Inter-Domain Traffic Engineering for Balanced Network Load

Levente HUSZÁR, Csaba SIMON, Markosz MALIOSZ

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar tudósok krt. 2., 1117 Budapest, Hungary, e-mail: {huszar | simon | maliosz}@tmit.bme.hu

Manuscript received October 01, 2010; revised October 30, 2010.

Abstract: Inter-domain cooperation at control plane level offers the possibility to balance the traffic in a much more flexible way compared to the single domain Traffic Engineering (TE) methods. This solution is especially effective if the offered traffic and the network load permits to avoid congestion without recomputing the routes. Current networking technologies and the emerging trends in aggregation-access-core network architectures provide means to realize such a cooperative control plane. We consider a dual opto-electronic network model where traffic grooming is also possible. In the particular case of the layer 2 switched aggregation network, which feeds a core network using a GMPLS control plane, this solution means redistributing the traffic between the spanning trees of the aggregation domain. We propose an inter-domain control plane cooperation and investigate it by means of simulations on several core network topologies deploying IP over WDM technology. In the simulations we analyze the effects of the traffic redistribution rate on the throughput, on the number of lightpaths and on the number of opto-electronic conversions. We show that if congestion occurs in the core, we can eliminate the congestion just with a proper coordination between the control planes of the aggregation and core domains, redistributing the traffic prior entering the core.

Keywords: Traffic engineering, knowledge plane, GMPLS, MSTP, CSPF.

1. Introduction

Emergent technologies like GMPLS, WDM and carrier-grade Ethernet will replace legacy ones in future Internet domains. Combining of these different data plane technologies and different services at different layers into an efficient interworking environment is a challenging task. The resulting system should offer a trade-off for service providers to operate their networks.
A common trend in communication networks is the widespread of fiber technologies. The optical networks are favored by the operators because they offer higher network capacities. At the same time they also come with more advanced control and management features. However, these, combined with the new services demanded by the market, increase the complexity of the networks and advanced network optimization mechanisms become a must. The most common optimization mechanisms deal with the traffic, several Traffic Engineering (TE) proposals are known in the literature. Typically such TE solutions deal with a single domain, only [1], [2], [3]. Alternatively a joint optimization of traffic over a cascade of core networks has been proposed [4], [5], but they consider the same TE algorithm all over the domains. Nevertheless, the traffic originated by the end users reaching the core networks should cross the aggregation and access domains, which use different TE mechanisms. This paper investigates the effect of cooperation at control plane level between the different TE mechanisms of the core and access or aggregation domains. The advantage of this cooperation is that it allows much more flexibility in maintaining balanced core network usage. Additionally, other optimization goals can be satisfied, if we can redistribute the traffic among the edge nodes connecting the access and the core. In our paper we investigate the impact of network load optimization on the efficiency of a dual opto-electronic network model [6], [7].

In the followings, we will investigate the networking technologies covered by the paper and summarize the trends in aggregation-access-core network architectures. Then we will present our solution on the joint access-core network optimization and we investigate it by means of simulations. Finally we conclude our paper.

2. Networking technologies

2.1. Data and control plane technologies

Most of the applications deployed over the info-communication networks are based on IP, while the access networks are predominantly Ethernet-based. From the access network the Ethernet traffic is concentrated at the edge nodes and is forwarded to packet-based transport in the core domain. The trends evolve towards the wide deployment of WDM (Wavelength Division Multiplexing) network devices in the core, which enables the transmission over the established connection-oriented lightpaths in the optical domain.

These lightpaths form a virtual topology over the physical topology that can be reconfigured dynamically in response to traffic changes and/or network
Inter-Domain Traffic Engineering for Balanced Network Load

The combination of IP directly with WDM results in an efficient assignment of optical network resources to forward IP traffic [8], [9].

The versatile Generalized MultiProtocol Label Switching (GMPLS) protocol enables the integrated control of both IP and WDM. Integrated routing and wavelength assignment based on GMPLS is currently the most promising technology, which also delivers effective TE capabilities. The typical routing protocol in such networks is the Constrained Shortest Path First (CSPF) [10], [11].

