Optical IP Switching for dynamic traffic engineering in next-generation optical networks

Marco Ruffini, Donal O'Mahony, Linda Doyle

Centre for Telecommunication Value-Chain Research
University of Dublin, Trinity College
Dublin 2, Ireland
{ruffinim@tcd.ie, Donal.OMahony@cs.tcd.ie,lodoyle@tcd.ie}

Abstract. WDM technology has increased network link capacity dramatically, moving the network bottleneck from the transport to the routing layer. Hybrid electro-optical architectures seem at the moment a reliable and cost-effective solution for near-future implementations of the routing/switching layer. In this paper we present a novel approach to dynamic optical circuit switching based on the Optical IP Switching (OIS) model. OIS nodes classify IP packets by destination prefix, aggregating them into dynamically created optical paths. We report simulation results based on real traffic traces collected from the pan-European GÈANT network.

Keywords: Dynamic Optical Circuit Switching, Optical Traffic Engineering.

1 Introduction

Wavelength division multiplexing (WDM) and erbium-doped fiber amplifiers (EDFA) technologies, 10 years ago, have started a deep revolution in networking. Combined together, they increased the overall bandwidth availability and drastically reduced the cost of data transfer. These technologies allow the transport of hundreds of gigabits of data on a single fiber for distances over 1000km without need of electro-optical conversion. Such dramatic progress has not yet occurred at the routing layer; as a result the network bottleneck has moved from the optical-transport to the routing layer, as conventional electronic routing does not seem capable of offering a cost-effective solution to the increasing bandwidth demand.

During the past 10 years a lot of work has been done on optical packet switching, with the aim of bringing transparent optical operation to the network layer. However, even though many solutions have appeared in different optical laboratories around the world, what is missing is a break-through technology capable of delivering a cost effective implementation suitable for large scale deployment. More and more people in the research community begin to believe that it is unlikely that all-optical switching will reach the market in the near future.

I. Tomkos et al. (Eds.): ONDM 2007, LNCS 4534, pp. 309–318, 2007.

Under these circumstances, the idea of a hybrid electro-optical solution becomes instead more feasible: many optical architectures have already been proposed that implement the concept of dynamic circuit switching.

The basic idea behind dynamic optical circuit switching is to group and switch all the packets sharing a common route into dedicated all-optical channels, bypassing some of the intermediate IP hops. This process can produce consistent cost saving, as expansive router cards can be replaced by much cheaper transparent optical ports. Such savings however are only possible if data is efficiently aggregated into the optical paths.

In this paper we first investigate how existing optical architectures implement dynamic optical circuit switching. Then we introduce the Optical IP Switching concept, presenting a novel method of forwarding IP traffic through dynamically established optical circuits. In section 5 we report the results of our simulations, based on real traces, analyzing the efficiency of OIS to switch data at the optical layer. Finally we conclude the paper.

2 State of the art

Optical bypass of the IP layer is a well-known technique that allows saving router resources and OEO conversions, bringing potential economic advantages to service providers. This technique has been used in the past few years on Sonet/SDH networks where the deployment of optical add-drop multiplexers (OADM) enabled the addition or extraction of selected wavelengths from an incoming WDM bundle.

Add drop multiplexers however could not be dynamically configured, and network operators could not adapt their topology to the actual traffic demand, since any modification or update required a large amount of time and manual interaction.

Recent advances in optical technology brought to market new devices like reconfigurable OADM (ROADM) and MEMs-based photonic switches. Their capability of switching wavelength and fibers on a sub-second scale is ideal for fast network reconfiguration and allows to implement novel bandwidth on-demand services. However reconfiguring a network topology is a critical task, which needs to be supported by a robust control plane. The Internet Engineering Task Force is currently working on this issue, carrying on the standardization of the Generalized Multiprotocol Label Switching, a protocol suite for the optical control plane.

GMPLS provides intra and inter-domain discovery and signaling allowing dynamical, on-demand creation and deletion of data circuits (either in the electrical or optical domain). We remind the reader however that GMPLS is not a traffic engineering tool: it provides the signaling features needed to modify the topology at different layers, but does not include a planning capability that automatically suggests how connections should be updated. Researchers and network operators have so proposed different approaches to dynamic optical circuit switching, using GMPLS as signaling protocol.

