Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: April 23, 2017

I. Hussain, Ed.
R. Valiveti
K. Pithewan
Infinera Corp
Q. Wang, Ed.
ZTE
L. Andersson, Ed.
F. Zhang
M. Chen
J. Dong
Z. Du
Z. Haomian
X. Zhang
J. Huang
Q. Zhong
Huawei

October 20, 2016

GMPLS Routing and Signaling Framework for Flexible Ethernet (FlexE) draft-izh-ccamp-flexe-fwk-00

Abstract

Traditionally, Ethernet MAC rates were constrained to match the rates of the Ethernet PHY(s). OIF's implementation agreement [OIFMLG3] was the first step in allowing MAC rates to be different than the PHY rates. OIF has recently approved another implementation agreement [OIFFLEXE1] which allows complete decoupling of the MAC data rates and the Ethernet PHY(s) that support them. This includes support for (a) MAC rates which are greater than the rate of a single PHY (satisfied by bonding of multiple PHY(s)), (b) MAC rates which are less than the rate of a PHY (sub-rate), (c) support of multiple FlexE client signals carried over a single PHY, or over a collection of bonded PHY(s). The FlexE SHIM functions which bond multipe Ethernet PHY(s) to form a large "pipe" view the connectivity between two FlexE aware devices as a collection of multiple point-to-point links (one link per Ethernet PHY). These logical point-to-point links can either be direct links (without an intervening transport network), or realized via a Optical transport network. This draft catalogs the usecases that capture the FlexE deployment scenarios -- including the cases that include/exclude OTNs.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Copyright Notice

Table of Contents

1. Introduction 3
1.1. Requirements Language 4
2. Terminology 4
3. Usecases 5
3.1. FlexE unware transport 5
3.2. FlexE Aware 7
3.2.1. FlexE Aware Case - No Resizing 7
3.3. FlexE Termination - Transport 11
3.3.1. FlexE Client at Both endpoints 11
3.3.2. Interworking of FlexE Client w/ Native Client at the other endpoint 12
3.3.3. Interworking of FlexE client w/ Client from OIF_MLG . 14
3.3.4. Back-to-Back FlexE 15
3.3.4.1. FlexE Client BW Resizing 15
4. Requirements 16
5. Framework 17
6. Architecture 17
7. Solution 17
8. Acknowledgements 17

9. IANA Considerations 17
10. Security Considerations 17
11. References 17
11.1. Normative References 17
11.2. Informative References 18
Appendix A. Additional Stuff 18
Authors' Addresses 18

1. Introduction

Traditionally, Ethernet MAC rates were constrained to match the rates of the Ethernet PHY(s). OIF's implementation agreement [OIFMLG3] was the first step in allowing MAC rates to be different than the PHY rates standardized by IEEE. OIF has recently approved another implementation agreement [OIFFLEXE1] which allows complete decoupling of the MAC data rates and the Ethernet PHY(s) that support them. This includes support for (a) MAC rates which are greater than the rate of a single PHY (satisfied by bonding of multiple PHY(s)), (b) MAC rates which are less than the rate of a PHY (sub-rate), (c) support of multiple FlexE client signals carried over a single PHY, or over a collection of bonded PHY(s). The capabilities supported by the OIF FlexE implementation agreement version 1.0 are:

a. Support a large rate Ethernet MAC over bonded Ethernet PHYs, e.g. supporting a 200G MAC over 2 bonded 100GBASE-R PHY(s)

b. Support a sub-rate Ethernet MAC over a single Ethernet PHY, e.g. supportnig a 50G MAC over a 100GBASE-R PHY

c. Support a collection of flexible Ethernet clients over a single Ethernet PHY, e.g. supporting two MACs with the rates 25G, 50G over a single 100GBASE-R PHY

d. Support a sub-rate Ethernet MAC over bonded PHYs, e.g. supporting a 150G Ethernet client over 2 bonded 100GBASE-R PHY(s)

e. Support a collection of Ethernet MAC clients over bonded Ethernet PHYs, e.g. supporting a 50G, and 150G MAC over 2 bonded Ethernet PHY(s)

All networks which support the bonding of Ehernet interfaces (as per [OIFFLEXE1]) include a basic building block -- which consists of two FlexE SHIM functions (located at opposite ends of a link) and the (logical) point to point links that carry the Ethernet PHY signals between the two FlexE SHIM Functions. These logical point-to-point PHY links can be realized in a variety of ways:

a. These are direct point-to-point links with no intervening transport network.

b. The Ethernet PHY(s) are transparently transported via an Optical Transport Network. Optical Transport Networks (defined by [G709] and [G798]) have recently expanded the traditional bit (or codeword) transparent transport of Ethernet client signals, and included support for the usecases identified in the OIF FLexE implementation agreement.

c. Realized by tunneling the Ethernet PHY(s) over some other type of network (e.g. IP/MPLS). Thus, for example, the Ethernet PHY(s) signals could be carried over a pseudowire (or a LSP)in the IP/MPLS network. Note that the OIF implementation agreement [OIFFLEXE1] only includes support for 100G Ethernet PHY(s). As a result of this encapsulation into a PW, the bandwidth of the PW will be much larger than the bit rate of the Ethernet PHY (i.e. 100G), and such a pseudowire cannot be transported in networks that only include 100G Ethernet links. This scenario is realizable when (a) higher rate Ethernet PHY(s), e.g. 200G/40G are supported) or (b) OIF extends the FlexE groups to include lower rate Ethernet PHY(s), e.g. at the 25G/50G rate. Further study is needed to ensure that these scenarios are realizable, practical, and beneficial to operators. With this in mind, the current draft doesn’t include any coverage for this scenario.

