IPG Considerations

Faisal Dada

Norbert Folkens

Contributions

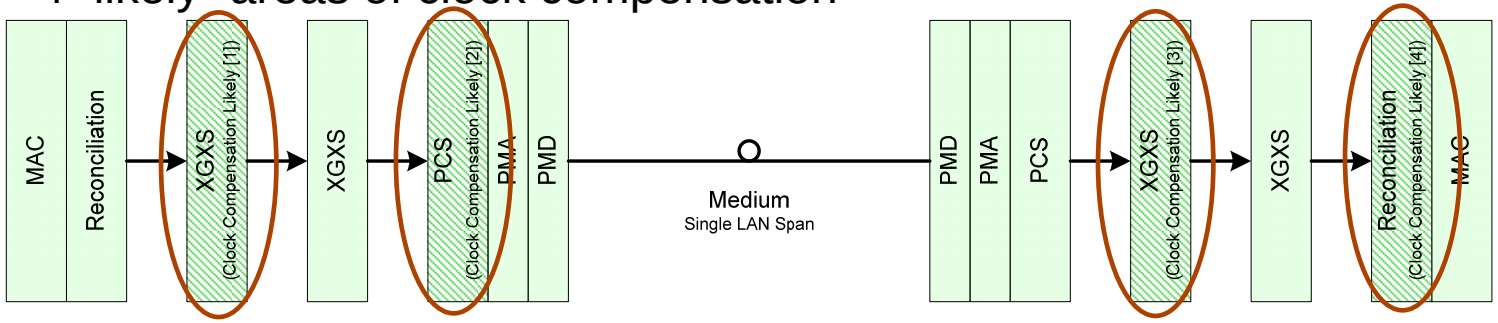
- Gary Nicholl & Mark Gustlin
- Steve Trowbridge
- Pete Anslow
- David Law
- Brad Booth
- Doug Massey

Background Information

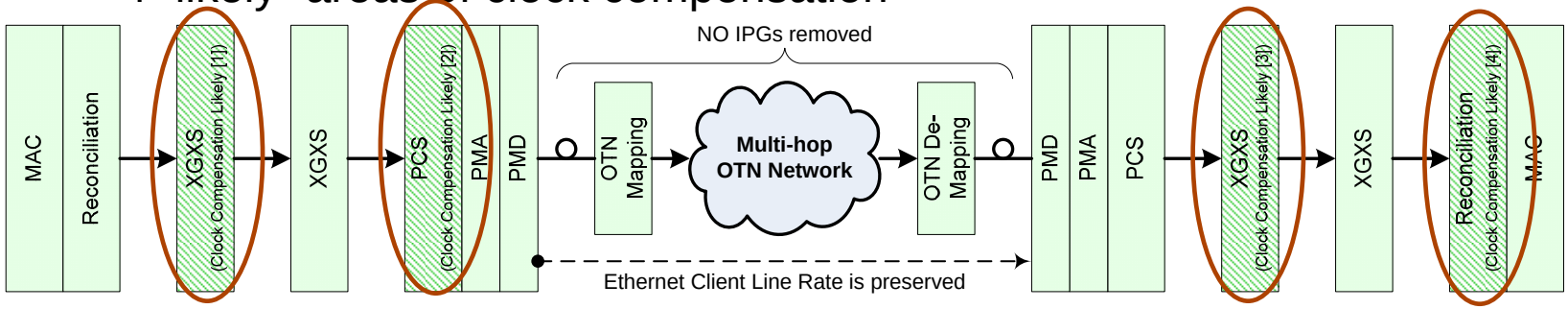
- Ethernet clocks can have frequency offsets of up to +/- 100PPM
- On transmission from a MAC a minimum number of IPGs (12 octets) are added between packets.
- In 10GBASE-R at least 5 octets of IPG must be preserved before the receiving MAC (IEEE 802.3 2005 Section 49.2.4.7: /T/ + 4/I/ are preserved)
- A Deficit Idle Counter may reduce the number of IPGs to 9 octets after the transmitting RS.
- Clock compensation is only required from a MAC to MAC path:
  - single LAN span
  - across an OTN network.

