

# An Evolutionary Path to B400G PMDs

**Ali Ghiasi**

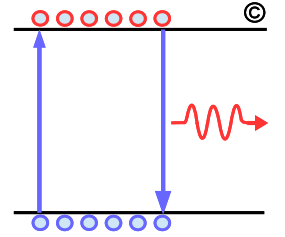
**Ghiasi Quantum LLC**

**B400G Study Group**

**Virtual Meeting**

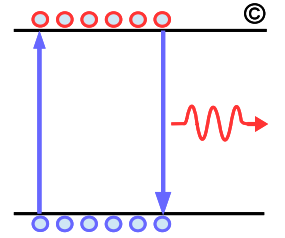
**March 22, 2021**

# Overview



- ❑ **B400G rates and PMDs**
- ❑ **Facebook network architecture and evolution**
  - Higher radix flatter network
  - Operating in breakout mode
- ❑ **Switch evolution and ASIC BW limitation**
- ❑ **Facebook optics evolution**
- ❑ **Some thoughts on FEC options**
  - Monster end-end FEC
  - Inner product optics FEC
  - Segmented FEC
- ❑ **Rome was Not built overnight**
- ❑ **B400G PMD sets timeline.**

# B400G Rates



## ❑ 800 GbE/800G – paired 51.2 Tb switches

- Implementing 800 GbE/800G relatively speaking is free when coupled with CK 100G I/O, QSFP-DD800/OSFP, and Ethernet Technology 800G PCS
- The primary application of 800 GbE/800G are:
  - 8x100 GbE – fabrics and servers
  - 4x200 GbE – fabrics and servers
  - 2x400 GbE – fabrics, Routers, and DCI (400ZR)
  - 1x800 GbE – Routers and DCI (800ZR)

## ❑ 1600 GbE/1600G – paired 102.4 Tb switches

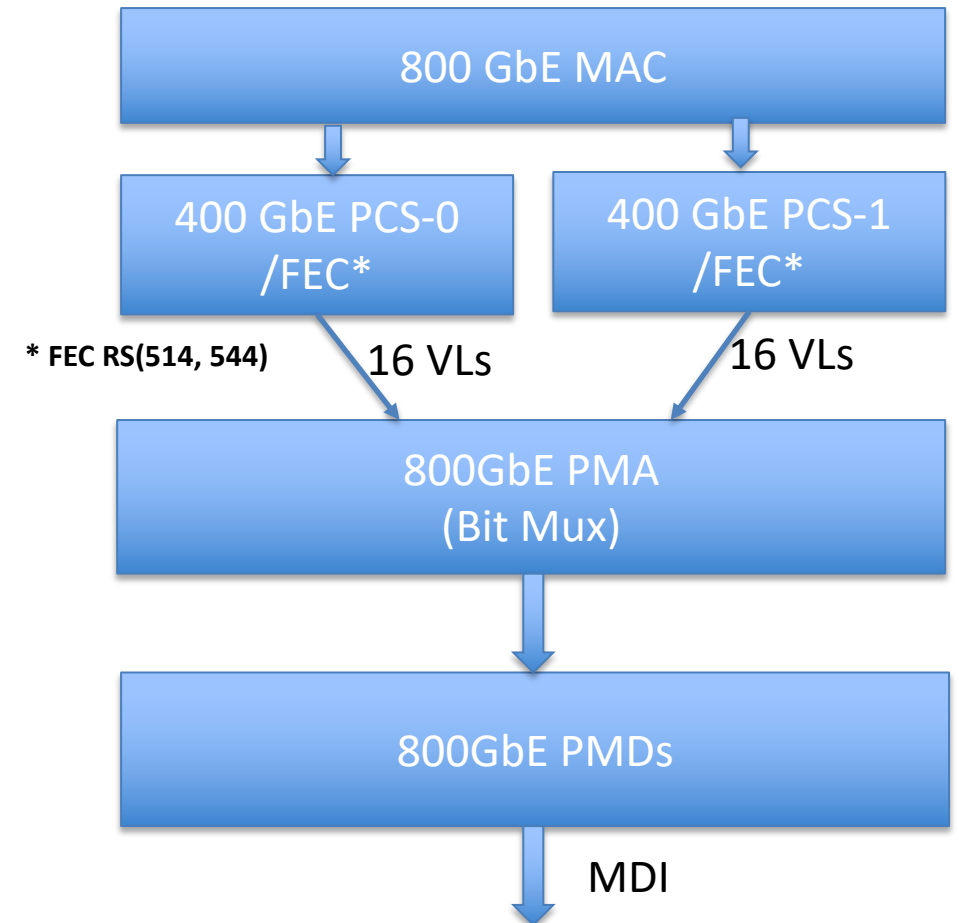
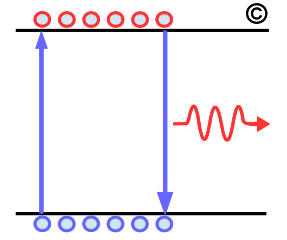
- Will require 200G/lane electrical or co-packaged optics where neither one is ready for prime time and not to mention 102 Tb switches
- The primary application of 1600 GbE/100G are:
  - 8x200 GbE – fabrics and servers
  - 4x400 GbE – fabrics and servers
  - 2x800 GbE – fabrics, Routers, and DCI (2x800ZR)
  - 1x1600 GbE – Routers ?, 1600ZR?? (may not offer increased fiber capacity and shorter reach)
- For this generation it is even less clear what are the real 1600 GbE applications
- If 1600 GbE PCS/FEC is free for 102.4 Tb switches then likely 1600 GbE support will be included even if there is no broad market potential!

