Data Center architecture trends and their impact on PMD requirements

Mark Nowell, Matt Traverso – Cisco

Kapil Shrikhande – Dell

IEEE 802.3 NG100GE Optics Study Group

March 2012

Supporters

- Scott Kipp – Brocade
- David Warren – Hewlett-Packard
- Gary Nicholl – Cisco
- Jeff Maki – Juniper
- Pete Anslow – Ciena
- Shimon Muller - Oracle

Overview

- How Data Center architectures are changing and how that impacts technology requirements
- Implications for NG100G Optics Study group
- Recommendations

Not All Data Centers Are The Same

Enterprise Data Centers

- Lower port counts (Than MSDC)
- Network provides workload mobility
- L2 - L3 forwarding agnostic

MSDC

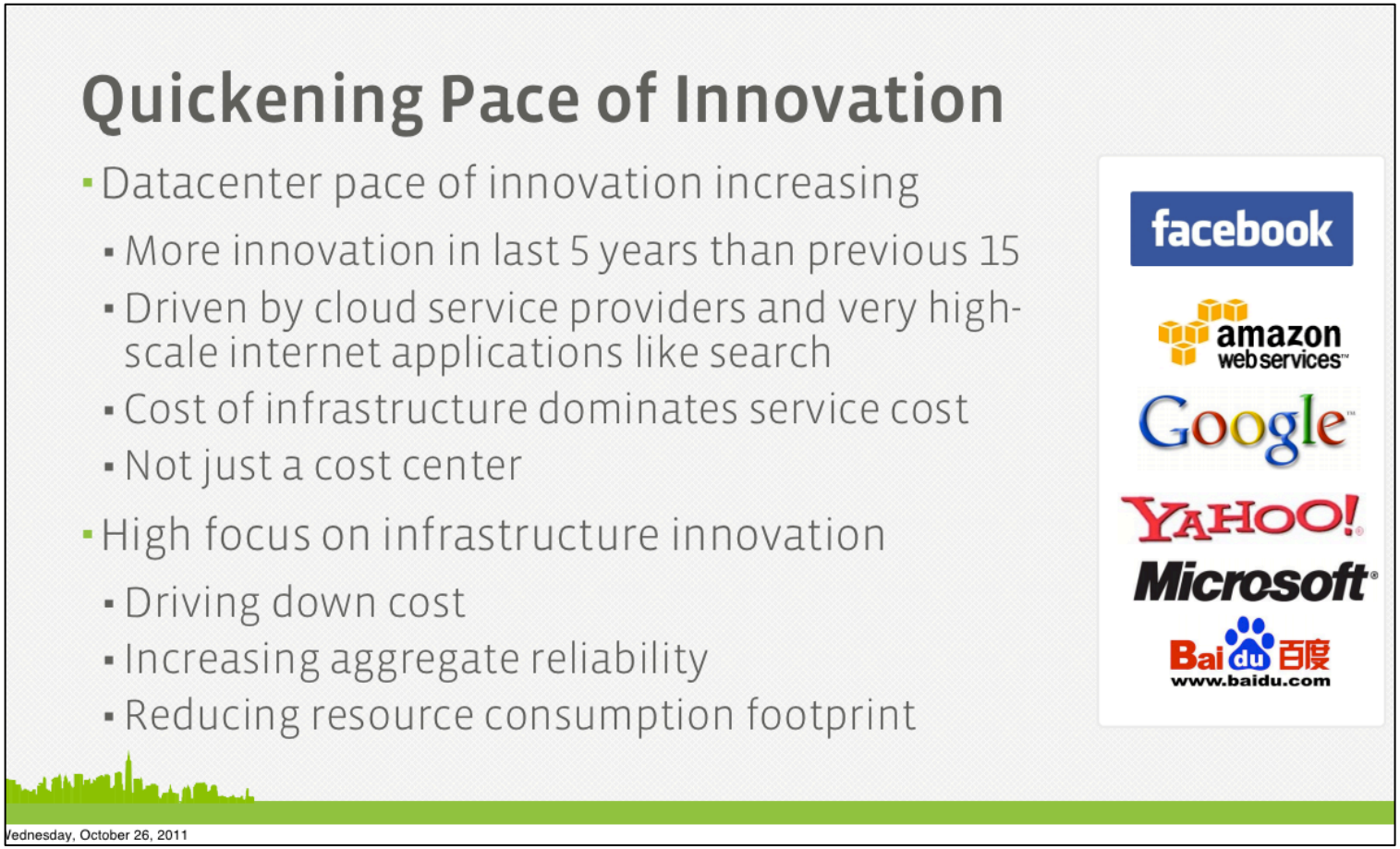
- High port count :30K-100K
- Single tenant
- Server virtualization not used or hidden from network
- L3 forwarding only
- POD sub-unit

