AN OPTICAL NETWORK INFRASTRUCTURE SUITABLE FOR GLOBAL GRID COMPUTING

D. Simeonidou, R. Nejabati, M. J. O'Mahony

Department of Electronics Systems Engineering, University of Essex, Colchester, CO4 3SQ, UK

Email: dsimeo, rnejab, mikej@essex.ac.uk

A. Tzanakaki, and I. Tomkos

*Optical Communications Systems and Networks Group, Athens Information Technology Center
Markopoulo Ave., PO. Box 68, 190 02, Peania, Athens, Greece,

Email: atza, itom@ait.edu.gr

Abstract

This paper presents a novel Grid network scenario based on an optical infrastructure using optical burst switching. The functional blocks required are identified as Core Router the Grid User Network Interface and the Grid Resource Network Interface. The details of the associated challenges in terms of functionality and technology are discussed and solutions are proposed.

1. Introduction

Over the past few years it has become evident that local computational resources cannot keep up with the demands generated by some users/applications (either high-volume jobs with high demands for processing and storage or large number of medium/small jobs requesting distributed resources) and distributed computing using the concept of a computational Grid is proposed as the solution. Distributed computing is not a new paradigm but until recently networks could not support the features and capabilities to offer efficient use of remote resources. In such networks, the Grid application characteristics and requirements have major influence on the choice of the suitable infrastructure (i.e. physical layer technology, transport and switching format, as well as control, signalling and management). Grid applications can differ with respect to the following characteristics: granularity of traffic flows, required data transaction bandwidth, QoS, acceptable delay, throughput and packet loss, storage capacity, processing power etc. As the bandwidth and speed of networks based on optical technology utilising wavelength division multiplexing (WDM) have increased significantly, the interest in distributed computing as a realistic solution has significantly increase.

When trying to identify the features and capabilities of a suitable infrastructure it is important to understand the nature and requirements of the specific applications that will be supported by global grid networks [1] in terms of bandwidth requirements and their unique properties [2]. For example, *particle physics* due to its large international collaborations and experiments involving enormous amounts of data requires very advanced network infrastructures that can support processing and analysis of Petabytes per year through globally distributed computing resources [2]. *Very Long Baseline Interferometry* (VLBI) is used by radio astronomers to obtain detailed images and relevant experiments bring data from a network of distributed instruments to a central point to correlate the signals from individual telescopes [3]. *High Performance Computing* and *Visualization* focuses on adapting and developing parallel codes for execution on parallel processors. In these systems remote visualisation of terabytes of data is needed, which can generate data flows requiring high bandwidth links ~ 1 Gbit/s or few hundred Mbit/s in case of compressed video and these requirements increase with the number of remote observers if software multicasting is used [2]. Another important application is *e-Health,* with the example of *Mammography* that introduces increased capacity requirements due to both size and quantity of images produced by scans. As described in [2] some initial calculations have shown that for 100 patients to be screened remotely, the network would have to carry 1.2GB of data every 30 seconds. It is important to note that the necessary features for this type of application are speed of data transfer and security. .

Due to the nature of the applications supported by global Grid computing, suitable network infrastructures are required to offer a set of features that are very different compared to traditional telecommunications infrastructures. In telecommunications networks when a traffic demand arises, independent of the transport protocol, the routing mechanism or the associated algorithms used, there is always a predetermined pair of two discrete points that need to communicate in order to satisfy the traffic demand. However, this is fundamentally different in grid network scenarios, in which a particular end user or application may require to access available network resources offering processing power or storage and needs to identify the availability of network resources. In this case only the source point is predetermined, while the destination has to be discovered and identified using intelligent mechanisms supporting advanced signalling schemes.

An additional fundamental difference of grids compared to conventional telecommunication networks relates to the bidirectionality of connections. In traditional telecommunication networks connections are most commonly bidirectional and the two directions demonstrate strong symmetry characteristics. In the grid paradigm the two directions are also required. One direction enables the source to discover and access the network resources in order to submit a job or request for a data set. The opposite direction is required in order to extract the results or the requested data and deliver them to the user. However, the two directions can be decoupled and set-up independently as due to the network scenario and resource allocation scheme they have a different set of requirements and features.

