Grid Optical Burst Switched Networks (GOBS)

Grid High Performance Networking Research Group	Editors: Dimitra Simeonidou (University of Essex) Reza Nejabati (University of Essex) Nicola Ciulli (Nextworks s.r.l)
GRID WORKING DRAFT	Lina Battestilli North Carolina State University
	Gino Carrozzo Nextworks s.r.l
	Pierro Castoldi Scuola Superiore Sant'Anna
	Franco Callegati, DEIS – University of Bologna
draft-ggf-ghpn-GOBS-1	Piet Demeester, Marc De Leenheer Bart Dhoedt, University of Gent
	Yue feng Ji Beinjg univeristy of Post and Telecommunication (BUPT)
	Ken-ichi Kitayama Photonic Internet Forum, Japan
Category: Informational Track	Gigi Karmous-Edwards MCNC, Research & Development Institute
	Jintong Lin Beinjg univeristy of Post and Telecommunication (BUPT)
http://forge.gridforum.org/projects/ghpn-rg/	Anna Tzanakaki, AIT
	Emmanuel (Manos) Varvarigos, Kyriakos Vlachos University of Patras
	Luca Valcarenghi, Scuola Superiore Sant'Anna
	Zhoujun Yu, Jian Wu, Hongxiang Wang Beinjg univeristy of Post and Telecommunication (BUPT)
	Georgios Zervas University of Essex

Status of this Memo
This memo provides information to the Grid community in the area of high performance networking. It does not define any standards or technical recommendations. Distribution is unlimited.

Comments: Comments should be sent to the GHPN mailing list (ghpn-wg@Gridforum.org).

1. Introduction

Optical networking for the Grid computing is an attractive proposition offering huge amount of affordable bandwidth and global reach of resources [1]. Currently, Grid computing using optical network infrastructure is dedicated to a small number of well known organizations with extremely large jobs (e.g. large data file transfers between known users or destinations [1]. Due to the static or semi-static nature of this type of Grids, long-lived wavelength paths between clients and Grid resources with centralized job management strategies are usually deployed (Lambda Grids). This type of Grid networking relies on carrier provision of optical network resources while the Grid users have no visibility of the lambda infrastructure. In other words, the Grid user is not able to setup paths over the optical Grid network.

As Grid applications evolve, the need for user controlled network infrastructure is apparent in order to support emerging dynamic and interactive services. Examples of such applications may be high resolution home video editing, real-time rendering, high-definition interactive TV, e-health and immersive interactive learning environments. These applications need infrastructures that makes vast amount of storage and computation resources potentially available to a large number of users. Key for the future evolution of such networks is to determine early on the technologies, protocols, and network architecture that would enable solutions to these requirements.

In an attempt to address this problem, in this draft novel network paradigms and solutions based on the optical burst switching are discussed.

1.1. Optical burst switching, a realistic technology for Grid networking

Optical burst switching (OBS) is a promising technology for the future networks where the bandwidth needs to be accessible to users with different traffic profiles. The OBS technology combines the advantages of optical circuit switching and optical packet switching [2]. An optical burst is usually defined as a number of continuous packets destined for a common egress point. The burst size can vary from a single IP packet to a large data set at milliseconds time scale. This allows for fine-grain multiplexing of data over a single wavelength and therefore efficient use of the optical bandwidth through sharing of resources (i.e. light-paths) among a number of users. The fundamental premise of OBS technology is the separation of the control and data planes, and the segregation of functionality within the appropriate domain (electronic or optical). Prior to data burst transmission a Burst Control Packet (BCP) is created and sent towards the destination by an OBS ingress node (edge router). The BCP is typically sent out of band over a separate signalling wavelength and processed at intermediate OBS routers. It informs each node of the impending data burst and setup an optical path for its corresponding data burst. Data bursts remain in the optical plane end-to-end, and are typically not buffered as they transit the network core. The bursts' content, protocol, bit rate, modulation format, encoding are completely transparent to the intermediate routers. The main advantages of the OBS in comparison to the other optical networking schemes are that: a) unlike the optical wavelength switched networks the optical bandwidth is reserved only for the duration of the burst; b) unlike the optical packet switched network it can be bufferless.
The OBS technology has the potential to bring several advantages for Grid networking:

- Native mapping between bursts and Grid jobs: the bandwidth granularity offered by the OBS networks allows efficient transmission of the user's jobs with different traffic profiles
- Separation of control and data plan: this allows all-optical data transmission with ultra-fast user/application-initiated light-path setup
- Electronic processing of the burst control packet at each node: this feature can enable the network infrastructure to offer Grid protocol layer functionalities (e.g. intelligent resource discovery and security)

2. Grid-OBS network elements

2.1 Core OBS router

As future optical technology moves to 40Gb/s and beyond, networking solutions must be designed to be compatible these bit rates, in order to reduce the cost per bit [3]. OBS has been introduced as a switching technology relaxed on fast switching requirements, as the relatively slow switch set-up times (milliseconds rather than nanoseconds) are small compared to the payload duration (usually hundreds of milliseconds or seconds) and therefore throughput is almost unaffected [4]. However, the introduction of Grid services over OBS implies new constrains for the switching speed requirements, which become particularly important when high speed transmission is considered.

A flexible Grid network will require also the support of users with small job requests. For example, a relatively small burst, 300ms, transmitted at 10Gb/s can be switched by a MEMS based switch typically within 20ms. Considering only the switching time, the throughput of the system is 93.3%. If the same burst is transmitted at 160Gb/s then its duration is 18.75ms and routing through the same switch would decrease the system's throughput to less than 50%. This becomes more severe when users with even smaller job requests are treated. These small jobs are implied by the small bursts and may be with short offset time. These types of bursts with small length (typical 100 to 1000 bytes), requires ultra-fast switching in nanoseconds. Additionally, the support of multicasting is particularly advantageous, in order to enable parallel Grid processing services latency [5] as well as resource discovery. For these reasons the deployment of fast switching technology is essential for future high speed OBS networks that can support Grid applications. It should be noted though, that the core OBS for the Grid computing may require intensive and intelligent processing of control information and BCP (i.e. performing some Grid network functionality, e.g.: taking part in resource discovery), which can only be performed by specially designed fast electronic circuits. Recent advances in the technology of integrated circuits allow complicated processing of bursty data directly up to 10Gb/s [6]. This sets the upper limit in the transmission speed of the control information and BCP. On the other hand the much longer transparently switched optical bursts (i.e. no conversion to electronic domain) are those that determine the capacity utilisation of the network. The optical bursts can be transmitted at ultra-high bit rates (40 or 160Gb/s), providing that the switching elements can support these bit rates. Faster bursts indicate higher capacity utilisation of the existing fibre infrastructure and significantly improved network economics.

The fast switching solutions that have been proposed are based on the use of fast active components, like Semiconductor Optical Amplifiers (SOAs). Switching is achieved either by broadcasting the signal (passive splitting) and selecting the appropriate routes using fast gating [7,8] or by converting the signal's wavelength and routing it to an output port of a passive routing device (AWG) [9,10,11]. The gating solution is independent of the signal's bit rate and also supports multicasting but scales poorly to a large port-count switch. The wavelength conversion and selection solution is scalable but bit-rate dependent on the utilised conversion technique.

The deployment of fast switching assists the efficient bandwidth utilisation but provides an expensive solution when it scales to many input port. On the other hand, there is no additional benefit for long bursts of data (e.g. originated from large GRID users) if fast switching is utilised. Therefore, a proper OBS networking solution needs to consider a combination of fast (e.g. SOA-based) and slow (e.g. MEMS-based) switches.

One solution can be based on the use of OXCs that has a number only of output ports connected to a fast optical switch that follows. Several OXCs and fast switched can be placed in parallel in a scalable wavelength modular architecture. At the switch input the wavelength channels per input fibre are separated. When a BCP appears the control mechanism must first recognise if the BCP belongs to a long, a short burst. In the first case the OXC is reconfigured so that when the long burst arrives it automatically routed to the appropriate output port. In the other two cases the short and the active bursts are routed directly to the fast switch (through pre-defined paths) and switched immediately to the next node. This set-up requires all the switching paths inside the OXC to be initially connected to the fast switch ports and special design constrains must be considered to avoid collision. The benefit of the proposed scheme is that it reduces the requirements on fast switching and therefore smaller and cost efficient matrices are only required

2.2 Edge Grid-OBS router

An Edge Grid-OBS Node must be able to fulfil Grid application requirements and make efficient use of network resources by using OBS technology as a solution towards ubiquitous photonic Grid networking. The router architecture should introduce a mechanism that can process Grid-IP traffic for GridDiffServ provisioning and maps it onto optical bursts. In Grid-OBS networks, a data burst and its burst control header are transmitted separately on different wavelength channels and switched respectively in optical and electronic domains. Thus, in an OBS network an ingress edge router able to initiate a burst control header and also map user traffics traffic into the optical domain in the form of variable length optical bursts is mandatory.
An edge Grid-OBS router must be able to perform the follow functionalities:

a) Grid Job Classification
b) Traffic aggregation and optical burst assembly
c) Optical burst transmission
d) Grid user to network as well as Grid resource to network signaling

- **Grid Job classification**

The Job classification at the edge of the network must provide fair and specialized services – Grid Differentiated Services (GridDiffServ). Application performance and Grid Network utilization can be enhanced by efficiently matching computational and network resources to user/application requirements. A flexible and scalable Grid job classification mechanism can process jobs based on Grid requirements. Such a classification will trigger the Grid-OBS edge routers' intelligent mechanisms i.e. job scheduling, queuing and resource discovery, for GridDiffServ provisioning. The job classification can combine three independently parallel processed schemes; The

Network-Oriented Classification scheme, the Grid-Oriented Classification scheme and the Time-Oriented Classification Scheme, all positioned at the ingress edge routers and synchronously triggered by Grid Job Requests only. The speed of job classification is vital for providing intelligent services at edge router level. Lack of wire speed classification will result in queuing Grid-IP job requests before they are processed. Important traffic will be dropped or unfair queuing will occur [12].

- **Burst aggregation**

The burst aggregation algorithm at the edge router can greatly impact the overall OBS network operation because it sets the burst characteristics and therefore shapes the burst arrival traffic. The algorithm has to consider the following parameters: a pre-set timer, a maximum burst length, and a minimum burst length. The timer determines when the end-device is to assemble its collected traffic into a new burst. The maximum and the minimum burst length parameters shape the size of the bursts. It is necessary to set a maximum burst length since very long bursts hold on to the resources of the network for a long time and, thus, they cause the unfair loss of other bursts. On the other hand, the minimum burst length is necessary because very short bursts may give rise to too many control packets, which can overload the control unit of the OBS node. The burst aggregation algorithm may use bit-padding if there is not enough data to assemble a minimum size burst.

- **User and resource network interface functionality:**

To facilitate on demand access to Grid services, interoperable procedures between Grid users and optical network for agreement negotiation and Grid service activation have to be developed. These procedures constitute the Grid User Optical Network Interface (G-OUNI). The G-OUNI functionalities and implementation will be influenced by number of parameters as follows:

- Service invocation scenarios
- Control plane architecture

The GUNI in a grid enabled OBS network needs to provide the following main functionalities:

- Flexible bandwidth allocation
- Support for claiming existing agreements
- Automatic and timely light-path setup
- Traffic classification, grooming, shaping and transmission entity construction

On the other hand, geographically distributed processing and storage resources across the network constitute fundamental elements of the large scale Grid network. In such network scenario the Grid resources (i.e. storage and processing) can dynamically enter and leave the OBS network based on pre-established agreements. This fact imposes the necessity of a dedicated signalling and control interface between such resources and the Grid network. Like the GUNI, the Grid resource network interface (GRNI) must perform interoperable procedures between external network elements and the OBS network. But unlike the GUNI, the interface will be between resources-end elements (processing and/or storage distributed across network) and the optical network. The similarity

between GUNI and the GRNI makes it possible to extend the GUNI model to provide required functionalities for the resource network interface. Main functionalities of such an interface can be:

- Support for existing agreements
- Job submission to local Gird resources
- Support for advance resource reservation schemes
- Propagation state of the local resources (available storage/ processing resources)
- Propagation of service related events
- Sending back results to source or multiple alternative destinations

AS both GUNI and GRNI with aforementioned functionalities are related either to the Grid users or Grid resources (i.e. Grid network end elements), thus their functionalities must be integrated into an edge OBS router device. Such edge router must be an agile and user-controlled interface able to map user traffic into optical domain at sub-wavelength granularity (i.e. in the form of optical bursts).

3. Control plane and signalling for Grid-OBS

The utilization and improvement of the GMPLS control plane (i.e., routing and signalling protocols) allows Grid-OBS to provision Grid application with the required QoS. The GMPLS control plane would contribute not only on improving Grid-OBS resilience but it will indeed impact Grid-OBS ability of providing QoS connectivity. Currently deployed optical networks are still based on permanent and semi-permanent optical connections terminated at each network node by optoelectronic transponders. Because of their high cost, fixed bit data rate, and fixed protocol data format, optoelectronic transponders limit the network evolution. Novel emerging technologies, such as Optical Burst Switching (OBS), can boost the network evolution from the technological viewpoint by allowing the introduction of all-optical sub-networks at whose edges optical data signals undergo optoelectronic conversion.

High performance applications, such as several Grid applications, may significant benefit from the introduction of advanced network features provided by OBS networks, e.g. data transparency at extremely high bandwidth. For some Grid applications however, there is the need for all bursts to travel the same route through the network. These applications are particularly sensitive to jitter and out-of-order delivery of packets. In these cases the setup of persistent routes can guarantee the required level of Quality of Service (QoS). Persistent OBS connections require a session declaration separated from the cross-connect setup phase and the data burst transmission phase.

During the session declaration phase the routing decision is taken for the burst data flow and an identifier (or label) is associated to the flow in such a way that every burst belonging to that flow is treated in the same way from source to destination.

The cross-connect setup phase refers to the signaling messages that travel out-of-band ahead of the data burst. These messages notify how to configure the switch for the incoming burst (explicit or estimated setup/release).

Data burst transmission phase refers to the transparent flow of optical data bursts.

The management of persistent connection in OBS networks however seems to have many similarities to connection setup and data forwarding in Generalized Multi-Protocol Label Switching (GMPLS) networks, where every data packet is characterized by a label defined during the initial path setup phase. Because of the flexible structure that characterizes the GMPLS protocol suite, GMPLS seems to be a qualified candidate to incorporate the aforementioned OBS session declaration phase.

3.1 Connection Setup Mechanisms

- **Signaling**

In most OBS variants, the signaling of connections is accomplished using a one-way signaling scheme whereby the burst is transmitted after an offset without any knowledge of whether the optical path has been successfully established end-to-end. Therefore it is

possible that a burst may be lost if the control packet is not able to reserve resources at any of the OBS nodes along the burst's path. The OBS network, however, does not retransmit lost bursts as this is left to the upper network layers. Note also that it is very important that the offset is calculated correctly. If the offset is too short then the burst may arrive at a node prior to the control packet and thus be lost. On the other hand, offsets that are too long reduce the throughput of the end-device

- **Routing**

An OBS network needs an effective routing algorithm. One approach is to route the control packets on a hop-by-hop basis, as in an IP network, using a fast table look-up algorithm to determine the next hop. The second approach is to use the multi-protocol label switching (MPLS) techniques. In MPLS, a packet is marked with a label, which is used to route the packet through the network. At each node, the label of an incoming packet is looked up in a table in order to obtain the destination output port and a new label valid on the next hop. A third routing approach is to use the constrained-routing version of MPLS, which can be used to explicitly setup routes. This explicit routing is very useful in a constrained-based routed OBS network, where the traffic routes have to meet certain Quality of Service (QoS) metrics such as delay, hop-count, BER or bandwidth.

