Support for an objective of 25 Gb/s over MMF
Draft 0.1e

September 2014, Ottawa, Canada

(team of many)

Contributors

- Alan Flatman, LAN Technologies
- Jonathan King, Finisar
- Scott Kipp, Brocade
- Paul Kolesar, Commscope
- John Petrilla, Avago Technologies

Supporters

- Chris Cole, Finisar
- Jack Jewell, independent
- Robert Lingle, OFS
-
-
-

Contents

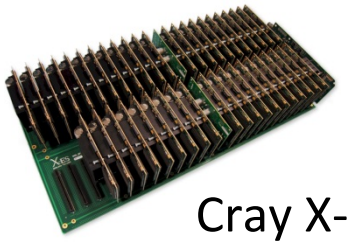
- Why include 25Gb/s over MMF in this project ?
  - Server designs, rack space and switch capacity
  - Middle-of-Row (MoR), End-of-Row (EoR), and Cabinet-to-cabinet
  - 'Home run' architectures – direct connect from server to core, eg large enterprise environments
  - Structured cabling
  - These server interconnect architectures represent a significant portion of market that is not addressed by a 5 m reach PMD
  - Ideally, optics and copper would plug into same socket (e.g. SFP28, QSFP)
    - Common electrical connector, compatible TP1, TP4 spec's
- Proposed 25Gb/s over MMF objective
- How adding an optical objective augments the 5 Criteria responses
  - **Broad Market Potential**
  - Compatibility
  - Distinct identity
  - Technical feasibility
    - Low technical risk, drawing on 32GFC and 100GBASE-SR4
  - **Economic feasibility**

Why include 25Gb/s over MMF in this project ?

- Original CFI included just Top-of-Rack (ToR) server-to-switch architectures
  - ToR is not sufficient for all applications
- Middle-of-Row (MoR) and End-of-Row (EoR) architectures include ~40% of total server-to-switch links, representing a very substantial market potential
  - Not addressed with a 5 m reach cable.
  - Some of these links may be addressed with AOCs.(spell out)
    - Market resistance to AOCs > 10 m (for pragmatic reasons).
  - A pluggable optic is needed for EoR and MoR architectures.
- The development of 32G Fibre Channel optical modules and consequent market interest shows that economic feasibility is achievable.
- An objective for 25Gb/s over MMF significantly broadens the market potential of the 25G Ethernet project.

Server Designs

- Microservers – ARM Servers
- Blade Servers
- 1/2U Servers
- 1U Servers
- 2U Servers
- 4-12U Servers
- Rack and multi-rack Servers



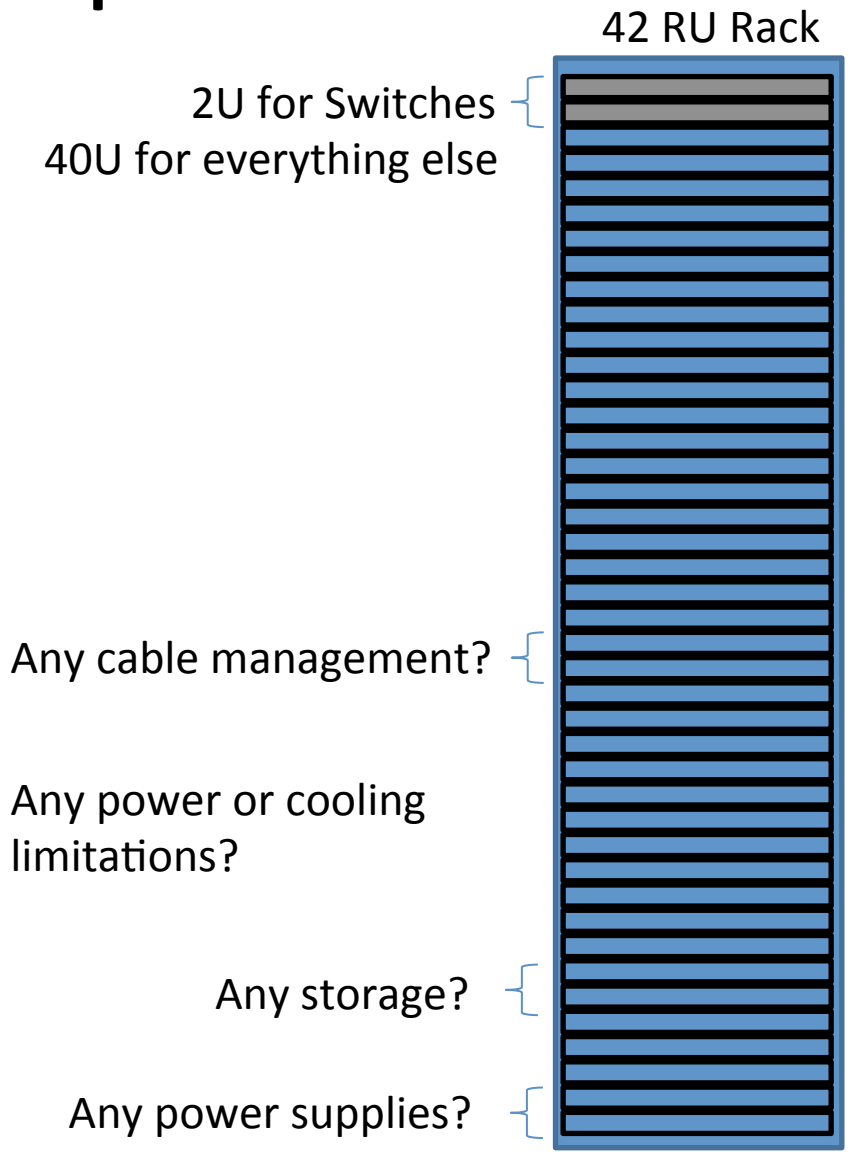
Cray X-ES Microserver

8U Storage Server

Mainframe

Rack Space

	Max Servers / 40RU
Micro-Server	>100
Blade Server	>100
1/2U Server	80
1U Server	40
2U Server	20
4U Server	10
8U Server	5
12U Server	3
Mainframe	<1



25GbE Switches

- Switch ASICs are increasing speed from 10GbE to 25GbE and more than doubling the port counts from 64 ports to 128+ ports

