Scalable Impairment-Aware Anycast Routing in Multi-Domain Optical Grid Networks

C. Develder[1], M. De Leenheer[1], T. Stevens[1], B. Dhoedt[1], G. Markidis[2] and A. Tzanakaki[2]
1: Dept. of Information Technology, Ghent University – IBBT, G. Crommenlaan 8 bus 201, 9050 Gent, Belgium
2: Athens Information Technology (AIT), P.O. Box 68, Markopoulo Ave., 19002 Peania, Athens, Greece
Tel: +32 9 3314961, Fax: +32 9 3314899, e-mail: chris.develder@intec.ugent.be, atza@ait.org

ABSTRACT

In optical Grid networks, the main challenge is to account for not only network parameters, but also for resource availability. Anycast routing has previously been proposed as an effective solution to provide job scheduling services in optical Grids, offering a generic interface to access Grid resources and services. The main weakness of this approach is its limited scalability, especially in a multi-domain scenario. This paper proposes a novel anycast proxy architecture, which extends the anycast principle to a multi-domain scenario. The main purpose of the architecture is to perform aggregation of resource and network states, and as such improve computational scalability and reduce control plane traffic. Furthermore, the architecture has the desirable properties of allowing Grid domains to maintain their autonomy and hide internal configuration details from other domains. Finally, we propose an impairment-aware anycast routing algorithm that incorporates the main physical layer characteristics of large-scale optical networks into its path computation process. By integrating the proposed routing scheme into the introduced architecture we demonstrate significant network performance improvements.

Keywords: Grid computing, routing algorithms, optical networks, physical impairments, anycast routing.

1. INTRODUCTION

Today, the need for network systems to support storage and computing services for science and business, is often satisfied by relatively isolated computing infrastructure (clusters). Migration to truly distributed and integrated applications requires optimization and (re)design of the underlying network technology to create a Grid platform for the cost and resource efficient delivery of network services with substantial data transfer, processing power and/or data storage requirements. Optical networks offer an undeniable potential for the Grid, given their proven track-record in the context of high-speed, long-haul, networking. Not only eScience applications dealing with large experimental data sets (e.g. particle physics) but also business/consumer oriented applications can benefit from optical Grid infrastructure [1]: both the high data rates typical of eScience applications and the low latency requirements of consumer/business applications (cf. interactivity) can effectively be addressed.

When using transparent WDM as such network technology, signals are transported end-to-end optically without being converted to the electrical domain in between. Connection provisioning of all-optical connections (lightpaths) between source and destination nodes is based on specific routing and wavelength assignment algorithms (RWA). Traditional RWA schemes only account for network conditions such as connectivity and available capacity, without considering physical layer details. However, in transparent optical networks covering large geographical areas, the optical signal experiences the accumulation of physical impairments through transmission and switching, possibly resulting in unacceptable signal quality [2]. Considering that the future optical network is likely to have a much larger domain of transparency and higher bit-rates (40Gbps is a given and 160Gbps is on the horizon), it will therefore require more detailed consideration of the physical impairments in RWA algorithms, leading to so-called Impairment Aware RWA [3].

Another emerging and challenging task in distributed and heterogeneous computing environments, is job scheduling: when and where to execute a given Grid job, based on the requirements of the job (for instance a deadline and minimal computational power) and the current state of the network and resources. Traditionally, a local scheduler optimizes utilization and performance of a single Grid site, while a meta-scheduler is distributes workload across different sites. Current implementations of these (meta-)schedulers only account for Grid resource availability [4]. A different approach can be identified in the concept of anycast routing, optimizing both network and resource usage concurrently. This requires the routing algorithm to be aware of both network and resource states, which forms a major scalability issue in large-scale networks with a multitude of resources.

In this paper we propose a novel architecture to support impairment-aware anycast routing for large-scale optical Grid networks. Section 2 discusses general approaches to support multi-domain networks. We then proceed to introduce a novel architecture, which can provide anycast Grid services in a multi-domain scenario (Section 3). Simulation analysis is used to demonstrate the improved scalability without incurring significant performance loss. Furthermore, Section 4 shows how to incorporate physical layer impairments, to further improve the performance of optical Grid networks. Conclusions are presented in Section 5.