Internet-draft examines the usescases that arise when the logical links between FlexE capable devices are (a) point-to-point links without any intervening network (b) realized via Optical transport networks. This draft considers the variants in which fhe two peer FlexE devices are both customer-edge devices, or customer-edge/provider edge devices. This list of usecases will help identify the Control Plane (i.e. Routing and Signaling) extensions that may be required).

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Terminology

a. Ethernet PHY: an entity representing 100G-R Physical Coding Sublayer (PCS), Physical Media Attachment (PMA), and Physical Media Dependent (PMD) layers.

b. FlexE Group: A FlexE Group is composed of from 1 to n 100GBASE-R Ethernet PHYs. Each PHY is identified by a number in the range [1-254].

c. FlexE Client: an Ethernet flow based on a MAC data rate that may or may not correspond to any Ethernet PHY rate (e.g., 10, 40, m x 25 Gb/s).

d. FlexE Shim: the layer that maps or demaps the FlexE clients carried over a FlexE group.

e. FlexE Calendar: The total capacity of a FlexE group is represented as a collection of slots which have a granularty of 5G. The calendar for a FlexE group composed of n 100G PHYs is represented as an array of 20n slots (each representing 5G of bandwidth). This calendar is partitioned into sub-calendars, with 20 slots per 100G PHY. Each FlexE client is mapped into one or more calendar slots (based on the bandwidth of the FlexE client).

3. Usecases

3.1. FlexE unware transport

The FlexE shim layer in a router maps the FlexE client(s) over the FlexE group. The transport network is unware of the FlexE. Each of the FlexE group PHY is carried independently across the transport network over the same fiber route. The FlexE shim in the router tolerates end-to-end skew across the network. In this usecase, the router makes flexible use of the full capacity of the FlexE group, and depends on legacy transport equipment to realize PCS-codeword-transparent transport of 100GbE. It allows striping of PHYs in the FlexE group over multiple line cards in the transport equipment. It is worth mentioning that in this case, the FlexE SHIM layer is terminated at the routers, and the coordination of operations related to FlexE clients, e.g. creating new FlexE clients, deleting existing FlexE clients, and resizing the bandwidth of existing FlexE clients (if desired) happens between the two routers. Note that the transport network is completely transparent to the FlexE signals, and doesn’t participate in any FlexE protocols.

