Moore's Law and Networking

Andreas Bechtolsheim
Arista Networks Inc

June 4, 2012

Original Prediction made in 1965

ITRS Packaging Roadmap

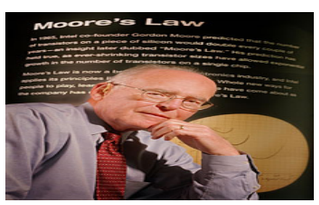
M. Jürgen Wolf

Fraunhofer IZM,
Berlin, Dresden, Germany

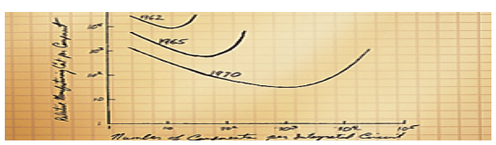
wolf@izm.fraunhofer.de

Fraunhofer IZM

Moore´s Law

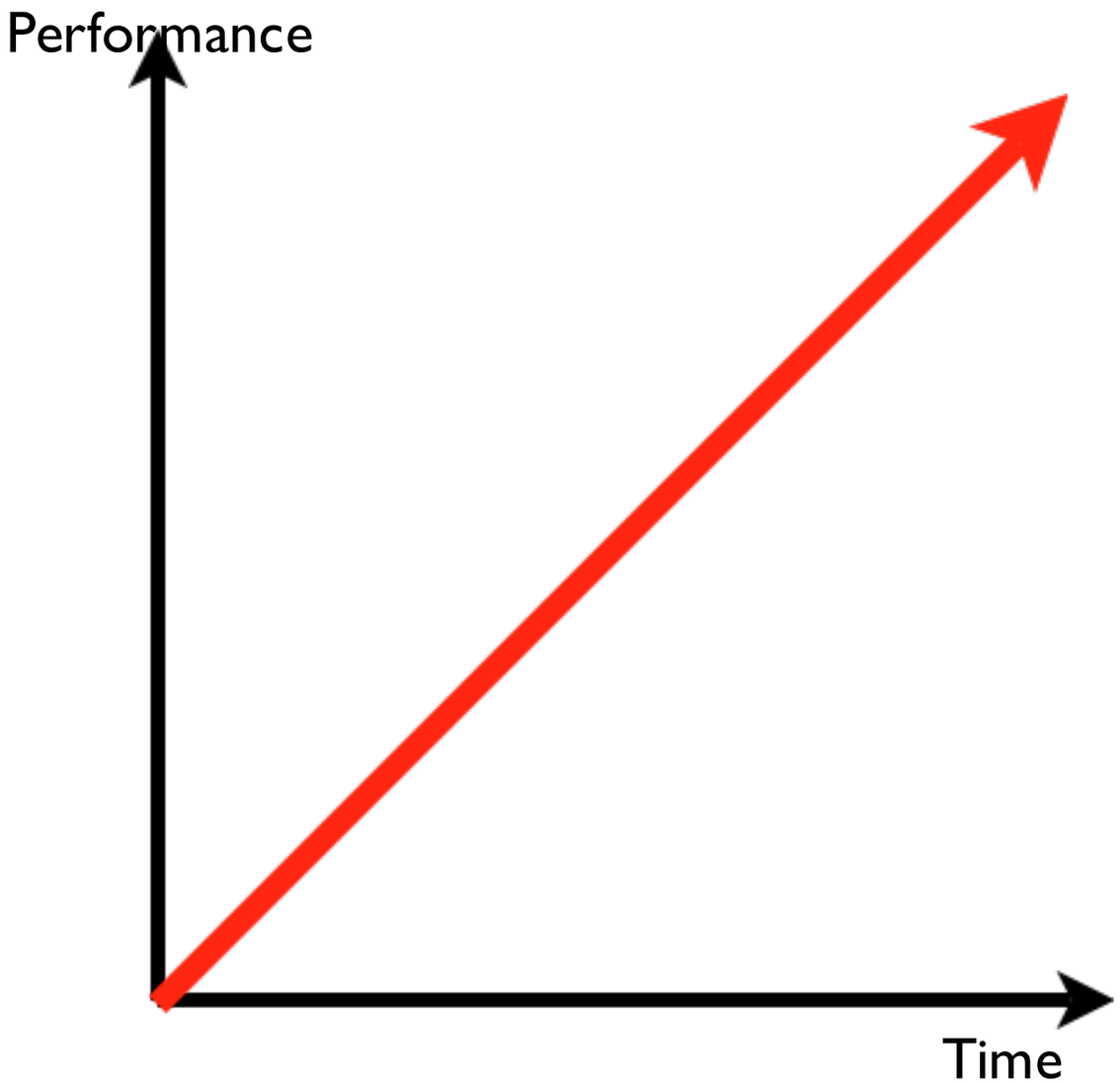


The observation made in 1965 by Gordon Moore, co-founder of Intel, that the number of transistors per square inch on integrated circuits had doubled every year since the integrated circuit was invented. Moore predicted that this trend would continue for the foreseeable future.