2. MULTI-DOMAIN ROUTING

To efficiently manage large-scale networks, they are generally composed of smaller sub-networks, usually referred to as domains. The control and management of a single domain is performed locally, and information concerning state and availability is in general not shared with other domains. Special agreements such as Service Level Agreements (SLAs) are usually required between different domains to create peering connections and allow transit data transfers. The domain size and heterogeneity make it difficult to collect all information needed to make optimal decisions for multi-domain control and management. In general, two extreme approaches are possible for the control of such networks, each having specific advantages and disadvantages:

- *Centralized*: A single control entity is aware of the full network and resource state of the multi-domain network, receiving all communication requests and responsible for all scheduling decisions. The main strengths of this approach are its straightforward deployment and reconfiguration possibilities. However, this approach is not scalable for large-scale networks, both in terms of control traffic and computational complexity inherent in job scheduling. Also interoperability issues may arise since domains may have different control plane protocols. Furthermore, confidentiality policy violation may arise, since each optical grid site advertises all network and resource state and configurations to all parties involved.
- *Fully Decentralized*: Resources send updates to all clients directly and clients individually perform the network and resource scheduling. An important assumption is that this approach requires total transparency between domains, which in reality is difficult to achieve. It also implies that the number of status updates sent between each client-resource pair will increase dramatically compared to the centralized setup. An advantage of this setup is the removal of the single point of failure (the centralized scheduler).

In the following section, we propose an alternative to these approaches, which tries to combine the advantages of both techniques while minimizing the relevant issues.

3. ANYCAST PROXY ARCHITECTURE

The architecture is based on the use of proxy servers, which form an overlay to control job scheduling and routing in a multi-domain optical Grid network. Two types are available: *client proxies* contacted by clients for routing and scheduling a job request, and *resource proxies* collecting resource states and sharing this in an aggregated form with the client proxies. A resource only forwards state information to its closest resource proxy using anycast communication. Typically, this proxy belongs to the same domain as the resource node, and from the resource perspective it also behaves like a local scheduler. Likewise, a client with a job to submit to the multi-domain Grid, forwards its request to the nearest client proxy using anycast communication. Upon reception of the job request, the client proxy selects the most suitable resource proxy to forward the request to, based on aggregated information. Communication between proxy pairs can use a general inter-domain routing algorithm: Section 4 presents such an algorithm specifically aimed at exploiting physical layer information.

Using this proxy-based approach has the following benefits:

- Allows domain information hiding, and domain-specific policies for client and resource proxies;
- Control plane scalability: the intelligent state aggregation results in reduced control plane traffic;
- Scalability of scheduling complexity: client proxies base their selection of a suitable resource proxy on aggregated values, while the resource proxy is responsible only for resources in its own domain;
- Optimization over all interconnected domains of the Grid network is possible, and considers both network and resource states. Examples are minimal job blocking, global load-balanced resource utilization, etc.;
- Can support any subset of parameters available to the routing protocol, i.e. computational resource states, physical parameters of photonic network, etc.

Previous work has studied both optimal (based on Integer Linear Programming) and heuristic algorithms to optimally dimension the proxy infrastructure [2]. The proposed algorithms perform concurrent placement of proxy servers and determination of their capacities, and results show that a small number of proxy servers suffices even in large-scale networks, implying that it is a practical and efficient system.

3.1 Performance Analysis

To further motivate the efficiency of the proposed architecture, we performed discrete event simulation for the three aforementioned control plane scenarios. The simulations are based on a realistic network and job scenario, based on the multi-domain Phosphorus network. At each edge node, the number of clients and resources is chosen to be proportional to the number of inter-domain links. Intra-domain topologies are abstracted to simplify the simulation setup: resources and clients are connected to the domain edge node by an aggregation tree. The job model is configured in a similar way as described in [6]; job duration is assumed to be distributed hyper-exponentially and the job inter-arrival times follow an exponential distribution (i.e. Poisson arrival process). Furthermore, we assume that resources update their state information at a frequency higher than the one used by proxies. Since proxies generally aggregate the state information of multiple resources, their overall rate of change is much lower, thus allowing a lower state update frequency. For the simulation results shown, we assumed the resource update interval is half of the resource update interval.

Results for the job loss rate related to the job IAT (and corresponding average generated system load on the second axis) are depicted in Figure 1. We can conclude that there is no significant difference in job acceptance rate between the three alternative approaches; the less-frequent distribution of aggregated resource state by the proxy system does not prevent efficient resource allocation.

