Flex Ethernet Implementation Agreement

IA # OIF-FLEXE-01.1

June 21, 2017

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

The OIF is an international non profit organization with over 90 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 IEEE 802.1, IEEE 802.3, IETF, IP-MPLS Forum, IPv6 Forum, ITU-T SG13, ITU-T SG15, MEF, ATIS-OPTXS,ATIS-TMOC, TMF and the XFP MSA Group.

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 Group: **Physical and Link Layer**

TITLE: **Flex Ethernet Implementation Agreement 1.1**

SOURCE:

TECHNICAL EDITOR

Stephen J. Trowbridge, Ph. D.
Nokia
5280 Centennial Trail
Boulder, CO 80303 USA
Phone: +1 972 477 8172
Email: steve.trowbridge@nokia.com

WORKING GROUP CHAIR

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

ABSTRACT: The Flex Ethernet (FlexE) Implementation Agreement provides a generic mechanism for supporting a variety of Ethernet MAC rates that may or may not correspond to any existing Ethernet PHY rate. This includes MAC rates that are both greater than (through bonding) and less than (through sub-rate and channelization) the Ethernet PHY rates used to carry FlexE. This can be viewed as a generalization of the Multi-Link Gearbox implementation agreements, removing the restrictions on the number of bonded PHYs (MLG2.0, for example, supports one or two 100GBASE-R PHYs) and the constraint that the FlexE clients correspond to Ethernet rates (MLG2.0 supports only 10G and 40G clients).

Notice: This Technical Document has been created by the Optical Internetworking Forum (OIF). This document is offered to the OIF Membership solely as a basis for agreement and is not a binding proposal on the companies listed as resources above. The OIF reserves the rights to at any time to add, amend, or withdraw statements contained herein. Nothing in this document is in any way binding on the OIF or any of its members.

The user's attention is called to the possibility that implementation of the OIF implementation agreement contained herein may require the use of inventions covered by the patent rights held by third parties. By publication of this OIF implementation agreement, the OIF makes no representation or warranty whatsoever, whether expressed or implied, that implementation of the specification will not infringe any third party rights, nor does the OIF make any representation or warranty whatsoever, whether expressed or implied, with respect to any claim that has been or may be asserted by any third party, the validity of any patent rights related to any such claim, or the extent to which a license to use any such rights may or may not be available or the terms hereof.

© 2017 Optical Internetworking Forum

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction other than the following, (1) the above copyright notice and this paragraph must be included on all such copies and derivative works, and (2) this document itself may not be modified in any way, such as by removing the copyright notice or references to the OIF, except as needed for the purpose of developing OIF Implementation Agreements.
By downloading, copying, or using this document in any manner, the user consents to the terms and conditions of this notice. Unless the terms and conditions of this notice are breached by the user, the limited permissions granted above are perpetual and will not be revoked by the OIF or its successors or assigns.

This document and the information contained herein is provided on an “AS IS” basis and THE OIF DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY, TITLE OR FITNESS FOR A PARTICULAR PURPOSE.

1 Table of Contents

2 List of Figures

3 List of Tables

4 Document Revision History

Working Group: Physical and Link Layer

SOURCE:

Editor's Name
Stephen J. Trowbridge, Ph. D.
Alcatel-Lucent
5280 Centennial Trail
Boulder, CO 80303 USA
Phone: +1 972 477 8172
Email: steve.trowbridge@nokia.com

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

DATE: **January 2016**

Issue No.	Issue Date	Details of Change
oif2016.246.01	10/26/2016	Incorporate Implementation Guide oif2016.367.00

5 Introduction

The Flex Ethernet (FlexE) implementation agreement provides a generic mechanism for supporting a variety of Ethernet MAC rates that may or may not correspond to any existing Ethernet PHY rate. This includes MAC rates that are both greater than (through bonding) and less than (through sub-rate and channelization) the Ethernet PHY rates used to carry FlexE. This can be viewed as a generalization of the Multi-Link Gearbox implementation agreements, removing the restrictions on the number of bonded PHYs (MLG2.0, for example, supports one or two 100GBASE-R PHYs) and the constraint that the FlexE clients correspond to Ethernet rates (MLG2.0 supports only 10G and 40G clients).

5.1 Requirements

The general capabilities supported by the FlexE implementation agreement are:

- Bonding of Ethernet PHYs, e.g., supporting a 200G MAC over two bonded 100GBASE-R PHYs.
- Sub-rates of Ethernet PHYs, e.g., supporting a 50G MAC over a 100GBASE-R PHY.
- Channelization within a PHY or a group of bonded PHYs, e.g, support a 150G and two 25G MACs over two bonded 100GBASE-R PHYs.

Note that hybrids are also possible, for example a sub-rate of a group of bonded PHYs, for example, a 250G MAC over three bonded 100GBASE-R PHYs.

The general approach is illustrated in Figure 1.

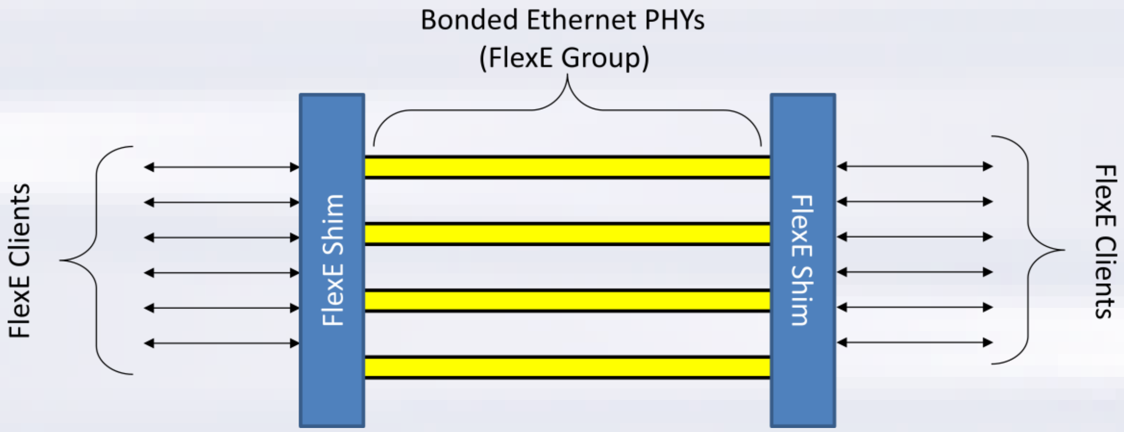


Figure 1: General Structure of FlexE

The *FlexE Group* refers to a group of from 1 to n bonded Ethernet PHYs. This version of the Implementation Agreement supports FlexE groups composed of one or more bonded 100GBASE-R PHYs. New higher rates (e.g., 400GbE under development in the IEEE P802.3bs project) are intended to be included once those standards are complete.

A *FlexE Client* is an Ethernet flow based on a MAC data rate that may or may not correspond to any Ethernet PHY rate. The FlexE client MAC rates supported by this implementation agreement are 10, 40, and m × 25 Gb/s.

The *FlexE Shim* is the layer that maps or demaps the FlexE clients carried over a FlexE group. Similar to terminology of MLG, the *FlexE mux* refers to the transmit direction which maps the FlexE clients over the FlexE group. The *FlexE demux* refers to the receive direction which demaps the FlexE clients from the FlexE group.

5.2 Relationship to IEEE 802.3 Stack

The FlexE shim can be envisioned as being in the middle of the PCS in the 100GBASE-R stack as illustrated in [802.3] Figure 80-1. Each FlexE client has its own separate MAC, Reconciliation Sublayer, and xMII above the FlexE shim which operate at the FlexE client rate. The layers below the PCS (100GBASE-R PMA, optional FEC, PMD) are used intact as specified for Ethernet.

5.2.1 FlexE mux functions

The functions of the FlexE mux (the FlexE shim functions in the transmit direction) are illustrated in Figure 2.

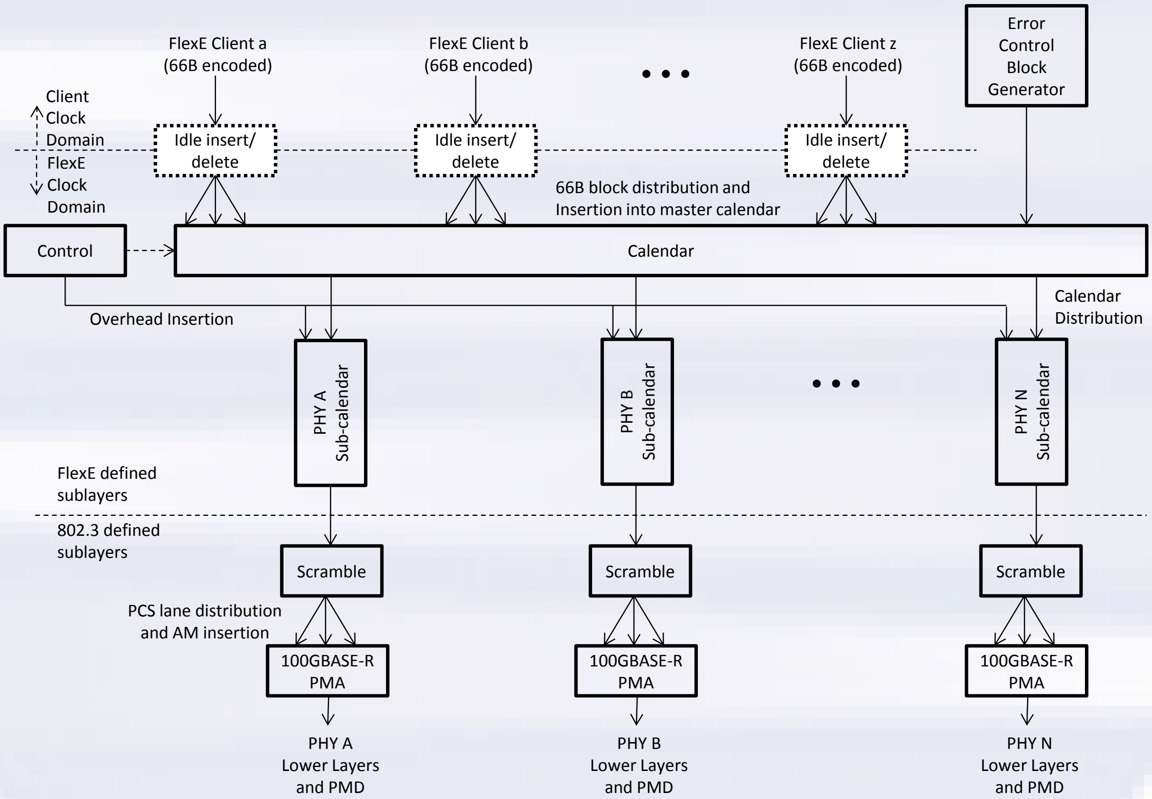


Figure 2: FlexE mux functions

5.2.1.1 FlexE Client

Each FlexE client is presented to the FlexE shim as a 64B/66B encoded bit-stream according to [802.3] Figure 82-5. How this bit-stream is created is application specific (see clause 7.2 for details), but should appear to the FlexE shim as having been created from an Ethernet MAC operating at a rate of 10, 40, or m × 25 Gb/s, through a logical RS layer which performs the link fault signaling functions described in [802.3] clause 81.3.4. The content of FlexE client stream may be LF in the case of an upstream failure of the FlexE client. The start control character is aligned to an 8-byte boundary, feeding through a logical xMII at the corresponding rate, and a 64B/66B encoder resulting in a FlexE client rate of:

$$\frac{66}{64} \times \text{FlexE Client MAC rate } \pm 100\text{ppm}$$

5.2.1.2 Idle insert/delete

All FlexE clients must be rate-adapted to match the clock of the FlexE group. This is accomplished by idle insertion/deletion according to [802.3] clause 82.2.3.6 and/or ordered set deletion according to [802.3] clause 82.2.3.9. The nominal rate of the adapted signal is slightly less than the nominal rate of the FlexE client to allow room for the alignment markers on the PHYs of the FlexE group and insertion of the FlexE overhead.

