BROCADE

nX50G PCS WITH m PMDS n = 1, 2, 4 AND m = KR, CR AND SR

By Scott Kipp

09-15-15

kipp_nge_01_0915

Supporters

- Contributors
  - Vineet Salunke, Cisco
  - Rob Stone, Broadcom
  - Ali Ghiasi, Ghiasi Quantum
  - Chris Cole, Finisar
  - Jeff Maki, Juniper
  - John D'Ambrosia, Dell
  - Kapil Shrikhande, Dell
- Supporters
  - Jonathan King, Finisar
  - Martin Skagen, Brocade
  - David Skirmont, Brocade
  - Scott Sommers, Molex
  - Tom Palkert, Molex
  - Paul Kolesar, Commscope
  - Nathan Tracy, TE Connectivity
  - Mark Gustlin, Xilinx
  - Peter Stasser, Huawei
  - Doug Coleman, Corning
  - Steve Swanson, Corning
  - Mike Dudek, QLogic

Agenda

- Why standardize nX50G?
- Sample architectures
- PMD Requirements
- Who needs nX50GbE?

The Right Sized Project

- We’re looking for the goldilocks project
  - Not too big – takes too long
  - Not too small – limits market adoption
  - Just right in the groove – meets most needs
- 50G technology usher in a new generation of Ethernet speeds and the best tradeoffs for the first project are:
  - Number of Lanes: 1, 2, 4, 8, 16
  - Number of PMDs: KR, CR, SR, PSM, CWDM, LR, ER
  - Number of FECs: KR, RS(528,514), RS(544, 514)
  - Red text shows optimal selection

Why include nX50G?

- Starting a 50GbE project without 100GbE and 200GbE is short sighted and ignores optimizations regarding:
  - Common logic architecture
  - FEC selection and implementations
  - Port commonality between multiple implementations such as: SFP, QSFP CFP4, COBO and other form factors
- Need to leverage 50G signaling technologies across multiple speeds
- 200GbE is for the Switch market when 50GbE servers are prevalent
- Minimal additional work for good ROI

Why not use LAG or FlexE for 200GbE?

- LAG
  - Reduces efficiency specially in case of intermix traffic
  - Higher latency due to round robin frame distribution
- FlexE is touted as more efficient implementation of LAG
  - Initiated to support variable rate 25G increment to maximize Coherent optics links capacity depending on the fiber length and/or OSNR
  - Another application of FlexE is more efficient than LAG data distribution over multiple physical lanes
  - FlexE likely will define 200G MAC rate
- QSFP56 natively supports 50, 100, 200GbE without the FlexE overhead

Keeping up with the Jones

- The whole technology ecosystem is growing faster and data centers can’t wait for 400GbE costs to come down
  - 40GbE was adopted when 100GbE was too expensive
- Intel new processor delivers 200Gb/s throughput
- Flash memory improving application performance drastically – often 30%
- OTN using 200G wavelengths
- Lower cost/bit and power/bit than 25/100GbE
- 200GbE will be lower cost/port than 400GbE
- 50/200GbE is the right evolutionary step

