

STRONGEST – Document

## **Deliverable D4.1**

# **“Report on implementation and demonstration plans”**

Version and status:	Version 3.0, final	
Date of issue:	11/01/2011	
Distribution:	Public	
Author(s):	Name	Partner
	Rafael Canto	TID
	Juan Fernandez-Palacios	TID
	Oscar González	TID
	Eduardo Azañón	TID
	Lars Dembeck	ALU
	Ulrich Broniecki	ALU
	Stephan Bunse	ALU
	Wolfram Lautenschlaeger	ALU
	Jens Milbrandt	ALU
	Svetoslav Duhovnikov	NSN-G
	Cyril Margaria	NSN-G
	Andrew Lord	BT
	Yu Rong Zhou	BT
	Francesco Paolucci	CNIT
	Filippo Cugini	CNIT
	Piero Castoldi	CNIT
	Luca Bincoletto	TI
	Tullio Loffredo	TI

Loris Marchetti	TI
Luis Velasco	UPC
Davide Careglio	UPC
Raul Muñoz	CTTC
Ramon Casellas	CTTC
Ricardo Martinez	CTTC
Ricard Vilalta	CTTC
Norberto Amaya	UEssex
Qin Yixuan	UEssex
Dimitra Simeonidou	UEssex
Georgios Zervas	Uessex

## Abstract

This deliverable describes the network functionalities to be implemented and experimentally validated in STRONGEST. D4.1 also outlines the architectural design of both medium and long term network prototypes.

## Table of Contents

<b>Abstract</b>	<b>2</b>
<b>Table of Contents</b>	<b>3</b>
<b>Executive summary</b>	<b>5</b>
<b>1 Introduction</b>	<b>11</b>
<b>2 Data plane solutions</b>	<b>12</b>
2.1 Multi-granular photonic node and network	13
2.1.1 Testbed description	15
2.1.2 Node architecture to be implemented	18
2.1.3 Implementation and demonstration plans	21
2.2 100 Gbit/s packet processing for the long-term scenario	25
2.2.1 Testbed and functionalities to be implemented	26
2.2.2 Implementation and demonstration plan	28
2.3 MPLS-TP and WSON integration for the mid-term scenario	29
2.3.1 Testbed description	29
2.3.2 Network functionalities to be implemented	31
2.3.3 Implementation and demonstration plans	38
<b>3 Control plane solutions</b>	<b>39</b>
3.1 MPLS-TP and WSON control plane integration	39
3.1.1 Testbed description	39
3.1.2 Network functionalities to be implemented	40
3.1.3 Implementation and demonstration plans	42
3.2 Multi-technology and multi-domain PCE interworking	42
3.2.1 Testbed description	42
3.2.2 Multi-domain PCE architecture to be implemented	47
3.2.3 Implementation and demonstration plans	49
3.3 Multi-layer algorithms	49
3.3.1 Testbed description	49
3.3.2 Algorithms to be implemented	53
3.3.3 Implementation and demonstration plans	58
3.4 Interface between Control Admission and GMPLS	60
3.4.1 Traffic monitoring/management in Terabit/s packet networks	60
3.4.2 Control plane architecture for RACS-PCE interworking	62
<b>4 List of acronyms</b>	<b>66</b>
<b>5 References</b>	<b>71</b>

6 Document History

73

## Executive summary

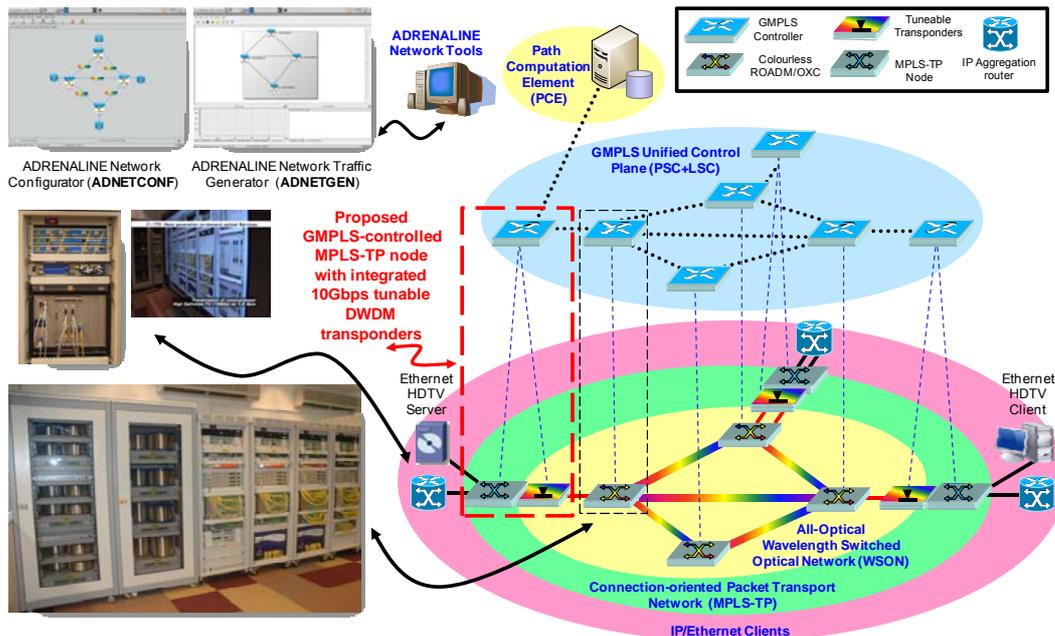
Main WP4 objectives are the implementation, integration and experimental validation of the new metro and core networking solutions designed in STRONGEST WP2 and WP3. Deliverable D4.1 describes the implementation and demonstration plans for the innovative data plane and control functions envisaged in STRONGEST.

### Data plane solutions

STRONGEST data plane implementation and demonstration will focus on two innovative and evolutionary networking approaches, the medium-term and the long-term ones, better explained in the following.

**Medium-Term networking solutions** based on the integration of wavelength switched optical networks (WSON) and connection-oriented packet transport networks (e.g MPLS-TP) for Ethernet service delivery. According to this choice, the STRONGEST mid-term data plane prototype (Figure 1) combines six key technologies, namely:

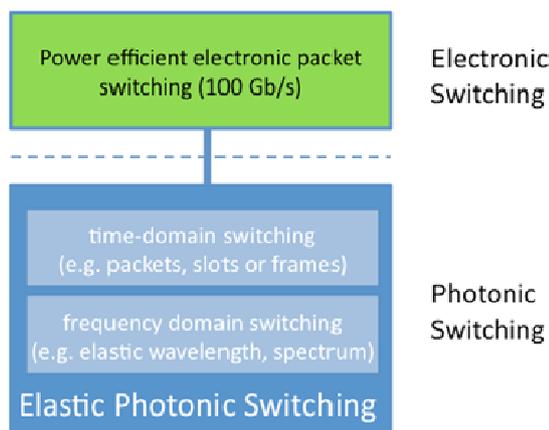
- Connection-oriented Packet Transport Network (PTN) based on Multiprotocol Label Switching - Transport Profile (MPLS-TP). This technology provides connection-oriented transport for packet services, allowing flexible packet aggregation and grooming (statistical multiplexing) by means of the electronic switching.
- Pseudo-wire emulation edge-to-edge (PWE3). This technology emulates the operation of a transparent wire carrying an Ethernet service over a packet switched network (PSN) based on MPLS-TP.
- Wavelength Switched Optical Network (WSON), providing reconfigurable high-bandwidth end-to-end optical connections, transparent to the format and payload of client signals.
- GMPLS-enabled unified control plane for MPLS-TP (packet) and WSON (lambda) transport technologies. A single control plane instance is applied in a ubiquitous way to the entire data plane switching layers within the same domain. The applicability of a single GMPLS control plane governing multiple switching technologies provides a unified control and automatic management for both LSP provisioning and recovery. This unified control plane (an enhancement, with respect to current solutions mainly based on IETF standards) will be developed in the control plane experimental tasks, and will be adopted in the data plane experiments.
- Field programmable gate array (FPGA) based reconfigurable software/hardware co-design. This technology provides a quick platform for data plane which will be able to send/receive and process the data at line speed (i.e. 10 Gbit/s). This platform benefits from the high performance of hardware and also relishes the flexibility of software.
- Software/hardware defined adaptable network (SHDAN) framework, which provides a facility for network elements to dynamically tune, update and add network function blocks based on different network/application profiles and provider/user requests.



**Figure 1: Logical view of the enhanced single-domain dual-region (MPLS-TP and WSON) mid-term testbed architecture for IP and Ethernet services.**

A first application of Multi-granular optical node architectures could support gridless elastic services, as well.

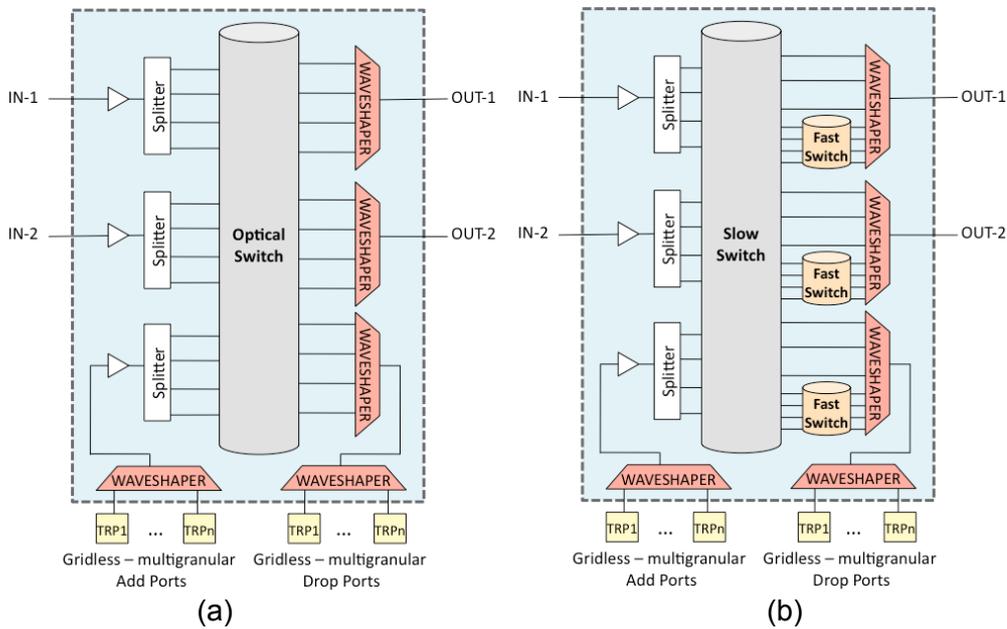
**Long-Term networking solutions** based on Multi-granular photonic nodes and power efficient ultra high capacity packet processing. STRONGEST Multi-granular photonic node to be implemented in WP4 (Figure 2) will support elastic optical circuit switching as well as enhanced network dynamics and finer (subwavelength) BW granularity enabling scalability to multiple tens of Terabit throughput per node with optimized cost and energy efficiency.



**Figure 2: Multi-granular photonic node**

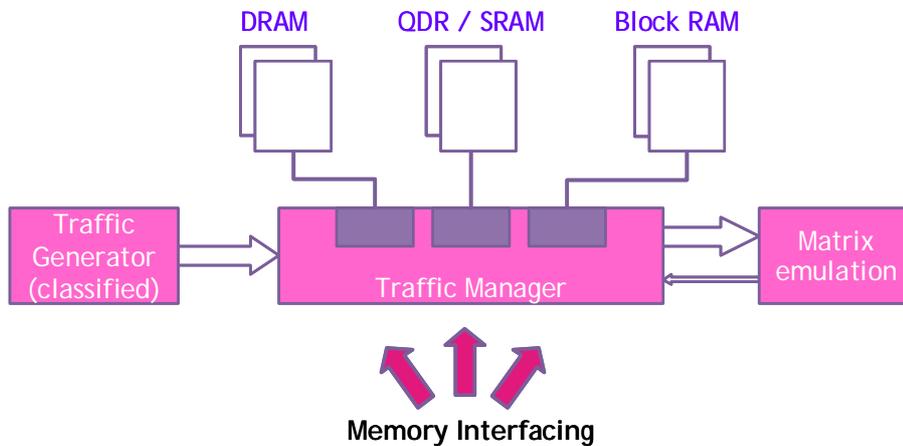
The envisaged implementation activities are closely coupled with WP2 architectural studies, where STRONGEST is following an innovative approach with different technological options targeting the required scalability and power efficiency. For instance, the photonic part of the hybrid node architecture (Figure 3) could be based on novel multi-granular optical node architectures capable of supporting gridless elastic services in the

mid-term scenario and flexible time and spectral domain allocation in the long-term scenario.



**Figure 3. (a) mid-term and (b) long term gridless multi-granular photonic node architectures**

The above gridless multi-granular photonic node could also be combined with the power efficient ultra high capacity electronic packet transport solutions designed in WP2 (Figure 4).

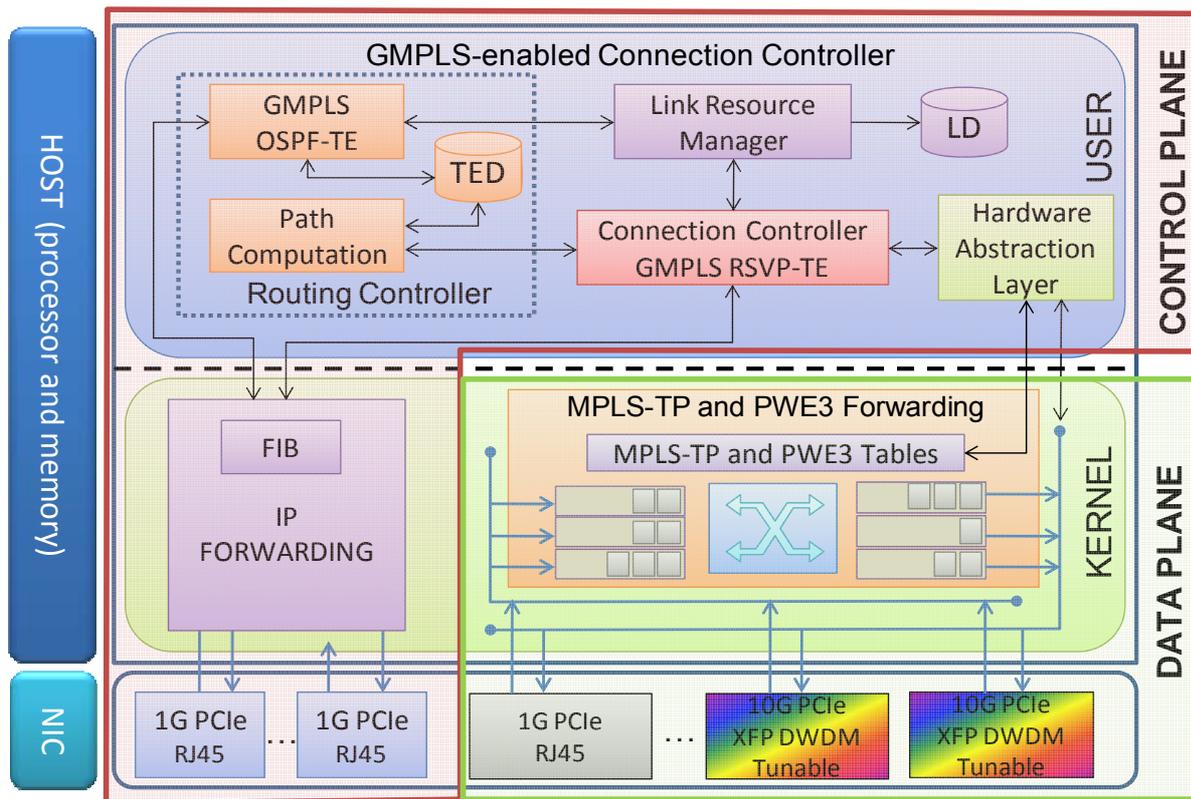


**Figure 4: 100 Gbit/s testbench scenario**

**Control plane solutions**

WP4 control plane activities, for both mid- and long-term scenarios, will be focused on the implementation of:

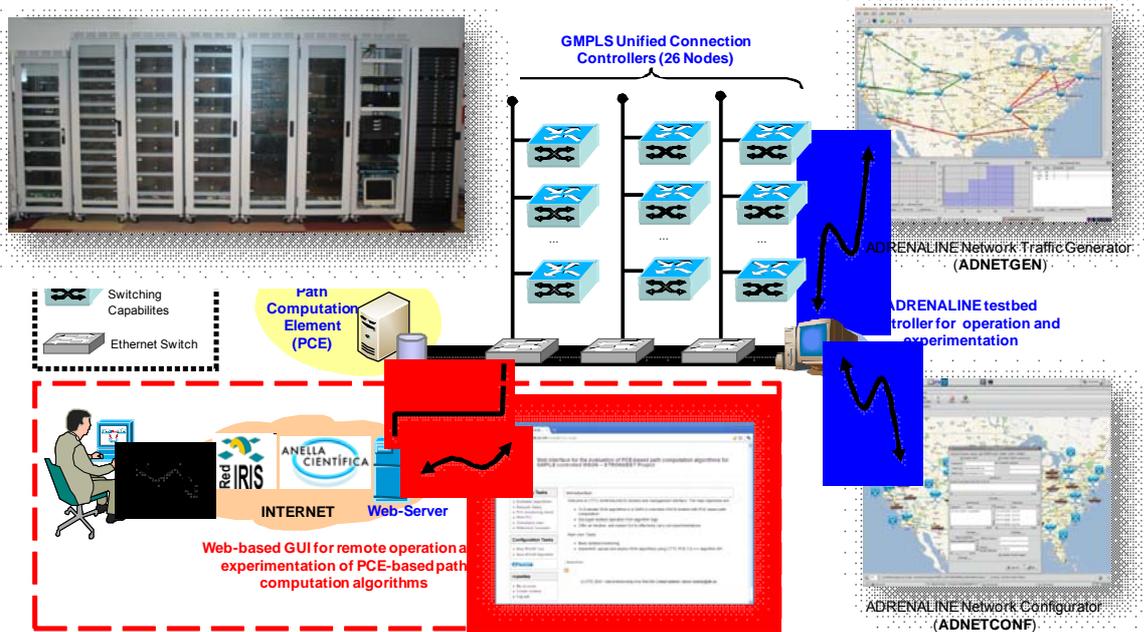
- 1.- An open, GMPLS-controlled, single-domain, dual-region (MPLS-TP and WSON) control plane according to WP3 architectural designs. Figure 5 shows the logical architecture of the proposed GMPLS-controlled MPLS-TP node with integrated 10Gbit/s tuneable DWDM transponders.



**Figure 5: Architecture for the proposed GMPLS-controlled MPLS-TP node with integrated 10Gbit/s tunable DWDM transponders**

The control plane is responsible for handling dynamically and in real-time MPLS-TP bandwidth in order to manage the establishment, maintenance, deletion and survivability of MPLS-TP connections (RSVP-TE Connection Controller), and for disseminating and discovering, from the neighbor GMPLS connection controllers, the network topology and MPLS-TP resource availability (PSC TE links) that are stored in the Traffic Engineering Database (TED) repository (OSPF-TE Routing Controller). Focusing on the peer or unified control plane interworking model, a GMPLS controller's TED contains the information relative to the WSON and MPLS-TP layers (i.e., LSC TE links and PSC TE links respectively) present in the dual-region network.

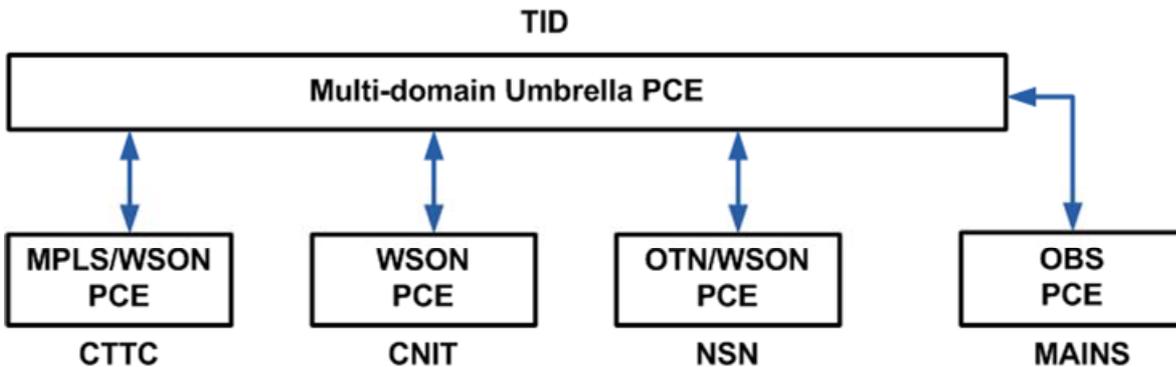
2.- Multilayer routing, grooming and restoration algorithms, as designed in WP2. The multilayer control plane architecture to be implemented in WP4 (Figure 5) will allow third parties to remotely operate the deployed testbed and to evaluate dynamic multilayer algorithms with PCE path computation through an open API (Figure 6).



**Figure 6: Open GMPLS-enabled control plane testbed**

3.- A Multi-domain and multi-technology PCE testbed (Figure 7) based on:

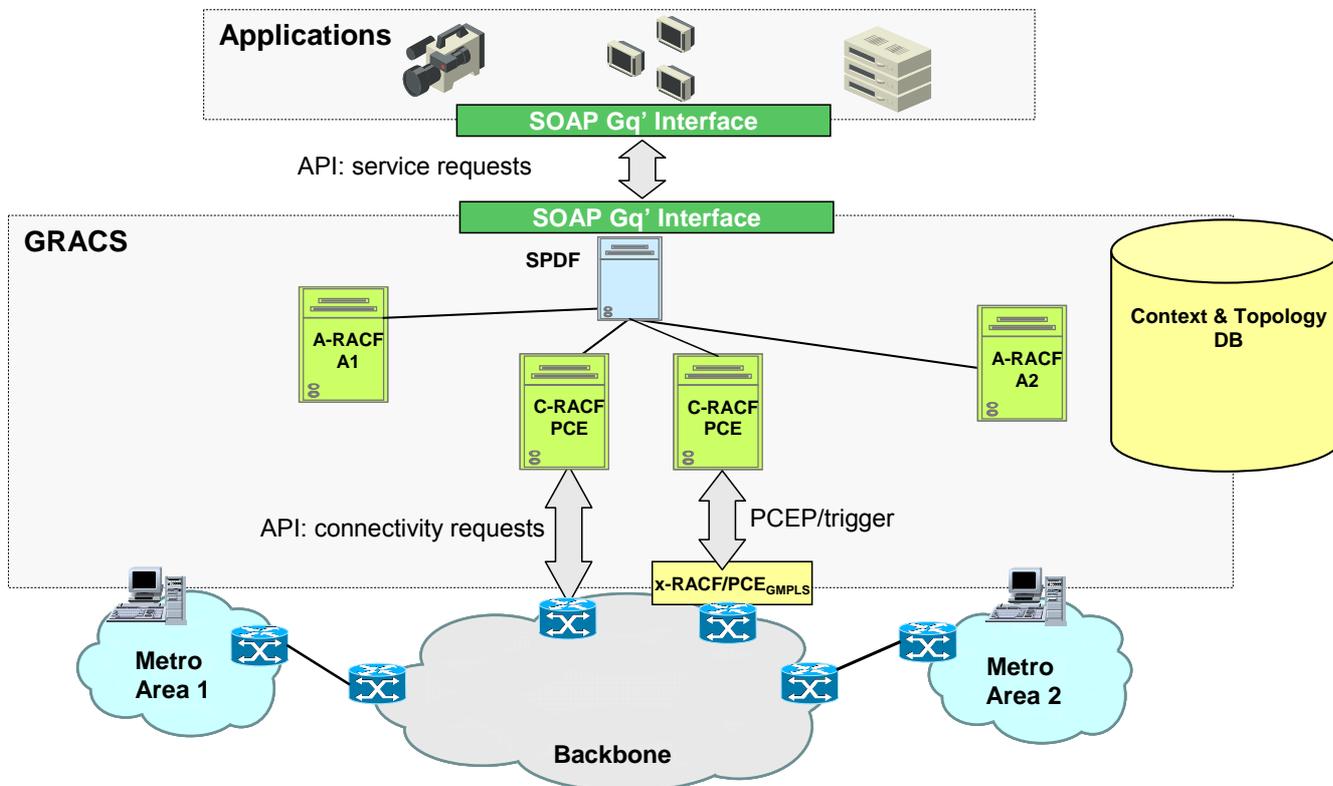
- Multiple PCE testbeds from different partners interconnected by means of IP tunnels.
- Different technologies forming several regions: MPLS, OTN, WSON, Subwavelength (cooperation with MAINS)
- Hierarchical PCE architecture designed and standardized in STRONGEST



**Figure 7: Multi-domain PCE testbed**

4.- A test-bed on "traffic monitoring/management in Terabit/s packet networks". A traffic gate for the measurement of statistical data will be implemented; this gate signals various forecast parameters to the GMPLS control plane, and particularly: actually required capacity, application stream bit rate granularity as a metric of traffic volatility, and expected packet loss ratio. Furthermore traffic admittance functions, which rely on the statistical data in the traffic gate, will be implemented. The goal is to establish Service Level Agreement (SLA) limits for various traffic parameters, detect potential violations, and feed back the traffic impairments to the originators.

5.-The RACS-PCE control plane architecture as designed in WP3. The architecture that will be implemented is depicted in Figure 8. It enhances the available RACS prototype with multi-layer routing capabilities by interacting with a GMPLS control plane based on hierarchical PCE.



**Figure 8: RACS-PCE control plane architecture**

## 1 Introduction

STRONGEST main goal is to design and demonstrate an evolutionary ultra-high capacity multilayer transport network, compatible with Gbit/s access rates, based on optimized integration of optical and packet nodes, and equipped with a multi-domain, multi-technology control plane. This innovative solution will overcome the problems of current networks that still provide limited scalability, are neither cost effective nor energy efficient, and do not properly guarantee the end-to-end quality of service.

Two specific Project work packages, WP2 and WP3, aim at designing, respectively, innovative data plane architectures, resulting in optimized transport solutions, and advanced control functions, to effectively support end-to-end services delivery across heterogeneous domains.

Beside such studies and design activities, STRONGEST is consistently committed to demonstrate that the devised solutions will actually and successfully work in the technology and market scenarios that are expected in a ten years time frame. To this aim an ad hoc work package, WP4 “Network prototypes implementation and demonstration”, is devoted to the implementation, integration and experimental validation of the metro and core networking solutions designed in WP2 and WP3.

The present deliverable D4.1 describes the Project plans for implementing and demonstrating network prototypes, based on advanced data plane functions (such as: multi-granular photonic node, 100 Gbit/s packet processing, MPLS-TP and WSON integration) and control functions (such as: MPLS-TP and WSON control plane integration, multi-technology and multi-domain PCE interworking, multi-layer algorithms, interface between control admission and GMPLS), that are envisaged by STRONGEST, in both medium and long-term prospects for networks evolution.