5.2.1.3 66B Block Distribution and Insertion into Calendar

The 66B blocks from each FlexE client are distributed sequentially into the calendar in the order described in clause 6.3.

5.2.1.4 Calendar Distribution

The 66B blocks from calendar are distributed to each PHY of the FlexE group according to the ordering describe in clause 6.3. The FlexE overhead is inserted into the sub-calendar of each PHY as described in clause 7.4.

5.2.1.5 100GBASE-R PHY functions (Scramble, lane distribution, AM insertion, PMA, FEC, PMD)

The stream of 66B blocks of each PHY is distributed to the PCS lanes of that PHY with insertion of alignment markers, and this is presented at the PMA service interface in the 100GBASE-R stack. Lower layers and interfaces of the 100GBASE-R Ethernet PHY (e.g., CAUI, FEC, PMA, PMD) are used as specified in [802.3].

5.2.1.6 Error Control Block Generator

Error Control blocks are generated for insertion into calendar slots that are unused or unavailable. See Figure 14.

5.2.1.7 Control

The control function manages which calendar slots each FlexE client is inserted into and inserts the FlexE overhead on each FlexE PHY in the transmit direction.

5.2.2 FlexE Demux Functions

The functions of the FlexE demux (the FlexE shim in the receive direction) are illustrated in Figure 3.

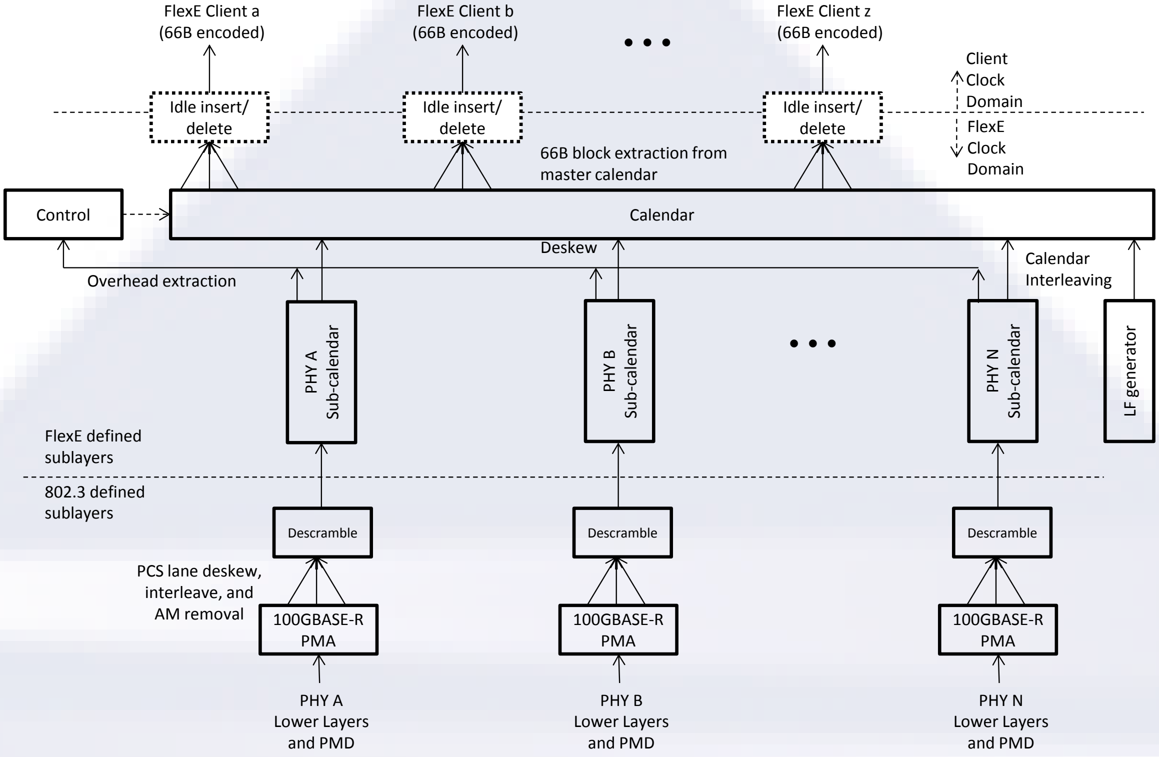


Figure 3: FlexE demux functions

5.2.2.1 100GBASE-R PHY functions (PMD, PMA, FEC, lane deskew, interleave, AM removal, descramble)

The layers of each 100GBASE_R PHYs below the PCS are used exactly as specified in [802.3]. The PCS lanes are recovered, deskewed, reinterleaved, and the alignment markers are removed. The aggregate stream is descrambled.

5.2.2.2 Calendar Interleaving and Overhead Extraction

The calendar slots of the sub-calendars on each PHY are logically interleaved in the order specified in clause 6.4. The FlexE overhead is recovered from each PHY.

5.2.2.3 LF Generator

In the case that any PHY of the FlexE group has failed (PCS_Status=FALSE) or overhead frame lock or overhead multiframe lock has not been achieved on the overhead of any of the PHYs, LF is generated towards all FlexE clients in the group.

5.2.2.4 66B Block Extraction from Calendar

The 66B blocks are extracted from the calendar positions assigned to each FlexE client in the order described in clause 6.3.

5.2.2.5 Idle Insertion/Deletion

Idle insertion/deletion according to [802.3] clause 82.2.3.6 and/or ordered set deletion according to [802.3] clause 82.2.3.9 may be performed to rate-adapt the extracted 66B flow when necessary to adjust to the FlexE client rate. Note that the nominal rate of the

66B flow carried in the calendar slots is slightly less than the nominal rate of the FlexE client as the available space is reduced by the FlexE PHY PCS lane alignment markers and the FlexE overhead.

5.2.2.6 Control

The control function manages which calendar slots each FlexE client is extracted from and extract the FlexE overhead from each FlexE PHY in the receive direction.

5.3 Sample Applications

FlexE can support a variety of applications. A non-exhaustive list includes:

- Router to Transport Connection (more examples below).
- Intra-Data Center “Fat Pipe” application: bonded PHYs for flows exceeding the PHY rate, or carrying traffic that doesn’t distribute efficiently with LAG.
- Generalized MLG applications, e.g., an n × 100G PHY as an umbilicus to a satellite shelf of lower rate ports.

One case of router to transport connection is where the transport network is unaware of FlexE. This case is illustrated in Figure 4. This may be used with legacy transport equipment that provides PCS-codeword transparent transport of 100GbE, but provides no special support for FlexE.

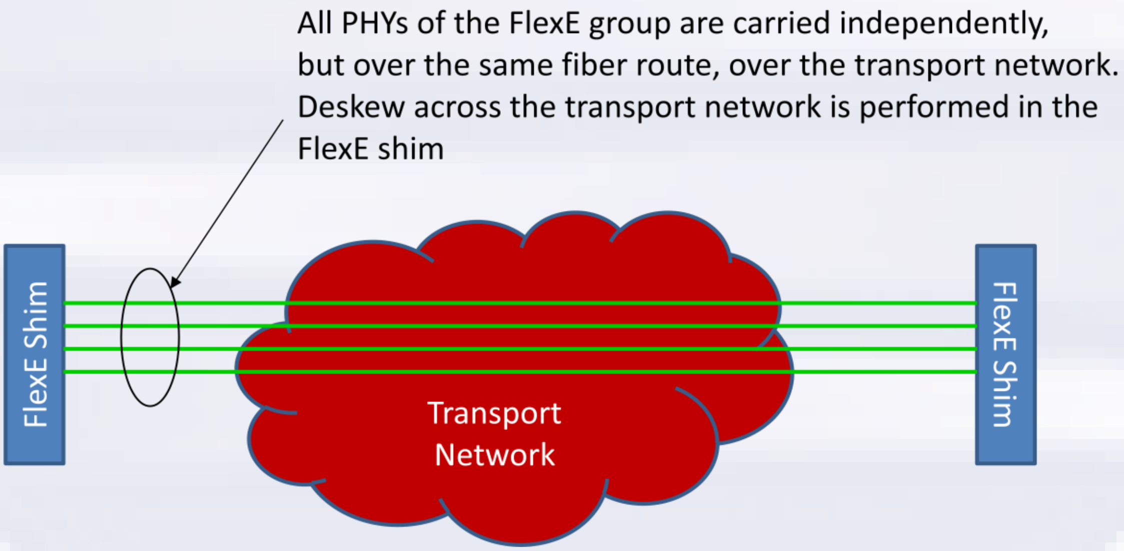
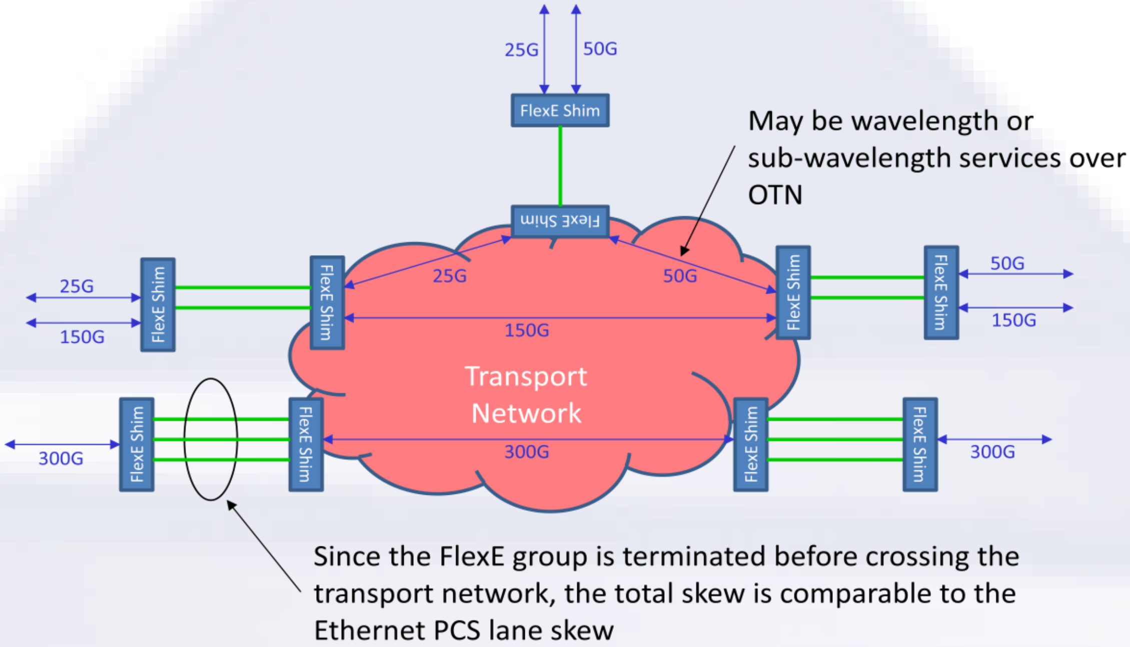

