
This paper was published in Journal of Optical Networking and is made available as an electronic reprint with the permission of OSA. The paper can be found at the following URL on the OSA website: <http://www.opticsinfobase.org/JON/abstract.cfm?uri=JON-7-10-846>.

Systematic or multiple reproduction or distribution to multiple locations via electronic or other means is prohibited and is subject to penalties under law.

Control Plane Issues in Multilayer Traffic Engineering

Bart Puype,^{*} Didier Colle, Mario Pickavet and Piet Demeester

Department of Information Technology, Ghent University – IBBT – IMEC,

Gaston Crommenlaan 8, bus 201, B-9050 Gent, Belgium

^{}Corresponding author: bart.puype@intec.ugent.be*

Multilayer traffic engineering utilizes functionality in multiple network layers to optimize network performance and resource usage. For IP-over-optical networks in particular, automatic optical switching allows provisioning on demand lightpaths serving as capacity for IP/MPLS links, allowing online reconfiguration of the IP/MPLS layer logical topology. Multilayer traffic engineering builds on the interaction between IP/MPLS routing and logical topology configuration. Technical issues in the IP/MPLS layer limit traffic measurement accuracy as well as routing flexibility and number of allowable topology updates. Through simulation, this paper examines the extent in which both measurement inaccuracies and routing granularity degrade multilayer traffic engineering performance.

OCIS codes: 060.4251, 060.4256, 060.4253, 060.4259.

1. Introduction

Layers in networking provide a physical and logical separation into distinct communication technologies. In current broadband Internet backbone networks, traffic is transported using packets in the electrical domain. Long haul bandwidth however is provided using optical networking technology. Traffic engineering (TE) within an IP network domain has

been made possible by adding label switching functionality to IP nodes, using the MPLS (Multi-Protocol Label Switching) protocol instead of, or in addition to more simple routing protocols such as OSPF.

Originally, single wavelength point-to-point fibers were used to carry traffic between IP routers. The arrival of WDM technology has vastly increased the available capacity on a single fiber, far beyond bandwidths manageable by a single optical linecard. Concatenating fiber wavelength channels into paths through an optical network (lightpath) allows traffic to bypass intermediate nodes in the IP domain. A lightpath establishes a direct connection (forwarding adjacency in GMPLS [1] terminology) between two IP routers. The set of configured forwarding adjacencies is called the IP/MPLS logical topology (Fig. 1). The term IP/MPLS link will be used for these forwarding adjacencies in the remainder of this text,

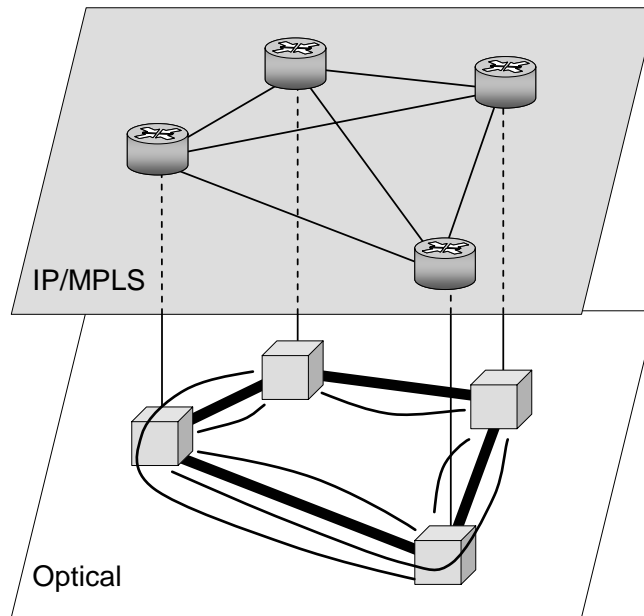


Fig. 1. basic IP/MPLS-over-optical multilayer network with full mesh logical topology

Fast switching optical technology (e.g. based on MEMS, etc.) enables dynamic configuration of such lightpaths, as opposed to the manual configuration of bypasses using for example optical patches. Architectures such as ASON [2] and GMPLS help to standardize this

optical layer flexibility and integrate it into general networking infrastructure, using the tried label switching concept. As such, the optical technology domain becomes a true configurable networking layer, built upon the physical layer.

Multilayer traffic engineering (MLTE), like traditional IP layer traffic engineering, optimizes some performance metric of the network by acting upon the traffic flows offered to the network. Firstly, such traffic flows need to be identified. This can be done by separating the packet IP traffic into flows based on TCP connections, based on hashing algorithms on the IP packet header, based on IP node pair matching, based on MPLS labels (a Label Switched Path), etc. Secondly, the (ML)TE acts on the identified traffic flows. Most often, a specific route is given to packets belonging to a certain flow, however, TE can also consist of traffic shaping by rate limiting (or dropping) some flows. Some TE algorithms will classify traffic into several traffic classes, and perform distinct TE actions on each those classes. This allows establishing quality of service (QoS) for different types of traffic (or customers, applications, etc.), leading to multi-service TE.

In any case, the QoS concept, often just a bandwidth guarantee, is a typical objective of (ML)TE, raising the performance of the network above general best-effort connectivity. Of course other performance metrics exist such as resource usage, total network throughput, traffic path length (i.e. number of hops), that yield conflicting and interacting TE objectives. Multilayer TE is special in this case, in that it can leverage flexibility from the underlying automatic switched optical network layer to help in accommodating traffic flows. For example, it is possible to upgrade the capacity of IP/MPLS links dynamically, by provisioning additional optical connections (lightpaths) parallel to the already established IP/MPLS link optical capacity.

In a more general case, the complete IP/MPLS logical topology (i.e. the set of packet/label switch capable forwarding adjacencies) can be provisioned and even reconfigured online.

As such, MLTE shows some similarities with the logical topology provisioning performed in process called network grooming. The term grooming designates the offline provisioning of a logical topology, a typical example being the provisioning of a SDH/SONET based topology for IP, ATM or other types of traffic. One may in fact perform networking grooming on a regular basis, in order to keep the logical topology suitable to changing (e.g. increasing) traffic demands, in which case the term dynamic grooming is used. MLTE however is an online network process, just like traditional TE. Therefore, some typical grooming approaches such as ILP based algorithms are unsuited for MLTE objectives. MLTE algorithms need to be agile, fast-acting and robust. Additionally, one will often prefer online and distributed mechanisms in MLTE (similar to TE), whereas grooming is almost always performed offline and centralized (e.g. controlled within a network operations center), and requires manual evaluation of the resulting topology before network provisioning or migration is effected.

The main benefit of MLTE in such a point of view consists of the very fast reaction time to changes in traffic patterns. Whereas grooming, as an offline process, is not time-critical, several technical issues will impose a limit on MLTE schemes. For example, while optical switching can be performed on a sub-second timescale, provisioning a lightpath may suffer from signaling incurred by the routing and wavelength assignment process in the optical layer. Worse, routing updates and logical topology changes can take a long time to propagate through the IP/MPLS network, or in a worst-case scenario, cause some routing convergence problems. Some of these issues are caused by the signaling hierarchies in a multilayer network.

Additionally, while operators have gone to great lengths to reduce the amount of layers in a typical backbone network, several distinct switching paradigms remain in current state-of-the-art networks (Fig. 2). This is reflected in the types of labels that can be used in GMPLS. In the remainder of the paper, we will assume an IP/MPLS over optical network. The layer 2 framing used in the network stack will remain out of scope, but can for example be Ethernet. IP/MPLS logical link capacities will have the provisioned optical lightpath bandwidth, i.e. 1 Gbit/s, 10 Gbit/s (corresponding to Ethernet standards), but may also be 2.5 Gbit/s, 40 Gbit/s etc., depending on the exact optical transponder technology.

