COIN – ACOFT 2007
24 – 27 June 2007, Melbourne, Australia

Scheduling in Optical Grids: A Dimensioning Point of View

Chris Develder, Marc De Leenheer, Tim Stevens, Bart Dhoedt, Filip De Turck, Piet Demeester

Department of Information Technology (INTEC), Ghent University,
G. Crommenlaan 8 bus 201, BE–9050 Gent, Belgium
email: {chris.develder, marc.deleenheer}@intec.ugent.be

***Abstract* — Optical Grids promise cost and resource efficient delivery of (distributed) services. We address the fundamental question of how to dimension such optical Grids, and the impact of Grid scheduling algorithms on the resulting resource dimensions.**

I. Introduction

Grid networks promise to offer a platform for the cost and resource efficient delivery of network services to execute tasks with (possibly high) data rates, processing and data storage requirements, in a geographically widely distributed environment. Realization of that Grid requires integration/ interaction of Grid logic into/with the network layers. Given the high data rates involved, optical networks offer an undeniable potential for the Grid. Apart from (re)designing the architecture of a flexible optical layer, delivering the Grid promise implies answering a series of fundamental questions [1] including the development of the necessary design techniques for e.g. dimensioning, algorithms for routing and control offering both QoS and resilience guarantees.

This paper focuses on how to dimension such an optical Grid, given the network topology and the amount of Grid jobs offered. Fundamental differences with respect to dimensioning "classical" optical networks lie in the following concepts: (i) *The anycast routing paradigm:* A Grid job does not care where it is executed; (note that this does not apply to the job's processed result, which should be sent back to the job submitter); (ii) *Burst starvation:* bursts can not only be lost because of network contention (e.g. no available wavelengths), but also through lack of Grid resources (CPU, disk space) preventing timely execution of a job; and (iii) *Advance reservation:* Jobs may be announced relatively long in advance. This notion of advance reservations of resources is not present in classical IP-based OBS. Here we will only deal with (i) and (ii).

Related work on dimensioning Grids is quite scarce. In [2], the authors use analytical ILP and heuristic approximations to cater for excess load scenarios, starting from a given Grid configuration. Other dimensioning efforts assume that the fraction of generated jobs (originating at a particular site) going to a given computational Grid site is known, thus fixing a priori the arrival rates of jobs at each job execution site. In contrast, we will assume flexible scheduling strategies to incorporate (i).

II. Optical Grid Architecture

In a Grid environment, users submit jobs to the network through a Grid User network Interface (GUNI), thus providing the job's characteristics (processing, storage, priority/policy requirements, etc.). Likewise, grid resources announce their capabilities (storage space, processing power, etc.) through a Grid Resource Network Interface (GRNI). In this study, we will make abstraction of these interfaces and only consider the user's job arrivals (average rate λ) and server capacities (average processing rate μ, number of servers n). Note that also the network characteristics, such as topology and bandwidth will need to be known to the Grid scheduling and/or routing algorithms. The latter will be discussed in more detail in subsequent sections.

That optical technology will leverage Grid networks is irrefutable, but whether to adopt a Optical Circuit Switching (OCS) or rather an Optical Packet/Burst Switching (OPS/OBS) paradigm is still debatable. Depending on the ratio signaling time/job transmission time, OCS can be acceptable [3]: only if jobs require sufficiently long data transmissions—hence light path holding times that are long compared to the setup and tear-down process—OCS makes sense. For small jobs, some form of rather complex grooming/aggregation at the OCS edges will be required to warrant efficient use of light paths. As job data size reduces and/or latency-sensitivity increases, OBS will be more efficient [4]. Another advantage of a packet switching paradigm such as OBS [5] is its ease in dealing with highly dynamic traffic patterns (both in space and time).

The methodology followed in this paper can be used for both OBS and OCS choices, as will be illustrated in the following.

