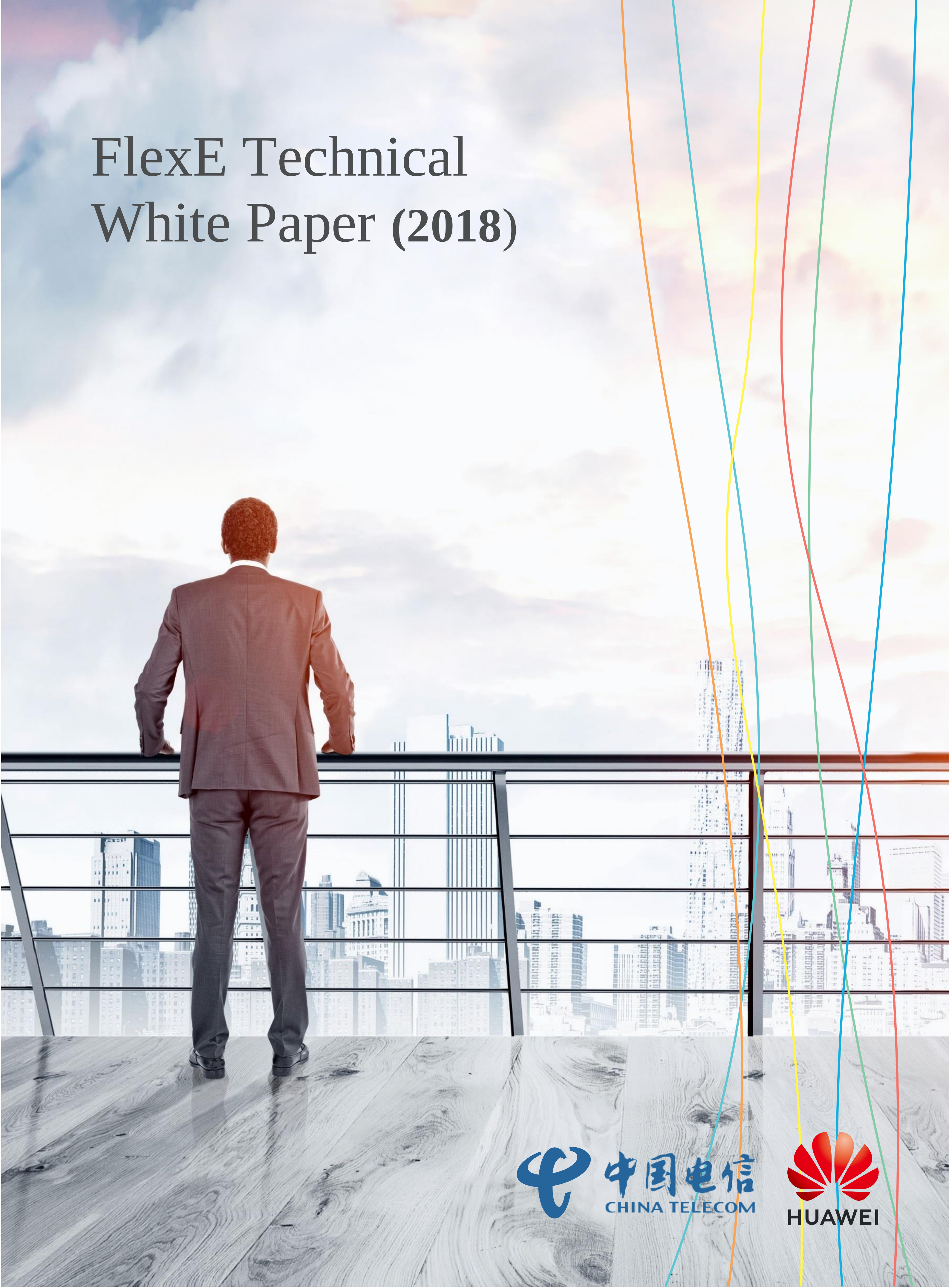
FlexE Technical White Paper (2018)



中国电信 CHINA TELECOM

HUAWEI

1 Overview

Based on high-speed Ethernet interfaces, Flexible Ethernet (FlexE) is a cost-efficient carrier-grade interface technology providing high reliability and dynamic configuration through decoupling of the Ethernet Physical layer (PHY) and Media Access Control layer (MAC) rates. Leveraging the most widely used and powerful Ethernet ecosystem, FlexE addresses the challenges in developing services such as video streaming, cloud computing, and 5G, attracting wide attention from the industry since it was proposed in 2015.

2 Birth of FlexE

Built on Ethernet technologies, FlexE meets requirements for high-speed transmission and flexible bandwidth configuration.

Ethernet was first proposed by Xerox in 1972 and was gradually improved on the basis of CSMA/CD technology. Since the 1980s, the development of Ethernet has been standardized by IEEE 802.3/1. Ethernet developed rapidly to meet industry service demand and has become the most widely used OSI Layer 2 (L2) interconnection technology, and arguably the most complete ecosystem in the IT industry.

Ethernet technologies comply with MAC/PHY layer standards defined by IEEE 802.3 at the interface layer. Prior to 2010, Ethernet was developing in iterations of approximately tenfold bandwidth increases, from 10 Mbit/s, 100 Mbit/s, 1 Gbit/s, 10 Gbit/s, 40 Gbit/s, to 100 Gbit/s. However, in recent years, propelled by higher service requirements and the development of technologies like serializing/deserializing circuitry (SERDES), Ethernet supports evolution from 25 Gbit/s, 50 Gbit/s, 200 Gbit/s, 400 Gbit/s, to 800 Gbit/s. Figure 1 shows the evolution of Ethernet interfaces (see reference [1]).

