# Considerations for multi-lane implementations of higher rate Ethernet Interfaces

IEEE 802.3 High Speed Study Group

September 18, 2006

Steve Trowbridge
Lucent Technologies
+1 303 920 6545
sjtrowbridge@lucent.com

Lucent Technologies
Bell Labs Innovations

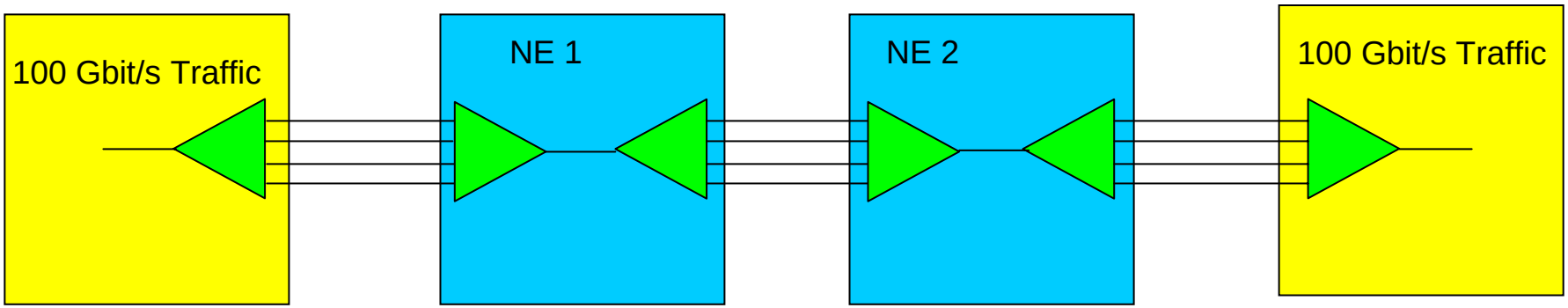
# Issues to be considered for Multi-Lane implementations of High Rate Interfaces

- What is the maximum Distance and maximum permissible Differential Delay between Endpoints?
- Single or Multi-Hop (at layer 1) for transport of individual lanes?
- Are individual lanes carried over New or Existing physical interfaces?
- Is the number of lanes (and overall bandwidth) fixed or variable? Over what range?
- Can the interface operate at reduced bandwidth with failure of one or more lanes?
- What is the basis for inverse multiplexing (packet vs. bit/byte/column slicing)?
- What existing technologies can be reused or emulated?
- Is a flexible multi-lane specification a good candidate to address with a different PAR than specification of the next serial interface rate?

# Two Fundamental Approaches to Multi-Lane Transport

| Short reach single span (low latency LAN environment, no FEC) | MAN/WAN environment (longer reach, higher latency, FEC) |
|---|---|
| e.g., 10G Base-CX4 | e.g., 10G LAG, ITU-T Virtual Concatenation |
| ▪ Physical interface may be unique (used for no purpose except the particular transport of this lane – 10G BASE-CX4 uses 4 lanes of 3.125 Gigabaud/s).<br>▪ Point to point lanes (can't be carried separately over a server layer network – requires demux/remux for every segment of a long reach connection)<br>▪ Limited (sub-microsecond) deskew between lanes<br>▪ May not be able to operate with reduced bandwidth in the presence of individual lane failures | ▪ Lanes use existing physical interfaces at lower bitrates<br>▪ Lanes can be carried independently across a sever layer network and not reaggregated until the far end<br>▪ Inverse multiplexing methodologies:<br>– Packet based - flow based distribution is needed to maintain LAN invariants.<br>– Bit/byte/column slicing, longer differential delay compensation required than with point-to-point multi-lane approaches<br>▪ Variable number of lanes and total bandwidth<br>▪ Add or remove lanes/bandwidth in-service<br>▪ Operate at reduced bandwidth in the presence of lane failures. |

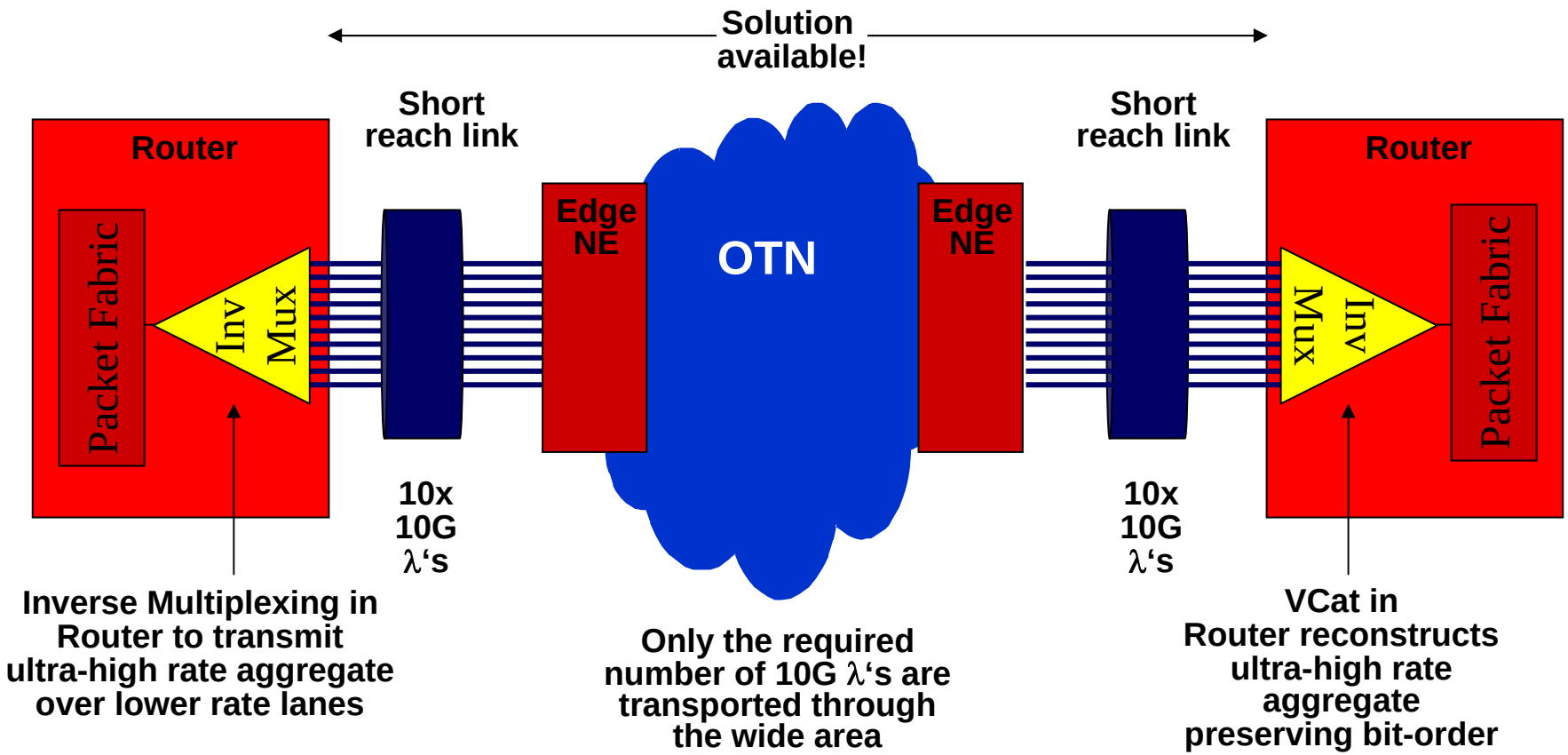
# Point to Point Multi-lane Approach e.g., 4x25 Gbit/s