64 10GbE port ASIC enables
48 SFP+ and 4 QSFP+
620Gb/s of Throughput

128 25GbE port ASIC enables
32 QSFP+
3.2 Tb/s of Throughput

10GbE Switch Designs

- Blade Switches

4 SFP+

- 1/2U Switches

12 QSFP+ = 48 25GbE

- 1U Switches

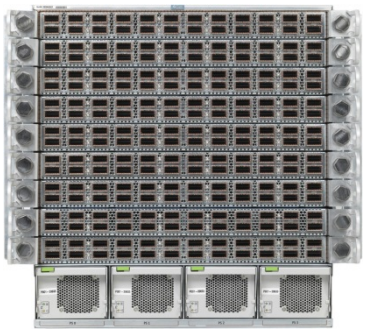
64 SFP+

36 QSFP+
= 144 25GbE

- 2U Switches

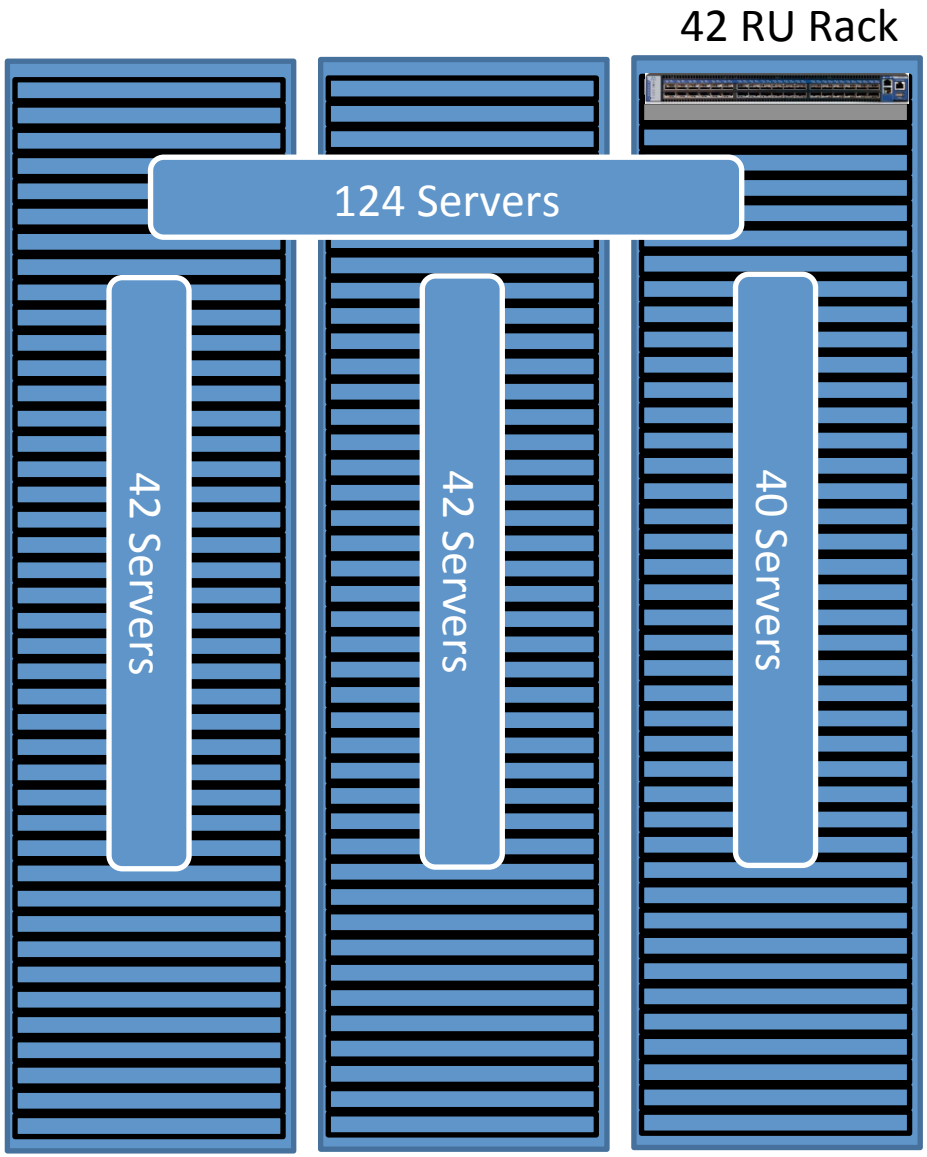
96 SFP+

- 4-12U Modular Switches



216 QSFP+
= 864 25GbE

1U Server ToR Designs



36 QSFP+
= 144 25GbE

20 ports left for uplinks…

One ToR switch can support multiple racks of 1U servers

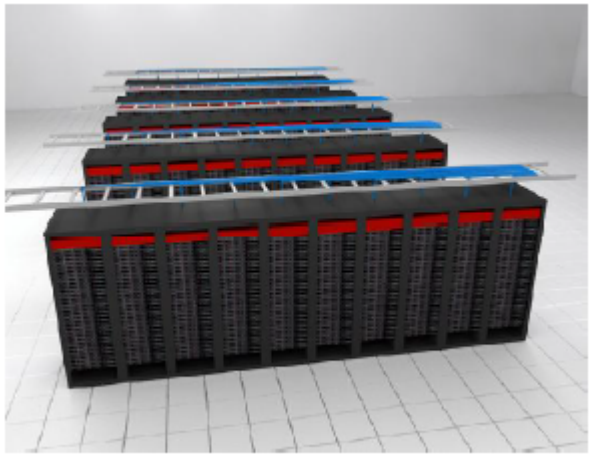
If the servers are 2U and not fully filled, it can support many more racks

ToR, MoR, EoR

ToR
(up to 5 m)

MoR
(up to 15 m)

EoR
(up to 50 m)

