Rule-Based Advertisement and Maintenance of Network State Information in Optical-Beared Heterogeneous Networks

János Szigeti, Tibor Cinkler

High-Speed Networks Laboratory
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics **
{szigeti,cinkler}@tmit.bme.hu

Abstract. A more flexible routing of better performance can be achieved in multilayer networks when the controllers of different network layers are "peering" instead of "overlaying", i.e., instead of the simple overlay model the *peer interconnection* or the *vertically integrated* model is used, and Label Switched Paths are searched in a Wavelength Graph that represents the network state very accurately. However, routing in "Peer network" does not only require large databases for storing the Wavelength Graph and complex path-search algorithms, it also increases the load of control channels, as more link state changes must be advertised. Much of the link state information is, however, redundant as the state changes are not independent. In this paper we propose a method for topology information advertisement and maintenance to significantly reduce the amount of control messages without deteriorating the quality of routing.
Key words: path computation, multilayer, vertically integrated, topology advertisement

1 Introduction

The current core & metro communication networks are based on optical transmission and optical switching equipment. They are able to carry a large amount of data, however, optical packet level forwarding of data is still problematic. The new paradigms are OBS (Optical Burst Switching), and in longer term optical label switching and stripping (OLS) and optical packet switching (OPS) but OCS (Optical Circuit Switching) is still viable as in core networks the volume of aggregated traffic between node pairs does not fluctuate too dramatically, and relatively long-term connections must be set up to serve them. The connection provisioning is the task of the *Control Plane* (CP) and of the *Management Plane* (MP).

Recently, two mayor standardization bodies have made efforts towards the standardization of the Control Plane of Optical Networks: ITU-T has defined

** This work has been supported also by the EC within the IST FP6 NoE e-Photon/ONe+ (www.e-photon-one.org) research framework.

I. Tomkos et al. (Eds.): ONDM 2007, LNCS 4534, pp. 279–288, 2007.

ASON [1] while IETF has created the GMPLS framework [2]. The latter one has even split the tasks of the controller (*Optical Crossconnect Controller*, OCC in ASON terminology) by separating MPLS and GMPLS Traffic Engineered LSP calculation from other control functions and delegating it to the *Path Computation Element* (PCE) [3].

These PCEs must cope with the increasing number of devices, fibers, wavelengths, connections, etc. in the network. One way is to expand their storage and calculation capacity by square or higher polynomial as the network grows, while the other, the preferred way is to reduce the information to be processed.

2 Routing in Multilayer Networks

Multilayer networks consist of multiple networking technologies and techniques stacked one over the other, e.g., IP/MPLS/OTN or IP/Ethernet/ngSDH/OTN.

Nowadays, in a multilayer network all the lower layers are statically configured either manually or via the MP, while the uppermost layer is switched via the CP. This is typical for both, IP networks and PSTNs today. However, to reduce the OPEX and to speed up provisioning of new services a CP is needed at lower layers as well.

In such networks the simplest case is when all the layers have their own CP, and the upper (client) layer adapts to the changes of the lower (server) layer. This is referred to as *Overlay Model*.

2.1 Vertically Integrated (Peer) Network Model

If enough information is exchanged between the CPs of the layers and the CPs are interfaced well to each other, then they can take routing decision jointly. This is referred to as *Peer Model* since the layers are equal (peer).

Beside the Overlay and Peer Interconnection models discussed before, there is the so-called vertically integrated model where the network layers are typically run by the same operator, and instead of one CP per layer there is a single integrated unified CP for all the layers.

Now, all the information necessary for routing is spread to all network components that are responsible for making routing decisions. Using the IETF terminology, these components are the PCEs. They maintain a *Traffic Engineering Database* (TED), and based on the TED, they calculate Traffic Engineering *Label Switched Path*s (TE LSPs) and provide them to their clients (*Path Computation Client*, PCC), whenever a traffic demand arrives in the network and the ingress node requests a TE LSP.

