White Paper

The Evolution of ITU-T G.709 Optical Transport Networks (OTN) Beyond 100Gbit/s

Microsemi

Power Matters.™

1　Revision History

1.1　Revision 1.0

First release of the white paper

Contents

Figures

Tables

Abstract

The ITU-T OTN protocol underpins today's 10G, 40G and 100G DWDM packet optical transport networks. OTN has proven to be extremely flexible for accommodating new client signals and line rates. However, there were several new considerations to take into account when extending OTN to rates beyond 100Gbit/s (B100G). The B100G signals, referred to as "OTUCn," also needed to support new types of client signals such as 400GbE in a manner that is efficient both in terms of signal bandwidth use and implementation complexity. The ITU-T has just published the first phase of this work in the 2016 revision to G.709. This white paper provides a tutorial description of this work, including discussion of the new considerations that motivated the modifications and new features associated with the new OTN rates and OTUCn format. This tutorial also covers the new companion flexible OTN (FlexO) interface protocol that was defined in new G.709.1 for use with OTUCn.

Preface

In 2016, ITU-T released a major update to the G.709 OTN standard that covered rates beyond the 100Gbit/s OTU4. This new "beyond 100G" (B100G) family of interfaces are officially referred to as "OTUCn" and were defined as an $n \times 100$Gbit/s modular structure. The OTUCn signals reuse much from current OTN, but also incorporate aspects influenced by IEEE 802.3 Ethernet. The Ethernet influence includes the introduction of the new flexible OTN (FlexO) interface to go along with OTUCn. OTUCn is also defined to support new types of client signals in a manner that is efficient both in terms of signal bandwidth use and implementation complexity. This white paper provides a tutorial description of OTUCn and FlexO. Some of this work is ongoing and will be addressed in future releases of this white paper.

1   Introduction

The ITU-T G.709 Optical Transport Network (OTN) standard has defined the optical transport backbone of the worldwide service provider networks. As illustrated in Figure 1, OTN has been optimized to be the transport protocol to carry a converged mix of different types of client signals in a manner that efficiently uses the optical layer resources. See the white paper " A Tutorial on ITU-T G.709 Optical Transport Networks (OTN)" [17] for a full tutorial on G.709 for rates up through 100Gbit/s. This present white paper describes the work and new release of G.709 to extend it for rates beyond 100Gbit/s. The ITU-T standards group responsible for OTN standards is Question 11 of Study Group 15 (Q11/15). As discussed at several points in the white paper, this Q11/15 work was intentionally tied closely to the parallel work in IEEE 802.3bs to define a 400Gbit/s Ethernet signal (400GbE).

Figure 1   Converged transport over OTN
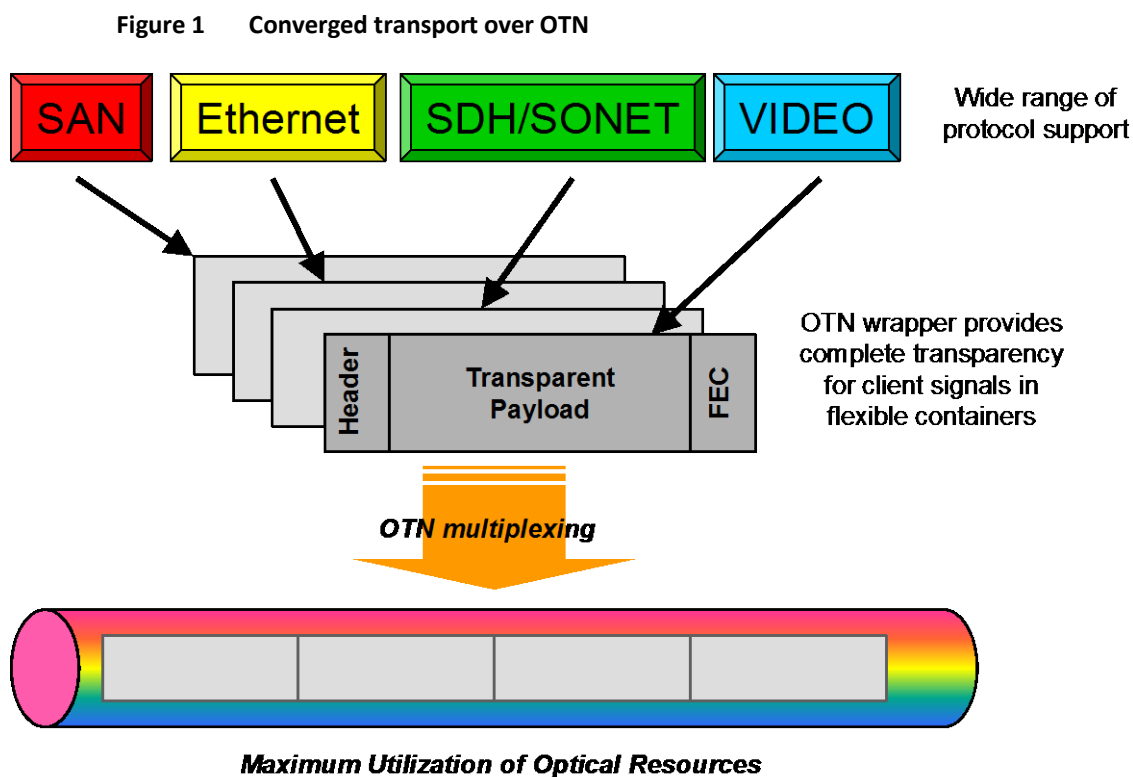
Taking OTN to rates higher than 100Gbit/s posed multiple challenges. Some of these were encountered when the 100Gbit/s OTU4 OTN signal was defined, but some were new challenges. Other objectives and challenges came up during the course of developing the B100G standard. These challenges and objectives included:

- The Shannon channel capacity limits are catching up to optical transport network capabilities. Specifically, the 50 GHz channel spacing of the standard wavelength grid currently used for dense wavelength division multiplexing (DWDM) imposes limits on transporting signals over reasonable distances when they have rates much over 100Gbit/s. (200Gbit/s was the practical limit per wavelength for distances of interest in telecom networks at beginning of the B100G project, although higher rates per wavelength will become possible driven by advances in Coherent DSP technology and new higher order modulation formats.)

- The old paradigm of adding new discrete rates for OTN had largely reached its practical limits, making a modular rate and frame structure approach more attractive.

- While G.709 had previously focused on Inter-Domain interfaces (IrDI) (e.g., interfaces between service providers or service provider domains), these same interfaces have been used as the basis for the Intra-Domain interfaces (IaDI) between equipment within a service provider network domain. Service providers have increasingly requested that the IaDI also be specified to a level that would allow interoperability between equipment from different vendors.[1]

- The IEEE 802.3bs Task Force working on 400Gbit/s Ethernet was examining several new approaches that were different from its previous interfaces. The new OTN format needed to not only carry 400GbE, but also re-use its technology and PHY components in order to benefit from the Ethernet component cost curves.

- The higher bit rates and increased use of multi-lane interfaces poses additional considerations regarding Forward Error Correction (FEC) and performance monitoring.

- Different B100G interface types have different FEC performance capability requirements. Consequently, while the current OTUk frame format had fixed dedicated FEC overhead, it was more appropriate to specify the FEC on a per-interface basis and not make the FEC overhead in integral part of the frame structure.

- Continuing to use 1.25Gbit/s Tributary Slot (TS) sizes in OTN would be impractical for B100G rates, so a larger Tributary Slot size was desirable. As will be discussed below, this decision was complicated by the emergence of the IEEE 802.3 work on 25GbE.

- B100G interfaces should re-use as much IP from the 100Gbit/s OTN interfaces as practical.

- There should not be a new switching layer in OTN associated with B100G.

- The high data rates and the introduction of new data client signals such as the OIF's Flexible Ethernet (FlexE) have made the current byte-oriented mapping approach for data clients impractical. This motivated the desire for a wide-word type of mapping.

- In order to optimize the use of each wavelength, including the transmission reach, there was a desire to transmit the OTN signals at the rate required for the client payload being carried rather than at the full discrete rate of the OTUk signal.

Each of these challenges will be addressed in this white paper, along with its implications for the B100G standard and influence on early working assumptions.

---

[1] Rather than defining interfaces in IrDI vs. IaDI terms, the ITU-T has come to realize that it is more meaningful to define them in terms of their application. For example, applications that had used IrDI were typically interfaces for short-reach or client side interface applications. Largely for this reason, the IrDI and IaDI terms have been removed from the latest (2016) version of G.709. Since the IrDI and IaDI terminology is known from previous versions of G.709, and was still in use during the early development of OTN B100G, this white paper continues to use the terminology as appropriate in the introductory/background paragraphs, but not with reference to the final frame format.

2 Optical Layer Considerations

The ITU-T (Q6 of Study Group 15) defined a standard wavelength grid for dense wavelength division multiplexing (DWDM) with the wavelengths on 50GHz (0.39nm) spacing.[9] Extending OTN up to 100Gbit/s was relatively straightforward in that the signal could be transmitted without using the entire 50GHz channel. In other words, the Shannon limit of the 50GHz channel was not exhausted. While NRZ (Non-Return to Zero) line codes were typically used for simplicity at the lower rates, a complex modulation scheme was used for all 100Gbit/s and most 40Gbit/s OTN signals. As will be explained in this section, the combined use of more complex modulation schemes and multiple wavelength channels will be required for B100G.

As noted, the NRZ line code was used for OTN rates below 40Gbit/s. While NRZ is the simplest line code, it has the drawback of relatively high frequency components. The impacts of dispersion in a fiber increase with the frequency content of the signal being carried, and for practical telecom network distances, this limits NRZ use to roughly 25Gbit/s. For higher rates, a more spectrum efficient modulation is required. In other words, a modulation method was required that uses a lower baud rate[2]. The most commonly used line code for OTU3 and OTU4 is DP-QPSK (dual-polarization – quadrature phase shift key), also known as PM-QPSK (Polarization Multiplexing-QPSK). While NRZ transmits one bit/symbol, DP-QPSK divides the transmitted signal onto two different polarization modes and uses two bits/symbol for each polarization. Consequently, the spectrum used by DP-QPSK is approximately ¼ of the spectrum used by NRZ.

DP-QPSK was also attractive because it gave the best performance relative to the power dissipated in the optical module. The dual polarization takes advantage of fiber's ability to preserve the relative polarization of the input signals as they transit the fiber, thus allowing different signals to be transmitted with different polarizations and recovered at the receiver. A 90-degree polarization difference is typically chosen for robustness. For some very long distance links, such as undersea cables, DP-BPSK (dual-polarization – binary phase shift key) lines codes have been used. The16QAM (quadrature amplitude modulation) line coding is being considered for some metro applications, due to its potentially lower cost. However, it requires linearity, and hence DSP, which increases the power dissipation in the module.

As rates increase beyond 100Gbit/s, fitting within the 50GHz wavelength slot[3] requires increasingly complex modulation, which lowers the signal to noise ratio (SNR). For example, DP-16QAM (Dual Polarization – 16QAM) is expected to be used for rates around 200Gbit/s in order to fit within the 50GHz optical channel. As noted above with respect to metro applications, this requires DSP at the receivers; however, DSP receivers are already commonly used for DP-QPSK receivers due to the implementation advantage they give over traditional receiver designs. The lower SNR of the 16QAM relative to the QPSK reduces the achievable reach of the signal. As it turns out, 200Gbit/s is effectively the maximum rate that can be transmitted over a 50GHz channel for most distances of interest in telecom networks. Consequently, higher rate B100G signals will typically be divided across multiple

---

[2] Recall that baud = bit/symbol, and the baud rate is hence sometimes referred to as the symbol rate.

[3] Recall that since the optical signal frequency (υ) is the speed of light divided by its wavelength (υ=c/λ), WDM is essentially a form of FDM, where the transmitted wavelength is modulated similar to the carrier frequency in an FDM system. As with all FDM systems, as the symbol rate increases for a given modulation format, the frequency spectrum used by that signal increases. When the frequency spectrum used by a given optical signal increases to the point where it overlaps with the signal in an adjacent channel, crosstalk occurs.

optical channels[4]. While interfaces using multiple wavelengths were already an option for OTU3 and OTU4, they will typically be required for a 400Gbit/s OTN signal.

One alternative to addressing the limits of the 50GHz DWDM channel slots is to either move to wider slots, or remove the wavelength grid altogether. A standardized wider slot is unattractive, since it would not be efficient for current signals, and could be too coarse for some of the IaDI options. Hence, it would typically not make efficient use of the entire available spectrum. Removing the grid altogether is not practical, since it would require such a wide range of optical sources. The most practical approach is going to a narrower grid where multiple adjacent slots can be grouped together to create larger slots. This flexible DWDM approach, sometimes referred to as "FlexGrid," is defined in ITU-T G.694.1. Specifically, G.694.1 defines a grid with 12.5GHz channels, and illustrates how adjacent slots can be combined to create wider channels. The FlexGrid approach allows the network operator to allocate the optimum wavelength channel size to each transport signal, thus maximizing the use of the available spectrum. However, the inevitable churn from network reconfiguration will quickly produce a fragmented spectrum on the fiber where it will be difficult to find new openings for larger channel slots. While it may be feasible in some applications, FlexGrid does not provide a good general solution. Consequently, the 50GHz grid will continue to be used for some time.

It should be noted that coherent receivers are assumed here. The reason coherent receiver technology became popular is that it provides a linear transformation between the optical and electrical domains. This allows using the electronics to filter/cancel out the effects of both polarization mode dispersion (PMD) and chromatic dispersion. Using coherent receivers does not increase the spectral efficiency or reach, rather it just simplifies the receiver implementation since the dispersion compensation is done by DSP in the electrical domain rather than the optical domain.

One of the early OTN B100G working assumptions was that the minimum bit rate for a multi-vendor IaDI (MV-IaDI) optical tributary signal will be 50-56Gbit/s. For example, while an MV-IaDI OTUCn rate may be 200Gbit/s, it will always use transmission rates of at least nominally 50Gbit/s per wavelength for that interface. This assumption is consistent with the IEEE 802.3bs agreements for 400GbE.

For the multi-lane electrical interfaces, it was agreed that a B100G signal will use at least 25-28Gbit/s per lane, which is also consistent with 400GbE.

Note that for commonality with 400GbE, OTN B100G electrical and optical 50Gbit/s signals will use PAM-4 (4-level pulse amplitude modulation) instead of NRZ. Since PAM-4 transmits 2-bits/symbol, the symbol rate is 25GBaud.

The lane structure and information distribution across the electrical and optical lanes will depend on the interface type. For example, in order to maintain commonality with multi-lane OTL4.4 interfaces for OTU4, an equivalent OTLC.4 interface was defined for B100G. [14] (In other words, both OTL4.4 and OTLC.4 are based on 4 × 28Gbit/s physical lanes composed of 5Gbit/s logical lanes.) However, the "FlexO" interface described below in section 10 is based on the 400GbE format with its 28Gbit/s logical lane structure.

---

[4] For the purposes of this white paper, the term "optical channel" is being used in the loosest sense, and essentially equates to a wavelength or a wavelength slot/band or channel slot within the ITU-T DWDM grid. Strictly speaking the correct technical term for this a network media channel. Also, in the strict OTN sense, an optical channel is the set of wavelengths that carry both the OTN payload signal and the associated optical supervisory channel (OSC) on a separate wavelength.

3 Signal Formats and Frame Structure

This section describes the signal format and rates for the digital portion of the OTN B100G signal.

3.1 Introduction and Background

Traditionally, the standard SDH and OTN rates had increased at each step by a factor of four. The one exception to this was OTN Optical Data Unit level 4 (ODU4), which was chosen to be optimized for carrying 100GbE Ethernet clients rather than choosing a 160Gbit/s rate for carrying four of the 40Gbit/s ODU3 signals. There were several advantages to this choice. By that time, it had become apparent that Ethernet is an increasingly important client signal for transport over OTN. Also, a rate around 112Gbit/s was much more cost effective for the bandwidth relative to 160Gbit/s, especially for the optical component technology available at that time. This was especially true since the OTU4 could then re-use the Ethernet 100GbE optical interface modules. A final consideration was the potential for OTU4 to become the default long-reach WAN interface for 100GbE. The ITU-T is using a similar line of reasoning for rates beyond 100Gbit/s.