- **Wavelength Allocation**

As in any other type of optical network, each OBS network has to assign wavelengths at the different WDM fibers along the burst route. This wavelength allocation in OBS depends on whether or not the network is equipped with wavelength converters, devices that can optically convert signals from one wavelength to another. In an OBS network with no wavelength converters, the entire path from the source to the destination is constrained to use the same wavelength. In an OBS network with a wavelength conversion capability at each OBS node, if two bursts contend for the same wavelength on the same output port, then the OBS node may optically convert one of the signals from an incoming wavelength to a different outgoing wavelength. Wavelength conversion is a desirable characteristic in an OBS network as it reduces the burst loss probability, however it is still an expensive technology. An OBS network will most likely be sparsely equipped with wavelength converters, i.e., only certain critical nodes will have that ability.

- **Pre-transmission Offset Time**

An OBS user first transmits a control packet and after an offset time it transmits the burst. This offset allows the control packet to reserve the needed resources along the transmission path before the burst arrives. Furthermore, the OBS nodes need this offset time to set up their switching fabrics so that the data burst can ``cut-through'' without the need for any buffers. Ideally, the offset estimation should be based on the number of hops between the source and the destination and the current level of congestion in the network. Obviously, an incorrect offset estimation would result into data loss because the burst may arrive at an OBS node before the optical cross-connect has been completely set up. Therefore, determining this offset is a key design feature of all OBS networks and its effectiveness is measured in terms of the burst loss probability. There are variations in the

OBS literature on how exactly to determine the pre-transmission offset time and how to reserve the needed resources at the core OBS nodes. Despite their differences, however, all of the proposed OBS architectures have a dynamic operation, which results in high resource utilization and adaptability.

- **Scheduling of Resources: Reservation and Release**

Upon receipt of a control packet, an OBS node processes the included burst information and allocates resources in its switch fabric that will permit the incoming burst to be switched out on an output port toward its destination. The resource reservation and release schemes in OBS are based on the amount of time a burst occupies a path inside the switching fabric of an OBS node.

There are two OBS resource reservation schemes, namely, immediate reservation and delayed reservation. In the immediate reservation scheme, the control unit configures the switch fabric to switch the burst to the correct output port immediately after it has processed the control packet. In the delayed reservation scheme, the control unit calculates the time of arrival tb of the burst at the node, and it configures the switch fabric at tb.

There are also two different resource release schemes, namely, timed release and explicit release. In the timed release scheme, the control unit calculates when the burst will completely go through the switch fabric, and when this time occurs it instructs the switch fabric to release the allocated resources. This requires knowledge of the burst duration. An alternative scheme is the explicit release scheme, where the transmitting end-device sends a release message to inform the OBS nodes along the path of the burst that it has finished its transmission. The control unit instructs the switch fabric to release the connection when it receives this message.

Combining the two reservation schemes with the two release schemes results in the following four possibilities: immediate reservation/explicit release, immediate reservation/timed release, delayed reservation/explicit release and delayed reservation/timed release.

- **Burst scheduling issues for congestion resolution**

It has already been outlined that, because of contemporary requests for a given output port by different bursts, a congestion resolution issue arises in OBS networks. The time, the wavelength and/or the space domains can be exploited to solve congestion. As always happens any alternative offers some performance improvement at some cost in network complexity and/or resource utilization. The best results can be obtained by exploiting all means in an integrated way, designing suitable burst scheduling algorithms.

For instance load balancing over the wavelength set of a fiber has been shown to provide a significant performance improvement that gets bigger and bigger as the wavelength set grows in size [13][14]. This solution requires wavelength converters at the OBS node,

possibly tunable to guarantee maximum flexibility and therefore trades hardware complexity with performance.
Load balancing on the wavelengths is even more effective when combined with some limited buffering in the time domain. In this case the OBS switch control logic, by processing the BCP, chooses the forwarding path, i.e. the output fiber, and also addresses the congestion resolution issue, by deciding:

- the wavelength on the designated output fiber that will be used to transmit the packet, in order to properly control the output interface;
- the delay, if any available, that will be assigned to the packet in case all wavelengths are busy at the time of packet arrival;
- to drop or re-route the burst, if no wavelength and delay are available.

The wavelength and delay scheduling (WDS) is addressed by the WDS algorithm [15], i.e. some sort of optimization, where bursts are scheduled in a given time window over a set of wavelengths. A number of WDS algorithms have been proposed that are based on heuristics that can be classified as:

- delay oriented algorithms (D type), that aim at minimizing the latency and choose the earliest available wavelength;
- gap oriented algorithms (G type), that aim at minimizing the gaps between bursts (i.e. maximizing the line utilization) and choose the minimum gap with the previous bursts.

It is interesting to note that a G type algorithm does not necessarily imply a larger latency in the OBS node, because the better utilization of the available transmission resources may turn in shorter waiting times.
Moreover the algorithm may or may not try to fill the gaps (voids) between bursts, with a technique known as void-filling. Therefore WDS algorithms can be:

- D or G type without void filling (noVF), just exploring the scheduling times after any other scheduled burst;
- D or G type with void filling (VF), exploring all scheduling times, including those between other scheduled bursts, to see whether the newcomer may fit in between.

The problem in implementing these algorithms, on top of the additional hardware required to implement a delay buffer (delay lines etc.) is that they need to scan a data structure recording the time of arrival and departure of already scheduled bursts. The complexity of this data structure may be fairly large, depending on the number of wavelengths and delays and varies according to the traffic conditions, therefore the time needed to perform the scheduling algorithm is not easily predictable and may turn to be large enough to make it a system bottleneck. Effective solutions to implement this search has been addressed, for instance, in [16],[17].

Performance can be improved even further combining the flexibility of adaptive routing with the efficiency of packet multiplexing over a large set of wavelengths. For instance at each node, traffic is normally forwarded along the shortest path but alternative paths of equal or higher hop count are also identified and are used in a Multi-Path Routing (MPR) strategy, that dynamically uses alternatives when shortest path (also called the default link) becomes congested.
A number of alternatives exist to choose the alternative paths, for instance all paths with the same hop count could be considered as alternatives etc. Again performance is traded

with complexity (more processing in the network nodes) and cost (more traffic in the networks, especially when longer alternative paths are considered).
Finally, the most integrated approach is to see a set of possible routes as a shared pool of resources to which a WDS scheduling will be applied thus increasing as much as possible the dimensions of the resource set over which to balance the load.
Last but not least it is worth mentioning that this kind of problems are not peculiar to OBS networks, but also apply to faster switching technologies, such as Optical Packet Switching, as log as the information units to be forwarded are asynchronous and variable in length. Therefore effective scheduling algorithms could be applied through different technologies, that just scale in term of switching time.

- **Limited Buffering Using Fiber Delay Lines**

One of the main design objectives for OBS is to build a bufferless network, where the user data travels transparently as an optical signal and ``cuts-through'' the switches at very high rates. Bufferless transmission is important to OBS because electronic buffers require optical-to-electronic-to-optical conversion, which slows down the transmission, and optical buffers are still quite impractical. In fact, as of today, there is no way to store light and so the only possible optical buffering is to delay the signal through very long fiber lines. Fiber delay lines (FDLs) can potentially improve the network throughput and reduce the burst loss probability. In the presence of FDL buffers, the OBS reservation and release schemes have to be revised. In addition to scheduling the wavelengths at the output ports, the OBS nodes also have to manage the reservation of their available FDL buffers.

- **Variations on Burst Dropping**

Most of the OBS literature specifies that if all the resources are occupied at the moment of the burst arrival then the entire data burst is lost. An interesting OBS variation, is to divide each burst into multiple segments and in the case of resource contention, instead of dropping the entire burst, either the head or the tail segment is deflected to an alternative route to the destination.

- **Classes of Traffic**

In an OBS network, the filtering of upper layer data and the assignment of classes to bursts will occur at the edge of the network during the burst assembly process. In order to minimize the end-to-end delay of the high priority traffic, the burst assembly algorithm can vary parameters such as the pre-set timers or the maximum/minimum burst sizes. However, selecting the values for these parameters is a difficult task because of the throughput interdependency between the different classes of traffic. Here are some of the proposed solutions:

a) Classes Based On Extended Offsets: The higher priority traffic is assigned a longer offset between the transmission of its control packet and its corresponding data burst. The burst blocking probability decreases as the offset time increases. One of the main constraints of this scheme is the maximum acceptable upper layer delay, i.e., certain high priority applications cannot tolerate long pre-transmission offsets.

b) Classes based on the Optical Signal Properties and Preemption: This scheme is based on the physical quality of the optical signal such as the maximum bandwidth, the error rates, the signal to noise ratio and the spacing between the different wavelengths. These parameters are included in the control packets. A connection is established only if all of these requirements can be met, possibly using a constrained-based routing algorithm. In addition to the intrinsic physical quality, it is possible to implement priorities based on a preemption mechanism, where a lower priority burst, which is in the process of being transmitted, can be preempted by a higher priority one.

- **Multicast**

In OBS, as in wavelength-routed networks, the multicasting is achieved through light splitting, which inherently results in signal losses. Therefore, there is a limit on the number of times the signal can be split and the number of hops it can traverse. In addition, the multicasting in all WDM network is tightly coupled with wavelength allocation and is greatly dependent on the availability of wavelength converters. It is important to note, however, that the dynamic nature of OBS makes it suitable for optical multicasting because the resources of the multicast tree are reserved on a per-burst-basis.

3.2 QoS provisioning in Grid-OBS networks

The aim of this section is to evaluate benefit and limits of the OBS session declaration phase managed using GMPLS and to investigate the requirements and the extensions that should be introduced into the GMPLS protocol suite. In particular ReSerVation Protocol with Traffic Engineering extensions (RSVP-TE), Link Management Protocol (LMP) and Open Shortest Path First with TE extensions (OSPF-TE) protocol require new objects and procedures, such as new properly formatted label, new interface switching capability descriptors and proper routing and signaling procedures to allow Grid applications to exploit the benefit of the emerging powerful OBS technology.

Optical networks have been identified as the network infrastructure that would enable the widespread development of Grid computing, i.e. global Grid computing. However just offering large bandwidth connections is not sufficient for the requirements of Grid computing applications. Thus not Optical Networks but Intelligent Optical Networks must be considered as the suitable network infrastructure for global Grid computing. Intelligent Optical Networks, i.e. optical networks equipped with the Generalized Multiprotocol Label Switching (GMPLS) protocol suite, are able to dynamically adapt to both network and applications changes to satisfy the Grid computing application requirements. Intelligent optical networks are also able to offer different optical bandwidth granularities. Indeed while wavelength routed optical network research is already tackling its advanced issues, Optical Burst Switching (OBS) is gaining momentum in the optical network research field. OBS is able to offer finer optical granularity connectivity service to Grid computing applications than wavelength-routed networks. This would allow users to pay just for what they need for running their applications. Indeed, while applications that need to move large amount of data, e.g. data Grids, might require the entire bandwidth offered by all-optical connections, i.e. light

paths, other applications would just require fraction of the bandwidth. OBS represents the solution for providing Grid computing applications with the fraction of bandwidth they need while maintaining the protocol transparency advantages of wavelength routed networks. Thus by offering both wavelength routed, OBS, connectivity services the optical network infrastructure would allow not only users to pay what they asked for but also optical network service connectivity providers to better optimize their network utilization.

However different bandwidth granularities cannot be the only service offered by Intelligent Optical Networks. In particular Grid computing applications pose strict constraints on delay and delay jitter. Thus, at each granularity (i.e., wavelength routed, OBS), Intelligent Optical Network connectivity services must guarantee the suitable quality of service (QoS) considering also delay and delay jitter constraints. In addition, connectivity service differentiation must be guaranteed within each connection granularity. On the one hand guaranteeing connectivity service differentiation at the lightpath granularity appears to be achievable through the utilization of GMPLS protocol extensions for traffic engineering . On the other hand guaranteeing QoS of service at the OBS granularity is still matter of thorough research. In particular the synergy between GMPLS with traffic engineering extension control plane with OBS protocols appears to be necessary.

Finally another important issue to be addressed is the matchmaking of the application requirements to the connectivity services. For example, applications requiring a fraction of lightpath bandwidth, thus suitable for OBS, but requiring stringent constraints on delay and delay jitter might be better served by over provisioning them with a lightpath than utilizing for them an OBS connection.

Quality of Service (QoS) support for GRID Applications requires several characteristics referring to different elements such as networks, CPUs and storage devices. Typical network requirements are: end-to-end delay the traveling packet time from the sender to the receiver, delay jitter the variation in the end-to-end delay of packets between the same node pair, throughput (i.e., bandwidth) the rate at which the packets go through the network and packet loss rate the rate at which the packets are blocked, loss or corrupted [18 , 19]. Optical Burst Switching (OBS) networks will be able to satisfy GRID Applications high bandwidth requirements combining the strengths of both Wavelength Routed (WR) and Optical Packet Switching (OPS) networks, moreover several approaches for QoS provisioning in OBS networks have been proposed in the literature. The main aim here, is to provide relative service differentiation with regards to packet loss probability, nevertheless they are based on relative QoS model in which the service requirements for a given class of traffic are defined relatively to the service requirements of another class. It is possible to distinguish in:

- *Offset-based schemes* [20,21] that introduce an extra-offset time between control burst (CB) and data burst (DB) to differentiate bursts in several service classes. These technique have been proposed utilizing Just-Enough-Time (JET) protocol in buffer-less OBS networks, and it has been proved that, opportunely setting the

extra-offset time (the higher priority, the higher extra-offset time), high class bursts loss rate can be independent from lower classes traffic. The main drawback of these schemes is represented by the aware increase of end-to-end delay for high priority burst.

- *Strict priority schemes* [22], minimize high priority bursts loss rate allowing them to preempt reservations of lower priority bursts. Therefore a specific burst can be only blocked by reservations of higher class bursts or in-going transmission of lower priority bursts, in this case the end-to-end delay is proved to be less with respect to offset-based schemes, but the lower class burst loss rate is still strongly dependant on the higher priority traffic as in offset-based schemes.
- *Segmentation-based schemes* [23, 24] avoids bursts collisions in core nodes providing preemptive high class bursts combined with low class bursts segmentation and deflection. In particular when a contention occurs, lower class contending burst is divided into multiple segments and only overlapping segments are dropped or deflected. This approach can decrease low priority bursts loss rate but it significantly increases the physical layer architecture.

Other schemes propose to differentiate bursts classes allowing each class to utilized different network functionalities (e.g, extra-offset, wavelength conversion, deflection routing) considering class specific QoS requirements [25]. The burst scheduling outlined in the previous section can be used to this end, for instance partitioning the resources to be allocated by the WDS algorithm or allowing higher priority classes to use more domains for congestion resolution thus implementing some form of priority. Studies on the effectiveness of this approach has been carried on in [26] and [27], proving that significant QoS differentiation can be achieved in particular by partitioning the wavelength domain and/or allowing for more extensive use of multi-path routing.
The usefulness of end-to-end re-routing with respect to deflection routing is investigated in [28], it improves network throughput reducing nodes congestion and decreases delay jitter avoiding unpredictable delays typically introduced by deflection routing; moreover end-to-end re-routing is able to more efficiently provide network resilience in case of node or link failures.
Other proposal for OBS networks [29,30,31,32], aim to provide quantitative QoS guarantees with regard to packet loss rate, worst case end-to-end delay and throughput. These kind of QoS schemes seems to be more suitable to be applied in a Grid environment where each application needs specific QoS requirements. Proportional QoS schemes are proposed in [29,30], to adjust the service differentiation of a particular QoS metric to be proportional to particular weights that a network service provider can set; these schemes feature in advance discard of lower class optical bursts. In [31] an early dropping mechanism, which probabilistically drops lower class bursts, and a wavelength grouping mechanism, which provisions necessary wavelengths for high class busts are proposed. In [32] a possible architecture to provide quantitative QoS guarantees with respect to worst case end-to-end delay, throughput, and packet loss probability in buffer-less Labeled OBS networks is proposed. In particular [32] shows that deploying fair scheduling algorithms in both the data plane of the edge nodes and the control plane of core nodes it is possible to support a wide range of service guarantees with regards to throughput, end-to-end delay and packet loss probability.