The information necessary for routing is the state of the network components. Spreading the information is referred to as *Link State Advertisement* (LSA), that is performed by the *Interior Gateway Protocol* (IGP).

Clearly, in the case of the Peer Model, not only the node connectivity, but all the wavelength information is to be spread unless full wavelength conversion capability (e.g., electronic switch core) is assumed in all nodes (*Wavelength Interchangeable* (WI) network). That is not a typical case.

Practically, a *Wavelength Graph* (WG) is used to represent node connectivity, wavelength usability and conversion/grooming capability. In this case a node is modeled as a directed subgraph [4]. However, flooding all this information of all layers to the PCEs loads the Control Channel significantly.

Although the vertically integrated model offers more detailed topology information to the PCE and there are no interaction limitations between the optical and the electronic (MPLS) layers, there are some problems with the conventional WG which need to be solved.

- *Edge state-dependency.* Switching an edge in the WG affects other edges as well, making them unreachable (see the dotted edges in Fig. 2.a). The background of this behaviour is that the optical crossconnect may switch the λ-channel coming at a given input port to *any* output port at the same λ wavelength. Each switching possibility must be advertised. Having N output ports means N advertised links. Whenever the input port is switched to a specified output port, the other $N-1$ output ports become unreachable from the given input port, and vice versa.
  The goal of the proposed model is to overcome this problem and reduce or eliminate the advertisement of affected edges. Of course, there are also devices which can duplicate, split and merge λ-channels in the optical layer. The model should also support these advanced OXCs, however, primarily we focus on simpler optical devices.
- *Electronic port assignment based on stalled information.* Usually, the OXC has either no or just a limited opto/electronic conversion capability [5], i.e., not all of the wavelength channels can be simultaneously converted into electronic signal and back. The O/E converters are relatively large and expensive, and if the routing algorithms tend to avoid O/E conversions, there is no need for full conversion capability. An OXC with limited O/E/O capability is realized as depicted in Fig. 1.a.

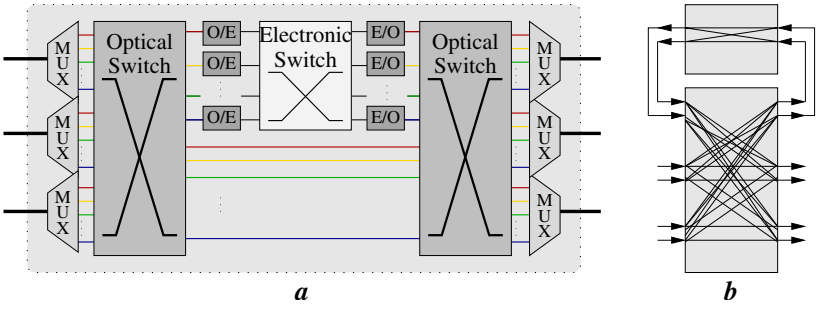

Fig. 1. OXC with limited O/E/O conversion capability: a) physical, b) logical view

The corresponding directed subgraph in the WG is shown in Fig. 1.b.

When we route a connection into the electronic layer, we do not care about which O/E port is assigned to the conversion, as each O/E converter has the same attributes and the traffic in the electronic layer can be re-arranged (groomed). However, the allocation must name a definite O/E port. If the LSA is delayed by any information update strategy [7][8], it may occur, that the next connection – as the decision is based on stalled information – is routed via an already used O/E port (Fig. 2.b) and its SETUP will fail (Fig 2.c) even though there are plenty of other available O/E ports in the OXC. This is the second problem in the conventional WG we want to solve.