The wide application of Ethernet interface technologies has driven development and improvement of Carrier Ethernet technology on operators' metro networks and WANs since 2000. Carrier Ethernet provides high reliability, operability, and maintainability for operators' networks. Organizations including MEF, IEEE, and BBF have formulated standards on Carrier Ethernet, empowering it with carrier-grade functions such as OAM, protection switching, high-performance clock, and QoS/QoE guarantee. This has allowed Carrier Ethernet to be widely deployed on metro networks, WANs, mobile backhaul networks, and private lines.

With the rise of cloud computing, video, and mobile communication services, requirements on IP networks have gradually shifted from bandwidth to service experience, service quality, and networking efficiency. To meet these requirements, Ethernet, the underlying connection technology, must develop the following capabilities in addition to its existing advantages of cost-efficiency, high reliability, and easy O&M:

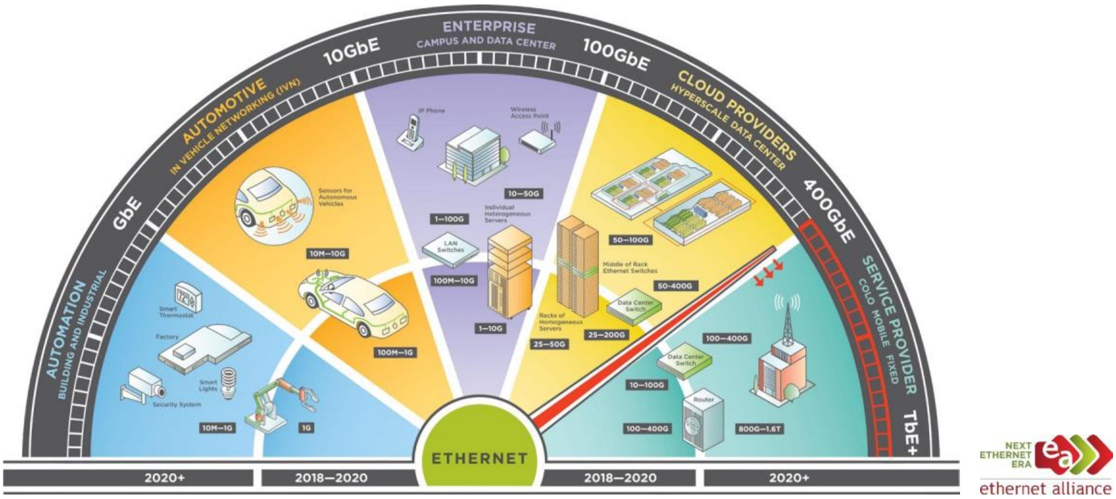


Figure 1 Evolution of Ethernet interfaces released by Ethernet Alliance

- **Multi-granularity and flexible bandwidth adjustment**: As service and application scenarios become increasingly diverse, the industry expects Ethernet interfaces to provide more flexible bandwidth granularities, without being confined to the tiered bandwidth system of 10 Gbit/s, 25 Gbit/s, 40 Gbit/s, 50 Gbit/s, and 100 Gbit/s defined in IEEE 802.3. The industry even seeks ultra-high-speed 800 Gbit/s and 1.6 Tbit/s Ethernet interfaces. However, the standards for such interfaces have yet to be formulated, calling for alternative solutions.
- **Decoupling Ethernet interface capabilities and optical transmission capabilities**: Ethernet interface capabilities and optical transmission device capabilities are not developing in pace with each other. On an IP network, a device's high-speed Ethernet interfaces are often constrained by the optical transmission capability of the network. If Ethernet interface rates are decoupled from optical transmission network rates (in other words, the DWDM link rate of an optical transmission network is not strictly required to match the rate of an Ethernet UNI interface), existing optical transmission networks can be maximally used to transmit traffic on ultra-high bandwidth Ethernet interfaces.
- **IP and optical converged networking**: By decoupling Ethernet interface capabilities and transmission capabilities, networking can be simplified and given more flexibility, using simple mappings between Ethernet interfaces and optical transmission networks. (This applies to the cross-region inter-IDC networking scenario, which is also the first usage scenario of FlexE.) Moreover, flexible traffic grooming and scheduling optimization can be achieved (see reference [2]).
- **Enhanced QoS capabilities for multi-service bearer**: Enhanced user experience in multi-service bearer scenarios is a development goal of high-speed Ethernet technologies. If Ethernet can provide channelized hardware isolation on physical-layer interfaces, services can be isolated by slicing at the physical layer. In addition, Ethernet can work with upper-layer networks or applications, in combination with high-performance programmable forwarding and hierarchical QoS scheduling, to enhance QoS capabilities in multi-service bearer scenarios.

FlexE is developed to meet these requirements.

3 Key Technical Implementation of FlexE

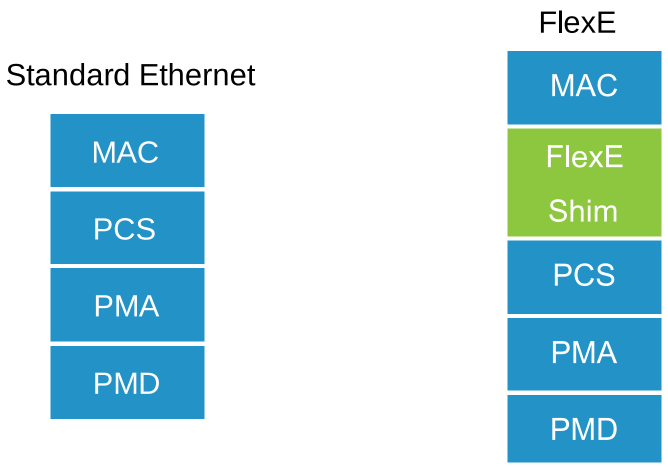


Figure 2 Structures of standard Ethernet and FlexE

Based on IEEE 802.3, FlexE introduces the FlexE Shim layer to decouple the MAC and PHY layers (as shown in Figure 2), achieving flexible rate matching.