In conclusion there are different ways to provide QoS in OBS networks, the key issues in providing QoS for Grid applications is to understand the requirements for each specific application and find out the right strategy to quantitatively provide them.

Providing Grid computing applications with resilient connectivity appears one of the QoS requirements of increasing importance. In addition maintaining, even upon failure occurrence, QoS differentiation among the connections utilized by the applications, i.e. differentiated resilience (reliability), is required. Resilience in OBS network has just started to be addressed by the optical network community [33,34]. In general OBS dynamic routing, on which hop-by-hop OBS routing is based, is able to overcome "by nature" network failure. However because of the high recovery time [33,35], mainly due to the routing table updates [36], dynamic OBS rerouting is not able to guaranteed the required QoS.

Already proposed pre-planned global rerouting based on Labeled Optical Burst Switching has shown to be promising for balancing the network load and recovery bursts after a physical network kink failure [37]. However resilient schemes based on deflection routing have shown the ability of improving the performance, in terms of burst blocking probability, of resilient schemes based on global routing updates during the failure recovery phase. In both cases the utilization of schemes based on traffic engineering extensions to GMPLS already developed for wavelength routed network might help in improving OBS network performance, in terms of burst loss probability, upon failure occurrence [34]. Previously proposed schemes are based on proposed extensions to routing and signaling protocols of the GMPLS protocol suite. Therefore routing and signaling protocols are also important for Grid-OBS resilience.

For example a better choice for the deflection path taken by the bursts involved in the failure can be obtained by utilizing a weighted stochastic approach, such as the one utilized in [34]. The approach proposed in [34] represents a scheme fairly simple to be implemented applicable to both local and global rerouting. In addition failure notification based on RSVP-TE signaling might improve failure notification time.

The main issue in utilizing resilient schemes already proposed for wavelength routed network consists in the different dynamic characteristics of OBS and wavelength routed networks. Indeed OBS network parameters, such as load, change much more quickly that the correspondent ones in wavelength routed network. A possible solutions therefore would be to apply schemes typical of OBS in the short time scale and periodically improving their performance by changing their behavior through the feedback obtained by wavelength routed alike resilient schemes.

3.2.1 Grid Differentiated Service (GridDiffServ) provisioning

The model of a grid open environment assumes that services and customers of different types, including completely new ones, can be added in or removed at run time. On such environment, the availability of resources can change at any time, and also new types of resources are continuously added to the pool as older technology is removed. Thus,

fragile mechanisms that depend on the unique characteristics of specific computing and networking platforms are likely to have a negative rather than positive impact on the long-term efficiency of the physically heterogeneous and distributed Grid environment.
A flexible, scalable and robust resource reservation and allocation scheme is required to handle any type of application (e.g. distributed supercomputing, data intensive, collaborative applications) and in turn any type of network and computational resource requirements and provide a package of fair specialized services - the Grid Differentiated Services (GridDiffServ). Application performance depends on carefully selecting the type and number of computational resources used (based on application requirements), the available network bandwidth and latency, and the location and volume of input and output data. Furthermore, optimal load balancing across heterogeneous computing and network infrastructures is also critical for both Grid network resource availability and user/application efficiency. A QoS-aware Grid network infrastructure must not be limited in providing different priorities on buffering, edge delay, network jitter, protection, restoration, latency and bandwidth, etc. but also provide variable multicast services, magnitude-aware bandwidth provisioning (steady-state, and peak demands). Moreover, must consider other critical requirements, such as user-resource and resource-resource distance, occupancy and availability, user identification priorities, the security needs and other QoS needs. For all the abovementioned criteria of GridDiffServ provisioning there is a need for a Grid–OBS infrastructure where the job requests can be classified at edge node level. Then further algorithms and processes such as control plane resource discovery mechanisms, burst aggregation, and scheduling can be applied at class of service level.
The service provisioning based on fair and specialized resource reservation services can be initialized by a job request description mechanism. This mechanism is constructed to enable jobs to be described in a standard way so their description maybe ported to and understood by different systems. For a user to be able to make use of multiple systems, therefore, it is currently necessary for them to have several job descriptions, one for each of the proprietary systems that they wish to use [38].
.

3.2.2 QoS Grid Resource Management

The ability to provide an agreed upon Quality of Service (QoS) is important for the success of the Grid, since, without it users, may be reluctant to pay for Grid services or contribute resources to Grids, which would hinder its development and limit its economic significance.
The resource manager of a Grid receives information about the job characteristics and requirements and determines when and on which processor each job will execute. The objectives that we set for the resource manager is to assign computational resources to computational tasks in an efficient and fair way, while meeting to the degree possible the QoS requirements of the individual tasks.
Efficiency in the use of resources is clearly important because this is what motivated the Grid in the first place.
Fairness is important because it is inherent in the notion of sharing, which is the raison d'etre of the Grid. Meeting the requirements of the users is important because otherwise the users will not want to use, pay, or contribute resources to the Grid.

In order to provide the agreed QoS to the users, while using the available resources efficiently (that is, on demand), the Grid resource manager has to be able to reserve (parts of) resources for the execution of specific tasks. The requirement of on demand and efficient use of resources implies that resources (or parts of resources) should be allocated to a task only for the time period during which they are actually used, and should be available to other tasks for the remaining time.

This is not accomplished by existing resource reservation protocols. Delays incurred by the transmission channel are important and must be taken into account. These delays can be significant and comparable with the burst size or even with the task execution times. To this end and in order to reserve Grid resources only for the time needed, burst carrying data and execution instruction must arrive sequentially at the resources. Thus, communication delays apart from task execution times must be incorporated for job scheduling and efficient use of resources. Specific tasks characteristics that are important for resource management include:

- ***The estimated workload of the task.*** The workload can be categorized depending on the kind of the system resource we are referring to:
  - For *computer resources,* the workload can be measured, for example, by the number of instructions of the task.
  - For *network resources,* the workload can be measured by the number of bits that have to be transferred.
  - For *storage resources,* the workload can be measured by the number bits stored and the duration of time for which they have to be stored.

- ***The variance of workload.*** Since the workload is not generally known a priori and is better modeled as a random variable, it is useful for the resource manager to have a measure of the variability of the workload around its mean.

- ***The required Quality of Service.*** The resource manager also needs to be informed of the QoS the user requires. The QoS might, in addition to the estimated workload, include the following parameters:
  - Deadline (i.e. required completion time of task)
  - Probability to miss the deadline requirements
  - Reliability (fault tolerance) requirements; if this aspect is important, the task should be scheduled on more reliable resources or on more than one processor for execution.

- ***The relation between the tasks.*** Any temporal relations between tasks could be given in the form of a directed acyclic graph (DAG), giving precedence constraints for task execution.

- ***The cost that the user is willing to pay.*** Depending on the cost that the user is willing to incur, the scheduler may send the tasks to more or less expensive resources. Also, in case some tasks have to be rejected, the cost that a user is willing to incur will

influence the choice of the tasks that are rejected. The cost of a user is not necessarily an explicit amount that is charged. Instead it may be implicitly found from the resources the user is contributing to the Grid infrastructure.

3.3 Constrained based physical layer routing and signaling in OBS control plane.

The OBS routing protocols offer the opportunity to take into consideration the physical layer characteristics of the network infrastructure as part of the routing algorithm and the Grid service offering. In addition to the information relating to the traditional Grid resource characteristics, physical layer characteristics (i.e. chromatic dispersion, polarization mode dispersion, amplifier gains, amplifier noise, launch power level, span length, loss of a pan and node, crosstalk levels) will be considered.
Grid-OBS separate control packet resembles the
Based on these parameters information, carried by the burst control packet, a set of available Grid and network resources can be identified by the OBS routers. These costs will be taken into account when finding the possible paths to establish the Grid services as and when required across the network. The Grid service will be established across the path that satisfies the service policy requirements in terms of all critical parameters specific to the requested service. This is a novel way of implementing user controlled constrained based routing across the two network domains (Grid and optical).

Grid services can register themselves within the Grid-OBS infrastructure in distributed way by constrained based routing protocol as a multi metric algorithm . Related Grid service type and parameters like (Grid service type, CPU utilization, Storage type and size ...) are available through edge OBS routers to the end Grid users as they are advertised within the control plane by routing update messages. Constrained based routing protocol picks up the sites of Grid services that meet the metrics desirable by the application or the end user, the same way that it does this function for path selection. Explicit route objects are then generated and carried into the core of the OBS network in control packet in advance to the data burst to reserve the optical network and Grid resources. These communication between the end users ‘ sites and the OBS network is done through O-UNI s and O-NNI s signaling of the edge and core OBS routers and API s of the services and applications . At the destination end, work load estimation is calculated and the application resources are synchronously reserved and allocated with the optical network resources within the task duration. Any changes about the reserved or released resources are immediately flooded by the OBS edge routers within control packets into the network. The Grid service providers and users can benefit from GMPLS extension over OBS networks to have dynamic user centric optical infrastructure.

3.4 OBS-over-GMPLS (O_2G): a Control Plane architecture for Grid services support in wide-area multi-domain optical networks

The choice of the actual Control Plane solution for an optical network is mainly driven by the fulfilment of a number of requirements (originated both from the users and from the network operator) that in some cases may have conflicting facets: e.g., the user's need for an efficient utilization of his/her connection (i.e. availability of sub-session signalling dynamics) might go against the network operator's wish to maintain a controllable and manageable infrastructure.
The following table summarizes a set of the most influencing requirements in this scope.

req no.	Description	implications on Control Plane
1	Efficient bandwidth utilization (i.e. cost-effective transport connections[1])	OBS is needed for this, no alternatives currently available on Deterministic Multiplexing[2] technologies.
2	Low blocking probability for transport connections	OBS could help, but "full"[3] TE routing and/or crankback perform much better, even if on a different timescale.
3	Resilient transport connections: i.e. availability of recovery procedures on the network	OBS deflection routing could help during the set-up, to the same extent of a pre-planned GMPLS local repair procedure with incomplete TE information or by applying RWA procedures. ASTN/GMPLS can also deploy end-to-end recovery procedures, which are more "intelligent", but it might result in a too slow reaction. The failure identification – localization – notification – reaction chain can be implemented in similar ways in OBS and ASTN/GMPLS, although with some variations (e.g. centralized vs. distributed reaction engines). The key difference between the two approaches lies in the fact that, on a node or link fault, the ASTN/GMPLS failure reaction aims to heal both the flowing traffic (Data Plane) and SCN[4] connectivity, whereas OBS only aims to heal SCN connectivity[5].
4	Controllable network, with a "manageable" Control Plane:	ASTN/GMPLS and its Control Plane low-pace dynamics are needed for this.

[1] In this table, "transport connection" is used to refer to either an end-to-end chain of transport resources (as in a classical circuit) or a part of it along an end-to-end path (as in OBS resources allocation).
[2] In Deterministic Multiplexing techonologies (both TDM – SONET, SDH -- or WDM ones) the bandwidth allocation needs to be tuned on the peak rate of the traffic and, usually, the resource allocation (i.e. SDH time-slots or DWDM lambdas) have the scope of the end-to-end path.
[3] In the scope of this discussion, "full" TE routing means advertisement of TE routing information *including* various levels of details on the status of resources allocation for each TE link, e.g. either the overall amount or the detailed list of allocated wavelengths in a fiber. This information might need to be extended if bundling is applied to pools of fibers.
[4] Signalling Communication Network.
[5] Upon a network failure in OBS, the reaction will only tend to establish new FECs (destination-based forwarding information or NHLFE associations). This will result in a new SCN and Data Plane paths towards the destination, which will be used by future BCPs and DB, respectively. The ongoing bursts flowing through the failed resource (link or node) are not going to be restored on the new path. This approach is reasonable for *relatively* short-lived bursts but might be unacceptable for *relatively* long-lived bursts, where the overall performances of the end-to-end communications would be significantly impaired. Long bursts might be well worth having their connections restored, depending on various factors: the impairment on the application, the recovery times, etc. The main consequence of the OBS approach to restoration is a possible physical decoupling between current SCN and data paths after a failure and the data paths established before the failure.

	i.e. a Control Plane whose status is synchronized with that of the Data Plane, and which can be easily know by a Network Manager	
5	Optimize the traffic balancing within the network	This requires "full" TE routing. The relative slow dynamics of ASTN/GMPLS LSPs might be compensated for with automatic LSP rerouting.
6	Dynamic set-up of transport connections through boundaries between different administrative network domains	In principle, this feature is supported by both ASTN/GMPLS and some OBS JIT implementations [39] but ASTN/GMPLS is more mature because: • The ASTN/GMPLS network interface model for inter-domain signalling and routing is the result of wide consensus among standardization bodies (ITU-T and OIF) and industry. • Applying the LSP stitching approach, E-NNI G.RSVP-TE could be used to pre-plan trunks through the E-NNI, thus implementing cut-through domains.

Table 1: Identification of Control Plane architectures in Grid-OBS deployment scenarios.

The best Control Plane architecture able to fulfil the above requirements much depends on the specific context which characterizes the network infrastructure, its users and its operator. The following table depicts two main scenarios: a network directly owned and operated by a group of "power" users, and the more general case of a third-party advanced network infrastructure, supporting different kind of users (e.g. both Grids services and other kind of premium services).

		Grid users owning the network	**Grid users interconnected through "third-party" Network Operators and sharing the infrastructure with non-Grid users (e.g. business)**
Addition of a new Virtual Organization (VO)		This implies the deployment of brand new connections *at first in terms of fibers* then through the configuration of connections, in case circuit-oriented technologies are deployed.	This translates in the configuration of new connections, i.e. circuits more frequently than fibers.
Control Plane general requirement		The network can adapt to different Control Plane architectures (the network is owned directly by users) and even incomplete CP solutions can be acceptable (e.g. signalling-only, without any dynamic routing).	The Control Plane architecture must guarantee the co-existence of Grid users and business users with their respective QoS
Session duration	**Long-lived**	→ ASTN/GMPLS CP would be the optimum to manage the automatic Bandwidth on Demand services once the fibers are laid. → "Light" Control Plane procedures (more complex than just OBS signalling but less complex than full ASTN) could fit this case particularly in small-sized networks and with a high rate of setup/release of connections.	→ ASTN CP (GMPLS + O-UNI+ E-NNI) is best suited: it is complete even if with a demanding CP burden (two/three-tiers signalling protocols, intra-domain and inter-domain full routing, link management, crankback, recovery)
	Short-lived	→ OBS native CP (e.g. JIT based) is best suited: it is fast and with a not-	→ ASTN-like Control Planes provide best services for business users but

		demanding CP burden (simple one-tier signalling protocols implemented in hw, limited routing, no link management)	with too slow dynamics for Grid users. → ASTN for business users + an enhanced OBS Control Plane (more routing intelligence and knowledge, traffic engineering, more complex recovery, etc.) for Grid users could be the solution.

Table 2: Identification of Control Plane architectures in Grid-OBS deployment scenarios.

3.4.1 Current trends in OBS and ASTN/GMPLS coexistence in support of Grid applications

Focusing on the viable Control Planes for OBS networks, different solutions have been proposed in literature [39,40,41,42,43,44,45]. Some of them refers to Labelled OBS (LOBS) and propose OBS extensions to the GMPLS protocol suite (i.e. G.RSVP-TE, G.OSPF-TE, LMP); others, are much more focused on efficient one-tier signalling (e.g. JIT/JET) and completed with brand-new and light routing procedures, e.g. based on centralized routing engines and simple signalling with a limited number of messages (e.g. just 5 in the MCNC-RDI JIT implementation: SETUP, SETUP ACK, KEEP_ALIVE, CONNECT, RELEASE).