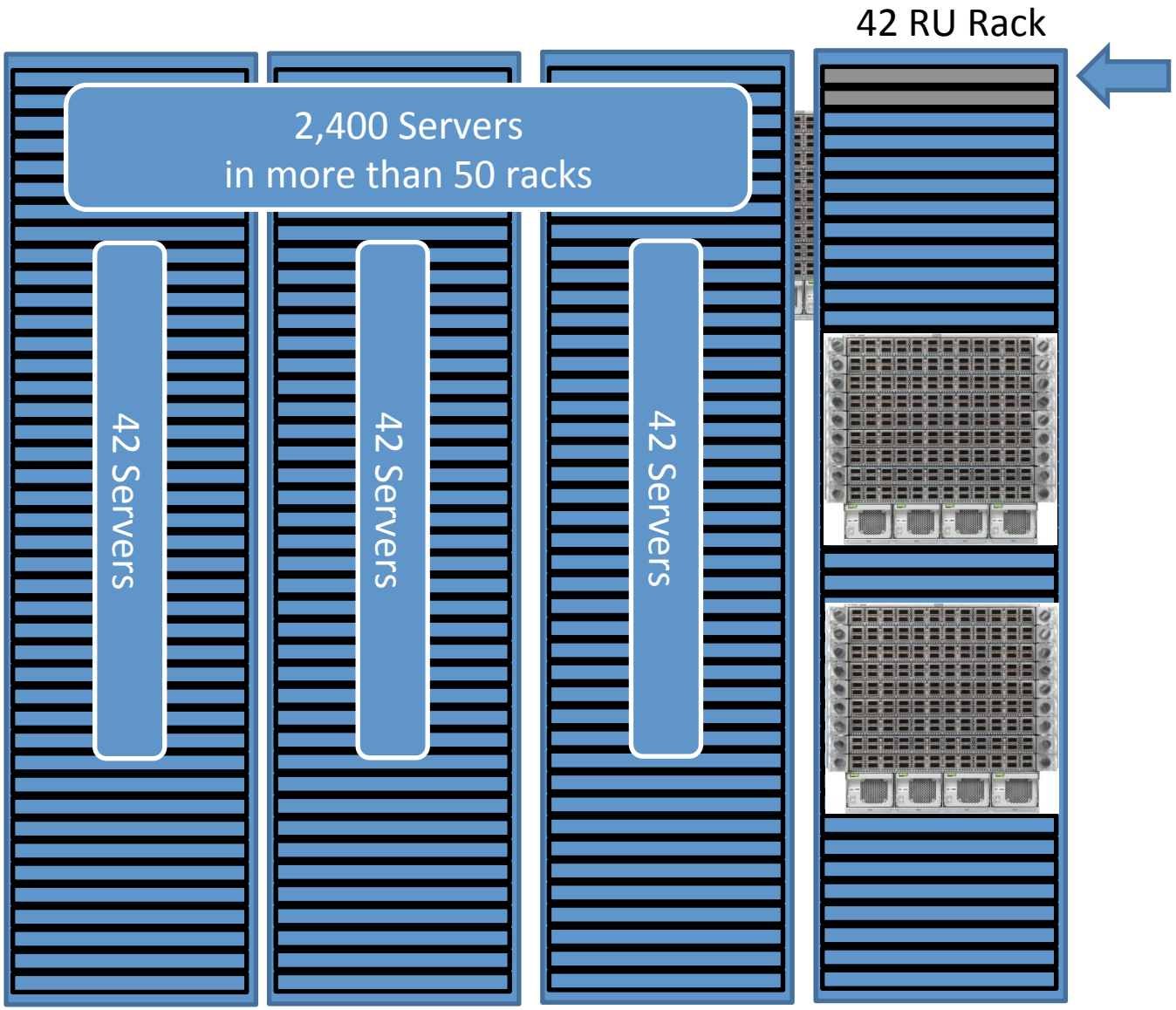


Addressed by 25Gb/s copper

Not addressed by 25Gb/s copper

Add attribution for pictures - jiminez

1U Server EoR Designs



216 QSFP+
= 864 25GbE
Each switch could support 800 servers with 1.6Tb/s of uplinks

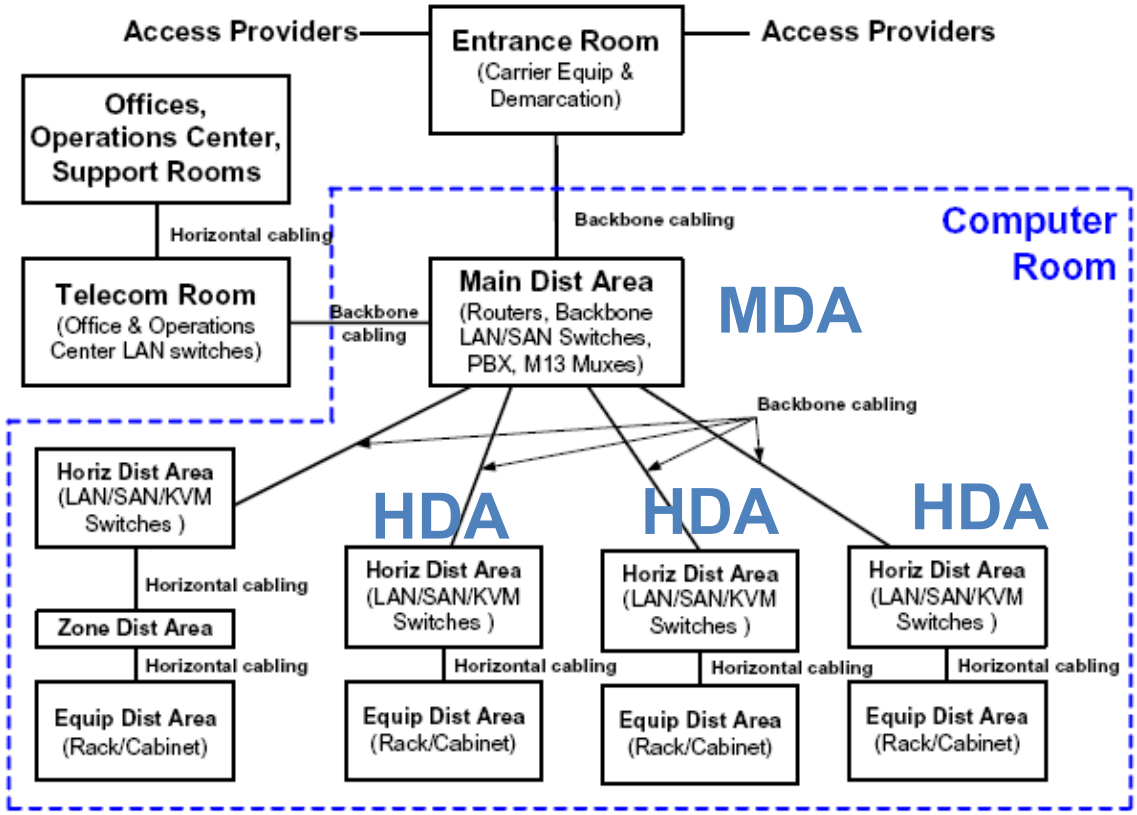
This rack of modular switches would support 2,400 servers

