Throughput of different Lane Aggregation Methods

Stephen J. Trowbridge
Alcatel-Lucent

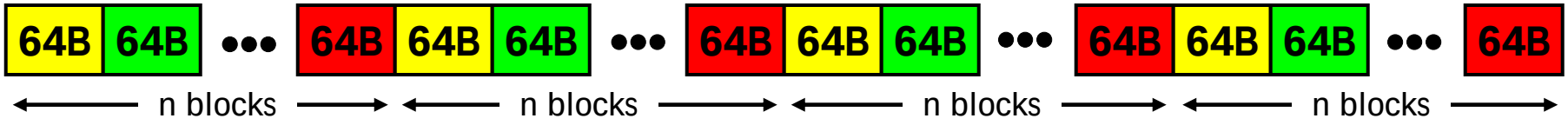
IEEE Higher Speed Study Group
November 2007

Outline

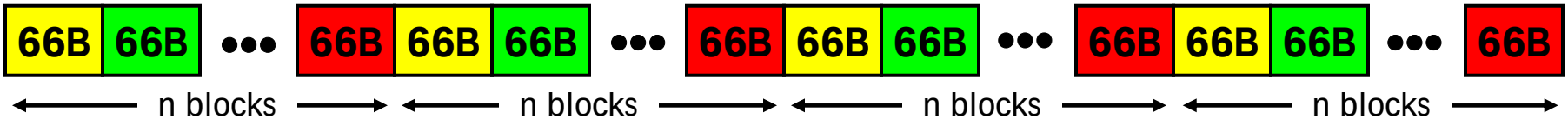
- Lane Aggregation Alternatives
  - (Virtual) Lane alignment markers inserted by increasing lane rate
  - (Virtual) Lane alignment markers inserted by stealing from IPG
  - Aggregation at the physical layer (APL) - distribute packet fragments similar to clause 61
- Lane Configurations examined
  - 100 GbE – 10 lanes
  - 100 GbE – 4 lanes
  - 40 GbE – 4 lanes

Lane alignment markers inserted by increasing lane rate

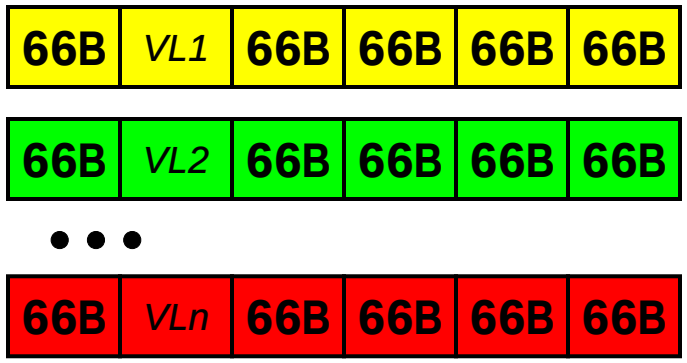
MII:



64B/66B encode:



Inverse multiplex into n (virtual) lanes, add lane alignment markers:

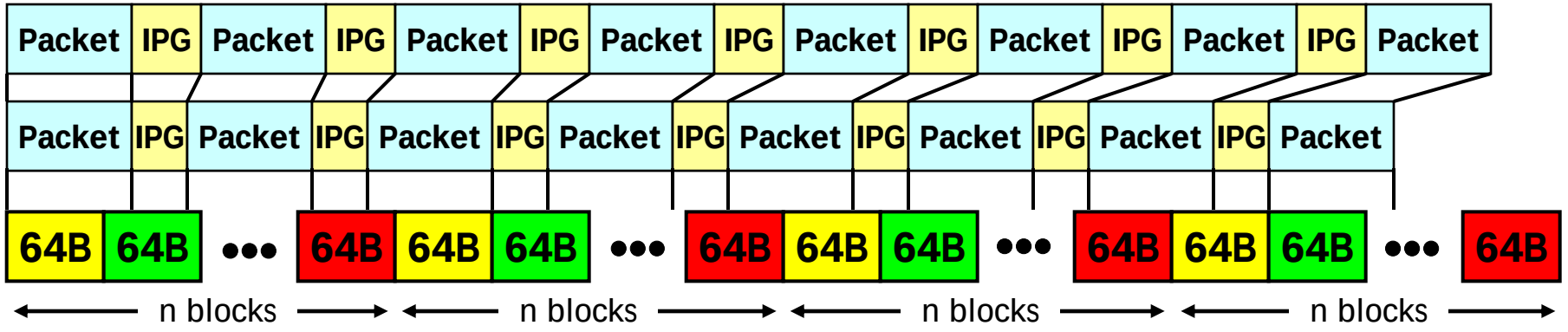


Lane rate =
(MAC + PCS) rate / n
x 1.000061

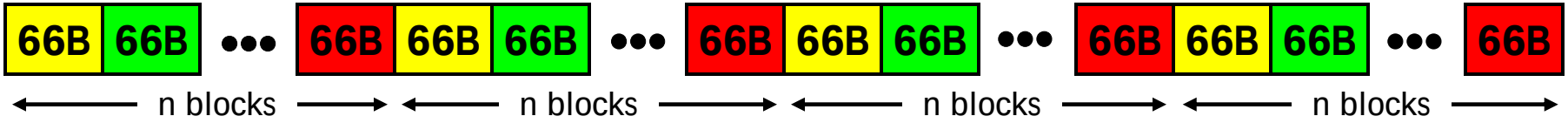
Alcatel·Lucent

Lane alignment markers inserted by stealing from IPG

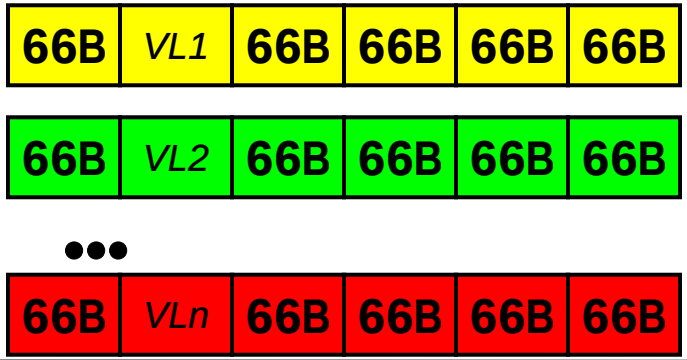
MII: Prior to PCS coding, reduce rate by 0.0061% by deleting from IPG



64B/66B encode:



Inverse multiplex into n (virtual) lanes, add lane alignment markers:

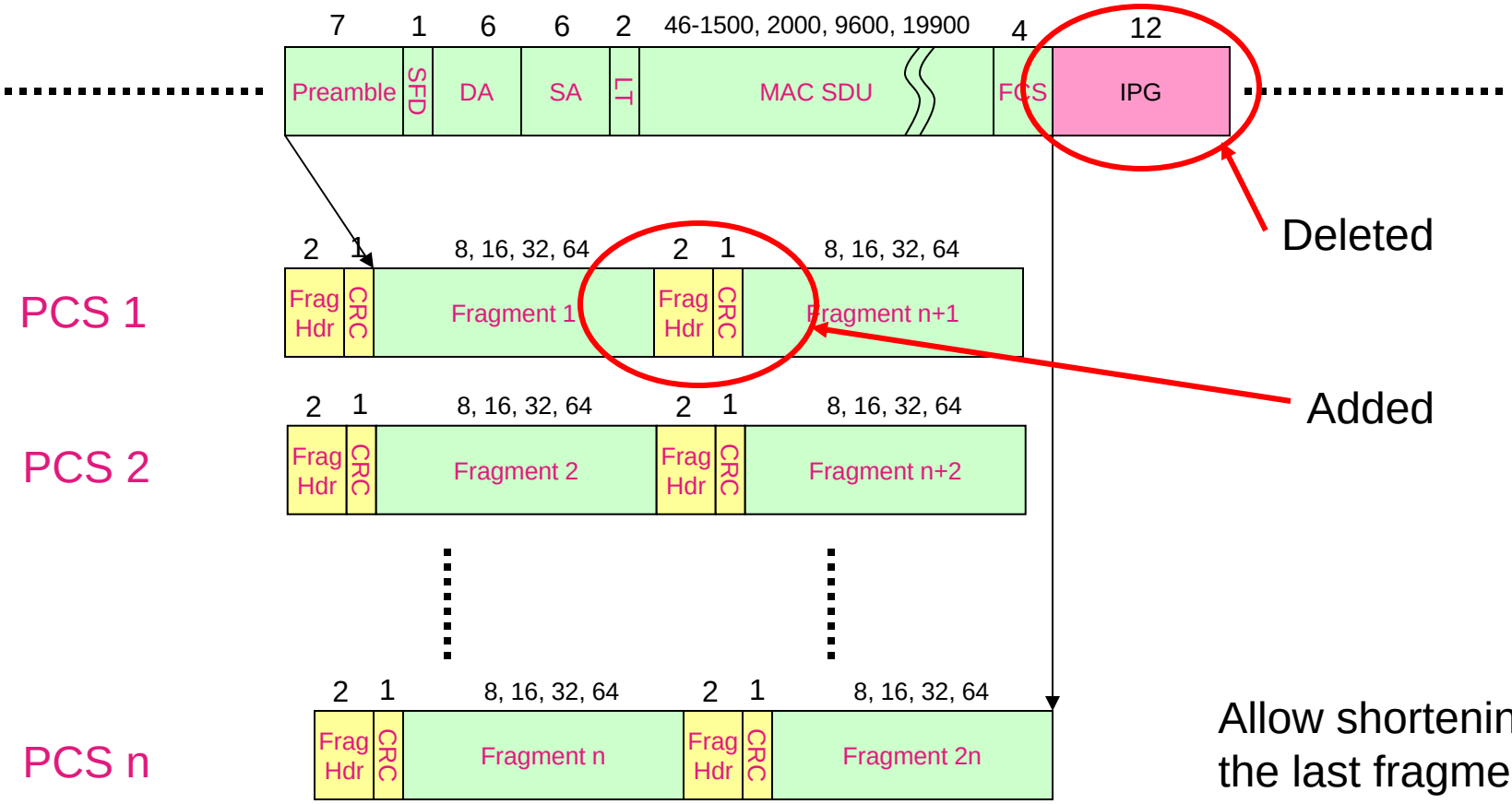


Lane rate = (MAC + PCS) rate / n

Alcatel·Lucent

Review of Physical Layer Aggregation (based on clause 61)

ref. frazer_01_1106.pdf



Allow shortening the last fragment

Alcatel·Lucent

Cases to be considered

	100 GbE 10 lanes	100 GbE 4 lanes	40 GbE 4 lanes
(Virtual) lane alignment markers inserted by increasing lane rate	A	B	C
(Virtual) lane alignment markers inserted by stealing from IPG	D	E	F
Aggregation at the Physical Layer (APL)	G	H	I

Case A – Ten lane 100 GbE Interface with increased lane rate to carry 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 PMD Rx PMD	Two virtual lanes per physical lane	10x10.31312946 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

Case B – Four lane 100 GbE Interface with increased lane rate to carry 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

Case C – Four lane 40 GbE interface with lane rate increased to carry 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

Cases A, B, C - Is there any throughput limitation with lane alignment markers carried by increasing lane rate?

- **Lane alignment markers don't take away from the MAC**
- **Only IPG shrinkage is due to clock alignment**

MAC SDU size		Preamble	SFD	DA	SA	LT	FCS	Packet Total	IPG	Total	+100ppm into -100ppm	IPG Shrinkage
46	Minimum	7	1	6	6	2	4	72	12	84	83.98	0.02
100		7	1	6	6	2	4	126	12	138	137.97	0.03
500		7	1	6	6	2	4	526	12	538	537.89	0.11
1500	Max Basic	7	1	6	6	2	4	1526	12	1038	1537.69	0.31
2000	802.3as	7	1	6	6	2	4	2026	12	2038	2037.59	0.41
9600		7	1	6	6	2	4	9626	12	9638	9636.07	1.93
19900	Max Jumbo	7	1	6	6	2	4	19926	12	19938	19934.01	3.99

Case D – Ten lane 100 GbE interface with lane alignment markers stealing from IPG

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 PMD Rx PMD	Two virtual lanes per physical lane	10x10.3125 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

Case E – Four lane 100 GbE interface with lane alignment markers stealing from IPG

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

Case F – Four lane 40 GbE interface with lane alignment markers stealing from the 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

Cases D, E, F - Is there any throughput limitation with lane alignment markers carried by stealing from IPG?

- **Need to add space required for lane alignment markers to IPG shrinkage required clock alignment**

MAC SDU size		Pre amble	SFD	DA	SA	LT	FCS	Packet Total	IPG	Total	+100ppm into -100ppm	Remove Space for lane markers	Total IPG Shrinkage
46	Minimum	7	1	6	6	2	4	72	12	84	83.98	83.98	0.02
100		7	1	6	6	2	4	126	12	138	137.97	137.96	0.04
500		7	1	6	6	2	4	526	12	538	537.89	537.86	0.14
1500	Max Basic	7	1	6	6	2	4	1526	12	1038	1537.69	1537.60	0.40
2000	802.3as	7	1	6	6	2	4	2026	12	2038	2037.59	2037.47	0.53
9600		7	1	6	6	2	4	9626	12	9638	9636.07	9635.48	2.52
19900	Max Jumbo	7	1	6	6	2	4	19926	12	19938	19934.01	19932.80	5.20

Cases G, H, I – Is there any throughput limitation from APL?

