# Experimental Evaluation of a PCE Transport SDN Controller for Dynamic Grooming in Packet over Flexi-Grid Optical Networks

Ricardo Martínez[1], Ramon Casellas[1], Ricard Vilalta[1] and Raül Muñoz[1]

[1] Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), ricardo.martinez@cttc.es

**Abstract** *We validate the implementation of a unified PCE-based Transport SDN controller for Multi-Layer Networks (packet over optical flexi-grid). A novel on-line MLN routing algorithm targeting grooming strategies is experimentally evaluated under dynamic and heterogeneous data-rate packet traffic.*

## Introduction

Cloud services are drastically increasing transport traffic needs. Thus, network operators are seeking for cost-effective solutions to increase transport capacity, aggregating heterogeneous packet data rate services (10, 40, 100 Gb/s) within coarse optical infrastructures in an automatic and agile way. In this multi-layer network (MLN), a centralized controller coordinates multiple layers to offer on-demand, flexible bandwidth services. This enables traffic grooming[1,2] (electrical and optical) to improve network resource utilization (optical spectrum, packet client ports, SBVTs). Herein, the MLN is formed by MPLS switches connected to a flexi-grid optical network (BV-OXCs and SBVTs[1]).

Electrical grooming allows transporting multiple packet flows into existing optical channels (packet virtual links). This enhances the utilization of packet ports and optical capacity. Optical grooming, supported by SBVTs, offloads electronic processing burden towards the optical layer[1]. Thereby, both transponder and spectrum utilization are improved by enabling that a set of optical flows from/to different sources/destinations are groomed using particular SBVT[2].

A unified Transport SDN (T-SDN) controller provides the automatic control / configuration of MLN connections. The T-SDN controller relies on a PCE architecture[3,4]. Upon receiving a packet connection, the T-SDN controller triggers a MLN routing algorithm exploiting both electrical and optical traffic grooming opportunities using the current MLN topology (physical and virtual links) and network resources: link bandwidth, optical spectrum, packet ports and SBVTs.

The implemented MLN T-SDN architecture is presented along with detailing a proposed on-line MLN routing algorithm. The whole system is experimentally evaluated upon dynamic traffic and heterogeneous data rates using as a performance metric the bandwidth blocked ratio.
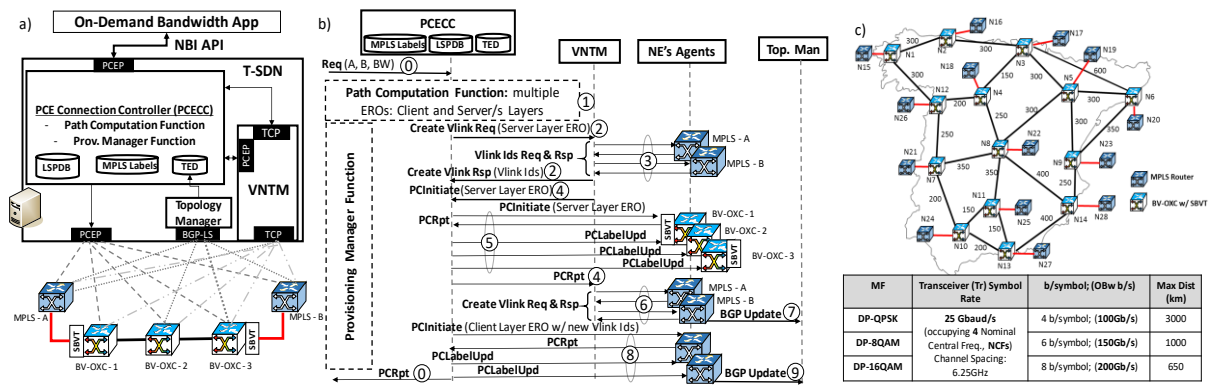
## PCE-Based T-SDN Controller for MLN

Fig. 1.a shows the implemented PCE-based T-SDN controller architecture. The core building block is the PCE Connection Controller (PCECC) which performs three basic functions: i) to process packet connections requests via a NorthBound Interface (NBI) using a PCEP API. Each request specifies (see step 0 in Fig. 1.b) packet endpoints and bandwidth (bits/s); ii) to perform path computation; iii) to handle Provisioning Manager function for configuring Network Elements (NEs) to set up/release packet and optical connections.

Path Computation executes the devised MLN routing algorithm (step 1 in Fig. 1.b). It uses the Traffic Engineering Database (TED) information (topology and network resources) updated by the Topology Manager (i.e., BGP-LS). The output of the Path Computation is encoded as $N$ (=> 1) Explicit Route Objects (EROs) describing the entire MLN path (i.e., nodes, links and network resources). One of these EROs is dedicated to the so-called Client Layer (packet) ERO which besides carrying the packet nodes and links contains the selected label (from the MPLS Label database). If $N>1$, the remaining EROs belong to the Server Layer. Such EROs are related to optical path segments[3] (nodes, links, frequency slot and SBVTs) Each server layer connection induces a packet virtual link which inherits TE properties from its underlying connection: TE metric (based on the number of traversed optical NEs) and available bandwidth derived from the SBVT configuration (symbol rate and modulation format). The available bandwidth of the virtual link may be larger than the requested one ($r$) fostering future electrical grooming chances[2].