In [1] for example a hybrid electro-optical architecture is presented, where electrical routers use GMPLS to create new optical paths, when the existing ones become congested.

NTT has produced a prototype, the Hikari router [2],[3], based on similar concepts: the router uses a photonic switch to create new optical paths, when the existing ones cannot allocate any other MPLS circuit.

Dynamic optical circuit switching seems also to be the key towards the implementation of grid network architectures. The idea of grid networks was developed to support the interaction of high-end applications distributed around the globe that need to exchange information at ultra-high data rates: distributed computing, e-VLBI, High-energy physics, e-Health applications, only to mention a few. Reconfigurable optical networks seem to suit this concept very well, as dedicated optical lightpaths can be established on demand to satisfy the large bandwidth demand of such applications. DRAGON [4] and MUPBED [5] are an example of optical grid networks [6]. The OptIPuter [7], in particular, is capable of creating dedicated end-to-end lightpaths in real-time, either by analyzing IP flows or following direct requests from applications.

The fast reconfiguration offered by ROADM and MEMs switches has also triggered the development of more unconventional approaches: in [8] for example, the authors propose the idea of an ad-hoc optical network, where nodes can connect to use and offer network services, with a plug-and play approach. Auto-discovery, self-configuration and signal monitoring are in this case essential features for the correct operation of the network.

In [9],[10] we have proposed an architecture, called Optical IP Switching (OIS), that creates and deletes optical paths depending on the traffic encountered at the IP layer. One of the distinguishing features is that optical paths are created in a distributed fashion, based only on local decisions. We believe that this approach better satisfies the requirements of inter-domain networking, since different domains can implement their own policies to decide, for example, if an incoming signal should be transparently switched or locally terminated. This is in our opinion more realistic compared to the idea of creating end-to-end optical paths, where each domain is supposed to accept modifications to their logical and physical topology demanded by competing network operators.

3 Prefix-based Optical IP Switching

In the OIS network architecture we have introduced in [9], each IP router constantly analyzes the traffic searching for large, long-lived IP flows. When a suitable aggregate of flows is identified the router activates the photonic switch to create an optical cut-through path between its upstream and downstream neighbours. Only three nodes participate initially to the new optical path, which can then be extended by other nodes in a distributed fashion.

In this paper we focus on a novel approach to Optical IP Switching that aggregates packets based on routing destination prefixes instead of considering distinct IP flows.

The main idea behind prefix-based OIS is to classify the forwarded IP traffic using the network prefixes stored in the IP routing table.

Packets are classified as follows:

1. First we differentiate the packets depending on their arrival (source) and departure (destination) interfaces: this is necessary because photonics switches do not have the capability of grooming traffic optically. The classification is operated building a matrix with number of lines and rows equal respectively to the number of incoming and outgoing interfaces (we assume for simplicity that different interfaces are linked to different destinations[1]). The generic matrix cell (i,j) will identify traffic incoming from interface i, relayed through interface j.
2. Within each of the previous classes we operate a finer classification by destination prefix: in each cell of the matrix we build up a list of destination prefixes reachable through the corresponding interface. This classification is also necessary because the upstream node, source of the new optical path, needs to be informed on which network prefixes it should route into the new lightpath.

In the second classification it is possible to include a "prefix threshold" that will discard all the prefixes bringing data below a certain value. As we will show in section 4, this technique can be used to diminish the amount of information exchanged and save resources at the upstream router.

Each OIS node examines incoming data sampling packets at a rate up to 1/1000, a value that, according to [11], allows a good statistical traffic characterization. For each packet the router checks its destination address and determines the output interface, using the longest prefix matching algorithm. This information is sufficient to classify the packets: the size of the packet payload adds up to the total amount of data carried by its matching prefix. In this implementation we only consider the packet size, but other attributes, for example a timestamp, could be used to improve the traffic characterization.

Each router collects data during an "observation" period before taking a decision. At decision time, the router analyzes the statistics collected, summing up the amount of data brought by the different prefixes within each cell. Only cells whose cumulative data is over a pre-established "path threshold" (100 Mbps in our case) are considered for Optical IP Switching.