- Simple, low latency (sub-microsecond) deskew for lanes over fiber ribbon cable or multiple wavelengths (CWDM?) over a common fiber
- Need to implement the new higher rate interface at every network element along the path of the ultra-high rate service to deploy the service

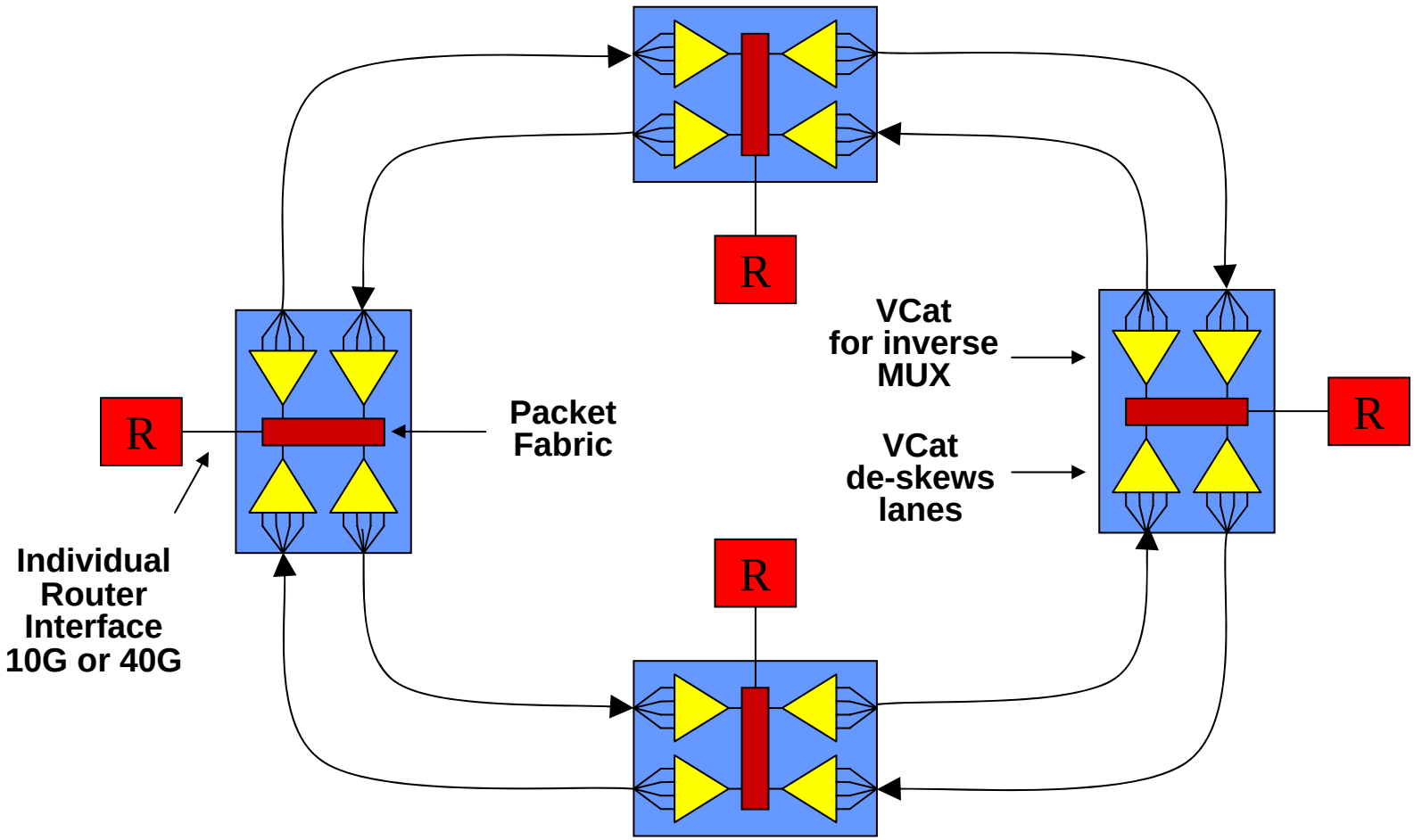
# Virtual Concatenation Example

Use existing lower rate interfaces for lanes. New capability only needed at endpoints.