Both Client and Server Layer EROs are passed to the Provisioning Manager function. This operates as an Active Stateful PCE[4] storing existing (packet and optical) connections in the LSP database (LSPDB), and instantiating / programming involved NEs via PCLabelUpd[3] (steps 5 and 8 in Fig. 1.b).

To manage (e.g., setting up) each optical path segment, the Provisioning Manager relies on a standalone process referred to as Virtual Network Topology Manager (VNTM). VNTM is responsible for: i) retrieving link identifiers for the new virtual link. To do this, the VNTM communicates with the respective packet NE's Agents connected by the virtual link using a

**Fig. 1:** a) MLN T-SDN architecture; b) Workflow for setting up a packet connection; c) MLN Topology and SBVT capabilities

| MF | Transceiver (Tr) Symbol Rate | b/symbol; (OBw b/s) | Max Dist (km) |
|---|---|---|---|
| DP-QPSK | 25 Gbaud/s (occupying **4** Nominal Central Freq., **NCFs**) Channel Spacing: 6.25GHz | 4 b/symbol; (100Gb/s) | 3000 |
| DP-8QAM | | 6 b/symbol; (150Gb/s) | 1000 |
| DP-16QAM | | 8 b/symbol; (200Gb/s) | 650 |

proprietary TCP protocol (step 3); ii) requesting to the Provisioning Manager (PCEP PCInitiate / PCRpt, in step 4) the allocation/removal of optical resources such as frequency slot and SBVT (transceivers) specified in the Server Layer ERO; iii) triggering the virtual link creation (via the proprietary TCP protocol, step 6) including all its TE properties (metric and available bandwidth) to the respective packet NEs' Agents. Finally, iv) VNTM occupies / releases the used bandwidth (reflecting the SBVT configuration) in the physical ports of the virtual link's NEs (step 6 in Fig. 1.b).

When a virtual link is created / modified / removed (step 6 and 9), respective packet NEs' Agents send a BGP Update message. Indeed, all NEs' Agent keep track of their own local resources being responsible for its configuration / dissemination. Link State information is stored by the Topology Manager in the TED for subsequent path computations. Herein, we considered that whenever a new virtual link is created the establishment of the packet connection over that link is arbitrarily delayed 5s. This is done to ensure that the virtual link is completely configured prior to proceed with the actual programmability of the packet connection.

**On-line Traffic Grooming Algorithm**

The devised on-line MLN routing algorithm aims at applying both electrical and optical grooming strategies to attain the most efficient use of MLN resources and is formed by 2 independent Steps:

Step 1 routes a request, using a shortest path cost (*SP*) algorithm, only over existing packet virtual network topology to favour electrical grooming. If it succeeds, the computed Client Layer ERO is delivered to the Provisioning Manager. Otherwise, Step 2 is executed.

Step 2 triggers a MLN routing combining packet ports, virtual and optical links and SBVTs. SBVTs support different modulation formats (16QAM, 8QAM, QPSK). The algorithm prioritizes selecting the most advanced modulation format (*MF*) as long as the maximum permitted distance (in km) is not exceeded. Bearing this in mind, supported *MF*s are sorted in

decreasing order (bits/symbol), and the SP is executed for each candidate *MF*. If a path is found, this is passed to the Provisioning Manager. Otherwise, next *MF* is explored. For a considered *MF*, the required number of transceivers (*NTr*) in a SBVT is computed to support the requested bandwidth (*r*). Using *NTr*, the fixed symbol rate (*S*) and *MF* (bits/symbol), it is computed the so-called "optical" bandwidth (*OBw*). In general, $OBw (NTr * S * MF) >= r$.
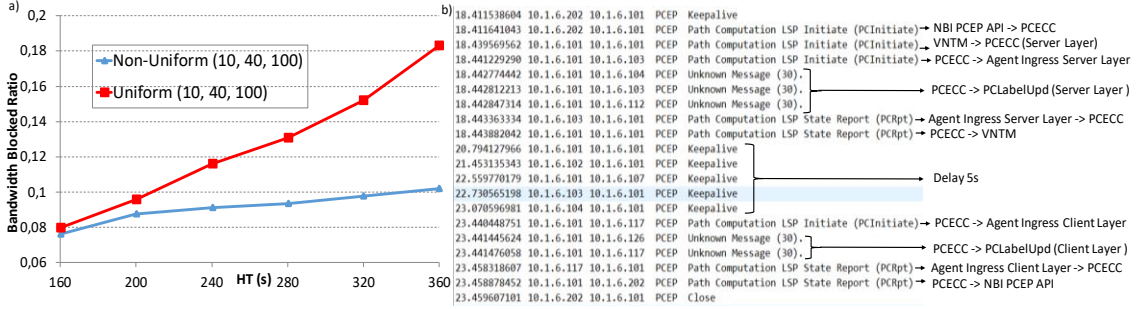
During the SP execution (path tree construction), for each eligible link (packet or optical) a set of restrictions are checked. Specifically, for both physical packet ports and virtual links, it is checked that computed *OBw* can be occupied. If not, the packet link / port is discarded. The cost of allocating *OBw* in a physical port is set to 1 whereas in a virtual link such a cost is matched to TE metric computed when creating the virtual link. In this work, this cost is set to *max {1, number of optical hops - 3}*.