Clouds

- 30K-100K ports
- Variants:
  - Multi-tenancy
  - Hundreds of thousands of tenants
- Workload and VM mobility
- L3 forwarding only or L2-L3 forwarding agnostic

- Architectures are similar
- Fastest growing market

What the DC architects are saying:



Source: James Hamilton, Amazon. Internet Scale Infrastructure Innovation, Open Compute Summit 2011
http://mvdirona.com/jrh/TalksAndPapers/JamesHamiltonOCP%20SummitFinal.pdf

What the DC architects are saying:

Source: James Hamilton, Amazon. Internet Scale Infrastructure Innovation, Open Compute Summit 2011
http://mvdirona.com/jrh/TalksAndPapers/JamesHamiltonOCP%20SummitFinal.pdf

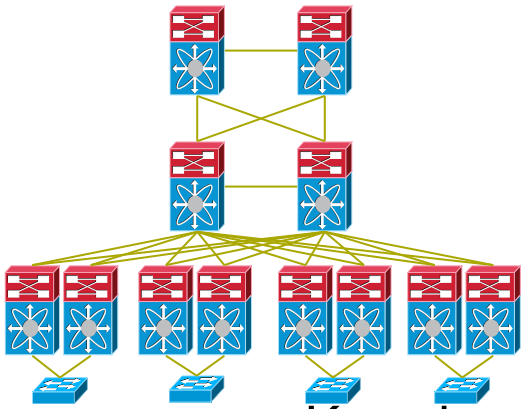
What's Driving the Evolution of DC Environments – Customer Perspective

- Need to achieve higher scalability
- Need for better high availability and lower fate sharing
- Need to accommodate diverse workloads concurrently
- Need flexibility on workload mobility
- Need to further simplify operational models
- Need for lower and or predictable latency / response time
- Need physical facilities to evolve with technology
- Need lower cost connectivity to support large environments and trends in traffic, bandwidth and speed

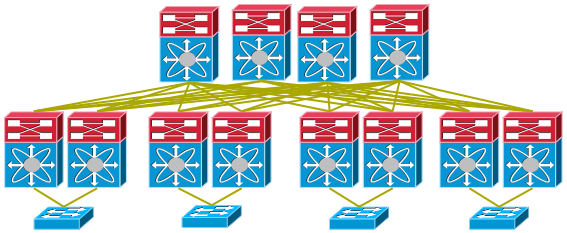
Data Center Market Transition
Core and Aggregation

