034115

PHOSPHORUS

Lambda User Controlled Infrastructure for European Research

Integrated Project

Strategic objective:
Research Networking Testbeds

Deliverable reference number: D.5.3

Grid Job Routing Algorithms

Due date of deliverable: 2007-06-31
Actual submission date: 2007-06-31
Document code: Phosphorus-WP5-D5.3

Start date of project:
October 1, 2006

Duration:
30 Months

Organisation name of lead contractor for this deliverable: **Athens Information Technology (AIT)**

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission	
RE	Restricted to a group specified by the consortium (including the Commission	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Grid Job Routing Algorithms

Abstract

This deliverable proposes enhanced routing algorithms capable of calculating optimal network paths considering a number of constraints such as the Grid user requirements and the physical layer characteristics. These algorithms aim to offer improved overall network performance and efficiency, optimize Quality of Service levels and increase user satisfaction in service level agreements (SLAs). In addition they target at providing some orchestration between the submitted jobs and the optical network components and resources capable of processing the jobs through the use of anycast-based routing in multi-domain Grid networks.

The document presents a number of mathematical models for expressing physical layer performance issues as well as addressing the Grid User quality of service requirements. Also methods for integrating the proposed models into the routing algorithms are introduced and benefits provided by such an approach are demonstrated through a number of simulation studies.

Finally the applicability of the algorithms in the GMPLS Control Plane is discussed and an architecture to support anycast-based routing in multi-domain Grid networks allowing control plane scalability is presented.

List of Contributors

George Markidis	AIT	Emmanouel Varvarigos	RACTI
Anna Tzanakaki	AIT	Tim Stevens	IBBT
Stelios Sygletos	AIT	Joachim Vermeir	IBBT
Ioannis Tomkos	AIT	Chris Develder	IBBT
Panagiotis Kokkinos	RACTI	Marc De Leenheer	IBBT
Konstantinos Christodoulopoulos	RACTI	Bart Dhoedt	IBBT

Table of Contents

Table of Figures

0 Executive Summary

This deliverable, entitled “Grid job routing algorithms” proposes a routing approach that takes into consideration both the physical layer characteristics of the network infrastructure as well as Grid-specific characteristics and requirements in order to offer improved QoS, satisfaction of service level agreements (SLAs) and overall optimized routing performance. In addition it aims to provide some orchestration between the submitted jobs and the optical network components and resources capable of processing the jobs through the use of anycast-based routing in multi-domain Grid networks.

Under this framework a physical layer analysis is presented addressing the most crucial and fundamental optical constraints that could be included in the routing procedure to allow efficient utilization of the network resources. As part of the work presented in this deliverable accurate analytical expressions are derived and discussed offering precise evaluation of the degradations introduced due to the presence of the physical impairments and their interactions. Also strategies and methods for the integration of the calculated and monitored physical layer impairments into the GMPLS control plane are considered focusing in different directions based on existing literature and demonstrated solutions.

In addition, this document analyzes the grid user quality of service (QoS) requirements that should also be considered by the Grid job routing algorithms when optical connections have to be established. Equations that can be used to model requirements like delay, bandwidth demand, delay Jitter and packet loss are identified and a method allowing their integration with physical layer constraints is proposed.

Furthermore this deliverable presents an architecture to support anycast-based routing in multi-domain Grid networks allowing control plane scalability, support of any subset of parameters that are available to the routing protocol and system-wide optimization of the Grid network. Algorithms to optimally dimension the proxy infrastructure by concurrent placement of proxy servers and determination of their capacities are described and simulations are performed to demonstrate the control plane scalability of the proposed approach

The algorithms and approaches proposed and discussed in this deliverable are evaluated and tested through simulation studies focusing on the PHOSPHORUS network topology presented in D6.1 [PHOSP-TestBed] and some required in the context of a realistic network infrastructure. These routing approaches will be incorporated in the optical Grid Simulator which will be developed in WP5 to evaluate the preliminary design of the control plane and will be presented in detail in deliverable [Phos-D5.6].

In subsequent deliverables ([Phos-D5.2] and [Phos-D5.4]) Grid job executions models associated with scheduling and workflow will be developed and evaluated.

The structure of this document can be described as follows:

In section 1 the objectives of the routing algorithms as well as the scope of the document in the implementation and the simulation environment of the PHOSPHORUS framework are stated.

In section 2 the terminology relevant to the Grid job routing algorithms is stated.

In section 3 an introduction to general routing issues is presented focusing on the challenges of routing in optical Grid networks.

In section 4 the infrastructure and the user related requirements are identified and several issues that should be taken into consideration by the routing algorithms are isolated and analyzed.

In section 5 the PHOSPHORUS network scenario on which the simulations and the simulation results presented in this deliverable are focused, is introduced. Also a short explanation of the PHOSPHORUS architecture is given with respect to routing issues.

In section 6 the developed routing algorithms are presented and the simulation results are explained and analyzed.

In section 7 some closing notes of this deliverable are provided.

Finally in Appendix A the Linear Programming (LP) formulation of the Routing and Wavelength Assignment (RWA) problem is analyzed in detail and in Appendix B the LP formulation is extended to deal with the case in which a demand requires more than one wavelength over the same path.

1 Objectives and Scope

This document investigates the use of constraint based routing algorithms in optical network infrastructures that support Grid computing applications. The general objective is to offer improved overall network performance and efficiency, optimize Quality of Service levels, and increase user satisfaction in service level agreements (SLAs), security and resilience. The routing approaches described in this deliverable offer the opportunity to take into consideration the physical layer characteristics of the network infrastructure and costs related to Grid-specific characteristics and requirements as part of the routing algorithm. Additionally, routing strategies that offer improved performance in multi-domain scenarios are proposed and analyzed. However, the translation of the routing strategies into practical routing protocols, is not considered as part of this deliverable as this will be the main focus of Deliverable 5.5 “Recommendations for Control Plane Design”. Also it should be noted that a discussion on the extensions specifically made for the simulation environment (which is being developed in this work package), will be reported in Deliverable 5.6 “Grid Simulation Environment”.

In summary, the following objectives will be treated in this document:

- Provide a detailed analysis of features and shortcomings of optical Grid networks, focusing on optical technology, available routing algorithms, and multi-domain issues.
- Describe requirements emerging in the physical domain (e.g. linear and non-linear impairments) and in the application domain, and analyze their effect on routing algorithms.
- Analyze the architectures, technology choices and network scenarios specific to the Phosphorus project testbed.
- Propose routing algorithms which can optimize metrics emerging from both the physical layer and the Grid layer.
- Propose an approach to create an efficient and optimized routing plane in a multi-domain, optical Grid environment.

2 Terminology

In this section some definitions particularly relevant in the context of Grid job routing algorithms are provided in case with reference to the originator document

Keyword	Source	Definition
Grid	**[OGF-GFD81]**	A system that is concerned with the integration, virtualization, and management of services and resources in a distributed, heterogeneous environment that supports collections of users and resources (virtual organizations) across traditional administrative and organizational domains (real organizations).
Optical Grid	**[OGF-GFD36]**	It is a new topological solution where the network topology is required to migrate from the traditional edge-core telecom model to a distributed model where the user is in the very heart of the network. In this type of network the user would have the ability to establish true peer-to-peer networking (i.e. control routing in an end-to-end way and the set up and teardown of light-paths between routing domains). To facilitate this level of user control, users or applications will be offered management/control or even ownership of the network resources from processing and storage capacity to bandwidth allocation (i.e. wavelength and sub-wavelength). These resources could be leased and exchanged between Grid users. This solution will have a direct impact on the design of optical network elements (optical cross-connects, add-drop multiplexers etc) and will impose new demands to the interface between the Grid user and network (GUNI). The network infrastructure, including network elements and user interface, must enable and support OGSA. Through OGSA the Grid user

		can only have a unified network view of its owned resources on top of different autonomous systems. The resources can either be solely owned or shared with other users.
Job	**[OGF-GFD81]**	A user defined task that is scheduled to be carried out by an execution subsystem. In OGSA-EMS, a job is modeled as a manageable resource, has an endpoint reference, and is managed by a job manager.
Domain	**[RFC4726]** **[OGF-GFD81]**	*Network*: A domain is considered to be any collection of network elements within a common sphere of address management or path computational responsibility. *Grid:* A group of computers and resources under a common administration
Protocol	**[OGF-GFD11]**	A complete and unambiguous set of rules (formats, their semantics & syntax, parameters, timing, error handling, ...) defining the communication between two or more entities
Forwarding Adjacency - LSP	**[IETF – RFC4206]**	An LSP created by an LSR using GMPLS TE procedures and announce as a TE link into the same instance of the GMPLS Control Plane and therefore an FA is only applicable when an LSP is both created and used as a TE link by exactly the same instance of the GMPLS control plane. Also an FA is a TE link between two GMPLS nodes whose path transits zero or more (G)MPLS nodes in the same instance of the GMPLS control plane.
Broker	**[OGF-GFD11]**	A process which performs resource quoting (producer) or resource discovery (consumer) and selection based on various strategies, assigns application task(s) to those resources, and distributes data or co-locates data and computations. *Cost Models* may be used for negotiations before selecting/requesting resources.
Network Resource Provisioning System (NRPS)	**[PHOSPHORUS-D1.1]**	The module that has the main task of specifying, reserving, allocating and deploying the set of network resources (links, cross-connections, etc.) required to accomplish the task specified by a user.

Co-allocation	**[OGF-GFD11]**	Ensures that a given set of resources is available for use simultaneously.
User	**[OGF-GFD11]**	A person authorized to submit jobs to *High Performance Computing* resources.
Service Level Agreements (SLA)	**[OGF-GFD44]**	A contract between a provider and a user that specifies the level of service that is expected during the term of the contract. SLAs are used by vendors and customers, as well as internally by IT shops and their end users. They might specify availability requirements, response times for routine and ad hoc queries, and response time for problem resolution.
Advance Reservation	**[OGF-GFD11]**	Is the process of negotiating the (possibly limited or restricted) delegation of particular resource capabilities over a defined time interval from the resource owner to the requester.
Path Computation Element	**[Farrel06]**	An entity (component, application or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.
Routing Controller	**[Alangar03]**	Provide for the exchange of routing information between and within a RA. The routing information exchanged between RCs is subject to policy constraints imposed at reference points (E-NNI and I-NNI).
Resource Manager	**[OGF-GFD81]**	A manager that implements one or more resource management functions that may be applied to resources.
Grid Resource	**[OGF-GFD81]**	In OGSA, a resource is an entity that is useful in a Grid environment. The term usually encompasses entities that are pooled (e.g. hosts, software licenses, IP addresses) or that provide a given capacity (e.g. disks, networks, memory, databases). However, entities such as processes, print jobs, database query results and virtual organizations may also be represented and handled as resources.

3 Introduction to Routing Protocols and Algorithms

Routing is a process of path determination and data forwarding for traffic going through a network. In simple networks, routing tables can be manually configured or identified from the configuration of interfaces on the router. In more complex networks where a number of routers are arranged in a mesh topology, with a large number of links between them, each having different capabilities, manual configuration becomes difficult.

However, more importantly there is a need to **react dynamically to changes** in the network especially in Grid environments e.g., when a link or router fails, we need to update all of the routing tables across the whole network to take account of changes. Similar changes are desirable when failures are repaired or when new links and nodes are added. These dynamic processes complement Grid requirements for adjustable communication service parameters.

For these purposes we rely on **routing protocols** to collate and distribute information about network connectivity, reachability, adjacency and optimization of paths by examining a range of variables related to network conditions and configurations. The value of these parameters is provided to sophisticated route calculation algorithms that can be run against the view of the network to determine the best path along which to forward traffic. Routers use Layer 3 addresses, e.g. IP, for identification of source and destination of packets and have to determine out of which interface should a packet be sent, and on to which next hop (when interfaces lead to multi-access links), utilizing their routing tables which, comprise some form of look-up algorithms that take an IP address and derive an interface identifier and a next hop address.

The routing protocols that are used for routing table creation and allow routers to communicate with routing protocol messages can be categorized into either **distributed** or **centralized**. In distributed protocols each router in the network makes an independent routing decision based on available information whereas in centralized protocols routing decisions are made by a central node in the network and then are distributed to other nodes.

Moreover there are two types of routing protocols: **distance vector** (including path vector) and **link state** routing protocols. The distance vector protocol works by letting each node inform its neighbours about its best

idea of distance to every other node in the network. Once a node receives the distance vectors from its neighbours, it compares these with its own distance vector, each destination and, if necessary, computes/re-computes its best path to each destination and the next hop for that destination. The distance vector protocol has the advantage of simplicity, and with amendments, it also supports route aggregation. However, a distance vector protocol suffers from:

- slow distance vector convergence
- formation of routing loops during convergence
- the problem of counting to infinity.

An example of a distance vector protocol is Routing Information Protocol (RIP), where to address the counting-to-infinity problem, a destination is declared as unreachable when the path cost to the destination reaches 16.

A link state protocol requires that each router stores the entire network topology and computes the shortest path by itself. A link state database is maintained at each router and link state information is exchanged by the means of link state update packets. In contrast to a distance vector protocol, a link state protocol provides:

- faster network convergence
- the ability to support multiple routing metrics.

A link state protocol is more complex than the distance vector protocol and has a higher computational overhead. An example of a link state protocol is Open Shortest Path First (OSPF).

3.1 Classification of Optical networks

The superior properties of optical fibers (i.e. very high bandwidth, low loss, low cost, light weight and compactness, strength and flexibility, immunity to noise and electromagnetic interference and corrosion resistance) against copper cables forced the deployment of **optical networks**. Optical networks rely on wavelength division multiplexing (WDM) to efficiently exploit the massive available bandwidth and offer high capacity and long-reach transmission capabilities.

The first generation of optical networks consists of point-to-point WDM links (opaque networks). This means that at each switching point, the optical signal is converted to electrical form, and the processing and forwarding is done in the electrical domain offering at the same time regeneration of the optical signal. This regeneration can vary in terms of the signal quality improvement it offers:

- 1R Regeneration (reamplification): It is the simplest case of regeneration. Optical amplifiers belong to this category and in this case 1R regeneration is truly transparent. However noise is added to the amplified signal and any signal distortion due to e.g. nonlinearities and dispersion are not compensated for.

- 2R Regeneration (reamplification with reshaping): The signal is amplified and also reshaped but not retimed. Through this mechanism additional phase jitter maybe introduced that will eventually limit the number of stages that can be cascaded.

- 3R Regeneration (reamplification with reshaping and retiming): This includes signal amplification, reshaping and retiming that completely reset the effects of any impairment that the signal has experienced.

By using optoelectronic regeneration we can assure that the signal can reach large distances, since when transformation to the electrical domain also causes the signal to be cleaned and compensated for any noise, dispersion impairments and fiber nonlinearities. On the other hand the use of regeneration poses some limitations. Most of the regenerators used nowadays are optoelectronic as their all-optical counterparts are not feasible because of immature technology which imposes the use of bit rate and modulation format specific devices. Also regenerators are expensive devices and operate on a wavelength per wavelength basis which means that we require one regenerator for each wavelength on a link thus increasing thus the overall cost.

Second generation networks obviate the need for conversion to the electronic domain by providing switching and routing services at the optical layer. These networks are known as all-optical networks (transparent networks) and offer a reduction of unnecessary and expensive optoelectronic conversions, providing thus an ability for high data rate, flexible switching, and support of multiple types of clients (different bit rates, modulation formats, protocols, etc.) In all-optical networks the signals are transported end-to-end optically, without being converted to the electrical domain along their path. This reduces complexity and overheads and offers reduction of unnecessary and expensive optoelectronic conversions. However, due to the analogue nature of the optical networks as the optical signals propagate through the fibers, they experience several impairments degrading their performance. This has a direct impact on the dimensions that an all-optical network can support. Therefore when performing routing in all optical networks is of significant importance to capture and take into account as much as possible all the impairments that affect and deteriorate the quality of the signal in order to improve the overall network performance.

However, some optoelectronic conversion capability, at least to some limited extent and degree, may still be desirable for several reasons. These reasons have to do mainly with wavelength conversion, protection, grooming, aggregation, demarcation, network monitoring, etc. Therefore, independent of transparency requirements it seems that some limited optoelectronic conversion may be unavoidable for purposes that are not directly connected to the quality of the signal at the receiving nodes. For this reason islands of transparency were recently proposed [Wagner00] as a compromise between all-optical and opaque networks. In these networks selective regeneration is used at specific network locations as needed in order to maintain acceptable signal quality from source to destination. This approach reduces the number of regenerators required compared to the case of opaque networks, but requires more complicated monitoring of the signal quality and resource allocation schemes.

3.2 Routing in Optical networks

With recent technology advances optical networks are evolving from simple point-to-point links into transparent architectures supporting switching using optical add/drop multiplexer (OADM) and optical cross-connect (OXC) nodes.

OADMs are elements that provide capability to add and drop traffic in the network (similar to SONET ADMs). They are located at sites supporting one or two (bi-directional) fibre pairs and enable a number of wavelength channels to be dropped and added reducing the number of unnecessary optoelectronic conversions, without affecting the traffic that is transmitted transparently through the node (Figure 3.1).

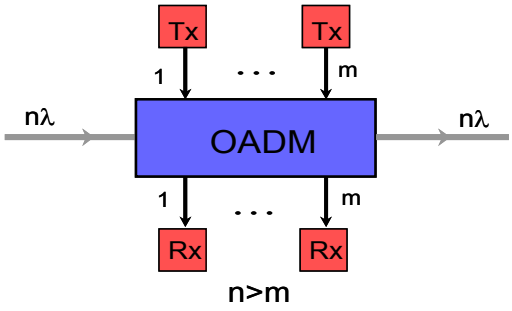

Figure 3.1: Generic Optical Add/Drop Multiplexer (OADM) architecture

Optical cross-connects (OXCs) are located at nodes cross-connecting a number of fibre pairs and also support add and drop of local traffic providing the interface with the service layer. To support flexible path provisioning and network resilience, OXCs normally utilise a switch fabric to enable routing of any incoming channels to the appropriate output port and access to the local client traffic. Various OXC architectures have been proposed and a common design is based on switches that are surrounded by wavelength multiplexers/demultiplexers as shown in Figure 3.2. Thus, an OXC can cross-connect the different wavelengths from the input to the output, where the connection pattern of each wavelength is independent of the others. By appropriately configuring the OXCs along the physical path, a logical connection may be established between any pair of edge nodes.

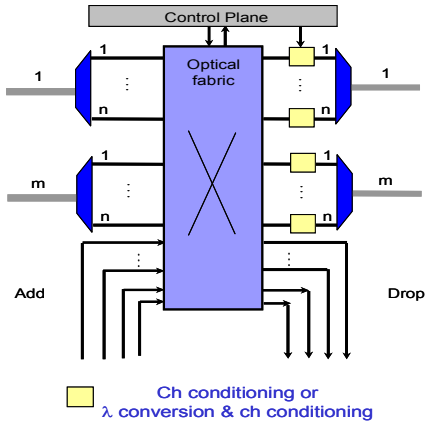

Figure 3.2: Transparent optical cross-connect (OXC) technologies

Routing in communication networks generally involves the identification of a path for each connection request between two discrete network locations (nodes). For routing in all-optical networks not only the path but also the wavelength should be determined, which results in the so called **"Routing and Wavelength Assignment" (RWA)** problem [Evo00] : *"Given one or more connections that need to be established in an all-optical domain, determine the routes over which each connection should be routed and also assign each connection a colour".* If the routes are already known, the problem is called the **"Wavelength Assignment" (WA)** problem in which two lightpaths must not be assigned the same wavelength on a given link.

The network shown in Figure 3.3 is called **wavelength routed network** and consist of several OADMs and OXCs interconnected through optical fibres and edge nodes which provide the interface between non-optical end systems (such as IP routers, ATM switches, or supercomputers) and the optical core. The optoelectronic conversion is done at the edge nodes and hence everything between the two edge nodes is pure optical. The main mechanism of transport in such networks is the lightpath (also referred to as λ-channel), which is an optical connection channel established over the network of OXCs and OADMs and which may span a number of fibre links (physical hops). Thus lightpaths are end-to-end paths in which signals propagate all optically as shown in Figure 3.3 as red and green direct lines.

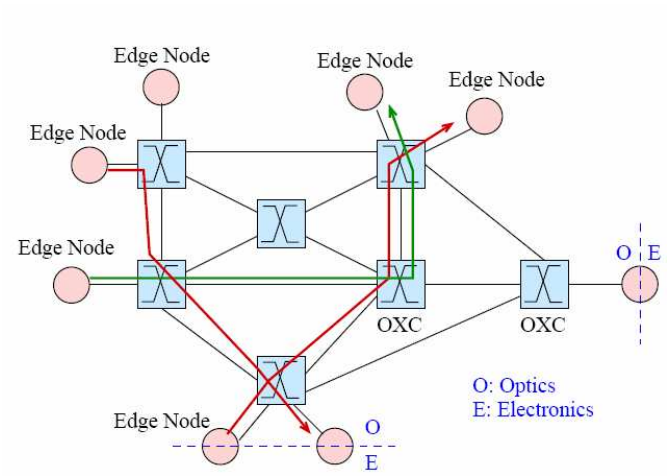

Figure 3.3: A wavelength routed WDM network

If **wavelength conversion** is allowed in the network, a lightpath can exit an intermediate node on a different wavelength than the one it has entered the node with. If no wavelength conversion is allowed then the wavelength continuity constraint is imposed to the generic RWA problem. This constraint specifies that a lightpath should occupy only a specific single wavelength, throughout the route from the source to the sink node and cannot change at any point.

3.3 Algorithms Description

For the routing problem there are generally three approaches that are used in the literature depending on whether the paths are pre-calculated or not. The **fixed routing** assumes a single, specific path for each connection. Obviously, this method can lead to high blocking probabilities, if the resources along the path are tied up, and it cannot directly handle faults in the network. The **fixed-alternate** approach considers multiple alternative routes in the network between each source and destination node. On the other hand **adaptive routing** is preferred, as the route between a source and a destination is chosen dynamically at every instance that is requested depending on the state of the network at that instant. When a connection request appears, an appropriate path is calculated. Here a connection is blocked only if there is no possible route between the nodes. It must be noted that in order to perform adaptive routing in a dynamic network, there is an overhead at the control and management layers. However the network performance is expected to be enhanced.

Most connection set-up algorithms calculating a path between two nodes are based on a standard **shortest path algorithm**. Often used is the well-known Dijkstra algorithm [Gross98] that - in binary heap implementation - has a complexity of $O(m+n\log n)$, where m is the number of edges and n the number of vertices. A shortest path algorithm minimizes the route weight (which is the sum of its link weights). The main motivation for this is that the weights can be used in order to consider some cost parameter when performing the routing. For example, all link-weights equal to one means that no link has precedence (e.g., to achieve minimum network load) or link-weights equal to the reciprocal of the free capacity means that more congested links are avoided (e.g., to achieve load balancing). Not only shortest path algorithms use weights, but in general also other algorithms use weights as their minimization goal. Note that the terms weight, length, cost, and metric are often used synonymously. Some path computation approaches are:

1. ***k*-Shortest paths algorithms** (see, e.g.[Eppstein94]) compute paths in the following way. Suppose w_1 is the minimum weight achieved by some path between the end nodes, w_2 the next larger weight, etc. The algorithm calculates all paths with w_1, all paths with w_2,… until w_k is reached. Note that for a specific w_i, multiple paths can exist. *k*-Shortest paths algorithms are used, e.g., to produce set of path options for iterative algorithms.

2. ***k*-Shortest edge-disjoint paths algorithms** and ***k*-shortest node-disjoint paths algorithms** (see [Bhandari99]) compute *k* paths (i) whose overall weight sum is at minimum and (ii) which are mutually edge-disjoint and node-disjoint, respectively. This is used, e.g., with *k*=2 for 1+1 protection.

3. **Shortest path algorithms subject to constraints** [Chen98] compute paths which are minimal in weight and which fulfill a set of further constraints. Common are constraints in terms of additive, multiplicative, and concave functions of further link-values. These constraints can be used to directly include physical effects in the computation.

4. **Shortest pair of disjoint paths algorithms** subject to constraints combine 2 and 3.

The above algorithms have to be extended with the wavelength assignment problem. The wavelength assignment problem is surveyed in [Zang00]. For the static networks with a given set of paths, the wavelength

assignment problem is subsequently directed to assign a wavelength to each path in a way that no two paths share the same wavelength on the same fiber link. For the dynamic cases, however, there are plenty of heuristics that have been proposed and can be combined with a path calculation algorithm. Here, minimizing the blocking probability is the main issue. The calculations are performed on line and make use of the current state information. Some more details on the wavelength assignment algorithms are provided in [Zang00]:

Random Wavelength Assignment (R). As the name implies, this scheme searches among the available wavelengths on the required route to chose one randomly (usually with uniform probability).

First-Fit (FF). Here, when searching for available wavelengths, a lower numbered wavelength is considered before a higher-numbered wavelength. The first available wavelength is then selected. This scheme performs well in terms of blocking probability and fairness, and is preferred in practice because of its small computational overhead and low complexity.

Least Used (LU). LU selects the wavelength that is the least used in the network, thereby attempting to balance the load among all the wavelengths.

Most-Used (MU). MU is the opposite of LU in that it attempts to select the most-used wavelength in the network. In a single-fibre network, MU becomes FF.

Least-Loaded (LL). The LL heuristic, like MU, is also designed for multi-fiber networks. This heuristic selects the wavelength that has the largest residual capacity on the most-loaded link along route.

MAX-SUM (MS). MS was proposed for multi-fiber networks but it can also be applied to the single-fiber case. It considers all possible paths (paths with their pre-selected routes) in the network and attempts to maximize the remaining path capacities after path establishment.

Relative Capacity Loss (RCL). RCL was proposed in [Zhang98] and is based on MS.

Wavelength Reservation (Rsv). In Rsv, a given wavelength on a specified link is reserved for a traffic stream, usually a multi-hop stream.

Protecting Threshold (Thr). In Thr, a single-hop connection is assigned a wavelength only if the number of idle wavelengths on the link is at or above a given threshold [Birman95]

3.4 Literature Review on RWA

The RWA has received extensive attention in the literature, mostly from a mathematical perspective. (e,g [Rama98, Zang00, Muckhe97]). The underling mathematical problem is very hard in general. The WA problem can be seen as the equivalent to the problem of coloring the nodes of a graph so that no two nodes connected by an arc of the graph have the same color: Simply represent each connection by a node, and connect every pair of nodes whose corresponding connections ride on the same link (and so need to be assigned different

wavelengths). This coloring problem is known to be **NP-complete** (Nondeterministic Polynomial-complete) [Garey79] which means that, in general, it is computationally intractable, but there are many fast but approximate (heuristic) algorithms for solving it. These algorithms try to optimize a properly selected cost function and reduce the complexity of the RWA problem.

An early work performed focusing on the RWA problem that initiated the study on the routing and wavelength assignment in WDM optical networks is presented in [Banerjee96]. The objective of this work is to minimize the number of the required wavelengths for the routing of a fixed set of connection requests. RWA problem is segmented in smaller sub-problems which are solved independently of each other using efficient, approximating techniques. For the determination of the paths of each connection, a multi-commodity flow formulation is used in conjunction with randomized rounding techniques. The wavelength assignment part of the problem is carried out by the use of similar techniques that are adopted for the graph-coloring problem.

In [Mokhtar98], the authors considered the RWA problem as a joint optimization problem and compared different schemes for wavelength assignment (first fit, also studied in [Chlamtac92], random fit, maximum wavelength utilization [Bala95]). It was shown that the scheme which tries to first allocate the most used wavelengths is more efficient. The criterion that was used for this purpose was the blocking performance of the network. An analytical technique for the computation of the blocking probability is also presented for fixed and alternate routing. The wavelength assignment in fixed routing optical networks is also studied in [Subram97]. In [Hyytia00] and [HyytiaVirtamo00], the first fit policy is proven to be the simplest and the one with the least complexity.

When dealing with RWA as a joint optimization problem, it is considered in an obvious way as a special case of the integer multi-commodity flow problem [Vazirani01] with additional constraints, where each lightpath corresponds to one flow unit, and is formulated as an integer linear program (ILP) [Papadimi98]. Typical RWA ILP formulations were initially proposed in [Banerjee96] and [Stern99]; they contain all necessary and sufficient types of constraints for a general RWA scheme to be valid (flow conservation, distinct wavelength assignment, wavelength continuity) and aim to minimize the maximum congestion (in terms of lightpaths) arising on network links. The dual scheme is discussed in [Rama95], that tries to maximize the number of connections established, while the traffic characteristics are a priori given and the network resources (number of available wavelengths) are constrained. A few newer and more sophisticated RWA ILP formulations are presented in [Krishna01-Design, Krishna01-Algo, Saad04]. Despite that these formulations are able to produce exact RWA solutions, ILP is generally NP hard. In addition, such approaches become space intractable when dealing with large networks, since the amount of ILP variables and constraints grows exponentially with network size. Time reduction is achieved by relaxing the integrality constraints and solving the resulting linear program (LP). Since fractional flows are not physically realizable in WDM optical networks (they are expressed in numbers of lightpaths), the LP solution must finally be converted to an integral one, that approximates the optimal value of the LP objective; that usually happens by utilizing appropriate rounding techniques. An RWA LP formulation recently proposed in [Ozdaglar03] has been shown to produce optimal integer solutions (without rounding) for a great fraction of RWA instances, despite the absence of integrality constraints. Space reduction is usually achieved by forcing lightpaths to be routed through a restricted subset of candidate paths, linear on the size of the network. This technique seems to decompose RWA; however, by selecting an appropriately big (but constant) number of candidate paths to serve each lightpath, the space of RWA solutions constructed is expected to be representatively large and contain an optimal RWA solution almost surely.

In [Birman95] various algorithms have been proposed for fixed and alternate routing for the selection of the path and the wavelength. [Birman96] gives approximate formulas for the blocking probability for fixed routing and random wavelength allocation. The same technique was extended in [Harai97] for the case of alternate routing and random wavelength allocation. In [Rama95], lower bounds for the blocking probability are calculated with and without wavelength conversion, by using an integer linear programming formulation. [Barry95] proposes a traffic model for circuit switched all-optical networks for the calculation of the blocking probability of a connection, using or not wavelength conversion capabilities.