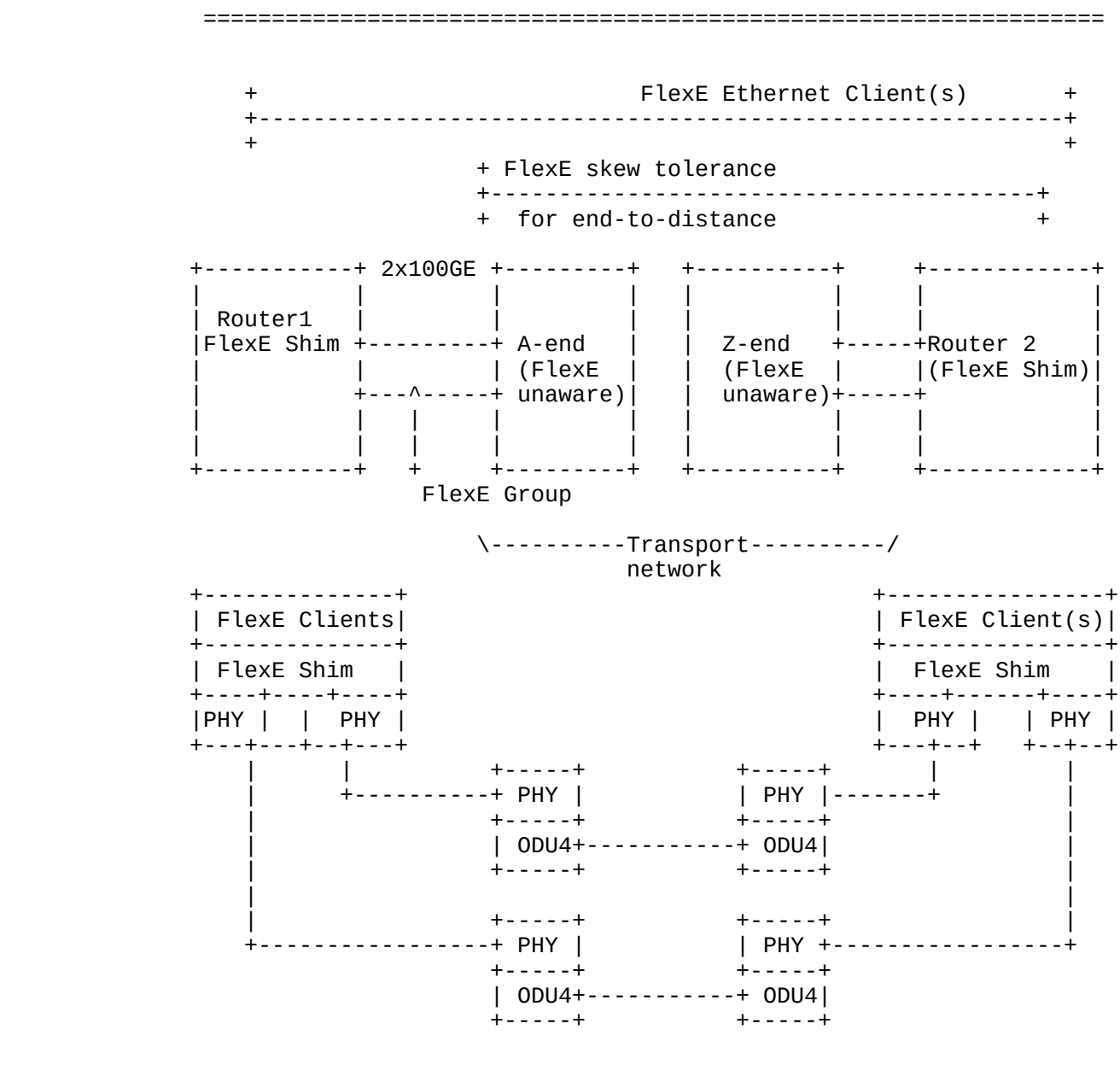

Figure 1: FlexE unaware transport

3.2. FlexE Aware

3.2.1. FlexE Aware Case - No Resizing

This scenario represents an optimization of the FlexE unaware transport presented in Section 3.1, and illustrated in Figure 1. In this application (see Figure 2), the devices at the edge of the transport network do not terminate the FlexE shim layer, but are aware of the (a) composition of the FlexE grpup (i.e. set of all contained Ethernet PHYs) and (b) format of the FlexE overhead. They "snoop" the FlexE overhead to determine the subset of the set of all calendar slots that are available for use (i.e. these calendar slots may be used, or unused). The transport network edge removes the unavailable calendar slots at the ingress to the network, and adds the same unavailable calendar slots back when exiting the network. The result is that the FlexE Shim layers at both routers see exactly the same input that they saw in the FlexE unware scenario -- with the added benefit that the line (or DWDM) side bandwidth has been optimized to be sufficient to carry only the available calendar slots in all of the Ethernet PHY(s) in the FlexE group. This mode may be used in cases where the bandwidth of the Ethernet PHY is greater than the bit rate supported by a wavelength (and it is known that that all calendar slots in the PHY are not "available").

The transport network edge device could learn of the set of unavailable calendar slots in a variety of ways; a few examples are listed below:

a. The set of unavailable calendar slots could be configured against each Ethernet PHY in the FlexE group. The FlexE demux function in the transport network edge device (A) compares the information about calendar slots which are expected to be unavailable (as per user supplied configuration), with the corresponding information encoded by the customer edge device in the FlexE overhead (as specified in [OIFFLEXE1]). If there is a mismatch between the unavailable calendar slots in any of the PHYs within a FlexE group, the transport edge node software could raise an alarm to report the inconsistency between the provisioning information at the transport network edge, and the customer edge device.

b. The Transport network edge could be configured to act in a "slave" mode. In this mode, the FlexE demux function at the Transport network edge (A) receives the information about the available/unavailable calendar slots by observing the FlexE overhead (as specified in [OIFFLEXE1]) and uses this information to select (a) the set of wavelengths (with appropriate capacities) or (b) the bandwidth of the ODUflex (or fixed rate ODUs) that could carry the FlexE PCS end-to-end.

c. The set of unavailable slots could be negotiated between FlexE Shim entity in the customer device and the partial rate ODUflex mapper located in the transport network element. Thus, for example, the transport network element could declare the maximum number of 5G slots that could be transported over a single wavelength, and the customer network device can choose the number of 5G slots that will be used between customer devices. This process could be accomplished through control protocols such as LMP,using the appropriate control channel for transporting the messages.

In the basic FlexE aware mode, the transport network edge does not expect the number of unavailable calendar slots to change dynamically.

Note that the process of removing unavailable calendar slots from a FlexE PHY is called "crunching" (see [OIFFLEXE1]). The following additional notes apply to Figure 2:

a. The crunched FlexE PHYs are independently transported through the transport network. The number of used (and unused) calendar slots can be different across the FlexE group. In particular, if all the calendar slots in a FlexE PHY are in use, the crunching operation leaves the original signal intact.

b. In this illustration, the different FlexE PHY(s) are transported using ODUflex containers in the transport network. These ODUflex connections can be of different rates.

c. In the most general form, G.709 Section 17.12 allows for a FlexE group consisting of m Ethernet PHY(s) to be crunched, combined, and transported using n ODUFlex containers (where n can range between 1 and m). In other words, the ITU G.709 recommendation allows for (but not require the support for) the degenerate cases in which (a) each Ethernet PHY within the group is transported using its own ODUflex, and (b) all the PHY(s) are crunched, combined and transported over a single ODUflex container. If all the sub-calendar slots in a given PHY are available, it is possible to transport the content of the PHY in one of two ways: (a) as shown in Figure 2, or (b) using a FLexE unware (i.e. PCS-codeword transparent transport) mode. The latter approach (of using FlexE unaware transport) for a few select (fully-utilized) PHYs is not attractive from the perspective of skew between the PHYs that comprise the FlexE group. For simplicity, the preferred mode of operation will be one in which the same mapping procedure is used for member PHYs of a FlexE group.