Past → CY2010+

3 Tier Architecture
Oversubscription between Access Aggregation and Core



Non-blocking Data Center Fabric
Oversubscription only in Access



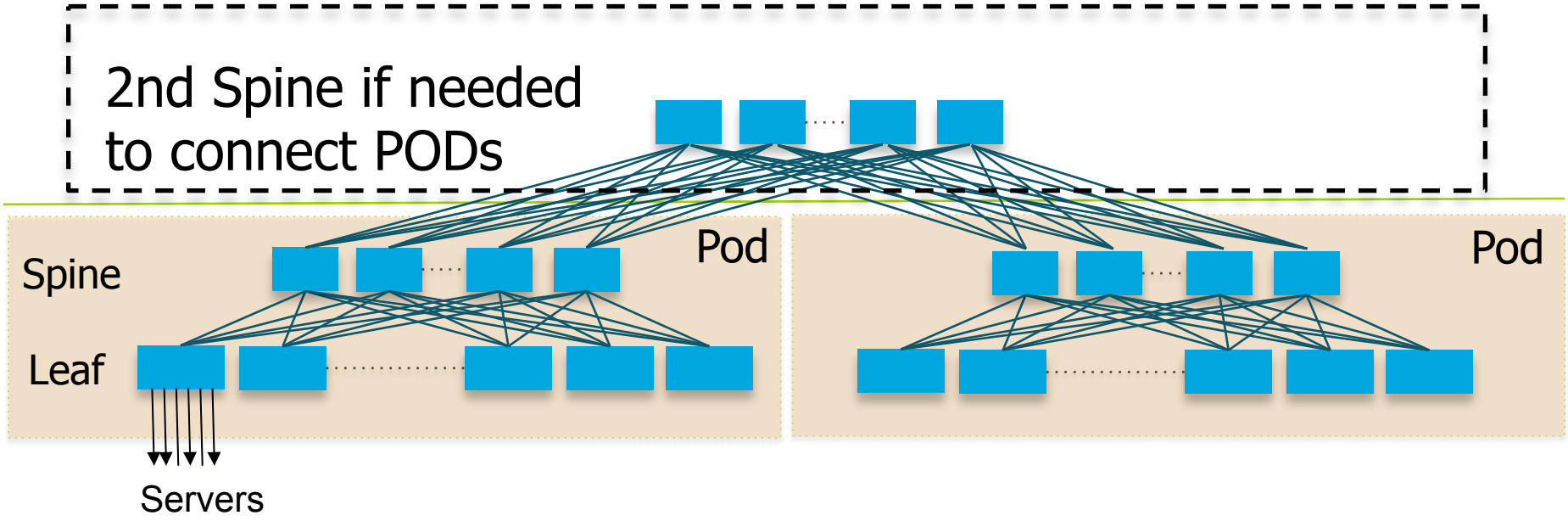
Large MSDC doing this on new network builds

Key changes at the connectivity layer:

- Tree-based architecture optimized for N-S traffic
- Over-subscription in access, aggregation and core
- Lower ports counts in Agg-Core

- Meshed architecture better suited for N-S and E-W traffic
- Over-subscription only in access, 1:1 in aggregation and core
- **Higher ports counts in Agg-Core!**