Recently the incorporation of physical impairments in network design problems has received more attention from researchers focusing on transparent optical networks. Most reported studies can be classified into two categories: effects of impairments on network performance and network design with impairment consideration. In the first category the RWA algorithm is treated in two steps: first a lightpath computation in a network layer module is provided, followed by a lightpath verification performed by the physical layer module. In this viewpoint Huang et al [Huang05], modelled their impairment-aware RWA algorithm taking into account the PMD and OSNR performance parameters separately and compare the estimated PMD penalty along the calculated route and the computed OSNR level at the end of the route against two thresholds concerning each of the performance parameters. In [Cardilo05] the authors, used the OSNR model considered in [Huang05], with some enhancements to take into account nonlinearities stemming from the Kerr Effect as well. Also in [Ramam99], crosstalk and ASE noise were considered to evaluate the BER in the receiving end of the path. In the other category the physical layer impairments are considered before the network layer module proceeds to the lightpath computation and a validation of the signal quality requirements follows. In this aspect in [Martins03], the authors proposed a dynamic routing algorithm which selects the route based on lowest physical impairments, including ASE accumulation, amplifier gain saturation and wavelength dependent gain along the path and then calculate BER to check for the required signal quality. In [Kulkarni05], a scheme which takes into account physical linear impairments including noise, chromatic and polarization mode dispersion, crosstalk and filter concatenation effects was considered in an integrated approach through the estimation of the signal Q-factor. Link Q penalties are evaluated and assigned as cost to the links instead of the traditional link lengths, forcing the network layer module to determine less degraded routes An ultra long haul network scenario was examined in [Markidis06], where both linear and non linear transmission impairments are more intense and therefore regeneration is inevitable. The ICBR scheme presented in [Kulkarni05] was enhanced in [Markidis06], by considering also nonlinear effects and the penalty due to the jitter accumulation arising from 2R regeneration to demonstrate the superiority of the ICBR algorithm compare with the traditional shortest path algorithm.

3.5 Communication types

The most common communication type that is presented in many network scenarios is a point-to-point (p2p) connection. In many applications though, users require connections beyond that service. For example, VLBI (Verty Long Baseline Interferometry) projects a number of distributed radio-telescopes are simultaneously sending large amount of data to a single computation point for hardware correlation. Also Grid tasks that are executed on separated cluster environments may require high bandwidth connections between each others to

synchronize computation data. Therefore, p2p service may be insufficient for some final user groups, which must be adjustable to the needs of demanding Grid users.

Three types of connections with more than two end points are identified:

- **Point-to-multipoint**

P2MP is the connection type where a single source of information is sending data to multiple destination points. A special case of this scenario is found in VLBI projects, where data is sent in the reverse direction from multiple sources to a single destination point. The point-to-multipoint connection can easily be represented as a network star topology model.

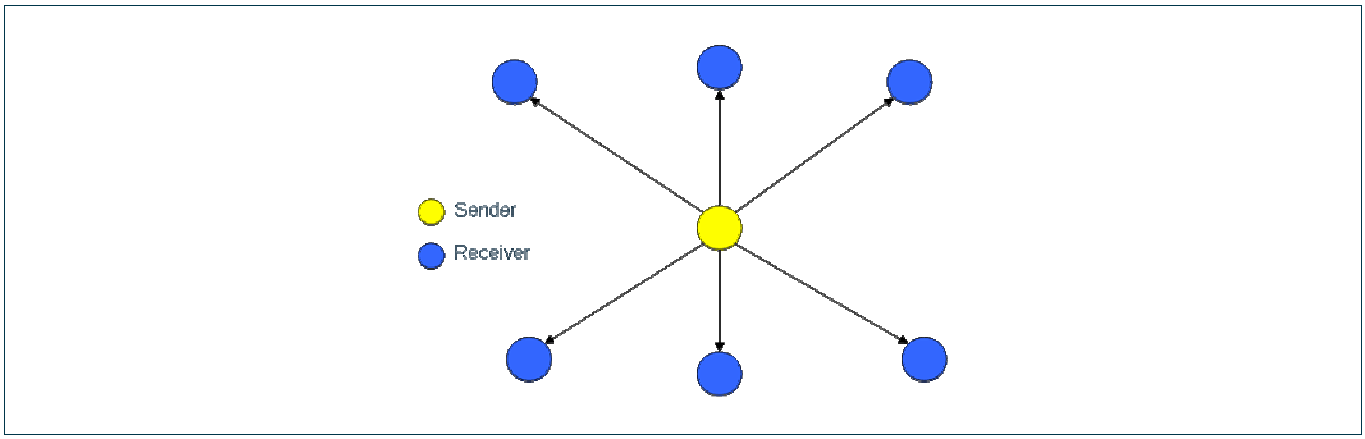

Figure 3.4: Point-to-multipoint connection.

- **Multicasting**

Multicasting is the connection type where a single source of information distributes the same data to multiple destination point simultaneously, optimizing link usage. Optimization in this case means that data is transferred over each link only once and is copied only where links split to deliver content to single receivers. The main idea of multicasting is to deliver the same content, at the same time to multiple users with minimum link utilization. Therefore multicasts transmission is often used for multimedia data streaming but also for computation data distribution (ftp) to multiple cluster infrastructures.

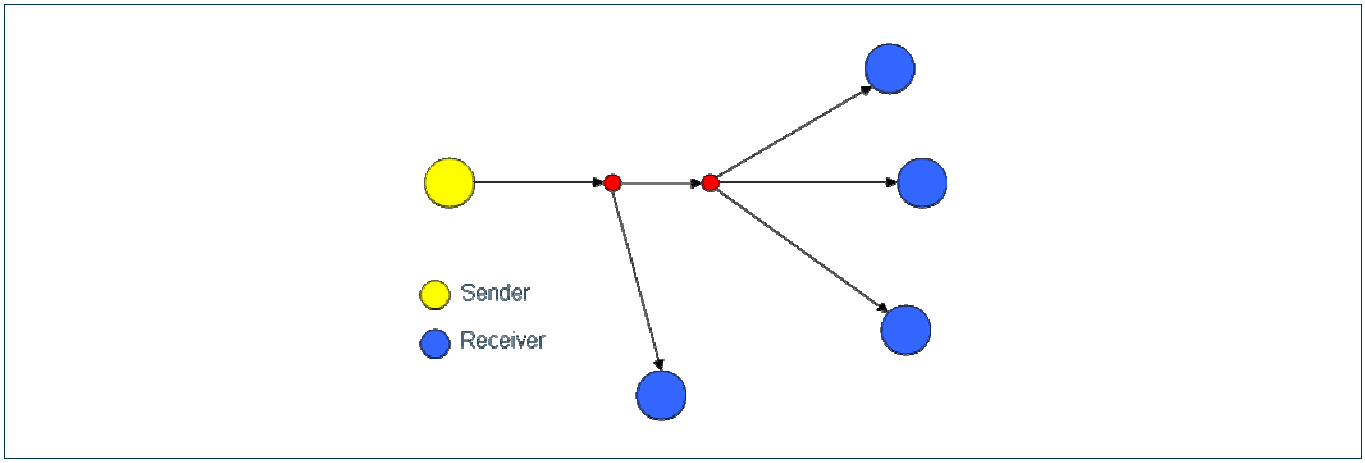

Figure 3.5: Point-to-multipoint connection.

- **Anycasting**

Anycast is a type of data transmission where data is sent from a single source to the nearest or best destination point. The idea behind anycast is that a client wants to send packets to any one of several possible servers offering a particular service or application but does not really care which one. The group of receivers is identified with a single routing address, however only one is receiving the data at the same time.

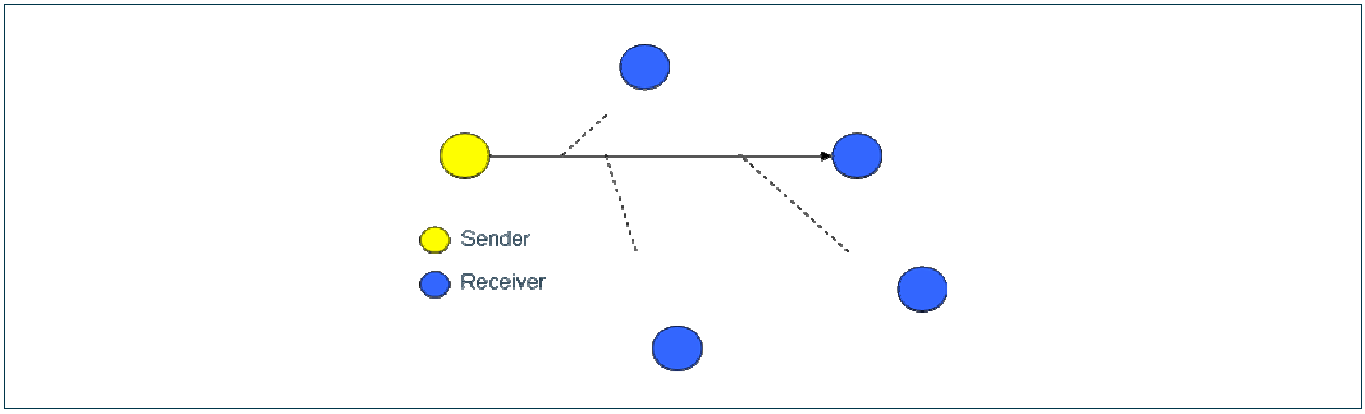

Figure 3.6: Anycasting.

3.6 Multi-domain routing

An important problem of global communication networks is the difficulty to efficiently manage such networks. Indeed, large-scale networks are generally composed of smaller sub-networks, usually referred to as ***domains***. The control and management of a single domain is performed locally, and information concerning state and availability is in general not shared with other domains. Special agreements (SLA or Service Level Agreement)

are usually required between different domains to create peering connections and allow transit data transfers. Problems arise for the control and management of interconnections of domains, i.e. a *multi-domain network*, since their size and heterogeneity make it difficult to collect all information needed to make optimal management decisions. The scale of the network directly influences the number of events related to network state and availability; transferring this data to the controlling entities, and in turn processing it can generate a considerable overhead, leading to inefficient network operation. Controlling the timing of when to send the state information, together with aggregation of this information (e.g. sending average values, aggregating information of multiple network links into a single value, and so on), can significantly reduce control plane overhead.

In essence, two different approaches are possible for the control of such networks, each having specific advantages and disadvantages.

- ***Centralized****:* A single control entity is aware of the full network and resource state of the multi-domain network. It receives all communication requests and is responsible for all scheduling decisions (i.e. when data transfer can start, which network route must be used, what level of reliability is available …). The main strengths of this approach are its straightforward deployment and reconfiguration possibilities. However, this approach is not scalable for larger networks, and suffers from a single point of failure.

- ***Distributed****:* in this case, resources send updates to all clients directly and clients individually perform the network control. An important assumption is that this approach requires total transparency between domains (which in reality is difficult to achieve). This means the number of status updates sent will increase dramatically compared to the centralized setup. An advantage of this setup is the removal of the single point of failure.

In section 6.2, we propose an alternative to these approaches, which tries to combine the advantages of both techniques while minimizing their respective problems.

3.7 Applicability to the control plane

The role of the control plane continues to evolve as increased intelligence is added to network elements and edge devices, to control the establishment and maintenance of connections in the network. Current control plane functions include: routing (intra-domain and inter-domain), automatic topology and resource discovery, path computation, signalling protocols between network switches for the establishment, maintenance, and tear-down of connections, automatic neighbour discovery and local resource management to keep track of available bandwidth resources.

Implementing specific control functions in the distributed control plane rather than in the centralized management plane could speedup the reaction time for most functions, improve the control plane's scalability, reduce operational time and costs and enable more agility in the behaviour of the optical network.

Currently, the GMPLS protocol suite [Mannie04] which is being developed by the IETF has gained significant momentum as a contender to become the basis of a uniform control plane that could be used for multiple

network layers and be responsible for switching of packets, flows, layer 2 paths, TDM channels, wavelengths, and fibers. GMPLS framework enables the capability of dynamically setting up transparent end-to-end connections but it still does not offer a way of guaranteeing the end-to-end optical signal quality.

A number of attempts that focus on different directions have been made to address the integration of physical layer impairments into the GMPLS control plane to provide optimize connection requests. The first approach deals with enhancing GMPLS signalling protocols to encompass physical impairments in GMPLS. Cugini et al. [Cugini05] presented a novel approach by introducing the required extensions into the signaling (Resource reservation Protocol with Traffic Engineering extensions, RSVP-TE) and management (Link Management Protocol, LMP) protocols. In the proposed scheme, lightpath routes from source to destination are dynamically computed by exploiting current OSPF-TE implementation, without taking into account the physical impairments. Only upon lightpath establishment, through the reservation protocol, the amount of impairments is dynamically computed. The lightpath set up request can be either accepted or rejected based on the amount of accumulated impairments already at some intermediate node or at the destination node. Link Management information is introduced in order to guarantee the link property correlation between adjacent nodes. This approach requires the introduction of a local database in each OXC to store the physical parameters that characterize the OXC itself and the links connected to its interfaces. OXCs automatically maintain local database synchronization and dynamically evaluate the lightpath signal quality by means of extensions to the LMP (Link Management Protocol) and the RSVP-TE (Resource Reservation Protocol with Traffic Extensions) protocols. Local database synchronization is guaranteed between adjacent OXCs by exchanging LMP packets over the out of band control plane. The Link Summary message of LMP is extended to contain the information regarding the link physical parameters specified in the local database whereas the RSVP-TE signalling protocol is extended to dynamically estimate the lightpath signal quality during the set up process. The lightpath setup process collects the physical parameters that characterize every traversed OXC and link from the source node to the destination node. Every considered physical parameter is separately evaluated to verify whether it falls within the acceptable range. Specifically, the source node generates an extended version of the RSVP PATH message containing the physical information of the transmitting interface and of its outgoing link. Every traversed network element, before propagating the PATH message, updates these parameters by adding up its own local values. Admission control at intermediate or at the destination node compares the overall accumulated parameter values with the local parameter ranges that characterize its interfaces. If the accumulated parameter values are within the acceptable range, the PATH message is propagated and eventually a RSVP RESV message is sent back to the source node. Otherwise the lightpath request is rejected and a proper RSVP ERROR is sent to the source node. In case the request is rejected, further set up attempts following different routes are triggered in order to avoid blocking induced by physical impairments. Periodic impairment-aware PATH and RESV Refresh messages are also utilized to keep the physical parameters updated and to automatically detect physical changes.

In the second approach additional information regarding the network physical parameters are inserted into the distributed routing protocol i.e. Open Shortest Path First with Traffic Engineering extensions (OSPF-TE). The computation of a path request is driven by the source node of the connection by interacting with the Traffic Engineering Database (TED) which collects the physical layer information. The TED is a repository located in each node with an updated picture of not only its local network resources (e.g. adjacent links) but also information related to remote links. Network-wide information stored in every TED serves as the input information for the RWA algorithms in order to compute optimal routes by using updated network-layer

attributes. This information should be frequently updated in order to provide a high degree of accuracy and correctness to the IA-RWA algorithms implemented in the distributed control plane. For this purpose, the existent GMPLS-based routing protocols need to be extended to flood optical performance parameters as traffic engineering (TE) attributes are disseminated [Strand01]. Impairment-related parameters are carried on the TE Link State Advertisements (TE-LSA) [Kompella05]. In particular, information is encapsulated within the top-level Link Type/Length/Value (TLV) as a common sub-TLV, which is referred as impairment sub-TLV. The contents of the impairment sub-TLV are: Type, which is used to identify uniquely this sub- TLV, Length, which contains the total length (in bytes) including the header of the sub-TLV, and Value, which contains the link parameters considered (e.g., OSNR, PMD). The construction of the proposed Impairment sub-TLV is similar as standardized TE information (link metric, unreserved bandwidth, etc.) Therefore, an on-line monitoring system can inform the Link Resource Manager (LRM) about changes in physical parameters such as ASE noise or PMD penalties in adjacent links which in turn informs the Routing Controller (RC) in order to flood the new physical value to the entire network by using the appropriate extensions to OSPF-TE [Martinez06]. As a result of the flooding mechanism every node's TED will be aware of the new impairment parameter value even if the link is not adjacent. This global information will be used by the IA-RWA algorithm during the path computation.

The possible path computation procedures identified in the context of PHOSPHORUS as define in [G^{2}MPLS-ARCH] should be in compliance with the IETF Path Computation Element *(*PCE) architectural model [Farrel06] since most of the protocol-specific issues are defined and solved in this framework. The integration of physical layer parameters in the control plane following this approach can be accomplished by introducing a separate component responsible for the inter- and intra-domain path computation based on specified constrains. This component can be identified as an application (different building block) residing within or externally to a network node, providing optimal routes and interacting with the control plane for the establishment of the proposed paths. The Path Computation Element (PCE) is an entity (component, application or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints during the computations [Farrel06]. The deployment of a dedicated PCE will relax the processing power needed by a network node to run constrained based routing algorithms and implement highly CPU-intensive optimization techniques. Also it may eliminate the need for the network nodes to maintain the memory demanding Traffic Engineering Database (TED) by establishing it on a separate node and making it available for path computation through the PCE. Another incentive that makes the solution of a separate PCE attractive is the optimal inter-domain routing which can be handled through distributed computation with cooperation among PCEs within each of the domains or even by a central PCE that has access to the complete set of topology information. Additionally the PCE can in an efficient manner consider local policies that impact the path computation and selection, in response to a path computation request, and also it can be used to compute backup paths in the context of fast reroute protection. Finally the sophisticated constraint routing algorithms utilize by the PCE can in a convenient way address issues like: i) resource coordination (e.g. CPUs, storage) ii) advance reservation iii) physical layer impairments in transparent optical networks iii) different connection types (unicast, multicast or anycast) and iv) QoS, separately or simultaneously in an integrated manner.

The PCE could represent a local Autonomous Domain (AD) that acts as a protocol listener to the intra-domain routing protocols e.g. OSPF-TE, and is also responsible for inter-domain routing. PCEs peer across domains and exchange abstract or actual topology information to enable inter-domain path computation and also utilize a modified version of OSPF-TE to share a link state database between domains. The constraint path computation process performed by the PCE can be described briefly in the following steps. Upon a request

arrival the user specified parameters carried by the LSP request are parsed into constraints inside the PCE which takes the responsibility to provide the required end to end path if possible. The coordination of Grid applications requires the rapid discovery of appropriate connections which are the result of very complex and intensive path computations. PCE can assist to this direction through the abstracted information of the global topology stored in the TED, of each domain, especially in cases where the network Management Plane is not able to provide this functionality. From the Grid constraints (Grid resource scheduling and coordination) described above the PCE constructs a reduced topology of the network based on which the IA-RWA algorithms that are implemented on the PCE proceed to the path calculation taking into consideration physical layer parameters to provide improved performance for the connection requests.

4 Infrastructure Considerations and User Related Requirements

In this section the most critical physical layer characteristics that should be considered by the routing algorithms to provide optimized performance and overcome certain infrastructure related restrictions are presented and thoroughly analyzed. In addition, a number of user related requirements that should also be taken into account by the routing algorithms to offer adequate network QoS are introduced.

4.1 Physical layer performance considerations

In all-optical networks it is usually assumed that all routes have adequate signal quality. Generally, this is obtained through some link budgeting procedures that, finally, impose limits to the geographic size of the all-optical domain by evaluating all the physical features of the network (amplifiers, fibers, wavelengths, etc.). However, the increasing bit rates of the line transmissions impose an increase in the injected power, which consequently implies a major impact on optical impairments.

Physical layer impairments may be classified as linear and non-linear. Linear impairments are independent of the signal power and affect each of the optical channels individually, whereas nonlinear impairments affect not only each optical channel separately but they also cause disturbance and interference between them. In this section we give a description of the origin and the impact of the impairments considered in the simulations.

4.1.1 Linear impairments

Amplified Spontaneous Emission noise

The advent of practical optical amplifiers capable of simultaneously amplifying multiple signal wavelengths that occupy an appreciable range of the optical spectrum was the key technological advance that ushered in the WDM revolution. Optical amplifiers are used at the end of each fiber span to boost the power of the WDM signal channels to compensate for fiber attenuation in the span.

Unfortunately, optical amplification is not possible without the generation of amplified spontaneous emission (ASE) noise which can be classified among the most severe impairment that limit the reach and capacity of WDM all-optical networks. Each optical amplifier contributes ASE, and these contributions add cumulatively along the amplifier chain. This accumulated ASE gives rise to signal-spontaneous beat noise at the receiver, which is the fundamental noise limit in an optically amplified transmission system. Each EDFA contributes an amount of ASE [Rama98]:

$$P_{ASE} = 2hvB_0 n_{sp}(G-1) \tag{1}$$

where P_{ASE} is the power in an optical bandwidth B_0 , h is Planck's constant , v is the optical frequency, n_{sp} is the spontaneous emission factor, and G is the optical amplifier gain. The spontaneous emission factor n_{sp} is determined by the inversion of the amplifiers Er ions. The contribution of each amplifier's ASE to the accumulated ASE is characterized by the amplifier's noise figure (NF), which at high gain can be approximated by $NF \approx 2n_{sp}$.

Following the analysis presented by [Ramam99] the ASE power through the inline amplifiers can be expressed as follows:

$$\begin{aligned} P_{ase}(k,\lambda_i) = {} & P_{ase}(k-1,\lambda_i)L_f(k-1,\lambda_i)G_{in}(k,\lambda_i)L_{tap} \\ & + 2.nsp.\left[G_{in}(k,\lambda_i)-1\right]hv_iB_o.L_{tap} \end{aligned} \tag{2}$$

and the ASE power through the nodes can be expressed by :

$$\begin{aligned} P_{ase}(k,\lambda_i) = {} & P_{ase}(k-1,\lambda_i)L_f(k-1,\lambda_i)G_{in}(k,\lambda_i)L_{dm}(k)L_{sw}(k)L_{mx}(k)G_{out}(k,\lambda_i)L_{tap}{}^2 \\ & + 2.n_{sp}.[G_{in}(k,\lambda_i)-1].hv_iB_o.L_{dm}(k)L_{sw}(k)L_{mx}(k)G_{out}(k,\lambda_i)L_{tap} \\ & + 2.n_{sp}.[G_{out}(k,\lambda_i)-1].hv_iB_oL_{tap} \end{aligned} \tag{3}$$

where $P_{ase}(k,\lambda_i)$ corresponds to the ASE noise power at the k^{th} amplifier and λ_i wavelength and $L_x(k,\lambda_i)$ and $G_x(k,\lambda_i)$ are the losses and gain of the various elements through the amplifier chain. The ASE noise variance at the end of the chain is described by:

$$\sigma^2_{ASE} = 4R_\lambda{}^2 b_i P_{avg}(N,\lambda_i)P_{ASE}(N,\lambda_i)B_e / B_o \tag{4}$$

where b_i is zero or two if i=0 or i=1, R_λ the responsivity of the receiver (1.25 A/W), P_{avg} the average signal power and B_e the electrical bandwidth of the receiver. The ASE noise variance will be used to calculate the Q factor degradation due to ASE.

Another constraint on the maximum number of optical amplifiers can be set, that is proportional to the average optical power P_{avg} launched at the transmitter and inversely proportional to an acceptable optical SNR_{min}, Planck's constant h, carrier frequency u, optical bandwidth B_0, amplifier gain G and amplifier spontaneous emission noise n_{sp} and given by

$$N \le \left\lfloor \frac{P_{avg}}{2huB_0(G-1)n_{sp}SNR_{\min}} \right\rfloor \quad (5)$$

A more generalized ASE noise constraint can be expressed as

$$\sum_{j=1}^{N} n_{sp.j}(G, j-1) \le \left\lfloor \frac{P_{avg}}{2huB_0 SNR_{\min}} \right\rfloor \quad (7)$$

to account for different fiber losses and different types of optical amplifiers.

Chromatic Dispersion (CD)

Chromatic dispersion or group velocity dispersion (GVD) has been considered for many years the most serious linear impairment for systems operating at bit rates from 2.5 Gbps to 10 Gbps and is causing different frequencies of light to travel at different speeds. This linear process causes broadening of the optical pulses, resulting in inter-symbol interference which impairs system performance. In this respect, chromatic dispersion imposes the limitation of the maximum transmission distance.

Chromatic Dispersion arises for two reasons. The first is the dependence of the optical fibre's index on the optical wavelength (material dispersion) and the second is due to waveguide dispersion where the power distribution of a mode between the core and the cladding of the fiber is a function of the wavelength.

In order to minimize the chromatic dispersion, various dispersion compensation techniques, which use a dispersion compensation fiber (DCF), have been studied [Breuer95, Rothnie96, Hayee97], and the dispersion shifted fibers (DSF) have been deployed. It is noteworthy that the effect of GVD combined with fiber nonlinearities, such as self- and cross-phase modulation (SPM/XPM) and four wave mixing (FWM), is much more complicated since GVD can increase or alleviate the effects of fiber nonlinearities.

For this reason, for part of the simulation studies, presented in this deliverable, we handle chromatic dispersion semi-analytically together with Self Phase Modulation (SPM) and evaluate an eye closure penalty induced from the combination of the two impairments.

Also in another part of the simulations, the chromatic dispersion has been considered as an upper bound on the maximum length of an M-link segment depending on the bit rate B, the chromatic dispersion D_{cd} and the modulation format use under the following formula:

$$\sum_{l=1}^{M} D_{CD,l} d_l < \frac{d}{B^2} \quad (8)$$

where d_l is the length of the l link in kilometers and d is a constant depending on the modulation format .

Crosstalk (XT)

Crosstalk is introduced in WDM systems when leakage of optical signals generated through multiplexing/demultiplexing, switching, and other optical components interferes with the data channels, imposing power penalties in the system. Therefore, the level of crosstalk introduced across an optical path is closely linked with the node architecture and the technologies of the elements comprising the nodes. In our simulation network we considered an OXC based on the wavelength selective architecture that is depicted in Figure 4.1. Two types of XT arise in WDM systems: inter-channel (inter-band) and intra-channel (intra-band). The former is when the interfering XT element is on a sufficiently different wavelength compared to the wavelength of the desired signal so that the difference in wavelength between the signal and the XT element is larger than the receiver's electrical bandwidth. The latter occurs when the interference is on the same wavelength or sufficiently close with the desired signal so that the difference in wavelengths is within the electrical bandwidth of the receiver.

We consider only intra-channel crosstalk as its effect is much more severe compared to inter-channel crosstalk and we study crosstalk in conjunction with ASE, since both impairments are very closely related to the power of the signal traversing the OXCs. The amount of energy that leaks to neighbouring wavelengths is described by the signal-to-crosstalk ratio (X_{sw}) and is expressed as [Ramam99] :

$$P_{XT}(k,\lambda_i) = \sum_{j=1}^{J_k} X_{sw} p_{in}(j,k,\lambda_i) L_{sw}(k) L_{mx}(k).G_{out}(k,\lambda_i) L_{tap} \qquad (9)$$

where $p_{in}(j,k,\lambda)$ is the power of the *jth* co propagating signal at the switched shared by the desired signal, and J_k is the total number of crosstalk sources at the *kth* node.

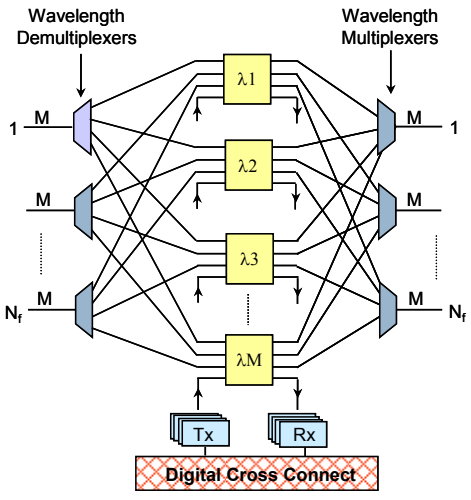

Figure 4.1: Node architecture

Finally the noise variance of crosstalk is described by [Ramam99]:

$$\sigma_{XT}^2 = 2\xi_{pol} R_\lambda^2 b(i) P_{ave} P_{XT} \quad (10)$$

where ξ_{pol} is the polarization mismatch factor between the signal and the crosstalk lightwaves.

Filter Concatenation (FC)

A serious signal impairment that is unique to all-optical networks is distortion-induced eye closure, an effect that is produced by signal passage through multiple WDM filters between the source and the receiver and originates mainly from spectral clipping due to the narrowing of the overall filter pass-band [Tomkos01]. This effect is essentially relatively small in a point-to-point optical system since a given signal passes through at most two filters: a MUX and a DMUX.

However, in a transparent optical network, a signal may be demultiplexed and remultiplexed at many network elements throughout its path before it is finally received. Thus the signal experiences the concatenation of the entire set of filters in its path. The effective spectral transfer function of the filter set is the multiplication of each of the individual filters' transfer function, and can therefore be much narrower in spectral width than that of a single filter [Antoniades02]. This in turn can lead to a time-domain distortion and a distortion-induced eye closure penalty that is related to Q penalty.

Polarization Mode Dispersion (PMD)

Polarization Mode Dispersion is the most important polarization effect for high capacity, long haul systems with high bit rates. PMD arises from the birefringence in the fiber that gives rise to the differential group delay between the two principal states of polarization. PMD is manifest as a time varying and statistical pulse broadening and pulse distortion because the perturbation of the fiber symmetry that gives rise to the birefringence varies randomly in orientation along the fiber and is also dependent on environmental variations, particularly temperature. Because of the statistical nature of PMD, the differential group delay increases with the square root of the length of the fiber and is expressed in units of $ps/\sqrt{km}$. The penalty induced by the PMD is model using [Cantrell03]:

$$Q_{PMD} = 10.2.B^2 D_{PMD} L \quad \text{in dB} \quad (11)$$

where D_{PMD} is the fiber dispersion parameter (0.5 or 0.1 $ps/\sqrt{km}$ for old and new fibers respectively) , L is the length of the transmission fiber and B is the signal bit rate.

Also an upper bound on the maximum lengths of an M-link segment can be defined by

$$\sqrt{\sum_{l=1}^{M} (D_{PMD,l})^2 d_l} < \frac{f}{B} \quad (12)$$

where *f* is a fraction of the bit duration (typically 0.1) and *B* denotes the bit rate.

4.1.2 Nonlinear Impairments

In WDM systems there are, in general, two ways to increase the system channel capacity. One is to increase the number of WDM channels, and the other is to increase the channel bit rate for each wavelength. Both attempts are associated with higher total injected power into the fiber, leading to the intensification in the fiber nonlinearities which fall into two categories. One is stimulated scattering (Raman and Brillouin), and the other is the optical Kerr effect due to an harmonic motion of bound electrons in the material resulting in an intensity dependent refractive index with optical power [Agrawal95]. While stimulated scatterings are responsible for intensity dependent gain or loss, the nonlinear refractive index is responsible for intensity dependent phase shift of the optical signal.

