Optical grid networking exploiting path computation element (PCE) architecture

Filippo Cugini

National Laboratory for Photonic Networks,
Consorzio Nazionale Interuniversitario per le Telecomunicazioni
(CNIT), 56124 – Pisa, Italy
Fax: +39-050-5492250
E-mail: filippo.cugini@cnit.it

Sugang Xu and Hiroaki Harai

Network Architecture Group,
New Generation Network Research Center,
National Institute of Information and
Communications Technology (NICT),
4-2-1 Nukui-Kitamachi,
Koganei, Tokyo, 184-8795, Japan
Fax: +81-42-327-6680
E-mail: xsg@nict.go.jp
E-mail: harai@nict.go.jp

Francesco Paolucci, Luca Valcarenghi and Piero Castoldi*

Center of Excellence for Information,
Communication and Perception Engineering (CEIICP),
Scuola Superiore Sant'Anna, 56124 – Pisa, Italy
Fax: +39-050-5492250
E-mail: fr.paolucci@sssup.it
E-mail: l.valcarenghi@sssup.it
E-mail: castoldi@sssup.it
*Corresponding author

Abstract: A dynamic and joint optimisation of computational and network resources may significantly improve the performance of grid-enabled applications. In this paper, the path computation element protocol (PCEP) is first proposed as standard grid network service interface between grid resource manager (GRM) and network resource manager (NRM). Then, two different schemes exploiting the PCEP protocol and the PCE capability to provide synchronised and optimal path computations are proposed. The schemes allow the GRM to exploit performance metrics representative of the expected network resource utilisation and to perform a jointly optimal choice of both computational and network resources.

Simulation results show that the proposed schemes significantly improve the overall amount of successfully established grid services. Experimental

implementations on the 'JGN2plus' testbed are presented to show that the proposed schemes avoid control plane extensions or interfaces specifically designed for grid purposes and do not substantially affect the overall grid service delivery time.

Keywords: grid computing; grid network service; GNS; GMPLS; path computation element; PCE; PCE protocol; PCEP; wavelength-switched optical network; WSON.

Reference to this paper should be made as follows: Cugini, F., Xu, S., Harai, H., Paolucci, F., Valcarenghi, L. and Castoldi, P. (2010) 'Optical grid networking exploiting path computation element (PCE) architecture', *Int. J. Communication Networks and Distributed Systems*, Vol. 5, No. 3, pp.246–262.

Biographical notes: Filippo Cugini received his MS in Telecommunication Engineering from the University of Parma, Italy. Since 2001, he has been with the National Laboratory of Photonic Networks, Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Pisa, Italy. In 2007 and 2008, he was a Visiting Researcher at the National Institute of Information and Communications Technology (NICT), Japan. His main research interests include theoretical and experimental studies in the field of optical communications. In particular, the focus is on ethernet, GMPLS and PCE protocols and architectures, survivability and traffic engineering in IP over WDM networks, and grid networking. He is the co-author of more than 60 publications in international conferences and journals.

Sugang Xu received his BE and ME in Computer Engineering from Beijing Polytechnic University, Beijing, China, in 1994 and 1997, respectively, and his PhD in Information and Communication Engineering at the University of Tokyo, Tokyo, Japan, in 2002. He joined the Global Information and Telecommunication Institute, Waseda University, in 2002 as a Research Associate. Since 2005, he joined the National Institute of Information and Communications Technology (NICT), Tokyo, Japan, as an Expert Researcher. His research interests include algorithms, network architectures, photonic network control, optical grid network systems, and parallel and distributed processing. He is a member of IEEE and IEICE.

Hiroaki Harai is currently a Group Leader of the National Institute of Information and Communications Technology (NICT), Tokyo, Japan, where he is leading the AKARI Architecture Design Project for New Generation Network. His research interest includes design of new generation network architecture and optical circuit and packet switch integrated network. He received his ME and PhD in Information and Computer Sciences from Osaka University, Osaka, Japan. He was elected Outstanding Young Researcher in the 3rd IEEE ComSoc Asia-Pacific Young Researcher Award. He is concurrently a Visiting Associate Professor at The University of Electro-Communications, Tokyo, Japan.

Francesco Paolucci received his MS in Telecommunications Engineering from the University of Pisa, Italy, and his PhD in Information and Communication Technologies from the Scuola Superiore Sant'Anna, Pisa, Italy. He is currently a Junior Researcher with Scuola Superiore Sant'Anna, Pisa, Italy. His main research interests include simulative and experimental activities in the optical communications and network services areas, mainly focused in traffic engineering in GMPLS-based networks, PCE architectures in multi-layer and multi-domain networks, grid network services, multi-layer network reliability

for QoS-based services, integration between GMPLS-based networks and WDM-PON-based metro-access networks.

Luca Valcarenghi holds a Laurea degree in Electronics Engineering (1997) from Politecnico di Torino, Italy, an MS in Electrical Engineering (1999), and a PhD in Electrical Engineering Telecommunications (2001) both from the University of Texas at Dallas (UTD). Since September 2002, he is an Assistant Professor at the Scuola Superiore Sant'Anna of University Studies and Doctoral Research of Pisa, Italy. His main research interests are optical networks design, analysis, and optimisation; communication networks reliability; IP over WDM networking; QoS in network infrastructures for grid computing; fixed and mobile network integration; energy saving in communications networks

Piero Castoldi is an Associate Professor in Telecommunications at the Scuola Superiore Sant'Anna in Pisa. Since January 2005, he is the Director of the CNIT National Photonic Networks Laboratory in Pisa, Italy. He is a Research Leader of the 'Networks and Services' Area at the Integrated Research Center for Photonic Networks and Technologies (IRCPhoNeT), a research centre jointly supported by Scuola Superiore Sant'Anna, CNIT and Ericsson. His scientific activity has covered the areas of digital transmission, telecommunication networks and systems both wired and wireless. He is the author of more than 170 publications in international journals and conference proceedings.

1 Introduction

The increased capacity of networks connecting geographically distributed computational and storage resources is pushing grid computing from local area networks (LANs), i.e., local or cluster grid, to wide area networks (WANs), i.e., global grid. However, while cluster grid-enabled applications may generally exploit almost dedicated network resources, in global grid network resources are shared among different, even non-grid, applications with possibly contending quality of service (QoS) requirements. Moreover, in global grid, the large distances and the heterogeneity of network resources may additionally contribute to provide unpredictable network behaviour, affecting the whole grid performance in terms of end-to-end transmission delay, service blocking probability and overall service delivery time.