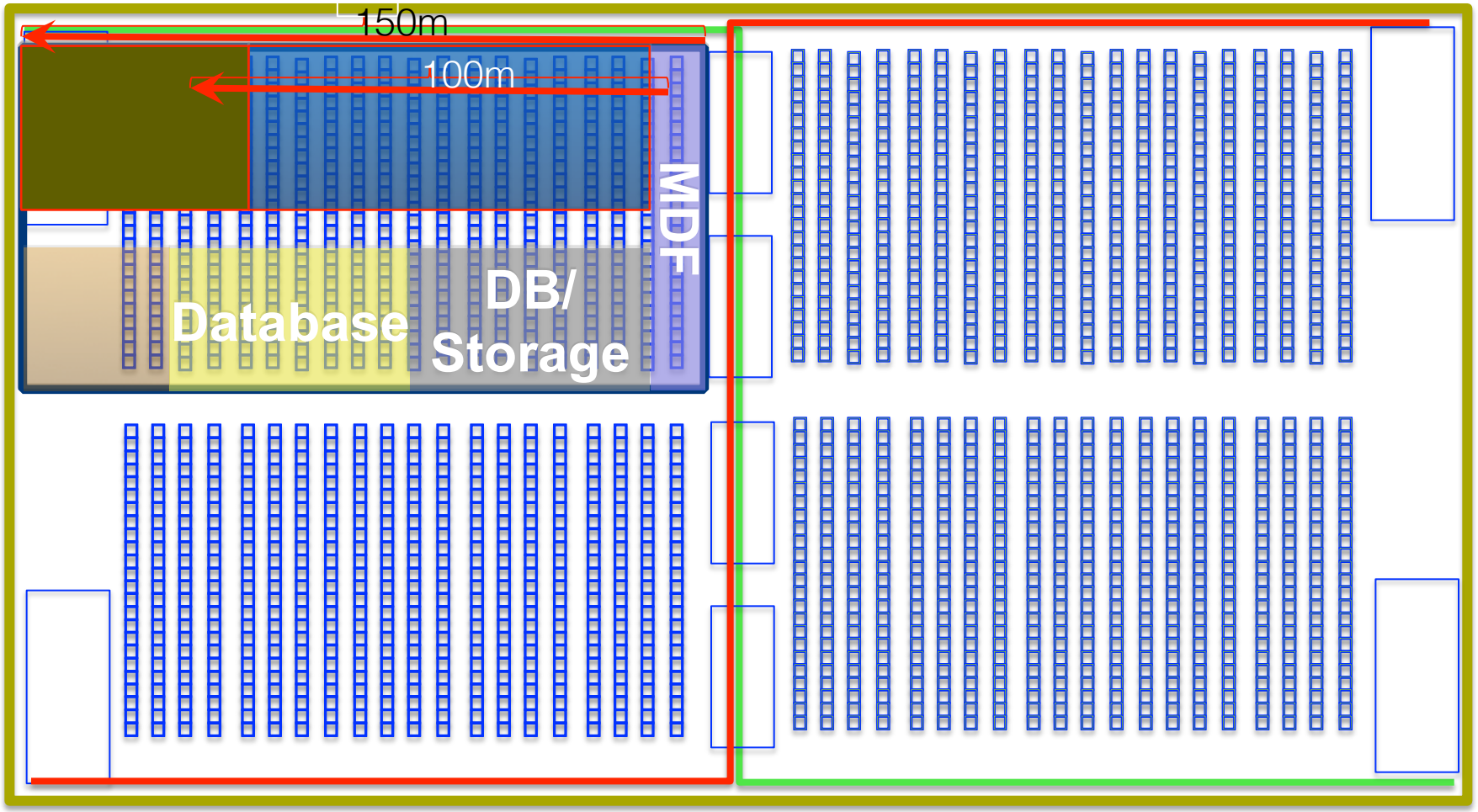
DC Fabric Concepts: Leaf, Spine & Pod



- Pod: East-West communication is equidistant across a 2-tier topology
- HA Model: N+1 on spine and paths vs 1+1 on classic model
- All switches within a tier provide equal port density

- Pod's max density: # of spine switches x switch port density
- Max # spine switches: ½ the port density of a leaf switch
- Larger than single pod capacity: requires an additional tier

Scaling the network
Mapping Topology to Infrastructure



Schematic representation only

- Fiber runs up to 2 km corner to corner
- Actual deployment dependent on numerous factors such as facility constraints, scale requirements

Challenges

- Most DC architectures built around 1-10GE MMF reach (1-300 meters inside DC for MMF)
- MSDC environments PODs side at 100-150m with interconnect requirements beyond 150m on MMF optics
- MDSC Inter-POD > 2km
- 40/100GE MMF reach challenges compared to 10G
- Reach challenges with 40G/100G MMF drive need for lower cost single mode optics
- Highly meshed interconnect drives need for high port density on equipment.
- When using ribbon fiber, are there ribbon TAPs?
- Cable Management. Automated patch panels: Need SMF to enable.

