# Analysis of Flex Ethernet Architectures in Optical Transport Networks

André Marcelo Fernandes Pereira
Instituto Superior Técnico, Lisbon, Portugal

November 2017

**Abstract**

The rise of data-center traffic led to the development of Data-Center Interconnect (DCI) network solutions, whose market has been rising significantly in the past few years. DCI modules are simple devices that provide an interface between data-centers and the underlying optical transport network while ensuring scalability and cost efficiency. These modules are responsible for mapping flows into flexible optical interfaces and providing switching funtionalities. However, in order to maintain simplicity and low costs, any client flexibility must be handled outside this equipment. The introduction of Flex Ethernet (FlexE) addresses this problem, introducing client flexibility by decoupling flows from the physical interfaces connecting routers and transport boxes. FlexE introduces a new layer in the Ethernet stack allowing for the virtualization of flows across Ethernet interfaces, allowing single flows to span multiple interfaces and multiple flows to be grouped into one interface. This thesis studies the use of FlexE solutions in a Dense Wavelength Division Multiplexing (DWDM) transport network scenario, analyzing how the degree of transport box awareness to FlexE clients impacts the overall network efficiency. A network simulation was developed in order to assess the multiple proposed FlexE scenarios influence the provisioning of FlexE client flows in a DCI scenario and consequently, the efficiency of resource usage, such as router cards, transport equipment, client interfaces and DWDM transport interfaces.
**Keywords:** Optical transport networks; FlexE; DCI; DWDM.

## 1. Introduction

Telecommunication networks play a central role in the interconnection of today's society. In recent years, cloud applications have grown in popularity, allowing for remote computation and storage. This increased need for cloud resources has led to the emergence of large scale data-center facilities that house multiple companies, applications and user data. Cloud traffic is expected to reach 14.1Z Bytes/year (1 Zettabyte = $10^{21}$ Bytes = $10^{12}$ GBytes)[3], adding up to about 92% of total data-center traffic.

In order to cope with these amounts of traffic, data-center operators are focusing efforts and money in developing their Data-Center Interconnect (DCI) network solutions, ensuring security backups and redundancy of critical data, guarantying scalability, efficiency and low latency while remaining affordable and simple [2]. DCI networks are therefore becoming evermore important to operators and big corporations, with the DCI market growing almost 50% in 2016 alone [1].

Employing a flexible client solution in DCI networks provides a simple yet efficient way of dealing with large amounts of traffic, independently of bit-rate, source and destination. Therefore, this study proposes the use of Flex Ethernet (FlexE) as one solution for DCI.

## 2. Background

A data-center consists in a group of servers that can host a number of different applications and store their data remotelly. Data-centers have a hierarchical architecture with multiple layers of Layer 2 switches, topped with a Layer 3 IP/MPLS router that routes traffic to and from the data-center.

Large companies typically own several geographically separated data-center facilities in order to ensure data redundancy. This means that data-centers must communicate with each other, forming a Data-Center Interconnect (DCI) network. In a DCI network, data-centers are connected to one another via IP/MPLS routers, placing a heavy burden on core transport networks.

DCI networks (Figure 1), use specific devices to provide an interface with the underlying transport network. IP/MPLS routers are connected to a DCI module via multiple Ethernet links. This DCI module provides an interface between these routers and the network elements of the transport network, the ROADMs by multiplexing client signals arriving from a router (client-side) and converting them to the optical domain. The resulting optical channels (line-side) are then switched across the DCI network, to a certain destination router.

DCI modules must maintain contact with the trans-

port network's control plane, requesting available lightpath capacity for incoming clients. The control plane is then responsible for the establishment of lightpaths at the DCI modules, taking into account the overall network's state [5].

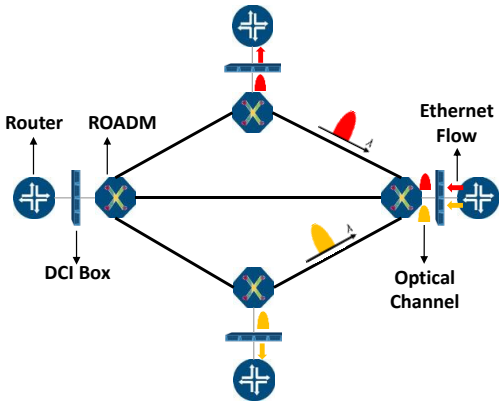


Figure 1: DCI network scenario [5].

As data-center traffic continues to rise, so does the diversity of applications and services, which results in a large number of traffic flows with variable bit rates. The use of fixed Ethernet rates in a data-center scenario can become inefficient in terms of resource usage, since it limits the number of clients in one interface, not taking full advantage of its full capacity. Moreover, as the used Ethernet standards continue to increase in bit-rate (e.g. up to 100Gb/s and 400Gb/s), the efficient mapping of such large clients into an optical transport network in a DCI scenario becomes inefficient and complex. Therefore, there is a need to introduce flexibility at the service layer as a means of efficiently mapping diverse clients, with variable rates, allowing for a more efficient use of network resources. Flex Ethernet (FlexE) was introduced to address such problems, providing client flexibility for DCI networks [6]. FlexE allows for the dissociation of client rates and the physical medium by adding a FlexE shim between the MAC and PCS layers in the Ethernet protocol stack as well as for the grouping of Ethernet physical interfaces, PHYs, in order to achieve higher rates then the existing fixed ones. FlexE also provides a means of grouping multiple client signals into one single PHY or even spread different clients across multiple PHYs, independenly of their rate [8]. In order to do so, FlexE uses TDM techniques for the multiplexing of individual, fixed size, calendar slots which can be filled with any MAC flow [6].

### 2.1. Flex Ethernet Architectures

An implementation of FlexE is proposed in [7], where PHYs in the router-to-transport module or transport box (R-T) connection can carry multiple clients smaller than their capacity, in what is called sub-rating, or allow for the demultiplexing of larger clients across several PHYs, through PHY bonding. This allows for the channelization of flows, meaning that each PHY can carry different clients. This scenario is represented in Figure 2 a). Logical groups of PHYs form FlexE groups that are defined between two FlexE shims. These shims map/de-map client flows to or from the group's PHYs. FlexE's general architecture is presented in Figure 2 b).