Clock Compensation Locations

- Two possibilities of a MAC to MAC path
  - Single LAN Span
    - 4 “likely” areas of clock compensation



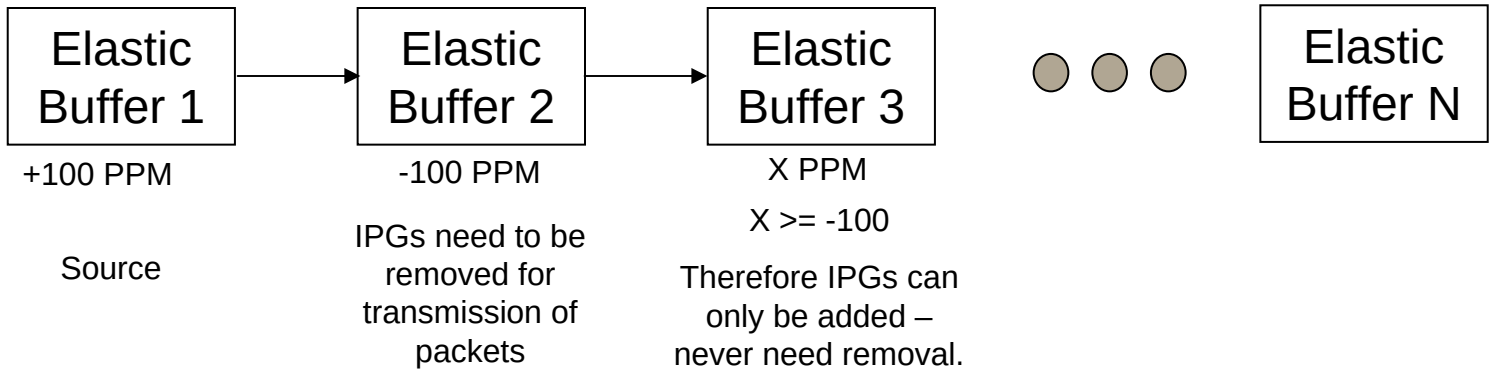
  - Across OTN Network:
    - Ethernet client rate into and out of OTN preserved
    - OTN network does not compensate by removing or adding IPGs
    - 4 “likely” areas of clock compensation



Clock Compensation – Case Study

- There are 4 likely locations for clock compensation (elastic buffers) between MACs in 10G BASE-R
- This presentation will show a case study for the amount of buffering required for 10G BASE-R
  - The case study will use N elastic buffers.
  - This presentation will provide a case study for maximum storage required using packets of X bytes.
- The presentation will extrapolate the case study to show the storage requirements for elastic buffers for 100G
  - Start of packets aligned to 8 bytes with Alignment Codes
  - It is our opinion that for MLD, clock compensation can occur only on the transmitting and receiving PCS.
    - Extension Layers will not be allowed to compensate for clocks unless the MLD layer is re-processed
  - The value of N (elastic buffers between MACs) may be set to 2

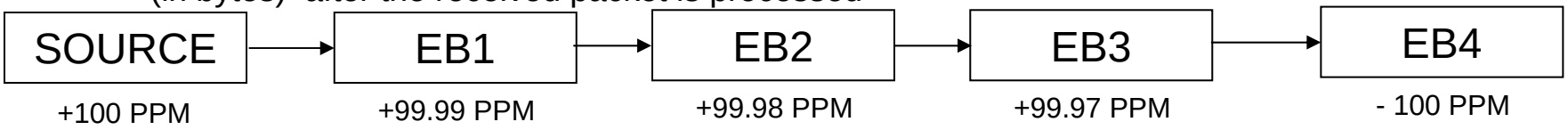
10G IPG Removal – Maximum PPM Offset



- From Elastic Buffer 1 (EB1) to EB2 there is a 200 PPM difference
- Every byte of data requires 200/1,000,000 = 1/5000 bytes of storage.
- For a X byte packet each packet fills the buffer by H {X / 5000} bytes.
- In this case EB3 can be the same speed or faster than EB2 but never slower
  - EB3 may only insert IPGs never remove them.
- IPGs are removed in groups of 4 bytes. Hence every Y {4 / H} packets an IPG adjustment is made
- The maximum storage occurs after Y packets and is 4 bytes

