FlexE Neighbor Discovery Implementation Agreement

IA # OIF-FLEXE-ND-01.0

September 12, 2018

Implementation Agreement created and approved
by the Optical Internetworking Forum
www.oiforum.com

The OIF is an international non-profit organization with over 100 member companies, including the world's leading carriers and vendors. Being an industry group uniting representatives of the data and optical worlds, OIF's purpose is to accelerate the deployment of interoperable, cost-effective and robust optical internetworks and their associated technologies. Optical internetworks are data networks composed of routers and data switches interconnected by optical networking elements.

With the goal of promoting worldwide compatibility of optical internetworking products, the OIF actively supports and extends the work of national and international standards bodies. Working relationships or formal liaisons have been established with CFP-MSA, COAST, Ethernet Alliance, Fibre Channel T11, IEEE 802.1, IEEE 802.3, IETF, InfiniBand, ITU-T SG13, ITU-T SG15, MEF, ONF, Rapid I/O, SAS T10, SFF Committee, TMF and TMOC.

For additional information contact:
The Optical Internetworking Forum,
5177 Brandin Ct.
Fremont, CA 94538
510-492-4040 Φ info@oiforum.com

www.oiforum.com

Working Groups: Physical and Link Layer WG
Networking and Operations WG

TITLE: **FlexE Neighbor Discovery Implementation Agreement 1.0**

SOURCE:

TECHNICAL EDITORS

Qichang Chen, Ph. D.
Huawei Technologies Co., Ltd.
Phone: +86 755 28976940
Email: chenqichang1@huawei.com

Qiwen Zhong
Huawei Technologies Co., Ltd.
Phone: +86 755 28976962
Email: zhongqiwen@huawei.com

Tad Hofmeister, Ph.D.
Google
Email: tad@google.com

WORKING GROUP CHAIRS

David R. Stauffer, Ph.D.
Kandou Bus, S.A.
EPFL Innovation Park Bldg. I
1015 Lausanne Switzerland
Phone: +1 802 316-0808
Email: david@kandou.com

David Ofelt, Ph.D.
Juniper Networks
Phone: +1.408.745.2945
Email: ofelt@juniper.net

Jonathan Sadler
Coriant
Phone: +1.630.798.6182
Email: jonathan.sadler@coriant.com

ABSTRACT: This Implementation Agreement specifies the OIF organization specific type, length and value (TLV) extension to enable 802.1ab Link Layer Discovery Protocol (LLDP) for FlexE neighbor discovery which enables remote FlexE PHY capability and deskew capability discovery, PHY connectivity discovery and verifications, and FlexE Group subgroup integrity verification.

1 Table of Contents

2 List of Figures

3 List of Tables

4 Document Revision History and Input Contributions

Issue No.	Issue Date	Details of Change
oif2017.376.00	07/2017	Initial draft
oif2017.376.01	10/2017	2017 Q4 meeting in Shanghai - significant overhaul of the content arrangements and use cases;
oif2017.376.02	10/2017	● Added considerations for the dual-mode switching process in case of PHY down/restart and other typical switching scenarios; ● Incorporated several future work items;
oif2017.376.03	10/2017	Minor non-technical editorial changes;
oif2017.376.04	04/2018	2018 Q1 meeting in San Antonio – ● Introduced the deskew capability TLV; ● Minor non-technical text changes;
oif2017.376.05	04/2018	2018 Q2 meeting in Nuremburg – ● Incorporated suggested editorial changes from Microsemi; ● Incorporated detailed LLDPDU encapsulation process suggested from Ciena as Appendix A;
oif2017.376.06	07/2018	2018 Q3 meeting in Vancouver – Include 1st straw ballot comments resolution. Update document header & reference as clean version. Final version for principal member company ballot;

5 Introduction and Requirements

FlexE Neighbor Discovery is intended for remote FlexE PHY capability and deskew capability discovery, PHY connectivity discovery and verifications, and FlexE Group subgroup integrity verification.

5.1 FlexE Connectivity Discovery (FECD)

With the physical links set up between the PHYs of different FlexE nodes, it is very important for a node's control plane/centralized SDN controller to find out certain characteristics of such links between the PHYs of a local node (Local System) and the remote nodes (Remote Systems). It would be helpful for the system administrator/a SDN controller to create and initialize FlexE Groups out of a given set of PHYs for any given FlexE nodes based on the following discovered connectivity information:
(1) Identification of the remote nodes' PHYs which are connected to the local node's PHYs;
(2) Number of the individual PHY connections between a local node and a remote node.

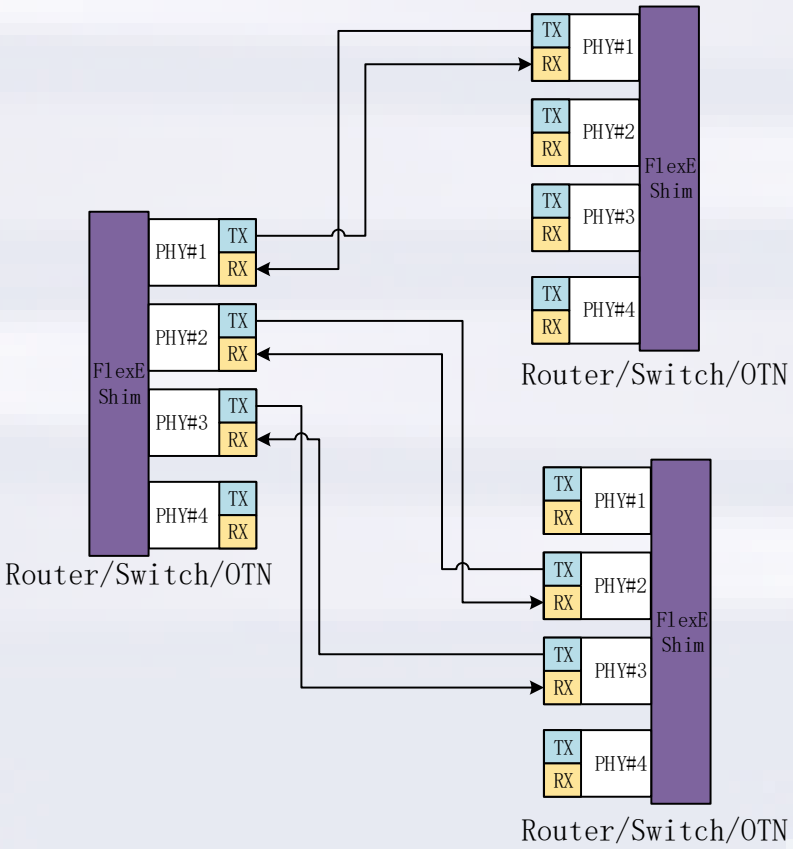

Figure 1 FlexE device connection example

In addition, it is useful to find out the following scenarios:

- One PHY is looped back to itself;
- Any two PHYs on the local node are linked to each other;

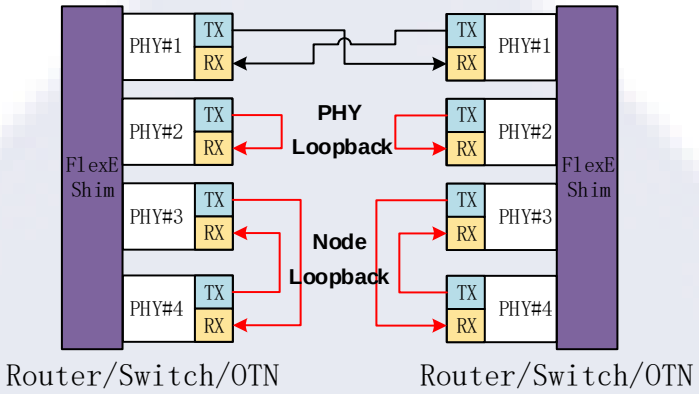

Figure 2 PHY Loopback and Node Loopback Example

5.2 FlexE Connectivity Verification (FECV)

Once the PHY connections between the local and remote nodes are discovered, the local control plane or a SDN controller needs to verify whether the actual connections match the intended logical connections between the configured FlexE Groups in the local and remote nodes in order for the FlexE Group to enter “active” state. There is a connectivity verification requirement for locating the mismatch between the physical connectivity of the PHYs which are intended to form a FlexE Group and the planned logical connectivity of the PHY configuration of the intended FlexE Group if the group is not initialized.