BROCADE

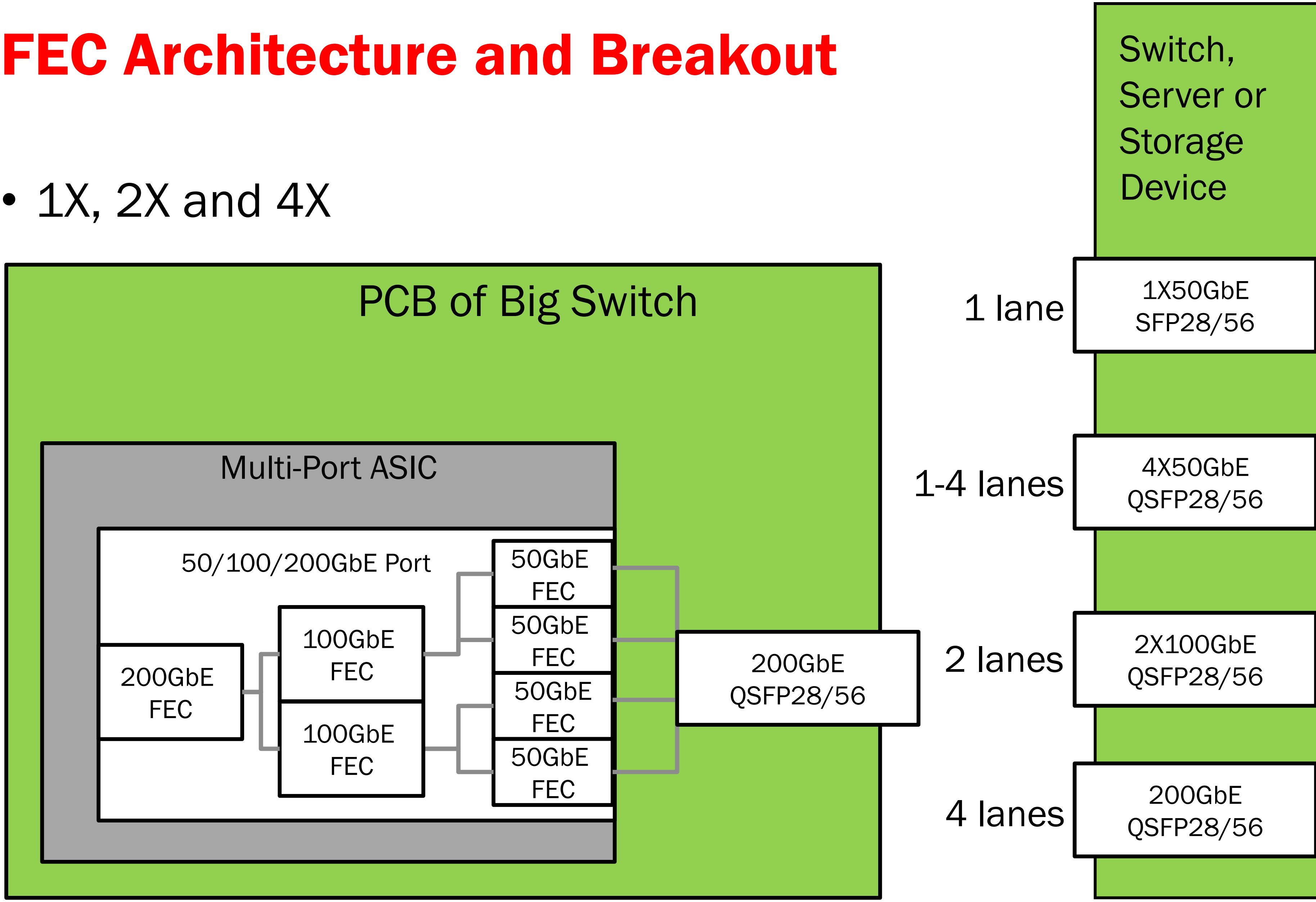
Sample Architectures

How nX50GbE could be implemented?

- High port count ASICs (>100 ports) and switches require higher densities than SFP, so ports must be grouped
- QSFP is excellent way to enable high port count switches with:
  - 128-144 ports/1U
  - Lower cost/bit
- Taking a limited view of a 50GbE implementation leads to inconsistencies and later challenges
- We have the opportunity to do it right