The path creation process (Figure 1) begins considering the generic cell (i,j) with the highest data rate: the router signals the upstream and downstream neighbors (using interfaces i and j) checking their capability to support a new optical cut-through path and proposing a suitable wavelength. After both neighbors have acknowledged the request, the router passes upstream the list of prefixes to be switched through the new optical path. Once the path is created the upstream node updates its routing table and starts injecting the suitable packets into the new optical path. The same operation repeats for the remaining matrix cells with traffic above the path threshold.

The advantage of using a prefix-based approach compared to the flow-based one is twofold: first it lowers the amount of information exchanged between the router and its upstream neighbor, as each prefix counts for a large number of flows. Second, it simplifies the routing for the upstream node. While the flow-based approach requires adding class D IP addresses to the routing table (which might create a problem in

[1] Interfaces linked to the same destination would be considered as a unique entry in the matrix.

terms of the size of the routing table), the prefix-based approach only requires a reordering of the IP table. The prefix based approach works as a highly dynamic traffic engineering tool, maximizes the amount of switched data and generally improves the QoS levels. On the other hand however it lacks the granularity of the flow-based method, and cannot be used to guarantee a deterministic QoS for individual flows.

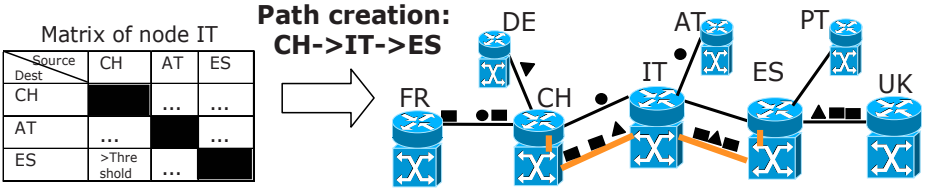

Figure 1 Path creation process

3.1 Path extension mechanism

As previously stated, the decision of creating an optical cut-through path is local, as it only involves a node and its direct neighbors. The path can be set up only if all 3 nodes give a positive acknowledge (this decision could depend on hardware capabilities, inter-domain policies and SLAs between the operators).

A path is created in three different circumstances:

1. The path is newly established. In this case the operation proceeds exactly as described in the previous paragraph.
2. The outgoing interface of the selected cell is the source of an already existing path. In this case (Figure 2) the node creates an upstream extension to an already existing path.
3. The incoming interface of the selected cell is the destination of an already existing optical path. This mechanism is complementary to the one above and creates a downstream path extension.

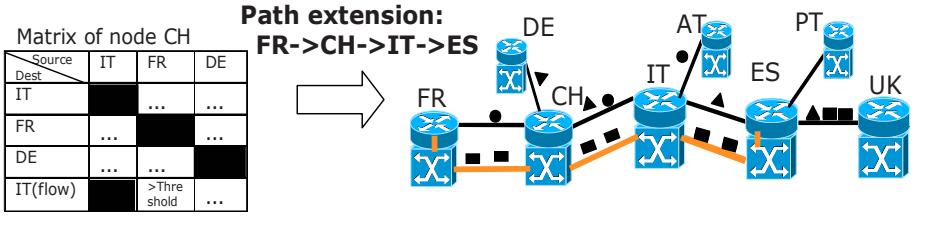

Figure 2 Path extension mechanism

The path extension mechanism presents some differences from the path creation. The purpose of the extension algorithm is to select a subset of the prefixes switched by the original paths. Only this subset will be carried by the new extended path, while the remaining data will be routed through the default links.

The extension algorithm plays an important role in the trade-off between length of the optical path and amount of data carried by the path. An optical cut-through path

can aggregate together only packets sharing a common path. When an existing path is extended, statistically, only a subset of the original packets will share the new longer path, diminishing the amount of data transported by the optical channel and consequently the channel efficiency. On the other hand however longer cut-through paths increase the number of transparent hops, enhancing the cost-saving potentials of optical switching. From this perspective, the extension algorithm has the task to optimize the cost-efficiency problem delineated by this trade-off.