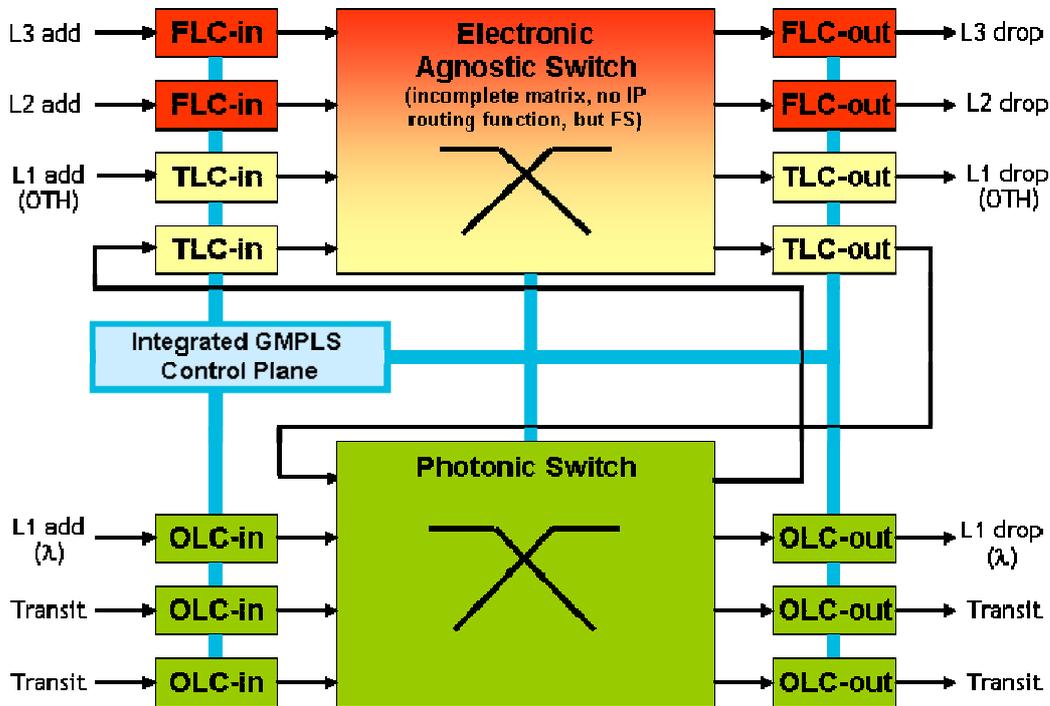
The document is therefore comprised of two specific chapters, dedicated to the plans for experiments regarding data plane and control functions, respectively. For each topic, after presenting the rationale of the technical proposal, dedicated sub-sections describe in details the experimental test bed, the novel functions and architectures to be realized, and the detailed implementation and demonstration plans.

These planned experimental activities shall be carried out and completed within the three years Project lifetime.

## 2 Data plane solutions

Current traffic forecasts predict a traffic growth by a factor of 100 over 10 years in core networks. Therefore, future node architectures have to deal with 100 Terabit/s throughput. Current IP centric network architectures will not scale to handle these huge amounts of data. Power consumption  $P$  of IP routers is increasing with capacity  $C$  according to  $P \sim C^{2/3}$ . A study for Japan [Tucker\_2008] predicts that in 2020 routers would consume 50% of the nation's 2005 total electricity generation, i.e. today's network architecture approaches are becoming unsustainable. Forecast technology improvements, e.g. CMOS technology evolving from 45nm today towards 13nm will be insufficient even to stay flat with respect to power consumption. That means that a paradigm shift is required for network and node architectures. Primary requirements for future node architectures are scalability and energy efficiency.

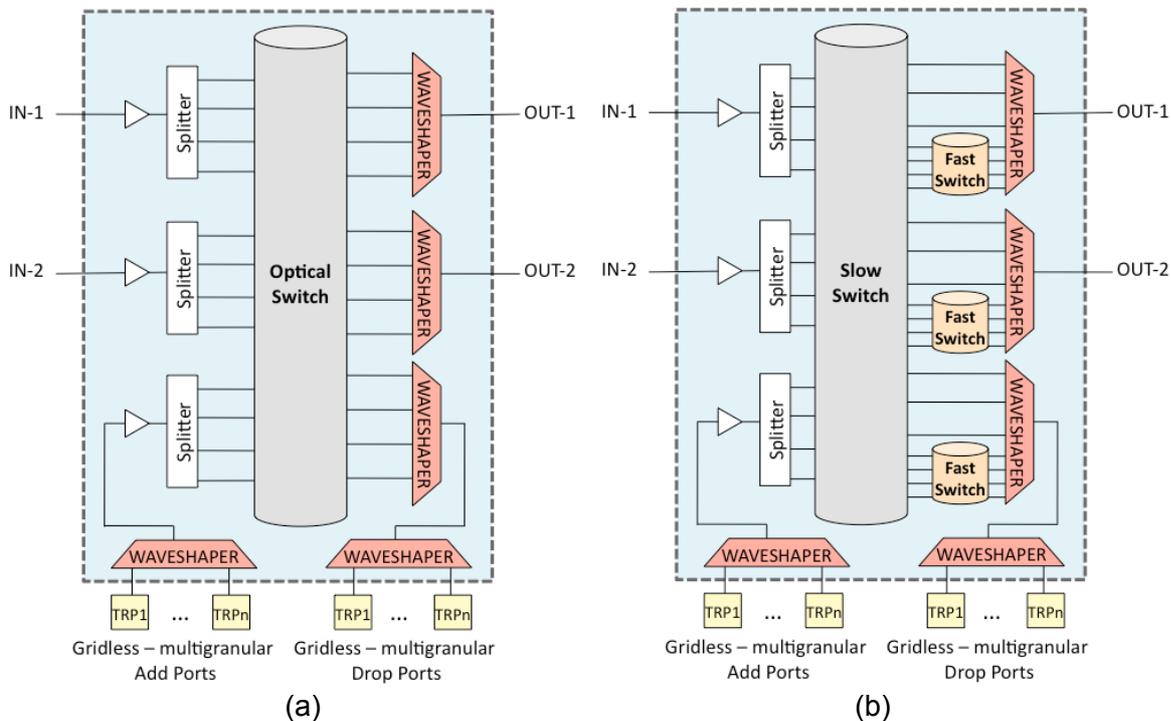
From an individual node perspective most of the traffic in a node is transit traffic (between 80% and 90%) which can be treated in the lowest possible layer, i.e. in the photonic layer. Photonic technologies can handle the same amount of data compared to electronics with only one twentieth of power consumption or, in other words, an energy reduction of 95% can be achieved. Since packet processing is still required for the remaining add/drop traffic, we are following a hybrid approach to the node architecture, which combines a small electronic part for efficient electronic packet or circuit switching with a bigger photonic part for switching wavebands or wavelength channels (see Figure 9), shifting the majority of IP processing to the edges of the network, i.e. avoiding any IP routing in the core.



FLC: Flexible, reprogrammable-Linecard, TLC: TDM-LC, OLC: Optical-LC

**Figure 9. Hybrid node architecture**

The envisaged implementation activities are closely coupled with WP2 studies, where STRONGEST is following an innovative approach with different technological options targeting the required scalability and power efficiency. For instance, the photonic part of the hybrid node architecture could be based on novel multi-granular optical node architectures capable of supporting gridless elastic services in the mid-term scenario and flexible time and spectral domain allocation in the long-term scenario (Figure 10). This novel approach is presented in section 2.1.



**Figure 10. (a) mid-term and (b) long term gridless multi-granular photonic node architectures**

## 2.1 Multi-granular photonic node and network

Optical core transmission technology is currently stabilizing around 100 Gbit/s. There are now several 100 Gbit/s products available from vendors, and most of the others will follow suit in the next year. The synchronization of Ethernet and ITU standards has enabled the component industry to harmonise development efforts at 100 Gbit/s. This standards interworking together has enabled development costs for this highly complex new technology to be minimised. Nevertheless, these costs are still extremely high and will result in an expensive end product for some time to come, especially given that the demand for this amount of capacity is still uncertain. The main attraction of the 100 Gbit/s solution is the fact that it is coherent. This coherent transmission gives the power to compensate for large amounts of chromatic dispersion and PMD as well as having a much better OSNR performance. In fact, coherent technology has enabled the 100 Gbit/s solution to perform on a par (or in some cases better) with existing, legacy 10 Gbit/s systems.

Meanwhile, 10 Gbit/s has become a highly cost effective, commodity product, with recent, continued price reductions coming from the development of pluggable components. One ramification of this is that 40 Gbit/s hasn't so far had very much industry traction: it has had to compete with a continually eroding 10 Gbit/s price. It has found application in one

or two exceptional cases where the increased additional total capacity per fibre has been a critical factor.

One interesting spin off from the 100 Gbit/s development is that we might see future 40 Gbit/s coherent products which provide the same performance as their 100 Gbit/s counterparts, but should be cheaper and certainly occupy less spectrum.

For the future, we are now seeing significant research effort exploring the next potential data rates above 100Gbit/s. There is a great deal of work looking at 200 Gbit/s, 400 Gbit/s and 500 Gbit/s. These bit rates should be able to re-use some of what has been developed for 100 Gbit/s, thereby avoiding another new development cycle, which the industry currently would not be able to afford. The R&D investment into 100 Gbit/s will take years to recoup and so it is expected that forthcoming increases to the data rate will rely on the 100Gbit/s developments that have taken place. It is felt that a 1Tb/s solution would not be able to re-use much existing technology, and essentially would have to start again. Currently there isn't a large market driver for these kind of rates, so we expect to see lower bit rate products instead.

One thing is clear: there is a wide range of requirements for transmission rates both between world regions and individual operators. A specific operator will also have a wide range of requirements in different parts of their network. Operators also carry multiple services; some of these will need to carry very high bit rate traffic (up to 100Gbit/s ) from data centres, whilst simultaneously the operator will need to carry 100 Mbit/s, 1 Gbit/s and 10 Gbit/s circuits for their customers. An operator may have legacy transmission technology, but want to upgrade in service, without having to layer a completely new system on top.

Finally we are seeing many examples of operators filling up the C band of their optical fibres. This has steered the 100 Gbit/s developments towards spectrally efficient modulation formats: 100 Gbit/s wavelengths still able to squeeze into the ITU wavelength grid, by using multi level coding. So DP-QPSK makes use of two polarisations, and then for each polarisation it uses a quadrature phase modulation (i.e. 4 levels – giving a halving of spectral width) resulting in a 4 times reduced optical spectrum. Higher bit rates will possibly explore yet more multi levels.

Although 100 Gbit/s can be highly spectrally efficient, that isn't the case for 10Gbit/s and to some extent 40 Gbit/s . The main reason for this is the ITU wavelength grid. This grid has standardised specific wavelengths, spaced by either 50 or 100Gbit/s Hz. Component manufacturers have been able to focus on making components at precisely these frequencies, and the result has been a very successful standard. Recently, there have been an increasing number of papers suggesting that the ITU grid is holding the industry back.

If we desire to mix bit rates on fibres: i.e. to run networks with 10 Gbit/s , 40 Gbit/s , 100Gbit/s and other bespoke rates then being restricted to wavelengths on the ITU grid will lead to an inefficient use of optical spectrum. We are now starting to see technologies emerge that don't need to be restricted to the ITU grid.

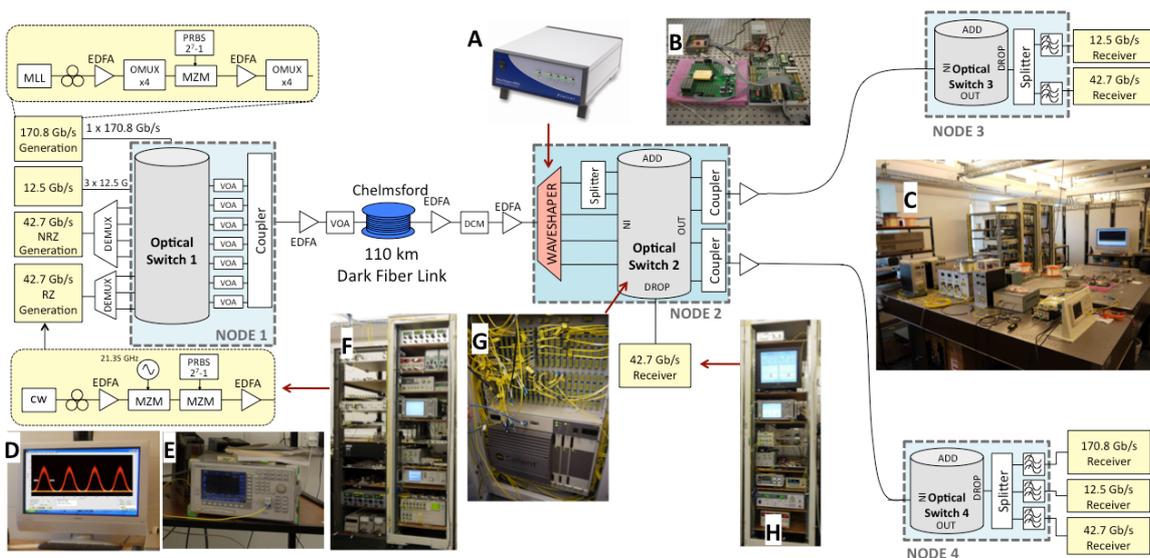
New ROADMs making use of LCOS arrays to give very fine spectral resolution (1GHz) and allowing a wide range of add/drop spectral shapes. We propose to make use of this exciting new technology in WP4 – in particular the Waveshaper being manufactured by Finisar.

Variable bit rate transmitters are a promising technology option, as well. OFDM technology is analogous to ADSL in that the aggregate baud rate for an optical channel comprises the addition of multiple sub carriers – all closely spaced. One advantage of this is that carriers can be added or removed – thus changing the overall bit rate. Therefore, if a smaller bit rate is needed, then this is all that is used, thus freeing up additional spectrum for other channels.

Unsurprisingly, there are significant challenges with this new concept of gridless or elastic transport: challenges arising from the components to the transmission performance to the overall network control. Just in the last year we have seen a dramatic increase in the number of published papers and research teams looking into this (2 years ago it was virtually zero).

STRONGEST is committed to a thorough study of this technology for its long-term data plane work. The details of our approach are described below. There is a high degree of novelty in what we are proposing here. The main concept of Multi-granularity seeks to impose a control onto the concept of elastic networks, by giving some separacy to different services, thereby making it easier to guarantee performance as well as design practical switches. The following sections describe our ideas and work so far.

### 2.1.1 Testbed description



**Figure 11: Flexible multi-granular testbed (University of Essex)**

The long-term data plane experimental testbed, located at the University of Essex, is depicted in Figure 11 and a general view of the lab facilities utilized in STRONGEST is shown in Figure 11 inset C. The configuration consists of a source node (Node 1), a transit node (Node 2) and two destination nodes (Node 3 and Node 4). Nodes 1 and 2 are connected via a 110 km field fiber link with full dispersion compensation. Node 2 is then linked to Nodes 3 and 4 using local fiber in the lab. At Node 1, the transmitter (Figure 11 inset F) generates signals of multiple bit rates and modulation formats as described in the next section. They are subsequently sent over the field fiber link (University of Essex Lab – Chelmsford) to Node 2 where they are either dropped or routed to different destination nodes. At Node-2, a four-port Finisar WaveShaper (Figure 11 inset A) is used as a dynamic gridless WSS, and a 3D-MEMS-based optical switch (Figure 11 inset G) is used to provide the optical routing function. Instrumentation equipment is used to monitor the signals in the time domain with a 1-ps optical sampling scope (Figure 11 inset D) and

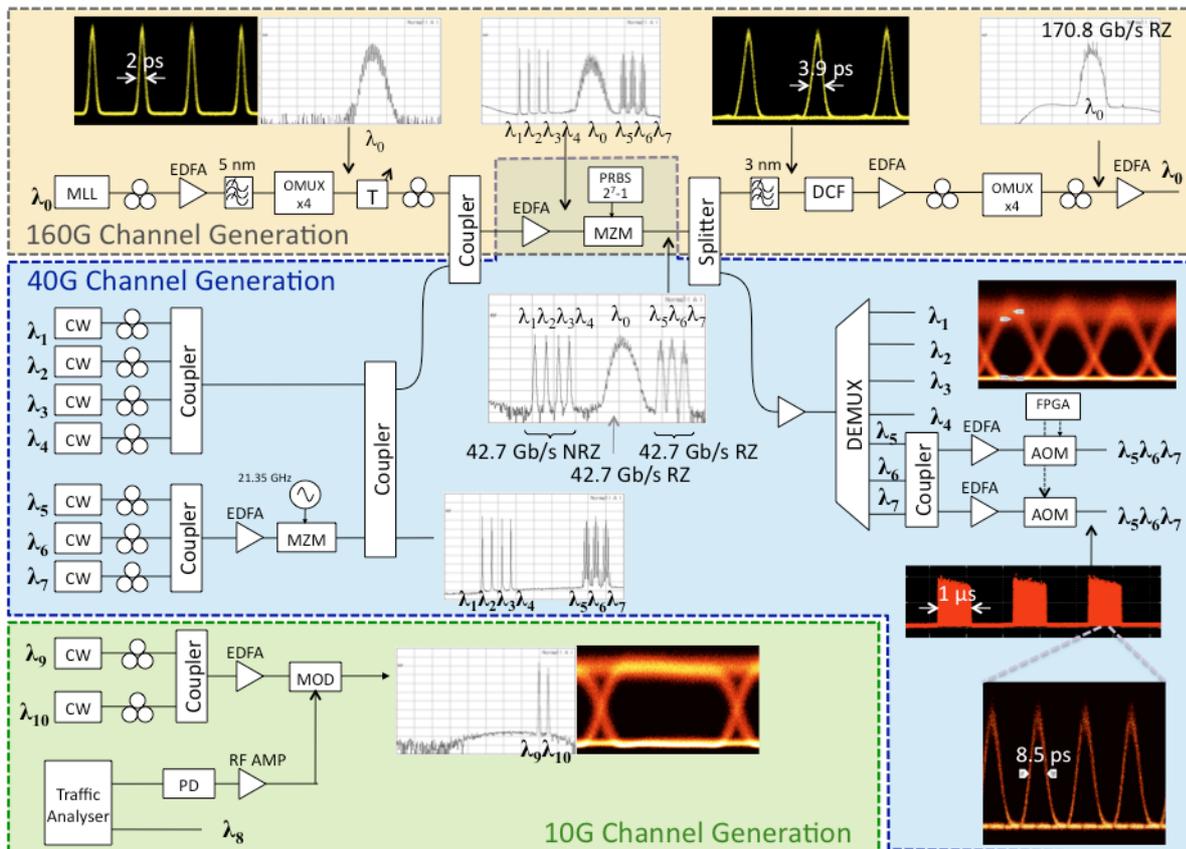
spectral domain Optical Spectrum Analyser (Figure 11 inset E). At the system output signal integrity is assessed by means of OSNR and BER measurements performed by the 160G/40Gbit/s receiver (Figure 11 inset H).

Although this is the initial testbed configuration it is possible to change the setup to demonstrate different aspects of the gridless multi-granular photonic node/network architectures. For instance, fast switches (Figure 11 inset B) may be introduced to provide the functionality of sharing spectral resources in the time domain.

### **2.1.1.1 Multiple bit rate and modulation format transmitter**

The transmitter configuration, as shown in Figure 12, is made up of three sections, one for each bit rate (10Gbit/s, 40Gbit/s, 160 Gbit/s). In general, these sections are independent with the exception of the MZM data encoder, which is shared between the 160 Gbit/s and 40Gbit/s channel generation. With this setup it is possible to generate the following channels: one 170.8 Gbit/s RZ, four 42.7 Gbit/s NRZ, three 42.7 Gbit/s RZ carrying 1 $\mu$ s bursts and three 12.5 Gbit/s. This capability is not static, however, as the 42.7 Gbit/s channels can be easily reconfigured between NRZ and RZ modulation formats and additional 10 Gbit/s channels may be added if required. Also, all the channel centre wavelengths are tunable with the exception of one 10 Gbit/s channel generated directly from the traffic analyzer ( $\lambda_8$ ). As a result of this tunability, with this transmitter it is possible to generate suitable signals for operation in a gridless environment.

The 160 Gbit/s channel generation consists of a Mode-Locked Laser (MLL), 5 nm filter, 4x Optical Multiplexer (OMUX), data encoder and dispersion compensating fibre (DCF) and second OMUX. The MLL is driven at 10.675 GHz and generates a stream of 2-ps pulses at the same rate. These pulses are amplified, filtered (to limit their spectral bandwidth) and input to the 4x OMUX. Here, the pulse rate is multiplied by a factor of four by generating two copies of the original pulse stream, introducing delay to one of the copies and recombining them again to obtain a pulse stream with double the rate. The process is repeated twice inside the OMUX so the pulse rate is multiplied by a factor of four to produce a pulse stream of 42.7 Gpulse/s. Next, the signal is modulated with a 42.7 Gbit/s pseudo-random bit sequence of length  $2^7-1$  (PRBS7) using a LiNbO<sub>3</sub> Mach-Zehnder Modulator (MZM). This MZM data encoder is shared between the 160 Gbit/s and 40 Gbit/s channel generation sections hence at the MZM output the wavelength from the MML is extracted for subsequent processing with a 3 nm flat-top filter. Then, it is passed through 18m of DCF to compensate for dispersion, amplified and input to a second OMUX which once more multiplies the bit-rate by four to generate a final bit rate of 170.8 Gbit/s. Throughout this section polarisation controllers are used before polarization-sensitive devices e.g. OMUX, MZM and a variable delay line is used before the MZM data encoder to synchronise the pulse stream and corresponding modulating bit stream.



**Figure 12:160 Gbit/s, 40 Gbit/s and 10 Gbit/s transmitter configuration (University of Essex testbed)**

The 40Gbit/s channel generation section consists of continuous wave (CW) lasers, polarisation controllers, couplers, a MZM pulse carver, a shared MZM data encoder and a demultiplexer. It produces seven 42.7 Gbit/s channels that can be easily switched between the RZ or NRZ modulation formats. The 42.7 Gbit/s NRZ channels are generated by directly modulating four CW lasers at different wavelengths with a PRBS7. On the contrary, for the 42.7 Gbit/s RZ signals a MZM driven at 21.35 GHz is used as a pulse carver before the process of data encoding. Both 42.7 Gbit/s RZ and NRZ signals are coupled together with the MLL pulses so that the same MZM can be used for data encoding at 42.7 Gbit/s. At the output of the MZM modulator individual wavelengths are extracted using either a fixed 200-GHz DEMUX for all the signals or a combination of 200-GHz fixed and 100-GHz tunable DEMUX for the RZ and NRZ signals respectively. Finally, the 42.7 Gbit/s RZ channels are combined using a 4x2 coupler, and the two copies of the combined channels are further amplified and passed through corresponding Acousto-Optic Modulators (AOM) controlled by an FPGA to generate sub-wavelength channels of 1μs-long bursts of data.

The 10Gbit/s channel generator consists of an Anritsu traffic analyzer (MD12306), photodetector (PD), 10Gbit/s RF amplifier, 10Gbit/s modulator, CW lasers and polarization controllers, 3 dB coupler and EDFA. The traffic analyzer is used to generate two 12.5 Gbit/s NRZ signals modulated with PRBS23. One of these signals is input to a photodetector and the resulting electrical output is amplified and used to drive the 10Gbit/s modulator. The optical input to the modulator consists of two CW wavelengths coupled together and amplified. At the modulator output two copies of the original 12.5 Gbit/s channel are generated at wavelengths that can be arbitrarily tuned. This feature may be used, for instance, for demonstrations where the 10 Gbit/s channels are required to have

channel spacing lower than the standard 50 GHz ITU grid. Multiple bit rate and modulation format receiver

The two-stage receiver, depicted in Figure 13, comprises a variable attenuator, isolator, a first amplification and 2.4 nm filtering stage followed by a second amplification stage and 0.8 nm filter, a splitter, clock recovery sub-system (CR), 40 Gbit/s photo-detector (PD), 40Gbit/s electrical de-multiplexer (DEMUX) and error detector (ED). This configuration is used to directly measure BER for the 42.7 Gbit/s channels and by adapting it (adding an EAM to its front-end and reconfiguring the clock recovery system) it is also used for measuring BER for the 170.8 Gbit/s channel. The variable optical attenuator at the front end is used to vary the input power to the receiver. The two amplification and filtering stages compensate for these power variations so a rather constant power is obtained after the second filter (0.8 nm). Then, a splitter is used to generate copies of the optical signal, which are input to a power meter, the CR and 40 Gbit/s data PD. The latter two in turn generate electrical signals that are subsequently input to the BERT set made up of a 40 Gbit/s electrical DEMUX and error detector (ED). BER readings are taken from the ED for different levels of signal attenuation and both power and BER are recorded. Thus, sensitivity measurements and BER curves can be produced to evaluate the integrity of the optical signal. A number of devices are also used for monitoring the signal and optimising the performance of the receiver i.e. power meters, optical sampling scope, OSA.

