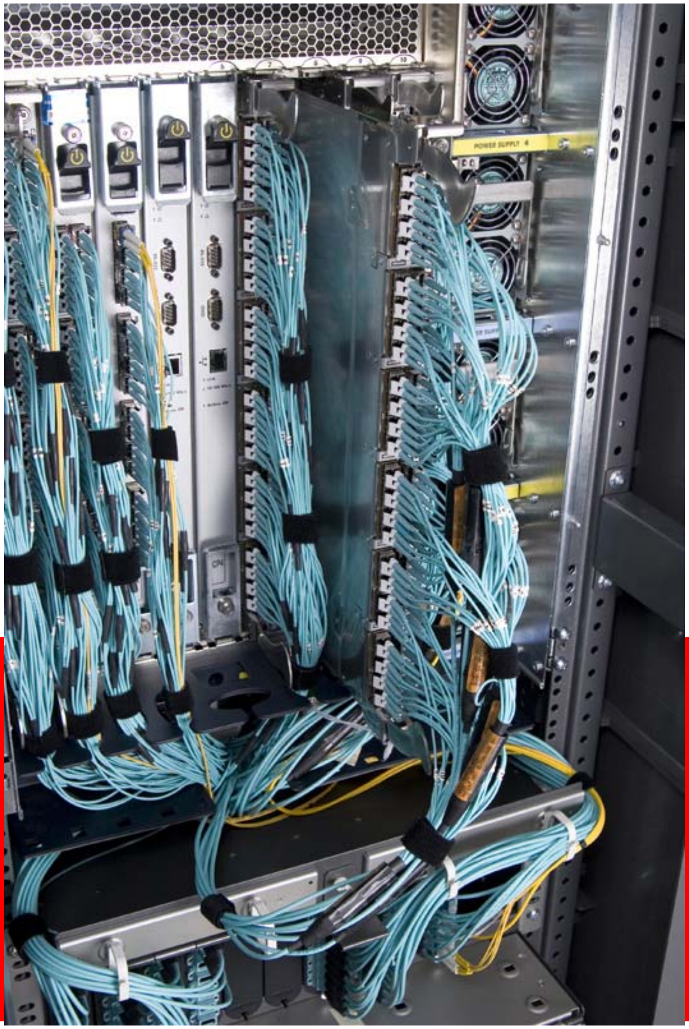
BROCADE

Switch Perspectives on XR Links



Scott Kipp

Office of the CTO

September 3, 2008

Discussions Today

- Problem Statement
- Data Center Designs
- Switch and line card designs

- Big Disclaimer – Every data center is different with I/O requirements depending on applications
- This presentation is full of generalizations that might not be true in specific cases

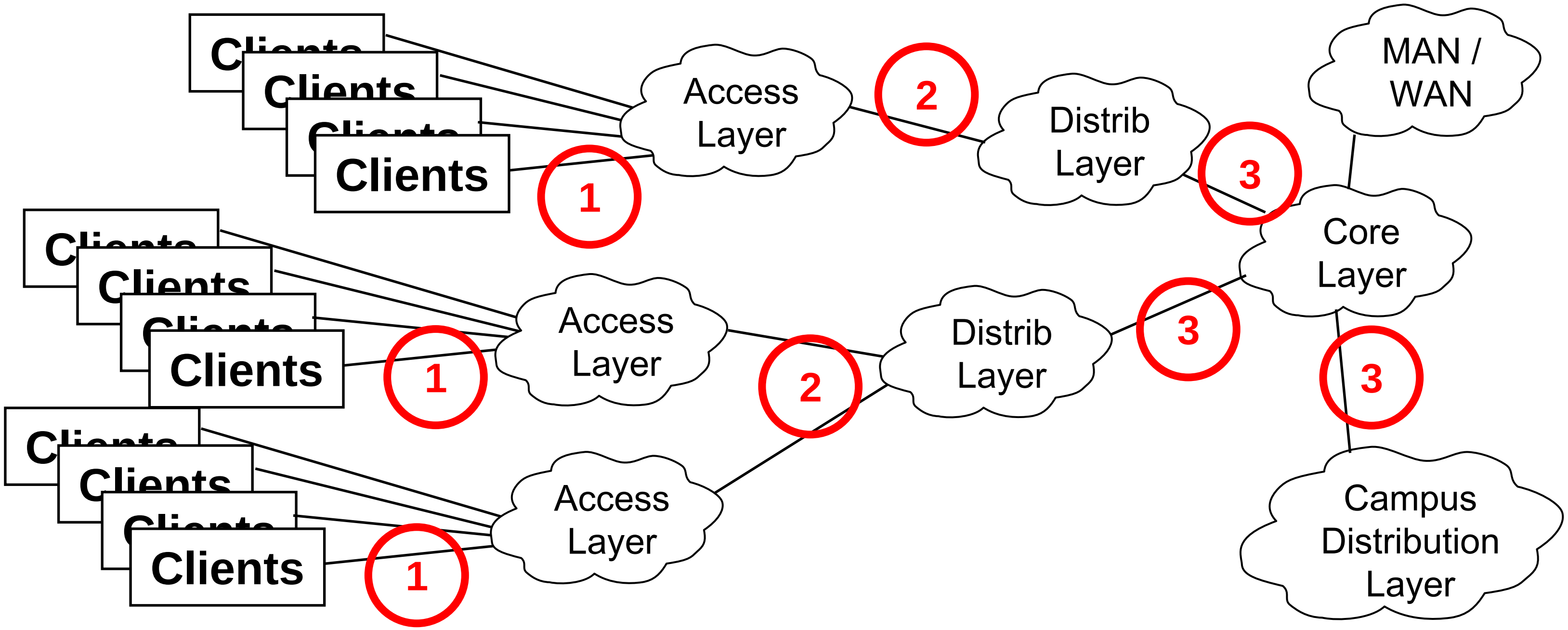
Layers of Ethernet Networks

# of Links in Flatman_01_0108

1. Client to Access Layer (C-A) – 250,000 Links
2. Access Layer to Distribution Layer (A-D) – 16,000 Links
3. DistribCore Layer – 3,000