To meet the QoS requirements of global grid-enabled applications, a dynamic and deterministic joint optimisation of both computational and network resources (Travostino et al., 2006; Foster and Kesselman, 2004; Valcarenghi et al., 2006) is, therefore, necessary. The joint optimisation is particularly needed and potentially beneficial, when bandwidth-greedy applications require the provisioning of dedicated high-speed connections (i.e., lightpaths) among grid resources connected through a wavelength-switched optical network (WSON). A way of achieving this joint optimisation is the cooperation between the grid resource manager (GRM) and the network resource manager (NRM).

The recent introduction of the path computation element (PCE) within the control plane of optical networks represents an appealing solution to provide efficient utilisation of network resources. The PCE is defined as a functional element devoted to

constraint-based path computation (Farrel et al., 2006). Within the PCE-based architecture, the PCE communication protocol (PCEP) (Vasseur and Le Roux, 2008; Bradford et al., 2008) has been defined to allow a path computation client (PCC) to require the PCE to perform path computation. In particular, the PCEP 'PCReq' message is defined to carry the path computation requests from the PCC to the PCE while the PCEP 'PCRep' message carries the replies computed by the PCE.

In this paper, we propose to exploit the PCEP protocol as the standard interface between the GRM, which behaves as PCC, and the NRM, which embeds a PCE implementation. Then, we propose two novel schemes for grid resource selection specifically designed to exploit the standard PCEP features. The schemes resort to the currently defined PCE architecture and PCEP implementation to provide a feedback to the GRM on the expected network resources utilised by alternative choices of computational resources. In this way, the GRM can evaluate in advance the impact on the network performance of different alternatives and, among them, select those that minimise the grid service delivery time and the overall network resource utilisation.

We evaluate through simulations the benefits in terms of network resource utilisation provided by the proposed grid networking schemes exploiting the PCE capability to perform concurrent and synchronised path computations. Moreover, we extensively detail the implementation of the proposed PCEP-based schemes in the WSON 'JGN2plus' network testbed (http://www.jgn.nict.go.jp/). The proposed schemes are particularly suitable in case of grid services for e-science applications [e.g., electronic very-long-baseline interferometry (e-VLBI)] (Xu, 2007), that require high bandwidth connections and high level of connectivity among multiple grid resources (e.g., full mesh of lightpaths) and expect significant time duration (e.g., tens of minutes or above). Differently from current network-aware grid implementations, the proposed solution does not require the introduction of new control plane extensions or interfaces specific for grid purposes. In addition, the proposed schemes preserve the adequate level of confidentiality on detailed network information between GRM and NRM (i.e., link bandwidth availability and strict routes information are not disclosed by NRM). Therefore, the proposed PCEP-based solution is easily implementable at NRMs, including independent internet service providers (ISPs) not belonging to grid virtual organisations (VOs).

2 Grid network services

Within the Grid High Performance Networking Research Group (GHPN-RG) of the Open Grid Forum (OGF), grid network services (GNS) have been proposed (Clapp et al., 2004; Ferrari, 2005) for enabling the joint optimisation of both computational and network resources. GNS are grid services that handle the monitoring and the allocation of grid network resources. Among the proposed services, the informational GNS (also referred to as network information and monitoring service – NIMS) provides the grid middleware with an updated snapshot of the network resource status.

Two classes of NIMS can be identified. The first class of NIMS resorts to distributed methodologies and network sensors for measuring network performance. Such network measurements are triggered directly by the GRM and do not require the cooperation with the NRM, which can be an independent ISP. Examples of existing network sensors include the network weather service (NWS) (Wolski et al., 1999), a distributed tool that periodically monitors and dynamically forecasts the performance of various network and

computational resources over a given time interval. Currently, the tool includes sensors for end-to-end TCP/IP performance (bandwidth and latency), available CPU percentage and available non-paged memory. 'TopoMon' (Burger et al., 2002) improves NWS by providing topology, link bandwidth and link latency information. Information on the network topology and link latency is obtained through 'traceroute'. The European Enabling Grids for E-sciencE (EGEE) project (http://www.eu-egee.org) has developed its own architecture for network monitoring. They focus on collecting the following performance measurements: instantaneous connectivity of an internet path, packet loss, two-way delay (i.e., round trip time), TCP throughput, domain name system (DNS) lookup, port scan. The following sensors are used: 'PingER' for two way delay and loss measurements, 'UDPmon' for one way delay and one way loss measurements, 'Bandwidth Test Controller (BWCTL)' for TCP throughput. In addition to proposed grid monitoring tools, several distributed monitoring systems for the internet exist that either measure network performance or explore network topology (Huffaker et al., 2002; Ficara et al., 2007; http://www.caida.org/projects/ark/). However, all the current tools exploiting distributed methodologies or distributed sensors suffer from several drawbacks. For example, they provide data about the nodes involved in the monitoring efforts only. Such information could not be sufficient if many sites communicate with each other simultaneously. Indeed if two pairs of communicating sites are contemporarily sharing some common links, network sensors typically predict performance for each pair separately. In addition, collected data may not precisely represent the actual network resource availability. Moreover, and most important in this study, they cannot be applied in WSON since the lightpath connection is only activated at the time of service set up.

The second class of NIMS implementations resorts to direct cooperation between GRM and NRM to retrieve network information. The various solutions and projects proposed so far (Zervas et al., 2008; Lehman et al., 2006; Takefusa et al., 2006; Battestilli et al., 2007; Habib et al., 2006; ARGON Specification, 2005; Baraglia et al., 2006; Tomkos et al., 2007; Palmieri, 2006; Danelutto, 2004; De Leenheer et al., 2006) can be typically classified according to two different approaches.