# B400G Rates, Cont.

## ❑ The decision regarding 800 GbE MAC-PCS/FEC is urgent given some of the potential ASICs in flight

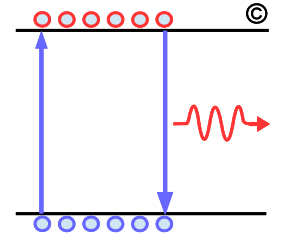
- We either need to adopt Ethernet Technology Consortium (ETC) proposed 800 GbE and if we define a new 800 GbE PCS/FEC can be disruptive to product in flight
  - PCS/FEC based on clause 119 but with unique identifier to indicate PCS-0/1
- ETC 800 GbE implementation is based on
  - Dual 400 GbE instance of PCS/FEC
  - With 32 virtual lanes instead of 16
  - With additional set of markers to allow interleaving odd/evens codewords
- ETC 800 GbE implementation is practically free

## ❑ 800 GbE must be part of fast-tracked task force!



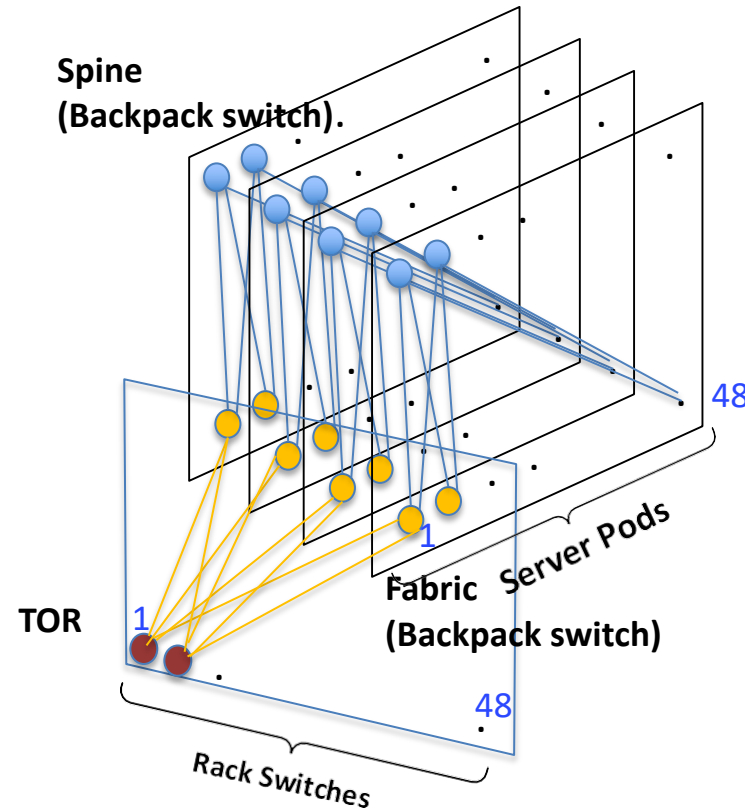
[https://ethernettechnologyconsortium.org/wp-content/uploads/2020/03/800G-Specification\\_r1.0.pdf](https://ethernettechnologyconsortium.org/wp-content/uploads/2020/03/800G-Specification_r1.0.pdf)

# Facebook F4 Architecture

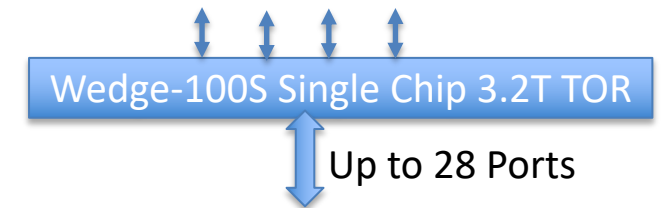
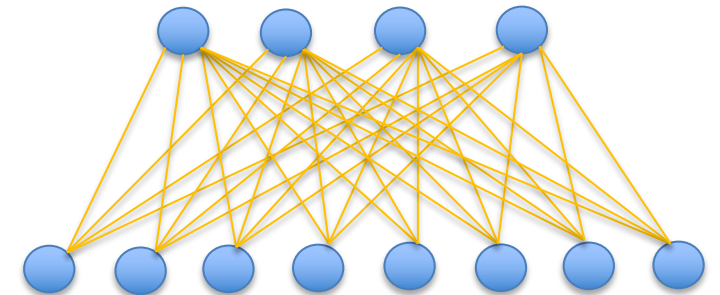


## Highlight of F4 network

- 12x3.2T switches to build 12.8T (128x100 GbE) fabric/spine switch
- Fabric/spine links operating at 100 GbE
- Supports up to 48 racks switches/Pods with up to 48 Pods
- Each fabric switch connects to 48 rack switches and 48 spine switches
- Each rack switch within a pod connects to each of 4 fabric switches
- Rack switches are heavily oversubscribed with 4 uplinks with as many as 28 downlink/server connections.