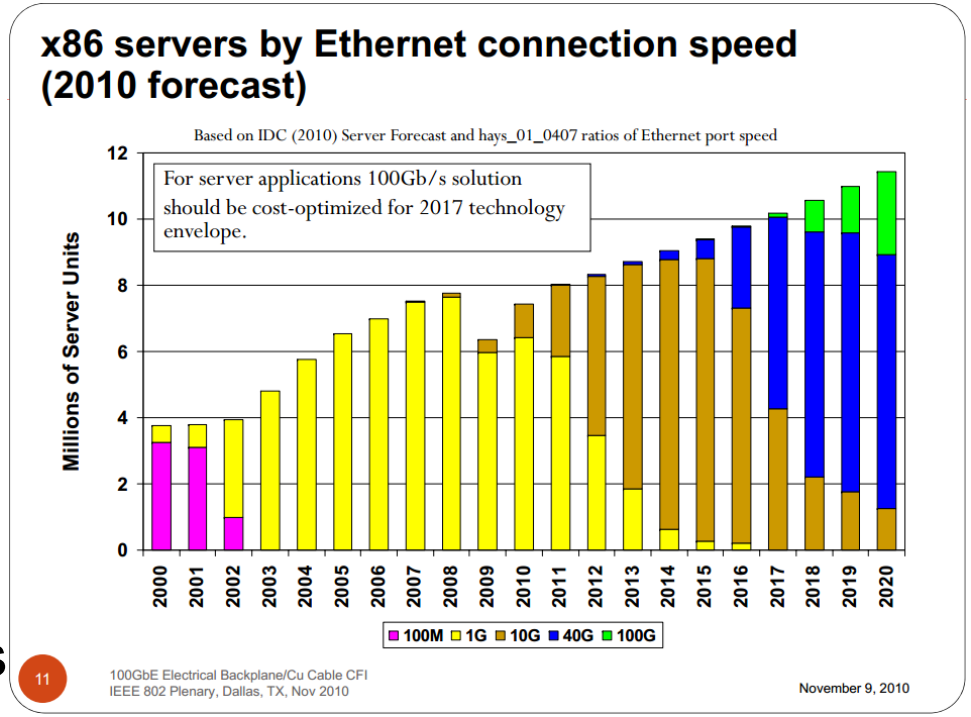
Implications for NG100G Optics Study group

- Boiling this down to PMD requirements that the Study group needs to consider:
  - System port density is critical (size/power challenge on PMD)
  - Economics is critical (cost challenge on PMD)

Broader Market Context

- This project is likely to complete in 2014 timeframe - Cost optimization thus should be targeted within 2-3 years
- Coincides with forecasted emergence of 100G Server Market

- Market transition to 25G SerDes technology taking place
- IEEE 802.3bj Task Force…
- Multiple announcements and developments within CMOS
  - During 802.3ba timeframe, 25G SerDes relied upon SiGe



Source = CFI_01_1110.pdf

LINKS:

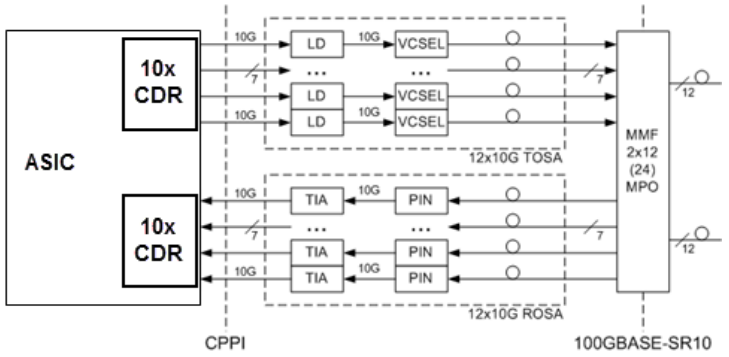
- Altera… Demonstrating 25-Gbps Transceivers in Programmable Logic, **Sept 2010**
- Xilinx… World's First Single-FPGA Solution for 400G Communications Line Cards, **Nov 2010**
- Inphi samples chips to power 100G ports, **Sept 2011**
- Avago Technologies Demonstrates Industry's First 28-nm 25-Gbps Long Reach-Compliant ASIC SerDes, **Feb 2012**

