40 GbE and 100 GbE PCS Considerations

Stephen J. Trowbridge
Alcatel-Lucent

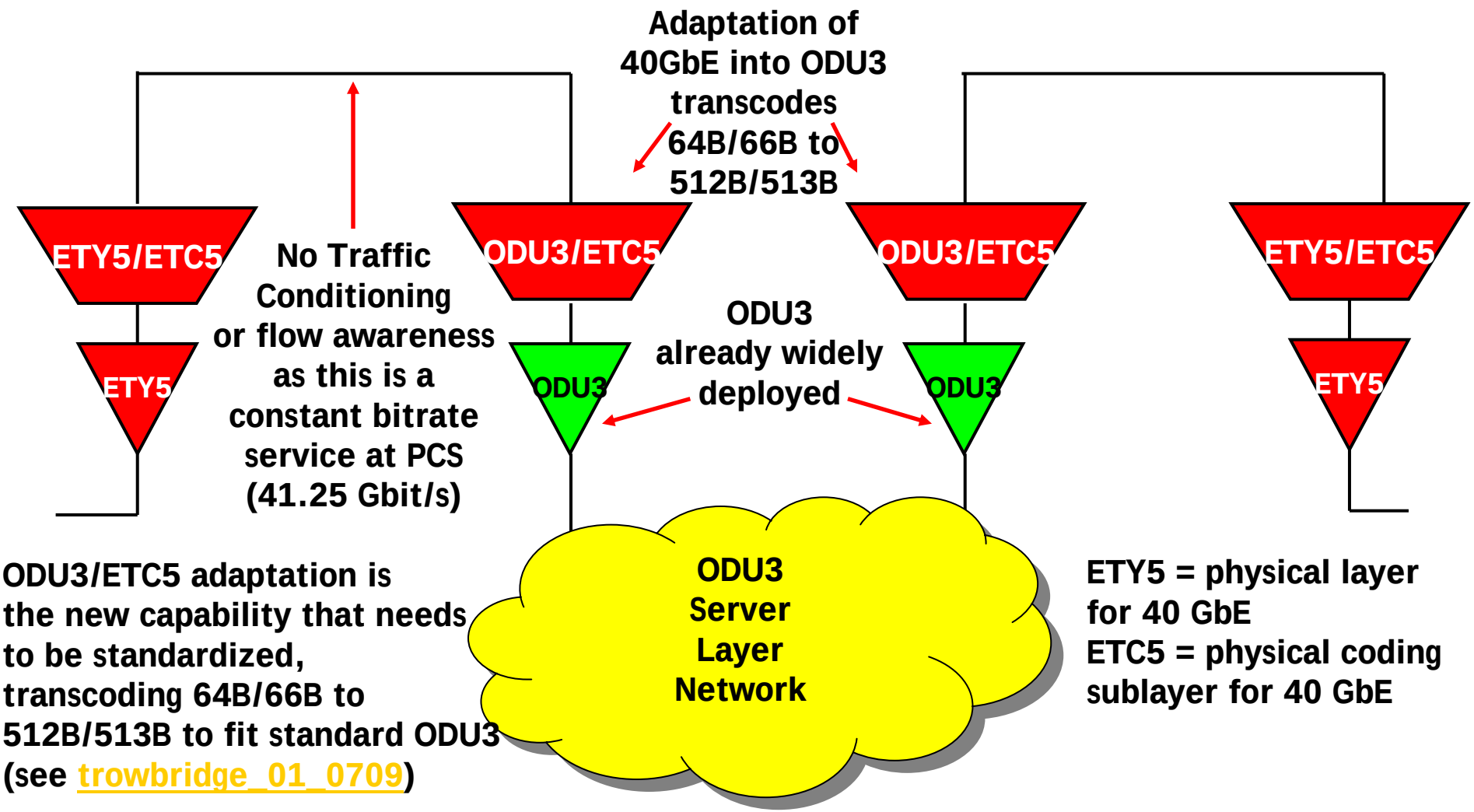
Supporters

- Gary Nicholl (Cisco)
- Pete Anslow (Nortel)
- Mark Gustlin (Cisco)
- Med Belhadj (Cortina)
- Frank Chang (Vitesse)
- Martin Carroll (Verizon)
- Ralf-Peter Braun (Deutsche Telekom)
- George Young (AT&T)
- Thomas Fischer (Nokia Siemens Networks)