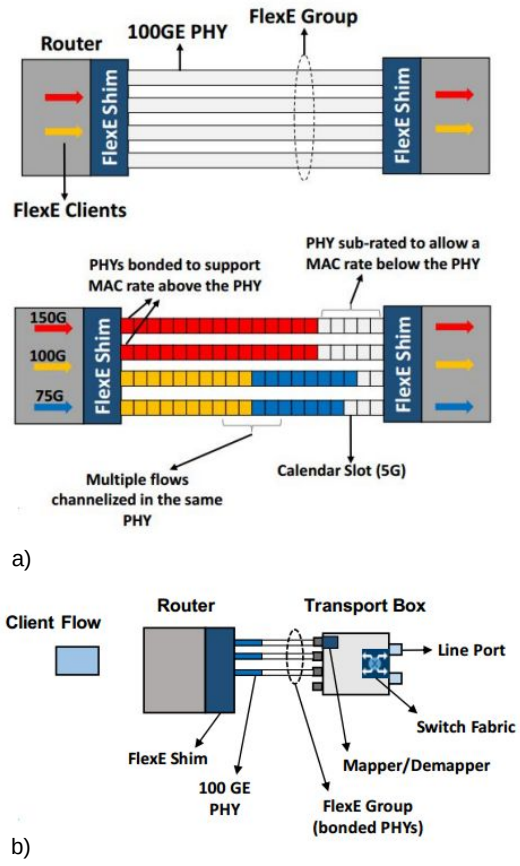


Figure 2: Flex Ethernet Structure: a) PHY bonding, sub-rating and channelization; b) Architecture legend [5].

FlexE foresees three different scenarios in terms of the complexity of the transport equipment and its awareness to FlexE clients: FlexE-Unaware, FlexE-Partially-Aware and FlexE-Aware. In the FlexE-Unaware scenario (Figure 3), the transport box has no visibility over the contents of the FlexE groups, allocating all the PHYs' capacity line-side, independently of the actual transported payload. The shims are placed exclusively at the routers, and the transport boxes simply map the PHYs into appropriate transport containers. This architecture implies that all the clients mapped into a FlexE group must have the same end-points. Moreover, the individual clients are only recovered at the FlexE shim in the destination router, meaning each FlexE group consists of a symmetrical set of PHYs on both end-nodes. A FlexE group must also be associated with a unique pair of shims, meaning

its PHYs must be connected to a unique router card on the source and destination nodes. Furthermore, PHYs belonging to the same FlexE group must take the same path in the network, eliminating the need for delay compensation at the shim. This imposes a co-routing constraint that must be accounted for.

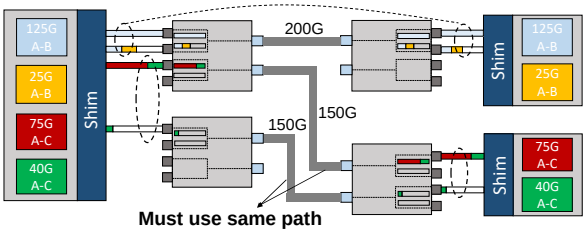


Figure 3: FlexE-Unaware scenario architecture [4].

The FlexE-Partially-Aware scenario, in Figure 4, is similar to the Unaware scenario. However, it foresees a limited level of FlexE awareness in the transport box. In this instance, transport boxes are only aware of which slots are being used by client flows, and which ones are not. This is achieved via a mapper/demapper at each transport box's client ports. However, this introduces additional complexity in the transport equipment. Since the client mapping is no longer fixed, arbitrary sized payload is mapped into ODU*flex* containers, requiring the presence of switching fabric for container aggregation within the transport box itself.

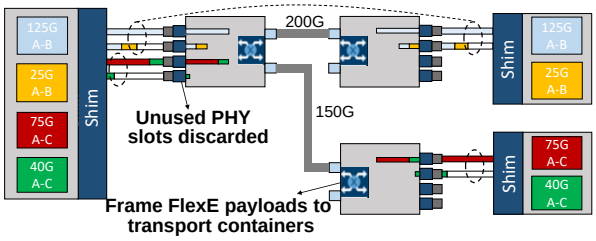


Figure 4: FlexE-Partially-Aware scenario architecture [4].

The FlexE-Aware transport scenario, in Figure 5, confines FlexE groups to a single R-T connection by placing additional FlexE shims at the transport box. A single FlexE group can carry clients with multiple destinations, as they will be individually recovered at the transport box. Consequently, the transport box must have inherent grooming capabilities to aggregate and switch the flows onto the desired optical channels, adding complexity to its design. Contrary to previous scenarios, FlexE groups can no longer span across multiple transport boxes.

## 3. Implementation

The purpose of this analysis is to minimize cost of the FlexE solution by minimizing the number of required hardware, while ensuring resource usage efficiency.

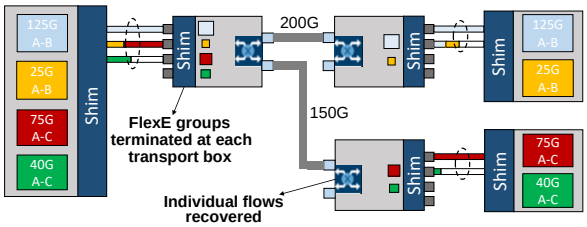


Figure 5: FlexE-Aware scenario architecture [4].

Given a specific network $\Gamma(V, E)$, each node has multiple router cards, $r_k^i$, and transport boxes, $t_k^i$, depending on the number of demands it processes. Router cards and transport boxes are physically connected via Ethernet PHYs. This analysis considers two main FlexE settings, regarding the rate of client interfaces: 100 GbE interfaces and 400 GbE client interfaces. In the first case, each router card has 16 client ports, with a total capacity of 1.6 Tb/s, and transport boxes have a total capacity of 400 Gb/s distributed across 4 client cards and 2 optical line ports. In the second case router cards have 16 client ports, with a total capacity of 6.4 Tb/s, while transport boxes remain with a total capacity of 400 Gb/s, using one single 400 GbE client card.