All PHYs of the FlexE group are carried independently, but over the same fiber route, over the transport network. Deskew across the transport network is performed in the FlexE shim

Figure 4: Router to FlexE unaware Transport Network connection

In the FlexE unaware case, the FlexE shim, e.g., in a router, maps the FlexE client(s) over a group of bonded Ethernet PHYs. Each of the Ethernet PHYs is carried independently over the transport network using a PCS codeword transparent mapping. The Ethernet PHYs are intended to be carried over the same fiber route: diverse routing is not envisioned. All of the PHYs of the FlexE group need to be interconnected between the same two FlexE shims. The FlexE shim will need to tolerate and accommodate considerably more skew than if the FlexE shims were only separated by an Ethernet link distance of 40km or less, as the transport network could carry the signal over thousands of kilometers. In Figure 4, it is the PHYs of the FlexE group which are carried over the transport network.

Another case of router to transport connection is where the transport network equipment terminates the FlexE group. This case is illustrated in Figure 5.



May be wavelength or sub-wavelength services over OTN

Since the FlexE group is terminated before crossing the transport network, the total skew is comparable to the Ethernet PCS lane skew

Figure 5: FlexE terminating transport network equipment

In the FlexE terminating case, the distance between any pair of FlexE shims is limited to the Ethernet link distance (about 40km maximum), so the amount of skew that needs to be tolerated and compensated is considerably less. The other important distinction here is that it is the FlexE clients, rather than the PHYs of the FlexE group, which are carried over the transport network. The FlexE client could be constructed to be the complete size of the payload that can be carried over a single wavelength (e.g., construct 200G to fill a DP-16QAM wavelength with the bonding of two 100GBASE-R PHYs), or could be a smaller client which is multiplexed and switched at a sub-wavelength level.

The final router to transport example described is one where the transport network is aware that it is carrying FlexE PHYs (as opposed to 100GbE), but the FlexE group is not terminated on the transport equipment. This may be used to support cases where the Ethernet PHY rate is be greater than the wavelength rate, the wavelength rate is not an integral multiple of the PHY rate, or there is a reason (for example, wavelengths terminated on different transponder line cards) that it is not possible to terminate the FlexE group in the transport equipment. This kind of example is illustrated in Figure 6.

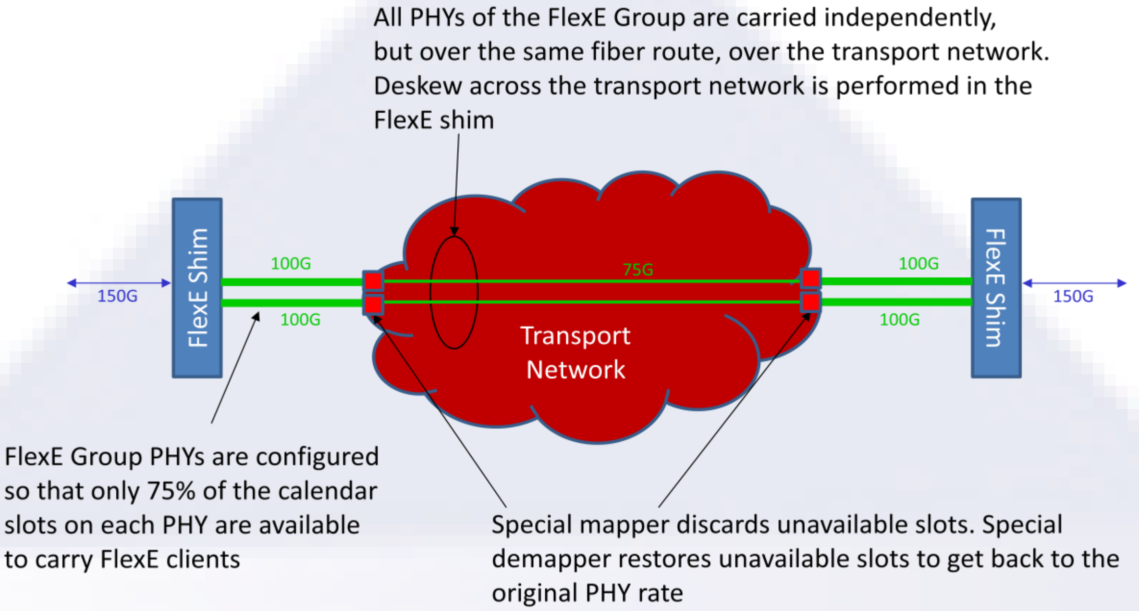

All PHYs of the FlexE Group are carried independently, but over the same fiber route, over the transport network. Deskew across the transport network is performed in the FlexE shim
150G
FlexE Shim
100G
100G
75G
Transport Network
100G
100G
FlexE Shim
150G
FlexE Group PHYs are configured so that only 75% of the calendar slots on each PHY are available to carry FlexE clients
Special mapper discards unavailable slots. Special demapper restores unavailable slots to get back to the original PHY rate

Figure 6: Example of FlexE aware transport of Ethernet PHYs of a FlexE group

6 General Mechanism

6.1 FlexE Group

The 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]. The values of 0 and 255 are reserved. A PHY number may correspond to the physical port ordering on equipment, but the FlexE shim at each end of the group must identify each PHY in the group using the same PHY number. PHY numbers within the group do not need to be contiguous.

Each of these PHYs uses the bulk of the PCS functions described in [802.3] clause 82 including PCS lane distribution, lane marker insertion, alignment, and deskew. All PHYs of the FlexE group must use the same physical layer clock. Each PHY of the FlexE group is able to deliver a logically serial stream of 64B/66B encoded blocks from the FlexE mux to the FlexE demux at a data rate of:

$$103.125\ \text{Gb/s} \times \frac{16383}{16384} \pm 100\text{ppm}$$

While the protocol supports a number of PHYs in the FlexE group up to 254, practical implementations are likely limited to the range of 4-8 PHYs. The fraction applied to the base rate reflects the fact that 1/16K of the space of the interface is occupied by PCS lane alignment markers which are not space available to carry FlexE blocks. The FlexE blocks carried over each PHY of the FlexE group has the format of a logically serial stream of (mostly) legal 64B/66B blocks with the format described in [802.3] Figure 82-5, although the blocks do not appear in a sequence that would make sense if interpreted as an Ethernet interface. The actual PHYs of the FlexE group may transcode these blocks to 256B/257B format according to [802.3] clause 91.5.2.5 according to the PHY type, but they are trans-decoded back to 64B/66B blocks prior to delivery to the FlexE demux.

6.2 FlexE Client

The FlexE Client presented to the FlexE shim is in the format described in clause 5.2.1.1.

All FlexE clients to be transmitted over the same FlexE group are aligned to a common clock and rate-adapted to the available space in the FlexE calendar (whose nominal rate is slightly less than that of the FlexE client due to space needed for the FlexE PHY PCS lane alignment markers and FlexE overhead) according to the process described in clause 5.2.1.2.. The rate-adapted FlexE client operates at a rate of:

$$\text{FlexE Client MAC rate} \times \frac{66}{64} \times \frac{16383}{16384} \times \frac{1023\times 20}{1023\times 20+1} \pm 100\text{ppm}$$

This nominal rate is about 0.011% less than the nominal rate of the FlexE client, which is well within what can be accomplished with idle insertion/deletion without packet loss. Note that this doesn't actually correspond to any clock that needs to be generated in an implementation, as the idle insertion deletion process will simply operate by filling the allocated block positions in the FlexE group from a FlexE client FIFO, inserting or deleting idles in the process of filling the block positions in the FlexE group according to the calendar (see below).

6.3 FlexE Calendar

The FlexE mechanism operates using a calendar which assigns 66B block positions on sub-calendars on each PHY of the FlexE group to each of the FlexE clients. The calendar has a granularity of 5G, and has a length of 20 slots per 100G of FlexE group capacity. Two calendar configurations are supported: an “A” and a “B” calendar configuration. At any given time, one of the calendar configurations is used for mapping the FlexE clients into the FlexE group and demapping the FlexE clients from the FlexE group. The two calendar configurations are provided to facilitate reconfiguration. When a switch of calendar configurations adds or removes FlexE clients from the FlexE group, existing clients whose size and calendar slot assignments are not changed by changing the calendar configuration are not affected.

For a FlexE group composed of n bonded 100GBASE-R PHYs, the logical length of the calendar is 20n. The blocks as allocated per the calendar are distributed to n sub-calendars of length 20 on each of the PHYs of the FlexE group according to Figure 7.

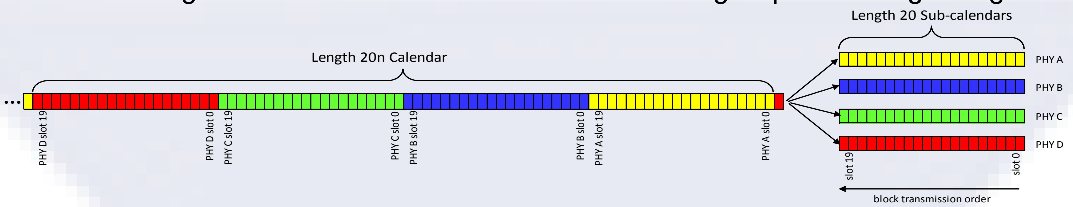


Figure 7: Illustration of FlexE Calendar Distribution

The order of distribution of twenty blocks at a time is selected over simple “round robin” distribution of 66B blocks. Calendar slots are identified by their PHY number and the slot [0-19] (within that PHY). The “logical” sequence number of a calendar slot is 20×the PHY number plus the calendar slot number within the PHY. The sequence is ascending order. Note that the sequence numbering is not necessarily consecutive when the assigned

PHY numbers are not contiguous. This logical order only matters when calendar slots on different PHYs are assigned to the same FlexE client.