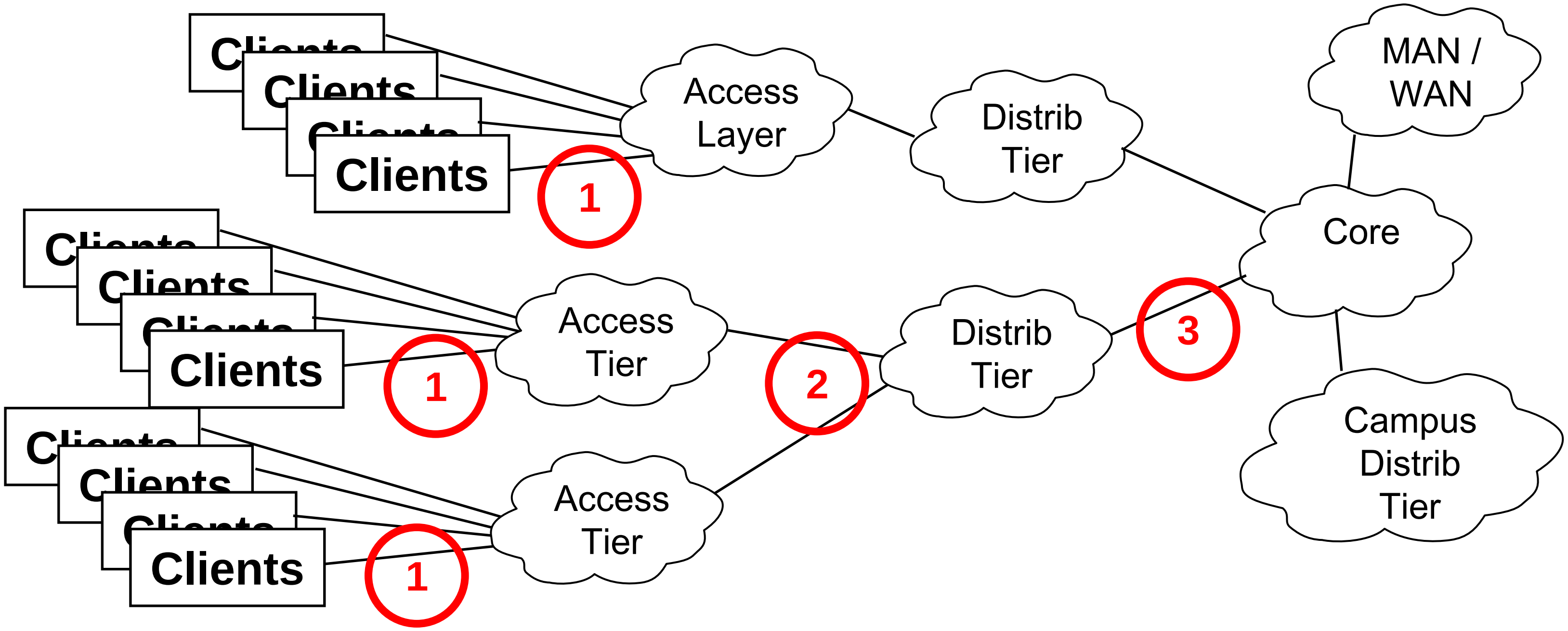
15.6 : 1

5.3 : 1



Problem Statement

- 10+ % of links longer than 100 meters in Layer 2 and 3 links



1. Clients to Access Tier
1-10G eventually 40G
100% satisfied by 100 meters
0% need longer lengths

2. Access to Distribution Tier
1-10G eventually 40G
89% satisfied by 100 meters
11% need longer lengths

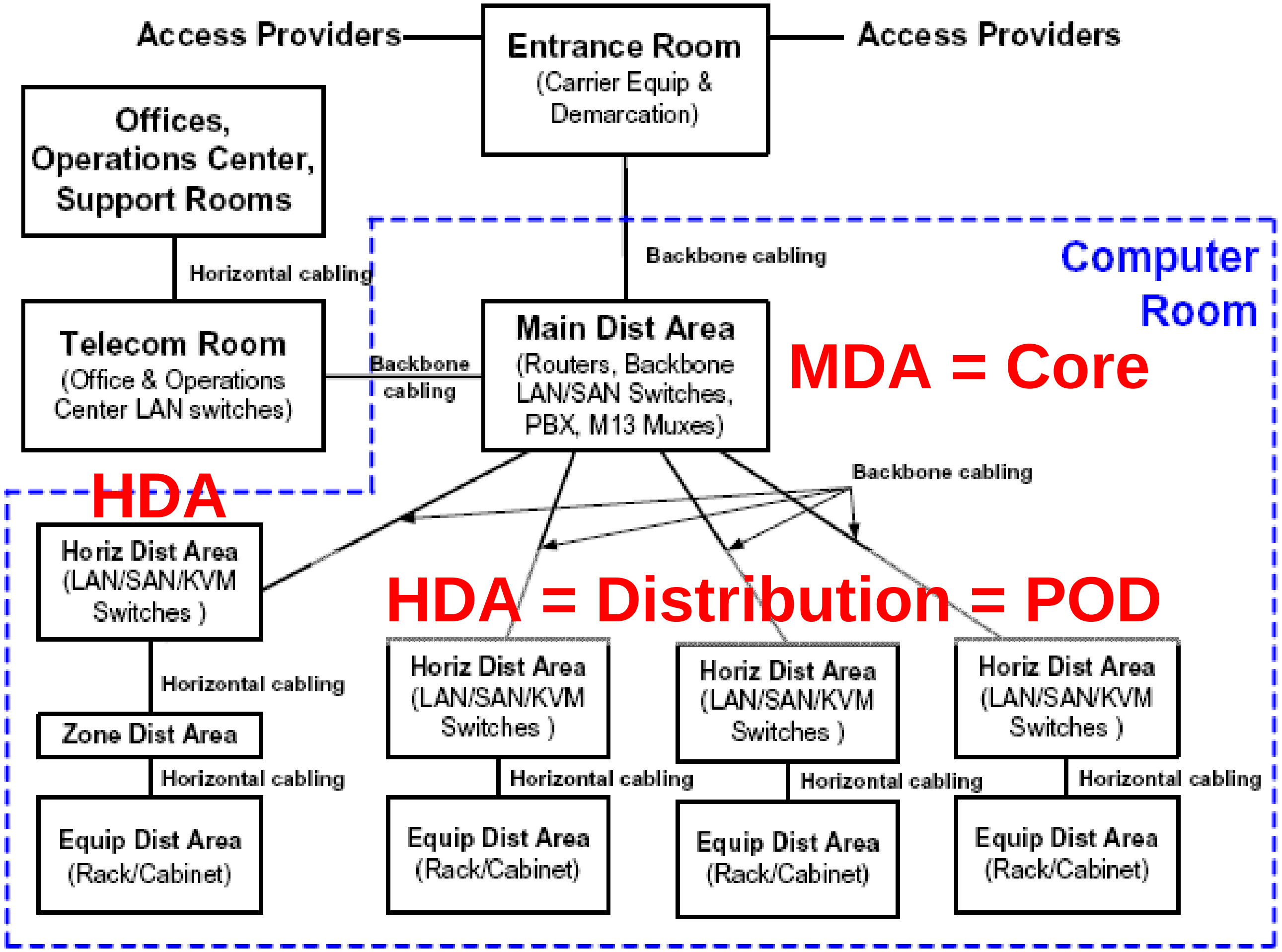
3. Distribution to Access Tier
10-40G eventually 100G
85% satisfied by 100 meters
15% need longer lengths

% are from flatman_01_0108.pdf

TIA-942 – Standardized Cabling

TIA-942 - Telecommunications Infrastructure for Data Centers standard defines the MDA (Main Distribution Area) that fans out to HDAs (Horizontal Distribution Areas) via backbone ribbon cables in a star topology