10G IPG Removal – Non Uniform Distribution

- Extremely small difference initially – but large difference on the last hop
  - Each packet adds ~0.001 bytes to the EB1, EB2 & EB3 and H {X/5000} bytes to EB4
  - Assume that on Packet 1 all Elastic Buffers need clock adjustment
  - FIFO Threshold T is reached before Packet 1 arrives. The table shows the status of the FIFO Fill (in bytes) after the received packet is processed

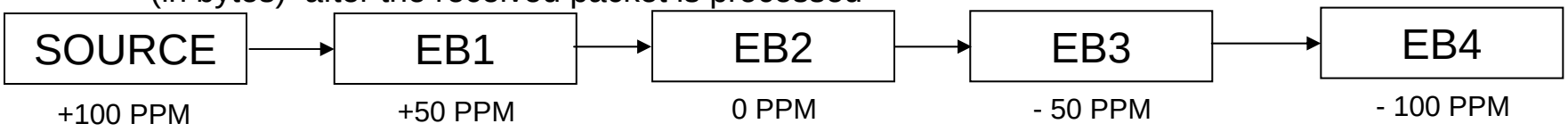


Before Packet 1	FIFO Fill is at T	FIFO Fill is at T	FIFO Fill is at T	FIFO Fill is at T
Packet 1	4 bytes IPG removed FIFO is (T – 3.999)	IPG not removed FIFO is (T + 0.001)	IPG not removed FIFO is (T + 0.001)	IPG not removed FIFO is (T + H)
Packet 2	No IPG removed FIFO is (T – 3.998)	4 bytes IPG removed FIFO is (T – 3.998)	IPG not removed FIFO is (T + 0.002)	IPG not removed FIFO is (T + 2H)
Packet 3	No IPG removed FIFO is (T – 3.997)	No IPG removed FIFO is (T – 3.997)	4 bytes IPG removed FIFO is (T – 3.997)	IPG not removed FIFO is (T + 3H)
Packet 4	No IPG removed FIFO is (T – 3.996)	No IPG removed FIFO is (T – 3.996)	No IPG removed FIFO is (T – 3.996)	4 bytes IPG removed FIFO is (T – 4 + 4H)
Packet 5	No IPG removed FIFO is (T – 3.995)	No IPG removed FIFO is (T – 3.995)	No IPG removed FIFO is (T – 3.995)	4 bytes IPG removed FIFO is (T – 4 + 5H)

Using this example we can see that the maximum storage needed at the last Elastic Buffer is N * H bytes just before the adjustment.

10G IPG Removal – Evenly distributed offsets

- Difference of -50PPM between links
  - Each packet adds L {X/20000} bytes to the buffer
  - Assume that on Packet 1 all Elastic Buffers need clock adjustment
  - FIFO Threshold T is reached before Packet 1 arrives. The table shows the status of the FIFO Fill (in bytes) after the received packet is processed



Before Packet 1	FIFO Fill is at T	FIFO Fill is at T	FIFO Fill is at T	FIFO Fill is at T
Packet 1	4 bytes IPG removed FIFO is (T – 4 + L)	IPG not removed FIFO is (T + L)	IPG not removed FIFO is (T + L)	IPG not removed FIFO is (T + L)
Packet 2	No IPG removed FIFO is (T – 4 + 2L)	4 bytes IPG removed FIFO is (T – 4 + 2L)	IPG not removed FIFO is (T + 2L)	IPG not removed FIFO is (T + 2L)
Packet 3	No IPG removed FIFO is (T – 4 + 3L)	No IPG removed FIFO is (T – 4 + 3L)	4 bytes IPG removed FIFO is (T – 4 + 3L)	IPG not removed FIFO is (T + 3L)
Packet 4	No IPG removed FIFO is (T – 4 + 4L)	No IPG removed FIFO is (T – 4 + 4L)	No IPG removed FIFO is (T – 4 + 4L)	4 bytes IPG removed FIFO is (T – 4 + 4L)
Packet 5	No IPG removed FIFO is (T – 4 + 5L)	No IPG removed FIFO is (T – 4 + 5L)	No IPG removed FIFO is (T – 4 + 5L)	No IPG removed FIFO is (T – 4 + 5L)

Using this example we can see that the maximum storage needed at the last Elastic Buffer is N * L bytes just before the adjustment.