FlexE uses the client/group architecture, where multiple FlexE Clients can be mapped to a FlexE Group for transmission, implementing bonding, channelization, sub-rating, and other functions (see reference [3]).

- **FlexE Client**: an Ethernet flow based on a MAC data rate that may or may not correspond to any Ethernet PHY rate. FlexE Clients correspond to various user interfaces that function in the same way as traditional service interfaces on existing IP/Ethernet networks. They can be configured flexibly according to bandwidth requirements. They support a variety of Ethernet MAC rates (for example, 10 Gbit/s, 40 Gbit/s, N x 25 Gbit/s, and even non-standard rates) and can be transmitted to the FlexE Shim layer as 64B/66B-encoded bit streams.
- **FlexE Shim**: a layer that maps or demaps the FlexE Clients carried over a FlexE Group. It decouples the MAC and PHY layers and implements key functions of FlexE through the calendar slot distribution mechanism.
- **FlexE Group**: a group composed of from 1 to N Ethernet PHYs. Because FlexE inherits IEEE 802.3-defined Ethernet technology, the FlexE architecture provides enhanced functions based on existing Ethernet MAC and PHY rates.

In a FlexE point-to-point connection scenario, multiple Ethernet PHYs are bound to a FlexE Group to carry one or more FlexE Clients distributed and mapped through the FlexE Shim.

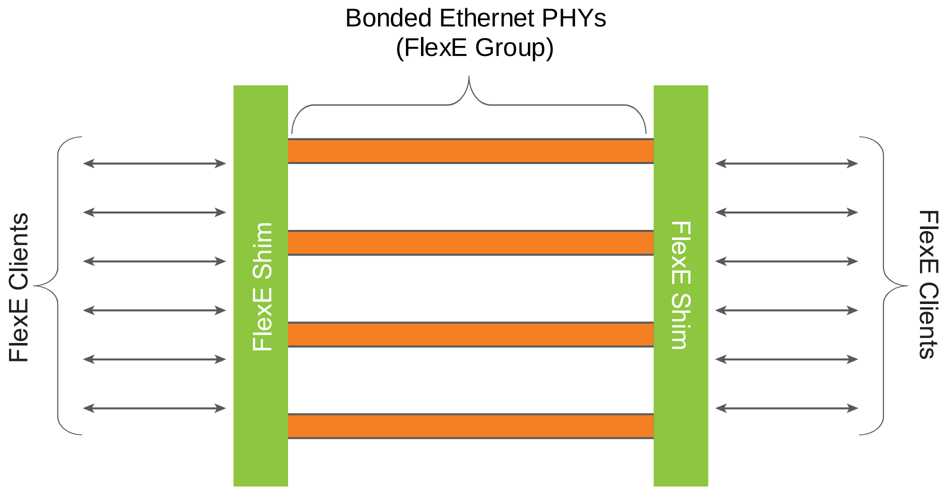


Figure 3 General structure of FlexE

The FlexE Shim implements the core functions of FlexE. It partitions each 100G PHY in a FlexE Group into a group, called a sub-calendar, of 20-slot data channels. The sub-calendar provides 5 Gbit/s bandwidth per slot. The Ethernet frames of FlexE Clients are partitioned into 64B/66B blocks, which are distributed to multiple PHYs of a FlexE Group based on slots through the FlexE Shim.

According to OIF FlexE standards, the bandwidth of each FlexE Client can be set to 10, 40, or N x 25 Gbit/s. Because the bandwidth of each slot of a 100GE PHY in a FlexE Group is 5 Gbit/s, a FlexE Client theoretically supports multiple rates by combining any number of slots.

The calendar mechanism enables the FlexE Shim to map and carry FlexE Clients with different rates in a FlexE Group and to allocate bandwidth to the clients.

FlexE allocates available slots in a FlexE Group according to the bandwidth required by each FlexE Client and the distribution of slots in each PHY, mapping each client to one or more slots. FlexE then uses the calendar mechanism to enable a FlexE Group to carry one or more FlexE Clients. Each 64B/66B-encoded block in a FlexE Client is carried over a slot (a basic logical unit carrying the 64B/66B block), as shown in Figure 4. In the calendar mechanism, FlexE uses every 20 blocks (slots 0 to 19) as a logical unit (a sub-calendar) represented by the green data blocks in Figure 4.