In order to assess the outcome of employing a FlexE solution, the multiple FlexE scenarios described in Section 2.1 were implemented and simulated. However, the implementation of these scenarios must take into account a set of constraints.

Considering a set of $N$ FlexE client flows $F = \{f_1, f_2, ..., f_N\}$, each client, with a certain source and destination, $f_j(s, d)$, and a bit-rate of $\gamma_{f_j}$ [Gb/s], must be assigned to a single FlexE group $g_y$. Each group is also characterized by its source and destination nodes, $g_y(s, d)$. A certain FlexE group contains $M$ FlexE flows, $g_y(s, d) = \{f_1^y, f_2^y, ..., f_M^y\}$ and has a total rate $rate_{g_y}$ [Gb/s]. This rate must be met with sufficient PHYs of a certain capacity, $PHY\_cap$ [GbE]. In this sense, a group's total PHY capacity, $cap_{g_y}$ [GbE] must respect $cap_{g_y} \geq rate_{g_y}$. As PHYs may be sub-rated, a FlexE group might have available free capacity, $free_{g_y}$ [Gb/s] in its PHYs, given by $cap_{g_y} - rate_{g_y}$. Note that $cap_{g_y}$ must be provisioned on both the source and destination nodes. The number of PHYs connected to a given router card must be limited by the number of available router ports, $Router\_Ports$, and the number of PHYs connected to a given transport box must also be limited to the number of available client ports, $Tp\_Clt\_Ports$.

A transport box at node $i$, $t_k^i$, has a number of line ports, $Tp\_Line\_Ports$. Each line port can deploy a lightpath $l_{g_y,t_k}^{p,\lambda}$, using a physical path $p$, assigned with wavelength $\lambda$ and connected to the transport box $t_k$ of group $g_y$. A transport box's line-side capacity must suffice for mapping all connected FlexE clients while limiting the number of deployed lightpaths to the avail-

able number of line ports. Note that each lightpath deployed at a source node, must have a corresponding lightpath associated to a transport box at the destination node, using the same path $p$, wavelength $\lambda$ and serving the same FlexE group $g_y$. Lightpaths that share the same wavelength $\lambda$ may not take physical paths with links in common due to the wavelength continuity constraint. Finally, a co-routing constraint for all lightpaths carrying PHYs of the same FlexE group must be ensured, meaning that different lightpaths assigned to the same FlexE group $g_y$ must always take the same path $p$ through the network.

In order to overcome the impracticality of finding an optimal solution, heuristic methods where developed for the provisioning of FlexE client flows for each FlexE scenario. In this sense, a sequential flow assignment algorithm based on a greedy approach is proposed, with the objective of minimizing the overall required hardware, hence the overall cost of the solution.

### 3.1. FlexE-Unaware and FlexE-Partially-Aware flow provisioning algorithms

The proposed heuristic for the FlexE-Unaware scenario foresees three outcomes for the assignment of a flow $f_j$ to a FlexE group: assign $f_j$ to an existing group, expanding an existing group and assign $f_j$ to it or assign $f_j$ to a new FlexE group. The first outcome occurs if there is an existing FlexE group with the same end-nodes as $f_j$ with enough free capacity in its PHYs for all of it, respecting the co-routing constraint. If that group only has enough free capacity for a part of $\gamma_{f_j}$, it may be extended by adding new PHYs. Lastly, if there is no existing group available, a new one must be created. The process of creating a new FlexE group or extending an existing one must first verify if there is an available path in the network for $f$, accounting for the co-routing constraint. If there are no available paths, $f_j$ may be blocked. However, this process imposes that an initial path $p_1$ must be found for $f_j$, in order to search groups that also use $p_1$. However, because it is not possible to know how many wavelengths will be needed beforehand, the worst case scenario is assumed (i.e. two additional wavelengths will be needed).

Algorithm 1: Flow provisioning for the FlexE-Unaware Scenario with 100GbE client interfaces.

**Input:** Network $\Gamma(V, E)$; Flow $f_j$ with capacity $\gamma_{f_j}$ between nodes $sd$; Set of existing FlexE groups $G$.

From $G$, extract all FlexE groups between nodes s d: G(s, d).
**for all** group $g_y \in G(s,d)$ **do**
    Check the free available capacity in the PHYs of $g_y$, $free_{g_y}$, respecting the co-routing constraint.
**end for**
**if** $\max free_{g_y} \geq \gamma_{f_j}$ **then**
    Assign flow $f_j$ to group $g_y$ with the highest $free_{g_y}$.
**else**
    **for all** $g_y \in G(s,d)$ **do**
        Determine how many new PHYs, $n_{PHYs}$ must be added to $g_y$, such that it supports $\gamma_{f_j}$.
        Determine how many new lightpaths, $n_{LPs}$, are required to $n_{PHYs}$ PHYs.
        **if** *Router_Ports* of router $r_k^i$ serving $g_y < n_{PHYs}$ **or** available wavelengths in $p < n_{LPs}$ **then**
            Remove $g_y$ from $G(s,d)$.
        **end if**
    **end for**
    **if** $G(s,d)$ is not empty **then**
        Select the group $g_y$ with the smallest $n_{PHYs}$. As a tie-breaker, select the one with the highest $free_{g_y}$.
        Provision the required $n_{PHYs}$ and connect them to the source/destination routers of $g_y$.
        **for** $i = 1 : n_{PHYs}$ **do**
            Search for an available transport box client port connected to an existing lightpath $l^{p,\lambda}_{g_y,t_k}$ with sufficient capacity for the PHY $i$. If found, connect PHY $i$ to this client port map it into $l^{p,\lambda}_{g_y,t_k}$.
            If PHY $i$ has not yet been assigned, search for an available transport box with an unused line port. If found, connect PHY $i$ to this client port and provision a new lightpath $l^{p,\lambda_{new}}_{g_y,t_k}$ using the First-Fit algorithm to assign it the first available wavelength $\lambda_{new}$.
            If PHY $i$ has not yet been assigned, provision a new transport box, and connect PHY $i$ to its first client port. Provision a new lightpath $l^{p,\lambda_{new}}_{g_y,t_k}$ using the First-Fit algorithm to assign it the first available wavelength $\lambda_{new}$.
        **end for**
    **else**
        Using the $k$-Shortest Path algorithm, compute the first path $p$ between nodes $s-d$ with sufficient available wavelengths to carry $\left\lceil \frac{\gamma_{f_j}}{PHY\_cap} \right\rceil$ PHYs.
        **if** a path $p$ was found **then**
            Create a new FlexE group $g_y$ with $\left\lceil \frac{\gamma_{f_j}}{PHY\_cap} \right\rceil$ PHYs. Connect the PHYs to a single router card on the source and destination nodes. If necessary, provision a new router card(s).
            **for** $i = 1 : \left\lceil \frac{\gamma_f}{PHY\_cap} \right\rceil$ **do**
                Search for an available transport box with an unused client and line port. If found, connect PHY $i$ to this client port and provision a new lightpath $l^{p,\lambda_{new}}_{g_y,t_k}$ using the First-Fit algorithm to assign it the first available wavelength $\lambda_{new}$.
                If PHY $i$ has not yet been assigned, provision a new transport box, and connect PHY