In the first approach, the NRM implements, at the network layer, the exchange of computational layer related information. In particular, computational information is introduced at the network layer through extensions of the network control plane, i.e., the routing and/or signalling protocol. An example of such approach is pursued in the PHOSPHOROUS project (Zervas et al., 2008). In Zervas et al. (2008), novel extensions to the OSPF-TE routing protocol are proposed to advertise the presence of grid sites, grid services, computing elements, sub-clusters and storage elements. In addition, extensions to the RSVP-TE protocol are proposed in the CALL procedure to encompass information associated to the grid job description. A further example is represented by the DRAGON project (Lehman et al., 2006) where the OSPF-TE protocol is extended with authentication, authorisation, accounting (AAA) information and scheduling parameters. Moreover, non-standard topology aggregation information is exchanged among different domains to cover the set up of services traversing multi-domain networks. In this way, the control plane has the capability to serve grid users' requests by computing and establishing end-to-end connections that satisfy both computational and network requirements. The approach based on control plane extensions is particularly suitable in all the cases where the NRM belongs to the grid VO and a single network domain is present and organised in a single routing area. However, in different scenarios, for instance where the NRM is an independent ISP, the required control plane extensions

may be hardly supported. In addition, their implementation will be available only upon full support of router vendors which may difficultly introduce control plane extension (with the related potential convergence and scalability issues) just for a specific application (i.e., grid).

In the second approach, the GRM retrieves network information directly from the NRM through the definition and implementation of specific interfaces. An implementation of such approach is provided in the G-LAMBDA project (Takefusa et al., 2006) where a web service interface is defined between the GRM and the NRM. The communication between GRM and NRM is achieved through SOAP/HTTP-based communication. Exchanged information encompasses network resource availability information provided by NRM to GRM and network reservation requests triggered by the GRM to NRM. However, no performance metrics are so far exchanged between the GRM and the NRM, thus preventing the possibility to perform joint optimisation of both computational and network resources. Moreover, even though a preliminary standardisation process of such interface has started within the Open Grid Forum, its adoption within the NRMs, particularly by independent ISPs not belonging to grid VO, seems still quite far to be achieved.

In this paper, we focus on this second approach. The proposed solution resorts to the currently defined PCEP protocol as standard interface between the GRM and the NRM. Thus, differently from previous solutions, no control plane extensions or interfaces between GRM and NRM are specifically required for grid purposes. This approach has the advantage of operating over optical networks with 'legacy' control plane. The proposed PCEP-based solution, however, suffers from the main limitation that currently, no advance reservation for scheduling purposes has been standardised (e.g., start and end time for a required network connection). However, such capability may be introduced in future PCEP versions for general applications, i.e., without requiring specific protocol extensions for grid purposes.

3 PCE architecture

This section briefly summarises the main concepts and definitions that characterise the PCE architecture and the PCEP protocol that will be utilised in the following sections. The PCE is defined as an entity that is capable of computing a network path or route based on a network graph, and of applying computational constraints during the computation (Farrel et al., 2006). The PCEP protocol (Vasseur and Le Roux, 2008) defines the communication between the PCC and the PCE. The PCEP 'PCReq' message allows the PCC to request the PCE to perform path computation under various constraints and preferences, such as end points (i.e., the source and destination of the path to be computed), bandwidth requirements, set up and holding priorities, objective functions, link and node diverse path computation for multiple (synchronised) requests. Additional requirements may be considered in WSON. For example, whether the same wavelength assignment is strictly required for primary and backup paths or in both directions of a bidirectional path.

The PCEP 'PCRep' message allows the PCE to return to the PCC the results of the path computation. It includes the explicit route object (ERO) indication and additional optional information such as metric values and, in case of WSON, the computed wavelength to use during the signalling process of the lightpath set up. In particular, for

what concerns the ERO information, three possibilities are available: the ‘strict ERO’, the ‘loose ERO’ and the ‘encrypted ERO’. The strict ERO details the sequence of nodes to traverse. The loose ERO specifies just few nodes of the computed path (e.g., end points) without providing details on the strict route indications. The encrypted ERO exploits the path key (PK) (Bradford et al., 2008) extensions to preserve the strict ERO confidential at the PCC but avoiding re-computations in case of successive connection set up.

Additional PCEP messages (Vasseur and Le Roux, 2008) are defined to set up and terminate the PCEP session (i.e., PCEP ‘OPEN’ and ‘CLOSE’ messages) and to maintain active the established session (i.e., PCEP ‘KEEPALIVE’ message).

4 PCEP-based grid networking

The typical objective function applied by GRM consists in the maximisation of the overall amount of grid services successfully established and in the minimisation of each service delivery time subject to the required QoS constraints. Service delivery time is defined here as the time required to complete the computational task included the time required to complete the data transfer among the involved grid resources. Thus, all the VO computational resources (e.g., CPUs) as well as the network resources (e.g., available bandwidth) connecting the grid resources should be considered in the service allocation. However, current GRMs typically rely just on computational resource information to select the grid resources to exploit. When a WSON network providing lightpath connections is considered, the computation of each service delivery time is simplified given the fixed amount of bandwidth made available to grid resources, i.e., the fixed time required to send all grid data. In this way, the minimisation of each service delivery time is roughly achieved even if network resources are not considered. However, the lack of network information at the GRM does not guarantee the efficient network resource utilisation and, in turn, the maximisation of the amount of established grid service.

4.1 Grid networking scheme not utilising the PCE (Dg)

The typical scheme currently implementable by GRM, here referred to as *Dg*, first selects the least loaded *g* grid resources and performs ‘distributed’ path computation among them (no joint optimisation is provided). In *Dg*, the GRM performs grid resource selection according to information available at the computational layer only, i.e., among the available *G* grid resources, the set of *g* least loaded grid resources is selected. In case of equally loaded grid resources (according to a pre-defined load granularity), a random choice is performed. Then, the NRM is requested to establish the necessary connections among the *g* network nodes to which grid resources are connected. The NRM is responsible for lightpath activation, i.e., it has the rights to access ingress nodes and configure them for lightpath set up by specifying also the ERO. Lightpath set up is achieved by exploiting the GMPLS protocol suite. The PCE is not utilised and a distributed set up is exploited: ingress nodes are independently configured with just the indications on the path destinations (loose ERO is provided). Network resource information provided by the routing protocol allows the ingress node to locally compute the strict routes towards the destinations. The signalling protocol is then responsible for

the connection set up. The signalling protocol also identifies the specific wavelength to use by exploiting the currently defined GMPLS features (e.g., label set, acceptable label set, crankback).