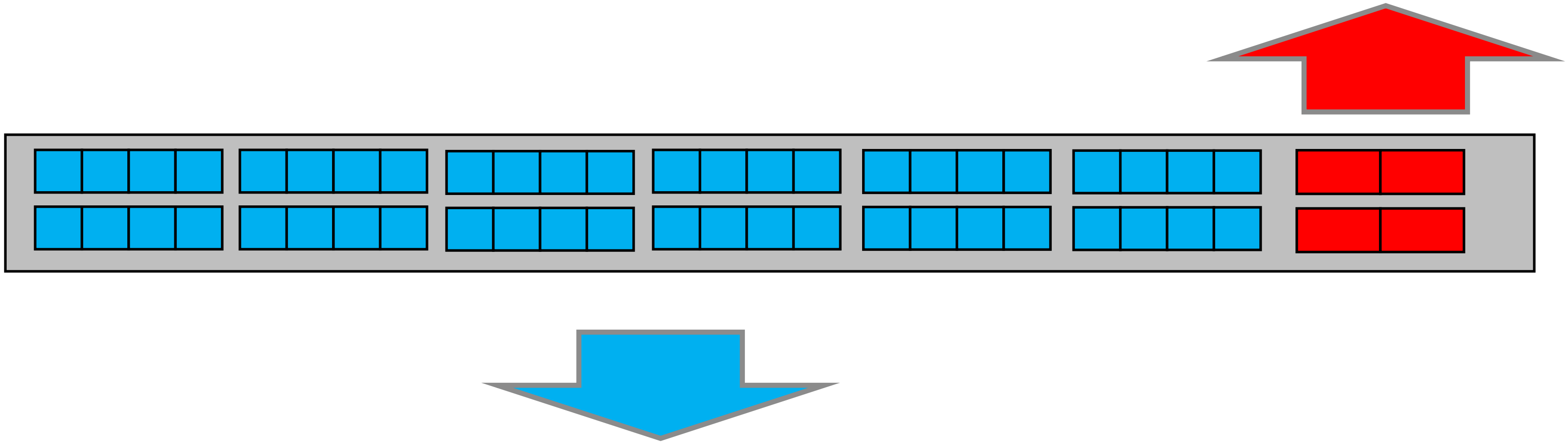
FEC Architecture and Breakout

- 1X, 2X and 4X



64 Port Switch Designs

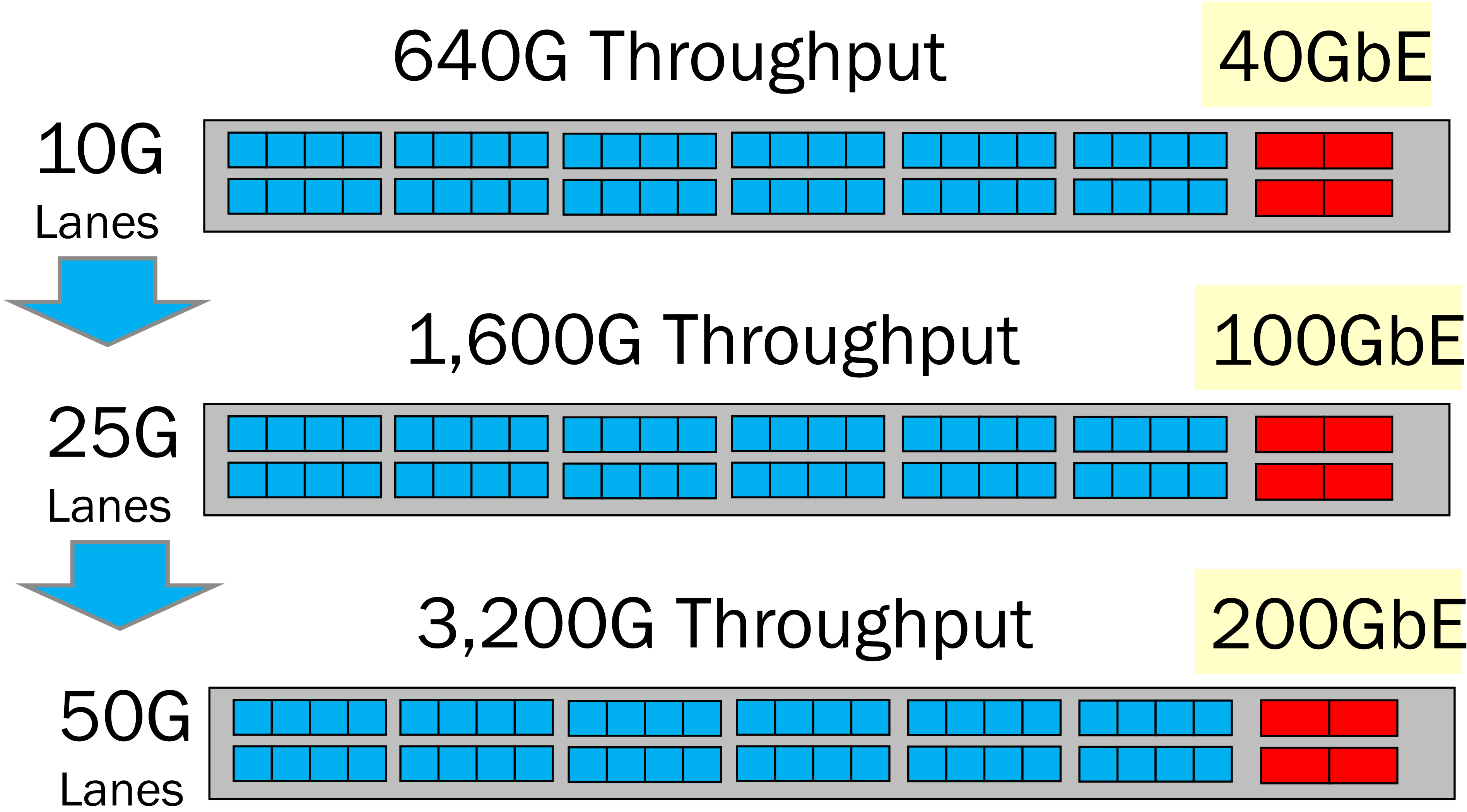
- 64 port ASICs have been very successful because they offer a balanced solution of up and down ports