1975 Revision that became known as Moore's Law: The Number of Transistors will double every 2 years

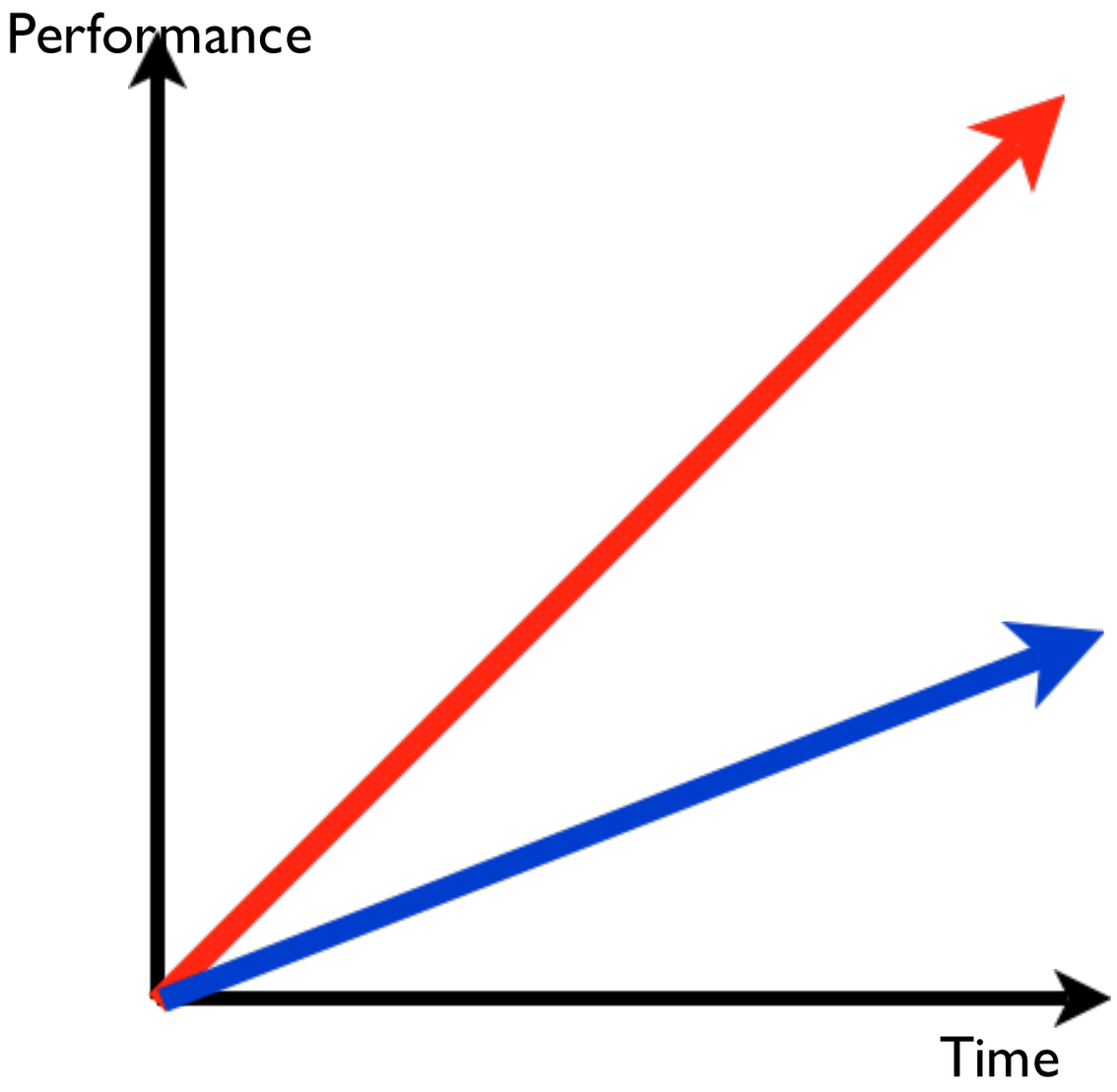
Moore's Law and Networking



CPU: 2X/2Y =

64X/12Y

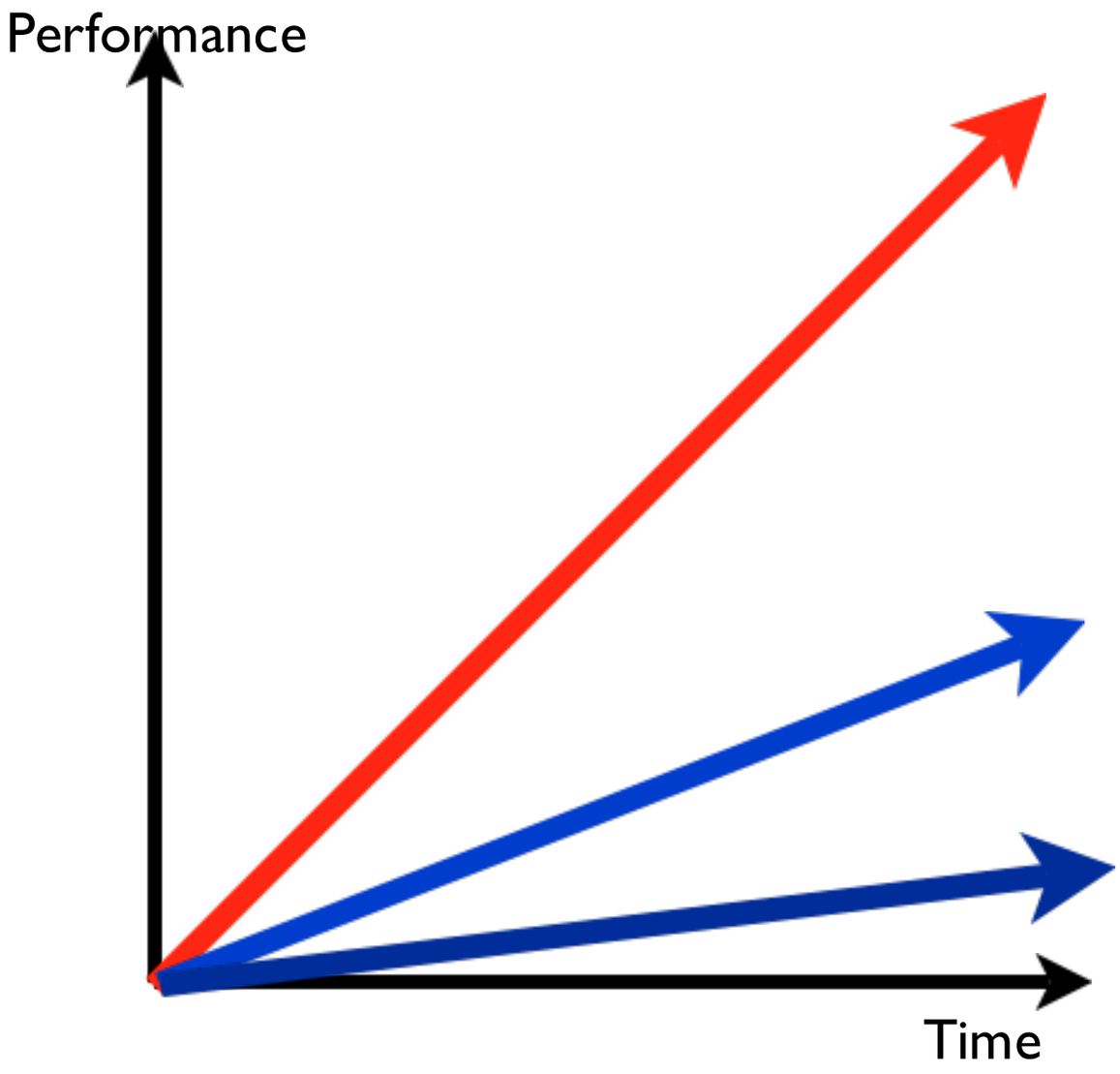
Moore's Law and Networking



CPU: 2X/2Y =