4.2 Grid networking scheme utilising the PCE (Pg)

An alternative scheme, here referred to as *Pg*, is considered when the NRM exploits the PCE. As in *Dg*, in *Pg*, the GRM selects the least loaded *g* grid resources and requests the NRM to perform lightpath activation (no joint optimisation is provided). The difference with respect to *Dg* is that, in this case, NRM resorts to the PCE to perform multiple concurrent and synchronised path computations. The PCEP protocol is used only between the NRM and the PCE. The PCE has full visibility of network resources and provides all strict route indications. The ingress nodes are then configured with the strict ERO information. In addition, when a WSON network is considered, the PCE may provide also indications on the specific wavelength to use thus avoiding conflicts during the GMPLS signalling phase and resource reservation. To provide such information, the PCE exploits either the detailed route information stored in its database (if the PCE works under stateful condition) (Farrel et al., 2006) or the detailed wavelength availability information advertised through the routing protocol (Otani, 2008).

4.3 PCEP-based scheme with feedback information (Pkg)

The first scheme proposed in this paper fully exploiting the PCE architecture is here referred to as *Pkg*. The *Pkg* scheme introduces a direct communication between the GRM (which behaves as PCC) and the PCE controlled by the NRM. The *Pkg* scheme selects the least loaded *g* grid resources and exploits PCE not only for concurrent path computation but also to provide a feedback on *k* grid resources ($k > g$). Upon a new grid service request, two steps can be identified within the *Pkg* scheme.

Step 1 In the first step, the least loaded *g* grid resources are first identified by the GRM. In case of equally least loaded grid resources, up to *k* equally least loaded resources are randomly selected. By utilising the PCEP communication, the GRM requests the PCE to perform the *p* path computations [e.g., $p = g(g-1)/2$ in case of full mesh] for every combination *c* of the required *g* out of the identified *k* resources [i.e., $c = k!/(g!(k-g)!)$ combinations]. Each of the *p* requests included within the PCEP 'PCReq' message is identified by: source, destination, bandwidth (e.g., equivalent to one lightpath) and additional optional parameters (e.g., protection). The PCE performs synchronised path computation and replies with a 'PCRep' message which includes, for each path computation request, the 'ERO information' and the 'metric value'. Metric values can be representative of QoS parameters (e.g., communication delay) or economic value of the occupied resources which can be considered, as done in this study, proportional to the amount of utilised network resources (e.g., number of occupied links). By exploiting the metric values, the GRM performs joint optimisation of both computational and network resources. In particular, the least loaded *g* grid resources that guarantee the minimum combined metric value are selected for service (i.e., lightpath) set up.

Step 2 The GRM then performs the second step of the *Pkg* scheme by providing the NRM with the specific *p* lightpath activation requests. Depending on the relationship between GRM and NRM, two scenarios can be identified for the definition of the ERO to be included by the PCE within the 'PCRep' message and to be forwarded by the GRM to the NRM. In the first scenario, mainly suitable in case of high level of cooperation between GRM and NRM (e.g., they belong to the same VO), the strict ERO is provided in its explicit form. In the second scenario, suitable in case of limited cooperation between GRM and NRM (e.g., NRM is an independent ISP which maintains network information confidential), the ERO is passed in its encrypted form by resorting to the path-key (PK) (Bradford et al., 2008) attribute. In this case, the PK decryption is then required at the NRM prior to lightpath activation. Finally, the lightpath set up is performed as in the *Pg* scheme. The *Pkg* scheme with $k = g$ provides the same results as the *Pg* scheme.

4.4 PCEP-based scheme with feedback information and a-priori communication (PGkg)

A further enhanced scheme exploiting the PCE architecture, here referred to as *PGkg*, is proposed. The *PGkg* scheme expands the *Pkg* scheme by introducing a further step (i.e., step-0) prior to the two steps that characterise the *Pkg* scheme.

The 'step-0' consists in an offline PCEP-based communication between the GRM and the PCE which takes place prior to all grid service requests. The PCE performs a non-synchronised independent path computation between all node pairs connecting the grid resources potentially involved in the grid service. The PCE replies with loose ERO information and the related computed metric values. The step-0 allows the GRM to be notified about static metric values associated with all the possible p' connections between the G grid resources (e.g., $p' = G(G-1)/2$). Each static metric value represents a lower bound of the actual metric value: it is computed by considering all network resources available, i.e., it is optimistic compared to the actual value used during the network operations. For example, such metric value is computed by considering the shortest path between the grid resources, which however may not be available in case of high network load.

The metric values provided by step-0 are used by the GRM upon new grid service requests occurrence. In particular, the GRM first identifies, among the set of least loaded grid resources, the k resources that minimise the considered metrics. In this way, the *PGkg* scheme avoids the possible random selection of the k least loaded grid resources performed in the *Pkg* scheme. Then, the *PGkg* scheme behaves as *Pkg* by applying the two steps required to identify and establish the grid service among the identified g grid resources.

Table 1 summarises the four considered grid networking schemes.

Table 1 Considered grid networking schemes

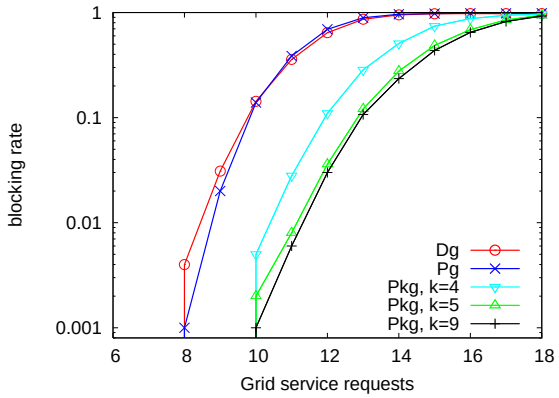
Scheme	*Description*	*GRM*	*NRM*	*Complexity (# path computations)*
Dg	Grid networking scheme not utilising the PCE	g	Distrib.	$p = g(g-1)/2$
Pg	Grid networking scheme utilising the PCE	g	PCE	$p = g(g-1)/2$
Pkg	PCEP-based scheme with feedback information	k (>g)	PCE	$cp = k!/(g!(k-g)!\ g(g-1)/2$
		g		$p = g(g-1)/2$
PGkg	PCEP-based scheme with feedback information and a-priori communication	G	PCE	$cp' = k!/(g!(k-g)!\ G(G-1)/2$
		k (>g)		$cp = k!/(g!(k-g)!\ g(g-1)/2$
		g		$p = g(g-1)/2$

Notes: The GRM column reports the amount of least loaded grid resources considered by GRM and for which the NRM is requested to perform either distributed or PCE-based path computation. 'Schemes *Pkg* and *PGkg* exploit two and three steps respectively'.