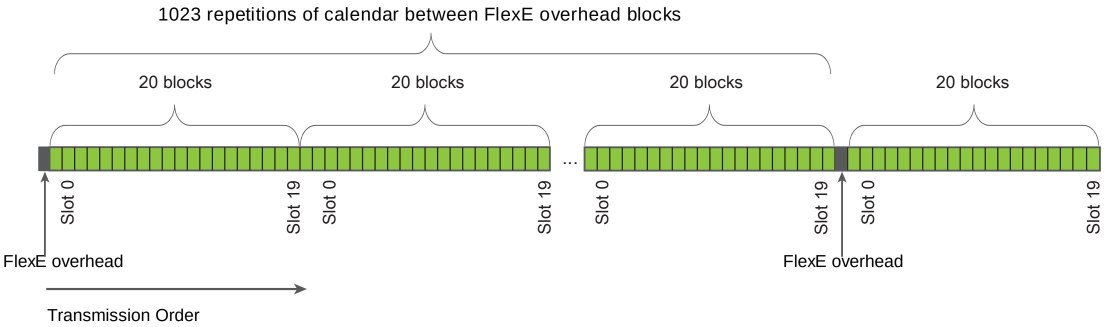


Figure 4 Frame structure of FlexE

Overhead frames and multiframes are defined for the FlexE Shim to configure slot mappings and implement the calendar mechanism. The FlexE Shim provides inband management channels through overhead, allowing configuration and management information to be transmitted between two interconnected FlexE interfaces to automatically set up links through auto-negotiation. Specifically, one overhead multiframe consists of 32 overhead frames, each of which contains eight overhead slots, depicted by the black blocks in Figure 4. An overhead slot, which is a 64B/66B block, occurs once every 1023 repetitions of 20 blocks. Overhead slots have different fields. An ordered set with block type code 0x4B and O code 0x5 marks the first block of an overhead frame.

After the overhead frame and multiframe are locked, an inband communication channel is established to carry configuration and management information between two FlexE interfaces. For example, after a FlexE Client's slot mapping information is sent from the transmit end to the receive end through this communication channel, the receive end can restore the FlexE Client based on this information. FlexE inband management also allows interconnected interfaces to exchange link state information and OAM information, such as remote PHY fault (RPF) information.

FlexE achieves dynamic bandwidth adjustment for clients by allowing slot/calendar configurations to be modified. To reflect FlexE Client mappings in a FlexE Group, interconnected interfaces use an overhead management channel to transmit two calendar configurations (A and B configurations, represented by "0" and "1", respectively). A and B calendar configurations can be dynamically switched to one another to achieve bandwidth adjustment.

Because a FlexE Client may have different bandwidth values in A and B calendar configurations, with the help of higher level system application control, client bandwidth can be adjusted seamlessly by calendar switching. The overhead frame provides a request/acknowledge mechanism for switching between the two configurations.

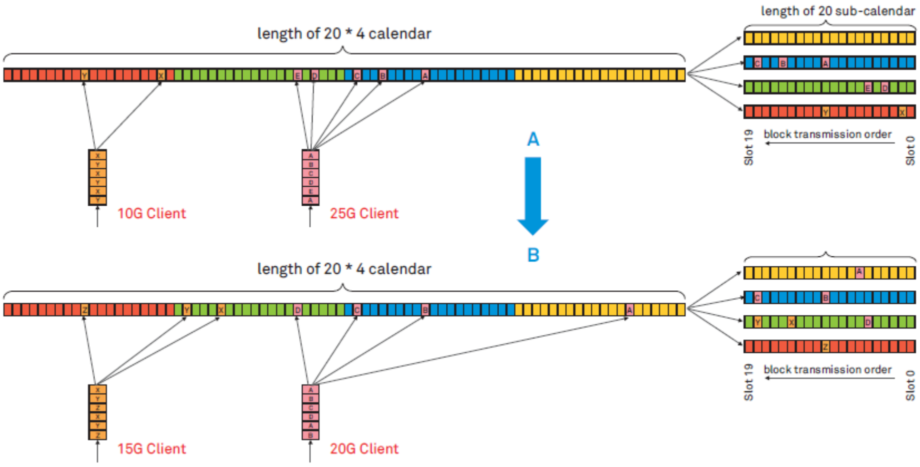


Figure 5 Bandwidth adjustment through FlexE calendar configuration switching

4 FlexE Applications

FlexE Clients can flexibly provide upper-layer applications with bandwidth unconstrained to the rates of Ethernet PHYs. Figure 6 shows the three key features provided by FlexE from the perspective of client-group mapping:

- **Bonding**: This allows multiple links to be bonded into a higher-speed link. For example, eight 100G PHYs can be bonded to support an 800G MAC.
- **Channelization**: This allows one PHY or a collection of bonded PHYs to carry several lower-rate MACs. For example, one 100G PHY can carry four MACs with rates of 25G, 35G, 20G, and 20G, and a collection of three 100G PHYs bonded together can carry three MACs with rates of 125G, 150G, and 25G.
- **Sub-rating**: This allows a single lower-rate MAC to be carried over one PHY or a collection of bonded PHYs, by filling the empty slots with Ethernet error control blocks. For example, a 50G MAC can be carried over a 100G PHY.

Sub-rating can be regarded as a subset of channelization. This function allows FlexE interfaces connecting to an optical transport network to match DWDM link rates through a simple mapping process. Specifically, if the rate of a MAC is lower than the rate of a PHY, the FlexE overhead frame marks the unused timeslots as unavailable and fills in error control blocks in the corresponding timeslots of the calendar. In FlexE aware transport mode, the optical transport equipment discards the timeslots marked as unavailable.