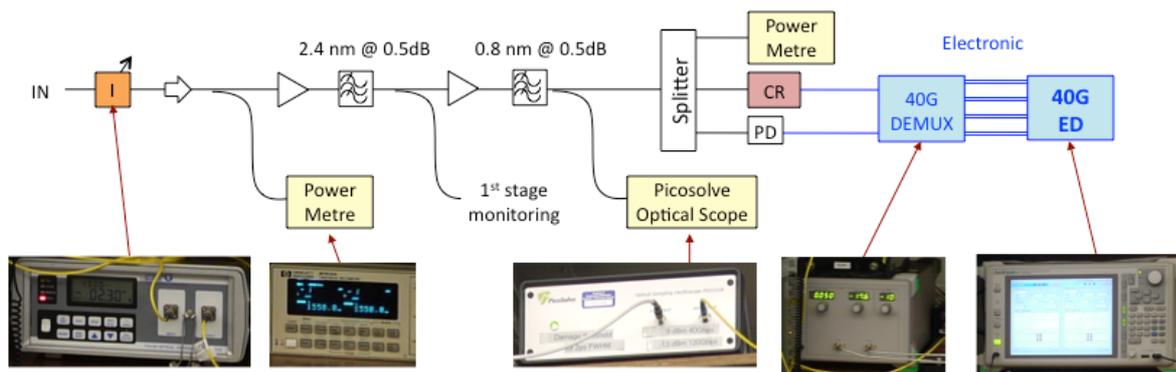
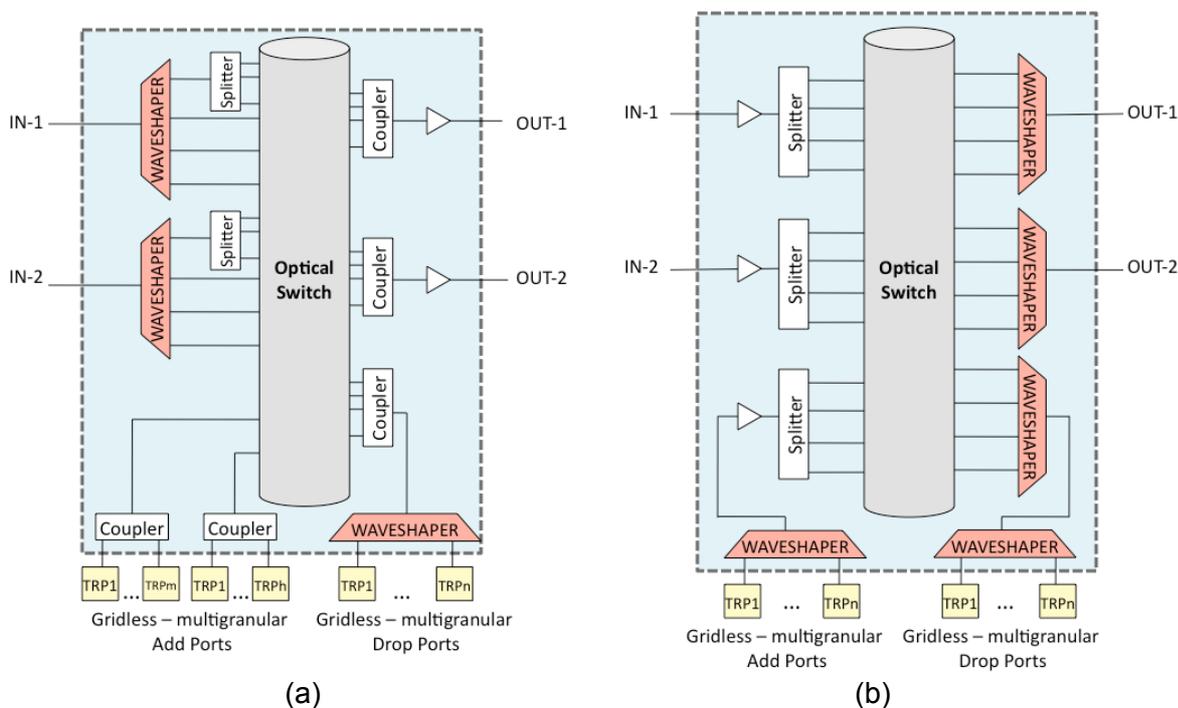


Figure 13: 160Gbit/s 40Gbit/s receiver and BER measurement (University of Essex testbed)

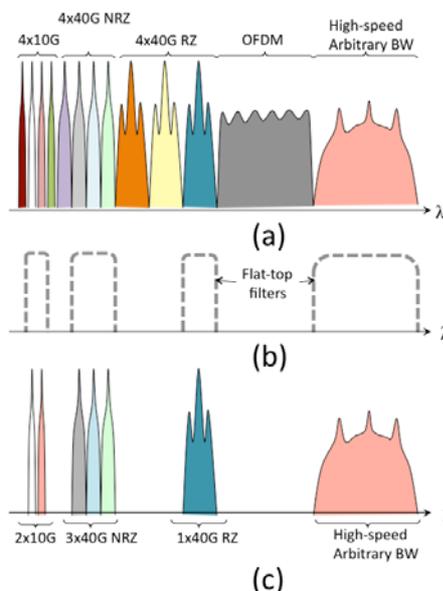
### 2.1.2 Node architecture to be implemented

The optical node architectures to be implemented are depicted in Figure 14 and Figure 16. They support flexible spectrum switching whereby spectral resources are dynamically allocated. In the first architecture, shown in Figure 14a, incoming signals (Figure 15a) are input to a programmable Waveshaper that features a fine granularity, e.g. 1 GHz resolution with a minimum 12.5 GHz spectral width. Here, customized logical filters (LCOS) with arbitrary bandwidth and centre wavelength are configured to de-multiplex a combination of signals onto different output ports. As shown in Figure 15b, several independent logical filters are supported on a single output, thus, multiple wavelengths or wavebands can be extracted from the input signal and conveyed on a single Waveshaper output port irrespective of their centre wavelength or bandwidth (Figure 15c). Therefore, a substantial level of flexibility is achieved together with a significant scalability improvement over traditional networks including waveband-switching networks. Then the Waveshaper outputs are connected to an optical switch where their respective spectrum is routed through to the output fibres. It is important to mention that the Waveshaper, as a Gridless WSS, allows for any combination and number of frequency bands (12.5 GHz spectral width and 1 GHz

resolution) to be routed to any of the 4 demux output ports. However once a band is used on a specific port (one out of four) it can't be internally routed at the same time-period (multicast) to the other ports. To provide multicasting functionality, an optical splitter is used to replicate the signals from a Waveshaper port onto several switch inputs. At the node outputs signals with a common destination are combined, amplified and transmitted.



**Figure 14. Architectures for gridless multi-granular (elastic) networks**

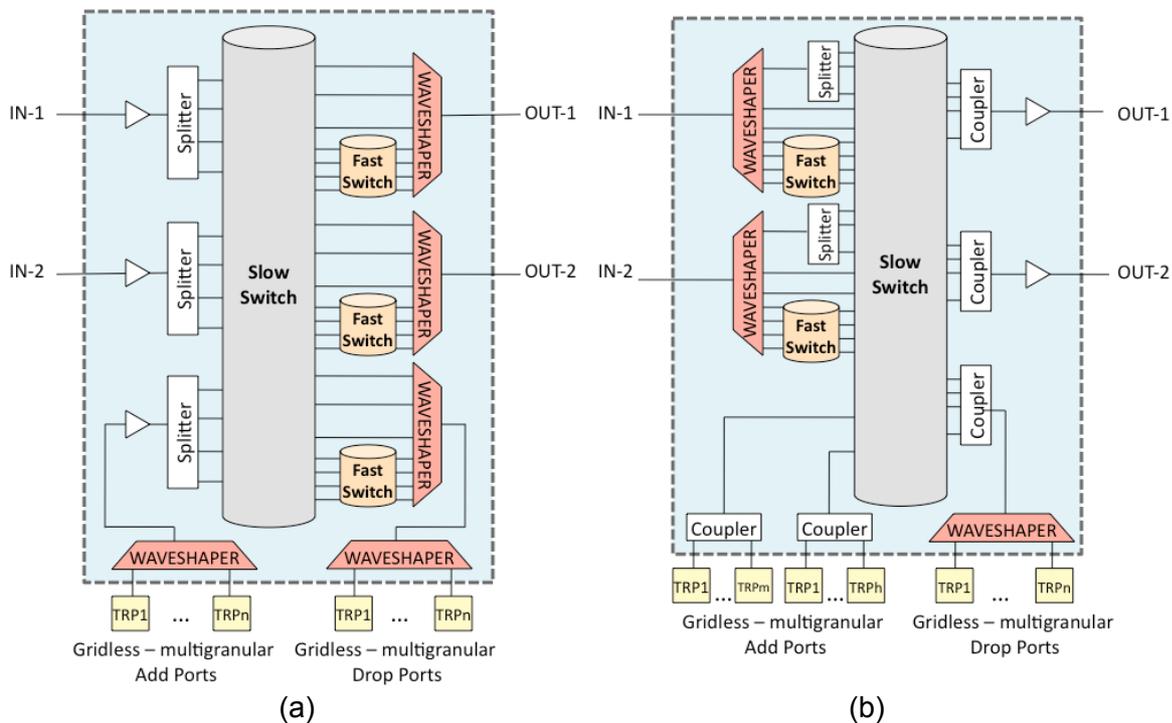


**Figure 15: Principle of operation of gridless multi-granular (elastic) demultiplexing**

The second architecture, shown in Figure 14b, is a partial broadcast-and-select arrangement where the distinct Waveshaper ports are used as inputs, connected to the optical-switch-array outputs, and the common port is connected to the output fiber. Here, replicas of the incoming signals can be switched from the optical switch inputs to the inputs

of the Waveshaper, where wavelengths destined to their corresponding fiber are passed while the rest are blocked. Hence, multicasting is naturally supported as wavelengths from a single input can be passed onto several outputs by appropriately configuring their respective Waveshaper ports. Dynamic routing of arbitrary optical spectrum is achieved by appropriately configuring the spectral shape of the Waveshaper ports and the cross-connections in the optical switch.

The network flexibility provided by these architectures may be used, in a heterogeneous network, to seamlessly accommodate wavelengths carrying high-speed traffic (e.g. 100 Gbit/s and beyond) and to increase channel density for low-speed wavelengths (e.g. 10 Gbit/s ) while maintaining compatibility with existing channel plans. Although wavelengths can operate on a completely gridless manner, for practical implementation and management, a finer grid than the standard ITU grid may be used i.e. 25 GHz or 12.5 GHz.



**Figure 16: Architectures for elastic time and spectral domain allocation**

The addition of fast switches to these photonic node architectures (as shown in Figure 16) enables sharing spectral resources in the time domain. Thus, waveshapers are used to allocate spectral resources and fast switches used to manage the time allocation of these resources. Flexible time and spectral domain allocation is achieved by modifying the transfer function of the Waveshaper ports connected to the fast switch and by controlling the cross-connections in the fast switch. The fast switch can be treated as a single 4x4 element or as four individual 1x1 (optical on/off gates) switches to realize different architectural configurations.

Several architectures may be produced depending on the way fast and slow switches are arranged. For instance, fast switches may be connected in series with the slow switch and placed either at the input side or at the output side of the node. Also, a parallel arrangement may be produced whereby signals are filtered, by the Waveshaper, to/from either the fast or slow switch. Hence, different architectures are possible, each with different characteristics and hardware requirements.

## 2.1.3 Implementation and demonstration plans

### 2.1.3.1 Fast switch evaluation

- Participants: University of Essex
- Status: Survey of fast switches available in the market and testing of three devices has been carried out (see section 2.1.1)..
- Plans: Sourcing 4x4 and 2x2 switches for STRONGEST implementations / demonstrations
- End: December 2011

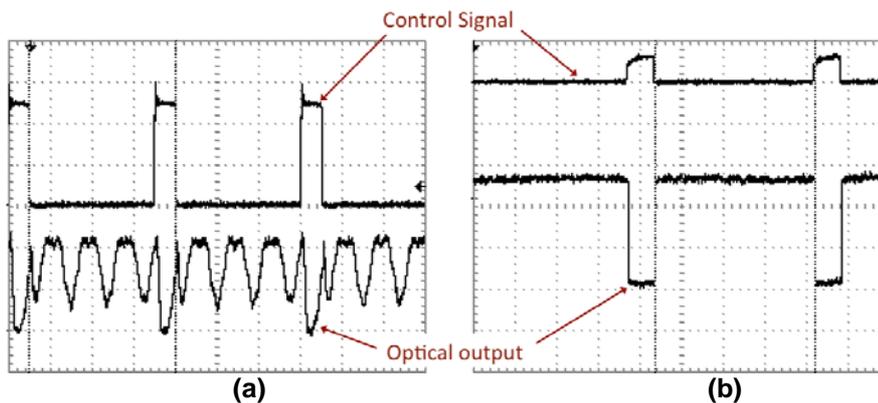
A survey of fast switches available in the market has been carried out and the main results are shown in Table 1.

**Table 1: Fast switches available in the market**

Company	Technology	Size	IL (dB)	XT (dB)	PDL (dB)	Speed (ns)	Repetition Rate	Switch voltage (V)	Control signal (V)	Power Consumption (W)	
										Switch	Including driver
Agiltron	Solid-state all-crystal	2x2	1.3	20	0.35	300	300 kHz	400	5	NIP**	12
Agiltron	Solid-state all-crystal	1x2	1	20	0.35	300	300 kHz	400	5	NIP**	12
Bati	Opto-ceramic	1x2	1	20	0.2	50	1 MHz	200	5	0.5	7.2
Bati	Opto-ceramic	2x2	1	20	0.2	50	1 MHz	200	5	0.5	12
Brimrose		1x2	4	40	NIP**	200	NIP**	NIP**	5	1	55
Eospace	Lithium niobate	1x2, 2x2 s/mod	4		SP	0.01	Ext driver	5	Driver dep	NIP**	Driver not provided
Eospace	Lithium niobate	1x2, 2x2	4	20	SP	10	Ext driver	5	Driver dep	NIP**	Driver not provided
Eospace	Lithium niobate	1x2, 2x2	4	18	NIP**	10	Ext driver	15	Driver dep	NIP**	Driver not provided
Eospace	Lithium niobate	1x2, 2x2	3	18	NIP**	10	Ext driver	15	Driver dep	NIP**	Driver not provided
EpiPhotonics	PLZT	2x2	6	25	1	10	Ext driver	10	3.3	0.025**	9.5
EpiPhotonics	PLZT	4x4	10	25	1	10	10 MHz	10	3.3	0.1**	9.5

\* Information not provided by the manufacturer  
 \*\* 1 MHz repetition rate  
 SP: Single polarisation

Three of these fast switches were tested at the University of Essex lab namely Bostonati's (BATI) 2x2 50ns switch [Jiang\_2004] and EpiPhotonic's 2x2 and 4x4 10ns switches [Nashimoto\_2010]. Bati's switch shows resonance at some repetition rates close to 1MHz caused by the piezoelectric effect (Figure 17a). As a result, the optical signal suffers excessive attenuation at regular time intervals and unacceptable penalty is introduced. EpiPhotonics 2x2 and 4x4 switches showed a stable behaviour at all frequencies tested (Figure 17b). Also, the power consumption of the 2x2 switch is the lowest available at 25 mW at a 1-MHz repetition rate. However, the greatest contribution to power consumption is from the electronic driver, with a total value of 9.5 W measured in the lab. One additional advantage of the EpiPhotonics switches is that they are controlled with a LVTTTL (3.3V) signal hence they can be directly controlled by an FPGA without the need for an external voltage translator.



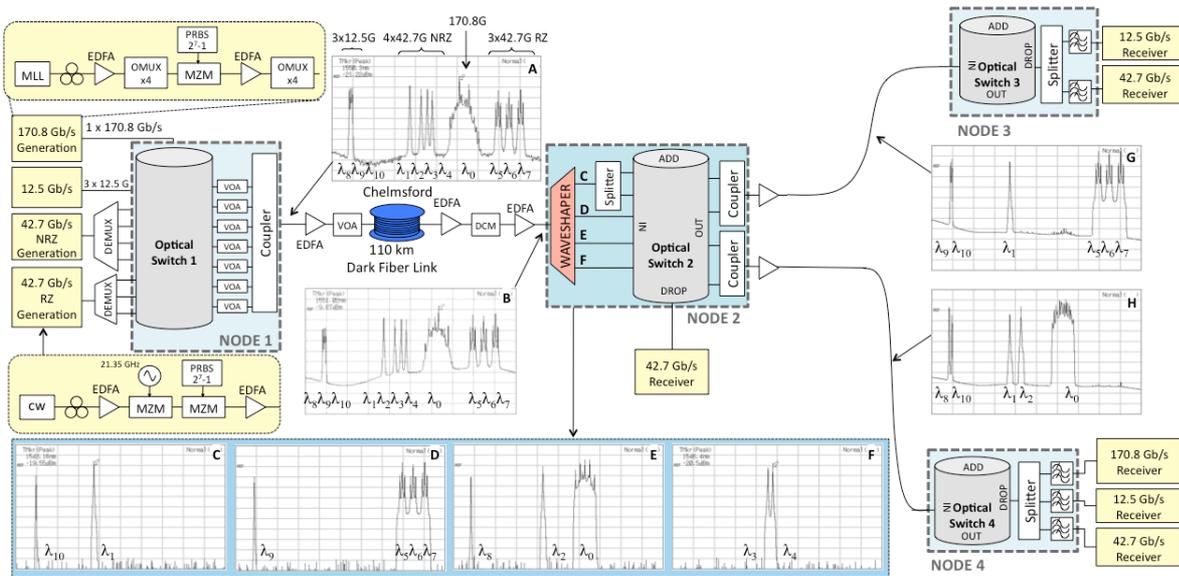
**Figure 17: Control signal and optical output for (a) Bati's 50ns 2x2 switch when resonance occurs and (b) EpiPhotonic's 10ns 2x2 switch.**

### 2.1.3.2 Waveshaper performance evaluation as gridless WSS

- Participants: University of Essex, British Telecom
- Status: Preliminary implementation shows feasibility of using the device as gridless WSS.
- Plans: Additional tests required in order to determine limits for acceptable performance e.g. minimum channel spacing and channel bandwidth required/supported for various bit rates and modulation formats.
- End: December 2011

### 2.1.3.3 Gridless Multi-granular node

- Participants: University of Essex, British Telecom
- Status: Initial implementation of transmitter, receiver and network with a single gridless multi-granular node completed and results have been submitted for publication.
- Plans: Sourcing own waveshapers in order to continue with implementation of a larger node evaluating several architectures and/or network with several gridless multi-granular nodes.
- End: December 2011



**Figure 18: Initial gridless multi-granular demonstration results (University of Essex, BT)**

An initial implementation of a gridless multi-granular network has already taken place at the University of Essex, in collaboration with BT, using the testbed setup presented in Figure 11. Results from this demonstration have been submitted for publication [Amaya\_2011].

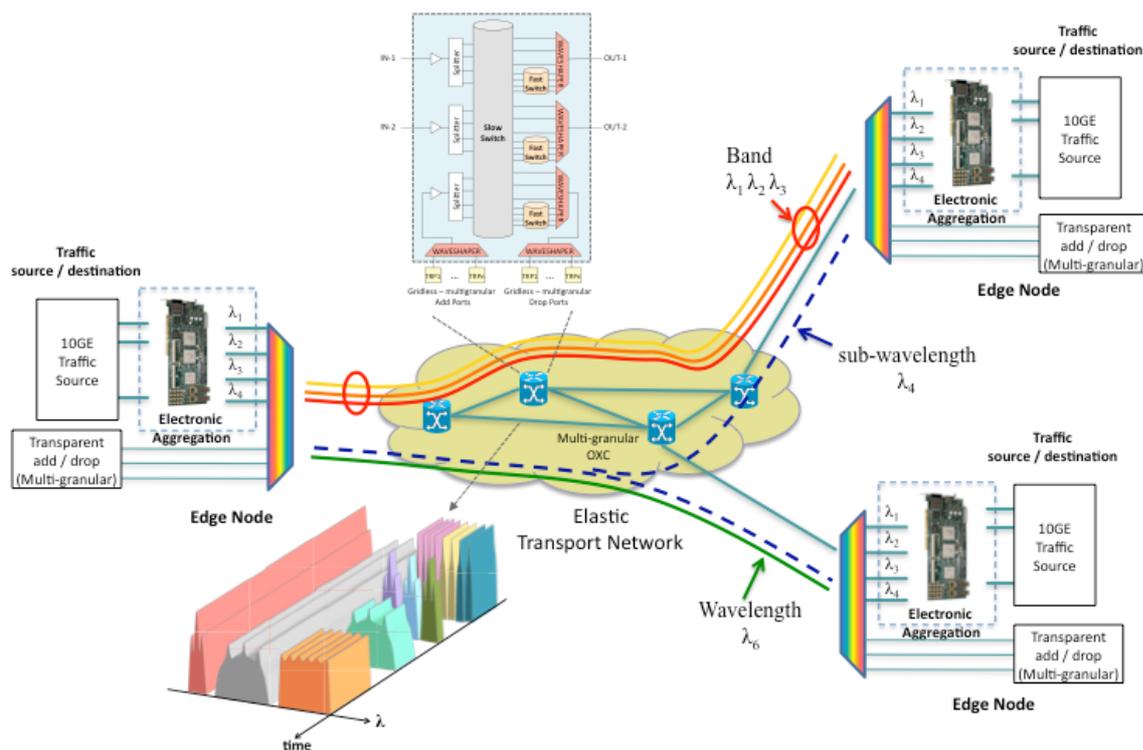
In the demonstration, eleven wavelength signals with different bit rates and modulation formats are generated at the source node (Figure 18 inset A), a 170.8 Gbit/s RZ signal with a bandwidth of 350 GHz ( $\lambda_0=1550.52$  nm), four 42.7 Gbit/s NRZ signals with centre wavelengths 25 GHz off the 100-GHz ITU grid ( $\lambda_1=1543.13$  nm,  $\lambda_2=1544.73$  nm,  $\lambda_3=1545.52$  nm,  $\lambda_4=1546.32$  nm), three 42.7 Gbit/s RZ signals with a wavelength spacing of 200 GHz ( $\lambda_5=1555.75$  nm,  $\lambda_6=1557.36$  nm,  $\lambda_7=1558.98$  nm) and three 12.5Gbit/s NRZ signals with a spacing of 25 GHz ( $\lambda_8=1534.44$  nm,  $\lambda_9=1534.64$  nm,  $\lambda_{10}=1534.83$  nm). The 12.5 Gbit/s NRZ signals are modulated with a pseudo-random bit sequence of length  $2^{23}-1$  (PRBS23) and the others are modulated with PRBS7. Next, all wavelength signals are added to Node 1, where they are switched through OXC1, combined at the output and transmitted over the field-fiber link to Node 2. At Node 2 (Figure 18 inset B), the composite signal is input to the WaveShaper, where wavelengths/wavebands are separated onto output ports according to their destination and traffic type by software-programming an appropriate spectral transfer function for each port. In order to improve spectral efficiency and optimize the channel performance, a custom bandwidth is allocated for each wavelength/waveband: for the 170.8 Gbit/s, 42.7 Gbit/s NRZ and 12.5 Gbit/s signals flat-top filters with respective bandwidths of 400 GHz, 100 GHz and 20 GHz are used, and the 42.7 Gbit/s RZ signals ( $\lambda_5-\lambda_7$ ) are switched as a waveband with a flat-top filter of 600 GHz bandwidth.

At the WaveShaper output port 1 (Figure 18 inset C) are wavelengths  $\lambda_1$  and  $\lambda_{10}$ , which are the multicast traffic toward Node 3 and 4. Wavelengths  $\lambda_5$ ,  $\lambda_6$ ,  $\lambda_7$  and  $\lambda_9$  going toward Node 3 are output on port 2 (Figure 18 inset D). The WaveShaper port 3 outputs signals at  $\lambda_0$ ,  $\lambda_2$  and  $\lambda_8$  going further to Node 4 (Figure 18 inset E), and wavelengths  $\lambda_3$  and  $\lambda_4$  at output port 4 (Figure 18 inset F) are dropped at Node 2. The multicast wavelengths, at port 1, are replicated using a 3dB optical splitter and the resulting copies are input to separate ports of OXC2 where they are switched to different destinations. The unicast wavelengths and one copy of the multicast wavelengths are routed through OXC2, combined using a 3dB coupler and transmitted to either Node 3 (Figure 18 inset G) or Node 4 (Figure 18 inset H).

### 2.1.3.4 Long-term photonic node architecture and elastic transport network

- Participants: University of Essex, British Telecom
- Status: University of Essex and British Telecom labs are already connected
- Plans: sourcing fast switches and waveshapers that will be used in the implementation of long-term prototype.
- End: March 2012

An initial proposed configuration for the long-term architecture demonstration is depicted in Figure 19. The aim is to demonstrate an elastic network made up of long-term nodes with the time and frequency switching capability. The elastic transport network will enable sharing resources in the time and spectral domains improving network utilisation. Also, multi-granularity is supported by the system with a significant scalability improvement as signals going in the same direction can be grouped and carried as wavebands using a single slow switch and waveshaper port regardless of their centre wavelength or required bandwidth. Optical aggregation is carried out in the core network by grooming sub-wavelength or sub-waveband channels. At the network edge transparent transport of optical signals is possible but also electronic aggregation of layer-two traffic may be performed.



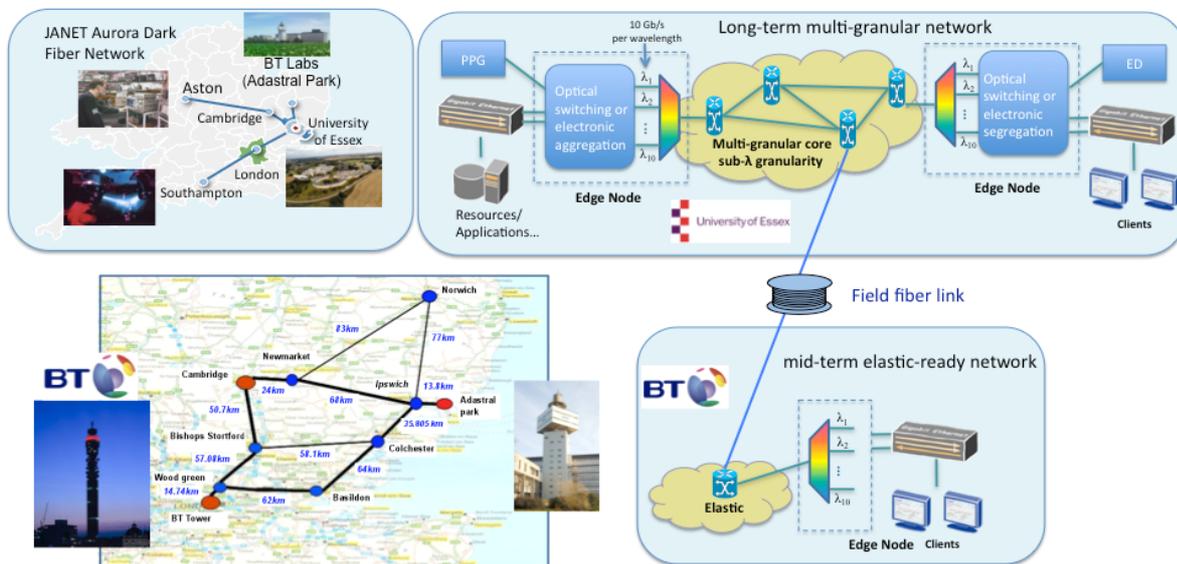
**Figure 19: Long-term elastic transport network demonstrator (University of Essex)**

### 2.1.3.5 Mid and long-term interworking

- Participants: University of Essex, British Telecom
- Status: Fibre for interconnection of the two testbeds has been patched at both ends.
- Plans: Testing dark fibre performance.

- End: June 2012

Interworking between mid-term and long-term architectures is planned to be demonstrated in a distributed experimental testbed located at the University of Essex and BT. The initial experiment configuration is shown in Figure 20. The long-term testbed, located at University of Essex, has the functionality of flexible time and spectral domain allocation whereas the mid-term testbed, located at BT, is elastic-ready and can handle gridless multi-granular traffic but does not provide the function of multiplexing in the time domain. The two testbeds are interconnected by a 49 km field fibre link deployed between University of Essex in Colchester and BT in Ipswich. In addition, some of the nodes in the long-term testbed may be interconnected using a 110 km field fibre link with full dispersion compensation between Colchester and Chelmsford as shown in Figure 11. A BT field fibre link from Adastral park to BT Tower could also be used for the optical networking testbed, which consists of field fibres of G652 and various amplifier sites at intermediate exchanges, the total fibre link has a round trip distance of ~360km as shown in Figure 13.