d. When the crunched FlexE PHY(s) have a rate that is identical to that of a standard Ethernet PHY, it is possible that the transport network may utilize standard ODU containers such as ODU2e, ODU4 etc. As currently defined by ITU G.709 Section 17.12, the crunched, sub-rate signal is always mapped to an ODUflex, and the mapping to a fixed rate ODU signal is not required. This option could be dropped if it results in any significant simplification.

Note: The figure may need further editing to accurately depict the signal hierarchy.

==========

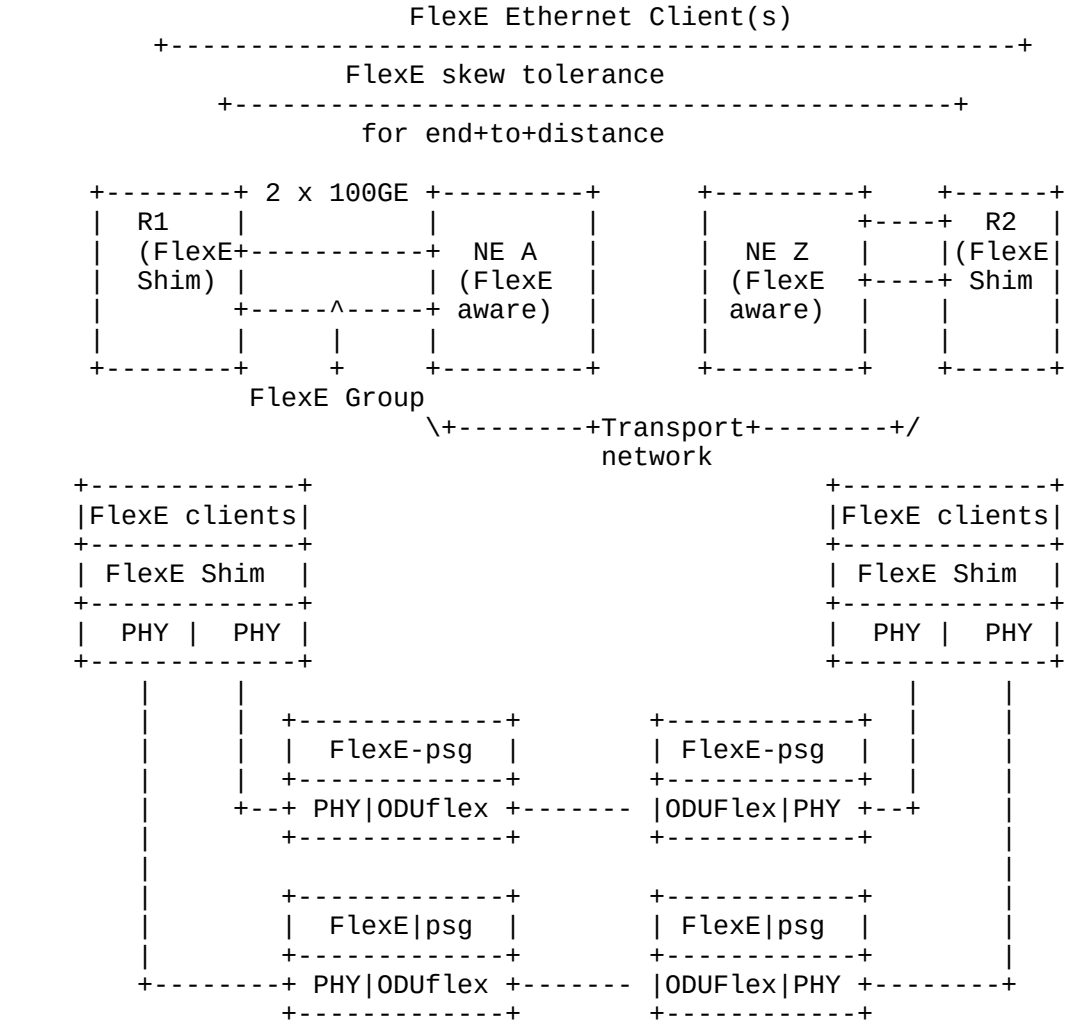

Figure 2: FlexE Aware Transport

3.3. FlexE Termination - Transport

These usecases build upon the basic router-transport equipment connectivity illustrated in Figure 1. The FlexE shim layer at the router maps to the set of FlexE clients over the FlexE group, as usual. This section considers various usecases in which the equipment located at the edge of the transport network instantiates the FlexE Shim function which peers with the FlexE shim on the customer device. In the router to network direction, the transport edge node terminates the FlexE shim layer, and extracts one or more FlexE client signals, and transports them through the network. That is, these usecases are distinguished from the FlexE unaware cases in that the FlexE group, and the FlexE shim layer end at the transport network edge, and only the extracted FlexE client signals transit the optical network. In the network to router direction, the transport edge node maps a set of FlexE clients to the FlexE group (i.e. performing the same functions as the router which connects to the transport network).The various usecases differ in the combination of service endpoints in the transport network. In the FlexE termination scenarios, the distance between the FlexE Shims is limited the normal Ethernet link distance. The FlexE shims in the router, and the equipment need to support a small amount skew.

