Considerations for 25GbE

John D'Ambrosia, Dell
Jon Lewis, Dell
Kapil Shrikhande, Dell

IEEE 802.3 25GbE Study Group
IEEE 802.3 Sept 2014 Interim
Kanata, Canada

Ethernet is Evolving

- Changing environment
  - "10x the Performance @ 3x the cost"
  - "The Ethernet of Everywhere" – being used everywhere for everything
    - Pick one – "Web Scale Data Center, Enterprise Data Center,, Enterprise, Campus, Client Side Connections, Etc"
    - Architectures Top-of-Rack, End-of-Row, Middle-of-Row
  - IEEE 802.3 Ethernet Bandwidth Assessment – 32% to 95% CAGR
  - Connections from ≈0m to 40km
  - "Fixed Ports" – really? – Form factor – yes – Function - no
  - Market Timings
  - PoE Certification discussions
- Per the 25GbE CFI Consensus Presentation:
  - **Web-scale data centers and cloud based services need**
    - **Servers with >10GbE capability**
    - **Cost sensitive for nearer-term deployment**
- Remember that Ethernet products designed for this space will move into other applications!

Things to Consider for Objectives

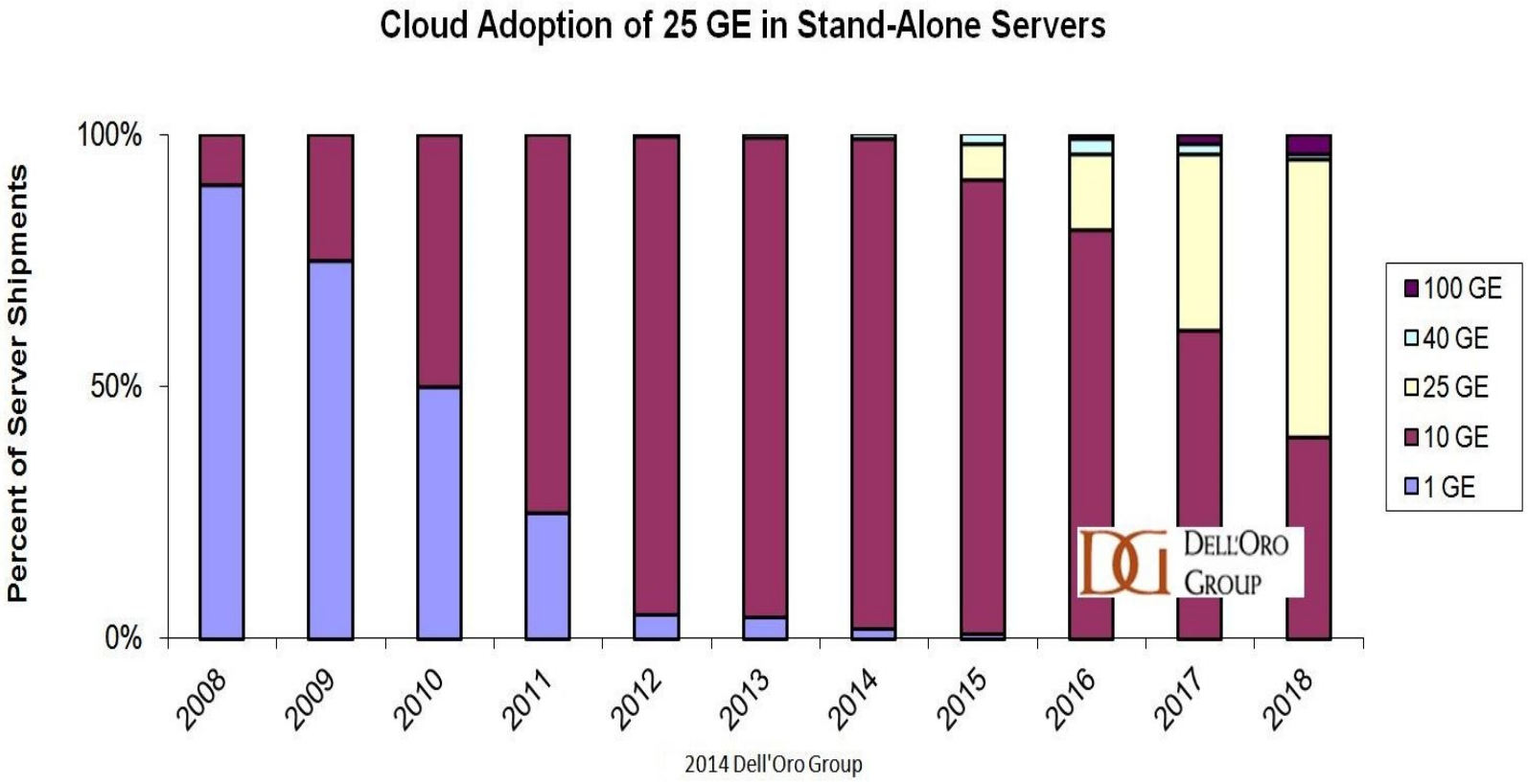
- Cu Cable Reach
- Need for an MMF Objective?
- Need for electrical interfaces?