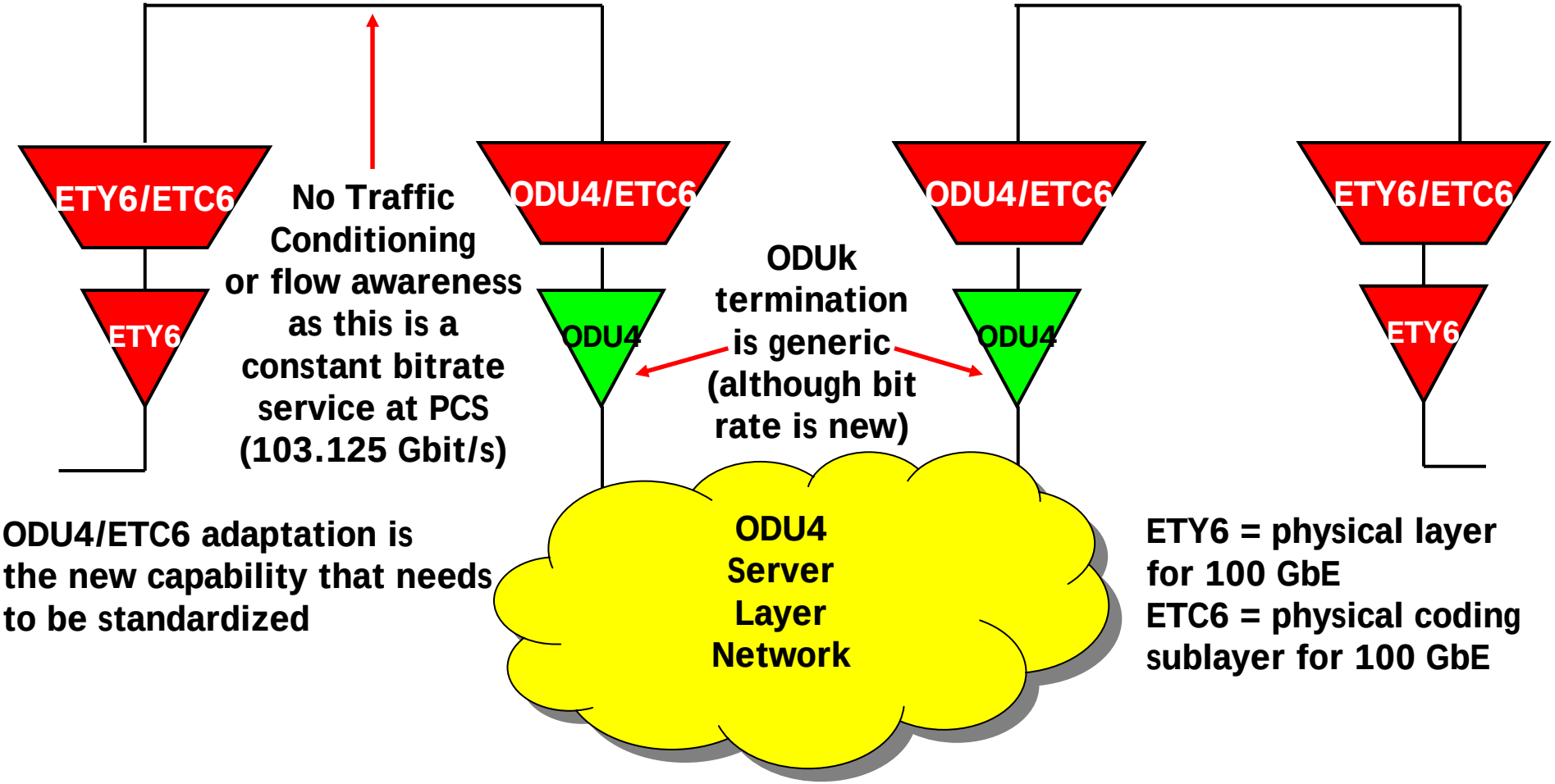
PCS Considerations

- See earlier presentation trowbridge_01_0307.pdf *Mapping of 100 GbE into OTN and the need for a lane-independent PCS*
- Desire to have a single "socket" for adaptation to a server layer network – the mapping into OTN should not depend on the Ethernet PMD or the number of lanes
- Ingress and Egress of OTN network should be free to use different Ethernet PMDs
- Need for a common understanding of the meaning of "transparent" transport of 40 GbE and 100 GbE

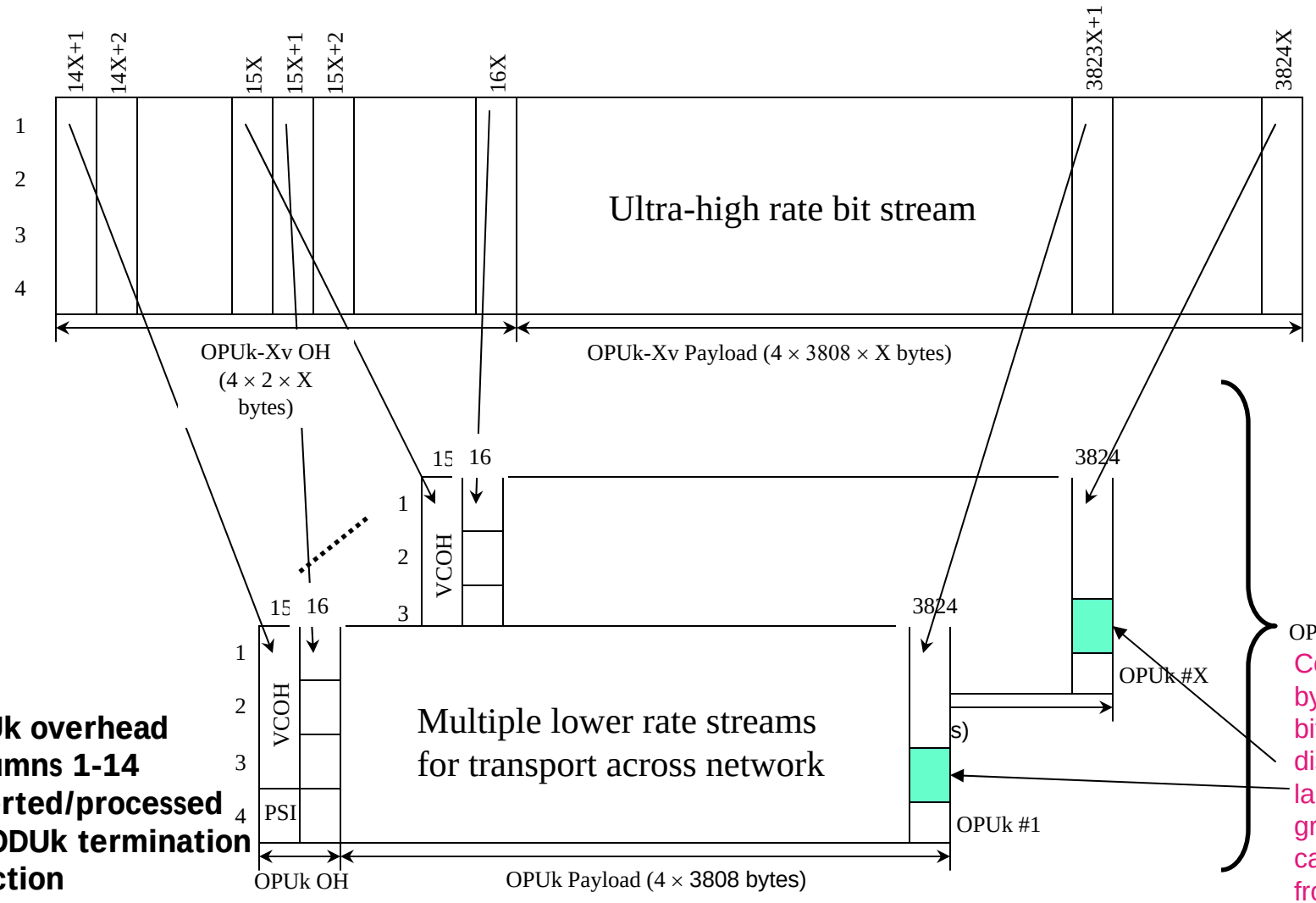
Ethernet Private Line Service – Type 2+ codeword transparent mapping of 40 GbE (ETY5) into ODU3 Server Layer Network



Ethernet Private Line Service – Type 2+
100 GbE (ETY6) into new ODU4 Server Layer Network