Optical networks can provide the infrastructures required for this type of applications offering large amount of low cost bandwidth provided that are based on a truly peering model and support a distributed control plane. This will allow control/access and even ownership of network resources by the users/applications in contrast to traditional telecommunications networks. The common requirements in this type of networks are summarised as follows:

- High capacity for bulk data transfer as well as low cost bandwidth on demand for short or long periods of time between discrete points across the network (i.e. point and click provisioning and open GRID service architecture (OGSA) [4])
- Service granularity at the wavelength and sub-wavelength level
- Multicasting capabilities for intelligent resource discovery (key enabler for the OGSA)
- Hardware flexibility to support wide range of different distributed resources
- Resilience to different layers, from the application layer to the wavelength layer
- Network security both at user-network level and network-network level
- Ability to provide management and control of distributed network resources to the user/application

This paper proposes a novel Grid network scenario based on an optical infrastructure where the Grid resource discovery, allocation and path setup will be fully performed either in a self-organised manner by the optical layer (for small/medium size jobs) or by the user/applications (high-demand, high-volume jobs) through user-owned network resources (wavelength, switch ports, etc).

2. Optical Networks Infrastructures for Grid Applications

The features listed above can be supported by optical network solutions offering enormous transmission bandwidth and advanced capabilities such as fast and dynamic path provisioning, reconfiguration, and multicasting at the optical layer. Optical networks offer bandwidth manipulation at the wavelength (wavelength switching) and sub-wavelength level (optical burst and packet switching) supporting not only high service granularity but also the capability to accommodate a wide variety of traffic characteristics. Optical technologies adapted to this new set of requirements through the development of novel hardware designs and solutions accompanying the programmability of the network infrastructure can facilitate dynamic control of bandwidth to support new application centred network infrastructures. The control and management of network resources can be provided through the use of existing protocols like Generalised Multiprotocol Label Switching (GMPLS) [5] in combination with new protocols such as the Optical Border Gateway Protocol (OBGP) [6] and will enable set-up, control and tear-down of end-to-end lightpaths across multiple domains (figure 1).

In these networks resource request, discovery and allocation functions have to be performed initially when a processing requirement (i.e. bandwidth demand) arises. For example if a number of users (workstations) have jobs to be posted on the grid, each user constructs a "request" indicating the amount of processing required. These requests are received by the ingress point of the network and assuming that the total volume of the data to be processed is large enough, a bandwidth request arises at the optical layer. In case of acceptance, the data itself is immediately allowed in the network making use of available resources i.e. wavelengths. The intermediate routers are not notified in advance of the arrival (i.e. no advanced reservation), but decide on the fly where to forward the data.

The intermediate optical routers will require some application and network-level awareness in order to configure the resources that are best suited for the particular task. This can be achieved utilising two types of information:

- The data itself needs to be accompanied by some information on the nature of the job. This may be an indication of the estimated computational and storage capacity required, as well as QoS related information (e.g. maximum completion time) that can be considered as feed-forward information.
- In addition, information about the status of the network needs to be included. The idea is that the grid resources themselves provide periodically information about their status and availability. This includes free storage capacity, computational load, and network resources that can be considered as feed-back information.

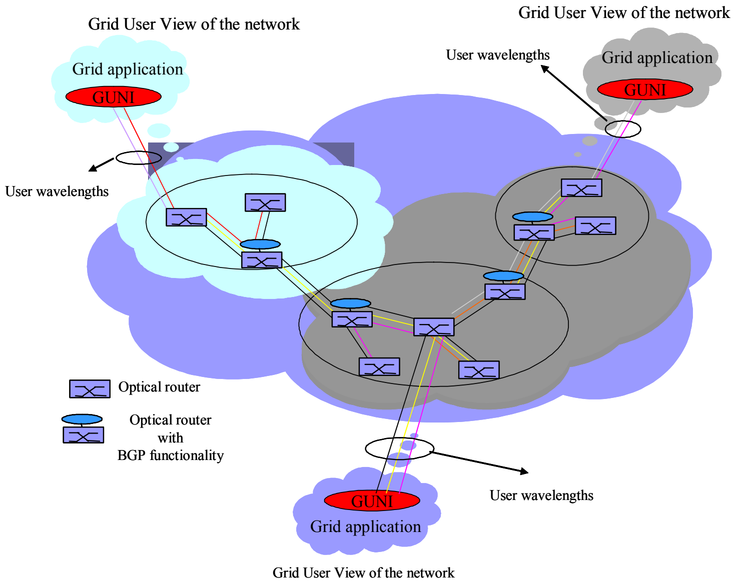
Grid User View of the network
Grid application
GUNI
User wavelengths
Grid User View of the network
User wavelengths
Grid application
GUNI
Optical router
Optical router with BGP functionality
GUNI
Grid application
User wavelengths
Grid User View of the network

