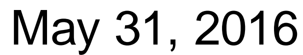
Emerging Optical interconnect Technology for the Cloud



Topics to be Covered

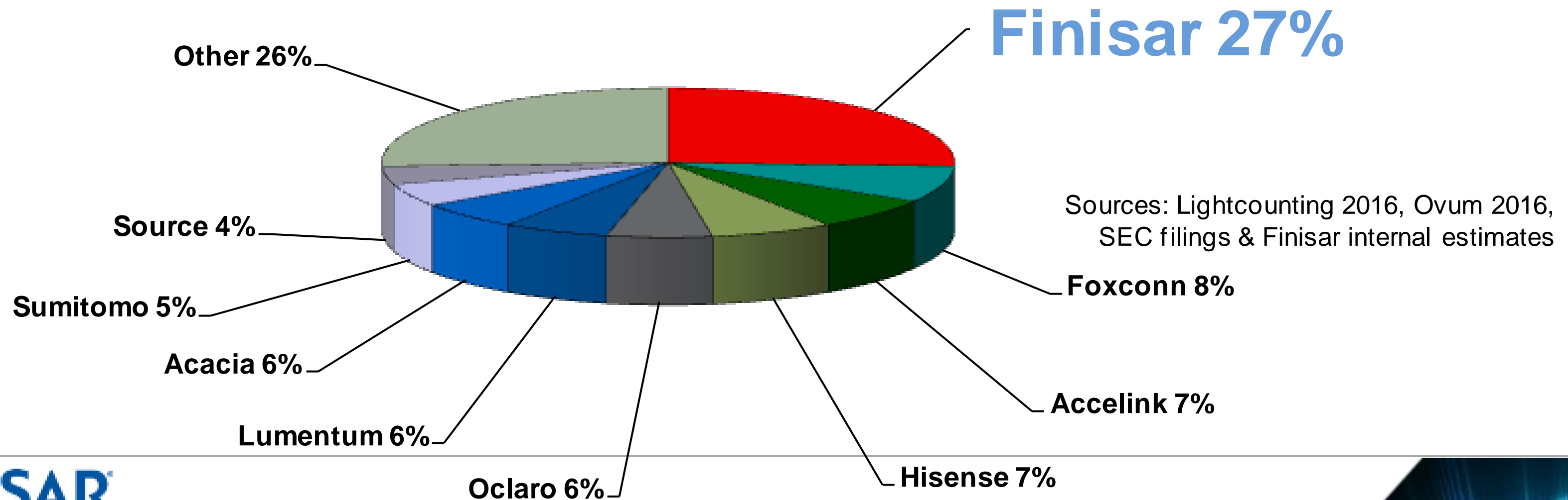
- Introduction to Finisar
- Basic Optical Packaging Needs
- Trends in Optical packaging

The goal of this session is to highlight recent and emerging advances in optical interconnect packaging: topics to include standards based form factors in data centers and their differentiators; mid-board optics; high-density integration; photonics manufacturing initiatives & ecosystem.

Who is Finisar? Largest provider of Fiber Optic Transceivers

Wide portfolio of products reaching from:

<1m to >120km ; 1G to 56G/optical channel)
Multi-Mode (VCSEL), Traditional Single mode, Silicon Photonics, Bidirectional, Short Wave WDM, CWDM, L and C band WDM



Metrics of Optical Interconnect

- Data at a distance
  - Measured by Bandwidth (BW) x Reach product
  - Example: 25GB/s x 12 ch x 100m = 30 TBitm/s
- Signal integrity
  - Measured in **B**it **E**rror **R**atio (BER) (number or errors / number of bits transmitted)
  - Examples: 1E-15 (error free) –or-- 3E-5 for Forward Error Correction (FEC) enabled links
- Density
  - Measured in BW / RU (**R**ack **U**nit)
  - Example: QSFP28… 36 x 100G / RU = 3.6TB/RU
- Cost
  - Measured in $US / Gbits (full duplex)
  - Example: $US600 / 300G = $2/GBit/s
- Power dissipation
  - Measured in pJ/bit (mW/GBit/s)
  - Example: 7W / 300Gbit = 23pJ/bit

RU = 19" **R**ack **U**nit (1.75" high x ~16")

Optical Packaging

Last 20 years fiber optics have been :

Largely **P**lugable

Examples: GBIC, SF**P**, XF**P**, QSF**P**, CX**P**, CF**P**, CF**P**2, CF**P**4, CF**P**8, uQSF**P**, QSF**P**-DD, etc.