JDSU

10G Summary - Buffering Requirements

- The maximum storage required is when the clock distribution is non uniform. For N clock crossings with X bytes packets this is

  Max Storage = MAX of [N * {X/5000} OR 4]

- Setting maximum N (EB's) as 8:

  – For standard smallest packets of 64 bytes maximum storage is

  MAX of [8 * {64/5000} OR 4] = 4 bytes

  – For standard largest packets of 2000 bytes maximum storage is

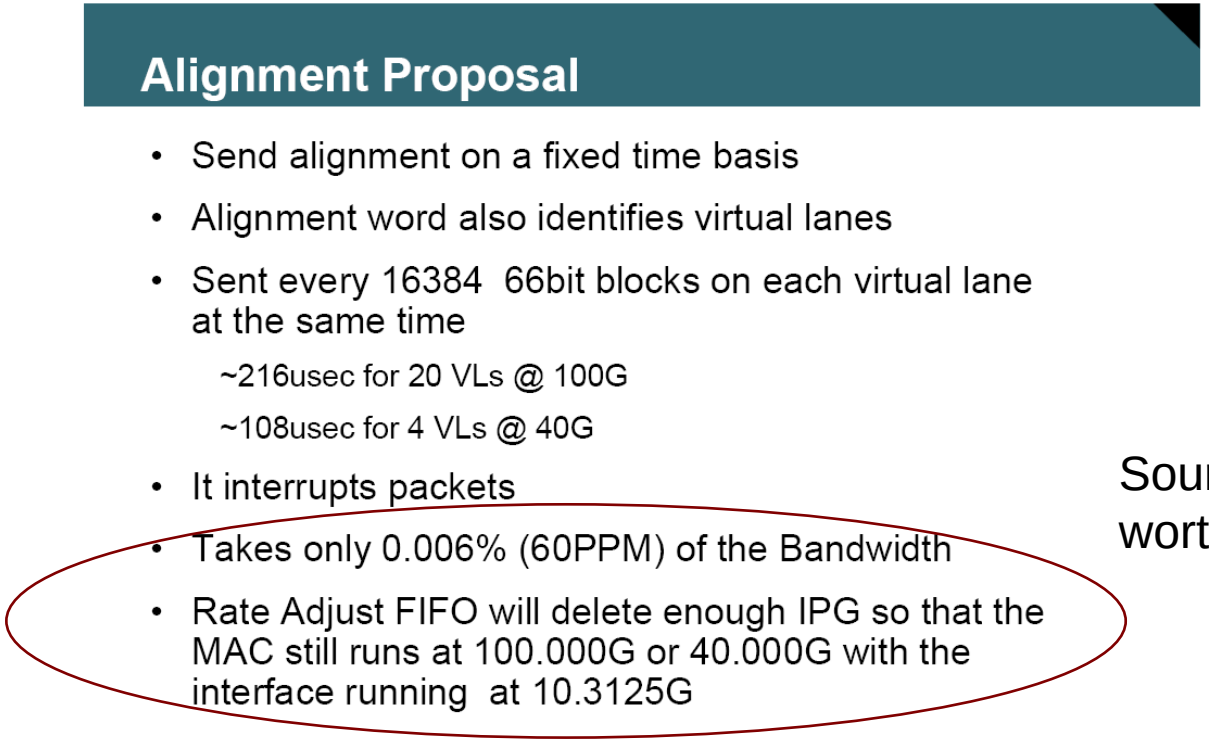
  MAX of [8 * {2000/5000} OR 4] = 4 bytes

100G Background - Start of packets at 8 byte boundaries

- Removal of IPG in 8 byte blocks for clock compensation
- Removing IPGs will not result in min. of 5 IPGs between packets – a min. 1 byte will separate packets
  - /T/ uses one IPG (would not be possible to remove control block 0x87)
- Deficit Idle Counter (DIC) will run from 0 to 7
  - IPG removal cannot occur between all packets. Removing 8 IPGs when the RS reduced the IPG to < 9 bytes is not possible
  - IPG removal can occur every other packet
    - Example: A 80 byte packet with 12 byte IPG and 8 byte preamble is 100 bytes on the line. This will get adjusted to 96 bytes or 104 bytes by the DIC. IPG can only be removed when the 104 byte packet is sent
- Buffering requirements on the clock compensation Elastic buffers will increase over the 10G case
  - Every other packet can be adjusted by the DIC such that a IPG removal may not occur
  - Worst case will be 2 times the buffering requirements of 10G because IPG removal may only happen every 2nd packet

The Alignment Code Block - Based on MLD

- Gustlin_01_0108 proposes:



Source has 60PPM worth of IPG removed.