# Virtual Concatenation Example II

## Ultra High Rate Service Without Ultra High Rate Interfaces!



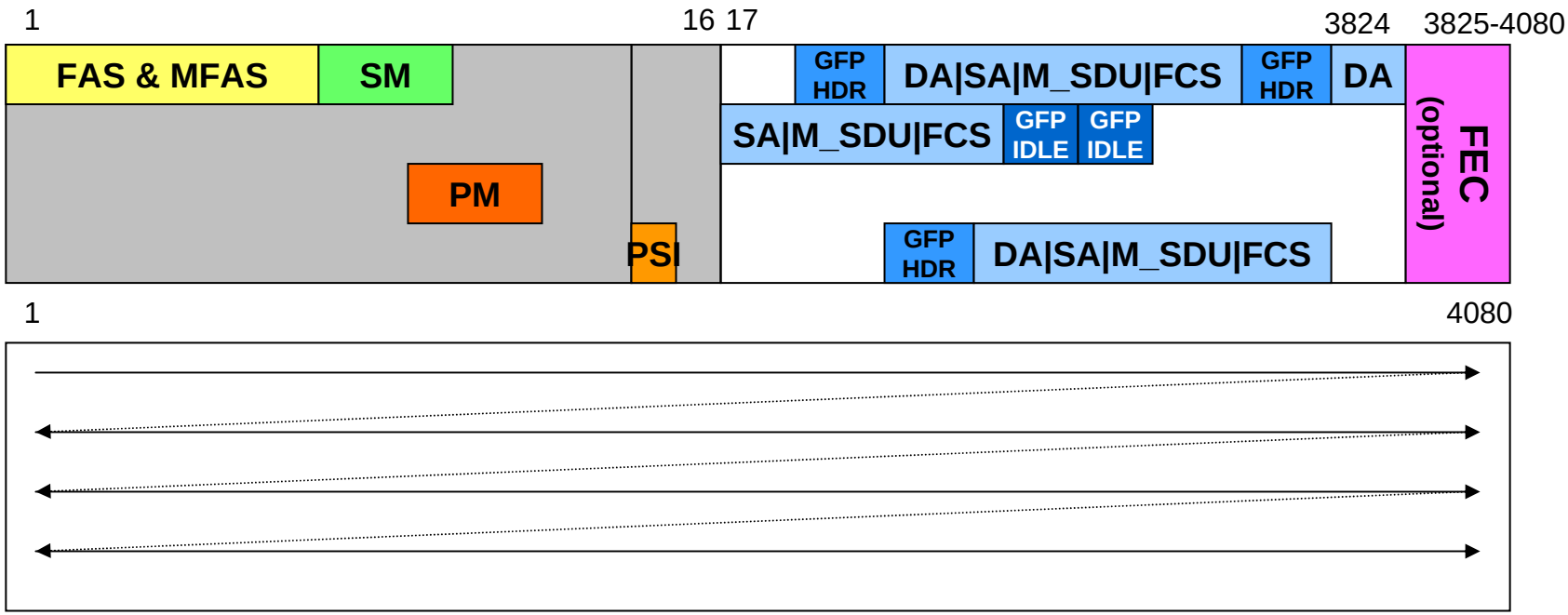
# Advantages of Virtual Concatenation (or Virtual Concatenation-like mechanism) as the Multi-Lane Mechanism for High Rates

- Existing Specifications – can reuse or emulate deployed and proven technology from transport networks
- Not constrained to a single, arbitrary bitrate (e.g., 100G, 160G) for services
- “Pay as you Grow” - can add capacity in cost-effective 10G increments without having to cost-justify “next” 100G
- Decouple the interface rate from the service rate. Can deploy high-rate services now based on cost-effective, 10G interfaces, and deploy higher rate interfaces when they become cost effective.
- Networkable components - high rate pipes can be provide with only new functionality at the endpoints using only EXISTING functionality at intermediate nodes.
- Inherent reliability - operate at decreased bitrate with individual lane failures rather than losing the entire pipe

# Inverse Multiplexing – Ethernet LAG vs. VCAT

| **Ethernet LAG**<br>Distribute traffic to lanes at packet level | **Virtual Concatenation**<br>Distribute traffic to lanes by byte/column slicing |
|---|---|
| ▪ Each lane is an independent MAC<br>▪ Need to choose same lane for every packet in a flow to maintain ordering, but once this is done, latency differences between lanes are irrelevant (do not need to be compensated by the receiver)<br>▪ Large aggregate flows create load balancing problems – need to try to parse individual flows from inside various tunneling mechanisms<br>▪ Individual flows may be hidden, e.g., with IPSec | ▪ Inverse multiplexing done below the packet layer in the protocol stack<br>▪ No packet awareness within an individual lane<br>▪ Receiver buffers traffic to align with highest latency lane<br>▪ Remultiplexing reconstitutes the original bitstream from which the packets can be extracted<br>▪ Never a load balancing problem – N bit/second data stream is distributed evenly into X bitstreams of N/X bits/second. |

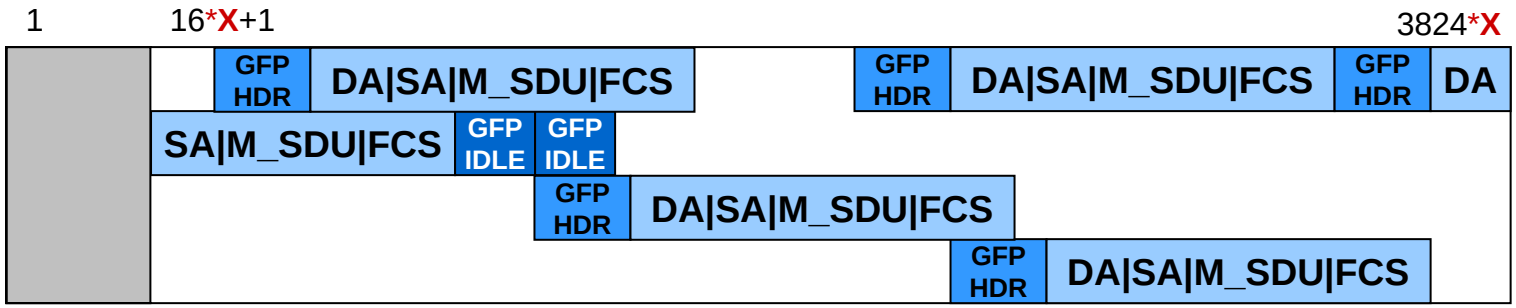
# OTUk Frame Format and Transmission Order