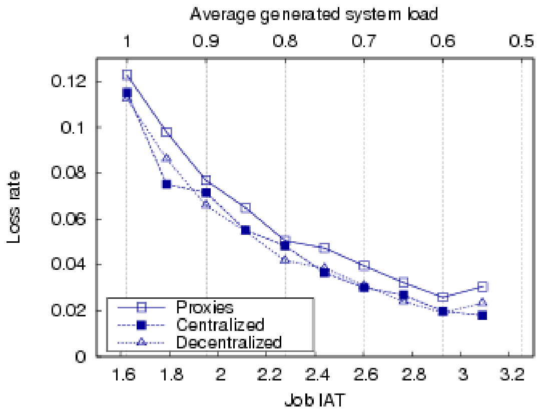

Figure 1: Job loss rate for varying IAT (load).

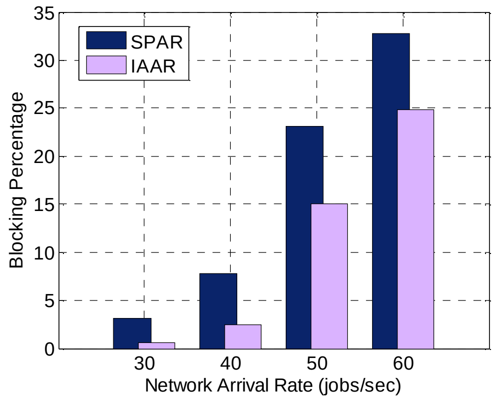

Figure 2: Blocking percentage of anycast routing algorithms with (IAAR) and without (SPAR) impairment-awareness

4. IMPAIRMENT-AWARE ANYCAST ROUTING

As the optical signal propagates through an all-optical path, it experiences the impact of a variety of effects that introduce different types of signal distortions, both linear and non-linear. Linear effects such as amplified spontaneous emission (ASE) noise, polarization mode dispersion PMD, chromatic dispersion (CD), in-band crosstalk (XT), and filter concatenation (FC), are independent of the signal power and affect each of the optical channels individually. In contrast, nonlinear impairments like self-phase modulation (SPM), cross-phase modulation (XPM) and four wave mixing (FWM) affect not only each optical channel separately but they also cause disturbance and interference between them. The induced signal distortions that are considered for this work can be categorized in distortions of almost “deterministic” type related only to the single channel’s pulse stream, such as the interplay of SPM and group velocity dispersion (GVD) or the optical filtering introduced by the (de)multiplexing (MUX/DEMUX) elements at the optical cross-connects/optical add drop multiplexers (OXC/OADMs). The other category includes degradations of pertubative nature, introduced by ASE noise and the WDM nonlinearities (FWM and XPM).

This work demonstrates an algorithm for anycast routing which takes the physical performance of the optical network into consideration. More specifically not only the availability of optical connections but also the quality of these connections in terms of the quality factor Q is considered before they can be established. The analytical Q-factor introduced in [6] for the performance evaluation of a static unicast IA-RWA, has been used to integrate different types of degradations and thus to reflect the overall signal quality in the context of anycast routing. For Q-factor evaluation, we ignored the interplay among the different types of degradations and pertubative nature distortions (ASE, XPM, FWM) are assumed to follow a Gaussian distribution. SPM/GVD and optical filtering effects were introduced through an eye closure penalty metric calculated on the most degraded bit-pattern.

4.1 Algorithm

Initially the dynamic Impairment Aware Anycast Routing (IAAR) algorithm collects all relevant physical and network layer parameters required for the Q-penalty evaluation of each bidirectional link. Upon a job request arrival the algorithm performs the following steps:

1. All available Grid resources in other domains are identified, and considered as possible destinations.
2. A path from the client domain to each available Grid resource domain is calculated according to the Dijkstra algorithm for which Q-penalties are assigned as weights. For each resource domain W paths are tried to be discovered, one for each wavelength according to their availability (where W is the link capacity).
3. For each possible destination, the lightpath which has the lowest Q-penalty is chosen and passed to the next step (the maximum number of paths chosen in this phase equals the number of Grid resource domains).
4. Only lightpaths (one for each possible destination) with BER < 10^{-15} are considered as candidates to accommodate the job request.
5. The first least-loaded lightpath (the least-loaded with the minimum wavelength number: first fit wavelength assignment scheme is used) is chosen and its network resources as well as the specific destination Grid

resources are reserved for a certain job execution time. (Least-loaded path is that where the minimum number of free wavelengths over all its links is the maximum such number over all paths.)

To compare the proposed IAAR algorithm, we compare it with a straigthforward Shortest Path Anycast Routing (SPAR) algorithm has the following changes. In Step 2 the Dijkstra algorithms considers link lengths as weights instead of the Q-penalties for the path computation process. Also, instead of selecting a lightpath based on the minimum Q-penalty (for each wavelength), the shortest path (in distance) is selected (Step 3). A similar approach is used for Step 5. Finally, note that the job is considered lost if its BER value is greater than 10^{-15}.