Home Run

- A home run is a server connect architecture where a server is connected straight into the core of the network
  - Common for storage servers or NAS that shares massive files – feedback to EA at HPC'13
  - Mainframes and large enterprise servers may connect straight into the core
  - These links need high speed
- Used when servers, storage and switches are consolidated into different areas
- Usually associated with structured cabling

TIA-942 – Data Center Cabling and Design

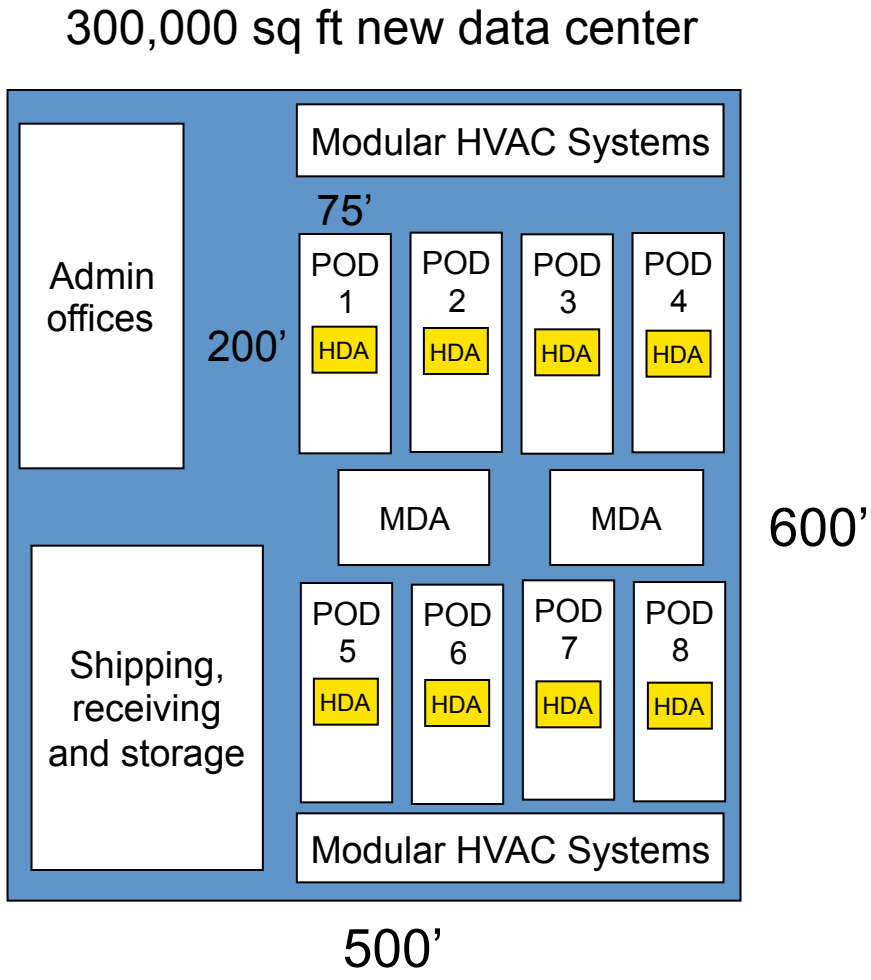
- TIA-942 - Telecommunications Infrastructure for Data Centers defines:
- MDA (Main Distribution Area) that fans out to
- HDAs (Horizontal Distribution Areas)



25G optics would address horizontal cabling

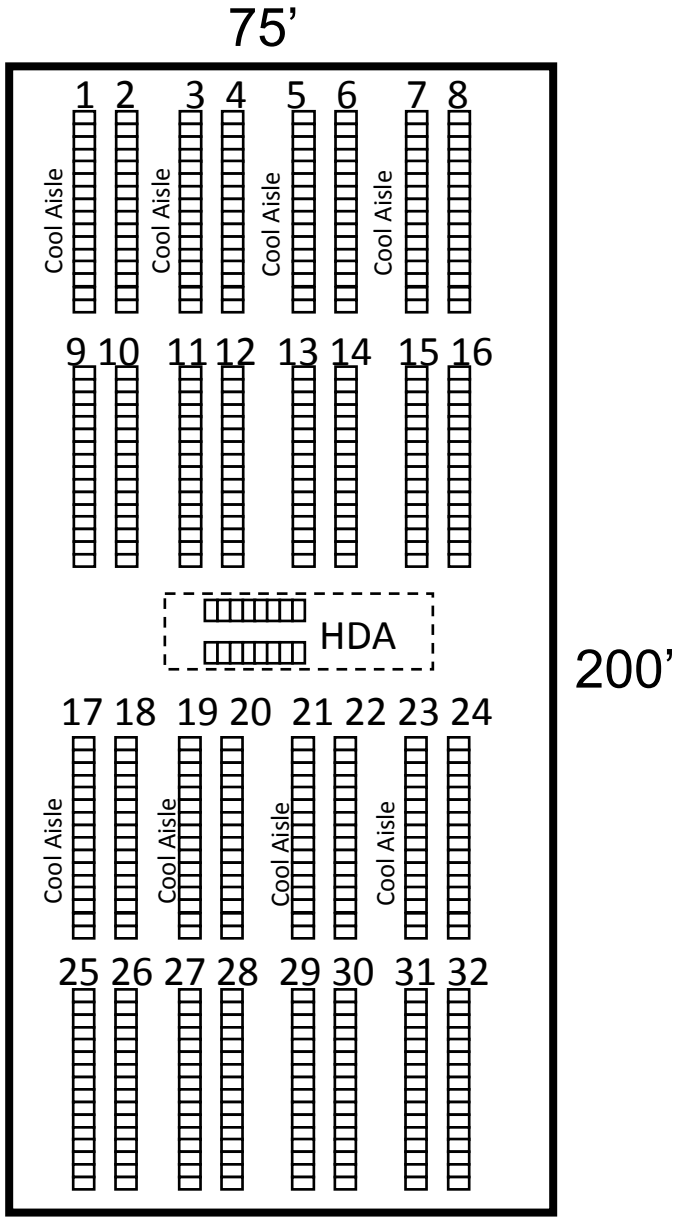
New Mega Data Center Design

- Most new data centers are being designed with a Pod (or Cell) Architecture
- Pods usually 15-20,000 sq ft
- HDA (Horizontal Distribution Area) is where distribution switches are located
- The Main Distribution Area (MDA) interconnect PODs and connects to the WAN and telecom networks