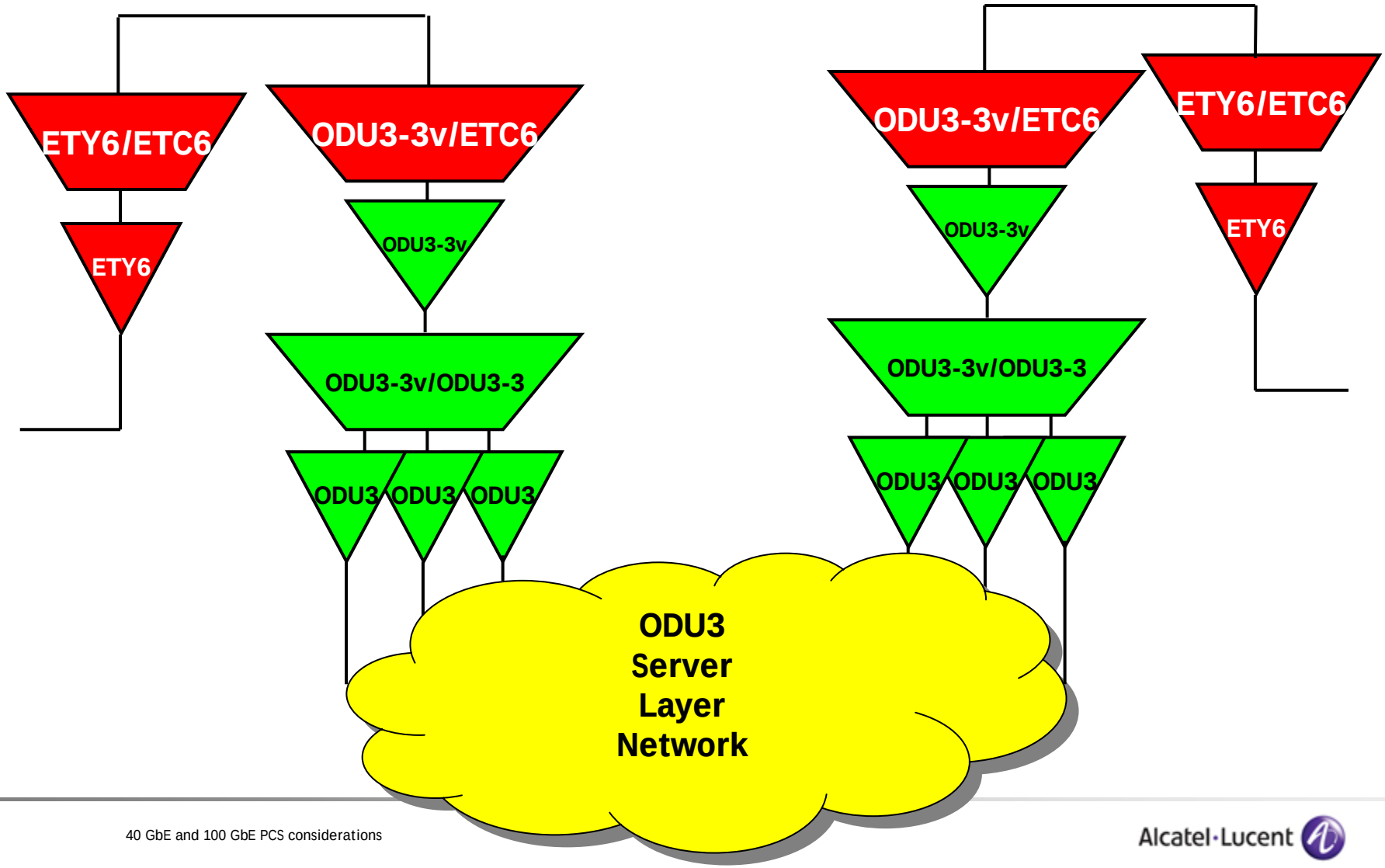


ODUk-Xv/ODUk-X-L Adaptation



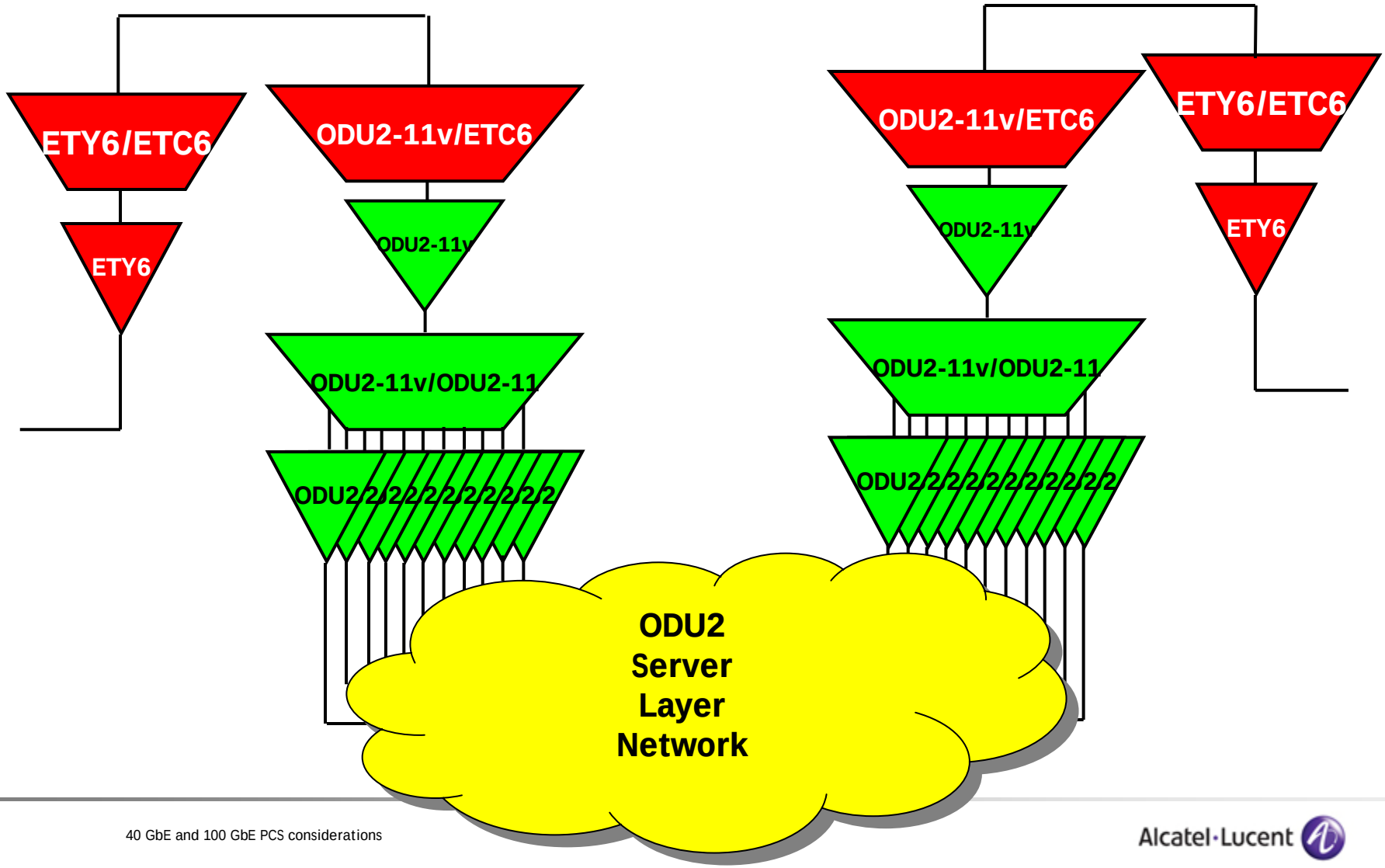
OPUk overhead columns 1-14 inserted/processed by ODUk termination function

Ethernet Private Line Service – Type 2+ using Virtually Concatenated Server ODU3-3v



Alcatel·Lucent

Ethernet Private Line Service – Type 2+ using Virtually Concatenated Server ODU2-11v