Some Embedded: "Optical Engines":

Examples: SNAP12, POP4, BOA10, uPOD, MiniPOD, BOA25, MBOM, etc.

<u>Going forward</u> the Cloud Computing Data Center market needs:

Higher BW Density

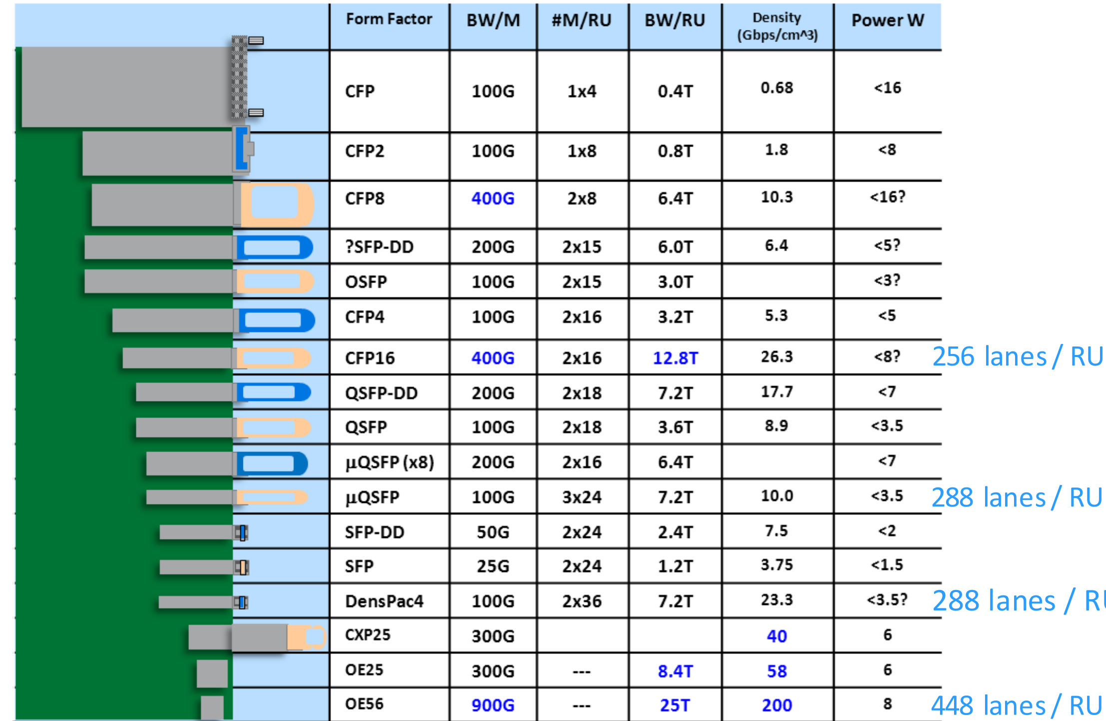
Lower Power Dissipation per bit

Lower Cost per bit

Higher Bandwidth Density … more bits/mm^2

- Plugable modules can deliver up to 12.8TB/sec / 1RU
- Embedded or 'On Board Optics' can provide 28TB / 1RU
  - Limit of embedded modules around an ASIC set by losses in the host board.
  - Thermal Management limited by air flow blockage by fibers
- Higher densities achievable through aggregation of optics and ASIC

25G and 56G Form Factors



Form Factor	BW/M	#M/RU	BW/RU	Density (Gbps/cm^3)	Power W
CFP	100G	1x4	0.4T	0.68	<16
CFP2	100G	1x8	0.8T	1.8	<8
CFP8	400G	2x8	6.4T	10.3	<16?
?SFP-DD	200G	2x15	6.0T	6.4	<5?
OSFP	100G	2x15	3.0T		<3?
CFP4	100G	2x16	3.2T	5.3	<5
CFP16	400G	2x16	12.8T	26.3	<8?
QSFP-DD	200G	2x18	7.2T	17.7	<7
QSFP	100G	2x18	3.6T	8.9	<3.5
µQSFP (x8)	200G	2x16	6.4T		<7
µQSFP	100G	3x24	7.2T	10.0	<3.5
SFP-DD	50G	2x24	2.4T	7.5	<2
SFP	25G	2x24	1.2T	3.75	<1.5
DensPac4	100G	2x36	7.2T	23.3	<3.5?
CXP25	300G			40	6
OE25	300G	---	8.4T	58	6
OE56	900G	---	25T	200	8