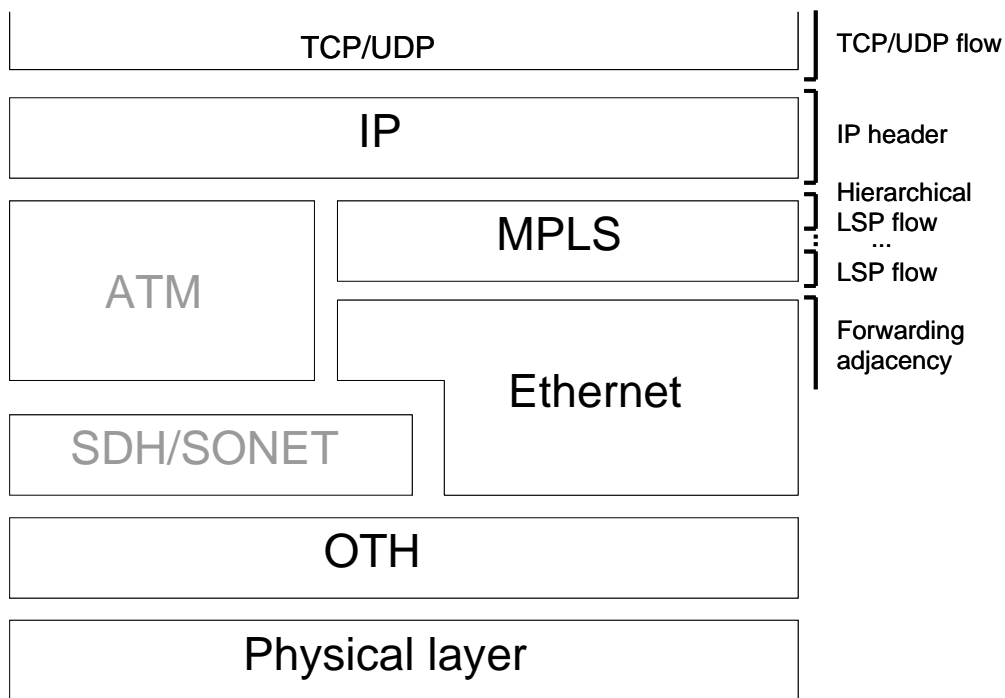


Fig. 2. Network layers

Also, networks can be separated into several network planes, as shown on Fig. 3 for the IP/MPLS network layer. We can identify the data plane and control plane. The data plane carries and processes the actual network payload, i.e. user data. The control plane consists of all control functions such as routing protocols. For the IP/MPLS layer, these consist typically of OSPF [3] which is used to construct shortest path connectivity automatically and of course MPLS which is

used to set up to control LSPs. The routers may run RSVP-TE [4] to manage MPLS label distribution.

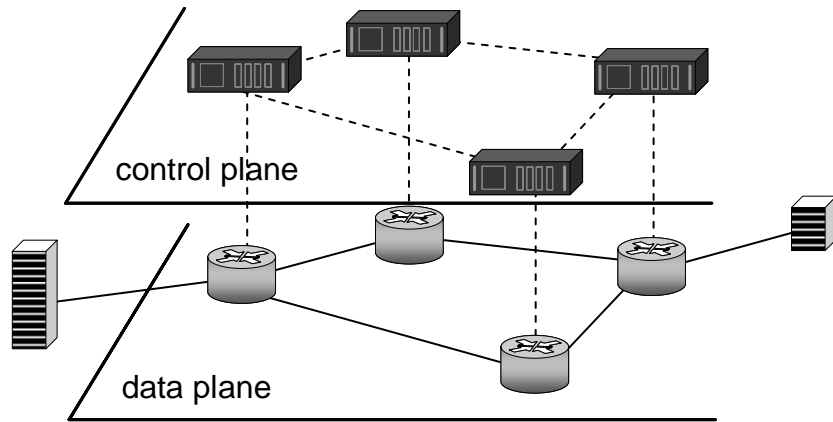


Fig. 3. Network planes for the IP/MPLS layer

The control plane also includes monitoring functionality such as traffic measurement. For other network layers, other types of control plane functions are required. For example, the optical layer has routing (e.g. neighbor discovery) and reservation (RWA, routing and wavelength assignment) protocols. Interlayer functions such as the User-Network Interface (UNI) allow requesting optical connections from the IP/MPLS layer. Each network layer may include a separate control plane, however standardization efforts such as GMPLS propose an integrated control plane taking care of all control functions of the multilayer network. In this paper however, we will concentrate on IP/MPLS layer control plane issues, more specifically on traffic measurement and routing policies.

Firstly, traffic measurement can be performed in multiple ways. We will focus on two methods. A first method uses basic IP/MPLS interface monitoring to establish the actual load on each IP/MPLS link in the logical topology. A second method collects MPLS reservation requests into a database, in order to calculate a theoretical link occupancy value for each link. Of course

this latter method requires MPLS functionality and, more specifically, for all traffic to be carried in LSP tunnels.

Secondly, the term routing policy is used to indicate the type of routing that is incorporated into the MLTE scheme. Each layer in the Fig. 1 has its own specific properties as far as routing and forwarding is concerned. Routing actions may be performed on TCP/UDP flow, LSP or IP/MPLS forwarding adjacency granularity.

This paper will focus on IP/MPLS measurement and routing granularity issues and their impact on MLTE performance. To this end we present in Section 2 a MLTE scheme that can naturally operate in several routing granularity modes. In Section 3 we look at how measurement inaccuracies degrade MLTE throughput. Section 4 describes an evaluation scenario to extract and discuss some routing granularity performance metrics.

2. Multilayer traffic engineering scheme

Multilayer traffic engineering consists of multiple interdependent mechanisms such as routing, online grooming, logical topology construction and capacity adjustment. These mechanisms can be implemented and combined in various ways. For example, in related literature, [5] presents an integrated routing approach across IP/MPLS and WDM layers. Multilayer routing there is performed by considering a single graph model. The mechanism provides both dynamic grooming and shared protection for multilayer networks.

[6] presents the concept of lightpath fragmentation and de-fragmentation, which tries to solve the logical topology construction problem through finding optimal optical-electronic-optical termination nodes for optical connections.

In [7], MLTE is separated in long-term offline logical topology design versus short-term, fast-responding online dynamic routing and capacity adjustment mechanisms.

A straight-forward approach [8] establishes high and low traffic watermarks on links which are used as decision trigger to start logical topology reconfiguration.

This section outlines the proactive MLTE strategy [9-11] and algorithm that will be used later in some performance studies below, paying special attention to how the strategy relies on traffic measurement and how different types of routing techniques and associated routing granularities fit in with this scheme.

MLTE mechanisms

Regardless of the specific MLTE scheme utilized, some basic mechanisms become apparent. Firstly, capacity needs to be allocated in order to carry the offered traffic. Secondly, routes need to be assigned so this traffic has paths over the logical topology. The routing mechanism builds on MPLS or OSPF etc. functionality that is present also in networks with a static IP/MPLS topology. We will classify the allocation of optical capacity into lightpaths that either change the logical topology (i.e. IP/MPLS router adjacencies are modified), or only upgrade existing IP/MPLS logical links (i.e. additional optical bypasses between already adjacent routers).

It should be noted that in terms of network stability and convergence, these mechanisms have very different time constants. Logical link capacity up or downgrades are available immediately, limited only by optical layer switching and connection establishment times. Rerouting traffic in the MPLS (and especially IP) layer should be done carefully to counter instability issues. Logical topology reconfiguration takes a long time to converge, as it interacts with the routes of many traffic flows.

Routing cost function

Since the underlying switched optical layer in an IP/MPLS-over-Optical network is assumed to be fully flexible, the goal is to leverage this property in order to provide capacity only where needed. To this end a worst-case over-provisioning scenario is considered: a logical topology which is a full mesh. Starting from this, we attempt to reduce the logical topology into fewer IP/MPLS links, while still being able to accommodate the offered traffic.