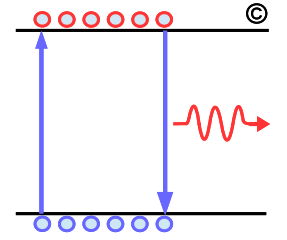


Backpack Fabric Switch 128x100 GbE  
(3 stage Clos based on 12x3.2T Switches)



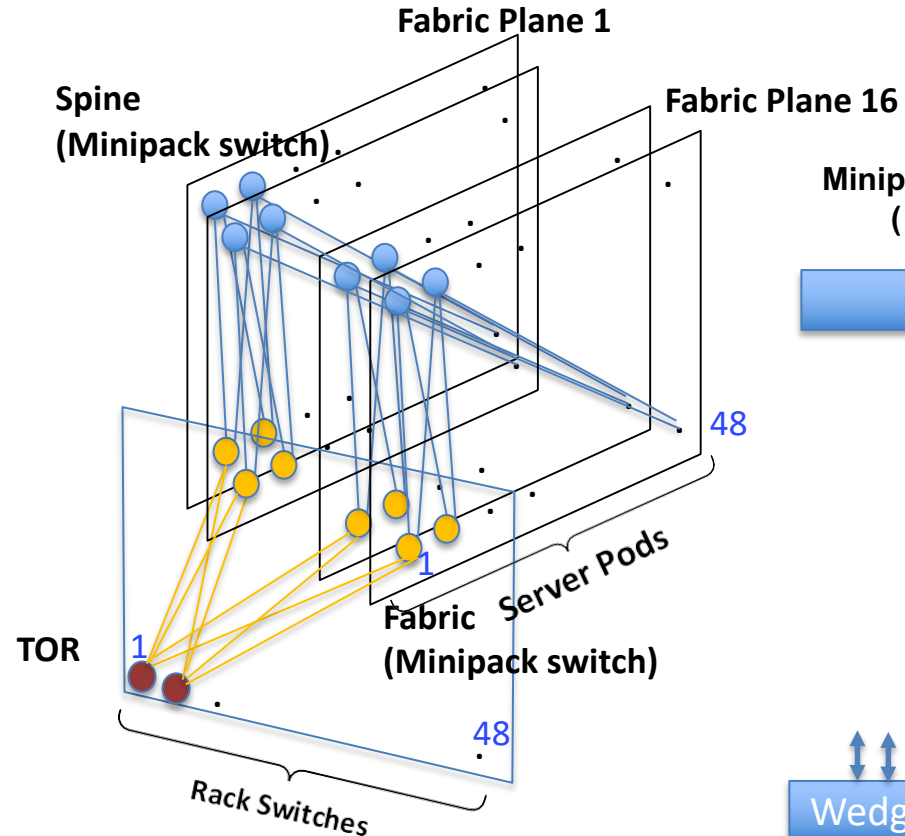
Source: Alexey Andreyev OCP Summit 2019

# Facebook F16 an Improved DC Network

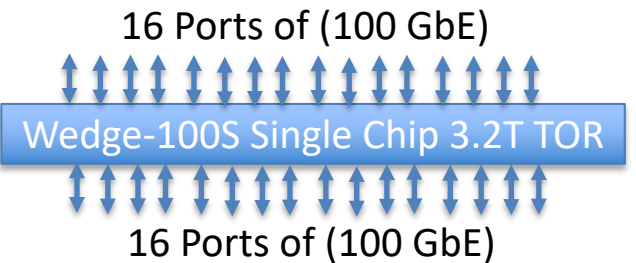
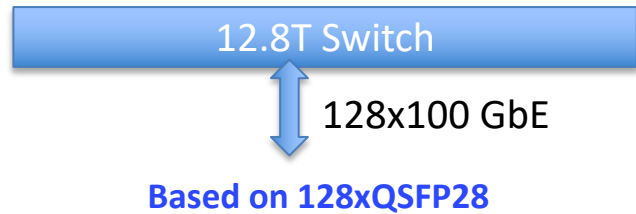


## Facebook F4 to F16 is an excellent example of improving network efficiency and fabric BW leveraging 4x increase of switch ASIC BW without requiring faster MAC rate

- Fabric/spine links still operate at 100 GbE
- 4x improvement in fabric BW but significantly greater efficiency improvement
  - 4x the number of spine planes
  - Each spine switch is a single ASIC instead of 12 ASICs stage Clos
  - F16 is 3-layers network where F4 was 5 layers
- Minipack fabric operating with 48/48 up/down links with just 1.33 oversubscriptions
- Wedge-100S TOR operates without oversubscription where previously operating ~7 to 1
  - Network without oversubscription may operate in cut-through mode instead of store-forward which increases latency.

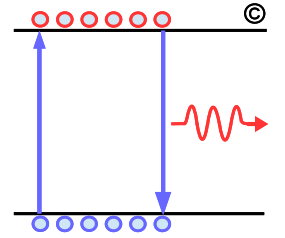


Minipack Fabric Switch 128x100 GbE  
(Single ASIC 12.8T Switch)



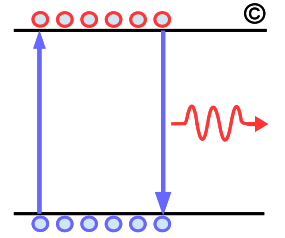
Source: Alexey Andreyev OCP Summit 2019

# 400G Ports Today Primarily Operate as 100/200 GbE



- ❑ **F16 architectural will allow 4x BW improvement with availability of 51.2T switches but now only will require 400 GbE optics**
- ❑ **51.2T switches will have 800 GbE MAC and possibly 1600 GbE MAC as long as adding higher speed MAC/PCS/FEC are free**
  - If 1600 GbE has significant overhead “no longer free” then 1600 GbE MAC rate will be delayed till broad market is developed.
- ❑ **During 802.3bs and during B400G CFI the topics of breakout were discussed but we were not sure how the breakout market will be**
  - As Lightcounting and other firm starting to breakdown 100 GbE, 200 GbE, vs 400 GbE use case when possible
  - Implementing 400G ports and supporting 50/100/200/400 GbE has been a great success story
  - It was overall architecture of 802.3bs/cd that created an eco-system which allows implementing 200/400 GbE with no or little cost adder where most applications are breakout
- ❑ **B400G should build on 802.3bs architecture to allow seamless support of breakout applications, such 400/200/100 GbE.**