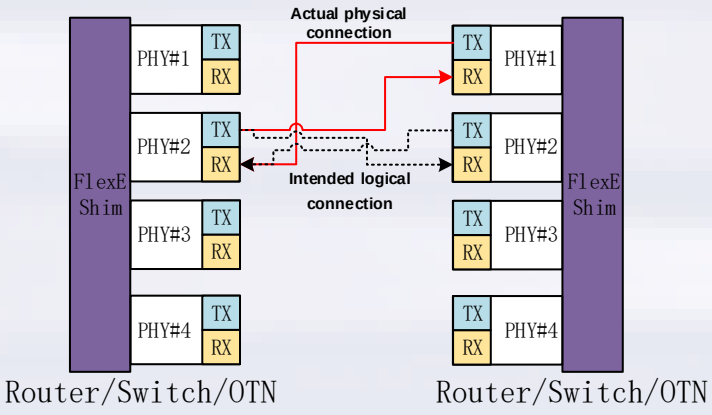

Figure 3 Intended and Actual PHY Connectivity Verification Case

5.3 FlexE Remote Capability Discovery (FERCD)

When there exists more than one physical links between two adjacent FlexE nodes or a FlexE node is connected to multiple FlexE-aware nodes via different links, it is necessary for a node’s control plane/SDN controller to discover the capability of the remote

system/node before deciding how those linked PHYs could and should be configured to become PHYs of a FlexE Group.

With a centralized SDN controller, the controller can learn about the FlexE capability of each node respectively in the network and there is no need for FlexE Remote Capability Discovery requirement. However, this IA focuses on FlexE Remote Capability Discovery (FERCD) where the nodes on both sides of the physical PHY link(s) have only local control planes.

5.4 FlexE Subgroup Integrity Verification (FESIV)

In the FlexE-aware transport scenario, a FlexE Group could be transported by two or more independent transport network. In such scenarios, the FlexE PHYs on the FlexE-aware transport equipment is connected to a subset of PHYs. Performance monitor tailored to a FlexE sub-group is desirable.

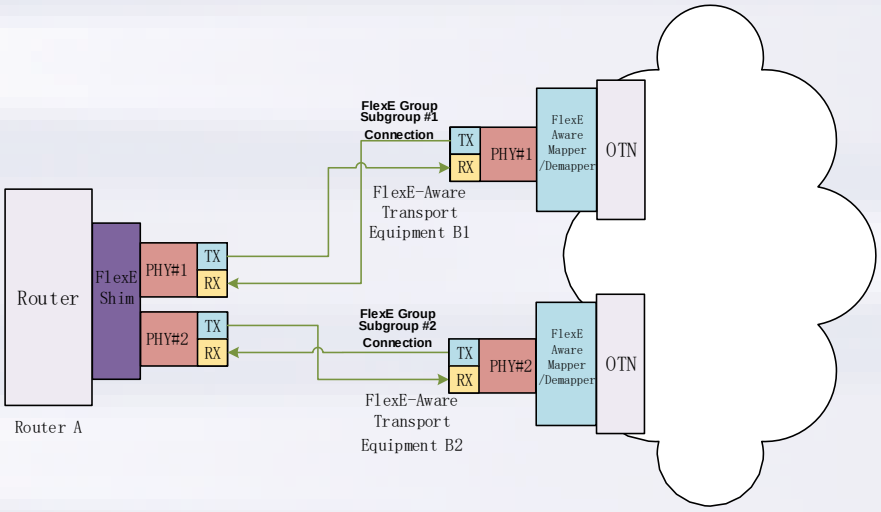

Figure 4 FlexE Subgroup Integrity Verification case with two FlexE subgroups

This IA defines the notion of FlexE Group Subgroup which represents a part of the FlexE Group that is carried by one independent transport network in the case of multiple transport segments carrying one FlexE Group. It is important for both ends of the physical links to discover and verify the FlexE Group Subgroup via the neighbor discovery process. With the presence of FlexE Subgroup Integrity Verification (FESIV), the PHY maps can be monitored for mismatch on a sub-group basis.

5.5 FlexE Deskew Capability Discovery (FEDCD)

For a FlexE Group consisting of multiple bonded PHYs, the demux shall be able to deskew multiple FlexE PHYs within its permissible skew tolerance. A FlexE Group would fail to operate properly if the skew between its member PHYs is beyond the demux's deskew capability.

Although FlexE IA specifies 300ns low skew tolerance and suggests as high as 10μs high skew tolerance, many FlexE vendor implementations could come with different skew tolerances. Therefore, there could be varying skew tolerance capabilities.

1. No PHY bonding capability and therefore no skew tolerance is provided.
2. Only 300ns low skew tolerance and therefore limited to short-reach applications.
3. 300ns low skew tolerance and application-specific high skew tolerance ranging from 300ns to 10μs which is intended for medium-distance or long-haul applications.

It is useful to verify that both ends of a FlexE Group link have same/similar deskew capabilities and their deskew capabilities match the requirements of the intended application. For example, it is inappropriate to put a FlexE device with the above-mentioned capability 2(only 300ns low skew tolerance) into a long-haul DCI application where a FlexE device with the above-mentioned capability 3 is needed. For two FlexE devices both with capability 3 used in a long-haul application, it is helpful to find out the exact deskew capability of both ends for the FlexE Group link to make sure the skew tolerances on both ends are adequate with regard to the link's length.

5.6 Summary of Requirements and Specifications

The FlexE Neighbor Discovery (FlexE ND) Implementation Agreement is intended to provide mechanisms to facilitate the setup of FlexE Group. Specifically, this IA defines

1) Connectivity discovery mechanism for a node to identify the links with its neighbor nodes and the loop links.
2) Connectivity verification mechanism for a configured FlexE Group for locating the inconsistency between the physical links and the intended logical connection configuration of the FlexE Group.
3) Remote capability discovery mechanism for a node to learn about FlexE capability of the remote nodes and identify the PHYs which could be bonded to form a FlexE Group.
4) FlexE Group subgroup integrity verification mechanism for a FlexE-aware transport node to verify the FlexE subgroup connectivity in the case of FlexE over multiple transport segments.
5) FlexE PHY deskew capability discovery for verifying that the FlexE device's deskew tolerance is sufficient for the intended application.

Furthermore, this IA defines the message format and advertising mechanism of OIF FlexE specific Link Layer Discovery Protocol (LLDP) type, length and value (TLV) fields as OIF extension to 802.1ab LLDP.

6 Overview of FlexE Neighbor Discovery

6.1 Overview of Protocol Stack

For a FlexE Group comprised of n 100G/200G/400G PHYs, there are two optional management channels per PHY, the section management channel which is available for communication between two FlexE section terminating points, and the shim to shim management channel, which is available for communication between two FlexE Shim terminating points. The message over management channels shall be 64B/66B encoded per [802.3] clause 82.2.3. The management channel can be abbreviated as MC and we will use the term MC referring to the management channel in subsequent text in this IA.