6.4 FlexE Overhead and Alignment

The alignment of the data from the PHYs of the FlexE group is accomplished by the insertion of FlexE overhead into the stream of 66B blocks carried over the group. The FlexE overhead is delineated by a 66B block which can be recognized independently of the FlexE client data. An illustration of the FlexE overhead on each PHY of the FlexE group is given in Figure 8.

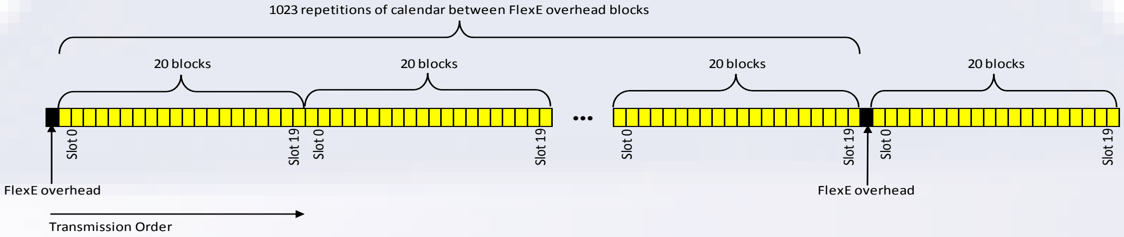


Figure 8: Illustration of insertion of FlexE overhead on each PHY of the FlexE group

On a 100GBASE-R PHY, a FlexE overhead block will occur approximately once per 13.1µs. The actual format of the FlexE overhead blocks is such that they occur in a repeating sequence of eight blocks, so the sequence has a period of approximately 104.77µs. This sequence of overhead blocks is inserted in the same positions in the sub-calendar sequence on each PHY and is used to align all of the PHYs of the FlexE group at the FlexE demux to reconstruct the sequence in the order of the calendar so that the FlexE clients can be recovered.

The FlexE aware transport network scenario illustrated in Figure 6 allows for marking a certain number of the calendar slots as unavailable. This is different from “unused”, in that it is known, due to transport network constraints, that not all of the calendar slots generated from the FlexE mux will reach the FlexE demux and therefore no FlexE client should be assigned to those slots. The intention is that when a PHY of the FlexE group is carried across the transport network, the mapping is able to compress the signal to less than the PHY rate by dropping the unavailable calendar slots. A case where 25% of the calendar slots are unavailable is illustrated in Figure 9. Unavailable slots are placed at the end of each relevant sub-calendar (the highest numbered slots).

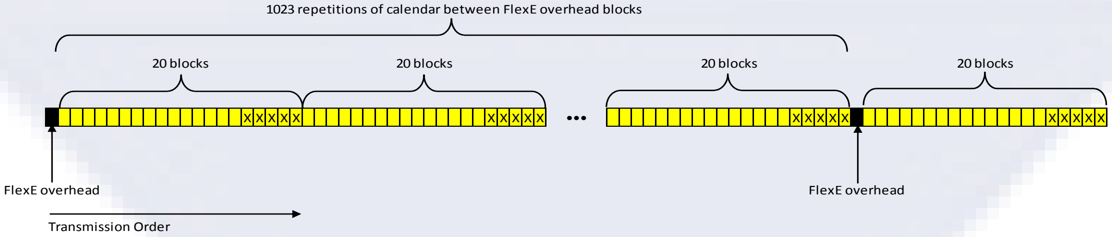


Figure 9: Illustration of Unavailable Calendar Slots to Facilitate Transport at lower rates

The anchor position of the FlexE overhead on each PHY is encoded as an ordered set (control block type 0x4B). A distinct “O” code is selected (0x5) which is different from that for the sequence ordered set used by Ethernet or the signal ordered set used by Fibre channel. The information to be transmitted in the FlexE overhead is encoded into the bytes D1, D2, and D3 of the ordered set block as indicated in Figure 10.

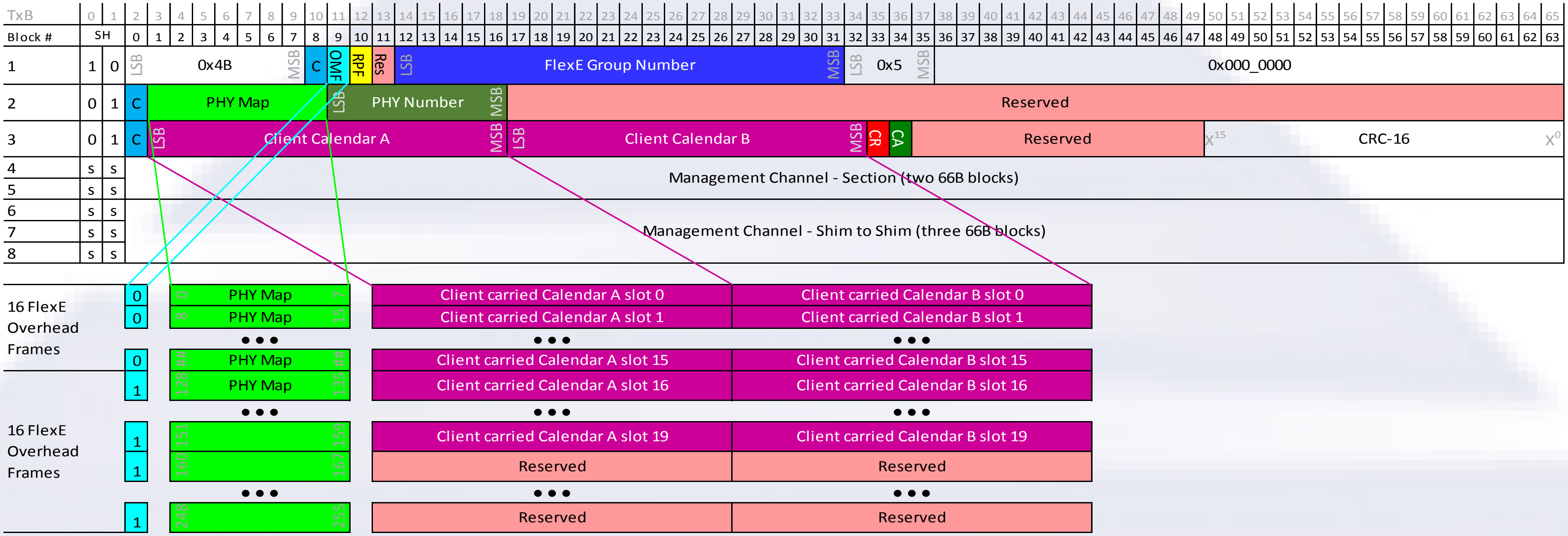
1	0	0x4B	D1	D2	D3	0x5	0x000_0000

Figure 10: Encoding of Ordered Set block for FlexE overhead

The information which needs to be included in the overhead includes:

- Which PHYs are part of this FlexE group
- The number identifying this PHY within the FlexE group
- A way to transmit the programming of the sub-calendar configurations for each PHY from the FlexE mux to the FlexE demux
- A way to indicate which calendar configurations (“A” or “B”) is in use at this time
- The FlexE group number (if necessary, for the case where the same PHYs may be configured to be part of different FlexE groups)
- Fields to support protocols necessary for changing the configuration of FlexE clients into calendar slots.
- Two optional management channels. These may not be necessary in all applications (for example, if a network management system has direct access to the FlexE shim at both ends of the connection), but may be useful for applications such as using FlexE for an n × 100G umbilicus to a remote shelf of lower-rate ports. One management channel (the 4th and 5th blocks of the FlexE overhead frame) is available for communication across a section (for example, from a router with a FlexE shim to FlexE aware transport equipment which does not terminate the FlexE shim), and the other management channel (the 6th-8th blocks of the FlexE overhead frame) is used for communication between the FlexE shims.

The amount of information to be conveyed from the FlexE mux to the FlexE demux exceeds the 24 bits available in a single ordered set block per PHY. This is addressed by spreading the relevant overhead across a sequence of eight FlexE overhead blocks on each PHY, each separated by 20 × 1023 FlexE data blocks. This group of eight overhead blocks is referred to as the FlexE overhead frame. The FlexE overhead frame is illustrated in Figure 11. The meaning, interpretation and processing of this overhead is explained in clause 7.



C = Calendar configuration in use (see 7.3.2)
OMF = Overhead Multiframe Indicator (see 7.3.1)
C = Calendar configuration in use (see 7.3.2)
RPF = Remote PHY Fault (see 7.3.8)
CR = Calendar Switch Request (see 7.3.4)
CA = Calendar Switch Acknowledge (see 7.3.4)
s s = Valid sync header bits (01 or 1 0)

PHY Map = Control of which PHYs are members of this group (see 7.3.3)
PHY Number = Identity of this PHY within the group (see 7.3.3)
FlexE Group Number - See 7.3.6
Client (being programmed) - See 7.3.4
Reserved - See 7.3.7
Management Channels - See 7.3.5
CRC-16 - See 7.3.8

Figure 11: FlexE Overhead Frame and Multiframe

The first block of the FlexE overhead frame is encoded as an ordered set as shown in Figure 10. The next two FlexE overhead blocks are encoded as data 66B data blocks. The final five FlexE overhead blocks are reserved for the two optional management channels, and may carry any legal 66B block per [802.3] Figure 82-5 (excluding Ordered Sets with O code=0x5, as this is reserved for FlexE PHY overhead).

The first block in the FlexE overhead frame serves as a marker to be used for alignment and re-interleaving of the sub-calendars from each of the PHYs of the FlexE group at the FlexE demux. One FlexE overhead frame consisting of eight 66B overhead blocks is transmitted in approximately 104.77µs. Subject to the amount of buffer provided in a given implementation, skew detection and compensation across the PHYs of the FlexE group can be compensated up to nearly half of this amount.

Test vectors for the overhead format are found in Appendix A.

7 Detailed Functions

The detailed processing to implement the functionality described in clause 6 is provided in this clause.

7.1 FlexE Group Functions

The FlexE group is composed of from 1 to n 100GBASE-R PHYs. Each 100GBASE-R PHY reuses nearly all of the functions as described for 100GBASE-R in [802.3]. This includes a subset of the functions of the PCS as described in clause 82, and all of the functions from sub-layers below the PCS as described in clauses 83, 86, 88, and 91-95 as appropriate according to the PHY type.