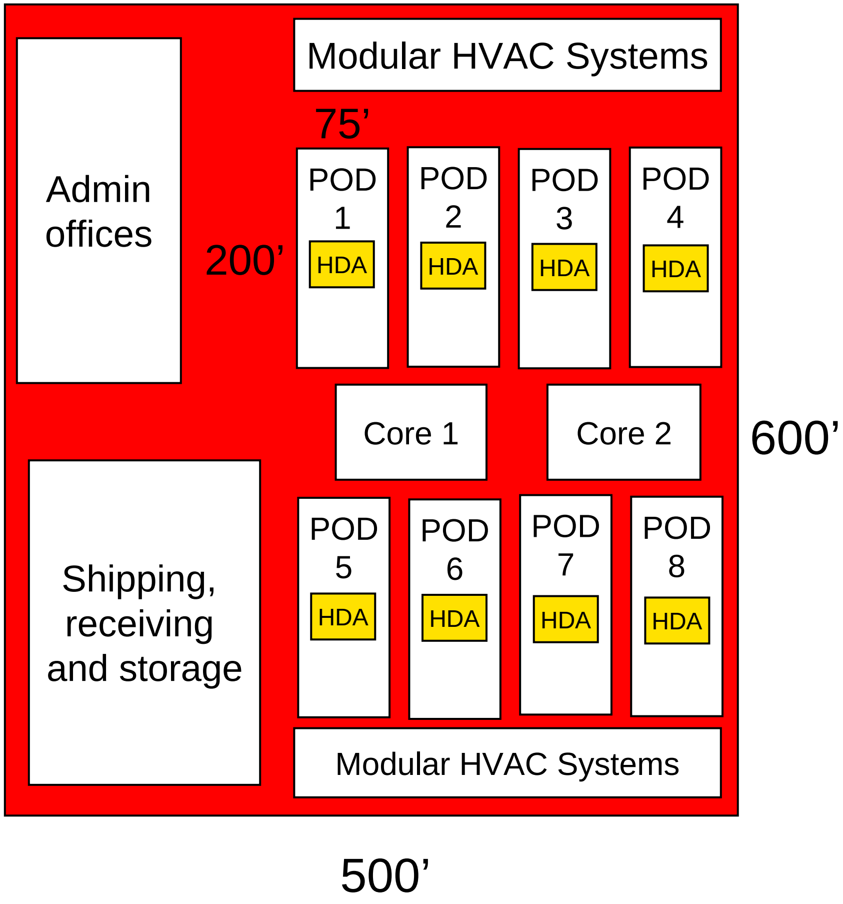
ZDA or EDA = Access



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 is where distribution switch is located
- Core layer is where Core Switches are located to interconnect PODs and to connect to the telecom networks
- 3,000 Servers / Pod for 24,000 servers total in data center

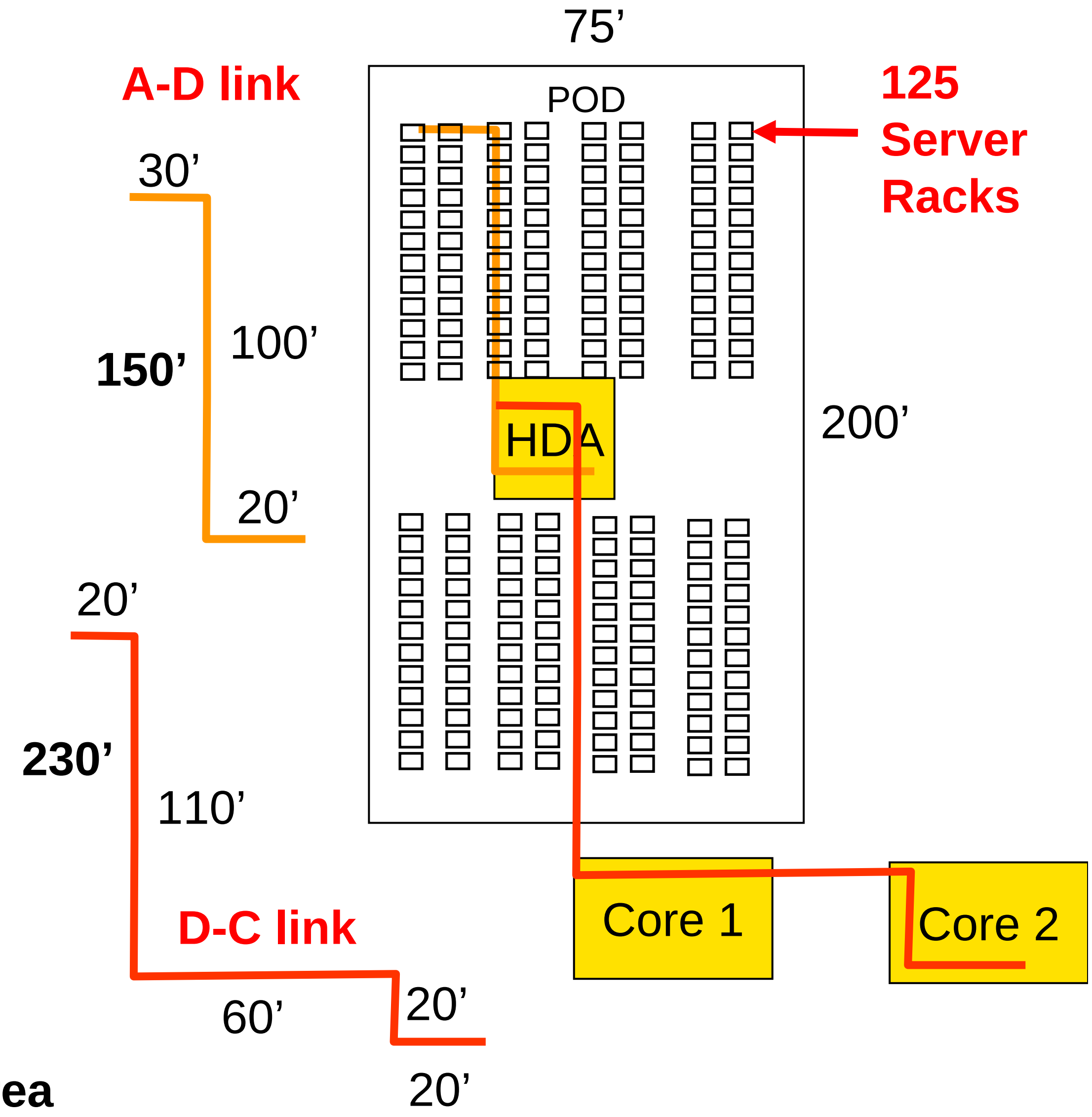
300,000 sq ft new data center



Most Links less than 100 meters

- Max link distance within the pod (Access to Distribution) should be less than 330’or 100 meters.
- This show 150’ of horizontal cabling in this example and there could be 20+’ of vertical cable length.

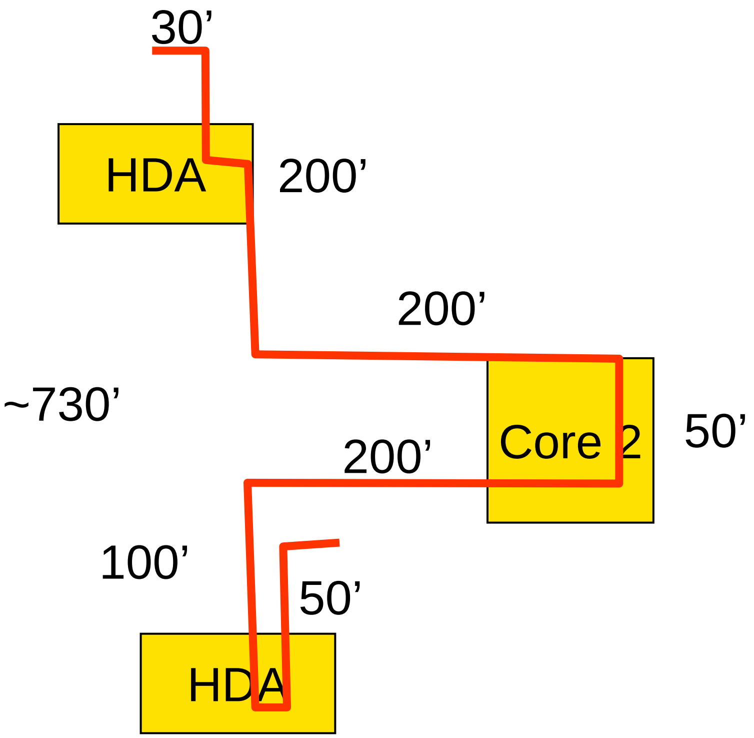
- Max link distance from the pod (Distribution-to-Core) should be less than 330’or 100 meters.
- This show 230’ of horizontal cabling in this example and there could be 20+’ of vertical cable length.