LLDP has been widely in use for connectivity discovery in traditional Ethernet (802.3 std Ethernet with full BW MAC) network. The FlexE neighbor discovery enhanced LLDP runs over the section management channel in the FlexE overhead frame per PHY. The LLDP based FlexE neighbor discovery protocol stack is illustrated in figure 5. The section management channels are abbreviated to MC-s. Each PHY's section management channel offers a bandwidth of 1.2Mbps to carry LLDPDUs. The LLDPDUs are MAC encapsulated per 802.1AB-2015 and 64B/66B block encoded per [802.3] clause 82.2.3.All LLDPDUs shall use the "nearest bridge" destination MAC address 01-80-C2-00-00-0E. The source address shall be the individual MAC address of the sending station or port. The EtherType used to identify the LLDP protocol shall be 0x88CC.

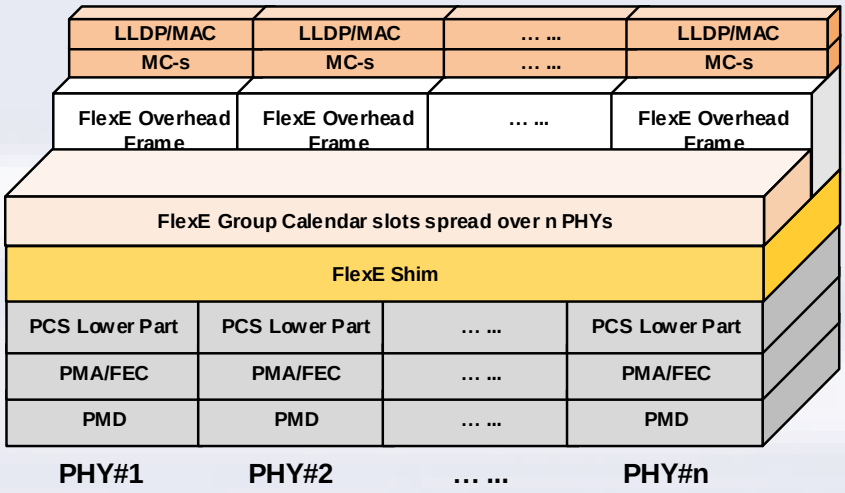

Figure 5 LLDP based neighbor discovery protocol stack

6.2 Unaffiliated PHY and Operational PHY

An unaffiliated PHY is a PHY which does not belong to any FlexE Group, however is transmitting and trying to receive FlexE Overhead Frame(working in FlexE mode). The PHY number field and the PHY map field shall be all 0's while the FlexE Group Number field shall be 0xFFFFE in the overhead frame as shown in table 1.

Table 1: Field values of the FlexE overhead frame for an unaffiliated PHY

Overhead Marker	Group Number	PHY Number	PHY Map
0x4B + 0x5	0xFFFFE	0x00	All 0x00

The section management channel of an unaffiliated PHY remains available for neighbor discovery.

An operational PHY refers to a PHY in a FlexE group. Fields in the overhead frame shall carry the values according to the FlexE IA.

6.2.1 FlexE and StdE Compatibility

With the emergence of FlexE, an Ethernet PHY can either operate in FlexE or standard Ethernet (StdE) mode. The Ethernet PHYs can be classified into one of the following three categories:

- StdE capable only;
- FlexE capable only;
- StdE and FlexE dual-mode capable.

A dual-mode capable interface is able to receive the LLDPDUs from a PHY under StdE mode or a PHY in FlexE mode. Dual-mode capable PHYs examine the incoming 64B/66B block stream for FlexE anchor block (0x4B + 0x5) to check if the far end is in FlexE mode or StdE mode. They will parse the LLDPDU in section management channel for a PHY working in FlexE mode if there is FlexE anchor block, or parse the LLDPDU in StdE mode if no anchor block is detected. LLDPDU can run in section management channel for the PHYs under FlexE Mode with or without OIF organization specific TLVs or in the entire PHY under StdE mode.

This IA suggests that a dual-mode capable interface should initially link up in FlexE mode as unaffiliated PHYs before a FlexE Group is configured on both end systems. Those FlexE capable only interfaces are up as unaffiliated PHYs with section management channel available for FlexE neighbor discovery (notification). If those interfaces detect FlexE anchor block (0x4B + 0x5) on the incoming 64B/66B stream, they learn that the far end is a PHY operating in FlexE mode. If the dual-mode capable interfaces fail to locate any FlexE anchor block within 10k overhead frame cycles (roughly 1s) after powering up, they learn that the remote PHY is under StdE mode.

Note: Some FlexE capable only PHYs such as the PHYs on OTN transport equipment are either FlexE-aware or FlexE-terminate capable and possess no L2/L3 full BW MAC processing capability but are capable of handling MAC packets over section management channel.

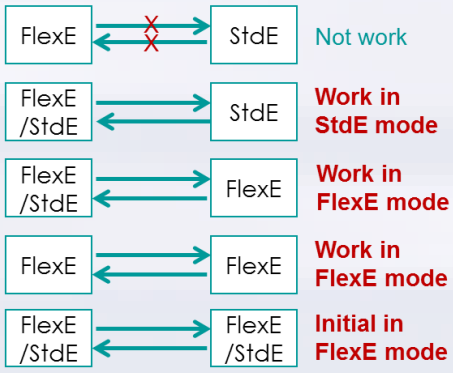

Figure 6 Power-up modes for FlexE, StdE, and dual-mode capable PHYs

As shown in figure 6, a dual-mode capable interface should work in StdE mode when connected to a StdE capable only interface or a dual-mode capable interface manually configured in StdE mode. A dual-mode capable interface should work in FlexE mode as unaffiliated PHY when connected to a FlexE capable only interface or a dual-mode capable interface. As shown in figure 7, FlexE Group should then be configured on both interfaces as required based on the discovered connectivity and remote system capabilities. If a FlexE Group is not configured due to some administrative decision, dual-mode capable interfaces would then switch to StdE mode. Once the dual-mode capable interface enters StdE mode,

it remains in StdE mode unless the link goes down (A link down event can be triggered by a manual configuration forcing the mode change or the cable/fiber disconnection).

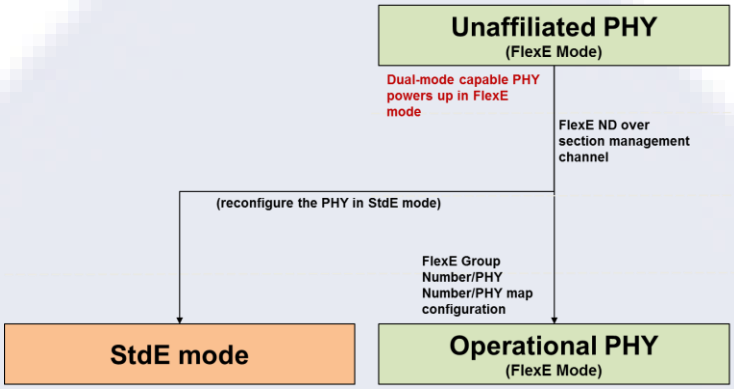

Figure 7 Dual-mode capable PHY initialization scenario

6.3 LLDP OIF Organizationally FlexE Specific TLV Extensions

Figure 8 provides an overview of LLDP TLV extension based on the 802.1AB Organizationally Specific TLVs and carrying TLV type of 127 and using OIF OUI as 0x00-0F-40.