As preliminary work began in the ITU-T, it had become apparent that the next higher Ethernet rate would be 400GbE. The ITU-T needed to balance the desire to have a rate optimized for carrying and re-using technology from 400GbE, and the need for a more modular rate to carry multiplexed signals at lower rates over longer reach channels. The initial plan had been to standardize a limited number of new IrDI rates, which were primarily matched to new Ethernet rates, and to define a modular structure that also allows constructing Intra-Domain Interface (IaDI) signals to match the channel characteristics/quality.

Limiting the different IrDI rates, and tying them to new Ethernet rates, was intended to reduce cost by having fewer, higher-volume options. As noted above, subsequent developments blurred the distinction between IrDI and IaDI. More importantly, around the same timeframe, there was an independent proposal to define a modular flexible approach to Ethernet PHYs. This approach, called "Flexible Ethernet" (FlexE), was specified in an Optical Internetworking Forum (OIF) Implementation Agreement (IA). [15] The FlexE concept inspired an analogous approach for OTN B100G interfaces that is called "Flexible OTN" (FlexO). The initial version of the G.709.1 FlexO Recommendation [16] specified an OTN B100G short reach PHY as being a set of individual n 100Gbit/s PHYs that are grouped together to carry the $n \times$ 100Gbit/s OTN signal. It was defined in a manner that allows using a 100GbE/OTU4 optical module for each constituent 100Gbit/s optical signal. FlexO is described in section 10.

The OTN ODUk structure had used an asynchronous multiplexing approach, with the next higher discrete rate carrying a combination of the lower rate signals. The frame of each OTN ODUk rate was identical, but transmitted at a higher signal rate. In contrast, SONET/SDH created its higher rate signals by interleaving an integer multiple of its base rate signals (SONET STS-1 or SDH STM-1). For rates beyond 100Gbit/s, the ITU-T Q11/15 chose to use a hybrid of these two approaches. A new base signal frame was established at around 100Gbit/s, with multiples of this base frame interleaved to create the higher rate signals. The terminology chosen was to call the base frame an ODUC (100 Gbit/s ODU slice), and to call signal constructed from $n \times$ ODUC slices an ODUCn, where the "C" corresponds to the Roman numeral for 100.

3.2 Frame Format

As illustrated in Figure 2, the base ODUC frame uses a frame structure identical to the ODU4, except that there are no fixed stuff columns in the payload area. Unlike SONET/SDH, the payload area of each signal is a single contiguous payload container rather than an interleaving of the base-rate signal payload containers. This payload structure is illustrated in Figure 6 in section 4.2. The resulting OPUCn payload container is analogous to the contiguously concatenated SONET/SDH signals (e.g., SONET STS-12C or SDH VC-4C), except that this approach is always taken with OPUCn rather than being the special case.

Figure 2   OTUC1 base frame format and OTUCn conceptual frame format
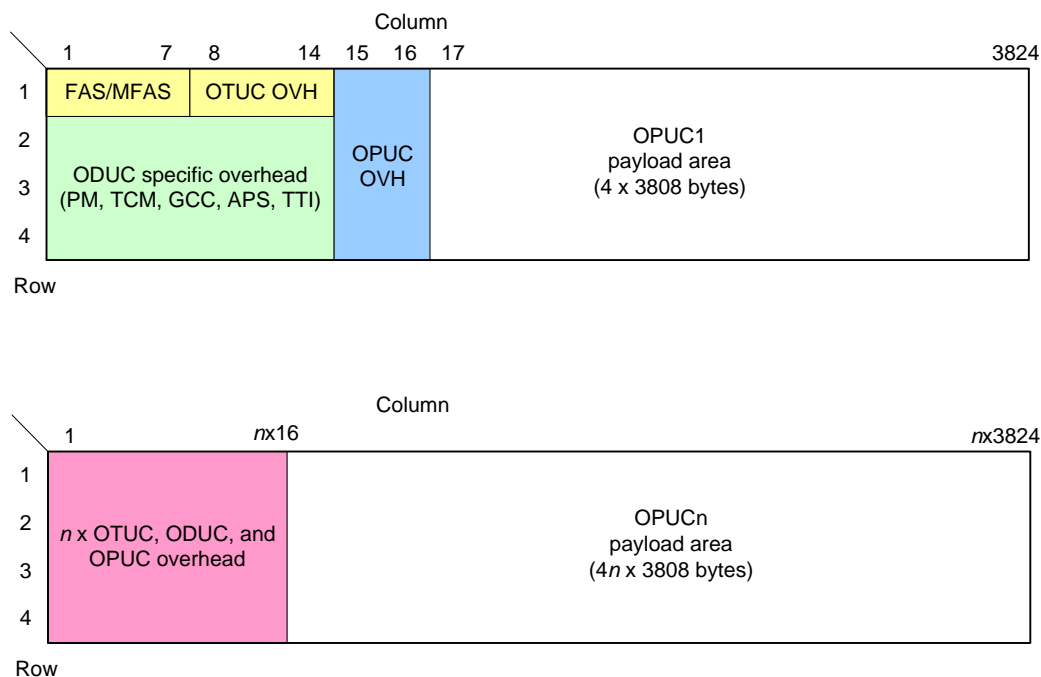
The physical layer of the OTUCn signal will depend on the interface. For example, it can be transmitted as a single serial stream, as n 100Gbit/s streams, or $n/2$ 200Gbit/s streams in the optical domain, or as multiples of 25Gbit/s or 50Gbit/s with electrical domain interfaces. Rather than defining a serial interleaving format for the OTUCn stream, analogous to that of SONET/SDH, it is left to be specified by the individual interface. The OTUC slices are interleaved in a defined manner so that the OPUCn Tributary Slots have a known order. For transmission, however, each OTUC is treated individually as a 100Gbit/s entity. For example, an OTUC can be transmitted over 4 of the 25Gbit/s lanes or 2 of the 50Gbit/s lanes of an OTUCn interface. In these cases, each OTUC uses a specific subset of the electrical PHY lanes, independent of the PHY lanes used by the other OTUC elements. As another example, the "FlexO" interface discussed below in section 10 transmits each OTUC element as a separate 100Gbit/s optical signal using 100GbE/OTU4 optical modules, and hence requires no OTUC interleaving in the PHY layer. Yet another example is an optical module that combines two of the OTUC elements into a 200Gbit/s stream for transmission over the optical channel (wavelength), in which case the method of interleaving the OTUC elements is arbitrary as long as the optical source and sink agree on the method. Alternatively, the 200Gbit/s interface could use the complex modulation format to multiplex the two OTUC streams (e.g., each transmitted independently on a separate polarization mode of the optical line code). Since the OPUCn tributary slots and word locations are numbered (see Figure 6), there is no ambiguity in constructing the OPUCn at the source and recovering it at the sink.

In addition to the modular frame structure, the B100G signals also differ from current OTN signals in another important way. For network simplification, an ODUCn signal is only carried point-to-point between network nodes. In other words, the ODUCn signal is only a Multiplex Section layer entity that cannot be switched. It only

exists to carry lower rate ODUk signals between a pair of nodes, with all switching being done at the ODUk level. The implications of this decision include: [5]

- No client signals are directly mapped into the OPUCn. They must first be mapped into an ODUk (including ODUflex), which is then mapped or multiplexed into the OPUCn. (This topic is discussed below in the client mapping section 5.)

- All the components of the interface signal go through the same fiber and optical switches (i.e., the same Optical Multiplex Section trails) such that the OTUCn signal can be managed as a single entity. Consequently, very limited deskew is required[6].

A fundamental difference between the current OTUk frame format and the OTUCn frame format is that the OTUCn frame has no dedicated area for FEC. In other words, the OTUCn and ODUCn frame formats are identical, except for the population of the OTUC-specific overhead fields. As discussed above, the reason for this choice is the recognition that there will be different interface types for the OTUCn, and each interface will have its own requirements in terms of the strength of FEC that is required. Since aspects of the FEC choice were expected to borrow from the IEEE 802.3bs work, removing dedicated FEC overhead space from the OTUCn frame format removed this specific B100G dependency on IEEE 802.3bs.

The rate of the base OTUC signal was chosen in order to meet the following criteria:

- An OPUC1 must be capable of carrying an ODU4 client

- An OPUC4 must be capable of carrying a 400GbE client.

- The resulting signal rate should be reasonable efficient within the constraints of the first two criteria. For example, when the signal is divided down to the nominally 25Gbit/s rate, the resulting electrical lane interface rate should be compatible with the OIF CEI-28G specification.

The resulting rates are shown in Table 1.

Table 1   OTN B100G signal and payload rates

OTUCn/ODUCn signal rate	OPUCn payload area rate	OTUCn/OPUCn frame period
$n \times (239/226) \times 99.5328$ Gbit/s	$n \times (238/226) \times 99.5328$ Gbit/s	
$= n \times 105.258138$ Gbit/s	$= n \times 104.817727$ Gbit/s	1.163 µs
Note: All rates are ±20 ppm.		

The OPUC, ODUC, and OTUC frame structure is shown in Figure 3. The overhead for each level is further illustrated in Figure 3, Figure 4 and Figure 5. The ODUC frame is structured as four rows of 3824 columns. As shown in Table 1, as the signal rate rises, the frame period remains constant. The OPUC payload area consists of columns 17-3824 for all four rows. As illustrated in Figure 6, when an OPUCn signal is constructed, there are four rows of $n \times 3808$ columns. A discussed above, the interleaving of OTUC slices into an OTUCn frame is not illustrated in G.709.

---

[5] This decision not to have ODUCn switching would appear to imply that TCM is not required. However, as discussed below in the overhead section, TCM is still very useful with ODUCn.

[6] The primary contribution to skew is the different propagation speeds of different wavelengths in a fiber when different parts of the OTUCn signal are transmitted on different wavelengths.

3.3   OTN B100G OTUC and ODUC Signal Overhead

As with the current OTN signals, the OTUC shares overhead columns with the ODUC overhead. However, as discussed above, unlike the current OTN signals, there is no reserved area in the OTUC for FEC. The OTUC overhead is shown as the **A** and **B** areas in Figure 3. The **A** field contains the frame alignment signal (FAS) and the multiframe alignment signal (MFAS). The MFAS field is a binary counter that shows the phase of the current frame within the 256-frame multiframe used by some of the overhead. For example, the Payload Structure Identifier (PSI) overhead shown in Figure 3 and Figure 5 uses the MFAS to determine the meaning of the byte during that frame. The MFAS is also used in removing skew between portions of the OTUCn signal that were transmitted over different wavelengths or different FlexO interface PHYs. The **B** area of Figure 3 provides General Communication Channel (GCC) and section monitoring (SM) information for the OTUCn. The SM overhead includes the trail trace identifier (TTI), a BIP-8 for error detection, a backward error indication (BEI), a backward defect indicator (BDI), an incoming alignment error (IAE) indicator, and a backward IAE (BIAE). As with current OTN, the TTI is used for connectivity fault detection, the BEI is sent by the OTUCn sink to the OTUCn source as a (binary) count of the number of errors detected by the previous BIP-8, the BDI is used by the sink to inform the source that it is seeing a signal failure. The IAE indicates that a frame alignment error was detected on the incoming signal, and the BIAE informing the source that an IAE was seen. The IAE and BIAE are used to disable the error counting in their respective directions during frame alignment loss conditions.

Figure 3   OTN B100G base signal frame, including OTUC1 overhead structures
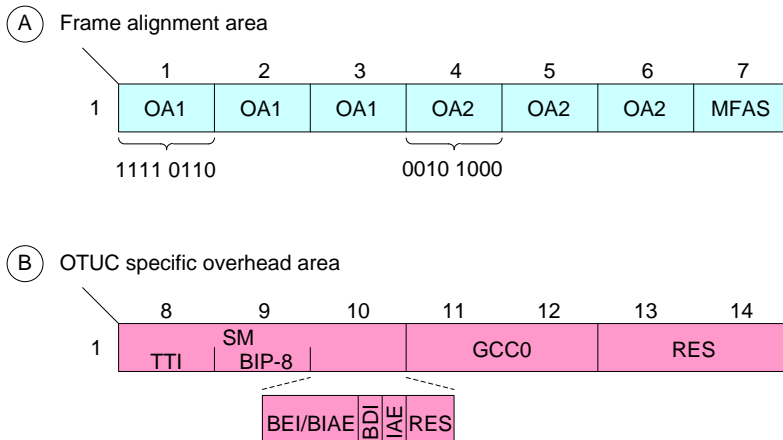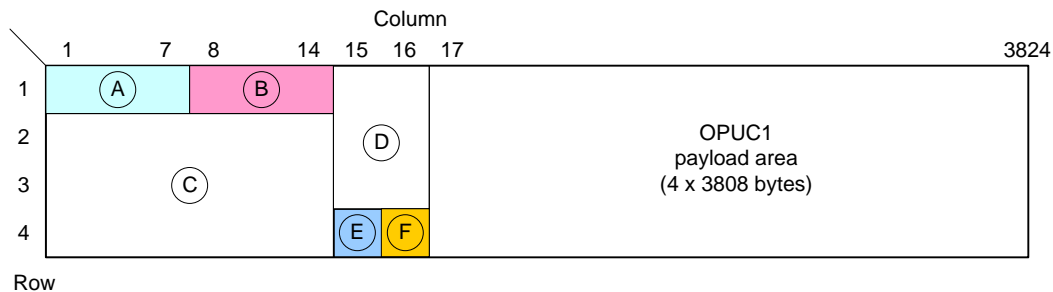
As can be seen from Table 2, some of the overhead bytes are active on only the first OTUC slice, while others are active on all slices. Overhead like TTI, BDI, and STAT pertain to the whole interface, so they only need to appear once. The frame alignment is repeated in all slices in order to accelerate the framing recovery time, and also to have it available for frame recovery on all lanes of a multi-lane interface. The BIP-8 and BEI are active in all slices in order to provide better (and modular) error detection capability.

Note that scrambling is still required to protect against malicious users who may attempt to interrupt the operation of an OTUCn link by sending long information strings in the client payload that would cause problems for the receiver clock and data recovery circuitry. However, for OTUCn signals, the scrambling is specified as part of the PHY signal (e.g., FlexO) rather than specified in a general sense for the OTUCn.

Table 2   OTUC overhead usage

OTUC Overhead	Active on OTUCi slices:
Frame Alignment Signal (FAS) OA1 & OA2	1-n
Multi-frame Alignment Signal (MFAS)	1-n
SM – TTI	1
SM – BIP8	1-n
SM – BEI	1-n
SM – BDI	1
SM – BIAE	1-n
SM – STAT	1
GCC0	1-n*
* The GCC0 channels on all n OTUC slices are merged into a single higher-rate channel for the OTUCn interface. For vendor-specific interfaces, GCC0 can optionally be active on only OTUC slice 1.	

The ODUCn consists of the OPUCn and the ODUCn overhead, which is functionally a Multiplex Section overhead. The ODUCn overhead is area **C** in Figure 3, with Table 3 showing in which ODUC slice each is active. It contains the overhead for path performance monitoring (PM), two generic communications channels (GCC), an automatic protection switching and protection communications channel (APS/PCC), and a set of bytes reserved for experimental purposes. The PM overhead consists of trail trace identifier (TTI, for connectivity fault detection), a BIP-8 for error detection, status information (to indicate whether this is a normal signal or a maintenance signal), and a BEI. The PM overhead also includes a delay measurement function to perform round trip delay measurement at the Path level. (See 8.3.) As with the OTUCn, the BEI is sent by the ODUCn sink to the ODUCn source as a (binary) count of the number of errors detected by the previous BIP-8. Note that the former fault type and fault location (FTFL) overhead has been removed from all OTN signals, with the byte's status being changed to "Experimental."