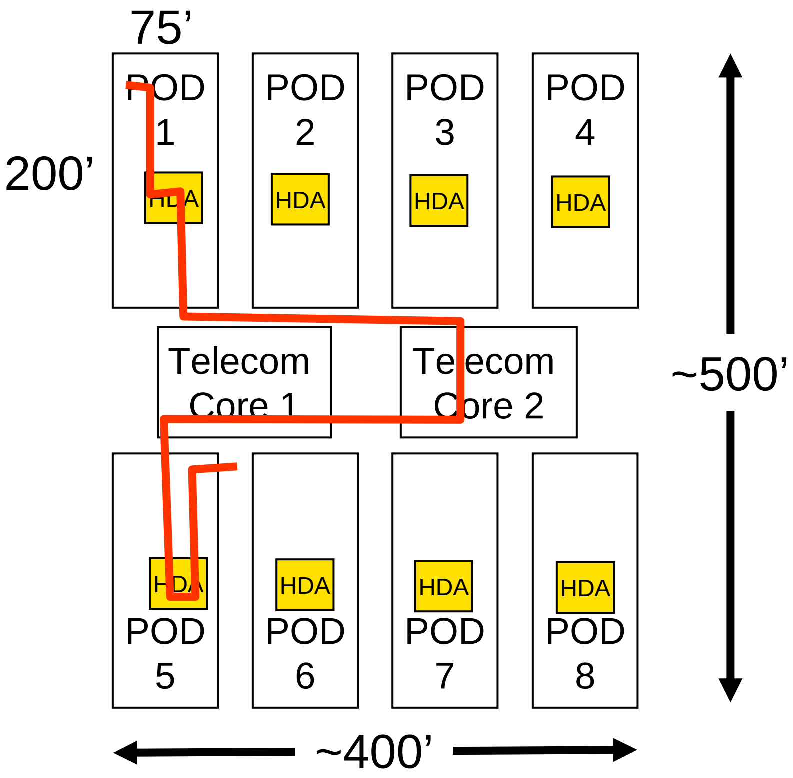


HDA = Horizontal Distribution Area

Inter-Pod distances quickly exceed 100 meters

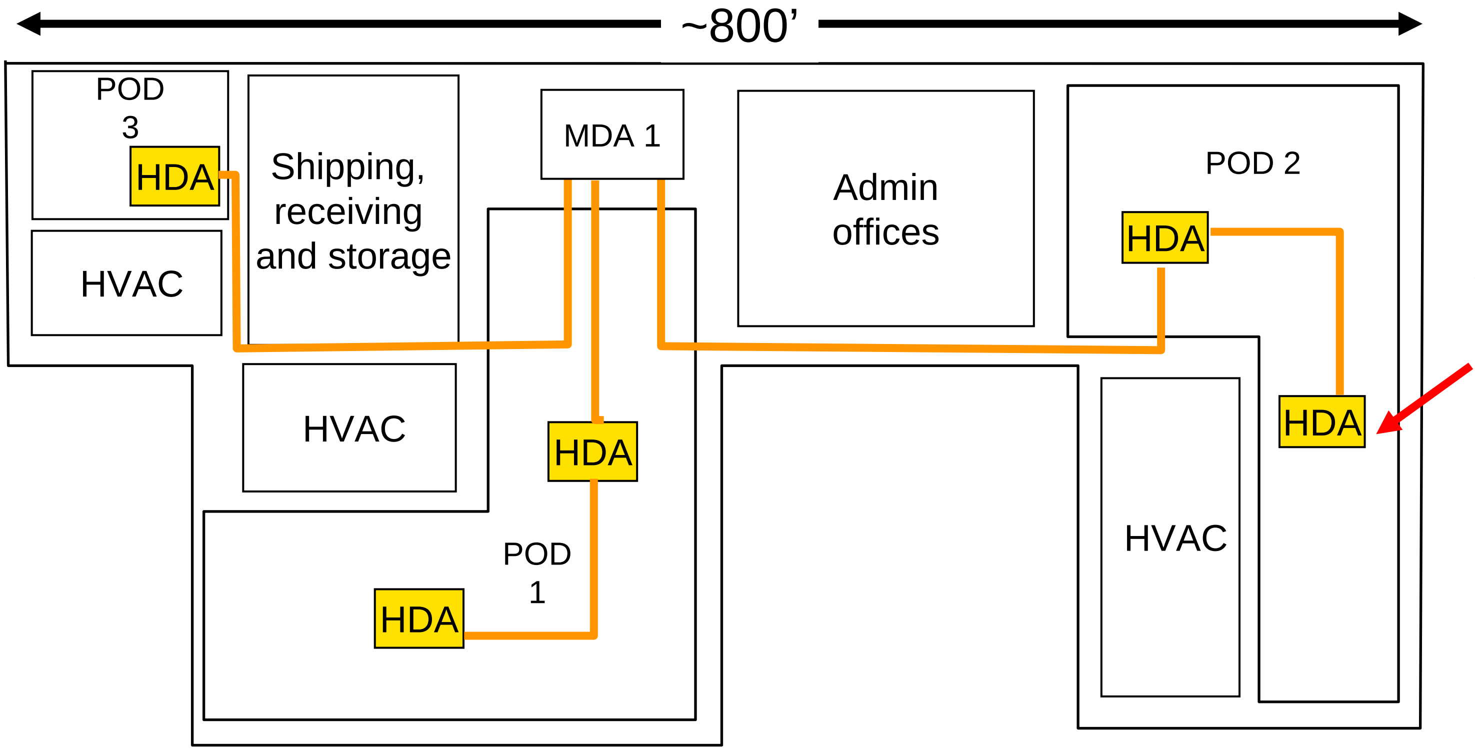
- If a direct link from pod to pod is needed, it quickly exceeds the 100 meters and may go through 3 patch panels
- So use singlemode





Existing Data Centers

- Existing Data Centers usually have grown organically and work around many obstacles
- Multifloor designs, retrofitted HVAC and cabling, and very large, irregular HDAs cause links to exceed 100 meters
  - PODs/HDAs larger than 30,000 sq ft will require more links over 100 meters
  - Recommend XR to support up to 250 meters on OM4 fiber for these situations