Figure 1: Optical Grid network infrastructure

Both information mechanisms suffer by some degree of inaccuracy as feed-forward information is based on predictions of the future state of the network, while feed-back information reports on the past status of the network. The degree of accuracy depends on e.g. absolute or relative distance to the grid resources, frequency and nature of jobs, etc. However, appropriate algorithms that utilise this information and optimise the network performance can be used.

It is important to note that a job in general does not need to access resources in one network location, but the network will be able to allocate any available resources in a single site or multiple sites in order to fulfil the application's storage and/or processing requirements. The end user doesn't specify the network location that the job is being processed. The impact of this is that a job is not explicitly scheduled, but implicitly through its progression in the network. This makes the grid completely distributed and thus more scalable.

Once a job is completed, results may have to be reported back. This is an entirely different problem, since in this case the destination is fixed and known. In effect, the return path presents a more traditional forwarding problem that should be supported by the same technology and within the same network infrastructure. A variety of options may be applicable and will depend on parameters such as:

- the requirement to report results back (e.g. data replication applications do not need to report back results)
- the processing time (i.e. for short processing times the same path could be used whereas for longer processing times a new path from the Grid resources to the user may need to be established)
- storage availability (lack of storage resource may force faster release of processed data back to the user)
- size of results compared to the size and frequency of the jobs posted/processed

In optical infrastructures supporting Grid networks high bandwidth users and applications can manage and control network resources, in a distributed and truly peer-to-peer manner. Optical networks offering enormous transmission bandwidth can support bandwidth manipulation at the wavelength (wavelength switching) and sub-wavelength level through technologies such as optical burst and packet switching offering not only high switching granularity but also the capability to accommodate a wide variety of traffic characteristics and distributions. Optical technologies adapted to this new set of requirements through the development of novel hardware designs and solutions accompanying the programmability of the network infrastructure can facilitate dynamic control of bandwidth to support new application centred and user owned network infrastructures.

A transport protocol suitable for optical infrastructures supporting Grid applications that satisfy the above requirements is optical burst switching (OBS) [2], [7] and [8]. OBS can be used to support multi-service traffic, offering high granularity and high spectral efficiency. It can accommodate bursty traffic with improved network economics and provide convergence of electronic and optical technologies. It can also enable control and management integration and simplification offering a distributed control plane supporting advanced signaling schemes.

OBS is based on the idea of separating the forwarding and switching functions at the network nodes, while forwarding is determined by the information carried by the burst control packet, offering out-of-band encoding of control information on a common signalling wavelength on the link. Each control packet is transmitted ahead of its corresponding burst, and contains the timing information of the burst. The overhead information, encoded typically at a lower bit-rate than the data burst, is processed independently of the data burst most commonly in the electronic domain. However, the data burst is transparently switched and routed through the network without the requirement of any optoelectronic conversion reducing the overall network cost and offering transparency to data rate and format. OBS is based on the asynchronous operation mode, where the optical bursts are of variable length depending on the nature of the application generating the original data packets.

Therefore, OBS is offering the capability to deal with a wide variety of applications and accommodate small, medium and large jobs, as the burst length can vary in order to support the application requirements in terms of duration, latency etc. In addition, using appropriate schemes e.g. varying the offset time between the control packet and the data burst appropriate QoS parameters can be assigned. Alternative solutions that can offer QoS assignment is the utilisation of the optical network physical parameters and performance features. Another important feature of OBS, and in particular of its variations that do not require resource reservation, is that it does not impose any strict requirement for symmetry in the two directions of transmission.