III. Dimensioning Optical Grids

A. Problem Statement

The problem we are trying to solve can be summarized as follows:

Given:

– A graph representing the network topology (nodes representing Grid sites and switches, links the optical fibers interconnecting them),
– The arrival process of jobs originating at each Grid site,
– The job processing capacity of a single Grid server, and
– A target maximum job loss rate,

Find:

– The amount of Grid servers at each site, and
– The amount of link bandwidth to install,
– While meeting the maximum job loss rate criterion.

For the job arrival process, we will assume Poisson arrivals, since measurements at a Grid level have shown that the interarrival times for jobs arriving to a large scale Grid are indeed exponentially distributed [6]. To dimension the Grid server sites and links, we will take an iterative approach, first calculating the amount of server sites needed, and subsequently derive the inter-site job rates, hence bandwidth.

B. Dimensioning the server sites

In this work, targeting a first assessment of the impact of job scheduling on resource requirements, we do not take into account job buffering: upon arrival of a job, the scheduler tries to find a free server, and drops the job if none is found. Thus, assuming Poisson arrivals (mean job arrival rate λ), and exponentially distributed job processing times (mean processing time μ), we can use the well-known Erlang B formula to calculate the total number of servers n required to achieve a maximal loss rate L:

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

Numerically solving the Erlang B formula for n only gives the total amount of servers to install. To decide where to place the servers, we will consider three strategies:

(i) *unif*: uniformly distribute the servers among all Grid sites (i.e. if there are N sites, then put n/N at each site);

(ii) *prop*: distribute the servers proportionally to the arrival rate at each site, i.e. if λ_i is the job arrival rate at site i then put $n \cdot \lambda_i / \lambda$ servers at site i;

(iii) *lloss:* distribute the servers to try and achieve the same "local loss rate" at each site. Therefore, calculate n_i as the number of servers to install locally at site i to achieve the loss rate L (i.e. solve $L = ErlangB(n_i, \lambda_i, \mu)$) and then install $n \cdot n_i / (\sum n_i)$ servers at site i).

C. Scheduling non-local traffic

When a job arrives at a site, the scheduler needs to decide at which server to execute it (we assume jobs will not be migrated amidst their execution). All scheduling approaches studied here will always choose a local server (i.e. at the job arrival site) if one is free. The only difference in the scheduling approaches is in electing a remote server for job execution (if a free one can be found). If upon arrival of a job no servers are free, the job is considered lost (cf. bufferless assumption for using ErlangB). We will discuss three scheduling approaches:

(i) *rand*: randomly choose a free server (i.e. if there are K free servers, each has 1/K chance of being chosen);

(ii) *SP*: the closest free server in terms of hop count is chosen, thus striving to minimize network usage;

(iii) *mostfree:* choose a free server at site S, where S is the site with the highest number of free servers, in an attempt to avoid overloading sites and thus limiting non-local job execution.

D. Dimensioning the links

Given the server locations determined by one of the dimensioning strategies in Section B and a scheduling strategy in C, the amount of jobs transmitted between each two sites can be found. Because of the high interdependency of job arrival patterns and blocking among different sites, analytical approximation based on e.g. fixed point Erlang approximations [7] did not yield satisfactory results. Hence, we resorted to simulation. Given the number and location of the servers, and the job arrival rates, we determined the amount of jobs exchanged between each two Grid sites. Using shortest path routing, we could then calculate the required link bandwidths.

IV. CASE STUDY

To assess the impact of the scheduling algorithm and site dimensioning strategies on resource requirements, we considered a case study on European network topology.