**Figure 20: Mid and long-term interworking demonstrator (University of Essex, BT)**

### 2.1.3.6 Photonic node architectures, input from WP2

- Participants: University of Essex
- Status: Novel multi-granular photonic node architectures (section 2.1.2) proposed by University of Essex in last plenary meeting.
- Plans: Expecting input from WP2 on architecture studies and design. As this is an iterative process between WP2 and WP4 it will finish at the end of the project.
- End: December 2012

## 2.2 100 Gbit/s packet processing for the long-term scenario

According to predictions of the semiconductor industry, the bandwidth of large scale buffer memories will only scale with a factor of 14 within the next 10 years compared to a factor of 100 for the traffic growth. This means that the amount of memory modules/pins per module must grow by a factor of 7 (=100/14). Taking into account that today's memory busses

already have a width of 300-400 bits, this will not scale for implementation. Therefore we are working on a new concept for a memory interface which will lead to a significant reduction of the required bandwidth. We will assess the performance of our concept firstly by simulations. The potential of demonstrating the benefits of the concept on the basis of the available test bed infrastructure and considering the effort for implementation will be investigated carefully.

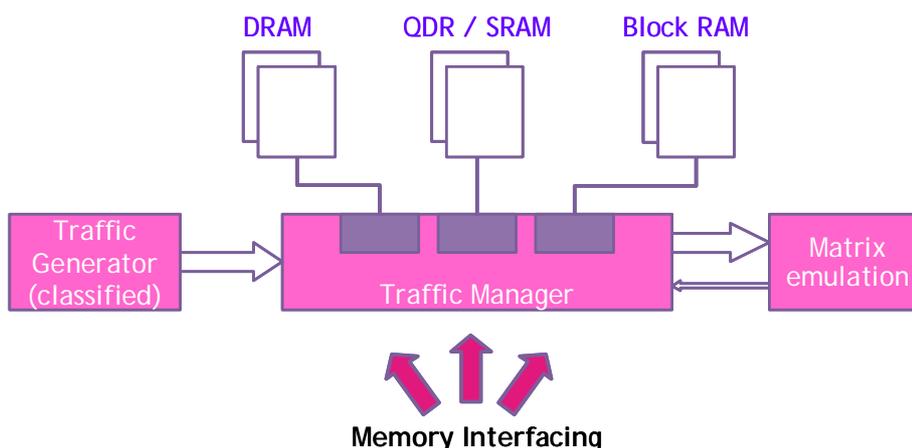
To evaluate new concepts and strategies, we will set-up a test bench to experiment on selected key functions of a novel, efficient traffic manager for electronic packet processing in the context of a hybrid multi-granular node. The objective is to assess the performance of diverse traffic management and memory architectures and to build the foundations for an implementation on the 100 Gbit/s experimental platform available in the Alcatel-Lucent lab, representing the major 100 Gbit/s Layer 2 processing functions of a line card (without physical layer/transmission parts).

### 2.2.1 Testbed and functionalities to be implemented

#### Input Traffic Manager

Two application scenarios must be considered for the test bench. The first one is given in Figure 21. The traffic generator on the left provides classified packets for the simulation. This is similar to a real-world situation. A packet enters a packet line card on a fibre and the transport protocol is terminated. Then the packet is classified, which means that the packet is assigned to one out of a fixed number of flows. In subsequent blocks all packets of the same flow are handled in the same manner, e.g. they are sent to the same output, to the same queues, provided with the same MPLS or Ethernet header etc. Classification can either be simply done by address look-up or by sophisticated DPI techniques. The assignment of a packet to a certain flow is typically described by a flow ID which is placed in front of the packet.

In our test bench we will model the incoming traffic for the traffic manager in such a way, that the packets already have a flow ID in front of them.

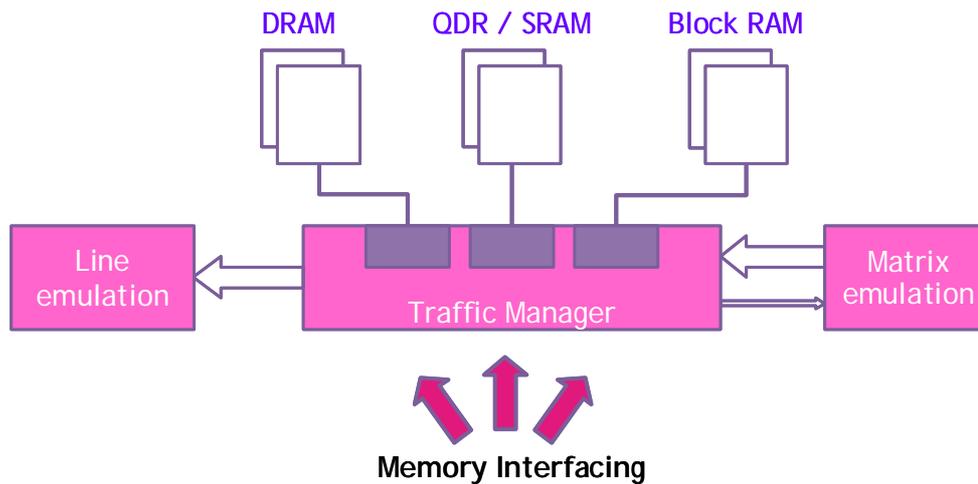


**Figure 21: First testbench scenario for the path from line termination to matrix**

The traffic manager then processes the packets, for which a significant amount of memory is required. There are three different types of memories available: Dynamic RAMs (DRAM), static RAMs (SRAM) partly with a quad data rate interface (QDR), and FPGA internal Block RAM (BRAM). In an ASIC implementation, there is fast on-chip memory available instead

of the block RAM. A severe problem is the memory interface as described above so we want to investigate and simulate several memory architectures.

After the packets have left the traffic manager they will enter the matrix. Since the matrix has no significant storage capacity, it performs a backpressure indicated by the arrow from the matrix to the traffic manager in Figure 21. In this way, the matrix determines which queues of the data buffer in the traffic manager are served. In order to achieve a realistic simulation of the traffic manager these backpressure events must also be modelled and taken into account.



**Figure 22: Second testbench scenario for an output traffic management**

### Output Traffic Manager

If the traffic manager resides on the output path of the line card, the application scenario is rather different. In general the input traffic manager has the task to traverse all packets through the matrix, which is eased by the fact that the matrix typically provides a speed-up, i.e. the bandwidth between matrix and traffic manager is higher than on the input side of the traffic manager. For the output traffic manager, the situation is vice versa, the input data bandwidth exceeds the output data rate – packets must be stored. In addition, the output traffic manager must carefully respect the properties of the output lines in order to avoid exceeding service level agreements or even to cause packet drops simply by overloading a fixed data rate output line. Output traffic management is more difficult and requires more effort in terms of number of queues and bandwidth regulation. For this reason this application scenario must be evaluated in addition to the input traffic manager scenario.

A test bench modelling this application scenario is given in Figure 22. The packet stream now enters the traffic manager on the right hand side. The data stream out of the matrix emulation block can be reduced on reception of backpressure signals from the traffic manager. The traffic manager has an output line which is virtually divided into several channels. The traffic on these channels already respects the property of the output link for which this traffic is destined (e.g. does not exceed 1 Gbit/s if the output line is a 1 GigaEthernet circuit or respects SLAs). The line emulation controls if the traffic manager really performs the demanded shaping of the traffic.

## 2.2.2 Implementation and demonstration plan

The objective of the investigations is to lay the foundations for an implementation of novel memory interfaces for traffic management on a 100 Gbit/s line card.

- Participants: Alcatel Lucent
- Status: we have a flexible FPGA based 100 Gbit/s line card available, which was developed within another project, for the implementation of the traffic manager.
- Plans: first, simulations and assessments of different memory architectures will be performed. Then, following these functional assessments, the identification of additional research results, which could be expected by dedicated hardware implementation and demonstration, will be carried out.
- End: October 2012

The demonstration plan is comprised of 5 steps:

### **Step1: Evaluation of 100 Gbit/s line card properties**

To define a test bench which allows a later transfer of the results to an implementation it is inevitable to achieve a profound understanding of the technology on that 100 Gbit/s line card, e.g. both an in-depth appreciation of the different memory technologies and the functional blocks of a traffic manager.

### **Step 2: Creation of a test bench for investigation of different memory interface building blocks**

The second step comprises the possibly most time consuming tasks. Here the test bench for the evaluation of the different memory interfaces is built.

### **Step 3: Creation and investigation of different memory interfaces**

This step presents the core tasks: the implementation and investigation of the different memory interfaces. Of particular interest are the memory usage patterns, i.e. utilization of the memory and the drop probabilities, i.e. which percentage of the traffic is lost. The results will be used as basis for the decision on which key functions may then be verified in the hardware environment.

### **Step 4: Preparation of the Implementation on the 100 Gbit/s line card**

During this step we will prepare a detailed demonstration plan of a new traffic management, i.e. which parts of the traffic manager can be implemented on the 100 Gbit/s line card. The main goal is to demonstrate the feasibility of the approach.

### **Step 5: Implementation on the 100 Gbit/s line card**

On condition that there is a clear benefit on implementing the new memory architecture and depending on both the estimated effort for realisation of the traffic management on a FPGA based 100 Gbit/s line card and the remaining effort in the STRONGEST project, we will start an implementation of selected key functions of the traffic management with new memory interfaces.

## 2.3 MPLS-TP and WSON integration for the mid-term scenario

### 2.3.1 Testbed description

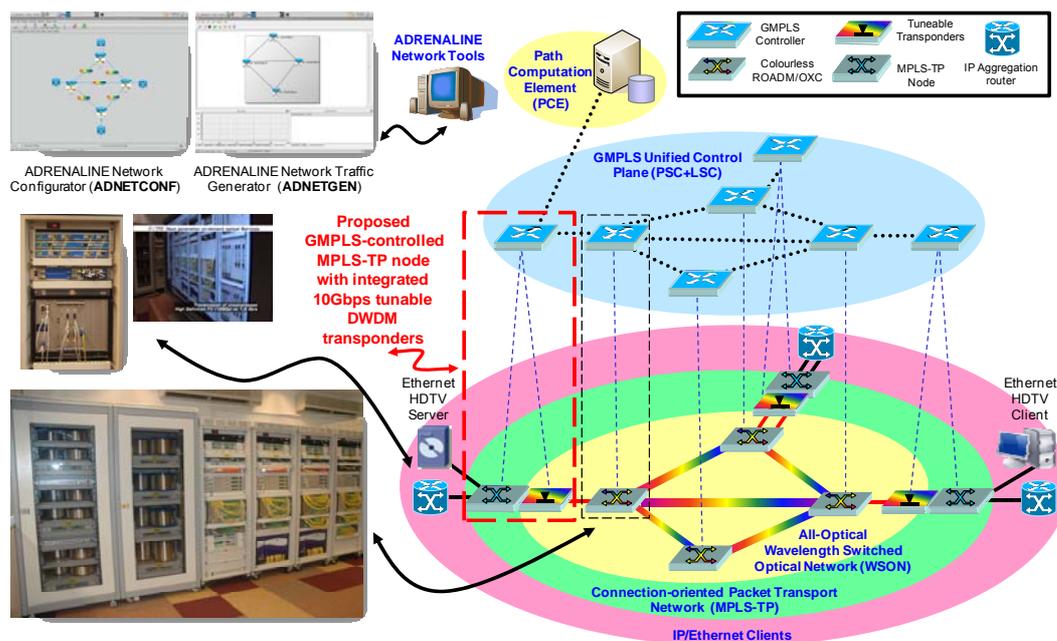
This activity aims at deploying a GMPLS-controlled single-domain dual-region (MPLS-TP and WSON) transport network for demonstration of dynamic and flexible IP and Ethernet services. The objective is to replace the TDM aggregation and transport infrastructure, based on legacy SONET/SDH technology, to offer cost-effective TDM services, by a new dynamic network infrastructure that delivers the high-bandwidth transport and deterministic performance of the optical circuit technology along with the efficient aggregation and statistical multiplexing of the packet switched technology to support packet-based services (e.g., Ethernet, Voice over IP, Layer 2/Layer 3 Virtual Private Networks, IP Television, etc.). The deployment of the proposed testbed combines six key technologies, namely:

- Connection-oriented Packet Transport Network (PTN) based on Multiprotocol Label Switching - Transport Profile (MPLS-TP). This technology provides connection-oriented transport for packet services, allowing flexible packet aggregation and grooming (statistical multiplexing) by means of the electronic switching.
- Pseudo-wire emulation edge-to-edge (PWE3). This technology emulates the operation of a transparent wire carrying an Ethernet service over a packet switched network (PSN) based on MPLS-TP.
- Wavelength Switched Optical Network (WSON), providing reconfigurable high-bandwidth end-to-end optical connections, transparent to the format and payload of client signals.
- GMPLS-enabled unified control plane for MPLS-TP (packet) and WSON (lambda) transport technologies. A single control plane instance is applied in a ubiquitous way to the entire data plane switching layers within the same domain. The applicability of a single GMPLS control plane governing multiple switching technologies provides a unified control and automatic management for both LSP provisioning and recovery. This unified control plane (an enhancement, with respect to current solutions mainly based on IETF standards) will be developed in the control plane experimental tasks, and will be adopted in the data plane experiments
- Field programmable gate array (FPGA) based reconfigurable software/hardware co-design. This technology provides a quick platform for data plane which will be able to send/receive and process the data at line speed (i.e. 10 Gbit/s ), This platform benefits from the high performance of hardware and also relishes the flexibility of software.
- Software/hardware defined adaptable network (SHDAN) framework, which provides a facility for network elements to dynamically tune, update and add network function blocks based on different network/application profiles and provider/user requests.

The proposed testbed will be deployed based on adding, extending and enhancing the existing GMPLS-controlled WSON infrastructure of the CTTC ADRENALINE (All-optical Dynamic REliable Network hAndLING IP/Ethernet Gigabit traffic with QoS) testbed ®. Specifically, three new GMPLS-enabled MPLS-TP nodes with integrated 10 Gbit/s bps tuneable DWDM transponders will be designed, implemented, connected and validated in the ADRENALINE testbed, as shown in Figure 23. The ADRENALINE testbed is a

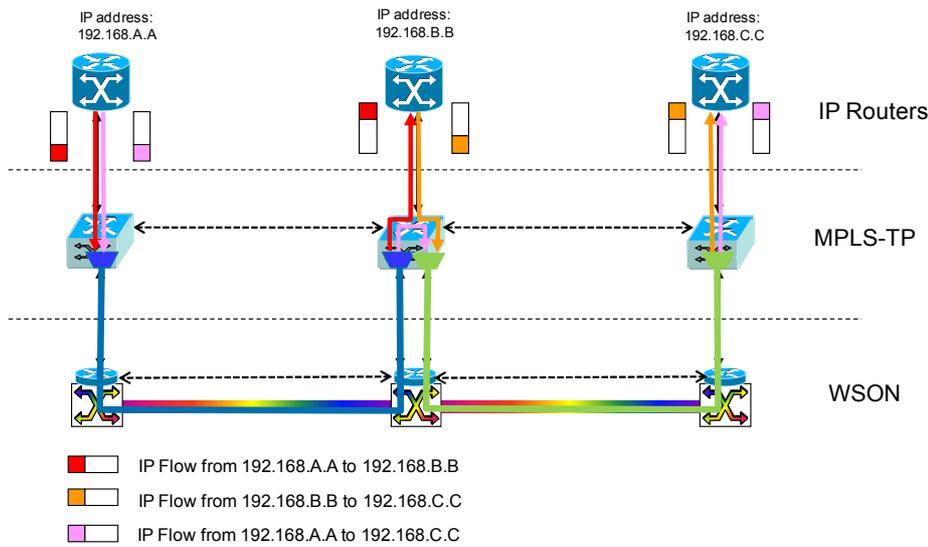
GMPLS-controlled Intelligent Optical Network composed of an all-optical DWDM mesh network with two colour-less ROADM nodes and two OXC nodes, providing reconfigurable (in space and in frequency) end-to-end lightpaths. The optical node architecture is based on using AWG as DWDM (de-) multiplexers (8 and 16 wavelengths with 50 and 100Gbit/s Hz channel spacing, respectively), and MEMS as the switching technology. Arrays of power meters and VOAs are used for optical power equalization at output fibers. The ADRENALINE testbed deploys a total of 610 km of G.652 and G.655 optical fiber divided in 5 bidirectional links, in which EDFA optical amplifiers are allocated to compensate for power losses during optical transmission and switching at C-band.

Each optical node is equipped with a GMPLS Connection Controller for implementing a distributed GMPLS-based control plane, in order to manage automatic provisioning and survivability of lightpaths (RSVP-TE signaling protocol for wavelength reservation, and OSPF-TE routing protocol for topology and optical resource dissemination), allowing traffic engineering algorithms with QoS. The ADRENALINE testbed includes a Path Computation Element (PCE), which is a dedicated network entity responsible for performing advanced path computations. The PCE serves requests from Path Computation Clients (PCCs), and computes constrained explicit routes over the topology that constitutes the optical transport layer.

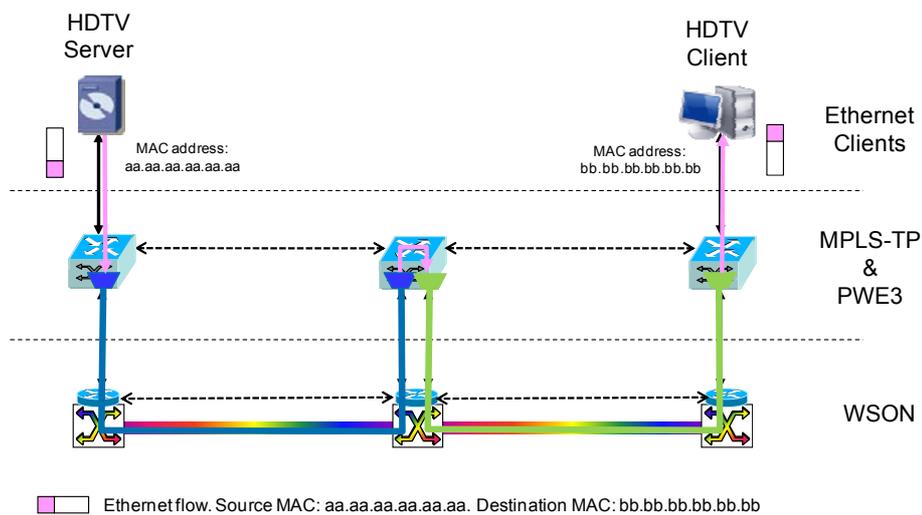


**Figure 23: Logical view of the enhanced single-domain dual-region (MPLS-TP and WSON) ADRENALINE testbed architecture for IP and Ethernet services.**

Two demonstration activities are planned addressing the transport of both IP and Ethernet services. The former (Figure 24.a) will be based on offloading pass-through traffic from the IP layer to the MPLS-TP and WDM layer (i.e., router bypass). It allows reducing the IP flows forwarded by the intermediate IP routers, and thus reducing the router processing unit and the required number of ports per IP router. The second demonstration activity (Figure 24.b) considers Ethernet HDTV video distribution. To this end, an HDTV server and a client will be introduced in the testbed, offering point-to-point Ethernet HD video distribution, without requiring forwarding in the IP layer.



a) Example of IP service for Traffic offloading



b) Example of Ethernet service for HDTV distribution

**Figure 24: Planned demonstration scenarios**

## 2.3.2 Network functionalities to be implemented

### 2.3.2.1 Implementation approach considerations

The control plane will be deployed in a software development platform based on a multi-core PC architecture with Linux as operating system consisting of a host and a series of Network Interface Cards (NICs). The main advantages of using a Linux software platform is that it is a cost-effective solution (since it relies on PC and open-source Linux kernel), and provides maximum flexibility in terms of development of new functionalities at user-level (applications) but also at Kernel-level. In particular, it allows reusing and extending the GMPLS protocol stack developed for the WSON network of the ADRENALINE testbed®, reducing the development time in comparison with a new implementation from the scratch. Moreover, a wide range of open-source software is available and ready to be used. For

example, the Click Modular Router (CMR) software can be used to develop a MPLS-TP and PWE3 forwarding engine in an easy way, reducing the development time in comparison with a hardware implementation based on FPGA. However the main disadvantage of deploying CMR based MPLS-TP and PWE3 is the performance bottleneck due to the communication between the host (processor and memory) and the NICs, and the software/hardware interruptions. Thus, it is not possible to reach line card packet processing for high data bit rates (e.g., 10 Gbit/s or above). The impact of this bottleneck in the control plane is almost negligible, since the bandwidth consumed by the control channels are of few tens of Kb/s. However, it may be critical for the data plane implementation. In order to evaluate the pros and cons of the software versus the hardware implementation, the data plane of one or two of the developed nodes will be implemented in hardware (FPGA + 10 GigaEthernet XFP port + 1 GigaEthernet port), and the other two or one node will have a data plane based on software implementation (host with CMR + 10Gbit/s E PCI NIC with XFP port + 1 GE PCI NIC).

### 2.3.2.2 Software based data plane implementation

As for the software implementation of the data plane, Figure 25 shows the architecture of the proposed MPLS-TP and PWE3 forwarding engine with the integrated 1 GigaEthernet interfaces and 10Gbit/s tunable DWDM transponders. The MPLS-TP forwarding engine architecture considers two different services: transport of Ethernet frames or transport of IP packets. In the proposed architecture, all Ethernet packets from a defined set of Network Interface Cards (NIC) used for the data plane are sniffed. These packets are classified into two types:

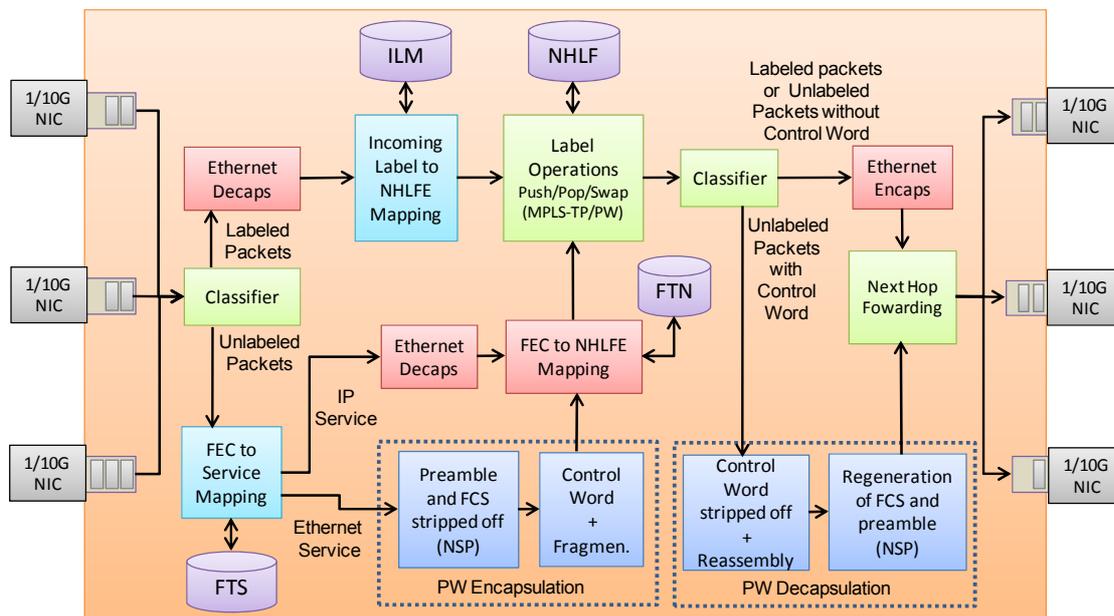
- MPLS-TP labeled packets. The Incoming Label Map (ILM) maps each incoming packet label and port to an entry of the Next Hop Label Forwarding database (Next Hop Label Forwarding Entry, NHLFE).
- Unlabeled packets. The FEC to Service (FTS) determines whether the required service of the packet received from outside the MPLS-TP domain at the ingress node is Ethernet or IP. For Ethernet services, a pseudo-wire is encapsulated; the ingress Native Service Processing (NSP) function strips the preamble and frame check sequence (FCS) from the Ethernet frame. After that, the control word is prepended to the resulting frame, and optionally, fragmentation can be used since the maximum frame size that can be supported is limited. Finally, a FEC to NHLFE (FTN) maps each FEC to a NHLFE. As for IP service, the Ethernet header of the received frames is stripped off (i.e., Ethernet decapsulation) and the FEC is mapped into a NHLFE.

At this point, both labeled and unlabeled packets have a NHLFE. It is used for forwarding a packet, based on the NHLF database. It contains the next hop for the packet, and the operations to be performed on the label stack of the packet. Three operations have been defined:

- Replace the label at the top of the label stack with a specified new label (*swap*).
- Strip off the label stack (*pop*).
- Push one specified new MPLS-TP label (for IP service) or two specified new MPLS-TP and PW labels (for Ethernet service) onto the label stack (*push*).

After performing the NHLF operations, the labeled packets and unlabeled packets without control word (i.e., IP service) are encapsulated with the proper Ethernet header and forwarded to the next hop through the corresponding NIC. As for the unlabeled packets with control word (i.e., Ethernet Service), the control word is stripped off, and optionally

packet reassembly is performed. Then, the NSP regenerates the FCS and the preamble of the Ethernet header before forwarding the frame to the attachment circuit.



**Figure 25: Architecture of the MPLS-TP and PWE3 forwarding engine of the software implementation.**

### 2.3.2.3 Hardware based data plane implementation

As mentioned in subsection 2.3.2.1, an alternative candidate approach for implementing data plane is to deploy FPGA based reconfigurable hardware. Traditionally, the approach of using general purpose processors for data plane offers great flexibility at the cost of other design factors like size and power consumption. Moreover, due to the general purpose nature and architecture, they may not be the strongest performer compared to other type of hardware like special purpose processor or coprocessor. There are four basic steps for an instruction to be executed by the general purpose processor when running a software procedure: fetch, decode, execute, and write back. These steps are normally pipelined in order to improve efficiency and speed. Apart from pipelining, parallelism is also a natural method to increase performance. However, parallelism does not improve the speed of a single application. It only helps when multiple programs are running on a single processor. Not until recently, have modern processors had more than one processing 'core' so that instructions can be processed in parallel by different cores. However, this kind of parallelism is coarse-grained and is not scalable since it is restricted by the number of cores.

In contrast, the FPGA based data plane not only offers the programmability of software systems, indeed, it also provides the parallel architectures within the hardware that can be (re)-defined to suit the application. For example, while a 3 GHz Pentium class processor is a highly optimized general purpose programmable "software system", it can be outperformed by an "FPGA based system" working at 300 MHz that has been designed for specific applications. An FPGA based system offers the ability to give both high-performance and flexibility. Table 2 summarizes performance results of different types of algorithms in hardware accelerator based on FPGA compared to a general-purpose processor.