Next Gen MMF PMD

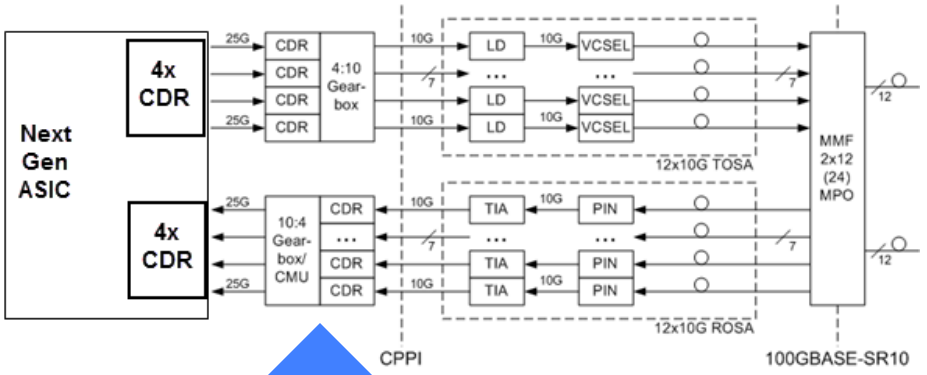
- 4x25G (aka SR4) is only proposal
- Definite port density advantages over 100GBASE-SR10
  - Optical lane rate will align with electrical lane rate – no GB or reverse GB needed
  - Cable reduction – great – reduces infrastructure cost
- Reach – unable to meet true DC needs (i.e to be compatible with reaches supported by 10GBASE-SR)
  - 100m definitely needed – cost/power/reach tradeoff above 100m needs to be understood
  - Is a second (shorter) reach required?
    - What reach?
    - Can AOC address? How would the standard address this case?
    - Further study to define?

System Need: MMF PMD

- Built today, an –SR10 interface is optimized to work with system chipsets with 10G SerDes technology



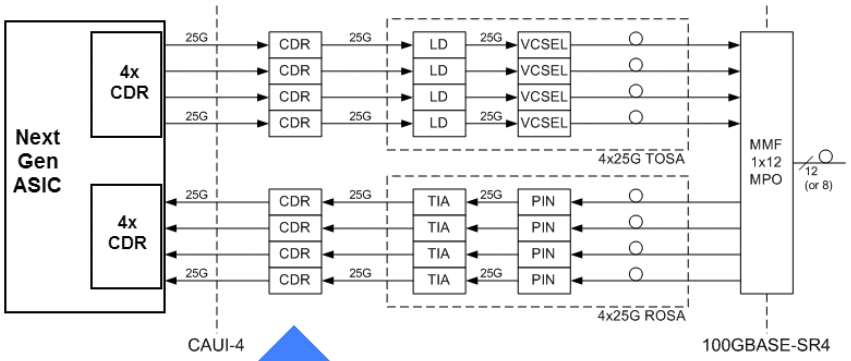
- Next Gen ASIC technology with 25G SerDes will need a Reverse Gearbox block function to interface to SR10



RGB adds cost & power to system

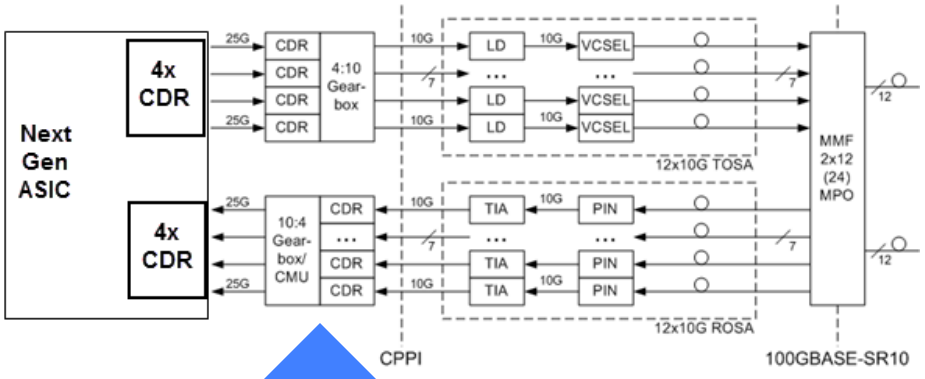
System Need: MMF PMD (2)