Courtesy of William Wang, Finisar

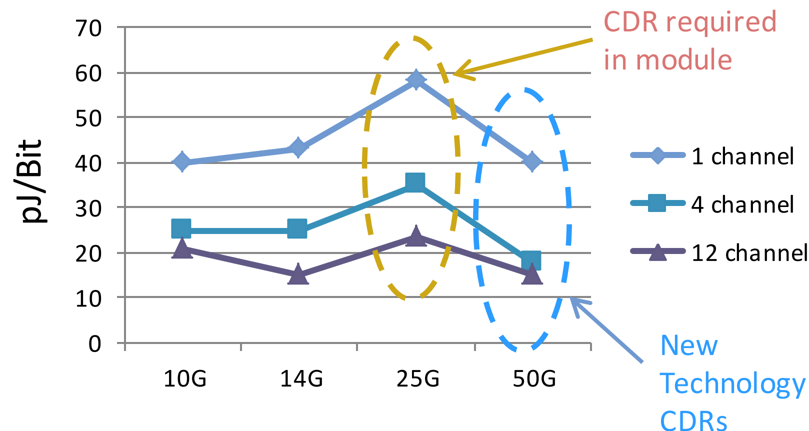
RU = 19" Rack unit (1.75" x ~16")

Lower Power… lower pJ/Bit

- ⬆ Higher speeds often require more power due to electrical losses in the PCB
- ⬆ CDRs in the module became a required function at 25G. Often unused.

- ⬇ At 56G a new IC node reduces power.
- ⬇ Higher Bit rate amortizes the power consumed

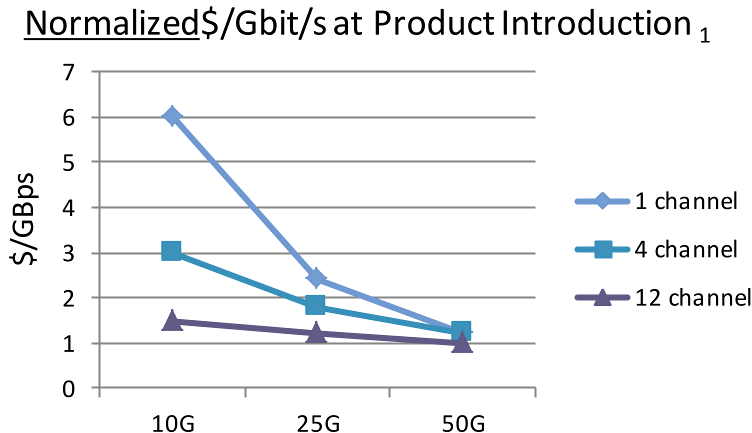
- ☑ So far the combination has been a ‘wash’:
  - Increased power due to board E-loss at higher speed is made up by increased speed and more efficient technology node.



Need to reduce power/bit by increasing speed without increasing the PCB losses…
→ Eliminate crossing the PCB with the high speed trace.

Lower Cost… lower **$/Gbit**

- Higher channel count provides lower cost per bit through integration



- Higher data rates (eventually) provide lower cost per bit
  - Example: 1G module costs ~ 4G module costs ~ 8G module costs today.
  - Design of product needs to be for cost parity with today's products.

- Reducing path loss leads to lower cost through simplification
  - Less need to correct distortions

FINISAR

1. Normalized to Expected 56GPAM introduction cost

Summary… Trend for Fiber Optic Packaging

- **Historical: Front Panel Plugable Optics: This was the standard for 15 years**
  - Maximum at <15TBps / RU ?
- **Growth Area: Mid-Board Optics**
  - Increased BW density beyond what plugables can achieve
    - Greater than 25TB / RU is easily achievable
  - Reduced power (reduced trace loss)
    - Elimination of CDR on Optical RX side
(Difficult for plugables due to the long signal trace)
  - Improved Signaling due to shorter traces
    - Ability to offer error free links (as opposed to 25GE SR4 which requires FEC
- **Emerging area for Innovation: First Level Packaging of Optics**
  - Optical Devices integrated on the same level of packaging as the ASIC
  - Anticipated gains:
    - Further improvement in BW density
    - Further reduction in power: Elimination of CDRs & lower E-losses
    - Lower cost per bit Simplified product

Packaging needs for Large Scale (Cloud) data Centers

- Fine pitch 30GHz package interconnect
  - 500um pitch
  - Sockets but more importantly: array interconnect

- High Speed Fine Pitch Substrates
  - Glass (with through vias)
  - Multi layer
  - Silicon w/ low cost high density TSV

- Reflowable transparent plastics for optics
- Precision Molding of optical features

- Optical PCB materials and connectors: reduce fiber congestion

- Cost effective Multi Chip module Technology

© Finisar Corporation Confidential.