5 Numerical results

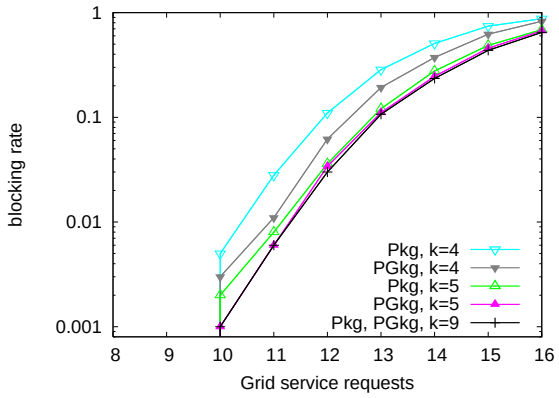
The *Dg*, *Pg*, *Pkg* and *PGkg* schemes have been evaluated through simulations considering two different network topologies: the 3 × 3 mesh and the NSF transparent network topology made of $N_1 = 9$ and $N_2 = 14$ nodes, respectively. Each link carries $W = 8$ wavelengths. One grid resource per node is considered (i.e., $G = N$). Each grid service s_i requires, at the network layer, a full mesh of lightpaths between $g = 3$ least loaded grid resources. Various values of g have been evaluated. Here the case with $g = 3$, and thus $p = 3$ bidirectional lightpaths, is reported. Services are sequentially established and never torn down. Grid resource load (e.g., computational load) has initial null value and it is incremented, upon service establishment, by a load l. Values of l are assumed for simplicity with fixed granularity, i.e., each service s_i randomly utilises a grid resource load equal to either $l = 1$ or $l = 2$ The PCE-based path computation is based on integer linear programming (ILP) formulation. In the considered scenario, ILP-based computation performs extremely fast; in larger network scenarios heuristics could be adopted to provide the best trade-off between optimality and scalability (Zang et al., 2000). However, since PCE is devoted to path computation, it is expected that the adopted algorithm will provide the optimal achievable solutions. The implemented objective function minimises the overall amount of used network resources and, as secondary objective function, minimises the load of the most loaded link. Service blocking (i.e., rejection) occurs if the network resources are insufficient to guarantee all the required p lightpath establishments. All simulation results are collected upon 1,000 independent trials.

Figure 1 Service blocking rate (see online version for colours)



Notes: *Dg*, *Pg* and *Pkg* schemes, with $g = 3$ in the 3×3 mesh network topology.

Figure 2 Service blocking rate (see online version for colours)



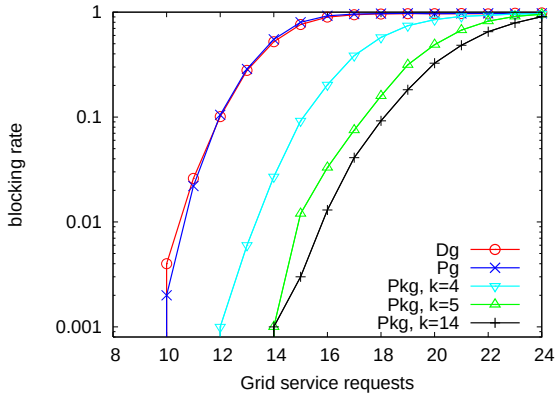
Notes: *Pkg* and *PGkg* schemes, with $g = 3$ in the 3×3 mesh network topology.

Figures 1 and 2 show the service blocking rate as a function of the requested services for the considered schemes when the 3×3 mesh network is considered. Figure 1 shows the performance of *Dg*, *Pg* and *Pkg* schemes. Results show that, at reasonable blocking rate values, only a slight advantage is achieved by the concurrent computation of $p = 3$ lightpaths performed by *Pg* with respect to the distributed computation and set up performed by *Dg*. Results show that significant reduction of service blocking rate is achieved by *Pkg* already with $k = g + 1 = 4$, which requires $c = 4$ combinations of $p = 3$ computations. Figure 1 also shows that $k = g + 2 = 5$ further improves the overall performance, closely approximating the bound achieved by considering $k = N_1 = 9$. The case with $k = 9$, however, may introduce scalability issues due to the large number of potentially required path computations ($c = 84$).

Figure 2 shows the performance of the *PGkg* scheme in case of $k = 4$ and $k = 5$, and compares them with the previously presented *Pkg* results obtained with the same values of k. Results show that with $k = 4$, a further significant service blocking rate improvement is achieved compared to the *Pkg* scheme. Moreover, with $k = 5$, the *PGkg* scheme

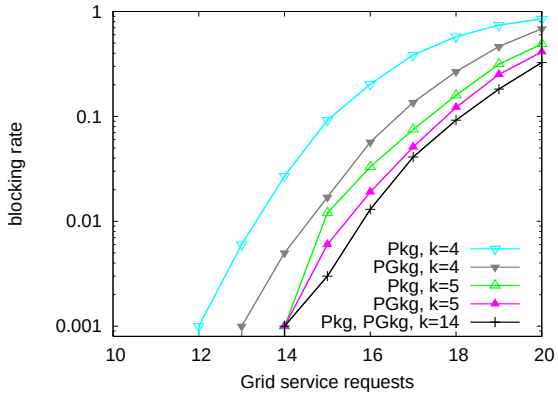
performance is almost equivalent to the optimal performance achieved by *Pkg* or *PGkg* schemes using $k = N_1 = 9$.

Figure 3 Service blocking rate (see online version for colours)



Notes: *Dg*, *Pg* and *Pkg* schemes, with $g = 3$ in the NSF network topology.

Figure 4 Service blocking rate (see online version for colours)



Notes: *Pkg* and *PGkg* schemes, with $g = 3$ in the NSF network topology.