Since the ODUCn is not switched, it does not require TCM in quite the same way that a switched ODUk does. However, for applications where the OTUCn signal goes through 3R repeaters, TCM is valuable for determining the performance of the optical signals on each side of the repeater. A "carrier's carrier" application can use multiple repeaters along the connection. Due to the multiple optical segments and the different service providers along the OTUCn path, multiple levels of TCM are still required. Consequently, it was decided that the B100G overhead will continue to support the same six levels of TCM overhead that are defined for current ODUk.

Figure 4   OTN B100G ODUC1 overhead structures
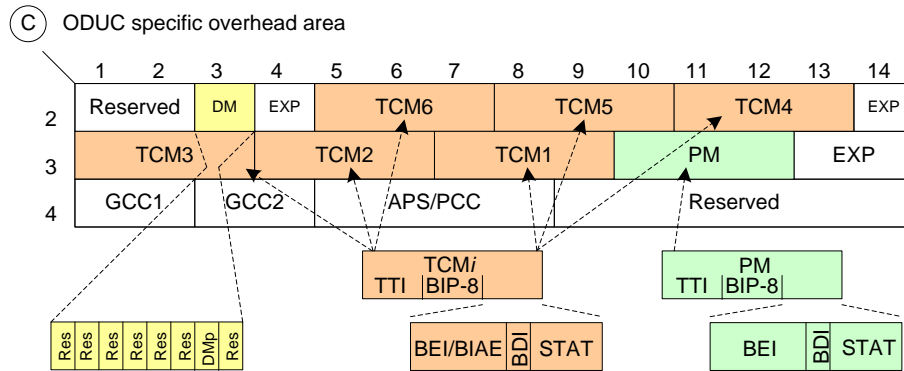
Table 3   ODUC overhead usage

ODUC Overhead	Active on ODUC*i* slices:
Delay Measurement (DM)	1
PM – TTI	1
PM – BIP8	1-*n*
PM – BEI	1-*n*
PM – BDI	1
PM – STAT	1
PM – Delay Measurement	1
Expansion (EXP)	1-*n*
GCC1	1-*n**
GCC2	1-*n**
APS/PCC	1
* The GCC1 and GCC2 channels on all *n* ODUC slices are merged into a single higher-rate GCC1 and GCC2 channel for the ODUCn interface. The GCC1 and GCC2 fields can also be optionally combined/merged to create a single larger channel from both.	

3.4   OTN B100G OPUC Signal Overhead

The overhead information for the OPUCn is contained in the **D**, **E** and **F** areas of Figure 3, and is shown in more detail in Figure 5. As with current OPUk overhead, the OPUCn overhead is primarily concerned with the mapping/multiplexing and demapping/demultiplexing of the client signals. For all applications, the Payload Structure Indicator (PSI) byte contains indicators for the payload type (PT) and Multiplex Structure Identifier (MSI). Note that since no client signals are mapped directly into an OPUCn (i.e., the clients are always first mapped into their own ODUk/ODUflex), PT = 22 ("ODU multiplex structure supporting ODTUCn.ts") is the only PT used for B100G signals. Since the entire OPUCn is a combined payload area, it is only necessary to indicate the PT in OPUC

#1. The MSI indicates for each TS of the ODUC whether that TS is available, whether it has been allocated, and the tributary port number associated with the client signal using that TS. The reason for indicating the availability of the TS will become clear in the discussion of "OTUCn-M" below in section 7. As shown in Figure 5, the set of PSI bytes appear in row 4 of column 15 of each ODUC slice. The correspondence between specific OPUCn overhead bytes and their respective payload area Tributary Slots is discussed below in section 5 under payload mappings.

The frequency justification required in order to adapt between the client signal rate and the OPUCn channel payload rate is located in the OPUCn overhead area. The details of the JC bytes are described in section 6, with the correspondence between their location within the multiframe and the associated TS illustrated in Figure 6.

Hitless Adjustment of ODUflex(GFP) (HAO) is not supported by the B100G signals.

Figure 5   B100G OPUC1 signal overhead structure with the generic mapping procedure
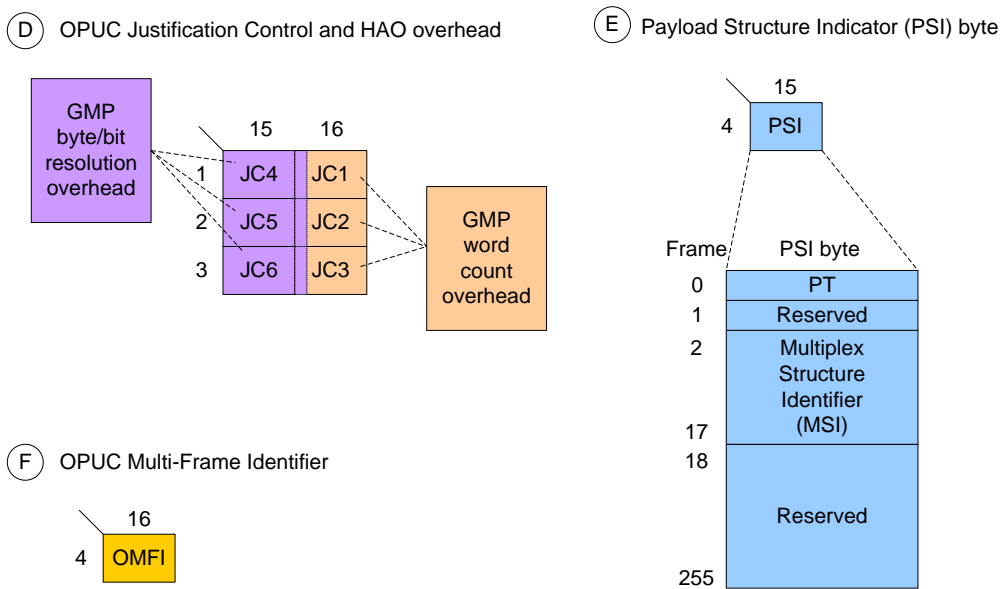
Table 4   OPUC overhead usage

OPUC Overhead		Active on OPUC*i* slices:
PSI	PT = 22	1
	MSI	1-*n*
GMP JC1-JC3		1-*n*
GMP JC4-JC6		1-*n*

4 OPUCn Structure

4.1 Tributary Slot (TS) Size Considerations

The current OTN signals use a 1.25Gbit/s Tributary Slot (TS) size that results in 80 TS in the OPU4[7]. For B100G signals that may eventually have rates of multiple Terabit/s, this 1.25Gbit/s TS size becomes impractical both in terms of the device complexity required to support it, and the associated network management complexity. Consequently, Q11/15 chose a 5Gbit/s TS size for OPUCn.[8] The 5Gbit/s rate is four times the 1.25Gbit/s rate of current OTN, and also provides the flexibility for efficient transport of important existing and emerging clients. For example, 10GbE signals are an increasingly important existing client signal. Examples of emerging client signals that can be mapped efficiently with the 5Gbit/s TS granularity include 25GbE, 50GbE, and 16Gbit/s Fiber Channel (16GFC), which can be carried in three of the 5Gbit/s TS.

In order to further minimize the added IC and network management complexity associated with large numbers of TS, client multiplexing was restricted such that an OPUCn signal will only be required to carry a maximum of $10n$ clients.[9]

It was expected that clients with rates lower than 5Gbit/s (e.g., ODU0 or ODU1) would typically first be multiplexed into a current ODUk (k≥2), which would then be efficiently mapped into the OPUCn. GMP also makes it feasible to carry ODU0 and ODU1 clients by directly mapping them into the 5G TS rather than into a higher rate ODUk first.

4.2 OPUCn Payload Area Structure

The individual TS are interleaved into the OPUCn contiguous payload area on a 16-byte (128-bit) block basis. In other words, each of the $20n$ TS occupies 16 bytes at a time in a fixed round-robin manner. This 16-byte TS interleaving granularity maintains a consistent 16-byte-modularity for both the OTUC/ODUC/OPUC overhead and the TS. Moving to 16-byte TS interleave granularity rather than the current single byte granularity is much more convenient with the high B100G data rates and the associated wide data buses within ICs that process them.

The OPUCn structure, including an illustration of the TS naming convention, is shown in Figure 6 and Figure 7. The TS number is designated as TS A.B. The "A" refers to the OPUC slice associated with that column/block (reflecting

---

[7] Of course, a 2.5Gbit/s TS is also used for OPU1, OPU2 and OPU3 with the AMP frequency justification control.

[8] Q11/15 initially chose a 10Gbit/s TS for B100G. However, the subsequent launch of the IEEE 802.3by project for 25GbE caused Q11/15 to move to a 5Gbit/s TS size in order to allow efficient 25GbE transport over B100G without requiring a 2-stage multiplexing of 25GbE clients first into a legacy ODUk. While the primary motivation for the 25GbE project involved efficient interconnections between computer servers and the top-of-rack (ToR) Ethernet switches in the server racks of Data Centers, once an Ethernet signal rate is standardized there is nothing to prevent its use in a variety of applications, including WAN connections. For this reason, the ITU-T must reasonably expect that there will be a customer desire for 25GbE connections over OTN. For example, a data center or enterprise customer may want to use 25GbE as a modular WAN interface where different numbers of the 25GbE portions are activated to meet the traffic demands. Other client types (e.g., CPRI) were expected to choose 25Gbit/s rates in order to leverage the lower costs of higher-volume 25GbE PHYs.

[9] The restriction of a maximum of $10n$ clients results in the same total as would have resulted from a 10Gbit/s TS approach, but allows the efficient mapping flexibility of the 5Gbit/s TS.

the fact that the OPUCn logically consists of n OPUC slices). The "B" term refers to the order of appearance of each group of 16-byte TS blocks within slice A, up to 20 groups (i.e., B = 1, …, 20). The order of TS appearance within the OPUCn frame is first by the OPUC slice and then by the TS number within an OPUC slice.[10] In other words, as illustrated in Figure 7, the TS A.B appearance order is TS1.1, TS2.1, TS3.1,…, TSn.1, TS1.2, TS2.2,…, TSn.2, TS1.3,…, TS1.20,…, TSn.20. See Figure 8 for an example illustration of mapping a client into OPUCn.

The actual TS rate can be calculated as follows: (See Figure 6 and Table 1.)

TS rate = [OPUC rate]*[ratio of bits/row/TS to the OPUC bits/row]
= 5.24089 Gbit/s per TS[11]

The OTN Generic Mapping Procedure (GMP) is used as the Justification Control to map client signals into the OPUCn. The GMP overhead is described in Section 6. With the interleaving of n OPUC elements to form the OPUCn, there will be n appearances of Justification Control overhead (JC1-JC6) per frame. (See Figure 3, Figure 5, and Figure 6.) A 20-frame multiframe provides a unique Justification Control overhead location associated with each of the 20n TS. The locations of the JC1-JC6 bytes associated with each TS are illustrated in Figure 6. In general, following the TS(A.B) naming convention, the location of the JC1-JC6 bytes associated for a given TS a.b are located in columns 14 and 15 of frame b - 1 of OPUC number a.

---

[10] Since there are 3808 payload columns per row in the OPUC, each of the 20n TS in the OPUCn appears exactly

[(3808n columns/OPUCn row)] / [(16 columns/TS)*(20n TS)] = 11.9 times/row

Consequently, the set of TS will repeat their row alignment every 10 rows. (i.e., exactly 8 times per 20-frame multiframe). Since 10 rows is 2.5 frames, the TS locations repeat their alignment to the OPUCn frame every 5 frames (i.e., 4 times per 20-frame multiframe).

[11] TS rate = [104.817727 Gbit/s]*[(row/3808*8 bits)(128 bits/TS/occur.)(11.9 occur./row)]
= 5.24089 Gbit/s per TS

Figure 6   OPUC tributary slot interleaving and overhead structure (for Slice A)

OMFI bits 4-8	Row	15	16	17	32	33	48	...	321	336	337	352	...	3808	3824
00000	1	JC4-6 for TS A.1	JC1-3 for TS A.1	TS A.1	TS A.2				TS A.20	TS A.1					TS A.18
	2			TS A.19	TS A.20			...	TS A.18	TS A.19			...		TS A.16
	3			TS A.17	TS A.18				TS A.16	TS A.17					TS A.14
	4	PSI	OMFI	TS A.15	TS A.16				TS A.14	TS A.15					TS A.12
00001	1	JC4-6 for TS A.2	JC1-3 for TS A.2	TS A.13	TS A.14				TS A.12	TS A.13					TS A.10
	2			TS A.11	TS A.12			...	TS A.10	TS A.11			...		TS A.8
	3			TS A.9	TS A.10				TS A.8	TS A.9					TS A.6
	4	PSI	OMFI	TS A.7	TS A.8				TS A.6	TS A.7					TS A.4
00010	1	JC4-6 for TS A.3	JC1-3 for TS A.3	TS A.5	TS A.6				TS A.4	TS A.5					TS A.2
	2			TS A.3	TS A.4			...	TS A.2	TS A.3			...		TS A.20
	3			TS A.1	TS A.2				TS A.20	TS A.1					TS A.18
	4	PSI	OMFI	TS A.19	TS A.20				TS A.18	TS A.19					TS A.16
⋮								⋮							
i-1	1	JC4-6 for TS A.i	JC1-3 for TS A.i	TS A.x_i											
	2							...					...		
	3														
	4	PSI	OMFI												TS A.z_i
⋮								⋮							
10011	1	JC4-6 for TS A.20	JC1-3 for TS A.20	TS A.9	TS A.10				TS A.8	TS A.9					TS A.6
	2			TS A.7	TS A.8			...	TS A.6	TS A.7			...		TS A.4
	3			TS A.5	TS A.6				TS A.4	TS A.5					TS A.2
	4	PSI	OMFI	TS A.3	TS A.4				TS A.2	TS A.3					TS A.20

Column

$$x_i = [(i\text{-}1)(12)+1]\bmod 20$$
$$z_i = [(i\text{-}1)(12+12]\bmod 20$$

Figure 7   Illustration of the OPUCn tributary slot order (from first row of OPUCn frame)

Figure 8   Word numbering illustration for GMP multiplexing – ODUflex tributary using three TS (#2.3, 1.4 and 1.15) into an OPUC2

Note: In this example, different colors are used for each word shown for that client. The sequence number of the 16-byte segment of the 3 x 16-byte word is shown in italics.

As discussed below in section 6, the GMP word count overhead (C_m carried in bytes JC1-JC3) gives the number of data words carried in the TS(s) during that multiframe. With the current OTN, the word size associated with the GMP count was eight bits (one byte) per TS used by that client. This is consistent with each TS using an 8-bit column width. In other words, it uses C_8 as the base for the multiplexing clients into the TS of a HO OPUk. For example, an ODU0 client uses a single 1.25G TS and has a corresponding single byte GMP word size, while an

ODU2 client uses eight 1.25G TS and consequently uses an 8-byte GMP word size. As discussed above and illustrated in Figure 6, for B100G the TS appearances are structured as 16-byte columns, with 16 bytes of client data mapped into each TS appearance. Since 16 bytes is 128 bits, it made sense to use C_{128} as the base granularity for GMP. A client signal using m TS will have a $128m$-bit GMP word.

As shown in Figure 6, each 5Gbit/s TS appears in 3808/20 = 190.4 columns of the OPUC. With the 16-byte base GMP word size (C_{128}), there are 952 words/TS/multiframe.[12]

---

[12] (190.4 bytes/row/TS)(4*20 rows/MF)/(16 bytes/word) = 952 words/TS/MF

5   Payload Mapping and Multiplexing

As noted above, no clients are directly mapped into an OPUCn payload. In other words, the ODUCn is always a transport server layer signal that only carries client signals that have been already mapped into a lower rate ODUk. (k = 0, 1, 2, 3, 4, flex[13]) Consequently, there is no essential difference between mapping and multiplexing from the B100G standpoint. The basic B100G multiplex hierarchy is illustrated in Figure 9.

CBR clients with rates >100Gbit/s (i.e., rates too large for an OPU4) are mapped into an ODUflex, which is then multiplexed into an OPUCn. The most important of these clients will be 200 Gbit/s and 400Gbit/s Ethernet, and FlexE. Note that generic CBR clients are not shown as "new" in Figure 9 since they use the same mapping as any other CBR client into an ODUflex(CBR). The only difference here is that the ODUflex(CBR) rate will have the 5Gbit/s granularity of the OPUCn TS rather than the 1.25Gbit/s granularity associated with current ODUk TS rate. The 200GbE and 400GbE clients are shown as new since they are expected to need some special processing associated with preserving the client rate when Ethernet lane alignment markers are removed by the Ethernet PCS layer ahead of the OTN mapper.