$i$ to its first client port. Provision a new lightpath $l^{p,\lambda_{new}}_{g_y,t_k}$ using the First-Fit algorithm to assign it the first available wavelength $\lambda_{new}$.

30: **end for**
31: **else**
32: Block the flow $f_j$
33: **end if**
34: **end if**
35: **end if**

**Output:** Updated set $G$.

The FlexE-Partially-Aware scenario is very similar to the FlexE-Unaware scenario, and the developed heuristic for it operates in the same way. The only difference between the two scenarios is that instead of transparently mapping client PHYs, the FlexE-Partially-Aware scenario only maps actual traffic.

Another considered scenario uses a transport box configuration with one single 400 GbE client-port and the same two 200 Gb/s line-ports, shown in Figure 6. This configuration alters the overall FlexE architecture and consequently, the heuristic approach taken. Since transport boxes now have a single client interface, line-ports must be provisioned with lightpaths when the equipment is created. In the Unaware scenario, transport boxes must be created taking into account the allowed lightpath rates. For example, if lightpaths are restricted to 100 Gb/s, in order to transparently map a 400 GbE client, four lightpaths will be required, therefore, transport boxes must have four line ports. In the Partially-Aware scenario, the client's capacity can be simply limited according to lightpath rates. Therefore, for the Unaware scenario, the number of required wavelengths cannot be known beforehand, similarly to what happens in the previous scenario using transports with multiple 100 Gb/s client interfaces. In this sense, a worst case scenario of requiring four wavelengths is assumed at the beginning. Moreover, since R-T connections can only carry flows with the same destinations, the pair of lightpaths originating in every transport box must take the same path over the network. The heuristics developed for both the Unaware and Partially-Aware scenarios are very similar. However, the in the Partially-Aware scenario, it is always known that two wavelengths will be required, when initializing new transport equipment.

### 3.2. FlexE-Aware flow provisioning algorithms

In the FlexE-Aware scenario, transport boxes can now process individual FlexE clients, independently of their destination. In this sense, the approach taken now is to first find a suitable router-transport-transport-router combination and then make the lightpath provisioning. This adds some complexity to the developed heuristic, specially when the number of hardware escalates, since all combinations of transport boxes must be considered. The following heuristic starts by testing all possible combinations of transport

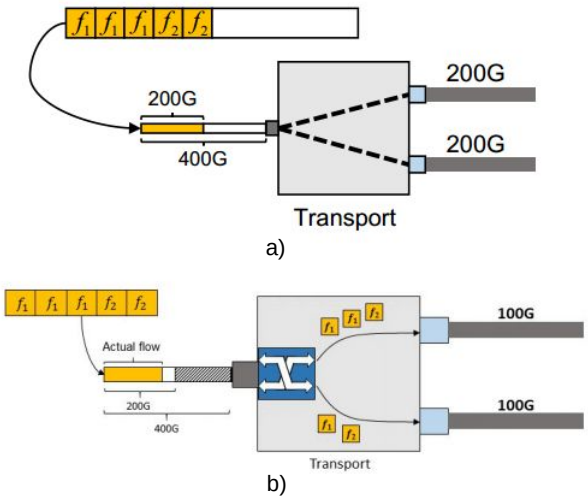


Figure 6: Transport boxes with 400 GbE client interfaces: a) FlexE-Unaware; b) FlexE-Partially-Aware.

boxes for available capacity for a client $f_j$. Four possible outcomes may occur: using two existing transport boxes (source/destination), using and existing transport box at the source and a new one at the destination and vice-versa, or using two new transport boxes, if none of the previous combinations are available. For each of the possible solutions obtained, it is also necessary to check whether or not new router cards must be created. The computed solutions are then ordered by least new router cards, then by least new transport boxes and finally by most PHYs left unused. The best solution found is then considered for the provisioning of new lightpaths. It is important to note that the same drawback stated in 3.1 regarding the uncertainty of the number of wavelengths needed still applies.

Algorithm 2: Flow provisioning for the FlexE-Aware Scenario with 100 GbE client interfaces

**Input:** Network graph $\Gamma(V, E)$; Flow $f_j$ with capacity $\gamma_{f_j}$ between nodes $s - d$.

1: **for all** transport box at the source node, $t_s$ **do**
2: **for all** transport box at the destination node, $t_d$ **do**
3: Check how much capacity is available in the existing lightpaths between $t_s$ and $t_d$ and store it in $cap_{ex}(t_s, t_d[Gb/s])$.
4: Check how much potential capacity is available through the unused line-ports of $t_s$ and $t_d$ and store it in $cap_{pot}(t_s, t_d)[Gb/s]$.
5: Set the number of new transport boxes required $new_{TB}(t_s, t_d) = 0$.
6: **if** $t_s$ is a virtual (not yet deployed) transport box **then**
7: Set $new_{TB}(t_s, t_d) = new_{TB}(t_s, t_d) + 1$.
8: **end if**
9: **if** $t_d$ is a virtual (not yet deployed) trans-