- Proposed –SR4 solution would offer path to lowest component count interface
  - *Care must be taken in defining electrical i/o !*



CDR may be pulled into ASIC – especially for Server & low port count implementations

- Next Gen ASIC technology with 25G SerDes will need a Reverse Gearbox block function to interface to SR10



RGB adds cost & power to system

Next Gen MMF PMD status in SG

- Broad Market Potential: meeting DC requirements addresses BMP. System port density is critical to achieve those requirements.
  - However, two PMDs complicates BMP response.
- Economic Feasibility: SG has data already. More can not hurt
- Technical Feasibility: Solid data establishing feasibility. Extra work needed to justify two PMDs
- Distinct identity: Two 4x25G MMF PMDs could complicate the response.

Next Gen SMF PMD

- Three proposals under consideration is SG:
  - 4x25 parallel
  - PAMn
  - Do nothing - economies of scale best
- 40G/100G MMF reach limitations are heightening the pressure on SMF to meet DC requirements
- Architecture trends demand high port count, low cost interface solutions
- Reach – DC scale requires reaches up to 2km, but 300-500m should be optimization point for SG
  - Parallel proposal has increased cable costs as reach increases
  - Monitor taps, automated path panels – require duplex SMF

Next Gen SMF PMD status in SG

- Broad Market Potential: meeting DC requirements addresses BMP
- Economic Feasibility: SG has data already. More can not hurt
- Technical Feasibility: key focus for SG this meeting and next!
- Distinct identity: should limit to only one PMD not both.

Recommendations

- Note this is NOT proposed language for objectives – rather guidance.
- MMF objective:
  - Define a PHY supporting 100m MMF
    - Fiber type to be defined in TF.
  - More study needed on impact of shorter reach differences (power/ size/cost) relative to 100m option
- SMF Objective
  - Define single PHY supporting reach of >= 300m
    - Final reach to be determined in TF after detailed analysis of technology breakpoints
  - Discussion point to consider:
    - Should reach objective be defined as a minimum, maximum or range

Backup

Data Center Market Transition

Bandwidth Scale

Mainstream Adoption

Servers

ToR

Core

40x 1G Servers
40G Rack

Uplinks
2-4x10G

10x Increase

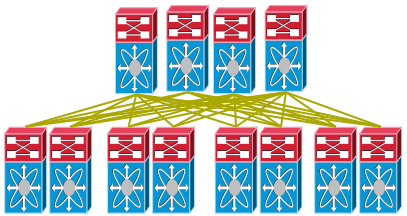
40x 10G Servers
400G Rack

Uplinks
2-10x 40G
2-4x100G

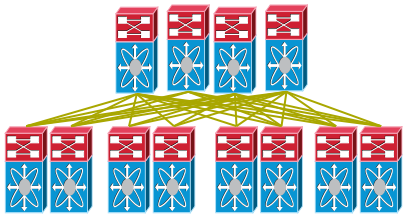
4x Increase

40x 40G Servers
1.6T Rack

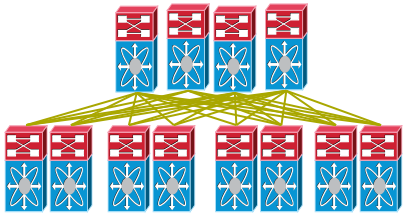
Uplinks
2-16x 100G
2-4x400G



Non-Blocking 10G Core



Non-Blocking 40/100G Core



Non-Blocking 400G/1Tb Core