As explained in section 5.5, some high speed packet data clients use the new ODUflex(IMP), which replaces ODUflex(GFP) for this application. These clients include FlexE.

Current ODUk signals can be multiplexed directly into the new ODUCn. Clients with rates below the OPU4, including ODU0 or ODU1-mapped clients, can also be carried in the OPUCn by two-stage multiplexing. The two-stage multiplexing allows aggregating multiple low rate clients into a higher rate current ODUk (k = 2, 3, 4), which can then be multiplexed into the OPUCn. The two-stage multiplexing can either be advantageous to increase bandwidth efficiency through aggregation of low-speed clients, and/or enabling interworking with current OTN equipment within the network.

An ODUk is multiplexed into an OPUCn by first asynchronously mapping it into an intermediate structure called an Optical Channel Data Tributary Unit for B100G (ODTUCn). The overhead of the ODTUCn is the information required for timing justification between the client ODUk rate and the ODTUCn rate, which is the derived from the OPUCn rate. As illustrated in Figure 6, this overhead is contained in the JC1-JC6 bytes, which are time-shared among the TS across an OPUCn multiframe. As explained above (see Section 4.2), the OPUCn payload area is divided into Tributary Slots (TS), with each client ODUk occupying an integer number of TS. The ODTUCn is directly byte synchronously mapped into a set of OPUCn TS in 128-bit sized words. For that reason, it is referred to as an "OPUCn.ts" where the "*ts*" refers to the number of TS used by that ODTUCn. The payload byte capacity of an ODTUCn.ts is:
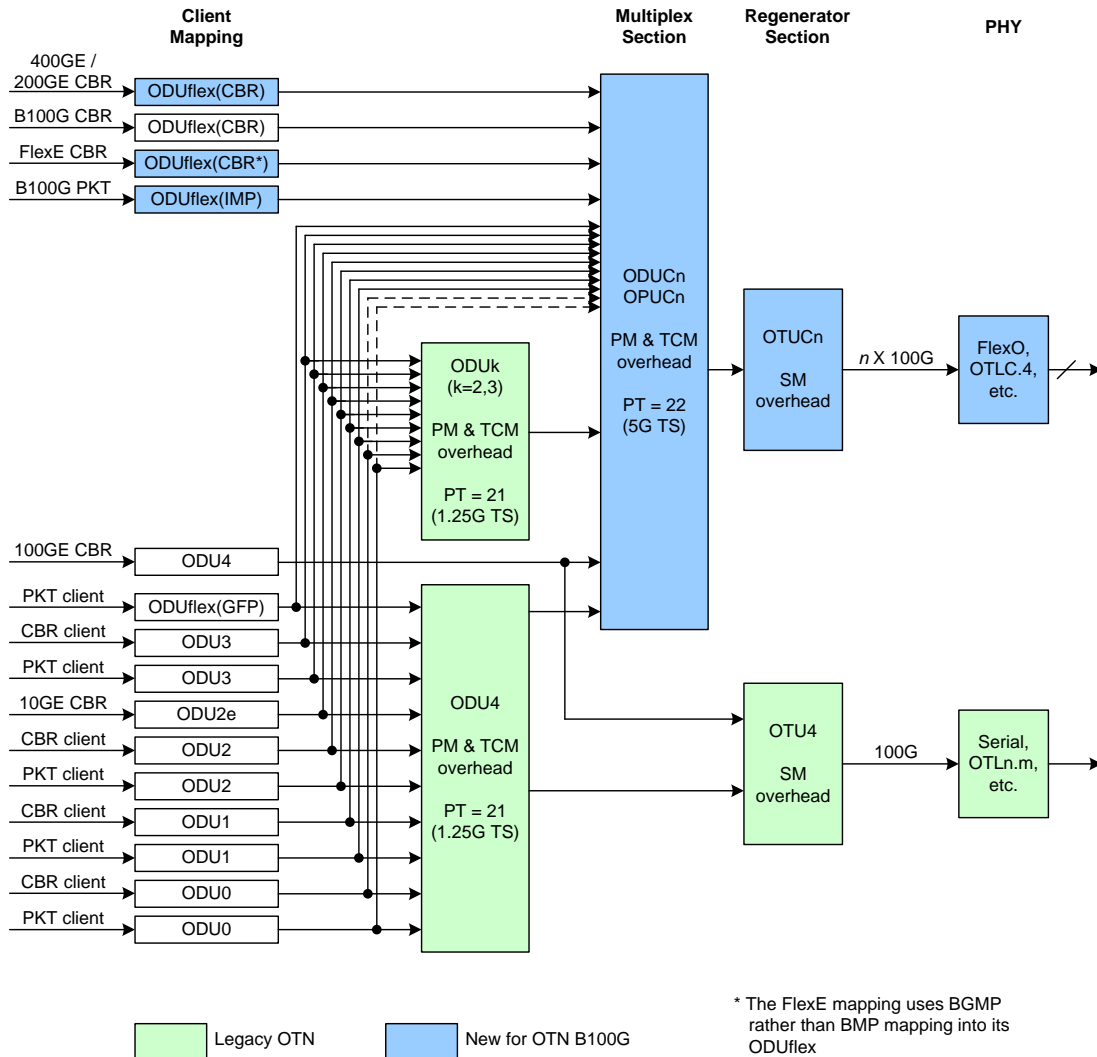
$$(ts)(952 \text{ words/TS/multiframe})(16 \text{ bytes/word}) = (15232)(ts) \text{ bytes/multiframe},$$

where $1 \leq ts \leq 20n$.

---

[13] As noted above, while an ODU0 or ODU1 can be multiplexed directly into a 5Gbit/s TS, due to the associated bandwidth inefficiency this would be very rare.

Figure 9   B100G multiplex hierarchy illustration

5.1   CBR Signal Multiplexing Using GMP

The concept behind GMP is that the JC1-3 bytes of each multiframe are used to communicate the number of payload words that will be mapped into the OPUk/OPUCn payload area during the next multiframe. The transmitter and receiver use modulo arithmetic based on this count value to determine the location of payload and stuff words within the payload area of the frame. See Section 6 and Microsemi white paper [17] for a complete description of the GMP process, including some specific mapping examples. As described above, the word size was chosen to be 16 bytes (128 bits) per 5Gbit/s TS used by that client. For example, an ODU2 client uses two TS and hence a 32-byte word, while an ODU3 client uses eight TS and hence a 128-byte word.

5.2   ODUflex(CBR) Mappings

There is no change to definition of ODUflex(CBR) signals for B100G. The ODUflex(CBR) is created by wrapping the client signal with a current OTN frame. In other words, a BMP (Bit-synchronous Mapping Procedure) process is

used that simply adds OTN overhead to the client signal stream to create a signal with a rate that is exactly (239/238)(client rate). The ODUflex rate is thus directly derived from the client signal rate. They are mapped/multiplexed into a number of OPUCn TS adequate to carry that signal. See Microsemi white paper [17] for a further description of ODUflex(CBR) signals and client mappings into them.

A variation on this BMP process is used when the CBR client consists of a stream of 64B/66B characters. Conceptually it is the same as ordinary BMP, except that a 64B/66B block will always start on an odd-numbered bit within a payload byte. 25GbE was the first client defined to use this mapping.

5.3   Considerations for Flexible Ethernet (FlexE) Client Signals

During the time when the ITU-T was developing the B100G standard, the OIF (Optical Interworking Forum) was defining a new interface called Flexible Ethernet (FlexE). The motivation behind FlexE is to decouple the Ethernet MAC and Physical Medium Dependent (PMD) sublayers, especially in terms of rates. It allows both multiple MAC flows that are each less than the PMD rate, and MAC flows that are each more than the PMD rate, to share a set of PMDs. In other words, FlexE effectively extends the OIF MLG (Multi-Lane Gearbox) to provide a variety of sub-rates and also an efficient mechanism for bonding multiple parallel links. While a full description of FlexE is beyond the scope of this white paper, it is summarized here in terms of the aspects relevant to carrying a FlexE signal over OTN. (See [18] for a tutorial on FlexE.)

FlexE defines a CBR signal that is constructed as a FlexE Group carrying one or more FlexE client signals. The FlexE Group consists of between 1 and 254 100GBASE-R Ethernet PHYs[14], all sharing a common PHY clock source.

The FlexE signal carried on each PHY consists of a round-robin repeating set of "slots" for 64B/66B characters. Specifically, the frame format for each PHY is a slot carrying the FlexE signal overhead followed by 1023 sets of 20 "calendar" slots for carrying FlexE client data. The calendar associated with each PHY is called a sub-calendar of the overall calendar for the FlexE Group. The characters in the overhead slot provide for frame alignment on each PHY and alignment across all the PHYs, in addition to carrying other required overhead for the FlexE signal. The nominal bandwidth of each calendar slot is:
(100 Gbit/s/PHY) / (20 calendar slots/PHY) = 5 Gbit/s.

Ethernet inter frame Idle insertion/deletion is used to adapt the FlexE client rate to the exact rate of the calendar slot set. The FlexE data flow is illustrated in Figure 10.

A FlexE client is a stream of 64B/66B characters associated with an Ethernet MAC packet flow, and occupies one or more of the repeating calendar slots. Sub-rate (i.e., <100 Gbit/s) clients can thus be time division multiplexed onto the same PHY by assigning them the set of calendar slots needed for that client's bandwidth. Note that the calendar slots for a sub-rate client can be spread across multiple PHYs. Clients with bandwidth greater than an individual PHY are accommodated by bonding multiple PHYs to carry the client (i.e., spreading the client's calendar slots across multiple PHYs as needed).

---

[14] OIF plans to add support for using 200GBASE-R and 400GBASE-R PHYs when IEEE 802.3 approves the associated standards. In any case, all the PHYs of a FlexE Group will use the same rate.

Figure 10   FlexE Transmit data flow illustration
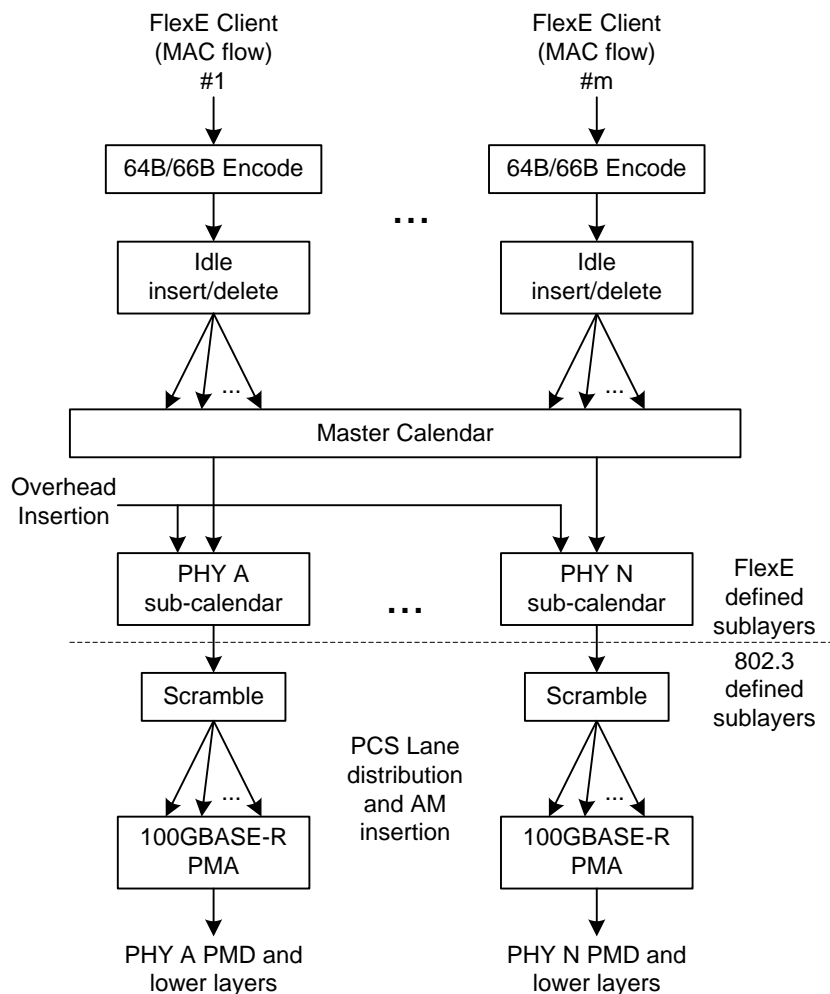
There are three options for carrying the FlexE information over OTN:

1. Terminate the FlexE and carrying the FlexE clients as Ethernet clients over OTN (either by using ODUflex(IMP) or by using the associated specified CBR mapping).

2. Simply treat each FlexE Group PHY as a 100GBASE-R signal and transport it accordingly.

3. Exploit calendar fill information within the FlexE stream, as explained below, to allow carrying a lower rate signal over OTN.

Of course, in the first option the FlexE signal itself is not actually transported over OTN. The second option is referred to as the "FlexE unaware" mapping, since the OTN is not required to know that the 100GBASE-R is carrying FlexE. The last option is known as a "FlexE aware" mapping and is described below.

The FlexE feature exploited by the FlexE aware mapping is the following. If it is known, or decided, that the clients carried in a FlexE Group do not require the entire Group capacity, the sub-calendars can be provisioned with the unneeded calendar slots marked as unavailable. A FlexE PHY can discard the Unavailable calendar slots in order to use a lower bit rate (e.g., in order to meet an optical reach objective for that PHY). The unavailable calendar slots are always located at the end of the sub-calendars (i.e., the highest numbered slots, furthest in time from the last

overhead slot). FlexE fills these unavailable slots with Ethernet Error control blocks. A FlexE aware OTN node can discard the characters from the unavailable FlexE calendar slots during the mapping process in order to construct a lower rate ("crunched") CBR client signal. Since the number of unavailable calendar slots is static, the resulting "crunched" FlexE signal is still a CBR stream that can be mapped into OTN as a partial rate signal with respect to the ingress FlexE signal. Hence, the OTN transport bandwidth requirement for the FlexE signal can be optimized for the amount of bandwidth actually required for the set of constituent FlexE clients. The OTN demapper restores the unavailable calendar slots when it creates the egress FlexE signal.

Bit-synchronous GMP (BGMP) is used for mapping the FlexE signal into an ODUflex. BGMP generates the C_m values deterministically (e.g., based on a sigma-delta algorithm) rather than deriving them from the input client rate and buffer fill.

5.4   GFP Mapping of Packet Clients

Rather than mapping GFP frames directly into the OPUCn payload area, GFP clients are mapped first into an ODUk (k ≥ 2) or an ODUflex(GFP).

GFP is used for packet clients with rates ≤100 Gbit/s, as described in [17]. Packet clients with rates >100 Gbit/s are mapped into an OPUCn using the new ODUflex(IMP) method described below in Section 5.5 rather than using GFP. It is also possible to use this new method for packet clients with rates ≤100 Gbit/s.

5.5   Packet Client Mapping Using ODUflex(IMP)

For data rates over 100 Gbit/s, it becomes unattractive to do the packet data mapping on the byte-by-byte basis defined for GFP. Most packet data protocols already use word sizes of at least 8 byte words for rates ≥10 Gbit/s. A wide-word version of GFP was considered, that would be based on 64-bit (8-byte) words. However, since Ethernet is the dominant packet client and virtually all wide word clients use the Ethernet 64B/66B line coding, it simplified the IC data paths if both native Ethernet CBR and packet clients were mapped as streams of 64B/66B words. Consequently, rather than using ODUflex(GFP) for packet clients, Q11/15 chose to define a new approach based on mapping packet clients into an ODUflex as a 64B/66B code word stream. Non-Ethernet packet clients are first encapsulated into Ethernet, with the resulting 64B/66B stream using the ODUflex(IMP) mapping.