This leads to the identification of two possibly competing research directions:

1) To build the OBS Control Plane through an extension of the [G]MPLS protocols for all-optical networks supporting VBR traffics. These approaches (e.g. LOBS) aim to inherit the traffic engineering/QoS and recovery procedures of [G]MPLS by improving the OBS performances achievable with the one-tier signalling protocols (i.e. burst blocking probability, burst recovery in case of network failures or blocked wavelengths, etc.).

2) To improve the intelligence of the OBS native signalling (i.e. one-tier based) through the implementation of light protocols to be run in very fast hardware devices. These light protocols are aimed at adding the logic for building and modifying dynamically the routing tables on OBS Network Elements or for managing the inter-domain connections, etc.

3.4.2 "OBS-over-GMPLS" (O_2G) Control Plane architecture

The two approaches described above (i.e. LOBS and OBS JIT/JET) are competing ones, if the final objective is to implement an integrated OBS/GMPLS Control Plane according to a *peer-to-peer* model, as referred by most of the literature. This is not the case of the following proposal, which starts from the assumption, partially discussed in the previous sections, that the ASTN/GMPLS and OBS Control Planes (1) have different purposes and fulfil different requirements, (2) do different jobs, (3) in different timescales.

ASTN/GMPLS is mainly targeted to long-lived connections and – from a Network Operator perspective – it is mainly aimed to speed-up and automate the procedures for setting up and healing circuits across its network and the neighbouring ones, in a multi-domain framework (through the ASTN implementation: GMPLS-based O-UNI, I-NNI, E-NNI). In this context the burden of a two/three-tier signalling, of bundled routing advertisements and link management is acceptable, if compared with the capability of

implementing a resource-based approach that automatically and in a distributed way can provide full TE and recovery, above all in the inter-domain scope
On the contrary, OBS Control Plane proposals and implementations are the optimal solution for short-lived connections, and perform better in small-sized networks. These solutions can work also in long-lived connections and dense networks, but with an impact on the achievable performances. Indeed, due to the proportional relationship between the offset time and the time spent in an OBS node to process the setup message under any kind of reservation scheme (JIT, JET, Horizon, etc., ref. [46] for details), it is evident that even with the OBS-specific one-tier signalling protocols the fast signalling efficiency decreases with the number of hops to be traversed during the setup phase.
Thus, whereas OBS can certainly fulfil the user requirement of an efficient bandwidth utilization, it would be good to limit the number of nodes involved in the OBS signalling transactions, and rely on a different Control Plane architecture to handle recovery and network resource optimization issues.
An OBS-over-GMPLS (O_2G) framework is proposed here, where an OBS-ruled edge network section runs on an ASTN/GMPLS-based core network section according to an *overlay* model, as depicted in figure 1.

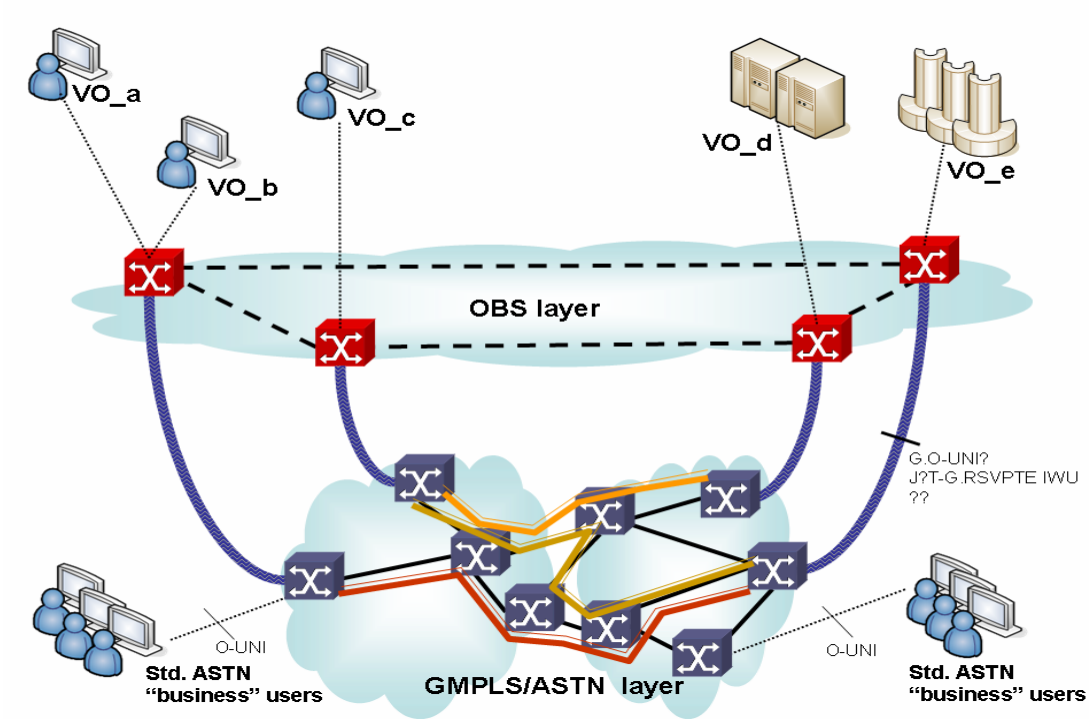

Figure 1 – O_2G overlay framework

3.4.3 Overview of the architecture

In this overlay architecture, the ASTN/GMPLS and OBS Control Planes coexist. The ASTN/GMPLS manages the network resources with a circuit granularity (lambda, waveband or fiber) and in "circuit" timescales, thus simplifying the logical topology to be exposed to and used by OBS; whereas the OBS Control Plane manages the network resources with an optical burst granularity. This overlay allows the two Control Planes to

live together and manage network resources in different network regions and in different timescales, i.e. those for which they were natively conceived, in order to blend the best of each solution (full TE and LSP recovery from ASTN/GMPLS, fast provisioning for burst from OBS CP).
According to the O_2G architecture, the edge network OBS nodes are connected throughout LSPs (acting as "virtual links", or GMPLS Forwarding Adjacencies – FAs) across the GMPLS core. Each "virtual link" between edge nodes is supported by one LSP (or a part of it, in case of waveband switching), and is perceived by the OBS edge as a link. The relationship between a virtual link and an FA-LSP is *many-to-one*: a single FA-LSP might support multiple virtual links if the LSP is FSC (fiber-switched), or just a single virtual link if the LSP is LSC (lambda-switched).
In the O_2G architecture the following types of network elements are identified, 2 belonging to the network OBS edge, 2 belonging to the network ASTN/GMPLS core, and 1 belonging to both sections:

- **Edge OBS Node (EON)**: this is a plain OBS edge node, located in the network OBS edge, which performs traffic aggregation into bursts, and originates BCP signalling.
- **Core OBS Node (CON)**: this is a plain OBS core node, located in the network OBS edge, which process BCP signalling and configures optical resources for DBs accordingly.
- **Edge GMPLS Node (EGN)**: this is a hybrid OBS/GMPLS node, located at the ASTN/GMPLS network boundary, and able to understand BCP signalling and run a GMPLS stack.
    - When operating in GMPLS timescales, it takes part in the GMPLS procedures concerning the FA-LSPs (set-up, recovery and rerouting). The GMPLS protocols run by this node need to be extended in order to provide a suitable characterization of the FA as an optical link. It also implements the O-UNI for "standard" ASTN end-users.
    - When operating in OBS timescales, this node normally provides OBS functionality: it processes incoming BCPs and allocates transport resources based on FEC information. In this context, the only difference with respect to a plain core OBS node (CON in this architecture) is that some of the outgoing links or lambdas are *virtual* (the *virtual* lambda is *locally* a physical resource but its next-hop OBS peer is located on the other side of the ASTN/GMPLS network section).
- **Core GMPLS Node (CGN)**: this is ASTN/GMPLS core node, located in the ASTN/GMPLS network core. The standard GMPLS functionality of this node needs to be extended in order to feed the EGNs with the optical layer specific information for "virtual link" emulation over the FA-LSP. This node supports (if necessary) E-NNI functionality for inter-domain connection set-up (and, optionally, recovery) procedures.
- **Core Hybrid OBS/GMPLS Node (CHN)**: the role of this node, located in the ASTN/GMPLS network core, is to support standard OBS signalling throughout the whole network if no FA-LSP cut-through is available (yet). The Control Plane

of this node is the sum of those of CON and CGN. The main difference between a CHN and an EGN is that the EGN can join OBS' and GMPLS' lambdas to inject the bursts into the FA tunnel: its role is to cross-connect OBS' lambdas with GMPLS' ones. The CHN, on the contrary, will make all the physical resources available to both OBS and GMPLS in a flexible way, but *exclusively*: the lambda cross-connections must be homogeneous in ownership: either GMPLS-to-GMPLS or OBS-to-OBS. Comparing the OBS/GMPLS overlay with IP/MPLS, the EGN is homologous to a LER, whereas the CHN to an hybrid IP+MPLS node (i.e. an IP router with plain destination-based forwarding plus an LSR).

In O_2G, the coexistence of OBS and GMPLS is a key issue in EGNs. This node has different kind of adjacencies in place: fixed OBS adjacencies with CONs and EONs, fixed GMPLS adjacencies with CGN and flexible OBS or GMPLS adjacencies with CHNs.
The "virtual-link" FA-LSPs throughout the GMPLS core can be:

1. Pre-planned by the network operator on the basis of traffic forecasts (and, of course, re-planned at regular intervals), and
    - planned by an ASTN/GMPLS network manager and set-up via management according to an SPC model
    - planned by the LSP ingress node and set-up via signalling
2. Automatically set-up by EGNs on the basis of specific events or some statistics on the DBs traffic, e.g.:
    - when the amount of BCP set-ups passing through a fixed couple of EGN nodes is above a specified threshold, these EGNs might decide to establish one or more FA-LSPs between them and "tunnel" the OBS traffic through these virtual links. This option requires that some sections in the network core feature CHNs;
    - in case of a BCP request blocking due to busy or in recovery state LSP.
3. Set-up (and torn down) on user demand, when a session[6] between two or more user termination nodes begins.

This model brings several advantages:

- Efficient usage of core circuit resources thanks to the edge OBS multiplexing
- Perfectly manageable core thanks to GMPLS
- A smaller number of nodes have resources (i.e. lambdas) which "fluctuate" according to the OBS timescales $\Rightarrow$ the network status is more manageable
- Resilient core: the "virtual links" between edge nodes are implemented by resilient LSPs, which results in always healthy links between OBS nodes. The only problems arise if an LSP failure occurs during a burst transit.

[6] [6] A "session" is intended here as the connectivity relationship between two or more user nodes where the traffic is exchanged (in bursts).

- Capability to create trunks in the core network and (which is more important) through the boundaries between different administrative domains (e.g. trunking with waveband-switched LSPs)
- Smaller number of burst multiplexing points ⇒ operational applicability of finer OBS multiplexing techniques such as in JET or, on the contrary, similarity in blocking probability performance of different OBS signalling paradigms (with/without Delayed Reservation and void filling).

The price paid for these improvements is a loss of lambda switching flexibility within the core network, where a pool of OXCs is replaced by a rigid link (i.e. LSP).
Figure 2 depicts the O_2G overlay architecture; specific details on the various functional aspects of it are discussed in the following sections.

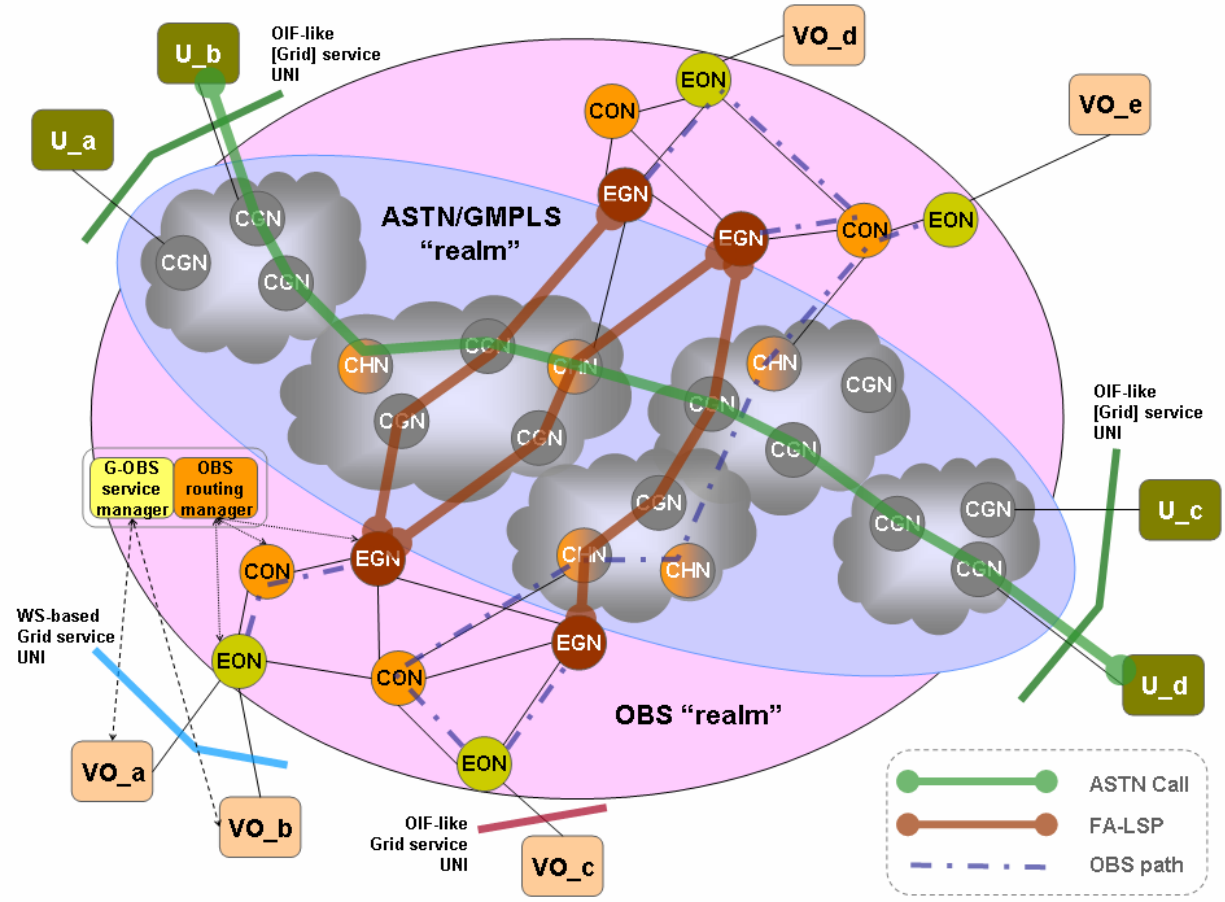

Figure 2 – O_2G architecture

3.4.4 O_2G end-to-end "session" definition

The end-to-end transport connection between the end-users is here referred to as *"session"*, in order to abstract from any specific concept (either closer to a circuit-switched or burst-switched paradigm) or directionality. The "session" between two or more end-users is a long-lived data transfer relationship (long with respect to the bursts

timescale), which O_2G practically implements in different segments, each one according to different switching paradigms. The reference splitting is a core segment supported by an FA-LSP (and thus based on circuit-switching), with two burst-switching edges, implemented in the OBS domain.
Although each single segment might be unidirectional only (as natively imposed by OBS) or bidirectional (as allowed by ASTN/GMPLS), the concept of "session" is inherently bidirectional: its circuit-switched core segment will use a bidirectional FA-LSP, whereas each edge segment will be based on a couple of (possibly differently routed) OBS paths with towards the end-user.