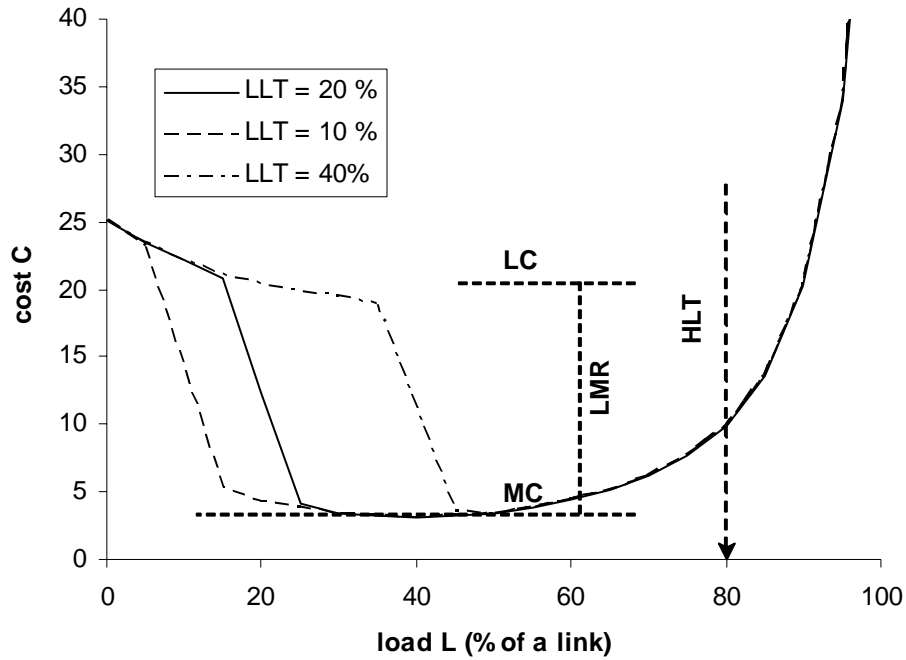


Fig. 4. Load based routing cost function

The logical topology and flow routing part of the presented strategy is combined into a single mechanism by using shortest path routing under a load based routing cost (Fig. 4) [9]. This cost is such that, when flows (regardless of their granularity) are routed over a full mesh, links will remain unused, effectively reducing the mesh into a sparser topology. The routing cost is used so that full mesh links not carrying any traffic flows are not provisioned by the MLTE strategy. This is similar to a traffic grooming approach, but since simple shortest path is used on a flow-by-flow basis, the mechanism can be implemented online and in a distributed way: the set

of traffic flows can be partitioned and assigned to several routing entities (implemented in for example an IP/MPLS router and associated path computation element).

The load based routing cost function has a high cost for high loads in order to prevent IP/MPLS link overload. More importantly, there is also a high cost 'bump' for links with a load below a certain threshold (LLT). This prevents routing over lowly loaded links (unless necessary, of course), which effectively reduces the number of such links. The ratio between cost values for lowly and moderately loaded links will induce multi-hop paths through the networks, grooming traffic onto fewer links, reducing the logical topology size. The ratio in fact sets a limit to the maximum length (in hops) of such groomed paths. The cost for overload should further be high enough so that the overload avoidance mechanism dominates the grooming functionality. Changing LLT causes the grooming to be more or less aggressive, by reducing or increasing the range of acceptable link loads. Full mesh links that are not established in the actual network (i.e. they carry no traffic) are considered to have no load.

It should be noted that the load based rerouting of traffic flows will itself modify the load of some of the links in the logical topology. Even with perfect link state advertisement, this creates a feedback mechanism that can lead to longer network-wide routing convergence times, and possibly instability. To counter this, the algorithm will take into account the load of the currently rerouted flow when generating the routing costs. The link load used in calculating the cost is not the value reported from the routing protocol LSAs, but instead the load a link would have if the flow would be routed over it. Whether we need to add the flow bandwidth to a link load depends on whether the flow currently has that link in its path (the routing entity knows the current route of flows it manages).

In [9], some other cost function shapes were examined for this MLTE strategy, but found to be inferior to the one used in this discussion. In addition to worse performance in terms of throughput or capacity efficiency, those cost functions resulted in problems converging to a proper grooming solution. Therefore, control plane issues would become even more apparent for such cost functions.

Link capacity up/downgrade

Separate from the logical topology construction mechanisms, IP/MPLS link capacity up/downgrade also build on lightpath provisioning flexibility. Link capacity up/downgrade works by concatenating multiple parallel lightpaths into a single IP/MPLS forwarding adjacency. Mechanisms such as virtual concatenation (VCAT) or generic framing procedure (GFP) include similar concepts where multiple layer 2 or lower bandwidth tunnels are combined into a single logical link. It should be noted that higher speed Ethernet (beyond 10 Gbit/s) is also envisioned as multiple bandwidth channels, e.g. [12]

In the previous cost function discussion, it was assumed that link loads are expressed in percentage of a IP/MPLS link. E.g. if no on-demand upgrade mechanism is available then at most one lightpath can be establish between two IP/MPLS routers. 100% link load in such a case corresponds with the full capacity of a lightpath, 2.5 Gbit/s, 10 Gbit/s, etc. Usually, link loads should not be allowed to exceed 80% or so, in order to avoid packet loss caused by router buffering problems or small timescale traffic peaks (load measurements will always entail some averaging). When router connectivity consists of either one or no lightpath, we can see problems arise for very large traffic demands. The logical topology mesh 'explodes' and approaches a full mesh. In most high traffic scenario however, some high traffic nodes or links can still be identified. Additionally, high node degree logical topologies are unrealistic, because this requires

a large number of router interface (line cards), which have to be installed on beforehand of course, and cannot be 'provisioned' on-demand. A sparser mesh is more suitable in this case; this can be implemented by using the link up/downgrade mechanism [11]. The 100% in this case corresponds with the capacity delivered by the maximum number of parallel lightpaths. This typically depends on the technology used. An example would be a logical link concatenation of upto four 2.5 Gbit/s lightpaths; in that case 100% means 10 Gbit/s.

Optical metrics

The routing cost function as such does not take into account the physical layer. This means that choices in logical topology design may result in sub-optimal provisioning in the optical layer. For example, the logical topology may be constructed such that many long lightpaths need to be established unnecessarily. When taking the physical layer (i.e., fiber lengths, topology etc.) into account, the logical topology more closely resembles the physical topology, which yields better optical resource usage.

The network scenario envisioned in designing the MLTE strategy is an overlay, or possibly augmented network. Overlay networks separate network layers into distinct entities, limiting the information exchange between layers. This may be done because of technological reasons (manageability) or confidentiality issues etc. Physical and optical layer information is brought into the MLTE strategy, or rather, the routing cost function, by using an optical metric.

The optical metric consists of a value for each IP/MPLS router pair (similar to the full mesh of link loads), which is reported from the optical layer. For example the metric discussed in [10] is linear with optical layer hop count (of a lightpath connecting the IP/MPLS router pair in question). The optical metric allows reducing optical resource usage significantly. The metric is multiplied with the routing cost function, in order to assign the shortest path routing cost for the

MLTE strategy. Setting optical metric parameters (in this case slope and Y-axis intercept of the linear optical hop count function) pushes optimization more or less towards optical resource usage, this is an operator settable parameter.

3. Traffic measurements

The proactive MLTE strategy basic mechanisms require accurate knowledge of the traffic load in IP/MPLS links (which are supported through optical layer lightpaths). The load is used in a cost function used by a shortest-path algorithm, in order to route IP/MPLS traffic flows as well as construct the IP/MPLS logical topology. In an ideal situation where the impact of (re)routing an IP/MPLS flow on the IP/MPLS link loads is considered to be known immediately, the MLTE strategy can converge very fast; in most cases, we found that the routing and logical topology will have stabilized as soon as each IP/MPLS flow has been routed once. Normally, a full mesh of IP/MPLS flows is considered; representing the traffic matrix over the IP/MPLS nodes.

In a realistic and therefore sub-optimal case where link state updates are non-instantaneous, convergence slows down, and the algorithm may need to cycle through the routing process several times (routing each flow multiple times).

Network load measurements

Looking at how load can be extracted from the network, two main methods can be identified, the ideal and the instant-knowledge case. Firstly, load can be measured directly on the IP/MPLS router interfaces and flooded over the network. Secondly, load can be inferred from monitoring traffic flows at ingress points. The first method is how one would flood IP/MPLS link state information using for example the OSPF protocol.