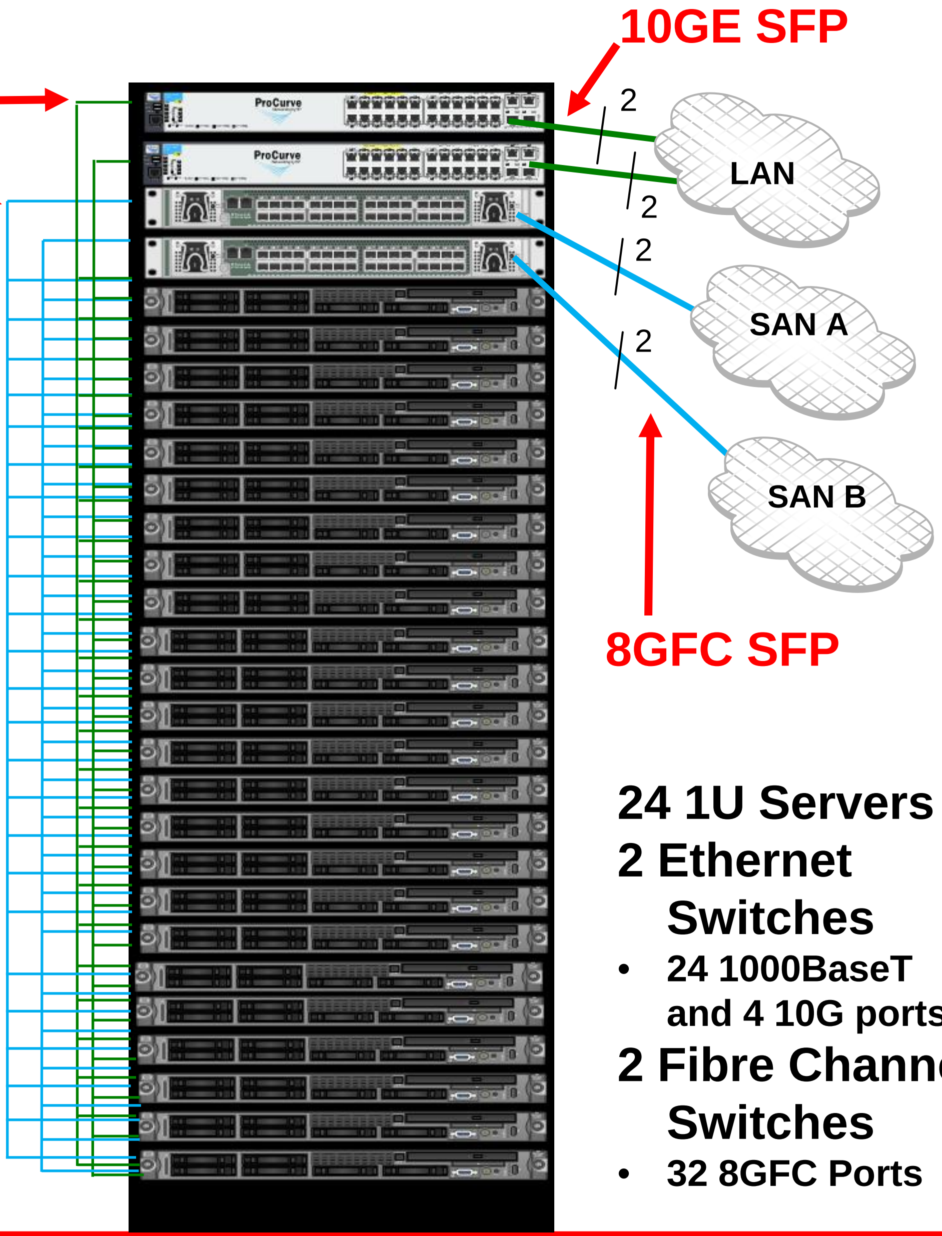
Links from this Distribution Layer should be covered by 200 meters of OM3

Access Layer links

- Many new installations using top-of the rack switches that only need a few meters of cabling in the Client-to-Access (C-A) Layer on the left
- Access-to-Distribution (A-D) links on right as uplink ports to Distribution Switches in HDA
- 12:1 Ratio of C-A and A-D
  - 24 Server Links / 2 uplinks
  - Flatman ratio 15.6:1
  - Some Access switches have 48 ports to 2 uplinks for 24:1 ratio
- Blade Servers with integrated switch go straight to Distribution
  - C-A links are internal