- A 60 PPM offset is created from the source when a packet is transmitted
  - Buffering Requirements need to accommodate for this offset.

JDSU

Removing Idles for MLD Alignment

- Alignment Markers are inserted on each VL every 16k code blocks
  - 20 Markers are inserted once every 327,680 (16k*20) code blocks
  - Markers are inserted instantaneously, hence require a buffer in the PCS to store 1320 (20 x 66) bits.
  - This buffer needs to deplete over time and would require 20 idle code words to be deleted.
- Removing Idles can be done in two ways:
  - Remove idles *continuously* until the PCS buffer is depleted
    - Down stream buffers will have to handle a stream of packets with at most 40 (2x20) packets with irremovable IPG
  - Remove idles in a *weighted* fashion such that the down stream nodes only see a 60PPM offset.
    - Down streams buffers will see a 60PPM increase in clock compensation requirement. (Modeling only - really down stream nodes only see 200PPM of offset with some IPGs not removable).

100G Buffering Requirements –
8 Byte alignment with VL Marker; Continuous IPG removal

- IPG Removal can only occur every 2nd packet
- The first 20 IPG removal locations are used up for the PCS buffer
- The maximum storage required is when the clock distribution is non-uniform. For N clock crossings with X bytes packets this is

  Max Storage = MAX of [(N +20) * {2 *X/5000} OR (4 * 2)]

- Using the proposal that N be defined as 2

  – For standard smallest packets of 64 bytes maximum storage is

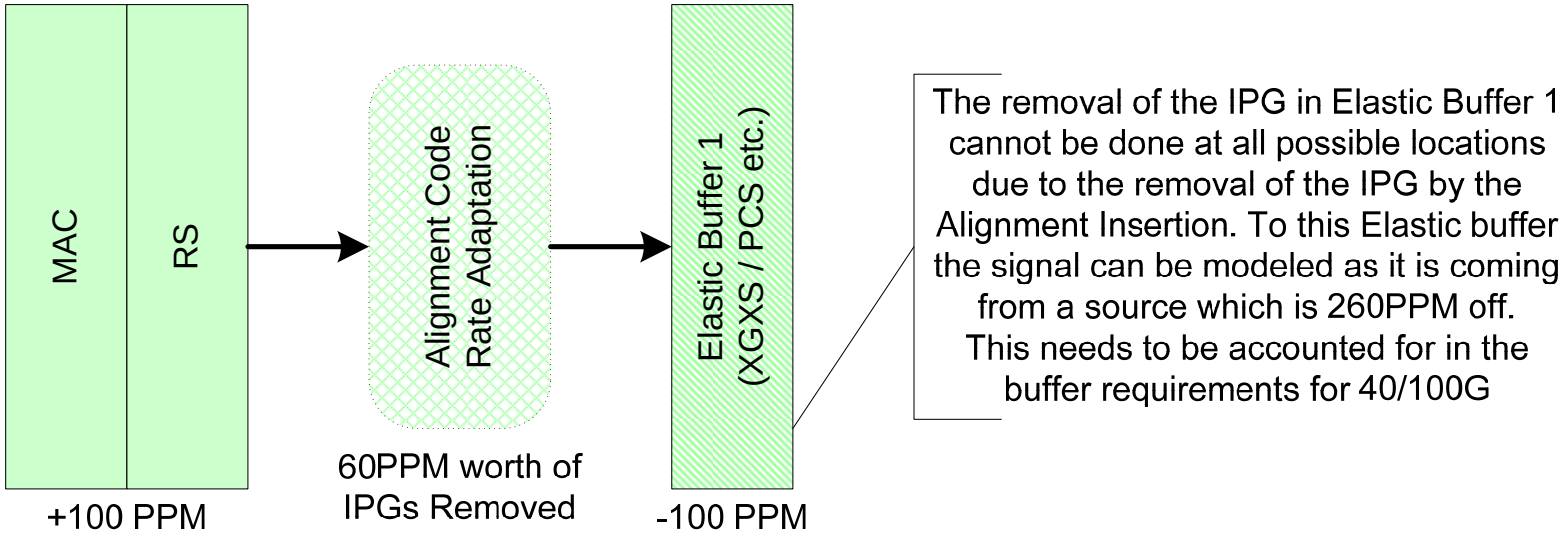
  MAX of [22 * {2 * 64/5000} OR 8] = 8 bytes