The second method requires traffic monitoring for each of the IP/MPLS flows (the minimum number of flows being a full mesh between the IP/MPLS routers), but also keeping track of the routes of such flows, in order to derive IP/MPLS link loads. This information is used in the MLTE strategy. In the case of a centralized mechanism providing IP/MPLS routes and logical topology updates, flow routes can be stored, and flow bandwidth requested from the ingress IP/MPLS routers. In a distributed case, these bandwidths and routes will need to be flooded over the network, since, for example, MLTE processing of flows is then performed by the ingress node itself.

Simulation and study setup

In order to evaluate the impact of these control plane limitations, the MLTE simulation environment was altered such that delays in link state advertisement were taken into account.

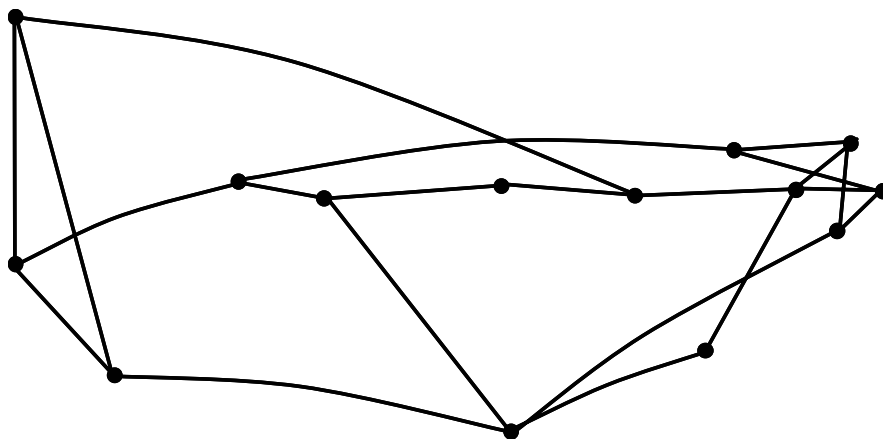


Fig. 5. NSFNET 14 node, 21 span topology

Both the direct interface measurement method and the flow based method were examined. For the simulation run, the 14 node, 21 span NSFNET topology was used (Fig. 5), with uniformly distributed traffic matrices. Volumes of the traffic matrices (versus IP/MPLS link

or lightpath capacity) were such that a range of sparsely through densely meshed logical topologies were setup by the MLTE strategy.

A Java based simulation environment was used which allows applying these traffic patterns on the proposed physical topology. The environment includes discrete event techniques in order to simulate distributed control plane behavior. However, the data plane itself is not modeled on a low level, that is, we did not use packet simulation. Instead, the traffic patterns are abstracted as a set of traffic flows with a source, destination and bandwidth requirement. The MLTE strategy is then simulated by setting up logical topology links and assigning routes to these flows in the simulated IP/MPLS over optical network. The environment runs as a single process (on one PC).

Impact of the logical topology mesh size

The impact of control plane limitations is dependent on the meshedness of the logical topology (or similarly, the volume of traffic demands). A densely meshed network offers more capacity, but also caters to more diverse traffic patterns. A more sparsely meshed logical topology will be highly optimized by the MLTE strategy towards the specific traffic pattern. This means that for sparser topologies, greater (relative) changes occur in the IP/MPLS logical topology, when traffic patterns shift. In such cases, the traffic measurement delays prove to be problematic. Here we have introduced minimal delay into a flow based measurement regime and examined performance impact for various mesh sizes; further on we look at sensitivity to larger delays.

On Fig. 6, the throughput of traffic demands offered to the IP/MPLS layer is shown against offered traffic volume (relative %). As mentioned in [10], mesh size (number of logical topology mesh edges) is almost linear with increasing traffic volume, until a full mesh is

reached. A full mesh in this simulation case requires 91 links to connect 14 nodes. The 100% traffic volume is the case where throughput degradation is no longer seen (a volume corresponding to a topology roughly 2/3 of a full mesh). Throughput is measured against the ideal measurement scenario, where no traffic is lost. Since in a network with multi-hop flow routes, loss in a link or a traffic flow will influence loss in other links and flows, traffic loss was determined by iterative approximation [11, 13].

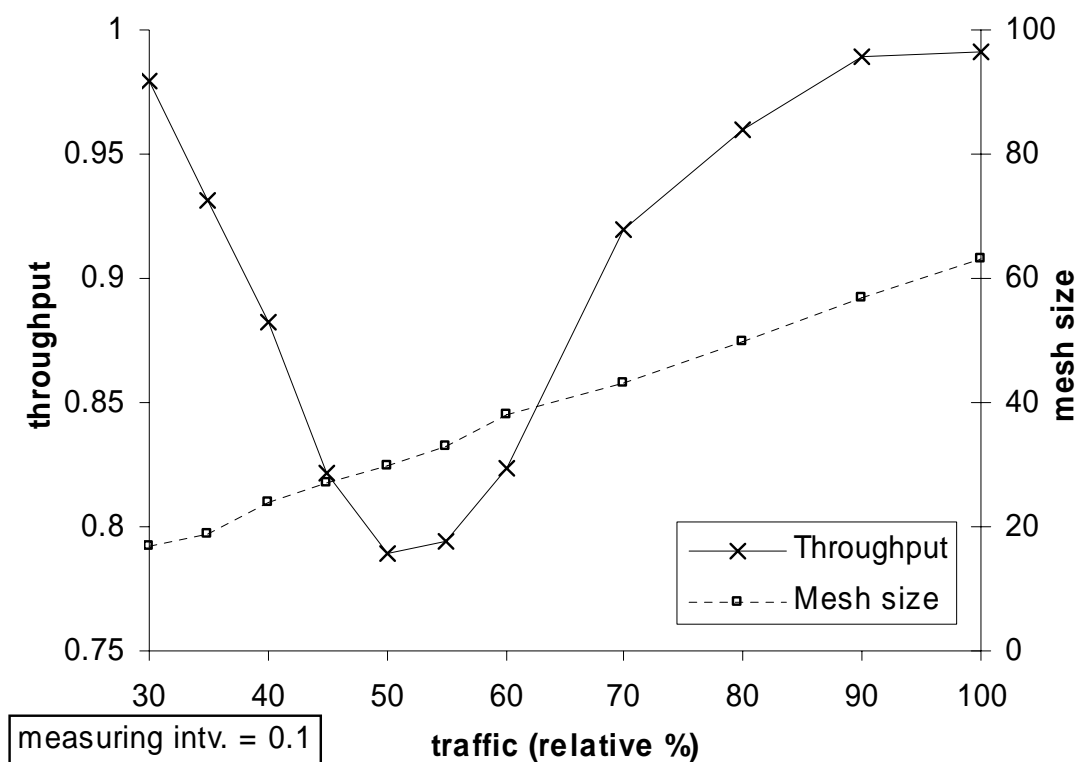


Fig. 6. IP/MPLS flow throughput versus traffic volume

This means that for delayed measurements, some severe throughput degradation is seen around 50% traffic volume, i.e., a topology corresponding to 1/3 of a full mesh, and in the 14 node network case, this is therefore about 30 IP/MPLS (bidirectional) links. For very sparse meshes (approaching a spanning tree, 14 links), which appear for loads < 50% relative, the logical topology becomes such that all traffic patterns can be carried on the mesh, even for sub-

optimal routes; for low volumes, the logical topology becomes more and more over-provisioned. This explains the higher throughput seen for low traffic volumes.

Impact of delayed traffic measurements

In any case, for the regular operating range of MLTE; that is, somewhere between a spanning tree and a full mesh topology, some problems arise when control plane measurement delays are taken into account. These are caused by inaccurate IP/MPLS flow route assignments which lead to non-optimally filled as well as overloaded IP/MPLS links. We examine delay duration impact for the worst case mesh size (i.e. about 1/3 of a full mesh, as suggested by Fig. 4).