B. Dual opto-electronic model

In the above model the lower layer is an ‘optical’ one, the upper layer is typically an ‘electronic’ one, capable of performing joint time and space switching. Paths of the lower layer correspond to a single link in the upper layer. Lightpaths are special routes: they arise and terminate in the electronic layer. The upper, electrical layer can perform multiplexing of different traffic streams into a single wavelength path (λ-path) or lightpath via simultaneous time and space switching. Similarly it can demultiplex different traffic streams of a single lightpath. Furthermore, it can perform re-multiplexing as well: some of the de-multiplexed traffic will be again multiplexed into some other wavelength paths and handled together along this other path. This is often referred to as traffic grooming and we will refer to it as grooming [7].

The question is how these layers can be operated together. Both IETF and ITU-T propose models and solutions how to operate these two (or more) layers together [21], [22]. We consider the case when both layers are handled via a distributed control plane to ensure full and joint on-line adaptivity of both of the layers. By using dynamic optical layer, it is possible to create an adaptive set of lightpaths that satisfies emerging traffic demands. Those two physical nodes that are connected by a lightpath are seen as adjacent by the upper layer. Multiplexing and demultiplexing the traffic of a lightpath is impossible by applying only optical devices. In these cases lightpaths have to be torn down, their traffic has to be taken up to the electronic layer that increases the number of lightpaths. In addition, the number of applicable opto-electronic converters per node is limited.

C. Traffic Engineering mechanisms

The traffic in the network domains is engineered in the control plane. The major goal of such a Traffic Engineering (TE) mechanism is the enhancement of the performance of an operational network, at both the traffic and resource levels. Traffic engineering optimizes the use of network resources to achieve specific goals, such as to avoid congestion, to minimize delay, to differentiate
services, etc. Most of these methods are valid on intra-domain level, because internal network information is typically limited outside of administrative domains [17]. Consider the domain as a closed entity, with a given traffic matrix. In such solutions the TE affects only the output, but it does not take into consideration the possibility to influence the input. [2], [3], [18], [19]. Even if it does, it considers that all domains – the one that provides the traffic, and the one that conveys the traffic – use the same TE method (e.g. Path Computation Element based TE - SPF) [4], [5].

3. Network Models

A. Reference network model

Our proposal is based on the presumption that both the aggregation and access domains are, or in the near future are expected to be, based on switched layer 2 (L2) technologies [12], which offer lower bit costs [13]. L2 switched networks deploy Spanning Tree Protocol variants (STP, MSTP, etc.) [14] to convey the traffic towards the core.

Fig. 1 gives a picture of the reference network model developed within the CELTIC TIGER2 project [15]. The reference network [16] reflects the view of major service providers and vendors on the evolution of networking infrastructure and the way it will assimilate the new technologies. As seen in Fig. 1, the networks are divided in three segments: the Access, the Metro and the Core. Depending of the country and/or local geographic specificities as well as the Internet Service Provider (ISP) choices, part of the sub-segments depicted in Fig. 1 may be missing, but based on the current practices and medium-term forecasts, this generic model describes all networks.

![Figure 1: TIGER2 generic network reference model.](image)

The Access network is a local area network, and is widely studied in the literature. It connects the end users to the first Central Office (CO). Typically they have a tree-based topology, which aggregates the traffic to the COs. Core
networks are also well studied and in this model we define it as the national or wider area domain. Typically they have a meshed topology. As seen in Fig. 1, the metro network, which links the access to the core, is split into three subsegments. In legacy infrastructures, these sub-segments together form a hierarchy. Access areas may be connected to any metro sub-segment by COs, and each metro is connected to the core through a Point-of-Presence (PoP).