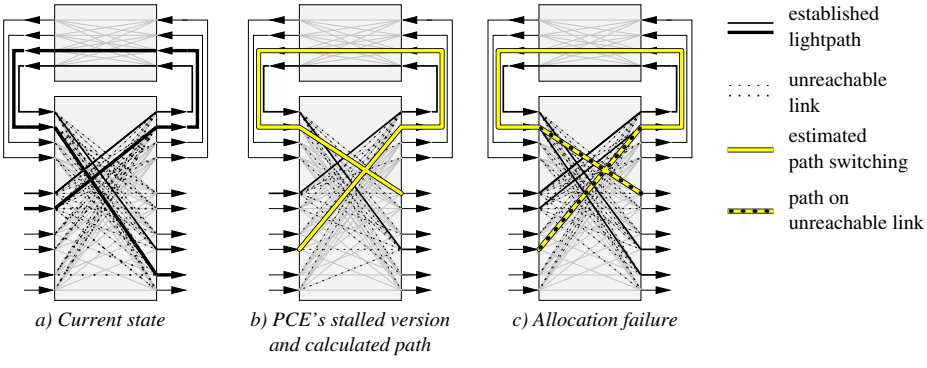

Fig. 2. Failed allocation due to inaccurate link state information

Although efforts were made to reduce the amount of advertised LSA messages while keeping the same level of accuracy in the network view [8], these intra-OXC aggregations are still redundant and do not provide a general solution for any type of OXC.

3 Proposed Solution

Assuming a dynamically configured network, unlike the overlay networks, where established lightpaths appear as direct links between non-adjacent nodes in the upper layer's CP, in the peer model the topology of an optical crossconnect and the corresponding subgraph in the WG does not change due to connection setup/release. The only thing that changes, is the state of the optical switches and, consequently, the attributes of edges in the WG.

Whenever a connection is set up or released, the controllers of the physical devices along the connection path are notified about it by the signaling protocol. These notifications trigger a reconfiguration of switching units inside the OXC and also a *Link State Advertisement* procedure to let the PCEs know about the new state of the device.

Knowing the current state of the OXC and the received signaling message, the new state of the OXC (meaning the new attributes of each link advertised about the OXC) is predictable. If the logic, deciding which advertised edges of the OXC become unusable or available after a SETUP or RELEASE call, is given to the PCE, the whole LSA procedure can be replaced by a simple notification which must be sent to the PCE(s). For compatibility reason, this notification can be encapsulated into a single LSA message.

To maintain an accurate TED in the PCE the following information is needed:

1. Logical topology about the physical devices and the fiber cable connections between them
2. Resource allocation/deallocation information
3. Switching rules to decide link reachability

In this enumeration 1 and 3 is invariant information, determined by the type of the physical device, while 2 is network event-dependent.

3.1 Description of the Messages

Static invariant data, i.e., the topology and the switching rules – as they are coherent – can be advertised together and *initially*. The basic block of the static data describes a port, represented by a node in the WG, and its attributes, that are beside the TE attributes the identifiers of incoming and outgoing links. For both groups of incoming and outgoing links a discrete value n in the range $0 < n \leq N$ – where N is the size of the set – denotes how many links of the group can be used simultaneously (e.g., $n = 1$ for simple crossconnects). In this message specification we exploit the fact that the TE metrics of a link are determined by the ports at its endings. Defining the switching rules that way, the topology of the network is also given, as each link is assigned to an input (head) and an output (the link's tail) group.

The simplest realization of the dynamic data is an LSA message naming the concerned link and the allocated bandwidth. Whenever it becomes 0 the PCE knows that the link is not switched.

3.2 Examples

In the following examples we show how the switching rules of the OXC are described beside its topology. All the figures show OXCs with 3 ports and 2 wavelengths on each port.

Figure 3 shows a simple OXC (OXC#0) on the left, where each port behaves the same way allowing 1 – 1 link selection (as shown under the magnified node-box) in both directions. On the right, an extended device (OXC#1) can be seen with 1 λ-converter per wavelength. Figure 4 depicts OXCs with grooming capability: OXC#2 has unlimited grooming capability (all of the 6 output ports may be switched simultaneously from the input port, as denoted under the magnified input node-box), while OXC#3 has 3 grooming ports, which is resolved by a single "grooming" node allowing 3 input and 3 output ports to be simultaneously switched.