4 40GbE QSFP+ Uplinks to core

48 10GbE SFP+ Downlinks to servers

Progression in Speeds for 64 Port Switches



640G Throughput

40GbE

10G
Lanes

1,600G Throughput

100GbE

25G
Lanes

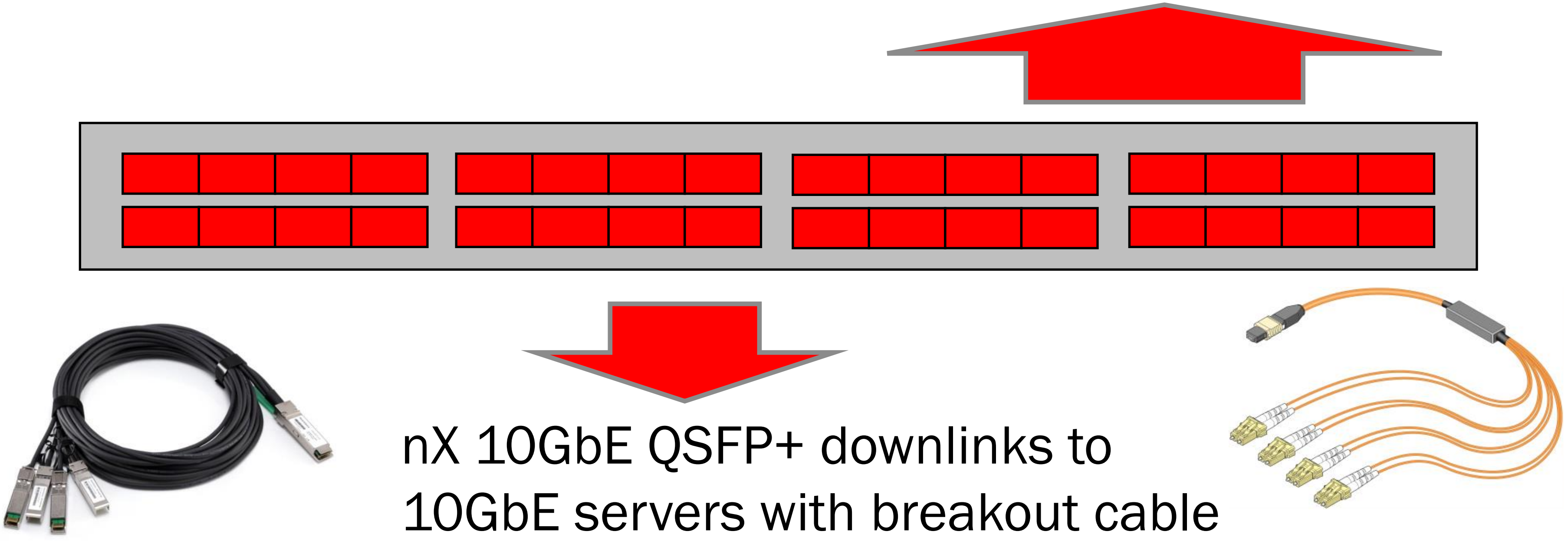
3,200G Throughput

200GbE

50G
Lanes

128 Port Switch Designs

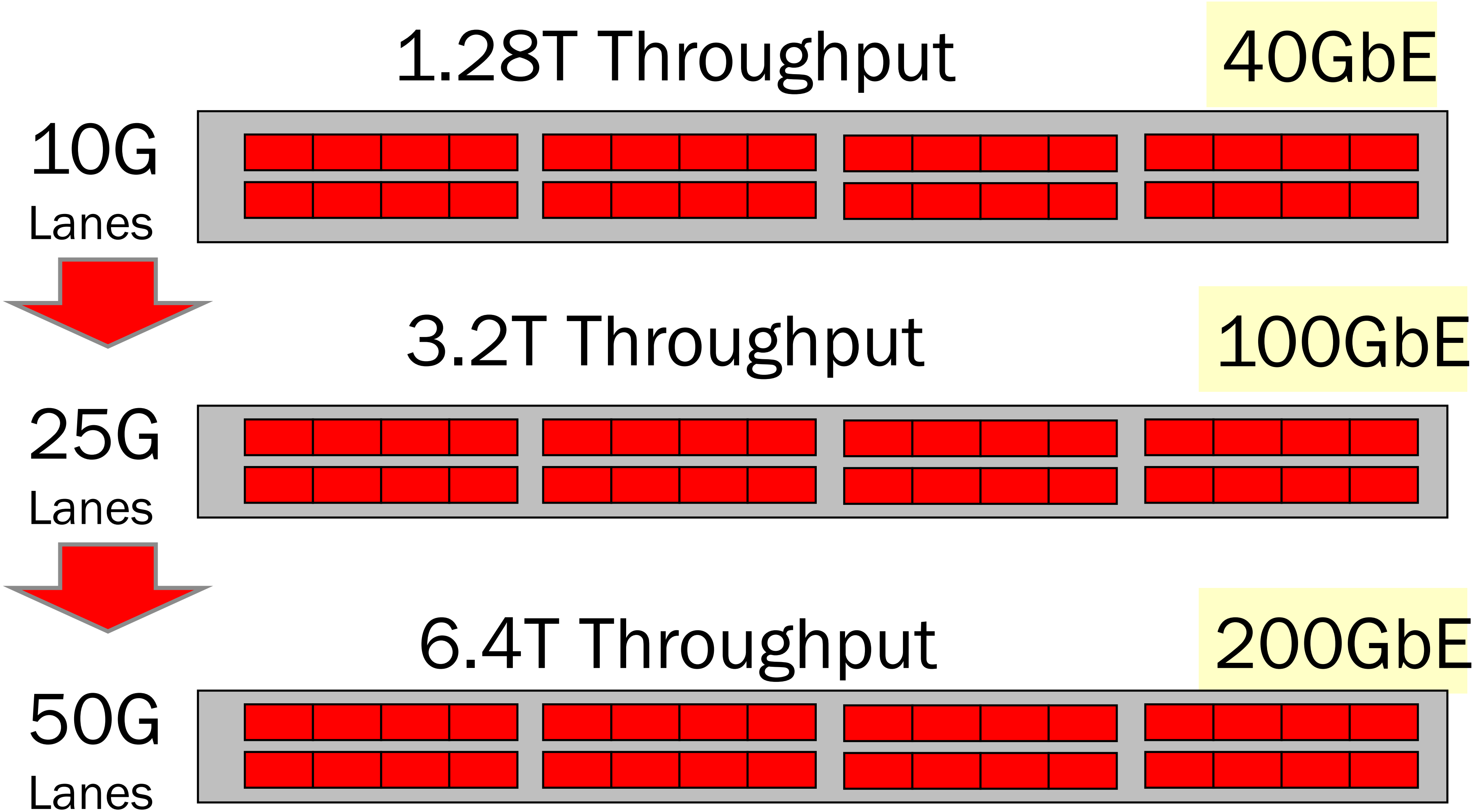
- 128 port ASICs are emerging as new high port count switch in 32 QSFP port configuration



n 40GbE QSFP+
Uplinks to core

nX 10GbE QSFP+ downlinks to
10GbE servers with breakout cable

Progression in Speeds - Easy to Understand!



1.28T Throughput

40GbE

10G

Lanes

3.2T Throughput

100GbE

25G

Lanes

6.4T Throughput

200GbE

50G

Lanes

200GbE Backplane Support

- Many backplane implementations are based on 4-lane technologies and well suited to be upgraded from 100GbE to 200GbE
  - Some say 8 lanes can’t be effectively routed across backplanes
  - Doubling the backplane capacity is necessary to deliver 6.4 Tb per line card while non-blocking
    - 32 ports of QSFP56 will be the natural progression to deliver 6.4 Tb
- 50/200GbE backplanes will double the bandwidth of 25/100GbE backplanes
- 400GbE backplanes aren’t being standardized and might be delayed until 4X100G lanes become technically feasible