64X/12Y

1GigE-10GigE: 10X/12Y

Moore's Law and Networking

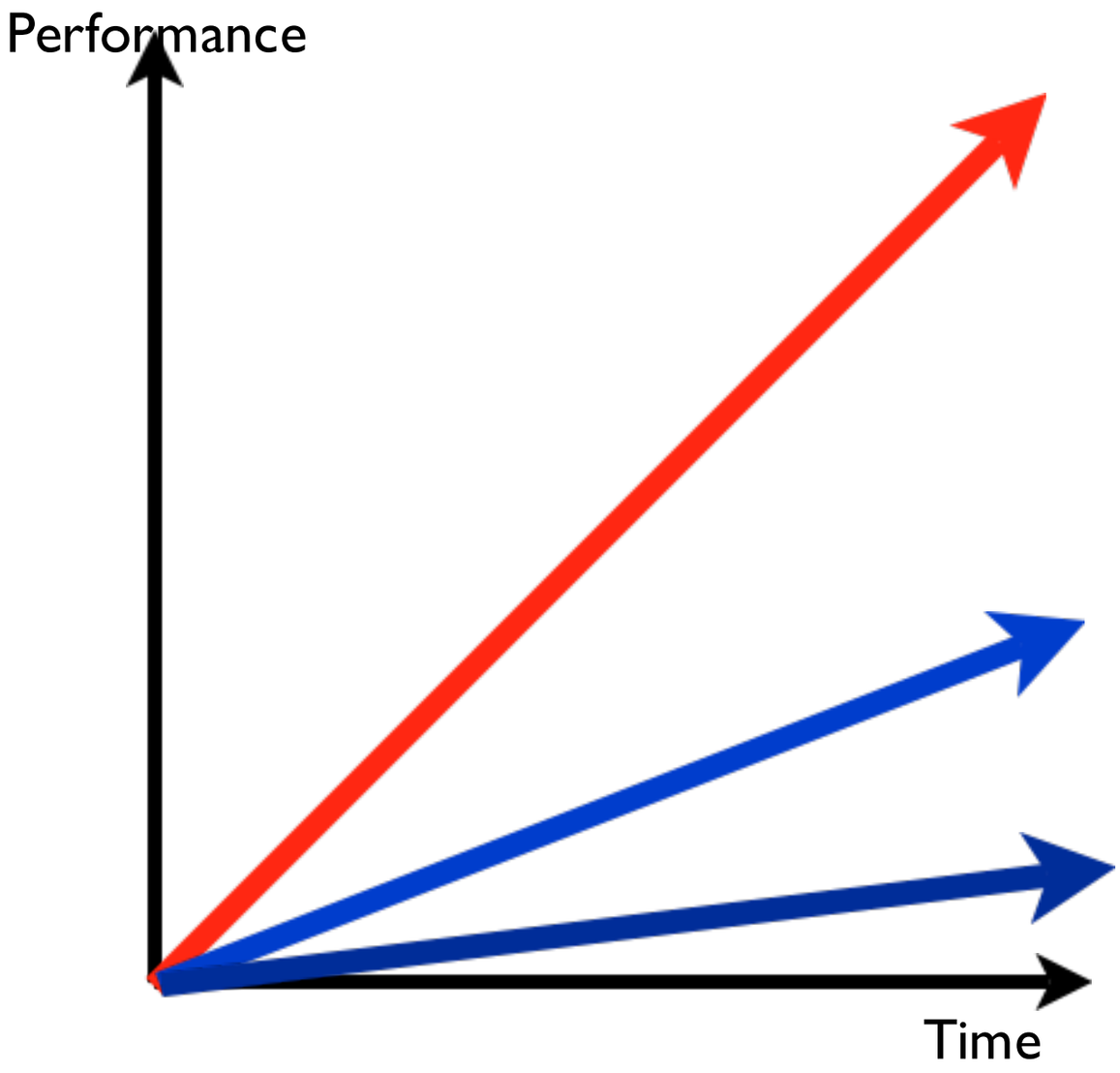


CPU: 2X/2Y =
64X/12Y

1GigE-10GigE: 10X/12Y

WAN Routers:
4X/12Y

Moore's Law and Networking



CPU: 2X/2Y =
64X/12Y

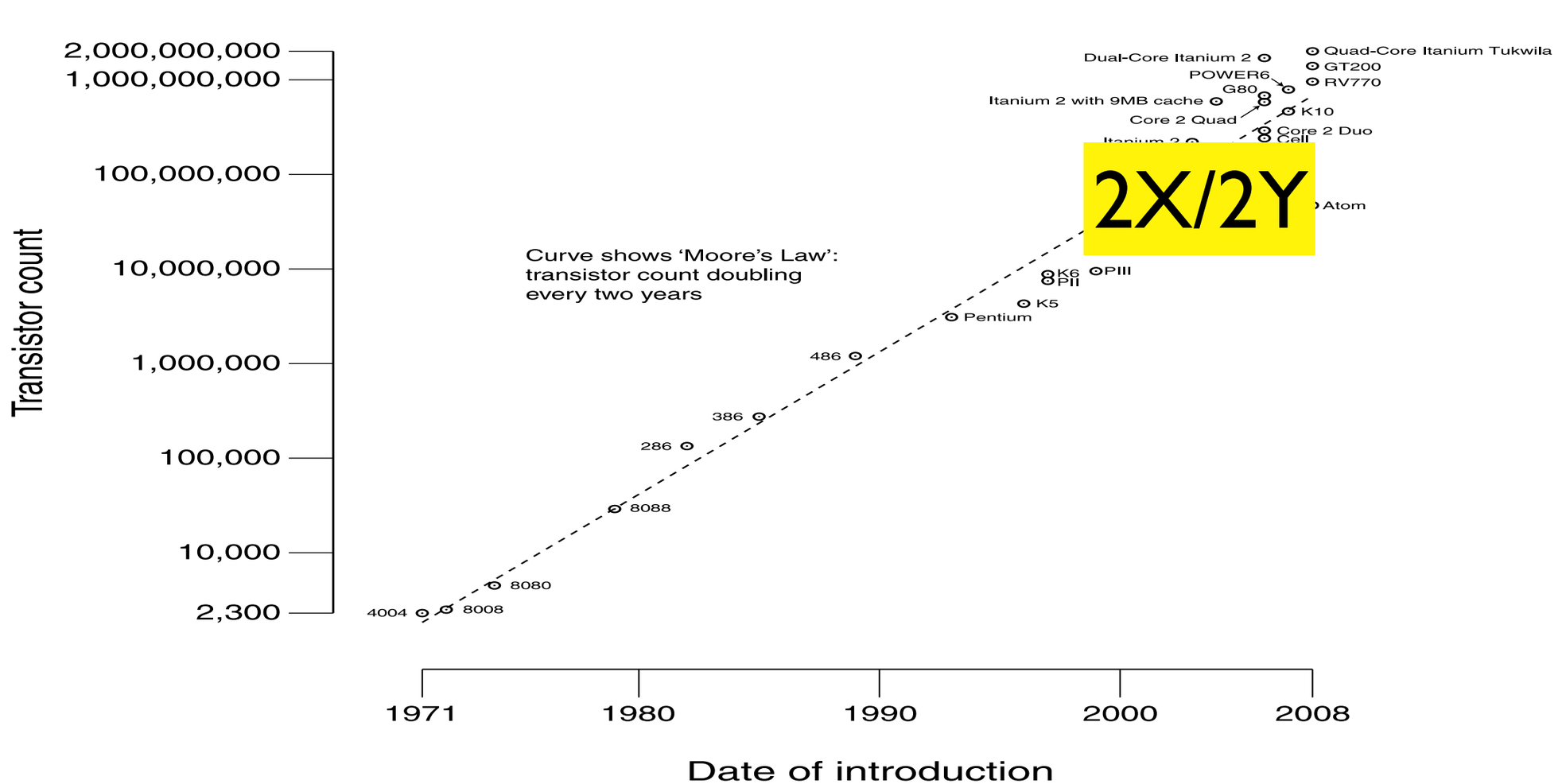