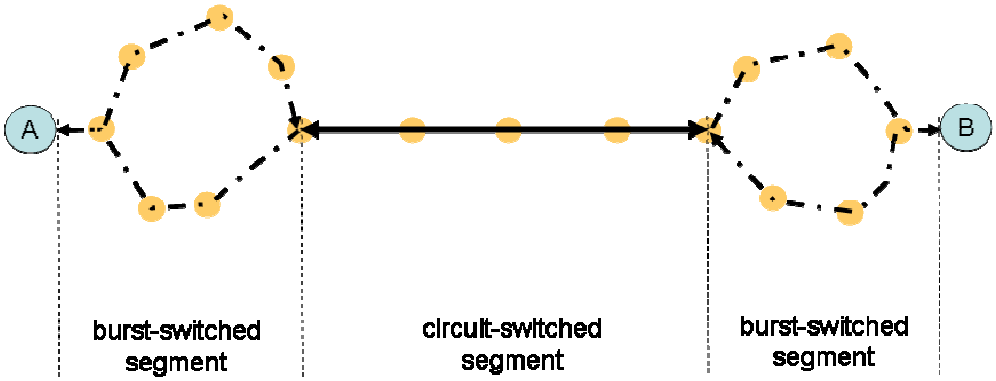

Figure 3 – O_2G end-to-end session model

3.4.5 OBS SCN over ASTN/GMPLS: the OBS Control Plane end-to-end continuity

The OBS signalling has to flow transparently through the ASTN/GMPLS network section between each couple of EGNs. This can be accomplished in two ways:

1. establishing an "OBS Control FA LSP" (OC-LSP) between each couple of EGNs, to be shared for the transport of BCPs related to all the FA-LSPs interconnecting those EGNs;
2. establishing a GMPLS Control Plane adjacency between the each EGNs couple sharing some FA-LSP, and transport the BCP signalling on the GMPLS SCN.

In the former case, the BCP signalling is conveyed in a really transparent way (optically) through the ASTN/GMPLS network, with no OEO conversion delays in SCN routing nodes. This is a much faster approach for BCP forwarding, if compared to OBS on the same *physical* network, i.e. if the whole network were OBS, the BCPs would have experienced an OEO (+ message processing, of course) delay on each traversed core node. This approach can be further refined with the adoption of GMPLS recovery techniques to enhance the resilience of the OC-LSP.
The latter case, i.e. GMPLS Control Plane adjacency, introduces an end-to-end (between ingress EGN to the egress EGN) forwarding delay of BCPs due to OEO conversions + message processing + routing decisions on the involved GMPLS SCN nodes. However, due to the above considerations, this delay might still be tolerated in the burst offset budget. The resilience of this communication relies on the SCN reactivity to routing failures and it is expected to have worse performances with respect to the previous approach. On the contrary, this approach is more flexible in terms of resource allocation,

as it does not require the deployment of dedicated GMPLS data plane resource to support OBS Control Plane communication.
Furthermore, the approach based on GMPLS Control Plane adjacency needs the adoption of the LMP Control Channel Management procedure, for the following purposes:

- the classical availability of an heartbeat on the SCN communication between the two nodes (i.e. keep-alive procedure)
- an end-to-end address resolution for OBS between the data plane port addresses and corresponding control interfaces on the same node. This is both a requirement for the LMP to set-up the Control Channel (CC), and a service offered by it to the OBS Control Plane on EGN when a BCP has to be sent to the other side of the ASTN/GMPLS network. The binding between remote EGN data plane ports and control interfaces can be learnt by both EGNs during the set-up of the first FA-LSP between them, if proper information is conveyed in G.RSVP-TE opaque objects.

Concerning the problem of address resolution, the first approach based on OC-LSPs has one more advantage; of course it requires the creation of a couple of virtual control interfaces at the two ends of the OC-LSP, but with no need to exchange addressing information. Since the control channel built over the OC-LSP is a point-to-point link, each EGN will send OBS control information to the remote end using a default multicast destination address (e.g. 224.0.0.1).
The main discrepancy between the two proposed approaches largely depends on which technology the ASTN/GMPLS SCN is based on, with respect to an OBS SCN where BCPs travel on dedicated lambdas with OEO conversion at OXC. In cases where the ASTN/GMPLS SCN is implemented on completely electrical technology with poor performance in terms of delay and bandwidth, the first approach (OC-LSP) might be mandatory to preserve OBS Control Plane performances.

3.4.6 Addressing

According to the selected overlay model, the OBS and the ASTN/GMPLS network sections have two separate addressing spaces. The address resolution of OBS control interfaces across the ASTN/GMPLS core network much depends on the selected approach for overlaying the OBS SCN on the core network, and might work according to the requirements and procedures introduced in section 3.4.5.
In principle, the O_2G architecture does not preclude any choice for the addressing schemes within the two data planes. When applicable, the usage of IPv4 and Unnumbered data plane addressing in the ASTN/GMPLS network is suggested, in order to help reduce the overall size of routing and signalling messages.[7]
The only specific addressing requirement in O_2G concerns the EGN "tributary" data interfaces, i.e. those interfaces going towards the OBS edge section, and made available for multiplexing onto FA-LSPs. In the perspective of supporting administrative heterogeneity in the ASTN/GMPLS network core, these data interfaces need to be tagged and exported as *Transport Network Addresses (TNAs)*. These TNAs will have to be

[7] G.RSVP-TE messages tend to become quite cumbersome when dealing with paths with multiple IPv6 data interfaces, and adopting procedures such as explicit source routing, recovery, crankback and detailed route recording (due to extended ERO, RRO and XRO objects).

advertised across the ASTN/GMPLS network core, through E-NNI boundaries, up to the "internal UNI" (i.e. other EGNs or UNI-N nodes). The standard G.OSPF-TE procedures need to be extended in order to distribute this detailed information on tributary data interfaces inside a routing domain and, then through the E-NNI (which already supports this).

3.4.7 Issues on administrative ownership of network sections

The administrative ownership of the various network sections in the O_2G architecture is for further study. However, the architecture has no specific requirements in terms of stiff administrative boundaries, e.g.:

- A complete O_2G architecture (i.e. with both an OBS edge and ASTN/GMPLS core) can fit a single administrative domain; this could be the case of an ASTN/GMPLS network operator that enhances its network services for Grid users by adding an OBS aggregation layer.
- The administrative boundary can run between the OBS edge and the ASTN/GMPLS core; this could be the case of a group of Grid users managing their own purely OBS network, and seeking for long-distance transport services from an ASTN/GMPLS third-party network operator, and sharing their capacity by multiplexing different OBS sessions on procured GMPLS LSPs.

3.4.8 User transport service interfaces

As depicted in figure 2, in the O_2G architecture both OBS and ASTN/GMPLS Control Plane can have a direct "adjacency" with the end-users. For this reason, the architecture needs to expose different kinds of User Network Interfaces in order to support the diversity of users' requirements in terms of bandwidth flexibility (i.e. burst- or circuit-switched connections), and to preserve interoperability with end-points supporting either an OBS or a ASTN/GMPLS UNI.
Three models have been identified to convey the transport service requests; the first two proposals refer to users served through the OBS edge network section, the third proposal addresses those users served directly by the ASTN/GMPLS network.

1. Standard Grid service transactions (e.g. WS-based), translated into network resource management by a "G-OBS service manager", whose operations are coupled with an "OBS routing manager" (e.g. a *Routing Data Node – RDN[45]*). The UNI service request results in a proper (centralized) configuration of the OBS SCN for the subsequent transport of BCPs (and DBs, due to the topological identity between SCN and data network in OBS).
2. An OIF-like UNI extended to support OBS burst signalling and, optionally, Grid applications specificity. In this perspective, the service management is handled by the UNI-N node, according to a distributed model. The Grid users interact directly with the network SCN through the UNI for *session* set-up / tear-down, BCP signalling initiation requests and, possibly, for SCN advanced usages (e.g. advertisement of Grid resources).
3. A standard OIF UNI for ASTN/GMPLS end-users, optionally extended to support Grid services advertisement and declaration.

Case 2 and 3, when Grid service-specific extensions are supported, have a relevant consequence on the Control Plane of the UNI-N node. In fact, be it either a CGN or a "lighter" EON, it must be equipped with a proper routing engine to manage the distributed advertisement of Grid-level resources.
As discussed above, the request of an end-to-end connection (aka session) between network users will be supported by core FA-LSPs, and might (or might not) result in their set-up. In the latter case, a number of issues need to be considered, depending on the selected UNI model.
In case 3, the user's connection is completely based on the end-to-end ASTN call between the UNI-N nodes where the end-point are attached; thus, the set-up of the transport connection evolves according the standard mechanisms for call set-up in ASTN.
In case 1 the user's connection is based in its core part on a FA-LSP, whose set-up needs to be coordinated with the end-to-end session set-up. This action is handled by the management entities made available by the OBS network and in charge of setting up the session (the Grid service manager and the OBS routing manager), which instructs the two EGNs involved in the FA-LSP to signal it throughout the GMPLS.
Case 2 is the most challenging from this point of view. The information on the end-to-end session is available at the UNI-N node where, as a consequence, the planning of the consequent new FA-LSP takes place. This has two main implications on the Control Plane of the UNI-N engines:

- The UNI-N has to take part in both ASTN/GMPLS routing plane (at least in order to gain summarized information about the ASTN/GMPLS core network TE topology) and OBS routing information (e.g. by interacting with the OBS routing manager). The main objective is to determine the "best" three segments of the requested session, i.e. the "best" couple of EGNs to be used to traverse the ASTN/GMPLS core.

- The UNI-N has to "remotely" trigger an FA-LSP set-up on the selected ingress EGN. Two alternatives are proposed here:

    1. Opaquely piggybacking, on OBS signalling, of GMPLS-specific information about the requested FA-LSP.

    2. Explicitly request the FA-LSP at the ingress EGN, by means of extensions to the GMPLS signalling (e.g. using a Notify message whose context is focused on the requested LSP[8]).

  The two proposed approaches have different pros and cons: the first one does not require the presence of a G.RSVP-TE protocol stack on the UNI-N engine, and might result in a simpler implementation. However, in this case, the FA-LSP set-up is not decoupled from the first BCP signalling in the considered session, and might produce a "default" discarding of a number of BCPs for the first bursts. On the contrary, the second case can allow to set-up the FA-LSP contextually to the session set-up, with no binding to the session bursts.

[8] The acceptance of such Notify message by the destination node (i.e. the ingress EGN) would result in a non-compliant behaviour, since the LSP does not exist yet. Specific extensions to Notify processing rules and objects are needed to support this approach.

3.4.9 Inter-domain operations

In the O_2G architecture the connection set-up across the boundary between different administrative domains is operated according to the ASTN model for inter-domain (E-NNI) routing and signalling procedures.

3.4.10 Relationship with LOBS and/or static routing

The pre-planned or dynamic set-up of FA-LSPs, as discussed in 0, implies a reduced flexibility in the OBS network routing.

In details, once a set of FA-LSPs (<EGN_x1 ↔ EGN_y1>, ..., <EGN_x*n* ↔ EGN_y*n*>) has been set-up to serve as a cut-through for an end-to-end session between users attached to EON_a and EON_b (and, for sake of multiplexing, between other EONs couples), the routing of BCPs in the edge network section will have to be pin-holed to EGNs belonging to the FA-LSPs set, i.e. it needs to guarantee that the actual path for DBs will always include EGNs in that set of couples.

This "route planning" capability is one of the features that could be provided by Labelled OBS (LOBS,[39]), but the burden introduced by an MPLS Control Plane on the OBS SCN seems to be excessive for the targeted benefit.

As an alternative to LOBS, the end-to-end session set-up procedure can be extended to configure properly a pool of available static routes across the OBS network section between the EON and the eligible EGNs, based on destination-based routing tables design. This route planning, if periodically refreshed, can offer a simple and viable solution.

3.4.11 O_2G extensions to existing Control Plane functionality

Although GMPLS natively supports Lambda Switch Capable (LSC) and Fiber Switch Capable (FSC) interfaces, some more extensions are needed for the DWDM operation with an overlay OBS network. These extensions are aimed at enhancing the routing of LSPs at the Data Plane and also at improving the information on the resulting Forwarding Adjacencies available as a link for the edge OBS signalling. The key challenge here is to model the whole FA-LSP as an optical link, to be summarized by the EGN and fed into the OBS TE routing plane.

Some of these information elements are:

1. The number of free/allocated wavelengths on a fiber, which could enhance the routing decisions by limiting the number of wavelength conversions
2. Instructions on wavelength converters, optical transmission impairments (e.g. PMD) and signal quality (e.g. OSNR)[47,48], which could enhance the routing decisions when computing an FA-LSP into the GMPLS domain and could be used by signalling for configuring specific node behaviours for the resources to be allocated (e.g. possible setup of a OEO regeneration due to estimated signal degradation, forced wavelength conversion, etc.).
3. LMP extensions for optical link monitoring and bundling (applying some results from IETF CCAMP work as specified in RFC 4209 [49]), in conjunction with the availability of runtime estimates for BER, detection of LOS (Loss of Signal) conditions, possible estimates / measurements of optical impairments and/or optical jitter.

4. Some possible extensions in addition to setup/holding priorities, which could be useful in the Grid-DiffServ model.
5. The FA-LSPs cross-connection set-up does not follow the standard ASTN/GMPLS procedures; in particular, the two end-point resource allocations at EGNs, i.e. cross-connection between the tributary fiber termination point and the line fiber termination point, do not have to take place: in fact, their cross-connection occurs when bursts are dynamically multiplexed onto the FA-LSP. In those EGNs supporting both FA-LSPs and legacy LSPs, this behaviour coexists with normal LSP set-up procedures and specific ruling information need to be conveyed from the LSP ingress to egress node by means of proper G.RSVP-TE signalling extensions.
6. GMPLS signalling speed-up under some circumstances, e.g. when FA-LSPs set-up is triggered by EGNs to overcome some emergency condition (e.g., as discussed above, blocking conditions on ports entering the ASTN/GMPLS core). When conflicting procedures are disabled (e.g. crankback), the GMPLS signalling (G.RSVP-TE Finite State Machine) can in principle allow a 1-tier set-up of network resources[9], with an offset-delayed transmission of data traffic. This approach is still different from OBS signalling, since the result will still be a complete circuit set-up, but can achieve better set-up performances with respect to standard GMPLS set-up procedures.
7. G.RSVP-TE opaque extensions for EGN-to-EGN address resolution purposes, as explained in section 3.4.3

Exporting some of these information elements into the OBS layer as FA features can improve the network knowledge available at the burst setup phase. The main consumers of such information are the OBS routing engines (be them centralized – as the OBS routing manager – or distributed) that, in this scenario, are responsible for selecting the best fitting LSP pair (Control + Data) for the optical burst. However, some possible extensions to signalling could also derive from the low-level definition of the O_2G architecture (per-protocol extensions). This much depends on the optical layer information selected for being conveyed by GMPLS and OBS signalling extensions, and on the level of integration with Grid service-specific information.
Concerning the last issue, an eligible reference model can be the G-UNI semantics, to be translated into specific extensions for real network interfaces (e.g. between the VO and the OBS domains, between the OBS and the GMPLS domains, between peering domains, etc.). The application of the G-UNI semantics throughout the network will promote the

[9] This behaviour has been already adopted in several implementations, by using an "advanced reservation" scheme where the switch cross-connections are planned and requested soon after processing the Path message. However, the start of data transmission at the ingress and egress nodes is usually delayed until Resv or ResvConf messages, respectively, have been successfully received and processed. This behaviour is recommended in SONET or SDH networks, where the circuit set-up sequence is: (a) creation of cross-connections, (b) enabling alarms detection in the various sections and (c) start injecting traffic. In these TDM networks, if alarms enabling is performed before all the path cross-connections have been completed, it might happen that spurious alarms are generated and hit those interfaces where alarms detection has been enabled. The 2nd and 3rd signalling tiers (Resv and ResvConf G.RSVP-TE messages) are usually adopted to ensure that all the cross-connections are successfully completed (Resv upstream direction) and alarms detection has been enabled on all nodes (ResvConf downstream direction).

seamless integration of Grids and network layer, thus enabling the Grid-OBS scenario. Consequently, the OBS Control Plane engine in the mixed OBS/GMPLS nodes (EGNs and CHNs) does not interact directly with the switch fabric via SNMP/TL1 interfaces as usual, but through the mediation of the GMPLS resource control functionality. Therefore, in such nodes some interworking functions between the OBS and GMPLS procedures need to be defined.