In the simulations we consider only nonlinearities stemming from the Kerr effect which occur due to the nonlinear relationship between the induced polarization ***P*** and the applied electric field ***E*** when higher powers and/or bit rates are applied as shown in (Eq.13)

$$P = \varepsilon_0 \left(\chi^{(1)} E + \chi^{(3)} E^3 + \ldots \right) \tag{13}$$

where ε_0 is the permittivity of vacuum and $\chi^{(j)}$ the *j-th* order susceptibility. The linear susceptibility $\chi^{(1)}$ is the dominant contribution to the polarization ***P*** and its effects are included through the refractive index n [Agrawal95]. The cubic term $\chi^{(3)}$ is responsible for phenomena like third-harmonic generation, four wave mixing and nonlinear refraction. The first two processes (processes that generate new frequencies) are usually not important unless phase matching conditions are satisfied. Nonlinear refraction instead is always present and deeply affects the propagation of intense light in an optical fiber. The electromagnetic wave passing along the optical fiber induces a cubic polarization which is proportional to the third power of the electric field (13). This is equivalent to a change in the effective value of $\chi^{(1)}$ to $\chi^{(1)} + \chi^{(3)}E^2$. In other words the refractive index is changed by an amount proportional to the optical intensity

$$n(I) = n + n_2 I \tag{14}$$

where *n* is the linear part , *I* is the optical intensity and n_2 is the nonlinear-index coefficient related to $\chi^{(3)}$ by

$$n_2 = \frac{2}{\varepsilon_0 c n} \frac{3}{8} \chi^{(3)} \tag{15}$$

This intensity dependence of the refractive index (optical Kerr effect) is responsible for numerous nonlinear effects. Note that even if the value of the nonlinear coefficient n_2 is quite small, nonlinear effects in optical fibers assume a relevant importance due to the fact that the magnitudes of these effects depend on the length of the fiber along which the wave travels and on the ratio n_2/A_{eff}, where A_{eff} is the effective area of the lightmode. Despite the intrinsically small values of the nonlinear coefficient for silica, the nonlinear effects in optical fibers

can be observed even at low powers considering that the light is confined in a relative small area over long interaction lengths due to the extremely low attenuation coefficient and the event of optical amplifiers. This is the reason why nonlinear effects can not be ignored when considering light propagation in optical fibers.

One manifestation of the intensity dependence of the refractive index occurs through self-phase modulation (SPM), a phenomenon that leads to spectral broadening of optical pulses travelling along a fiber. Cross-phase modulation (XPM) is the analogue of SPM but this time the induced phase depends not only on its own intensity, but also on the one of the other co-propagating lightwaves. Another nonlinear effect that can be relevant in optical fibers is four wave mixing (FWM). Due to this phenomenon new frequencies are generated that coincide with the transmission channel and induce dependent interferences which degrade the transmitted signal.

The propagation of the signal across the fibre is described by the nonlinear Schrödinger differential equation (NLSE)

$$\frac{\partial}{\partial z}A(z,t)+\frac{\alpha}{2}A(z,t)+\frac{i}{2}\beta_2\frac{\partial^2}{\partial t^2}A(z,t)-\frac{1}{6}\beta_3\frac{\partial^3}{\partial t^3}A(z,t)-i\gamma\left|A(z,t)\right|^2A(z,t)=0 \qquad (16)$$

where $A(z,t)$ is the envelope of the transmitted signal which is assumed to vary slowly compared to the carrier wave. The equation as shown takes into account only the Kerr effect, however it can be modified to include Raman scattering as well. This equation cannot be solved in the general case, and therefore a numerical solution has to be used. The most common technique used for this purpose is the Split Step Fourier method, where the linear and non-linear parts of the differential equation are solved independently for a small section of the fibre, and their result is added together and used for the calculation of the next small step. The complication is that the linear part is best solved in the frequency domain, whereas the non-linear part is best solved in the time domain, requiring the calculation of a considerable number of Fourier Transforms. This is usually a time consuming process, particularly for large WDM systems. Therefore, **analytical models** for the WDM effects like the ones presented here are of considerable interest.

Self Phase Modulation (SPM)

The first effect of the nonlinear refractive index that must be considered in both single and multi-channel transmission is self phase modulation, that is, the change of the optical phase of a channel by its own intensity.

The NLSE can also be expressed as [Agrawal95]

$$i\frac{\partial A}{\partial z}-\frac{1}{2}\beta_2\frac{\partial^2 A}{\partial T^2}+\frac{1}{2}iaA+\gamma\left|A\right|^2A=0 \qquad (17)$$

where A is proportional to the slowly varying amplitude of the electric field of the pulse envelope, β_2 is the second-order GVD, and γ is the nonlinear coefficient, defined as $\left(\gamma=\frac{2\pi n_2}{\lambda A_{eff}}\right)$. T=t-z/$u_g$ is the frame of reference moving with the pulse at the group velocity, u_g. It can be found from (Eq.17) that the second term is

related to chromatic dispersion, the third term is contributed by the fiber loss, and the forth term is associated with the fiber nonlinearity. Assuming the absence of fiber loss and dispersion, the electric field of the pulse envelope is in the following form [Agrawal95]

$$E(z,T) = E_0(T)e^{i\varphi_{NL}(z,T)} \tag{18}$$

where $E_o(T)$ is the electric field at $z = 0$, and φ_{NL} is the nonlinear phase shift, defined as

$$\varphi_{NL}(z,T) = \gamma P(T) z \tag{19}$$

where $P(T)$ is the power which is proportional to $|E_o|^2$. It can be seen in (Eq.19) that the time dependence of $\varphi_{NL}(z, T)$ is related with the instantaneous optical frequency, $\delta\omega(T)$ that is given by [Agrawal95]

$$\delta\omega(T) = -\frac{\partial\varphi_{NL}(z,T)}{\partial T} \tag{20}$$

δω(T) can be considered as a frequency chirp or new frequency component. Consequently, SPM induces the pulse spectral broadening while the pulses propagate in the fiber.

It is important to take into account the effect of SPM combined with chromatic dispersion since both affect the signal quality in terms of pulse broadening but each operates in different way. These combined effects have been studied extensively in [Hayee97, Stern90]. SPM process produces new frequency components as the pulse propagates through the fiber. The new frequency components are a positive linear frequency chirp. In the anomalous regime ($\beta_2 < 0$), chromatic dispersion produces negative frequency chirp and it tends to negate the positive chirp induced by SPM. Consequently, interacting between SPM and dispersion results in the reduction of the pulse broadening. On the other hand, in the normal dispersion region ($\beta_2 > 0$), chromatic dispersion also generates positive linear frequency chirp. Therefore, the effect of the pulse broadening is much more accelerated since the pulse has been spread out by both of SPM and dispersion as depicted in Figure 4.2.

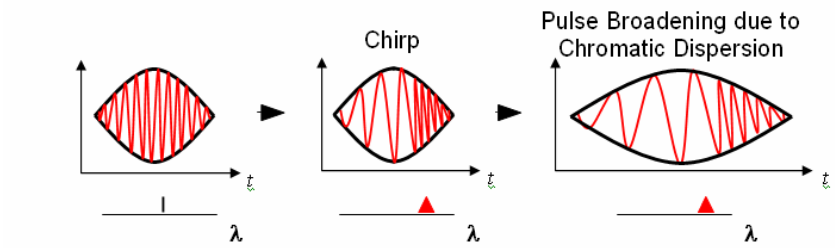

Figure 4.2: SFM effect combined with chromatic dispersion

Instead of solving the NLSE we approximate SPM effects and chromatic dispersion analytically. We assume that the transmitter has an extra frequency chirping that causes the same amount of distortion as if there was a

transmission through a nonlinear optical fiber (causing the same amount of distortion due to chromatic dispersion).

The frequency chirping of the transmitter can be modelled by:

$$a_{SPM} = -\frac{k}{B}\sum_{i=1}^{N}\gamma_i\left(\frac{P^i D^i}{a^i}\left\{l^i - \frac{1}{a^i}\left(1-e^{-a_i l_i}\right)\right\} + \frac{P^i}{a^i}\sum_{k=i+1}^{N} l^k D^k\left\{1-e^{-a_i l_i}\right\}\right) \tag{21}$$

Where *k* is the chirp parameter, *B* is the total chromatic dispersion of the link, *N* is the number of fiber segments, *P* the input power of each segment, *D* the chromatic dispersion, a the attenuation and *l* the length of the respective segment. The summation in (Eq.14) can be simplified to the summation of the following terms each one representing the fibers segments that formulate the link:

$$\begin{aligned}
I_{span} = {} & M\gamma_{SMF}\frac{P_{inSMF}D_{SMF}}{a_{SMF}}\left(l_{SMF} - \frac{1-e^{-a_{SMF}l_{SMF}}}{a_{SMF}}\right) \\
& + M\gamma_{SMF}\frac{P_{inSMF}}{a_{SMF}}\left(1-e^{-a_{SMF}l_{SMF}}\right)\left(D_{DCF}l_{DCF} + D_{post}l_{post}\right) \\
& + \left(M^2 - \frac{M(M+1)}{2}\right)\gamma_{SMF}\frac{P_{inSMF}}{a_{SMF}}\left(1-e^{-a_{SMF}l_{SMF}}\right)\left(D_{DCF}l_{DCF} + D_{SMF}l_{SMF}\right) \\
& + M\gamma_{DCF}\frac{P_{inDCF}D_{DCF}}{a_{DCF}}\left(l_{DCF} - \frac{1-e^{-a_{DCF}l_{DCF}}}{a_{DCF}}\right) \\
& + \left(M^2 - \frac{M(M+1)}{2}\right)\gamma_{DCF}\frac{P_{inDCF}}{a_{DCF}}\left(1-e^{-a_{DCF}l_{DCF}}\right)\left(D_{SMF}l_{SMF} + D_{DCF}l_{DCF}\right) \\
& + M\gamma_{DCF}\frac{P_{inDCF}}{a_{DCF}}\left(1-e^{-a_{DCF}l_{DCF}}\right)\left(D_{post}l_{post}\right)
\end{aligned} \tag{21}$$

$$\begin{aligned}
I_{PRE} = {} & \gamma_{PRE}\frac{P_0 D_{PRE}}{a_{PRE}}\left(l_{PRE} - \frac{1-e^{-a_{PRE}l_{PRE}}}{a_{PRE}}\right) \\
& + \gamma_{PRE}\frac{P_0}{a_{PRE}}\left(1-e^{-a_{PRE}l_{PRE}}\right)\left\{M\left(D_{SMF}l_{SMF} + D_{DCF}l_{DCF}\right) + D_{post}l_{post}\right\}
\end{aligned} \tag{22}$$

$$I_{POST} = \gamma_{POST}\frac{P_{inPOST}D_{POST}}{a_{POST}}\left(l_{POST} - \frac{1-e^{-a_{POST}l_{POST}}}{a_{POST}}\right) \tag{23}$$

where *M* is the number of spans referring to SMF and DCF segments, pre to the pre compensation fiber and post to the post compensation fiber.

Using this phase shift at the transmitter we can simplify the NLSE to a linear problem which we solve by convolving the transfer function of the fiber with the chirped signal generated by the receiver and measure the eye closure penalty at the end of the transmission.

Cross Phase Modulation (XPM)

Another nonlinear phase shift originating from the Kerr effect is cross-phase modulation (XPM). While SPM is the effect of a pulse on it own phase, XPM is a nonlinear phase effect due to optical pulses in other channels. Therefore, XPM occurs only in multi-channel systems. In a multi-channel system, the total nonlinear phase shift of the signal at the center wavelength λ_i is expressed by [Agrawal95],

$$\phi_i^{NL} = \gamma L_{eff} \left\{ P_i + 2\sum_{j \neq i}^{M} P_j \right\} \tag{24}$$

where γ is the nonlinear coefficient $\left(\gamma = \frac{2\pi n_2}{\lambda A_{eff}} \right)$, L_{eff} is the effective fiber length, and *M* is the total number of channel in the system. The first term in (Eq.24) is responsible for SPM, and the second term for XPM. This equation might lead to a speculation that the effect of XPM could be at least twice as significant as that of SPM. However, XPM is relevant only when pulses in the other channels are synchronized with the signal of interest. When pulses in each channel travel at different group velocities due to dispersion, the pulses slide past each other while propagating. Figure 4.3 illustrates how two isolated pulses in different channels collide with each other. When the faster travelling pulse has completely walked through the slower travelling pulse, the XPM effect becomes negligible. The relative transmission distance for two pulses in different channels to collide with each other is called the walk-off distance, *Lw* [Agrawal95].

$$L_W = \frac{T_0}{\left| v_g^{-1}(\lambda_1) - v_g^{-1}(\lambda_2) \right|} \approx \frac{T_0}{\left| D\Delta\lambda \right|} \tag{25}$$

where *To* is the pulse width, v_g is the group velocity, and λ_1,λ_2 are the center wavelength of the two channels. *D* is the dispersion coefficient, and $\Delta\lambda = |\lambda_1-\lambda_2|$.

When dispersion is significant, the walk-off distance is relatively short, and the interaction between the pulses will not be significant, which leads to a reduced effect of XPM. However, the spectrum broadened due to XPM will induce more significant distortion of temporal shape of the pulse when large dispersion is present, which makes the effect of dispersion on XPM complicated.

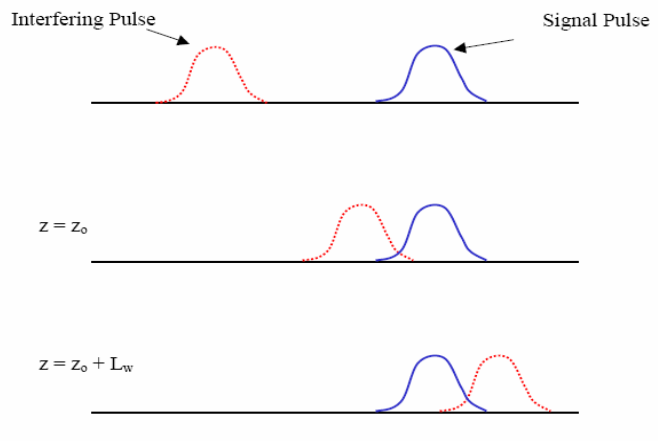

Figure 4.3: Illustration of walk-off distance

In this section we investigate the impact of XPM on the performance of a single link modifying Cartaxo analytical model [Cartaxo99] to be compatible with the structure of the link presented in Figure 5.4 The XPM induced intensity modulation (IM) frequency response is described by

$$H_{XPMik}^{IM}(\omega)=\frac{P_{XPM,ik}(\omega)}{P_k(\omega)}=2P_i(0).g_i^{net}(L_T).\exp\left[-j\omega\sum_{l=1}^{N}\frac{L^{(l)}}{u_{gi}^{(l)}}\right].\sum_{l=1}^{N}\gamma_l^{(l)}\exp\left[j\omega\sum_{l=1}^{N}d_{ik}^{(n)}L^{(n)}\right]$$

$$.\prod_{n=1}^{l-1}e^{-a^{(n)}L^{(n)}}\,g_k^{(n)}.\left\{\begin{array}{l}\dfrac{1}{\left(a_{ik}^{(l)}\right)^2+\left(b_i^{(l)}+q_k^{(l)}\right)^2}\left[a_{ik}^{(l)}.\sin(B_i^{(l-1)}-Q_k^{(l)})-\left(b_i^{(l)}+q_k^{(l)}\right).\cos(B_i^{(l-1)}-Q_k^{(l)})\right.\\ \left.\left[a_{ik}^{(l)}.\sin(Q_k^{(l+1)}-B_i^{(l)})+\left(b_i^{(l)}+q_k^{(l)}\right).\cos(Q_k^{(l+1)}-B_i^{(l)})\right].e^{-a_{ik}^{(n)}L^{(l)}}\right.\\ +\dfrac{1}{\left(a_{ik}^{(l)}\right)^2+\left(b_i^{(l)}-q_k^{(l)}\right)^2}.\left[a_{ik}^{(l)}.\sin(B_i^{(l-1)}+Q_k^{(l)})-\left(b_i^{(l)}-q_k^{(l)}\right).\cos(B_i^{(l-1)}+Q_k^{(l)})\right.\\ \left.+\left[-a_{ik}^{(l)}.\sin(Q_k^{(l+1)}+B_i^{(l)})+\left(b_i^{(l)}-q_k^{(l)}\right).\cos(Q_k^{(l+1)}+B_i^{(l)})\right].e^{-a_{ik}^{(n)}L^{(l)}}\end{array}\right\} \quad (26)$$

for multi-segment fiber links, where *i* stands for the probe signal and *k* for the pump, $P_{XPM,ik}(\omega)$ represents the XPM-induced IM originated by pump channel *k*, $P_k(\omega)$ the pump channel , $P_i(0)$ the power of channel *i* at the fiber input, $g_i^{net}(L_T)$ the net power gain for the i channel from the transmitter up to the receiver, L_T the total system length, *N* the total number of spans, $\gamma^{(l)}$ the nonlinearity constant for the l span ,$d_{ik}^{(l)}$ is the walk–off parameter between channels i and k in the l[th] segment (given by $d_{ik}=(u_{gi}^{(l)})^{-1}-(u_{gk}^{(l)})^{-1}$), *a* the attenuation constant and

$$a_{ik}^{(l)} = a^{(l)} - j\omega d_{ik}^{(l)} \quad b_i^{(l)} = \omega^2 D_i^l \lambda_i^2 /(4\pi c) \quad q_k^{(l)} = \omega^2 D_k^l \lambda_k^2 /(4\pi c)$$

$$B_i^{(l)} = \omega^2 \lambda_i^2 \sum_{n=l+1}^{N} L^{(n)} D_i^{(n)} /(4\pi c) \quad Q_k^{(l)} = \omega^2 \lambda_k^2 \sum_{n=1}^{l-i} L^{(n)} D_k^{(n)} /(4\pi c) \tag{27}$$

Using (Eq. 26) we can obtain an analytic expression for the XPM noise-like variance given by [Pachnicke03]

$$\sigma_{XPM}^2 = \overline{P}(0)^2 \sum_{k=1,k\neq i}^{N} \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| H_{XPM,ik}^{IM}(\omega, L) \right|^2 . \left| H_{opt.filter}(\omega) \right|^2 . PSD_k(\omega) d\omega \tag{28}$$

where *P(0)* is the average channel power, $H_{XPM,ik}(\omega)$ the transfer function due to XPM as described above, $H_{opt,filter}(\omega)$ the transfer function of the optical filter at the receiver and $PSD_k(\omega)$ is the power spectral density of channel *k*.

Four Wave Mixing (FWM)

The final nonlinear impairment that we examine in this study is Four-Wave Mixing (FWM) which is a major factor of the degradation of the system performance in multi-channel lightwave systems. The generation of FWM depends on many characteristics of the system such as channel spacing, fiber length, and chromatic dispersion. In particular, the influence of FWM is larger when the optical channels are equally spaced, and when wavelength division multiplexed (WDM) systems operate in the zero dispersion region.

FWM is the phenomenon where three waves (or two waves) are mixed during the propagation in the fiber and generate the new optical frequency (fourth wave or third wave), which may fall into other channels and cause the in-band crosstalk. Through the FWM process, a fourth (or third) wave will be generated at a frequency

$$f_{ijk} = f_i + f_j - f_k \, (i, j \neq k) \tag{29}$$

By modifying [Zeiler96, Inoue94] to be compatible with the link architecture we described in section 5.3 the FWM signal power can be expressed as

$$P_{FWM}=\frac{1024\pi^6.3^2}{n_o^4\lambda^2c^2}D^2P_0^3\left|\begin{array}{l}\frac{x_{222}}{A_{eff2}}.\frac{e^{-a_{pre}.L_{pre}+j\Delta k_{pre}L_{pre}}-1}{j\Delta k_{pre}-a_{pre}}+\frac{x_{111}}{A_{eff1}}.G_{\inf SMF}.\frac{e^{-a_1L_1+j\Delta k_1L_1}-1}{j\Delta k_1-a_1}.\\ e^{-a_{pre}.L_{pre}+j\Delta k_{pre}L_{pre}}.\left(\sum_{m=1}^{M}e^{j(m-1)(\Delta k_1L_1+\Delta k_2L_2)}\right)+\frac{x_{222}}{A_{eff2}}.G_{\inf SMF}.G_{intSMF}.\frac{e^{-a_2L_2+j\Delta k_2L_2}-1}{j\Delta k_2-a_2}.\\ e^{-(a_{pre}.L_{pre}+a_1L_1)+j(\Delta k_{pre}L_{pre}+\Delta k_1L_1)}.\left(\sum_{m=1}^{M}e^{j(m-1)(\Delta k_1L_1+\Delta k_2L_2)}\right)+\frac{x_{222}}{A_{eff2}}.G_{\inf SMF}.G_{intSMF}.G_{intDCF}.\\ \frac{e^{-a_{post}.L_{post}+j\Delta k_{post}L_{post}}-1}{j\Delta k_{post}-a_{post}}e^{-(a_{pre}.L_{pre}+a_1L_1+a_2L_2)+j(\Delta k_{pre}L_{pre}+M(\Delta k_1L_1+\Delta k_2L_2))}\end{array}\right|^2 \quad (30)$$

where n_o is the refractive index , D is the degeneracy factor ($D=3$ if i=j or $D=6$ if i≠j), P_0 is the launched signal power per channel (equal power per channel assumed, and also same polarization assumed) x_{111},x_{222} is the third-order nonlinear susceptibility and A_{eff1}, A_{eff2} the effective area of the SMF and DCF segments respectively, L is the length of a fiber segment and α is its corresponding attenuation, G is the gain of the amplifiers, M is the number of SMF spans which is the same with the number of the DCF spans and Δk_i is the phase mismatch which is related to signal frequency differences and chromatic dispersion D_i and may be expressed as

$$\Delta k_i=\frac{2\pi\lambda^2}{c}\left|f_i-f_k\right|\left|f_j-f_k\right|.\left\{D_i+\frac{dD_i}{d\lambda}\left(\frac{\lambda^2}{2c}\right)\left(\left|f_i-f_k\right|+\left|f_j-f_k\right|\right)\right\} \quad (31)$$

where $dD_i/d\lambda$ is the dispersion slope.

In the WDM systems, the total FWM power generated at the frequency, f_m is given by [Inoue92]

$$P_{FWM}(f_m)=\sum_{f_k=f_i+f_j-f_m}\sum_{f_j}\sum_{f_i}P\left(f_i+f_j-f_k\right) \quad (32)$$

In the equally spaced channel WDM systems, the channels in the middle of the signal band would be affected severely because the number of FWM signals is maximum at the center channel [Inoue92]. However, it is noteworthy that this is not always true if the wavelengths in the WDM system are far from the zero dispersion wavelength, and if there is substantial difference in chromatic dispersion between two channels. This is because FWM power, in general, rather than the number of FWM signals on each channel degrades the system performance. Methods of reducing the effect of FWM is to increase the channel spacing, apply proper dispersion maps and allocate the channels unequally [Kikuchi97], letting new optical frequency fall out of the channel band.

4.1.3 Performance Metrics

The performance of a digital lightwave system is commonly specified using the Q-factor. The Q-factor is the electrical signal-to-noise ratio at the input of the decision circuit in the receiver's terminal. This is shown schematically in Figure 4.4 using a typical eye diagram. For the purpose of calculation the signal level is interpreted as the difference in the mean values, and the noise level is the sum of the standard deviations. The Q-factor is formed by the following ratio:

$$Q = \left(\frac{|\mu_1 - \mu_0|}{\sigma_1 + \sigma_0} \right) \tag{33}$$

where μ_0 and μ_1 are the mean values of the "zeros" and the "ones" , and σ_0 and σ_1 are their standard deviations at the sampling time. The Q factor given here is a unitless quantity expressed as a linear ratio, or it can be expressed in decibels as 20log(Q). The factor of 20 is used to maintain consistency with the linear noise accumulation model.

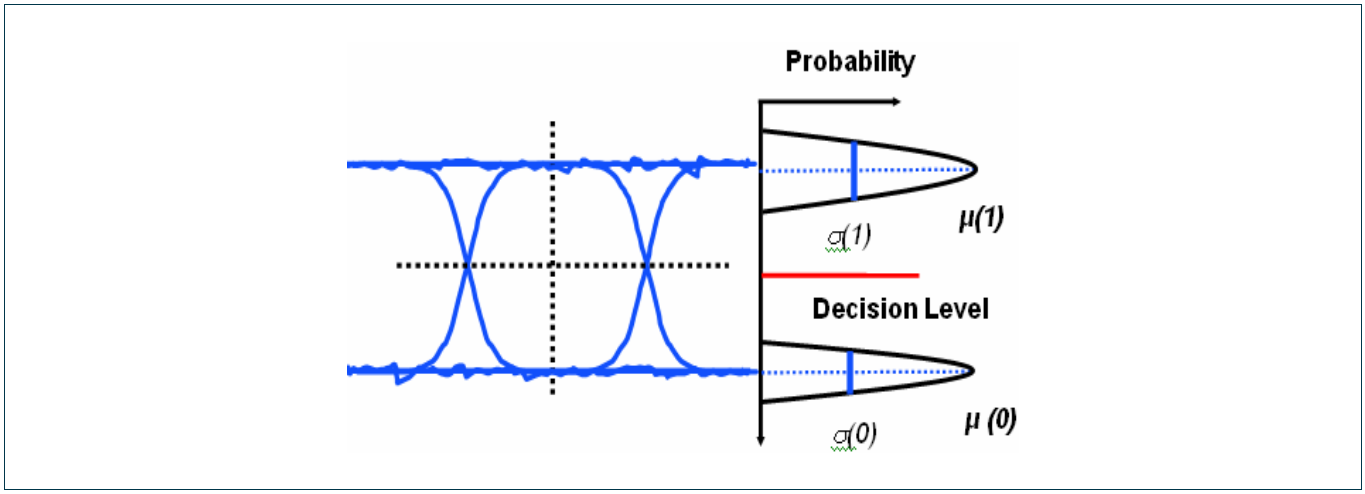

Figure 4.4: A received eye diagram and voltage histogram indicating the parameters that are included in the definition of Q-factor.

The Q factor is related to the system's bit error ratio through the complementary function given by:

$$BER(Q) = \frac{1}{2} erfc\left(\frac{Q}{\sqrt{2}} \right) = \frac{1}{\sqrt{2\pi}} \int_q^{\infty} e^{\frac{-a^2}{2}} da \tag{34}$$

A useful approximation for converting BER back into the Q-factor is given by:

$$Q \cong t - \left[\frac{2.307 + 0.2706t}{1 + t\left(0.9923 + 0.0448t\right)} \right] \quad where\ t = \sqrt{-2\log_e(BER)} \tag{35}$$

4.1.4 Methods of impairments suppression

To overcome the problems caused by the impairments at the physical layer, dynamic impairment management techniques may be implemented in-line (e.g. optical means of impairment compensation) or at the optical transponder interfaces (e.g. electronic mitigation of impairments). From the network layer view, the implementation of certain RWA algorithms that consider signal impairments and constrain the routing of wavelength channels according to the physical characteristics of the optical network paths can further improve the performance and minimize the blocking probability of connection requests. These algorithms are reported in the literature as Impairment Constraint Based Routing (ICBR) algorithms and ensure that connections are feasible to be established considering not only the network conditions (connectivity, capacity availability etc) but also the equally important physical performance of the connections. In section 6.1, two impairment constraint routing approaches are developed and demonstrated.

4.2 User requirements in grids

Two types of QoS attributes can be distinguished: those based on the quantitative, and those based on the qualitative characteristics of the Grid infrastructure. Qualitative characteristics refer to aspects such as service reliability and user satisfaction. Quantitative characteristics refer to aspects such as network latency, CPU performance, or storage capacity. Although qualitative characteristics are important, it is difficult to measure these objectively. Our focus is primarily on quantitative characteristics.

The quantitative requirements can be further distinguished to strict and to non-strict. A user with strict requirements requests that a task is scheduled in the Grid only if all of these (strict) requirements are satisfied, or else the task should not be scheduled. On the other hand a user with non-strict requirements requests only a best-effort performance from the Grid. In general a user may have both strict and non-strict QoS requirements.

The following are quantitative requirements for network QoS:

- Delay: the time it takes for a packet to travel from the source (sender) to the destination (receiver),
- Delay jitter: the variation in the delay of packets taking the same route,
- Bandwidth: the rate at which packets are transmitted,
- Packet-loss rate: the rate at which packets are dropped, lost, or corrupted.

Computational QoS requirements can be specified based on how the resource (CPU) is being used – i.e. as a shared or an exclusive access resource (time or space shared). In a time shared approach (more than one user-level application shares one CPU), the application can specify that it requires a certain percentage access

to the CPU over a particular time period. In space shared approach (one user-level application has exclusive access to one or more CPUs), the application can specify the number of CPUs as a QoS parameter. In space shared approach only one application is allowed to use the CPU for 100% of the time, over a particular time period.

Storage QoS requirements are related to access devices such as primary and secondary disks or other devices such as tapes. In this context, QoS is characterised by bandwidth and storage capacity. Bandwidth is the rate of data transfer between the storage devices and the application program for reading or writing data. Bandwidth dependents on the speed of the bus connecting the application to the storage resource, and the number of such buses that can be used concurrently. Capacity, on the other hand, is the amount of storage space that the application can use for writing data.

In a usual Grid scenario the user requests a minimum end-to-end delay. This is one of the most important user requirements, and it is quite difficult for the Grid network to satisfy it. The end-to-end delay is defined as the time between the task's creation at the user and the time the corresponding output returns to him, after the task is executed in the Grid. The end-to-end delay includes the delay induced by the network for transferring the needed data, by the computational resources for executing the needed tasks and by the storage resources for reading or writing data.

Resource reservation and/or Service Level Agreements (SLA) are mechanisms that can be employed to satisfy the QoS requirements posed by an application user. In this way, the application user can get an assurance that the resource will provide the desired level of QoS. The reservation process can be immediate or undertaken in advance, and the duration of the reservation can be definite (for a defined period of time) or indefinite (from a specified start time until the completion of the application).