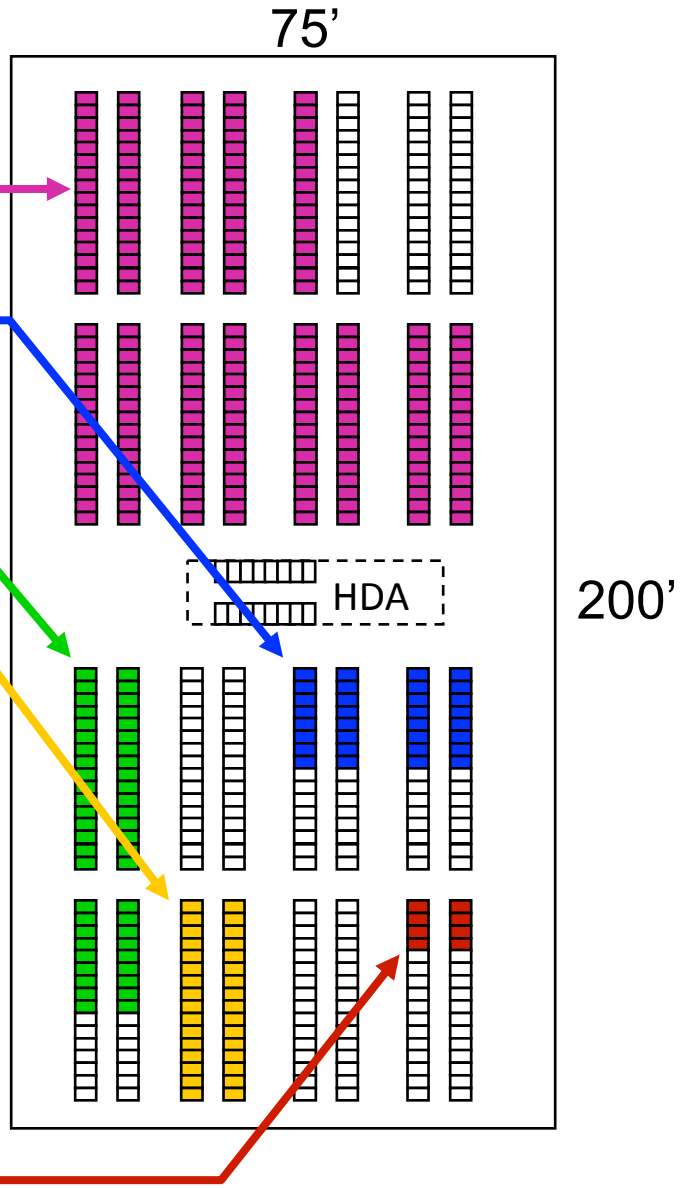
POD Architecture

- 15,000 sq ft POD
- Up to 5,000 servers / POD
- 512 Racks possible
  - 32 Rows of racks
  - Each row has 16 racks
- Horizontal Distribution Area (HDA) connects all of the racks



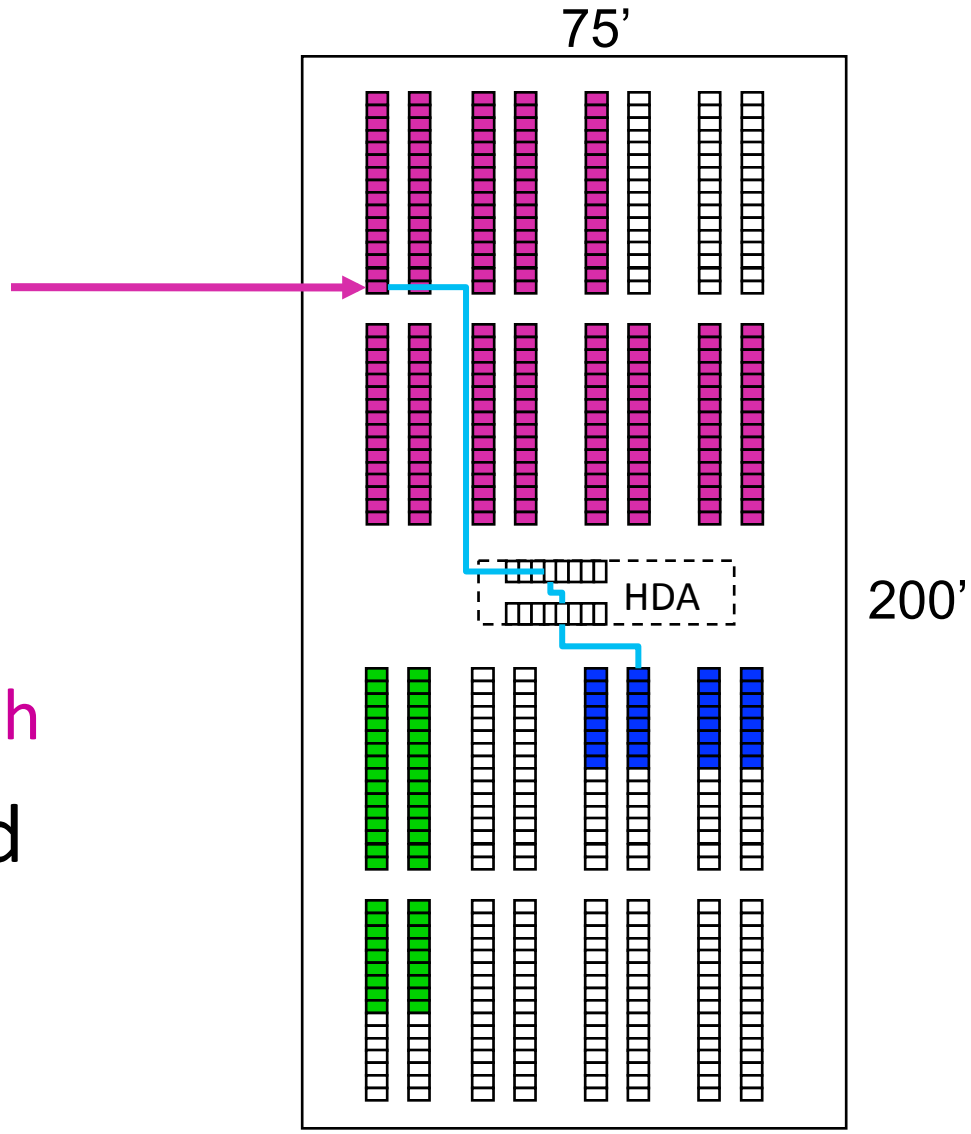
POD Design

- 200 Server Racks
- 32 Switch Racks
- 50 Storage Racks
- 32 Tape Racks
- 8 Mainframe Racks