3.3.1. FlexE Client at Both endpoints

In this scenario, service consists of transporting a FlexE client through the transport network, and possibly combining this FlexE client with other FlexE clients into a FlexE group at the endpoints. The FlexE client signal can be transported in two manners within the OTN: (i) directly over one or more wavelengths (ii) mapped into an ODUflex (of the appropriate rate) and then switched across the OTN. Figure 3 illustrates the scenario involving the mapping of a FlexE client to an ODUflex envelope; this figure only shows the signal "stack" at the service endpoints, and doesn’t illustrate the switching of the ODUflex entity through the OTN. The ODUflex mapping will be beneficial in scenarios where the rate of the FlexE client is less than the capacity of a single wavelength deployed on the DWDM side of the OTN network, and allows the network operators to packet multiple FlexE client signals into the same wavelength -- thereby improving the network efficiency. Although Figure 3 illustrates the scenario in which one FlexE client is transported within the OTN, the following points should be noted:

a. When the FlexE Shim termination function recovers multiple FlexE client signals (at node A), the FlexE signals can be transported independently. In other words, it is not a requirement that all the FlexE client signals be co-routed.

b. Conversely, at the egress node, FlexE clients from different endpoints can be combined via the FlexE shim, eventually exiting the transport edge node over an Ethernet group.

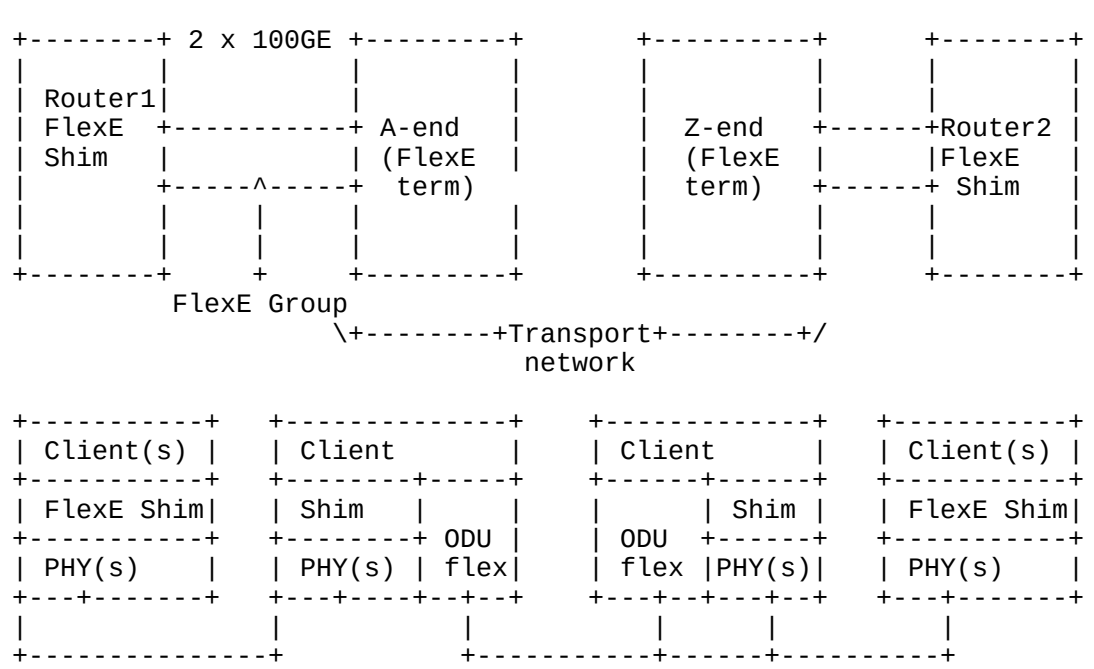

Figure 3: FlexE termination: FlexE clients at both endpoints

3.3.2. Interworking of FlexE Client w/ Native Client at the other endpoint

The OIF implementation agreement [OIFFLEXE1] currently supports FlexE client signals carried over one or more 100GBASE-R PHY(s). There is a calendar of 5G timeslots associated with each PHY, and each FlexE client can make use of a number of timeslots (possibly distributed across the members of the FlexE group. This implies that the FlexE client rates are multiples of 5Gbps. When the rates of the FlexE client signals matches the MAC rates corresponding to existing Ethernet PHYs, i.e. 10GBASE-R/40GBASE-R/100GBASE-R, there is a need for the FlexE client signal to interwork with the native Ethernet client received from a single (non-FlexE capable) Ethernet PHY. This

capability is expected to be extended to any future Ethernet PHY rates that the IEEE may define in future (e.g. 25G, 50G, 200G etc.). In these cases, although the bit rate of the FlexE client matches the MAC rate of other endpoint, the 64B66B PCS codewords for the FlexE client need to be transformed (via ordered set translation) to match the specification for the specific Ethernt PHY. These details are described in Section 7.2.2 of [OIFFLEXE1] and are not eloborated any further in this document.