In a Grid environment the user forwards these QoS requirements to a scheduler whose role is to process them in order to find if their satisfaction is possible. This process includes a number of algorithms for selecting a suitable computation site for the execution of the task, for selecting a feasible path over which to route the task, for coordinating the resources and for utilizing mechanisms that employ in-advance reservations. QoS-aware scheduling algorithms are the topic of Deliverable 5.2. In Section 6.1.2 of this deliverable we propose ways to formulate the Grid user requirements and especially the network requirements in order to use them in Grid job routing algorithms, described in Section 6.1.

5 PHOSPHORUS Network Scenarios

In this section network architectures with respect to routing are briefly described according to [G^{2}MPLS-MODELS] and the PHOSPHORUS network topologies are defined based on [PHOSP-TestBed]

5.1 PHOSPHORUS Network Architectures

Two Control Plane models are identified in the PHOSPHORUS architecture: the overlay and the integrated model. In this section we will briefly discuss the different routing approaches adopted by these models to identify the location of the routing functions in the two models. More detailed information will be provided in deliverable D5.5 “Recommendations for Control Plane design”

5.1.1 Overlay model

In the GMPLS overlay model [IETF-RFC4208], a GMPLS User to Network Interface (UNI) is identified and strict separation of routing information between network layers is operated. The topological view from an Area is limited to the intra-area details, plus some reachability information about far-edge nodes, either statically configured (configuration-based reachability [GMPLS-ARCH]) or derived by some routing interactions (a kind on partial peering reachability [GMPLS-ARCH]). The core-nodes in an Area act as a closed system and the edge-nodes do not participate in the routing protocol instance that runs among the core nodes.

Following this approach in the PHOSPHORUS Overlay model, the Grid layer has Grid and network routing knowledge in order to provide Grid and network resource configuration and monitoring. The Control Plane in these cases acts as an information bearer of network and Grid resources and as a configuration “arm” just for the network service part.

This model is intended to be mainly deployed when most of the computational and service intelligence need to be maintained in the Grid layer for specific middleware design and functional behaviours. The leading role in this case is played by the Grid scheduler, which is the overall responsible for initiation and coordination of the reservation process through the participating Grid sites and the network in between. The Grid topology overlays the network topology, but the Grid scheduler needs to know both in order to send detailed connection requests towards the G^{2}MPLS, e.g. by specifying the ingress and egress network attachment points or possibly the explicit route to follow.

5.1.2 Peer (integrated) model

In the GMPLS peer model, routing advertisements are distributed to the whole network and all the nodes run the same instance of GMPLS Control Plane, even if they have different switching capabilities. This is the native GMPLS deployment model, which in some cases may encounter scalability issues.

In such a framework, the basic construct is the Forwarding Adjacency Label Switched Path (FA-LSP). It is an LSP created either statically or dynamically by one instance of the Control Plane and advertised as a TE link into the same instance of the Control Plane (for the use by upper layers or neighbouring regions). The topological view in a peer model is obtained by a mix of TE-links with different switching capabilities descriptors and dynamical virtual TE-links bound to FA-LSPs.

In the PHOSPHORUS peer model, Grid sites are modelled as special network nodes with specific additional Grid resource information. The resulting topology is flat and integrated with respect to the positioning of the Grid layer against the network layer. The Grid scheduler functionality is still needed to support the many user applications that rely on specific Grid infrastructure but most of the functions are implemented by the Control plane i.e. the full path between the selected Grid job providers is determined by the G^{2}MPLS.

5.2 Network scenarios to evaluate Grid job routing algorithms

The PHOSPHORUS global testbed will consist of multiple local testbeds located in several places in Europe, United States and Canada. For the integration of the whole PHOSPHORUS testbed all local testbeds must be interconnected on the data plane as well as on the control/provisioning plane. The data plane connections will be used to transit user data between Grid resources located in different local testbeds while the control/provisioning plane connections will be used for integration of the control planes (GMPLS, G^{2}MPLS) of local testbeds as well as integration of NRPSes to allow for signalling between them and multi-domain processing of users' requests.

The data plane connectivity will be based on dedicated lightpaths capable of transmitting huge amounts of data (the amounts which will be generated by PHOSPHORUS applications) and will be comprised of switching resources in local testbeds and a set of transmission links between local testbeds. In order the PHOSPHORUS

testbed to allow the demonstration of the project developments, it was decided that the data plane would be built as an optical network with switching capabilities in local testbeds and transparent lightpaths between local testbeds.

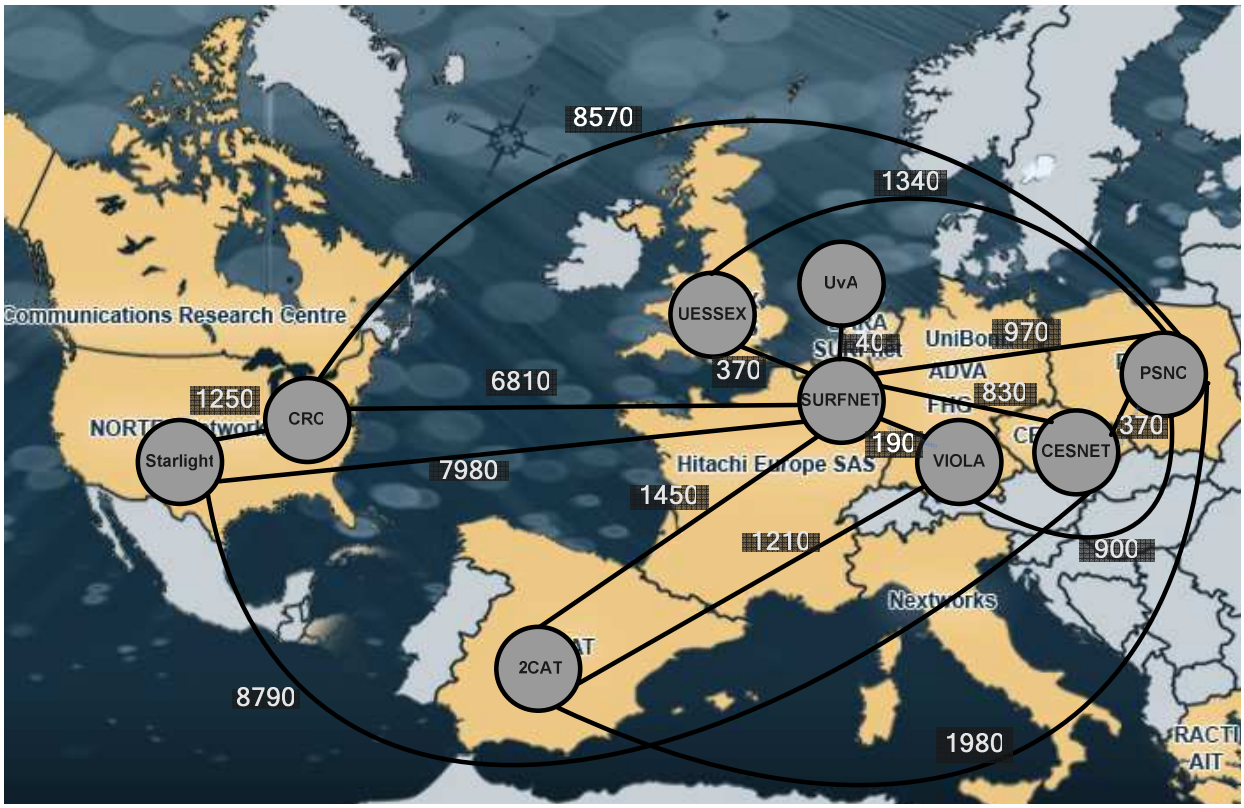

Figure 5.1: PHOSPHORUS Global Testbed

The topology of interconnections between local testbeds is shown in Figure 5.1 for the global network and in Figure 5.2 for the European network as they will be utilized for the simulations that will be presented in this document. The exact global network topology can be found in [PHOSP-TestBed].

The links connecting local testbeds are listed in Table 5-1 along with their estimated distance.

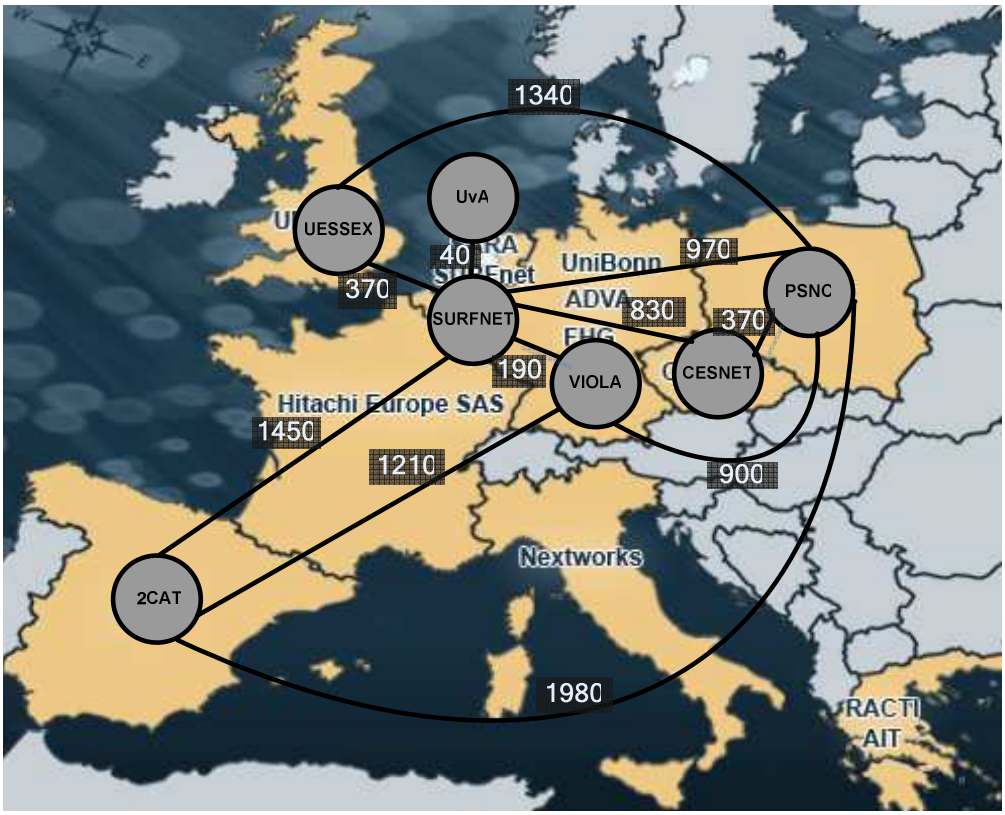

Figure 5.2: PHOSPHORUS European Testbed

Test-beds		Estimated Distance (km)
PSNC	CESNET	370
PSNC	I2CAT	1980
PSNC	UESSEX	1340
PSNC	VIOLA	900
PSNC	SURFnet	970
SURFnet	I2CAT	1450
SURFnet	VIOLA	190
SURFnet	UESSEX	370
SURFnet	UvA	40
SURFnet	CESNET	830
I2CAT	VIOLA	1210

STARLIGHT	CESNET	8790
STARLIGHT	SURFnet	7980
STARLIGHT	CRC	1250
CRC	PSNC	8570

Table 5-1: Network connections for simulations

5.2.1 Multi-domain networks

The network considered in multi-domain routing consists of the core network depicted in Figure 5.3 enlarged with local networks at every core switch. These local networks are generated using the Barabási-Albert (BA) algorithm. The BA algorithm is based on two generic mechanisms [Barabasi99]:

1. **growth**: networks expand continuously by the addition of new vertices.

2. **preferential attachment**: new vertices attach preferentially to sites that are already well connected.

The algorithm works as follows: we start with a small number (m_0) of vertices, at every time step we add a new vertex with $m(\leq m_0)$ edges that links the new vertex to *m* different vertices already present in the network. To incorporate preferential attachment, we assume that the probability Π that a new vertex will be connected to vertex *i* depends on the connectivity k_i of that vertex, so that $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$. After *t* time steps, the model leads to a random network with $t + m_0$ vertices and mt edges. This network evolves into a scale-invariant state with the probability that a vertex has *k* edges, following a power law ($K^{-\gamma}$) with an exponent $\gamma_{\mathrm{mod}el} = 2.9 \pm 0.1$. This algorithm has been proposed in [Barabasi99] to model complex networks like the World Wide Web and the nervous system.

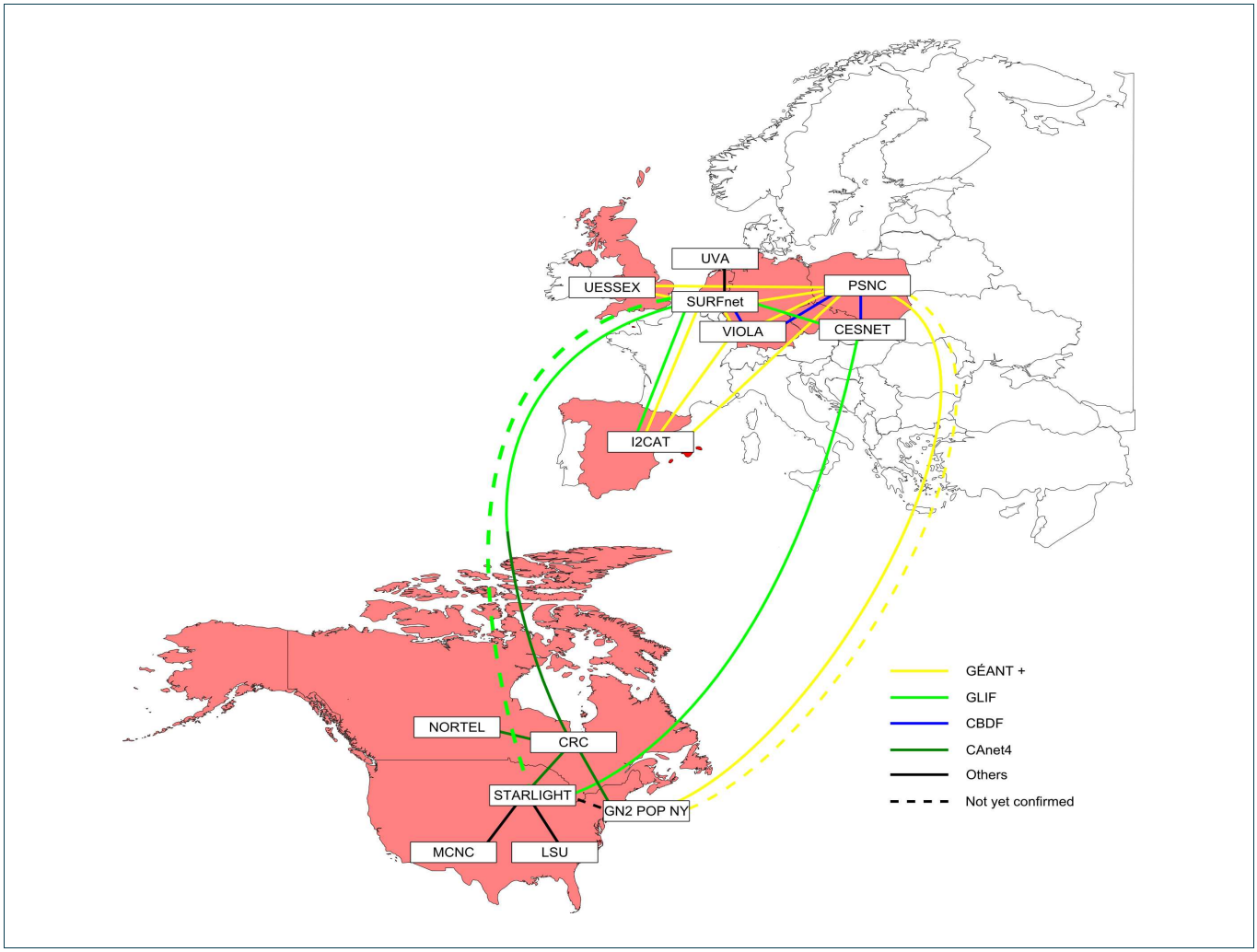

Figure 5.3: PHOSPHORUS Global network extended

5.3 PHOSHORUS link and node architectures and characteristics

A number of links in the PHOSPHORUS testbeds follow the structure presented in Figure 5.4 where inline amplifiers are positioned after each fiber span to compensate for fibre losses.

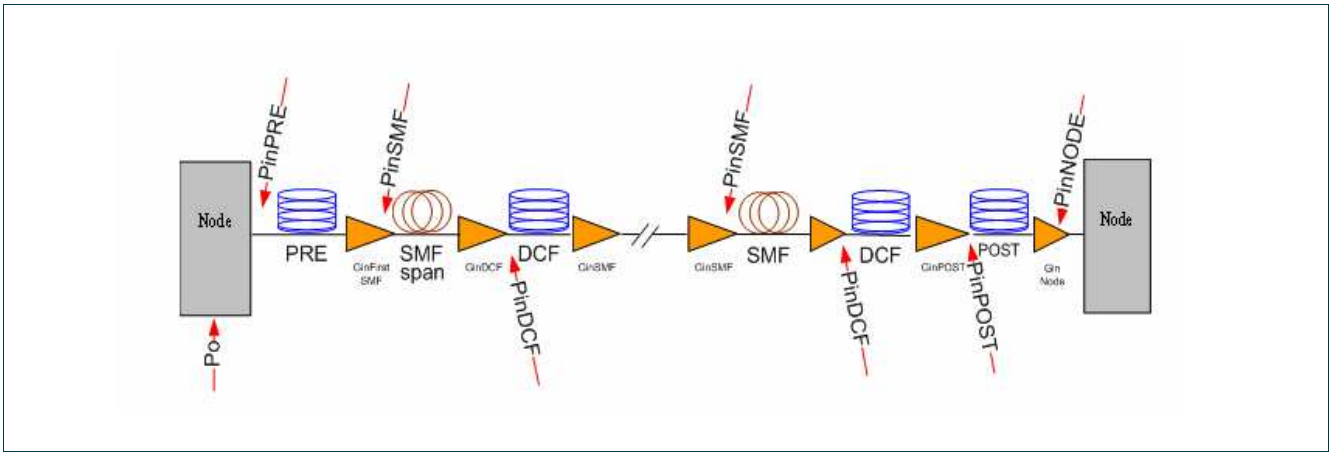

Figure 5.4: Link architecture used for the simulations

The link consists of a number of spans (Single Mode Fiber, SMF) according to the link length where each span is followed by a dispersion compensation fiber (DCF) used to compensate at the requested degree the chromatic dispersion introduced by the SMF. At the beginning of the link a pre-compensation fiber (PRE) is used to insert initial chromatic dispersion whereas at the end, a post-compensation fiber (POST) is used to collect the chromatic dispersion evolved through the link. The inline amplifiers of the link are used to compensate the losses of each fiber segment and boost the power to the appropriate levels at the entrance of each segment. The parameters concerning the two fiber types applied to the link (DCF and SMF) are reported in Table 5-2

Parameters	SMF	DCF
Attenuation a (dB/km)	0.25	0.5
Nonlinear index coefficient n (m^2/W)	$2.6*10^{-20}$	$3.5*10^{-20}$
Chromatic Dispersion Parameter D(s/m^2)	$17*10^{-6}$	$-80*10^{-6}$
Dispersion Slope $dD/d\lambda$(s/m)	$0.085*10^{3}$	$-0.3*10^{-3}$
Effective Area A_{eff}(m^2)	$65*10^{-12}$	$22*10^{-12}$

Table 5-2: Fiber Characteristics

6 Enhanced Grid Job Routing Algorithms for Optimum Path Computation

In this section Grid job routing algorithms that consider network and Grid requirements providing optimum path discovery and efficient resource utilization are introduced and analyzed. A set of simulation results that are produced based on these algorithms that are applied on the PHOSPHORUS network topology are presented. In addition, an optimal architecture offering anycast routing in multi-domain Grid networks is proposed and studied through simulations to demonstrate the flexibility and scalability of the proposed solution.

6.1 Optimal routing considering network and Grid requirements/constraints

Physical layer characteristics and Grid user requirements are important parameters that should be considered in the routing calculation procedure to achieve optimum routing performance. The incorporation of these parameters in the routing algorithms is described below.

6.1.1 Physical layer impairments

As described in section 3.4 two methods for considering physical layer impairments into the routing process appear in the literature. In the first one, usually the RWA algorithm is treated in two steps: first a lightpath computation in a network layer module is provided, followed by a lightpath verification performed by the physical layer module. The other method integrates the physical layer impairments into the routing process and therefore the lightpath computation is performed according to the optical parameters.

Following the latter approach two different techniques for introducing physical layer parameters into the routing algorithm have been developed and are analyzed in the following sections. The first technique considers linear impairments individually into its routing process using a set of discrete criteria, each corresponding to a specific physical impairment, that have to be simultaneously satisfied for a connection to be possible to be established. The alternative technique combines a wide range of linear and nonlinear impairments under a unified performance parameter and performs routing based on the value of this parameter. If for any discovered path

the value of this parameter is not above a certain predefined threshold the connection is not feasible to be established.

6.1.1.1 *Network design with Individual impairment consideration*

For the following algorithm description a fixed network topology is represented by a connected, simple graph $G=(V,A)$. V is the set of generalized nodes; each node is a terminal station and can perform routing capabilities as well. Whenever specified, nodes may also be equipped with wavelength conversion capabilities. A denotes the set of (point-to-point) single-fiber links; they can be directional, bidirectional (consisting of two opposite directional links) or unidirectional (both directions are available, but all different data streams must occupy different wavelengths no matter their direction). Each fiber is able to support a common set C of W distinct wavelengths; $C = \{1, 2, \ldots, W\}$. The static version of RWA defines an a priori known traffic scenario; this is given in the form of a matrix of nonnegative integers **R**, called the traffic matrix. Then, $\boldsymbol{R}(s, d)$ or $\boldsymbol{R}_{sd}$ denotes the number of requested connections (lightpaths to be established) from source-node s to destination-node d.

The algorithm

The algorithm given a specific RWA instance; that is, a fixed network topology, its nodes' and links' characteristics and a static traffic scenario, returns the instance solution in form of routed lightpaths and assigned wavelengths. More specifically, the types of input and output are:

Inputs:

- The considered network topology, represented by graph $G=(V,A)$. Let $|V|=N$ and $|A|=L$.
- A detailed description of the conversion capabilities network nodes are equipped with.
- The type of network links; i.e., directional, unidirectional or bidirectional.
- The number of available wavelengths, W.
- An $n \times n$-dimensional traffic matrix of nonnegative integers, **R**.
- A positive integer k, denoting the number of candidate paths to serve each requested connection.

Outputs:

- A solution of the considered RWA instance, if such exists, in form of routed lightpaths and assigned wavelengths.
- The optimization objective value. In case of infeasibility, a negative value is returned.
- The throughput of the process; that is, the fraction of the requested connections that are established. Alternatively, the blocking probability of the process, equal to 1−throughput.

The algorithm consists of three phases. The first (pre-processing) phase computes a set of candidate paths to route the set of the requested connections. The second phase utilizes the Simplex algorithm to solve the Linear Program (LP) that formulates the given RWA instance. The third phase, finally, handles the infeasible instances, in order to establish some (since all is impossible) of the requested connections.

Phase 1: In this phase, *k* candidate paths to serve each requested connection are identified. These are selected as short as possible and also in a way to share as less links as possible, for achieving lower link congestion. Given a requested connection between source-node *s* and destination-node *d*, the *k*-paths selection process is as follows: Initially, all network links are assigned unit cost values. Then the minimum-cost path (and obviously the shortest one) is identified using Dijkstra's algorithm. Once this path is identified, a two unit cost value (i.e., twice the value of the initial cost) is added to all of its links and the new minimum-cost path is computed. Similarly, a four unit cost value (i.e., twice the value of the previously added cost) is added to all links of the newly computed path. This process iterates, until *k* different paths are identified. Each iteration selects the shortest possible path, by also favouring links not often previously used. Then a cost value is added to its links (twice the cost value added in last iteration), decreasing the probability of them being selected in future iterations. This process can easily be generalized to include weighted links. After a subset P_{sd} of candidate paths for each commodity pair *s*−*d* is computed, the total set of computed paths, $P = \bigcup_{s-d} P_{sd}$, is inserted to the next phase. If *k* candidate paths are not available to serve a requested connection, it can be identified within at most *D*−1 Dijkstra's iterations, where *D* is the diameter of the network. The time complexity of the pre-processing phase is clearly polynomial.

Phase 2: Taking into account the network characteristics (topology, type of links, node conversion capabilities and number of available wavelengths), the traffic scenario and the set of paths identified in phase one, Phase 2 formulates the given RWA instance as an LP. The LP formulation used is presented in 9Appendix AAppendix A. This LP is solved using Simplex algorithm that is generally considered efficient for the great majority of all possible inputs. If the instance is feasible, the algorithm terminates by returning the optimized value of the LP objective and the feasible solution that achieves this optimization, in the form of routed lightpaths and assigned wavelengths. Feasible solutions exist, iff all the requested connections are able to be established concurrently using the available number of wavelengths; in that case, the process is said to have throughput equal to 1. Otherwise, this phase returns 'infeasible' (a negative cost function value) and the algorithm proceeds to the third phase.

Phase 3: This phase is utilized, when the considered LP instance is infeasible. Infeasibility is overcome by iteratively increasing the number of available wavelengths by 1 and re-executing Phase 2, until a feasible solution is found and all requested connections are able to be concurrently established. This process may render the whole algorithm inefficient; thus, a constraint on the number of LP iterations has to be set. However, if the initial value of *W* is chosen realistically big with respect to the number of the requested connections, only a few additional LP executions suffice for a feasible solution almost surely to be found. Let *C'* = {1, 2, . . . ,*W'*} be the minimum sufficient set of wavelengths, in order all the requested connections to be concurrently established. The resulting RWA solution must be converted to a final one that uses only *W* wavelengths; therefore, *W'*−*W* wavelengths must be removed and the lightpaths occupying them have to be blocked. The removed wavelengths are those occupied by the minimum number of lightpaths, in order the minimum number of requested connections to be blocked. The algorithm terminates and outputs the routed lightpaths and assigned wavelengths, along with the throughput of the process, which is a fractional value between 0 and 1.

The flow cost function

The exact LP formulation is provided in Appendix A, and the flow cost function F_l that is used to express the amount of congestion arising on each network link, given a specific routing of the requested connections is

analyzed in this section. Let *c* be the number of lightpaths crossing (or number of wavelengths occupied by) link *l*. In terms of our formulation, that is $c = \sum_p \sum_c \lambda^c{}_{pl}$ and D_l is a properly increasing function of *c*. D_l is also chosen to be convex (instead of linear), implying thus a greater amount of 'undesirability', when a single link becomes highly congested.

We utilize the following flow cost function:

$$D_l(c) = \frac{c}{W+1-c} \tag{36}$$

Obviously, when $c = 0$, also $D_l = 0$; thus, empty links are assigned zero flow cost function values. When a link becomes totally congested ($c = W$), D_l obtains its maximum value *W*. In addition, D_l is more suddenly increasing at higher levels of congestion. For example, when a link is one lightpath left to be fully congested ($c = W - 1$), D_l is less than half of its maximum ($D_l = \frac{W-1}{2}$). By simply adding one flow unit, D_l is over-doubled and reaches its maximum. The argument is that we prefer, in terms of network performance, few low-congested links to be added one flow unit, than a single link to be totally congested, since in the latter case, a significant number of candidate paths is probably blocked and routing options are limited, while in the former, space for future lightpaths is left.

The above (nonlinear) function is inserted to the LP in the approximate form of a piecewise linear function; i.e., a continuous non-smooth function, that consists of *W* consecutive linear parts. The piecewise linear function is constructed as follows: Set $i = 1, \ldots, W$ and begin with $D_l(0)=0$. Then,

$$D_l^i(c) = a_i c + \beta_i \ , \ i-1 \le c \le i, \tag{37}$$

where $\alpha_i = D_l(i) - D_l(i-1)$ and $\beta_i = (1-i)D_l(i) + i\,D_l(i-1)$. Observe that the piecewise linear function is exactly equal to the corresponding D_l for each of their integral arguments ($c = 0, 1, \ldots, W$) and greater in any other (fractional argument) case. Inserting a sum of such piecewise linear functions to the LP objective, therefore, results in the identification of integer optimal solutions by Simplex, since the vertices of the polytope constructed by the constraints set tend to correspond to the corner points of each piecewise linear function and thus consist also of integer components. Clearly, the LP must include a constraint for each of those *W* linear parts; that is, for every link *l* and also for every *I*,

$$D_l \ge \alpha_i \sum_p \sum_c \lambda^c{}_{pl} + \beta_i \tag{38}$$

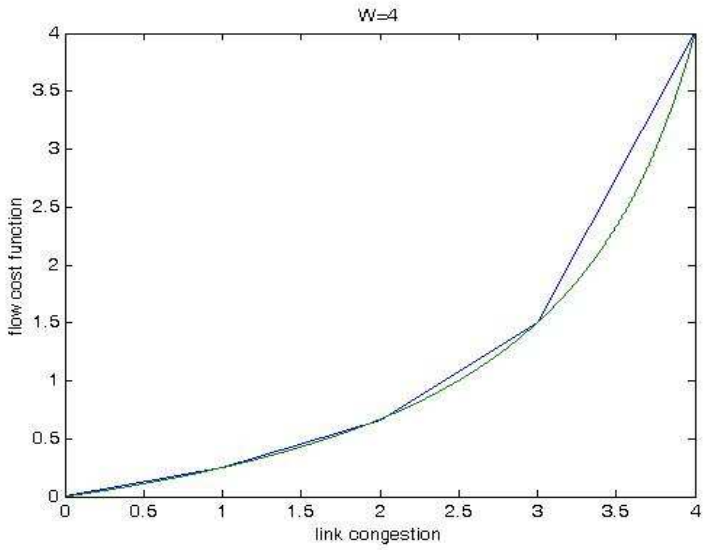

Figure 6.1: The flow cost function (curve line) and the corresponding piecewise linear function, in case W = 4.

Impairment-constraint based RWA experiments

In impairment-constraint based RWA, all lightpaths that do not satisfy any of the considered impairment constraints have to be removed from the final RWA solution. In the experiments presented here we have considered the following linear impairments: Amplified Spontaneous Emission (ASE), Chromatic Dispersion (CD), Polarization Mode Dispersion (PMD), as referred in section 4.1.1. These impairments can easily be handled by assigning an impairment-induced cost to each link, linear with respect to its length, and posing the corresponding upper bound on the acceptable impairment- induced total cost of each candidate path to serve a requested connection. The cost value assigned to fiber l of length d_l equals to

- $(D_{PMD})^2 d_l$, when considering PMD;

- the number of optical amplifiers placed on the fiber, when considering ASE noise;

- $D_{CD} d_l$, when considering CD.