  – For standard largest packets of 2000 bytes maximum storage is

  MAX of [22 * {2 * 2000/5000} OR 8] = 17.6 bytes

Weighted Removal of IPGs - Buffering Requirements for Alignment Codes

- A standard source would add the required IPGs between packets.
- Alignment Code insertion removes 60PPM worth of IPG.
- The first Elastic Buffer after the Alignment Insertion will not see all the IPGs; the signal will look like its has been rate adapted for 60PPM
- Eg. if the source is +100PPM and EB1 is -100PPM the rate adaptation at EB1 needs to accommodate 200PPM with 60PPM worth of IPG removed.
  - This can be modeled as 260 PPM of correction.



100G Buffering Requirements –
8 Byte alignment with VL Marker; Weighted IPG removal

- IPG Removal has to accommodate extra 60PPM and can occur only every 2nd packet
- A bit of storage is needed every 1,000,000/260 ~ 3800 bits
- The maximum storage required is when the clock distribution is non uniform. For N clock crossings with X bytes packets this is

Max Storage = MAX of [N * {2 *X/3800} OR (4 * 2)]

- Using the proposal that N be defined as 8
  - For standard smallest packets of 64 bytes maximum storage is

MAX of [2 * {2 * 64/3800} OR 8] = 8 bytes

  - For standard largest packets of 2000 bytes maximum storage is

MAX of [2 * {2 * 2000/3800} OR 8] = 8 bytes

Conclusion - 100G Buffering Requirements

- Start of packet to 8 byte boundaries & account for the Alignment characters with continuous IPG Removal

  *Max Storage = MAX of [(20+N) * {2X/5000} OR 8] bytes.*

  – Maximum Storage for largest packet (X = 2000) with 2 buffers (N = 2) is

  16.97 bytes

- Start of packet to 8 byte boundaries & account for the Alignment characters with a weighted removal of IPGs

  *Max Storage = MAX of [N * {2X/3800} OR 8] bytes.*

  – Maximum Storage for largest packet (X = 2000) with 8 buffers (N = 8) is

  8 bytes

Summary

- For 802.3ba it should be explicitly stated that only one MLD layer may occur between MACs
  - Fixes the number of EB's locations to 2
- An appendix should be provided to specification which highlights the minimum storage required for each elastic buffer
  - This should be presented as an equation rather than a hard number
  - If a number is to be provided it should be large enough to accommodate reasonable non-standard packet types
    - We propose 256 bytes
- A clear definition of the IPG removal for MLD alignment must be provided to determine buffering requirements
  - Using a continuous IPG removal will cause some packet jitter

Subtle Observations

- 8 Byte removal can be done between MLD and PCS as a 66B control word (IPG) will be removed.
  - PCS Scrambling would still need to be redone
  - There seems to be a clear benefit to having 8 byte IPG removal for bypassing the PCS layer.
- We need to consider if at some point in the future there are possibilities of having more than one MLD layer between MACs (ie. More than 2 EB's)
  - In this case we should scale the size of the IPG buffer

Thank You

Questions?

BACKUP – Deficit Idle Counter

100G Buffering Requirements – Conclusion
Jumbo Packets

- If we keep the start of packet at 4 byte boundaries

*Max Storage = MAX of [N * {X/5000} OR 4] bytes.*

  - Maximum Storage for largest packet (X = 10K) with 8 (N = 8) buffers is ***16 bytes***

- If we change the start of packet to 8 byte boundaries

*Max Storage = MAX of [N * {2X/5000} OR 8] bytes.*

  - Maximum Storage for largest packet (X = 10K) with 8 buffers (N = 8) is ***32 bytes***

- If we change the start of packet to 8 byte boundaries & account for the Alignment characters with continuous IPG Removal