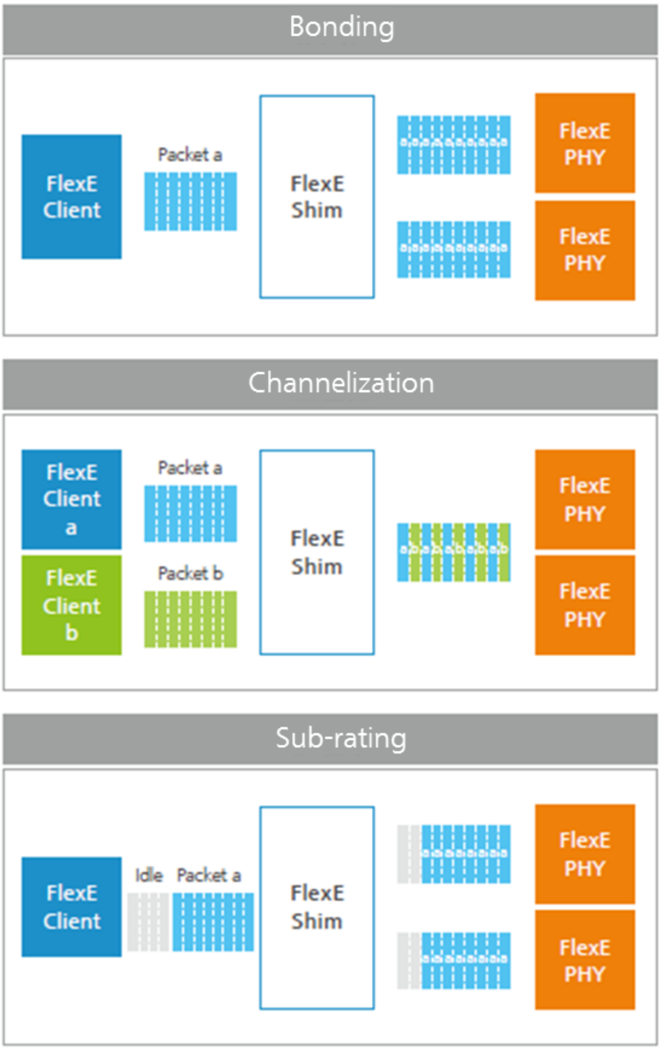


Figure 6 FlexE functions

4.1 Using FlexE Bonding to Provide Ultra-High Bandwidth Interfaces

Converged transport has become popular in IP network construction. New services and applications keep emerging, but IP networks featuring statistical multiplexing still follow the pace of standardization in the communications industry. As a result, IP networks face growing challenges in on-demand, fast networking, flexible resource configuration, and dedicated resource guarantee.

To overcome these challenges, FlexE provides functions such as link bonding, network slicing, and interface channelization to ensure on-demand bandwidth allocation, hard pipe isolation, and low latency guarantee on IP networks. FlexE can also work with SDN technologies to support the future service experience-centric network architecture, meeting the development requirements of future bandwidth-intensive services, such as video, AR/VR, and 5G.

Currently, FlexE makes inroads into scenarios requiring ultra-high bandwidth interfaces, IP+optical synergy, VIP private lines, or network slicing.

Ethernet standardization in the IEEE 802.3 working group is driven by service requirements and technology development and has some periodicity. In addition, fixed Ethernet standard rates (such as 10GE and 100GE) cannot provide bandwidth flexibility for networking. It is essential to use FlexE to bond links for higher bandwidth (such as 5x100GE or 10x100GE).

FlexE bonding is essentially an L1 link aggregation technology. Because FlexE bonds links based on fine-sliced 64B/66B blocks, it does not suffer traffic imbalance caused by per-flow or per-packet traffic distribution among physical links in a traditional link aggregation group (LAG[1]). Compared with LAG, FlexE provides completely balanced bandwidth allocation and reduces bandwidth waste by 10% to 30%.

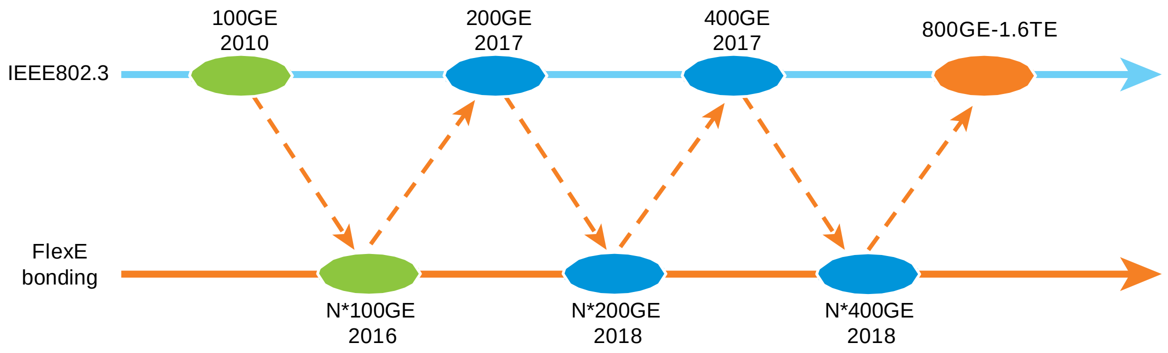


Figure 7 Relationship between IEEE 802.3 standardization periodicity and FlexE bonding evolution

[1] LAG combines a number of physical Ethernet interfaces into a logical interface to provide higher cumulative bandwidth and link protection. Currently, the widely used link aggregation standard is IEEE 802.1ax. Data flows are allocated to different physical Ethernet interfaces in a LAG based on a load-balancing algorithm (usually a hash algorithm). Due to bandwidth and behavior uncertainty of data flows, load imbalance exists among these physical Ethernet interfaces, and the bandwidth cannot be fully utilized.

4.2 Using FlexE for Flexible IP+Optical Networking

When being deployed on UNIs connecting routers to optical transport equipment, FlexE can map data flows on UNIs to the WDM links of NNIs on the optical transport equipment in a one-to-one manner. This greatly simplifies the mapping process and reduces device complexity, reducing CAPEX and OPEX.

The OIF FlexE standard defines three methods of FlexE signal mapping over transport networks: unaware, termination, and aware (see reference [3]).