Thus, the objective of the impairment-constraint based RWA process is then reduced to discarding all paths with unacceptably high impairment-induced costs and routing appropriately the requests using the remaining subset of candidate paths. This is easily implemented in the pre-processing phase of our algorithm by assigning impairment-induced (instead of unit) costs to network links and truncating all the unacceptable candidate paths.

We executed a great amount of experiments, in order to obtain comparative results of network performance under various network and impairment parameters. The network topology used for our simulations was the European PHOSPHORUS testbed presented in Figure 5.2 that consists of 7 nodes and 11 bidirectional links, whose lengths range between 40 km and 1980 km (average distance 877 km). Network performance was

measured through the use of the average blocking probability of 100 RWA executions corresponding to different random static traffic instances of a given traffic load. *W* was fixed to 4 and *k* was equal to 3; these values suffice for almost all RWA random instances of traffic load 0.2 and in absence of impairment constraints to be executed with 100% throughput. The blocking probability presented in the following sections is the average blocking probability of the experiments in which the solution of the proposed LP formulation was integer. Thus, the blocking probability is affected only by the physical impairments of the network. Finally, no wavelength conversion was considered to be available. All experiments were executed in MATLAB. For LP-solving, the GLPK-4.8 MATLAB library [glpk] was utilized.

PMD studies

In our studies, all network fibers are considered of the same type. RWA simulations were executed for D_{PMD} values ranged in 0.05−0.6 $\mathrm{ps}/\sqrt{\mathrm{km}}$ (small values correspond to newer fibers, while big values correspond to older ones), bit rates of *B* 10, 20, 30 and 40 Gbps and traffic loads of 10, 20, 30 and 40 percent of the number of the total possible connections. Figure 6.2 shows, that the PMD effect seems negligible at 10 Gbps for fibers with D_{PMD} values less than 0.25 $\mathrm{ps}/\sqrt{\mathrm{km}}$ (which is a typical value for modern fibers). However, at higher bit rates the situation is much worsened, especially when dealing with older fibers. At bit rates of 20, 30 and 40 Gbps, the corresponding curves rise (where the network's throughput starts to divert from 100%) at the critical D_{PMD} values of 0.15, 0.1 and 0.1 $\mathrm{ps}/\sqrt{\mathrm{km}}$, respectively. This is explained as follows: these *B*-D_{PMD} pairs lead to the same approximate maximum acceptable (due to PMD) path length of 625 km and there are only few paths in the topology under consideration that satisfy this constraint. At bit rates of 30 and 40 Gbps, the curves almost reach the exceptionable blocking probability value of 100% at the critical PMD parameter values of 0.55 and 0.4 $\mathrm{ps}/\sqrt{\mathrm{km}}$, respectively; these *B*-D_{PMD} pairs lead to the same approximate maximum acceptable path length of 39 km which is smaller than the minimum link length of 40km (Figure 5.2). Notice that the blocking probability value of each bit rate curve is close to zero when the PMD parameter equals to 0.1 $\mathrm{ps}/\sqrt{\mathrm{km}}$, and zero for 0.05 $\mathrm{ps}/\sqrt{\mathrm{km}}$, denoting that newer fibers are able to compensate totally the PMD effect in our experimental network, even when utilizing bit rates of 40 Gbps.

ASE noise studies

In our studies, all network fibers are considered to be of the same type; thus, they are characterized by the same attenuation coefficient value *a*. Assume, also, that optical amplifiers of the same type (characterized by common gains *G* and common noise figures) are placed on networks' nodes and fibers in the following uniform way: one optical amplifier is placed on each network node and one optical amplifier is placed on each (*G*/*a*) kilometres of fiber. The values used are P_{avg}=4 dBm (equal to 2.5 mW), *a* = 0.25 dB/km and hvB_0 = −58 dBm. We executed a great amount of RWA simulations for n_{sp} values ranged at 1.5−5 (corresponding to amplifiers of various noise figures), amplifier gain values of 17.5, 20, 22.5 and 25 dB and traffic loads of 10, 20, 30 and 40 percent of the number of the total possible connections. We considered both scenarios where FEC was not used (corresponding to SNR_{min}=25 dB) and FEC was used (corresponding to SNR_{min}=20 dB).

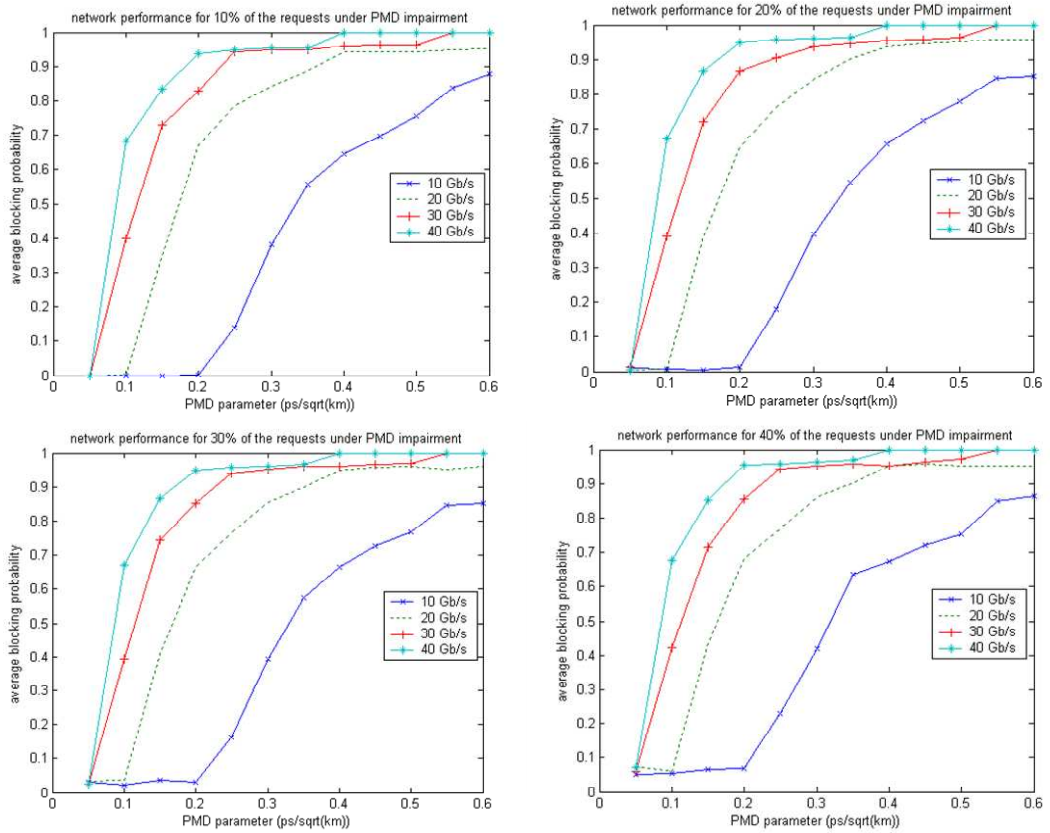

Figure 6.2: PHOSPHORUS testbed performance under PMD impairment for various loads.

Currently, network designers utilize two types of amplifiers: semiconductor optical amplifiers (SOAs) and erbium doped fiber amplifiers (EDFAs). SOAs are cheaper and smaller devices than EDFAs; however, they are characterized by bigger noise figure and thus bigger n_{sp} values. Typical SOAs are characterized by n_{sp} values up to slightly less than 7, while typical EDFAs are characterized by n_{sp} values in the range between 1.5 and 3.

Considering Figure 6.2 and Figure 6.4, the ASE noise effect seems negligible at G=17.5 dB with FEC, while when FEC is not used the ASE noise effect becomes an issue even at G=17.5 dB. Generally, the performance of the network is not satisfactory when FEC is not used and for this reason we won't analyze it further here. For G=17.5 dB with FEC, the performance curve rises at n_{sp} = 4, denoting that the network's throughput diverts from 100% in presence of 35 or more such SOAs per path. By increasing G to 20 dB the curve rises at n_{sp}=3, indicating that an average placement of 26 or more SOAs of n_{sp}=3 per path leads to positive blocking probability, and rises up to 40% for n_{sp} = 5 that corresponds to an average placement of 16 such SOAs per path. At G=22.5 dB and FEC, the curve rises almost from the beginning (at n_{sp}=2), indicating that an average placement of 22 or more such SOAs per lightpath is required to yield to a non zero blocking probability. At G=25 dB and FEC, more than 20% of the requested connections are blocked, even in presence of a sufficiently large number of wavelengths and SOAs with optimal noise figures (for n_{sp} =1.5 we have 16 amplifiers), and 12

amplifiers of n_{sp} = 2 per path is sufficient to cause a blocking probability greater than 45%. The blocking probability curve reaches its maximum value of 75%, when any amplifier of at least n_{sp}=4 is placed on each path.

Notice that when utilizing FEC techniques with EDFAs (instead of SOAs) with noise figures that approach the fundamental quantum limit we are able to utilize the gain of 20 dB, without causing any degradation to our network's performance.

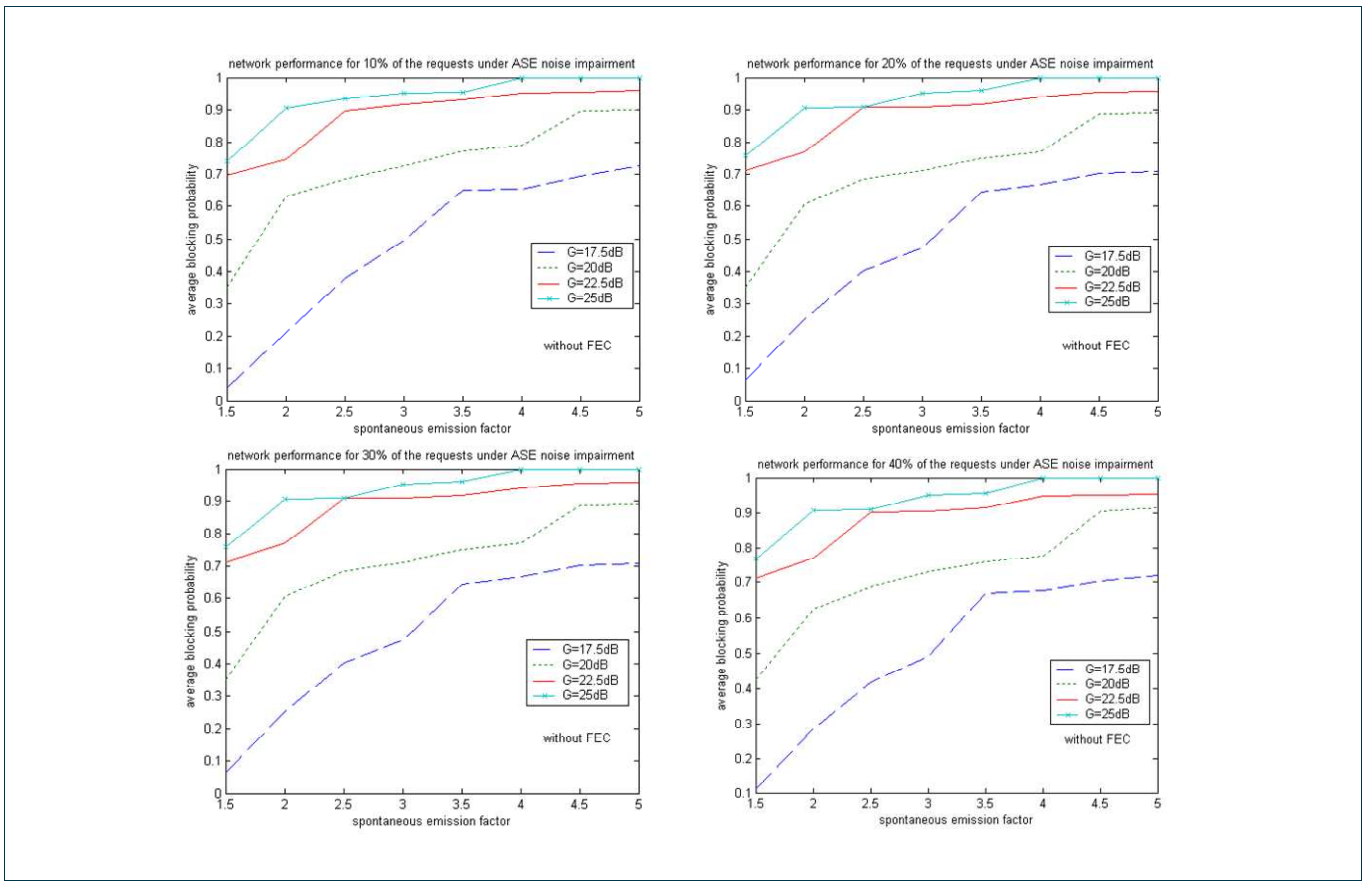

Figure 6.3: PHOSPHORUS testbed performance under ASE noise impairment for various loads. No FEC is used.

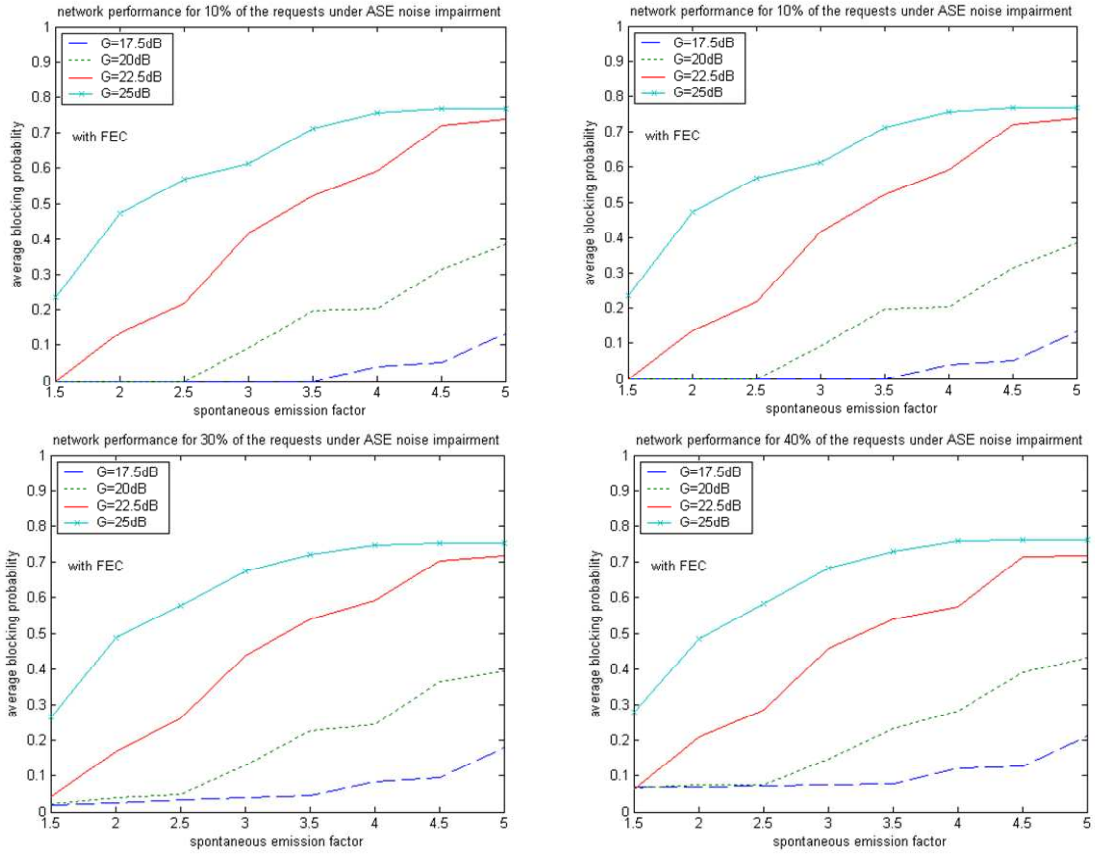

Figure 6.4 :PHOSPHORUS testbed performance under ASE noise impairment for various loads. FEC is used.

CD studies

In our experiments, all network fibers are characterized by the same D_{CD} value. B is considered constant and equal to 10 Gbps. We executed a great amount of RWA simulations for D_{CD} values ranged at 5−20 ps/(nm·km) (small values correspond to newer fibers, while big values correspond to older ones) and traffic loads of 10, 20, 30 and 40 percent of the number of the total possible connections. Both cases of LPF and NRZ modulation format used were considered. In our experiments, we assumed that on each network node there was a DCM that compensated 95% of the CD introduced by the previous fiber link.

In cases where DCM is not used a blocking probability close to 100% is obtained. It is experimentally shown that for B=10Gbps and D_{CD} = 17 ps/(nm·km), the threshold of 2 dB appears at about 50km if NRZ modulation format is used and at about 140km if LPF modulation format is used. However, since the links of the network under consideration are long, CD impairment would block almost all connections. For that reason, current day networks of speeds up to 40 Gbps require compensation techniques to be utilized, an assumption that we consequently also made in our simulation experiments.

As Figure 6.5 indicates, when the LPF modulation format is used the CD effect becomes negligible and the throughput of the network doesn't degrade even for high traffic loads. The network performance degrades at large D_{CD} values and when NRZ modulation format (instead of LPF) is used. With NRZ modulation format, the blocking probability curve rises at about 7 ps/(nm·km) and more than 30% of the requests have to be blocked when the fibers have 13 ps/(nm·km). The blocking probability may even reach the exceptionable value of 60% at D_{CD} =20 ps/(nm·km); Thus, LPF modulation format is essential for optimizing the network's performance under the CD impairment.

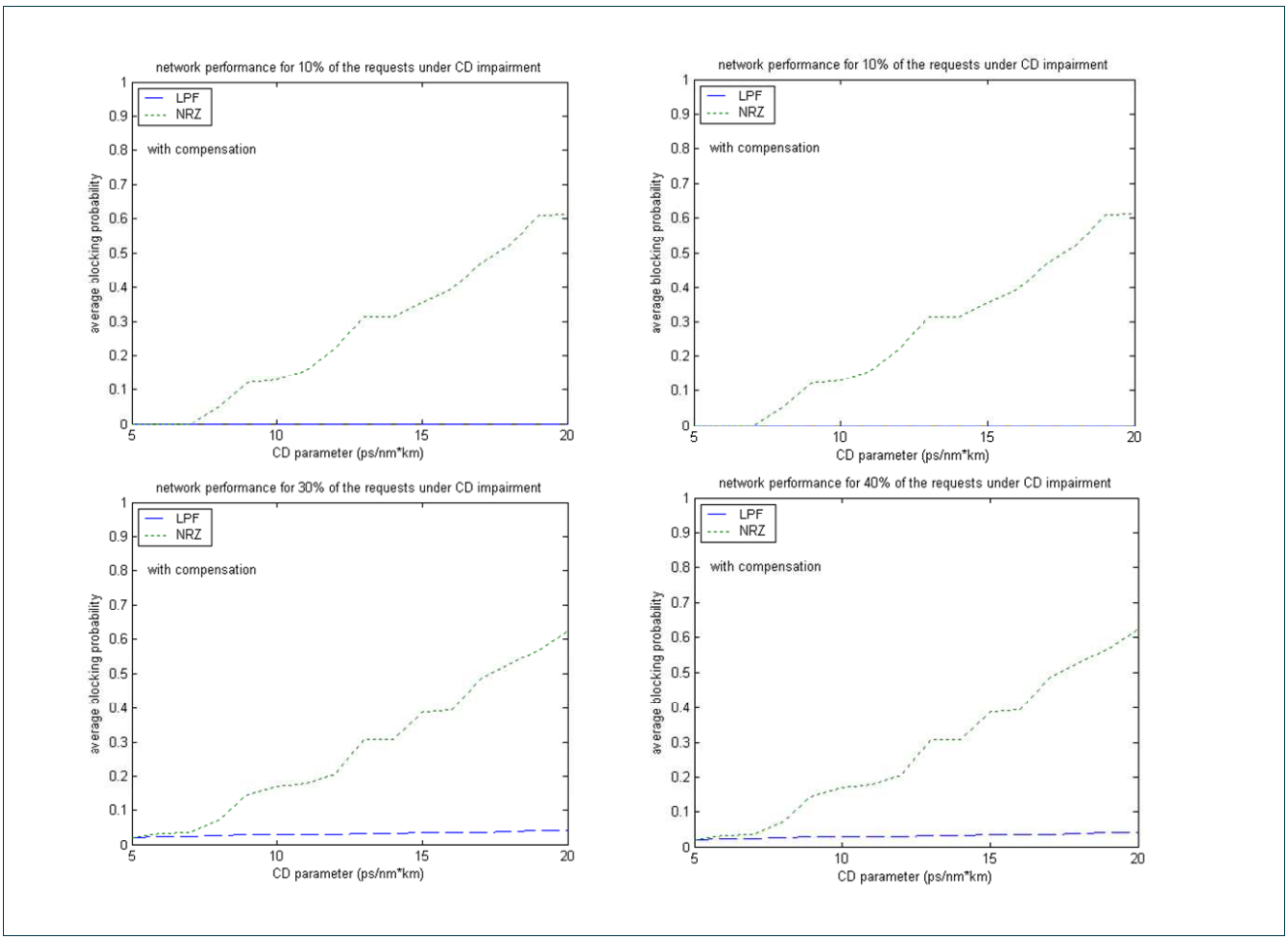

Figure 6.5 : PHOSPHORUS testbed performance under CD noise impairment for various loads. DCMs are used.

Similar experiments were performed using the NSFnet topology that consists of 14 nodes and 21 unidirectional links. The general conclusions for the NSF topology are similar to the ones presented here. Further details on the simulation setup and results is reported in [Iakoum07].

6.1.1.2 *Network design combining impairments*

Since impairments cause deterioration of the quality of the signal, as it propagates through the links, in this part of the simulations we relate each impairment to a corresponding link cost. A higher cost for a link implies more severe signal degradation due to the particular impairment, while the signal traverses through the link. On the other hand, a smaller link cost indicates that the link is more immune to this impairment and favouring thus the routing through this link. The link costs are assigned to be the Q-factor penalties caused by the corresponding link impairments and will constitute the criteria for the routing procedure as described in this section. Figure 6.6 presents the flowchart of the IA-RWA scheme [Markidis07] which can be separated into three phases.

Initially the pre-processing phase collects all the information related to the network and the traffic demands. Information such as the topology of the network, the link capacities, the fiber characteristics like the PMD parameters, dispersion map applied (pre, post and inline dispersion compensation), span lengths, attenuation of each span, the launched powers at each fiber segments, the noise figure of the amplifiers, the nodes architecture, the channel spacing and the link capacities are required by the algorithm for the physical impairments evaluation. Moreover information concerning the number of requests, the bit rate and the source destination pairs are required to identify the traffic demands.

The pre-processing phase also assigns costs to the links based on the above parameters. Q-factor penalties due to Chromatic Dispersion, Amplified Spontaneous Emission, Self Phase Modulation, Four Wave Mixing and Cross-Phase Modulation are calculated as explained in section 4.1 and are assigned as costs on each link of the network according to:

$$Q_k = \frac{pen_k \cdot P}{\sqrt{\sigma^2_{ASE,k} + \sigma^2_{crosstalk,k} + \sigma^2_{XPM,k} + \sigma^2_{FWM,k}}} \quad (39)$$

where pen_k is the relative eye closure penalty caused by optical filtering and SPM/GVD phenomena on link k, $\sigma^2_{XPM,k}$ and $\sigma^2_{FWM,k}$ are the electrical variances of the XPM and FWM induced degradations and finally $\sigma^2_{ASE,k}$ and $\sigma^2_{crosstalk.k}$ are the electrical variances of the ASE noise and the generated crosstalk.These costs are assigned as weights to the links before the Dijkstra algorithm which is a part of the RWA phase starts calculating paths.

As illustrated in the flowchart the RWA phase is initiated once the link costs have been found. This phase assigns paths (incorporating the physical impairments as weight of the links) and wavelengths to all the demands. Conventional RWA modules can be used for this phase e.g. shortest path or minimum hop by applying proper link costs. The RWA problem is treated as a single optimization problem to properly identify the area in which the optimal solution lies in. The exact formulation of the RWA problem used to assign the most suitable paths and wavelength to the connection requests is described in detailed in Appendix A.

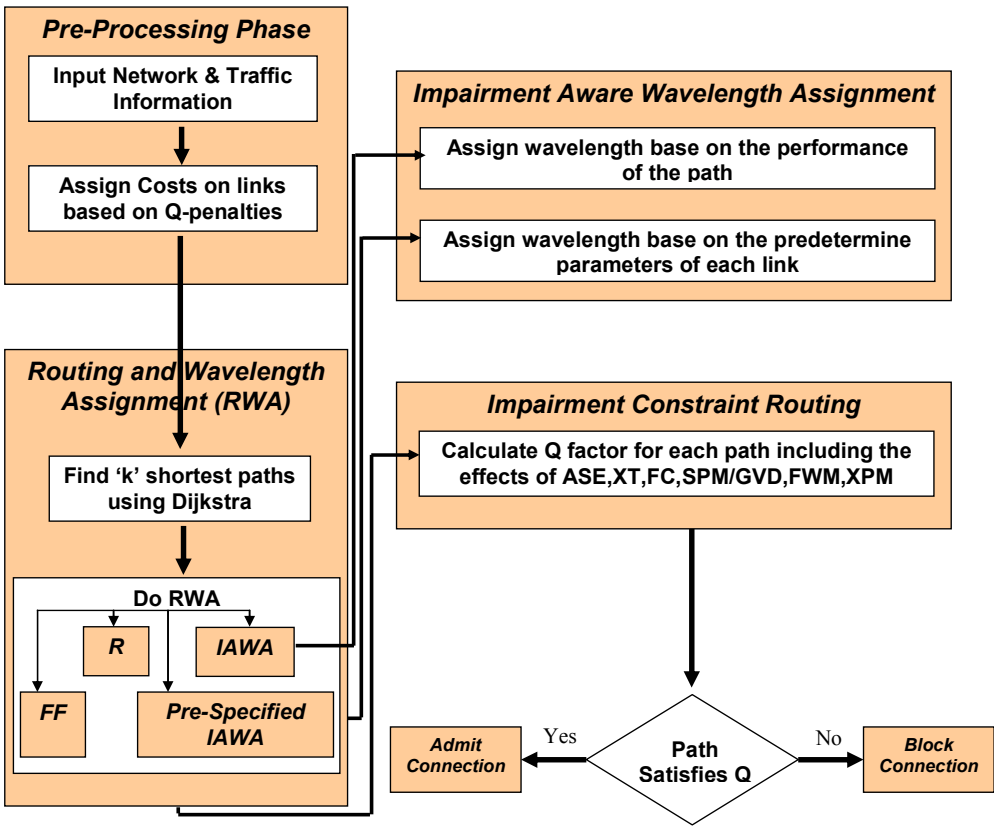

Figure 6.6: Impairment Aware Routing and Wavelength Algorithm flow chart

In order to reduce the computational complexity and simplify the optimization, for each connection request, a set of k-paths is identified that are as diverse as possible and are fed as input in the "Do RWA" module. In this way, the algorithm acquires the flexibility to select the optimum path among the k input paths that ensures minimization of the flow cost function of each link which is given by:

$$D_l(c) = \frac{c}{W + 1 - c} \tag{40}$$

where c is the number of used wavelengths in link l and W is the total number of wavelengths.

If the RWA formulation is feasible it specifies the paths that should be established and the minimum number of wavelengths required to carry the request. The lightpath establishment that follows is based on the selected wavelength assignment scheme. This may include either conventional strategies that are unaware of the physical performance status of the connection, (such as first fit (FF), and random fit (RF)) or strategies that take into consideration the corresponding impairments. For the later case two different schemes were examined, a direct implementation of the Impairment Aware Wavelength Assignment-(IAWA), as well as, a Pre-Specified (IAWA) scheme.

In the Impairment Aware Wavelength Assignment (IAWA) scheme, the lightpaths are established according to their Q-factor performance. More specifically, each potential lightpath, among those available on the path, is characterized in terms of Q-factor taking into consideration the already established wavelength connections. The one having the optimum performance is finally selected.

The Pre-Specified Impairment Aware Wavelength Assignment scheme (PS-IAWA) offers more advanced performance in terms of computational efficiency since according to this scheme, and prior to the wavelength assignment process, all the wavelength locations on per link basis are characterized and ordered in terms of their Q-factor value. Then the algorithm defines the paths and for each one of them the space of the common wavelengths that are available across its links. The final selection will be made based on the pre-specified order created at the beginning.

After the wavelength assignment is completed the control is transferred to the Impairment Constraint Based Routing module which verifies the Q-factor constraint considering all the physical impairments involved across the path. This module evaluates the Q-factor at the end of the route for the path designated as the best from the k candidates that satisfy the specific request. In [Cantrell03] a formula that separates the Q impairment into a product of eye and noise is proposed, and the ICBR module employs it accordingly for all the impairments presented in section 4.1:

$$Q_{end} = \left(\frac{\langle I_1 \rangle_{end} - \langle I_0 \rangle_{end}}{\sigma_{1,end} + \sigma_{0,end}} \right) = \left(\frac{\langle I_1 \rangle_{end} - \langle I_0 \rangle_{end}}{\langle I_1 \rangle_{start} - \langle I_0 \rangle_{start}} \right) \times \left(\frac{\sigma_{1,start} + \sigma_{0,start}}{\sigma_{1,end} + \sigma_{0,end}} \right) \times Q_{start}$$
$$= (Eye\,impairments) \times (Noise\,impairments) \times Q_{start} \qquad (41)$$

where eye impairments include PMD, SPM – GVD,FC and noise impairments consist of ASE, XT , FWM and XPM.

A path is accepted when the Q-factor value at the destination node is higher than 11.6dB, which corresponds to a BER of 10^{-15} after forward error correction (FEC) is utilized at 10Gbps and the connection is established, in any other case the path is rejected and the connection is blocked.

Simulation results based on the PHOSPHORUS European scenario

In this section the Impairment Aware Routing and Wavelength Assignment algorithm is deployed in the PHOSPHORUS European topology illustrated in Figure 5.2 and the simulation results are further discussed.