*Max Storage = MAX of [(20+N) * {2X/5000} OR 8] bytes.*

  - Maximum Storage for largest packet (X = 10K) with 8 buffers (N = 8) is ***112 bytes***

- If we change the start of packet to 8 byte boundaries & account for the Alignment characters with a weighted removal of IPGs

*Max Storage = MAX of [N * {2X/3800} OR 8] bytes.*

  - Maximum Storage for largest packet (X = 10K) with 8 buffers (N = 8) is ***42.1 bytes***

Proposal for 100G – Starting at 8 byte boundaries

- The Deficit Idle Counter (DIC) will move between 0 and 7.
- Define a Transmission Unit (TU) as preamble + MAC packet + IPG
- For TUs which are not multiple of 8 bytes we need to remove or add idles to align to 8 bytes.
- DIC Resets to 0 at start up.
- Same Example as before (TU of 97 bytes)

Packet Number	1	2	3	4	5	6	7	8
TU after RS	96	96	96	96	96	96	96	104
DIC	1	2	3	4	5	6	7	0
Action	- 1 IPG	- 1 IPG	- 1 IPG	- 1 IPG	- 1 IPG	- 1 IPG	-1 IPG	+ 7 IPG
IPG b/w pkts	11	11	11	11	11	11	11	19

- For TUs that are multiple of 4bytes we still need DIC
  - Eg. TU of 100 will be output as 96 and 104 bytes.

Starting at 8 byte boundaries – Summary

- Removal of IPG in 8 byte blocks for clock compensation.
- Removing Idles will not result in minimum of 5 IPGs between packets.
- This means insertion of clock compensation will not be after a full column of idles in the RS.
- IPG removal for clock compensation cannot occur between all packets. Removing 8 IPGs when the RS reduced the IPG to < 9 bytes is not possible.
  - /T/ uses one IPG (would not be possible to remove control block 0x87)
- Buffering requirements on the clock compensation FIFOs will increase over the 10G case.
  - Worst case will be 2 times the buffering requirements of 10G because IPG removal will happen every 4th packet for Jumbo packets.
  - Every other packet can be adjusted by the DIC such that a IPG removal may not occur.
  - Maximum buffering at each hop will be ~ 4N bytes where N is the number of clock crossings between MACs.

Deficit Idle Counter (DIC) @ 10G

- Used in the RS to align packets to 32 bit boundaries.
  - Resets to 0 at start.
  - Constrained between 0 and 3.
  - When idles are removed the counter is incremented by the value.
  - When idles are inserted the counter is decremented by the value.

Example:

Define a Transmission Unit (TU) as preamble + MAC packet + IPG

TU of 97 bytes are received continuously – DIC starts at 0

Packet Number	1	2	3	4	5	6	7	8
TU after RS	96	96	96	100	96	96	96	100
DIC	1	2	3	0	1	2	3	0
Action	- 1 IPG	- 1 IPG	- 1 IPG	+ 3 IPG	- 1 IPG	- 1 IPG	-1 IPG	+ 3 IPG

- In the case with a DIC of upto 3 it can be seen that the after the RS we still always have at-least 9 IPG bytes between packets [TU of 99 bytes goes out as 96, 100, 100, 100, 96]

Removing Idles in 10G

- For Jumbo Packets (10K bytes) we need to remove 2 bytes after each packet with the worst clock difference.
- We must always have at least 5 IPG bytes between packets. Hence in the case of a full line rate IPG can only be removed once.
- The removal of the IPG are a factor of the difference in the clock rate between hops.
- If we encounter the slowest clock rate then IPG removal is not needed in the next hop.
  - The next hop can only be the same speed or slower.
- If all hops need to remove IPG at the same time – additional storage in each FIFO is required.

IPG Removal 10G – Conclusion (2)

- For the LAN – OTN (multiple hops) – LAN case.
  - Steve's email explains that OTN re-generates the transmit clock to very good accuracy.
  - I could only think of one place where this is mentioned – ITU-T 7041 Section 8.4.1.1.2 (GFP-T egress clock recovery).
  - If the OTN clock is recovered then there is no need to account for any IPG removal between OTN hops – they would be essentially transparent to the LAN.
  - Options 2 & 3 from trowbrige_01_0108 will provide for this mechanism.