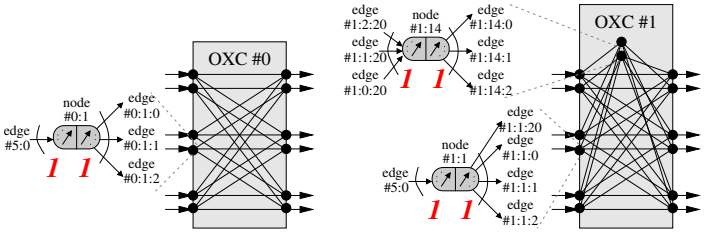

Fig. 3. Topology and switching rules of OXCs without grooming

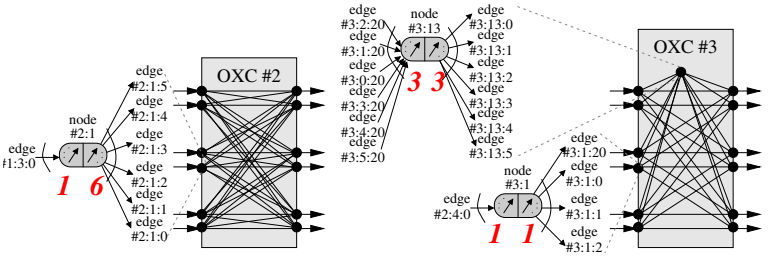

Fig. 4. Topology and switching rules of OXCs with grooming

3.3 Exceptions

In the examples of 3.2 the common attribute of the presented physical devices is that they are *passive* in the sense that their state does not change unless they receive control command from the CP. Their behaviour due to the command is also determined.

There are, however, also *active* physical components, whose state change is not predictable. E.g., the OXC with the so-called *lambda fragmentation/ defragmentation* capability[9] can decide on its own whether existing lightpaths should be converted into electronic signal and back for achieving better network performance (throughput) or not.

Another network components considered as *active* are the foreign domains in an inter-domain environment. For scalability and strategic reasons, domains will refuse advertising their detailed inner topology. Besides, a domain is also active in the sense, that it can initiate either intra- or inter-domain connection, which changes its link states without external influence.

4 Performance analysis

We examine the proposed method and compare it to the conventional method from three aspects:

– amount of required topology information messages,

– data storage capacity and
– path computation complexity.

Assuming standard topology update strategy (LSA sending is not delayed), the required number of messages does not depend on how the PCE stores the data and how it calculates the paths; it is implementation-independent. In turn, both data storage and path calculation are implementation-dependent.

Let us assume an implementation where the TED stores raw topology and TE information and the path computation is based on Dijkstra's algorithm where the unreachable links are forbidden by assigning very large (numerical equivalent of infinite) weight to them.

4.1 LSA Message Density Reduction

The analysis of the required number of LSA messages is split into two: first, we show that the advertisement of invariant data does not load the control network more than the conventional initial topology advertisement, next, we compare the LSA message requirements of the proposed method to the conventional solution. Both analysis are carried out on two types of OXCs:

– sOXC meaning *simple OXC* without electronic ports and
– OXC_K an OXC with K grooming ports.

In the analysis we use the following notation: N is the number of fiber ports, each carrying L wavelengths, K is the number of O/E ports and there are also K E/O ports. W_e, W_n, and W_a are the size of *link label*, *node id*, and *TE attributes*, respectively.

Invariant Data The amount of the invariant data (T^*) can be set against the number of initial LSA messages(T).

In the rule-based model the topology and rule initialization of a simple OXC requires for each input and output optical port a basic block and each block is $W_n + 2 + (1 + N) * W_e + W_a$ units long.

OXC type	basic blocks in T^* (new model)	LSAs in T (old model)
sOXC	$2 * N * L$	$N^2 * L$
OXC_K	$2 * N * L + 1$	$N * (N + 2 * K) * L + K^2$

In case of an OXC with K O/E ports, one additional basic block, denoting the electronic conversion and grooming capability, is required which is $W_n + 2 + 2 * N * W_e + W_a$ long. The blocks corresponding to optical ports are $W_n + 2 + (2 + N) * W_e + W_a$ long.