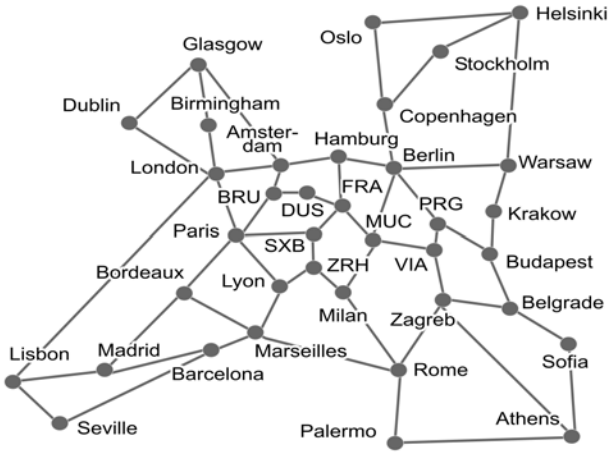

Fig. 1. European network topology

A. Scenario

The network topology considered is a meshed network covering Europe with 37 nodes and 57 bidirectional links (i.e. average node degree of 3.08), as illustrated in Fig. 1. The job arrival rates at each site were chosen randomly (each rate λ_i was with 30% chance uniformly chosen in [1,15] and 70% from [30,60]). The resulting arrival rates amounted to a total over all sites of 824.45 jobs per time unit (min/max/average per site: 1.93/59.26/22.28). To achieve max. 5% job loss, using the ErlangB formula and a processing time of 1 job per time unit per server, we found that 799 servers were required.

B. Local processing of traffic

The first criterion to judge the scheduling and dimensioning strategies by is the amount of jobs, taken over all sites, that is processed locally, as shown in Fig. 2. Note the relatively low fraction of locally processed jobs, which is due to the absence of buffering and the high resource load (cf. scaling the arrival rates down to 90%, we achieve ~70% local processing).

With respect to dimensioning strategies, as intuitively expected, the *prop* and *lloss* strategies (placing more servers at sites where more jobs originate) achieve higher local processing rates. When looking at the variation on local processing rates over all sites (see Fig. 3), we note that *lloss*

dimensioning achieves its aim of reducing variation in local processing rates, esp. for the *mostfree* scheduling strategy.

From the scheduling perspective, *mostfree* confirms our intuition by achieving the highest local processing rates. Still, the difference with the other ones is rather limited. *SP*, by its deterministic order in choosing sites for remote processing, systematically (over)loads the same servers thus achieving lowest local rates of all considered approaches.

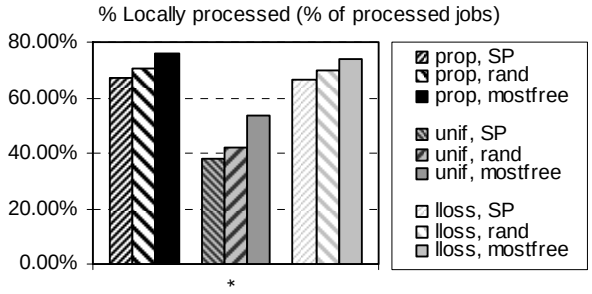

Fig. 2. Fraction of jobs that are processed locally (i.e. at originating site)

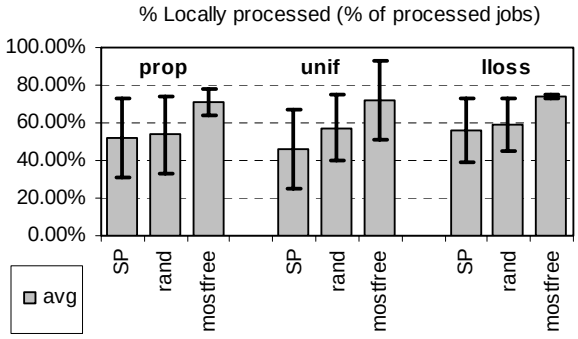

Fig. 3. Local processing fraction averaged over all sites, error bars indicate stdev.

C. Network Dimensioning

The last step in the dimensioning process is determining link bandwidths. The required input is a traffic matrix, giving the site-to-site job rates. Using these, either an OBS or OCS network can be appropriately dimensioned using conventional methods, e.g. using the ErlangB formula to calculate the number of wavelengths on each link. (Given that OBS allows full link bandwidth sharing, we found that in this particular case study using shortest path routing, the amount of wavelengths for OCS is a factor 5 higher).