4 Lane is Norm for High Throughput Links

- 4 lane implementations offer excellent forward and backward compatibility for line cards and backplanes
  - Let's standardize 4X100G Lanes when the time is right!
- QSFP28/56 is best selling parallel optical module and should support 200GbE and:
  - Is well established in the industry
  - Has a wide range of suppliers
  - Has volume production of connector and cage to create lower cost/bit than SFP28/56
- 8 and 16 lanes require different modules that are in lower volume and thus higher cost

Number of Lanes

- Key attributes for number of lanes

Number of Lanes	Does it fit In QSFP?	Does it use <=12 fibers?
1 – 50GbE	Yes – X4	Yes
2 – 100GbE	Yes – X2	Yes
4 – 200GbE	Yes	Yes
8 – 400GbE	No	No
16 – 800GbE	No	No

Just Right with n = 1,2,4

Cutoff Line

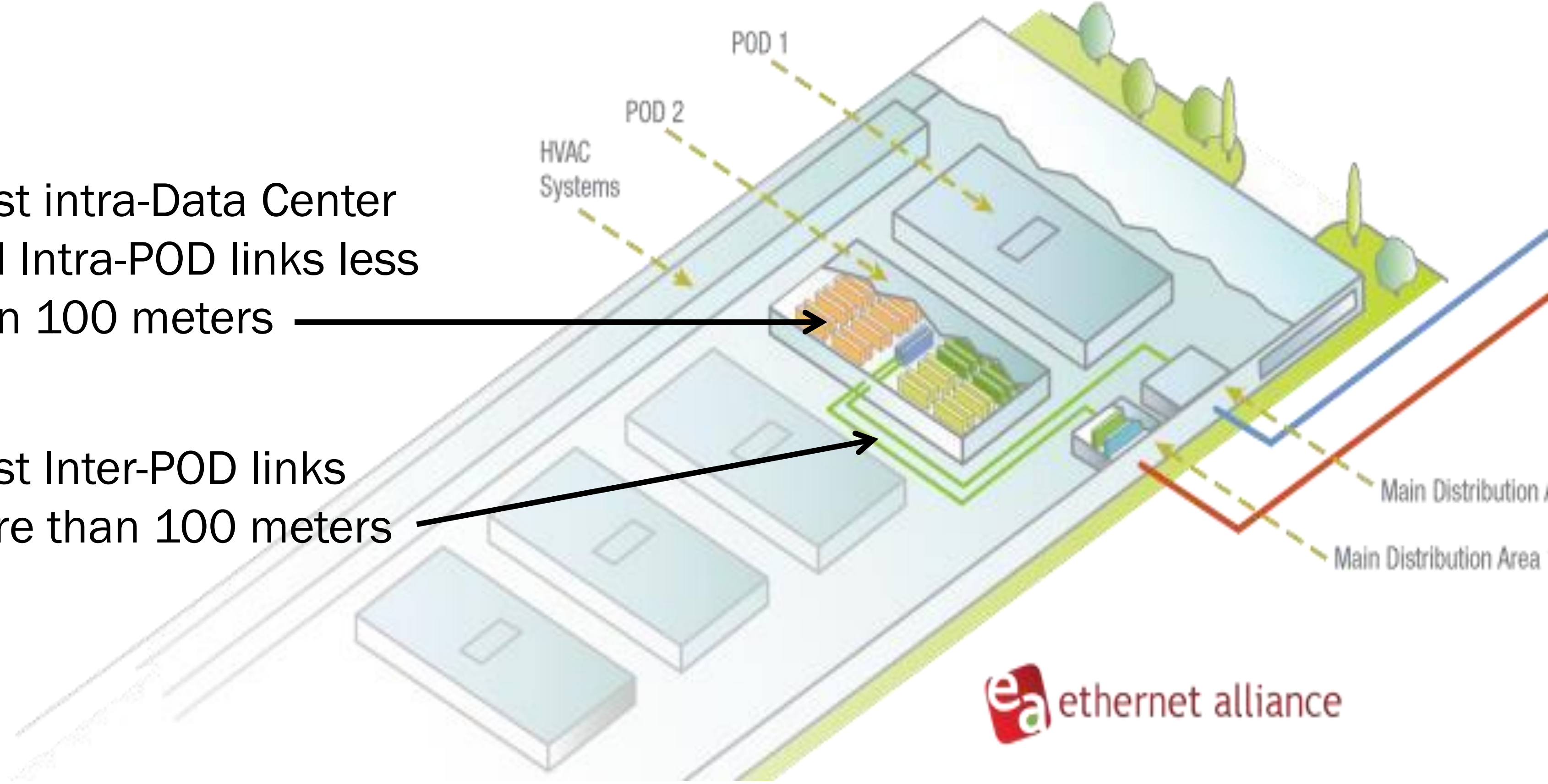
PMD Requirements

Best Solution for PMDs

- Best solution is for a port to eventually support from 3m to 10km+
  - Some call it a Universal Port
  - 10km solutions will be produced with or without IEEE standard
    - T11 is standardizing a 10km 64GFC link
- We don’t have to standardize all of the PMDs in the first project
  - Many PMDs have been defined in MSAs successfully
  - Future IEEE projects can broaden the PMD support