On the contrary, in the old model, on each λ each input port was connected to each output and electronic input port, and the electronic layer had its all-to-all switching capability. This results in far more initial LSA messages, however, these were much shorter, $2 * W_n + W_l + W_a$, denoting the two endings and the TE attributes of an identified link. Figure 5.a compares the total transmitted data size of T and T^*.

Dynamic Information In this section we investigate the impact of a single event (a single resource allocation or deallocation) and the total number of required LSA messages is the sum of LSA messages triggered by single events (M).

The amount of LSA depends on many environmental factors in the conventional model. The most important factor is whether links become unreachable or reachable due to an event or no change occurs at all. If there is no change, then the scope of the change is limited to the affected links. It is the case, when routing is done over an already configured lightpath. Let the ratio of events concerning a link along an already configured lightpaph be γ.

In other cases, the number of affected links is N in a mere OXC and $N+K$ in an OXC with K grooming ports. However, the more lightpaths are established, the more links became unreachable in advance, and their state does not change through a new lightpath setup. Let us denote the average input/output port usage ratio with η. Now the number of messages is:

OXC type	M^* (new)	M (old model)
sOXC	1	$\gamma * 1 + (1-\gamma) * (1-\eta) * 2 * N$
OXC_K	$1+\delta$	$\delta + (\gamma * 1 + (1-\gamma) * (1-\eta) * 2 * (N + \delta * K * L))$

In the case of sOXC, the number of affected links is 1 if the event aims an already configured lightpath (with γ ratio), and 2-times N (links to input and output ports) otherwise, but η rate of these affected links is already unreachable.

In OXC_K, this number is increased by the amount of links that a grooming event additionally affects.

In the rule-based topology advertisement these numbers are 1 and $\delta * 2 + (1 - \delta) * 1 = 1 + \delta$ since grooming requires 2 link allocations whereas no grooming only 1 (see Fig. 4).

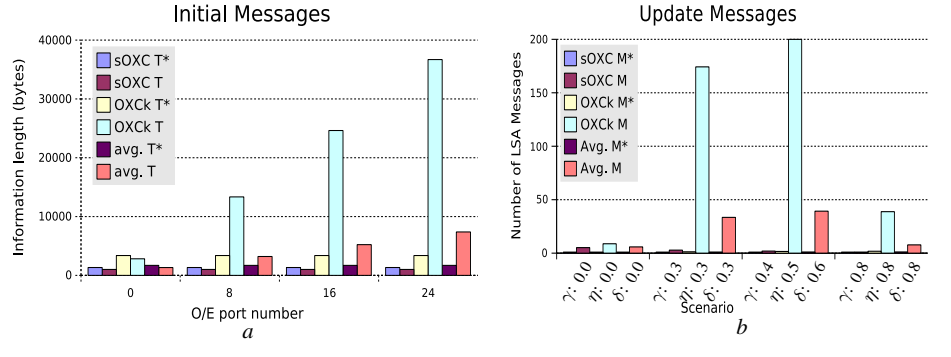

Fig. 5. Numerical results on COST266 Reference network

The amount of invariant and dynamic data is demonstrated in Fig. 5 on the COST266 European Reference Network consisting of 28 nodes, with 24 λs and with grooming capability at the 5 most connected nodes. The figures show 6 columns for each scenario comparing the values in non-grooming nodes, grooming nodes and network-wide average. Fig. 5.a illustrates that the old model was unable to cope with the increasing number of grooming ports, whilst the new model scales well. Fig. 5.b shows 4 scenarios with different network load from empty to near fully loaded (high γ, η, δ values) state. One can see that in our model the number of LSA messages is low and less state-dependent.

4.2 Data Storage Capacity

The data storage requirement is defined by the initial topology advertisement. The dynamic state updates do not influence the storage capacity requirement. The old and the new models require roughly the same amount of storage space:

D^*_{sOXC}	$2 * N * L * (W_n + 2 + (1 + N) * W_e + W_a)$
D_{sOXC}	$N^2 * L * (2 * W_n + W_l + W_a)$

In case of $\mathrm{OXC_K}$ the storage capacity can be calculated in the same way from T^* and T.