Home Runs

- Home Runs go
  - From select servers
  - To HDA
  - Patchcord within HDA
  - To centralized Switch
- 100 meters required



Is this ethernet , and is it just 25G
Is this a niche application

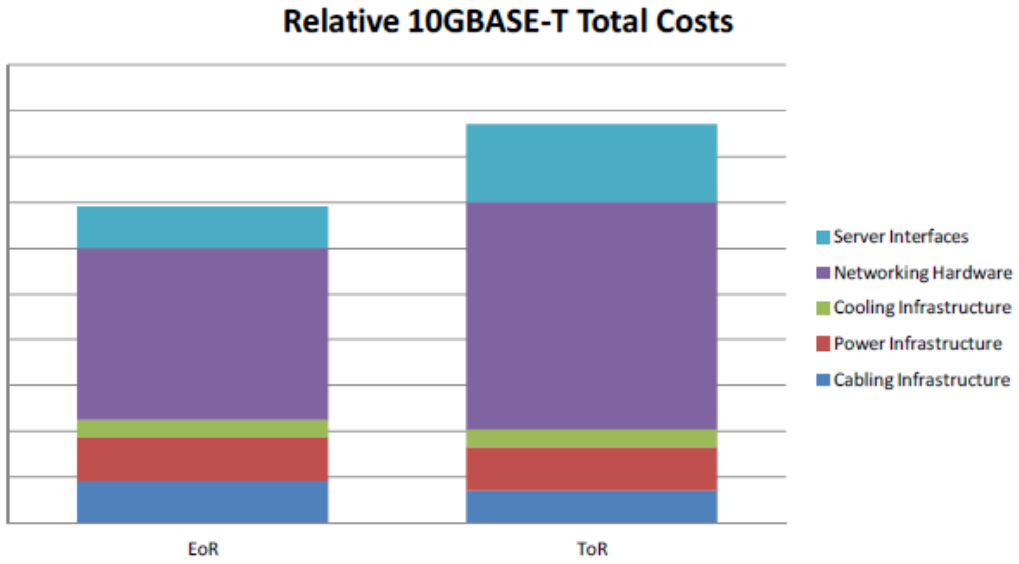
Structured cabling environment

Placeholder for Paul

TIA 942

From “40GBASE-T advantages and use cases” (jiminez_3bq_01_0711.pdf, 802.3bq)

EoR vs. ToR – Relative Costs (Detailed)



- Reduced cabling costs with ToR
- Increased cost of ToR architecture driven by network electronics and server interfaces

Source: Anixter Inc.

IEEE P802.3bq

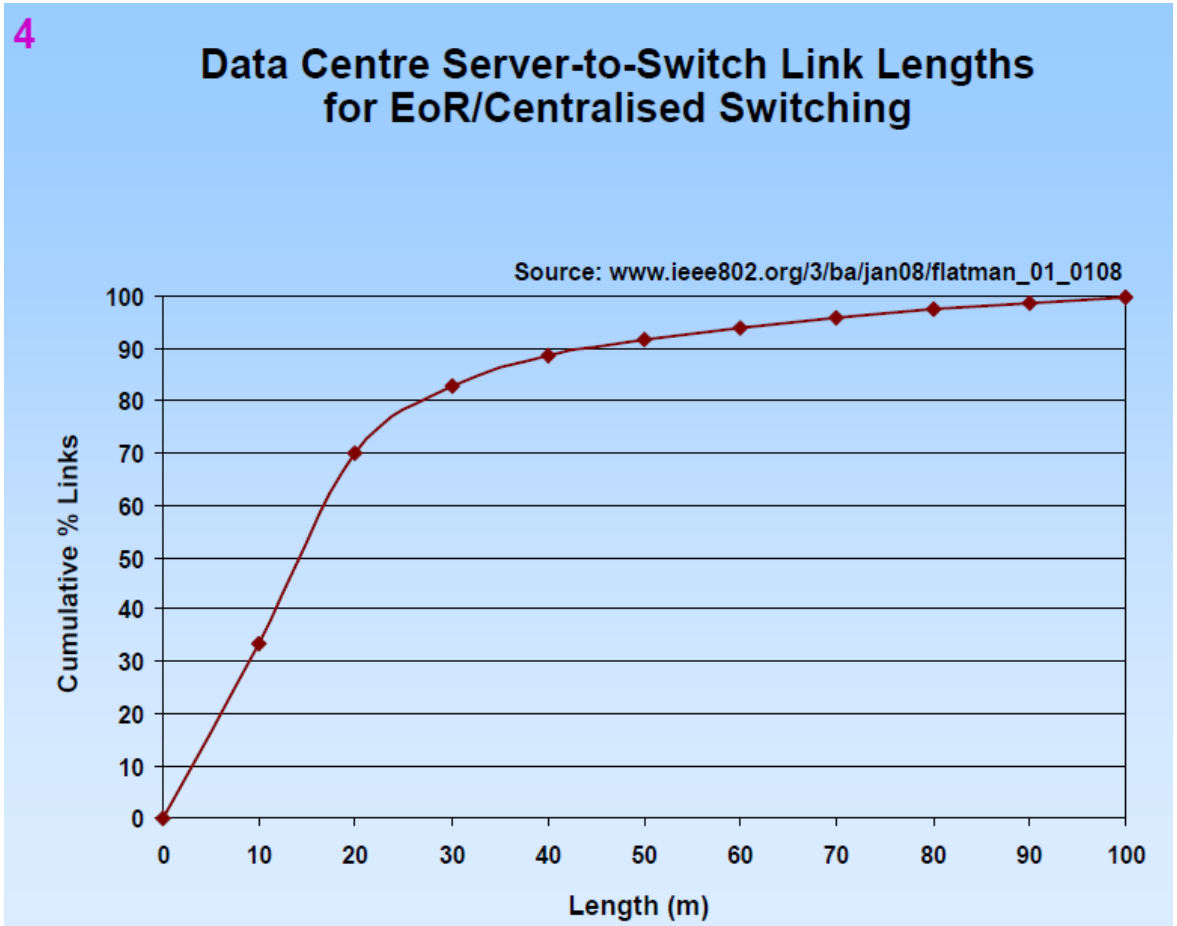
10

Port utilization argument.
Do we need this slide

- EoR is lower cost in some circumstances.