port box **then**
Set $new_{TB}(t_s, t_d) = new_{TB}(t_s, t_d) + 1$.
**end if**
**end for**
**end for**
Get the set T of all transport box pairs $(t_s, t_d)$, such that either $cap_{ex}(t_s, t_d) \geq \gamma_{f_j}$ **or** $cap_{pot}(t_s, t_d) \geq \gamma_{f_j}$.
**if** T is not empy **then**
**for all** transport box pair $(t_s, t_d)$, where $cap_{ex}(t_s, t_d) \geq \gamma_{f_j}$ **or** $cap_{pot}(t_s, t_d) \geq \gamma_{f_j}$ **do**
**for all** each router at the source node, $r_s$ **do**
**if** $r_s$ has sufficient port capacity to carry $\gamma_f$ to $t_s$ **then**
Store the number of router ports avvailable after this assignment, $ports_{r_s}$.
**end if**
**if** $r_s$ is a virtual router **then**
Set $new_{TB}(t_s, t_d) = 1$.
**end if**
**end for**
Select the router with the lowest value of $new_R(r_s)$and as a tie-breaker the one with the highest value of $ports_{r_s}$.
Store the selected source router $r_s$ for the transport box pair $(t_s, t_d)$, and set the router count to $new_R(t_s, t_d) = new_R(r_s)$.
Repeat steps **17-26** for the destination router $r_d$. Store the selected destination router, and update the router count: $new_R(t_s, t_d) = new_R(r_s) + new_R(r_d)$.
**end for**
Select the solution $(t_s, t_d, r_s, r_d)$ that, by decreasing priority:
a) Minimizes the number of new routers: $new_R(t_s, t_d)$;
b) Minimizes the number of new transport boxes: $new_{TB}(t_s, t_d)$;
c) Maximizes the amount of leftover unused router ports: $ports_{r_s} + ports_{r_d}$.
**else**
Block flow $f$.
**end if**

**Output:** Selected solution $(t_s, t_d, r_s, r_d)$ or blocking status.

## 4. Results

The heuristics described in Section 3 for flow provisioning were implemented using MATLAB in order to assess the performance and efficiency of the different FlexE scenarios. For this analysis two transport network topologies were considered: the German Backbone Network (GBN) and the United States Backbone Network (UBN), both represented in Figure 7.

The considered optical reach thresholds for the established lightpaths are 600km for 200Gb/s and 1300km for 150Gb/s rates. Beyond that, 100Gb/s rates are imposed. These rate limits imposed by path distance result from the applied modulation schemes, namely 16-QAM, 8-QAM and QPSK, respectively, all considering a 50GHz inter-channel spacing. Each lightpath is assumed to have a total capacity of 80 channels.

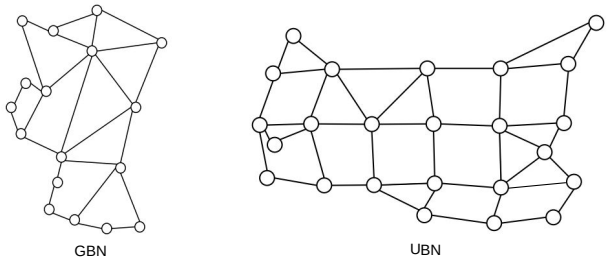


Figure 7: GBN and UBN network topologies.

Client flows can have rates of 10, 40 and $m\times$ 25Gb/s, with $1 \leq m \leq 8$, with calendar slots of 5Gb/s. As for client flow distribution, it can be uniform when all flows have the same occurring probability or weighted, if each flow's occurring probability is inversely proportional to its rate. This results in a smaller number of larger flows for the uniform distribution and in a larger set of smaller flows for the weighted client distribution. The total network traffic loads considered were 10, 30, 50, 70 and 100 Tb/s.

The traffic patterns range from single point-to-point connections to a full logical mesh topology, where each node is connected to all the others. These logical configurations are intended to emulate multiple DCI scenarios, from data-center replication to a more distributed and diverse traffic scenario. This variable is referred to as number of destinations per node. For the GBN network were considered 1, 4, 8, 12 and 16 destinations per node and 1, 6, 12, 18 and 23 destinations per node for the UBN network. Lastly, this analysis also considers three $k$-shortest path limits, namely $k = 1, 3$ and 5.

The case using 100 GbE client interfaces was studied first. In this case, router cards are considered to have a capacity of 1.6 Tb/s, distributed over 16×100 GbE Ethernet interfaces and transport boxes are considered to have a capacity of 0.4Tb/s, distributed over 4×100 GbE Ethernet interfaces on the client-side and 2×200 Gb/s line interfaces, referred to as 0.4T. Another scenario is considered where transport boxes have a capacity of 1.6 Tb/s distributed over 16×100 GbE client interfaces and 8×200 Gb/s line interfaces, referred to as 1.6T.

In the upcoming analysis a network load of 30Tb/s was fixed and the traffic topology was varied, from 1-1 communication to all-all communication, for both the GBN and UBN networks. Only the results for the shortest path were considered ($k = 1$), since this variable has little to no influence over the overall results when the total traffic load is fixed. Moreover, only the

cases using weighted client traffic are presented, since they make the differences between scenarios more evident.

The results of Figures 8 a) and b) show the amount of router cards required per Tb/s of carried traffic versus the number of destinations per node for the different FlexE scenarios while the results of Figures 8 c) and d) depict the amount of line ports required per Tb/s of carried traffic versus the number of destinations per node for both the GBN and UBN, respectively. For the GBN network, all scenarios but the Aware 1.6T decrease their efficiency with growing traffic pattern complexity, needing more router cards. As the number of destinations per node increases, so does the number of possible flow destinations. Therefore, the scenarios without full awareness of FlexE clients struggle with this increase of destinations, since they are oblivious to individual clients.

In the Aware scenarios, flows can be virtualized into a FlexE group independently of their destination, making them practically independent of the traffic pattern. However, in the Aware 0.4T scenario, the reduced transport box size overcomes this advantage, since FlexE groups cannot span multiple transport boxes. When the number of flows increases, in the weighted traffic scenario, the Aware 0.4T scenario actually performs worse than the other less-aware scenarios since new router ports must be used whenever a FlexE group cannot hold more flows. Because the Aware 1.6T scenario uses transport boxes with four times the size of the Aware 0.4T, this limitation is no longer felt, for the amount of traffic considered.