**Table 2 Speed Comparison between FPGA based and Software only processors [TechRep1]**

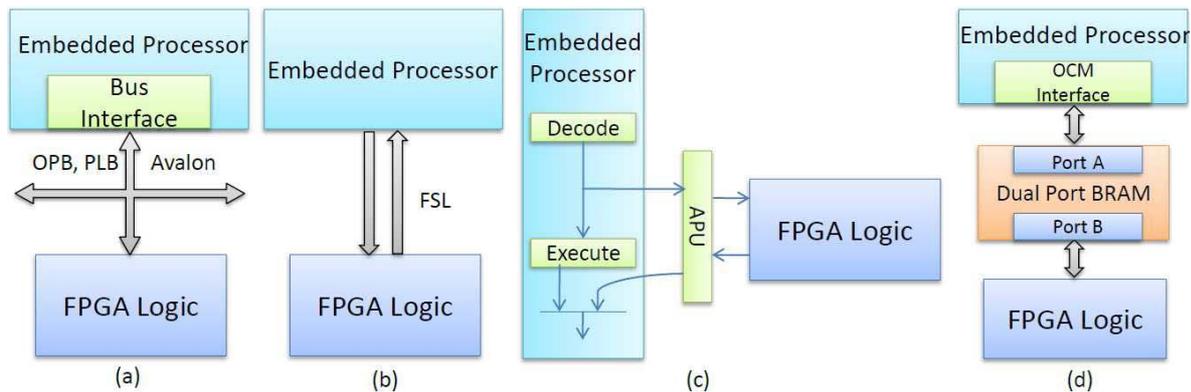
Application	FPGA based	Software only
Hough & intensive Hough processing	2 sec of processing time  @20 MHz <b>370× faster</b>	12 mins processing time  Pentium 4 - 3 GHz
AES 1MB data processing rate	424 ms/19.7 MB/s	5,558 ms/1.51 MB/s
Encryption	424 ms/19.7 MB/s	5,558 ms/1.51 MB/s
Decryption	<b>13× faster</b>	
Smith-Waterman ssearch34 from FASTA	100 sec FPGA processing <b>64× faster</b>	6161 sec processing time Opteron - 2.2 GHz
Multi-dimensional hypercube search	1.06 sec FPGA  @140 MHz Virtex II <b>113× faster</b>	119.5 sec  Opteron - 2.2 GHz
Callable Monte-Carlo Analysis 64,000 paths	10 sec of processing @200 MHz FPGA system <b>10× faster</b>	100 sec processing time Opteron - 2.4 GHz
BJM Financial Analysis 5 million paths	242 sec of processing @61 MHz FPGA system <b>26× faster</b>	6300 sec processing time Pentium 4 - 1.5 GHz
Mersenne Twister Random Number Generation	319M 32bit integers/sec  <b>3× faster</b>	101M 32bit integers/sec  Opteron - 2.2 GHz

### **FPGA + Embedded Processor implementation**

Given the nature of high interactive communication to the control plane, an FPGA with embedded processor method is adopted. The advantages of this method include: its potential to achieve the high performance packet processing at the MPLS-TP node, flexibility for future upgrades and modification, easy communication with control plane. However this approach has also few disadvantages including: requirement for a complex proprietary platform to combine embedded processor and FPGA, and requirement for a careful design to partition the data plane functionalities to two parts for implementation in logical gate and embedded processor of FPGA.

### **Inter-connection models between FPGA logic and embedded processor in FPGA fabric**

There are different solutions provided for deploying embedded processors inside the FPGA fabrics. Four approaches have been studied and analysed below:



**Figure 26: Different solutions for integrating an embedded processor with the FPGA logic**

**a) Embedded processor bus connected model**

As shown in Figure 26 (a), processor bus connected to the FPGA logics needs the embedded processor to move data and send commands through a bus. Consequently, a single data transaction can require many processor cycles and need bus arbitration. A direct memory access (DMA) can be deployed to enable the FPGA logic to operate on data located on bus connected memory without the interaction with the embedded processor at the cost of additional FPGA logics.

**b) I/O based model**

In this approach, Fast Simplex Link (FSL) as shown in Figure 26 (b) which is a more efficient interface than embedded processor bus through a dedicated point to point communication channel to exchange data between the FPGA logic and embedded processor without any arbitration overhead. The reduced control complexity enables lower latency and higher rate data movement but one channel only supports one direction data movement.

**c) Extended instruction set model**

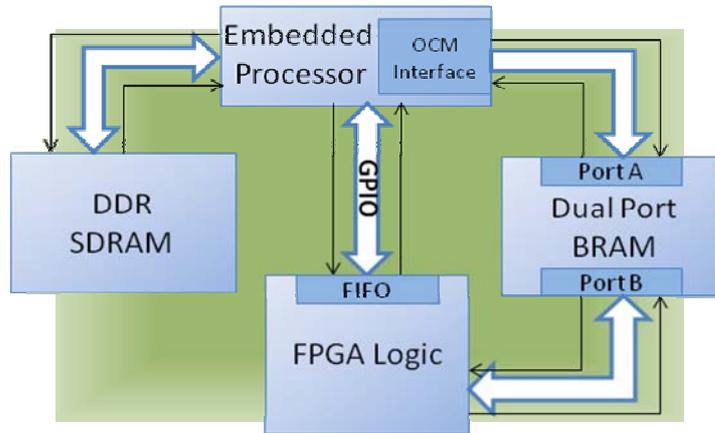
In this approach, Auxiliary Processor Unit (APU) as shown in Figure 26 (c), it attached directly to the instruction pipeline of the embedded processor and extends the instruction set. As the most highly integrated interface, this approach also clocks faster than an embedded processor bus. The brief data flow of APU is that the instructions from cache/memory are simultaneously presented to the embedded processor decoder and the APU controller. If the embedded processor recognizes the instruction, it is executed, otherwise, the APU has the opportunity to acknowledge the instruction then send to the FPGA logic to execute it.

**d) Shared memory model**

On Chip Memory (OCM) approach as shown in Figure 26 (d). An on chip dual port block RAM (DPBRAM) is used between the FPGA logic and embedded processor which both can directly and separately access the same data memory, e.g. one port can be employed for embedded processor data side interface connection, while the second port for external FPGA logic data access. This approach does not suffer any additional signalling overhead.

**Chosen interconnect models for STRONGEST**

Based on the discussion above, a combination of the shared memory approach and embedded processor bus connected approach as shown in Figure 27 will be deployed.



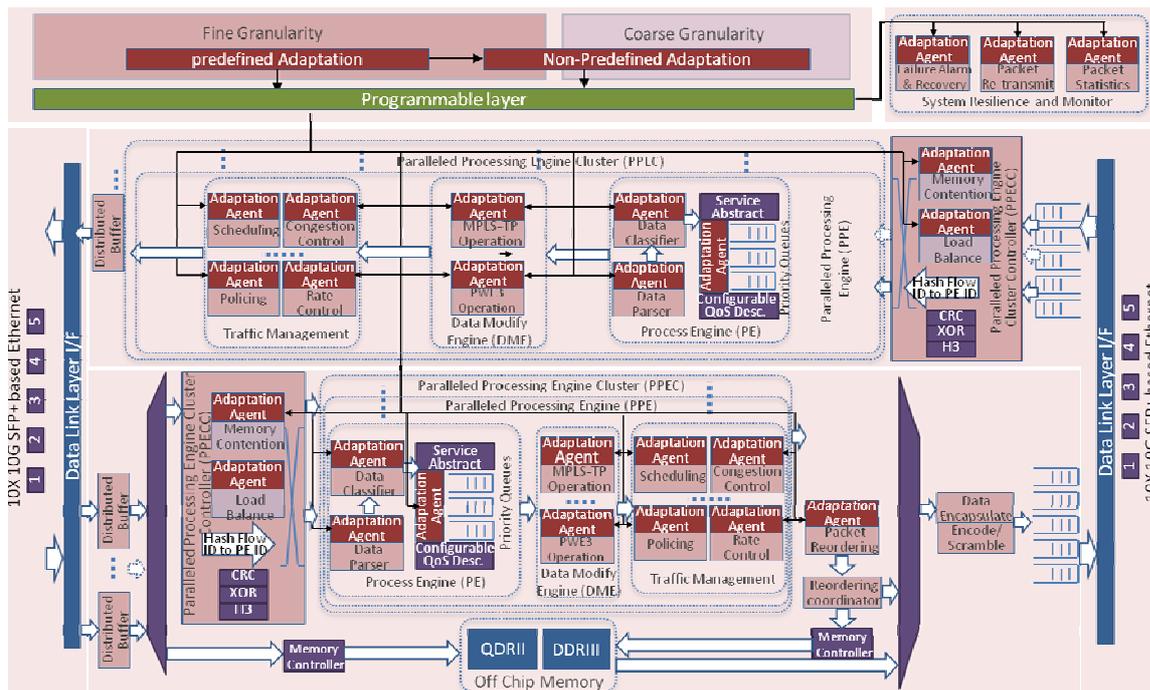
**Figure 27: Selected hardware architecture**

### Hardware based MPLS-TP and PWE3 forwarding engine architecture design

The hardware version MPLS-TP and PWE3 forwarding engine cannot promise the exact same functionality as the software version shown in Figure 25, but it will try to retain the same interfaces between data plane and control plane, for example, it will have lots of benefits if the control plane software modules can communicate with the hardware based data plane to access variable packet processing tables via the hardware abstraction layer. As it can be envisioned, in this hardware version, neither kernel space CMR nor 10Gbit/s XFP DWDM tunable interface driver are off-the-shelf so that rather large amount of man-month is needed.

### Hardware based data plane architecture design

Apart from MPLS-TP and PWE3 forwarding functions, depending on the implementation complexity, the hardware based data plane can provide traffic management function and SHDAN feature which enables system modules dynamic tuning and updating based on network environment, application/service QoS requirement, under the control of operator. An overall view is given in Figure 28. The system contains five paralleled process engines (PPE) at each direction forming a PPE cluster (PPEC). Each PPE is independent on physical interface (Ethernet interface) and contains almost all the functionalities such as process engine (PE), data modify engine (DME), traffic management (TM). The PPEs are selected dynamically according to the system status over the coordination of PPEC controller (PPECC) which makes the decision to maintain system load balance and minimise memory contention. Since there is a cluster of PPEs and each PPE is not bundled with fixed 10Gbit/s E port, packet might be in wrong order after processing, so that the packets re-order module is required to guarantee the correct order before sending to the transmission buffer.



**Figure 28: Architecture of the data plane of the hardware implementation**

As one of three major function blocks of PPE, PE, consists of data parser, data classifier and priority queues which respects upper layer service abstract and supports reconfigurable QoS description. The second major function of PPE, DME, is responsible for MPLS-TP and PWE3 forwarding. It's worth noting that DME can support more function rather than the mentioned, for example, VLAN tag modification. Also in the other end, small portion of the MPLS-TP and PWE3 forwarding engine will be in other module such as PE. The last major function of PPE, TM, will perform all the traffic management functions, e.g. scheduling, policing, congestion control, and rate control.

For the sake of system resilience and monitoring, corresponding modules are designed, which detect the system logical error and recover to healthy state, re-transmit packet after failure recovery and packet loss, and collect system statistical information, e.g. system load, throughput, number of errors, etc.

The system deploys both on-chip memory and off-chip DDR/QDR memory. On-chip memory is based on distributed memory hierarchy, which means each 10Gbit/s Ethernet port is equipped with dedicated buffer rather than shared memory, and PE directly accesses packet payload at its local memory, removing bottlenecked shared memory.

It is worth mentioning that Figure 28 is only an architectural representation that considers multiple modules for a complete solution but the actual implementation might not correspond to every module, and just consist of a basic number of modules.

**•Software/hardware defined adaptable network**

Since the emerging applications and services demand new network architectures and innovative transport technologies, the STRONGEST data plane nodes must support dynamic adaptation in different granularities in order to satisfy requirements from application/service and/or cognitive network environment. The flexibility can be achieved

through either predefined adaptation or non-predefined adaptation, via the intelligent adaptation agent on top of each of the function blocks. The former is suitable for system feature/function adjustment, the latter fits more the need of significant function changes and system updates. The predefined adaptation can be the outcome of non-predefined adaptation. This dynamics adaptation feature is enabled by a programmable layer which controls when and how a new configuration needs to be selected from a predefined adaptation “pool” and deployed, or a completely new non-predefined adaptation is needed.

### **2.3.3 Implementation and demonstration plans**

#### **Implementation Plans**

##### Data Plane (Software Implementation)

- Partners: CTTC
- MPLS-TP and PWE3 forwarding engine
  - Current status: Preliminary implementation with full functionalities.
  - End: 31st December 2010
- Integrated 10 Gbit/s tunable DWDM transponders
  - Current status: Test and validation of 10Gbit/s PCI Express NIC with XFP transceiver port
  - End: 31st December 2010 with fixed XFP at 850nm, and 31st December 2011 with tunable XFP.

##### Data Plane (Hardware Implementation)

- Partners: CTTC and University of Essex
- Preliminary version of MPLS-TP and PWE3 forwarding engine
  - Current status: High level function blocks designed.
  - Plans: Hardware implementation
  - End: 31 March 2012
- Integrated 10 Gbit/s tunable DWDM transponders (subject to XFP daughterboard commercial availability)
  - Current status: Sourcing FPGA daughterboard compatible with the ordered FPGA boards that can host tunable XFP module.
  - Plans: Tunable transponder implementation
  - End: 31 April 2012 with tunable XFP.

#### **Demonstration plans**

##### Data Plane (Software Implementation):

- Demonstration scenarios: IP services for Traffic offloading and Ethernet services for HDTV distribution.
  - End: 31 March 2011 (with fixed XFP) and 31 March 2012 (with tunable XFP)

##### Data Plane (Hardware Implementation):

- Demonstration scenarios: IP services for Traffic offloading and Ethernet services for HDTV distribution.
  - End: 31 April 2012 (with tunable XFP)

##### Integrated Control and Data Plane (Software + Hardware implementation):

- IP services for Traffic offloading:
  - End: 30 June 2012
- Ethernet services for HDTV distribution:
  - End: 30 September 2012

## 3 Control plane solutions

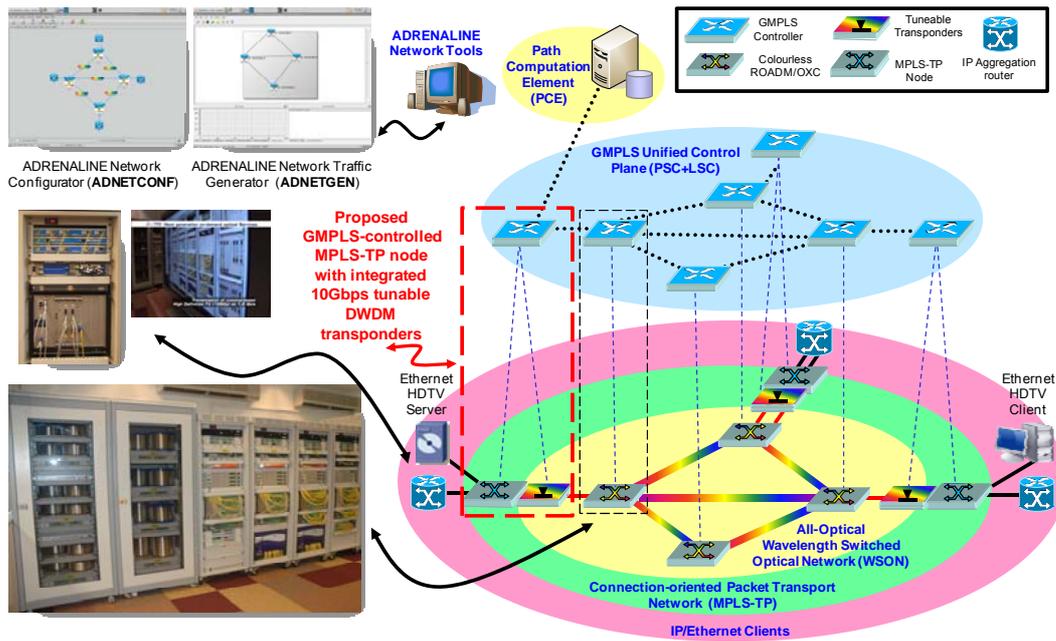
### 3.1 MPLS-TP and WSON control plane integration

#### 3.1.1 Testbed description

This activity aims at deploying a GMPLS-enabled unified control plane for MPLS-TP (packet) and WSON (lambda) transport technologies. A single control plane instance is applied in a ubiquitous way to the entire data plane switching layers within the same domain. The applicability of a single GMPLS control plane governing multiple switching technologies provides a unified control and automatic management for both LSP provisioning and recovery. This unified control plane (an enhancement, with respect to conventional solutions mainly based on IETF standards) will be developed in the control plane experimental tasks, and will be adopted in the data plane experiments

The proposed testbed will be deployed based on adding, extending and enhancing the existing GMPLS-controlled WSON infrastructure of the CTTC ADRENALINE (All-optical Dynamic RELiable Network hAndLING IP/Ethernet Gigabit traffic with QoS) testbed ®. Specifically, three new GMPLS-enabled MPLS-TP nodes with integrated 10Gbit/s bit/s tuneable DWDM transponders will be designed, implemented, connected and validated in the ADRENALINE testbed, as shown in Figure 23. The ADRENALINE testbed is a GMPLS-controlled Intelligent Optical Network composed of an all-optical DWDM mesh network with two colour-less ROADM nodes and two OXC nodes, providing reconfigurable (in space and in frequency) end-to-end lightpaths. The optical node architecture is based on using AWG as DWDM (de-) multiplexers (8 and 16 wavelengths with 50 and 100Gbit/s Hz channel spacing, respectively), and MEMS as the switching technology. Arrays of power meters and VOAs are used for optical power equalization at output fibers. The ADRENALINE testbed deploys a total of 610 km of G.652 and G.655 optical fiber divided in 5 bidirectional links, in which EDFA optical amplifiers are allocated to compensate for power losses during optical transmission and switching at C-band.

Each optical node is equipped with a GMPLS Connection Controller for implementing a distributed GMPLS-based control plane, in order to manage automatic provisioning and survivability of lightpaths (RSVP-TE signaling protocol for wavelength reservation, and OSPF-TE routing protocol for topology and optical resource dissemination), allowing traffic engineering algorithms with QoS. The ADRENALINE testbed includes a Path Computation Element (PCE), which is a dedicated network entity responsible for performing advanced path computations. The PCE serves requests from Path Computation Clients (PCCs), and computes constrained explicit routes over the topology that constitutes the optical transport layer.

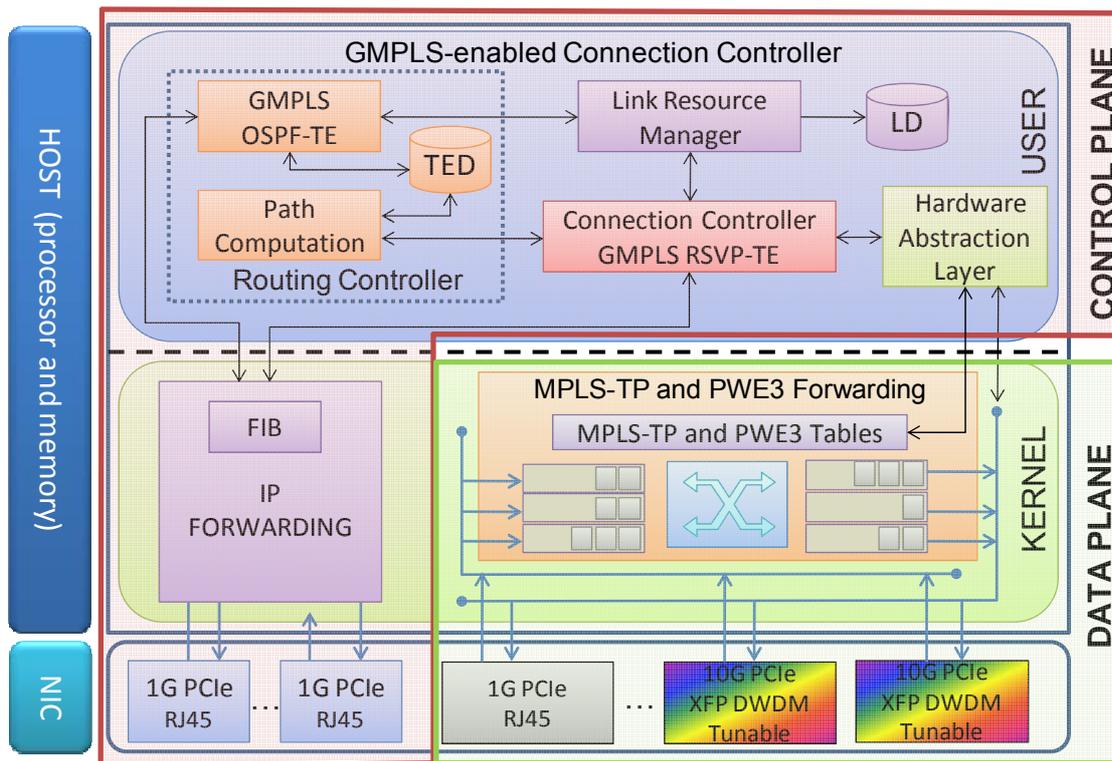


**Figure 29: Logical view of the enhanced single-domain dual-region (MPLS-TP and WSON) ADRENALINE testbed architecture for IP and Ethernet services.**

### 3.1.2 Network functionalities to be implemented

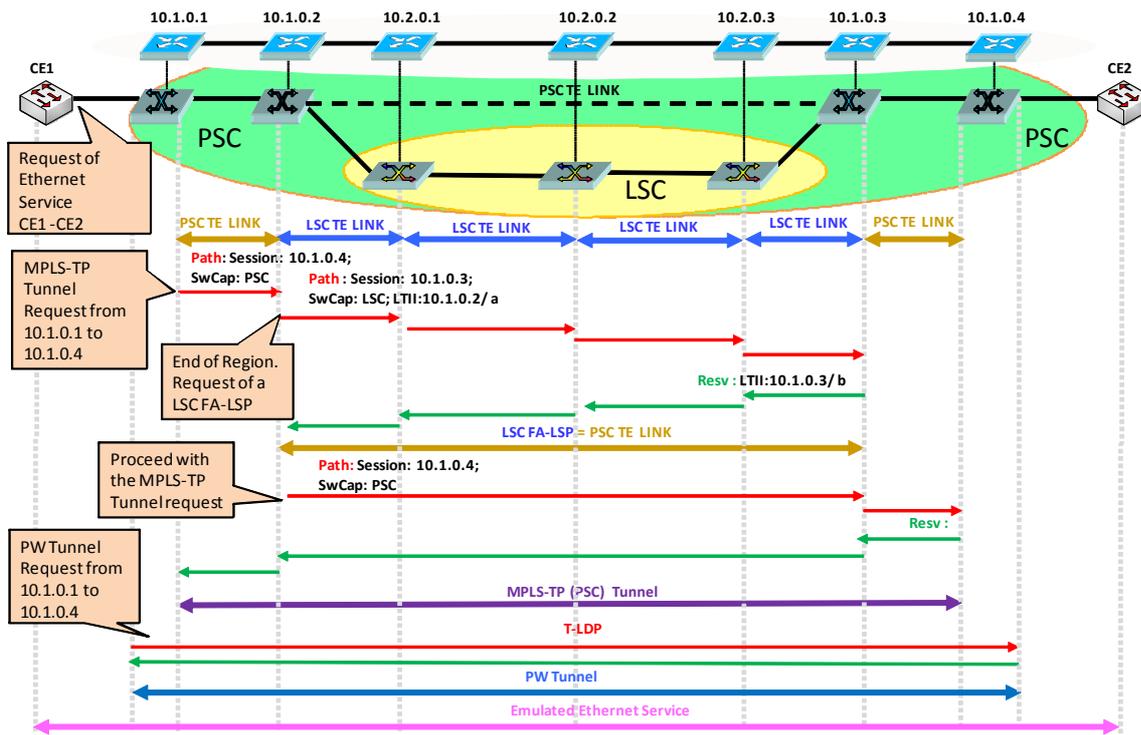
#### 3.1.2.1 Control plane functionalities

Figure 30 shows the logical architecture of the proposed GMPLS-controlled MPLS-TP node with integrated 10Gbit/s tuneable DWDM transponders. In general, the node architecture is divided into two main elements, the control plane and the data plane.



**Figure 30: Architecture for the proposed GMPLS-controlled MPLS-TP node with integrated 10Gbit/s bit/s tunable DWDM transponders**

The control plane is responsible for handling dynamically and in real-time MPLS-TP bandwidth in order to manage the establishment, maintenance, deletion and survivability of MPLS-TP connections (RSVP-TE Connection Controller), and for disseminating and discovering, from the neighbor GMPLS connection controllers, the network topology and MPLS-TP resource availability (PSC TE links) that are stored in the Traffic Engineering Database (TED) repository (OSPF-TE Routing Controller). Focusing on the peer or unified control plane interworking model, a GMPLS controller's TED contains the information relative to the WSON and MPLS-TP layers (i.e., LSC TE links and PSC TE links respectively) present in the dual-region network. Therefore, since a path across multiple regions can be computed, GMPLS defines the concept and notion of Forwarding Adjacency (FA) to favour the use of efficient multi-layer routing algorithms exploiting the grooming strategies. That is, assuming a local control policy, a GMPLS node may advertise an LSC LSP as a PSC TE link. Such a link is referred to as FA. The corresponding LSP is thus referred to as a FA LSP. The goal of using FAs is that, after the routing protocol floods the TE attributes (i.e., end-point LSRs, bandwidth, etc.) of a particular FA, this allows other MPLS-TP nodes to use the FAs as a regular PSC TE link for path computation purposes. Figure 31 shows an example of the establishment of a LSC FA LSP which is triggered when setting up a PSC connection (LSP in the context of the GMPLS). Finally, if an Ethernet service is requested, according to the PWE3 architecture, it is necessary to set up the PW tunnel through the just created PSC LSP. To do this, the source and destination MPLS-nodes exchange the PW labels using the T-LDP protocol. It is worth mentioning that the control plane is also responsible for creating and managing the forwarding tables of the data plane.