4.2 Performance Analysis

The performance of the IAAR and SPAR algorithms has been tested under a practical optical Grid scenario; the Phosphorus network topology was used, while considering the job model described in [6] and assuming the number of clients and Grid resources in each domain to be proportional to the inter-domain links. The number of wavelengths was set to 16 on each link with a channel spacing of 50MHz, and the Q-penalties are calculated for the middle channel, assuming that all channels are occupied (for the non-linear impairments that are wavelength dependant). Also dual-stage EDFAs of various noise figures were assumed for a span length of 80km [3]. Finally a certain dispersion map was considered [8] with inline dispersion equal to 20 ps/nm and pre-compensation of 600 ps/nm for all fiber links.

As Figure 2 reveals, significant performance improvement are achieved by including physical impairments into the anycast routing decision. The benefit ranges from 3 to 8% depending on the network load.

5. Conclusions

This paper discussed job routing and scheduling challenges in a multi-domain, optical Grid environment. We presented a novel anycast proxy architecture, which extends the anycast principle in a multi-domain scenario. The main purpose of the architecture is to perform aggregation of resource and network states, and as such improve computational scalability and reduce control plane traffic. Simulation analysis was used to demonstrate the improvement in control plane scalability, without significant decrease in performance. Furthermore, we showed that this approach is especially useful for Grids based on optical networks, by incorporating physical impairments into the routing protocol. Results indicate that the introduction of physical layer parameters into the routing algorithm can considerably improve anycast routing performance for optical Grid scenarios.

ACKNOWLEDGEMENTS

Part of this work has been funded by the EU through the FP6 IST Phosphorus project (www.ist-phosphorus.eu) and the FP6 NoE ePhoton/ONe+ project (www.e-photon-one.org). C. Develder is supported by the Research Foundation Flanders (FWO, www.fwo.be) as a postdoctoral fellow. M. De Leenheer thanks the IWT (www.iwt.be) for his Ph.D. grant.

REFERENCES

[1] M. De Leenheer, P. Thysebaert, B. Volckaert, F. De Turck, B. Dhoedt, P. Demeester, D. Simeonidou, R. Nejabati, G. Zervas, D. Klonidis, and M.J. O'Mahony, "A View on Enabling Consumer Oriented Grids through Optical Burst Switching", IEEE Communications Magazine, 44(3):124-131, Mar. 2006.

[2] I. Tomkos, D. Vogiatzis, C. Mas, I. Zacharopoulos, A. Tzanakaki, E. Varvarigos, "Performance engineering of metropolitan area optical networks through impairment constraint routing", IEEE Communications Magazine, 42(8): S40-S47, Aug. 2004.

[3] G. Markidis, S. Sygletos, A. Tzanakaki and I. Tomkos, "Impairment Aware Based Routing and Wavelength Assignment in Transparent Long Haul Optical Networks", Proc. Conf. on Optical Network Design and Modeling (ONDM), May 2007.

[4] K. Czajkowski, I. Foster, N. Karonis, C. Kesselman, S. Martin, W. Smith, S. Tuecke, "A Resource Management Architecture for Metacomputing Systems", Proc. IPPS/SPDP Workshop on Job Scheduling Strategies for Parallel Processing, pp. 62-82, Mar .1998.

[5] S. Figuerola, N. Ciulli, M. De Leenheer, Y. Demchenko, W. Ziegler, A. Binczewski, "PHOSPHORUS: Single-step on-demand services across multi-domain networks for e-science", Proc. SPIE Asia-Pacific Optical Communications (APOC), Wuhan, China, Nov. 2007.

[6] K. Christodoulopoulos, M. Varvarigos, C. Develder, M. De Leenheer, B. Dhoedt, "Job Demand Models for Optical Grid Research", Proc. Conf. on Optical Network Design and Modelling (ONDM), May 2007

[7] T. Stevens, M. De Leenheer, C. Develder, F. De Turck, B. Dhoedt, P. Demeester, "ASTAS: Architecture for Scalable and Transparent Anycast Services", Accepted for publication in Journal of Communications Networks.

[8] I. Tomkos, M. Vasilyev, J.-K. Rhee, A. Kobyakov, M. Ajgaonkar, and M. Sharma, "Dispersion map design for 10 Gb/s ultra-long-haul DWDM transparent optical networks," at the OECC July 2002, PD-1-2.