1GigE-10GigE: 10X/12Y

WAN Routers:
4X/12Y

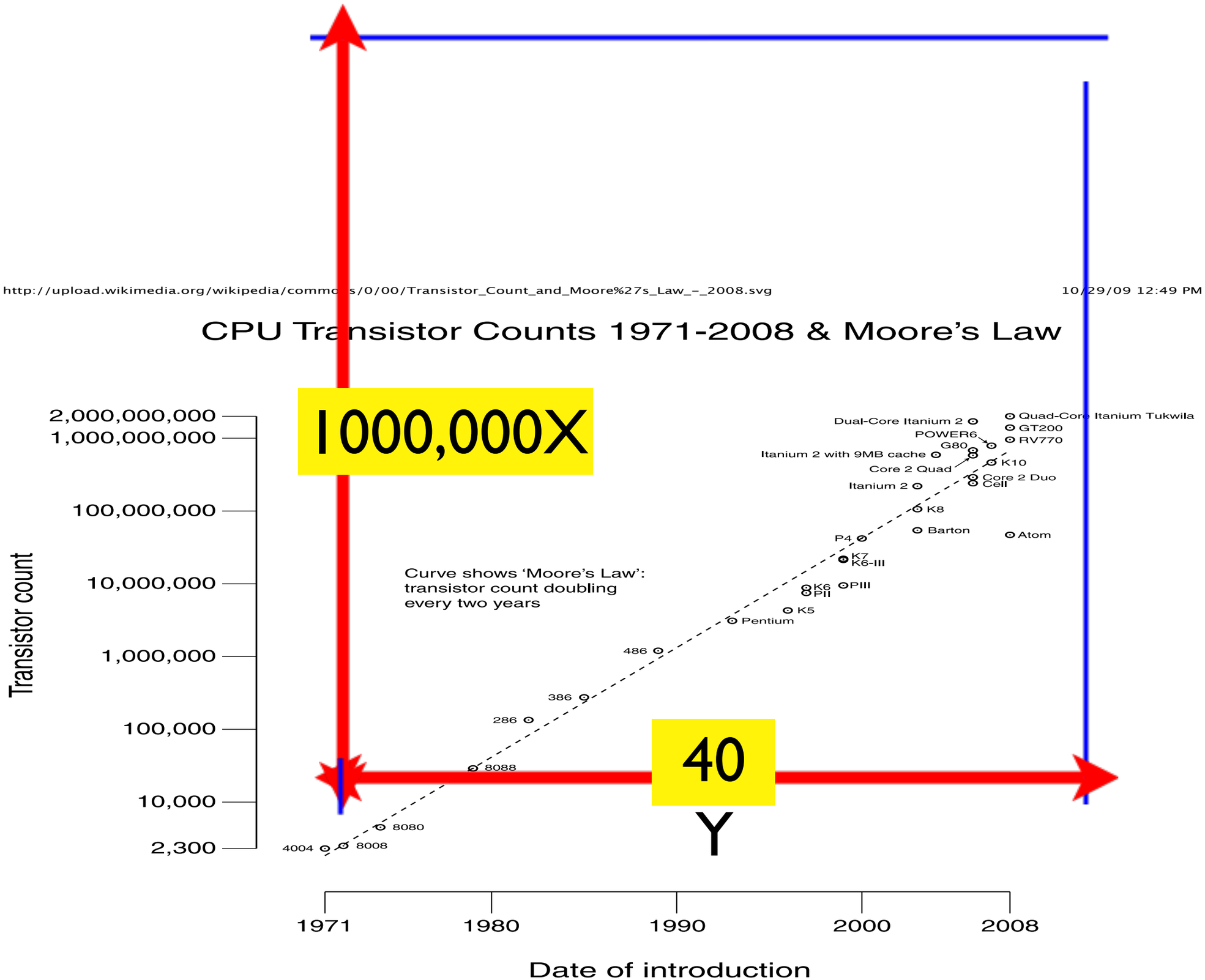
What happened???

Moore's Law 1971-2011

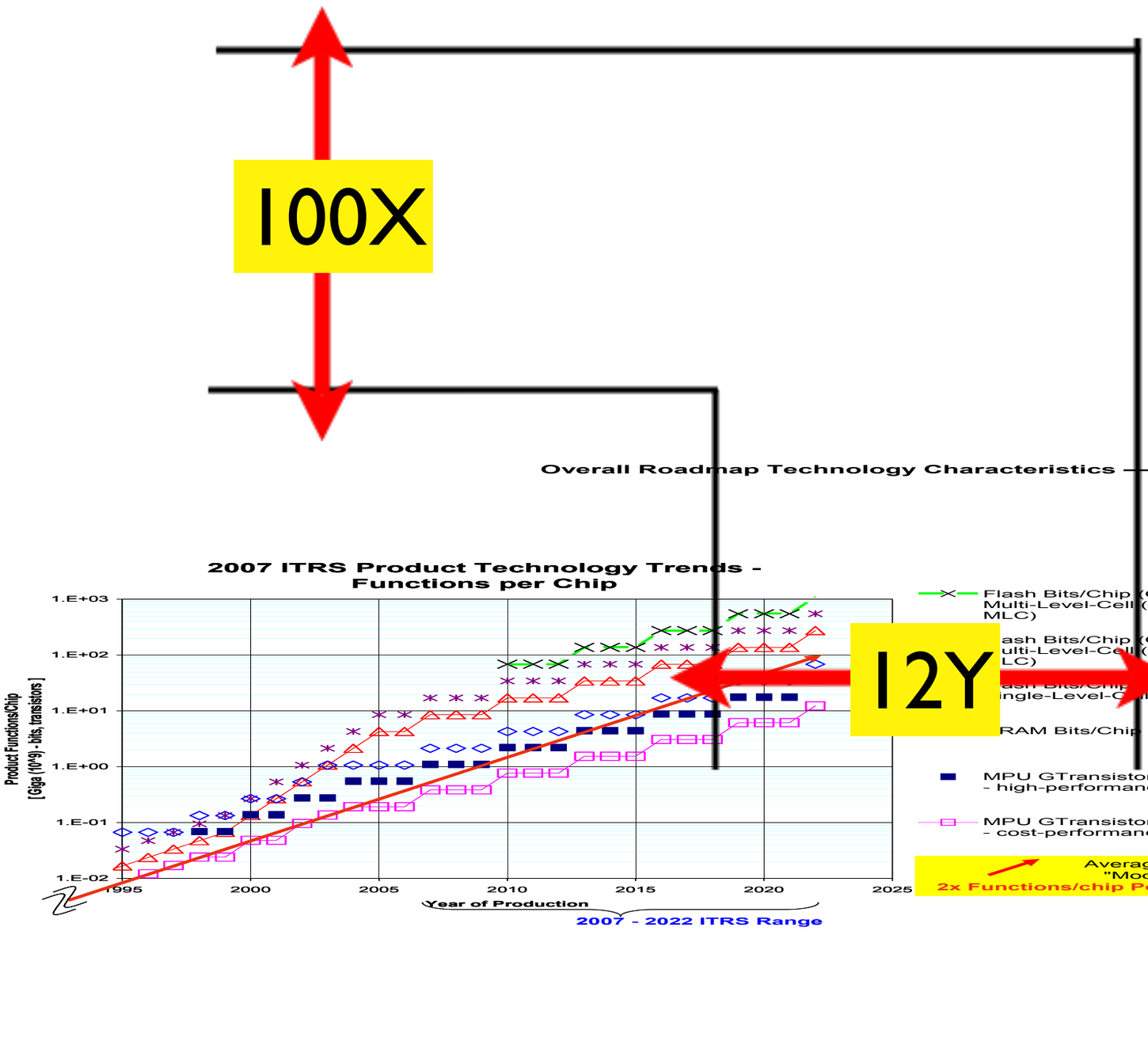
http://upload.wikimedia.org/wikipedia/commons/0/00/Transistor_Count_and_Moore%27s_Law_-_2008.svg 10/29/09 12:49 PM