The unaware mode uses the PCS codeword transparent mapping used by traditional Ethernet interfaces on the optical transport network. This is similar to using the optical transport network to transparently transmit data from FlexE interfaces. The unaware mode provides FlexE support without requiring hardware upgrade, fully utilizing legacy optical transport equipment. In addition, it can use FlexE bonding to provide E2E ultra-high bandwidth channels across optical transport networks.

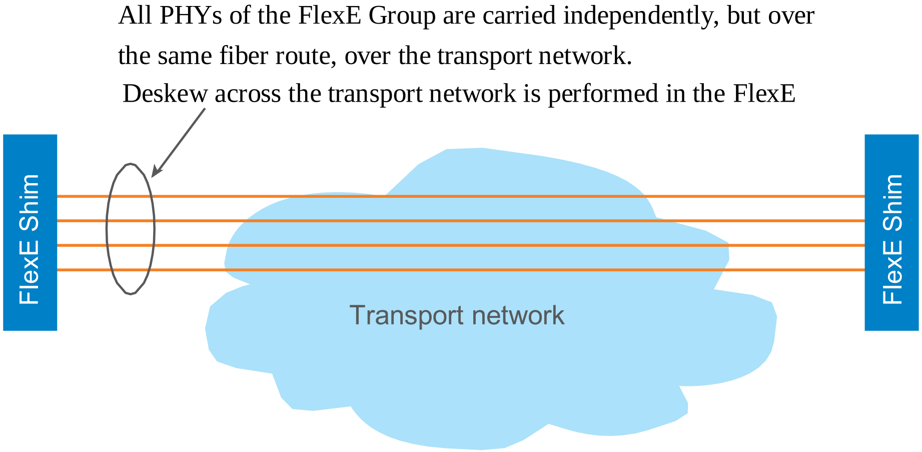
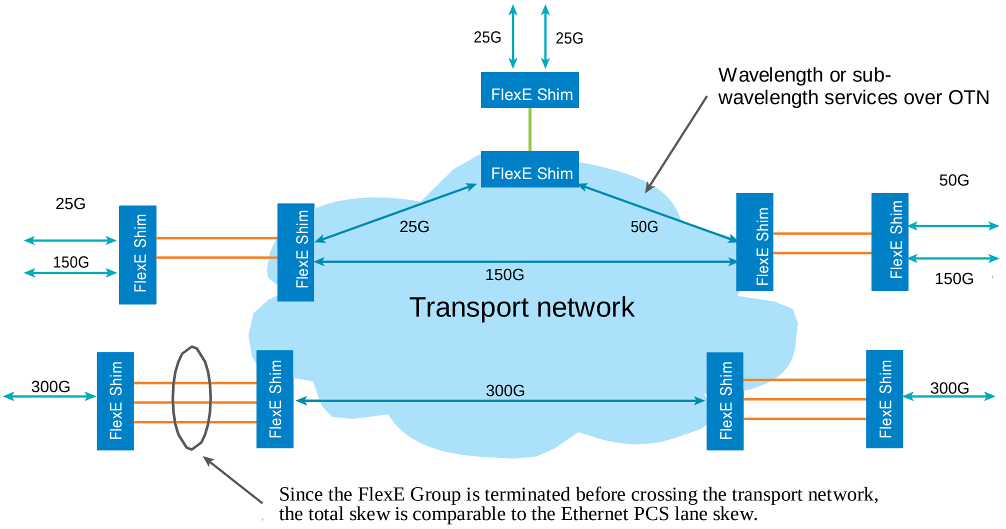

All PHYs of the FlexE Group are carried independently, but over the same fiber route, over the transport network.
Deskew across the transport network is performed in the FlexE
FlexE Shim
Transport network
FlexE Shim

Figure 8 Router-to-FlexE-unaware transport network connection

In termination mode, the optical transport equipment is aware of FlexE UNIs and can restore FlexE Clients and map them to the optical transport network for transmission. In this mode, FlexE interfaces function in the same way as traditional Ethernet interfaces to divert FlexE Clients to the optical transport network.