Figure 4 illustrates a scenario involving the interworking of a 10G FlexE client with a 10GBASE-R native Ethernet signal. In this example, the network wrapper is ODU2e.

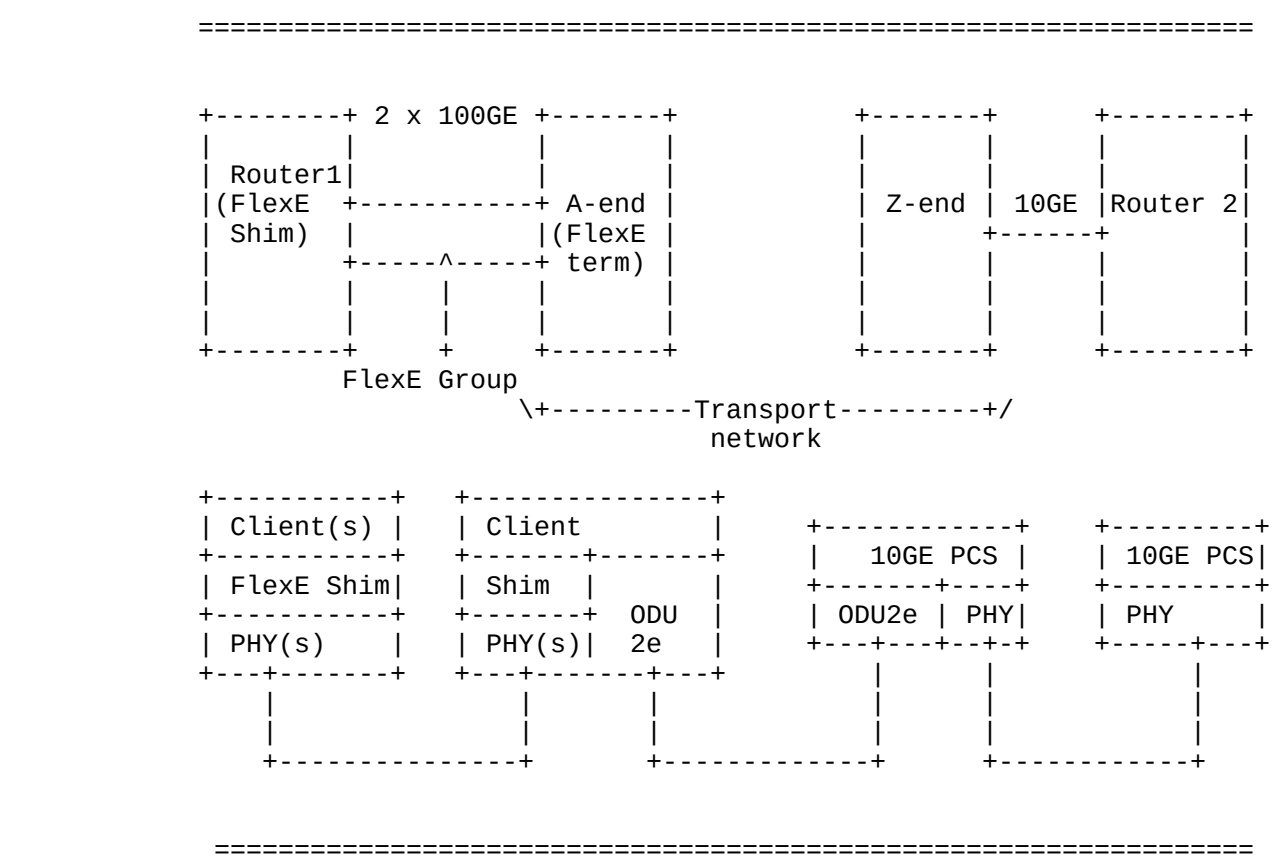

Figure 4: FlexE client interop with Native Ethernet Client

3.3.3. Interworking of FlexE client w/ Client from OIF_MLG

As explained in the Introduction section (Section 1 OIFMLG3 [OIFMLG3] introduced support for carrying 10GE and 40GE client signals over a group of 100GBASE-R Ethernet PHY(s). While the most recent implementation agreement doesn't call it out explicitly, it is expected that the FlexE clients (as defined in [OIFFLEXE1]), and 10GBASE-R/40GBASE-R clients supported by OIFMLG3 [OIFMLG3]) will interoperate.

Figure 5 illustrates a scenario involving the interworking of a 10G FlexE client with a 10GBASE-R client supported by an OIFMLG3 interface. In this example, the network wrapper is ODU2e.