Moore's Law 1971-2011



Semiconductor Technology Roadmap



Snapshot: Logic Density

1000 B

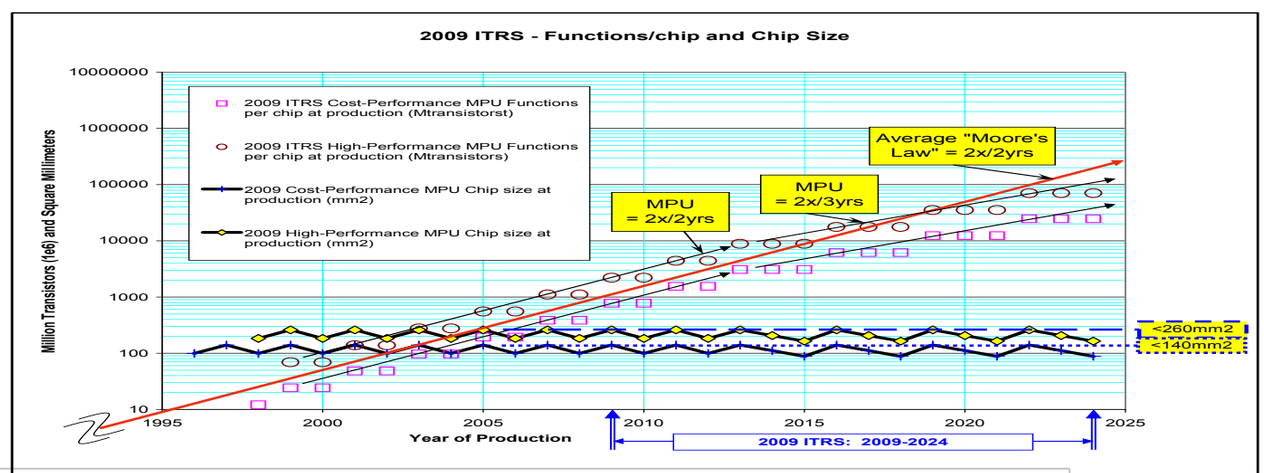
100 B

10 B

1 B

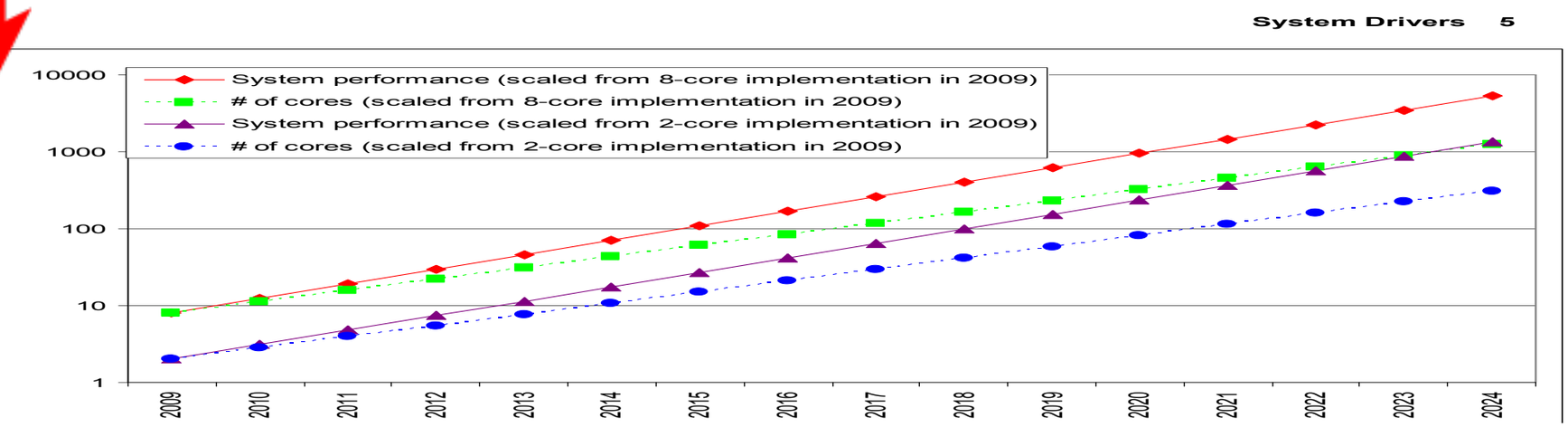
0.1 B

Roadmap Technology Characteristics —2010 Update

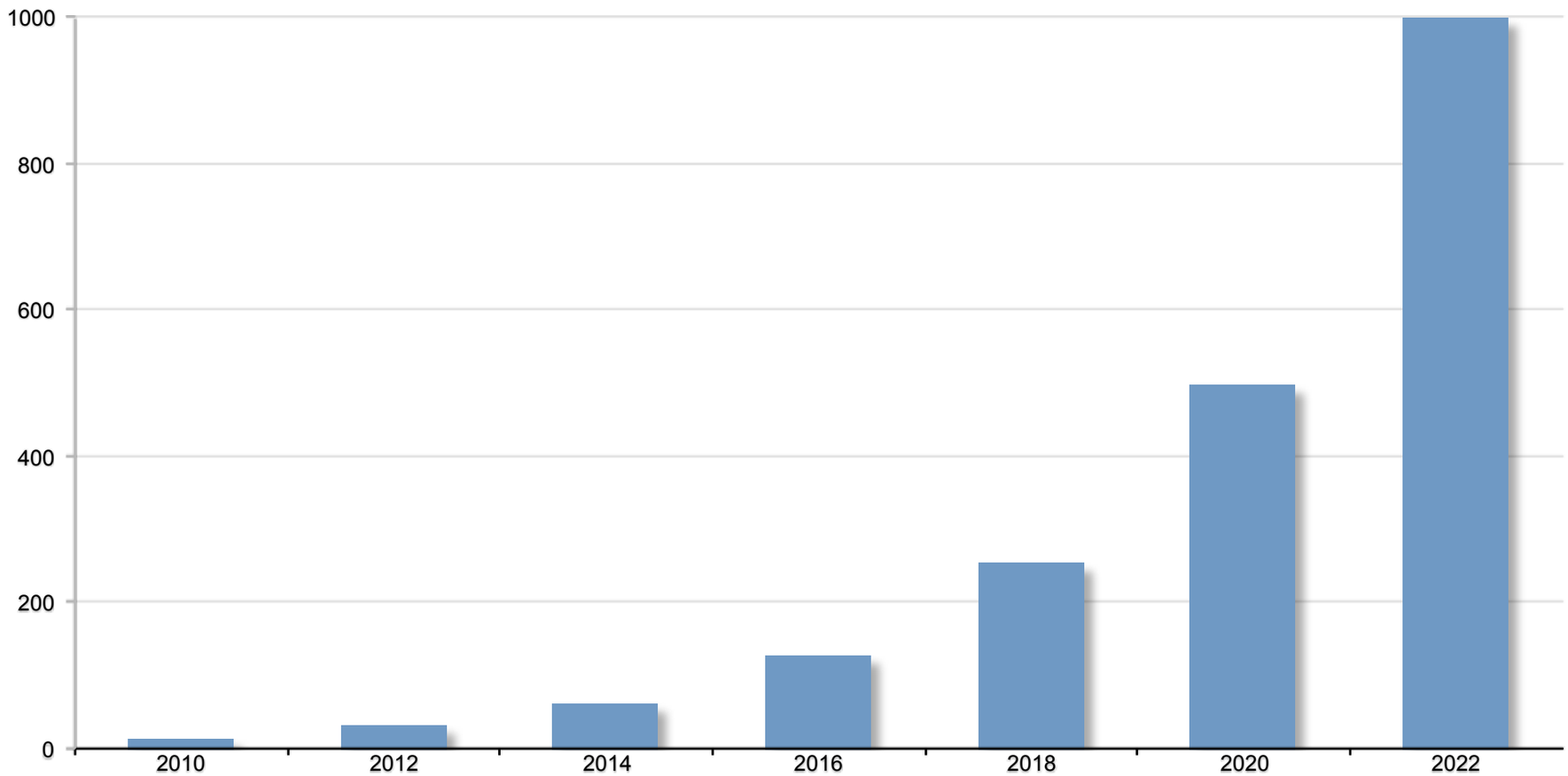


System Roadmap Projection

100X



64-bit CPU Cores over Time



100X Performance by 2022

Memory Hierarchy is Not Changing

NEED FOR NEW COMPUTING MODEL

Technological Transition

Computer architecture has changed. Today's multicore, multi-CPU server provides fast communication between processor cores via main memory or shared cache. Main memory is no longer a limited resource. In 2012 servers with more than 2 terabytes of RAM are available.