A Few Words First…

- Lane Rate / Maximizing Switch Efficiency / Breakout to lower rates driving new issues
- Examples –
  - Success of 40GbE or 10GbE?
    - Breakout from QSFP has been a noted success.
    - Challenges in quantifying application volumes
    - "Fixed Ports" – Form factor – yes, Media / Rate – no
  - On-going debate in IEEE P802.3bs 400GbE in relation to 100GbE breakout
  - Formation of IEEE 802.3 25GbE Study Group

- **From 25GbE Consensus Presentation**
- **25Gb/s technology standardized, developed, productized for 100GbE can be leveraged now!**

Technology	Nomenclature	Description	Status
Backplanes	100GBASE-KR4 100GBASE-KP4	4 x 25 Gb/s (NRZ) 4 x 25 Gb/s (PAM-4)	IEEE Std 802.3bj™-2014 Ratified
Cu Twin-Axial	100GBASE-CR4	4 x 25 Gb/s	
Chip-to-Chip	CAUI-4	4 x 25 Gb/s	IEEE P802.3bm in Sponsor Ballot
Chip-to-Interface	CAUI-4	4 x 25 Gb/s	
Module Form Factor	SFP28	1 x 28 Gb/s	Summary Document SFF-8402
	QSFP28	4 x 25 Gb/s	Style 1 - MDI for 100GBASE-CR4 Summary Document SFF-8665
	CFP2	4 x 25 Gb/s	
	CFP4	4 x 25 Gb/s	Style 2 MDI for 100GBASE-CR4


Cloud Adoption of 25 GE in Stand-Alone Servers
Percent of Server Shipments
100%
50%
0%
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
100 GE
40 GE
25 GE
10 GE
1 GE
DELL'ORO GROUP
2014 Dell'Oro Group

Big Driver: Total Cost of Ownership

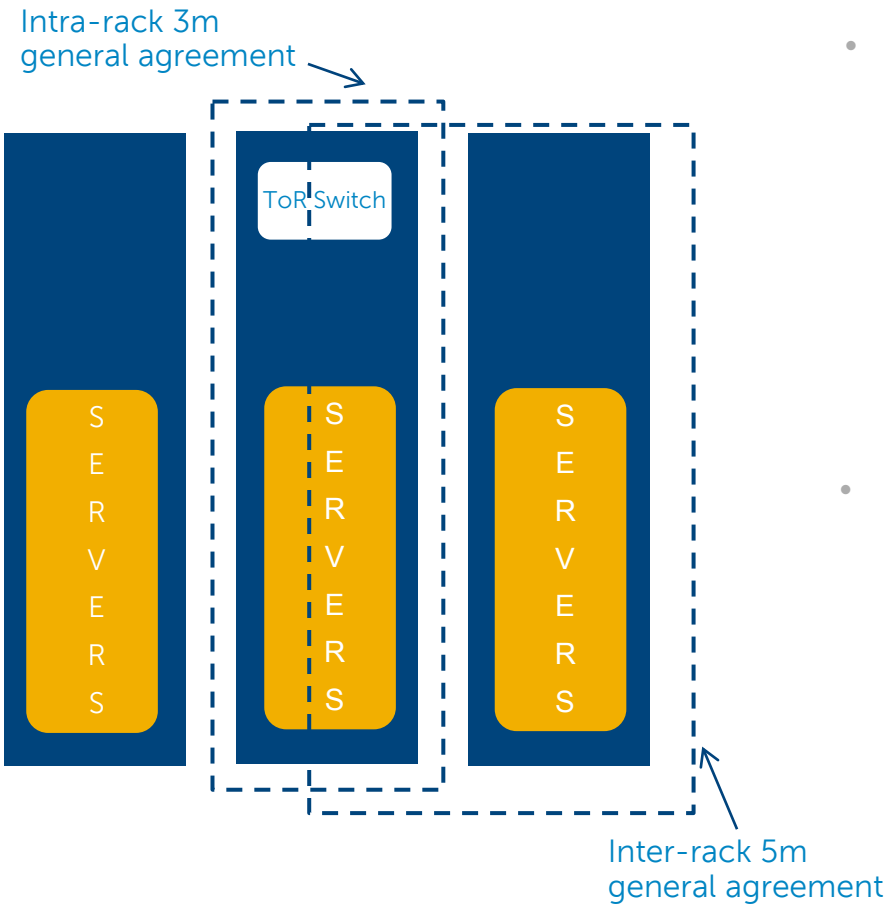
Consider Today's Cloud Scale Data Centers