Alcatel·Lucent

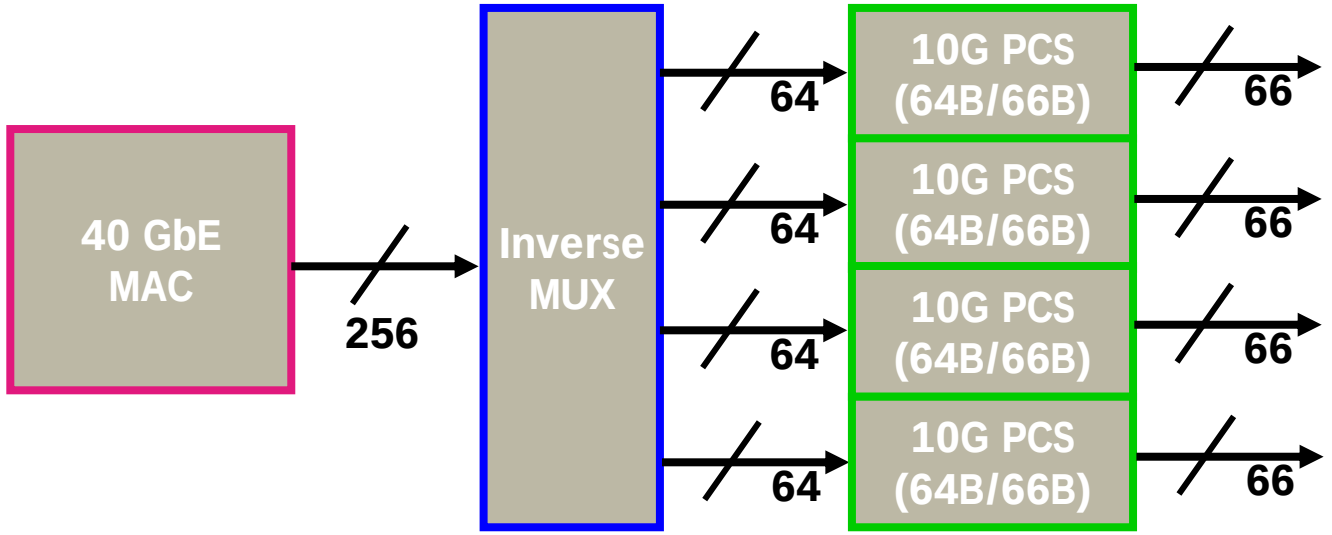
Assumptions/Assertions

- The 40 GbE and 100 GbE PCS will be 64B/66B coding similar to IEEE 802.3 clause 49
- Inverse multiplexing across lanes (or virtual lanes) of the Ethernet PMD will be to an 8-byte or 66B basis – no difference based on whether PCS calculation is done on an aggregate or per-lane (or virtual lane) basis
- The PCS to be carried over OTN is a 64B/66B encoded bitstream (or for 40 GbE, a 512B/513B transcoded equivalent of the 64B/66B encoded bitstream) that carries the equivalent of the CGMII (or XLGMII)

Some questions

- Does it matter whether the PCS is calculated per interface or per lane?
  - If the inverse multiplexing happens in units of 8 (or 8n) characters, it makes no difference
  - Inverse multiplexing at a different multiplexing granularity would add unnecessary complexity (e.g., if consecutive characters within a particular 66B codeword are not consecutive characters within the MAC)
- Will lane alignment markers be transported over a serial LAN interface in the future?
- Should lane alignment markers be transported over a server layer such as OTN?
- Should lane alignment markers be inserted inband (steal bandwidth from the MAC by deleting from IPG) or out of band (increase the lane bitrate)?

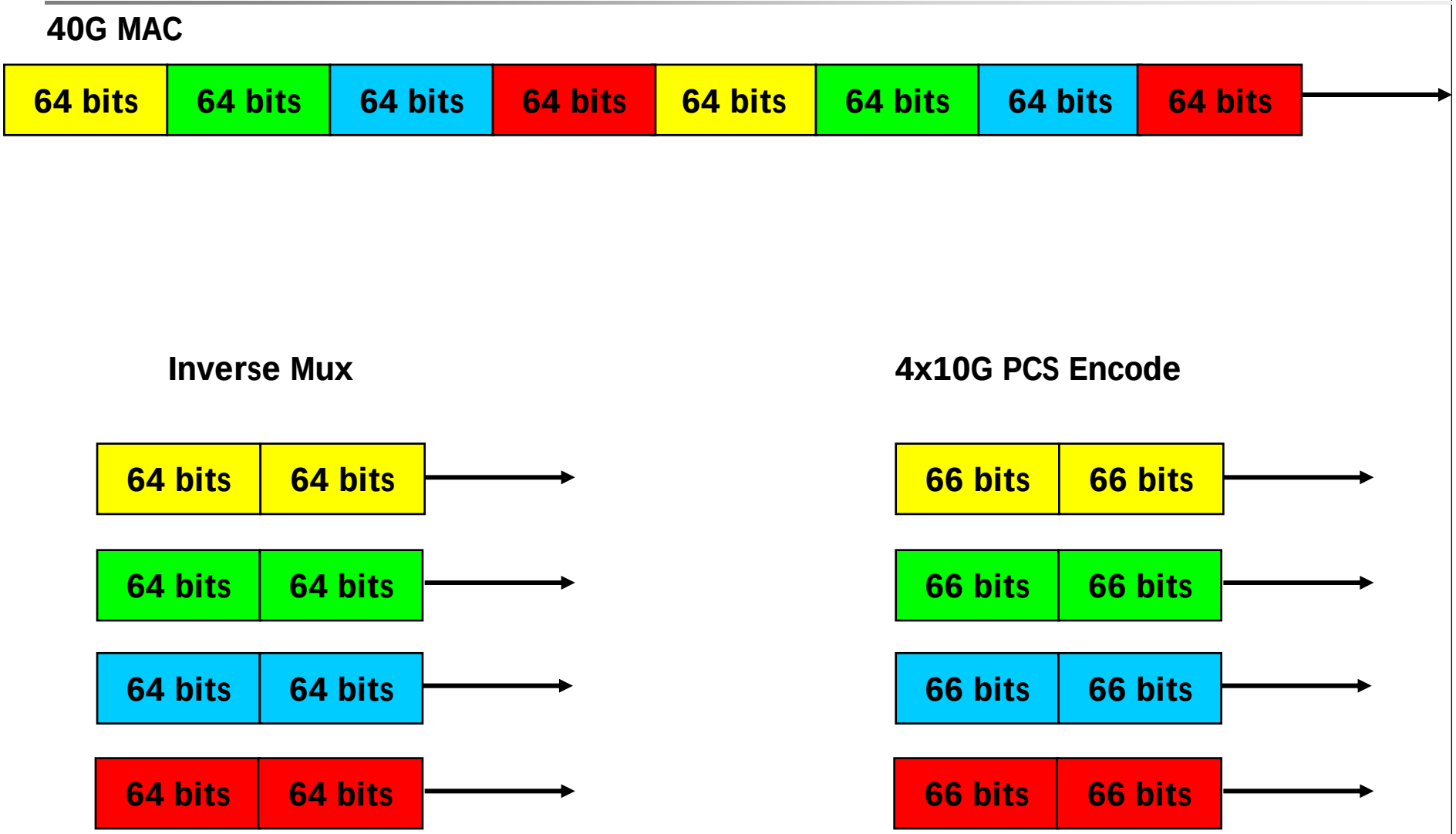
Inverse multiplexing granularity
Per-lane computation of 64B/66B



- 40 GbE MAC with internal 256-bit wide bus
- 4-lane inverse mux
  - 256-bit input @ 40.000 Gb/s
  - 4x64-bit output @ 10.000 Gb/s
  - Distribute blocks of 64 consecutive bits across 4 lanes
- 4 independent 10G PCS blocks
  - 64-bit input @ 10.000 Gb/s
  - 66-bit output @ 10.3125 Gb/s