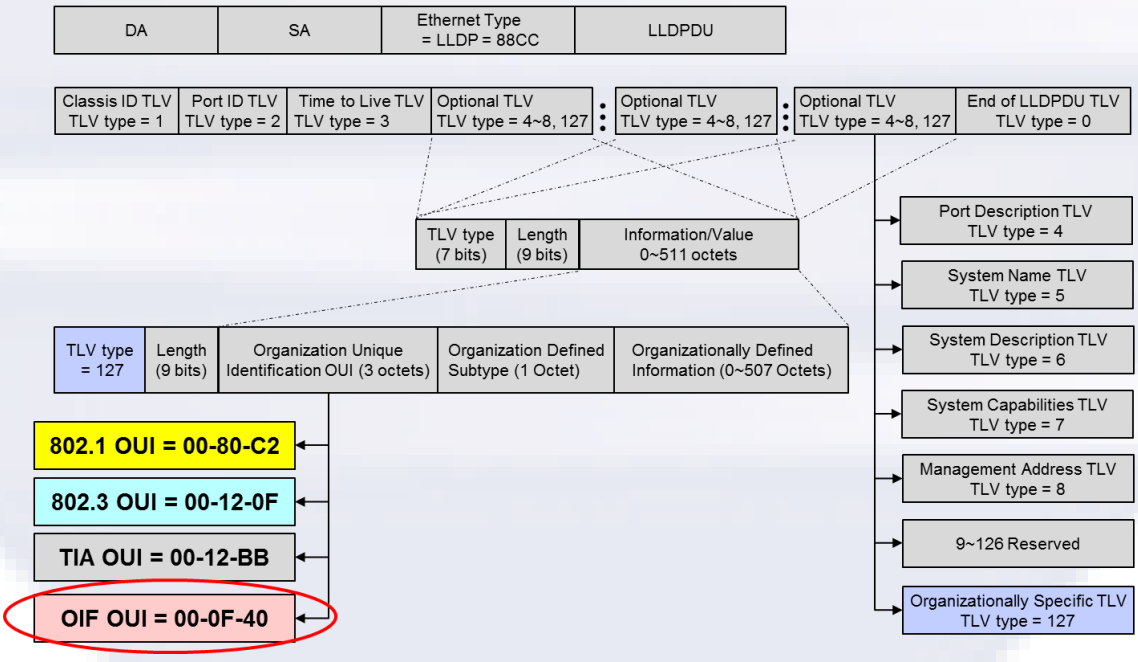

Figure 8 LLDP TLV extensions

The Organization Defined Subtype field values defined by this IA are listed in Table 2.

Table 2: OIF Organizationally Specific TLVs

OIF subtype	TLV name	reference to sub clause
1	FlexE Group Capability	7.1
2	FlexE Group Status	7.2

3	FlexE Deskew Capability	7.3
0, 4 to 255	Reserved	–

Note: the OIF FlexE Specific TLVs are optional for LLDPDUs over the FlexE section management channel or the entire PHY under StdE mode. LLDPDUs without the TLV extensions defined here are sufficient for connectivity discovery and verification purposes.

7 LLDP OIF Organization TLV Extensions

7.1 FlexE Group Capability TLV

The FlexE Group Capability TLV indicates whether the port/PHY can be organized into a FlexE group with other ports in the same local chassis, whether the port is capable of FlexE-aware/FlexE-terminate, as well as the max number of PHYs that can be supported in a FlexE group and the max number of groups that can be supported across multiple PHYs attached to the same chip or line card in the same chassis holding the same FlexE Group Capability ID.

The FlexE Group Capability TLV includes a 16-bits TLV header field and 11-octets TLV information string which contains OIF OUI, OIF Subtype, FlexE Group Capabilities, FlexE Group Capability ID, max number of PHYs in a FlexE Group and max number of FlexE Groups. Detailed format of FlexE Group Capability TLV is illustrated in Figure 9.

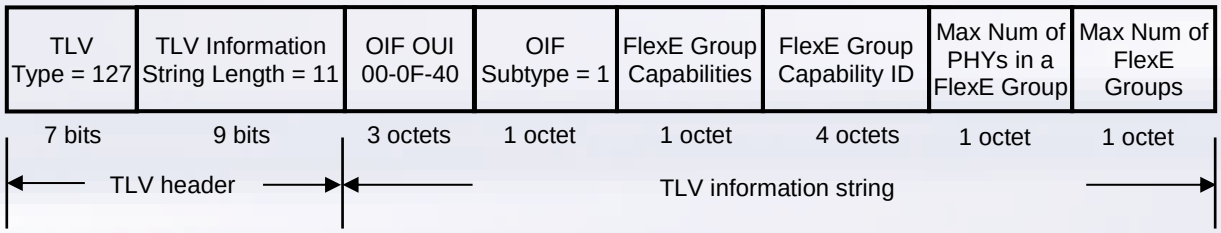

Figure 9 FlexE Group Capability TLV Format

7.1.1 FlexE Group Capabilities

The FlexE Group Capabilities field contains a bitmap indicating the supported FlexE capabilities. The bitmap is illustrated in Table 3.

Table 3: FlexE Group Capabilities

Bit	Function	Notes
0	FlexE Terminate Capability	1: supported; 0: not supported
1	FlexE Aware Transport Capability	1: supported; 0: not supported
2	Client Terminate and MAC Process Capability	1: supported; 0: not supported
3	FlexE Synchronization Channel Capability	1: supported; 0: not supported
4~7	Reserved	–

- Bit 0(LSB) indicates the capability whether FlexE-terminate is supported;
- Bit 1 indicates the capability whether FlexE-aware is supported;

- Bit 2 indicates the capability whether a FlexE Group over the PHY supports termination of FlexE clients and L2 MAC processing;
- Bit 3 indicates the capability whether synchronization (transport of frequency or time information via FlexE overhead frame's 6th block) over the PHY for a given FlexE Group is supported.

7.1.2 FlexE Group Capability ID

The FlexE Group Capability ID field contains a 4-octet values for those ports which can be used to build a FlexE group on the same chassis. If some ports in the same chassis could be used as PHYs of a given FlexE group/subgroup, they should carry the same 4-octet value in the FlexE Group Capability ID field. The values of 0x00000000 and 0xFFFFFFFF are reserved.

7.1.3 Max Num of PHYs in a FlexE Group

The field of <Max Num of PHYs in a FlexE Group> contains an integer value indicating the maximal number (#) of ports/PHYs can be used to build a FlexE Group across multiple PHYs holding the same <FlexE Group Capability ID>. The value can be in the range [1-254] as [0x01~0xFE].

Typical values are 1 ~ 8. For instance, one switch linecard has eight 100GE ports/PHYs and a field value of 8 for those PHYs indicates that all those eight 100GE ports/PHYs (Max Num of PHYs) can be bonded to build a FlexE Group with a total bandwidth of 800G.

7.1.4 Max Num of FlexE Groups

The field of <Max Num of FlexE Groups> contains an integer value indicating the maximal number of groups can be set up across multiple PHYs holding the same <FlexE Group Capability ID> in a node. The value can be in the range of [1-254].

Typical values are 8 and upwards. For example, this field value shall be 8 for a switch linecard with eight 100G FlexE ports and each port can be used to build an individual FlexE group (no deskew requirement). Another example is that a linecard has 8 100G FlexE ports but does not support 8 FlexE Groups with one PHY per Group due to the internal FlexE Group implementation limitations and therefore this field value shall be a positive integer less than 8 and a number that the linecard implementation supports.

An LLDPDU shall contain no more than one (0 or 1) FlexE Group Capability TLV field.

7.2 FlexE Group Status TLV

The FlexE Group Status TLV field indicates the current status of this port. This status includes indications whether it is in a given FlexE group/subgroup, whether it is in the FlexE mode, whether it is in FlexE-aware or FlexE-terminate mode, and the FlexE group/subgroup parameters. Figure 10 shows the FlexE Group Status TLV format.