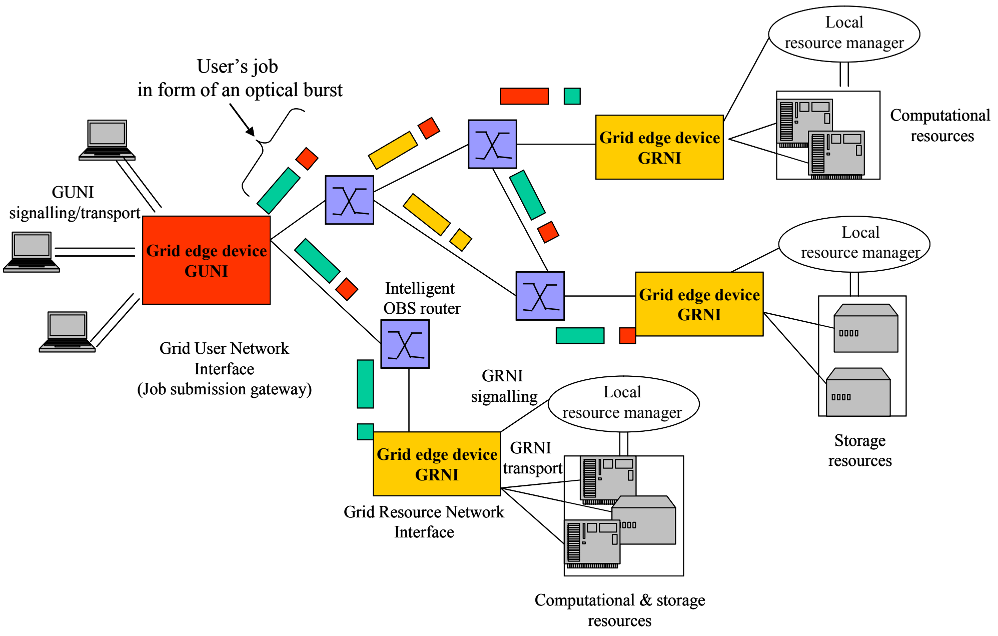

Figure 2: OBS Grid network scenario

2.1 Core Nodes

In the OBS network scenario (Figure 2) the core optical routers are active elements that process the control packets and offer intelligent resource discovery/allocation. Active optical core routers perform an important role especially in a self-organised optical Grid network scenario, where the intermediate nodes are not notified in advance (i.e. no advanced reservation), and decide on the fly where the bursts are forwarded. The core nodes based on the control information configure the switch matrix to route the incoming burst to the appropriate output port, and resolve any contention issues that may arise. In some cases control packet rewriting may be also required. A generic structure of an optical burst switch consists of an input interface, a switching and an output interface. The input interface performs delineation (i.e. identification of the packet start and end) and retrieves control information, encoded in the control packets. The switching block is responsible for routing the packets to the appropriate output ports resolving any contention issues, while the output interface is responsible for the control packet update and any signal conditioning that may be required such as power equalisation, wavelength conversion or regeneration. However, these core nodes can also provide new advanced functionalities to accommodate new grid specific network requirements. One concept that can be applied in a OBS network scenario suitable for grid applications having significant impact in the overall network performance and efficiency is the concept of 1:N light-trees supporting optical multicasting. A light-tree is a clear channel originating at a given source node having multiple destination nodes i.e. is a point-to-multipoint channel [9, 10]. The use of light-trees can significantly reduce the number of hops (or lightpaths) that a packet has to traverse and therefore significantly improve the throughput of the network. Figure 3 illustrates how a burst can be transmitted from node A to nodes C, D, E and F with minimum hops as only one burst is initially routed from node A to node C. At node C it is locally dropped and optically multicasted to node E, where it is dropped locally and also routed to node F. This scheme is used instead of bandwidth consuming multicasting of the original burst from the source node A to all different destinations C, D, E and F, that may take place at the optical or a higher layer. In general avoiding multicasting at a higher layer:

- reduces requirements for optoelectronic conversions
- limits the need for store-and-forward functions
- enhances the virtual connectivity of the network

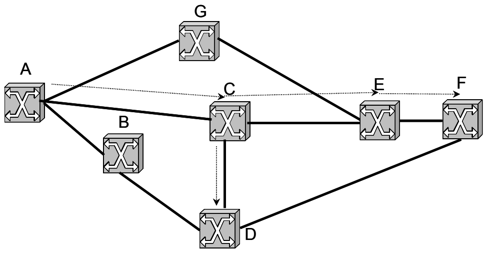

Figure 3: Network topology supporting light-trees

This feature is particularly suitable for OBS infrastructures appropriate for grid applications as discussed in the previous section.