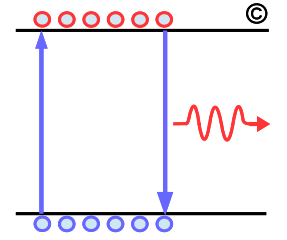
# What can we Learn from F4 to F16 Evolution



- ❑ Higher radix switches allow building flatter 2-layers or 3-layers networks which are more efficient, lower latency, lower power, and are lower cost
- ❑ Facebook could have upgraded F4 network by maintaining the 4 fabric planes but all fabric links would have been 400 GbE in 12.8 Tb switch generation
  - Would have resulted in 3x increase in # of ASICs
  - Would have required higher cost 400 GbE optics
- ❑ As important of 4x fabric BW increase are these additional enhancements to the DCN:
  - Elimination of 2 switch layers are as significant as increasing fabric BW by 4x
    - In F4 traffic from NIC-NIC goes through 11 switch ASICs instead of just 5 switches ASICs in case of F16
    - If one assumes latency increased linearly, 11 switches with  $\sim 750$  ns/ASIC would results in  $8.25 \mu\text{s}$  but in practice latency increases exponentially with # of switch layers
    - [Arista](#) reports 2-2.2  $\mu\text{s}$  latency for 2 layers when all ports operating in 100 GbE mode with cut-through but 3-4  $\mu\text{s}$  if operated in mixed 25GbE/100GbE modes that forces switch into store-forward
    - Facebook F4 likely operated with 50+  $\mu\text{s}$  and F16 likely is operating with under 5  $\mu\text{s}$
  - Operating TOR with 1:1 with potential cut-through operation instead of 7:1 over-subscriptions also is significant as increasing the fabric BW by 4x.



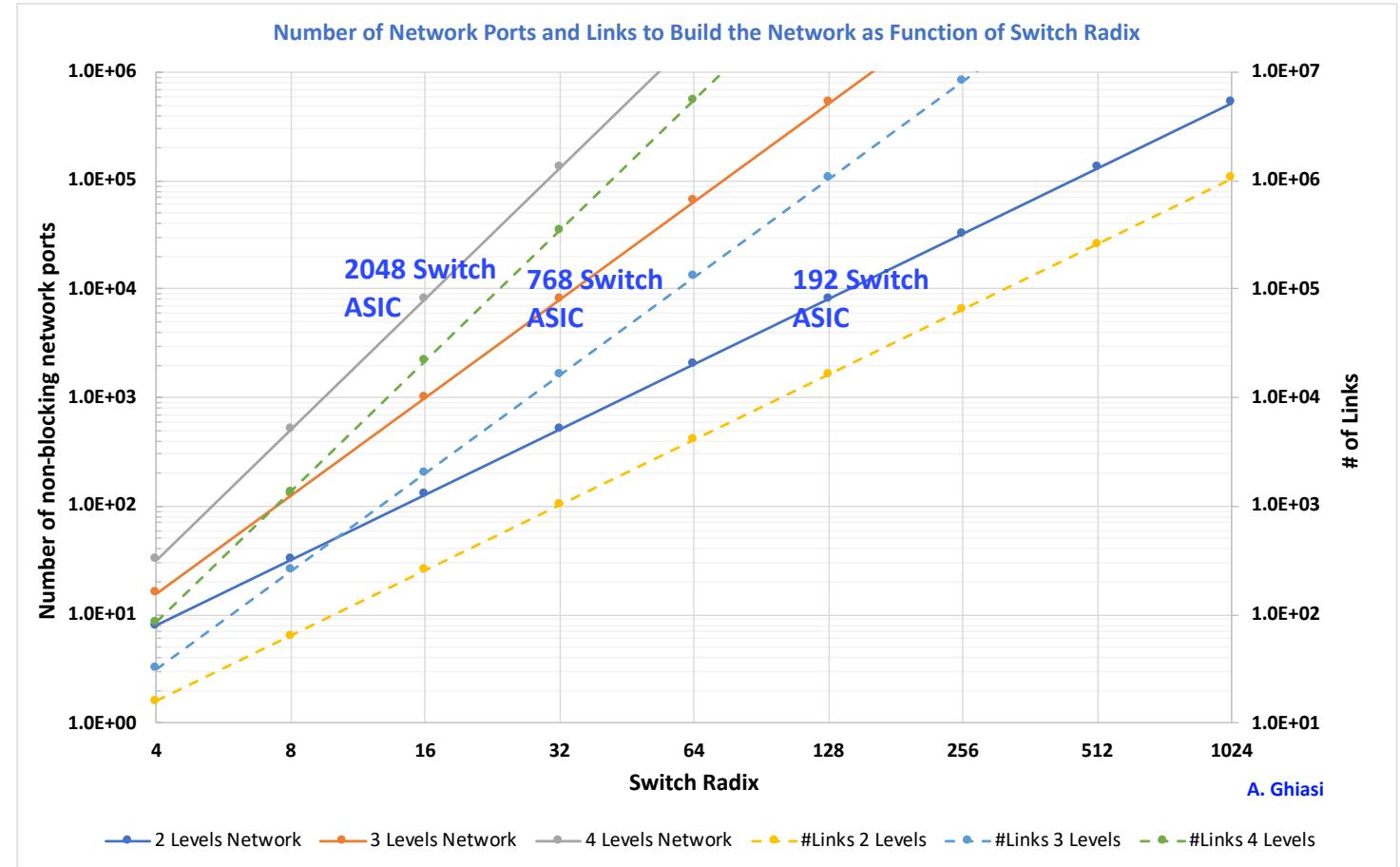
# Higher Radix Fabric Implementation Require More Interconnections but with Fewer ASICs



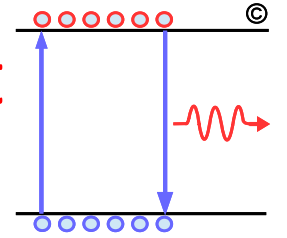
Higher radix fabric does increase # of inter-switch links and destinations but require fewer switch ASICs to build the same non-blocking network

- But increase # of links is not a true cost as often one 400GBASE-DR4 is operating as 4x100GBASE-DR link
- The only cost is the increase complexity of fiber plant due to breakout to additional end points

Higher radix networks not only improve performance but also power and cost are reduced.