Modern computer architectures create new possibilities but also new challenges. With all relevant data in memory, disk access is no longer a limiting factor for performance. In 2012 server processors have up to 80 cores, and 128 cores will come in the near future. With the increasing number of cores, CPUs will be able to process more and more data per time interval. That means the performance bottleneck is now between the CPU caches and main memory (see Figure 2). An optimized database technology should focus on optimizing memory access by the processing cores. Simple disk access optimization by caching data in memory may not yield breakthrough performances.

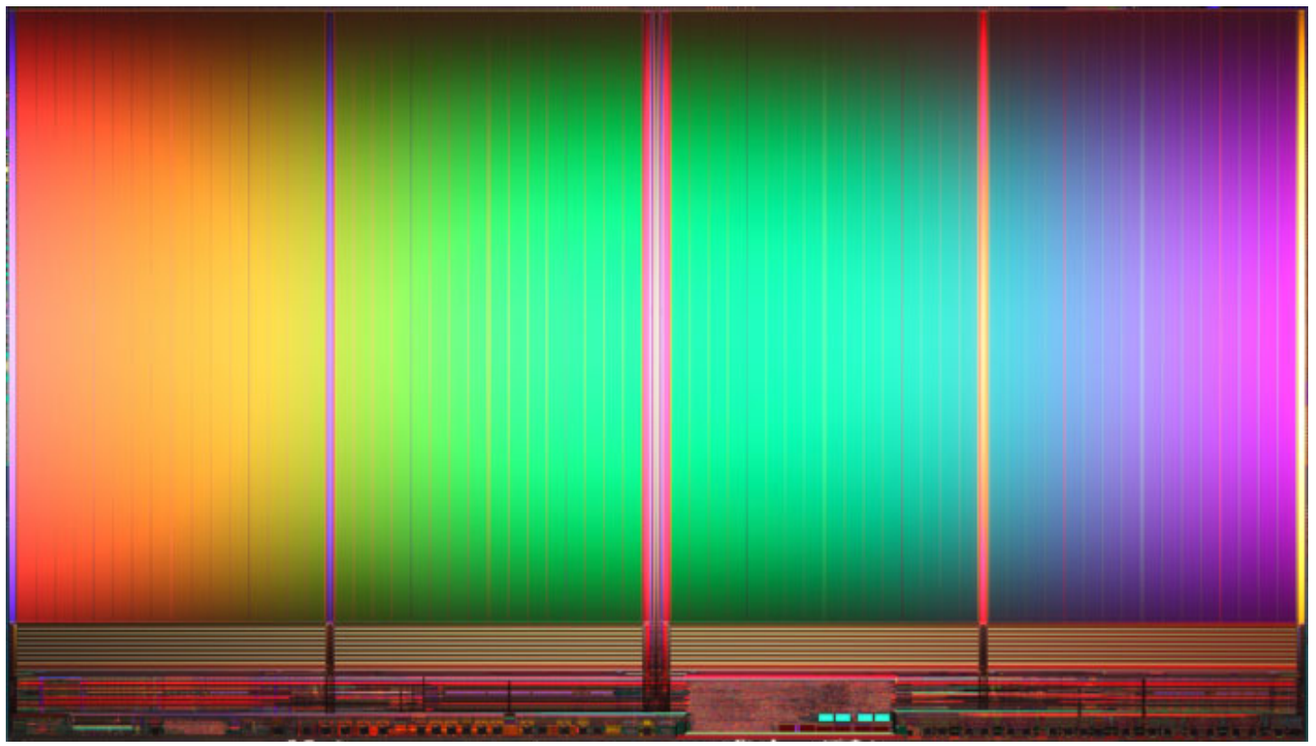
To provide an idea about sizes and access speeds of a current memory hierarchy, the table below compares the different layers in this memory hierarchy (CPU characteristics for Intel's Nehalem architecture).