The PHOSPHORUS European testbed topology consists of 7 nodes corresponding to the PHOSPHORUS local testsbeds distributed in 7 European countries and are interconnected with 11 bidirectional links. Figure 6.7 (right) presents the link length distribution where it is noticed that the link lengths cover a range from 40 to 2000km, with the average link length being 874km. In the majority of the simulations it is assumed that there are connection requests between every possible pair of nodes and therefore 21 end-to-end connections are requested to be established. The number of nodes that participate in these 21 connections is presented in Figure 6.7 (left) when ICBR is used for the discovery of the routes.

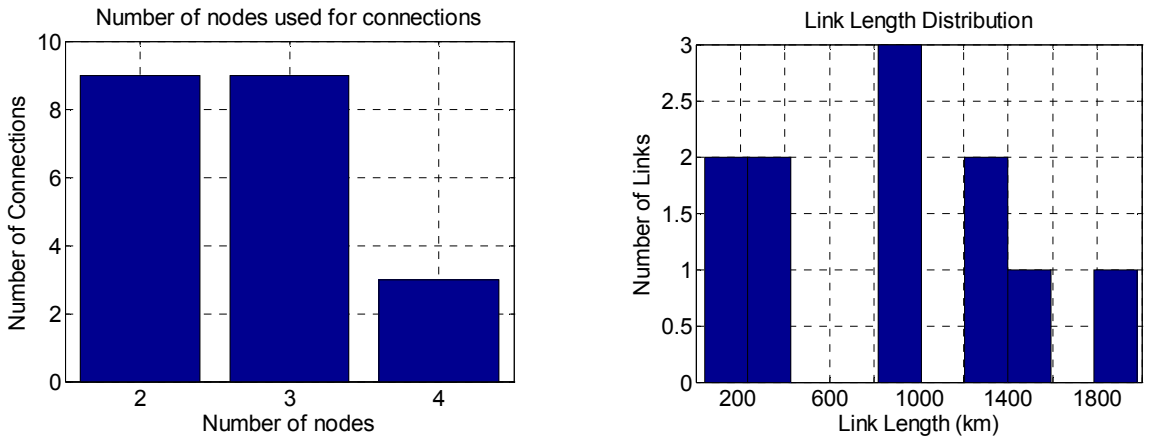

Figure 6.7 : The number of nodes participating in each connection and the distribution of link lengths.

The distribution of the lengths of these 21 connections is depicted in Figure 6.8 when ICBR and SP algorithms are used to establish the connections. The average connection length when SP is used is 968km and for the case where ICBR is used to satisfy the requests the average connection length is a bit higher reaching 983km.

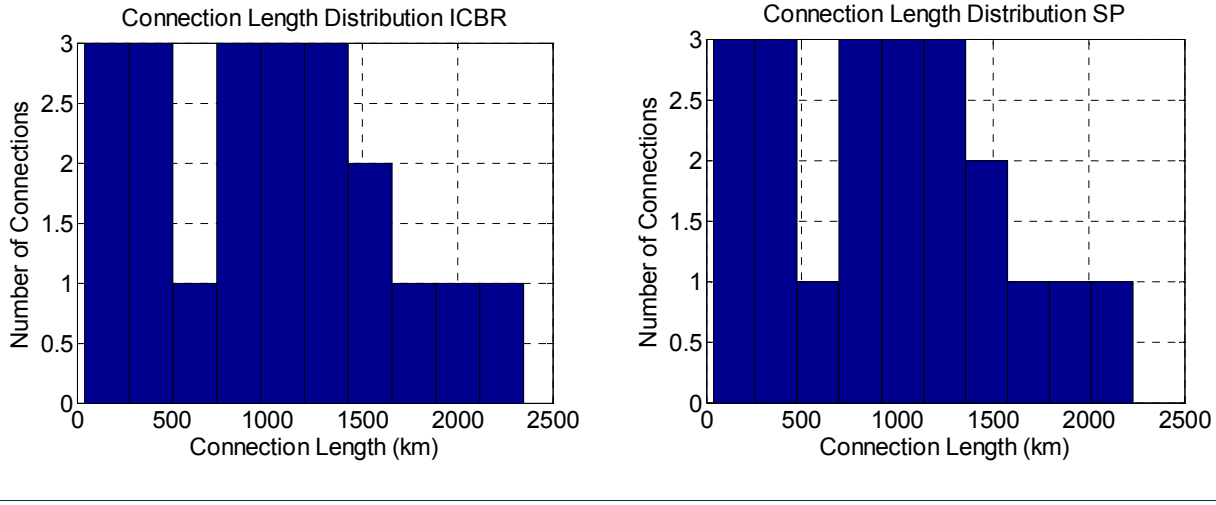

Figure 6.8 : The connection length distribution for ICBR and Shortest Path (SP).

In the next step of our analysis results concerning the blocking percentage for different lengths of the SMF fiber segment are presented. For these simulations we considered a dispersion scheme offering acceptable performance as identified through simulations, with a proper selection of power parameters. Figure 6.9 shows

the benefit that ICBR offers compared with SP routing for both heterogeneous and homogeneous environments. Here we have to notice that even when fiber and various element characteristics are the same in the whole topology, the network still demonstrates a degree of heterogeneity since two nodes of the network (PSNC and SURFnet) have a higher number of input/output ports (fiber counts) and they are directly connected to a larger number of other network neighbouring nodes compared to the rest of the nodes. This introduces a larger number of crosstalk interferers if all the connections through these two nodes are utilized. Therefore, when ICBR is utilized for both heterogeneous (Figure 6.9a) and homogeneous (Figure 6.9b) network scenarios an important improvement is observed.

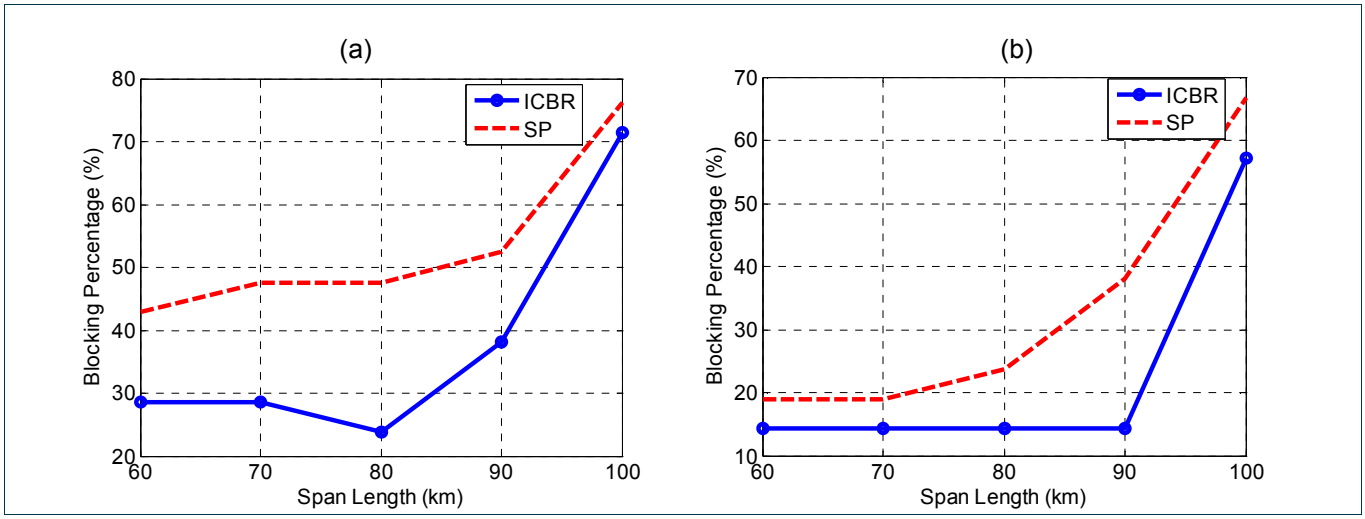

Figure 6.9 : Blocking percentage versus span length for ICBR and SP for the European PHOSPHORUS Scenario for (a) Heterogeneous and (b) Homogeneous fiber parameters

In Figure 6.10 we investigate how the ICBR routing algorithm responds to the increase of the traffic load. In the same figure we also included a plot of the blocking percentage when impairments are not considered in the network (black line) and therefore the blocking is only due to the traffic conditions in this case. The inevitable blocking percentage increase with the traffic load observed in Figure 6.10 is indicated by this black line. The ICBR scheme demonstrates similar blocking increase (around 20% for all loads) whereas the SP scheme appears inappropriate to satisfy the increasing traffic demands (the blocking percentage is increasing from 40% to 60%)

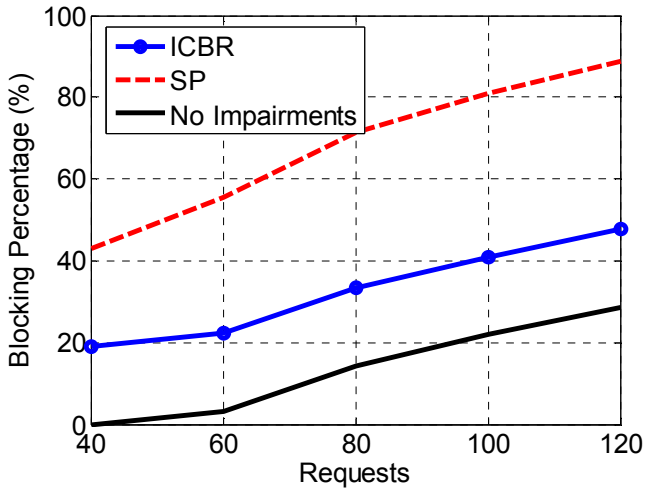

Figure 6.10 : Blocking percentage for different traffic demands

In terms of the physical parameters that influence the network performance, the behaviour of the two routing schemes for various dispersion approaches is examined. Figure 6.11a presents the blocking percentage when ICBR is used whereas Figure 6.11b demonstrates similar results for the SP case. According to these figures it can be concluded that the ICBR provides the capability of flexible engineering since a wider region of dispersion parameters offer improved network performance compared to the SP scheme.

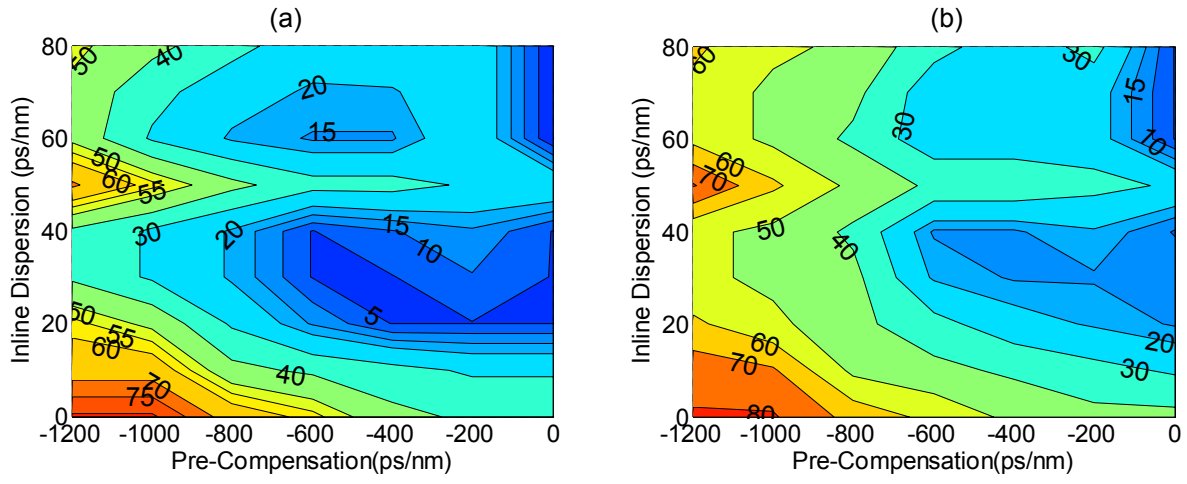

Figure 6.11 : Blocking percentage for various dispersion maps for ICBR and SP routing

In the following simulations we demonstrate the benefit of implementing an impairment aware wavelength assignment scheme in addition to impairment constraint routing. For a specific dispersion map where the residual dispersion -after each 80 km SMF-DCF segment – reaches 30ps/nm and the amount of pre-dispersion is -400ps/nm, the overall blocking percentage is calculated as a function of the channel power levels at the input of the inline modules. The corresponding results are presented in Figure 6.12 for both ICBR and SP routing schemes.

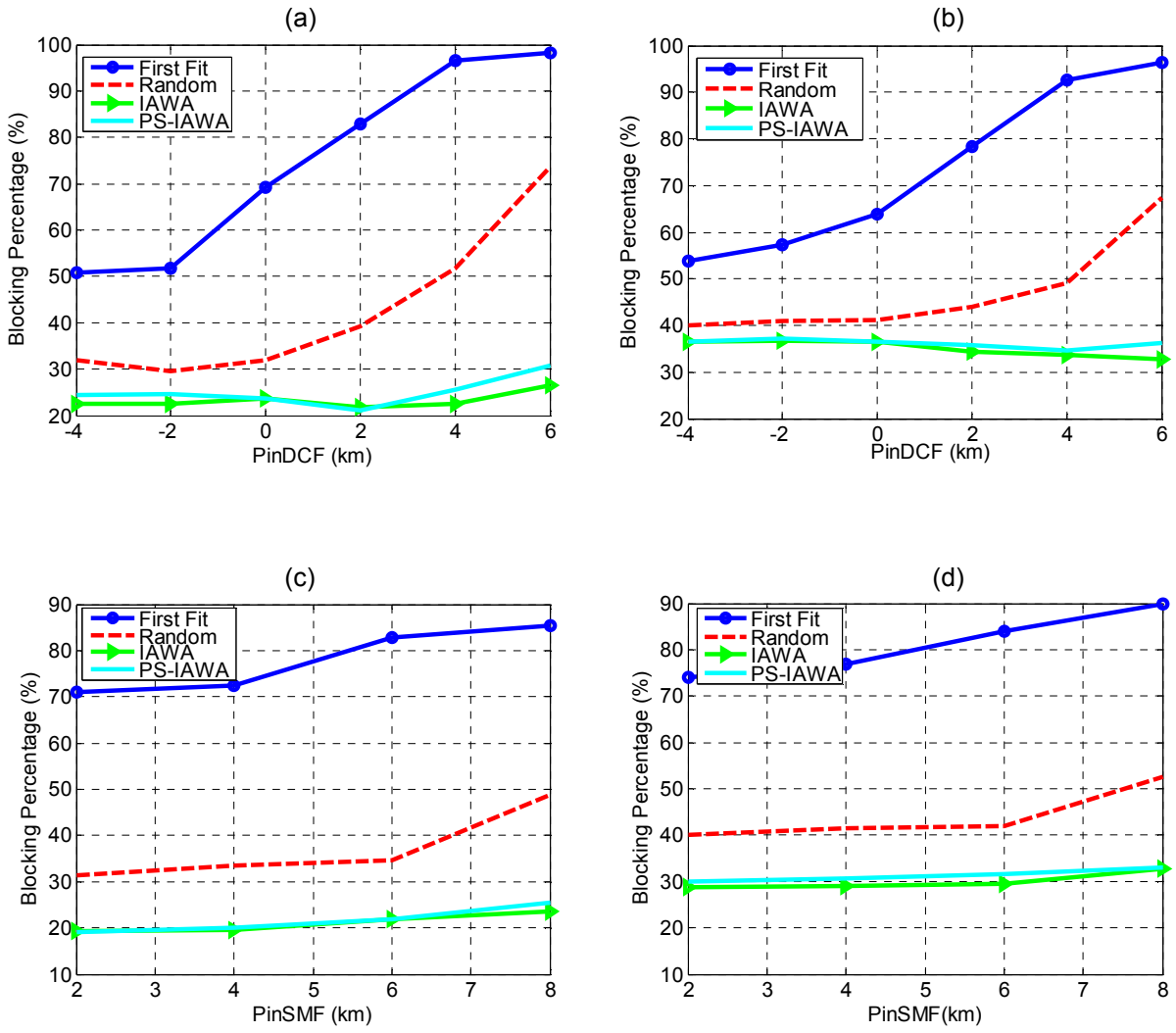

Figure 6.12 : Blocking percentage for ICBR and SP as a function of the power level at the DCF (a,b) and the SMF (c,d) segments for different Wavelength Assignment schemes.

It can be noticed that by selecting a proper wavelength assignment scheme, the overall blocking percentage of the network can be considerably reduced. The PS-IAWA outperforms first fit and random wavelength assignment schemes and exhibits similar performance behaviour with the more computational intensive IAWA scheme which calculates the Q-factor of all potential wavelengths that can be used to establish a lightpath each time a path is considered, whereas the PS-IAWA does not require any further calculations once the order of the wavelengths has been discovered at the beginning of the simulation. For low power levels in the DCF segment (-4dB to 2dB) the observed improvement is around 5% between the IAWA schemes and the random WA scheme, and more than 20% compared with First Fit scheme. The significant advantage of IAWA schemes is becomes more apparent when the power levels of the DCF increase. In such cases, a considerable gain is achieved by introducing the IAWA schemes which range from 20% to 40% compared with the random WA case and is even more for the first fit scheme. Similar conclusions can be drawn as the power in the SMF fiber increases, where the benefit is more than 10% for the majority of the input powers between IAWA schemes and random WA. Therefore by applying IAWA schemes in the network a wider range of input powers can be tolerated in both SMF and DCF segments. Also introducing our ICBR algorithm for the path computation procedure proves to be beneficial against the conventional shortest path as demonstrated by comparing figures Figure 6.12a and Figure 6.12c with Figure 6.12b and Figure 6.12d respectively. A noticeable improvement, around 10% is indicated at least for cases where the blocking percentage is in an acceptable level as it appears when IAWA schemes are implemented. Consequently, the combination of a proper WA assignment scheme with an ICBR algorithm provides significant performance improvement in the network.

Simulation results based on the PHOSPHORUS Global scenario

Here the evaluation of the Impairment Constraint Based Routing algorithm is performed by applying the algorithm to the PHOSPHORUS global topology illustrated in Figure 5.1.

The PHOSPHORUS Global network topology is generated by expanding the PHOSPHORUS European scenario to include two additional North American nodes and therefore it consists of 9 nodes interconnected with 16 bidirectional links. The link length distribution for this scenario is depicted in Figure 6.13 (right) where it is noticed that the link lengths cover a range from 40 to 9000km, with the average link length being 2692.54km. In the majority of the simulations it is assumed that there are connection requests between every possible pair of nodes and therefore 36 end-to-end connections are requested to be established. The number of nodes that participate in these 36 connections is presented in Figure 6.13 (left) when ICBR is used for the discovery of the routes.

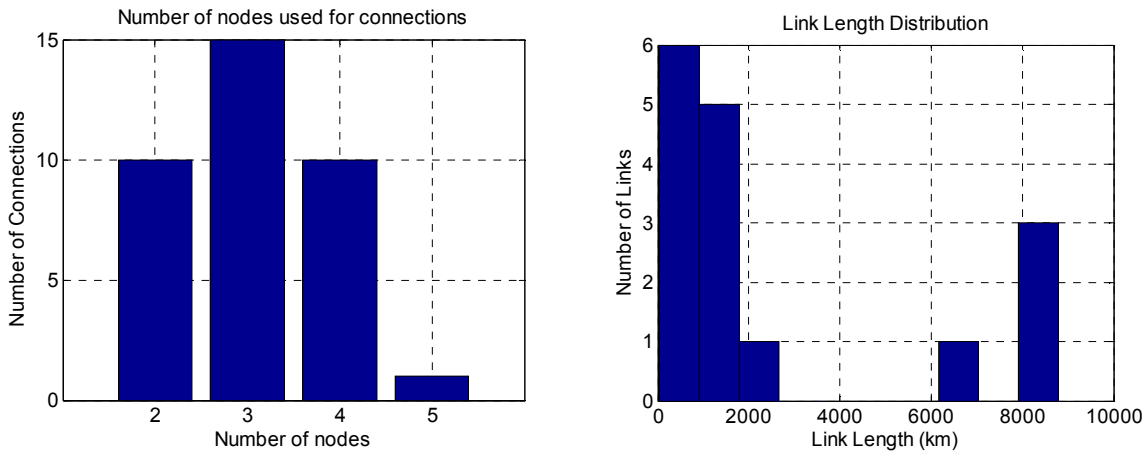

Figure 6.13 : The number of nodes participating in each connection and the distribution of link lengths.

Considering a transparent scenario, Figure 6.14 demonstrates high blocking percentage for a wide range of applicable dispersion maps for both the ICBR and the SP schemes, indicating that the existence of a completely transparent scenario in this case is not possible.

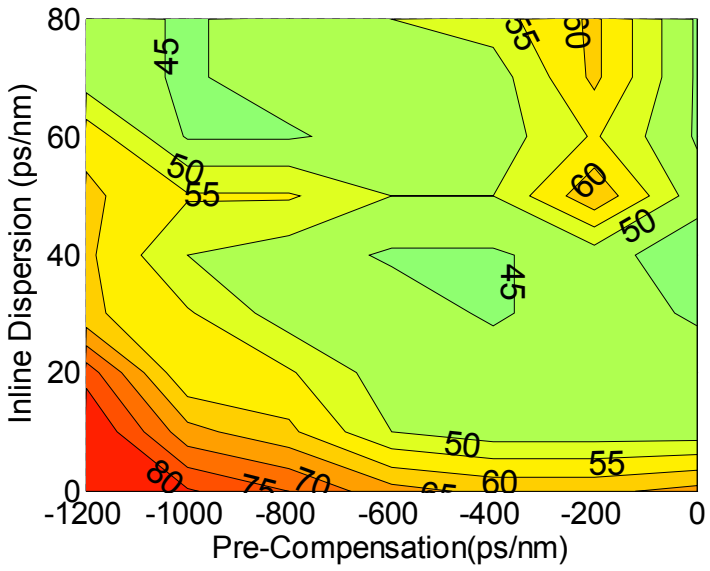

Figure 6.14 : Blocking percentage for different dispersion maps when ICBR is used in the transparent PHOSPHORUS global topology.

To overcome this problem, in the next step of our analysis regenerators are inserted at each node of the PHOSHPORUS global topology and in the long transatlantic connections reducing the maximum unregenerated length to 2000km. The results depicted in Figure 6.15 indicate the necessity of the regenerator deployment by demonstrating optimum performance for a wide variety of dispersion schemes. In addition a significant network blocking improvement is offered by the utilization of the ICBR algorithm in comparison with the conventional SP algorithm.

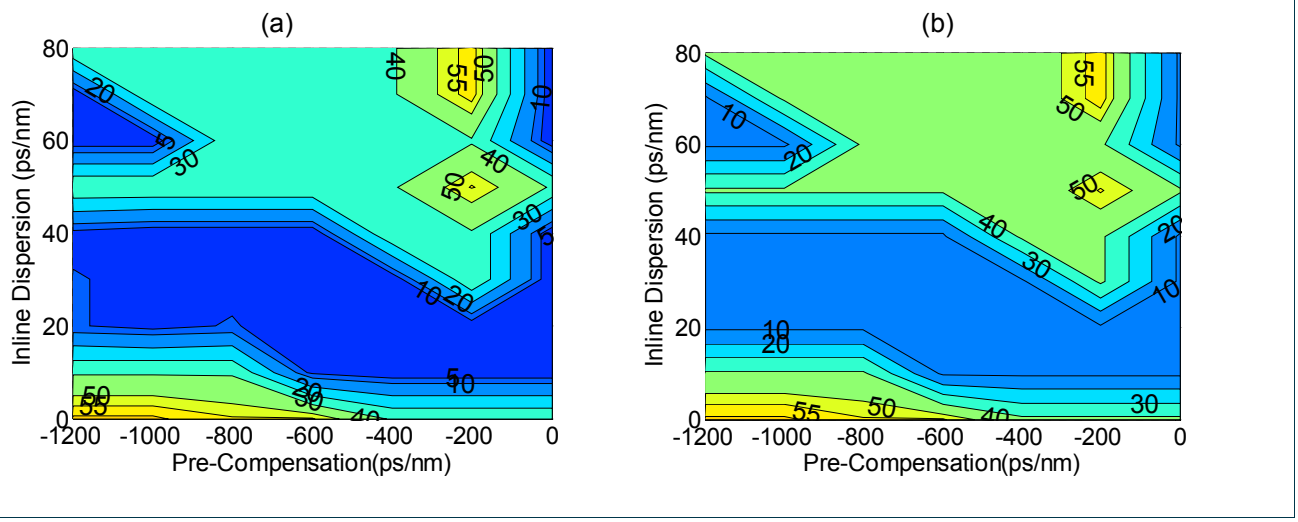

Figure 6.15 : Blocking percentage for different dispersion maps when (a) ICBR and (b) SP is used in PHOSPHORUS global topology employing 3R regeneration.

In Figure 6.16 the effect on the overall performance introduced by a network parameter such as the span length is examined. Choosing a specific dispersion map, which according to Figure 6.15 provides optimum performance for the 80km span length and for the same power parameters the impact of the span length as well as the benefit of the ICBR algorithm is demonstrated.

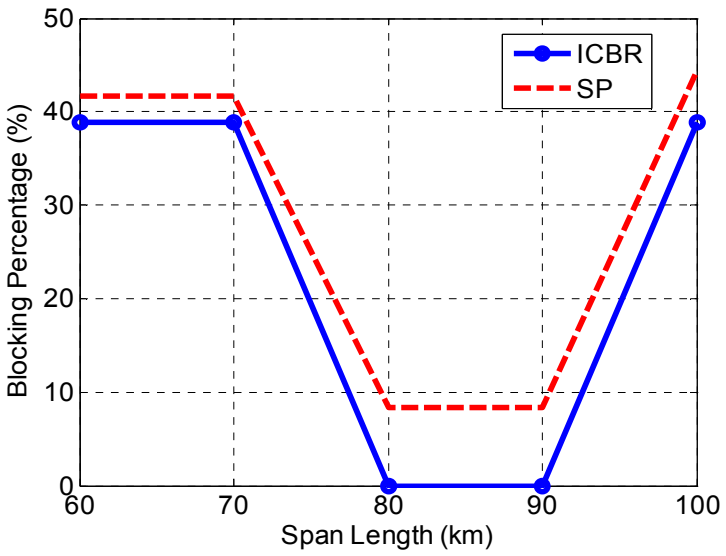

Figure 6.16 : Blocking percentage with respect to span length

In the rest of this section we consider the case of 2R regeneration instead of the 3R regenerators deployed for the previous simulations since in real networks and particularly for high data rates 3R regeneration maybe too expensive or even unavailable. When only 2R regeneration is employed, a penalty may arise due to the accumulation of jitter; this effect confines the maximum reach of the network and impacts its overall performance. Therefore, in the following simulations the overall system performance is evaluated through a closed-form BER expression that takes into consideration the interplay of amplified spontaneous emission (ASE) noise, optical filtering, self-phase modulation and group velocity dispersion (SPM-GVD), cross phase modulation (XPM) and four wave mixing (FWM) as well as the reshaping properties of the 2R regenerators [MarkidisPTL07].

The imperfect characteristics of the 2R regenerator are described based on the well-known approach that considers the effect of inter-channel nonlinearities (XPM, FWM) as independent amplitude perturbations [Ten99] and therefore the derived electrical noise variances are added to the corresponding variance of the accumulated ASE noise. Single channel penalties arising from the interplay between SPM, GVD, and optical filtering with the regenerative 2R reshaping are identified through numerical calculations.

The generalized O/E/O regeneration model used here has a similar performance with an all-optical solution from a system perspective. The 2R regenerator is schematically illustrated in Figure 6.17 where the filter could represent the low pass behaviour of the all-optical counterpart whilst the reshaping properties of the subsystem are modelled with a time invariant step-wise linear transfer function with slopes of γ around “mark” and “space” levels and a corresponding discontinuity at the threshold level. When γ=0 there is full suppression of the amplitude distortions, whilst when γ=1 no regeneration occurs. The position of the threshold (L in Figure 6.17) can be selected in order to optimize the reshaping capabilities under various system conditions. Although the signal is

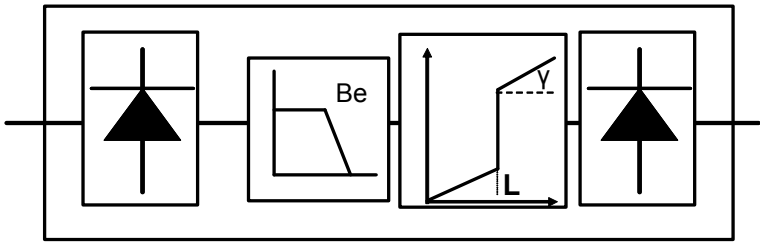

Figure 6.17 : A schematic diagram of a 2R regenerator

regenerated, some errors still occur and accumulate. The overall BER is derived taking into account the errors generated at each 2R stage due to amplitude distortions [Mørk03] as well as the error rate at the final receiving end taking also into account the jitter degradation [Öhlen 97]. This can be described by the following formula:

$$BER_{Total} = \sum_{i=1}^{N} BER_{i,ampl} + \frac{1}{2} Erfc\left(\frac{pen \cdot P}{\sqrt{2\left(\sigma_N^2 + \sigma_{jiter,N}^2\right)}} \right) \quad (42)$$

where N is the number of cascaded 2R regenerators, $\sigma_N^2, \sigma_{jitter,N}^2$ the electrical signal to noise beating terms attributed to the amplitude and random jitter distortions generated by the ASE noise and nonlinear impairments (FWM and XPM). The parameter “pen” for each path has been numerically identified, using a commercial simulation tool (VPI), and represents the eye closure penalty due to the accumulation of deterministic jitter. This residual type of degradation arises from the interplay between the SPM/GVD introduced pulse distortions and the regenerative reshaping of the 2R subsystem.

The results presented in Figure 6.18 are the outcome of a 2R regeneration scheme with a moderated suppression of the amplitude distortions using γ=0.5. Compared with the transparent case illustrated in Figure 6.14 it can be noticed that efficient 2R regeneration may not only improve the overall networking performance in terms of blocking but it can also relax the requirements of the design and engineering of the physical layer.