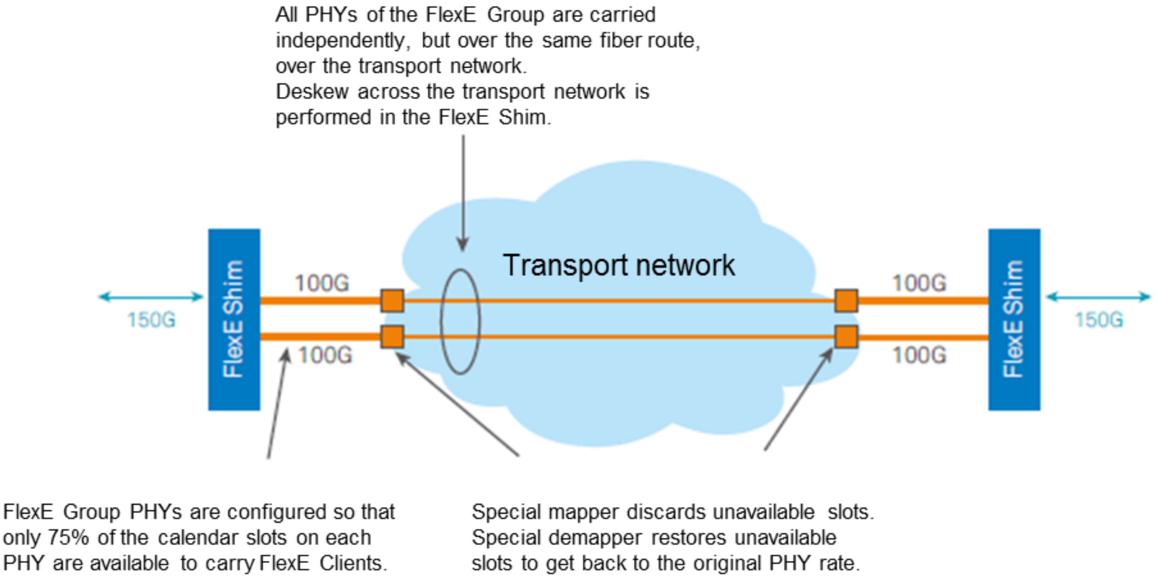


Wavelength or sub-wavelength services over OTN

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

Figure 9 FlexE-terminating transport network equipment

The aware mode mainly uses FlexE sub-rating. In this mode, FlexE identifies unavailable slots using special error control blocks. When a FlexE UNI maps data flows to the optical transport network in aware mode, it directly discards unavailable slots, extracts desired data according to the bandwidth of the original data flows, and then maps these data flows to DWDM links with matching rates on the optical transport network. The configurations on the optical transport equipment must be consistent with FlexE UNIs, allowing the equipment to be aware of FlexE UNIs and forward their data flows.



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.

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.

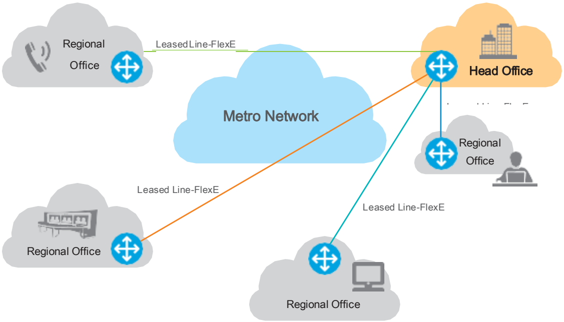
Figure 10 Example of FlexE aware transport of Ethernet PHYs of a FlexE Group

4.3 Using FlexE to Provide Hard Pipe VIP Private Lines

With hard pipe technology, IP networks can construct E2E hard pipes through FlexE channelization to carry VIP private lines and latency-sensitive services. On an IP network characterized by statistical multiplexing, E2E FlexE-based hard pipe private lines can fully utilize the existing network infrastructure to provide quality guarantee for certain high-value customer services.

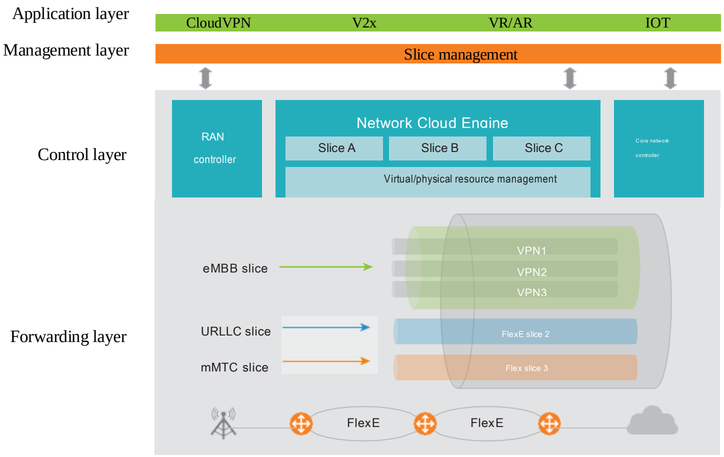
The Ethernet virtual private lines (EVPLs) have been widely used on enterprise networks and metro networks, especially for interconnecting geographically dispersed areas (such as enterprise headquarters and branches). The ever-increasing network services pose increasingly higher requirements for private line quality. For example, some services require exclusive bandwidth and extremely low latency, while other services demand strong privacy protection and high security. FlexE-based private lines can well meet these new requirements.

Figure 11 illustrates the application of FlexE on an enterprise network with relatively scattered geographical locations. Connections between offices in different regions are established through FlexE and can provide bandwidth according to data traffic requirements.



4.4 Using FlexE to Provide 5G Network Slicing

Network slicing divides network resources to meet the transport requirements of different services and guarantee SLA compliance (such as satisfying bandwidth and latency requirements). According to the *NGMN 5G White Paper* (see reference [4]), network slicing allows an IP network to carry diversified services, such as eMBB, autonomous driving, URLLC, and mMTC services. FlexE channelization provides physical division and isolation between FlexE Clients at the interface level and can construct E2E network slices based on the router architecture.



Management layer

» Each slice has an independent configuration and management UI.

» A slice can provide extended network functions as required.

Control layer

» Each slice has an independent network topology. The resource allocation modes and even control protocols are determined by the corresponding slice instance.

Forwarding layer

» FlexE is used to provide E2E physical-layer service isolation and differentiated SLA assurance.

Figure 12 FlexE-based 5G network

5 FlexE Standardization and Technology Development

The Optical Interworking Forum (OIF) launched the FlexE standard in early 2015 and released FlexE IA 1.0 in 2016. This standard defines support for 100GE PHYs. It is the first industry standard of FlexE and has attracted wide attention since being released. The OIF has also released FlexE IA 2.0 (see reference [5]). FlexE IA 2.0 now supports 200GE/400GE PHYs, and maintains the multiplexing frame format compatible with FlexE IA 1.0 and the padding mechanism for 100/200/400GE PHY rate adaptation. Additionally, FlexE IA 2.0 supports IEEE 1588v2 time synchronization in mobile backhaul application scenarios.

In addition to the OIF, standards organizations such as the ITU Telecommunication Standardization Sector (ITU-T), Internet Engineering Task Force (IETF), and Broadband Forum (BBF) have started FlexE standardization.