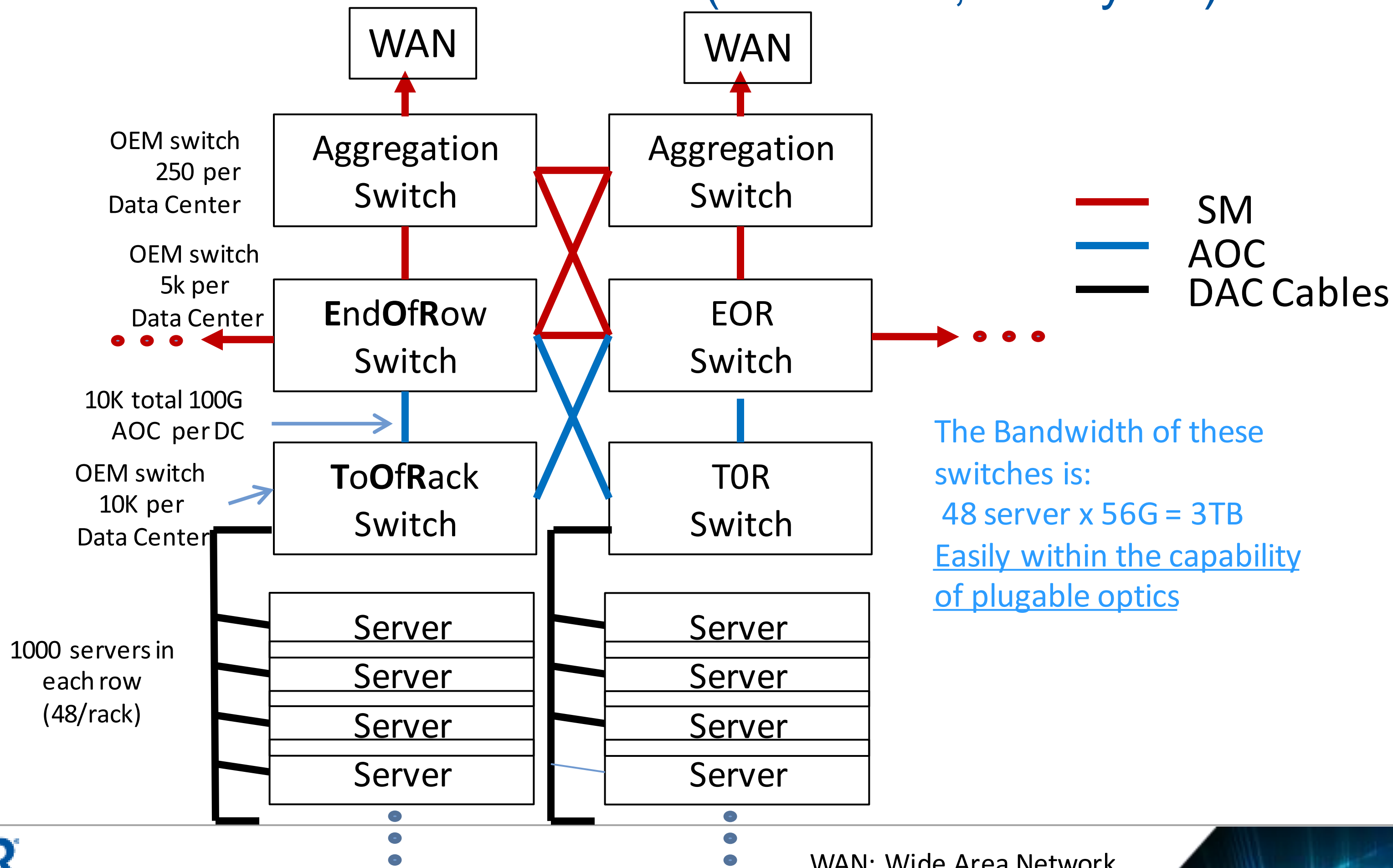
Thank You

© Finisar Corporation Confidential.