We have identified two algorithms for deciding whether a path should be extended:

- **Relative threshold extension algorithm**. The path extension threshold is expressed as a percentage of the data in the existing path, following the formula:

$$\text{Threshold}(\%) = \frac{\text{PrevNode} - 2}{\text{PrevNode} - 1} \cdot 100 \quad (1)$$

where PrevNode is the number of nodes in the existing path. Considering that packets excluded from the extended path need to be routed electronically over default routes, this approach makes sure that the extension process never increases the amount of data routed at the IP level.

Figure 3 reports an example of a cut-through path extended from 3 to 4 nodes, using a threshold selected according to (1).

- **Absolute threshold extension algorithm.** In this case the path can be extended if the amount of data switched after the extension is above the original path threshold. The original path can also be maintained if the data carried in it after the extension remains over the path threshold. This algorithm maximizes the total amount of switched data.

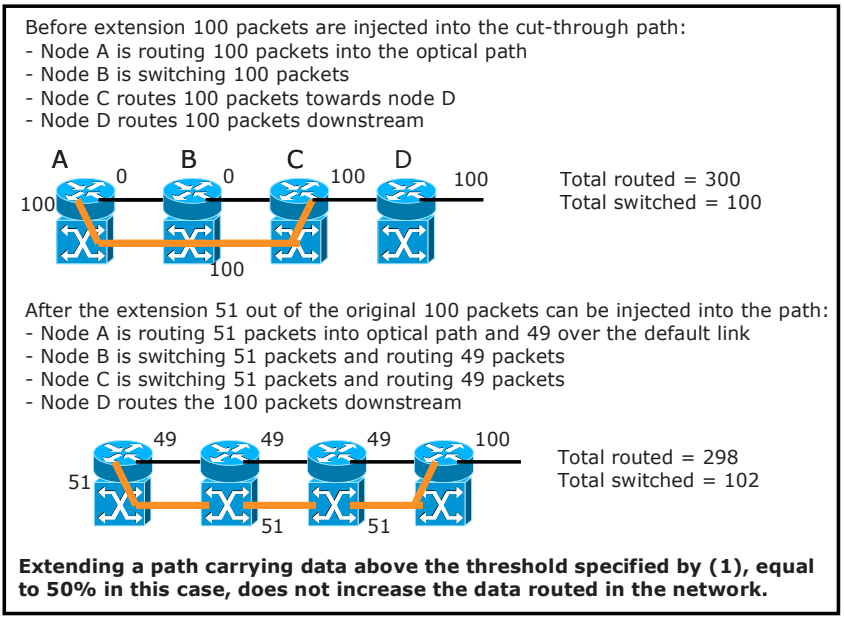

Figure 3 Example of relative threshold extension algorithm.

4 Elephant prefixes

The flow-based optical IP switching method introduced in [9] was developed considering the heavy tailed distribution of the Internet traffic, where a small number of "Elephant" flows carries most of the traffic [12],[13].

We have found a similar heavy tailed distribution in the routing prefixes: in a router's IP table a small number of prefixes routes most of the data. We can use these results to reduce the number of prefixes in the optical paths, with little impact on the amount of switched data.

We conducted our test on a dataset collected from the pan-European GÈANT network, using traces dating back to May 2005[2].

We have studied the network prefixes in the routing tables, classified them by the data rate at which they routed packets, and summed up the amount of data routed by all prefixes in each class. The results are shown in Figure 4[3]: the percentage of data carried by the prefixes routing traffic above a certain data rate (reported in the x axis), diminishes with much slower pace respect to number of prefixes considered. This implies that a large number of prefixes route a minor percentage of the total data. Excluding these prefixes from the optical path will save routing and network resources, without noticeably affecting the switched data.

In Figure 4, for example, we see that a threshold of 100 Kbps, would reduce the number of prefixes considered by 84%, causing only 8% of the data to be excluded from the optical path.