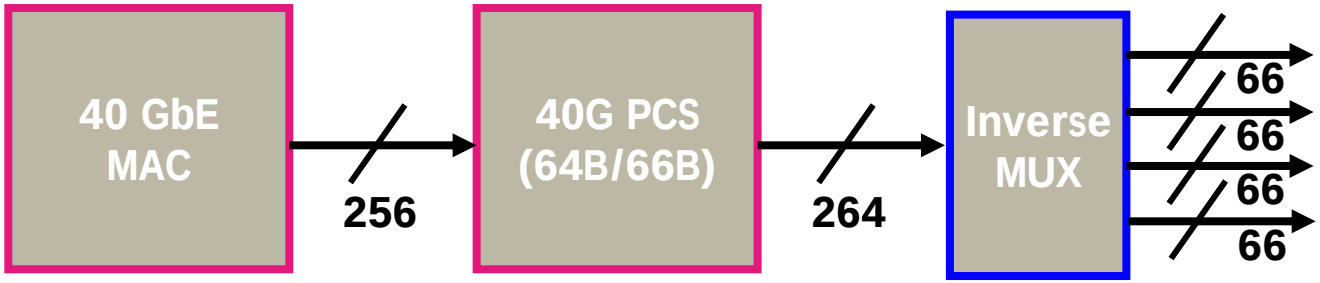
Inverse multiplexing granularity

Bitstream View – Per lane computation of 64B/66B



Alcatel·Lucent

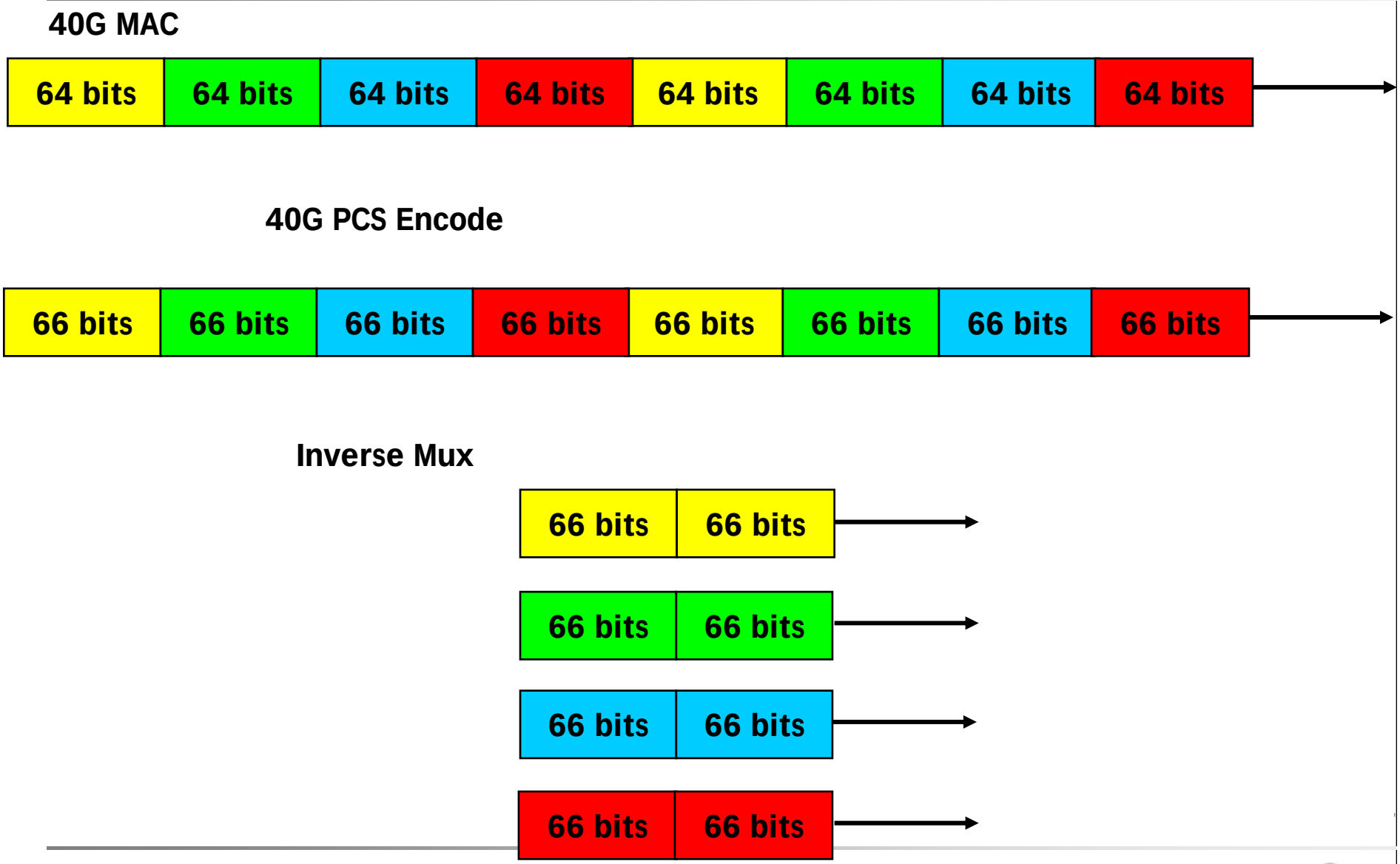
Inverse multiplexing granularity
Serial computation of 64B/66B



- 40 GbE MAC with internal 256-bit wide bus
- 64B/66B PCS block @ 40G:
  - 256-bit input @ 40.000 Gb/s
  - 264-bit output @ 41.250 Gb/s
- 4-lane inverse mux
  - 264-bit input @ 41.250 Gb/s
  - 4x66-bit output @ 10.3125 Gb/s
  - Distribute blocks of 66 consecutive bits across 4 lanes