Back Up Slides

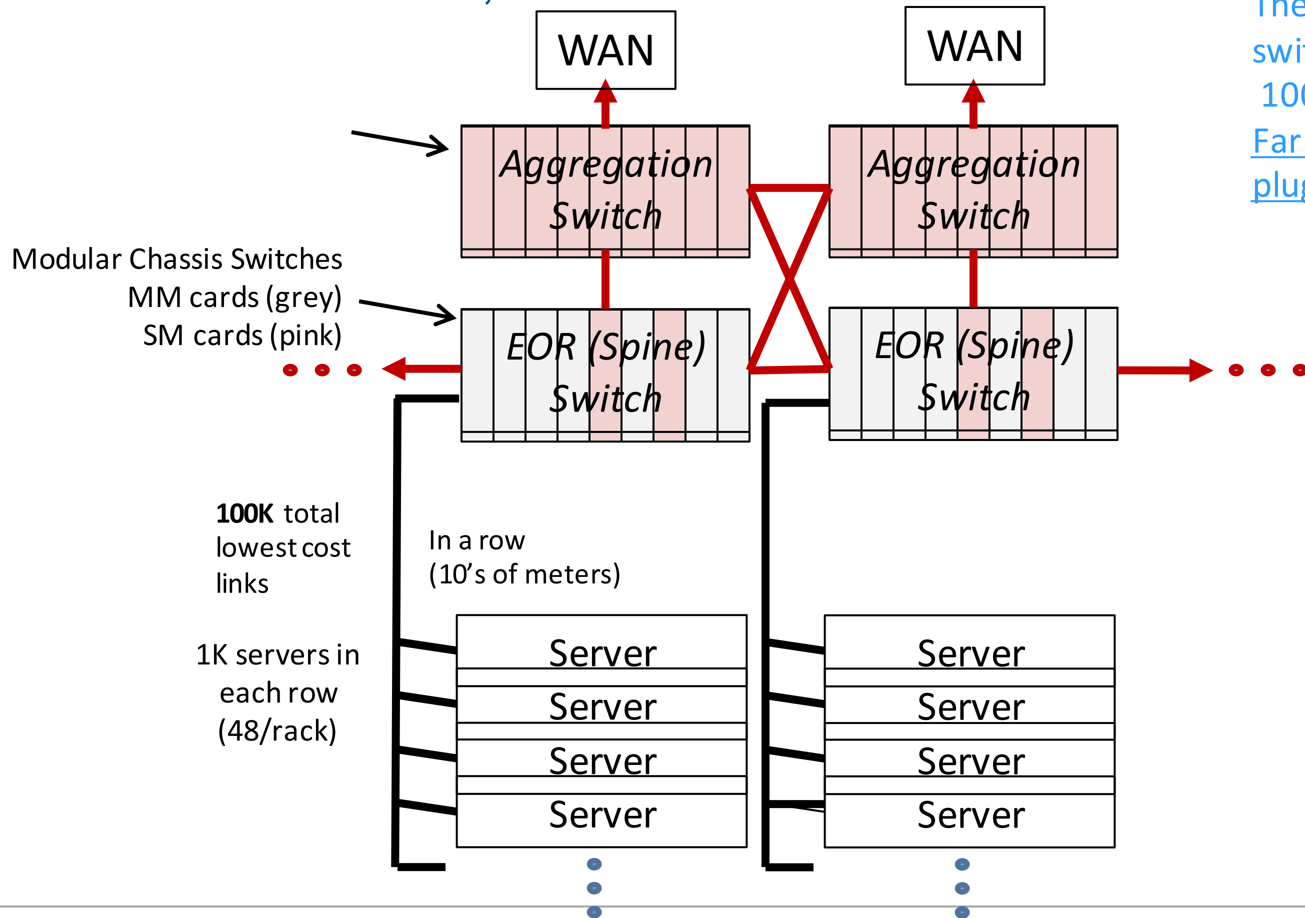
© Finisar Corporation Confidential.

Large scale Data Center Architecture (25G Lane; This year)



WAN: Wide Area Network

Next Gen Architecture; 57.6G Lane



The Bandwidth of these switches is:
1000 server x 56G = 56TB!
Far beyond the capability of plugables.