**OTN is not SONET!**

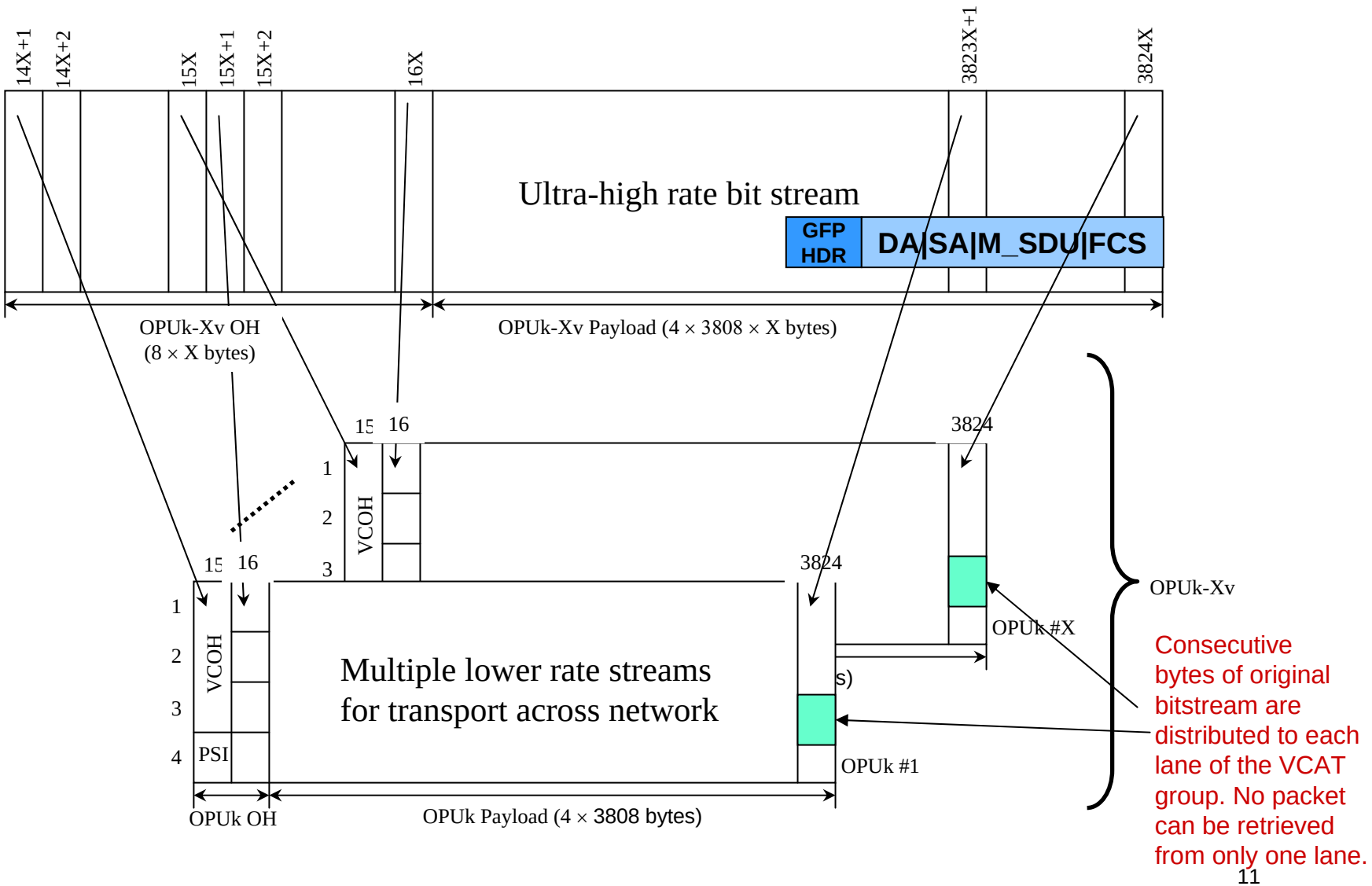
- Not optimized for voice (not based on 125μs frame)
- Not synchronous (no pointers, no clock traceability (S bytes))
- Provides a lightweight, bit rate independent “wrapper” to provide:
  - Section and path monitoring
  - Fixed length codeword identification for FEC computation
  - Byte position identification for multiplexing, inverse multiplexing

# For High Rate Services, Extend the Frame Size but maintain the frame repetition rate of a single lane




For example, the frame repetition rate of an OTU2/ODU2/OPU2 is 12.191 μs, with 4×3808 payload bytes transmitted per frame (9 995 276.962 kbit/s). To transmit approximately 100 Gbit/s, define a (logical) frame of 4×3808×10 payload bytes to be transmitted every 12.191 μs (99.995 276 962 Gbit/s)

# OPU Virtual Concatenation (G.709) - How does it work?



# OPUk Virtual Concatenation Attributes

- Can concatenate from 1-256 OPUk signals where k=1 (2.5G); k=2 (10G); or k=3(40G) providing bitrates from 2.5G to 10 Terabits
- Pipe size can be hitlessly increased or decreased in service in steps of the chosen base bitrate (2.5G, 10G, 40G)
- Individual VCAT signals (e.g., 2.5G or 10G) can be TDM multiplexed into higher rate signals (e.g., 10G or 40G)
- Automatic bandwidth decrease of payload pipe with failure of base bitrate signals rather than losing entire payload.
- Differential delay of up to 16 777 216 ODUk frames can be detected (+/-410.7s for k=1, +/-102.2s for k=2, +/-25.4s for k=3). Standard requires buffering to accommodate at least 125 µs of differential delay, although specific equipment may accommodate more.