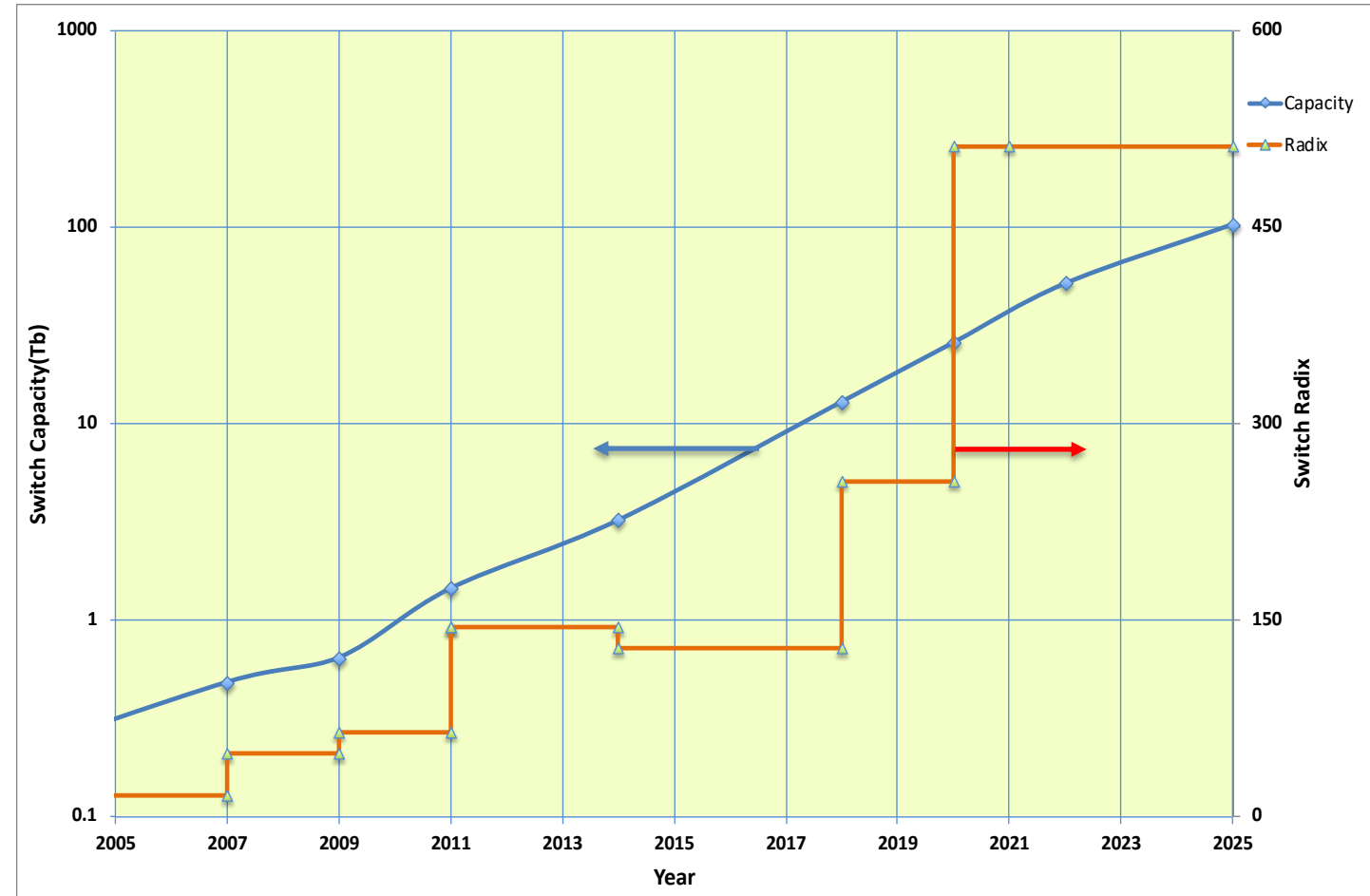


# Increased Switch Radix Redefining How Datacenter are Built

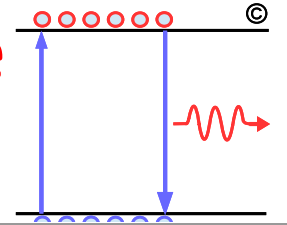


## □ The 256 and now 512 radix switches are redefining how datacenters are built

- The 256 or 512 radix switches allow eliminating a switch layer and still build same capacity fabric
- Fabrics are built with 4x100GbE instead needing 400 GbE optics now
- 3-4 layers DC fabrics are replaced with 2-3 layers without over-subscription and lower latency!