For a candidate optical link (between a pair of BV-OXCs) its cost/metric is always set to 1. To add an optical link to the *SP* it must be verified that: i) the spectrum continuity and contiguity constraints are ensured[3]; ii) the maximum distance permitted for the considered *MF* is not exceeded. If i) or ii) are not satisfied, the optical link is discarded. Finally, the allocation of *NTr* in a SBVT device has also an implicit cost to be accounted. Particularly, if a SBVT where the *NTr* are to be occupied is already being used by another optical flow, the cost of allocating resources in that SBVT is set to 1 (to promote optical grooming). Otherwise, the it is set to 2.

The output path may traverse different optical segments (using *NTr* in SBVTs) and virtual links. If no path is found for all *MF*s, the connection is blocked. Otherwise, Client and Server Layer EROs are sent to the Provisioning Manager.

**Experimental Performance Evaluation**

Fig. 2.b depicts the PCEP messages captured at PCECC for setting up a new packet connection. Upon receiving a PCInitiate message sent through the NBI PCEP API, the MLN path is

```
a) 0,2
Bandwidth Blocked Ratio
0,18                    ┌──────────────────────────┐
0,16                    │ ──  Non-Uniform (10, 40, 100) │
                        │ ─■─ Uniform (10, 40, 100)    │
0,14                    └──────────────────────────┘
0,12
0,1
0,08
0,06
      160   200   240  HT (s)  280   320   360
```

b)
| Time | Src | Dst | Proto | Info | |
|---|---|---|---|---|---|
| 18.411538604 | 10.1.6.202 | 10.1.6.101 | PCEP | Keepalive | |
| 18.411641043 | 10.1.6.202 | 10.1.6.101 | PCEP | Path Computation LSP Initiate (PCInitiate) | → NBI PCEP API -> PCECC |
| 18.439569562 | 10.1.6.101 | 10.1.6.101 | PCEP | Path Computation LSP Initiate (PCInitiate) | → VNTM -> PCECC (Server Layer) |
| 18.441229290 | 10.1.6.101 | 10.1.6.103 | PCEP | Path Computation LSP Initiate (PCInitiate) | → PCECC -> Agent Ingress Server Layer |
| 18.442774442 | 10.1.6.101 | 10.1.6.101 | PCEP | Unknown Message (30). | |
| 18.442812213 | 10.1.6.101 | 10.1.6.103 | PCEP | Unknown Message (30). | → PCECC -> PCLabelUpd (Server Layer ) |
| 18.442847314 | 10.1.6.101 | 10.1.6.112 | PCEP | Unknown Message (30). | |
| 18.443363334 | 10.1.6.103 | 10.1.6.101 | PCEP | Path Computation LSP State Report (PCRpt) | → Agent Ingress Server Layer -> PCECC |
| 18.443882242 | 10.1.6.101 | 10.1.6.101 | PCEP | Path Computation LSP State Report (PCRpt) | → PCECC -> VNTM |
| 20.794127966 | 10.1.6.101 | 10.1.6.101 | PCEP | Keepalive | |
| 21.453135343 | 10.1.6.101 | 10.1.6.102 | PCEP | Keepalive | |
| 22.559770179 | 10.1.6.101 | 10.1.6.107 | PCEP | Keepalive | → Delay 5s |
| 22.730565198 | 10.1.6.103 | 10.1.6.101 | PCEP | Keepalive | |
| 23.070596981 | 10.1.6.101 | 10.1.6.104 | PCEP | Keepalive | |
| 23.440448751 | 10.1.6.101 | 10.1.6.101 | PCEP | Path Computation LSP Initiate (PCInitiate) | → PCECC -> Agent Ingress Client Layer |
| 23.441445624 | 10.1.6.101 | 10.1.6.126 | PCEP | Unknown Message (30). | → PCECC -> PCLabelUpd (Client Layer) |
| 23.441476058 | 10.1.6.101 | 10.1.6.117 | PCEP | Unknown Message (30). | |
| 23.458318607 | 10.1.6.117 | 10.1.6.101 | PCEP | Path Computation LSP State Report (PCRpt) | → Agent Ingress Client Layer -> PCECC |
| 23.458878452 | 10.1.6.101 | 10.1.6.202 | PCEP | Path Computation LSP State Report (PCRpt) | → PCECC -> NBI PCEP API |
| 23.459607101 | 10.1.6.101 | 10.1.6.202 | PCEP | Close | |