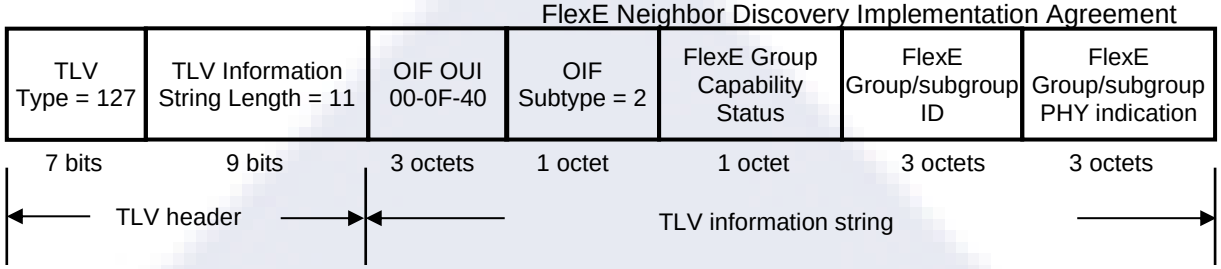

Figure 10 FlexE Group Status TLV Format

7.2.1 FlexE Group Capability Status

After a FlexE Group is initialized (the FlexE Group initialization process is out of scope of this IA), the FlexE Group Capability Status field shall contain a bitmap that indicates the currently enabled local FlexE Group Capability status on each operational PHY as defined in Table 4.

Table 4: FlexE Group Capabilities Status

Bit	Function	Value/meaning
0	FlexE Terminate	1 = in use; 0 = not in use
1	FlexE Aware Transport	1 = in use; 0 = not in use
2	Client Terminate and MAC Process	1 = in use; 0 = not in use
3	FlexE Synchronization Channel	1 = in use; 0 = not in use
4~7	Reserved	–

- Bit 0 indicates whether the PHY/group is in FlexE-terminate mode;
- Bit 1 indicates whether the PHY/group is in FlexE-aware mode; Note: the bit should be ignored if Bit 0 is 1 indicating the PHY/group is already in FlexE-terminate mode;
- Bit 2 indicates whether the PHY/group terminates FlexE Clients and provides L2 MAC processing;
- Bit 3 indicates whether the port/group is configured to support FlexE synchronization channel (i.e., transport the frequency or time via 6th overhead block of the FlexE overhead frame).

For example, an operational PHY working in FlexE-terminate mode with Client Terminate and MAC process capability and synchronization support configured shall have the FlexE Group Capability Status field value 0bxxxx11x1.

For an operational PHY on a FlexE-aware transport OTN node without Client Terminate and MAC process capability and synchronization support, the field's value should be 0bxxxx0010. For an operational PHY on a FlexE termination transport (OTN) node, the field's value should be 0bxxxx00x1.

For StdE capable only PHYs, the field's value should be 0bxxxx0100.

7.2.2 FlexE Group/Subgroup ID

The FlexE Group/Subgroup ID is a 3-octet field, which shall contain the configured FlexE group/subgroup ID. To be specific, the least significant 20 bits contain the FlexE Group ID of a given FlexE Group on this PHY, while the most significant 4 bits indicate the

absence/presence of FlexE subgroup and represent the subgroup ID if it is present. If the most significant 4 bits are 0x0, it indicates the absence of FlexE subgroup. If the most significant 4 bits are between 0x1 ~ 0xF, they identify a FlexE subgroup and represent the FlexE subgroup ID as illustrated in table 5.

Table 5: FlexE Group/Subgroup ID field

	Most Significant 4 bits	Least significant 20 bits
Group ID	0x0	FlexE Group ID of a given FlexE Group on this PHY
Subgroup ID	0x1	FlexE Group ID of a given FlexE Group on this PHY
Subgroup ID	…	…
Subgroup ID	0xF	FlexE Group ID of a given FlexE Group on this PHY

It's possible for a router to split a FlexE Group as multiple subgroups to be transported over transport network (e.g., OTN). For example, a router can be connected to two OTN transport devices and a FlexE-aware subgroup is created for the connected PHY on each OTN device.

As shown in table 5, the 4-bits Subgroup ID is in range of 0x1~0xF which can provide at most 15 subgroups. For a FlexE Subgroup, PHYs in the PHY Map of overhead frame are a subset of the PHYs of the FlexE Group. FlexE Group/Subgroup ID and the Chassis ID in LLDPDU over section management channel per PHY are sufficient to distinguish one FlexE Group/Subgroup from another FlexE Group/Subgroup.

7.2.3 FlexE Group/Subgroup PHY Indication

In a situation where a transport node is linked to a router in a subgroup transport configuration, both nodes need to check that all the PHYs of a subgroup are connected.

FlexE Group/Subgroup PHY indication field contains 3 octets as <Prev PHY Number, Current PHY Number, and Next PHY Number> which are illustrated in figure 11.

Octet 1 Prev PHY Number	Octet 2 Current PHY Number	Octet 3 Next PHY Number

Figure 11 FlexE Group/Subgroup PHY field

- Prev PHY Number field with the value 0x00 indicates that Current PHY is the first PHY with lowest PHY Number in the Subgroup. Any other value in Prev PHY Number field indicates there is another PHY with a lower PHY number belongs to the same Subgroup.
- Next PHY Number field with the value 0x00 indicates that Current PHY is the last PHY with the highest PHY Number in the Subgroup. Any other value in Next PHY Number field indicates there is another PHY with a higher PHY number belongs to the same Subgroup.

PHYs of a FlexE Group/Subgroup shall be fully present between two adjacent nodes and the FlexE Group PHY Map shall match FlexE Group/Subgroup PHY indication.

An LLDPDU shall contain no more than one (0 or 1) FlexE Group Status TLV field.

7.3 FlexE Deskew Capability TLV

The FlexE Deskew Capability TLV is used to indicate the local FlexE RX (receiving) demux's skew tolerance associated with current PHY to the remote end. FlexE Deskew Capability TLV shall be carried in an LLDPDU and transmitted periodically in default or user-specified cycle. An LLDPDU shall contain no more than one FlexE Deskew Capability TLV field. Figure 12 shows the TLV format.

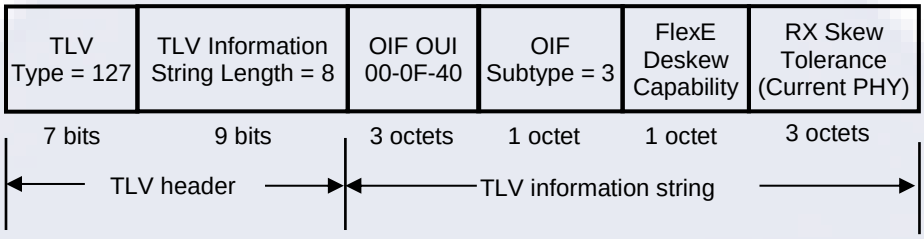

Figure 12 FlexE Deskew Capability TLV Format

7.3.1 FlexE Deskew Capability

The "FlexE Deskew Capability" field is a 1-octet bitmap indicating the supported deskew capability of the demux associated with current PHY. The bits are defined according to Table 6.

- Bit 0 indicates whether the local demux supports 300ns low skew tolerance.
  A binary value "1" indicates it is supported and the value "0" indicates no such support.
- Bit 1 indicates whether the local demux supports application specific RX high skew tolerance.
  A binary value "1" indicates it is supported and while the value "0" indicates no such support.

For example, the value 0bxxxxxx01 shows that the local demux supports only 300ns low skew tolerance while the value 0bxxxxxx11 indicates the local demux supports 300ns low skew tolerance and an application-specific high skew tolerance which is specified in the next field.