For the UBN network, the number of required router cards is significantly higher then with the GBN, since it is larger. Therefore, the reach constraints imposed to lightpaths due to path distances result in a less efficient use of the existing hardware, leading to the creation of more new equipment. The previous considerations regarding Aware scenarios also apply to the UBN, with the Aware 1.6T being the most efficient one.

Regarding the amount of required line ports, it is clear to see that the Aware 1.6T scenario has the worst performance of all scenarios. In part, this is due to the size of transport boxes, with more line ports than the other scenarios. However, the overall number of deployed transport boxes is slightly higher for the Aware 1.6T than for the Aware 0.4T scenario despite having larger transport boxes. This can only mean that lightpaths are not being efficiently filled, resulting in the provisioning of new transport equipment while there are still lightpaths with free capacity.

Regarding the Unaware and Partially-Aware scenarios, the latter outperforms the former in terms of line port efficiency, meaning Partially-Aware transport boxes fill lightpaths more efficiently. This is due to the fact that these transport boxes can map clients more efficiently into 150 Gb/s lightpaths, since they discard empty client calendar slots. On the other hand, Unaware transport boxes transparently map 100 Gb/s clients into lightpaths. In this sense, whenever there's a 150 Gb/s lightpath, only 100 Gb/s will actually be used. This effect is reduced when network size grows, simply because 150 Gb/s lightpaths become rarer while 100 Gb/s lightpaths become more common.

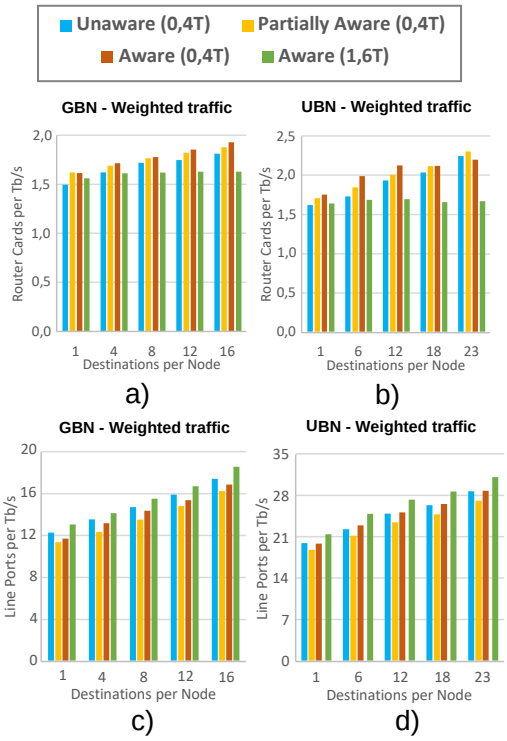


Figure 8: Simulation results: a) Router cards per Tb/s, GBN; b) Router card per Tb/s, UBN; c) Line ports per Tb/s, GBN; d) Line ports per Tb/s, UBN.

The results of Figures 9 a) and b) shows the average occupation ratio of deployed lightpaths versus the number of destinations per node, for the different FlexE scenarios while the results of Figures 9 c) and d) show the average occupation ratio of active client PHYs versus the number of destinations per node, for both the GBN and UBN, respectively.

The overall lightpath occupation ratio drops with the increase of traffic pattern complexity. For point-to-point communication, flows remain more concentrated on single FlexE groups, taking more advantage of the available lightpath capacity. However, with a more meshed traffic pattern, flows with different destinations are scattered across multiple FlexE groups resulting in the usage of empty line ports rather then already existing lightpaths. Even in the Aware 1.6T scenario, where FlexE groups only exist between single R-T connections, the increase in flow destinations requires new lightpaths serving them, leaving existing ones with free capacity, simply because they serve different destinations.

Contrary to the Aware 1.6T scenario, the Partially-

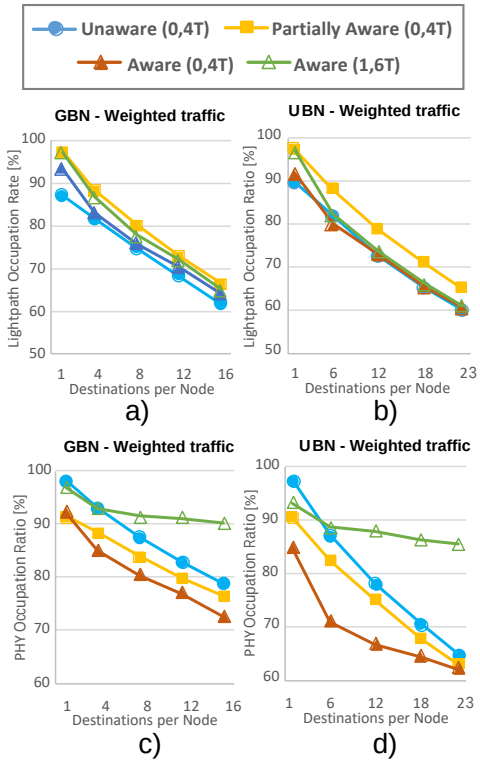


Figure 9: Simulation results: a) PHY occupation ratio, GBN; b) PHY occupation ratio, UBN; c) Lightpath occupation ratio, GBN; d )Lightpath occupation ratio, UBN.

Aware scenario's capability of spreading flows from the same FlexE group across multiple transport boxes, allied to the fact that it can discard unused traffic slots, makes it the most efficient scenario regarding line-side efficiency.

Analyzing the PHY occupation ratios, it's possible to identify the trade off for the line-side efficiency of both the Aware 1.6T and Partially-Aware scenarios. The decrease in client-side efficiency happens because when flows span multiple transport boxes, when the number of destinations increases, new transport boxes are quickly created and new PHYs are deployed, leaving existing ones with available free capacity. This happens because PHYs can only carry flows with the same destinations, for this scenario. On the other hand, the Aware 1.6T compensates for its lack of line-side efficiency with high client efficiency. With large enough transport boxes, the client PHYs are exploited much more efficiently, since they can carry flows independently of their destination, contrary to lightpaths.