The FlexE shim provides to each FlexE group PHY a set of 64B/66B encoded blocks that are encoded according to Figure 82-5. Within the PCS, clause 82, each FlexE group PHY reuses (with reference to Figure 82-2) in the transmit direction, the scrambler, block distribution, and alignment insertion processes. In the receive direction, each FlexE group PHY reuses the lane block sync, alignment lock and lane deskew (including BER monitor), lane reorder, alignment removal, and descrambling.

7.2 FlexE Client Generation

The format and bit rate of FlexE clients is described in clause 6.2. FlexE clients generally originate from one of the following sources.

7.2.1 FlexE Clients Generated internally within a system

A FlexE client may be generated internally within a system, for example from a Network Processing Unit (NPU) within a router. The packet flow is generated at the determined FlexE client MAC rate and 64B/66B encoded according to [802.3] Figure 82-5.

7.2.2 FlexE Clients received from an Ethernet PHY

FlexE clients at the rates of 10G, 40G, 100G, and in the future 25G, 50G, 200G and 400G (per existing P802.3by, P802.3bs, and proposed new projects) can be created from an Ethernet PHY at the corresponding rate with some processing to convert to the FlexE client format and rate.

A 10GBASE-R signal will be converted to a 10G FlexE client format before presenting to a FlexE mux by converting the coding from the 66B codeword set of [802.3] Figure 49-7 to the 66B codeword set of [802.3] Figure 82-5. This performs idle insertion/deletion in groups of four idles and or ordered set deletion where necessary to align the start control character to an 8-byte boundary (eliminating the use of control block types 0x2d, 0x66, and 0x33), converting instances where a control block contains two ordered sets to one (changing control block type 0x55 to 0x4b by deleting one of the two ordered sets encoded in the block), and by replacing the unnecessary idle control characters from the end of the Figure 49-7 control block type 0x4b with zeros to produce the Figure 82-5 control block format with control block type 0x4b. A 10G FlexE client may be converted to a 10GBASE-R signal by using the idle insertion/deletion process as described in [802.3] clause 49.2.4.7 to adapt to the 10GBASE-R nominal rate, ensuring that at least four idles appear between packets. Idles may be inserted or deleted in groups of four, and/or ordered sets may be deleted according to [802.3] clause 49.2.4.10, resulting in the start of packet on a four-byte boundary rather than the 8-byte boundary. Ordered sets encoded with control block type 0x4b have the zeros at the end of the block replaced with four idle control characters. The result is that the blocks are encoded according to Figure 49-7.

A 40GBASE-R signal can be converted to a FlexE client by serializing and deskewing the PCS lanes, removing the PCS lane alignment markers, and using the idle insertion/deletion process as described in [802.3] clause 82.2.3.6 and/or ordered set deletion as described in [802.3] clause 82.2.3.9 to adapt the signal to the 40G FlexE client rate. A 40G FlexE client coming from a FlexE demux is converted to a 40GBASE-R interface by using the idle insertion/deletion process as described in [802.3] clause 82.2.3.6 and/or ordered set deletion according to [802.3] clause 82.2.3.9, distributing the blocks round-robin to the four PCS lanes, and inserting PCS lane alignment markers.

A 100GBASE-R signal without FEC can be converted to and from a FlexE client in the same manner as 40GBASE-R described above (except that the number of PCS lanes is 20 rather than 4). A 100GBASE-R signal with FEC, in converting to a FlexE client, also will correct any errors per the FEC code, remove the FEC, and trans-decode from 256B/257B prior to the idle insertion/deletion process. To convert a 100G FlexE client coming from a FlexE demux to a 100GBASE-R signal with FEC involves the same processes as for 40GBASE-R, but in addition, transcoding the signal to 256B/257B, inserting the FEC lane alignment markers, and adding the FEC.

7.2.3 FlexE Clients from another FlexE shim

In the case of equipment which terminates the FlexE group, FlexE clients can be delivered from the one FlexE shim to another: for example, from a FlexE shim at the transport network ingress to another FlexE shim at the transport network egress. The FlexE client, a sequence of 64B/66B encoded blocks, is expected to be carried over the transport network without packet loss. As no timing information is carried by this stream, idle insertion/deletion and ordered set deletion are possible in the mapping over the transport network. The FlexE shim at the network egress will only need to perform idle insertion/deletion according to [802.3] clause 82.2.3.6 and/or ordered set deletion according to [802.3] clause 82.2.3.9, not due to any expected change in the nominal bit-rate, but simply to align the clock with the FlexE group clock.

7.2.4 Interconnect flexibility

Note that since the format of the FlexE client is simply a logically serial stream of 66B blocks at a given rate, FlexE clients do not need to be produced or received in the same manner at both ends of the connection. For example, a 10G, 40G or 100G FlexE client might be generated as a system internal signal in the main chassis of a system, connected using an n × 100G FlexE umbilicus to a satellite shelf, and connected to physical 10GBASE-R, 40GBASE-R or 100GBASE-R ports on the satellite shelf. In the case where the FlexE mux is receiving a FlexE client from a physical Ethernet port and the FlexE demux is delivering that FlexE client to a physical Ethernet port, the two ports obviously have to be the same nominal rate, but they may not have be the same PHY type.

7.3 FlexE Overhead Processing

The format of the FlexE overhead is indicated in Figure 11.

7.3.1 FlexE Overhead Frame and Multiframe Lock

The FlexE overhead is encoded as 66B blocks and are inserted on each PHY of the FlexE group. One overhead block is inserted after every 1023 iterations of the length 20 sub-calendar of 66B FlexE data blocks, so the sequence is one block of overhead followed by 1023 × 20 blocks of data followed by one block of overhead.

FlexE overhead frame lock is achieved at the receiver (FlexE demux) on each PHY by recognizing the FlexE block 1 of the FlexE overhead frame, encoded as a special ordered set (the sync header is 10, the control block type is 0x4B (ordered set), and the "O" code is 0x5), and then finding the FlexE ordered set block again (1023 × 20 + 1) × 8 block positions later. Once FlexE overhead frame lock is achieved, the next expected FlexE overhead block will be 1023 × 20 + 1 block positions later. While in FlexE overhead frame lock, bytes D1-D3 of the ordered set block, plus the 66Bblocks occurring at 20461, 40922, 61383, 81844, 102305, 122766, and 143227 blocks beyond the ordered set block will be interpreted as FlexE overhead frame. FlexE overhead is not interpreted if not in FlexE overhead lock. FlexE overhead lock will be lost if the sync header, control block type, or O code do not match at the expected position for 5 occurrences.

Certain information is transmitted in every FlexE overhead frame. Other information is distributed across a sequence of 32 FlexE overhead frames, referred to as the FlexE overhead multiframe. The OMF (overhead multiframe) bit has a value of "0" for the first sixteen overhead frames of the overhead multiframe, and a value of "1" for the last sixteen overhead frames of the overhead multiframe, as shown in Figure 11. The FlexE demux achieves overhead multiframe lock on each PHY when the OMF bit transitions from a "0" to a "1" or a "1" to a "0" in consecutive overhead frames with good CRC. There are two opportunities in the overhead multframe for the FlexE demux to achieve overhead multiframe lock.

7.3.2 Calendar Configuration in Use

There are two calendar configurations for each PHY of the FlexE group: the "A" calendar configuration (encoded as 0) and the "B" calendar configuration (encoded as one). The two calendars are used to facilitate reconfiguration. Clients can be added or removed from the FlexE group without affecting the traffic on other clients. While clients can be resized through calendar updates, there is no assurance that the resizing can be "hitless" in all network scenarios. Normally, changes are only made to the calendar which is not currently in use.

Exceptions would include initial link configuration or replacement of a failed circuit pack where it is necessary to download the calendar information into the replacement pack. The data flow through the FlexE mux/demux pair is indeterminate during any changes to the active calendar configuration, until the assignment for all slots in the active calendar configuration have stabilized and been received by the FlexE demux in FlexE overhead frames with good CRC.

The calendar configuration in use is signaled from the FlexE mux to the FlexE demux on each PHY in the three bit positions labeled "C" in Figure 11. While most of the FlexE overhead can be reliably protected by the CRC, the calendar configuration in use must be interpreted even if the CRC is bad, since the FlexE demux must be able to switch its calendar in use at precisely the same overhead frame boundary as the FlexE mux. So that this can be done reliably, three copies of the calendar configuration in use are transmitted, and interpreted by the receiver by majority vote. Since the three copies are separated into different FlexE overhead blocks across the overhead frame (the first and second copy are separated by 1,350,415 bits and the second and third copies are separated by 1,350,425 bits), the different copies will never be affected by the same burst error. Since each PHY should have a BER of 10^{-12} or better, the probability of two instances of the calendar in use being wrong is no more than 10^{-24}, which can safely be ignored.

When the calendar configuration in use changes from a 0 to a 1, or from a 1 to a 0, the calendar configuration used by both the FlexE mux and the FlexE demux will be changed beginning from the first data block following first block of the next FlexE overhead frame on each PHY.

7.3.3 PHY Map and PHY Number

The set of PHYs in the FlexE group (not necessarily consecutive PHY numbers) are indicated in the "PHY Map" field of the FlexE overhead. This is distributed as eight bits per overhead frame in each of the thirty-two overhead frames in the overhead multiframe (256 bits total), with each bit set to a one indicating a PHY number that is a member of the group, and all other bits of the PHY map set to zero. The PHY Map values are only accepted from overhead frames with good CRC. The full PHY map is sent on all PHYs of the FlexE group so that it is possible for the FlexE demux to verify that the same PHY numbers are configured at the FlexE mux as at the FlexE demux, and can tell whether all expected PHYs are being received.

The PHY number in the FlexE group (from 1 to n) is encoded in the second block of the FlexE overhead frame. Note that this is persistent information which does not change while the group is in service. The receiver accepts a value for "PHY Number" when the same value is received in two consecutive overhead frames with good CRC. Updates to the respective group of eight bits of the PHY map bit map are accepted from overhead frames with good CRC.

7.3.4 Calendar Configuration

The contents of both the A and B calendar configurations are transmitted continuously from the FlexE mux to the FlexE demux, with the contents of one calendar slot of the A and B sub-calendars for each PHY being transmitted in the first twenty overhead frames of the FlexE overhead multiframe. The client fields are ignored by the FlexE demux when not in overhead multiframe lock since the FlexE demux would not know which slot in which calendar that client belongs to.

The sub-calendar configurations on each PHY are transmitted by sending the clients assigned to each calendar slot in the same order as the corresponding 66B payload block positions occur in the transmission sequence on that PHY.