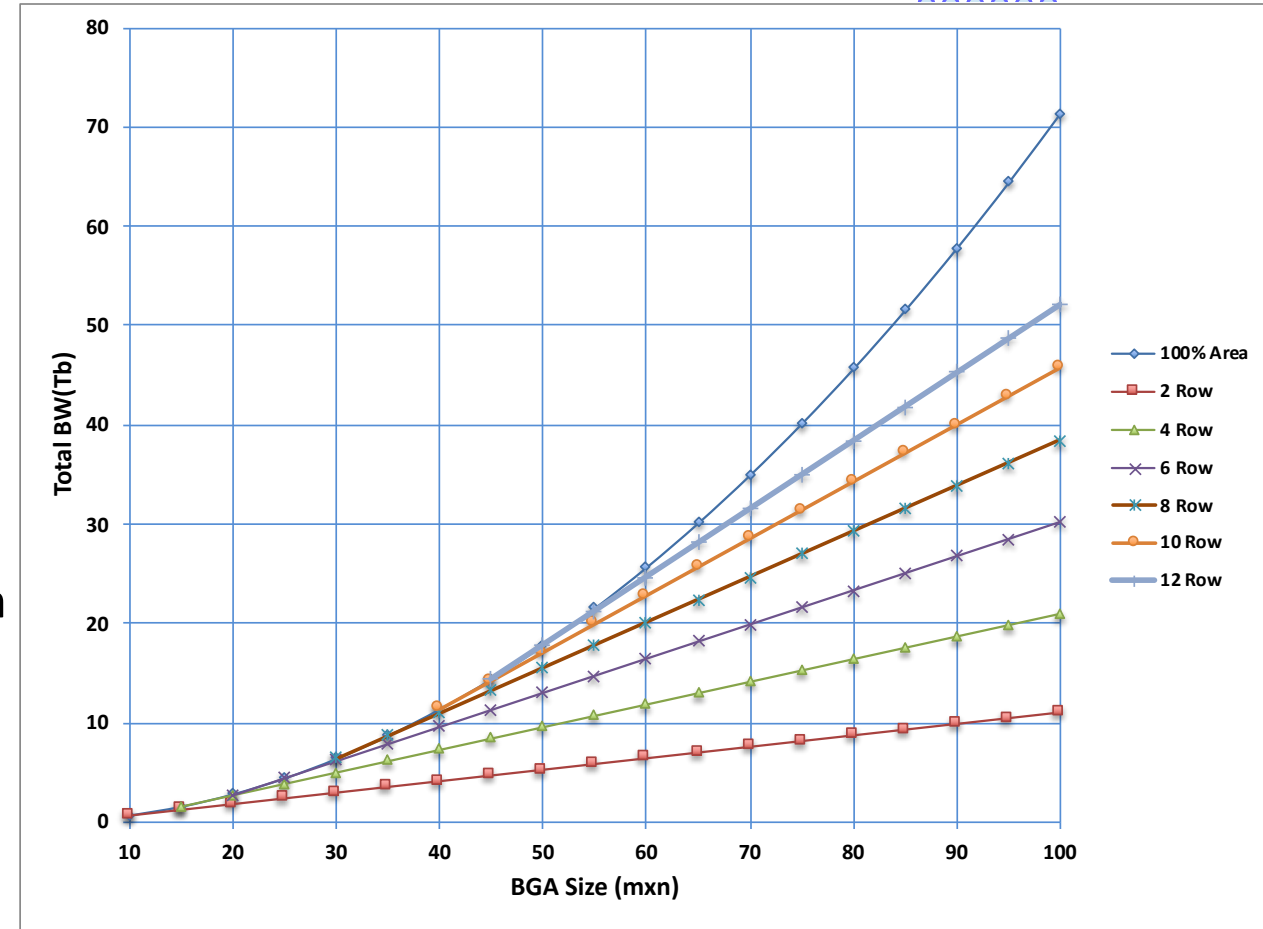
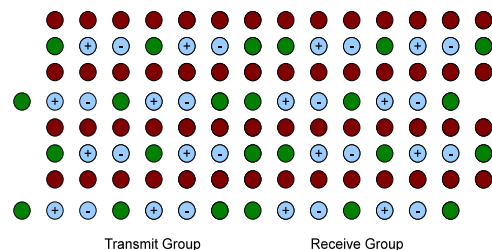


# 51.2+ Tb Systems Would Require Co-Packaged or 200G/Lane



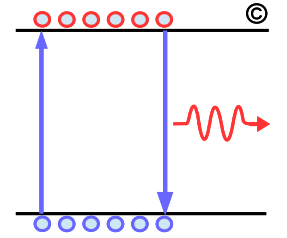
## 512 lanes SerDes which today already pushes the packaging boundaries

- Assumed ball map shown below should support 100G/200G PAM signaling
  - Aggregate BW listed is based on 100G SerDes
  - Some implementations may use more aggressive ball map reducing package size and and compromising on crosstalk
- Up to 512x50G switches exist today
- Ball map below double stacked (or rotated ) with 8 rows supports 25.6Tb in 70x70 BGA
- For 102.4 Tb generation co-packaged optics or 200Gx512 lanes would be required.



See A. Ghiasi <https://www.osapublishing.org/oe/fulltext.cfm?uri=oe-23-3-2085&id=310831>

# Different Possible Trajectories Facebook Could have Taken



□ Increased switch radix coupled to increased # of DC fabric planes pushes out the need for higher MAC rates!