Data Center Reach Requirements

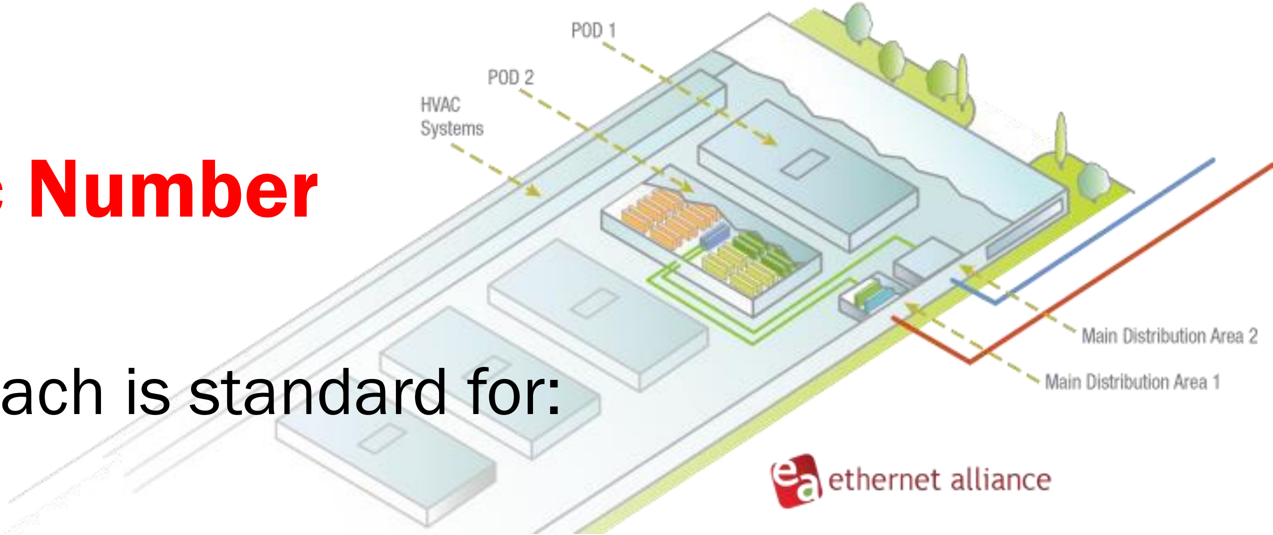
- Most Data Centers have been designed around 100 meter link reach
- PODs break Mega-DCs down into manageable pieces where intra-POD distances are less than 100 meters



Most intra-Data Center and Intra-POD links less than 100 meters

Most Inter-POD links more than 100 meters

100m is Magic Number



- 100 meter link reach is standard for:
  - 1000BASE-T
  - 10GBASE-T
  - 40GBASE-SR4
  - 100GBASE-SR10/4
  - 25GBASE-SR
- 100 meters should be feasible with 50G PAM-4 VCSELs with OM4 for 50,100,200GbE

SMF Possibilities

- Need SMF to go beyond 100 meters and we have too many solutions for a quick selection – no consensus
- For multi-lane WDM, there is no λ consensus
- We don’t want to delay the project with SMF PMDs
- Possible SMF PMD solutions include:

PMD Name	500m	2km	10km	20/40km
25G	DR	FR	LR	ER
50G	DR	FR	LR	ER
100G	DR, DR2	FR, FR2	LR, LR2	ER2, ER4f
200G	DR2,DR4	FR2, FR4	LR2, LR4	ER4

PMDs for This Project

- Compromise position is for common ports to support 3-100 meters in this project
- Evaluation of PMDs

PMD Name	Application	Distance	Include in project?
KRn	Backplane	~1m	Yes
CRn	Twinax Direct Attach Cable (DAC)	~3m with FEC	Yes, for server connectivity
SRn	MMF	100m	Yes, one way forward
PSMn	Parallel SMF	500m?	No, lack of consensus
CWDMn	Duplex SMF	2km	No, lack of consensus
LRn	Duplex SMF	10km	No, lack of consensus

Cutoff Line

Who needs nX50GbE?