The roles of the sub-segments should be specified in the context of the deployed technologies. In this paper we assume that carrier-grade Ethernet-based L2 technologies become dominant not only in the aggregation, but also in the access [12]. Based on the above assumptions we obtain the reference network used in this paper, derived from the generic network model [16]. In this specific model the metro sub-segments use L2 switching in the access and edge, while the core deploys L2/L3 TE mechanisms. Thus, the first two sub-segments of the metro represent successive aggregation levels of the user traffic. In the metro-access, the first aggregation level, the traffic from multiple COs is aggregated in Concentration Nodes (CN). In the metro-edge, the second level of aggregation, traffic from different CNs is processed by a L3/L2 metro node, and the PoPs at L2/L3 boundary are handling several tens of thousands users. As a summary, we consider that the metro-access and metro-edge networks form an aggregation domain, while the metro-core segment is a meshed distribution.

B. Core network models

A specific core network model has been proposed [16], starting from current ring topologies, widely deployed in optical networks. The model is a Double Rings with Dual Attachments (DRDA) and it can be used in core networks. In such topologies two rings, (the inner and the outer metropolitan rings) are interconnected in such a way, that every node in the outer ring is directly connected with its associated node in the inner ring, via double links (dual attachment). These provide high connectivity and multiple back-up paths for restoration purposes while reusing current network fiber deployments.

C. Investigated network topologies

Based on the reference networks presented in the previous section we designed a network that was used for our simulation based investigations, and its topology is presented in Fig. 2 (left). This network is divided in two main parts, an aggregation network using Multiple Spanning Tree Protocol (MSTP) [14] and a core part with Constraint based Shortest Path First (CSPF) TE [10].

The traffic sources are depicted on the left-most part of the figure, the aggregation network conveys the packets to the core network. At the boundary we have only three edges. In real-life networks the number of edges is kept as
low as possible for reasons of costs. The network has six destination nodes (sinks) represented by the exit points of the core network on the right side. The main function of the aggregation domain is to channel end-user traffic towards the core, thus its nodes are connected to two neighboring devices, at most.

The core domain has a meshed topology, with a 3 hop shortest distance between the ingress and the egress. The nodes of the core have a degree of connectivity of 3 or 4. This is a trade-off between cost effectiveness and the assurance of alternative paths. The aggregation domain uses Ethernet-switched technology, and the core uses WDM extended with an electronic control layer.

Apart from investigating the efficient network capacity usage and balanced load of the core domain, we also investigated the possibility to minimize the operations in the electronic layer and the usage of longer optical paths. These last two parameters are characteristics of the dual opto-electronic models.

![Figure 2: Topologies of aggregation and meshed core (left) and dual ring core (right).](image)

Our proposal supposes that the domains have a control plane that apart of running TE and other control functions are capable of communicating/cooperating with the control planes of the neighbouring domains. Such a control plane model is the Knowledge Plane [23] that can use MSTP in the aggregation and CSPF in the core domains.

We also investigated the behavior of the core if it deploys a dual ring topology (see Fig. 2 - right). We have kept the edge nodes and the output nodes from the previous topology in order to use the same aggregation network and to be able to compare the two results. In the following we refer to the first topology as meshed core, while to the latter one as dual ring core.

### 4. Inter-domain TE cooperation

Our proposal is to use shared intelligence between control planes, where the core intra-network functions are unchanged and only the inter-network control planes co-operate which enhances the performance.
In Fig. 2 the traffic reaches the core network through the aggregation domain. In case of any event (congestion on a link, link failure, etc.) the classical TE works with the assumption that the traffic matrix remains unchanged and it has to re-distribute the traffic volume relying on load redistribution inside the core. Our proposal is to use the Knowledge Plane and re-arrange the input traffic distribution outside the core edge routers. This means that – from the point of view of the core – we change the traffic matrix, since the load on the edges will be different.