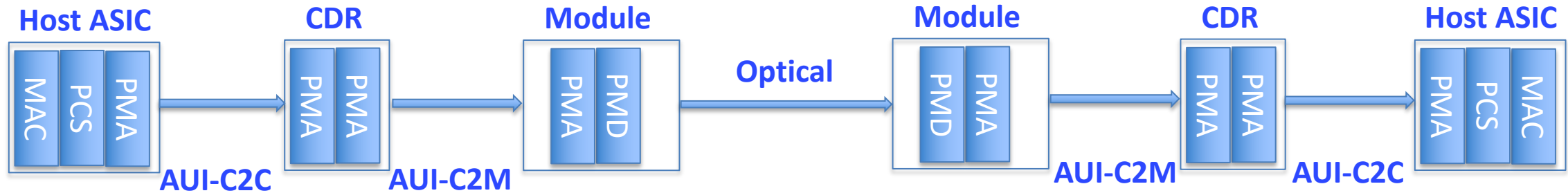
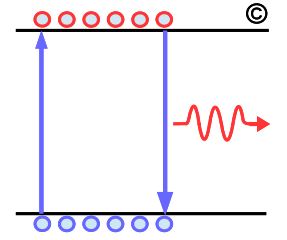
Platform Name	FB Initial Deployment Year / Status	Port Speed (Gb/s)	Electrical Lane Speed (Gb/s)	Switch Silicon Bandwidth (Tb/s)	Switch System Configuration (Radix)	Major Optical PMD	Optical Lane Speed (Gb/s/λ)	Optical Module Type			Optics for DC Topology	
								Front Panel Pluggable Optics	On Board Optics	Co-packaged Optics	F4	F16
OEM?	2016	40	10	1.28	128 x 40GbE	40GBASE-LR4	10	QSFP+	-	-	40 GbE	
Backpack Minipack	2017	100	25	3.2	128x100GbE	100G-CWDM4	25	QSFP-28			100 GbE	25 GbE
	2018	100	50	12.8	128 x 100 GbE	100G-CWDM4 (OCP)	25	QSFP-28	Mini-Photon	-	400 GbE	100 GbE
	2021	200	50	25.6	128 x 200 GbE	200G-FR4	50	QSFP-56	Next Gen OBO	-	800 GbE	200 GbE
	Planning – 2023	400	100	51.2	128 x 400 GbE	400G-FR4	100	TBD	Next Gen OBO	CPO Gen 1 (XSR)	1600 GbE	400 GbE
	Exploration	800	200	102.4	128 x 800 GbE	800G-FR4	200	-	-	CPO Gen 2	3200 GbE	800 GbE
	Exploration	1600	??	204.8	128 x 1600 GbE	?					6400 GbE	1600 GbE

Table from [https://www.ieee802.org/3/B400G/public/21\\_03/stone\\_b400g\\_01\\_210301.pdf](https://www.ieee802.org/3/B400G/public/21_03/stone_b400g_01_210301.pdf) but with additional annotations by the author.

- Goal: Preserve switch radix gen over gen while scaling port bandwidth
- Re-use existing fiber, power, cooling and physical infrastructure to enable “rolling upgrade” with minimal disruption

# 802.3bs FEC Architecture

- 802.3bs contribution from [Anslow](#) supports 4 AUI sub-links as shown below by stealing 0.1 dBo of optical budget to allow operation with one end-end FEC.

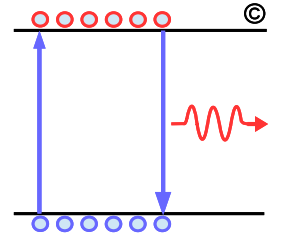


	RS(544,514) FLR = 6.2E-11			
	Electrical		Optical	
1:2 Same FEC, a = 0.75 worst skew	Burst	1.4E-6*	Random	2.4E-4
1:2 Same FEC, a = 0.75 worst skew	Burst	2.9E-6*	Random	2E-4
a = 0.75 misaligned	Burst	5.2E-6*	Random	2.4E-4
Random errors	Random	8.2E-5	Random	2.4E-4

Note – these values are the BER **including** the additional errors due to the bursts. To account for burst errors, the values marked with “\*” have been multiplied by 4 when a = 0.75.

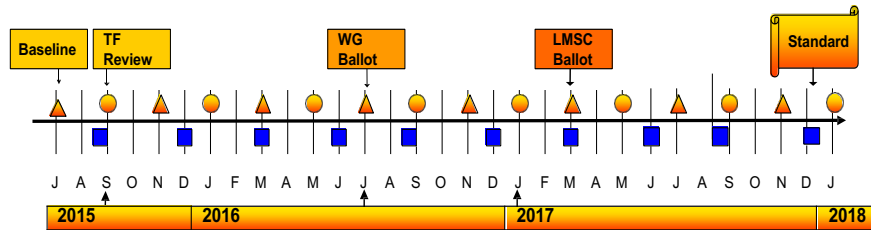
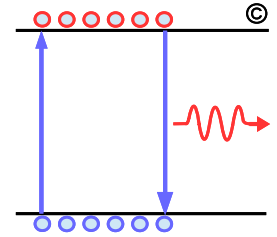