Need for nX50GbE

- Drivers for Higher Speed

Faster Processors

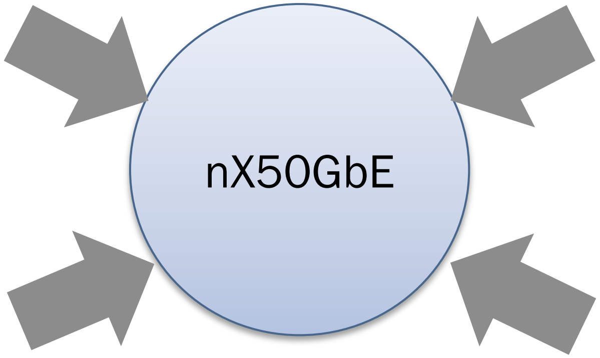
More Cores

Better I/O

More Applications

IoT

More Devices



Flash Memory

Low latency

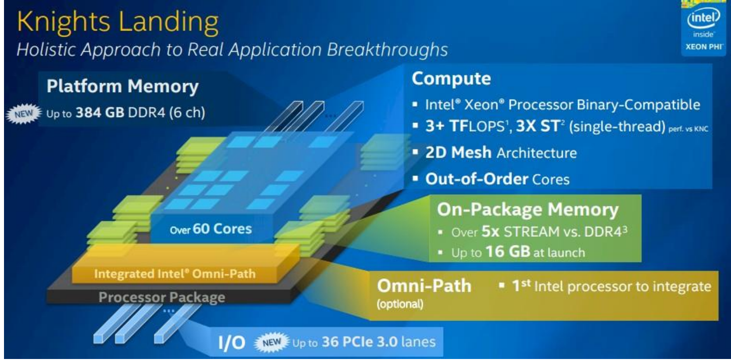
Better Performance

More Data

30-40% Growth/year

Faster Processors

>100G/processor


Knights Landing
Holistic Approach to Real Application Breakthroughs
intel inside XEON PHI
Platform Memory
NEW Up to 384 GB DDR4 (6 ch)
Compute
▪ Intel® Xeon® Processor Binary-Compatible
▪ 3+ TFLOPS[1], 3X ST[2] (single-thread) perf. vs KNC
▪ 2D Mesh Architecture
▪ Out-of-Order Cores
Over 60 Cores
On-Package Memory
▪ Over 5x STREAM vs. DDR4[3]
▪ Up to 16 GB at launch
Integrated Intel® Omni-Path
Processor Package
Omni-Path (optional)
▪ 1st Intel processor to integrate
I/O NEW Up to 36 PCIe 3.0 lanes

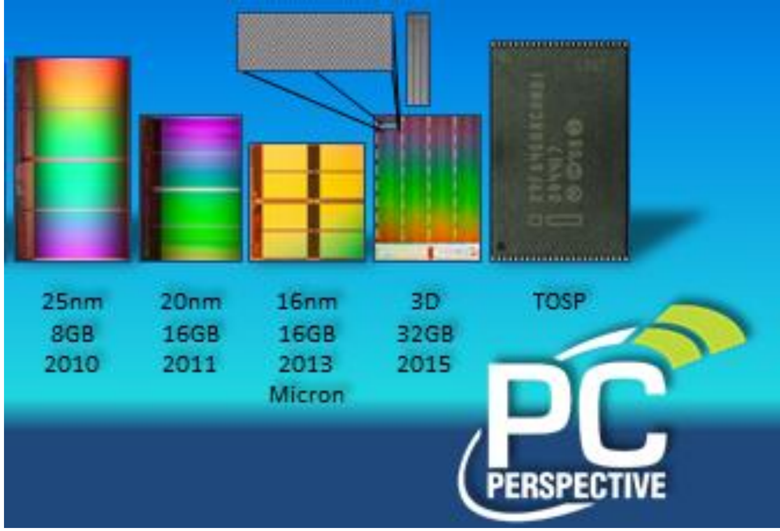
- Intel's Knights Landing supports 36 8Gb/s PCIe 3.0 lanes for aggregate bandwidth of 256Gb/s
  - Single processor could support 2X100GbE in one QSFP
  - 60 cores and over 8B transistors
- IBM's Power8 Processor supports 48 GB/s (384 Gb/s) of PCIe bandwidth
  - Single processor could support 3X100GbE
  - 12 cores and over 4.2B transistors
- Multi-Processor servers could easily support 200GbE for high performance computing
- Many 50GbE (2X25G) servers could transition to 100GbE (2X50G)

Source: http://www.theplatform.net/2015/03/25/more-knights-landing-xeon-phi-secrets-unveiled/

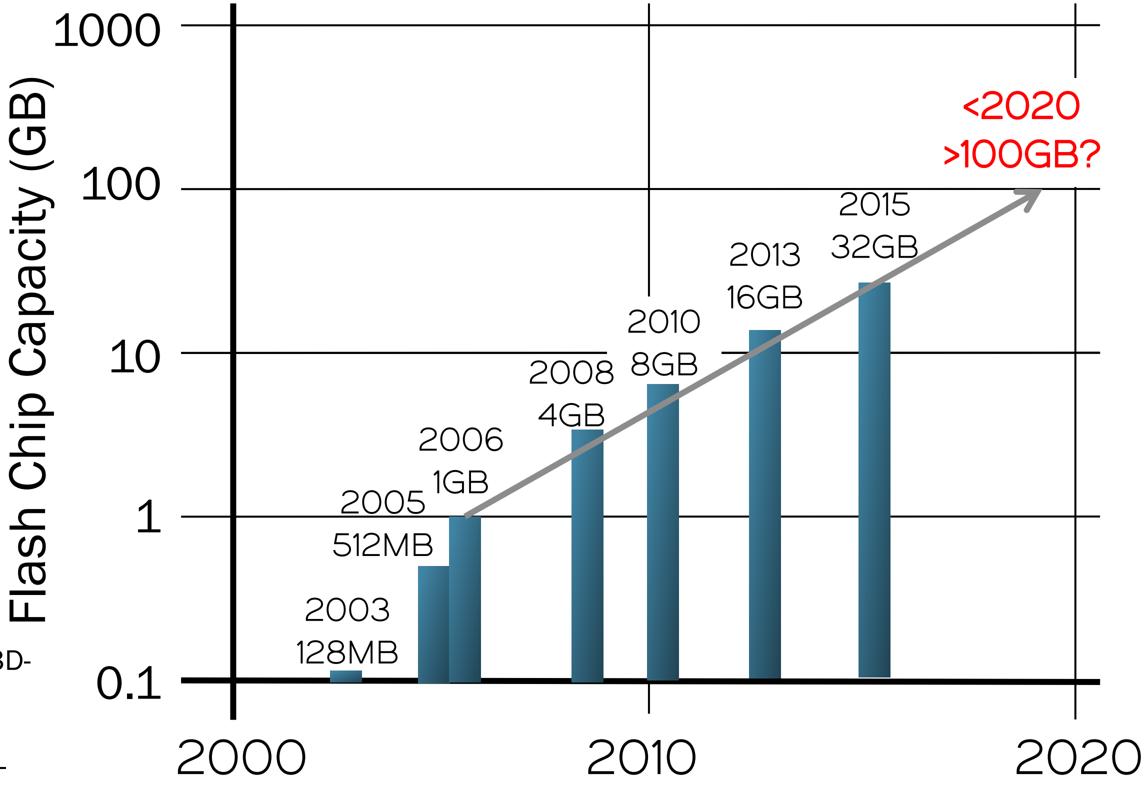
https://en.wikipedia.org/wiki/POWER8

Flash Explosion

Doubling Capacity Every Couple Years



- 3D NAND technology doubling capacity to surpass HDD[1]
- Solid State Disk (SSD) drive capacity to surpass HDD capacity in 2016[2]
  - 16TB in 2016
- Flash can improve application performance by 30%[3]