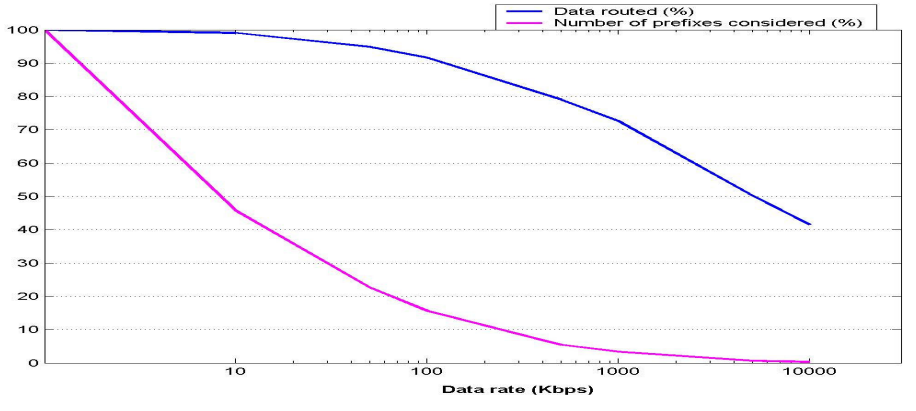

Figure 4 Heavy tail distribution of routing prefixes in GÈANT.

5 Simulation results

We have taken the pan-European GÈANT network as a reference model for our simulations, using empirical data collected from the access points. The GÈANT

[2] More details on the GÈANT dataset will be given in section 5.
[3] From a 15 minutes trace collected on the 4/5/2005 at 15.45.

dataset appeared to be especially suited for our case, as it complements the traffic traces with the BGP routes collected from the border routers. Data are made available by researchers from the Computing Science and Engineering dept. at the University of Louvain-la-Neuve, who also provide C-BGP[14], a network simulator capable of reconstructing the BGP network from the routes included in the dataset. The tool uses a clustering algorithm that reduces the number of BGP entries by more than two orders of magnitude, making it possible to simulate a real network scenario. The traffic traces, collected using Netflow with sampling rate of 1/1000, are summarized depending on their source/destination prefix and only the total number of bytes over a 15 minutes period is provided. This has the two-fold effect of saving storage space for the data files while keeping the traces anonymous. The disadvantage is that information about the precise timing of the flows is lost: a condition that however does not influence our study, since the mechanism that creates optical cut-though paths averages the observed traffic over a period of some minutes.

We have simulated the prefix-based OIS approach considering traffic traces and BGP tables on 4 different days, spaced approximately a week from each other. The dynamic optical links simulate 1 Gbps channels and the path threshold was set to 100 Mbps.

In Figure 5 we report the percentage of data that our OIS approach can switch optically using either of the extension algorithms introduced in paragraph 3. The results are averaged over 4 different traces and reported for different traffic levels, obtained by multiplying the original traces by progressively increasing factors.

The plot shows that the absolute threshold algorithm performs better compared to the relative threshold one. A higher traffic level moreover increases the number of channels above the path threshold, allowing creation of new optical paths. However, once all the possible paths have been created, the ratio between routed and switched traffic stabilizes.

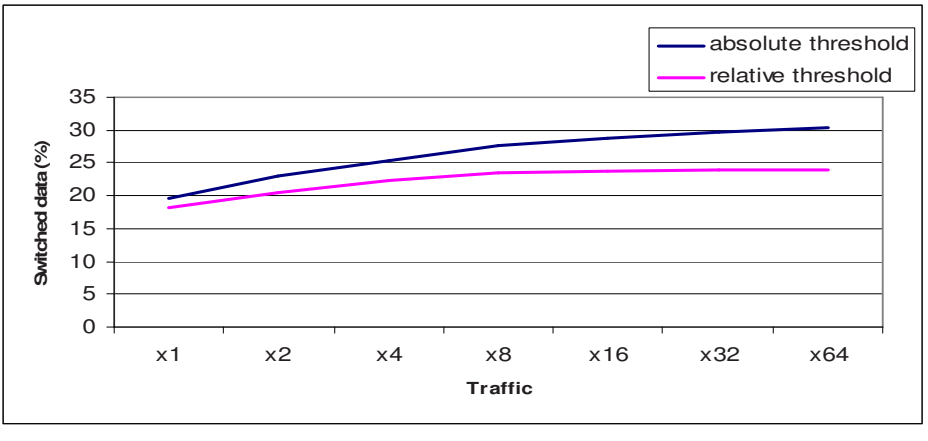

Figure 5 Data optically switched by the OIS architecture

Figure 6 reports the average optical channel usage. Under this perspective the relative threshold algorithm outperforms the absolute threshold one, allowing better exploitation of the optical bandwidth. According to these results we can state that the absolute threshold algorithm maximizes the total amount of switched data while the