Inverse multiplexing granularity

Bitstream View – Serial computation of 64B/66B

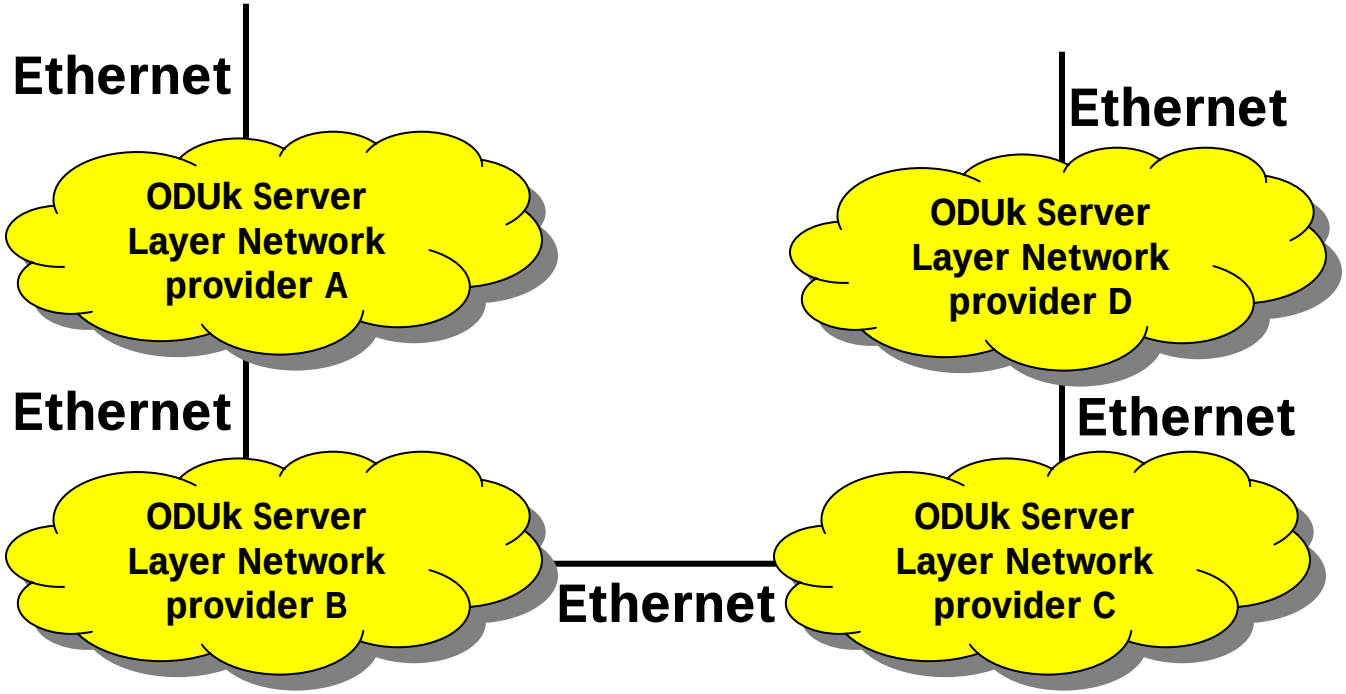


Inverse Multiplexing Granularity
Impact on multi-lane striping

- 10G Base-X interfaces do inverse multiplexing character by character across the lanes (10B codeword by 10B codeword).
- For 10 GbE, always know that you can have four consecutive bytes you can “steal” from the IPG for lane alignment marking
- With 66B codeword granularity, and four lanes (4x10 for 40 GbE or 4x25 for 100 GbE), the markers for alignment are separated by a minimum of 24 bytes from the first lane to the last (relative position within the XLGMII or CGMII).
- With 66B codeword granularity and 10 lanes (10x10 for 100 GbE), the markers for synchronization are separated by a minimum of 72 bytes from the first lane to the last (relative position within the CGMII)
- With 66B codeword granularity and 20 virtual lanes (20x5 for 100 GbE), the markers for synchronization are separated by a minimum of 152 bytes from the first lane to the last (relative position within the CGMII)
- No guarantee that IPGs will be large enough to accommodate lane alignment markers if inverse multiplexing is done on a 66B block basis, but an additional set of problems if inverse multiplexing is done on a byte basis
- Need to choose a lane synchronization marker that can either interrupt a packet, or can be transmitted out of band (Example: Cisco proposed CTBI/virtual lane approach would advocate the former)

Can you wait to deskew until after OTN network?

Extreme example – client layer handoffs across multiple server networks



- **Accumulated skew across multiple Ethernet links (minimum 2) likely exceeds deskew capability of the LAN**
- **Almost surely need to deskew lanes of the LAN at the OTN ingress**
- **The OTN will carry a serialized sequence of 66B codewords for 100 GbE, or 513B representation of 66B codewords for 40 GbE**

PCS options based on whether lane alignment markers steal from the MAC bandwidth and whether or not they are carried across the OTN

		OTN Interface Is space for lane alignment markers stolen from MAC bandwidth		
		No	Yes Are Lane alignment markers carried across OTN?	
			No	Yes
LAN interface Do lane alignment markers steal from MAC Bandwidth?	Yes	Option 1	Option 2	Option 3
	No	Option 4	unlikely	unlikely

Note: An implication of Options 2 and 3 is that (virtual) lane alignment markers will ALWAYS be present, even when a serial PMD exists for 40 GbE or 100 GbE. This may be needed anyway, if only to perform the electrical deskew across the CTBI/CFBI.

Option Summary - Where in the parallel LAN stack does the 64B/66B encoded data for OTN come from?

- Option 1: Deskew LAN virtual lanes, decode 64B/66B, reinsert into IPG to reach full MAC rate, re-encode 64B/66B (and transcode to 512B/513B for 40 GbE) to carry across OTN
- Option 2: Deskew LAN virtual lanes, remove lane alignment markers from serialized 64B/66B data, and transport rate reduced bitstream (sans lane alignment markers, -0.0061% reduction in bandwidth) over OTN
- Option 3: Deskew LAN virtual lanes, leaving lane alignment markers in place to be re-distributed across lanes of the LAN at the far end OTN egress
- Option 4: Rather than deleting from IPG to make room for lane markers, LAN lanes run at 0.0061% higher bitrate to make room for lane markers “out of band”. Deskew and remove lane markers at OTN ingress. This option will also allow for no lane markers once there is a serial PMD for 40 GbE or 100 GbE.

Example Data flow for 100 GbE, 4 lane LAN interface

Options 1, 2, 3 – Steal bandwidth for lane alignment markers from MAC

CGMII		100 Gb/s
Tx PCS	Delete from IPG to make room for lane alignment markers	99.993896484 Gb/s
	64B/66B coding	103.11870575 Gb/s
	Divide into 20 virtual lanes	20x5.155935287 Gb/s
	Add lane alignment marker to each 16383 data or control 66B blocks per lane	20x5.15625 Gb/s
Tx CTBI	Two virtual lanes per physical lane	10x10.3125 Gb/s
Tx gearbox PMA Tx PMD Rx PMD Rx gearbox PMA	Five virtual lanes per physical lane	4x25.78125 Gb/s
Rx CTBI	Two virtual lanes per physical lane	10x10.3125 Gb/s
Rx PCS	Deskew virtual lanes, remove lane alignment markers	103.11870575 Gb/s
	Decode 64B/66B	99.993896484 Gb/s
CGMII	Add to IPG to restore MAC rate	100 Gb/s

Data flow for 40 GbE, 4 lane LAN interface

Options 1, 2, 3 – Steal bandwidth for lane alignment markers from MAC

XLGMII		40 Gb/s
Tx PCS	Delete from IPG to make room for lane alignment markers	39.997558594 Gb/s
	64B/66B coding	41.247482300 Gb/s
	Divide into four virtual lanes	4x10.311870575 Gb/s
	Add lane alignment marker to each 16383 data or control 66B blocks per virtual lane	4x10.3125 Gb/s
Tx XLFBI	One virtual lane per physical lane	4x10.3125 Gb/s
Tx PMD Rx PMD	One virtual lane per physical lane	4x10.3125 Gb/s
Rx XLFBI	One virtual lane per physical lane	4x10.3125 Gb/s
Rx PCS	Deskew virtual lanes, remove lane alignment markers	41.247482300 Gb/s
	Decode 64B/66B	39.997558594 Gb/s
XLGMII	Add to IPG to restore MAC rate	40 Gb/s

Example Data flow for 100 GbE, 4 lane LAN interface with Option 4 Increase lane rate to accommodate lane alignment markers

CGMII		100 Gb/s
Tx PCS	64B/66B coding	103.125 Gb/s
	Divide into 20 virtual lanes	20x5.15625 Gb/s
	Add lane alignment marker to each 16383 data or control 66B blocks per lane	20x5.15656473 Gb/s
Tx CTBI	Two virtual lanes per physical lane	10x10.31312946 Gb/s
Tx gearbox PMA Tx PMD Rx PMD Rx gearbox PMA	Five virtual lanes per physical lane	4x25.78282366 Gb/s
Rx CTBI	Two virtual lanes per physical lane	10x10.31312946 Gb/s
Rx PCS	Deskew virtual lanes, remove lane alignment markers	103.125 Gb/s
	Decode 64B/66B	100 Gb/s

Data flow for 40 GbE, 4 lane LAN interface – Option 4

Increase lane rate to accommodate lane alignment markers

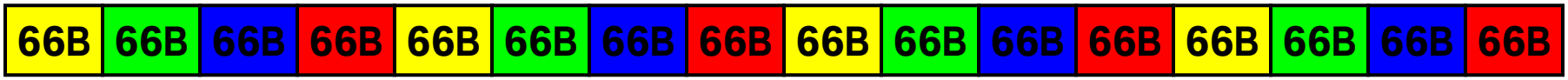
XLGMII		40 Gb/s
Tx PCS	64B/66B coding	41.25 Gb/s
	Divide into four virtual lanes	4x10.3125 Gb/s
	Add lane alignment marker to each 16383 data or control 66B blocks per virtual lane	4x10.3125 Gb/s
Tx XLFBI	One virtual lane per physical lane	4x10.31312946 Gb/s
Tx PMD Rx PMD	One virtual lane per physical lane	4x10.31312946 Gb/s
Rx XLFBI	One virtual lane per physical lane	4x10.31312946 Gb/s
Rx PCS	Deskew virtual lanes, remove lane alignment markers	41.25 Gb/s
	Decode 64B/66B	40 Gb/s