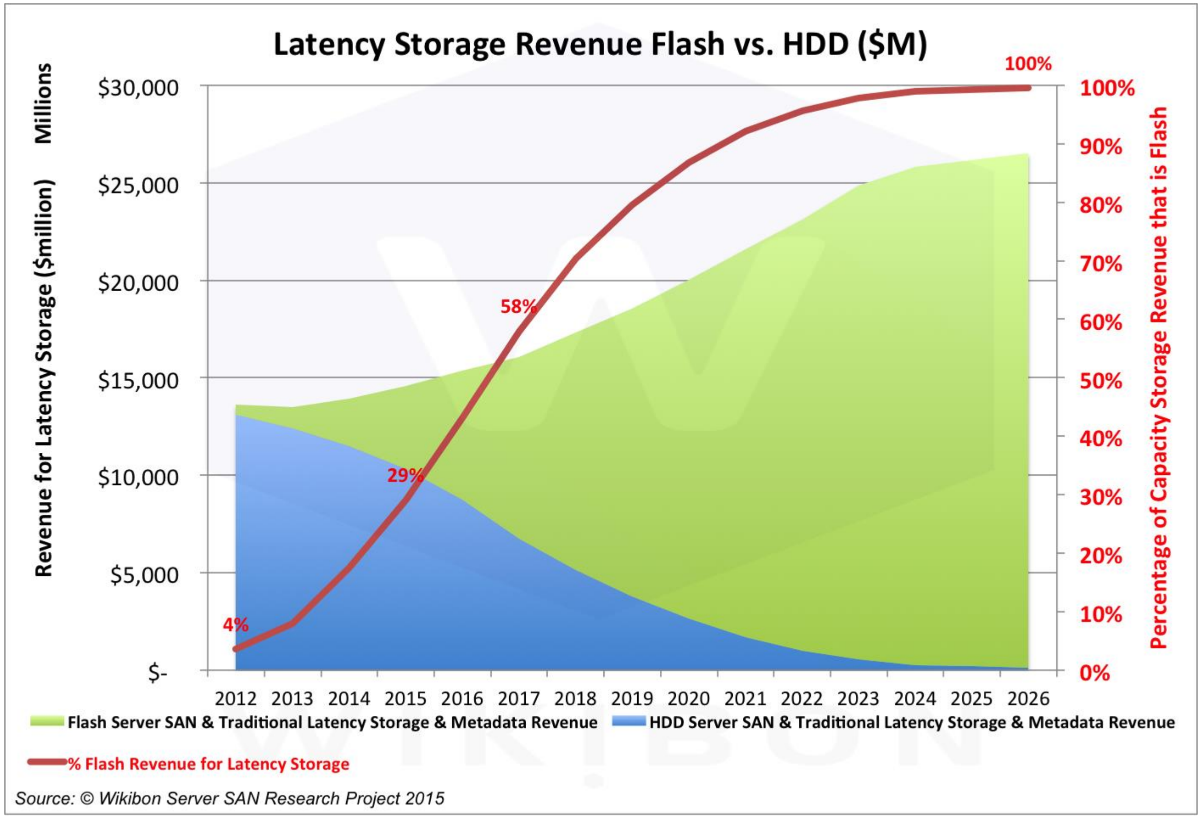


[1] http://www.pcper.com/reviews/Editorial/Intels-new-3D-NAND-will-cost-less-may-offer-effectively-similar-write-speeds
[2] http://www.ign.com/articles/2015/08/19/16tb-ssd-could-be-released-by-samsung-next-year
[3] http://www.computerweekly.com/feature/Optimising-application-performance-with-flash-storage

Flash Replacing HDD

Latency Storage required for high performance

- Single SSD driving up to 2.8GB/s or 22.4Gb/s[1]
- All Flash Arrays (AFA) already driving 8X10GbE and 40GbE links
- See more about Flash in: kipp_CU4HDDsg_01_0915



Source: © Wikibon Server SAN Research Project 2015

[1] http://www.intel.com/content/www/us/en/solid-state-drives/solid-state-drives-dc-p3700-series.html

nX50G PCS and m PMDs

Have you seen Goldilocks?

- 1x50G, 2x50G, and 4x50G are optimal PCS rate choices when targeting use of n x 50G optical-lane PMDs
  - 4X50G is ideal for high throughput links like uplinks and ports for high performance servers and storage
- (4X50G) 200GbE offers the following improvements over (4X25G) 100GbE:
  - Lower cost and power per bit
  - Lower latency from FEC
  - Better performance
- KRn, CRn and SRn PMDs are best choices for a quick project while meeting most data center reach requirements

BROCADE

THANK YOU