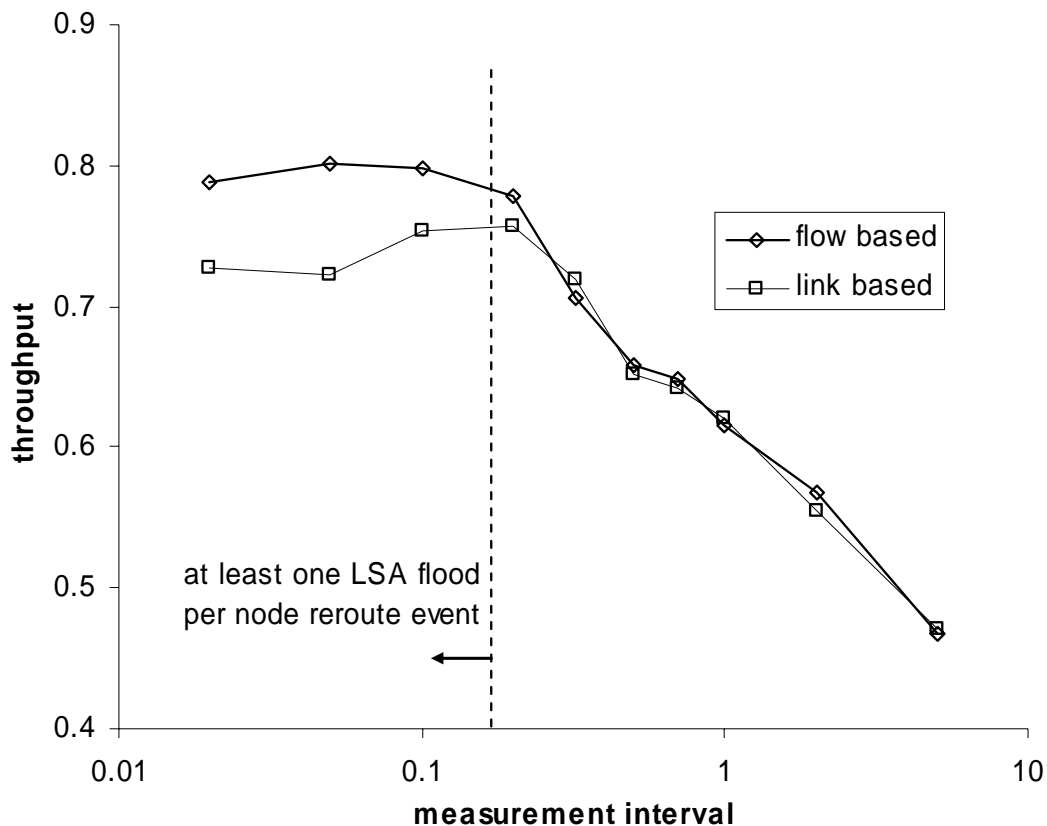


Fig. 7. Influence of measurement rate on throughput

On Fig 7, total traffic demand throughput is plotted against the measurement interval. The measurement interval is the time between two global IP/MPLS network load measurements (i.e. link state flooding). Throughput results for both the flow-based method and the link-based (direct interface measurement) method. An interval duration of 10 corresponds with the life-time of one traffic pattern; after that, a new traffic pattern is offered and the MLTE strategy will reconverge flow routes and logical topology. The IP/MPLS routing and logical topology construction process requires periodically rerouting the flows. This process is run distributed, this means that each node by itself will periodically reroute flows it has been assigned. Flows have been assigned uniformly, and node reroute events are also spread out uniformly in time. In the simulation study, the time between two node reroute events is about 0.17 (all 14 nodes are processed 4 times during the 10 second duration).

On the graph, one notices that results for both measuring methods are similar for measurement interval beyond these 0.17 time units. Throughput also degrades quickly beyond this point. This suggests that at least one network-wide measurement event needs to be performed for each node reroute event. This allows for accurate information to be available at the time flows are rerouted.

However, even for measurement intervals shorter than 0.17 time units, throughput still does not reach 100%. For these interval ranges, a discrepancy is noticed for both measurement methods, with the flow-based method enabling better throughput figures. This is as expected, since the flow-based method delivers accurate information based on flow bandwidth aggregation, which is what the MLTE strategy was originally designed for. The link-based method take into account traffic loss through the network, and therefore the measurements itself are inaccurate from a traffic engineering standpoint. For example, the flow-based method may aggregate to link

loads $> 100\%$, while actual link measurement will always return values $< 100\%$, which underestimates network overload events, and degrades the MLTE strategy's performance.

One also sees that throughput drops again for very short measurement intervals. Presumably, this is because in those cases some instability effects will arise between flow rerouting events within a single node. Also, some 'aliasing effects' arise between measurement rate and rerouting rate.

Concerning scalability, larger networks will have more routing nodes. Therefore, to converge traffic within a similar duration of absolute time, more routing events need to occur per time unit. This will pose stronger constraints on measuring rates. Additionally, more traffic flows are processed during each routing event (number of flows processes being proportional to network node count as well), this will further impact sensitivity of the MLTE algorithm to inaccurate measurements.

4. Routing granularity

The term routing granularity is used to indicate a characteristic of the routing policy that is incorporated into the MLTE scheme. Each layer has its own specific properties as far as routing and forwarding are concerned. Single TCP or UDP flows may in fact be assigned a unique route. This is done in applications using overlay networks (overlaid over the IP network) that employ their own hop-by-hop forwarding on top of the regular Internet. As mentioned before, IP header information may be used in classic TE; for example, in order to establish load-balancing flows over multiple parallel paths through an IP network. As these techniques acting on layer 3 and up are more the domain of single layer TE or user level optimization, these will not be considered.

In the layer stack, between layer 3 (IP, packet) and layer 2 (switching and framing), there is a so called layer 2.5 that has traditionally been implemented using ATM. The IETF MPLS scheme offers many of the same ATM concepts in a transparent manner. The ATM VCI and VPI concepts can be mimicked using label switched paths (LSPs). In fact there is a possibility to encapsulate hierarchical LSPs themselves into LSPs. This offers a wider range of flow granularities when using MPLS based routing policies. In any case, MPLS allows decoupling forwarding and routing, similar to ATM. Each LSP flow (i.e. with a unique separate label) can be assigned its own route or path through the network. Label Switching Routers are usually integrated with IP routing capability, into a single electrical domain IP/MPLS node that is both packet and label switch capable.

In the IP/MPLS logical topology (which in fact implements a dynamic layer 2), each logical link (or forwarding adjacency) can be assigned a coarse granularity routing weight [14] (or cost), such as done in e.g. OSPF. This allows some rudimentary TE possibilities at link granularity.

Routing granularity in MLTE

The MLTE strategy will be evaluated by offering traffic arrivals to be routed over the network. Each traffic arrival is represented as a traffic flow. Depending on the grouping of the set of flows, MLTE routing will have a finer or coarser granularity. The finest granularity is achieved by considering each flow individually.

When the number of flows is high, this requires a lot of processing. To alleviate this, we can group all flows between the same IP/MPLS router pair into a single node pair master flow. For example, we can encapsulate all LSP flows between a node pair inside a master LSP (i.e., a higher hierarchy LSP tunneling the lower hierarchy LSP flows). In that case, we only need to manage at most a full mesh of IP/MPLS master LSPs.

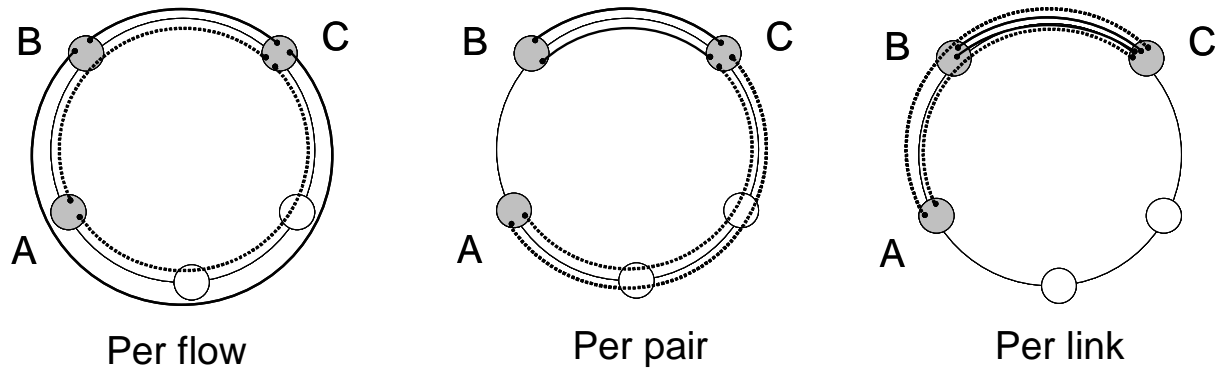


Fig. 8. Example flow routes for several routing granularities

Lastly, a routing policy coupling forwarding and routing delivers very coarse routing granularity by using traditional link weights (cf. traditional TE). In this case, there is no control over individual flows.