# Applicability of Virtual Concatenation

- Virtual Concatenation seems to be a compelling technology for carrying ultra-high bitrate signals over wider area networks
- Is virtual concatenation too heavyweight as a multi-lane approach for very short reach (e.g., router to router) enterprise applications?
  - Appealing to use one approach everywhere, but cost effectiveness should be investigated
  - Perhaps some “adjustments” (e.g., reduced buffering requirements for VSR applications where larger differential delay is not an issue) could make VCAT cost effective for even VSR applications rather than inventing a different multi-lane approach
  - Are other aspects of OTN framing and overhead (e.g., FEC, certain OA&M) that are useful even for VSR applications?
  - If a different multi-lane approach from VCAT is needed for VSR applications, where is the boundary (reach, domain, etc.) between applicability of the two approaches? What about interworking between the approaches once there is a need to carry a deployed enterprise technology over networks? (Don’t repeat 10G LAN PHY/WAN PHY problem - plan how to handle this from the start!)

# Backup and References



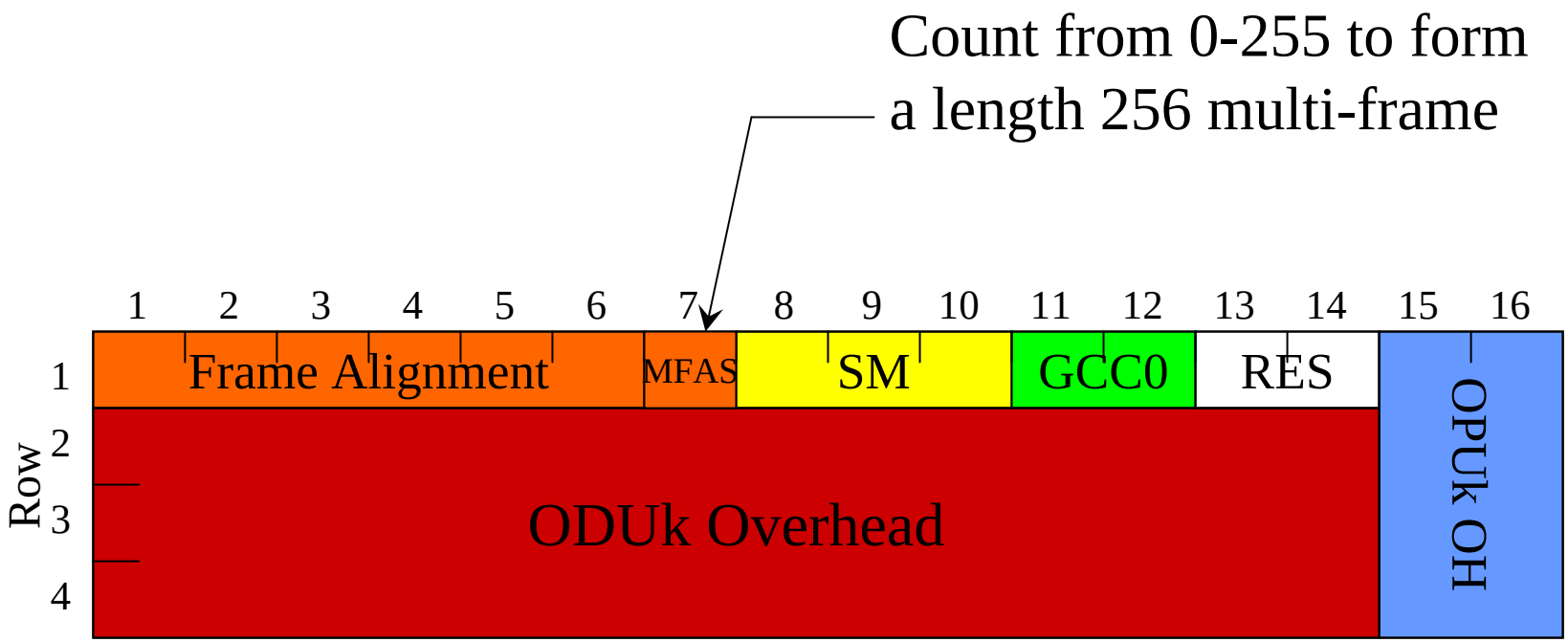


# References

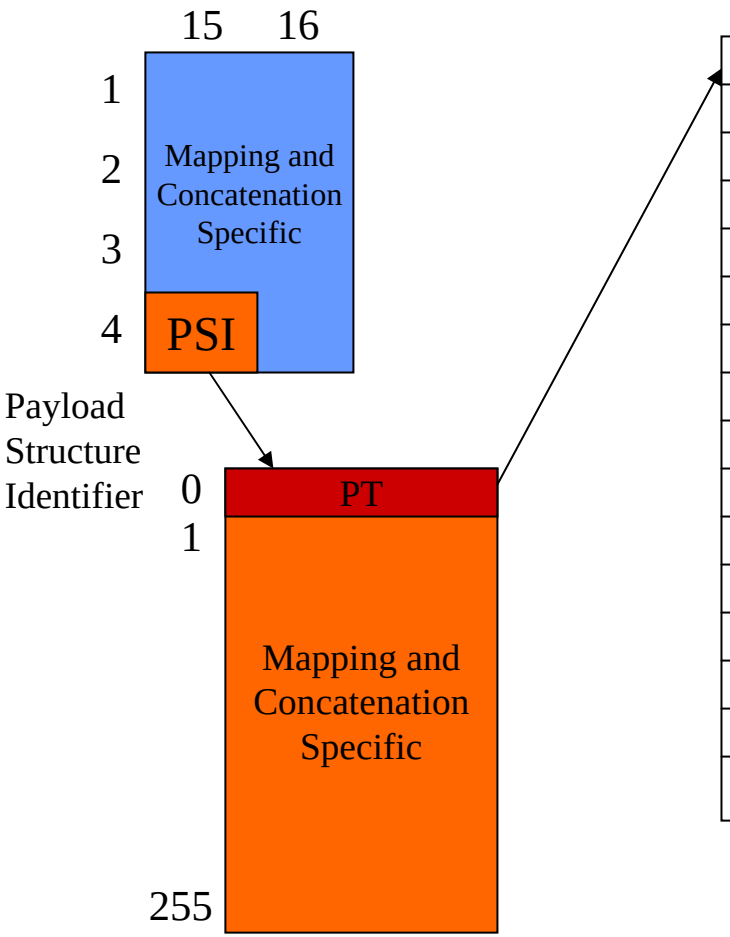