The OPUflex rate is a constant rate that is set to be higher than the maximum client rate. Rate adaptation between the packet client data and the OPUflex is performed by inserting Ethernet Idle code words into the 64B/66B stream as defined in IEEE 802.3 for Ethernet. For that reason, the new approach was given the name "ODUflex(IMP)," where IMP stands for Idle [insertion] Mapping Procedure. This is conceptually similar to ODUflex(GFP) in which the unused payload bandwidth of the fixed-rate ODUflex(GFP) signal was filled with GFP Idle frames. Note that IEEE 802.3 specifies inserting Idle characters only between Ethernet frames, in the inter-frame gap (IFG). Hence, while these mappings are Ethernet stream mappings, they are pseudo-packet aware rather than simple CBR mappings.

The packet client signal is mapped as an $n \times 25$ Gbit/s (n = 1, 2, …) Ethernet stream. Note that Ethernet clients with rates of 10 and 40 Gbit/s are also supported. The mapping process uses the following steps:

1. The data stream presented to the mapper consists of 64B/66B characters containing Ethernet MAC frames and inter-frame gap Idle (or Ordered set) characters.
2. The character stream is first 64B/66-encoded into a FlexE (Flexible Ethernet) signal, defined by the OIF. See 5.3 and [15].
3. The resulting scrambled FlexE signal is mapped into the ODUflex(IMP). Idle characters (or Ordered sets) are inserted between Ethernet frames at this point in order to match the rate of the FlexE stream with the OPUflex rate.

4. Next the FlexE character stream is scrambled, using a self-synchronous scrambler with the same $1+x^{39}+x^{58}$ method and polynomial specified in IEEE 802.3 Clause 49.2.6 [19].
5. Finally, the resulting ODUflex(IMP) is multiplexed/mapped into the OPUCn, where GMP is used for rate justification, as with all ODUk clients being multiplexed into an OPUCn.

The process is reversed at the demapper to recover the client packet stream.

The FlexE client signal rates are defined by the OIF to be
> $s \times 5,156,250.000$ kbit/s ± 100 ppm, with $s = 2, 8, 5n$ ($n \geq 1$).

Consequently, the ODUflex(IMP) bit rate is
> $(239/238)(s \times 5.15625$ Gbit/s ± 100 ppm)

As with ODUflex(GFP), although the approaches to deriving the rates are different, both ODUflex(CBR) and ODUflex(IMP) are multiplexed into the OPUCn by GMP, and are handled in the same manner within the network. Consequently, there is no need for intermediate switching nodes to know which type of ODUflex they are switching or multiplexing.

The OPUflex(IMP) overhead in columns 15 and 16 consists of reserved bytes in all except the PSI location. The PSI (Payload Structure Indication) has the normal 256-byte PSI format, but the only defined fields are the payload type (PT = 0001 1011) in byte 0 and the client signal fail (CSF) indication in byte 2. All other bytes and bits are reserved.

5.6   Considerations for 400Gbit/s and 200Gbit/s Ethernet Client Signals

The 400GbE and 200GbE signals are important clients with rate over 100Gbit/s. The mapping of these clients into OTN will not be formally defined until IEEE publishes the associated standards. However, IEEE 802.3bs work to date has allowed Q11/15 to begin its work on the mappings.

One of the early decisions by Q11/15 is that the same mapping of the 400GbE and 200GbE PCS into an OPUC4 and OPUC2, respectively, will be used regardless of lower layers (e.g., PMD, PMA and FEC) used by the 400GbE/200GbE signal on the interface to the OTN. This allows the OTN mapper and demapper nodes to operate independently and only handle their local Ethernet interface without concern for the Ethernet interface on the other side of the OTN. Note that this decision was based on the assumption that IEEE 802.3 maintains consistency for its initial and all future PMD types, which it did not do with 25GbE.

One of the implications of maintaining a common mapping method is that the RS(544,514) FEC used on the 200GbE/400GbE interface will be terminated at the OTN mapper (i.e., the Ethernet UNI port). FEC termination has a further implication. IEEE 802.3 agreed to have an FEC Signal Degrade parameter for 200GbE and 400GbE. When the FEC is terminated at the OTN mapper, the mapping must provide a means of communicating a received FEC SD condition to the OTN demapper so that the indication can be encoded into the egress 200GbE/400GbE stream. The method of communicating the FEC SD condition across the OTN was still under study as of the first release of this white paper.

A further implication of maintaining a common mapping is that the OTN mapper must terminate the 200GbE/400GbE lane alignment markers. This is required, for example, since the Ethernet interfaces on either side of the OTN network may use a different number of lanes. In order to maintain the identical bit rate (e.g., for carrying SyncE on 200GbE/400GbE), the OTN mapper needs to insert something (e.g., Ethernet Idle characters or special rate compensation blocks) into the Ethernet stream in place of the alignment markers after their removal. The replacement for the alignment markers was still under study as of the first release of this white paper.

6 Frequency Justification / Rate Adaptation

The Generic Mapping Procedure (GMP) rate adaptation/justification method is used for multiplexing/mapping all client signals into the OPUCn. GMP consists of two parts. The first part provides a count of the number of data words that will be sent for a given client signal during the next OPUCn multiframe. This count information is carried in the JC1-JC3 bytes. The second part simultaneously communicates information regarding the amount of remaining client data buffered at the GMP source that could not be transmitted yet because the amount was less than a data word. This latter part is carried in the OPUCn JC4-JC6 bytes and the MSBs of the JC1-JC3 bytes and allows finer justification control for better jitter/wander performance. Each is described separately here at a high level. For a more detailed description with examples, see white paper ESC-2081250. [17]

6.1 Basic GMP operation

The specifics of the GMP justification method works as follows. A count value, referred to as C_m,[15] is sent in the JC1-3 octets of tributary multiframe i (see Figure 6) to indicate the number of client signal payload words that will be transmitted during tributary multiframe $i+1$ in the TS of the OPUCn payload area used by that client. As explained below, the data and stuff words are distributed throughout the OPUCn payload container in a manner that the receiver can derive directly from the received count value.

As explained inSection 5, client ODUk signals are first mapped into an ODTUCn.ts structure, which includes the information to be carried in the "ts" set of OPUCn TS used by that client, and the associated JC-byte overhead illustrated in Figure 11. Since the word size was chosen to be 16-bytes (128-bits) per OPUCn TS used by that client, the ODTUCn.ts consists of a stream of $ts \times 16$-byte words. The GMP process maps either a client word or an all-zero stuff word into the next (16)(ts) ODTUCn.ts bytes. Each (16)(ts) byte word of the ODTUCn.ts is then word-synchronously mapped into the set of the OPUCn TS used by that client. The GMP overhead provides the mechanism to determine whether a word is client data or stuff. For example, consider a client that uses two OPUCn TS, and hence a 32-byte word. The mapping into the OPUCn begins at the first appearance of one of these TS within the OPUCn multiframe. (See Section 4.2 for description of the TS appearance order.) A 32-byte data word is then mapped into the first 32-byte appearances of that set of two TS. A 32-byte data or stuff word is then mapped into each subsequent consecutive 32 byte appearances of that TS set. See Figure 8 for an example illustrating the word-to TS mapping associated with a client using 3 TS.

As explained above, regardless of the number of TS used by a client, each multiframe will contain 952 words (either data or stuff) for that client. Consequently, for the purposes of GMP, the OPUCn payload words are numbered from 1-952. The method for determining the data and stuff word locations is based on modulo arithmetic. In modulo arithmetic, the modulo remainder of X divided by Y, which is expressed as (X)modY, is the integer remainder of X when it is divided by Y. For example, (49)mod13 = 10, since 49 = (3)(13) + 10. The number of payload word locations in the OPUCn multiframe (i.e., in the server payload area) is referred to as P_{server}, since the OPUCn is the server layer channel for the client signal. So P_{server} = 952 for OPUCn. Let j be the payload word location number (1 ≤ j ≤ 952) and let C_m be the count of the number of m-bit words to be transmitted in the next multiframe. The contents of word j in multiframe $i+1$ is determined by:

---

[15] The terminology "C_m" indicates that the data and stuff word size is m-bits, and hence the corresponding count increment is m-bits. As explained above B100G will use 128 bits as the base value. As a result we have, m = (128)(#TS), where #TS is the number of 5Gbit/s TS used by that client.

$$\text{Word } j = \begin{cases} \text{data} & \text{for:} \quad (j)(C_m) \bmod P_{server} < C_m \\ \\ \text{stuff} & \text{for:} \quad (j)(C_m) \bmod P_{server} \geq C_m \end{cases} \qquad [3]$$

$$P_{server} = 952$$

The result is evenly spaced groupings of payload words and all-zero stuff words within the set of TS used by that client. The average number of payload words per multiframe is determined by the ratio of the encoded client signal rate to the payload container rate:

$$C_{m\ average} = (P_{server})(\text{client stream rate / OPUCn.ts payload container rate}) \qquad [4]$$

Another way to describe this relationship is that remainder of $(j)(C_m) \bmod P_{server}$ is incremented by an amount equal to $(P_{server} - C_m)$ each time the payload octet number is incremented by one. When this $(j)(P_{server} - C_m) \bmod P_{server}$ remainder reaches a value that is less than the previous remainder (i.e., less than C_m when the remainder is taken modulo P_{server}) a stuff word is inserted. For example, consider a client that use will use 833 words in the next multiframe, which is exactly $(7/8)(952)$. The remainder value increments by 119 for each subsequent word location, which results in a remainder of zero for every eighth word location. This type of modulo count is easily implemented in hardware.

When a tributary uses multiple TS, the JC1-JC3 GMP overhead associated with that client signal is located in the last JC overhead appearance in the multiframe associated with that set of TS. See Figure 6 for the illustration of mapping between TS number and the associated JC byte location within the multiframe, and the discussion in Section 4.2 for the order of TS appearance.

Since each GMP overhead count value is never more than 952, regardless of the size of n in the OPUCn, the GMP overhead for B100G uses a 10-bit GMP count field rather than the 14-bit count field used for ODUk. This frees four bits to extend the ∑CnD field for additional resolution. The GMP field encoding and JC octet format is illustrated in Figure 11.

The average value of C_m will rarely be an integer. Consequently, C_m must occasionally be adjusted from multiframe to multiframe. Since a mismatch between the source and sink C_m value would cause significant data corruption, it is critical to communicate C_m and its adjustments in a very robust manner.

The bit inversion mechanism allows C_m to be adjusted by ±1 and ±2. The source signals the sign and magnitude of the adjustment by transmitting the C_m with different subsets of its bits inverted. These bit inversion patterns, shown in Figure 12, were chosen to have a per-octet Hamming distance of at least three between every pattern. The JC field also includes explicit Increment Indicator (II) and Decrement Indicator (DI) bits. The JC1-3 fields are protected by a CRC-6 error check code, which allows per-multiframe changes to the C_m of any magnitude without the need for persistency checking.

Figure 11   GMP justification control overhead
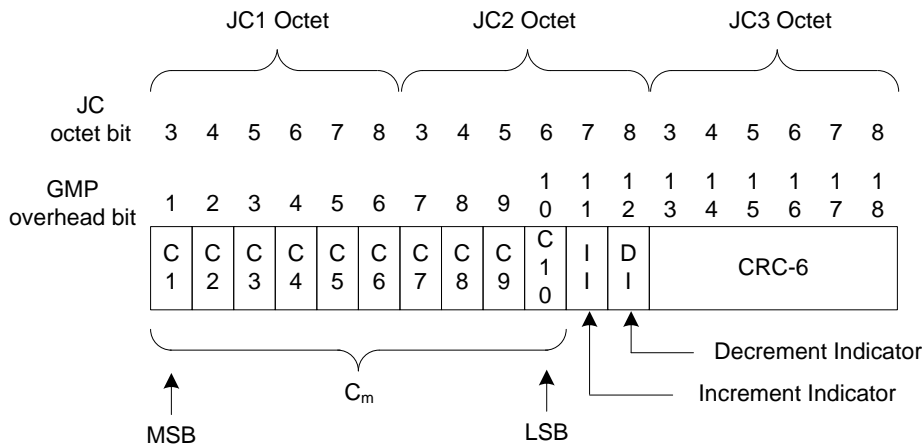
a) GMP justification control overhead fields
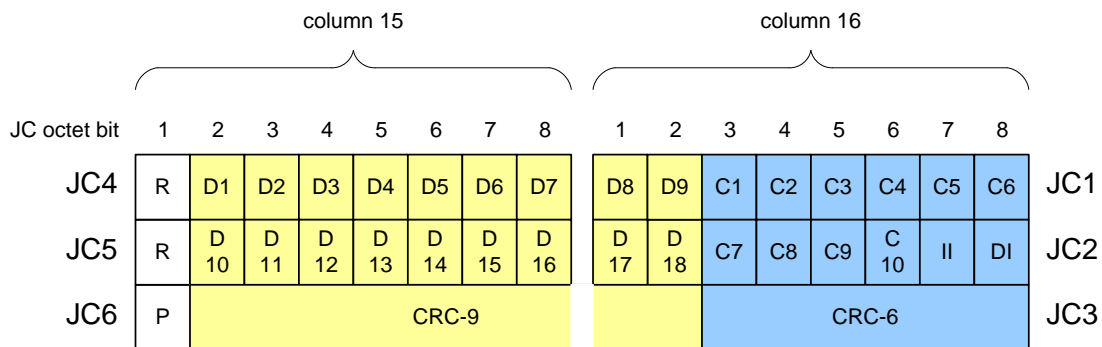
b) Octet illustration of all GMP overhead fields

As explained in detail in [17], this JC1-JC3 encoding provides extremely robust communication of the C_m values[16]. The CRC-6 is capable of detecting any 6-bit burst error, and hence can protect against the corruption of any single JC octet.[17] [18] Consequently, the combination of using the count value inversion patterns (including the II and DI) and the CRC-6 allow the receiver to correctly interpret the received C_m value in the presence of any error pattern affecting a single JC octet. The combination of the Increment and Decrement Indicators and the CRC also allow

---

[16] The robust JC octet format adopted for GMP was originally developed and proposed by the author of this white paper.

[17] Due to the spacing between JC bytes, it is assumed that an error burst will affect no more than a single JC octet per multiframe.

[18] The CRC and bit inversion patterns for the OPUCn 10-bit Cm were chosen to provide the same burst error robustness and implementation flexibility as the OPUk 14-bit Cm.

communicating an entirely new C_m, in any situation in which it is necessary. This type of change will typically only occur upon initialization, or upon entering or exiting a client signal fault condition.

Figure 12 C_m bit inversion patterns to indicate increment and decrement

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	II	DI	Δ
U	U	U	U	U	U	U	U	U	U	0	0	0
I	U	I	U	I	U	I	U	I	U	1	0	+1
I	U	U	I	U	I	I	U	U	I	0	1	−1
U	I	U	I	I	U	U	I	U	I	1	0	+2
U	I	I	U	U	I	U	I	I	U	0	1	−2
Binary value										1	1	>±2
NOTE												
− I indicates inverted Ci bit												
− U indicates unchanged Ci bit												

Initial source-sink C_m synchronization or recovery of from corruption of the sink's expected C_m value can be achieved within two frames, even in the presence of continuous increment and decrement actions.[19]

6.2   Fine-Grain Phase/Frequency Information with GMP

Since GMP operates on increments of multi-byte words, it potentially becomes more difficult to implement receiver desynchronizers with the proper jitter and wander performance. In order to solve this problem, an additional mapping "phase" information encoding capability was added in the JC4-6 bytes. As noted above, for OPUCn, the fields for this information were extended into the JC1-JC3 bytes in order to provide the fine resolution for multi-Terabit/s signals.