On Fig. 8, flow routes for these three routing policies and associated granularities are shown for an example. In the per flow case (fine granularity MPLS policy), we have individual control over each flow. For example, we can choose to route the two A-C flows in different directions over the logical topology ring, in an attempt to load balance traffic (similar for the two B-C flows). For the per pair case (coarser granularity MPLS), all traffic flows between a node pair take the same path. However, we can still load balance traffic: B-C traffic is routed in the opposite direction of A-C traffic. In the per link case (OSPF style policy) finally, only link weights can be set. If we assume all link weights are equal (shortest path routing), we get the routing as shown on the figure. Segment B-C can possibly become overloaded as all traffic takes the shortest path. One may try adjusting the weights on the ring segments to better spread traffic, for example, one could assign a high weight to segment A-B (which would cause flow A-C to be routed counter-clockwise instead). However, weight setting may interact with the path of other traffic flows, so this is non-trivial. In any case, routing granularity is clearly coarse in this case.

Evaluation scenario

We will evaluate the presented MLTE strategy for the three routing granularities. The per flow and per pair modes can be implemented simply by having the MLTE algorithm route either every individual traffic arrival, or an aggregated master traffic flow per node pair.

The per link mode, which is based on link weights, can be implemented by directly using the cost function as, e.g., OSPF routing weights in the logical topology. Note that since the cost function depends on link load (and therefore on the routes of the traffic), we have to first converge those flow routes in an offline process (i.e., without actually modifying routes on the actual network). The resulting link weights can then be assigned to the actual IP routing tables in the network simulation (no MPLS functionality will be used in the per link case).

The 14 node, 21 edge NSFNET network was again used for the simulation study, as in chapter 3. As this section of the paper discusses interaction between IP/MPLS and optical layer and is not limited to the IP/MPLS logical topology construction as in chapter 3, we will include an optical metric in the MLTE strategy. This metric is the generic optical hop count based linear metric with slope 1 (that is, node pair cost equals number of optical hops connecting the pair). We refer to [10] for more information on optical metrics and their influence on optical layer resource usage.

Traffic

Traffic arrivals were generated with exponentially distributed inter-arrival and service times. To vary traffic demand, the mean inter-arrival time was modulated according to a day-night rhythm. In this case, a simple sinusoidal variation was used, with the highest demand during peak hours (e.g. 8 pm). Parameters were chosen such that low demand periods (mornings) have a total demand one tenth of the peak load.

Service times were not modulated, but chosen so that the required number of traffic arrivals was generated for a simulation (taking into account the bandwidth of a traffic arrival). For the peak traffic demand, traffic between two nodes was chosen 30% of the lightpath capacity (i.e., 3 Gbit/s for 10 Gbit/s lightpaths). This choice operates the peak traffic demand very close to saturated logical topology links, for the per flow mode.

For the aggregate routing modes that reroute periodically, we chose a 15 minute interval between routing updates. Traffic service (or holding) times were in the order of a few minutes. We will examine performance of the three routing granularity modes, looking at IP layer and optical layer impact.

Impact on IP layer performance

Firstly on Fig. 9 the number of overloaded (load higher than 80%, as per the cost function) IP/MPLS links is shown vs. offered traffic demand (scaled where 1.0 is the peak load). Results show averages extracted from a sufficient number of logical topologies generated by the MLTE strategy, for each traffic volume. Note that even for similar volumes, actual traffic matrices may be quite different. Comparing per flow and per pair modes, one notices a penalty of about 5 additional overloaded links that is fairly independent of offered load. This aggregation penalty is caused by the less efficient TE as flows between two nodes always need to follow the same route in the per pair mode.

Next, when comparing per pair and per link, the overload penalty appears to grow with increasing traffic. This is a de-grooming penalty as traffic flows can no longer be routed individually over the network. This causes traffic to group together over a few heavily loaded IP/MPLS links, whereas the other TE modes have the option to spread traffic over the network, increasing logical topology meshing. This problem is especially apparent at high loads.

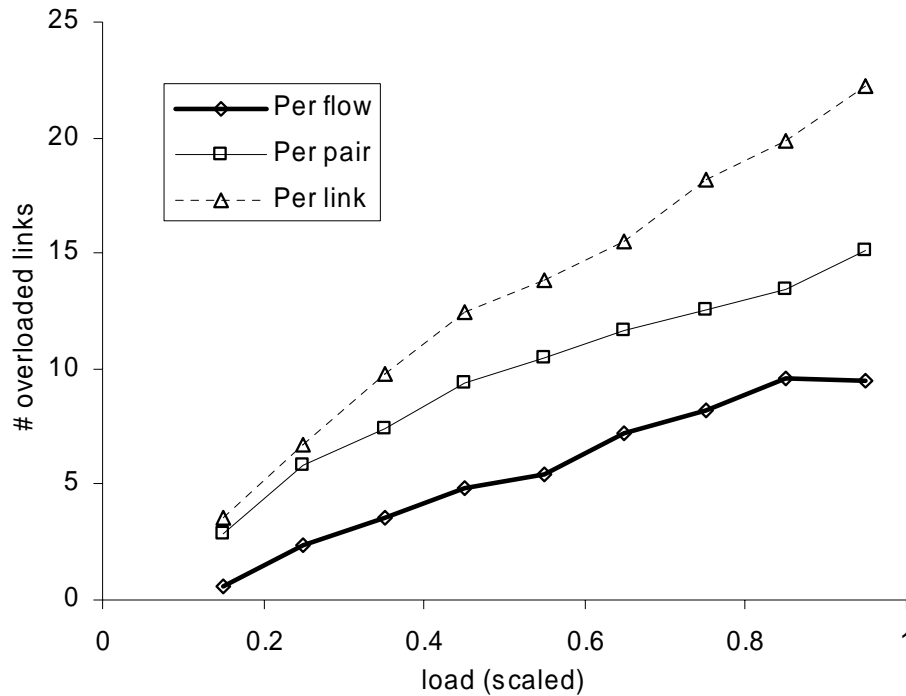


Fig. 9. Number of overloaded links versus load

Impact on optical resource usage

Next, the average number of required lightpaths (versus load) is shown on Fig. 10, again for the three MLTE granularity modes. We have opted to enable the link capacity up/downgrade mechanism in this case, with the assumption that IP/MPLS link overload can be solved by establishing an additional overload lightpath parallel to the overloaded IP/MPLS link.

Note the fact that the per link scheme requires less lightpaths than the per pair mode. This is not too surprising, given the fact that the per link mode concentrates overload in a few high capacity backbone IP/MPLS links (which will have very high overload). Overload lightpath filling will be much more efficient consequently, explaining the apparently low number of required lightpaths.