The Aware 0.4T is clearly the worst performer, since its transport box size is not enough to overcome the fact that FlexE groups cannot span multiple transport boxes.

Following the analysis of the multiple FlexE scenarios using transports with multiple 100 Gb/s, a new case using transport boxes with single 400Gb/s client interfaces is now considered. Router cards now have a capacity of 6.4Tb/s distributed over 16×400 GbE Ethernet interfaces.

In order to assess the effect of PHY bonding in the overall efficiency of the multiple scenarios, the obtained results are compared to the the ones obtained by using transport boxes with 4×100 GbE client interfaces, previously denoted as 0.4T. In this sense, the scenarios using transport boxes with 1×400 GbE client interfaces are referred to as 400G scenarios and the ones using transport boxes with 4×100 GbE client interfaces are referred to as 4×100G scenarios. The same approach previously described for the 4×100G scenarios was used in the following analysis.

The results of Figures 10 a) and b) show the amount of router cards required per Tb/s of carried traffic versus the number of destinations per node, for the different FlexE scenarios while the results of Figures 10 c) and d) show the amount of line ports required per Tb/s of carried traffic versus the number of destinations per node, for both the GBN and UBN, respectively.

With the changes made to the configuration of transport boxes, the amount of router cards required dropped significantly, since in the 400G scenarios, each client interface now corresponds to four individual interfaces in the 4×100G. However, the number of required router cards is not four times less than the 4×100G scenarios because client efficiency is not the same. The number of router cards required is consistently higher for fully meshed traffic, more noticeably, in the Unaware and Partially-Aware 400G scenarios. This is because client interfaces can only carry flows with the same destination. As the number of destinations per nodes increases, so does the number of required interfaces, hence the number of router cards is higher. Since the Aware scenarios don't have this limitation, they are more immune to the traffic pattern complexity.

For the GBN, the Unaware and Partially-Aware 400G scenarios have roughly the same performance. However, this changes for the UBN because the it is larger and path distances almost never allow for 200 Gb/s lightpaths. Therefore, the Unaware 400G scenario practically only uses 100 Gb/s lightpaths, since it downgrades the 150 Gb/s, which means that almost every transport box requires four line-ports. On the other hand, the Partially-Aware 400G scenario can limit the total capacity of the client interface. In this sense, almost no client interfaces have their full 400 Gb/s capacity available. This is why the Partially-Aware 400G scenario tends to require more client interfaces and, consequently, more router cards than the Unaware 400G scenario, for larger networks.

Regarding line port efficiency, the most striking result is the considerably large gap in the amount of line ports required in the Unaware 400G scenario compared to the 4×100G one. However, this result was

rather predictable since the number of line ports in a transport box depends on the lightpath rate. Therefore, four line ports are needed for 100 Gb/s lightpaths and two are needed for 200 Gb/s ones. Moreover, 150 Gb/s lightpaths are downgraded to 100 G/s, requiring even more line ports. This contrasts with the Unaware 4×100G scenario, which always uses transport boxes with two line ports. This gap between the two scenarios is even more noticeable with the UBN, since there are almost no 200 Gb/s lightpaths.

Both the Unaware and Partially-Aware 400G scenarios require an increasing amount of transport boxes for increasing traffic pattern complexity, compared to their respective 4×100G scenarios. This is due to the fact that transport boxes can no longer process multiple FlexE groups, since there is only one available client interface. Therefore, when the number of destinations increases, so does the amount of transport boxes.

Since the Aware scenarios can still process flows with different destinations in each transport box, the new 400G setting performs as well as the previous one. However, the Partially-Aware 400G scenario is the best performer of all in terms of required line ports, performing slightly better then the Aware scenarios. This happens because Aware transport boxes serve two different destinations, while Partially-Aware transport boxes serve one destination with two line ports, providing a more efficient use of the available R-T connection.

The results of Figures 11 a) and b) show the lightpath occupation ratio versus the number of destinations per node, for the different FlexE scenarios while the results of Figures 11 c) and d) show the PHY occupation ratio versus the number of destinations per node, for both the GBN and UBN, respectively.

The effect of having transport boxes serving one single destination is clearly noticeable here. Since, on both the Unaware and Partially-Aware 400G scenarios, all lightpaths in a new transport box must be deployed at once, all serving the same destination, when the number of destinations per node rises, this method is not effective, since there will be more lightpaths than needed for a single destination. Moreover, the Unaware 400G scenario behaves even worse for large networks because it needs to create even more lightpaths with smaller rates while Partially-Aware 400G transport boxes simply limit the client PHYs' capacity.

The new transport box setting does not affect the line-side efficiency of the Aware scenario, since it can still process multiple flows with different destinations in a single client interface, as before.

The results for the PHY occupation ratios clearly show the price to pay for using very large client interfaces. The fact that all 400G scenarios must create a single, large client PHY rather then creating smaller ones as needed, leads to an inevitable low client-side

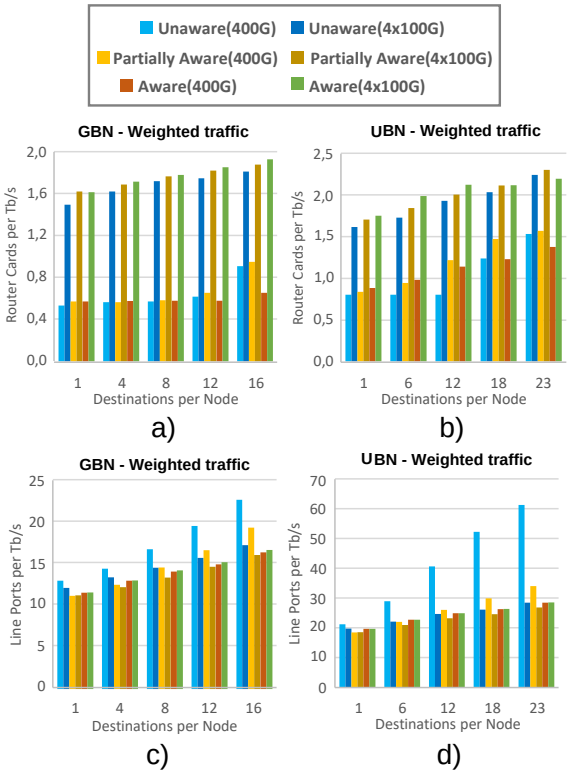


Figure 10: Simulation results for the 400G scenarios: a) Router cards per Tb/s, GBN; b) Router card per Tb/s, UBN; c) Line ports per Tb/s, GBN; d) Line ports per Tb/s, UBN.

efficiency, for the considered network load of 30 Tb/s.