**Figure 31: Example of Ethernet service provisioning through a Multi-layer GMPLS Unified Control Plane in a multi-region (MPLS-TP and WSON) transport network**

### 3.1.3 Implementation and demonstration plans

**Participants:** CTTC

**Current status:** Preliminary development and tests of the control plane Link Resource Manager (LRM).

**Plans:**

- Control plane implementation with full functionalities (31st December 2011)
- Demonstration of Multi-layer GMPLS-based Unified Control Plane for IP and Ethernet Services (31 March 2012)

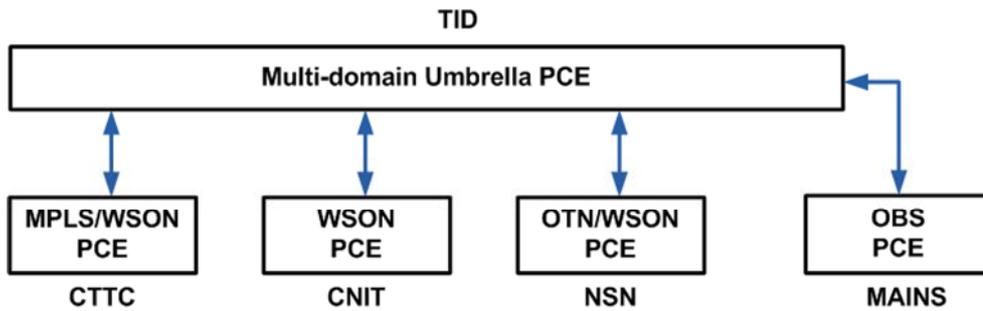
## 3.2 Multi-technology and multi-domain PCE interworking

### 3.2.1 Testbed description

#### 3.2.1.1 Overall testbed

STRONGEST Multi-domain PCE testbed is based on:

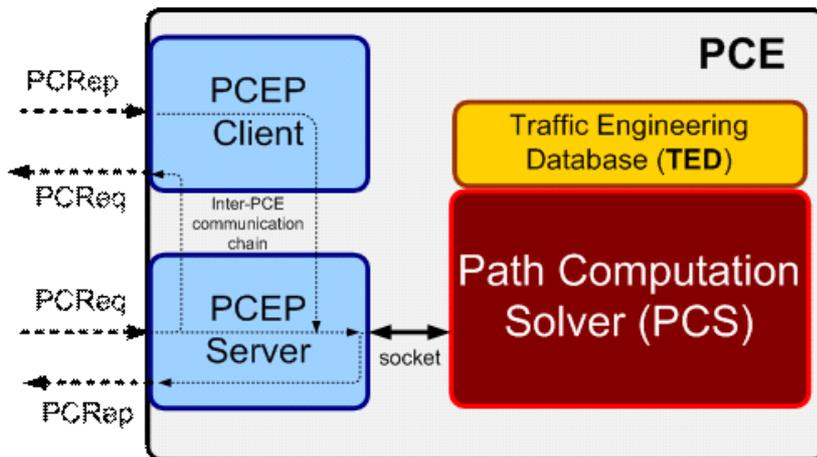
- Multiple PCE testbeds from different partners interconnected by means of IP tunnels.
- Different technologies forming several regions: MPLS, OTN, WSON, OBS (cooperation already agreed with MAINS)
- Hierarchical PCE architecture (Figure 4)



**Figure 32: Multi-domain PCE testbed**

### 3.2.1.2 CNIT PCE testbed

The Path Computation Element software tool developed at CNIT is called PaCE and is based on the standard PCE architecture described in [PCE]. The tool is written in C++ fully resorting to the C++ standard library and is implemented for Linux-based machines. The tool includes both the PCE and the Path Computation Client (PCC), that can run either separately (acting as basic external PCC) or in the PCE-integrated mode (acting as internal client, for example in inter-PCE chain). Both PCE and PCC communicate by means of the PCEP protocol, implementing the Finite State Machine as described in [PCEP]. The PCE module is based on the architecture depicted in Figure 33. The PCEP server accepts parallel PCEP sessions from different PCCs and handles the incoming and outgoing PCEP messages. The PCEP server internally communicates with the Path Computation Solver (PCS) by means of a persistently open TCP socket.



**Figure 33: CNIT PCE implementation: architecture.**

The PCS module can be applied for path computation in two main scenarios: MPLS networks and WSON.

Concerning MPLS, PCS is able to resort to MPLS Traffic Engineering Database (TED) synchronized with one of the commercial routers available at CNIT testbed (including Juniper and Cisco Routers connected via optical Gigabit Ethernet (1000BASE-LX), ATM over STM1 and Fast Ethernet interfaces). TED synchronization is performed through proprietary UNI communication.

Concerning WSON, PCS supports Multi-Rate (10-40-100Gbit/s ) lightpath requests implementing integrated RWA strategies to encompass impairments due to BER, OSNR and XPM. Currently, the PCS module resorts to an emulated TED referred to a transparent WSON domain of 17 optical nodes and 33 bidirectional WDM links. PCS is equipped with the following path computation (PC) functions: 1) k-Shortest PC , 2) k-Shortest least-congested PC, 3) k-Shortest least-congested PC with impairment validation. Wavelength suggestion functions (random, first-fit, ad-hoc IV-based) are also available. The PCE may act both in IV+RWA (with two separate PCE instances) and in IV&RWA architecture [WSON-IMP]. In the former case, the IV PCE has the capability to perform impairment-validated candidate path computation by resorting to the proprietary Candidate Path object. In this way, the PCC may ask for k pre-computed impairment-validated routes having the same end-points. Impairment validation is performed applying worst case penalties without resorting to dynamic resource allocation info in the TED. In the IV&RWA approach, joint path computation and wavelength suggestion is performed based on dynamic TED and considering both worst case penalties and guard bands constraints.

Currently the PCC supports the following optional PCEP objects: LSPA (FastReroute flag), SVEC (link/node/SRLG diverse computation), METRIC, BRPC flag [BRPC] and Bidirectional LSP flag in the RP object. The PCE supports the BANDWIDTH and METRIC objects and the following ERO sub-objects: IPv4, Unnumbered Interface, Label Control.

### **3.2.1.3 NSN-G PCE testbed**

NSN-G will provide a path computational element (PCE) for wavelength switched optical networks (WSON). The PCE implementation consists of a set of routing and wavelength assignment (RWA) algorithms and a module to assess the optical performance. It is based on centralized architecture approach and includes a locally stored traffic engineering database (TED). Furthermore private PCEP protocol implementation is available for PCE-PCE and PCE-PCC communication. In the following, all parts are described in more details.

#### 1) Routing and wavelength assignment (RWA) algorithms

The RWA engine is invoked after a valid path calculation request with end-points in the local PCE domain is received. Path search and wavelength assignment are jointly optimized with a cost function minimizing the required optical equipment for provisioning the requested service. For a tie-breaker minimum distance or hop count is used. The path search is based on the branch-and-bound class generic algorithms. Wavelength assignment can be: (1) first fit according to predefined sequence; (2) minimizing the wavelength changes between consecutive optical channels; (3) maximizing the availability of continuous wavelengths for future services. The wavelength continuity constraint is supported. Furthermore if disjoint paths for protection are required the same wavelength on both working and protection path can be demanded.

#### 2) Optical performance

We support two different approaches for evaluating the optical performance of an end-to-end service request: (1) exact which is based on pre-calculated reachability matrix and (2) simplified which uses OSNR threshold. For (1) NSN SURPASS TransNet planning tool for hiT7300 and hiT7500 is used. The physical topology layer is designed including all linear and non-linear fiber impairments and matrices with all feasible optical channels are generated for every transmission data rate – 2.5 Gbit/s, 10 Gbit/s , 40 Gbit/s , and 100 Gbit/s . A feasible optical channel is a “dummy” end-to-end service which does not require

signal regeneration. For (2) the same tool as in (1) is used, but the optical performance is estimated based on OSNR metric. If the OSNR along the optical path is below a given threshold, then optical signal regeneration is required.

### 3) PCEP implementation

PCEP library based on rfc5540 is used for communication between PCE-PCE and PCE-PCC. PCE discovery mechanism is not available at the moment. Rfc5521 is added for support of XRO objects and “draft-margaria-pce-gmpls-pcep-extensions-01” is implemented to enable WSON specific requirements.

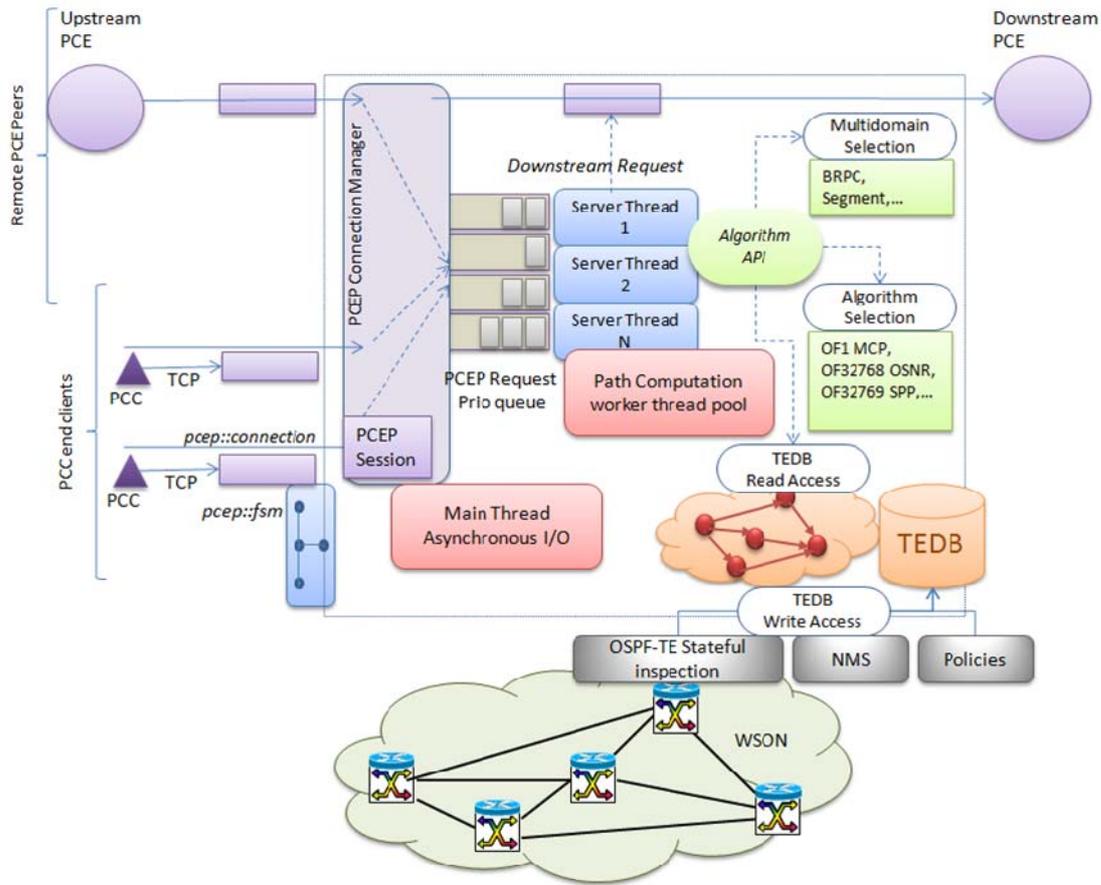
### 4) Future work

We plan to enable TED updates using OSPF-TE LSPs and implement PCE discovery routine.

#### **3.2.1.4 CTTC PCE of the ADRENALINE Testbed**

The ADRENALINE network includes a Path Computation Element (PCE), which is a dedicated network entity responsible for doing advanced path computations. The PCE serves requests from Path Computation Clients (PCCs), and computes constrained explicit routes (EROs) over the topology that constitutes the optical transport layer. The selected PCE deployment model is based on deploying a single PCE per OSPF-TE area, co-located in a GMPLS-enabled Controller node and coupled to a Routing Controller. The preferred synchronization mechanism, by which the PCE constructs a local copy of the Traffic Engineering database (TEDB) is non-intrusive: by sniffing OSPF-TE traffic, it constructs a dedicated (i.e. not shared) database using stateful inspection of the TE Link State Advertisements (LSAs) contained within the OSPF-TE Link State (LS) update messages, thus passively reusing the OSPF-TE dissemination mechanism, and not requiring the creation of an additional listener adjacency.

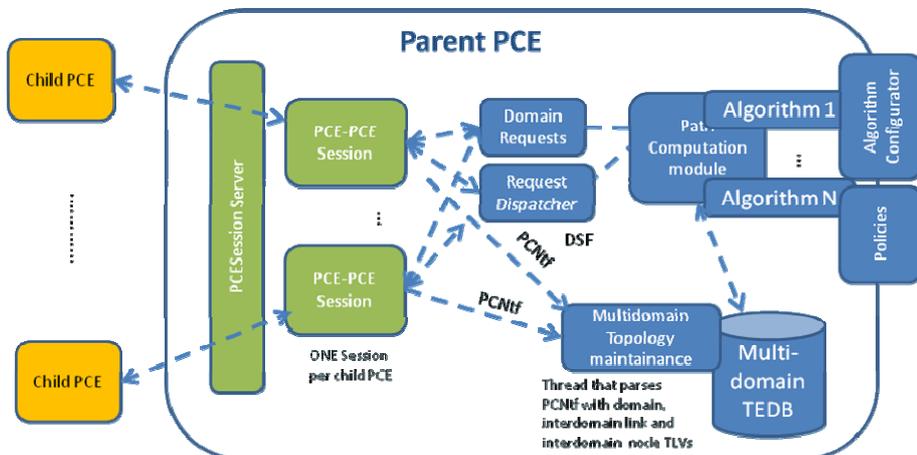
The functional architecture of the PCE is shown in Figure 34. The PCE is a multi-threaded, asynchronous process, serving requests in a client/server approach. Upon acceptance of a client connection, the Finite State Machine drives the PCE Communication Protocol (PCEP) and, after the initial handshake, PCCs may send Path Computation Requests. A dedicated thread is responsible for updating the Traffic Engineering database, while a thread pool is used for the actual path computation, using a writer/readers lock. Algorithms are implemented in shared libraries, using an abstracted algorithm Application Programming Interface (API). This allows algorithm implementations to consider the network topology and the Traffic Engineering database as a directed graph, and to request path computation to other PCE peers for distributed or collaborative Path Computation.



**Figure 34: Functional architecture of the Path Computation Element (PCE) in the ADRENALINE testbed**

### 3.2.1.5 TID Parent PCE Testbed

TID provides a Multi-domain parent PCE (P-PCE) following the hierarchical PCE approach described in D3.2 (Figure 7). For communication between child-parent PCE it follows RFC 5440, with extensions for GMPLS as described in [draft-ietf-pce-gmpls-pcep-extensions-01].



**Figure 35: Functional architecture of the Parent Path Computation Element (P-PCE)**

The parent PCE advertises its capabilities and starts listening for incoming TCP connections. It establishes a PCEP session with each children PCE, which is maintained by periodical keepalives. Upon establishment of the PCEP session, the parent PCE requests the children PCEs to report the domain connectivity information. During the life of a PCEP Session, the parent PCE forwards the notifications of the different children with both domain connectivity information as well as inter-domain information to the “Multi-domain Topology Maintenance” module. This module is a dedicated thread that parses the domain and interdomain information TLVs and maintains the interdomain topology from these PCEP notifications. Currently, only updates of the topology through the PCEP notifications are supported.

After the establishment of the PCE session, the parent PCE is ready to accept interdomain requests. All incoming requests are queued, and processed on a sequential basis. The parent PCE has two modes of operation, depending on the policy and the level of per domain information available. When running in “full topology view mode”, the parent PCE is able to compute the whole end to end path on its own, creating an ERO without the help of the children. When running in “part topology view mode”, the parent PCE computes in a first stage the set of interdomain links and border nodes, and then requests from the different children PCE the details of each domain.

### 3.2.2 Multi-domain PCE architecture to be implemented

The multi-domain PCE scenarios to be implemented in WP4 are based on the STRONGEST reference scenarios, which are described in detail in [STRONGEST\_D3\_1]. In particular, there are two main scenarios: multi-domain, single carrier, WSON regions, and multi-domain, single carrier, multi-region (WSON and MPLS-TP regions in STRONGEST, and OBS region from the MAINS project).

#### 3.2.2.1 Multi-domain, single-region WSON, single-carrier scenario

The first Implementation scenario is depicted in Figure 36. There are several WSON domains interconnected by inter-domain links and, in some cases, border nodes belonging to a pair of adjacent domains. In this scenario, 3R regeneration is mandatory in the Domain Border Elements. Each WSON domain has a PCE for the intra-domain path computation. However, when a multi-domain connection is requested, the request is forwarded to the Parent PCE, which is able to compute the optimal sequence of domains and the end-to-end path. Different levels of summarization as described in WP3 may be tested in these scenarios. Depending on the level of available information in the Parent PCE, the path will have to be completed in the Child PCEs (the WSON domain PCEs) or can be done all in the Parent PCE. Thus, there are two sub-scenarios:

**Abstracted topology view:** Parent PCE computes sequence of domains and interdomain links. Child PCE complete the end-to-end path. Child PCEs provide summarized view.

**Full topology view:** Parent PCE computes full end-to-end path. Child PCEs provide summarized view. Child PCEs provide enhanced topology view.

In the PCE scenario to be implemented, end-to-end lambda LSPs would be requested. However, as the focus of the implementation is on the multi-domain PCE architecture, the actual establishment of the LSP, to be done by means of E-NNI interface will only be performed depending on the partner availability, as no efforts will be devoted to E-NNI implementation.

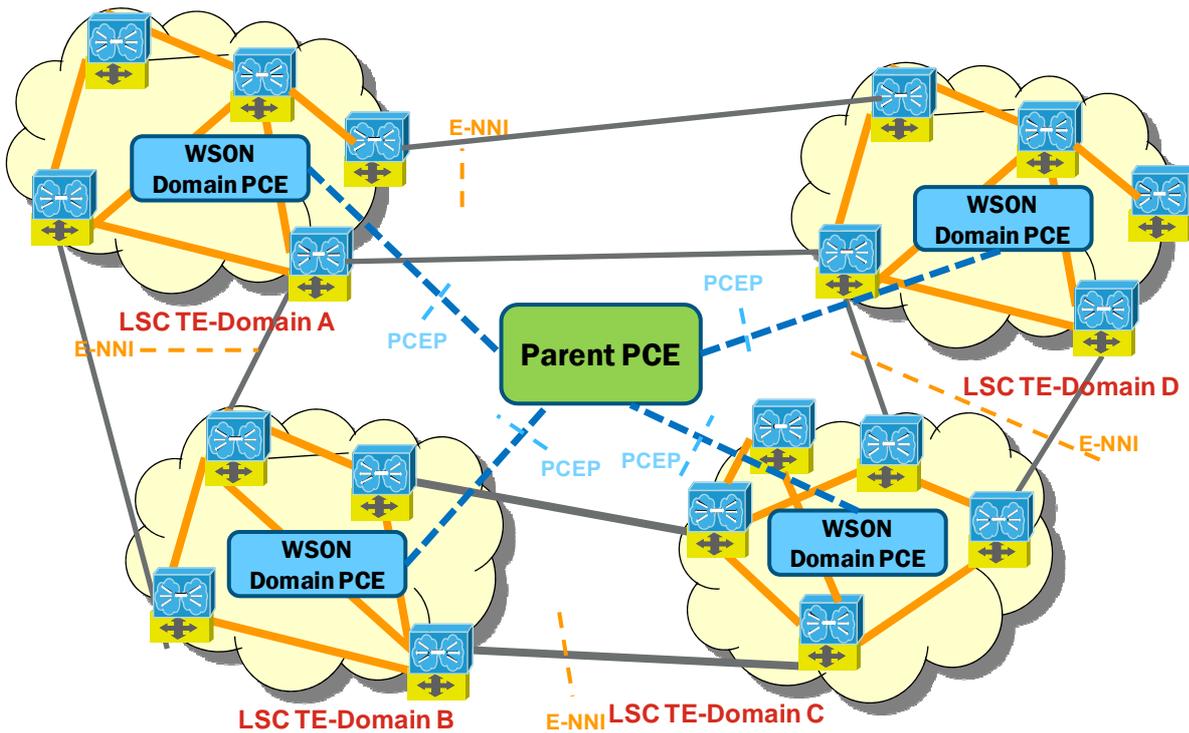


Figure 36 Multi-domain single-region WSON PCE Implementation Scenario

### 3.2.2.2 Multi-domain, multi-region WSON / MPLS-TP single-carrier scenario

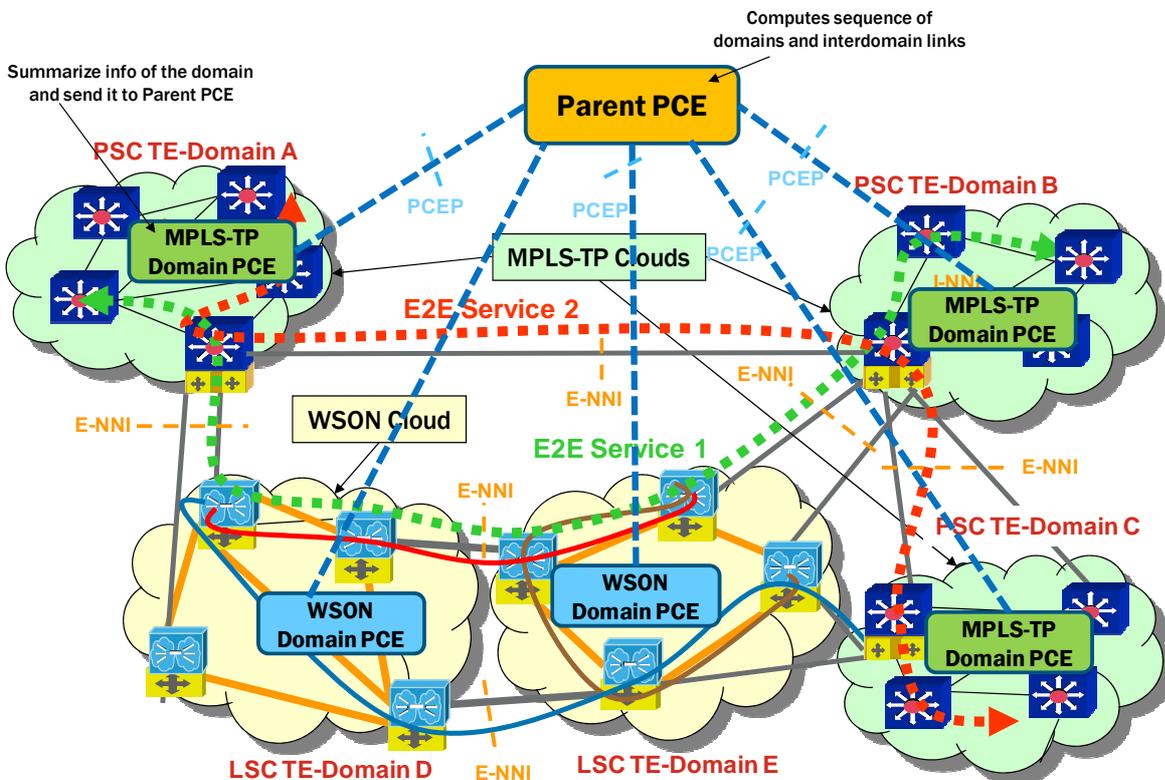


Figure 37: Multi-domain, multi-region, single-carrier hierarchical PCE implementation scenario.

In the PCE implementation scenario depicted in Figure 37, there are two different kinds of regions involved, WSON and MPLS-TP. In this scenario, a hierarchical PCE architecture will be implemented where the first level is composed of Child-PCEs in all the single domains, WSON as well as MPLS-TP, and a parent PCE exchanging routing and topology information with a higher level that has a summarized view of the whole network. Thus, the parent PCE will be in charge of selecting the set of domains and interdomain links. All the details of the intra-domain connections of the e2e path are provided by the different PCEs. As in the previous scenario, the focus is on the multi-domain/multi-region path computation, without efforts on signalling.

### **3.2.2.3 Multi-domain, multi-Region WSON / MPLS-TP/WSON single-carrier scenario**

This scenario involves collaboration with the MAINS project. One of the domains is formed by an OBS cloud, and has a PCE that summarizes its internal details, and allows to know the cost of traversing the cloud. The OBS PCE communicates with the Parent PCE of the previous scenario.

## **3.2.3 Implementation and demonstration plans**

**Participants:** TID, CTTC, NSN, CNIT

**Status:** Multi-technology individual PCE testbeds from different partners already connected

### **Plans**

- Step1 (2011) Multi-domain, single-region WSON, single-carrier scenario
- Step 2 (Q1-Q2 2012) Multi-domain, multi-region WSON / MPLS-TP single carrier scenario
- Step 3 (Q1-Q2 2012) Multi-domain, multi-region WSON / MPLS-TP/WSON single carrier scenario

### **End:**

- Step1: December 2011
- Step 2: June 2012
- Step 3: June 2012

## **3.3 Multi-layer algorithms**

### **3.3.1 Testbed description**

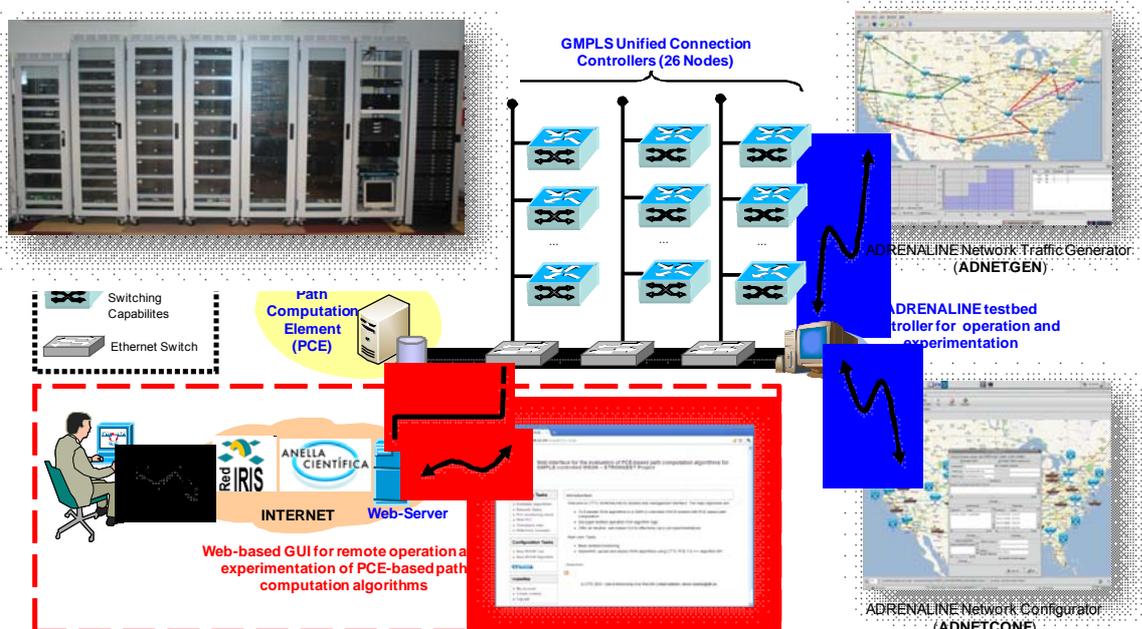
#### **3.3.1.1 Open GMPLS-enabled control plane testbed**

This activity aims at deploying an open, GMPLS-controlled, single-domain, dual-region (MPLS-TP and WSON) emulated transport network to allow third parties to remotely operate the deployed testbed and to evaluate dynamic multilayer algorithms with PCE path computation through an open API. The main features of the proposed testbed are:

- 26 GMPLS Unified Connection Controllers providing a unified control plane for packet (MPLS-TP) and lambda (WSON) switching capabilities. Each GMPLS controller has a global view of MPLS-TP and WSON network topology and resource availability (PSC and LSC TE Links), stored in the Traffic Engineering database (TED) disseminated by OSPF-TE routing protocol.