Specifically, the method for communicating the finer resolution timing information is to encode it in terms of phase relative to the data words. The GMP encoder makes a decision once per multiframe regarding how many data words it will transmit during the next multiframe. There will typically be some number of additional bytes remaining in the transmitter's buffers that cannot be transmitted since they constitute a fraction of a word. This fractional value is encoded in the D1-D18 field of the JC4-JC6 and JC1-JC3 bytes as a binary number representing the count of the remainder number of bytes. See Figure 11. This phase count value is referred to as the ∑CnD, which means the running Count of the Difference in the number of bytes that could be transmitted as a whole

---

[19] Achieving synchronization at the receiver requires receiving error free JC octets.

word and the bytes remaining untransmitted at the mapper/multiplexer.[20] Note that in practice, a transmitter can use an estimated filtered ∑CnD value rather than a strict measurement of the buffer fill remainder. This filtering removes the effects of incoming jitter and wander.

At the GMP receiver, the ∑CnD information helps the desynchronizer PLL control by providing a more accurate picture of the client frequency. Specifically, when the GMP word count value (C1-C10) is constant, the rate of change of the ∑CnD value indicates the frequency offset between the client signal and the OTN channel. Another way to view the ∑CnD is that it provides the receiver with an accurate indication to anticipate upcoming changes in the GMP word count value, and hence avoids multi-byte word sized steps in the PLL control. The field is protected by a CRC that would potentially allow the receiver to ignore or compensate for a value received with errors. However, since $\sum C_n D$ represents phase information, the receiver PLL can easily filter out the effect from receiving an errored $\sum C_n D$ value.

See [17] for additional discussion and specific examples for ∑CnD.

---

[20] A phase offset encoding approach was chosen for ∑CnD rather than a frequency offset encoding approach. A frequency offset approach encodes difference between the number of bytes to be transmitted in the next frame (i.e., [word size]×[C_m]) and the average number of bytes the client signal delivers during a frame period. As shown, above, the phase encoding is a running sum of the number of bytes that can't be transmitted in the next frame. Although both approaches are mathematically equivalent the phase encoding approach is more robust to transmission errors in ∑CnD. A transmission error with the frequency offset approach would result in a transient frequency error at the receiver. Transmission errors with the phase offset approach are more readily filtered and at worst result in a transient phase error. Hence, the phase offset approach is more robust to preventing short-term frequency deviations due to transmission errors.

7   Sub-rate OTUCn (OTUCn-M)

The distance over which a signal can be transmitted is a function of the signal rate. Power consumption is also a function of the signal clock rate. Consequently, there will be applications where it's desirable to transmit a B100G signal at a rate less than the discrete $N\times$OTUC rate in order to achieve the desired distance/reach for that channel. Such applications could include interconnections between two routers where the packet flow peak rate is less than the $N\times$OTUC rate, or the interconnections between two OTN cross-connects where the required capacity is less than the $N\times$OTUC rate. It could also address situations in which fiber impairments limit the capacity on a per link basis to less than the full $N\times$OTUC rate. For such applications, the B100G signal definition includes the option of transmitting a signal that has the full set of OTUCn/ODUCn overhead, but has an OPUCn consisting of only the active Tributary Slots. Specifically, an OTUCn-M signal consists of n copies of the OTUC, ODUC and OPUC overhead, and M of the 5Gbit/s TS. Since the overhead and TS each occupy 16 bytes of a row, and there are 3808/20/16 = 11.9 bytes/TS/row, the minimum OTUCn-M rate can be calculated as follows:

OTUCn-M rate = (OTUCn rate)(OTUCn-M row size/OTUCn row size)

$$= (\text{OTUCn rate})[(16)(n + 11.9M)]/[(16n)(1 + (11.9)(20))]$$

OTUCn-M rate = (OTUCn rate)$[(n + 11.9M)]/(239n)$ [5]

The availability of TS on an OTUCn-M interface are indicated in the OPUCn MSI fields. The specific values of M are a vendor-specific choice. Since OTUCn-M is a single-vendor interconnect application, the specific format of the transmitted signal (e.g., the manner in which the active TS are interleaved into a frame format) are not defined in G.709.

8   OAM&P

The key to saving network operational costs is having an effective OAM&P capability built into the signal format such that it's readily available to the service provider without impacting the client layer signal. The lack of this capability has been one of the reasons that Ethernet has been slow to take hold as a service provider transport technology.[21] While many client signals have their own native OAM&P capabilities, they are typically not comprehensive or powerful enough to provide carrier-grade functionality. Consequently, service providers prefer to map client signals into OTN for transport. Mapping all client signals into OTN also provides the possibility of a common network management framework and set of systems across their entire network.

8.1   Types of Overhead Channels

As noted above in the frame structure discussion of Section 3, the OTN B100G signals use a subset of the OAM overhead used by current OTN signals. The different OAM&P overhead channels used by B100G signals are summarized above in Table 2,Table 3, and Table 4. The functions of these different overhead types are summarized in Table 5.

8.2   Maintenance Signals

The maintenance signals used in B100G are summarized as follows:

- **AIS** – The OTUCn AIS signal is an all-ones pattern in all the frame bits except for the framing overhead bytes. The ODU AIS signal is sent at the ODUCn level in response to upstream failures. ODUCn-AIS is an all-ones pattern in the OPUCn (payload and overhead) and ODUCn overhead.

- **LCK** –The ODU LCK signal is a repeating 0110 0110 pattern in the entire ODUCn signal except for the frame alignment and OTUCn overhead.

---

[21] Service provider-type OAM&P capability has been added to Ethernet. Any Ethernet OAM&P, however, must travel in-band as an Ethernet frame in the same channel as the client data frames. This means that Ethernet OAM&P frames consume client signal bandwidth and require all NEs that make use of this OAM&P information to be capable of removing and inserting the OAM&P frames from the client data stream.

Table 5   OAM&P channel definitions

OAM&P channel	Used in	Function
APS / PCC	ODUCn	Automatic Protection Switching / Protection Communications Channel – While the use of this channel for B100G is for further study, it will be relatively simple since the ODUCn signal is never switched.
BDI	OTUCn, ODUCn PM & TCM	Backward Defect Indication – Sent from the overhead sink to the source to indicate that a defect has been detected in the forward direction.
BEI	OTUCn, ODUCn PM & TCM	Backward Error Indication – A binary count of the number of BIP-8 bits indicating errors, sent from the overhead sink to the source.
BIAE	OTUCn, ODUCn TCM	Backward Incoming Alignment Error – Indication sent from the overhead sink to the source that it received an IAE.
BIP-8	OTUCn, ODUCn	8-bit Bit Interleaved Parity- Used in the OTUCn SM, ODUCn PM.
DM	ODUCn PM, ODUCn TCM	Delay Measurement overhead to determine the round trip delay associated with the ODUCn path
GCC	OTUCn, ODUCn	General Communications Channel –One available in the OTUCn overhead (GCC0) and 2 in the ODUCn overhead (GCC1 and GCC2). The format is of the GCC channels is not specified in G.709.
IAE	OTUCn	Incoming Alignment Error – Indication sent downstream to inform the receiving NEs that a framing alignment error (e.g., a slip) was detected on the incoming signal. Primarily used to suppress BIP error counting.
MFAS	OTUCn	Multiframe Alignment Signal – Binary counter used to establish the 256-frame multiframe that is used for the time-shared overhead channels that spread their content over the course of a multiframe.
OA	OTUCn	Optical Alignment – Frame alignment signal for the OTUCn. OA1 = 1111 0110 and. OA2 = 0010 1000
TTI	OTUCn, ODUCn PM & TCM	Trail Trace Identifier –Used to check that the OTUCn and ODUCn being received are the ones expected.

8.3   Delay Measurement

Some services carried over OTN are sensitive to delay. Examples of such service applications include SAN signal transport and financial transaction communications for stock or commodity trading. Delay is important for SAN transport applications because the SAN signal throughput is a function of the delay in communicating a bandwidth grant from the sink to the source. The financial transaction application was brought about by computer aided trading. Studies have shown that a having a latency advantage of just several milliseconds over competitors can result in a substantial increase in profits from timely trades. While network engineers can calculate the delay for a circuit through their network, the users want a guarantee that the delay has been confirmed by measurement, and that it has not changed due to events like network protection switching. Consequently, an integrated delay measurement capability was added to OTN in 2009 at the request of multiple service providers. This capability has been carried forward into the B100G signals also.

The latency measurement is performed as follows. One node is provisioned to be the source of the delay measurement pattern, and the node at the other end of the path is provisioned to be the loopback node for that

measurement signal. The originating node then measures the delay between sending the pattern and receiving it back from the far-end node. Specifically, the source node initiates the latency measurement by toggling the PM delay measurement bit and initiating an OTN frame counter.   The loopback node for that signal transmits the received delay measurement bit value back to the source in the same overhead bit location. The source node measures the round-trip delay as the number of OTN frame periods between when it transmits the toggled bit and when it receives that bit value from the loopback node. The measurement resolution is approximately two ODUCn frames. The resulting approximately 2µs resolution is over two orders of magnitude better than the 500µs resolution required by the applications.

9   Carrying Frequency and Time-of-Day Synchronization information over B100G

One of the key early decisions for the original OTN standard was that it would not be required to transport network synchronization as part of the OTN signal. Since OTN client signals such as SONET/SDH can transport this synchronization, there was no compelling reason to add the extra complexity and stringent clock requirements to the OTN signals. The only constraint was that the OTN justification scheme for mapping SONET/SDH clients had to guarantee that these clients could be carried without causing them to violate the ITU-T Rec. G.825 jitter and wander specifications.[22]

However, since that time applications have emerged that can benefit from carrying precision timing information within the OTN signal itself rather than through one of its clients. One such application is being able to carry precision frequency and time-of-day information to the access edge of the OTN in order to time access links, including those to radio base stations. For B100G, the main application is to be able to distribute precision clock frequency and time-of-day information to the different COs within the network, without having to waste the bandwidth associated with using a client signal for it.

This capability was added to OTN at the November 2014 ITU-T SG15 meeting, and it involved the following elements:

- The timing information is carried in a digital overhead channel associated with the PHY, referred to as the OTN Synchronization Message Channel (OSMC).

- The overhead carries a packet-based precision timing protocol (PTP) using timestamps, which is an adaptation of the IEEE 1588v2 protocol.

- The Boundary Clock method will be used initially, since the Transparent Clock method may be too complex to be practical.

For OTUk, the OSMC is carried in an OTUk overhead channel. For OTUCn, however, the OSMC is carried in the FlexO signal overhead, since FlexO is the closest digital layer to the PHY. See Section 10 for the FlexO description.

---

[22] The jitter and wander requirements for OTN network interfaces are specified in ITU-T G.8251.

10 Flexible OTN (FlexO)

10.1 Introduction to FlexO

The FlexO approach was defined by the ITU-T Q11/15 in order to provide a flexible, modular mechanism to support different line rates with B100G signals. FlexO has conceptually similarities to the OIF FlexE. (See section 5.3 and [15] for more on FlexE.) Like FlexE, FlexO is a modular interface consisting of a set of 100Gbit/s optical PHY streams. For FlexO, a set of n 100Gbit/s PHYs are bonded together to carry an OTUCn, with each 100Gbit/s PHY carrying an OTUC slice. This flexibility allows using any value of "n" for an OTUCn interface rather than defining only certain discrete values of n (e.g., just n = 4 and 10). Further, if a set of m 100Gbit/s PHYs are available, as subset of n PHYs can be chosen to carry an OTUCn ($n<m$). For example, this would allow choosing the subset of PHYs with the best optical channel characteristics, or carrying multiple OTUCn signals over a set of PHYs.

FlexO takes advantage of being able to use existing 100GbE/OTU4 optical modules for the individual FlexO PHYs, thus benefitting from the lower cost of these optical modules. As future higher rate Ethernet modules become available (e.g., 200Gbit/s or 400Gbit/s PHYs), FlexO can be extended to also make use of them.

Rather than embedding the FEC within the OTUC frame as was done with OTUk, the OTUCn stream is mapped into a stream of FlexO FEC codewords. The alignment between the OTUC frame and the FEC codeword stream is arbitrary. This approach has already become common with strong soft-decision FEC coding for very long reach single-vendor IaDI applications.

FlexO makes partial reuse of the lane architecture and FEC structure from 100GbE and 400GbE Ethernet in order to leverage Ethernet IP and optical modules for lower cost. The primary reuse from 100GbE is the existing higher-volume 100GbE/OTU4 optical modules as the PHY. The lane structure is based on 400GbE, and the FEC is one that has been used in both 100GbE and 400GbE.

The IEEE 802.3 "KP4" Reed-Solomon RS(544,514,10) FEC is used for each 100Gbit/s FlexO stream. While the current OTN GFEC RS(255,239,8) Reed-Solomon code operates on 8-bit symbols, KP4 operates on 10-bit symbols. The overall performance of GFEC and KP4 are similar, however, the larger symbol size of KP4 allows achieving this performance with less overhead bandwidth.[23] Although the ODUC rate is higher than ODU4, the lower overhead rate of KP4 when applied to the ODUC achieves a bit rate slightly lower than the OTU4 using GFEC.[24] This allowed defining FlexO such that its rate is close enough to OTU4 to directly re-use optical modules that support OTU4.

10.2 FlexO Frame Format

The FlexO frame structure, as shown in Figure 13, consists of 128 FEC codewords. There are eight frames in the FlexO multiframe. The first FEC codeword of the frame begins with the lane Alignment Markers (AMs), which are reused from 802.3bs 400GbE. Since the optical PHY rate is 100Gbit/s, there are four 25Gbit/s logical lanes, and hence only the first four 400GbE 120-bit AMs are used. See Table 6 for the AM values.[25] In order to accommodate

---

[23] 255/239 = 1.0669, and 544/514 = 1.0584

[24] The ratio of the ODUC and ODU4 rates is (227/226) = 1.00442. The ratio of an ODUC directly using KP4 to an OTU4 with GFEC is (239/226)*(544/514)*(227/255) = 0.99635.

[25] Like most data communications protocols, IEEE 802 specifies bytes as being transmitted LSB first. Since OTN follows telecommunications convention, it transmits MSB first. Consequently, G.709.1 translates the 802.3bs hexadecimal AM values in order to reflect the MSB-first transmission order. For convenience, the table here shows both.

using future 200Gbit/s Ethernet optical modules, the FlexO AM field is sized to carry an additional four AMs. Consequently, the AM field at the beginning of the FlexO frame has (4+4)×(120-bits) = 960 bits. When only four AMs are required, they are located in bits 1-480, with bits 481-960 containing padding that is all-zeros (prior to scrambling).

Figure 13   FlexO frame and multiframe illustration
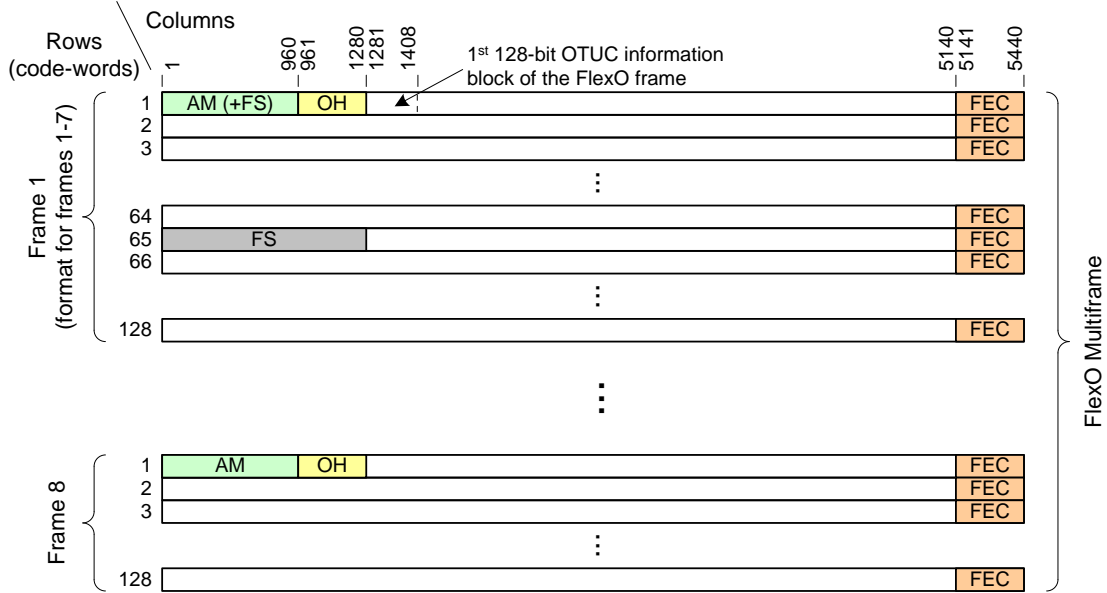
Table 6   FlexO and equivalent 803.3bs Alignment Markers

Field	CM0	CM1	CM2	UP0	CM3	CM4	CM5	UP1	UM0	UM1	UM2	UP2	UM3	UM4	UM5
Eth AM0	9A	4A	26	B6	65	B5	D9	D9	01	71	F3	26	FE	8E	0C
FlexO AM0	59	52	64	6D	A6	AD	9B	9B	80	8E	CF	64	7F	71	30
Eth AM1	9A	4A	26	04	65	B5	D9	67	5A	DE	7E	98	A5	21	81
FlexO AM1	59	52	64	20	A6	AD	9B	E6	5A	7B	7E	19	A5	84	81
Eth AM2	9A	4A	26	46	65	B5	D9	FE	3E	F3	56	01	C1	0C	A9
FlexO AM2	59	52	64	62	A6	AD	9B	7F	7C	CF	6A	80	83	30	95
Eth AM3	9A	4A	26	5A	65	B5	D9	84	86	80	D0	7B	79	7F	2F
FlexO AM3	59	52	64	5A	A6	AD	9B	21	61	01	0B	DE	9E	FE	F4
Note that the corresponding Ethernet and FlexO AM values are identical in how they appear on a serial interface. This table translates between the respective IEEE 802.3 and ITU-T approaches to expressing the bytes in hexadecimal format.															