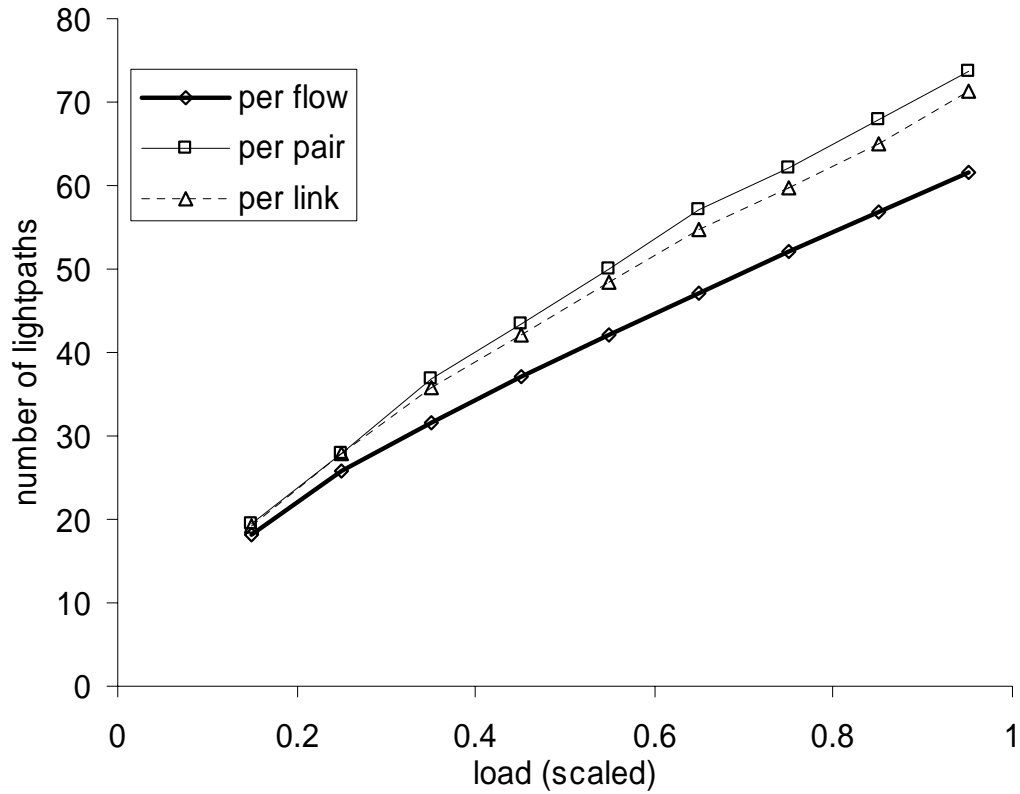


Fig. 10. Number of lightpaths

Then we break down the required lightpaths in regular and overload lightpaths. We show total amount of required optical capacity for a simulation run of one week (168 hours) on Fig. 11. Results are shown as lightpath hours; one lightpath hour corresponds with one lightpath worth of capacity being required during one hour.

For the per flow scheme, less than 1 in 1000 lightpath hours was setup to solve overload (not visible on the figure). We have chosen traffic demands to remain in the non-overloaded case for the per flow mode, as said before. Some heavily loaded (over 80%) IP/MPLS links were experienced, but very few IP/MPLS exceed 100% load (saturation).

The other two MLTE modes have a significant amount of overload capacity that needs to be established. Implementing a link capacity up/downgrade scheme can probably only be

rationalized when up/downgrade has a sufficiently large range in capacities to set up. Using up/downgrade only to solve overload (which comes with inefficient link filling) will not be very reasonable. Consequentially, one should not expect overload capacity up/downgrade to be available in realistic networks, and in fact all traffic in such overload lightpaths to be lost.

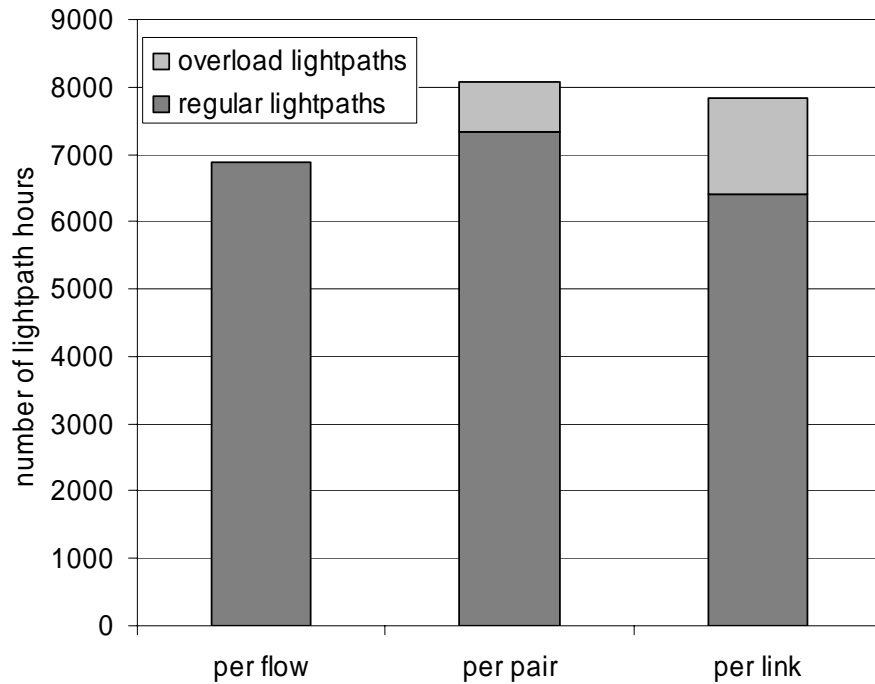


Fig. 11. Number of lightpath hours

Conversely, link capacity up/downgrade is more typically suitable in multilayer networks with a lower multiplexing factor, i.e. lower lightpath capacities vs. typical IP/MPLS bandwidths. For optical networks with lower lightpath capacities, even typical network operation will require setting up parallel lightpaths to carry all IP/MPLS traffic between two nodes. The availability of link capacity up/downgrade offers finer granularity control over IP/MPLS link bandwidth, and more efficient link filling. A drawback consists of the technological requirements for such a scheme (e.g. flexible IP/MPLS router interfaces and optical transponders).

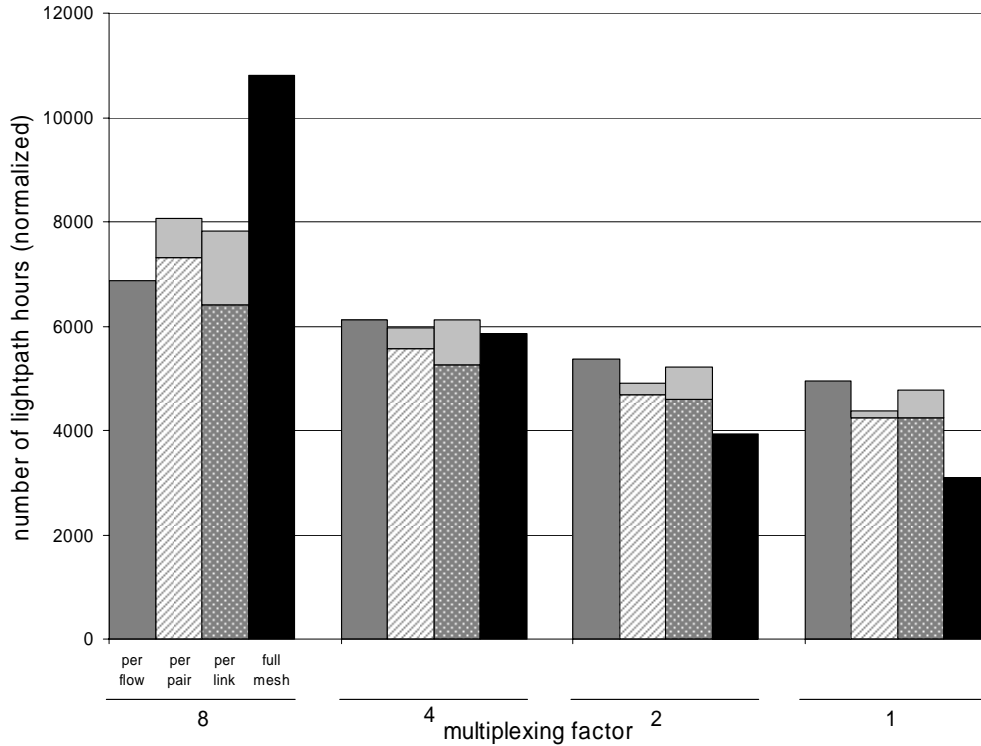


Fig. 12. Influence of multiplexing factor

On Fig. 12 we show similar results (again regular and overload lightpaths), but for several IP/MPLS - optical multiplexing factors (MF) 8, 4, 2 and 1. The multiplexing factor indicates capacity offered by lower layer bandwidth pipes in terms of upper layer traffic volume. For example, we could assume lightpath capacities of 40 Gb/s (MF 8), 20 Gb/s (MF 4), 10 Gb/s (MF 2), 5 Gb/s (MF 1) – in that case we assign multiplexing factors relative to the 5 Gb/s lightpath capacity case. The MF 8 case on Fig. 12 corresponds with results discussed up to now, meaning that for the results in the figure, we have examined the influence on MLTE performance of utilizing lightpaths that offer respectively only one half, a quarter or an eighth of the base capacity. Obviously lower MF scenarios will require more lightpaths to carry the same traffic volume.