- Single Path Computation Element (PCE) collocated in a GMPLS Unified Connection Controller, which is a dedicated network entity responsible for doing advanced path computations. The PCE constructs a local copy of the TED by sniffing OSPF-TE traffic. The PCE serves requests from Path Computation Clients (PCCs), and computes constrained explicit routes (EROs) over the topology that constitutes the MPLS-TP and WSON layer
- Fixed reference network topology for all the tests, based on the STRONGEST European Optical Network. It offers a framework for benchmarking and comparing the new proposed algorithms.
- Possibility of designing, implementing, uploading, compiling automatically, and deploying PCE path computation algorithms (i.e., routing and grooming) using the CTTC PCE algorithm API (developed in C++ ). This API encompasses a set of classes, methods and functions to access the traffic engineering Database (TED) with link and node attributes and the required PCE request parameters.
- Manual path computation client (PCC), scheduling automated experimentations requesting connections (LSPs in GMPLS), and obtaining algorithm performance indicators (i.e., blocking probability, setup delay, control bandwidth, etc.).
- Basic testbed monitoring (ping, system up time).

The proposed testbed will be deployed based on extending the GMPLS control plane emulator of the CTTC ADRENALINE testbed®, by developing an intuitive web-based Graphical User Interface (GUI) to allow third parties remote operation and evaluation of dynamic PCE-based multilayer path computation algorithms, as shown in Figure 38.

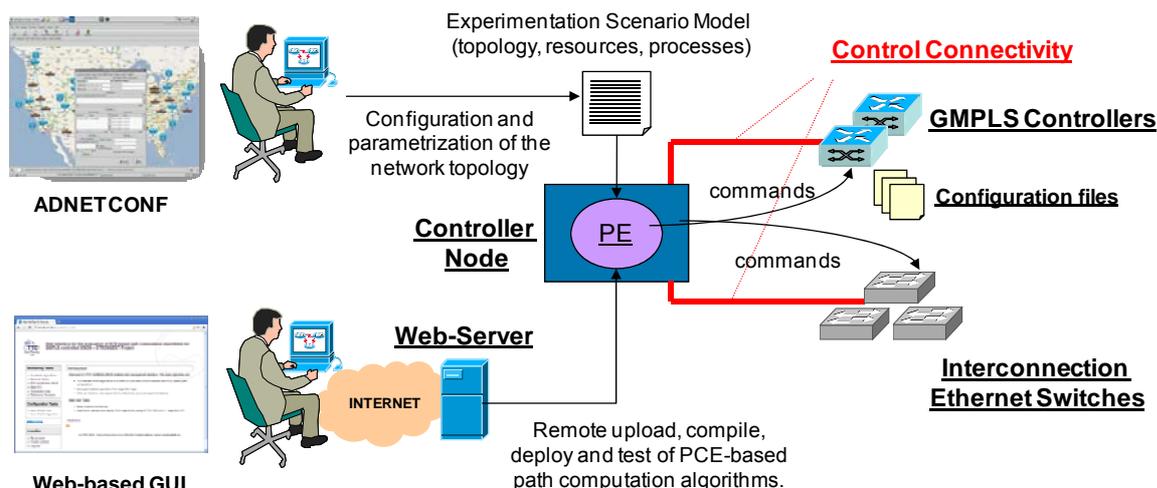


**Figure 38: Open GMPLS-enabled control plane testbed**

The GMPLS-enabled control plane testbed of ADRENALINE is used for the experimental performance evaluation of dynamic GMPLS and PCE-based path computation algorithms

and enhanced protocols (RSVP-TE, OSPF-TE, LMP and PCEP) in complex and diverse network topologies. The ADRENALINE Control Plane Testbed is composed of 74 GMPLS-enabled controllers without associated optical hardware (i.e., emulated switching layer capabilities), each one implemented in a Linux-based router. This set of GMPLS controllers introduces a new degree of flexibility in topology configuration, without restrictions regarding either the targeted optical network topology or the link resources (e.g., number of available wavelengths, fibers, etc.). Thus, the GMPLS controllers can be inter-connected following any devised topology, by means of Ethernet point-to-point links. The proposed solution allows the specification of control link parameters for realistic QoS constraints (fixed and variable packet delays, packet losses, bandwidth limitations, etc.) emulating optical links. To do this, it uses virtual local area networks (IEEE 802.1q VLANs), configured both in the layer 2 Ethernet switches and in the GMPLS-enabled controllers within the testbed, with optional GRE or IP/IP tunneling. Additionally, 4 virtualization servers with KVM/Xen virtualization software techniques have been introduced in the testbed, allowing to build virtualized GMPLS-enabled controllers in a single Linux-based physical hosts for prototyping and extension development.

The proposed architecture of the Web-based GUI for remote PCE-based path computation algorithm operation is shown in Fig 2. An experimentation scenario consists on a particular topology of GMPLS controllers (that will be set with the appropriated configuration of interconnection Ethernet switches; for example, setting the proper VLAN configuration in switches and GMPLS controllers operating systems) plus a particular configuration for each one of the processes running in those GMPLS controllers (i.e. RSVP-TE, OSPF-TE, PCE, etc). A scenario model is a formal specification of an experimentation scenario, written in XML language (with a particular syntax and semantic) so it can be processed automatically. Models include all the information regarding the network topology (network nodes involved in the scenario, the interconnection links among them, addressing issues, etc.) and the configuration of the processes running in each network node. This scenario model is generated locally, using ADNETCONF, a tool specifically designed for testbed model edition (alternatively, it could be written with a general purpose editor, since XML is text-based human-readable). Once the model has been created, it is ready to be processed. The processing engine (implemented with a software program) must run in a node (named controlling node in Figure 39) physically interconnected to the testbed interconnection Ethernet switches. In addition, the controlling node must have pre-existing IP connectivity (control connection in Figure 39) to all the GMPLS controllers and layer2 Ethernet switches. The processing of the model produces interaction with the network and backbone nodes through the control connection. There are two possible interactions: issuing commands (always) and installing configuration files (only in some cases during deployment). There are three different processing actions (deploy, undeploy and monitor). All of them are executed locally using ADNETCONF.



**Figure 39: Architecture of the proposed Web-based GUI for remote PCE-based path computation algorithm operation.**

Once the corresponding scenario is up in the GMPLS control plane testbed, it is ready to be operated remotely by third parties through a web-based GUI. The set of actions defined in the controller node for remote operation are:

- PCE-based algorithm configuration, consisting of: upload of the implemented algorithm (the third party locally creates an algorithm in a C++ file following the algorithm API), automated compilation (reporting whether the compilation was performed correctly or there was some error), and deploy the algorithm, given a pre-shared Objective Function (OF code) in the PCE (a C++ file is compiled as a library and deployed with the PCE). At this point, the algorithm belongs to the PCE pool.
- PCE-based algorithm test, consisting of: manual path computation requests, and scheduling of automated tests for dynamic statistical generation of connection requests in the GMPLS control plane testbed. The modeled request-arrival process is a Poisson one, and the holding time follows a negative exponential distribution, with request events uniformly distributed among all distinct source-destination node pairs. The web-based GUI allows researchers to upload a test parameters file with the definition of the traffic patterns: number of connections, mean inter-arrival time (IAT) in seconds, mean holding time (HT) in seconds, and the OF code of the deployed algorithm. To schedule a test, the researcher is also required to supply his email address, in order to be notified of the test results. Once the test is completed, the controller node logs the outcome of each connection (i.e. successful or failed establishment, average setup delay, etc.), providing data-mining and statistical processing of final results.
- Monitoring, consisting of: Ping and system up time of all the GMPLS controllers involved in the network scenario.

### 3.3.1.2 Dynamical optical bypass demonstrator

The basic idea is to bypass electronic IP routers by making use of photonic cross-connects. Photonic bypasses can be set up either statically, at network commissioning (state-of-the-art solution), or dynamically, when a threshold for the actual transit traffic is passed. By this the scalability of the network is drastically increased, because additional traffic can be treated in the optical layer without the necessity to enhance the capacity of IP

routers, which will lead to both Capex and Opex savings. Moreover, a significant reduction in energy consumption can be realized.

This will be achieved by:

- permanent analysis of transit traffic load in the packet switches of a network,
- selection of those traffic paths that carry the majority of load, and
- reconfiguration of the involved network resources to setup and release bypasses.

The testbed shall be set up to:

- verify the possibility to retrieve the required load data from the switches,
- develop and choose selection algorithms with respect to load minimisation and decision stability,
- verify the scalability of data retrieval and algorithms, and
- demonstrate the power of the dynamic bypass mechanism.

### **3.3.2 Algorithms to be implemented**

#### **3.3.2.1 Multilayer Failure Recovery**

To recover from failures, an online algorithm will be developed and placed in the centralized network management system (NMS). Once the failure is detected and localized, the algorithm is in charge of recovering the affected MPLS LSPs from the failure.

Two types of nodes can be distinguished at the MPLS layer: metro nodes performing client flow aggregation, and transit nodes providing routing flexibility. To minimize the number of ports, metro-to-metro connections are avoided being every metro node connected to one or more transit nodes. However, while it is typical that a transit node is co-located with an OXC node, metro nodes are usually closer to the clients, and thus, it is likely that some ad-hoc connectivity needs to be used to connect metro to transit nodes.

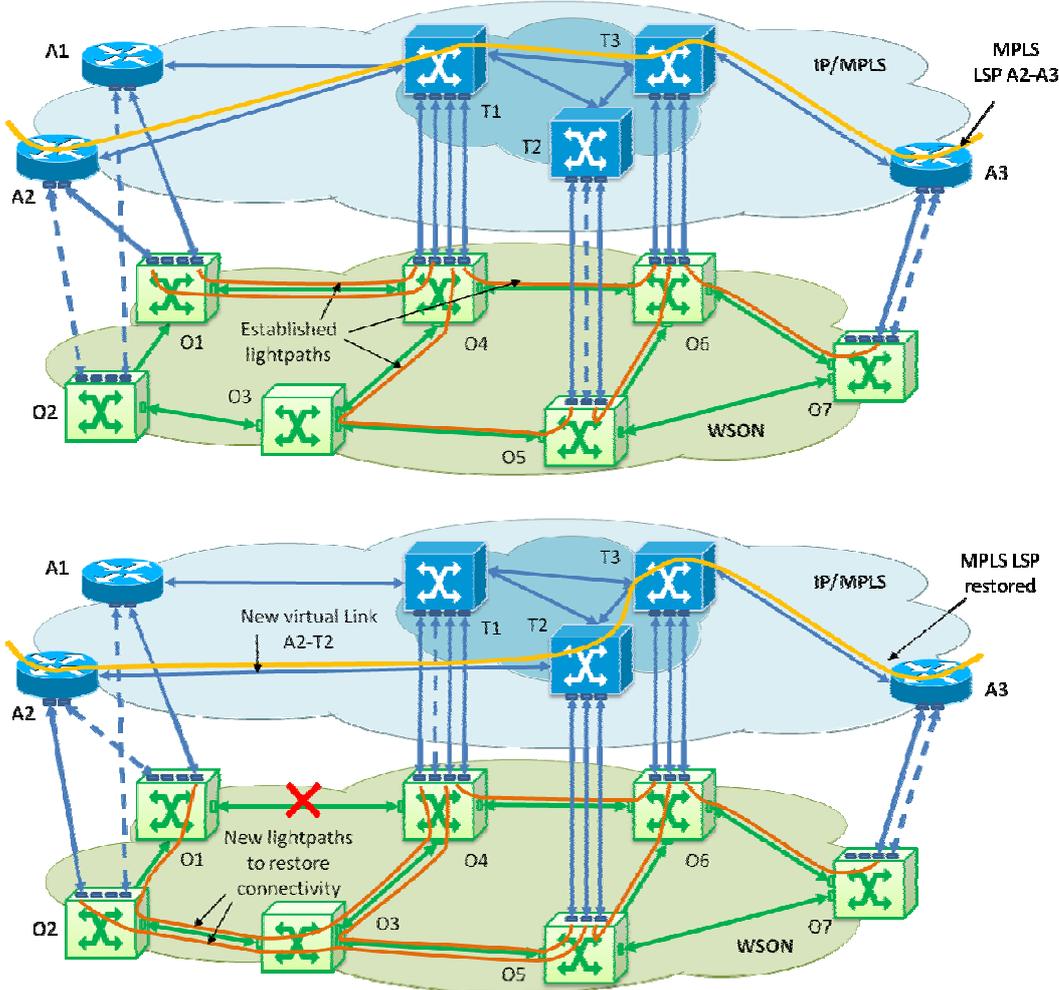
In this context, three failure types are under consideration:

- MPLS transit router failure (software failure),
- MPLS trunk ports or client OXC ports failure,
- WDM optical link failure.

From a general perspective, the algorithm's operation can be divided into two steps:

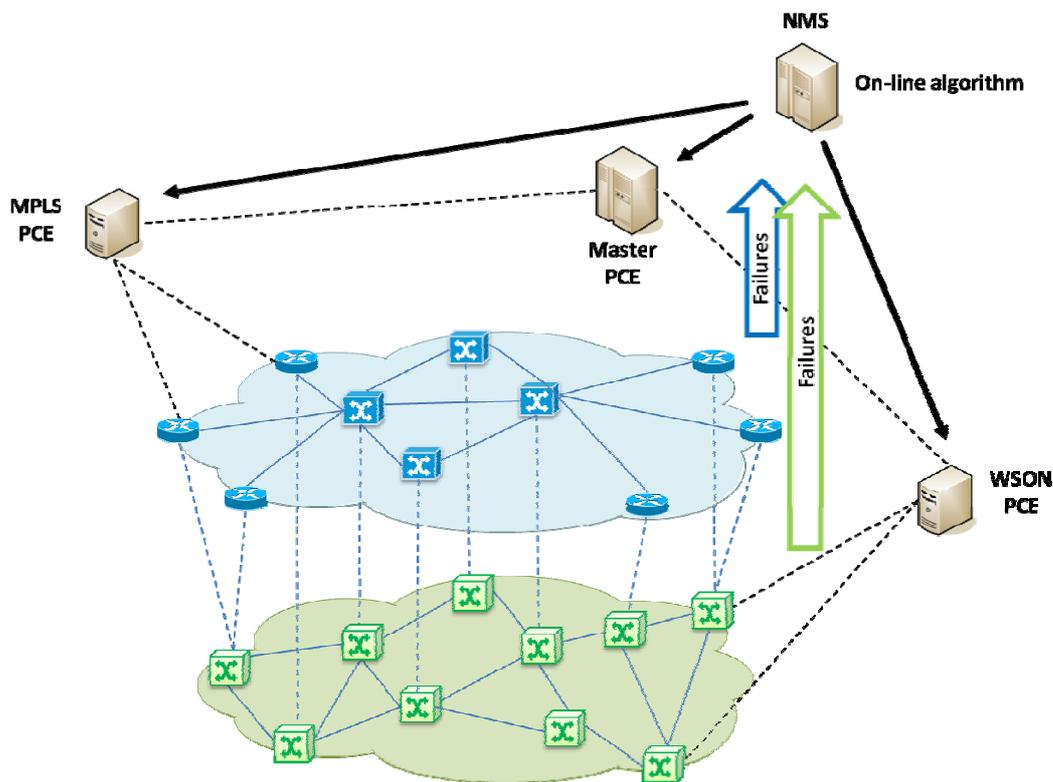
1. Restore the MPLS virtual topology connectivity. This is done by establishing new lightpaths at the optical layer.
2. Rerouting the affected MPLS LSPs over the reconfigured virtual topology.

For illustrative purposes, Figure 40 shows an example of the multilayer recovery algorithm operation. In the example, metro routers A1 and A2 are connected to transit router T1 and metro router A3 to transit router T3 through only one lightpath. To guarantee failure recovery, extra-capacity has been added in every router (dotted lines). When a WDM link fails, the multilayer recovery algorithm applies joint recovery schemes to recover the affected traffic. For instance, when the optical link O1-O4 fails, recovery actions are taken to restore metro-to-transit connectivity. If a lightpath can be restored at the optical layer, the connectivity at IP/MPLS is unaffected. In contrast, if no restoration is possible, a new lightpath has to be established to connect the IP/MPLS metro node to a different transit node and thus to restore the metro-to-transit connectivity. Both cases are shown in the figure for virtual links A1-T1 and A2-T2. Once the connectivity is restored, the MPLS LSPs can be eventually rerouted over the reconfigured virtual topology.



**Figure 40: Multilayer recovery**

The multilayer failure recovery algorithm must be designed as an on-line algorithm to be placed in the centralized multilayer NMS. Since a GMPLS control plane performs MPLS LSP set-up and tear down, whereas route computation is performed in the PCE, some coordination during the recovery period between NMS and PCE must be defined. Figure 41 shows a general scheme where a hierarchical PCE architecture has been considered. Note however that the hierarchical PCE architecture is only one of the possible solutions to coordinate the online recovery algorithm and the multilayer network.



**Figure 41 Hierarchical PCE and centralized NMS coordination**

### 3.3.2.2 Impairment aware load balancing in WSON

Different routing and restoration mechanisms for WSON, designed in WP2, are expected to be implemented in the ADRENALINE testbed. A detailed description of these algorithms will be reported in deliverable D2.2 “STRONGEST node & network architectures for energy efficiency and scalability”.

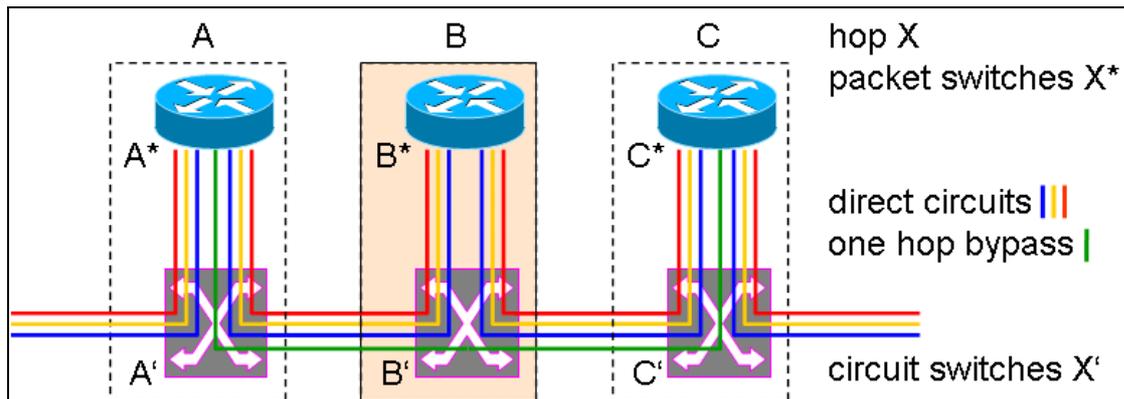
Specifically, such routing algorithms will encompass a wavelength assignment and routing algorithm, aware of physical layer transmission limitations (all major degradation sources for NRZ modulation), plus higher-layer functionality including TE (load-balancing) and restoration capability.

The implementation and experimental evaluation of these mechanisms aim to demonstrate the enhanced performance of STRONGEST in terms of survivability and network resources optimization.

### 3.3.2.3 Dynamic optical bypass

The dynamic optical bypass demonstrator shall be realised in two steps.

In a first step the focus will be directed to the retrieval of information from the packet switches and to the reconfiguration of the affected network resources. The demonstrator shall be implemented as a three-node network with a simple linear connection scheme and a possible one-hop-bypass (see Figure 42).



**Figure 42: Optical bypass demonstrator network (step 1)**

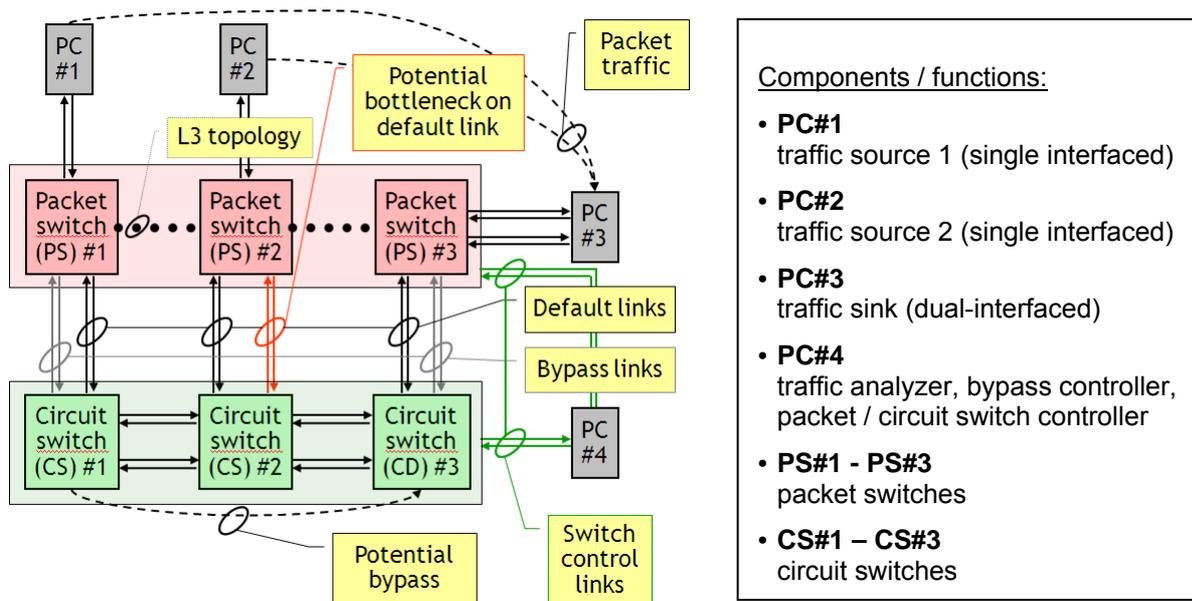
Different types of traffic like video streams and bulk data transfer shall be provided; the bypass triggering shall be based on a two-point decision model with upper and lower thresholds. Additionally, a sensitivity handling shall be implemented to avoid oscillations, i.e. unstable network conditions.

In the second step the limitations of a one-hop-bypass shall be overcome using an extended network topology. In that step the focus is on testing different algorithms developed in WP2 to select the most efficient bypass with its optimal length.

Investigations on more complex network topologies and on scalability issues towards 100Tbit/s networks shall be fostered by calculations and by simulations.

The network implementation for step 1 is illustrated in Figure 43. There are three switches ('#1', '#2' and '#3'), each consisting of a packet switch ('PS') and a circuit switch ('CS') part. Traffic will be generated by two sources ('PC#1' and 'PC#2') and sent to a common sink ('PC#3'). The transit traffic will be measured in the intermediate packet switch ('PS#2') and used as the criteria for bypass switching; the bypass traffic will be measured in 'PS#1'.

The demonstrator will be based on standard equipment. The three (logical) circuit switches will be realised in one physical device, and so will the packet switches, too.

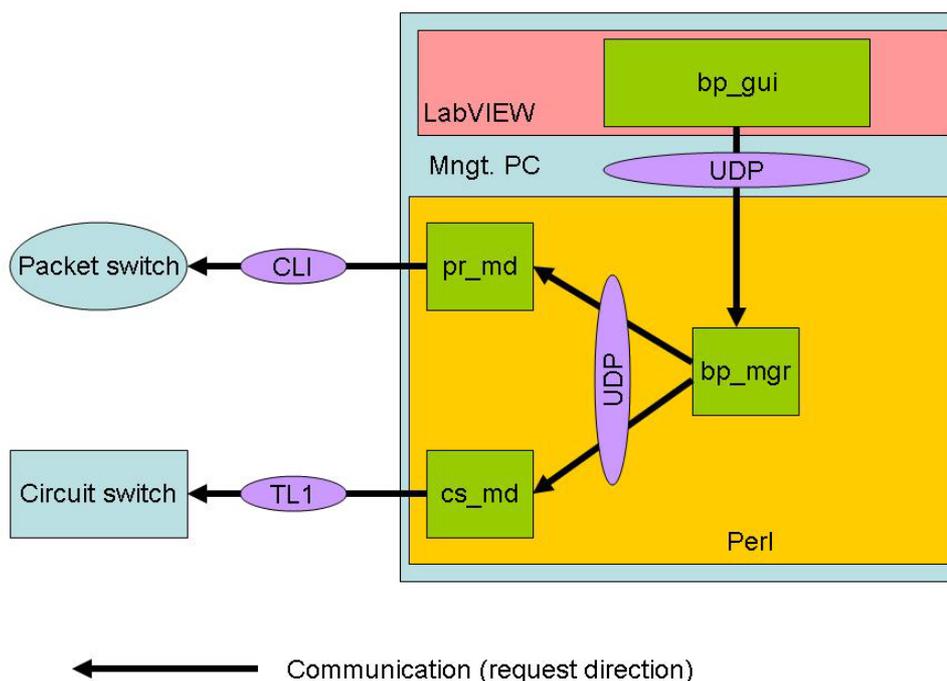


**Figure 43: Functional components of the bypass demonstrator (step 1)**

The measurement, the bypass triggering, and the equipment reconfiguration will be realised by in-house software, which shall be located at 'PC#4'. The functional components of the bypass management software are (see Figure 44):

- **Bypass manager (bp\_mgr):**  
The bypass manager periodically requests traffic counter values from the packet switches. It collects the data and provides these as time series. Based on the time series and built-in decision schemes, the bypass manager is able to set up and tear down a bypass.
- **Packet switch (pr\_md) and circuit switch mediation device (cs\_md):**  
The bypass manager communicates with the switch hardware through mediation device modules. The latter translate the monitoring queries and configuration commands to the different languages and workflows that are used to control the network devices.
- **Bypass management GUI (bp\_gui):**  
The graphical user interface provides a user-friendly view of the transported traffic volumes and controls/illustrates the state of the demonstrator, i.e. it displays the current states of counters and bypasses and provides means to control the demonstrator. It also makes the effect of dynamic optical bypassing visible.

Figure 44 also shows the interfaces between the software components as solid lines. For simplicity reasons, these interfaces are unidirectional and are pointing to the request direction as indicated by the arrows. Data queries and configuration commands can be sent only downstream, the response is then sent back to the requester.



**Figure 44: Structure of the bypass demonstrator software**

Each of the demonstrator software components is designed in a specific programming language and works as a separate process. The GUI is implemented in LabView while all other components are implemented in Perl. The components communicate with each other through UDP sockets, a method which makes it easy to replace a single component by another one. UDP, as a fast and commonly available protocol, is also used for internal inter-process communication. Both switches provide a terminal-based command line interface (CLI) that is used by the respective mediation device modules to query and configure the switches.

For step 2 of the demonstrator, the implementation according to Figure 43 shall be extended by at least two virtual nodes to provide concurrent paths and multi-hop bypass capabilities. The management software must be adapted to handle more than one physical switch as well as to manage several bypasses. Furthermore, it must be equipped with an appropriate decision algorithm.

Based on the results of step 2 and on the algorithms developed in WP2 calculations and simulations (software) shall be provided for larger networks. This work will be done in a co-operation among partners.