In order to achieve the above functionality it is important to identify an optimised optical burst switch architecture and the appropriate switching technology offering advanced features such as

- dynamic reconfiguration with high switching speed (~ μs)
- strictly non-blocking connectivity between input and output ports
- contention resolution mechanisms and technology choice
- multicasting capabilities
- capability to address contention issues and QoS differentiation
- scaleability
- upgreadability
- minimum performance degradation for all paths and good concatenation performance

In terms of optical switch architectures there are a number of options already proposed in the literature, but the different proposals need to be adjusted to the set of requirements imposed by this new application framework. Features such as multicasting are central and need to be addressed by the proposed solution. However, a full multicasting solution may not be necessary, while partial multicasting capability may be a more appropriate approach [11]. This will have a direct impact on the design of the OBS and will relax the requirements in implementation complexity and performance. A block diagram of a partially multicasting optical switch architecture is illustrated in Figure 4 [11].

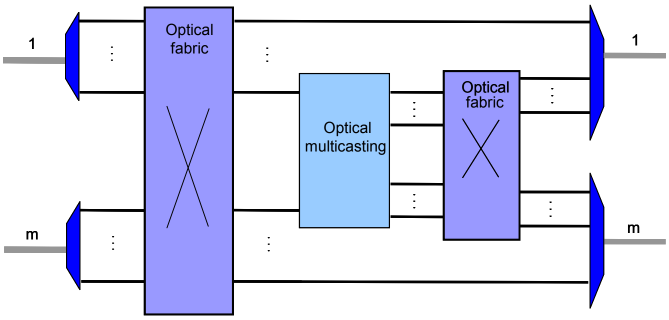

Figure 4: Optical switch architecture supporting partial multicasting

The obvious technology of choice to achieve multicasting operation is the use of passive splitters and couplers [12] however alternative solutions may be also attractive such as: multicasting switches [13] with fast response, wavelength converters supporting multicasting capabilities and other semiconductor devices offering similar features etc. This type of options may be in some cases even more attractive than conventional copy creation through passive splitters in terms of performance, functionality and scalability. It is also worth noting that apart from the architecture it is important to choose the switching technology so that it meets the requirement of asynchronous switching at the burst level (μs).

Optical burst switching relies on statistical multiplexing and as a consequence, temporary overload situations may cause contention and lead to burst losses. To resolve contention in the optical burst switch the following options and their combination can be used:

- Wavelength dimension: using wavelength conversion, a burst can be transmitted on a different wavelength on the designated output fibre
- Space dimension: using deflection routing, in which optical bursts are diverted through a different route to their destination nodes e.g. hot potato scheme
- Time dimension: using optical or electronic buffering. Optical buffering is commonly implemented using fibre delay lines or combinations of fibre delay lines and optical switches with fast reconfiguration speed

Optical buffers are very difficult to accurately control and can be a source of penalties in the system. A variety of mechanisms such as wavelength conversion or deflection routing (mentioned above) can be used to minimise contention and reduce the buffering requirements. Novel schemes related with the assignment of quality of service parameters and the fact that OBS switch resources are allocated just before the arrival of the burst at the node may be effectively used to eliminate or minimise buffering requirements e.g. quality of service implementation through the use of variable offset time between control packet and data burst in order to ensure sufficient time for resource allocation for high QoS traffic and applications. However, it is important to note that the use of some limited buffering may significantly improve the overall network performance and can be considered as a possible option.