Figures 3 and 4 show the service blocking rate for the considered schemes when the NSF network is considered. Figure 3 shows the performance of *Dg*, *Pg* and *Pkg* schemes. Results confirm that *Dg* and *Pg* achieve similar results. However, in this case, the two curves are almost completely overlapped. This is due to the lower amount of equal cost shortest routes that characterise the NSF network compared to the 3 × 3 mesh network. Thus the PCE cannot fully exploit its capability of performing synchronised path computation. Figure 3 also confirms that *Pkg* with $k = 4$ and $k = 5$ significantly improves the service blocking rate compared to *Dg* and *Pg* schemes. However, in this case, the *Pkg* scheme with $k = 5$ achieves a higher blocking rate compared to the case with $k = N_2 = 14$. This is mainly due to the higher number N_2 of nodes that characterise the NSF network compared to the number N_1 of nodes of the 3 × 3 mesh network. Figure 4 shows that the *PGkg* scheme, which performs upon service request $c = 10$ path computations, further

improves the *Pkg* performance, closely approximating the optimal performance achieved by *Pkg* or *PGkg* with $k = N_2 = 14$ (which however require $c = 364$ path computations).

6 Experimental validation

The proposed schemes have been evaluated on a metropolitan area WSON testbed, which is built on 'JGN2plus's ITU-T G.652 single-mode optical fibres. The WSON testbed is shown in Figure 5. The network connects three different locations in Tokyo (Japan): Koganei, Otemachi and Hakusan. The longest path among these three sites is about 60 km long and it connects Koganei and Kakusan. Each location is equipped with one transparent optical cross-connect (OXC). Two edge nodes are present in Otemachi and Koganei, one in Hakusan. The GMPLS control plane runs on control PCs exploiting generic routing encapsulation (GRE) connections in the address space 10.10.10.x over the real 192.168.x.x IP addresses (Xu and Harai, 2006). The data plane consists in w = 4 wavelengths generated by gigabit ethernet interfaces equipped with dense wavelength division multiplexing (DWDM) media converters. Detailed link capacity in terms of available wavelengths (from w0 to w3, i.e., 1548.5, 1549.3, 1550.1 and 1550.9 nm) before the grid service activation is depicted in Figure 5. The NRM consists of two sub-blocks: the PCE and the lightpath activator. The PCE is implemented in a Linux PC located in Koganei (192.168.100.1). The PCE implementation is derived from Paolucci et al. (2008). It is based on C code and ILP formulations to solve path computations. The PCEP module is based on C++ and TCP socket libraries. The five edge nodes are connected to equally least loaded grid resources (i.e., $G = 5$). The GRM (192.168.2.1) triggers the set up of one grid service s_i, requiring a full mesh of bidirectional lightpaths among $g = 4$ grid resources. It should be noted that in this experimental implementation, GRM and NRM lightpath activator are executed in node R2, namely, GRM and lightpath activator have the same IP address 192.168.2.1 of R2, shown in Figure 5.

When the *Dg* scheme is applied two conditions determine the service rejection. First, if the result of the random selection includes router R1, the service is rejected due to the lack of available wavelengths on link R1-X1. Second, even if the random selection excludes router R1, the service may also be rejected because of lack of network resources. This occurs if the distributed lightpath establishment does not allocate wavelength $w0$ for lightpath R2-R5 and wavelength $w3$ for lightpath R3-R4. The *Dg* scheme experiences a service rejection equal to 97.7%.

When the *Pg* scheme is adopted, the latter critical condition that affects the *Dg* scheme is avoided thanks to the concurrent path computation performed by the PCE. Indeed, the PCE identifies for each lightpath the proper available wavelength to use (i.e., $w0$ for lightpath R2-R5 and wavelength $w3$ for lightpath R3-R4). However, the former critical condition related to the random choice of router R1 persists. With the adoption of the *Pg* scheme, service rejection is reduced to 80%.

By applying the *Pkg* and the *PGkg* schemes with $k = 5$, the service is successfully established since both previous critical conditions are avoided. The $c = 5$ combinations of $p = 6$ lightpath computations allow to exclude router R1 and to allocate the wavelengths that avoid service rejection.

Figure 5 Experimental implementation on the JGN2plus testbed (see online version for colours)

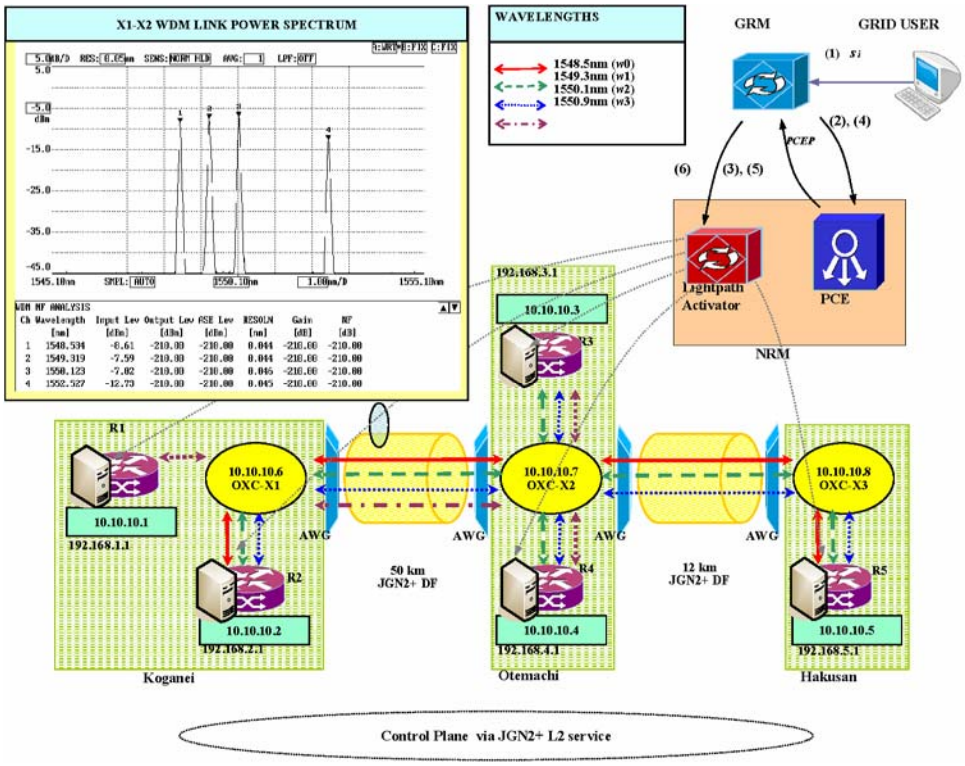


Figure 6 Messages exchange (see online version for colours)