5. Advanced network concepts, solutions and specific implementations

5.1 OBS for consumer Grid applications

For the average home user today, the network cannot sustain Grid computing. With a home access bandwidth of only a few Mbps, to at most 100 Mbps download speeds, and an order of magnitude smaller upload speeds, transmission of jobs would simply take too long. However, if the current trend holds and bandwidth availability (doubling each year) keeps growing faster than the computing power (at most doubling every 18 months) of an average end user, tapping into the Grid at home becomes viable.

Let us assume that in such a future Grid, home users are connected through a symmetrical access link offering a bandwidth of about 2.5 Gbps (in the optical range). While this kind of bandwidth is certainly not readily available to end-users at the time of writing, extrapolation of past trends shows that within 15 years such an evolution can be expected. Indeed, a typical broadband connection offers around 4 Mbps download speeds and 512 Kbps upload speeds. This means download bandwidth will have reached 2.5 Gbps within the next 10 years, and the same upload bandwidth will be available within 15 years. In analogy, if computational capacity doubles every 18 months, an increase in high-end desktop PC performance with a factor in the order of magnitude 210 should be envisaged. The resulting processing power will offer the possibility to process extremely demanding applications (by today's standards) on an ordinary desktop PC. However, as we will show, it is reasonable to assume that application demands will experience a similar increase in their requirements, making it unfeasible to execute them locally. The needed aggregate power for these applications is drawn from the Grid, where end users share their otherwise idle resources (most desktop computers have a low average processing load) and commercial providers offer dedicated computing farms (with a processing power comparable to that of hundreds or thousands of desktop PCs). This means that in this future Grid, a large user base will have direct access to a vast pool of shared resources as access bandwidth catches up with processing power.

In what follows we present some typical application requirements and their impact on the underlying Grid system, indicating that existing Grid infrastructures will fail to cater for their needs. A first application example comes from the area of multimedia editing; video editing applications are widely adopted, and allow users to manipulate video clips, add effects, restore films etc. Advances in recording, visualization and video effects technology will demand more computational and storage capacity, especially if the editing is to be performed within a reasonable time frame (e.g. allowing user feedback).

More specifically, 1080p High Definition Television (HDTV) [50] offers a resolution of 1920x1080, which amounts to about 2 MPixel per frame. Suppose now that a user would like to evaluate an effect for 10 different options, where applying an effect requires 10 floating-point operations per pixel per frame. It follows then that processing a 10 second clip (25 fps) already requires over 50 GFlop. This will take about 0.5 s to complete locally (we assume local processing power is 100 GFlops), while execution on a

provider's resource should only take 5 ms (assuming the capacity of providers is a factor 100 higher). Transmission time of 10 s of compressed HDTV video (bitrate 20 Mbit/s or a 25 MB filesize) on a 2.5 Gbit/s access link is 80ms. While the 2.5 Gbps is likely to be realized through optical technologies, it is unfeasible to assume that each end user is allowed to set up end-to-end wavelength paths for each multimedia editing operation. Indeed, unless wavelength path set-up times were to decrease sharply (currently in the range of 100 ms), the use of optical circuit switching (OCS) would waste a considerable amount of network resources and one would have to devise a mechanism able to handle path set-up and tear-down requests from vast amounts of users.

A second application example is the online visualization of (and interaction with) a virtual environment. Virtual environments are typically made up of various objects, described by their shape, size, location, etc. Also, different textures are applied on these objects. A user should not only be able to visualize selected scenes in the environment by adjusting his viewing angle, but should also be able to interact with the rendered objects. Usually the description of a scene can be realized in limited storage space, the size of a texture being limited to a few kilobytes. Thus a scene can be stored in a rather small storage space, typically around a few Megabytes. However, rendering the scene is a different problem altogether; if we demand a performance of 300 million polygons per second, computational capacities as large as 10000 GFlops are required [51]. Clearly, the rendering of these scenes, preferably in real-time, is unfeasible using only local resources. Suppose a user has at its disposal an archive of different scene descriptions, with a requested frame rate of 25 frames per second. This amounts to a latency smaller than 40 ms between the submission of the scene description, and the actual displaying of the scene. Assuming a scene is 2.5 MB in size, we obtain a transmission time of only 8 ms per frame (excluding overhead); this leaves us with about 30 ms for processing and retransmission of the final rendering, which should be possible with the given capacities of the (local) resource providers. Considering the delay associated with setting up an optical circuit, OCS can only be used when a user employs the same resource, hereby severely limiting the flexibility of the Grid concept. On the other hand, the lack of adequate QoS in the standard IP protocol makes it near-impossible to meet the strict real-time constraints.

When looking at the requirements of Grid technology for consumer applications, the following conclusions can be drawn:

The current network solutions (OCS for computational and through the current internet for peer-to-peer) unsuited for providing Grid access to everyone. The dedicated infrastructure will be too wasteful and inflexible, while the request/grant based architecture with electronic management of Grid resources will be too complex.

To overcome these problems, a new infrastructure will be required. There is no doubt that these will be based on optics, in particular OBS based architecture will be well suited to the task: low processing, with high resource utilization and simple control.

5.1.1 Self-organised OBS network for consumer Grids

It is usually assumed that OBS networks employ shortest path routing, seeking to minimize the end-to-end delay. It is however well known that this approach may lead to inefficient usage of network resources; certain links are hardly used, while others can become severely congested, which of course leads to sub-optimal network performance. This is especially true when the burst dropping probability is the main metric of interest, as is usually the case in OBS networks. Several approaches have been proposed to overcome this problem, such as deflection routing and multi-path routing. In any case, both the sender and the receiver are usually known in traditional data transfers. This differs from a Grid OBS network where the destination is not always known, as we'll show in the next section.

- **Anycast Routing in Grid OBS networks**

In the consumer Grid scenario [52,53,54], it doesn't matter where exactly the job is processed. Instead, the user is only interested in the fact that his job is processed within certain predetermined requirements. In general, there will exist multiple locations where a job can be executed, and the selection of a suitable resource is left to the routing protocol. This represents a shift in the nature of the employed routing algorithm; whereas previously bursts had an exact destination, now we only require the burst to be sent to any end node capable of processing the burst. The former approach is called unicast routing, while the latter is usually denoted by anycast routing. [55,56,57]

- **From User to Resource**

The basic operation of the Grid network is now as follows. First, the user realizes that a computing task cannot be completed within a reasonable timeframe on the local system, and decides to post it on the Grid to accelerate processing. The job is then transformed in an optical burst (containing code and data), accompanied by a header indicating various parameters (e.g. processing, storage and policy requirements). Note that a very important design decision has been made, i.e. the mapping of one job onto one optical burst. As discussed earlier, no destination address is needed, and thus the burst is simply handed over to the OBS network. The burst travels along a link, while the intermediate routers are not notified in advance of its arrival, much like JIT or JET based schemes. On arrival of the burst, an intermediate router decides on the fly where to forward the burst, based on information contained in the preceding header and on network and resource status information. Examples of such information are link load and blocking probability, delay requirements, estimated free computing or storage capacity which can be reached through a certain interface, and estimated computing and storage requirements of the burst. Since the end user doesn’t specify the network location where the burst will be processed, the job is scheduled implicitly through its progress in the network. This makes the Grid architecture completely distributed, which naturally implies better scalability and robustness. Note that an intermediate router does not need a detailed view of where the resources are located and how much (free) capacity they have. As long as there is enough information to push the burst closer to a suitable destination, a good decision can be made. This means that the aggregation of status information can be used to reduce control traffic.

- **Processing a Job**

Each intermediate router in the network goes through the same process, and eventually the burst arrives at a Grid resource. If this resource is able to handle the job contained within the burst, it will process it. If this is not the case, a deflection mechanism can be used to repost the job in the OBS network. It is also possible to drop a burst which cannot be timely processed.

- **From Resource to User**

Once the job is completed, its results must be delivered back to the user (most likely where the burst originated). Here the asymmetry of the Grid OBS network becomes clear; although posting a job uses the anycast paradigm, sending results back most likely will not. There is a distinct return address, and more traditional forwarding solutions have to be used. A variety of options and choices can be made, depending on such parameters as the processing time, storage availability, size of results, etc. For instance, a real time application requires its results to be transmitted as fast as possible, while for an offline calculation the results can be stored on the processing node until network availability improves. Also, we can consider a returning burst to be "more valuable" than one which hasn't been processed yet. Naturally, this notion gives rise to the introduction of different QoS classes in the network traffic.

- **Burst Correlation**

Up until now, we have assumed that all bursts are sent completely independent of each other in the network. However, we will show that it can be advantageous to dispose of a method to send consecutive bursts to the same resource.

a) The proposal of mapping one job onto one optical burst is mainly inspired by the simplicity and general application of this approach. However, this technique will prove insufficient whenever jobs are generated which are too large to fit into one optical burst. In this case, the original job has to be segmented into smaller sub-bursts, which are sent individually in the network. The routing algorithm must be adapted to make sure these sub-bursts arrive at the same resource. Also, resources must contain the functionality to reassemble the individual segments into the original job request.

b) A second scenario where burst correlation can be useful is for specific applications which can reuse input and output data of preceding bursts. For instance, in a virtual reality application, there is no need to re-render the complete scene when the user changes his viewing angle of the scene. Instead, it is better to make use of the rendering results of a previous burst, and incorporate only the changes generated by the user's actions. Note though that specific support for this feature will have to be built into the application logic.

Because of the architectural requirement to scale to large numbers of users, it is impossible to maintain the forwarding decision of each burst in all routers. An alternative approach is to let the user wait for the results of the first burst, extract the address of the

employed resource, and send all following bursts to the same destination address. Yet another possibility is that the first burst sets up a path which is followed by all later bursts, similar to the label switching technique. Aggregation techniques may be applicable too, such as merging common portions of serveral OBS paths, very like merging and stacking in label switching. As a logical extension this may result in OCS-like operation (wavelength switching), supporting the more static portions of the network.

- **Robustness**

Robustness of a network is typically evaluated based on the number of requests (jobs in our case) that cannot be handled whenever resources are failing. The heterogeneous nature of the Grid implies two types of resources can fail; the network resources (links and routers) and the server resources (the processing elements). We describe two methods to introduce robustness against failing resources of both types.

a) Spare capacity
Before deployment, a network is usually dimensioned based on load estimates or experienced job request rates. In case more network or server resources are introduced in the network than are strictly necessary, this remainder of capacity can be used in case certain Grid components fail. Research needs to be done on different restoration strategies, focusing on how and when this spare capacity will be utilized.

b) Duplicate Submission
If the same job is sent into the network more than once, the possibility that this job reaches a different server resource, or reaches the same server but arrived there over a different path, is non-negligable. Thus, this method can also introduce robustness in the Grid OBS network. Observe though that more capacity is used than strictly necessary.

5.1.2 Control plane issues for consumer Grid application

When looking at the requirements of Grid technology for consumer applications, the following conclusions can be drawn:

- It is economically unsound to build a dedicated network for each application. Although there exist several high bandwidth and computationally intensive applications, constructing a seperate network to which individual users connect, seems unrealistic. The current convergence of phone, television and data networks ("triple play") clearly proves this point.
- Grid service requests will be, in most cases, highly unpredictable, implying a dedicated, static infrastructure is not the most efficient solution.
- The sheer potential volume of requests makes electronic processing highly complex. In other words, we need to simplify intelligence in the network as much as possible, as well as use optics wherever appropiate to deal with the huge bandwidth requirements.
- In many cases, the transmission times (job sizes) will be rather short (few 100 μs to tens of ms). This means that using end-to-end circuit switched connections will prove to be too wasteful, as the holding time of a wavelength path will be too

small compared to its setup time. Real time applications place even further importance on this point.

We can easily deduce several essential requirements which the control and signaling plane should be able to satisfy:

- The ability for new application types to be deployed quickly and efficiently, which implies a flexible control plane is required. Indeed, as mentioned before, it is infeasible to build separate networks for each application type. As such, the basic infrastructure offered by the OBS network should be able to support all types of applications, each with its own typical resource usage patterns.
- Flexibility also indicates that the features offered by the control plane should be of relative simplicity. Features which are usable by only one application group introduce complexity in the signaling protocols and can usually be assembled from simpler, generally deployable components.
- Support for a huge number of users implies scalability of the control plane is essential. In light of this, research should focus on minimizing the control and signaling traffic. This point becomes even more important when users have a highly unpredictable traffic pattern.
- Support for highly dynamic user access patterns means the control plane should be adaptable to the Grid's status, e.g. by reducing signaling data in favor of more actual data transfers.
- Sufficient levels of speed and flexibility in the control plane are imposed by real time applications. As we mentioned repeatedly, the main disadvantage of traditional circuit switching is its inability to react quickly to dynamic traffic demands. Adding real time constraints to this setting is only possible with networks which have a minimal latency imposed by the control plane, thus leaving more time for the actual data transfers.

5.2 Wavelength Routed Optical Burst Switching for GRID

In this section wavelength routed optical burst switching (WR-OBS) for GRID application is presented. This solution utilizes traffic aggregation and wavelength routing technology. It aims to provide a kind of network architecture able to fulfill both existing data-intensive and future Grid application requirements and make efficient use of network resources. The solution is based on two-way resource reservation, in which the optical network can provide a more reliable service with longer end-to-end delay for GRID applications.

5.2.1 WR-OBS Network Architecture

In WR-OBS network there are edge routers and core routers, which have similar functionality compared with JET-OBS. Edge routers are responsible for accessing incoming traffic and building the packets into data burst and generating corresponding control packet (BCH). Core routers take charge of dealing with BCHs and setting up the optical switch. However, unlike JET-OBS, WR-OBS has a two-way resource reservation mechanism. Based on the control architecture, WR-OBS can be divided into centralized control WR-OBS and distributed control WR-OBS.

- Centralized control WR-OBS

In centralized control WR-OBS, there is a control node residing in the center of the core network to deal with all the bandwidth requests. All the core routers will distribute real time information about bandwidth allocation to this control unit. By this means, the control node has the ability to make exact decisions to all requests. The decision will be sent back to tell the source edge node whether send out the data burst or not. Concentrating all the processing and buffering within the edge of the network enables a bufferless core network simplifying the design of optical cross connects (OXCs) in the core network. Once a burst is released into the core network its further latency depends only on the propagation delay since there is no buffering in core nodes. This is especially important for time-critical traffic and cannot be achieved with the currently implemented IP-router infrastructure that provides hop-by hop forwarding only. However, the centralized control mechanism confines the scalability of the whole network.