2.2 Grid user-network interface (GUNI)

To facilitate on demand access to the self-organised optical network and consequently Grid services, interoperable procedures between Grid users and optical network for agreement negotiation and user job submission have to be developed. These procedures constitute the Grid User Network Interface (GUNI) (Figure 5). Although the network scenario is such that network nodes on fly decide how to forward the different job originating by the users and there is not any advance resource discovery or path setup, a user network interface is necessary to accept user job and submitted the job to the network

The GUNI will be part of the Gird edge device that performs interface functionality between Grid users and self-organised optical network. The GUNI-C (client side) provides a simple signalling mechanism between Grid users and the self-organised optical network. In the proposed Grid network the service invocation model follows the indirect service invocation scenario. It means the Grid users through a Grid service agent which is part of the GUNI-N (network side) request for a Grid service (i.e. permission for the job submission) and submit a job to the network. The Grid edge device (performing GUNI functionality) basically is an edge optical burst switched router with the Grid service provider functionality. At the edge device, the incoming Grid jobs from separate users are pre-processed and maybe merged together to construct the optical bursts that will be sent out to the network (anycast to several nodes). Also at the edge device, the returning results of a particular Grid job arriving from separately located resources are merged together and routed back to the user.

In summary the Grid edge device with GUNI functionality is responsible for:

- User job pre-processing and transmission entity construction:
  - Job classification, aggregation (grooming)
  - Transmission entity optical burst assembly
  - Flexible bandwidth allocation
- Pre-processing the coming back results from Grid resources
  - Send back results to the users
- Anycasting the optical burst to the core nodes
  - λ-selection for anycast
- User authentication and security check, job acceptance or rejection
  - Propagation of service and agreement related events
- Grid service billing, accounting and charging.
- Fault detection, protection and restoration

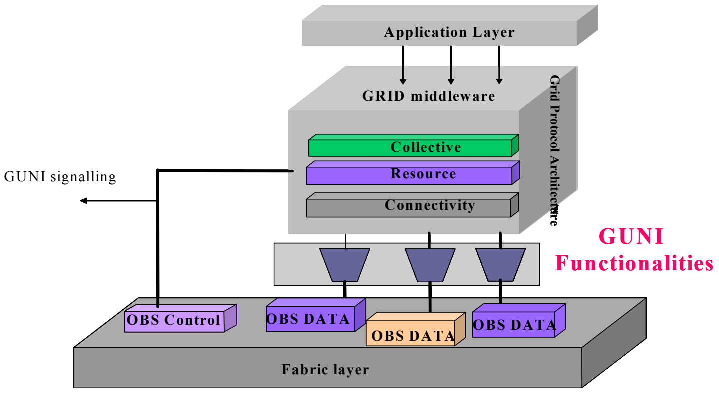

Figure 5: Grid user-network interface

2.3 Grid resource-network interface (GRNI)

The GRNI is a Gird edge device that performs interface functionality between local Grid resources (processing and/or storage) and self-organised optical network. The GRNI provides simple signalling between the local resource manager and the self-organised optical network. It also provides a data transport mechanism between Grid resources and self-organised optical network.

The GRNI is an edge optical burst switched router that receives the Grid jobs from the optical network in the form of optical bursts and routes them in a suitable format to the local resources (Figure 6). On the other hand it receives results of the completed jobs from local resources and routes them back through the optical network. The GRNI edge router will also provide updates on the state of the local resources (i.e. available processing and storage capacity) to the optical network through the GRNI signalling.

In summary the GRNI is responsible for:

- Pre-processing of the incoming optical bursts :
  - Optical burst segregation
  - Job submission to local Gird resources
- Advertising state of the local resources:
  - Broadcasting of the available processing/storage capacity to the optical network
- Sending back results of a completed job
  - Transmission entity construction (optical burst)
  - Bandwidth allocation and light-path setup