**For full details, refer to:**


1. ITU-T Recommendation G.709, “Interfaces for the optical transport network (OTN)”, March 2003, plus Amendment 1 (December 2003).
2. ITU-T Recommendation G.7042, “Link capacity adjustment scheme (LCAS) for virtual concatenated signals”, February 2004 plus Amendment 1 (February 2005), Amendment 2 (August 2005), Corrigendum 1 (August 2004).
3. ITU-T Recommendation G.7041, “Generic framing procedure (GFP)”, August 2005, plus Amendment 1 (March 2006).

# OTUk Overhead



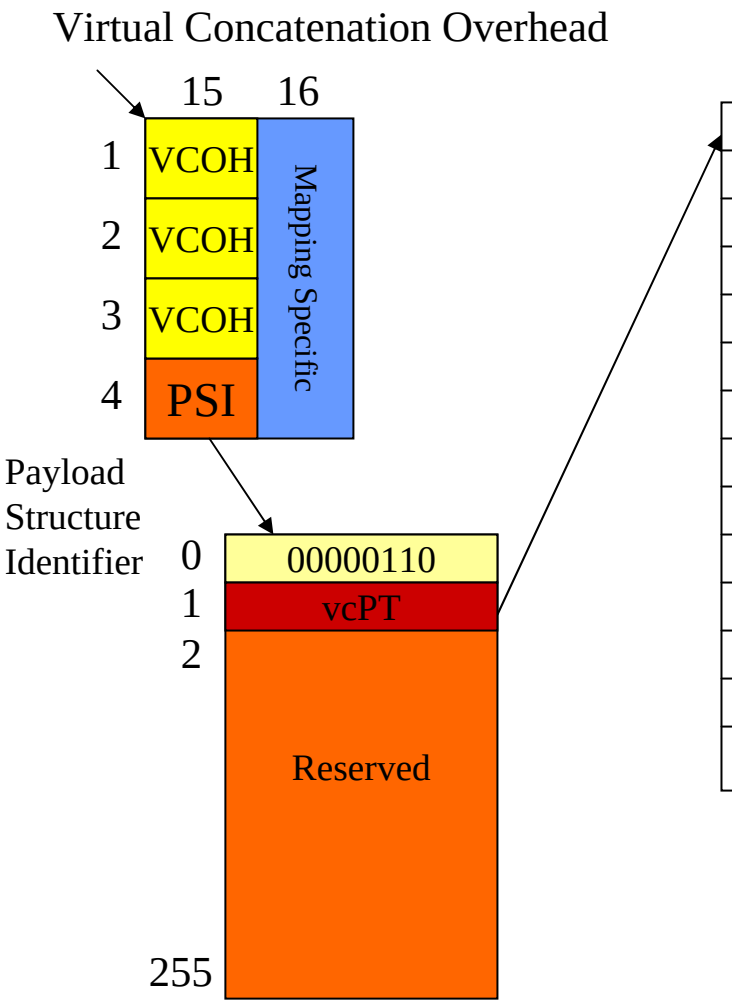
# OPUk Overhead




## Payload Type Code Points

| | |
|---|---|
| 00000001 | Experimental mapping |
| 00000010 | Asynchronous CBR mapping |
| 00000011 | Bit synchronous CBR mapping |
| 00000100 | ATM mapping |
| 00000101 | GFP mapping |
| 00000110 | Virtual Concatenated signal |
| 00001000 | Bit stream with octet timing mapping |
| 00001001 | Bit stream without octet timing mapping |
| 00100000 | ODU multiplex structure |
| 01010101 | Not Available (ODU maintenance signal) |
| 01100110 | Not Available (ODU maintenance signal) |
| 1000xxxx | Reserved codes for proprietary use |
| 11111101 | NULL test signal mapping |
| 11111110 | PRBS test signal mapping |
| 11111111 | Not Available (ODU maintenance signal) |
| 226 Spare Codes left for future international standardization | |