Option 1 – Pure 64B/66B of CGMII or XLGMII over OTN

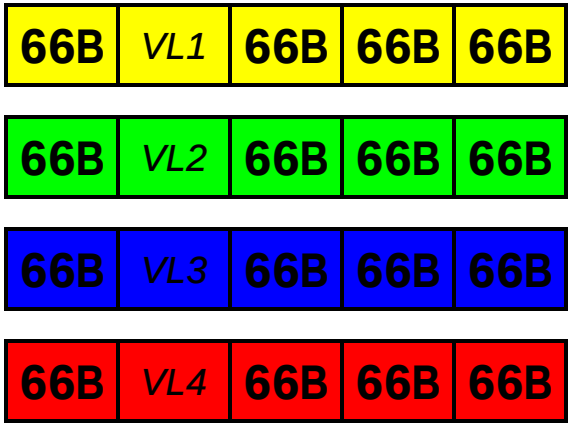
- This is the PCS that would likely be used if the only (or ANY) LAN interfaces being developed for 40 GbE and 100 GbE were serial interfaces
- Due to deletion of IPG to make room for lane alignment markers in the LAN, it is a different sequence of 66B blocks than is used in the LAN (the adaptation into OTN will reinsert the same amount of IPG, but not necessarily in the same places)
- May not be considered transparent by all customers
- Largest complexity of any of the options due to decode of 64B/66B, adjusting of IPG, and recoding of 64B/66B

Option 2 or 4 for 40 GbE - Deskew and remove lane alignment markers before mapping into OTN

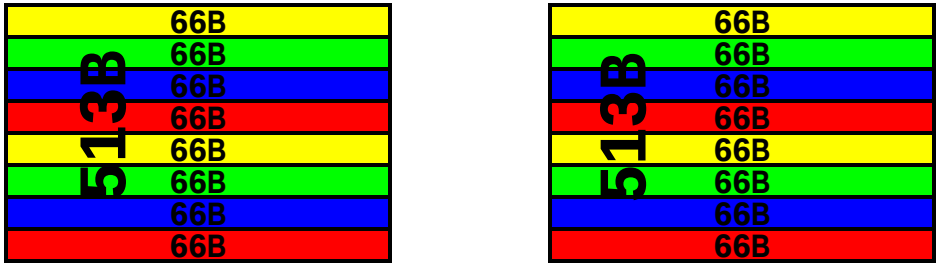
Delete from IPG, 64B/66B encode:



Inverse multiplex into lanes, add lane alignment markers:



Deskew, remove lane alignment markers, transcode to 512B/513B:

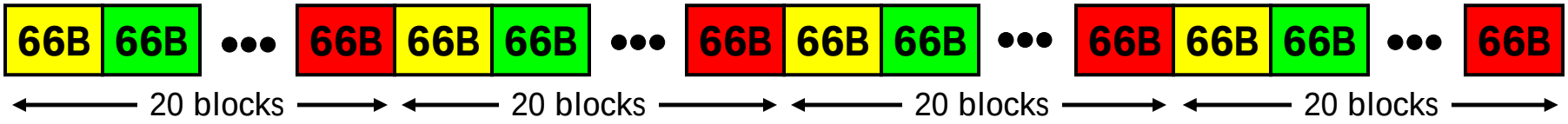


Data flow for Adaptation of 40 GbE, 4 lane LAN interface into ODU3 – Option 2 – remove lane alignment markers

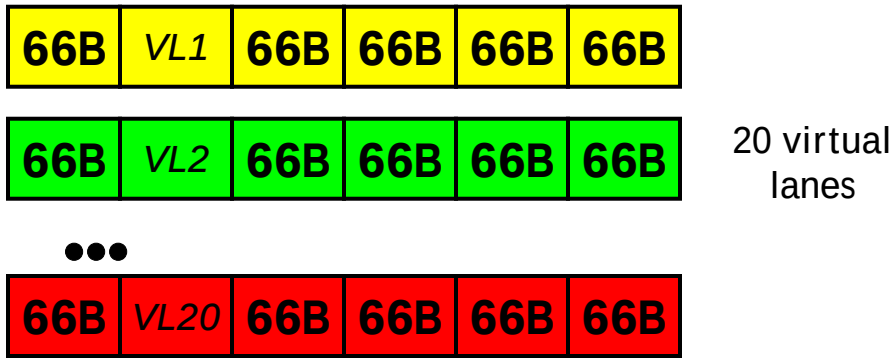
LAN PMD	One virtual lane per physical lane with lane alignment markers	4x10.3125 Gb/s
OPU3 adaptation	Deskew virtual lanes, remove lane alignment markers	41.247482300 Gb/s
	Transcode 64B/66B to 512B/513B	40.075678825 Gb/s
	Add Framing for 513B blocks	40.091302872 Gb/s
	Worst case +100ppm	40.095312002 Gb/s
OPU3 payload area	Worst case -20ppm	40.149716 Gb/s

Option 2 or 4 for 100 GbE - Deskew and remove lane alignment markers before mapping into OTN

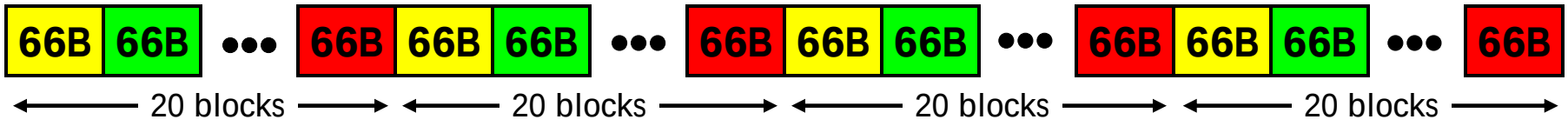
Delete from IPG, 64B/66B encode:



Inverse multiplex into 20 virtual lanes, add lane alignment markers:



Deskew, remove lane alignment markers, map into OPU4



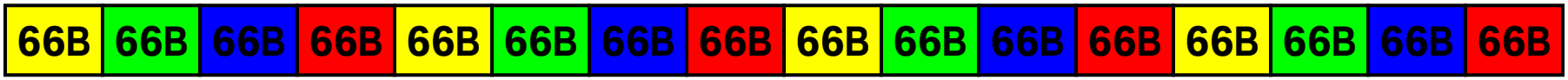
Rate reduced by 0.0061% due to shortening IPG to make room for lane alignment markers

Data flow for Adaptation of 100 GbE, 4 lane LAN interface into ODU4 – Option 2 – remove lane alignment markers

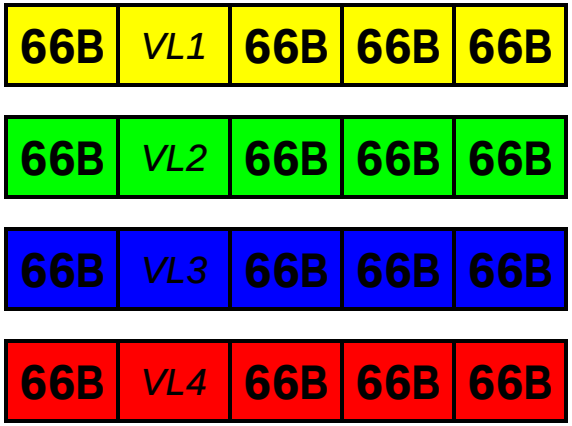
LAN PMD	Five virtual lanes per physical lane with lane alignment markers	4x25.78125 Gb/s
CTBI	Two virtual lanes per physical lane with lane alignment markers	10x10.3125 Gb/s
OPU4 adaptation	Deskew virtual lanes, remove lane alignment markers	103.11870575 Gb/s
	Worst case +100ppm	103.12901762 Gb/s
OPU4	Least possible nominal bitrate given ±20ppm clock tolerance	103.12695504 Gb/s