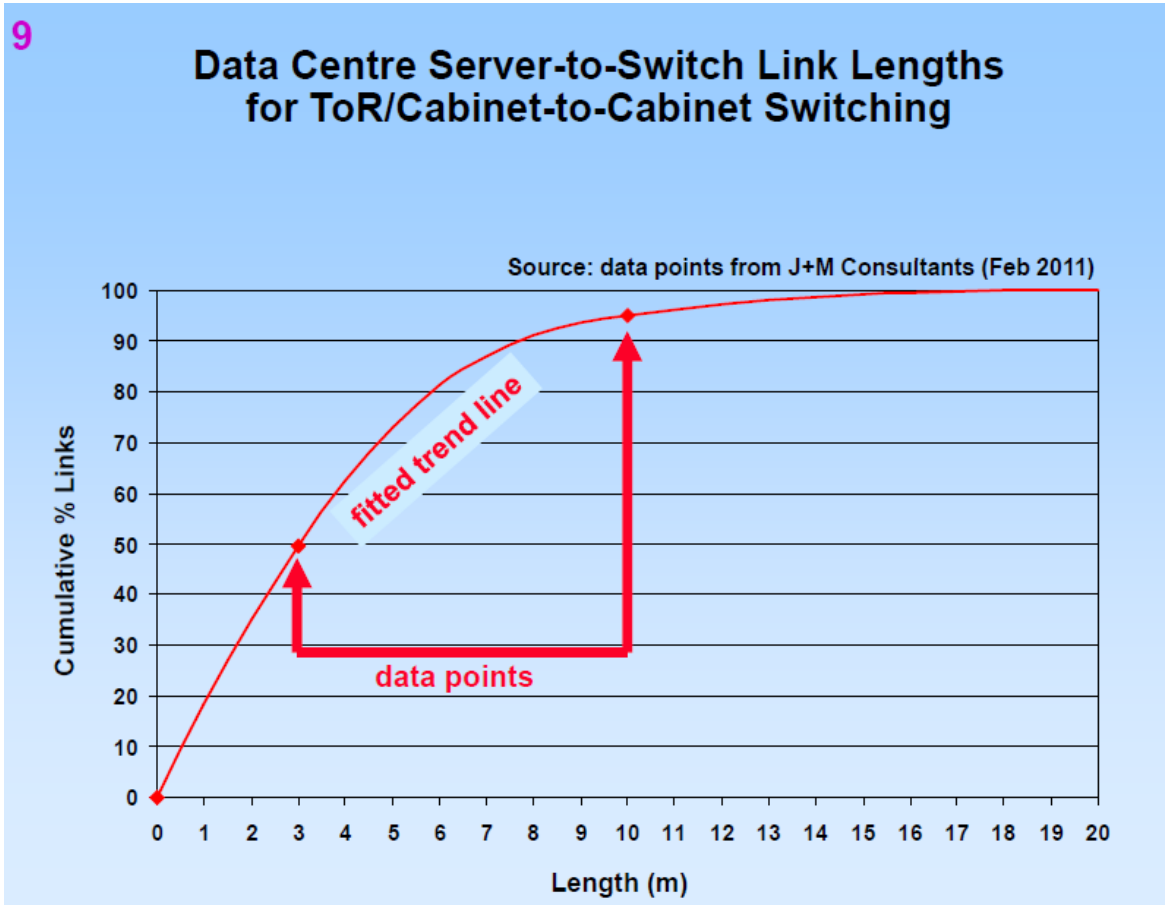
EoR link distributions

- Link lengths: ~85% > 5 m, ~ 90% < 50 m
  - From: *flatman_01_0911_NG100GOPTX.pdf*, reproduced with kind permission of Alan Flatman



ToR link distributions

- Includes cabinet-to-cabinet links
- Link lengths: ~30% > 5 m
  - From: *flatman_01_0911_NG100GOPTX.pdf*, reproduced with kind permission of Alan Flatman



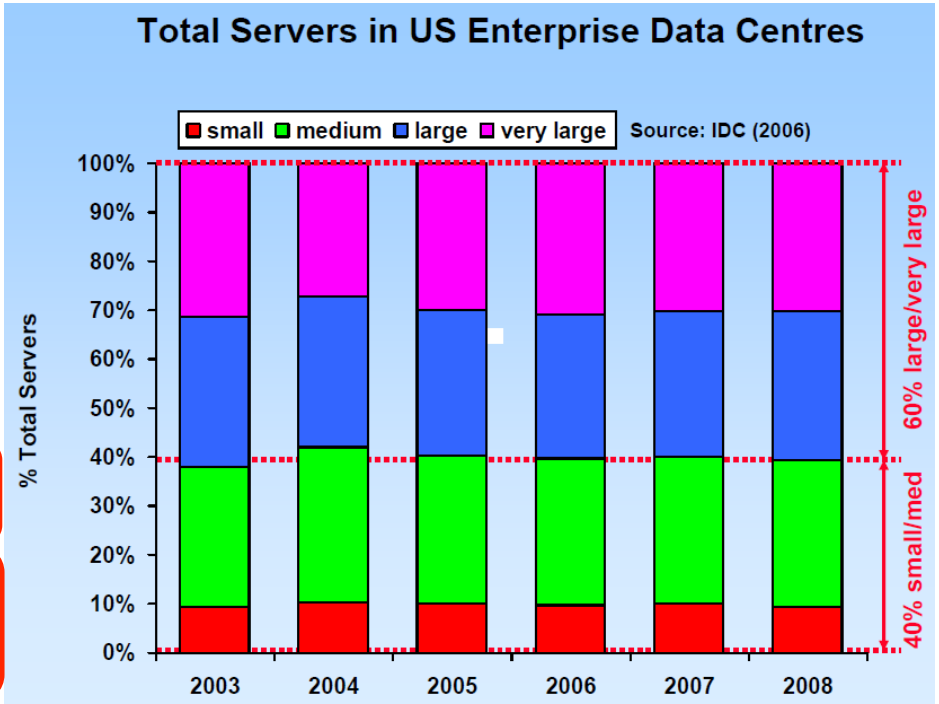
http://www.ieee802.org/3/100GNGOPTX/public/sept11/flatman_01_0911_NG100GOPTX.pdf