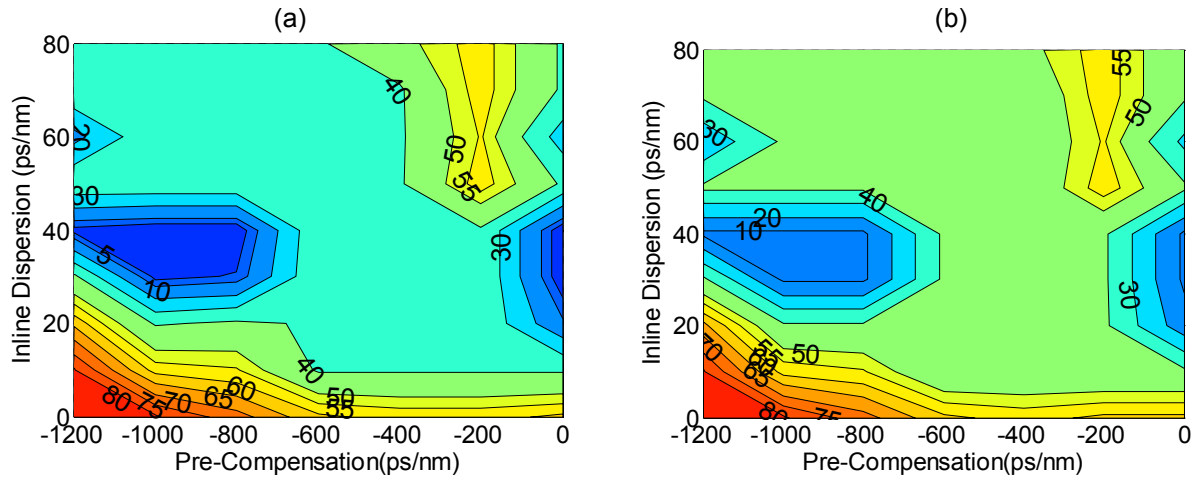

Figure 6.18 : Blocking percentage for different dispersion maps when (a) ICBR and (b) SP is used in PHOSPHORUS global topology employing 2R regeneration with γ=0.5.

Finally, the efficiency of the ICBR compared to the shortest path is also demonstrated as a function of the γ-parameter in Figure 6.19. The calculations have been performed considering the optimum threshold value L=0.3 and for two different pairs of pre and inline residual dispersion values. In both cases it is shown that the ICBR is more efficient compared to the SP for lower γ values where the suppression of the noise/amplitude distortion/jitter is more pronounced.

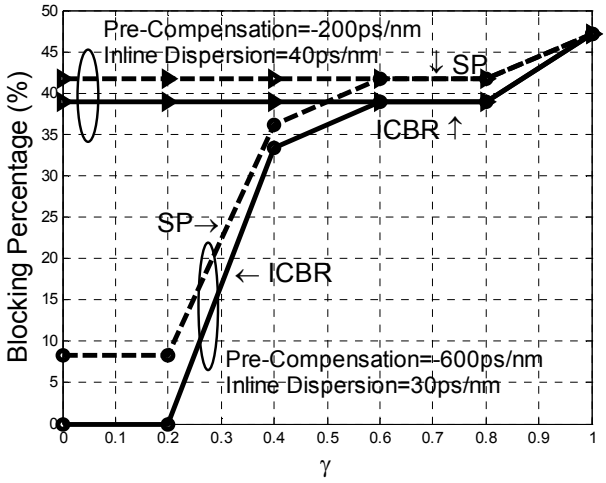

Figure 6.19 : Blocking percentage as a function of the γ-parameter for the ICBR and SP routing schemes.

6.1.2 Grid requirements

As described in Section 4.2 there are various Grid user QoS requirements that have to be addressed by the job routing algorithms. We have considered an underlying WDM optical network that utilizes wavelength routing for establishing connections. In the previous sections two techniques for introducing physical layer impairments into the routing algorithm were proposed. In this section we provide equations and extensions that can be used to enhance these two approaches, enabling them to cope with additional requirements introduced by a Grid User.

Network delay constraint

A user may specify an upper bound on the network delay that his jobs and their corresponding data must face during their transmission. In order to satisfy the network delay constraint that a user imposes we discard the paths that cannot satisfy this constraint and then apply one of the aforementioned RWA algorithms.

More specifically, we calculate for every candidate path its total delay, by aggregating the delays of the links that comprise it. The delay cost value of a fiber link l equals to $f \cdot d_l$, where f is the delay of a fiber per km (typical value: $f = 1/u_f = 5 \cdot 10^{-6}$ sec/ km) and d_l is the link length in km.

Assume that a user requires that the end-to-end connection delay is less than a specific value D_{comm}. A path, in order to satisfy the delay-constraint, must have delay less than the requested D_{comm}. Thus, every path p that satisfies the below inequality is a candidate solution:

$$\sum_{l \in E_p} f \cdot d_l = f \cdot \sum_{l \in E_p} d_l \leq D_{comm}, \quad \forall p \in P \tag{43}$$

where

l : link with length d_l

E_p : Set of links of path p

P : Set of candidate paths

D : The maximum acceptable communication delay-cost the user can tolerate

Assume that all the wavelengths of the network have capacity C and that the job has to receive data with size I in order to start its execution. If the transmission time of this data is considerable (i.e $1/C$ is comparable to D_{comm}), then we have to include in the above inequality (Eq.43) the transmission time:

$$f \cdot \sum_{l \in E_p} d_l + I/C \leq D_{comm} \quad (44)$$

The set of candidate paths is reduced by discarding all paths with unacceptable total delay-cost value, and the RWA algorithm (as presented in Section 6.1.1) uses the remaining set of candidate paths.

Bandwidth demand

We have considered an underlying WDM optical network that utilizes wavelength routing for establishing connections. This approach poses limitations on the granularity of the bandwidth that can be assigned to a connection request. One approach to provisioning fractional wavelength capacity is to multiplex traffic on a wavelength. The resulting networks are referred to as WDM grooming networks [Zang02]. We won't analyse these techniques further since they are out of the scope of this deliverable. However, with the algorithms presented here we can address cases in which a connection demands bandwidth that is multiple of the capacity of one wavelength (the user requires a connection with more than one wavelengths).

In the RWA algorithms (presented in this Deliverable) we give as input a $n \times n$ traffic matrix of integers denoted as **R**, whose elements represent the requested connections, e.g. $\mathbf{R}_{sd}$=m means that from (source) node *s* to (destination) node *d* there is a request that demands *m* (*m* ≥ 1) number of wavelengths.

There are two possible solutions to the demand of more than one wavelength (*m*>1):

(i) The LP formulation (Section 6.1.1 and Appendix A) solve this RWA problem so that the demanded *m* wavelengths can be routed by *j* different paths and $j \leq m$, $j \leq k$, where k is the number of candidate paths examined for each connection (used in the Dijkstra *k*-shortest path algorithm).

(ii) If there is a demand that the *m* (*m*>1) wavelengths of the requested connection have to be routed over the same path, then we have to modify the LP formulation presented in Appendix A. The changes are presented in Appendix B.

Delay jitter requirement

This user QoS requirement is not applicable to the kind of networks that we are examining. Delay jitter can be measured in packet switched networks where the buffering delays and routing decisions affect the delay and possible the order in which the packets arrive at the destination. However, in the networks discussed here, an end-to-end circuit switched connection is established in a form of a wavelength switched path over a WDM network. Therefore, assuming that the RWA solution gives a feasible solution to the connection request, any Delay jitter requirement of a Grid user can be met by this connection.

Packet loss ratio

A Grid user might request that his connection has packet loss ratio which is less than an upper bound. In our algorithms, we have expressed and ensured the quality of the end-to-end wavelength connection through

various physical impairments. If the impairment constrains are satisfied and the RWA algorithm returns a feasible solution to the request then we can assume that any user packet loss ratio requirement can be met.

End-to-end total Delay

A Grid user usually specifies an upper bound of the end-to-end delay that his tasks must face. The network delay constraint, we described previously, is a component of this end-to-end delay. Another component of this end-to-end delay is the delay induced by the computational resource (queuing and execution time of the job). In deliverable D5.2 [phos-D5.2] and in [kokvar07] we described a framework which provides to a user's task a delay guarantee on its execution on a specific resource. This execution time includes both queuing delay and the actual execution time. Using the aforementioned framework we can assign to each user-resource pair an upper bound of the computational delay.

So, in order to provide end-to-end delay to a user we have to combine the network and the computation delay guarantees. More specifically, (Eq. 43 or Eq. 44) gives the one way propagation delay $d_{i,j}^{comm}$ where *i* is the source (the node of the user) and *j* is the end of the path (a computation resource):

$$d_{i,j}^{comm} = f \cdot \sum_{l} d_l$$

If we assume that the QoS framework described in [kokvar07] is used, then the computation delay $d_{i,j}^{comp}$ bound of a task that user *i* sends for execution on resource *j* -assuming that the registration process of *i* to *j* has been successful (see [kokvar07]) is given by:

$$d_{i,j}^{comp} = T_{ij} + \frac{\sigma_{ij}}{g_{ij}} + \frac{J_{ij}^{\max}}{g_{ij}} + \frac{J_j^{\max}}{C_j},$$

where

T_{ik} : The time period for which the user *i* must locally withhold a task *k*, in order to preserve the framework's constraints.

σ_{ij} : The maximum workload of tasks (burstiness) that the user *i* will ever send to resource *j*.

g_{ij} : The computational rate the resource *j* provides to the user *i*.

$J_{ij}^{\max}$: The maximum task workload the user *i* will ever send to the specific resource *j*.

$J_j^{\max}$: The resource's *j* maximum acceptable task workload.

C_j: The computing capacity of resource *j*.

So the total end-to-end delay is given by:

$$d_{i,j}^{total} = d_{i,j}^{comm} + d_{i,j}^{comp} + d_{j,i}^{comm}$$

The set of candidate paths of the RWA problem must use the above equation in order to discard paths to resources that do not satisfy the total delay that the Grid user requests.

6.2 Multi-domain routing

In this section, we present an architecture to support anycast-based routing in multi-domain Grid networks. The main objective is to provide a scalable approach (i.e. to ensure control plane traffic remains feasible for large Grid deployments), while offering flexibility in the parameters available to the routing protocol. We present the architecture in detail, summarize algorithms for optimal planning of the architecture, and finally use simulation analysis to demonstrate the scalability in terms of control plane traffic and show the minimal performance loss when compared to an optimal (albeit non-scalable) routing strategy.

6.2.1 Anycast proxy architecture

In optical grids, orchestration between clients submitting a job, the optical network components and resources capable of processing jobs is inevitable. For small sites, resources could send status update messages directly to all grid clients, whereupon **clients autonomously** select the most appropriate processing resource and **reserve** optical network **resources** (through the optical control plane). In general, a **central scheduler** is used to shield grid and network complexity from end-users. In this case, clients forward a job description (duration, data size, etc.) to the scheduler, which selects the most suitable target resource(s) and contacts the optical control plane to reserve network resources. Once all reservations are made in the background, the central scheduler sends a notification to the client, who can then submit the job to the resource chosen by the scheduler.

In a multi-domain environment consisting of multiple grid sites, a single central scheduler might not be a feasible solution, however. First, different optical domains might employ different control plane protocols, potentially leading to interoperability issues. Secondly, this approach forces each optical grid site to advertise all network and resource state and configurations to all parties involved, which might violate confidentiality policies of the Grid site/network operator. Furthermore, scalability issues usually arise at the central scheduler entity due to computational complexity inherent in job scheduling. Finally, site-level state aggregation could be necessary to reduce control plane overhead

For the aforementioned reasons, this section introduces a proxy-based control plane as shown on Figure 6.20. Using this approach, a resource only forwards state information to its closest proxy using **anycast**[1] communications. Typically, this proxy belongs to the same domain as the resource node, and from the resource perspective it also behaves like a local scheduler. Likewise, a client who wants to submit a job to the multi-domain Grid forwards its request to the nearest proxy, also using anycast communications. Upon reception of the job request, the proxy selects the most suitable target proxy to forward the request to, based on **aggregated state information** of the resources connected to this proxy. We assume the proxy in the client domain will take the necessary control plane actions to set-up the light paths for actual job transmission in the data plane[2]. Once the job request is processed by the Grid and optical control planes, the client is notified about this and the actual job can be submitted.

Using this approach has the following benefits:

- Increased cooperation between independent optical grid sites due to the network and resource state distribution in aggregated form;
- Grid sites maintain their autonomy and configuration details are not revealed;
- Control plane scalability: the intelligent state aggregation results in reduced control plane traffic;
- Flexibility in migration and deployment: whenever a domain deploys a proxy, it can participate immediately in the multi-domain Grid network
- Adoption of novel data transport and control plane technologies is straightforward, by adding new interfaces to the anycast proxy servers which can understand and control these new protocols
- System-wide optimization of the Grid network is possible, e.g. minimal job blocking, global load-balanced resource utilization, etc.
- Can support any subset of parameters available to the routing protocol, i.e. computational resource states, physical parameters of photonic network, etc.

The following section briefly describes algorithms to optimally dimension the proxy infrastructure, by concurrent placement of proxy servers and determination of their capacities. Subsequently, control plane scalability of the proposed approach is demonstrated by means of simulation analysis.

[1] Anycast routing in this context means that the client does not specify the job request's destination: if multiple proxies are present in the client's domain, anycast routing will lead it to the most suitable, e.g. closest, proxy. See [Partridge03] for more details on anycast routing.

[2] Observe that we make abstraction of the precise form of this control plane and communication between control plane agents of the various domains

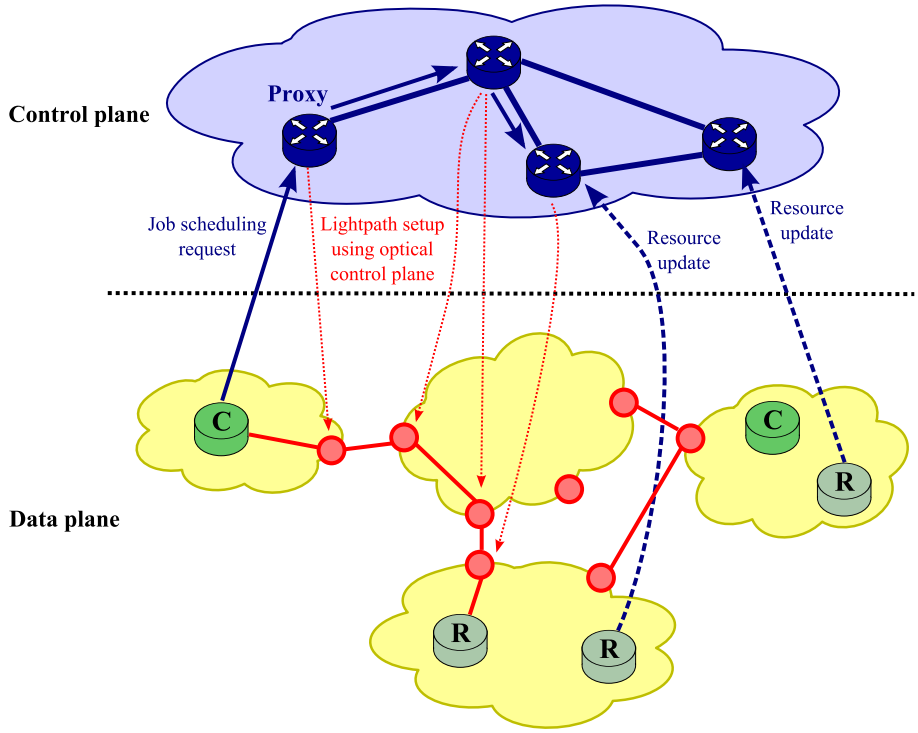

Figure 6.20 – Overview of the proxy-based anycast architecture

6.2.2 Dimensioning the anycast infrastructure

Equipped with the anycast architecture outlined in the previous section, we wish to determine how many proxies are needed and where they should be attached to the network for a given client and server configuration. More formally, given a network G(V , E), a set of source sites S ∈ V and their demands d_i , a set of server sites T ∈ V and their capacities c_j , edge weights w_e : e ∈ E, determine how many client proxies (CP) (resp. server proxies (SP)) are needed, and where they should be attached to the network. Additionally, determine which target sites need to be opened. The optimization process should balance network operational costs (related to flow unit processing costs for regular edges (w_e) and flow unit processing costs for proxies and servers), proxy infrastructure costs (determined by the fixed charge associated with each CP (resp. SP)), and server site opening costs.

We developed several techniques to solve this optimization problem. First, in [Stevens07], we addressed the **optimal** placement and dimensioning of such an anycast architecture using an **integer linear program** (ILP). Unfortunately, due to the complexity of the formulation, an exact solution can only be computed for relatively

small networks (up to 300 nodes). For this reason, we proposed two heuristic methods to solve this problem [Stevens07a): CP and SP **separated** and **combined** optimization. Contrary to the global optimization performed by the exact ILP, both heuristics decouple the proxy placement problem from the traffic engineering between the proxies, which results in a two-step optimization plan:

(1) Find suitable CP and SP locations and determine which target sites to use;

(2) Optimize the flow between CPs and SPs.

In fact, step (2) does not contribute to the proxy placement and dimensioning optimization, but allows us to examine the efficiency of the proxy locations determined in step (1).

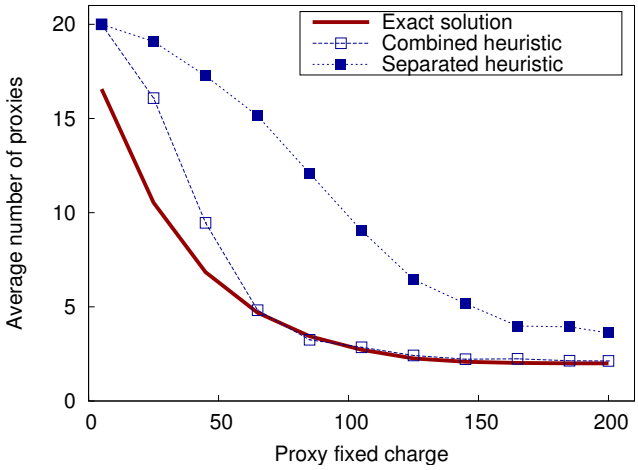

Figure 6.21: Dimensioning of proxy-based anycast architecture: number of proxies

Figure 6.21 and Figure 6.22 summarize the main results of our dimensioning and planning algorithms. Obviously, an increasing fixed charge for installing a proxy (either a CP or SP) leads to less proxies being installed and a growing path stretch. Additionally, the following conclusions can be drawn:

1) Both heuristics follow the same trend as the exact optimization, and both provide near-optimal results.

2) Separated optimization generally yields results with a smaller path stretch, at the expense of installing more proxies.

Combined optimization suggests a smaller number of proxies and a larger path stretch. On the contrary, the combined heuristic initially overestimates the infrastructure costs by coupling CP and SP functionality. Afterwards, the infrastructure costs can often be reduced when excess (unused) functionality is removed.

Further details concerning the modelling approach and simulation parameters, can be found in [Stevens07, Stevens07a].

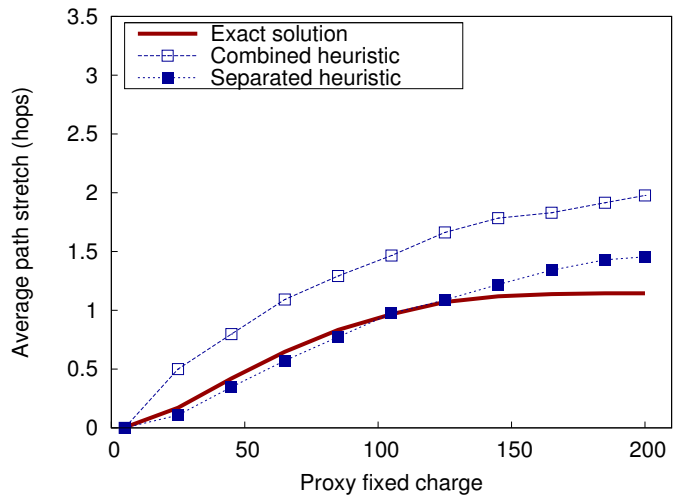

Figure 6.22 : Dimensioning of proxy-based anycast architecture: average path stretch

6.2.3 Resource state information: strategies for aggregation

This section focuses exclusively on the Grid control plane for multi-domain optical networks and investigates control plane scalability by means of discrete event simulation. The following assumptions are made:

- Resources send state updates at a fixed rate; if there are multiple receivers, the update messages are multicast;

- Proxy nodes send updates at a fixed rate (usually smaller than the resource update rate), using broadcast messages to all proxies.

For the inter-domain grids, three control plane scenarios were identified:

- *One central scheduler:* A single scheduling entity is aware of the full network and resource state of the multi-domain Grid. It receives all job requests and is responsible for all scheduling decisions. Figure 6.24 shows an overview of this approach. A single service node is installed somewhere in the core network. Every resource will send its status updates to this service node and every client will contact this service node for acquiring a resource to execute on. This means that a single entity (the service node) will be responsible for the scheduling of all jobs. This approach is not scalable and suffers from a single point of failure.

- *No scheduler:* in this case, resources send updates to all clients directly and clients individually select an appropriate resource. This requires total transparency between domains, and is depicted in Figure 6.23. Every client node will act as a service node and will be responsible for choosing a resource to execute on. The resource nodes will of course send their status updates to every client in the network. This means the number of status updates sent will increase dramatically compared to the centralized setup. An advantage of this setup is the removal of the single point of failure.

- *Proxy infrastructure:* Since both setups have their disadvantages a third setup has been considered. The proxy setup tries to combine the advantages of both setups while trying to avoid their disadvantages. In the proxy setup every local network has a server proxy and a client proxy. The server proxy will bundle the resources in the local network and acts as a single resource to the core network. Resource updates are sent to the server proxy in the same local network and are stored by the server proxy. The server proxy will send in its turn the combined status updates to every client proxy in the network. Since the status updates are bundled in the server proxy, less updates will traverse the core network. The client proxies act as a service node for the clients in the local network. The job flow is as follows: a client sends a job to the closest client proxy. The client proxy will check the status of the different server proxies and chooses the best one and forwards the job to that server proxy. The server proxy will then look at its resources and choose the best one and send the job to that resource. When the job is finished at the resource the result will travel the other way to the server proxy, client proxy and finally the client.

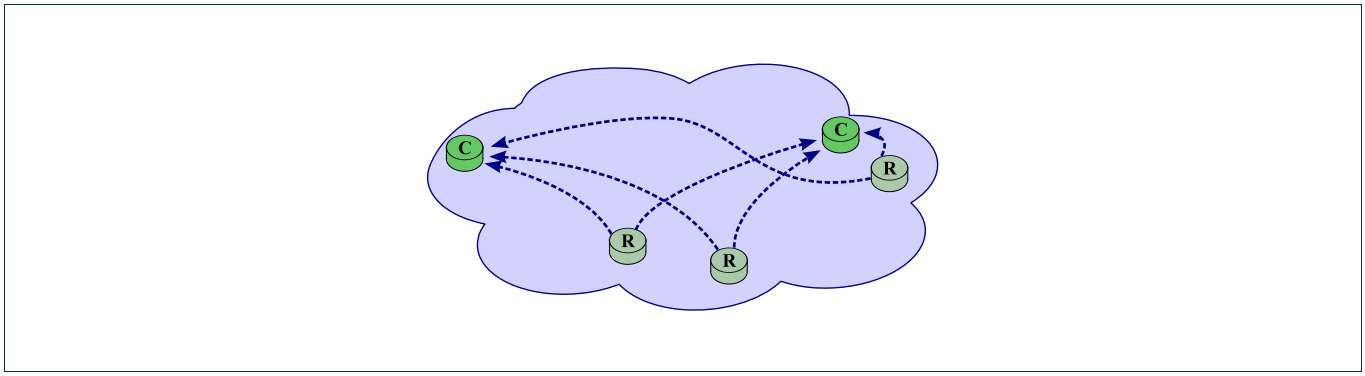

Figure 6.23: Fully distributed job scheduling

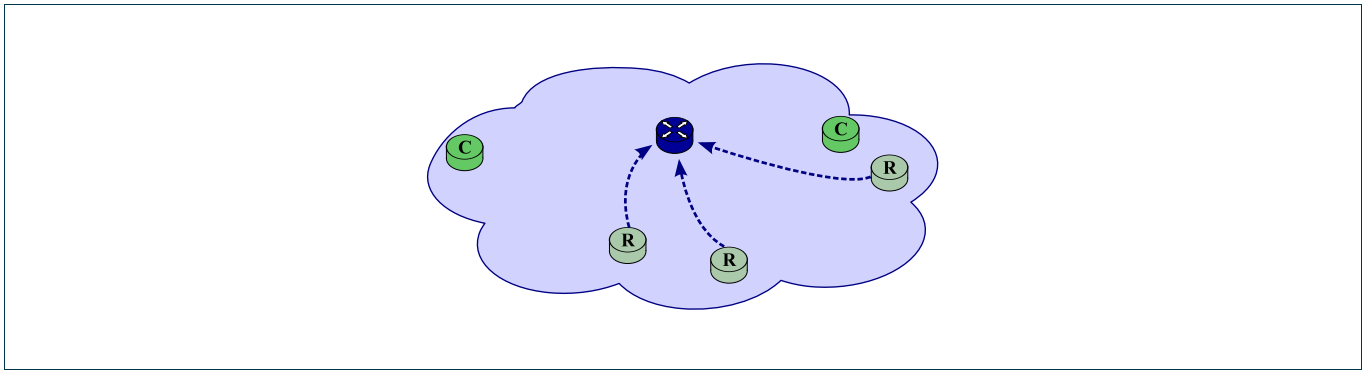

Figure 6.24: Centralized job scheduling

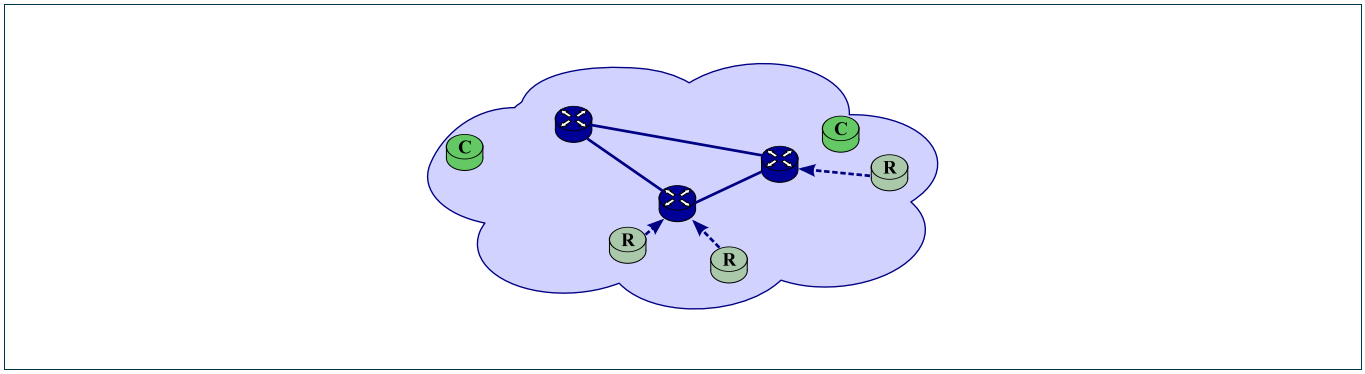

Figure 6.25: Proxy-based anycast job scheduling

6.2.4 Evaluation

In this section, discrete event simulation results for the three aforementioned control plane scenarios are discussed. The simulation network topology is the multi-domain PHOSPHORUS network (see Figure 5.3) and at each edge node the number of clients and resources is chosen to be proportional to the number of inter-domain links. Intra-domain topologies are abstracted to simplify the simulation setup: resources and clients are connected to the domain edge node by an aggregation tree. The job model is configured in a similar way as described in [Christo07]: the job duration is assumed to be distributed hyper-exponentially and the job inter-arrival times (IAT) follow an exponential distribution (i.e. Poisson arrival process). Furthermore, we assume that resources update their state information at a frequency higher than the one used by proxies. Since proxies generally aggregate the state information of multiple resources, their rate of change is much lower, thus allowing a lower state update frequency. Simulation parameters are summarized in the table below.

Parameter	Value
Topology	PHOSPHORUS topology. For each domain, the number of clients and resources is chosen to be proportional to the number of inter-domain links
Number of clients	61
Number of resources	41
Parallel jobs per resource	10
Job duration	Hyper-exponential distribution
Job IAT	Exponential distribution
Resource update interval	5 time units
Proxy update interval	10 time units

Table 6-1: Simulation parameters for proxy-based anycast architecture

Results for the job loss rate related to the job IAT (and corresponding average generated system load on the second axis) are depicted in Figure 6.26: We can conclude that there is no significant difference in job acceptance rate between the three alternative approaches; the less-frequent distribution of aggregated resource state by the proxy system does not prevent efficient resource allocation.

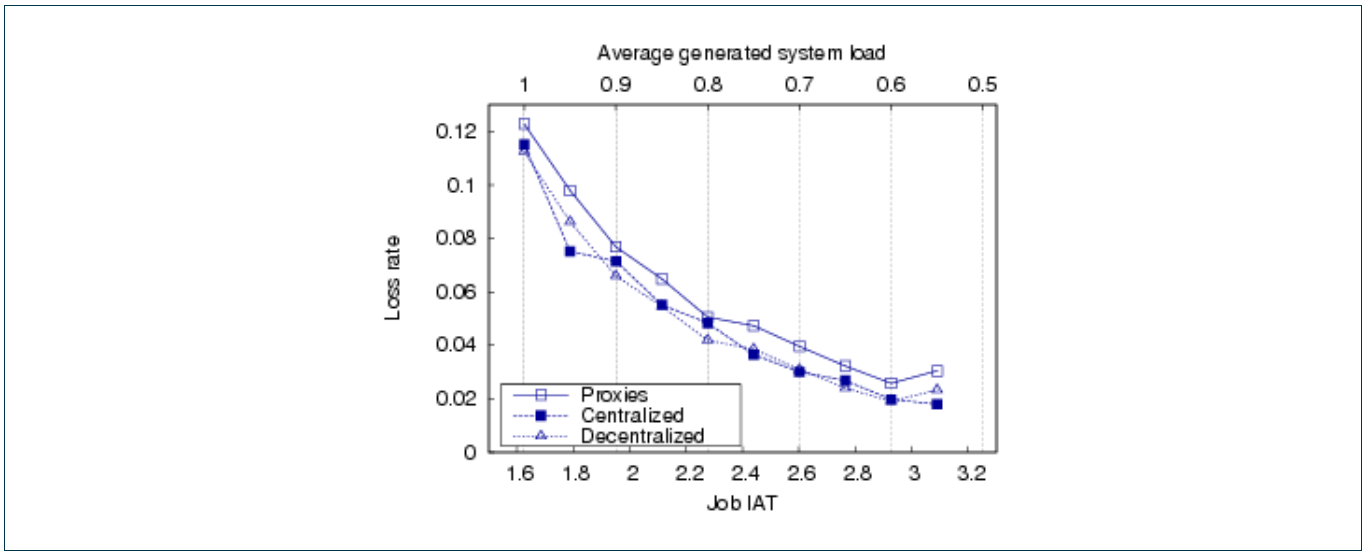

Figure 6.26: Job loss rate for varying job IAT (load)

Observing the corresponding number of events generated in the simulator (see Figure 6.27) we can conclude that a proxy-based inter-domain job allocation approach significantly reduces the control plane overhead, while job loss rates are comparable with those associated to the other strategies. As such, it offers a scalable

solution for a growing network with an increasing number of clients and computational resources. Indeed, when proxies aggregate state for a larger number of resources, their state will be even more accurate and less volatile. At the same time, proxies prevent frequent resource update messages from propagating through the network.