No.	Time	Source	Destination	Protocol	Info
1	0.000000	192.168.2.1	192.168.100.1	TCP	33643 > 3490
2	0.000536	192.168.100.1	192.168.2.1	TCP	3490 > 33643
3	0.000541	192.168.2.1	192.168.100.1	TCP	33643 > 3490
4	0.000967	192.168.100.1	192.168.2.1	TCP	3490 > 33643
5	0.001503	192.168.2.1	192.168.100.1	TCP	33643 > 3490
6	0.042824	192.168.100.1	192.168.2.1	TCP	3490 > 33643
7	0.050519	192.168.2.1	192.168.100.1	TCP	33644 > 3490
8	0.090143	192.168.100.1	192.168.2.1	TCP	3490 > 33644
9	0.096925	192.168.2.1	192.168.3.1	TCP	33645 > 11778
10	0.101584	10.10.10.6	10.10.10.2	RSVP	PATH Message.
11	0.103339	10.10.10.2	10.10.10.6	RSVP	RESV Message.
12	0.299702	192.168.2.1	192.168.4.1	TCP	33646 > 11778
13	0.313976	10.10.10.6	10.10.10.2	RSVP	PATH Message.
14	0.315669	10.10.10.2	10.10.10.6	RSVP	RESV Message.
15	0.502719	192.168.2.1	192.168.5.1	TCP	33647 > 11778
16	0.516133	10.10.10.6	10.10.10.2	RSVP	PATH Message.
17	0.517177	10.10.10.2	10.10.10.6	RSVP	RESV Message.
18	9.831721	50.50.5.1	50.50.255.255	ICMP	Echo (ping)

PCEP Open

PCEP KeepAlive

1st PCEP Req & Rep

2nd PCEP Req & Rep

1st Activation & Set up

2nd Activation & Set up

3rd Activation & Set up

1st Data packet

Figure 6 shows the sequence of the most significant packets exchanged in the experiment when the *Pkg* or the *PGkg* schemes with $k = 5$ are applied. A service s_i is first requested to the GRM [step (1) in Figure 5]. GRM requests the PCE to perform the path computation of the $p = 6$ lightpaths that connect nodes R1, R2, R3 and R4 [step (2) in Figure 5, packets 1–6 in Figure 6). In particular, packets 1 and 2 correspond to the PCEP

'Open' message exchange, packets 3 and 4 correspond to the PCEP 'Keepalive' message exchange, packet 5 to the PCEP 'PCReq' message and packet 6 to the PCEP 'PCRep' message. Due to lack of resources on link R1-X1, the path computation fails [step (3) in Figure 5] and no useful metric values are provided to the GRM. Then, GRM identifies a new set of $p = 6$ lightpaths among nodes R2, R3, R4 and R5 [step (4)]. Packet 7 includes the new PCEP 'PCReq' sent by the GRM to the PCE. The PCE in this case successfully solves the path computation and replies with a PCEP 'PCRep' message [packet 8, step (5)] which includes the metric values and the strict ERO information (for simplicity PK values are not used). Around 40 ms are required between each PCEP 'PCReq' and the related PCEP 'PCRep' (thus including the synchronised path computation of $p = 6$ lightpaths performed by the PCE). Packet 9 corresponds to the management requests for lightpath activation sent by GRM through the NRM lightpath activator to the ingress nodes [steps (6) and (7) collapsed for simplicity in a single message]. The three ingress nodes are sequentially configured for bidirectional lightpath activation: node R2 for lightpaths R2-R3, R2-R4 and R2-R5; node R3 for lightpaths R3-R4 and R3-R5, node R4 for lightpath R4-R5. Packet 10 shows the first 'RSVP Path' message relayed by node X1 to node R2 related to the activation of lightpath R3-R2. Packet 11 refers to the related 'RSVP Resv' message. Packets 12-17 show the messages exchanged for the activation of lightpaths R4-R2 and R5-R2. The time required to complete the signalling message exchange is measured in around 0.4 s. Packet 18 shows the first data plane packet (i.e., Ping message) received by node R2 which is sent from node R5 with a pre-configured data plane address 50.50.5.1. The successfully established lightpaths are monitored with a spectrum analyser placed in the fibre link between X1 and X2. In Figure 5, a capture of the spectrum analyser output visualises the three activated lightpaths corresponding to wavelengths of 1548.5, 1549.3 and 1550.1 nm, respectively. In addition, another wavelength 1552.5 nm which is outside of the lightpaths' wavelength range (i.e., from $w0$ to $w3$) is employed in the fibres between two adjacent nodes for erbium doped fibre amplifier (EDFA) stability concern. In the WSON test bed, at each end of one bidirectional fibre link, a laser source fixedly tuned on this wavelength is used to continuously inject light into the fibre in the direction of signal transmission.

The time elapsed between packet 17 (the last 'RSVP Resv' message) and packet 18 is measured in more than 9 s. This is mainly due to the node interface activation which can start only after the physical connection is activated, i.e., after all the intermediate cross-connections performed by the OXCs. Similar delays were experienced by some of the authors in another experimental implementation (Paolucci et al., 2008) based on different commercially available IP/MPLS routers. Due to such high lightpath activation delay, the additional message exchange required by the proposed *Pkg* and *PGkg* scheme implementations does not significantly increase the overall grid service delivery and confirms the limited dynamicity of grid services exploiting current optical network technologies.

7 Conclusions

In this study, the PCE communication protocol (PCEP) is proposed as standard interface between the GRM and the NRM. In this way, the GRM can retrieve from the NRM information related to the network layer. In particular:

1 the bandwidth available for a label switch path between remote grid resources (e.g., an entire lightpath connection)

2 a metric value representative of some predefined parameter (e.g., end-to-end delay or network cost).