4.3 Path Computation Complexity

The path computation algorithm checks each examined link whether it fits into the TE requirement of the connection and whether it is reachable. TE requirements can be checked in both models the same way. The difference is that in the proposed implementation, the reachability information is not stored but calculated, whenever the algorithm requires it. That means at most $n_i + n_o$ additional comparisons in each examined node, depending on the size of the incoming (n_i) and outgoing (n_o) groups.

Generally, it can be stated that the proposed model requires more computational steps. If the original algorithm performed in each node $d * X$ steps proportional to the nodal degree ($d = n_i + n_o$), these additional comparisons mean a *constant-times* computation complexity increment.

4.4 Fault tolerance

Finally, it must be also mentioned, that the presented model is less tolerant against network failures (basically lost LSA messages).

In the conventional model the lost LSA messages do not result in unrecoverable network failures, as most information update strategies advertise link states periodically even if no change occurs, secondly, whenever a controller receives an unfeasible resource allocation request, it knows, that the improper request is due to inaccurate information, and marks the link for advertisement.

Contrarily, in the proposed model the controller can notice the improper request, however, it cannot resolve, which lost event has caused the inaccuracy in the TED of the PCE. General protection against information loss, periodical TED synchronization or other supplementary methods are required which may be the subject of further research.

5 Conclusions

Vertically integrated multilayer networks provide far more flexible traffic engineering solutions than overlay networks do. However, there are also more difficulties we have to face with. In this paper we focused on topology advertisement. We pointed out that, in multilayer optical networks, traditional LSA may be applied, however, it increases the control traffic significantly, and the most link state changes are predictable. We presented our solution to reduce the control traffic (carrying topology information), which is a *rule-based state change advertisement*. After discussing the migration possibilities we have compared the proposed LSA method to the existing one, and showed, that the storage and computation requirements of the new method do not increase much while the LSA traffic is reduced significantly.

References

1. ITU-T *Architecture for the Automatic Switched Optical Network (ASON)*, G.8080/Y.1304, 2001
2. L. Berger *Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description*, RFC 3471, 2003
3. A. Farrel et al. *A Path Computation Element (PCE)-Based Architecture*, RFC 4655, 2006
4. T. Cinkler, D. Marx, C.P. Larsen, D. Fogaras: "Heuristic Algorithms for Joint Configuration of the Optical and Electrical Layer in Multi-Hop Wavelength Routing Networks", IEEE INFOCOM 2000, Tel Aviv, March 2000
5. T. Cinkler: "Traffic and λ Grooming", IEEE Networks, March/April 2003, Vol. 17., No.2, pp. 16-21
6. M. Perényi, J. Breuer, T. Cinkler, Cs. Gáspár: "Grooming Node Placement in Multilayer Networks", ONDM 2005, 9th Conference on Optical Network Design and Modelling, pp. 413-420, Milano, Italy, February 7-9, 2005
7. G. Apostolopoulos, R. Guerin, S. Kamat, S. Tripathi: "Quality of Service Based Routing: A Performance Perspective", ACM SIGCOMM '98, Vancouver, Canada, Aug. 1998
8. J. Szigeti, I. Ballók, T. Cinkler: "Efficiency of Information Update Strategies for Automatically Switched Multi-domain Optical Networks", IEEE ICTON 2005, 7th International Conference on Transparent Optical Network, Barcelona, Catalonia, Spain, July 3-7, 2005
9. T. Cinkler, G. Geleji, M. Asztalos, P. Hegyi, A. Kern, J. Szigeti: "Lambda-path Fragmentation and De-Fragmentation through Dynamic Grooming", IEEE ICTON 2005, 7th International Conference on Transparent Optical Networks, Barcelona, Catalonia, Spain, July 3-7, 2005