- Distributed control WR-OBS

In distributed control WR-OBS, each core node has its own intelligence to manage resource requests delivered to it and make its proper decision based on its own information about the whole network. Clearly, by employing this distributed control mechanism, it is not necessary to keep the powerful centralized control node and distribute real-time information to it, which makes up a more feasible and scalable network. In distributed WR-OBS the source edge node sends out a control packet to the core network for resource reservation along a hop-by-hop lightpath. Whatever there is enough bandwidth for the data burst along the whole lightpath, another control packet will be generated and sent to the source edge node. By this means, an end-to-end lightpath is reserved for the data burst. In this distributed control architecture, the data bursts have to experience a time delay for end-to-end resource reservation, which is the

round trip time (RTT) plus processing time for control packet in each intermediate node. In a network with the span of several hundred kilometers, this time delay has the order of several milliseconds, which is a typical forwarding time of IP routers.

- Comparison between WR-OBS and JET-OBS

WR-OBS network is quite similar with JET-OBS network. Both of them have edge routes for burst assembly and core routes for optical switching. The difference between them lies in the resource reservation process.

1. In JET-OBS, data burst will be sent out without the notification of whether the BCH successes in resource reservation, thus it is rather simple to be implemented. End-to-end delay is the propagation delay plus offset-time, which has the order of several hundred microseconds. Usually burst length will have the order of several hundred kilobytes.

2. In WR-OBS, data burst will be sent out only after a successful resource reservation, thus it is relatively complex. The time delay for resource reservation is on the timescale of milliseconds, so the burst assembly duration will have the order of milliseconds and the burst length will be several megabytes or more.

3. As a result of immature wavelength converter technology, OBS network always suffer from high burst blocking probability. However, in centralized control WR-OBS the control unit manages all resources and in distributed WR-OBS a backward control packet travels on the same lightpath taken by the forward control packet, which provides a chance to release resource locked improperly. Obviously better bandwidth utilization can be achieved in WR-OBS network and burst blocking probability will be reduced.

- Protocols and algorithms in WR-OBS

Most protocols and algorithms developed in JET-OBS, such as fixed assembly period and length (FPAL) for burst assembly, LAUC-VF for resource reservation and preemption algorithm for QoS, can be applied to WR-OBS. However, unique protocols and algorithms have been proposed based on the two-way resource reservation mechanism.

1. Parallel burst assembly algorithm. In this algorithm burst assembly process and resource reservation process are simultaneously implemented partially, end-to-end delay for packets is reduced efficiently. This characteristic is in common with other parallel burst assembly algorithm. The difference between them lies in the two-way resource reservation, that is using the information carried by the backward control packet, burst length estimation will be more accurate and the data burst will have another chance for resource reservation in the case of failure in the first attempt ion.

2. QoS provisioning. Two-way resource reservation mechanism can be used to supply end-to-end QoS provisioning naturally. By this means Diffserv QoS can be provided easily.

5.2.2 Applying Grid application in WR-OBS

In this section how a GRID job is generated and processed is described. WR-OBS needs a upgrading to support GRID applications. For centralized control WR-OBS, the control unit will act as a GRID resource manager, that is, all the GRID resource providers will register their service here and report the resource situation to it. For distributed control WR-OBS, the core routers will act as GUNI to support GRID functionalities.

Like JET-OBS, in WR-OBS a GRID job is created in the edge node and a BCH is sent to the network. However the following work is quite different. Once the data burst containing a job is built up a control packet is generated and sent to the core network. In centralized control WR-OBS network, this request will be sent to the control unit and the control unit will to find some resource to process this job. In distributed WR-OBS, the BCH will be sent to the core network using anycast protocol. The BCH will travel in the core network hop by hop until one core route find out where the corresponding job can be processed. All core routers will share the GRID service indexes among themselves and it will not take too much time to get the destination. By this means, in WR-OBS network a job can find the destination node, an end-to-end lightpath will be reserved for data burst and a backward control packet will be generated and sent to the source edge router to deliver this message. The backward control packet can also be used to reserve bandwidth for the completed job if necessary, which leaves out bandwidth reservation for it.

Clearly in WR-OBS network the source node will know whether the job can be done, where the job is processed, and when the result will be returned. All this information is quite important to GRID applications. To sum up, in WR-OBS jobs are totally under control and reliable service can be provided.

5.2.3 JET-OBS, WR-OBS and WRON for GRID application

In the future, optical network will be used to support all kinds of GRID applications. It is reasonable to construct a mixed optical network, that is, JET-OBS, WR-OBS and WRON will all be used for different GRID applications.

1. In JET-OBS, data burst has a length of several kilobytes and a shorter end-to-end delay compared with WR-OBS. The whole network is relatively easy to be implemented and provides a connectionless service. Thus JET-OBS is suitable for GRID applications with large number of users and small data transmission.

2. In WR-OBS, data burst has a length of several megabytes and a longer end-to-end delay. The whole network provides a connection-oriented service. Thus WR-OBS is suitable for GRID applications with high QoS requirement.

3. In WR-OBS, when the BCH reserve a whole wavelength for a job, WR-OBS transform to WRON. WRON can provide the best service and largest data

transmission. It is suitable for GRID applications such as large file transmission and so on.

5.3 Application aware programmable optical burst switched network

All the current research activities focus on applications that require long-lived wavelength paths and address the specific needs of a small number of well known organizations and users. A typical user is particle physics which, due to its international collaborations and experiments, generates enormous amounts of data (Petabytes per year) and requires very advanced network infrastructures that can support processing and analysis of these data through globally distributed computing resources. However, providing wavelength granularity BW services is not an efficient and scaleable solution for a wider base of user communities with different traffic profiles and connectivity requirements.
Examples of such applications may be: scientific collaboration in smaller scale (e.g. bioinformatics, environmental research), distributed virtual laboratories (e.g. remote instrumentation), e-health, national security and defense, personalized learning environments and digital libraries, evolving broadband user services (i.e. high resolution home video editing, real-time rendering, high-definition interactive TV). These applications need infrastructure that makes large amounts of bandwidth, storage and computation resources potentially available to a large number of users and they may require short lived connection set up. For example remote Mammography introduces high-capacity requirements due to size and quantity of images produced by scans

Optical burst switching (OBS) technology is a suitable candidate for implementing a scaleable network infrastructure to address the needs of emerging collaborative services and distributed applications. Its transport format can be ideally tailored to user's bandwidth requirements and can therefore provide efficient use of network resources. Furthermore, unlike the optical wavelength switched networks the optical bandwidth can be reserved for a short time, i.e. only for the duration of the burst.

As collaborative network services and applications evolve, it is infeasible to build a dedicated network for each application type. Consequently, there should be a dynamic and application-aware network infrastructure which is able to support all application types, each with its own access and resource usage patterns. This infrastructure should offer flexible and intelligent network components able to deploy new applications quickly and efficiently. Application-aware translates into faster and more flexible service provisioning, while optical networking offers high performance transport mechanism. The development of application-aware optical network allows the future network users to construct or choose their own application specific optical network topology and do their own traffic engineering. Therefore such network has ability to dynamically provision high performance data paths to support future and emerging network applications furthermore it will be able to discover network resources and computing resources based on application requirements and the user will be able to choose among discovered resources (i.e. light-path and computing resources).

The aim of this section is to propose an application-aware OBS network infrastructure able to dynamically interconnect computing resources and perform collaborative applications in a user-controlled manner. The OBS network will be able to discover

network resources and computing resources based on application requirements and the user will be able to choose among discovered resources (i.e. light-path and computing resources).

A typical collaborative networking scenario such as Grid networking using the application aware OBS infrastructure can be described as below:

The user/application sends the request for a service through user-network interface (edge router) by using dedicated optical bursts. The request is processed and distributed through the network for the resource discovery (both network and non-network resources) by core OBS routers using optical multicast or broadcast. After resource discovery, an acknowledgement message determines type and identity of computing resources (processing and storage) as well as associated network resources such as allocated light-path and the time duration that each light-path is available. Consequently the user can select among available resources to sends the job (application data) buy using another optical burst (non-active/normal burst) through the appropriate light-paths. Once the job has been completed (data has been processed), the results have to be reported back (if there are any results for the user (sender)). On the way back, based on the type of results as well as their requirements in terms of the network resources, a new path can be reserved using a new OBS signaling.
One of the advantages of this scenario is that both traditional data traffic and distributed application traffic can be supported by a common infrastructure. Core OBS routers perform burst forwarding when normal traffic transits across the network while in addition they support transport of traffic related to collaborative services by performing advance networking functionality such as resource discovery.

5.3.1 Programmable Optical Burst Switched Network

In this section a novel solution towards ubiquitous photonic Grid networking is proposed. This solution utilizes optical burst switching and active router technologies. It aims to provide a physical infrastructure able to fulfill both existing data-intensive and future Grid application requirements and make efficient use of network resources. The solution is based on programmable network architecture, in which the optical network topology is application aware and it can be programmed by Grid users and services.

The architecture is based on the novel concept of using active OBS routers for resource discovery and routing of the Grid jobs to the appropriate resources across the network. The network comprises active and non-active OBS routers. A non-active OBS router is a conventional OBS router and performs the burst forwarding functionality. The router is informed in advance about the data burst characteristics (duration, type, class of service, etc.) by the Burst Control Packet (BCP). Upon the data burst arrival the router, forwards the data to the appropriate output port. An active OBS router, in addition to the burst forwarding, can intercept with data carried by some optical bursts (active bursts) and perform dedicated Grid networking functionality. The proposed active OBS networking scheme has the potential to offer global reach of computing and storage resources to a large number of anonymous users with different traffic profiles. In such a network, OBS

offers efficient network resource utilisation while the active networking offers intelligent Grid functionality. One of the main advantages of the proposed scenario is that both traditional data traffic and Grid traffic can be supported by a common infrastructure. All OBS routers perform burst forwarding when normal traffic transits across the network while in addition some OBS routers (active routers) support transport of Grid traffic over the network.

- **Description of Transport format**

There are several major OBS variants differing in bandwidth reservation schemes [58]. Among all of them, the just-enough-time (JET) is the most appropriate protocol for the proposed Grid network architecture [59]. The JET protocol employs a delayed reservation scheme which operates as follows: an output wavelength is reserved for a burst just before the arrival of the first bit of the burst; if, upon arrival of the BCP, it is determined that no wavelength can be reserved at the appropriate time, then the BCP is rejected and the corresponding data burst dropped. The proposed network concept utilizes the JET scheme and extends it to support both active and non-active network operations. Non-Grid traffic is injected into the network in the form of a normal, non-active burst and active routers do not intercept the traffic. In this mode, once data is ready to be transmitted, a BCP is sent from the edge router into the optical network and the required resources are reserved for the duration of the burst. For efficient transmission of Grid traffic, we have developed a two-stage OBS networking scheme including an active stage and a non-active stage. Grid traffic is transmitted in two stages as follows: job specification is transmitted in the form of an active burst prior to the actual job (user data) which is transmitted in the form of a non-active burst. The user with a Grid job sends a request to the edge router informing about the job specification and resource requirements. The edge router then constructs and transmits the active optical burst for which the BCP only informs intermediate active routers that the incoming optical burst is active. After an offset time, the active burst is transmitted carrying information about the Grid job characteristics (i.e. processing and storage requirements). With this mechanism active routers prior to arrival of the job specification have been informed about the arrival of an active burst. Upon arrival of a job specification burst, an active router performs a resource discovery algorithm to find out whether there are enough Grid resources available within its Grid resource domain to perform the job. In addition, each active router multicasts both the BCP and data burst of an active burst towards the other active routers in the network. The user is informed about the result of resource discovery by each active router through acknowledgment or not-acknowledgment messages (optical burst). In case of resource availability the user transmits the actual job in the form of a non-active burst through the edge router.

In order to accommodate the requirements of this active Grid network scenario the JET scheme is modified. The job submission is divided into two steps:

1. The BCP of an active burst is sent to all active routers through intermediate nodes (active or non-active). After an offset time the active data burst is sent to the network. The result of the resource discovery algorithm in each active router produces an acknowledgment (Ack) or a notacknowledgment message (Nack). These messages are transmitted back to the user through an optical burst (non-

active burst). In case of acknowledgement, the active OBS router also informs the corresponding resource manager. At that point the resource manager reserves the local resources for a predefined and limited duration of time.

2. Receiving all ACK and NACK messages, the user can choose one or multiple appropriate destinations among all available resources across the network. The actual job is now sent within the reservation period to the appropriate destination in normal (non-active) optical burst format.

In summary, the proposed programmable OBS concept is a two mode networking scheme:

- It is an active network when the Grid job specification is routed through the network to discover the suitable Grid resources
- It is non-active when Grid jobs or normal data traffic are routed across the network

This combination provides bandwidth efficiency especially when a large data set needs to be transferred because the actual job is submitted to the network only when both the Grid resources and the network resources have been reserved. In addition it provides a secure and policy based Grid environment where the users have the ability to choose among the available resources in different Grid domains across the network. Furthermore, active routers in each domain can respond positively only to the requests that match with the applied policy in their corresponding domain.

- Grid enabled active OBS routers

Central to the programmable OBS network architecture is the possibility of using network processors (NPs) in active OBS routers, capable of analyzing data traveling through the network at wire speed. In the proposed network architecture active OBS routers utilize high-performance network processors (NPs) for routing the active jobs. The NPs are capable of executing specific processing functions on data contained within an active burst at line rates (e.g. Grid resource discovery algorithm). Active OBS routers are key enablers for the support of user-controlled networking functionalities: 1) quality of service (QoS) provisioning 2) reliable multicasting and 3) constrained base routing.
It has been shown in [60] that services and applications are concerned about QoS based on network, bandwidth and delay. In the proposed network architecture, a combination of the control protocol and active routers' processing power can be used to deploy an advanced burst-scheduling algorithm. This algorithm is able to reduce delay whilst maintaining high bandwidth efficiency and low burst loss rate.
In the active Grid network environment, multicasting performs an important role, where interactive and distributed applications are deployed. A reliable multicast protocol framework is deployed, in order to minimize the traffic load across the network and also reduce the recovery latency [61].

5.4 Optical Burst Ethernet Switched (OBES) Transport Protocol for Grid

A novel optical transport solution based on OBS and towards photonic Grid networks is being explained in this section. OBS network is able to promote traffic engineering to facilitate efficient and reliable network operations optimizing network resource utilization and traffic performance. In contrast, OBS face various implementation difficulties, such as router synchronization, header detection and extraction and thus sophisticated and complex bursty receivers are required at termination points. In order to resolve these OBS network inefficiencies, we propose a sub-wavelength transport technology, the Optical Burst Ethernet Switching (OBES), which can serve users with diverse traffic profiles (Grid users) over all-optical network and also provide the flexibility and robustness offered by Ethernet and its associated Data Transport Protocols (next-generation TCP and UDP).

In traditional OBS networks, a data burst consisting of multiple IP packets is switched through the network all-optically. Prior to data burst transmission a Burst Control Header (BCH) is created and sent towards the destination by an OBS ingress node. The BCH is typically sent out of band over a separate signaling wavelength and processed at intermediate OBS routers. It informs each node of the impending data burst and setup an optical path for its corresponding data burst. Data bursts remain in the optical plane end-to-end, and are typically not buffered as they transit the network core. As mentioned above, becomes clear that a bursty receiver is required on each intermediate node to detect and process the BCH.

Here, we propose a new burst switching transport format to tackle this problem and also become a transport solution towards ubiquitous photonic Grid networking. In the proposed optical OBES network, the BCH is transmitted synchronously in front of the data burst in Ethernet format and over a separate and dedicated wavelength channel while the data burst is transmitted asynchronously. More specifically, this synchronous BCH follows Ethernet transport format and is able to overcome the OBS core router synchronization and detection mechanism complexity (data and clock recovery) by exploiting the flexibility and robustness offered by Ethernet.