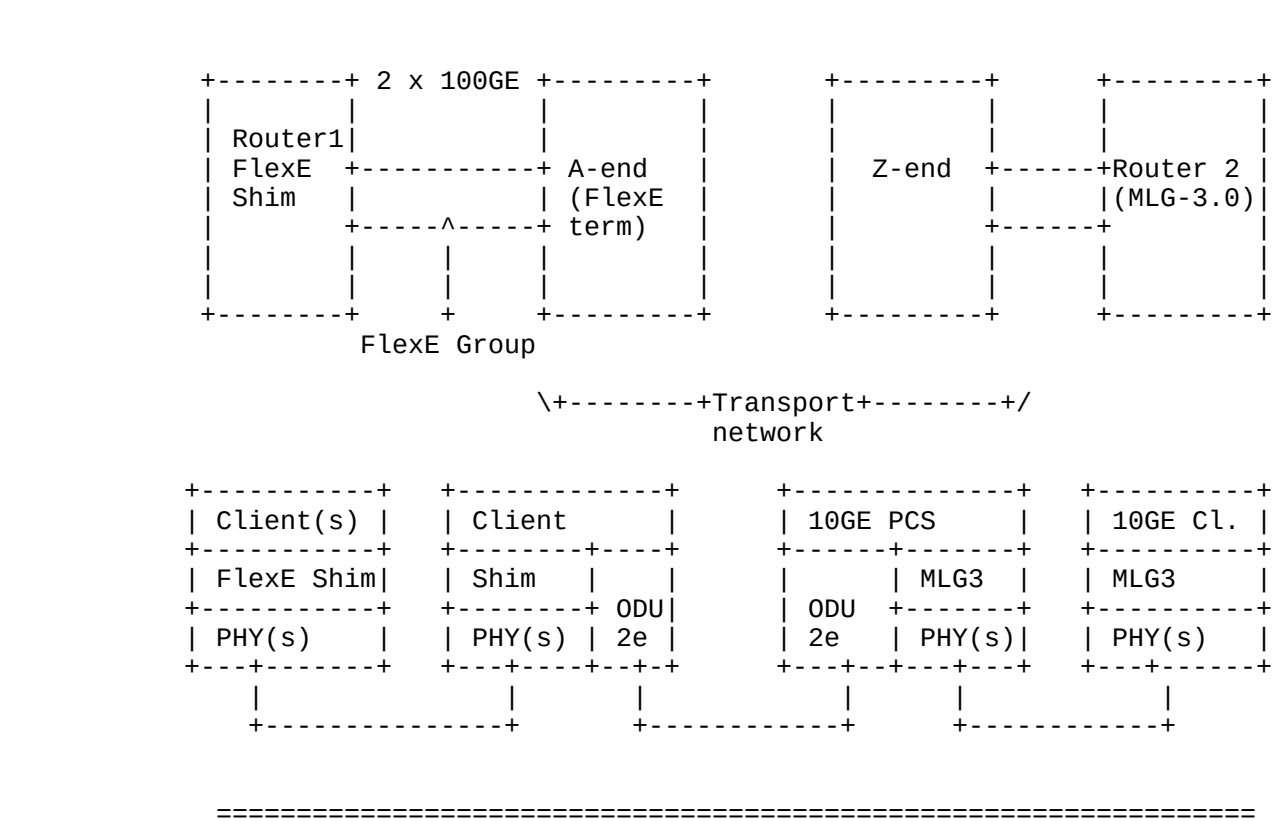

Figure 5: FlexE client interop with Ethernet Client supported by MLG3

3.3.4. Back-to-Back FlexE

This section covers a degenerate FlexE termination scenario in which router1, router2, and router3 are interconnected through back-to-back FlexE groups without an intermediate transport network (see Figure 6). In this example, the FlexE SHIM at Router2 extracts one or more FlexE client signals from the FlexE group connected to Router1, and mutliplexes these extracted FlexE signals into the FlexE group towards the appropriate router (e.g. Router3). Note that each of the extracted FlexE client signals can be indepdenently routed towards its respective FlexE group.

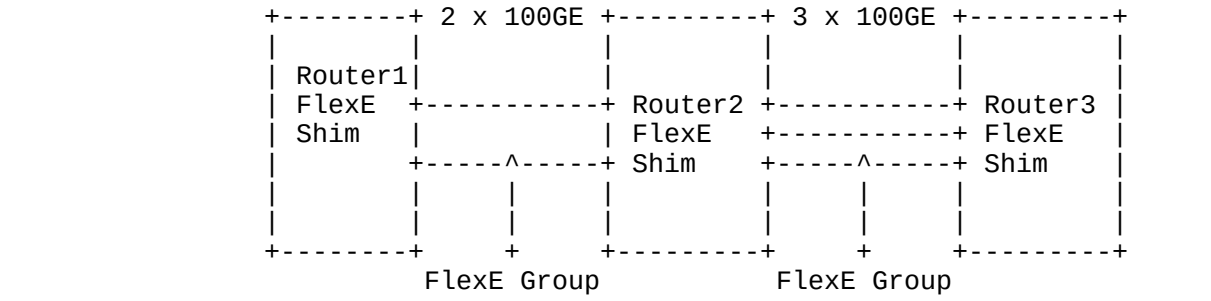

Figure 6: Back-to-Back FlexE

3.3.4.1. FlexE Client BW Resizing

In the scenario presented in Figure 6, it is possible to support the FlexE client signal resizing on an end-to-end basis. Thus, for example, the resizing of the end-to-end FlexE client circuit with a scope of Router1-Router2-Router3 is accomplished by correctly coordinating the resizing operations across these two segments: Router1-Router2, Router2-Router3. The hop-by-hop FlexE client signal resizing operations across each of these segments (or hops) are accomplished by using the following FlexE overhead (as per [OIFFLEXE1]):

a. Currently active FlexE calendar (containing a list of mapping between the 5G tributary slots and the FlexE client signals

b. Future calendar to which the sender wants to transition to.

c. Calendar switch request bit (CR)

d. Calendar switch acknowldege bit (CA)

It is expected that the exact sequence of FlexE client resizing operations will be different for the cases involving bandwidth increase/decrease.