Table 6: Bits Definition in FlexE Deskew Capability field

Bit	Function	Value/meaning
0	RX 300ns low skew tolerance	1 = support RX 300ns low skew tolerance; 0 = no deskew support/no PHY bonding capability;
1	Application specific RX high skew tolerance	1 = support application specific RX high skew tolerance which is specified in the next field "RX Skew Tolerance (Current PHY)"; 0 = not supported;
2~7	Reserved	—

7.3.2 RX Skew Tolerance (Current PHY)

The "RX Skew Tolerance (Current PHY)" field contains a 3-octets integer value which indicates current FlexE PHY's skew tolerance in terms of 64B/66B bit blocks count which is

equivalent to the count for 0.64ns skew tolerance for a 100GBASE-R PHY, 0.32ns for a 200GBASE-R PHY, and 0.16ns for a 400GBASE-R PHY.

For example, the value 500 means current FlexE PHY's RX deskew buffer is able to buffer 500 × 64B/66B blocks which translate to 320ns RX Skew Tolerance for a 100GBASE-R PHY. The value 15,625 represents 15,625 × 64B/66B blocks which translate to 10μs RX Skew Tolerance for a 100GBASE-R PHY.

Valid values should be in range of [0x000001 ~ 0xFFFFFE]. If bit 0 in FlexE Deskew Capability field is "0", this field is ignored and shall carry the value 0.

8 Appendix A: LLDPDU Encapsulation

To transmit a FlexE ND LLDPDU, the FlexE ND LLDP agent on the local node first collects the relevant information from the LLDP local system MIB and FlexE ND local system MIB to produce a LLDPDU with three mandatory TLVs(Chassis ID, Port ID, and Time to live) and one to a few FlexE ND TLVs. Each LLDPDU is first encoded into an Ethernet MAC frame by a LLC layer function (Link Layer Control service per IEEE 802.1ab-2016 standard) which is separate from the MAC layer in the data plane. The LLDPDU-encapsulated Ethernet MAC frame shall then be converted into a sequence of 64B/66B bit blocks by the PCS layer function per [802.3] clause 82.2.3. The resulting 64B/66B bit blocks are put into the section management channels (MC-s) of a local PHY in sequence. Figure 13 illustrates the encapsulation process. Note that the first 64B/66B bit block(/S/ block) of the LLDPDU MAC frame shall always appear in either 1st block or 2nd block of the section management channel. Figure 13 corresponds to the first case. Similarly, the last 64B/66B bit block(/T/ block) of the LLDPDU MAC frame may appear in either 1st block or 2nd block of the section management channel. Between two consecutive LLDPDUs, idle control blocks may appear in MC-s

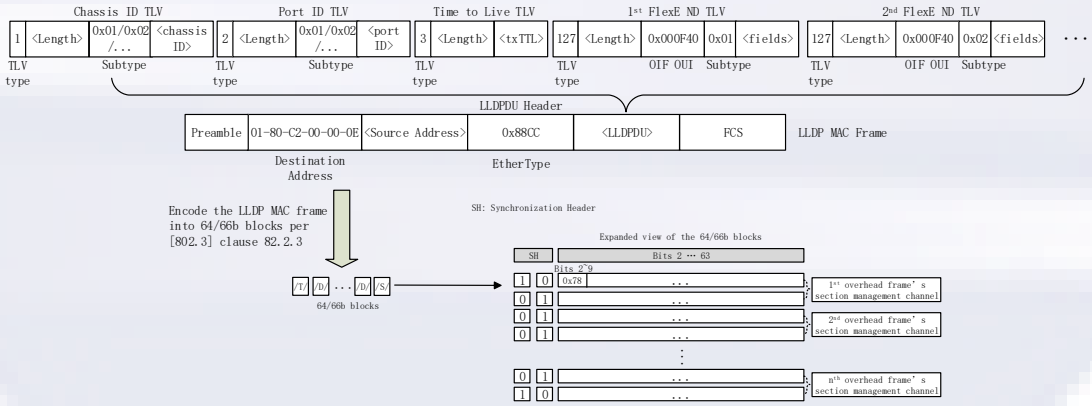

Figure 13 Encapsulation of a FlexE ND LLDPDU into the section management channel

On the receiving end, the 64B/66B bit blocks in the section management channels shall be extracted and merged in sequence to form a sequence of 64B/66B bit blocks. The sequence of 64B/66B bit blocks are processed by the PCS layer function per [802.3] clause 82.2.3 to generate Ethernet MAC frames. The LLC layer function examines each MAC frame's EtherType field and extracts the LLDPDU from the MAC frame if the MAC frame's "EtherType" value matches "0x88CC". The TLVs contained in the LLDPDU are used to

update the LLDP remote systems MIB and FlexE ND remote system MIB maintained on the local node and for the control plane decision-making.

9 Appendix B: Use Cases

9.1 Use Case: Connectivity Discovery for Group Configuration (FECD/FECV)

9.1.1 Manual FlexE Group Configuration

FlexE IA focuses on data plane operation after the FlexE Group and its PHYs are manually configured on both ends. The local end and the remote end must be configured with the same configuration.

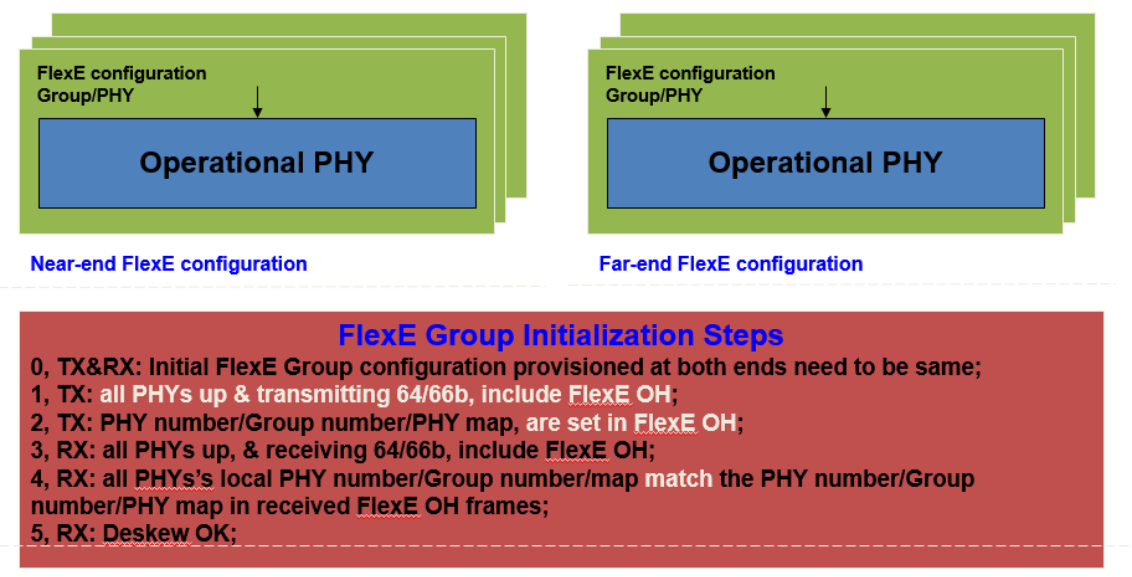

Figure 14 FlexE Group initialization steps

In the manual configuration mode as illustrated in figure 14, all PHYs on both ends are up and transmitting 64B/66B block stream including FlexE overhead frames. The PHY Map, PHY number and Group number are set and reflect the FlexE Group Configuration. At the receiver side/direction, the local control plane confirms that all PHYs' received incoming FlexE overhead frames are same with local configuration and all PHYs get successfully deskewed before the FlexE Group is up.