Type of Memory	Size	Latency
L1 cache	64 KB	~4 cycles [2 ns]
L2 cache	256 KB	~10 cycles [5 ns]
L3 cache (shared)	8 MB	35–40+ cycles [20 ns]
Main memory	GBs up to terabytes	100–400 cycles
Solid state memory	GBs up to terabytes	5,000 cycles
Disk	Up to petabytes	1,000,000 cycles

Figure 2: Hardware Architecture: Current and Past Performance Bottlenecks

Hard Disk drives are not keeping up
Flash solving this problem just in time

Flash Today: 8 GB per Die, 64 GB per Package



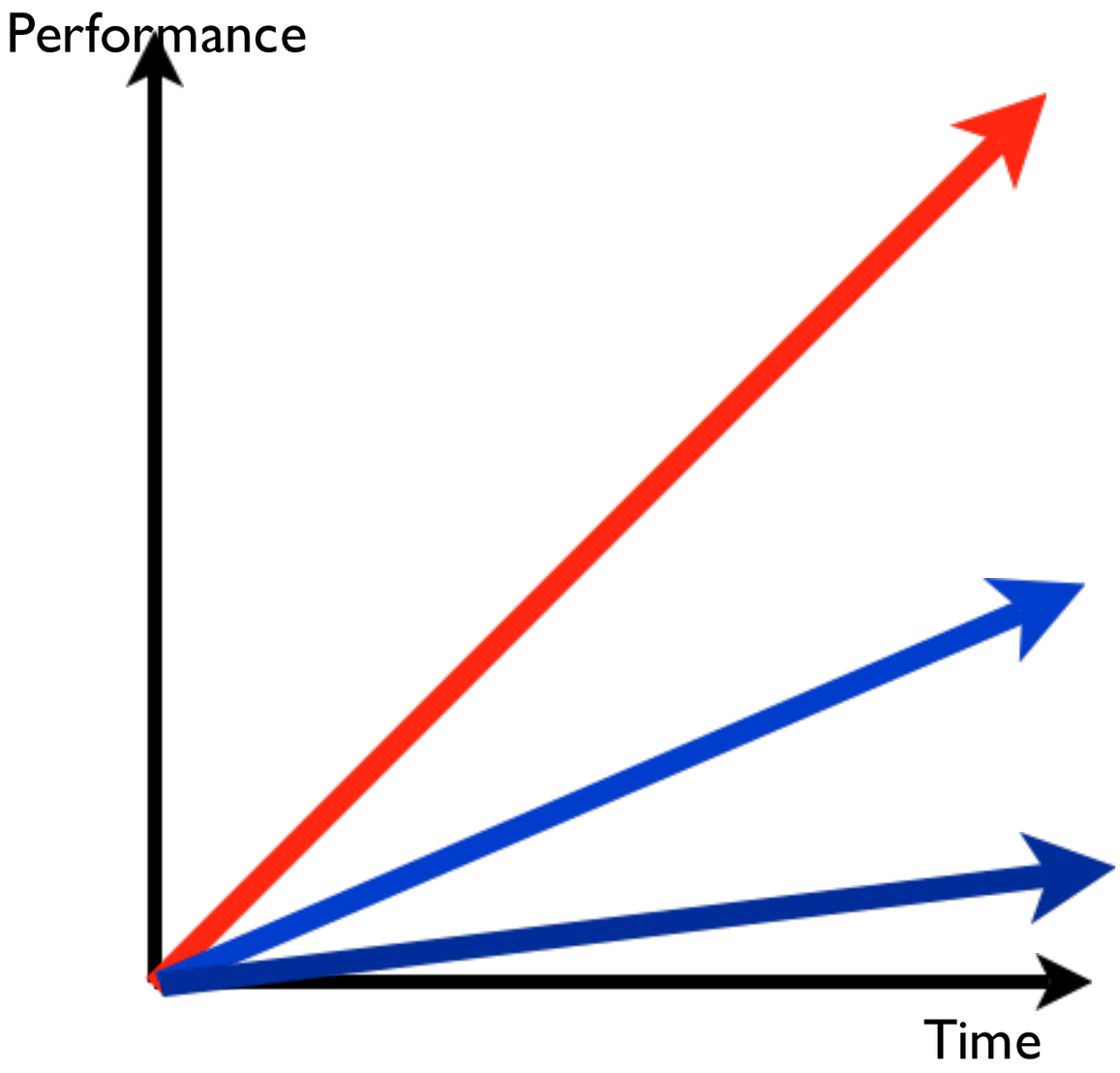
Expect to see 256 GB per package in 2013 and 1 TByte Flash per package in 2015

Moore's Law Summary

- **Moore's Law is alive and well**
  - 2X Density every 2 Years
- **Million-fold advance from 1971-2011**
  - Another factor of 100X next 12 years
- **Billion-fold advance expected 1971-2031**
  - Beyond that, it gets hard to forecast