Such information is then used by the GRM to perform joint optimisation between grid and network resources and guarantee the minimisation of the grid service delivery time. Two schemes exploiting the PCEP-based interface have been proposed to provide a feedback to the GRM on the expected network resources utilised by an extended set of grid resources. The two schemes have been evaluated through simulations. Results show that the proposed schemes guarantee the minimisation of the amount of used network resources, which in turn determines the maximisation of the overall amount of established grid services. In particular, the proposed *PGkg* scheme, which exploits the exchange of static network information prior to service requests, closely approximates the optimal performance without requiring a large number of synchronised path computations. Experimental implementation of the proposed schemes in the 'JGN2plus' testbed has been reported. Results show that, due to the large time required for lightpath activation, the proposed PCEP-based communication does not substantially affect the overall grid service delivery time. The proposed solution avoids control plane extensions or interfaces specifically designed for grid purposes. Moreover, it preserves the adequate level of confidentiality between GRM and NRM since it does not require the exchange of topology or detailed link bandwidth information.

This cooperative optimisation model presents a feasible way to fulfil a parallel joint-optimisation by integrating different optimisation capabilities and instances, which are engineered in dedicated problem spaces.

References

ARGON Specification (2005) *VIOLA Deliverable*, available at http://www.viola-testbed.de.

Baraglia, R. et al. (2006) 'A study on network resources management in grid', *CoreGrid 2006 Conf*.

Battestilli, L. et al. (2007) 'EnLIGHTened computing: an architecture for co-allocating network, compute, and other grid resources for high-end applications', *BROADNETS 2007 Conference*, September.

Bradford, R., Vasseur, J.P. and Farrel, A. (2008) 'Preserving topology confidentiality in inter-domain path computation using a key-based mechanism', draft-ietf-pce-path-key-05.txt, November.

Burger, M.d., Kielmann, T. and Bal, H.E. (2002) 'TopoMon: a monitoring tool for grid network topology', *International Conference on Computational Science (ICCS 2002)*, Amsterdam, 21–24 April 2002.

Clapp, G. et al. (2004) 'Grid network services', draft-ggfghpn-netservices-1.0, Informational, Grid High-Performance Networking Research Group (GHPN-RG), August.

Danelutto, M. (2004) 'Adaptive task farm implementation strategies', in *Proceedings of 12th Euromicro Conference on Parallel, Distributed and Network-Based Processing*, 11–13 February 2004, pp.416–423.

De Leenheer, M. et al. (2006) 'A view on enabling-consumer oriented grids through optical burst switching', *Comm. Magazine*, March.

Farrel, A., Vasseur, J.P. and Ash, J. (2006) 'A path computation element (PCE)-based architecture', rfc4655, August.

Ferrari, T. (Ed.) (2005) 'Grid network services use cases', Informational, Grid High-Performance Networking Research Group (GHPN-RG), 14 March.

Ficara, D. et al. (2007) 'The beacon number problem in a fully distributed topology discovery service', *IEEE GLOBECOM 2007*, pp.2591–2596.

Foster, I. and Kesselman, C. (2004) *The Grid 2*, 2nd ed., Morgan Kaufmann.

Habib, I.W., Song, Q., Li, Z. and Rao, N.S.V. (2006) 'Deployment of the GMPLS control plane for grid applications in experimental high-performance networks', *Communications Magazine*, March.

http://www.caida.org/projects/ark/

http://www.eu-egee.org

http://www.jgn.nict.go.jp/

Huffaker, B. et al. (2002) 'Topology discovery by active probing', in *SAINT Workshops 2002*, January–February, pp.90–96.

Lehman, T., Sobieski, J. and Jabbari, B. (2006) 'DRAGON: a framework for service provisioning in heterogeneous grid networks', *Communications Magazine*, March.

Otani, T. (2008) 'Generalized labels for G.694 lambda-switching capable label switching routers', draft-ietf-ccamp-gmpls-g-694-lambda-labels-02.txt, internet draft, July.

Palmieri, F. (2006) 'GMPLS-based service differentiation for scalable QoS support in all-optical grid applications', *Future Generation Computer Systems*, May, Vol. 22, No. 6, pp.688–698, Elsevier.

Paolucci, F., Cugini, F., Valcarenghi, L. and Castoldi, P. (2008) 'Enhancing backward recursive PCE-based computation (BRPC) for inter-domain protected LSP provisioning', *OFC 2008 Conf.*, February.

Takefusa, A. et al. (2006) 'G-lambda: coordination of a grid scheduler and lambda path service over GMPLS', *Future Generation Computer Systems*, October, Vol. 22, No. 8, pp.868–875, Elsevier.

Tomkos, I., Markidis, G. and Sygletos, S. (2007) 'Cross-layer optimized optical grid networks', *BROADNETS 2007 Conference*, September.

Travostino, F., Mambretti, J. and Karmous-Edwards, G. (2006) *Grid Networks*, ISBN: 0-470-01748-1, Wiley.

Valcarenghi, L., Foschini, L., Paolucci, F. and Castoldi, P. (2006) 'Topology discovery services for monitoring the global grid', *Comm. Magazine*, March.

Vasseur, J.P. and Le Roux, J.L. (2008) 'Path computation element (PCE) communication protocol (PCEP)', draft-ietf-pce-pcep-18.txt, November.

Wolski, R. et al. (1999) 'The network weather service: a distributed resource performance forecasting service for metacomputing', in *Future Generation Computing Systems*, Vol. 15, pp.757–768.

Xu, S. (2007) 'Challenges of e-science applications with on-demand optical grid networks: e-VLBI over GMPLS based lightpath networks', *iPOP 07 Conf*, June.

Xu, S. and Harai, H. (2006) 'Optical ring services in GMPLS based mesh networks: an implementation of optical GRID', *OFC 2006*, March.

Zang, H., Jue, J.P. and Mukherjee, B. (2000) 'A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks', *Optical Networks Magazine*.

Zervas, G., Escalona, E., Nejabati, R., Simeonidou, D., Carrozzo, G., Ciulli, N., Belter, B., Binczewski, A., Poznan, M.S., Tzanakaki, A. and Markidis, G. (2008) 'Multidomain optical networks: issues and challenges – phosphorus grid-enabled GMPLS control plane (GMPLS): architectures, services, and interfaces', *Communications Magazine, IEEE*, June, Vol. 46, No. 6, pp.128–137.