The Client fields indicate which of the FlexE clients is mapped into a given calendar slot in the A and B calendar configurations for the sub-calendar carried over that PHY. The size of a given FlexE client can be calculated based on the number of calendar slots that client is assigned to (i.e., how many calendar slots have the same numeric value in the Client field across the entire FlexE group). The Clients are indicated by 16-bit fields transmitted in the 3rd block of the FlexE overhead frame. The value 0x0000 indicates a calendar slot which is unused (but available). The value 0xFFFF (all ones) indicates a calendar slot that is unavailable, for the case indicated in Figure 6 where the full FlexE group PHY rate cannot be carried over the transport network in the FlexE aware transport use case. Any value other than 0x0000 or 0xFFFF may be used to designate a particular FlexE client carried by the group.

The Client fields are ignored in overhead frames with a bad CRC, leaving previous assignments to the clients in the relevant slot unchanged.

The full contents of both calendar configurations are transmitted from the FlexE mux to the FlexE demux approximately once every 3.35ms. The fact that the calendar configurations are transmitted continuously avoids any inconsistency between the calendars at the FlexE mux and the FlexE demux due to a lost message.

The normal process of reconfiguration (e.g., adding or removing FlexE clients to or from the FlexE group) will involve programming the new or modified FlexE client assignments into the calendar configuration which is not in use, then switching to the updated calendar configuration, and finally updating the original calendar configuration.

The switch from one active calendar configuration to another can be coordinated between the FlexE mux and the FlexE demux using the Calendar Request (CR) bit sent from the FlexE mux to the FlexE demux, and the Calendar Acknowledge (CA) bit sent from the FlexE demux to the FlexE mux. Normally, the CR bit has the same value as the active calendar configuration being sent from the FlexE mux to the FlexE demux, and the CA bit has the same value as the calendar configuration currently being used by the FlexE demux.

When the FlexE mux has completed the programming of the offline calendar configuration and is ready to switch, it informs the FlexE demux by changing the CR bit to the value of the offline calendar configuration on all PHYs of the FlexE group beginning with the same overhead frame in the overhead positions on each PHY.

When the FlexE demux is prepared to accept the switch of calendar configuration, it informs the FlexE mux by changing its CA bit to match the incoming CR bit on all PHYs of the FlexE group beginning with the same overhead frame in the overhead positions on each PHY. When this occurs is application specific. At the earliest, it occurs once the assignment of every calendar slot in the offline configuration has been received by the FlexE demux in a FlexE overhead frame with a good CRC after the change of the incoming CR bit. But the CA bit indication may be delayed for a variety of reasons: for example, software may need to be prepared for the incoming bandwidth change, for example in SDN applications.

The FlexE mux normally will switch calendar configurations only after receiving the CA bit acknowledgment after telling the FlexE demux it is ready to switch. The FlexE mux should set a timer (suggested default value as 1 second) after changing the outgoing CR bit. Appropriate values for the timer and action to be taken if the timer expires prior to receiving the CA bit are application specific. The timer could be as short as about 15ms (allowing for three complete transmissions of the calendar), but may be longer, for example, if software on the far end must be prepared to accept the switch to the updated calendar. The action to be taken on timer expiry without receiving the CA bit response is either to proceed with the switch or to raise an alarm and wait for corrective action to be taken. The FlexE mux indicates to the FlexE demux the change of calendar by changing the value of all three “C” bits to indicate the new calendar in the same FlexE overhead frame on all PHYs.

Note that the availability of the above information in the protocol is not intended to limit the ways in which FlexE can be used. The FlexE demux may not act as a “slave” of the FlexE mux in terms of calendar configurations in every application. For example:

- A static configuration (e.g., one composed of a fixed number of PHYs, perhaps performing simple bonding for a single client, or supporting only a fixed calendar configuration for something like a port expander) would not need to fully implement this protocol. Such a configuration would simply transmit the A and B calendar configurations as fixed, always indicate the A calendar configuration as the calendar configuration in use, and would alarm the configuration inconsistency if the received calendar configuration from the far end is not the values expected or if the far end attempts to switch calendars (e.g., sends the calendar in use bits or the CR bit indicating calendar B rather than calendar configuration A).
- An application where a management system or SDN controller has access to the FlexE mux/demux at each end of the FlexE group, that controller may configure the FlexE mux and demux configurations directly and instruct the two ends when to switch calendars. The information sent inband over the FlexE PHYs might just be used as a check for the consistency of the configuration rather than as control for how the FlexE demux is configured.

7.3.5 Management Channel(s)

Certain applications may require use of management channel(s). Two optional management channels are provided on each PHY of the FlexE group:

- A “section” management channel is carried from a FlexE shim to the adjacent FlexE aware node (which may be the far end FlexE shim in a simple router to router connection, or a FlexE aware transport network interface in the case of transport network does not terminate the FlexE group).

- A “shim to shim” management channel which is carried end-to-end between the shims that terminate the FlexE group.

When a management channel is not used, it is transmitted as an Ethernet idle control blocks as illustrated in Figure 12.

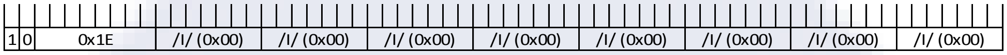


Figure 12: Ethernet Idle Control Block

The format of the management channel is application specific. The section management channel occupies two 66B blocks of each FlexE overhead frame, and the shim to shim management channel occupies three 66B blocks of each FlexE overhead frame. The total capacity of the section management channel is approximately 1.222 Mb/s (not counting the sync headers), or 1.260 Mb/s (counting the sync headers). The total capacity of the shim to shim management channel is approximately 1.833 Mb/s (not counting the sync headers), or 1.890 Mb/s (counting the sync headers). The only constraint on the management channel is that every 66B block is a legal format according to [802.3] clause 82. The protocol used over a management channel may be Ethernet based, using a combination of data and control blocks, or may be any other application-specific format using only data blocks.

A possible application is to carry the LLDP protocol [802.1AB] in the shim-to-shim management channel for verification of the connectivity of FlexE PHYs between FlexE shims, or in the section management channel to verify connectivity of PHYs between a router and transport equipment in a FlexE aware network configuration. The LLDPDUs are 66B block encoded per [802.3] clause 82.2.3.

Each PHY of the FlexE group can carry its own management channels. The management channels are not aggregated across the FlexE group.

7.3.6 FlexE Group Number

A 20-bit FlexE group number is available to allow checking that the correct PHY is being received from the correct group number.

When this field is used, it is normally be provisioned to the same value in both directions. When a non-zero value is provisioned, the received group number will be checked against the provisioned group number and any mismatch will be alarmed to indicate the misconnection.

7.3.7 Reserved Bits

The reserved bits in the FlexE overhead frame are reserved for possible future extensions to this implementation agreement. The reserved bits shall be transmitted as zero before scrambling. An implementation of this version of the IA should ignore these bits on receipt and leave the responsibility to an implementation of a newer version of the implementation agreement to recognize receipt of zeros as an indication of interconnection with an older version, and presumably the newer version knows whether it is interoperable with the older version.

7.3.8 Remote PHY Fault (RPF)

This is used to inform the far-end shim of a locally detected failure of the PHY. Since there is no 100G RS layer per PHY, the FlexE overhead is used to covey this information. See clause 7.5.2.

7.3.9 CRC-16

Primarily to avoid corrupting the content of the calendar configurations in the presence of bit errors, the FlexE overhead is protected by a CRC. The CRC is calculated over the following bits across the first three blocks of the FlexE overhead frame (in the order transmitted and received, not the order described):

- The D1, D2, and D3 bytes of the ordered set in overhead block 1.
- All eight octets after the sync header of overhead block 2.
- The first six octets after the sync header of overhead block 3.

The CRC is calculated using the polynomial $x^{16} + x^{12} + x^{5} + 1$ with an initialization value of zero, where x^{16} corresponds to the MSB and x^{0} corresponds to the LSB. This value is inserted by the FlexE mux into the transmitted overhead with the bit corresponding to x^{15} in the position transmitted first and the bit corresponding to x^{0} transmitted last. Note that while this is the opposite of normal Ethernet bit-transmission order, it is consistent with the order of transmission of the Ethernet FCS. It is calculated by the FlexE demux over the same set of bits and compared to the received value. Various overhead described in the previous clauses is either accepted or ignored based on whether the CRC matches the expected value.

7.4 FlexE Mux Data Flow

The FlexE Mux creates a logically serial stream of 66B blocks by interleaving FlexE clients according to a calendar of length 20n slots for a FlexE group composed of n 100GBASE-R PHYs. Each slot corresponds to 5G of bandwidth. A FlexE client is assigned a number of slots according to its bandwidth divided by 5G. The calendar configuration is distributed as described earlier in Figure 7.

Figure 13 presents an example of insertion of different bandwidth FlexE clients into a logical calendar. The slots assigned to a particular FlexE client do not all need to be on the same PHY of the FlexE group, and new clients can be added as long as there are sufficient slots available.

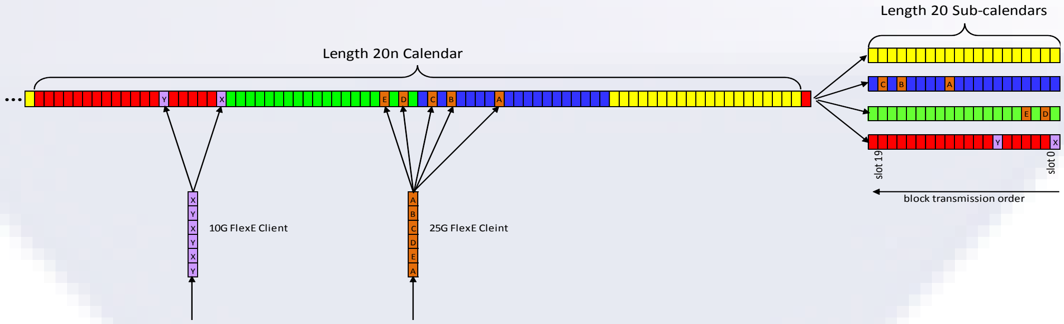


Figure 13: Illustration of Data Flow for FlexE Mux

Any slot in the calendar configuration which is either “unused” or “unavailable” will be filled with Ethernet Error control blocks with the format given in Figure 14. This ensures that any error in calendar slot assignment cannot appear to the FlexE demux as valid FlexE client data.

1	0	0x1E	/E/ (0x1E)	/E/ (0x1E)	/E/ (0x1E)	/E/ (0x1E)	/E/ (0x1E)	/E/ (0x1E)	/E/ (0x1E)	/E/ (0x1E)

Figure 14: Ethernet Error Control Block Format

These rules allow for creation of the complete data sequence on each PHY of the FlexE group. The FlexE overhead as described in clause 7.3 is inserted onto each FlexE group PHY after every 1023 repetitions of the sub-calendar sequence in the same relative position to the calendar sequence on every PHY. This provides a marker which allows the data from the different PHYs of the FlexE group to be re-interleaved in the original sequence so that the FlexE clients can be extracted. The 66B block stream is then converted into the format for the individual FlexE group PHY, which includes block distribution and alignment marker insertion, along with (if applicable) 256B/257B transcoding and FEC calculation and insertion.