	Top of Rack Box, Based on Single 128 I/O (3.2Tb) Silicon Switch Device					# TORs for a 100K Server Data Center
Server I/O	Oversubscription	Servers	100G Uplinks	Throughput (Tb/s) per ToR Switch	Utilization (%)	
40GbE (4x10G)	2.8:1	28	4	1.52	47.5	3572
40GbE (2x20G)	2.4:1	48	8	2.72	85	2084
25GbE Single Lane	3:1	96	8	3.2	100	1042

- Total Cost of Ownership - **Minimize cost per bit!**
  - CAPEX – Top of Rack Switches, Interconnect Structure
  - OPEX – Power / Cooling

Represents 100% port utilization and no stranded ports

Cu Cable Distribution



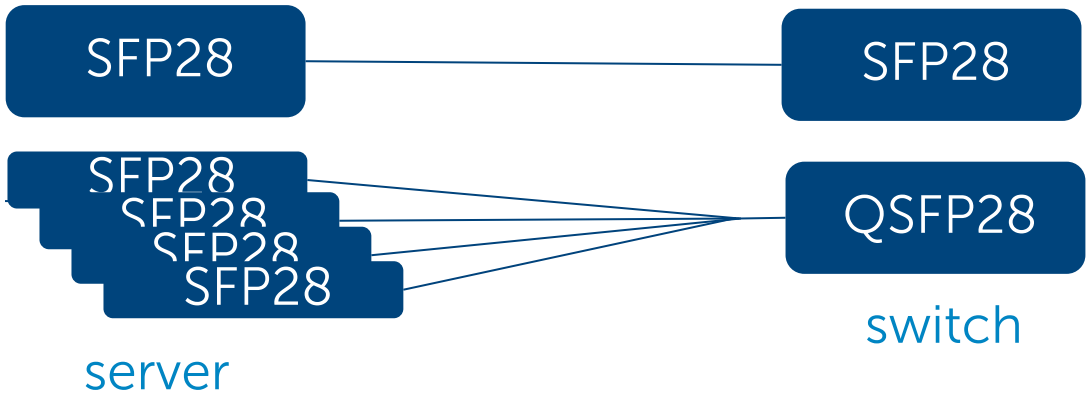
- Data obtained from
  - Two product groups within Dell (past 1 to 1.5 years)
  - 10GbE based products (servers & switches)
  - 40GbE based products (servers & switches)

Total (Cu Cable)	Division A	Division B
<=3m	79%	63%
5m	21%	28%
>5m	0%	8%

- Data obtained from
  - Two cabling companies (Molex,Tyco)