Lightpath hours were normalized towards MF 8, meaning that for example two 20 Gb/s lightpath hours are normalized as one 40 Gb/s lightpath hour, allowing to compare results for different multiplexing factors. In all cases, the cost function optimizes towards one MF 8 lightpath (100% load equals one MF 8 lightpath capacity).

If again average peak node pair traffic is 30% of a MF 8 lightpath, then the MF 8 case generally allows grooming of at least two node pair flows on the same lightpath. For MF 4, peak node pair traffic will fill up most of a single lightpath. For MF 2 and MF 1, even off-peak traffic demands will need parallel lightpaths in the IP/MPLS network. In other words, link filling will need to be optimized through grooming for the MF 8 case, whereas for lower MF this is less critical because of finer lightpath capacity granularity.

We have also shown lightpath hours for a fourth multilayer provisioning mode where all traffic flows are routed over direct IP/MPLS links, so a full mesh of lightpaths is established. Full mesh routing is especially inefficient at high multiplexing factors, but quite good at low MF, at least as far as optical resource usage is concerned. Since it requires a high number of router interfaces in every node in order to create the many forwarding adjacencies, the full mesh case is very rarely a viable option.

Looking at the three MLTE modes, we see that lowering multiplexing factor has a positive impact for all three modes in terms of optical layer resource usage. However, finer lightpath capacity granularity has a much larger improvement for the per pair and per link cases – or rather the benefit of per flow routing is less apparent for this case. These modes drop below the per flow mode in terms of total required lightpath hours once it is no longer necessary to groom multiple flows together in order to fill up a lightpath. Grooming (and specifically fine granularity grooming in the per flow case) solves suboptimal link filling and provides load

balancing using the IP/MPLS cost function. Lower MF however intrinsically solve link filling inefficiency; also, load balancing, while helping with IP/MPLS throughput, increases load in the optical layer of the network, as it routes traffic over longer paths. Since lower capacity lightpaths lead to better link filling, the need for fine granularity (per flow) load balancing or rather the goal of creating some uniform load in all IP/MPLS links as suggested by the cost function diminishes, in fact it only hurts optical layer resource usage.

Interestingly, it seems that the ratio of overload vs. total lightpath hours diminishes with lower MF. This is again because overload lightpath typically have bad lightpath filling efficiency. Wasted capacity can be reduced then by using lower capacity lightpaths. Note however that still it may not be possible to create overload lightpaths because of router interface or optical transponder limitations. Again the overload traffic may be lost, despite the good optical performance of the per pair and per link modes for low MF.

5. Conclusions

Control plane issues for MLTE were examined in two separate aspects: traffic measurement and path assignment.

For realistic, non-ideal traffic measurement, measurement delays due to the distributed nature of network link load extraction were found to be an issue. They degrade total demand throughput for typical operating modes of MLTE, i.e., moderately dense logical topology meshes. Measurement methods using direct per link interface monitoring were found to perform worse than methods that aggregate per flow bandwidth information extracted from IP/MPLS ingress nodes in order to derive logical topology link loads.

For the path assignment aspect of the control plane, several routing policies exist in considering traffic flows offered to a multilayer network. Flows can be considered individually,

aggregated per IP/MPLS node pair, or one can rely solely on OSPF style link weight setting, leading to distinct granularities in routing assignment. The suitability of MLTE modes with specific routing granularities depends on the multiplexing factor. For networks with coarse granularity (i.e., high capacity) lightpaths, fine granularity routing granularities solve the inefficient link filling optimally through multi-hop traffic grooming. For networks with finer granularity (low capacity) lightpaths, link filling optimization and the fine granularity load-balancing properties of the grooming part of the MLTE strategy become less of an issue. Instead, such networks lean more on the link capacity up/downgrade functionality in addition to grooming.

Acknowledgements

This work was partially funded by the European Commission through the projects IST-NOBEL II under contract FP6-027305 and the Network of Excellence BONE; by the Flemish Government through the FWO project 3G057808. Bart Puype thanks the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) for its financial support through his PhD grant.

References

1. E. Mannie (editor), "RFC3945: Generalized Multi-protocol Label Switching (GMPLS) Architecture," Network Working Group (2004)
2. ITU-T G.8080/Y.1304 (2001), Architecture for the automatically switched optical network (ASON)

3. J. Moy, "RFC2328: OSPF Version 2," Network Working Group (1998)
4. D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, "RFC3209: RSVP-TE: Extensions to RSVP for LSP Tunnels," Network Working Group (2001)
5. T. Cinkler, Cs. Gáspár, "Fairness Issues of Routing with Grooming and Shared Protection," in *Proceedings of 8th IFIP Working Conference on Optical Network Design and Modelling (ONDM 2004)*, Ghent, Belgium, pp. 665–668
6. T. Cinkler, G. Geleji, M. Asztalos, P. Hegyi, A. Kern, J. Szigeti, " λ -path fragmentation and de-fragmentation through dynamic grooming," in *Proceedings of 7th International Conference on Transparent Optical Networks (ICTON 2005)*, Barcelona, Spain, Vol. 2, pp 1–4
7. P. Iovanna, R. Sabella, M. Settembre, "A Traffic Engineering System for Multilayer Networks Based on the GMPLS Paradigm," *IEEE Network* 17(2) 28–37 (2003)
8. B. Gillani, R. Kent, A. Aggarwal, "Topology reconfiguration mechanism for traffic engineering in WDM optical network," in *Proceedings of 19th Symposium on High Performance Computing Systems and Applications (2005)*, Guelph, Ontario Canada, pp. 161–167
9. B. Puype, Q. Yan, D. Colle, S. De Maesschalck, I. Lievens, M. Pickavet, P. Demeester, "Multi-layer Traffic Engineering in Data-centric Optical Networks, Illustration of concepts and benefits," in *Proceedings of COST266/IST OPTIMIST workshop – 7th IFIP Working Conference on Optical Network Design and Modelling (ONDM 2003)*, Budapest, Hungary, pp. 211–226
10. B. Puype, Q. Yan, S. De Maesschalck, D. Colle, K. Steenhaut, M. Pickavet, A. Nowé, P. Demeester, "Optical cost metrics in Multi-layer Traffic Engineering for IP-over-Optical

- Networks,” in *Proceedings of 6th International Conference on Transparent Optical Networks* (ICTON 2004), Wroclaw, Poland, Vol. 1, pp. 75–80
11. B. Puype, D. Colle, M. Pickavet, P. Demeester, “Influence of Multilayer Traffic Engineering Timing Parameters on Network Performance,” in *Proceedings of IEEE International Conference on Communications* (ICC 2006), Istanbul, Turkey, Vol. 6, pp. 2805–2810
 12. H. Toyoda, S. Nishimura, M. Okuno, K. Fukuda, K. Nakahara, H. Nishi, “100-Gb/s Physical-Layer Architecture for Next-Generation Ethernet,” *IEICE Tran. on Communications*, Vol. E89B(3) 696-703 (2006)
 13. E. Van Breusegem, B. Puype, D. Colle, J. Cheyns, M. Pickavet, P. Demeester, “Evaluation of ORION in predimensioned networks,” in *Proceedings of 19th International Teletraffic Congress* (ITC19 2005), Vol. 6b, Beijing, China, pp. 1265-1274
 14. B. Fortz, M. Thorup, “Optimizing OSPF/IS--IS weights in a changing world,” *IEEE J. on Sel. Areas in Comm.* 20(4) 756–767 (2002)