7.5 FlexE Demux Data Flow

The FlexE Demux operates on a sequence of 66B blocks received from each PHY of the FlexE group. Recovering this sequence of blocks includes (if applicable), FEC error correction and FEC remove and trans-decoding to 64B/66B, PCS or FEC lane alignment, reinterleaving, and alignment marker removal. Once this has occurred, the PHYs of the FlexE group are re-interleaved so that FlexE clients can be recovered as illustrated in Figure 15.

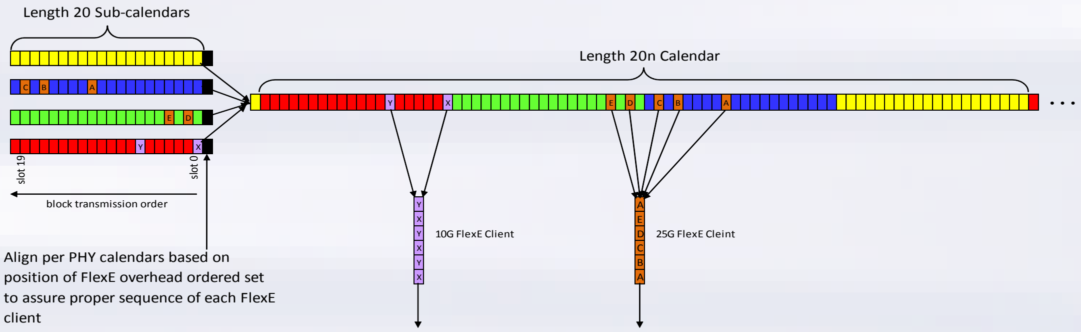


Figure 15: Illustration of FlexE Demux Data Flow

Note that the FlexE overhead frame repeats on a cycle of approximately 104.77μs, which allows measuring skew differences between PHYs of the FlexE group of approximately ±52μs.

7.5.1 Skew Tolerance Requirements

The amount of skew to be expected between the PHYs of the FlexE group are application specific. This implementation agreement specifies skew requirements for two classes of applications.

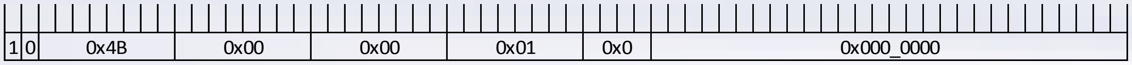
Low Skew Applications include intra-data-center applications, plus those transport network applications where the FlexE shim is implemented in the transport equipment and the FlexE clients rather than the PHYs of the FlexE group are carried across the transport network. The skew tolerance requirement for low skew applications is 300ns. Note that the intra-PCS-lane skew tolerance requirement for 100GBASE-R is 180ns. A larger skew budget is established for FlexE applications of similar reach to account for the fact that the PCS lane deskew is not synchronized across the PHYs of the FlexE group, and there may be other variation, such as cable length, or even heterogeneous 100GBASE-R PHY types which are not present within a single 100GBASE-R interface.

High Skew Applications include the broadest range of transport network applications where the PHYs of the FlexE group rather than the FlexE clients are carried over the transport network (FlexE aware/unaware transport). The skew tolerance for high skew applications may need to be as high as 10µs. This is established to account for about 6µs of dispersion-related skew if the PHYs are mapped over lambdas at opposite ends of the “C” band over large distances (e.g., trans-Pacific), with extra margin for things like split-band amplifiers and patch cords or the processing time to crunch and uncrunch the signal in the case where not all of the calendar slots can be carried over the transport network connection.

7.5.2 FlexE Demux Fault Handling

If the intra-PHY skew exceeds the skew tolerance of the implementation, the FlexE clients will not be demapped from the incoming PHYs, but will be sent continuous Ethernet Local Fault Ordered sets as illustrated in Figure 16 at the FlexE client rate.

If one or more of the PHYs of the FlexE group has failed (e.g., loss of signal, failure to achieve block lock or alignment lock, hi BER, or any other condition that results in PCS_status=FALSE), the Remote PHY fault bit is set to one in the reverse direction on that PHY. Note that this must be done in the FlexE overhead since there is no normal 100GBASE-R RS layer available to indicate remote fault to the far end. In addition, when one or more of the PHYs of the FlexE group have failed, if one or more of the PHYs fails to achieve overhead frame lock or overhead multiframe lock, if there is inconsistency among the PHY maps, PHY numbers, or FlexE group numbers received on the different PHYs of the group, or if the skew between PHYs exceeds the deskew buffer provided by the implementation, an appropriate alarm should be raised and all of the FlexE clients will be sent continuous Ethernet Local Fault Ordered sets as illustrated in Figure 16 at the FlexE client rate.



1	0	0x4B	0x00	0x00	0x01	0x0	0x000_0000

Figure 16: Ethernet Local Fault Ordered Set

7.6 FlexE Group Configuration

When a new FlexE Group is brought into service, the initial configuration must be provisioned from both ends, and the initial configuration must be the same. The group is configured (at least initially) to consist of from 1 to n PHYs. All of the PHYs which are configured as part of the group are brought into service, all transmitting the stream of 66B blocks including the FlexE overhead. The respective bits of the “PHY Map” field are set to indicate which PHYs the near end has configured as part of the group. The group is brought into service when all incoming PHYs are up, receiving FlexE overhead on each of the expected PHYs within the allowable skew tolerance of each other, and the received PHYs all indicate the expected PHYs as being part of the group. When the FlexE Group is first configured and brought into service, all of the calendar slots are normally set to “unused” or “unavailable” (as needed for partial-rate transport configurations), although there may be certain implementations such as a static configuration supporting simple bonding which are brought up with a single FlexE client filling all available calendar slots.

7.7 Energy Efficient Ethernet (EEE)

EEE is not supported on the PHYs of a FlexE group, as there is no good way to verify that every FlexE client is idle and put the entire group into a low power idle mode.

The “Fast Wake” mode of EEE can be supported for a FlexE client. The LPI control characters used during Fast Wake can simply pass through the 66B payload block positions allocated to that client by the calendar configuration, allowing an implementation to be aware when data is not arriving on a given FlexE client.

For FlexE clients generated internally to the system or created from an optical Ethernet PHY, “Fast Wake” is the only mode of operation possible, and negotiation between the FlexE client endpoints (above the FlexE client MAC) to confirm the use of “Fast Wake” will occur via LLDP. For a FlexE client generated from a copper Ethernet PHY, the AN capabilities should always indicate “No” for both “Deep Sleep” and “Fast Wake”, and the “Fast Wake” mode can then be enabled via LLDP negotiation.

8 Transport Network Mappings for Flex Ethernet Signals

Three different methods of mapping of FlexE signals over transport networks are possible.

8.1 FlexE Unaware Transport

The case of FlexE unaware transport involves the transport network mapping each of the Ethernet PHYs independently over the transport network using the existing PCS codeword transparent mapping. Since the FlexE mux and FlexE demux are separated by transport network distances, this requires a “high skew” implementation of the FlexE shim as described in clause 7.5.1.

Note that certain existing OTN mappers/demappers are not fully PCS codeword transparent with respect to LF and RF ordered sets, and may mistake an LF or RF sent for an individual FlexE client as an LF or RF for the entire link and bring the link down. This is not an issue in the case of FlexE unaware transport of simple bonding to carry a larger rate flow, as failure of the single high-rate FlexE client is equivalent to a failure of the group. But it may be an issue if FlexE unaware transport is used to carry a group of multiple lower-rate FlexE client services using less than fully PCS codeword transparent mappings.

8.2 FlexE termination in the Transport

The next case is where the FlexE shim is terminated by the transport network equipment, and rather than carrying the PHYs of the FlexE group over the transport network, the FlexE clients are carried over the transport network. The rate-adapted FlexE client bit-rate is given in clause 6.2.

Note that since the stream of blocks presented to the transport network from the FlexE shim does not have any timing information, the transport network is not required the signal at the adapted FlexE client bit-rate: idle insertion/deletion or padding may be used in the mapping specified by ITU-T if it provides a more convenient rate for the transport.

When a FlexE client is mapped in this manner, it may be connected, at the OTN egress, to another FlexE shim where it will be clock aligned with the FlexE group at the network egress. It may also be connected to an Ethernet PHY with the same nominal MAC rate as the FlexE client using the appropriate conversion as described in clause 7.2.2.

Note that in the case where it is necessary to connect a FlexE client to an Ethernet PHY across an OTN where the Ethernet PHY uses a legacy mapper/demapper, it may be necessary to perform the conversion of the FlexE client to the Ethernet PHY format according to clause 7.2.2 immediately after the FlexE shim and to map the FlexE client over OTN as if it were an Ethernet PHY of the corresponding rate.

This second case can use a “low skew” implementation of the FlexE shim as described in clause 7.5.1.

8.3 FlexE Aware Transport

The third case is where the transport network equipment is aware it is carrying PHYs of a FlexE group, but does not terminate the FlexE group in the transport network equipment.

The transport network equipment terminates the section management channel on each of the FlexE PHYs, extracting (but not forwarding) this channel at the network ingress and inserting this channel at the network egress. At the network ingress, any content of the section management channel is replaced with Ethernet idle control blocks (see Figure 12) into the transport network.

This may be used to support cases where the Ethernet PHY rate is be greater than the wavelength rate, the wavelength rate is not an integral multiple of the PHY rate, or there is a reason (for example, wavelengths terminated on different transponder line cards) that it is not possible to terminate the FlexE shim in the transport equipment. In cases where the Ethernet rate is greater than the wavelength rate or is not an integral multiple of the wavelength rate, the transport network equipment may “crunch” the PHY of the FlexE group by allowing bits or bytes to be discarded from the unavailable calendar slots at the transport network ingress and these bits or bytes re-inserted with fixed values at the transport network egress. The mapping of this requires serializing and deskewing the PCS lanes of the PHY, then discarding from the “UNAVAILABLE” calendar slots to reduce the bit-rate. For example, if only 15 of 20 calendar slots are available in a sub-calendar for a given PHY, there are effectively 1023 repetitions of a length 15 calendar after discarding the unavailable slots included in the information that must be transported. At the transport network egress, the bits or bytes removed from the unavailable slots are restored to the FlexE PHY stream of 66B blocks so that error control blocks occur in every unavailable slot as illustrated in Figure 14. The net bit-rate of this reduced-rate flow (the information that must be transported) when there are “n” available calendar slots on a given PHY is:

$$103.125\ \mathrm{Gb/s} \times \frac{16383}{16384} \times \frac{1+1023n}{20461} \pm 100\mathrm{ppm}$$

It is expected that the granularity of partial-rate transport is 25Gb/s. Note that what is specified in this IA is only the information to be carried over the OTN, and how it is mapped is over transport networks is specified by ITU-T Q11/15.