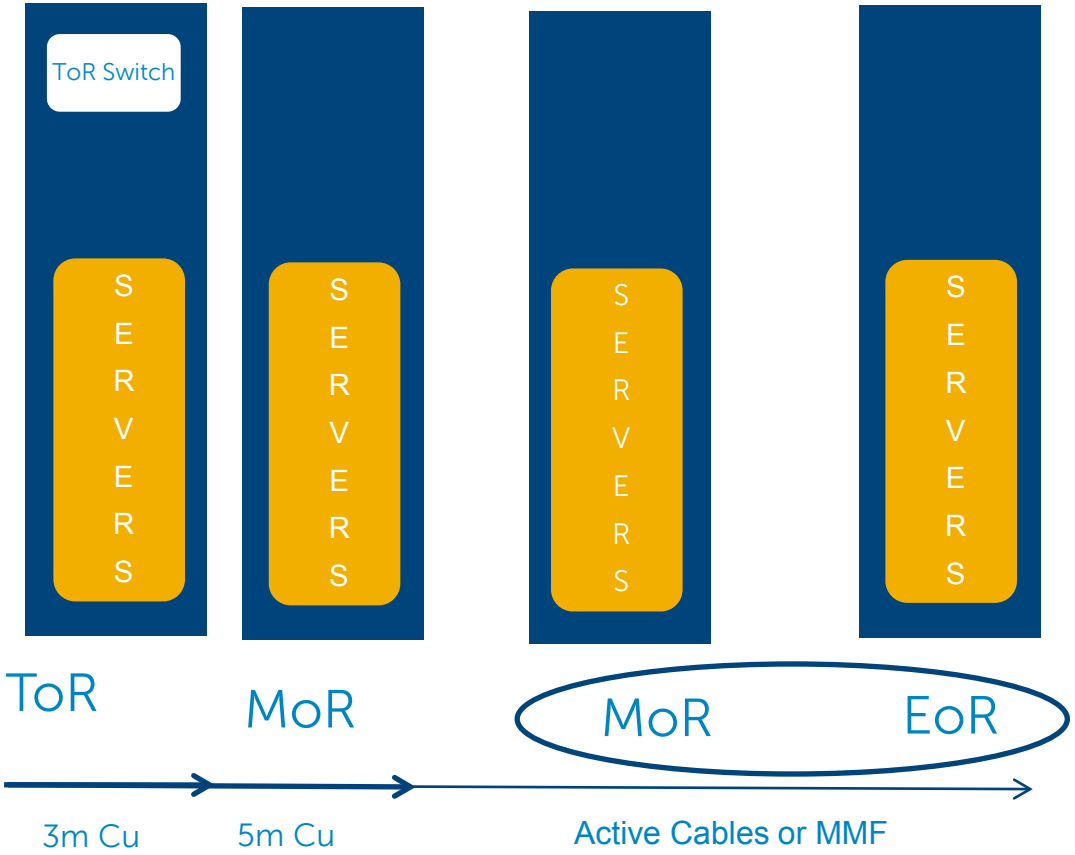
Total (Cu Cable)	Company A MDI1/MDI2	Company B
<=3m	62% / 69%	80%
5m	30% / 24%	15%
>5m (Passive)	1%	5%
>7m (Active)	7% / 6%	-

Thoughts Related to Cu Cable Objective



- Different applications easy to envision –
  - Lower density based on 25GbE / SFP28 from server to switch
  - Higher density based on breakout from 100GBASE-CR4 / QSFP28 on switch to SFP28 on server
- Channel budget?
  - Switch – must be constrained to meet 100GBASE-CR4 budget
  - Server – different options –
    - Different server port types to support budget for lower / higher density switch applications?
    - Add budget to server from cable?
    - Reduced budget for server NIC, 3m no FEC?

Beyond Top of Rack



I/O per 10G Server** ports

SFP+ (PHY Type)	
SR Optics	31.5%
CR (all passive, no active cables)	Some portion of 68.5%
Unknown*	Some portion of 68.5%

- - in some instances dual ports are used for physical redundancy, but one port may not be populated

** - data gathered from Dell general purpose server family

Comparison Between Options

- Option #1 – Reduce cable reach to 3m / assign budget to server
  - High density passive Cu switching applications limited to intra-rack / higher density server form factors
  - Potential for stranded ports on high density switches increases
    - More switches – CAPEX / OPEX
  - Forces use of active cable assemblies / optics for reaches beyond 3m
  - Limits broad market potential to intra-rack applications
- Option #2 –Choose 5m reach objective and TF can specify 3m cable with no FEC
  - Asymmetrical budget, NIC needs less loss than budgeted for host. Leave switch budget alone, 3m cable. No FEC
  - Reduces latency (for those applications)
- Option #3 –
  - Choose objective targeting 3m intra-rack applications
  - Choose objective targeting 5m inter-rack applications

Chip-to-Module (C2M) Interfaces

- Chip-to-module (C2M) interfaces will happen
  - SFP28 for 25GbE connections anticipated
  - QSFP28 for 4x25 GbE connections anticipated
  - Recommend SFP28 / QSFP28 for MDI
- C2M channel budget details need to be consolidated with host trace portion of Cu cable channel budget
- Chip-to-chip interface should be defined
- Leverage IEEE 802.3bm work
- Recommend adopting objectives for chip-to-chip and chip-to-module electrical interfaces

Summary

- Recommendations
  - Adopt a Cu cable objective for a 5m reach only (inter-rack)
  - Adopt a Cu cable objective for a 3m reach only (intra-rack)
  - Adopt a MMF Objective targeting xx m
    - Data on reach not available at this time
  - Consider SFP28 / QSFP28 for 25GbE MDIs
  - Adopt objectives for chip-to-chip and chip-to-module electrical interfaces