# OPUk-Xv Overhead

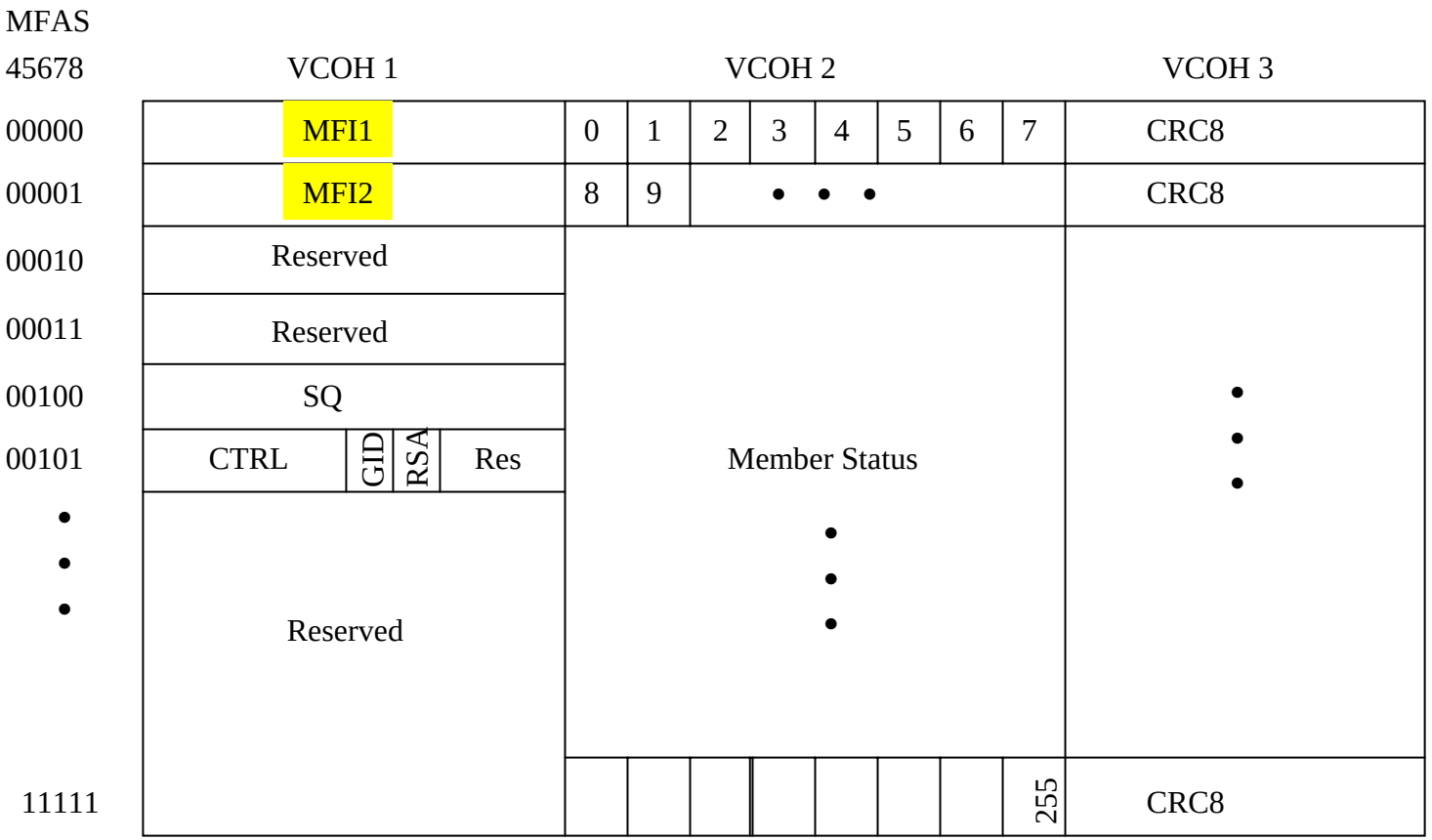



Virtually Concatenated Payload Type Code Points

| | |
|---|---|
| 00000001 | Experimental mapping |
| 00000010 | Asynchronous CBR mapping |
| 00000011 | Bit synchronous CBR mapping |
| 00000100 | ATM mapping |
| 00000101 | GFP mapping |
| 00001000 | Bit stream with octet timing mapping |
| 00001001 | Bit stream without octet timing mapping |
| 01010101 | Not Available (ODU maintenance signal) |
| 01100110 | Not Available (ODU maintenance signal) |
| 1000xxxx | Reserved codes for proprietary use |
| 11111101 | NULL test signal mapping |
| 11111110 | PRBS test signal mapping |
| 11111111 | Not Available (ODU maintenance signal) |
| 228 Spare Codes left for future international standardization | |

PSI is identical in all X OPUk frames

# Virtual Concatenation Overhead (VCOH)




MFI1 + MFI2 + MFAS provides a 24 bit (16 777 216 ODUk frame) sequence to align incoming frames
ITU-T G.709 requires minimum 125 µs of differential delay to be accommodated

# LCAS - Example of Failure and Restoration of a Member

| Step | Description | Member 0 | | | Member 1 | | | Member 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | CTRL | SQ | MST | CTRL | SQ | MST | CTRL | SQ | MST |
| 1 | Initial Condition | NORM | 0 | OK | NORM | 1 | OK | EOS | 2 | OK |
| 2 | Member 1 Failure detected at sink | NORM | 0 | OK | NORM | 1 | FAIL | EOS | 2 | OK |
| 3 | Source omits Member 1 from group | NORM | 0 | OK | DNU | 1 | FAIL | EOS | 2 | OK |
| | • • • | | | | | | | | | |
| 4 | Member 1 Restoration detected at sink | NORM | 0 | OK | DNU | 1 | OK | EOS | 2 | OK |
| 5 | Source resumes use of Member 1<br>Sink uses first frame tagged NORM | NORM | 0 | OK | NORM | 1 | OK | EOS | 2 | OK |