When using the KP4 FEC, Ethernet streams are transcoded from 64B/66B block codes to 256B/257B block codes for increased bandwidth efficiency.[26] The KP4 code was chosen by IEEE 802.3 such that an integer number of 256B/257B blocks fit into an FEC codeword (i.e., 5140 = 20×257). When 200GbE or 400GbE AMs are inserted into the FEC codeword stream, they are inserted at the beginning of an FEC codeword, followed by padding such that the AMs and padding end on a 257-bit boundary. Since OTUCn does not use the 64B/66B or 256B/257B Ethernet line coding, FlexO was defined to use the following data alignment method within the FEC codeword stream.

As explained above, OTUCn uses a 128-bit modularity for its frame overhead and payload Tributary Slots. Consequently, it was desirable to have the OTUC information begin on a 128-bit boundary in the first FEC codeword of the FlexO frame, and have each 8-codeword FlexO frame contain an integer number of 128-bit OTUC blocks[27]. As shown in Figure 13, this initial 128-bit alignment is enabled by having a 320-bit FlexO overhead field immediately following the 960-bit AM field, creating a 1280-bit combined field. Since the 5140-bit FEC codeword payload is not divisible by 128, the alignment of the codeword and 128-bit block boundaries shift in each successive codeword of the frame.

As can be seen from Figure 13, in each FlexO frame there are:

$(8)(128)(5440) = 5570560 = 85 \times 2^{16}$ total bits, and
$[(8)(128)(5140) - (7+8)(1280)] = 5244160 = 85 \times 2^8 \times 241$ OTUC payload bits.

Consequently, the FlexO rate is:

$(256/241)(\text{OTUC rate}) = (256/241)(239/226)(99.5328 \text{ Gbit/s}) = 111.809474 \text{ Gbit/s}.$

Since this rate is -4.46ppm relative to the nominal OTU4 rate, a FlexO signal can readily use an OTU4 PHY module.

10.3   FlexO Overhead

The overhead functions are associated with the FlexO PHY operations, including bonding multiple OTUC1 members. The overhead format is shown in Figure 14. As explained above, the overhead field is 320 bits (40 bytes) long. For increased field capacity, it is time shared across an 8-frame overhead multiframe, determined by the three LSBs of its MFAS field. The field definitions and functions can be summarized as follows:

GID (Group ID) is the number assigned to the group of PHYs carrying a particular OTUCn signal. For example, if multiple OTUCn signals are carried over a set of PHYs, the GID would distinguish between them at the receiver. (GID = 0 indicates that a PHY is not part of any FlexO group.)

PID (PHY ID) is the number assigned to that PHY within the FlexO group. The receiver uses this information to restore the OTUCn by correctly re-ordering the received signals. The PID values do not need to be sequential. For example, if the group is carried over a subset of the available PHYs, the PID values could correspond to the PHY numbers used by the equipment.[28] In any case, the order of the

---

[26] For robustness, the 64B/66B blocks use a pair of flag bits per 66-bit block to indicate whether the block contains only data or also contains control characters. Since the FEC also protects any flag bits, it becomes practical to use a single flag bit in a 256B/257B block rather than the eight flag bits that would be required for the equivalent four 64B/66B blocks.

[27] In a FlexO frame, there are:
[(128 codewords)*(5140 bits/codeword) – (1280)*(2) AM, OH & FS bits] / (128 bits/block)
= 5120 128-bit blocks/frame

[28] However, the same PID value must be used in both directions of transmission.

interfaces is always in rising PID value order from the lowest to the highest PID value used by the group. The OTUC slices are mapped from slice #1 to slice #n in order onto the interfaces (PHYs).

MAP The MAP field indicates which specific PHYs are used by that FlexO group. For example, since the PID values are not required to be sequential, the MAP confirms to the receiver which PID values it should be receiving for members of that group. The MAP field contains one bit corresponding to each of the 256 possible PID values, and a "1" in a bit position indicates that the corresponding PID is being used by this group.

AVAIL indicates the number of "available" and valid OTUC slices that are mapped into the FlexO frame for that PHY. This field is not especially useful with 100Gbit/s PHYs, since AVAIL can only be "0" or "1." However, future higher rate PHYs (e.g., 200Gbit/s and 400Gbit/s) would allow having only a subset of the PHY's capacity used for carrying actual OTUC slices. AVAIL then tells the receiver how much of the PHY capacity is in use.

STAT is similar to the OTUCn STAT in that it provides status information about the PHY connection. The only defined STAT function at this time is to report a far end (remote) PHY failure (RPF) condition.

CRC A simple persistency check could have been used for the FlexO overhead, since most of the fields contain static provisioned values and the receiver doesn't need to respond to the overhead in real-time. However, a CRC was added to cover the fields in columns 2-10 of Figure 14, since a CRC can provide a simpler method of regularly confirming the integrity of the received overhead values than performing frequent persistency checks.

FCC (FlexO Communications Channel) is included for the purpose of managing the interface. It is not a generic/general purpose communications channel like the OTN GCCx channels. The FCC channel rate is slightly less than 18Mbit/s.

OSMC (OTN Synchronization Message Channel). The OSMC is only active on the first member of the FlexO group (i.e., the member with the lowest PID value). The OSMC function is described in section 9.

Figure 14   FlexO overhead illustration

frames \ bytes	1	2	3	4	5	6	7	8	9	10	11	12	13	...	26	27	28	29...40
xxxx x000	MFAS	STAT	GID	GID	GID	RES	PID	MAP				CRC			FCC		OSMC*	
xxxx x001	MFAS	STAT	AVAIL					MAP				CRC			FCC		OSMC*	
xxxx x010	MFAS	STAT						MAP				CRC			FCC		OSMC*	
xxxx x011	MFAS	STAT						MAP				CRC			FCC		OSMC*	
xxxx x100	MFAS	STAT			RES			MAP				CRC			FCC		OSMC*	RES
xxxx x101	MFAS	STAT						MAP				CRC			FCC		OSMC*	
xxxx x110	MFAS	STAT						MAP				CRC			FCC		OSMC*	
xxxx x111	MFAS	STAT						MAP				CRC			FCC		OSMC*	

* The OSMC is only used on FlexO interface member #1. These bytes are Reserved for the other FlexO interface members.

10.4   FlexO Data Flow Processing

The transmit and receive data flows are illustrated in Figure 15 and Figure 16, respectively.

Unlike Ethernet, which uses a self-synchronous scrambler, FlexO takes advantage of its fixed frame structure to use a less complex frame-synchronous scrambler. The FlexO scrambler is the same $x^{16} + x^{12} + x^3 + x + 1$ scrambler used for OTUk scrambling, except that it is reset (to 0xFFFF) on the first bit of the FlexO frame. For a further simplification, the scrambler is specified to run over all the bits of the FlexO frame rather than pausing during overhead and FEC fields. In the transmit direction, this can be implemented as feeding the scrambler entire 5440-

bit blocks with dummy FEC and AM field bits. The actual AM information then overwrites the AM field bits between the scrambler output and FEC encoding input. The dummy FEC fields are overwritten by the FEC encoder.

Figure 15   FlexO transmit flow illustration
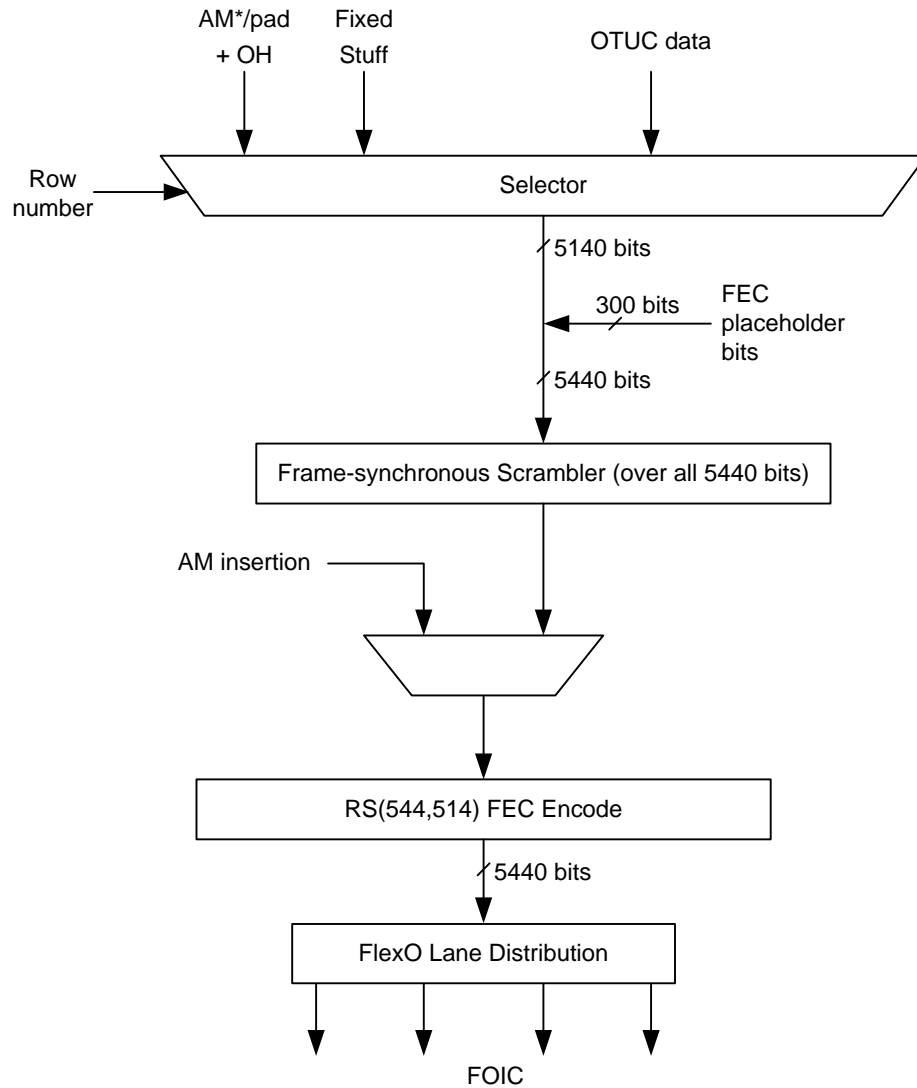
* The AM fields are inserted as all-zeros prior to scrambling, and then overwritten with the actual AM values prior to the FEC encoding

Figure 16   FlexO receive flow illustration
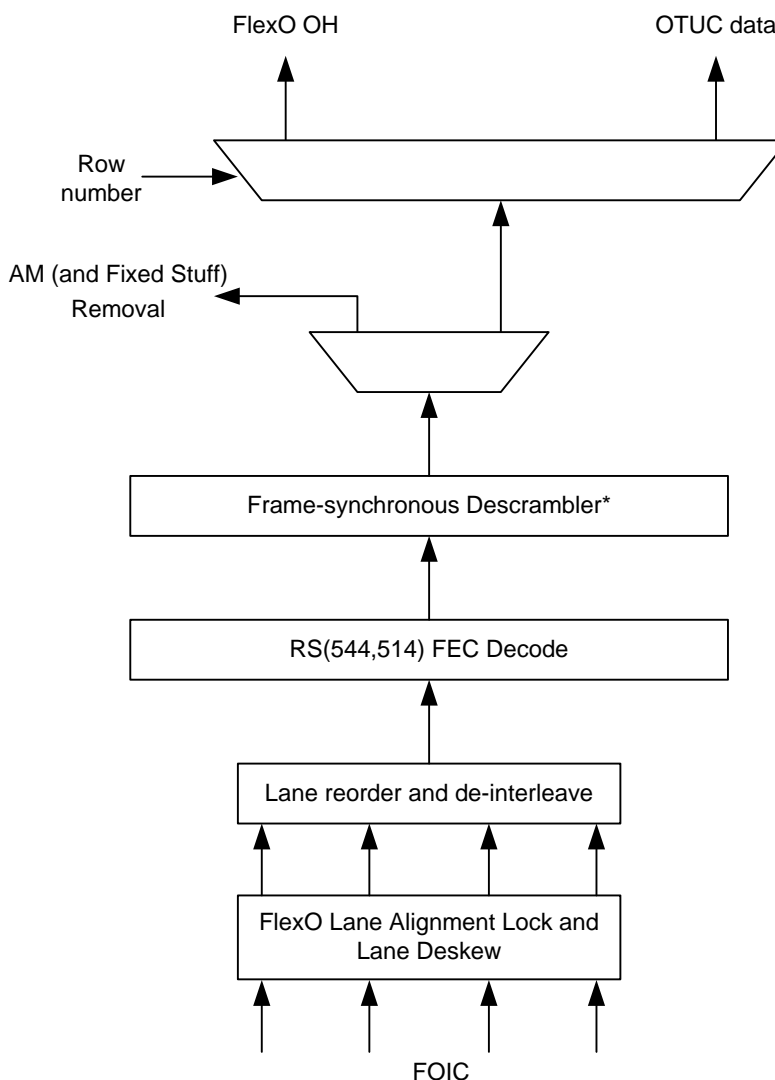
* The descrambler operates as if it were performed across entire
    5440-bit FEC codeword blocks, even if the the FEC decoder
    outputs only the 5140 FEC information bits.

10.5   FlexO Interface (FOIC)

The FlexO interface is designated as FOICx.y, where the "*Cx*" indicates the interface rate (i.e., *x* × 100Gbit/s, following the same convention as OTUCn), and *y* indicates the number of PHY lanes being used. The first release of

G.709.1 only specified the FOIC1.4 described here[29]. The output of the FEC encoder is distributed to the four 28Gbit/s lanes of the FOIC in the same 10-bit round-robin manner that it is specified in clause 91 of IEEE 802.3.

It should be noted here that the AM values of Table 6 are inserted into the AM field (Figure 13) such that they appear in the appropriate sequence on each logical lane. Specifically, since the FEC stream is distributed to the logical lanes in 10-bit increments[30], the FlexO frame AM field begins with the first 10 bits of AM0, followed by the first 10 bits of AM1, followed by the first 10 bits of AM2, followed by the first 10 bits of AM3, followed by the next 10 bits of AM0, etc. This AM field format results in the AM0 row (Table 6) appearing in order on only FOIC lane 0, the AM1 row appearing on FOIC lane 1, etc. The result is illustrated in Table 7.