The rates of the information that must be transported of the 64B/66B flows for a given number of available calendar slots to be carried over the transport network are given in Table 1.

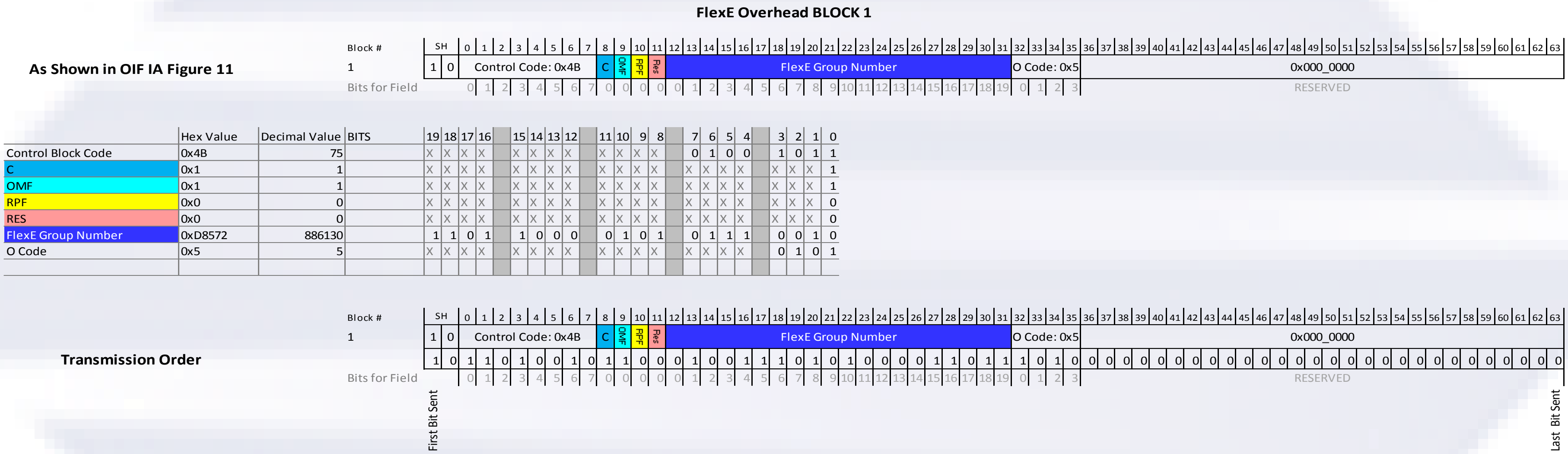
Table 1: 64B/66B Rates given number of Available Calendar Slots

PHY Available Slots	64B/66B Flow rate ±100ppm
5	25.78345626
10	51.56187276
15	77.34028925
20	103.1187057

As described in clause 6.4, unavailable slots are always at the end of the sub-calendar configuration for the respective PHY. It is expected that since the rate of a wavelength isn’t expected to change in-service, when a partial-rate signal is carried over the OTN, the mapper is statically configured to drop bits or bytes from a certain number of calendar slots at the ingress and to restore those same bits or bytes to the calendar slots at the egress to contain Ethernet error control blocks as described in clause 7.4 in the payload block positions for each unavailable calendar slot. The OTN mapper is not expected to dynamically react to which slots are marked as unavailable in the calendar configurations, but may non-intrusively monitor the FlexE overhead and detect as an error condition if a calendar slot the mapper has been configured to drop indicates that it is carrying FlexE client data rather than being marked as unavailable.

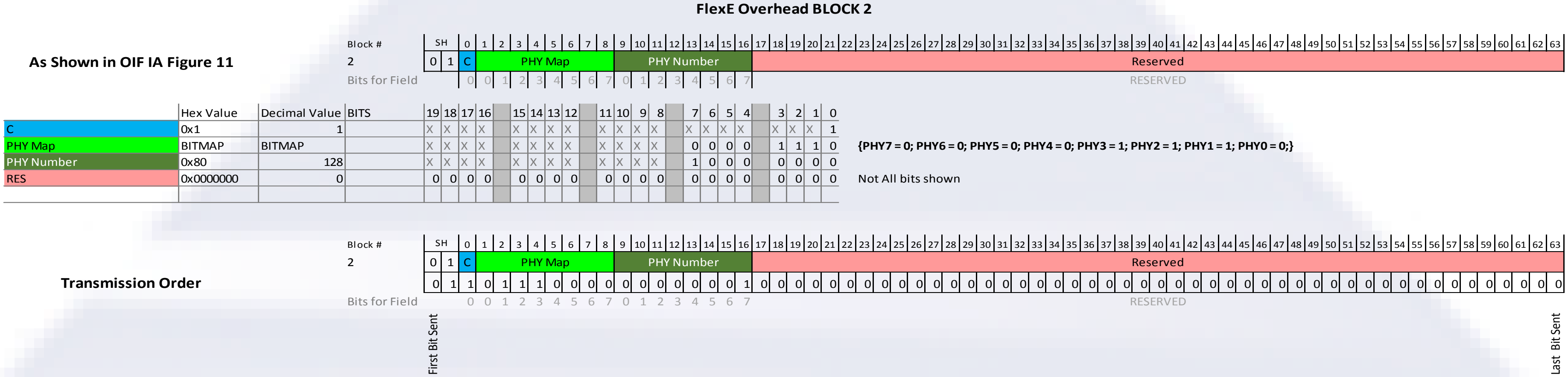
9 Appendix A: Test Vectors

Test vectors for the first three overhead blocks of the FlexE overhead frame are given in Figure 17, Figure 18 and Figure 19. Note that the last five overhead blocks of the FlexE overhead frame form the two management channels.



	Hex Value	Decimal Value	BITS	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Control Block Code	0x4B	75		X	X	X	X	X	X	X	X	X	X	X	X	0	1	0	0	1	0	1	1
C	0x1	1		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
OMF	0x1	1		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
RPF	0x0	0		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0
RES	0x0	0		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0
FlexE Group Number	0xD8572	886130		1	1	0	1	1	0	0	0	0	1	0	1	0	1	1	1	0	0	1	0
O Code	0x5	5		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0	1	0	1

Figure 17: Test Vector for first block of FlexE overhead frame



	Hex Value	Decimal Value	BITS	19	18	17	16		15	14	13	12		11	10	9	8		7	6	5	4		3	2	1	0	
C	0x1	1		X	X	X	X		X	X	X	X		X	X	X	X		X	X	X	X		X	X	X	1	
PHY Map	BITMAP	BITMAP		X	X	X	X		X	X	X	X		X	X	X	X		0	0	0	0		1	1	1	0	{PHY7 = 0; PHY6 = 0; PHY5 = 0; PHY4 = 0; PHY3 = 1; PHY2 = 1; PHY1 = 1; PHY0 = 0;}
PHY Number	0x80	128		X	X	X	X		X	X	X	X		X	X	X	X		1	0	0	0		0	0	0	0	
RES	0x0000000	0		0	0	0	0		0	0	0	0		0	0	0	0		0	0	0	0		0	0	0	0	Not All bits shown

Figure 18: Test Vector for second block of FlexE overhead frame

FlexE Overhead BLOCK 3

As Shown in OIF IA Figure 11

Block #	SH		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
3	0	1	C	Client Calendar A																Client Calendar B																CR	CA	Reserved													CRC-16															
Bits for Field			0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	0	0	RESERVED													15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

	Hex Value	Decimal Value	BITS	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
C	0x1	1		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
Client Calendar A	0xD647	54855		X	X	X	X	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1	1
Client Calendar B	0x6A74	27252		X	X	X	X	0	1	1	0	1	0	1	0	0	1	1	1	0	1	0	0
CR	0x0	0		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0
CA	0x0	0		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0
CRC-16	0x8563	34147		X	X	X	X	1	0	0	0	0	1	0	1	0	1	1	0	0	0	1	1

CRC is sent X15 first as requested by the IA

Transmission Order

Block #	SH		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
3	0	1	C	Client Calendar A																Client Calendar B																CR	CA	Reserved													CRC-16															
	0	1	1	1	1	1	0	0	0	1	0	0	1	1	0	1	0	1	1	0	0	1	0	1	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	1	1	0	0	0	1	1
Bits for Field			0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	0	0	RESERVED													15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

First Bit Sent

Last Bit Sent

Figure 19: Test Vector for third block of FlexE overhead frame

10 References

10.1 Normative references

[802.3]	IEEE Std 802.3^{TM}-2015 *Standard for Ethernet*.
[G.709]	ITU-T Recommendation G.709 (06/2016), *Interfaces for the Optical Transport Network*.
[802.1AB]	IEEE Std 802.1^{TM}AB-2016, Station and Media Access Control Connectivity Discovery.

11 Appendix B: List of companies belonging to OIF when document was approved

Acacia Communications
ADVA Optical Networking
Alibaba (China) Co. Ltd.
Amphenol Corp.
Anritsu
Broadcom Limited
Brocade
BRPhotonics
Cavium
China Telecom
Ciena Corporation
Cisco Systems
Coriant
Corning
Credo Semiconductor (HK) LTD
Dell, Inc.
Elenion Technologies, LLC
Fiberhome Technologies Group
Finisar Corporation
Foxconn Interconnect Technology, Ltd.
Fujikura
Fujitsu
Furukawa Electric Japan
Gigamon Inc.
Global Foundries
Google
Hewlett Packard Enterprise (HPE)
Hitachi
Huawei Technologies Co., Ltd.
Infinera
Inphi
Integrated Device Technology
Intel
Invecas
Ixia
Juniper Networks
Kandou Bus
KDDI Research, Inc.
Keysight Technologies, Inc.
Lumentum
MACOM Technology Solutions
Marvell Technology
MaxLinear Inc.
Mellanox Technologies
Microsemi Inc.
Microsoft Corporation
Mitsubishi Electric Corporation
Molex
MoSys, Inc.
MRV
NEC Corporation
NeoPhotonics
Nokia
NTT Corporation
O-Net Communications (HK) Limited
Oclaro
Orange
PETRA
Qorvo
Ranovus
Rianta Solutions, Inc.
Rockley Photonics
Samsung Electronics Co. Ltd.
Samtec Inc.
Semtech
SiFotonics Technologies Co., Ltd.
Silab Tech Private Ltd.
Socionext Inc.
Spirent Communications
Sumitomo Electric Industries
Sumitomo Osaka Cement
TE Connectivity
Tektronix
Teledyne LeCroy
TELUS Communications, Inc.
UNH InterOperability Laboratory (UNH-IOL)
Verizon
Viavi Solutions Deutschland GmbH
Xilinx
Yamaichi Electronics Ltd.
ZTE Corporation