Total server volumes: 40% EoR vs 60% ToR

- Relative volume inferred from total number of servers in small/med vs large/v.large data centers (?)
  - From *flatman_01_0911_NG100GOPTX.pdf*, used with kind permission of Alan Flatman

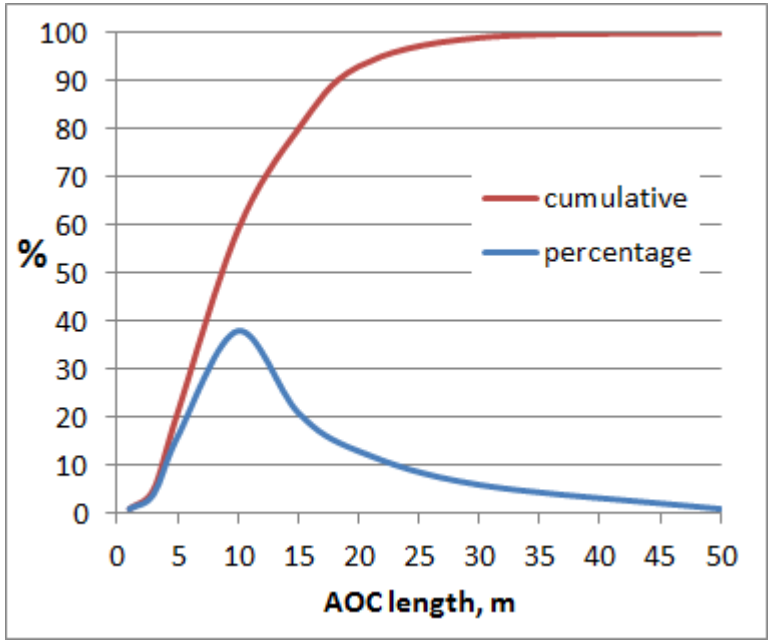
Flatman Data Centre Cabling Survey

- summary presented to IEEE 802.3ba in Jan 2008
  - www.ieee802.org/3/ba/jan08/flatman_01_0108
- 9 enterprise data centres from US, UK, Germany
- total data centre floor space = 715,000 square feet
- small, medium, large, v. large sizes (IDC classes)
- Flatman data good for EoR/centralised switching
  - expected to continue for small/medium data centres
- but now needs to take account of ToR switching & cabinet-to-cabinet links
  - being deployed mainly in large/v.large data centres
  - with much shorter server links than before



http://www.ieee802.org/3/100GNGOPTX/public/sept11/flatman_01_0911_NG100GOPTX.pdf

AOC length distribution and relative volumes



- Average AOC length < 10 m, 90% < 18 m
- Pluggable optics links exceed AOC volumes by ~ 3:1
- Add ref

Breakout bonus

?

Placeholder

Density advantage, 4x25 servers, across 3 or 4 racks feeding one 32 port QSFP switch

Economic Feasibility

Placeholder

Jiminez ppt ?

Port utilization efficiency: Big switches need multiple racks of servers to be fully utilized

Electrical connector

Placeholder

Can use same as twin-ax MDI

Proposed objective

- **Define a single-lane 25 Gb/s PHY for operation over MMF consistent with IEEE P802.3bm Clause 95**

How 5 Criteria responses may be modified by a 25 Gb/s over MMF objective

Broad Market Potential

- **An optical PHY utilizing a serial 25 Gb/s (1 x 25 Gb/s) electrical interface and optimized MMF interface will reduce cost, size and power for server interconnects in the data centers internet exchanges, co-location services, services provider and operator networks and provide a balance in cost between network equipment and attached stations.**

Supporting material:

Other infrastructure, e.g. in support of End-of-Row (EoR) or Middle-of-Row (MoR) will accelerate deployment and enhance deployment of Top-of-Rack (ToR)

From page 8 of Call For Interest Consensus presentation, "The term "TOR" has become synonymous with server access switch, even if it is not located "top of rack" ", acknowledging that a 5 m reach may not be sufficient.

Where longer than 5 m reaches are not sufficient, reliance on active cable assemblies does not provide satisfactory support in structured-cable installations.

Existing form factors supporting multiple lanes of similar electrical and optical interfaces provide high port density options

many other examples …

Compatibility

Inclusion of an objective for a single-lane 25 Gb/s PHY for operation over MMF is expected to have no specific Compatibility statement.

Distinct Identity

Each IEEE 802 LMSC standard shall have a distinct identity. To achieve this, each authorized project shall be:

a) Substantially different from other IEEE 802 standards.

b) One unique solution per problem (not two solutions to a problem).

c) Easy for the document reader to select the relevant specification.

d) Substantially different from other IEEE 802.3 specifications/solutions.

- **There is no standard that supports Ethernet over duplex multimode fiber cabling at a data rate of 25Gb/s. The IEEE P802.3 project will define a single 25Gb/s PHY over multimode fiber.**
- **The proposed amendment to the existing IEEE 802.3 standard will be formatted as a new clause, making it easy for the reader to select the relevant specification.**

Technical Feasibility

For a project to be authorized, it shall be able to show its technical feasibility. At a minimum, the proposed project shall show:

a) Demonstrated system feasibility.

b) Proven technology, reasonable testing.

c) Confidence in reliability.

- Component and cabling vendors have presented data indicating that 25Gb/s operation over multimode fibre cabling is feasible with known techniques similar to those used in existing 32G-FiberChannel and 802.3bm standards. Presentations have provided analyses of PHY feasibility based on measurements of installed cabling and proposed new cabling types from TIA and ISO/IEC aimed at this application.
- Systems and infrastructure supporting Ethernet operation over multimode fiber cabling have been deployed by the hundreds of millions at speeds ranging from 10Mb/s to 10Gb/s. The proposed project will build on Ethernet component and system design experience and the broad knowledge base of Ethernet network operation.
- The reliability of Ethernet components and systems can be projected in the target environments with a high degree of confidence.

Economic Feasibility

- The cost factors for Ethernet components and systems are well known.
- Prior experience with optical modules for 100GBASE-SR4 (4 lanes at 25.78 GBd per lane) and 32GFC (1 lane at 28.05 GBd) indicate that the specifications developed by this project will entail a reasonable cost for the performance of a single-lane 25 Gb/s PHY for operation over MMF.

Thank you !