- » The ITU-T Q11/15 and Q13/15 work groups are defining the mapping of the FlexE unaware, terminate, and aware modes on OTNs, which will be released through a supplementary version of the G.709 standard. The mapping of the FlexE unaware mode references the PCS codeword transparent transmission mode of 100GBASE-R on OTNs. In terminate mode, the existing transmission equipment carries Ethernet data, and the idle/padding mechanism can be used to adjust transmission rates. The mapping of the aware mode on OTNs is implemented through the latest idle mapping mechanism. The rate of client data flows on the UNI side and the DWDM link rate can be adjusted. A mechanism for FlexE time and frequency synchronization in the OTN mapping is also being discussed (see reference [6]).
- » The BBF launched the *Network Services in IP/MPLS Network using Flex Ethernet* standard project (see reference [7]) in May 2017. It defines how to implement the enhanced QoS function architecture through FlexE interfaces on IP/MPLS networks, and how to be compatible with tunneling technology that supports FlexE interfaces based on the existing networks to better bear bandwidth and latency guaranteed services. During the BBF conference in 2017, multiple FlexE-based proposals were accepted. These proposals include technical solutions and architectures for deploying FlexE on IP/MPLS networks, FlexE-based network slicing, and more flexible path provisioning/management based on Segment Routing.
- » The IETF has started to formulate the FlexE control plane standard. The objective is to extend FlexE from interface technology to end-to-end technology, providing port-level hardware-based isolation based on IP/MPLS. This would support solutions such as network slicing and VIP private line. Currently, the IETF focuses on the FlexE framework, mainly involving the architecture and scenarios of end-to-end FlexE, and the signaling and routing protocols that need to be improved/extended to implement its paths. Signaling extension focuses on RSVP-TE and Segment Routing. Routing protocol extension includes the extension of IS-IS, OSPF, and BGP-LS (see reference [8]).
- » With the emergence of new services such as 5G URLLC bearer and time-sensitive applications, deterministic networking (DetNet) was introduced to guarantee worst-case latency on IP/Ethernet networks. The Layer 2 technology IEEE 802.1 TSN and Layer 3 technology IETF DetNet define the congestion management mechanism on IP/Ethernet networks, scheduling algorithm based on latency information, explicit path establishment, and high-reliability redundant link technology. These technologies can work with FlexE technology to provide deterministic service bearers with lower-bounded latency and zero packet loss, which has also become a research focus (see references [9] and [10]).

With the official release of the OIF's FlexE IA 2.0 and FlexE technology's systematic application and architecture expansion in related standards organizations in the data communications field, FlexE technology has attracted wide industry attention. Chip/Device manufacturers are actively engaged in R&D, product testing, and demonstration. Network operators and large OTT service providers are also actively participating in standards promotion, technical cooperation, and solution verification. The related industry chain is being formed (see reference [11]).

6 Summary

As a technical architecture based on Ethernet and industry chain expansion, FlexE technology completely reuses the existing IEEE 802.3 Ethernet physical-layer standards. Through lightweight enhancement at the MAC/PCS logic layer, it implements flexible multi-rate interfaces and seamlessly interconnects with IP technology. FlexE meets requirements for large bandwidth, flexible rate, and channel isolation in IP/Ethernet scenarios, meeting the development demands of technologies and industries. Emerging services such as AR/VR and 5G, and the improvement of FlexE technology, are accelerating the formation of the FlexE industry chain. FlexE, as the foundation technology in the future IP/Ethernet system, will see dramatic growth and wide deployment.

References:

[1] http://www.ethernetalliance.org/roadmap/

[2] "How can Flexibility on the Line Side Best be Exploited on the Client Side?", OFC 2016 © OSA 2016, Tad Hofmeister, Vijay Vusirikala, Bikash Koley Google, Inc.

[3] OIF Flex Ethernet Implementation Agreement: IA OIF-FLEXE-0.10

[4] NGMN 5G White Paper, 2015 https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[5] OIF Flex Ethernet Implementation Agreement: IA OIF-FLEXE-0.20

[6] ITU-T G.709/Y.1331 Amendment 1 (11/2016): Interfaces for the optical transport network Amendment 1

[7] https://wiki.broadband-forum.org/display/BBF/BBF+Quarterly+Newsletter

[8] https://tools.ietf.org/html/draft-izh-ccamp-flexe-fwk-03

[9] http://www.ieee802.org/1/pages/tsn.html

[10] https://datatracker.ietf.org/wg/detnet/about/

[11] EE Times: Ethernet Flexes Network Muscles: http://www.eetimes.com/document.asp?doc_id=1330553&page_number=2

Acronyms and Abbreviations

Acronym or Abbreviation	Full Name
BBF	Broadband Forum
BGP-LS	Border Gateway Protocol-link state
DWDM	dense wavelength division multiplexing
eMBB	Enhanced Mobile Broadband
FlexE	flexible Ethernet
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IoT	Internet of things
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
MAC	medium access control
MEF	Metropolitan Ethernet Forum
MPLS	multiprotocol label switching
NGMN	Next Generation Mobile Network
OAM	operation, administration and maintenance
OIF	Optical Internetworking Forum
OTN	optical transport network
PCS	physical coding sublayer
PHY	physical layer
PMA	physical medium attachment
PMD	physical media dependent
QoS	quality of service
QoE	quality of experience
RSVP-TE	Resource Reservation Protocol-Traffic Engineering
SerDes	serializer/deserializer
UNI	user-to-network interface
NNI	network to network interface
URLLC	Ultra-Reliable and Low-Latency Communication