relative threshold makes better use of the channels it creates. Which of the two would represent a better approach in a practical implementation depends on the cost associated to electronic routing, optical switching and wavelength channels.

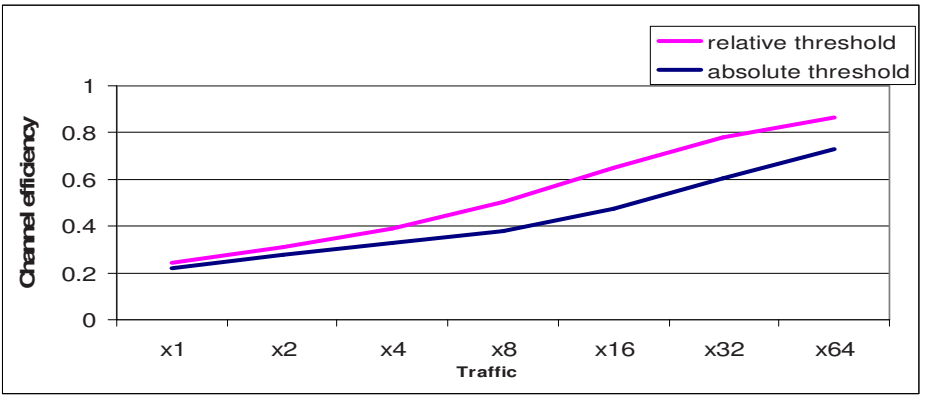

Figure 6 Average channel occupancy comparison.

In order to determine whether optical IP switching is more successful in networks where the node degree is lower, we altered the network topology deleting links to lower the average node degree. We have removed 5 main links from the original GÈANT topology, lowering the average node degree from 3.2 down to 2.8. Figure 7 shows the topology comparison, considering both the total switched traffic and the average channel usage (using the absolute threshold algorithm). The topology with fewer links switches over 20% more traffic, while the channel efficiency improves between 2 and 13%.

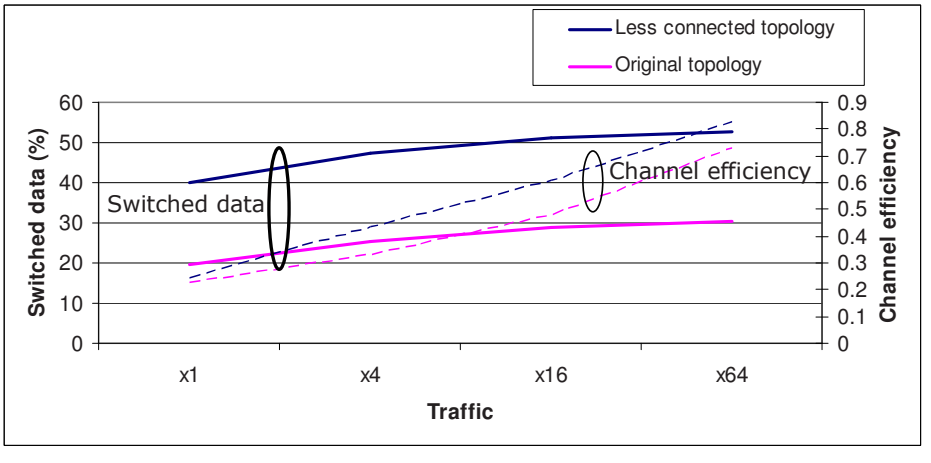

Figure 7 Comparison of different network topologies.

From this analysis we can deduce that the network topology can have a significant impact on the amount of traffic switched by the Optical IP Switching architecture. In particular a less connected topology increases the average number of IP hops, favoring the aggregation of packets at the optical level.