Let us take the topology presented in the left-side of Fig. 2. Now in the situation when the aggregation domain directs all the traffic to the e1_edge (the “northern” one), while e2_edge (the “middle” edge) and e3_edge (the “southern” edge) do not feed any traffic to the core. This is the worst case situation to overload the core and corresponds to the situation when only the tree rooted in e1_edge is used to collect the traffic in the aggregation domain. Now, if we take the opposite situation, when we use each of the trees in the aggregation domain to forward the same amount of traffic, then the aggregation domain distributes the traffic evenly among the three ingresses. In this case all regions of the core will be evenly loaded.

It is the task of the Knowledge Plane to map the traffic sources among the trees. In our simulations we used small individual flow throughputs. Each tree is collecting such individual demands and the sum of these represents the traffic load at the edges. Practically the granularity of the traffic is small enough to allow us to finely balance the load. In what follows we will use the term load balancing as the operation of load redistribution in the aggregation domain as described above. The goal of load balancing will be to decongest a certain area of the core network with a minimal redistribution of the original load.

5. Simulation results

A. Traffic model

During the simulations the traffic flows originated from the sources have the same bandwidth. We considered that we know the traffic matrix and the paths in the core are computed by a PCE using CSPF protocol. Additionally we generated background traffic, as well, which enter the core at the edge nodes and sink on the most right-hand side destination nodes. The links of the core networks had 200 Mbps capacity, which defines the load region where the core network is congested, but not overloaded of 400 Mbps to 800 Mbps.

In our investigations we used the e2_edge node where we directed all the traffic and tried to serve it using CSPF. The resulting paths were called the main branch. If the demand is high enough, the traffic demand cannot be served. If
we apply our solution to this situation that means that some part of the traffic will be shifted to the other two edges, \( e_1 \_edge \) and \( e_3 \_edge \). The paths that follow the flows entering on these two edges are called secondary branches.

We used the background traffic to “fill” the network up to the point where congestion might start to develop. We sent 200 Mbps background traffic on the main branch. Then we started to add new traffic demands until we reached the total one, which was set differently from case to case: all our simulations were run with the 500Mbps, 600Mbps, 700Mbps and 800Mbps total traffic demand. These are the situations when we can test the usefulness of our proposal and evaluate its impact on the efficiency of the opto-electronic core transport.

We used a flow level simulator, already used for the research of opto-electronic networks [24]. We generated individual flows, and the sum of these demands resulted in the overall traffic demand. Each link was divided into lightpaths of 10 Mbps capacity. This results in 20 lightpaths within each link that offers enough flexibility for multiplexing the flows within the core. Based on earlier work with the simulator we opted for 12 individual flows per lightpath, resulting in a flow capacity of 0.83 Mbps.

Within each scenario – that is for different overall traffic demands – we have simulated several sub-cases, where the load of the main branch was gradually re-distributed among the secondary branches. At first we started with the situation where 30% of the traffic was entering at edges \( e_1 \_edge \) and \( e_3 \_edge \) (15% on each of them). From there on, we stepwise directed more and more traffic towards to the secondary branches while the network was able to carry the traffic without loss. In order to be sure on that, we also simulated the next step following this point. The individual flow demands were scheduled randomly. For each situation we run ten simulations and averaged the results.

B. Spanning Trees in the aggregation domain

In order to get the input traffic at the ingresses of the core domain, we had to build the trees that bring the traffic to the edges of the core. For this, we had to build the set of trees that form the basis of the MSTP operation. We used a combination of the TOTEM [25] and BridgeSim [26] tools to simulate these trees.

With the combination of these two tools we could use the topologies created in TOTEM and apply the STP formation protocol implemented in BridgeSim. We generated all the spanning trees that can potentially be used in our scenarios, that is, all the trees that are rooted in one of the three edges and the traffic sources are their leaves. Fig. 3 presents the tree generated for the northern edge, the other two trees have a similar shape. The actual traffic distribution among the three trees is decided according to our solution and should be enforced by the Knowledge Plane, as mentioned earlier.
We did not investigate the behavior (delay, blocking, packet loss, etc.) of the aggregation domain, only used it to generate the MSTPs and determine the input traffic for the core domain. In what follows we will be interested only about the traffic distribution among the three ingress nodes.