**Fig. 2:** a) Bandwidth Blocked Ratio vs HT (s); b) PCEP messages capture for setting up Server and Client Layer EROs

computed. For each new virtual link, derived from the computed path, the VNTM invokes the Provisioning Manager (PCInitiate) to establish their associated Server Layer/s (optical segment/s). This entails programming involved optical NEs (PCLabelUpd). To avoid convergence issues, the actual packet connection configuration (PCInitiate / PCLabelUpd) is delayed 5s, ensuring that the optical segment/s and virtual link/s are actually set up and created. Finally, a PCRpt is sent through the NBI to report that the connection is active. Interactions via proprietary TCP protocol and BGP Updates are not shown due to the space limitations. Nevertheless, it is worth mentioning that BGP Update messages have been extended to carry within standard Path Attribute Link State details about supported link switching technology (packet or flexi-grid).

The performance evaluates the entire system (T-SDN controller and MLN routing algorithm) under dynamic packet traffic with heterogeneous data rates and models (uniform vs. non-uniform). Packet connection arrival process is Poisson (mean inter-arrival time is 20s) and the duration (holding time, HT) is exponentially modelled where mean HT is varied. For each HT data point, 3000 connections are generated. Uniform model considers that 10, 40 and 100 Gb/s connection requests are uniformly demanded. Non-uniform considers 10, 40 and 100 Gb/s data rates generated with a proportion of 6:3:1.

Fig. 1.c. shows the MLN topology: 14 MPLS NEs connected via SBVTs to 14 BV-OXCs forming an optical mesh network. MPLS NEs are equipped with a single 400 Gb/s port. Each SBVT supports three MFs (DP-QPSK, DP-8QAM, DP-16-QAM) having 10 transceivers. SBVT's features (fixed symbol rate and maximum permitted distance per *MF*) are detailed in Fig. 1.c along with optical link distance (km). Finally, optical links support 128 NCFs.

Fig. 2.a plots the bandwidth blocked ratio versus the HT for Uniform and Non-uniform models. As HT increases, blocked bandwidth degrades since resources become more used which complicates finding a feasible path. In this regard, the most limiting network resource is packet port. Indeed, when a new optical segment (part of a MLN path) is being set up, occupied port bandwidth is *OBw*. Recall that *OBw* is related to the SBVT configuration. Therefore, *OBw* gets bandwidth values (100, 150, 200Gb/s, Fig. 1.c) larger than the requested packet connection's bandwidth, i.e., *r*. Thereby, port bandwidth is exhausted faster than other resources (e.g., optical spectrum and SBVT) as HT grows. In light of this, grooming traffic flows becomes crucial to optimize port bandwidth usage. This fact is better appreciated observing that Non-uniform traffic does lower blocked bandwidth compared to Uniform. The reason behind that is that in Non-Uniform, low- and medium-rate requests (10-40 Gb/s) are proportionally larger than high-rate demands. Thus, chances to groom that connections into existing virtual topology (using unused port's bandwidth, *OBw*) are more likely to occur than in Uniform model. Hence, in the Non-uniform the number of established demands are higher increasing in turn the served bandwidth.

## Conclusions

The implementation of a unified T-SDN controller for MLN (packet over flexi-grid optical) is detailed. A new MLN routing algorithm is experimentally evaluated for dynamic and heterogeneous data-rate packet traffic aiming at optimizing MLN resources via grooming strategies.

## Acknowledgements

## References

[1] S. Zhang et al., "Evolving Traffic Grooming in Multi-Layer Flexible-Grid Optical Networks with Software-Defined Elasticity", J. Light. Tech., vol. 32, no. 16, p. 2905 (2015)

[2] J. Zhang et al., "Energy-Efficient Traffic Grooming in Sliceable-Transponder-Equipped IP-over-Elastic Optical Networks", J. Opt. Comm. Net., Vol. 7, no.1, (2015).

[3] R. Martínez et al "Distributed vs. centralized PCE-based transport SDN controller for flexi-grid optical networks" in proc. OFC2017.

[4] R. Casellas et al, "SDN Orchestration of OpenFlow and GMPLS Flexi-grid Networks with a Stateful Hierarchical PCE", JOCN, vol. 17, no. 1, (2015).