1000 BaseT

4GFC



24 1U Servers

2 Ethernet Switches

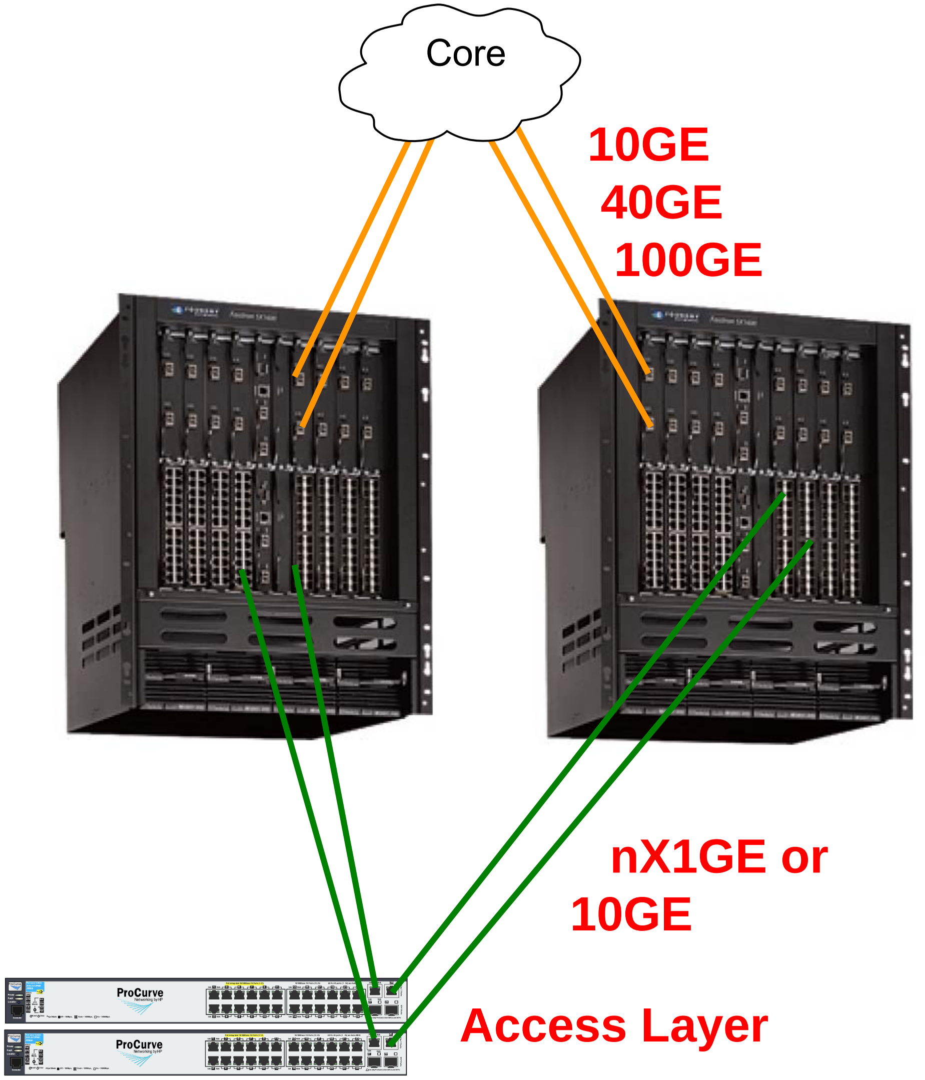
- **24 1000BaseT and 4 10G ports**

2 Fibre Channel Switches

- **32 8GFC Ports**

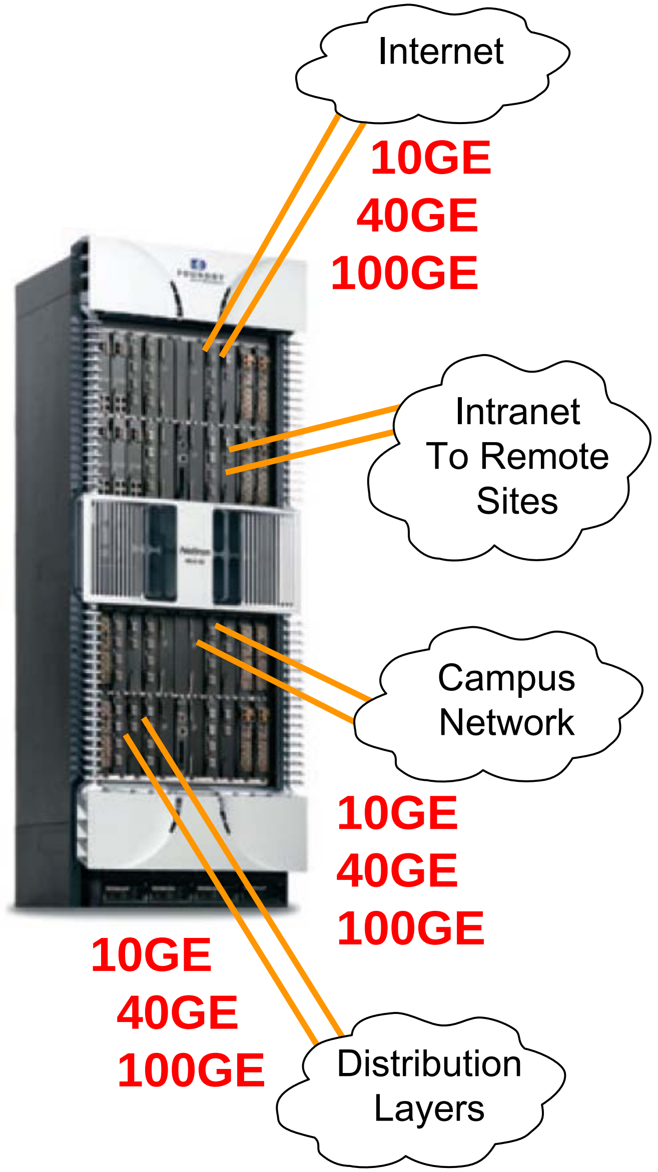
Distribution Layer Links

- Distribution Switches interconnects Access Layer and Core Layer
- Each Distribution Switch has 2 10GE (A-D) links to each server rack
- If 125 Server Racks in POD, then 500 A-D links to POD
- If A-D:D-C is 5.3:1, then there will be ~94 links from Distribution to Core



Core Layer Links

- The Core layer holds the largest modular switches connecting the Distribution layer to other networks
- About 100 links from each Distribution Layer in this example or 800 links to Core
- Other links go to offsite locations or outside of building and are usually singlemode fiber



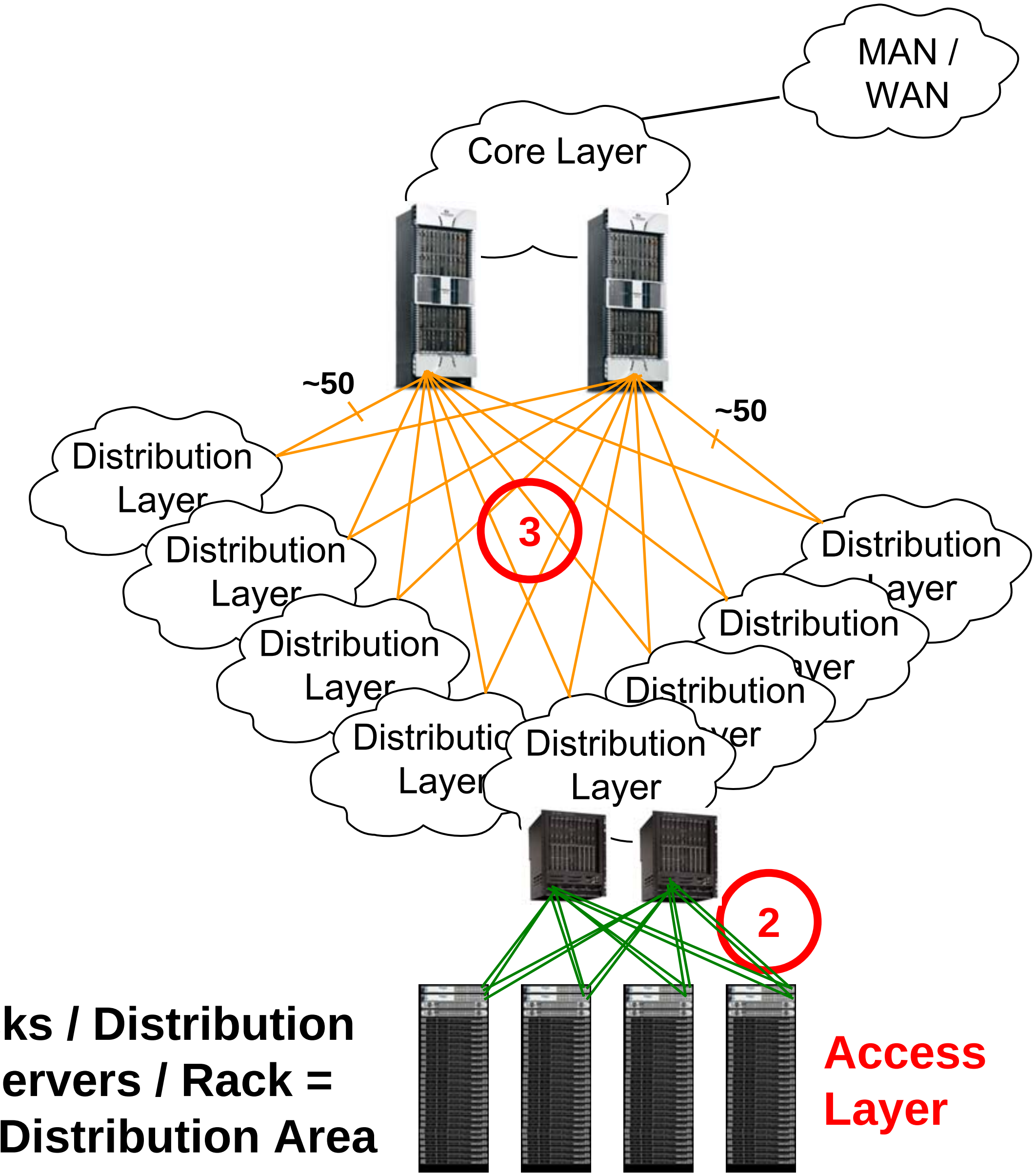
Link Ratios

1. Client-to-Access Links =

   48,000 for 24,000 servers

2. Access-to-Distribution Links = 500 links / Distribution Layer
   - 4000 A-D Links from 8 PODs

3. Distribution-to-Core Links = 800 Links
   - At a 5:1 ratio



125 Server Racks / Distribution Layer * 24 Servers / Rack = 3,000 Servers / Distribution Area

Switch Ratios

1. **Access Switches 1,000 – 2,000**
   48,000 C-A Links / 24 links / Switch = 2,000 Access Switches
   48,000 C-A Links / 48 links / Switch = 1,000 Access Switches
   4,000 A-D Links / 2 Links / Switch = 2,000 Access Switches
   4,000 A-D Links / 4 Links / Switch = 1,000 Access Switches
2. **Distribution Switches – 20-140**
   4,000 A-D Links / 2 10GE ports / line card = 2,000 Cards
   4,000 A-D Links / 4 10GE ports / line card = 1,000 Cards
   4,000 A-D Links / 8 10GE ports / line card = 500 Cards
   4,000 A-D Links / 16 10GE ports / line card = 250 Cards
   800 D-C Links / 8 10GE ports / line card = 100 Cards
   800 D-C Links / 4 40GE ports / line card = 50 Cards
   800 D-C Links / 8 40GE ports / line card = 25 Cards
   Total Line Cards needed range from 300 to 2200
   At 16 line cards / chassis – need from 20 to 140 chassis
3. **Core Switches - 2 – 4 for D-C Links**
   800 D-C Links / 8 10GE ports / line card = 100 Cards
   800 D-C Links / 4 40GE ports / line card = 50 Cards
   800 D-C Links / 8 40GE ports / line card = 25 Cards
   Total Line Cards needed range from 25 to 100
   At 32 line cards / chassis – need from 2 to 4 chassis for D-C Links
   More Chassis needed to connect to other networks