# LCAS - Example of addition of a Member

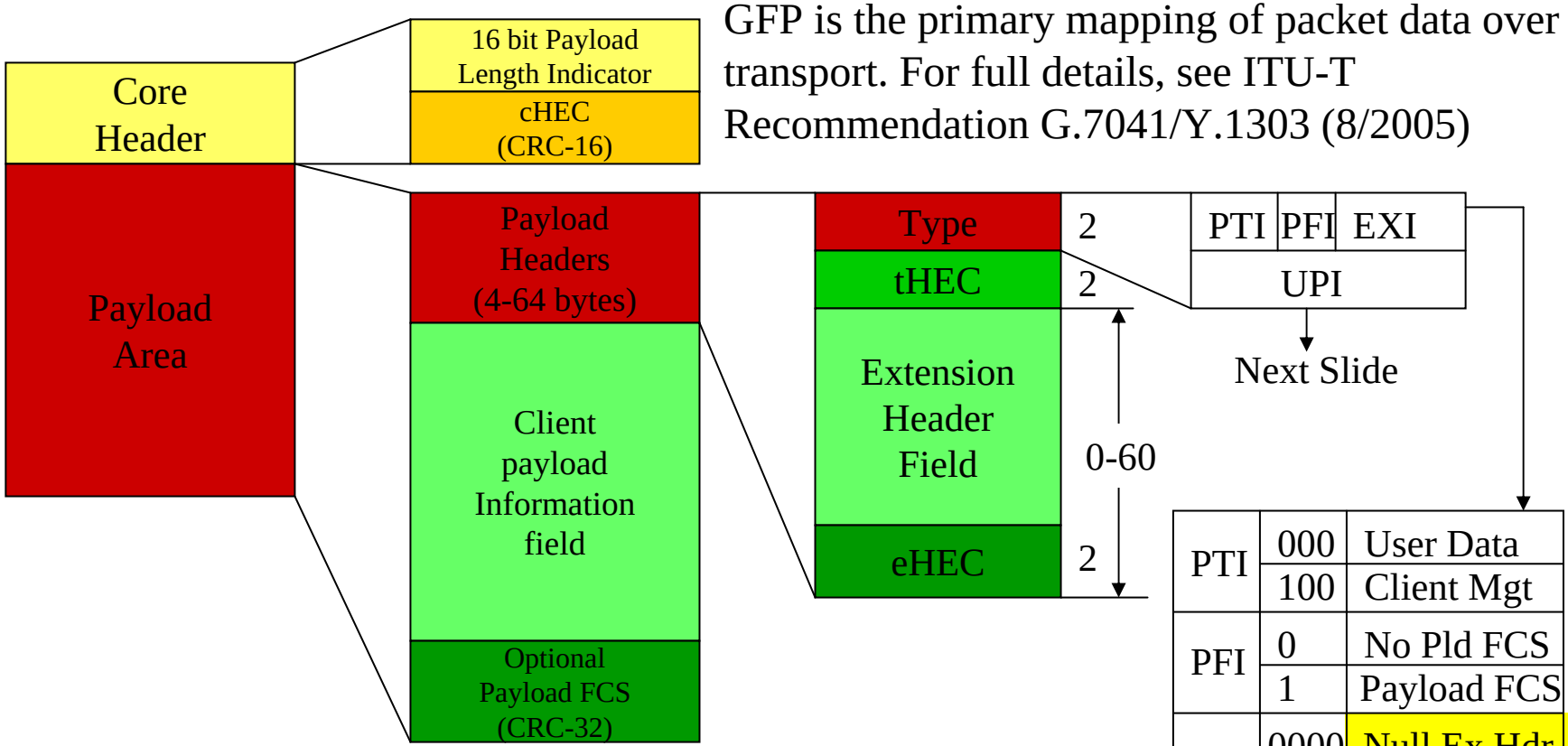
| Step | Description | Member 0 | | | Member 1 | | | Member 2 (New) | | | RSA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CTRL | SQ | MST | CTRL | SQ | MST | CTRL | SQ | MST | |
| 1 | Initial Condition | NORM | 0 | OK | EOS | 1 | OK | IDLE | FF | FAIL | 0 |
| 2 | Add NMS to So and Sk LCASC | NORM | 0 | OK | EOS | 1 | OK | IDLE | FF | FAIL | 0 |
| 3 | So (2) sends CTRL = ADD and SQ = 2 | NORM | 0 | OK | EOS | 1 | OK | ADD | 2 | FAIL | 0 |
| 4 | Sk (2) sends MST = OK to So | NORM | 0 | OK | EOS | 1 | OK | ADD | 2 | OK | 0 |
| 5 | So (1) sends CTRL = NORM;<br>So (2) sends CTRL = EOS and SQ = 2 | NORM | 0 | OK | NORM | 1 | OK | EOS | 2 | OK | 0 |
| 6 | RS-Ack bit inverted due to change in sequence | NORM | 0 | OK | NORM | 1 | OK | EOS | 2 | OK | 1 |

# LCAS - Example of deletion of a Member

| Step | Description | Member 0 | | | Member 1 | | | Member 2 | | | RSA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CTRL | SQ | MST | CTRL | SQ | MST | CTRL | SQ | MST | |
| 1 | Initial Condition | NORM | 0 | OK | NORM | 1 | OK | EOS | 2 | OK | 0 |
| 2 | NMS - Decrease command to So LCASC | NORM | 0 | OK | NORM | 1 | OK | EOS | 2 | OK | 0 |
| 3<br>4 | So (1) sends CTRL = IDLE, SQ > 1<br>So (2) sends SQ=1 | NORM | 0 | OK | IDLE | 2 | OK | EOS | 1 | OK | 0 |
| 5 | Sk (1) (unwanted) sends MST = FAIL to So | NORM | 0 | OK | IDLE | 2 | FAIL | EOS | 1 | OK | 0 |
| 6 | RS-Ack bit inverted due to change in sequence | NORM | 0 | OK | IDLE | 2 | FAIL | EOS | 1 | OK | 1 |
| 7 | NMS issues Decrease command to Sk LCASC | NORM | 0 | OK | IDLE | 2 | FAIL | EOS | 1 | OK | 1 |

# Generic Framing Procedure (GFP)

GFP is the primary mapping of packet data over transport. For full details, see ITU-T Recommendation G.7041/Y.1303 (8/2005)




GFP Client Frame

# GFP User Payload Identifiers (UPI) - for Client Frames

| PTI=000 | | | |
|---|---|---|---|
| UPI | GFP frame payload area | UPI | GFP frame payload area |
| 0000 0000<br>1111 1111 | Reserved and not available | 0000 1100 | Transparent Synchronous Fibre Channel |
| 0000 0001 | Frame Mapped Ethernet | 0000 1101 | Frame Mapped MPLS(Unicast) |
| 0000 0010 | Frame Mapped HDLC/PPP | 0000 1110 | Frame Mapped MPLS(Multicast) |
| 0000 0011 | Transparent Fibre Channel | 0000 1111 | Frame Mapped IS-IS |
| 0000 0100 | Transparent FICON | 0001 0000 | Frame Mapped IPv4 |
| 0000 0101 | Transparent ESCON | 0001 0001 | Frame Mapped IPv6 |
| 0000 0110 | Transparent Gb Ethernet | 0001 0010<br>through<br>1110 1111 | Reserved for future standardization |
| 0000 0111 | Reserved for future | 1111 0000<br>Through<br>1111 1110 | Reserved for proprietary use |
| 0000 1000 | Transparent multiple-access protocol over SDH (MAPOS | | |
| 0000 1001 | Transparent DVB ASI | | |
| 0000 1010 | Frame mapped IEEE 802.17 Resilient Packet Ring | | |
| 0000 1011 | Frame mapped Fibre Channel FC-BBW | | |