However, the manual FlexE Group configuration approach faces some realistic problems in real-world FlexE device deployment. In a FlexE node NodeA with multiple PHYs, configuring the multiple PHYs to form a single FlexE group with the neighbor node NodeB's PHYs requires some knowledge of physical PHY connections. However, it is very impractical to sort out the link pair manually without the help of some connectivity discovery process. It is very difficult to track which PHY of NodeA is connected to which PHY of NodeB in case of multiple PHY connections between NodeA and NodeB. For example, PHY#A1 of NodeA is connected to some PHY#B1 of the NodeB. It would require the network installation technician to mark and follow the connected fiber/cable to identify which PHY on NodeB is PHY#B1. This problem worsens as the number of physical connections grow between two

nodes and all those connections are intended to be bonded as a FlexE Group. The manual configuration approach requires marking all the to-be-bonded PHYs in advance and assigning those PHYs on both nodes with the same PHY number, Group number, PHY Map per PHY pair. Any changes to the physical connections or creating a new FlexE Group with more or fewer PHYs would require a lot of manual reconfigurations.

In a FlexE device consisting of multiple line cards, there could be restrictions that only the PHYs from the same line card can be bonded to form a FlexE Group with the remote node. It would take some manual efforts to identify and mark the PHYs from the same line card and create the FlexE groups only out of the PHYs from the same line card.

9.1.2 FlexE ND Enabling Efficient Group Configuration

With FlexE neighbor discovery, PHYs can be up as unaffiliated PHYs with section management channel running LLDP with OIF Organization specific TLVs. The local system can discover the PHY connectivity between the local end and remote end and discover the far end's FlexE capability. These connectivity information greatly facilitates the configuration for a FlexE group for the both ends. Figure 15 shows the connectivity discovery process per PHY before a PHY is configured.

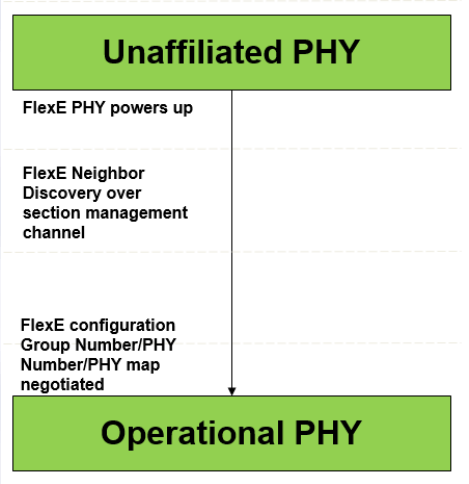

Figure 15 Per PHY Group configuration using FlexE ND

In the first phase of connectivity discovery, the PHY is an "Unaffiliated PHY". As the connectivity is discovered, then it becomes an "Operational PHY" in a FlexE group with the PHY number, Group number, and PHY map configured. On contrast, an "Operational PHY" becomes an "Unaffiliated PHY" when the group is removed. The PHY number, Group number, and PHY Map in "Unaffiliated PHY" shall carry the values as defined in section 6.2. Figure 16 shows the transition between Unaffiliated PHY and Operational PHY.

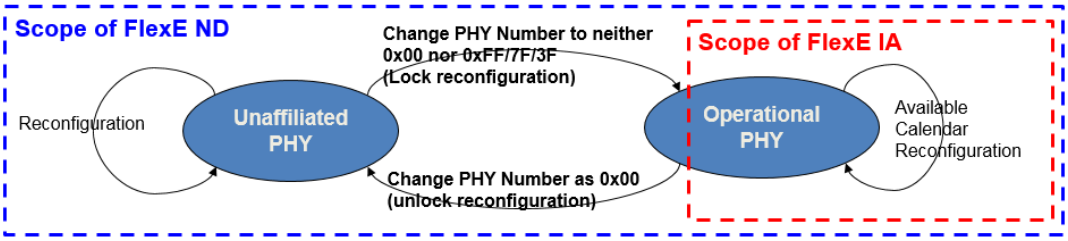

Figure 16 Transition diagram between unaffiliated PHY and operational PHY

9.2 Use Case: FlexE Group Remote Capability Discovery (FERCD)

Figure 16 shows a typical use case for FERCD, where “System A” is connected to “System B” and “System C” via three individual PHY connections. Assuming all three systems are FlexE-terminate nodes and all PHYs are FlexE capable, all the connected PHYs should be brought up as unaffiliated PHYs (see section 6.2). Each unaffiliated PHY is transmitting LLDPDU over section management channel of the overhead frame to its neighbor.

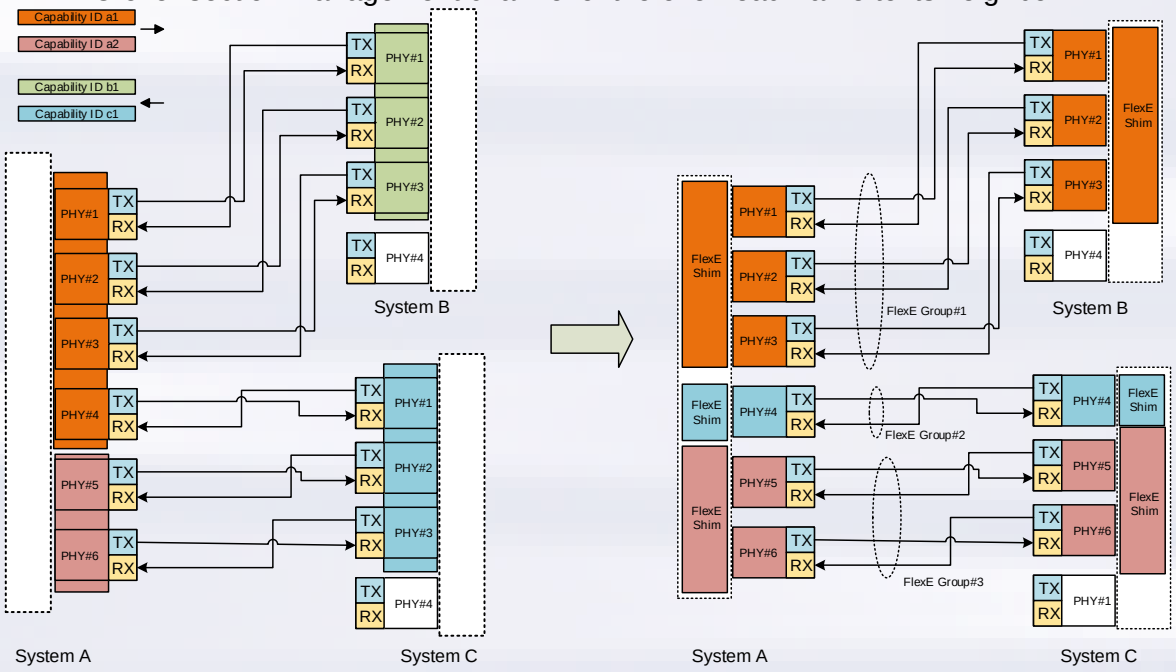

Figure 17 Remote Capability Discovery example

In Figure 17, “System A” can discover the connectivity between itself and “System B” as well as the connectivity between itself and “System C”. Furthermore, while System A and B notify each other its FlexE Group Capability TLV, “System A” discovers that all three ports on “System B” connected to “System A” hold the same FlexE Group Capability ID <b1> and match the local three ports’ capability. “System B” also discovers that all three ports on “System A” connected to “System B” hold the same FlexE Group Capability ID <a1>. “System A” discovers that all three ports on “System C” connected to “System A” hold the same FlexE Group Capability ID <c1> while “System C” will be able to discover that three ports on “System A” connected to “System C” hold different FlexE Group Capability IDs <a1> and <a2>. A FlexE Group cannot be set up across the PHYs with different FlexE Group

Capability ID <a1> and <a2> between System A and C, a possible FlexE Group configuration is illustrated in the right part of Figure 17.

