipei, Taiwan (2002)
6. Tornatore, M., Maier, G., Pattavina, A.: WDM network optimization by ILP based on source formulation. Proc. IEEE Infocom'02, vol. 3, pp. 1813–1821. New York, NY (2002)
7. Kolisch, R., Padman, R.: An integrated survey of project scheduling. OMEGA Int. J. Manage. Sci. **29**(3), 249–272 (2001)
8. Sgall, J.: On-Line scheduling—a survey, online algorithms: the state of the art, Lecture Notes in Computer Science, vol. 1442, pp. 196–231. Springer-Verlag, Berlin (1998)
9. Hall, L.A., Schulz, A.S., Shmoys, D.B., Wein, J.: Scheduling to minimize average completion time: off-line and on-line approximation algorithms. Math. Operations Res. **22**(3), 513–549 (1997)
10. Feitelson, D.G., Rudolph, L., Schwiegelshohn, U., Sevcik, K.C., Wong, P.: Theory and practice in parallel job schedul-

ing. Job scheduling strategies for parallel processing. Lecture Notes in Computer Science, vol. 1291, pp. 1–34. Springer-Verlag, Berlin (1997)
11. Bucur, A., Epema, D.: The maximal utilization of processor co-allocation in multicluster systems. Proc. 17th IEEE International Parallel and Distributed Processing Symposium (IPDPS'03), p. 60a. Nice, France (2003)
12. Bucur, A., Epema, D.: The influence of the structure and sizes of jobs on the performance of co-allocation. Lecture Notes in Computer Science, vol. 1911, pp. 154–173. Springer-Verlag, London, UK (2000)
13. Buyya, R., Murshed, M.: GridSim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing, concurrency and computation: practice and experience, vol. 14, no. 13–15, Wiley Press, USA, pp. 1175–1220 (2002)
14. Legrand, A., Marchal, L., Casanova, H.: Scheduling distributed applications: the simGrid simulation framework. Proc. of the 3rd International Symposium on Cluster Computing and the Grid, pp. 138–145. Tokyo, Japan (2003)
15. Volckaert, B., Thysebaert, P., De Turck, F., Demeester, P., Dhoedt, B.: Evaluation of grid scheduling strategies through a network-aware grid simulator. Proc. of the 2003 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'03), vol. 1, pp. 31–35. Las Vegas, NV, USA (2003)
16. Ranganathan, K., Foster, I.: Simulation studies of computation and data scheduling algorithms for data grids. J. Grid Computing 1(1), 53–62 (2003)
17. Cameron, D.G., Carvajal-Schiaffino, R., Millar, A.P., Nicholson, C., Stockinger, K., Zini, F.: Evaluating sheduling and replica optimisation strategies in OptorSim. Proc. of the 4th International Workshop on Grid Computing, pp. 52–59. Phoenix, AZ, USA (2003)
18. Hung, J.T., Kim, H.J., Robertazzi, T.G.: Scalable scheduling in parallel processors. Proc. of the 36th Annual Conference on Information Sciences and Systems (CISS'02), Princeton, NJ, USA (2002)
19. Yu, D., Robertazzi, T.G.: Divisible load scheduling for grid computing. Proc. of Parallel and Distributed Computing and Systems (PDCS'03), Marina del Rey, CA, USA (2003)
20. Marchal, L., Yang, Y., Casanova, H., Robert, Y.: A realistic network/application model for scheduling divisible loads on large-scale platforms, Rapport de recherche de l'INRIA-Rhone-Alpes (RR-5197), Montbonnot, France (2004)
21. Thysebaert, P., De Turck, F., Dhoedt, B., Demeester, P.: Using divisible load theory to dimension optical transport networks for computational grids. Proc. of the 2005 OFC/NFOEC Conference, Anaheim, CA, USA (2005)

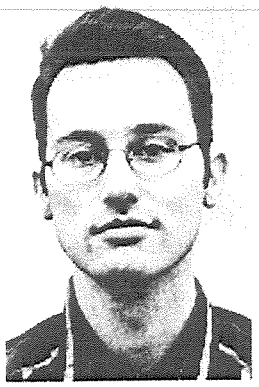


Pieter Thysebaert received his M.Sc. degree in Computer Science Engineering from Ghent University, Belgium, in June 2001. He is now a research assistant and Ph.D. student affiliated with the Department of Information Technology at Ghent University and has received a scholarship by the FWO (Fund for Scientific 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 with the Department of Information Technology at Ghent University and has received a scholarship by the IWT (Institute for Innovation in Science and Technology—Flanders). His main interests include modeling and optimization of Grid management architectures, specifically in the context of photonic networks.



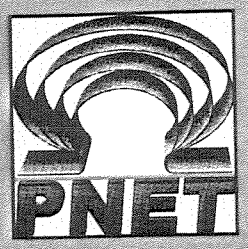
Bruno Volckaert received his M.Sc. degree in Computer Science from Ghent University, Belgium, in June 2001. He is now a research assistant and Ph.D. student affiliated with the Department of Information Technology at Ghent University and has received a scholarship by the IWT (Institute for Innovation 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 with the Fund for Scientific Research-Flanders, Belgium (F.W.O.-V.). At the moment, he is a part-time professor and a post-doctoral fellow of the F.W.O.-V., affiliated with the Department of Information Technology at Ghent University. Filip De Turck is 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 is responsible for several courses on algorithms, programming and software development. His research interests are software engineerin,g and mobile and wireless communications. Bart Dhoedt is 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 author of more than 400 publications in the area of network design, optimization and management. He is member of the editorial board of several international journals and has been member of several technical program committees (ECOC, OFC, DRCN, ICCCN, IZS).



Photonic Network Communications

Volume 12, Number 2, September 2006

ISSN: 1387-974X

Springer