**C. Balancing the load in the core**

![Figure 4](image)

*Figure 4*: The successfully served flow demands as the function of traffic re-distribution for the meshed core (left) and the dual ring core (right).

The left-hand side of Fig. 4 presents the ratio of successfully served traffic demands in the meshed core. It can be seen that 500 Mbps total traffic will be served if we redirect 35% of the traffic on the secondary branches. The traffic volumes that must be redirected to achieve a loss-free ratio for the 600Mbps, 700Mbps and 800Mbps traffic scenarios are 50%, 60% and 70%, respectively. These results confirm that if we redistribute the traffic before it hits the core edges, we can balance the core load, thus it is a viable mechanism to actively increase the efficiency of the core traffic engineering process.

We achieved similar results for the dual ring topology, as well (right-hand side of Fig. 4). The congestion-free core is achieved for the redistribution of the 35% of traffic for the 500Mbps case and 70% for the 800Mbps (worst) case.

**D. Dual opto-electronic model**

In the following, we present our simulation results on the effect of our proposal on the efficiency of the opto-electrical dual transport core network.
First we explain the results using the meshed core. The first parameter is the number of lightpaths. We can see in Fig. 5 (on the left) that as the rate of successfully served traffic demands is rising, but is still below 100%, the number of lightpaths is increasing. This is due to the fact that more and more individual flows are in the network and these are following new (alternative) routes. Thus, the increase of this parameter is not a consequence of the decreasing efficiency but of the growth of the core utilization.

This trend is reversing if we keep redistributing the traffic even after all the traffic reaches its destination. This corresponds to the situation depicted in Fig. 4 by the dots on the 100% line. As we already mentioned, for each overall traffic load scenario we simulated two cases when all the demands were served by the core: the “break-even” point and one following step where we increased the traffic redistribution by 10%. The results obtained for these cases are encircled in Fig. 5 and as we can see the number of paths is decreasing. The more the core is loaded, the more these trends are accentuated, therefore it can be seen the best result on the curve corresponding to the highest loads.

If the primary goal is the minimization of the operations in the electrical layer, the best option is to distribute the traffic. The drawback of this solution is that we have to use the alternative branches. In regular operation the alternative paths are associated with higher costs. Therefore it is the decision of the operator depending on its business model to find the balance between the path costs (that usually are translated into financial costs) and the efficiency of the opto-electronic layer (higher-layer delays are translated into worse QoS).

The right-hand side of Fig. 5 presents the number of opto-electronic conversions done in the core (values that yielded 100% success rate are highlighted with a circle). We have plotted in the same graph the results for both the meshed core and dual ring core. These conversions can happen only in...
the nodes that make a grooming operation and multiple conversions may happen in such a node. Based on our simulations there are several hundreds of such conversions per node. The trend observed for the number of the lightpaths is valid also here, for both core topologies.

6. Conclusion

This paper proposed a traffic management solution that improves the performance of the core network. The aggregation network is supposed to deploy L2 switched technologies, while the core network will use WDM in combination with GMPLS. We have prepared a meshed and a dual ring topology following the principles of the TIGER2 project’s reference network and investigated our proposal by means of simulations.

We have shown that if congestion occurs in the core, we can eliminate the congestion just with a proper coordination between the control planes of the aggregation and core domains, redistributing the traffic prior entering the core. We deployed MSTP protocol in the aggregation domain and the traffic redistribution was done using these spanning trees. This solution increases the ratio of successfully served traffic demands, increasing the utilization of the core. The traffic redistribution at the aggregation has positive effects even if there is no congestion in the network, because in such cases it increases the efficiency of the opto-electronic transport layer.

As a conclusion we can say that the cooperation of the control layer of the aggregation and core domains has multiple advantages. In the future we plan to investigate the trade-off between the cost of load balancing and opto-electronic efficiency.

Acknowledgements

This work has been partially funded in the framework of the CELTIC TIGER2 project (CP5-024) as part of the EUREKA cluster program.

References