9.3 Use Case: FlexE Group Remote Status Notification (FESIV)

Figure 4 shows the use case that PHYs belonging to a configured FlexE Group are divided as two subgroups and both subgroups are further FlexE aware transported over the transport network. The transport node B1 and B2 each will receive the PHY map indicating there are two PHYs in overhead frame from its neighbor node A for the FlexE Group. However, B1 and B2 are unable to locate another PHY for FlexE group integrity verification purpose besides the respective connected PHY. With FlexE Subgroup Integration Verification (FESIV), ports of node A notify node B1 and node B2 the subgroup integrity configuration by using FlexE Group Status TLV so that A-B1 and A-B2 can synchronize with each other about the subgroup's configuration integrity.

Table 7: FlexE Group Status TLVs for A, B1, B2 in figure 4

	PHY Number	FlexE Group Capability Status	FlexE Group/subgroup ID		FlexE Group/Subgroup PHY indication		
			Group ID	Subgroup ID	Prev PHY	Current PHY	Next PHY
A's PHY#1	1	0b00000001	0x00001	0x1	0x00	0x1	0x00
A's PHY#2	2	0b00000001	0x00001	0x2	0x00	0x2	0x00
B1's PHY#1	1	0b00000010	0x00001	0x1	0x00	0x1	0x00
B2's PHY#2	2	0b00000010	0x00001	0x2	0x00	0x2	0x00

9.4 Use Case: FlexE Deskew Capability Discovery (FEDCD)

To configure a FlexE Group over multiple PHY links, it would be useful for the network operator/admin to first check that the local PHYs' deskew capabilities and their remote PHYs' deskew capabilities are sufficient for the intended application deployment given the fact that there are FlexE implementations with varying deskew capabilities.

As illustrated in Figure 18, the intended application is a 2000km FlexE unaware transport long-haul DCI case. A FlexE Group is intended to be configured over the four 100GBASE-R PHY links between Node A and Node Z. Suppose the network operator/user has only access to Node A's control plane but also wants to find out whether Node Z's deskew capability matches the deskew requirement for the long-haul case. The FlexE Deskew Capability TLV from Node Z would perfectly serve the purpose and help alarm the network operator/admin if Node Z is a device with only 300ns low skew tolerance or a device with higher but still insufficient skew tolerance to meet the deskew requirement.

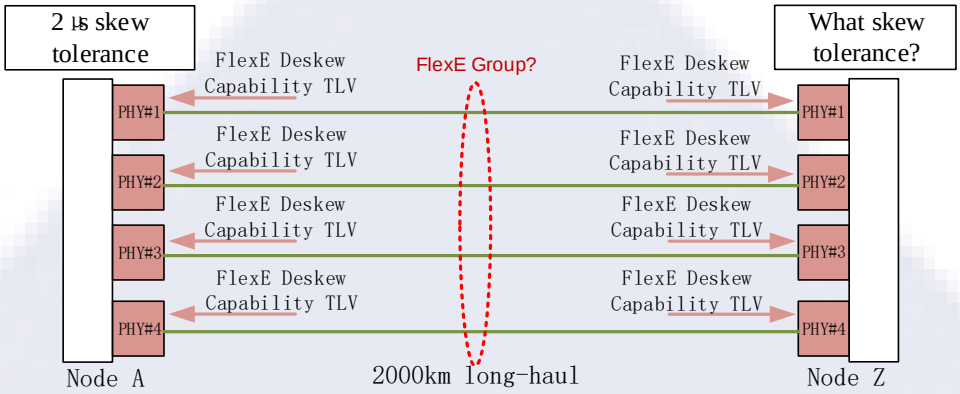

Figure 18 An example illustrating the use of FlexE Deskew Capability TLV

Table 8: Node A and Z's FlexE Deskew Capability TLVs for Figure 17

	FlexE Deskew Capability	RX Skew Tolerance (Current PHY)	Notes
A's Deskew Capability TLV over PHY#1	0b00000011	0d3,125	3,125 64B/66B blocks roughly translate to 2,000ns RX skew tolerance for a 100GBASE-R PHY
A's Deskew Capability TLV over PHY#2	0b00000011	0d3,125	
A's Deskew Capability TLV over PHY#3	0b00000011	0d3,125	
A's Deskew Capability TLV over PHY#4	0b00000011	0d3,125	
Z's Deskew Capability TLV over PHY#1	0b00000001	0d469	469 64B/66B blocks roughly translate to 300ns RX skew tolerance for a 100GBASE-R PHY
Z's Deskew Capability TLV over PHY#2	0b00000001	0d469	
Z's Deskew Capability TLV over PHY#3	0b00000001	0d469	
Z's Deskew Capability TLV over PHY#4	0b00000001	0d469	

10 Appendix C: List of OIF member companies when this IA was approved

Acacia Communications	Google	O-Net Communications (HK) Limited
ADVA Optical Networking	Hewlett Packard Enterprise (HPE)	Oclaro
Alibaba	Hitachi	Orange
Amphenol Corp.	Huawei Technologies Co., Ltd.	PETRA
Analog Devices	IBM Corporation	Precise-ITC, Inc.
Anritsu	Infinera	Qorvo
Applied Optoelectronics, Inc.	Innovium	Ranovus
Arista Networks	Inphi	Renesas Electronics Corporation
Barefoot Networks	Integrated Device Technology	Rianta Solutions, Inc.
BizLink Technology Inc.	Intel	Rockley Photonics
Broadcom Limited	Invecas, Inc.	Rosenberger Hochfrequenztechnik GmbH
Cadence Design Systems	IPG Photonics Corporation	Roshmere
Cavium	JCRFO	Samtec Inc.
CenturyLink	Juniper Networks	Semtech Canada Corporation
China Telecom Global Limited	Kandou Bus	SiFotonics Technologies Co., Ltd.
Ciena Corporation	KDDI Research, Inc.	Sino-Telecom Technology Co., Inc.
Cisco Systems	Keysight Technologies, Inc.	Socionext Inc.
Coriant	Lumentum	Spirent Communications
Corning	MACOM Technology Solutions	Sumitomo Electric Industries
Credo Semiconductor (HK) LTD	Maxim Integrated Inc.	Sumitomo Osaka Cement
Dell, Inc.	MaxLinear Inc.	Synopsys, Inc.
EFFECT Photonics B.V.	MediaTek	TE Connectivity
Elenion Technologies, LLC	Mellanox Technologies	Tektronix
Epson Electronics America, Inc.	Microsemi Inc.	Telefonica I + D
eSilicon Corporation	Microsoft Corporation	TELUS Communications, Inc.
Fiberhome Technologies Group	Mitsubishi Electric Corporation	UNH InterOperability Laboratory (UNH-IOL)
Finisar Corporation	Molex	Verizon
Foxconn Interconnect Technology	Multilane SAL Offshore	Viavi Solutions Deutschland GmbH
Fujikura	NEC Corporation	Xelic
Fujitsu	NeoPhotonics	Xilinx
Furukawa Electric Japan	Nokia	Yamaichi Electronics Ltd.
Global Foundries	NTT Corporation	

11 References

[FlexE1.1] OIF-FLEXE-IA-1.1.
[FlexE2.0] OIF-FLEXE-IA-2.0.
[802.3] IEEE Std 802.3™-2015 *Standard for Ethernet*.
[802.1AB] IEEE Std 802.1™AB-2016, Station and Media Access Control Connectivity Discovery.