6 Conclusions

In this paper we have presented a novel optical architecture capable of adapting the optical layer topology to the real traffic demand at the IP layer. The traffic is aggregated into dynamically created optical paths using a destination prefix based approach that reduces signaling overhead and saves router resources. Our simulation results, modeled on the pan-European GÈANT network, show that the OIS approach can switch optically about 20% of the total data, with the current network topology unchanged. Taking into consideration that an actual implementation might require higher values to be cost-effective, we have shown that OIS can switch much more data (over 50%) as traffic increases and considering less connected topologies.

Acknowledgments. We would like to thank Steve Uhlig and Bruno Quoitin for providing the GÈANT dataset.

References

1. S. Kano, T. Soumiya, M. Miyabe, A. Chugo. A Study of GMPLS Control Architecture in Photonic IP Networks. Workshop on High Performance Switching and Routing: Merging Optical and IP Technologies, 26-29 May 2002.
2. K. Sato, N. Yamanaka, Y. Takigawa, M. Koga, S. Okamoto, K. Shiomoto, E. Oki, W. Imajuku. GMPLS-Based Photonic Multilayer Router (Hikari Router): Architecture An Overview of Traffic Engineering and Signaling Technology. IEEE Communications Magazine, Vol. 40 , No 3 , March 2002.
3. E. Oki, K. Shiomoto, D. Shimazaki, N. Yamanaka, W. Imajuku, Y. Takigawa. Dynamic Multilayer Routing Schemes in GMPLS-Based IP+Optical Networks. IEEE Communications Magazine, Vol. 43 , No 1 , Jan. 2005.
4. P. Szegedi, Z. Lakatos, J. Spath. Signaling Architectures and Recovery Time Scaling for Grid Applications in IST Project MUPBED. IEEE Communications Magazine, March 2006.
5. T. Lehman, J. Sobieski, B. Jabbari. DRAGON: A Framework for Service Provisioning in Heterogeneous Grid Networks. IEEE Communications Magazine, March 2006.
6. I. W. Habib, Q. Song, Z. Li, N. S. V. Rao. Deployment of the GMPLS Control Plane for Grid Applications in Experimental High-Performance Networks. IEEE Communications Magazine, March 2006.
7. N. Taesombut, F. Uyeda, A. A. Chien, L. Smarr, T. A. DeFanti, P. Ppadopoulos, J. Leigh, M. Ellisman, J. Orcutt. The OptIPuter: High-Performance, QoS-Guaranteed Network Service for Emerging E-Science Applications. IEEE Communications Magazine, May 2006.
8. I. Cerutti, A. Fumagalli, R. Hui, A. Paradisi, M. Tacca. Plug and Play Networking with Optical Nodes. Proceedings of ICTON 06, Nottingham, UK, 18-22 Feb, 2006.
9. M. Ruffini, D. O'Mahony, L. Doyle. A Testbed Demonstrating Optical IP Switching (OIS) in Disaggregated Network Architectures. Proceedings of IEEE Tridentcom 2006,Barcelona, Spain, 1-3 March, 2006.
10. M. Ruffini, D. O'Mahony, L. Doyle. A cost analysis of Optical IP Switching in new generation optical networks. Proceedings of Photonics in Switching 2006. 16-18 October 2006, Herakleion, Greece.
11. T. Mori, M. Uchida, R. Kawahara, J. Pan, S. Goto. Identifying Elephant Flows Through Periodically Sampled Packets. IMC 04, Oct 25-27, 2004. Taormina, Sicily, Italy.
12. N. Brownlee, KC Claffy. Understanding Internet Traffic Streams: Dragonflies and Tortoises. IEEE Communications Magazine, October 2002.
13. K. Papagiannaki, N. Taft, S. Bhattacharya, P. Thiran, K. Salamatian, C. Diot. On the Feasibility of Identifying Elephants in Internet Backbone Traffic. Sprint ATL technical report, Sprint Labs, November 2001.
14. B. Quoitin, S. Uhlig. "Modeling the Routing of an Autonomous System with C-BGP". IEEE Networks Magazine, Nov/Dec. 2005.