### 3.3.3 Implementation and demonstration plans

#### 3.3.3.1 Open GMPLS-enabled control plane testbed

Participants: CTTC

Implementation Plans

- Deployment of the Web-based GUI for remote operation and experimentation of PCE-based path computation algorithms

- Current status: Preliminary implementation with full functionalities.
- End: 31<sup>st</sup> December 2010 (for WSON/PSC).
  
- GMPLS-enabled Unified Control Plane (LSC, PSC)
  - Current status: Preliminary development and tests of the Link Resource Manager (LRM).
  - End: 31<sup>st</sup> December 2011

#### Demonstration plans

- Remote operation and experimentation of PCE-based path computation algorithms for GMPLS-controlled single-domain WSON networks.
  - The platform will be ready to third parties from March 1<sup>st</sup> 2011.
- Remote operation and experimentation of PCE-based path computation algorithms for GMPLS-controlled single-domain MPLS-TP over WSON.
  - The platform will be ready to third parties from February 1<sup>st</sup> 2012.

### 3.3.3.2 Multilayer Failure Recovery

Participants: CTTC, UPC

Status: The algorithms to be implemented are already defined in WP2

Plans: The foreseen plan for the multilayer failure recovery algorithm is as follows:

- Algorithm design and implementation: Q2 2011
- Simulation: Q3-Q4 2011
- Experimental demonstration (if testbed available): Q1 2012

### 3.3.3.3 Impairment aware load balancing in WSON

Participants: CTTC, TID, UoP

Status: The algorithms to be implemented are already defined in WP2

Plans: The foreseen plan for the multilayer failure recovery algorithm is as follows:

- Algorithm design and implementation: Q2 2011
- Implementation in ADRENALINE testbed: Q3-Q4 2011
- Experimental demonstration: Q1 2012

### 3.3.3.4 Dynamic Optical bypass demonstrator

Participants: Alcatel Lucent

Status: The optical bypass architecture to be implemented is already designed

Plans: Dynamic optical bypass implementation 2011-2012

### 3.4 Interface between Control Admission and GMPLS

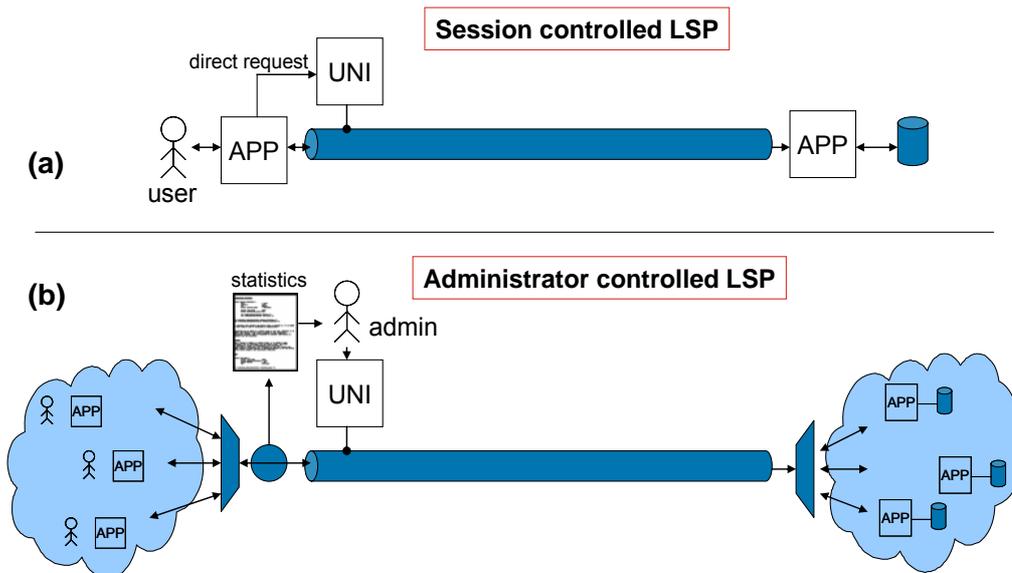
#### 3.4.1 Traffic monitoring/management in Terabit/s packet networks

Label Switched Paths (LSP) with resource reservation along the path are the one of the Quality-of-Service (QoS) promises of GMPLS. It is assumed that clients, prior to a particular transmission, request the necessary resources (transmission capacity) along the path to the destination. Once the resources are granted, a client can be assured of lossless communication. This seemingly simple approach raises a number of questions:

- (i) Who is requesting an LSP, how frequently does he so and how fast is he expecting an operational connection?
- (ii) How frequently an LSP setup request is being blocked due to exhausted resources (admission blocking)?
- (iii) How much capacity has to be requested, and how to assure that the real transmission does not exceed the granted resources?

The present test-bed is mainly addressing the last question, however, for clarification we have to look at question (i) first. Just for completeness, question (ii) is a rather economic dimensioning task around the Erlang B-formula, deciding whether or not it is worth to have sufficient unallocated reserves ready.

We distinguish between two types of LSP requestors (Figure 45): (a) end users who are running a network application that in turn requests the LSP and (b) network administrators who are acting on behalf of groups of users, and who are requesting tunnel LSPs for aggregated traffic of client networks (layers, domains, etc). A more detailed analysis of this classification of end-to-end services can be found in [STRONGEST\_D3\_2]. There is consensus in the consortium that the STRONGEST architecture is mainly focused on the administrator controlled LSP (Figure 45, alternative (b)).



**Figure 45: Session controlled (a) and administrator controlled LSP (b): Only the session controlled LSP knows the exact request time and size, while the administrator controlled LSP relies on statistics (averages)**

The administrator controlled LSP is typically set up in advance of the real traffic arrival, so that the set up delay and even admission blocking are out of the scope. The desired capacity is estimated by a combination of traffic statistics and forecast, which is inherently a slow process. Individual short term traffic load changes are reflected by a fluctuation overhead above the forecasted average traffic. If the fluctuation reserve is too small or if the forecasted values do not hold, there might be packet losses anyway, even though there is an exclusive capacity reservation.

### **3.4.1.1 Test-bed description**

In this test-bed Alcatel Lucent will implement a traffic gate for the collection of the statistical data. Based on this, the gate is intended to signal various forecast parameters to the GMPLS control plane, particularly the actually required capacity (effective bandwidth), the application stream bit-rate granularity as a metric of traffic volatility, and the expected packet loss ratio, see [STRONGEST\_D3\_2]. Furthermore we will implement experimental traffic admittance functions that rely on the statistical data in the traffic gate. The intention here is to establish Service Level Agreement (SLA) limits for various traffic parameters and detect potential violations and feed back the traffic impairments to the originators. The traffic admittance functions are part of the wholesale QoS domain concept [STRONGEST\_D3\_1].

### **3.4.1.2 Architecture to be implemented**

The test-bed will consist of several packet switching nodes with adjustable port rates that emulate the traffic engineered LSPs. The traffic gate will be implemented on an existing prototype board with two 10 Gbit/s Ethernet interfaces for data plane traffic and a Gigabit Ethernet interface for the control plane connection. The gateway logic (frame forwarding, collection of statistical data, packet marking, policing, etc.) will be implemented in FPGAs (Field Programmable Gate Arrays) on the prototype board. This enables transparent processing of arbitrary Ethernet traffic up to the full interface line rate of 10Gbit/s. For experiments with multiple traffic gate instances we will use additional software implementations with similar functionality but reduced performance (throughput).

The test-bed will be complemented by different types of traffic generators for payload and cross traffic generation. Quantitative analysis will be based on packet counters, capture devices, and Ethernet testers. Qualitative demonstration of the concepts will be done by commonplace applications like file transfer, or video and voice streaming.

Special attention will be devoted to the control plane interface. We will not implement a full featured control plane in this demonstrator, since we are not focused on specific control plane internal procedures. Instead, we will emulate the control plane communication to/from the data plane by light weighted graphical user interfaces for manual control and visualization. An exception from this rule will be the signalization to LSP end points. Here we are planning experiments with specially adapted end systems that make use of, or feed, the newly introduced OAM signalization.

### **3.4.1.3 Implementation and demonstration plans**

The realization of the FPGA based traffic gate can be done in parallel with the rest of the test bed. Implementation and most of the module testing (GUI, visualization, adaptation of end system protocol stacks) will be performed in a sub-equipped test bed with reduced load conditions.

The demonstration will consist of a selected number of traffic scenarios that highlight the newly introduced features under realistic traffic conditions.

Participants: Alcatel Lucent

Status: first software demonstrator available to calculate the required capacity on a link.

Plans: hardware realization and signalling of parameters to the Control Plane to be implemented.

End: December 2012

### **3.4.2 Control plane architecture for RACS-PCE interworking**

One of the most challenging objectives of the STRONGEST target architecture is to minimize the number of routers in the core network. This implies that a lot of small routers at the edge, implementing aggregation functions, should be connected, without higher hierarchy IP devices. With thousands of access nodes, a full mesh of the IP layer will lead to hundreds of thousands of virtual links and big routing tables in IP routers. Scalability concerns arise as the IP control plane technologies have not been designed for such scenarios, especially as regards maintaining a huge number of adjacencies, as well as incurring potential routing storms in case of link failure. These scalability problems can be solved in different ways, which are currently under study. Among others, STRONGEST devotes attention to an approach based on interworking between the Resource and Admission Control Subsystem (RACS) and the Path Computation Element (PCE).

In detail this approach can assure:

- Automated end-to-end service provisioning and QoS assurance.
- Scalable GMPLS-based control plane solutions in MPLS-TP/WSN networks composed by thousands of nodes.
- End-to-end OAM mechanisms enabling automated network failure detection between IP metro nodes.

The considered infrastructure consists of a hierarchical PCE architecture interacting with a GMPLS distributed control plane, and both embedded in a wider control framework provided by either the ETSI TISPAN<sup>1</sup> RACS or the 3GPP<sup>2</sup> PCC (Policy and Charging Control).

#### **3.4.2.1 Test-bed description**

The TI test-bed is distributed over two different laboratories: the “optical networking” lab and the “multiservice network control” lab. They are interconnected using internal infrastructure. The same infrastructure provides connection at the control plane layer with the test-beds of other partners.

---

<sup>1</sup> ETSI TISPAN:

[http://portal.etsi.org/portal/server.pt/community/default\\_community/redirect\\_page?TISPAN](http://portal.etsi.org/portal/server.pt/community/default_community/redirect_page?TISPAN)

<sup>2</sup> 3GPP: <http://www.3gpp.org/>

In the optical networking lab it is hosted the network prototype developed in the IST-NOBEL project, composed of six nodes with Fiber Switching Capability (FSC) and controlled by an ASON/GMPLS control plane running on six PCs with the RedHat Linux operating system [NOBEL\_D25]. Since the original implementation, the control plane has been extended and now can also be configured to run over a different data plane consisting of commercial Ethernet switches whose forwarding table is updated by means of RSVP-TE signalling instead of the classical “MAC-learning” feature. In this way, the GELS (GMPLS Controlled Ethernet Label Switching) framework is accomplished.

In the “multiservice network control” lab a RACS prototype developed within the IST-MUPBED project and a backup/restore application enhanced with the capability of requesting dedicated bandwidth to the network are deployed. This is accomplished by means of a Gq'-like interface (based on SOAP-Apache Axis 1.4) implementing the API between the NSR (Network Service Requester) and the NSP (Network Service Provider) [MUPBED\_D2\_5]. The RACS prototype, implementing the NSP function, is composed of a SPDF (Service Policy Decision Function) interacting with several x-RACFs (Resource and Admission Control Function) dedicated to specific intra-carrier network domains by means of an XML/HTTP Rq-like interface. The selection of the x-RACFs involved in the connection is made by a Context DB pre-filled with static topology information and dynamically updated as far as resource utilization is concerned.

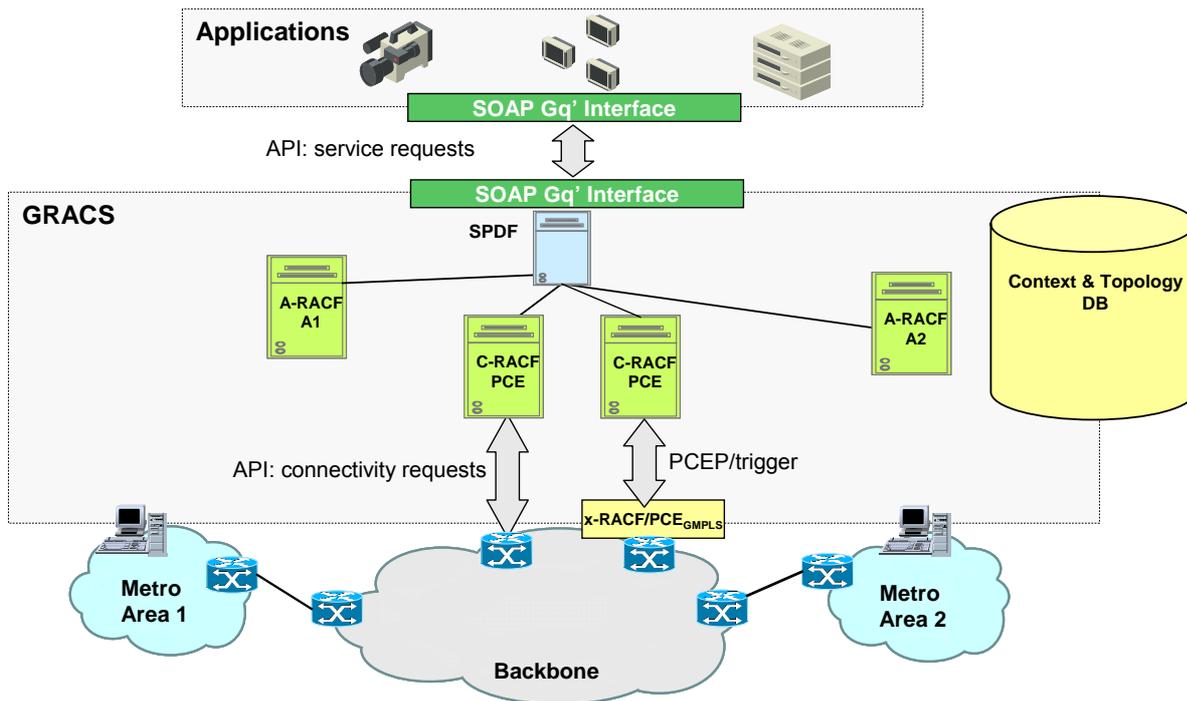
The x-RACFs modules interact with the ASON/GMPLS control plane through a telnet session.

### **3.4.2.2 Architecture to be implemented**

The architecture that will be implemented is depicted in Figure 46. It enhances the available RACS prototype with multi-layer routing capabilities by interacting with a GMPLS control plane based on hierarchical PCE.

The functional elements constituting such architecture are the Service Policy Decision Function (SPDF) module and the integration of the x-RACF with the PCE entities. The SPDF performs different functions such as: providing the network topology and technology abstraction to the applications, inter-domain/carrier communication, implementing a policy decision function and coordinating multiple x-RACF/PCE modules for multi-domain and multi-layer path computation and resource admission control purposes. Beside managing the inter-domain path admission control functions, the x-RACF/PCE module is responsible for interfacing with the GMPLS-enabled control plane which automatically sets-up the connections. The x-RACF can be either fully centralized, fully distributed (in the network nodes), or partially distributed (with many x-RACF instances, organized in a tree structure, controlling specific network domains).

The detailed network architecture together with the required interfaces (application-RACS, RACS-PCE, RACS-control plane) is currently under discussion in WP3.



**Figure 46: architecture to be implemented**

### 3.4.2.3 Implementation and demonstration plans

Participants: Telecom Italia, Alcatel Lucent

Status: The architecture to be implemented is already defined

Plans:

The implementation of the integrated RACS/GMPLS architecture will be mainly focused on a partially distributed approach that seems to be the most promising option since it offers a layered view with a clear separation between the transport functions (distributed in the low-tier modules) and the service functions (implemented in the top tier module). However, at this stage, it cannot be excluded that the other two approaches will be also supported, enlarging therefore the number of possible scenarios of the demo.

The demo scenario that is currently envisaged involves uniquely devices hosted in the TI laboratories. For this purpose, the current RACS implementation will be enhanced with multi-domain routing capabilities either integrating PCE functionalities or interfacing with external PCE devices. All the required interfaces (or a subset of them) will be implemented according to their definition studied in WP3. Moreover, the control plane prototype will be extended by adding PCE functionalities to the RC (Routing Controller) module and implementing the PCEP protocol features and objects that are strictly necessary for the demo. On the applications side, the already available backup/restore application deployed for the MUPBED project demonstration [MUPBED\_D3\_4] could be reused, but it's also under study the possibility to adopt other applications that should be accordingly adapted. The Wireshark [WRSK] network protocol analyzer will be used for the dissection of the messages exchanged between the network elements and for the analysis of the correct protocol message exchange. If required, it will be extended in order to decode new objects and/or messages that will be used.

Even if the kind of data planes available in the TI test-bed are only partially aligned with the mid-term network solution analyzed by the Strongest project (MPLS-TP over WSON), we believe that they can be positively adopted for the described demonstration since the functionalities of the integrated RACS/GMPLS architecture are independent of the underlying data-plane technology.

The possibility to interact with the test-beds of other partners is also under study, in order to show a more complete multi-domain / multi-carrier / multi-layer scenario, leveraging the VPN deployed in the project, which interconnects the sites of different partners. In order to meet this goal, the effort that is required to implement the necessary interfaces in all the candidate test-beds must be carefully evaluated.

End: December 2011

## 4 List of acronyms

3GPP	Third Generation Partnership Project
ADRENALINE	All-optical Dynamic RELiable Network hAndLING IP/Ethernet Gigabit traffic with QoS
API	Application Programming Interface
APU	Auxiliary Processor Unit
ASON	Automatic Switched Optical Network
ATM	Asynchronous Transfer Mode
AWG	Array Waveguide Grating
BER	Bit Error Ratio
BERT	Bit Error Ratio Test
BRPC	Backward-Recursive PCE (Path Computation Element)-based Computation (BRPC)
CAC	Connection Admission Control
CCC	Client Call Controllers
CLI	Command Line Interface
CMR	Click Modular Router
CP	Control Plane
CPE	Customer Premises Equipment
DCF	Dispersion Compensating Fibre
DDR	Double Data Rate
DMA	Direct Memory Access
DME	Data Modify Engine
DPBRAM	Dual Port Block RAM
DP-QPSK	Dual-Polarization Quadrature Phase Shift Keying
DWDM	Dense Wavelength-Division Multiplexing
EDFA	Erbium Doped Fibre Amplifier
E-NNI	External Network-to-Network Interface
ERO	Explicit Routing Object
FA	Forwarding Adjacency
FCS	Frame Check Sequence
FEC	Forwarding Equivalent Class
FPGA	Field Programmable Gate Array
FTN	FEC To NHLFE
FTS	FEC To Service
FSC	Fibre Switching Capability

FSL	Fast Simplex Link
GELS	GMPLS Controlled Ethernet Label Switching
GMPLS	Generalized Multiprotocol Label Switching
GRE	Generic Routing Encapsulation
GUI	Graphical User Interface
HD	High Definition
HDTV	High-Definition Television
HT	Holding Time
IAT	Inter-Arrival Time
IETF	Internet Engineering Task Force
ILM	Incoming Label Map
I-NNI	Internal Network-to-Network Interface
IP	Internet Protocol
IPIP	IP in IP
ITU-T	International Telecommunication Union-Telecommunication
IV	Impairment Validation
KVM	Kernel-based Virtual Machine
LAN	Local Area Network
LCOS	Liquid Crystal On Silicon
LDP	Label Distribution Protocol
T-LDP	Targeted Label Distribution Protocol
LER	Label Edge Router
LMP	Link Management Protocol
LRM	Link Resource Manager
LSA	Link State Advertisement
LSP	Label Switched Path
LSPA	LSP Attributes
LSR	Label Switch Router
MAC	Medium Access Control
MAC-DA	MAC Destination Address
MAINS	Metro Architectures enablINg Sub-wavelengths
MA-SA	MAC Source Address
MEMS	Micro Electro-Mechanical System
MLL	Mode-Locked Laser
MPLS	Multi Protocol Label Switching
MPLS-TP	Multi Protocol Label Switching – Transport Profile

MW	Middleware
MZM	Mach-Zehnder Modulator
NCC	Network Call Controllers
NHLF	Next Hop Label Forwarding
NHLFE	Next Hop Label Forwarding Entry
NIC	Network Interface Card
NMS	Network Management System
NRZ	Non-Return-to-Zero
NSI-WG	Network Service Interface Working Group
NSP	Native Service Processing
NSP	Network Service Provider
NSR	Network Service Requester
OBS	Optical Burst Switching
OBST	Optical Burst Switching Technology
OCM	On-Chip Memory
OF	Objective Function
OIF	Optical Internetworking Forum
OFDM	Orthogonal Frequency-Division Multiplexing
OMUX	Optical Multiplexer
OPST	Optical Packet Switching Technology
OSA	Optical Spectrum Analyzer
OSNR	Optical Signal-to-Noise Ratio
OSPF	Open Shortest Path First
OTN	Optical Transport Network
OXC	Optical Cross-Connect
PCE	Path Computation Element
PCEP	PCE Communication Protocol
PCC	Path Computation Client
PCC	Policy and Charging Control
PCI	Peripheral Component Interconnect
PCS	Path Computation Solver
PE	Process Engine
PMD	Polarization Mode Dispersion
PPE	Paralleled Process Engine
PPEC	Paralleled Process Engine Cluster
PRBS	Pseudo Random Binary Sequence

PSC	Packet Switching Cluster
PSN	Packet Switched Network
PTN	Packet Transport Network
PW	Pseudo-Wire
PWE	Pseudo-Wire Emulation
QDR	Quad Data Rate
QoS	Quality of Service
(x-) RACF	Resource and Admission Control Function (x=A → Access; x=C → Core)
RACS	Resource and Admission Control Subsystem
RAM	Random Access Memory
RP	Request Parameters
ROADM	Reconfigurable Optical Add-Drop Multiplexer
RSVP	Resource Reservation Protocol
RWA	Routing and Wavelength Assignment
RZ	Return-to-Zero
QPSK	Quadrature Phase Shift Keying
SDH	Synchronous Digital Hierarchy
SDO	Standards Developing/Development Organization
SHDAN	Software/Hardware Defined Adaptable Network
SONET	Synchronous Optical NETworking
SPDF	Service Policy Decision Function
SRLG	Shared Risk Link Group
SVEC	Synchronization VECtor
TBD	To Be Defined
TCP	Transmission Control Protocol
TDM	Time-Division Multiplexing
TE	Traffic Engineering
TED	Traffic Engineering Database
TLV	Type Length Value
TM	Traffic Management
TNA	Transport Network Assigned
UDP	User Datagram Protocol
UNI	User to Network Interface
UNI-C	User to Network Interface Client
UNI-N	User to Network Interface Network
VID	VLAN ID

VLAN	Virtual Local Area Network
VOA	Variable Optical Attenuator
WSON	Wavelength Switched Optical Networks
WSS	Wavelength Selective Switch
XFP	10 Gigabit/s Small Form Factor Pluggable
XML	Extensible Markup Language
XPM	Cross-Phase Modulation
XRO	Exclude Route Object

## 5 References

- [Tucker\_2008] R.Tucker, J.Baliga, R.Ayre, K.Hinton, W.Sorin "Energy Consumption in IP Networks", ECOC 2008
- [Amaya\_2011] N. Amaya, I. Muhammad, G. S. Zervas, R. Nejabati, D. Simeodinou, Y. R. Zhou, A. Lord, "Experimental Demonstration of a Gridless Multi-granular Optical Network Supporting Flexible Spectrum Switching," submitted to OFC 2011
- [Jiang\_2004] Jiang et al, "Transparent electro-optic ceramics and devices," SPIE Photonics Asia, November 2004
- [Nashimoto\_2010] K. Nashimoto, D. Kudzuma, and H. Han, "High-speed switching and filtering using PLZT waveguide devices," OECC 2010
- [E.493] ITU-T Recommendation E.493, "GRADE OF SERVICE (GOS) MONITORING", February 1996
- [E.720] ITU-T Recommendation E.720, "ISDN GRADE OF SERVICE CONCEPT", 1993
- [E.721] ITU-T Recommendation E.721, "Network grade of service parameters and target values for circuit-switched services in the evolving ISDN", May 1999
- [E.771] ITU-T Recommendation E.771, "Network grade of service parameters and target values for circuit-switched public land mobile services", October 1996
- [G.808.1]. ITU-T Recommendation G808.1: Generic protection switching – Linear trail and sub-network protection.
- [G841] ITU-T Recommendation G.841, "Types and characteristics of SDH network protection architectures", October 1998
- [NOBEL\_D25] IST project NOBEL deliverable D25: "Solutions for inter-domain and multi-layer NM & CP and Service Management concepts; NM and ASON prototype functional and design specification and test plan", 2005
- [MUPBED\_D2\_5] IST project MUPBED deliverable D2.5: "Final results on application-network interface on the MUPBED test bed", 2007
- [MUPBED\_D3\_4] IST project MUPBED deliverable D3.4: "Preliminary report on MUPBED test bed integration, interworking and operation", 2006
- [WRSK] <http://www.wireshark.org>
- [STRONGEST\_D3\_2] ICT project STRONGEST deliverable D3.2: "E-NNI extensions, PCE multi-domain architecture, OAM parameters and Traffic Admittance", 2010
- [STRONGEST\_D3\_1] ICT project STRONGEST deliverable D3.1: "Medium-term multi-domain reference model and architecture for OAM, control plane and e2e services", 2010
- [PCE] A. Farrel, J.-P. Vasseur, J. Ash, "A Path Computation Element-based Architecture", RFC 4655, IETF

- [PCEP] J.-P. Vasseur, J.-L. Le Roux, "Path Computation Element Communication Protocol (PCEP)", RFC 5440, IETF
- [WSON-IMP] Y. Lee, G. Bernstein, D. Li, G. Martinelli, draft-ietf-ccamp-wson-impairments-04, IETF
- [BRPC] J.-P. Vasseur, R. Zhang, N. Bitar, J.-L. Le Roux, "A Backward Recursive PCE-based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, IETF

## 6 Document History

0.01	29/09/2010	Juan Fernandez Palacios	D4.2 template distribution
0.04	21/10/2010	Juan Fernandez Palacios	D4.2 ToC 1 <sup>st</sup> proposal
0.06	12/11/2010	Juan Fernandez Palacios, WP4 contributors	D4.2 1 <sup>st</sup> draft (incomplete)
1.00	17/12/2010	Juan Fernandez Palacios, WP4 contributors	D4.2 pre-final draft
1.10	22/12/2010	Juan Fernandez Palacios	D4.2 version for final for quality check
1.20	30/12/2010	Emilio Vezzoni	Quality checked version
2.00	05/01/2011	Juan Fernandez Palacios, Emilio Vezzoni, Andrea Di Giglio	Final draft for General Assembly approval
3.00	11/01/2011	Emilio Vezzoni, Andrea Di Giglio	Approved version