At the receiver (see Figure 16), the FlexO overhead GID, PID, and MAP fields provide the group and PHY identity, and PHY sequence, required to reconstruct the OTUCn signal. Specifically, within a given FlexO group, OTUC slice #1 is mapped onto the interface with the lowest PID value, OTUC slice #2 onto the interface with the next highest PID value, etc., with OTUC slice #n mapped onto the group's interface with the highest PID value.

The receiver uses the OTUC FAS to perform deskew between the OTUC slices, as specified in G.709. [1] The skew tolerance requirement is 300ns.

Table 7   AM to FOIC logical lane mapping illustration

AM bits	Lane 0 10-bit symbol of AM0	Lane 1 10-bit symbol of AM1	Lane 2 10-bit symbol of AM2	Lane 3 10-bit symbol of AM3
1 – 40	0101100101	0101100101	0101100101	0101100101
41 – 80	0100100110	0100100110	0100100110	0100100110
81 – 120	0100011011	0100001000	0100011000	0100010110
121 – 160	0110100110	0010100110	1010100110	1010100110
. . .				
441 – 480	0100110000	0010000001	0010010101	1011110100
NOTE – Transmission order of each 10-bit word is left-to-right (MSB first). The transmission order within the FlexO frame is left-to-right across the row, and down the table. The transmission order for each lane is per-word and down the table.				

---

[29] Other future FOICx.y interfaces may include FOIC1.2 when 56G electrical lane interfaces become common, and FOIC2.y or FOIC4.y as 200GbE and 400GbE Ethernet modules become available. Note that FOIC2.y and FOIC4.y will require a different data flow than is illustrated here, since both 200GbE and 400GbE interleave the outputs of two separate FEC engines into the lanes rather than the single FEC engine used with FOIC1.y. Using two interleaved FEC streams provides additional robustness for the higher interface rates, and FlexO will adopt the same FEC interleaving. The 10-bit symbol oriented mapping onto the lanes is the reason why the AM plus overhead field was chosen to be divisible by both 10 and 128, rather than just by 128 (e.g., using 1280 bits rather than 1024 bits).

[30] The 10-bit words mapped onto the lanes correspond to the 10-bit symbols within the RS(544,514,10) FEC.

10.6   Future Extensions to FlexO

One extension discussed above is using 200Gbit/s and 400Gbit/s optical PHYs when they become available for 200GbE and 400GbE.

Q11/15 has also approved a project to define long-reach FlexO interfaces. The KP4-based FlexO described here has short-to-medium reach capabilities in a telecom network. A long-reach FlexO interface will require a stronger FEC. The initial assumption is that a hard-decision FEC will be chosen rather than a more complex soft-decision FEC.

10.7   Putting it Together

Figure 17 illustrates the frame format flow from client signals through the OTN B100G signal and FlexO and onto the fiber.

Figure 17   Illustration of mapping from client signals through OTUCn and FlexO onto the optical medium
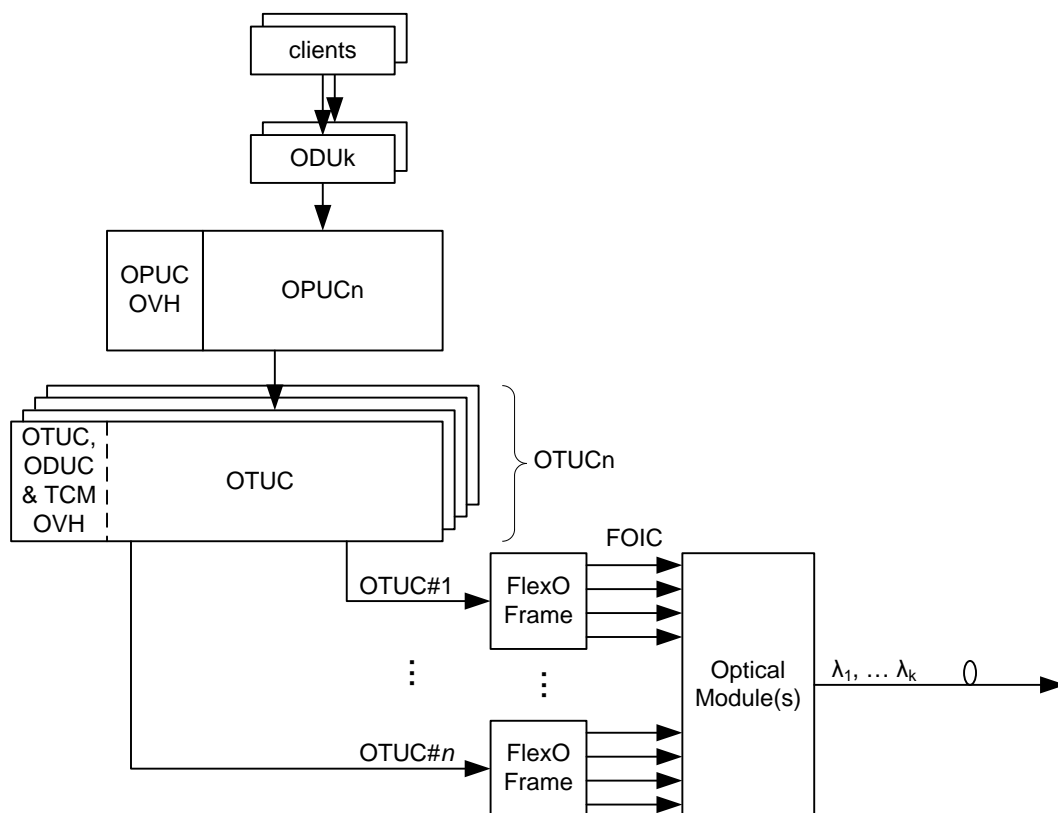
11 Conclusions

Originally introduced in 2000, OTN has proven capable of continuing to evolve and adapt to cover new applications and technology in the transport network. The OTUCn B100G standards work is the next stage of that evolution. The modular frame and rate structure readily allows:

- A convenient way for the OTUCn signal to be divided for transmission over multiple wavelengths.

- Supporting multi-vendor interfaces of different rates.

- A straightforward adaptation to carry future high-speed client signals, such as Terabit/s Ethernet.

- The introduction of FlexO for a modular interface that can exploit Ethernet/OTU4 PHY modules.

While the initial application for OTUCn signals will be long haul connections, it will see metro network applications as 400GbE and FlexE become available and data-center connectivity increases in importance.

Microsemi is a leader in OTN solutions for micro-OTP, ROADM, OTP, P-OTP, and CESR line cards, beginning with the innovative HyPHY, DIGI and META product families.[31] Microsemi's OTN solutions enable service providers to accelerate their network transition to OTN, delivering cost effective grooming and switching of services end-to-end in the new Packet Optical Transport Network. For more detailed information on the HyPHY, DIGI and META product families and the complete portfolio of OTN solutions from Microsemi, please visit the Microsemi Wireline Infrastructure solutions web page at www.microsemi.com.

---

[31] PMC-Sierra played a leading role as one of the most active participants in the evolution of the ITU-T OTN standards. PMC-Sierra was author or co-author on many of the key standards contributions and also served in key editor and leadership roles. These active participation and leadership roles have continued now that PMC-Sierra has become part of Microsemi.

Appendix A – G.709 Terminology Changes

When SG15 was developing the fifth edition of G.709 (2016), it had become apparent that some of the earlier terminology was no longer appropriate. For example, the IrDI (Inter-Domain Interface) was originally intended to be a relatively short reach interface between two service provider domains, while the IaDI (Intra-Domain Interface) would typically be a longer reach interface between two network elements from the same equipment vendor. However, the reach of the IrDI makes it adequate for many intra-domain applications, and a longer reach may be required for some inter-domain interfaces. Also, it has become common to specify multi-vendor versions of the IaDI (MV-IaDI) to allow at least limited interoperability. Both IrDI and IaDI typically use the same basic OTN frame format so that they can make use of the same framer and mapper circuits. Consequently, the terms IrDI and IaDI have largely lost their original meaning. G.709 (2016) has been updated to remove the IrDI and IaDI terms under the assumption that all OTN signals with any degree of electrical domain interoperability would use the same G.709 frame formats.

At this point in the evolution of OTN technology, the more meaningful distinction is that difference between whether optical domain multiplexing (WDM) is used on the interface, and whether there are multiple or a single wavelength used to carry the OTN signal.

12   References

[1]   ITU-T Recommendation G.709 (2016), *Interfaces for the Optical Transport Network (OTN)*

[2]   ITU-T Recommendation G.798 (2012), *Characteristics of optical network hierarchy equipment functional blocks*

[3]   ITU-T Recommendation G.872 (2012), *Architecture of optical transport networks*

[4]   M. Elanti, S. Gorshe, L. Raman, and W. Grover, *Next Generation Transport Networks – Data, Management, and Control Plane Technologies*, Springer, 2005.

[5]   ESC-2030895, "A Tutorial on SONET/SDH," Microsemi white paper by S. Gorshe.

[6]   S. Kartalopoulos, *Introduction to DWDM Technology*, IEEE Press, Piscataway, HJ, 2000

[7]   ITU-T Recommendation G.692 (2005), Optical interfaces for multichannel systems with optical amplifiers

[8]   ITU-T Recommendation G.8251 (2010), T*he Control of Jitter and Wander within the Optical Transport Network*

[9]   ITU-T Recommendation G.694.1 (2012) *Spectral grids for WDM applications: DWDM frequency grid*

[10]   ITU-T Recommendation G.694.2 (2003) *Spectral grids for WDM applications: CWDM frequency grid*

[11]   ESC-2061695, "ROADMs and the Evolution of the Metro Optical Core," Microsemi white paper by T. Mashologu and M. Orthodoxou

[12]   ITU-T Recommendation G.7041 (2016) Generic Framing Procedure (GFP)

[13]   ESC-2110752, "Enabling OTN Convergence – Solutions for the New Packet Optical Transport Network," Microsemi white paper by Mark Orthodoxou

[14]   ITU-T Supplement 58 (2016) Optical transport network (OTN) module frame interfaces (MFIs)

[15]   OIF ** (2016) Flex Ethernet Implementation Agreement

[16]   ITU-T Recommendation G.709.1 (2016) *Flexible OTN short-reach interface*

[17]   ESC-2081250, "A Tutorial on ITU-T G.709 Optical Transport Networks (OTN)," by Steve Gorshe

[18]   OIF FlexE Application Note (http://www.oiforum.com/documents/educational-white-papers/)

[19]   IEEE 802.3:2015, *IEEE Standard for Ethernet*

13 Glossary and Abbreviations

Term	Definition
3R	Re-amplification, Reshaping and Retiming
10GbE	10 Gbit/s Ethernet
40GbE	40 Gbit/s Ethernet
100GbE	100 Gbit/s Ethernet
AM	Alignment Marker (Ethernet and FlexO)
CBR	Constant Bit Rate
CESR	Carrier Ethernet Switch / Router
CMx	Portions of the Alignment Markers common to all lanes
CO	Central Office (of a telephone network provider)
CRC-n	n-bit Cyclic Redundancy Check error detection code
CWDM	Coarse Wavelength Division Multiplexing
DFB	Distributed Feedback laser
DWDM	Dense Wavelength Division Multiplexing
EDFA	Erbium-Doped Fiber Amplifier
FCC	FlexO Communications Channel
FEC	Forward Error Correction
FlexE	Flexible Ethernet (from OIF)
FlexO	Flexible OTN (G.709.1)
FOIC	FlexO Interface
GFP	Generic Framing Procedure (ITU-T Rec. G.7041)
GFP-T	Transparent mode of GFP
GID	FlexO Group ID
GMP	Generic Mapping Procedure
HAO	Hitless Adjustment of ODUflex(GFP) signals

Term	Definition
IaDI	Intra-Domain Interface
IMP	Idle Mapping Procedure
IrDI	Inter-Domain Interface
JC	Justification Control
JOH	Justification Overhead
KP4	Reed-Solomon RS(544,514,10) FEC from IEEE 802.3
MAP	FlexO overhead field for the OTUC to FlexO PHY mapping
MFAS	MultiFrame Alignment Signal
MSI	Multiplex Structure Identifier
OADM	Optical Add-Drop Multiplexer
OCC	Optical Channel Carrier
OCh	Optical channel with full functionality
ODU	Optical Channel Data Unit
ODUC	100Gbit/s element (slice) of an ODUCn
ODUCn	$n \times$ 100Gbit/s Optical Channel Data Unit
ODUflex(CBR)	Flexible rate ODU for carrying CBR client signals
ODUflex(GFP)	Flexible rate ODU for carrying packet client signals that use a GFP-F mapping into the OPUflex
ODUflex(IMP)	Flexible rate ODU for carrying packet client signals with Ethernet Idle characters used for rate adaptation when mapping into the OPUflex
ODUk	Optical Channel Data Unit-k
ODTUjk	Optical channel Data Tributary Unit j into k
ODUk-Xv	X virtually concatenated ODUk's
OH	Overhead
OMS	Optical Multiplex Section
ODTU	Optical channel Data Tributary Unit
ODTUCn	ODTU for multiplexing an ODU signal into an OPUCn

Term	Definition
ODTUCn.ts	ODTU for multiplexing an ODU into "ts" Tributary Slots of an OPUCn
OPU	Optical Channel Payload Unit
OPUC	100Gbit/s element (slice) of an OPUCn
OPUCn	$n \times$ 100Gbit/s Optical Channel Payload Unit
OPUk	Optical Channel Payload Unit-k
OSC	Optical Supervisory Channel
OSMC	OTN Synchronization Message Channel
OTP	Optical Transport Platform
OTU	Optical Channel Transport Unit
OTUC	100Gbit/s element (slice) of an OTUCn
OTUCn	$n \times$ 100Gbit/s Optical Channel Transport Unit
OTUk	completely standardized Optical Channel Transport Unit-k
OXC	Optical cross-connect equipment
P-OTP	Packet Optical Transport Platform
PID	FlexO PHY ID
PMD	(Ethernet) Physical Medium Dependent sub-layer
PSI	Payload Structure Identifier
PT	Payload Type
Q11/15	Question 11 of ITU-T Study Group 15, which is the standards group responsible for OTN standardization
ROADM	Reconfigurable Optical Add / Drop Multiplexer
TC	Tandem Connection
TCM	Tandem Connection Monitoring
UMx	Portions of the Alignment Markers unique to each lane
WDM	Wavelength Division Multiplexing

Notes

Microsemi
Power Matters.™

About the Author

Steve Gorshe, Ph.D. is a Distinguished Engineer at Microsemi, working on technology for optical transmission and access systems.

Steve is currently the Associate Rapporteur for Q11 of ITU-T Study Group 15, which is responsible for optical transport network standards. He is also technical editor for multiple ITU-T standards, including G.7041 (Generic Framing Procedure - GFP) and G.Sup43 (Transport of IEEE 10G Base-R in Optical Transport Networks (OTN)), G.Sup56 (OTN Transport of CPRI Signals) and co-editor for G.709 (OTN). He came to Microsemi through its acquisition of PMC-Sierra, where he has worked since 2000. He has worked since 1982 in research and development of telecommunications systems and ICs including holding the position of Chief Architect for NEC eLuminant Technologies. Steve is an IEEE Fellow, has served as Editor-in-Chief for *IEEE Communications Magazine*, and currently a Board of Governors Member-at-Large for the IEEE Communications Society. He has 38 patents issued or pending, over 24 published papers, and is co-author of two telecommunications textbooks and chapters in three additional textbooks. Steve received his Ph.D. and MSEE from Oregon State University and BSEE from the University of Idaho.

Microsemi
Power Matters.™

Microsemi Corporation (Nasdaq: MSCC) offers a comprehensive portfolio of semiconductor and system solutions for aerospace & defense, communications, data center and industrial markets. Products include high-performance and radiation-hardened analog mixed-signal integrated circuits, FPGAs, SoCs and ASICs; power management products; timing and synchronization devices and precise time solutions, setting the world's standard for time; voice processing devices; RF solutions; discrete components; enterprise storage and communication solutions; security technologies and scalable anti-tamper products; Ethernet solutions; Power-over-Ethernet ICs and midspans; as well as custom design capabilities and services. Microsemi is headquartered in Aliso Viejo, California, and has approximately 4,800 employees globally. Learn more at www.microsemi.com.