4. Requirements

This section summarizes solution requirements for the usecases described in this document to help identify the Control Plane (i.e. Routing and Signaling) extensions that may be required.

a. The solution SHALL support a FlexE group to address abovementioned usecases including FlexE unaware (where FlexE mux and demux can be separated by longer distances), FlexE aware (where FlexE mux and demux can be separated by shorter distances), and FlexE partially aware.

b. The solution SHALL support a flexible mechanism for configuring a FlexE group -- such as a signaling protocol or a SDN controller/ management system with network access to the FlexE mux/demux at each end of the FlexE group.

c. The solution SHOULD support the ability to add/remove Ethernet PHYs to/from a FlexE group. In the absence of this ability, it is acceptable to permit changes to the group members only when the group has been administratively locked (and hence not providing any service).

d. The solution SHOULD allow decoupling of FlexE group’s initial configuration and bring up operation from an addition (or removal) of FlexE clients to the FlexE group. For instance, it SHOULD be possible to configure and bring up a FlexE group without any FlexE client (e.g., with all calendar slots set to unused or unavailable).

e. The solution SHALL allow adding or removing a FlexE client to a FlexE group without affecting traffic on other clients.

f. The solution SHOULD allow resizing of FlexE client BW through coordination of calendar updates within a single FlexE group. There SHOULD be no expectation that FlexE client BW resizing be hitless in all network scenarios. This capability can be supported for the Back-to-Back FlexE scenario identified in Section 3.3.4.1

g. For the FlexE unaware case, each of the 100GBASE-R PHYs in the FlexE group SHALL be carried independently across transport network using a PCS codeword transparent mapping. All PHYs of the FlexE group SHALL be interconnected between the same two FlexE shims. The Ethernet PHYs SHOULD be carried over the same fiber route across the transport network (i.e., co-routed)

h. For the FlexE aware case, each of the 100GBASE-R PHYs in the FlexE group SHALL be carried independently across transport network. All PHYs of the FlexE group SHALL be interconnected between the same two FlexE shims. The Ethernet PHYs SHOULD be carried over the same fiber route across the transport network. In the transport network, in mux direction, the OTN mapper SHALL be able to discard unavailable slots (e.g., this can be based on static configuration as the rate of a wavelength is not expected to change in-service). In the transport network, in the demux direction, the OTN mapper SHALL be able to restore unavailable slots to match the original PHY rate.

i. For the FlexE termination case, the FlexE group SHALL be terminated at the transport network edge. It SHOULD be possible to carry (switch) each FlexE client extracted from the FlexE group independently across transport network using OTN mapping (e.g., ODUflex).

5. Framework

6. Architecture

7. Solution

8. Acknowledgements

9. IANA Considerations

This memo includes no request to IANA.

10. Security Considerations

None.

11. References

11.1. Normative References

[G709] ITU, "Optical Transport Network Interfaces (http://www.itu.int/rec/T-REC-G.709-201606-P/en)", July 2016.

[G798] ITU, "Characteristics of optical transport network hierarchy equipment functional blocks (http://www.itu.int/rec/T-REC-G.798-201212-I/en)", February 2014.

[OIFFLEXE1] OIF, "FLex Ethernet Implementation Agreement Version 1.0 (OIF-FLEXE-01.0)", March 2016.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

11.2. Informative References

[OIFMLG3] OIF, "Multi-Lane Gearbox Implementation Agreement Version 3.0 (OIF-MLG-3.0)", April 2016.

Appendix A. Additional Stuff

This becomes an Appendix.

Authors' Addresses

Iftekhar Hussain (editor)
Infinera Corp
169 Java Drive
Sunnyvale, CA 94089
USA

Email: IHussain@infinera.com

Radha Valiveti
Infinera Corp
169 Java Drive
Sunnyvale, CA 94089
USA

Email: rvaliveti@infinera.com

Khuzema Pithewan
Infinera Corp
169 Java Drive
Sunnyvale, CA 94089
USA

Email: kpithewan@infinera.com

Qilei Wang (editor)
ZTE
Nanjing
CN

Email: wang.qilei@zte.com.cn

Loa Andersson (editor)
Huawei
Stockholm
Sweden

Email: loa@pi.nu

Fatai Zhang
Huawei
CN

Email: zhangfatai@huawei.com

Mach Chen
Huawei
CN

Email: mach.chen@huawei.com

Jie Dong
Huawei
CN

Email: jie.dong@huawei.com

Zongpeng Du
Huawei
CN

Email: duzongpeng@huawei.com

Zheng Haomian
Huawei
CN

Email: zhenghaomian@huawei.com

Xian Zhang
Huawei
CN

Email: zhang.xian@huawei.com

James Huang
Huawei
CN

Email: james.huang@huawei.com

Qiwen Zhong
Huawei
CN

Email: zhongqiwen@huawei.com