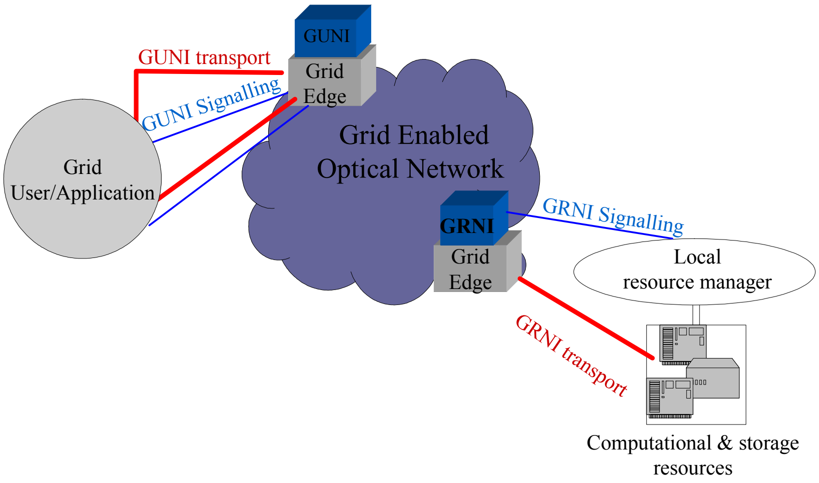

Figure 6: Grid resource-network interface

3. Conclusions

A variety of applications and the requirements they imposed on global Grid networks have been discussed. A novel Grid network scenario based on optical infrastructures has been proposed. The technology described is based on the optical burst switching paradigm in order to fulfil the Grid application specific traffic requirements and efficient sharing of the network resources. The fundamental functional blocks needed are identified as the Core Router, the Grid User Network Interface and the Grid Resource Network Interface. The core router based on optical technologies is able to support routing of the optical bursts on the fly and provide advanced features such as optical multicasting. The optical GUNI is able to support fast and dynamic wavelength allocation as well as traffic processing required by GRID applications. Interface functionality between local Grid resources (processing and/or storage) and self-organised optical network. The GRNI provides simple signalling between the local resource manager and the self-organised optical network while it also provides a data transport mechanism between Grid resources and self-organised optical network.

References

[1] M. Maeda, K. Thompson, “Cyberinfrastructure and Large-Scale Network Testbeds”, OFC2004, ThH1, Los Angeles February 2004

[2]D. Simeonidou (ed.), et al, “Optical Network Infrastructure for Grid”, Grid Forum Draft, Sept. 2003, http: /projects/ghpn-rg/document/draft-ggf-ghpn-opticalnets-1/en/1

[3] Mike J O’ Mahony, Tanya C Politi, Ann Ackaert, Didier Colle, Piet Demeester, Paul Lagasse, “Optical networking research: A European perspective”, OFC2004, WH1, Los Angeles February 2004

[4] Ian Foster, Carl Kesselman, Jeffrey M. Nick, Steven Tuecke, “The Physiology of the Grid”, Draft document, http://www.globus.org/research/papers/ogsa.pdf.

[5] A. Banerjee, J. Drake, P. Lang and B. Turner, “Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements”, IEEE Communications Magazine, pp 144-150, January 2001.

[6]“Optical BGP networks”, Canarie OBGP, Internet draft: http://obgp.canet4.net/

[7] C. Qiao and M. Yoo, "Optical Burst Switching - A New Paradigm for an Optical Internet", Journal of High Speed Networks, Special Issue on Optical Networks, Vol. 8, No. 1, pp.69-84, 1999.

[8] C. Qiao and M. Yoo. “Choices, features, and issues in optical burst switching”, Optical Networks, 1(2): 36-44, April 2000.

[9] L. H. Sahasrabuddhe and B. Mukherjee, “Light-Trees: Optical Multicasting for Improved Performance in Wavelength Routed Networks”, IEEE Comms Magazine, February 1999, pp. 67-73.

[10] X. Zhang, J. Y. Wei, C. Qiao, “Constrained Multicast Routing in WDM Networks with Sparce Light Splitting,” J. Lightwave Tech., vol. 18, no. 12, Dec. 2000, pp. 1917-1927.

[11] G. N. Rouskas, “Optical Layer Multicast: Rational, Building Blockes and Challenges”, IEEE Network, January/February 2003, pp. 60-65.

[12] K.D., Wu, J. C., Wu, C.S. Yang, “Multicast Routing with Power Consideration in Sparce Splitting WDM Networks,” Proc. IEEE ICC, 2001, pp. 513-517.

[13] www.lynx-networks.com.