SM

Lowest Cost
Lowest Power
(MM?)

(Est. # of low cost short reach links
is 8x the # of longer reach links
based on # of
T1 and T2 switches)

WAN: Wide Area Network

Large Scale Data Center:

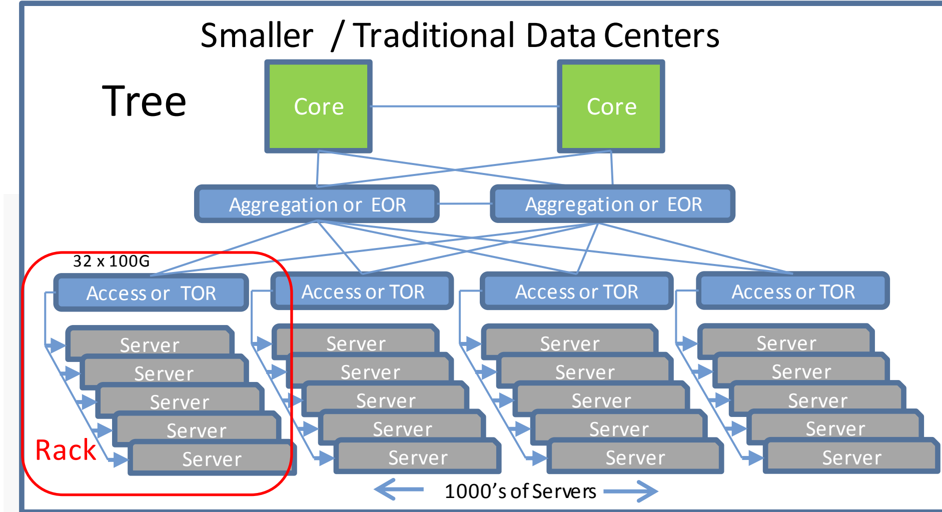
Building Block is 'The Rack'

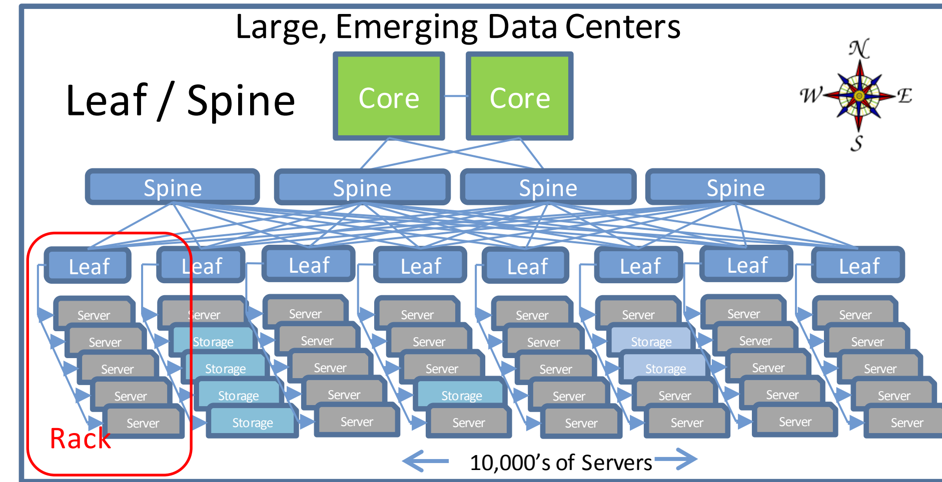
Object : Connect as many racks as possible

- High fan-out from switches (high radix)
- Over-subscription from Spine → Leaf(1:3)
- High 'East-West' traffic; Big, flat, homogenous

Spine Architectures are limited by the number of ports in the Spine Switch

- 2→10m for Servers to TOR/Leaf mostly DAC, some AOC at 10G; optics at 25G
- 20→300m MM runs for Leaf to Spine: Transceivers & AOC
- 100→500m+ SM runs from to spine to core
- Latency not as important; FEC acceptable
  - Time of flight on longer reaches dominates
- Non-standard OK; Multi source desired





FINISAR

TOR = Top Of Rack; EOR = End of Row; DAC= Direct attach Copper; AOC=Active Optical Cable; FEC = Forward Error Correction