1000s of
Access Switches

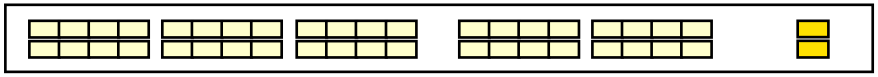
10s to 100+
Distribution Switches

2 to 10
Core Switches

***# of ports on a line card depends on oversubscription and backplane bandwidth**

Two Types of Switches

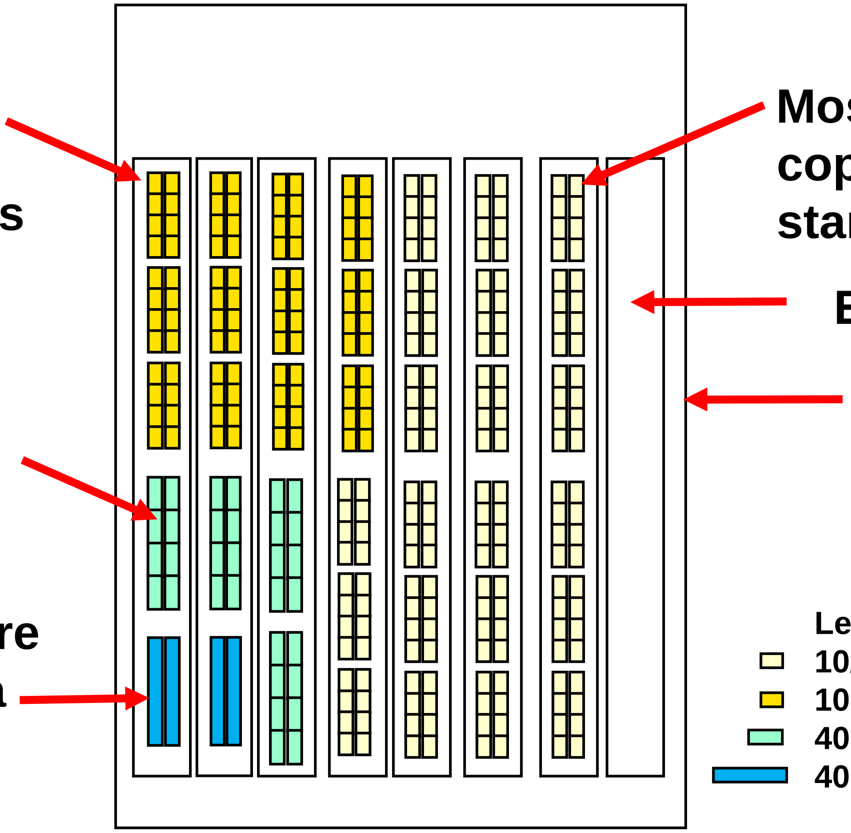
- Fixed Port – set type of ports in Switch – usually 1U



10/100/1000 BaseT ports or SFP

2 10GE Optical Uplink Ports

- Modular Chassis - very flexible depending on needs



10GE ports to servers and Access Switches

40GE MM ports to blade servers and Access Switches

40GE SM ports to Core Switches across data center or campus

Mostly 10/100/1000 copper ports to standalone servers

Empty Slot

Typically 8 Full length or 16 half length Port Blades

Legend:

□ 10/100/1000 Port (SFP or RJ45)

🟨 10G SFP+

🟩 40GE MM QSFP

🟦 40GE SM CFP

Transceiver Dimensions

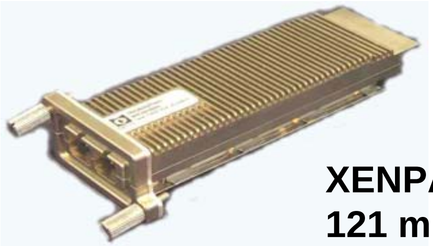
- 40GE SM PMD probably 4X wider than 40GE MM PMD

SFP+ / SFP
57mm L x 14 mm W x 9mm H

Likely Size of 40GE MM transceiver

XFP/ QSFP
75mm L x 18 mm W x +9mm H

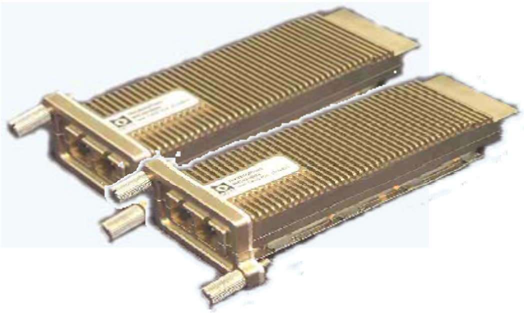
X2/XPAK
~76mm L x 36 mm W x +12/11mmH



XENPAK
121 mm L x 36 mm W x 11.98mm H

40GE SM
Probably CFP Size or
Double XENPAK

121 mm L x 72 mm W x 11.98mm H

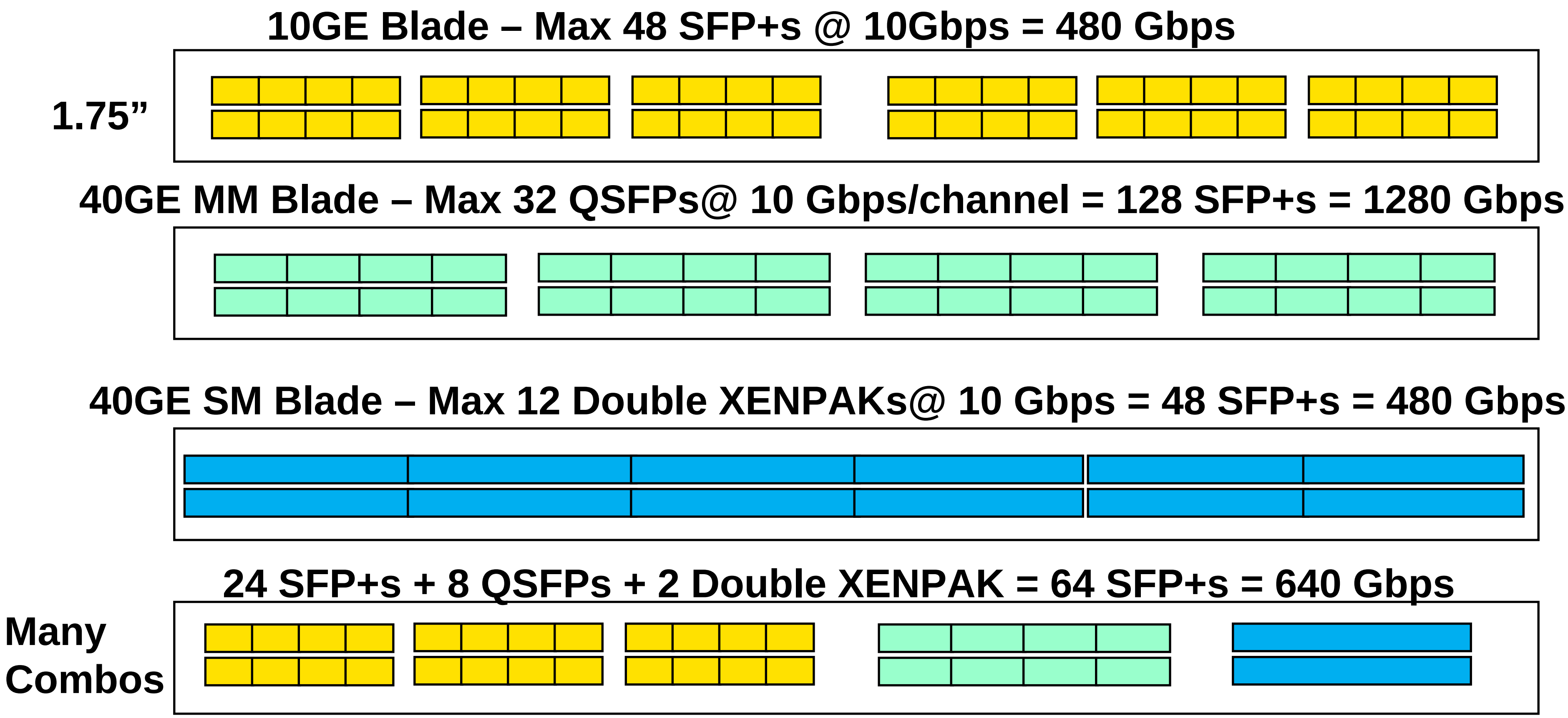


Size of First Generation CWDM module according to Chris Cole email (could be 1X XENPAK), second gen would probably be QSFP!