- **Need to add Fragmentation overhead**
- **Need to remove IPG**
- **May need to further restrict throughput due to clock differences**
- **Not analyzed: "LAG-like" effect of shortened final fragments producing uneven traffic distribution across lanes (insignificant for small fragments, but growing with larger fragment sizes)**

MAC SDU size		Preamble	SFD	DA	SA	LT	FCS	Packet Total to be distributed	IPG to be deleted
46	Minimum	7	1	6	6	2	4	72	12
100		7	1	6	6	2	4	126	12
500		7	1	6	6	2	4	526	12
1500	Max Basic	7	1	6	6	2	4	1526	12
2000	802.3as	7	1	6	6	2	4	2026	12
9600		7	1	6	6	2	4	9626	12
19900	Max Jumbo	7	1	6	6	2	4	19926	12

Cases G, H, I – Is there any throughput limitation from APL?

- **For each packet:**
  - **Divide into fragments**
  - **Add 3 bytes overhead per fragment**
  - **Remove 12 bytes IPG per packet to get net byte count**

Packet Total	IPG	Fragment Size											
		8			16			32			64		
		Frags	F OH	Net	Frags	F OH	Net	Frags	F OH	Net	Frags	F OH	Net
72	12	9	27	+15	5	15	+3	3	9	-3	2	6	-6
126	12	16	48	+36	8	24	+12	4	12	0	2	6	-6
526	12	66	198	+186	33	99	+87	17	51	+39	9	27	+15
1526	12	191	573	+561	96	288	+276	48	144	+132	24	72	+60
2026	12	254	762	+750	127	381	+369	64	192	+180	32	96	+84
9626	12	1204	3612	+3600	602	1806	+1794	301	903	+891	151	453	+441
19926	12	2491	7473	+7461	1246	3738	+3726	623	1869	+1857	312	936	+924

Cases G, H, I – How much throughput reduction is there from APL?

- **Bandwidth is restricted from two sources**
  - **Since IPG already removed during fragmentation, no spare capacity to absorb ±100ppm clock difference**
  - **Fragment overhead bytes in excess of IPG will further restrict throughput**

Packet Total	Fragment Size							
	8		16		32		64	
	Frag OH	Frag OH + clock	Frag OH	Frag OH + clock	Frag OH	Frag OH + clock	Frag OH	Frag OH + clock
72	-17.86%	-17.88%	-3.57%	-3.59%	3.57%	3.55%	7.14%	7.12%
126	-26.09%	-26.11%	-8.70%	-8.72%	0.00%	-0.02%	4.35%	4.33%
526	-34.57%	-34.60%	-16.17%	-16.19%	-7.25%	-7.27%	-2.79%	-2.81%
1526	-36.48%	-36.50%	-17.95%	-17.97%	-8.58%	-8.60%	-3.90%	-3.92%
2026	-36.80%	-36.83%	-18.11%	-18.13%	-8.83%	-8.85%	-4.12%	-4.14%
9626	-37.35%	-37.38%	-18.61%	-18.64%	-9.24%	-9.27%	-4.58%	-4.60%
19926	-37.42%	-37.45%	-18.69%	-18.71%	-9.31%	-9.34%	-4.63%	-4.66%

Conclusions

- Lane alignment marker methods do not restrict the packet throughput from the MII
- Lane alignment markers inserted by increasing the lane rate (rather than deleting from IPG) result in less IPG shrinkage
- APL by packet fragmentation will generally reduce packet throughput (significantly, if smaller fragments are used or user traffic consists of larger (jumbo) frames)
  - Small packets and large fragments result in the least restriction of throughput (only the smallest packets flow without any restriction of bandwidth)
  - Lose opportunity for clock rate matching within IPG since IPG is deleted due to fragmentation
  - Amount of bandwidth restriction depends on packet size, so no fixed limit can be enforced at MII to avoid packet loss
  - Will end users who didn't like LAG because of throughput restrictions based on distribution of flow sizes accept a lane distribution method that restricts throughput based on average packet size?