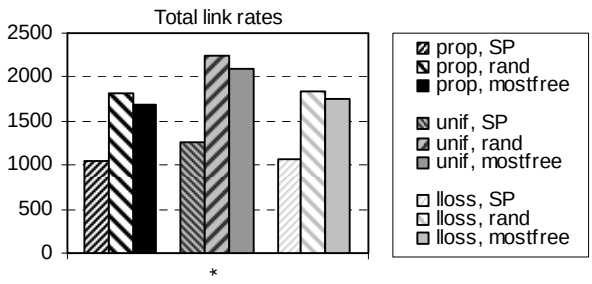

Fig. 4. Total link rates, i.e. number of jobs per time unit crossing each link summed over all links.

In Fig. 4 we present the total amount of jobs crossing each link. As expected, the *SP* scheduling achieves the lowest network load, by minimizing the path length that jobs have to cross. *Mostfree* obviously achieves lower network load than *rand* due to its higher local processing rates, but by ignoring the network topology never comes close to *SP*. Note that the impact of choosing an appropriate scheduling strategy is striking: relative differences are bigger than comparing different site dimensioning approaches.

V. Conclusion

We outlined a dimensioning approach to calculate both server site capacities and network dimensions for optical Grids. Using the methodology, combining analytics and simulation, we compared various site dimensioning strategies and job scheduling algorithms in terms of resource requirements. The impact of the scheduling mechanism on the required bandwidth is striking, and non-optimal job allocation in terms of processing resources (*SP* has lowest local processing %) is more than compensated by optimizing network use. Concerning dimensioning strategies, we found that *prop* leads to the cheapest network (lowest bandwidth), slightly outperforming the *lloss* strategy which pays a small price for achieving more fairness (i.e. less variation in local processing % amongst different sites).

Acknowledgement

Part of the work presented in this paper was supported by the EU through the IST Project Phosphorus (www.ist-phosphorus.eu). C. Develder thanks the Research Foundation – Flanders (FWO–Vlaanderen) for his post-doctoral fellowship. M. De Leenheer thanks the IWT for financial support through his Ph. D. grant.

References

[1] D. Simeonidou, et al., "Dynamic optical network architectures and technologies for existing and emerging Grid services," *J. Lightwave Techn*., vol. 23, no. 10, Oct. 2005, pp. 3347- 3357.

[2] P. Thysebaert, F. De Turck, B. Dhoedt, P. Demeester, "Using divisible load theory to dimension optical transport networks for Grid excess load handling," *Proc. Optical Fiber Commun. Conf. (OFC 2005)*, Anaheim, CA, USA, 6-11 Mar. 2005.

[3] F. Fahramand, et al., "A Multi-layered approach to Optical Burst-Switched based Grids," *Proc. 5th Int. Workshop on Optical Burst/Packet Switching (WOBS)*, Boston, MA, USA, Oct. 2005.

[4] M. De Leenheer, et al., "A view on enabling consumer oriented Grids through Optical Burst Switching," *IEEE Commun. Mag*., vol. 44, no. 3, Mar. 2006, pp. 124–131.

[5] C. Develder, M. De Leenheer, T. Stevens, P. Thysebaert, J. Baert, B. Dhoedt, Piet Demeester, "Delivering the grid promise with optical burst switching," (invited), *Proc. OBS Workshop at COIN-NGNCON*, Jeju, Korea, 8 Jul. 2006.

[6] K. Christodoulopoulos, M. Varvarigos, C. Develder, M. De Leenheer, B. Dhoedt, "Job Demand Models For Optical Grid Research," *submitted to ONDM 2007*.

[7] A. Zalesky, H. Le Vu, M. Zukerman, J. White, "Blocking probabilities of optical burst switching networks based on reduced load fixed point approximations," *Proc. Infocom 2003*, San Francisco, CA, USA, 30 Mar.-3 Apr. 2003, vol. 3, pp. 2008–18.