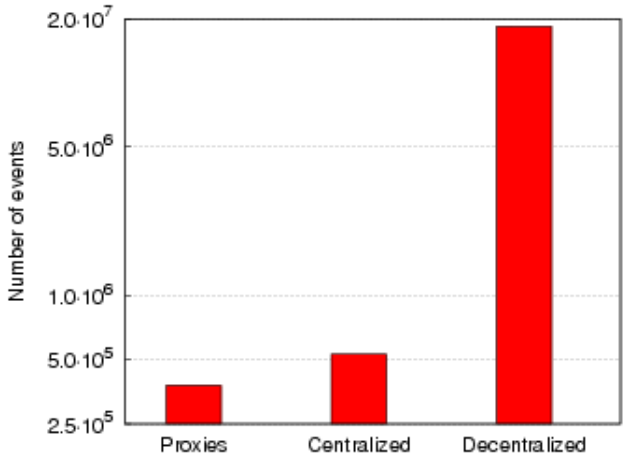

Figure 6.27: Number of control plane events for different multi-domain routing approaches

7 Conclusions

In this deliverable a number of routing approaches have been proposed that take into consideration physical layer characteristics and Grid-specific requirements, to provide optimum resource utilization and offer improved QoS. In addition, routing is used in order to provide enhanced coordination between the submitted jobs and the optical network components and resources capable of processing the jobs in the form of anycast-based routing in multi-domain Grid networks.

Accurate analytical models that evaluate physical layer degradations have been developed and integrated into the routing procedure to allow optimized routing performance. Two different impairment constrained based routing algorithms have been described and evaluated through extensive simulations focusing on the PHOSPHORUS network topology to demonstrate the importance and benefits of this type of routing approach. According to the first impairment constrained based routing approach a number of linear impairments have been considered individually as a set of performance metrics that have to be met before any connection can be established in order to exhibit their impact in the path estimation process. On the other hand the alternative approach a number of linear and nonlinear physical layer constraints have been considered, intergraded and included in the routing process through the estimation of the signal Q-factor. Simulation results revealed a noteworthy improvement in the network performance for a wide range of design parameters indicating the need to upgrade current routing approaches to include optical constraints.

In addition mathematical expressions for a variety of Grid user requirements dealing with networking issues have been described and discussed in detail in the document. More specifically a number of quantitative requirements for allowing network QoS have been investigated dealing with delay, delay jitter, bandwidth and packet loss rate. These requirements have been modelled and are incorporated in the job routing algorithms presented in this deliverable.

Finally a part of this deliverable is dedicated to the introduction of an efficient architecture that supports anycast-based routing in multi-domain Grid networks. The benefits of such approach as opposed to the

centralized and the fully distributed solutions have been recognized and analyzed. Mainly the proxy architecture offers control plane scalability due to the reduced control plane traffic occurring from the intelligent state aggregation, easy maintenance of administrative and security issues at the Grid sites since configuration details are not revealed, a straightforward adaptation to novel data transport and control plane technologies, system-wide optimization of Grid networks and support of any subset of parameters available to the routing protocol.

Algorithms for optimal planning and dimensioning of the proposed architecture have been described and evaluated through simulations. Also simulation analysis has been implemented through discrete event simulations for investigating the control plane scalability of multi-domain networks and demonstrated a significant overhead reduced of the proxy based architecture with respect to the other two approaches offering a scalable solution for a growing network with an increasing number of clients and computational resources.

8 References

[Agrawal95] G. Agrawal, Nonlinear fiber optics. Academic Press, 1995.

[Alangar03] W. Alanqar et al. “Requirements for Generalized MPLS (GMPLS) Routing for Automatically Switched Optical Network (ASON)“ December 2003

[Antoniades02] Antoniades et al. “Performance Engineering and Topological Design of Metro WDM Optical Networks Using Computer Simulation”, I IEEE Journal on Selected Areas in Communications vol. 20, No. 1, January 2002

[Bala95] K.Bala, T.Stern, K. Simchi, “Routing in Linear Lighwave Networks”, IEEE/ACM Transactions on Networking , vol. 3 pp 459-469, August 1995.

[Banerjee96] D. Banerjee and B. Mukherjee. “A practical approach for routing and wavelength assignment in large wavelength-routed optical networks”. IEEE Journal Selected Areas in Communications, vol 14,no.5 pp.903-908, June 1996.

[Barabasi99] Barabasi, A-L & Albert, R. (1999) *Science* 286, 509–512

[Barry95] R. Barry and P. Humblet , "Models of Blocking Probability in All-Optical Networks with and without Wavelength Changers", IEEE INFOCOM '95, Vol. 2, Boston, MA, April 1995, pp.402

[Bhandari99] R. Bhandari, “Survivable Networks”, Boston: Kluwer, 1999.

[Birman95] A.Birman and A.Kershenbaum, “Routing and Wavelength Assignment Methods in Single-Hop All-Optical Networks with Blocking” IEEE INFOCOM, Boston MA vol 2, pp.431-438, April 1995

[Birman96] A.Birman, “Computing Approximate Blocking Probabilities for a Class of All-Optical Networks” IEEE Journal on Selected Areas in Communications vol 14, no.5, pp852-857,Jun.1996

[Breuer95] D. Breuer, C. Kurtzke, and K. Petermann, “Optimum dispersion management of for nonlinear optical single-channels systems," in Proc. OFC'95, Feb 1995, pp. 196-197.

[Cantrell03] C. D. Cantrell, Transparent optical metropolitan-area networks, LEOS 2003, 16th Annual Meeting, Vol: 2, pp. 608-609

[Cardilo05] R. Cardillo V. Curri, M. Mellia, “Considering transmission impairments in wavelength-routed networks “ONDM, Milan, 2005

[Cartaxo99] A.Cartaxo “Cross-Phase Modulation in Intensity Modulation-Direct Detection WDM Systems with Multiple Optical Amplifiers and Dispersion Compensators” J.Ligthwave Technol, vol.17,no.2, pp.178-190,February 1999

[Chen98] S. Chen and K. Nahrstedt. An Overview of Quality of Service Routing for Next-Generation High-Speed Networks: Problems and Solutions. IEEE Network, Nov./Dec. 1998.

[Chlamtac92] I.Chlamtac, A.Ganz and G.Karmi, "Lightpath Communications: An Approach to High Bandwidth Optical WAN's" IEEE Transactions on Communications vol 40, pp 1171-1182 July 1992

[Eppstein94] D. Eppstein, "Finding the k Shortest Paths" 35th IEEE Symp. Foundations of Computer Science., Santa Fe, pp. 154-165, 1994.

[Christo07] K. Christodoulopoulos, M. Varvarigos, C. Develder, M. De Leenheer, B. Dhoedt , "Job Demand Models for Optical Grid Research", Proc. of the 11th Conference on Optical Network Design and Modelling (ONDM), Athens, Greece, May 2007

[Cugini05] F.Cugini, N. Andrioli , L. Valcarenghi , P.Castoldi "Physical Impairment Aware Signaling for Dynamic Ligthpath Set Up" ECOC 2005, Vol. 4, Th 3.5.6

[Farrel06] A. Farrel, J-P. Vasseur,J.Ash " A Path Computation Element (PCE) – Based Architecture" IETF RFC 4655 , August 2006

[Garey79] M.Garey and D.Johnson, "Computers And Intractability: A Guide to the Theory of NP-Completeness", New York: W.H. Freeman, 1979

[glpk] http://www.gnu.org/software/glpk/glpk.html

[GMPLS-ARCH] E. Mannie (Ed.), "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", IETF RFC 3945, October 2004.

[G^{2}MPLS-ARCH] PHOSPHORUS WP2, "The Grid-GMPLS Control Plane architecture", deliverable D2.1.

[G^{2}MPLS-MODELS] PHOSPHORUS WP2, "Deployment Models and Solutions of the Grid-GMPLS Control Plane", Deliverable D2.6.

[Gross98] J. L. Gross and J. Yellen. "Graph Theory and Its Applications." CRC Press, Boca Raton, 1998.

[Harai97] H.Harai,M.Murata and H.Miyahara, "Performance of Alternate Routing Methods in All-Optical Switching Networks" Proc. IEEE INFOCOM Kobe Japan vol 2, pp 517-525 , April 1997

[Hayee97] M. I. Hayee, and A. E. Willner, "Pre- and Post-Compensation of Dispersion and Nonlinearities in 10-Gb/s WDM system," IEEE Photon. Technol. Lett., vol. 9, no. 9, pp. 1271-1273, 1997.

[Huang05] Y. Huang et al "Connection provisioning with transmission impairment consideration in optical WDM networks with high-speed channels" , JLT, vol. 23, no. 3, pp. 982-993 March 2005.

[Hyytia00] E. Hyytiä and J. Virtamo, "Dynamic Routing and Wavelength Assignment Using First Policy Iteration" Fifth IEEE Symposium on Computers and Communications ISCC 2000,

[HyytiaVirtamo00] E. Hyytiä and J. Virtamo, "Dynamic Routing and Wavelength Assignment Using First Policy Iteration, Inhomogeneous case" International Conference on Performance and QoS of Next Generation Networking, P&QNet2000, pp. 301-316, 2000, Nagoya, Japan

[IETF-RFC4208] G. Swallow (Ed.), "Generalize Multiprotocol Label Switching (GMPLS) User-Network Interface (UNI): Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Support for the Overlay Model", IETF RFC 4208, October 2005.

[IETF-RFC4655] A. Farrel J.-P. Vasseur, J.Ash, "A Path Computation Element (PCE) – Based Architecture", IETF RFC 4655, August 2006.

[Inoue94] K. Inoue, K. Nakanishi, K. Oda "Crosstalk and Power Penalty Due to Fiber Four-Wave Mixing in Multichannel Transmissions" J. Lightw. Technol. Vol. 12, no. 8 pp.1423-1439, Aug 1994

[Inoue92] K. Inoue, "Four-Wave Mixing in an Optical Fiber in the Zero-Dispersion Wavelength Region," J. Lightwave Technol., vol. 10, no. 11, pp. 1553-1561,Nov. 1992.

[Kikuchi97] N. Kikuchi, K. Sekine, and S. Sasaki, \Analysis of cross-phase modulation (XPM) effect on WDM transmission performance," Electron. Lett., vol. 33, pp. 653-654, 1997.

[Kokvar07] P. Kokkinos, E. Varvarigos, “Resources configurations for providing QoS in Grid computing”, CoreGrid Symposium, June 12-13, Crete, Greece, 2007

[Kompella05] K. Kompella,Y Rekhter “OSPF extensions in support of GMPLS ” IETF RFC 4203, Oct. 2005

[Krishna01-Algo] R. M. Krishnaswamy and K. N. Sivarajan: ”Algorithms for Routing and Wavelength Assignment Based on Solutions of LP Relaxations”, IEEE Communications Letters pp. 435-437, 2001.

[Krishna01-Design] R. M. Krishnaswamy and K. N. Sivarajan: ”Design of Logical Topologies: A Linear Formulation for Wavelength Routed Optical Networks with no Wavelength Changers”, IEEE/ACM Transactions on networking, vol. 9, no. 2, pp. 186-198, 2001.

[Kulkarni05] P. Kulkarni, A. Tzanakaki, C. M. Machuca, I. Tomkos “Benefits of Q-factor based Routing in WDM Metro Networks”, ECOC’05 vol.4,pp.981-982

[Lakoum07] J. Lakoumentas, K. Manousakis, E. Varvarigos, "An LP approach to impairment-constraint based RWA in WDM optical networks”, to be submitted.

[Mannie04] E.Mannie “Generalized Multiprotocol Label Switching (GMPLS) Architecture”, IETF RFC 3945, October 2004

[Markidis06] G. Markidis,S. Sygletos, A. Tzanakaki, I. Tomkos “Impairment Constraint based Routing in Optical Networks Employing 2R regeneration”, ECOC’ 06, Cannes, France

[Markidis07] G. Markidis,S. Sygletos, A. Tzanakaki, I. Tomkos “Impairment Aware based Routing and Wavelength Assignment in Transparent Long HaulNetworks”, ONDM 2007 Athens, Greece

[MarkidisPTL07] G. Markidis,S. Sygletos, A. Tzanakaki, I. Tomkos “Impairment-Constraint-Based Routing in Ultralong-Haul Optical Networks With 2R Regeneration”, IEEE PTL Vol.19, Issue.6, March15, 2007 Page(s):420 - 422

[Martins03] F. Martins-Filho J. Martins-Filho, C. Bastos-Filho, S. Oliveira, E. Arantes, E. Fontana, F. Nunes, “Novel routing algorithm for optical networks based on noise figure and physical impairments”, ECOC 2003

[Martinez06] R.Martinez, C.Pinard, J.Comellas,G. Junyent “ On-line ICBR in a transparent GMPLS network: a reality check” WGN5, 30-31 March 2006,Girona,Spain

[Mokhtar98] A.Mokhtar and M.Azizoglu ,”Adaptive Wavelength Routing in All-Optical Networks”, IEEE/ACM Transactions on Networking vol 6, April 1998

[Muche97] B. Mukherjee. Optical Communication Networking. McGraw-Hill, 1997.

[Mørk03] J. Mørk, F. Öhman, S. Bischop, “Analytical Expression for the Bit Error Rate of Cascaded All-Optical Regenerators”, *IEEE Photonics Technology Letters,* vol. 15, pp. 1479-1481 Oct. 2003

[Ozdaglar03] A. E. Ozdaglar and D. P. Bertsekas: ”Routing and Wavelength Assignment in Optical Networks”, IEEE/ACM Transactions on Networking, 11(2), pp. 259-272, 2003.

[Pachnicke03] S. Pachnicke and E. Voges, "Analytical assessment of the Q-factor due to cross-phase modulation (XPM) in multispan WDM transmission systems", in Proc. SPIE, vol. 5247, Orlando, FL, Sep. 2003, pp. 61-70.

[Phos-D5.2] Phosphorous D5.2: ‘’QoS-aware resource scheduling’’

[Phos-D5.4] Phosphorous D5.4: ‘’Support for advance reservations in scheduling’’

[Phos-D5.6] Phosphorous D5.6: ‘’Simulation Environment’’

[PHOSP-TestBed] PHOSPHORUS WP6, “Test-Bed Design” deliverable D6.1

[PHOSPHORUS-D1.1] PHOSPHORUS Deliverable D.1.1, “Requirements and specifications of interfaces and architecture for interoperability between NRPS, GMPLS, Middleware”.

[Partridge93] C. Partridge et al., “Host Anycasting Service”, IETF RFC1546, Nov 1993.

[Papadimi98] C. H. Papadimitriou and K. Steiglitz: ”Combinatorial Optimization: Algorithms and Complexity”, Dover Publications, 1998.

[Rama95] R. Ramaswami and K. N. Sivarajan: ”Routing andWavelength Assignment in All-Optical Networks”, IEEE/ACM Transactions on Networking, vol. 3, no. 5, pp. 489-500, Oct. 1995.

[Rama98] R.Ramaswami and K.N.Sivarajan, “Optical Networks: A practical Perspective” , San Francisco, Morgan Kaufmann, 1998

[Ramam99] B. Ramamurthy, D. Datta, H. Feng, J.P. Heritage, and B. Mukherjee “Impact of Transmission Impairments on the Teletraffic Performance of Wavelength-Routed Optical Networks” JLT, vol. 17, no 10, October 1999.

[Rothnie96] D. M. Rothnie and J. E. Midwinter, “Improving standard fiber performance by positioning the dispersion compensating fiber," Electron. Lett., vol. 32, no. 20, pp. 1907-1908 Sept. 1996.

[Saad04] M. Saad and Z. Luo: ”On the Routing and Wavelength Assignment in Multifiber WDM Networks”, IEEE Journal Selected Areas in Communications, 22(9), pp. 1708-1717, 2004.

[Stern90] M. Stern, J. P. Heritage, R. N. Thurston, and S. Tu, “Self-Phase Modulation and Dispersion in High Data Rate Fiber-Optic Transmission Systems," J. Lightwave Technol., vol. 8, no. 7, pp. 1009-1016, 1990.

[Stern99] T. E. Stern and K. Bala: ”Multiwavelength Optical Networks: A Layered Approach”, Prentice Hall, 1999.

[Stevens07] T. Stevens, F. De Turck, B. Dhoedt, and P. Demeester, “Achieving Network Efficient Stateful Anycast Communications,”in Proc. of the 21st International Conference on Information Networking (ICOIN 2007), Estoril, Portugal, Jan 2007.

[Stevens07a] Tim Stevens, Joachim Vermeir, Marc De Leenheer, Chris Develder, Filip De Turck, Bart Dhoedt, Piet Demeester, “Distributed Service Provisioning Using Stateful Anycast Communications”, Accepted for publication in the Proc. of The 32nd IEEE Conference on Local Computer Networks (LCN), Dublin, Ireland, Oct 2007.

[Strand01] J.Strand, A.L. Chiu and R. Tkach, “ Issues for routing in the optical layer”, IEEE Communication Magazine, Feb. 2001, pp. 81-87

[Subram97] S. Subramaniam and R.A Barry, “Wavelength Assignment in Fixed Routing WDM Networks” Proc.ICC . Montreal Canada vol 1, pp 406-410, June 1997.

[Ten99] S. Ten, K.M. Ennser, J.M. Grochocinski, S.P. Burtsev, V.L. da Silva, “ Comparison of four-wave mixing and cross phase modulation penalties in dense WDM systems”, OFC’99, vol.3 pp.43-45

[Tomkos01] I.Tomkos et al. “Filter Concatenation in Metropolitan Optical Networks Utilizing Directly Modulated Lasers” IEEE Photonics Technology Letters, Vol. 13, No. 9, Sept. 2001

[Tomkos02] I. Tomkos et al., “Dispersion map design for 10 Gb/s ultra-long haul DWDM transparent optical networks,” in Proc. OECC’02, Yokohama, Japan, July 2002,

[OGF-GFD.11] M. Roehrig (Ed.),”Grid Scheduling Dictionary of Terms and Keywords”, OGF GFD-I.11 November.2002

[OGF-GFD.44] J. Treadwell (Ed.), ”Open Grid Services Architecture Glossary of Terms ”, OGF GFD-I.11 January .2005

[OGF-GFD36] D. Simeonidou (Ed.), “Optical Network Infrastructure for Grid”, GFD-I.036, August 2004.

[OGF-GFD81] J. Treadwell (Ed.), “Open Grid Services Architecture Glossary of Terms Version 1.5”, GFD-I.081, http://forge.gridforum.org/projects/ogsa-wg, July 2006.

[Öhlen 97] P. Öhlen and E. Berglind, “Noise accumulation and BER estimates in concatenated nonlinear optoelectronic repeaters”, *IEEE Photonics Technology Letters,* vol. 9, pp. 1011-1013, July 1997

[Vazirani01] V. Vazirani: ”Approximation Algorithms”, Springer Verlag, 2001.

[Wagner96] R. Wagner, R. Alferness, A. M. Saleh, and M. S. Goodman, “MONET: Multiwavelength optical networking”, Journal of Lightwave Technology, vol. 14, no 6, pp. 1349–1355, 1996.

[Wagner00] R. Wagner, “Evolution of Optical Networking”, LEOS 2000 Proceedings, TuC1, Puerto Rico, November 2000.

[Zhang98] X. Zhang and C. Qiao, ”Wavelength Assignment for Dynamic Traffic in Multi-fiber WDM Networks,”Proc., 7th International Conference on Computer Communications and Networks, Lafayette, LA, pp. 479-485, Oct. 1998.

[Zang00] H. Zang, J. P. Jue, and B. Mukherjee, “A Review of Routing and Wavelength Assignment Approaches for Wavelength-Routed Optical WDM Networks,” Opt. Net. Mag., vol. 1, Jan. 2000.

[Zang02] K. Zang and B. Mukherjee, ”A Review of Traffic Grooming in WDM optical networks: Architectures and Challenges,” *Optical Networks Magazine*, vol. 4, no. 2, Mar./Apr. 2003.

[Zeiler96] W. Zeiler, F. Di Pasquale, P. Bayvel, and J. E. Midwinter, “Modeling of four-wave mixing and gain peaking in amplified WDM optical communication systems and networks,” J. Lightw. Technol., vol. 14, no. 9, pp. 1933–1996, Sep. 1996.

9 Acronyms

AD	Autonomous Domain
ASE	Ampifier Spontaneous Emmision
BA	Barabási-Albert
BER	Bit Error Rate
CD	Chromatic dispersion
CP	Client Proxies
CPU	Central Processing Unit
DCF	Dispersion Compensation Fiber
DMUX	Demultiplexer
DSF	Dispersion Shifted Fiber
EDFA	Erbium Doped Fiber Amplifier
E-NNI	Exterior NNI
FA	Forwarding Adjacency
FC	Filter Concatenation
FEC	Forward Error Correction
FF	First Fit
FTP	File Transfer Protocol
FWM	Four Wave Mixing
GLPK	GNU Linear Programming Kit
GMPLS	Generalize Multi-Protocol Label Switching
G^{2}MPLS	Grid-GMPLS
GUNI	Grid User to Network Interface
GVD	Group Velocity Dispersion
IA-RWA	Impairment Aware Routing and Wavelength Assignment

IAT	Inter-Arrival Times
IAWA	Impairment Aware Wavelength Assignment
ICBR	Impairment Constraint Based Routing
IETF	Internet Engineering Task Force
I-NNI	Interior NNI
IM	Intensity Modulation
LMP	Link Management Protocol
LPF	LowPass Filter
LSP	Label Switched Path
MUX	Multiplexer
NF	Noise Figure
NLSE	Nonlinear Schrödinger Differential Equation
NNI	Network to Network Interface
NRPS	Network Resource Provisioning System
NRZ	Non Return to ZERO
OGF	Open Grid Forum
OGSA	Open Grid Service Architecture
OSNR	Optical Signal to Noise Ratio
OSPF	Open Shortest Path First
P2P	Point-to-Point
P2MP	Point-to-multipoint
PCE	Path Computation Element
PMD	Polarization Mode Dispersion
PSD	Power Spectral Density
PS-IAWA	Pre-Specified Impairment Aware Wavelength Assignment
QoS	Quality of Servise
RA	Routing Area
RC	Routing Controller
RF	Random Fit
RIP	Routing Information Protocol
RSVP	Resource reSerVation Protocol
SLA	Service Level Agreement
SMF	Single Mode Fiber
SOA	Semiconductor Optical Amplifiers
SP	Shortest Path
SP	Server Proxies
SPM	Self Phase Modulation
TDM	Time Division Multiplexing
TE	Traffic Engineering
TED	Traffic Engineering Database
TLV	Type/Length/Value
XPM	Cross Phase Modulation
XT	Crosstalk

Appendix A Linear Programming Formulation to Solve the RWA problem

In this Appendix a detailed description of the Linear Formulation for Routing and Wavelength Assignment problem is presented. The following parameters are considered to be known beforehand and are given as input to the algorithm for the routing of a given set of connections.

G(V,A) a unidirectional graph, where V is the set of vertices describing the nodes of the network and **A** the set of edges describing the links of the network.

N = |V| the number of the nodes of the network.

L = |A| the number of the edges of the network.

C the set of the available wavelengths.

W = |C| the total number of the available wavelengths.

R the traffic matrix in units of lightpaths, i.e. R_{12} = **2** indicates that there are 2 connection/lightpath requests between nodes 1 and 2 of the network.

U the total number of the distinct source-destination pairs.

k the total number of paths (main and alternate/protection) that have to be selected for each request.

P the set of all paths (main and alternate/protection) of all the connections.

Z the set of all nodes that have wavelength conversion capabilities.

Q the set of all available wavelength conversions at all the nodes of the network.

$\mathbf{Q_i}$ the set of all available wavelength conversions at nodes ***i*** of the network.

$\mathbf{T_i}$ the number of wavelength converters at node **i.**

D a properly chosen piecewise linear cost function. This function is a function of flow in every link and in its general form is a piecewise monotonically increasing convex function.

a the number of the piecewise linear segments comprising the piecewise linear cost function, 1 ***<a*** < ***W.***

We also introduce the following **variables**:

$\boldsymbol{\lambda^{c}_{p,l}}$ an indicator variable that has the value of 1 when path p occupies the link l and the wavelength C and 0 in all other cases.

$x_l = \sum_{\substack{p \in P \\ c \in C}} \lambda_{p,l} = \sum_{p \in P} \sum_{c \in C} \lambda_{p,l}$ the total flow on link l

The formulation of the problem is the following

Objective

$$\min \sum_{l} D_l(x_l) \qquad \forall l \in A$$

Subject to the constrains

1. $\sum_{p} \lambda^{c}{}_{p,l} \le 1 \quad \forall l \in A, c \in C$

2. $\lambda^{c}_{p,l_i} = \lambda^{c}_{p,lj} \quad \forall p \in P \, and \, (l_i, l_j) \in p$ are successive links in p

At nodes where no wavelength capabilities are available, passing lightpaths should be using at the egress the same wavelength as at the ingress so that the wavelength continuity constraint is satisfied.

3. $D_l \ge c_i x_l + \beta_i \quad \forall l \in A \;\; and \;\; \forall i, 1 \le i \le a$

In essence, the previous constraint is the mathematical expression of

$$D_l \ge \max\{c_i x_l + \beta_i\}$$

4. $\sum_{\substack{c \\ p \in P_{sd}}} \lambda^{c}{}_{p,l_i} = R_{sd} \qquad \forall l_i$ which is the first link in p

The sum of all lightpaths departing from node *s* (source) to node *d* (destination) has to be equal to the corresponding value R_{sd}, of the traffic matrix.

$$\sum_{\substack{c \\ p \in P_{sd}}} \lambda^{c}{}_{p,l_j} = R_{sd} \qquad \forall l_j$$ which is the last link in p

The sum of all lightpaths originating from node *s* (source) to node *d* (destination) has to he equal to the corresponding value R_{sd} **of** the traffic matrix

5. $$\sum_{c} \left| \lambda_{p,l_i} - \lambda_{p,l_j} \right| \le 2T_i \qquad \forall \text{int} ermediate\ node\ i\ that\ has\ wavelength\ conversion\ capabilitie, where\left(l_i, l_j\right) \in p\ are\ succesive\ \ link\text{s in } p$$

For nodes that have wavelength conversion capabilities, the number of lightpaths exiting in a different wavelength from the one they were using when entering the node must not exceed twice the number of the wavelength converters available at the specific node. In essence, this constraint is equivalent to

$$\max\left\{\sum_{c} \pm\left(\lambda^{c}{}_{p,l_i} - \lambda^{c}{}_{p,l_j}\right)\right\} 2T_i$$

As it has be shown, the total number of both optimization variables and the constraints of the formulation is on the order of $O(N^4)$, where *N* is the total number of nodes. This is the worst case; in practice, though, it is much lower, since the topologies are seldom fully mesh, which is the assumption for $O(N^4)$.

Appendix B Linear Programming Formulation to Solve the RWA Problem when Users Demand more than one Wavelengths over the same Path

In this Appendix we present extentions to the LP formulation presented in Appendix A, when there are connection requests that demand more that one wavelength over the same paths.

Assume that we have a request of *m*=2 wavelengths. We define a new variable $\lambda^{c'}{}_{pl}$ as follows: c' is a group of two consecutive wavelengths $c' = (c_i, c_{i+1})$, $c_i \in C$. Moreover, $\lambda^{c'}{}_{pl}$ is an indicator variable, equal to 1 if path p occupies wavelengths c' of link l. Then we define new constraints:

1. The condition $\sum_p \lambda^c{}_{pl} \le 1$ is transformed to $\sum_p \lambda^c{}_{pl} + \sum_p \sum_{c'} \lambda^{c'}{}_{pl} \le 1$, where $c \subset c'$ ($c' = (c, c+1)$ or $c' = (c-1, c)$). This condition means that a wavelength c (subset of c') or a group of wavelengths c', in a link l, can be used only once by any path p.

2. $\lambda^{c'}{}_{pl} = \lambda^{c'}{}_{pl'}$ and $\lambda^{c}{}_{pl} = \lambda^{c}{}_{pl'}$, where l and l' are consecutive links in path p. This condition means that a path p must occupy the same wavelength c (or wavelengths c') through the links it traverses.

3. $\sum_{p \in P_{sd}} \sum_{c'}{}^{c'} \lambda_{pl} = 1$, where l is the first link of the path p and $\mathrm{R}_{sd} = 2$. This condition means that the connection will use only one path p and only one wavelength pair from the first link l of this path p. On the other hand if $\mathrm{R}_{sd} = 1$, then $\sum_{p \in P_{sd}} \sum_{c} \lambda^c{}_{pl} = 1$.

4. $\sum_{p \in P_{sd}} \sum_{c'} \lambda^{c'}{}_{pl} = 1$, where l is the last link in path p and $\mathrm{R}_{sd} = 2$. This condition means that the connection will use only one path p and only one wavelength pair from the last link l of this path p. On the other hand if $\mathrm{R}_{sd} = 1$, then $\sum_{p \in P_{sd}} \sum_{c} \lambda^{c}{}_{pl} = 1$.

5. $F_l \geq f\left(\sum_{p}\sum_{c} \lambda^{c}{}_{pl}\right)$ and $F_l \geq f\left(\sum_{p}\sum_{c'} \lambda^{c'}{}_{pl}\right)$

The first set of constraints implies distinct wavelength assignment. The rest of the constraint sets denote flow conservation; the second set implies also wavelength continuity whenever required, since lightpaths are conserved using the same wavelength (or group of wavelengths).

Following the same methodology we can support requests of *m* > 2.