Ethernet-based BCH will carry resource discovery and invocation requirements and can be adapted by both Programmable OBS and wavelength routed OBS (WR-OBS) for Grid. In case of programmable OBS, both active and non-active BCP (and Ethernet-based BCH) will consist of a Grid identifier (Grid-ID) to trigger the appropriate Grid processes at Node level. Active BCP or active Ethernet-based BCH will carry a flag to identify that active Burst follows and also encapsulates the Grid Class of Service, which is mandatory for Grid Differentiated Service (GridDiffServ) provisioning. In case of WR-OBS for Grid, the BCH could have all resource requirements encapsulated.

The proposed transport format integrates Ethernet synchronization and OBS traffic-engineering advantages for a new era, the Optical Burst Ethernet Switched (OBES) Network infrastructure, which will offer robustness and flexibility in order to support current, evolving and emerging Grid network applications. Currently, the optical network

and Grid community utilizes standard transmission formats (mainly Ethernet and Gigabit Ethernet); therefore the proposed OBES transport format coupled with the generalized multiprotocol label switching (GMPLS) [62], can be used towards an integrated OBES-GMPLS robust and QoS-aware all-optical networks. The OBES-GMPLS can be utilized to provide Grid services such as resource monitoring, discovery and reservation over the a unified control plane.

6. Security issues in Grid-OBS networks

7. References

[1] D. Simeonidou et al., "Optical Network Infrastructure for Grid", Grid Forum Draft , GFD-I.036, Oct 2004

[2] C. Qiao, M. Yoo, "Optical Burst Switching - A new Paradigm for an Optical Internet", Journal of High Speed Networks, Spec. Iss. On Optical Networking, vol. 8, no. 1, Jan. 2000, pp. 36-44

[3] Mintera Optical Networks, "Myths and realities about 40G optical technology", white paper, www.mintera.com/WhitepaperWEBnew.pdf, 2002.`

[4] C. Qiao, M. Yoo, "Optical Burst Switching - A new Paradigm for an Optical Internet", Journal of High Speed Networks, Spec. Iss. On Optical Networking, vol. 8, no. 1, Jan. 2000, pp. 36-44

[5] M. Maimour and C. Pham, "Dynamic Replier Active Reliable Multicast (DyRAM)", Proceedings of the 7th IEEE Symposium on Computers and Communications (ISCC 2002)

[6]J. Gaither, "300-Pin MSA Bit-Error Rate Tester for the ML10G Board and RocketPHY Transceiver", XILINX, Application note: XAPP677 Virtex-II Pro family, Jan. 2004.

[7]C. Guillemot et. al. "Transparent optical packet switching: The European ACTS KEOPS project approach", IEEE Journal of Lightwave Technology, vol.16, no.12, pp. 2117-2134, Dec. 1998.

[8]A. Stavdas, S. Sygletos, M. O'Mahony, H.L. Lee, C. Matrakidis, A. Dupas, "IST-DAVID: concept presentation and physical layer modelling of the metropolitan area network", IEEE Journal of Lightwave Technology, vol.21, no.2, pp. 372-383, Feb. 2003.

[9]D.K. Hunter, M.H.M Nizam, M.C. Chia, K.M. Guild, A. Tzanakaki, M.J. O'Mahony, J.D. Bainbridge, M.S.C. Stephens, R.V. Penty, I.H. White, "WASPNET: A wavelength switched packet network", IEEE Communications Magazine, vol.37, pp.120-129, Mar. 1999.

[10]F. Xue, Z. Pan, H. Yang, J. Yang, J. Cao, K. Okamoto, S. Kamei, V. Akella, S.J.B. Yoo, "Design and Experimental Demonstration of a Variable-Length Optical Packet

Routing System With Unified Contention Resolution", IEEE Journal of Lightwave Technology, vol.22, no.11, pp.2570-2581, Nov. 2004.

[11]D. Klonidis, R. Nejabati, C.(T.) Politi, M.J. O'Mahony, D. Simeonidou, "Demonstration of a fully functional and controlled optical packet switch at 40Gb/s", in proc. 30th European Conf. on Optical Comm., Stockholm, Sweden, PD Th4.4.5, Sep. 2004.

[12] V. P. Kumar, T. V. Lakshman, and D. Stiliadis, "Beyond best effort: Router architectures for the Differentiated Services of Tomorrow's Internet", IEEE Commun. Mag., vol. 36, pp. 152-164, May 1998

[13] J. Turner, "Terabit burst switching", Journal of High Speed Networks, Vol. 1, No. 8, pp. 3-16 (1999).

[14] C. Guillemot et al., "Transparent Optical Packet Switching: the European ACTS KEOPS project approach", IEEE/OSA Journal on Lightwave Technology, Vol. 16, No. 12, pp. 2117-2134, 1998.

[15] F. Callegati et al., "Wavelength and Time Domains Exploitation for QoS Management in Optical Packet Switches", Computer Networks, Vol. 44, No 4, pp. 569-582, 2004.

[16] J. Xu et al., "Efficient Channel Scheduling Algorithms in Optical Burst Switching", IEEE Infocom, 2003.

[17] G. Muretto et al. "Effective Implementation of Void Filling in OBS Networks with Service Differentiation", Proc. of WOBS 2004, San Jose, CA, USA

[18] R. Al-Ali, G. von Laszewski, K. Amin, M. Hategan, O. Rana, D. Walker, N. Zaluzec, "QoS Support for High-Performance Scientific Grid Applications," in IEEE International Symposium on Cluster Computing and the Grid CCGrid, Chicago, 19-22 Apr. 2004, pp.134 – 143.

[19] D. Simeonidou et al., "Optical Network Infrastructure for Grid," GFD-I.036 draft, Informational GHPN-RG *http://forge.Gridforum.org/projects/ghpn-rg*

[20] M. Yoo, C. Qiao, and S. Dixit, "Optical Burst Switching for Service Differentiation in the Next Generation Optical Internet," *IEEE Communications Magazine*, pp. 98–104, Feb. 2001

[21] M. Yoo, C. Qiao and S. Dixit, " QoS Performance of Optical Burst Switching in IP-Over-WDM Netwoks," *IEEE Journal on Selected Areas in Communications,* pp. 2062-2071, Oct. 2000.

[22] A. Kaheel and H. Alnuweiri, "A Strict Priority Scheme for Quality-of-Service Provisioning in Optical Burst Switching Networks," in *Proceedings of IEEE Symposium on Computers and Communications-ISCC'03*, Turkey, Jun. 2003, pp. 16–21.

[23] V. M. Vokkarane and J. Jue, "Prioritized Routing and Burst Segmentation for QoS in Optical Burst Switched Networks," in *Proceedings of Optical Fiber Communication Conference- OFC'02*, Anaheim, March 2002, pp. 221–222.

[24] V. M. Vokkarane, Jason P. Jue, "Prioritized Burst Segmentation and Composite Burst-Assembly Techniques for QoS Support in Optical Burst-Switched Networks," *IEEE Journal on Selected Areas in Communications,* pp. 1198-1209, Sep. 2003.

[25] M. Casoni, M.L. Merani, "Resource Management in Optical Burst Switched Networks: Performance Evaluation of a European Network," in *Proceedings of 1st International Workshop on Optical Burst Switching*, 16 Oct. 2003, Dallas.

[26] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "Wavelength and Time Domains Exploitation for QoS Management in Optical Packet Switches", Computer Networks, Vol. 44, No 4 , pp. 569-582, 2004.

[27] F. Callegati, W. Cerroni, C. Raffaelli and M. Savi, "QoS differentiation in optical packet-switched networks", to appear on Computer Communications, 2006.

[28]M. Casoni, M.L. Merani, A. Giorgetti, L. Valcarenghi, P. Castoldi, "Guaranteeing Seamless end-to-end QoS in OBS Networks," in *Proceedings of OpNeTec,* Oct. 2004, Pisa.

[29]Y. Chen, M. Hamdi, and D. Tsang, "Proportional QoS over OBS Networks," in *Proceedings of IEEE Global Telecommunications Conference- Globecom' 01*, San Antonio, Nov. 2001, pp. 1510–1514.

[30] Y. Chen, M. Hamdi, D. H. K. Tsang, and C. Qiao, "Proportional differentiation – A scalable QoS approach," *IEEE Commun. Mag.*, vol. 41, pp. 52–58, Jun. 2003.

[31] Q. Zang, V. M. Vokkarane, J. P. Jue, B. Chen, "Absolute QoS Differentiation in Optical Burst-Switched Networks," *IEEE Journal on Selected Areas in Communications,* pp. 1781-1795, Nov. 2004.

[32] A. Kaheel, H. Alnuweiri, "Quantitative QoS Guarantees in Labeled Optical Burst Switching Networks", in in *Proceedings of IEEE Global Telecommunications Conference, Globecom'01*, Dallas, Nov. 2004.

[33]Yufeng Xin *et al.*, "Fault Management with Fast Restoration for Optical Burst Switched Networks", IEEEXplore

[34] L. Valcarenghi and A. Fumagalli, "Implementing Stochastic Preplanned Restoration

with Proportional Weighted Path Choice in IP/GMPLS/WDM Networks", in Photonic Network Communications, Special Issue on “Routing, Protection, and Restoration Strategies and Algorithms for WDM Optical Networks", Kluwer Academic Publishers, vol. 4, no. 3/4, July/December 2002

[35] Yufeng Xin *et al.*, “A Novel Fast Restoration Mechanism for Optical Burst Switched Networks”, IEEEXplore

[36]S. Darisala, A. Fumagalli, P. Kothandaraman, M. Tacca, L. Valcarenghi, M. Ali, and D. Elie-Dit-Cosaque "On the Convergence of the Link-State Advertisement Protocol in Survivable WDM Mesh Networks", in Proceedings of ONDM 2003, Budapest,

[37] Jing Zhang et al., “Pre-planned Global Rerouting for Fault Management in Labeles Optical Burst Switched WDM Networks”, ICC 2004

[38] A. Anjomshoaa, F. Brisard, M. Drescher, D. Fellows, A. Ly, S. McGough, D. Pulsipher, A. Savva, “Job Submission Description Language (JSDL) Specification v1.0”, Grid Forum Draft , GFD-R-P.056, 2005-12-01

[39] C. Qiao, “Labelled Optical Burst Switching for IP-over-WDM Integration”, IEEE Comm. Mag., vol. 38, no. 9, pp. 104-114, September 2000.

[40] S. Ovadia, C. Maciocco, M. Paniccia, “GMPLS-Based Photonic Burst Switching (PBS) Architecture for Optical Networks”, in Proc. of the First Intnl Workshop on Optical Burst Switching (WOBS2003) co-located with Opticomm'03, Oct. 16 2003, Dallas, Texas, USA.

[41] K.G. Vlachos, I.T. Monroy, A.M.J. Koonen, C. Peucheret, P. Jeppesen, “STOLAS: Switching Technologies for Optically Labelled Signals”, IEEE Comm. Mag., vol: 41, no. 11, pp. 43-49, November 2003

[42] H. Wang, R. Karri, M. Veeraraghavan, T. Li, “A Hardware-Accelerated Implementation of the RSVP-TE Signaling Protocol”, Proc. of IEEE ICC2004, June 20-24 2004, Paris, France

[43] K. Long, Z. Yi, Y. Xin, X. Yang, H. Liu, “Generalized MPLS (GMPLS) architecture's extensions for Optical Burst Switch network”, Internet Draft, draft-long-gmpls-obs-00.txt, work in progress, November 2005

[44] I. Baldine, G.N. Rouskas, H.G. Perros, D. Stevenson, “JumpStart: A Just-in-Time Signaling Architecture for WDM Burst-Switched Networks”, IEEE Comm. Mag., vol: 40, no. 2, pp. 82-89, February 2002

[45] I. Baldine, P. Mehrotra, G. Rouskas, A. Bragg, D. Stevenson, “An Intra- and Inter-Domain Routing Architecture for Optical Burst Switched (OBS) Networks”, in Proc. of

the Fifth Workshop on Optical Burst/Packet Switching, pp. 150-159, October 3, 2005, Boston, MA

[46] J. Teng, G.N. Rouskas, "A Comparison of the JIT, JET, and Horizon Wavelength Reservation Schemes on A Single OBS Node", in Proc. of the First Intnl Workshop on Optical Burst Switching, October 16, 2003, Dallas, Texas (co-located with Opticomm 2003).

[47] J. Strand, A.L. Chiu, R. Tkach, "Issues For Routing In The Optical Layer", IEEE Comm. Mag., vol: 39, no. 2, pp. 81-87, February 2001

[48] J. Strand, A.L. Chiu , "Impairments and Other Constraints on Optical Layer Routing", IETF RFC4054

[49] A. Fredette, J. Lang,"Link Management Protocol (LMP) for Dense Wavelength Division Multiplexing (DWDM) Optical Line Systems", IETF RFC 4209.

[50] L. Zong and N. G. Bourbakis, "Digital Video and Digital TV: A Comparison and the Future Directions", Proc. of the International Conference on Information Intelligence and Systems, Mar 1999.

[51] T. Ikedo, "Creating Realistic Scenes in Future Multimedia Systems", IEEE MultiMedia, 9(4):56-72, Oct 2002.

[52] P. Thysebaert, B. Volckaert, et al. "Towards Consumer-Oriented Photonic Grids", Workshop on Optical Networking for Grid Services (ECOC 2004), Sep 2004.

[53]] E. Van Breusegem, M. De Leenheer, et al., "An OBS Architecture for Pervasive Grid Computing", Workshop on Optical Burst Switching (Broadnets 2004), Oct 2004.

[54] M. De Leenheer, E. Van Breusegem, et al., "An OBS-based Grid Architecture", Workshop on High Performance Global Grid Networks (Globecom 2004), Nov 2004.

[55] C. Partridge, T. Mendez, W. Milliken, "Host Anycasting Service", IETF RFC 1546, Nov 1993.

[56] S. Bhattacharjee, M.H. Ammar, et al., "Application-Layer Anycasting", Proc. of IEEE Infocom'97, Apr 1997.

[57] E. Basturk, R. Engel, et al., "Using Network Layer Anycast for Load Distribution in the Internet", IBM Research Report (RC20938), Jul 1997.
[58] Rosberg, Z., et al.," Performance analyses of optical burst-switching networks", Selected Areas in Communications, IEEE Journal on, Volume: 21, Issue: 7, Sept.2003 Pages:1187 - 1197

[59] M. Yoo and C. Qiao, “Just-enough-time (JET): A high speed protocol for bursty traffic in optical networks”, In IEEE/LEOS Technol. Global Information Infrastructure, pages 26{27, August 1997.

[60] Foster, I., Kesselman, C. and Tuecke, S., “ The Anatomy of the Grid: Enabling Scalable Virtual Organizations”, International Journal of High Performance Computing Applications,15 (3). 200-222. 2001.

[61] M. Maimour and C. Pham, “Dynamic Replier Active Reliable Multicast (DyRAM)”, Proceedings of the 7th IEEE Symposium on Computers and Communications (ISCC 2002)

[42] Jingxuan Liu, Nirwan Ansari and Teunis J. Ott, “FRR for latency reduction and QoS provisioning in OBS networks”, IEEE Journal on Selected Areas in Communications, vol. 21, no. 7, Sep 2003, pp. 1210 – 1219.

[43] K. Christodoulopoulos, K. Vlachos et al.: “EBRP: A hybrid signaling protocol for efficient burst-level reservations and QoS differentiation in OBS networks, sumbited to Journal of Optical networking

[44] E. Varvarigos et. al. “Fair QoS Resource Management in Grids", invited as a book chapter in "Engineering the Grid" to be published by Nova Science Publishers, 2004.

[62] S. Sheeshia, C. Qiao, Jeffrey U.J. Liu, “Supporting ethernet in optical-burst-switched networks”, SPIE The Journal of Optical Networking, 1(8/9):299-312, 2002