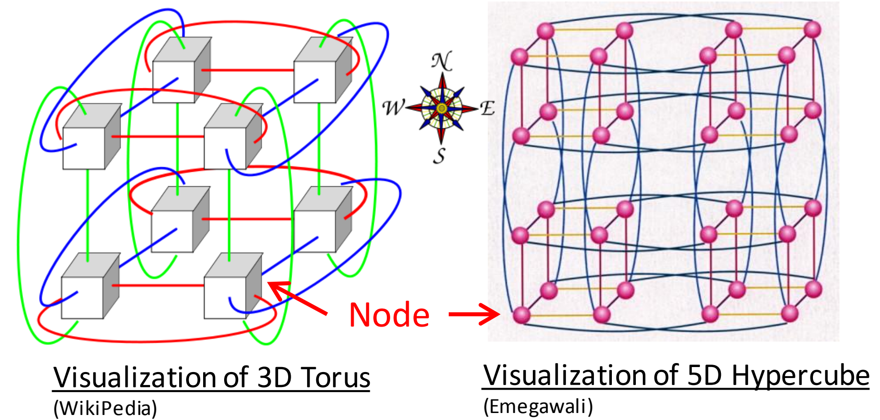
High Performance Computing:

The building block: <u>The 'Node'</u>

Object : <u>All to all connection</u>

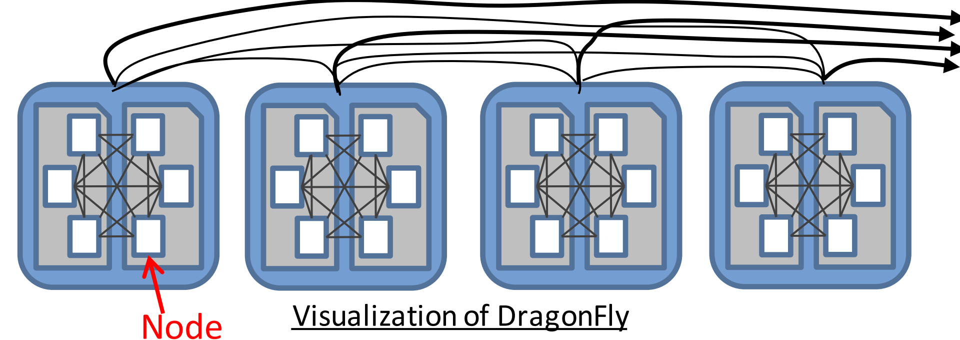
- Multi D Torus Structures; Hyper Cubes; Dragonfly Interconnect (N,S,E,W connection meaningless)
- Very Low over-subscription
- Breakout to processor occurs within the node
- < 2m MM runs to within the node (BOAs)
- 70-100m runs to far end of the machine possible. Most runs less than 30m (AOCs)
- Latency <u>very</u> important; FEC not acceptable
  - Reducing the latency from 100ns to 80ns improves system performance 25%
- Non-standard is OK but Multi source desired.
- High use of Embedded Optics (increases BW density off the card to get closer to all to all)

Sample Connection Scenarios for HPC



Visualization of 3D Torus (WikiPedia)

Visualization of 5D Hypercube (Emegawali)



Visualization of DragonFly

FINISAR

BOA= Board mounted optical Assembly (embedded optics); AOC=Active Optical Cable; FEC = Forward Error Correction

So…Key Take Aways…

- AGGREGATE Bandwidth doubles every 3 years
- Rate of Aggregate Bandwidth increase is the same for all form factors
- Cost / Gbps is lower for higher channels counts at the same point in time
- After 14 years the single optical channel part is the lowest cost solution
- 50G PAM4 will be the next data rate on these ramps
- 50G PAM4 expected in 2018
  This may be too soon for a new modulation scheme (we all have a lot to learn)
- Consolidation of plugable form factors is needed
  - > 10 new form factors fractures the market and prevents scaling for cost
- 50G PAM4 does not satisfy all 50G markets.
  - PAM4 requires FEC; Popular FEC coding induces 100ns latency
  - Low latency is required by Supercomputing, and increasingly important to LSDC
  - Implies NRZ or low latency FEC

FEC = Forward Error Correction; LSDC = Large Scale Data Center

Next Generation Products for Large-scale Datacenters and Supercomputers

- For >256 channels in one RU (>14Tbps):
  - Recommend: MM BOA 56G PAM4
- For <256 channels in one RU:
  - Recommend: SM and MM “XYZ-FP” 56GPAM4
- Longer term: 56G NRZ for Supercomputing