The difference in efficiency between the Unaware and the Partially-Aware scenarios results from the PHY capacity limitation imposed by the latter, which is even more evident in a large network such as the UBN. This gap in client-side efficiency is even more striking for point-to-point traffic in large networks, since Unaware 400G transport boxes transparently map the client's full capacity rather than limiting client capacity.

## 5. Conclusions

This article presented an analysis on FlexE and its proposed implementation. The main focus was to apply FlexE to a DCI network scenario, as a possible solution for providing client flexibility.

Through the obtained results for the multiple considered cases it was possible to identify the advantages and disadvantages of each FlexE scenario as well as the type of application to which they are most suited for. Regarding the Unaware and Partially-Aware scenarios, the former is more efficient in terms of router port utilization, more evidently, for high network loads, while the latter provides a higher efficiency in terms of required lighpaths. This is due to its capability of discarding empty calendar slots at the transport boxes. On the other hand, the Unaware scenario performs

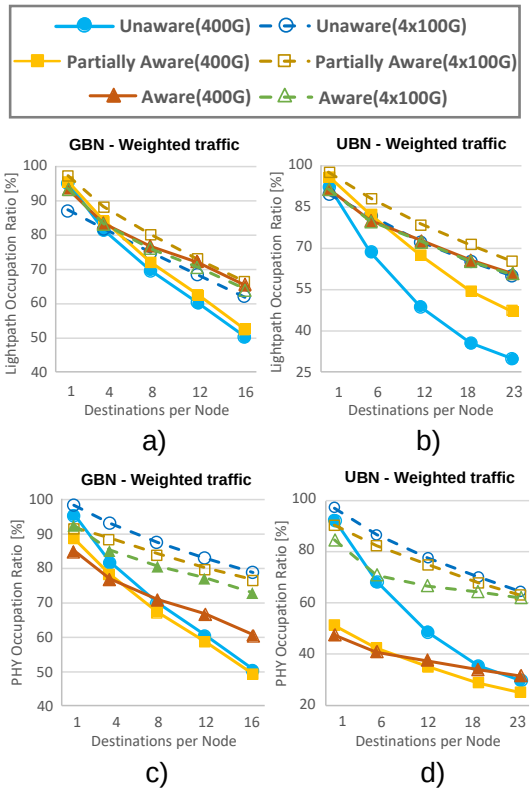


Figure 11: Simulation results for the 400G scenarios: a) PHY occupation ratio, GBN; b) PHY occupation ratio, UBN; c) Lightpath occupation ratio, GBN; d )Lightpath occupation ratio, UBN.

worse line-side since its transport boxes transparently map clients into lightpaths. The Aware scenario, by restricting FlexE groups to single R-T connections, provides a higher router card efficiency than the other scenarios. However, the use of such configuration is only beneficial when transport box capacity is the same as the router capacity. Smaller transport boxes impose a limitation in terms of provided line ports, hence, a limitation in the number of destinations a single transport box can serve. Moreover, the fact that large flows cannot span multiple transport boxes, in the Aware scenario, also limits its performance when using transport boxes with smaller capacity.

The effect of client interface size in the overall efficiency was also studied. The use of transport boxes with single client interfaces was compared with the previous multiple client interface setting, while maintaining the same overall capacity in order to assess the influence of client interface size in the multiple scenarios.

Overall, all scenarios benefit from using less client interfaces, in terms of used router cards. The Unaware and Partially-Aware scenarios are most influenced by this new transport box configuration, since PHYs can only carry flows with the same destinations. Therefore, they both turn out to have very poor client-side and line-side efficiency for more meshed traffic patterns and smaller traffic loads. The Unaware scenario provides a lower line-side efficiency than the Partially-Aware scenario because the transparent mapping of client PHYs into lightpaths results in a larger number of deployed lightpaths. However, this becomes an advantage for the Unaware scenario regarding client-side efficiency, since it doesn't limit the client's capacity, in contrast with the Partially-Aware scenario. In this sense, this new version of the Unaware scenario is better suited for point-to-point connections in terms of client-side efficiency while the Partially-Aware scenario proves to be a better option regarding line-side efficiency, for high network traffic loads. The Aware scenario's performance practically doesn't change, since it can still process single FlexE clients with different destinations.

## 6. Acknowledgements

The author would like to thank Prof. João Pires, from IST, and Eng. António Eira, from Coriant Portugal. This works was done in partnership with Coriant Portugal and was based on [4] by A. Eira, J. Pedro and J. Pires. Part of the presented analysis led to the writing of an article [5], which has been submitted to the Journal of Optical Communications and Networking (JOCN) and subjected to revision.

## References

[1] Optical dci market grows nearly 50% in 2016. Light Reading Network and Communications Industry, March 2017.

[2] Alcatel-Lucent. Data Center Interconnect Market Trends and Requirements, 2014.

[3] Cisco. Cisco Global Cloud Index: Forecast and Methodology, 20152020, 2016.

[4] António Eira and João Pedro. How Much Transport Grooming is Needed in the Age of Flexible Clients ? In *Optical Fiber Communications Conference*, 2017.

[5] António Eira, André Pereira, João Pires, and João Pedro. On the Efficiency of Flexible Client Architectures in Optical Transport Networks. *Journal of Optical Communications and Networking - Submitted*, 2017.

[6] Steven Gorshe. Otn interface standards for rates beyond 100 gbit/s. *Journal of Lightwave Technology*, 2017.

[7] Optical Internetworking Forum. Flex Ethernet Implementation Agreement. (March), 2016.

[8] Helen Xenos. What is flexethernet and why is it so important? Ciena, May 2016.