# How to Define 200G/lane Optical PMDs Prior to 200G/lane AUI

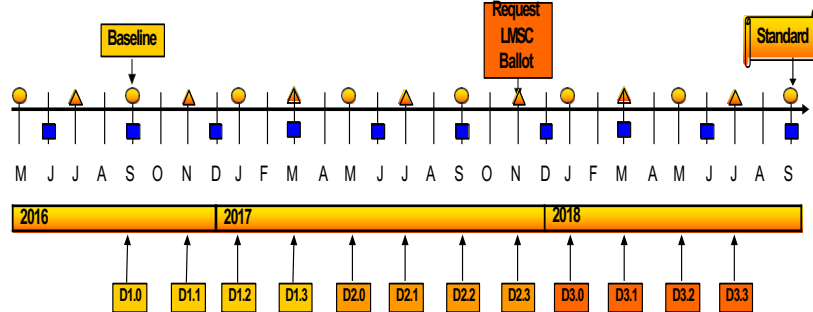


- ❑ **802.3bs successfully defined an architecture that operated with an end-end FEC by allocating 0.1 dBo to 4 AUI sub-links and without defining 100G-AUI**
  - With emergence of co-packaging there may not even exist an AUI link or depending on the implementation there may not even exist an XSR link
- ❑ **200G/lane AUI is substantially more complex due higher loss, ILD, and reflections and would full FEC gain applied to the AUI segment where allocating 0.1 dBo of end-end is no longer sufficient**
  - 200G/lane AUI could be based on PAM4 or based on PAM6/DSQ-32, [Ghiasi](#), depending on connector/channel improvements
  - Introducing a big monster end-end FEC that is costly to integrate into host ASIC in order to support 200G-AUI that may or may not even exist "unnecessary tax"
  - Introducing inner product FEC codes for optics also will tax co-packaging implementations
  - If the 200G-AUI or 200G-XSR improve sufficiently over time to operate with 0.1 dBo borrowed from end-end will allow overtime to eliminated the segmented FEC
    - But stealing anything >0.1 dBo will be too costly for the optics
  - If AUI don't improve sufficiently over time then segmented FEC is the best option without taxing everyone's!

# Rome was not Built Overnight



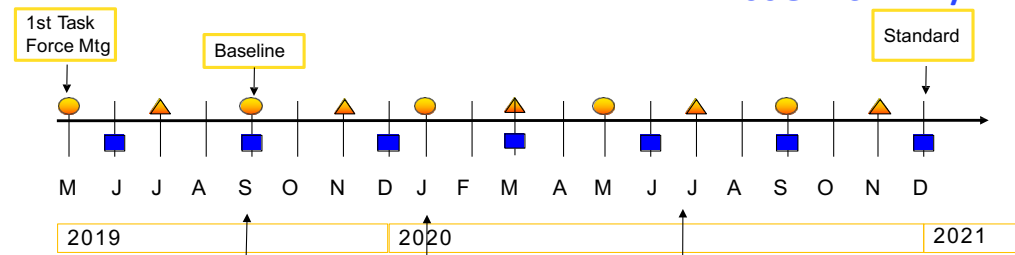
802.3bs Defined 200GbE/400GbE  
Defined only 50G-AUI, 200GBASE-FR4/LR4,  
400GBASE-FR8/LR8, and 400GBASE-DR4



802.3cd Defined lower speed PMDs and  
50G/lane CR/KR, 100GBASE-DR1

802.3cd defines 100GBASE-FR/LR and  
400GBASE-FR4/LR4-6

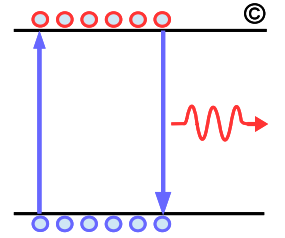
Remarks: Not all drafts Dx.y maybe listed.  
Not shown are 802.3cm PMDs 50G/lane MMF May-2018-  
Dec. 2019, 802.3cn 40 km PMDs July 2017-July 2020.  
It will take 7 years to complete the 200/400 GbE PMD set  
200GbE/400GbE optical PMDs will be shipping for 4+  
years before 100G/lane AUI availability!



802.3ck defines 100G/lane AUI/CR/KR



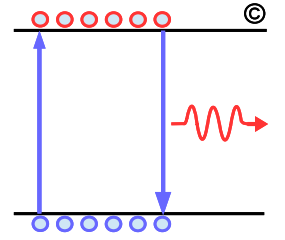
# B400G PMDs Sets



- ❑ **1<sup>st</sup> wave – taskforce start ~July 2021**
  - Consider adopting 800G Ethernet Tech. Con. MAC/PCS
  - Define 800G-DR4, 800G-SR8, 800G-FR8
  - 800G-AUI8, 800G-CR8/KR8
- ❑ **2<sup>nd</sup> wave – taskforce starts ~ Nov 2021**
  - 200G/lane SMF optics PMDs
  - 800G-ZR
  - 1600 GbE MAC/PCS
- ❑ **3<sup>rd</sup> wave – taskforce starts ~ July 2022 (after compilation of 802.3ck)**
  - 200G-AUI/C2C (let the MSA continue improving the connector as OIF investigates)
  - Other optical PMDs including more efficient MMF PMDs.



# Summary



- ❑ **To complete 802.3bs PMDs set will take ~7 years but B400G PMDs expected to be more complex**
  - Logical layers and optical PMDs require less time to get to complete draft than the CR/AUI PMDs
  - In every higher speed project starting with 10 GbE, 100GbE, and 200/400 GbE optical PMDs were defined first before associated AUI's were defined
- ❑ **Today's data center are built more efficient and flatter with less need to push the bitrate on individual links but instead switches with higher radix with increasing # of lower speed links deployed "breakout mode"**
  - For example 400G modules deployed today primarily operates as either 8x50GbE, 4x100 GbE, or 2x200 GbE and the exception are some of the longer reaches or 400-ZR
  - As long as implementing 800GbE/1600 GbE cost is about free then Ethernet eco-system will implement higher MAC rate where real use will be 8x100 GbE, 4x200 GbE, or 2x400GbE
- ❑ **The 200G-AUI links will be as complex as 200G-KR links and may require even more FEC gain than 100GBASE-KR**
  - 200G-AUI is an important development but channel/connector need further study/improvements to get to <math>< pJ/bit</math>
  - Given that 200-AUI may not even exist with emergence of co-packaging let's not tax everyone with a monster end-end FEC in order to support AUI sub-links or force an inner product FEC on the optics taxing co-packaged optics
  - Optical PMDs should be defined such that 0.1 dBo is reserved for XSR/AUI sub-links
  - Given the constraints segmented FEC is the best option for B400G but overtime XSR/AUI may operate with end-end FEC
- ❑ **B400G PMD set likely will span across 7+ years and we should only define these new objectives/PMDs when there is technical feasibility and broad market potential!**