Option 3 for 40 GbE – Deskew and keep lane alignment markers when mapping into OTN

Delete from IPG, 64B/66B encode:



Inverse multiplex into lanes, add lane alignment markers:



Deskew, keep lane alignment markers, transcode to 512B/513B:

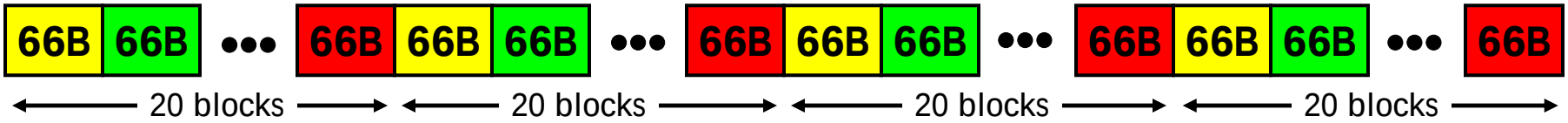
Alcatel·Lucent

Data flow for Adaptation of 40 GbE, 4 lane LAN interface into ODU3 – Option 3 – keep lane alignment markers

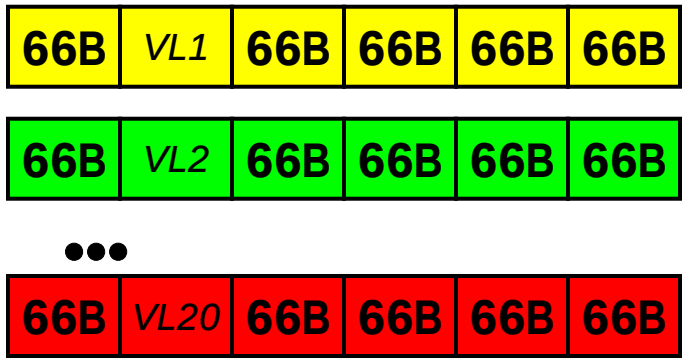
LAN PMD	One virtual lane per physical lane with lane alignment markers	4x10.3125 Gb/s
OPU3 adaptation	Deskew virtual lanes, keep lane alignment markers	41.25 Gb/s
	Transcode 64B/66B to 512B/513B	40.078125000 Gb/s
	Add Framing for 513B blocks	40.093750000 Gb/s
	Worst case +100ppm	40.097759375 Gb/s
OPU3 payload area	Worst case -20ppm	40.149716 Gb/s

Option 3 for 100 GbE – Deskew, leaving lane alignment markers in place, and map into OTN

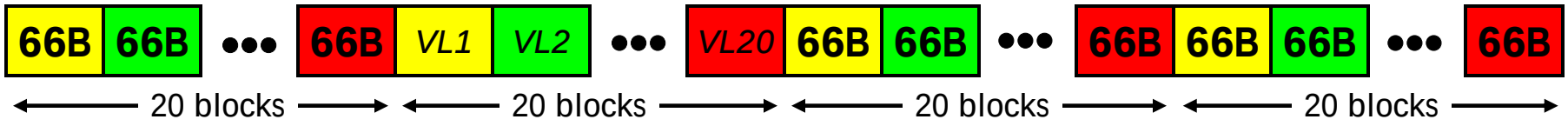
Delete from IPG, 64B/66B encode:



Inverse multiplex into 20 virtual lanes, add lane alignment markers:



Deskew, keep lane alignment markers, map into OPU4



Lane alignment markers replace 0.0061% of IPG – same PCS encoded rate

Data flow for Adaptation of 100 GbE, 4 lane LAN interface into ODU3 – Option 3 – keep lane alignment markers

LAN PMD	Five virtual lanes per physical lane with lane alignment markers	4x25.78125 Gb/s
CTBI	Two virtual lanes per physical lane with lane alignment markers	10x10.3125 Gb/s
OPU4 adaptation	Deskew virtual lanes, keep lane alignment markers	103.125 Gb/s
	Worst case +100ppm	103.13531250 Gb/s
OPU4	Least possible nominal bitrate given ±20ppm clock tolerance	103.13737525 Gb/s

Data flow for Adaptation of 100 GbE, 4 lane LAN interface into ODU4 – Option 4 – remove lane alignment markers

LAN PMD	Five virtual lanes per physical lane with lane alignment markers	4x25.78282366 Gb/s
CTBI	Two virtual lanes per physical lane with lane alignment markers	10x10.31312946 Gb/s
OPU4 adaptation	Deskew virtual lanes, remove lane alignment markers	103.125 Gb/s
	Worst case +100ppm	103.1353125 Gb/s
OPU4	Least possible nominal bitrate given ±20ppm clock tolerance	103.137375248 Gb/s

Data flow for Adaptation of 40 GbE, 4 lane LAN interface into ODU3 – Option 4 – remove lane alignment markers

LAN PMD	One virtual lane per physical lane with lane alignment markers	4x10.31312946 Gb/s
OPU3 adaptation	Deskew virtual lanes, remove lane alignment markers	41.25 Gb/s
	Transcode 64B/66B to 512B/513B	40.078125 Gb/s
	Add Framing for 513B blocks	40.09375 Gb/s
	Worst case +100ppm	40.097759375 Gb/s
OPU3 payload area	Worst case -20ppm	40.149716 Gb/s

Alignment marker format

- Should be 66 bits in length
- Should identify (virtual) lane number
- Could include other info (e.g., error control)
- Should be DC balanced and have average clock content if not scrambled

From gustlin_01_0107

10	BlockType = 0x1e	**0x4b**	**0x55**	**0x4b**	**0x55**	**0x4b**	**0x55**	**0x4b**	**VL#**

From gustlin_01_0907 (early version)

10	**Frm1**	**Frm2**	**TBD**	**TBD**	**~BIP**	**BIP**	**~VL#**	**VL#**

Frm1, Frm2 could be SONET A1, A2 (0xf628) or a new one (e.g., 0x5566)

- 2nd proposal has better DC balance
- 1st proposal stays within transcodable space of control block types, allowing for Option 3 transcoding to 512B/513B for 40 GbE keeping alignment markers
- 2nd proposal does not preserve 4-bit Hamming distance between control block types (may not be important depending on framing algorithm)
- Room for only one more control block type and still transcode to 512B/513B

Conclusions

- Inverse multiplexing across the lanes of a LAN interface should be done on an 8-byte (66B codeword) boundary
- Several viable options to establish relationship between the multi-lane LAN PCS and the 64B/66B or 512B/513B encoded bitstream to be mapped into OTN
- Easiest and most transparent options include:
  - Option 3: deskew but keep the virtual lane alignment markers in the OTN transported bitstream
    - ➢ This would require that virtual lane marking is done even when IEEE 802.3 defines a serial PMD for 40 GbE or 100 GbE;
    - ➢ but this may be necessary anyway for electrical deskew across a CTBI or XLFBI type interface
    - ➢ The alignment words need to be transcodable for 40 GbE
  - Option 4: deskew and remove lane alignment markers which are accommodated by increasing the LAN lane rate rather than stealing from the MAC
- Care should be taken in choosing the alignment word format to avoid expanding the 66B codeword space or to create difficulty for transcoding to 512B/513B in 40 GbE