Probable Size of 100GE MM and SM Module is unknown?

1 U Blade or 1U Switch Configurations

- Various Configurations on a 1U Switch or Line Card



10GE Blade – Max 48 SFP+s @ 10Gbps = 480 Gbps

1.75”

40GE MM Blade – Max 32 QSFPs@ 10 Gbps/channel = 128 SFP+s = 1280 Gbps

40GE SM Blade – Max 12 Double XENPAKs@ 10 Gbps = 48 SFP+s = 480 Gbps

24 SFP+s + 8 QSFPs + 2 Double XENPAK = 64 SFP+s = 640 Gbps

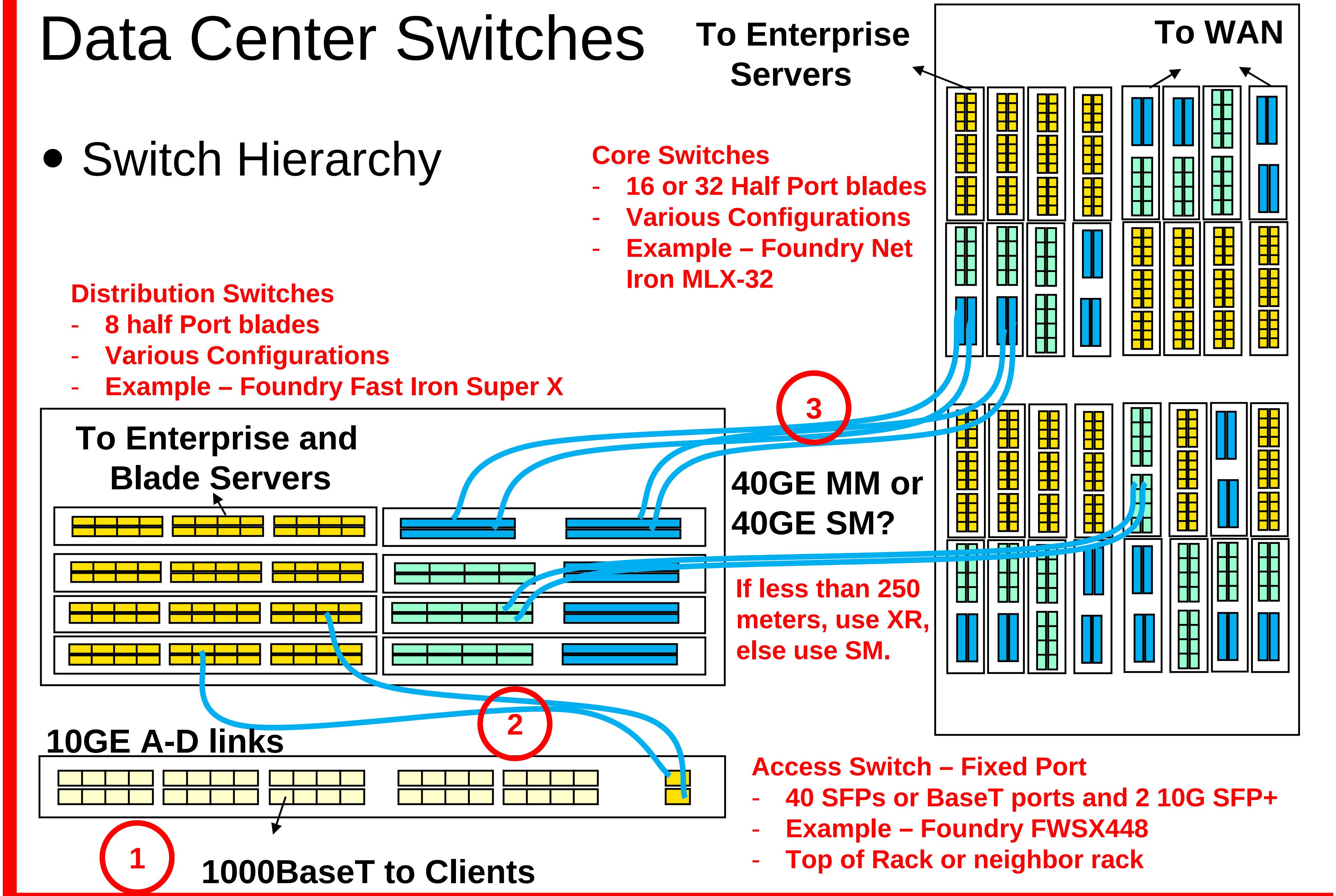
Many Combos

17.5” of a 1U server available for transceivers

***# of ports on a line card depends on oversubscription and backplane bandwidth. The number of ports and bandwidth is cut in half for a 1/2U blade**

Data Center Switches

- Switch Hierarchy



Core Switches
- 16 or 32 Half Port blades
- Various Configurations
- Example – Foundry Net Iron MLX-32

Distribution Switches
- 8 half Port blades
- Various Configurations
- Example – Foundry Fast Iron Super X

If less than 250 meters, use XR, else use SM.

Access Switch – Fixed Port
- 40 SFPs or BaseT ports and 2 10G SFP+
- Example – Foundry FWSX448
- Top of Rack or neighbor rack

Cost To User For Modular Switch

- Chassis
  - Power Supplies, Processor Cards, Switch Cards
- Line Card
  - Mixture of port types depending on layer
  - Volume is in 1000BaseT, then 1GE SFP, then 10GE
  - If special card is required for SM module, volume is less and cost more
- Modules
  - Copper, Optical SR, Optical XR?, Optical LR, Optical ER
- Software Licenses
  - Security Features, Advanced Features
- Warranty and Service
- Power, rack and space

Cost to User for a Link

Volume is key to cost

Link costs after chassis and overhead include:

- Fraction of a line card or switch
  - Cost of line card / number of Ports
  - More ports on line card leads to lower cost
- Modules
- Fiber
- Installation

802.3ba has large influence on the module cost

Cost = Line Card + Modules + Fiber + Installation

Conclusion

- Most volume is in short C-A links at 1000BaseT and these won’t be using 40GE or 100GE for a long time
  - C-A links are within the rack or a few racks over
  - Blade Servers may have Access switch built into chassis
- A-D links are moving to 10GE and 40GE is on the way
- D-C and core links are nx10GE now and moving to 40GE and 100GE for possibly longer distances
- Modern data centers that use a POD architecture with cells below 20,000 sq ft shouldn’t need XR links for A-D
- For the odd long link, we need an informative Annex for XR with a goal of 250 meters over OM4 fiber
- Singlemode links are needed for direct inter-POD links

BROCADE

™

Thank You