There has been nothing like this in the history of mankind

Moore's Law and Networking



CPU: 2X/2Y = 64X/12Y

1GigE-10GigE: 10X/12Y

WAN Routers: 4X/12Y

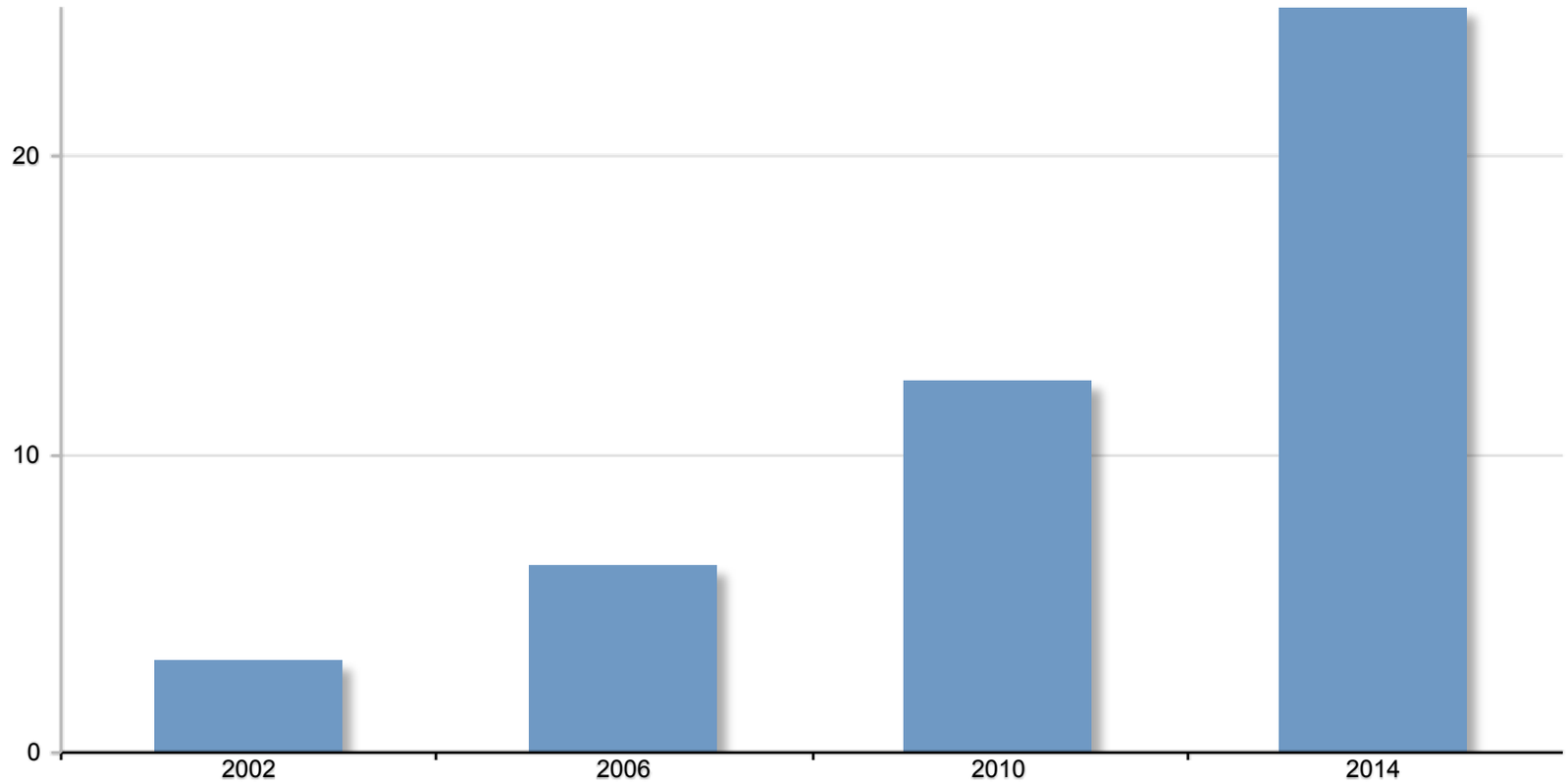
Why did Networking not Keep up with Moore's Law?

Three Major Problems

- **Moore's Law applies to Transistors, not Speed**
  - Transistor count is doubling every 2 years
  - Transistor speed is only increasing slowly
- **Number of IO pins per package basically fixed**
  - Limited by die area and package technology
  - Only improvement is increased I/O speed
- **Bandwidth ultimately limited by I/O Capability**
  - Throughput per chip = # IO Pins * Speed/IO
  - No matter how many transistors are on-chip

SERDES Speed (high-density CMOS)

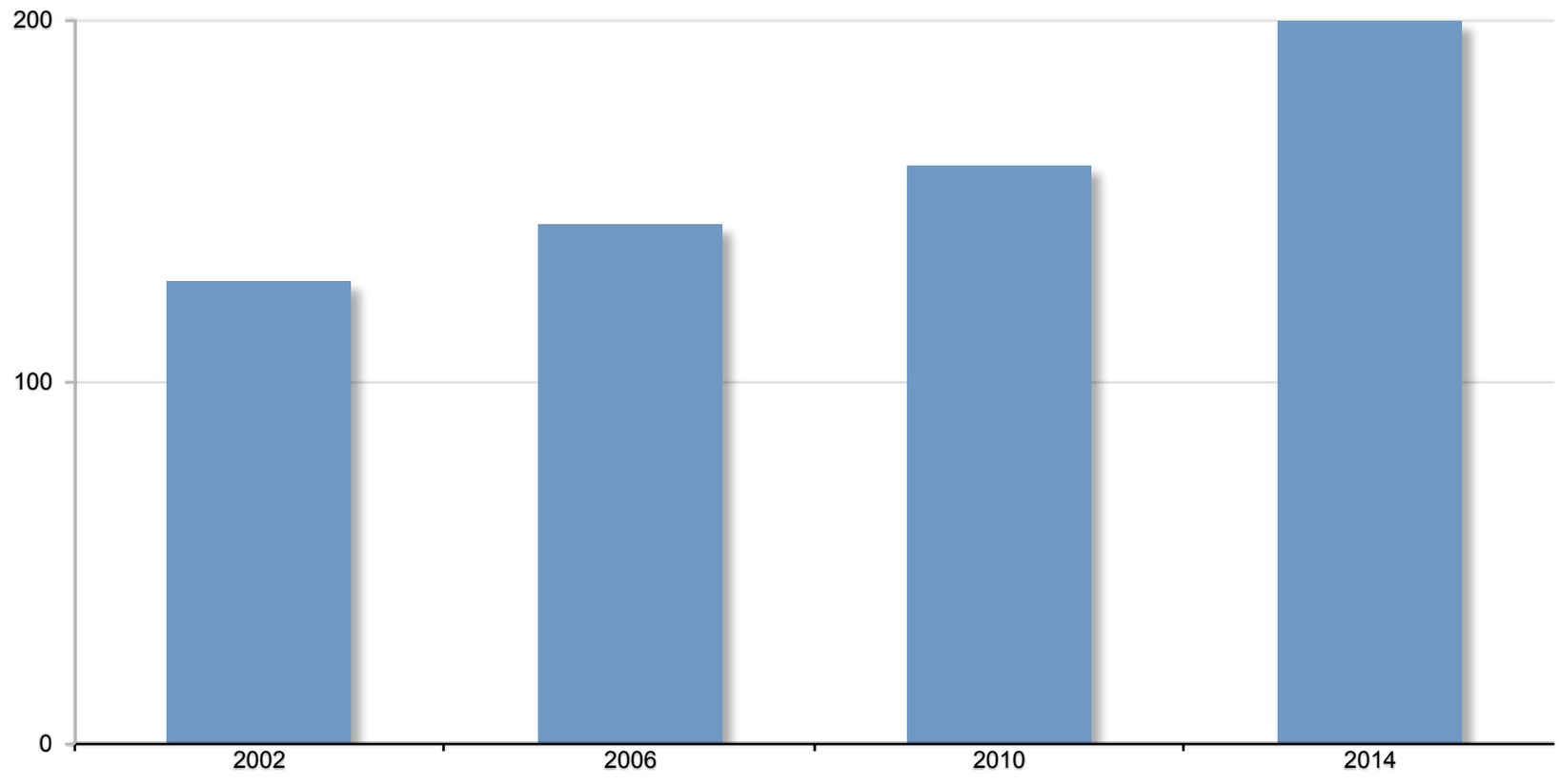
Gbps



8X in 12 Years = 2X every 4 Years

Number of SERDES per Package

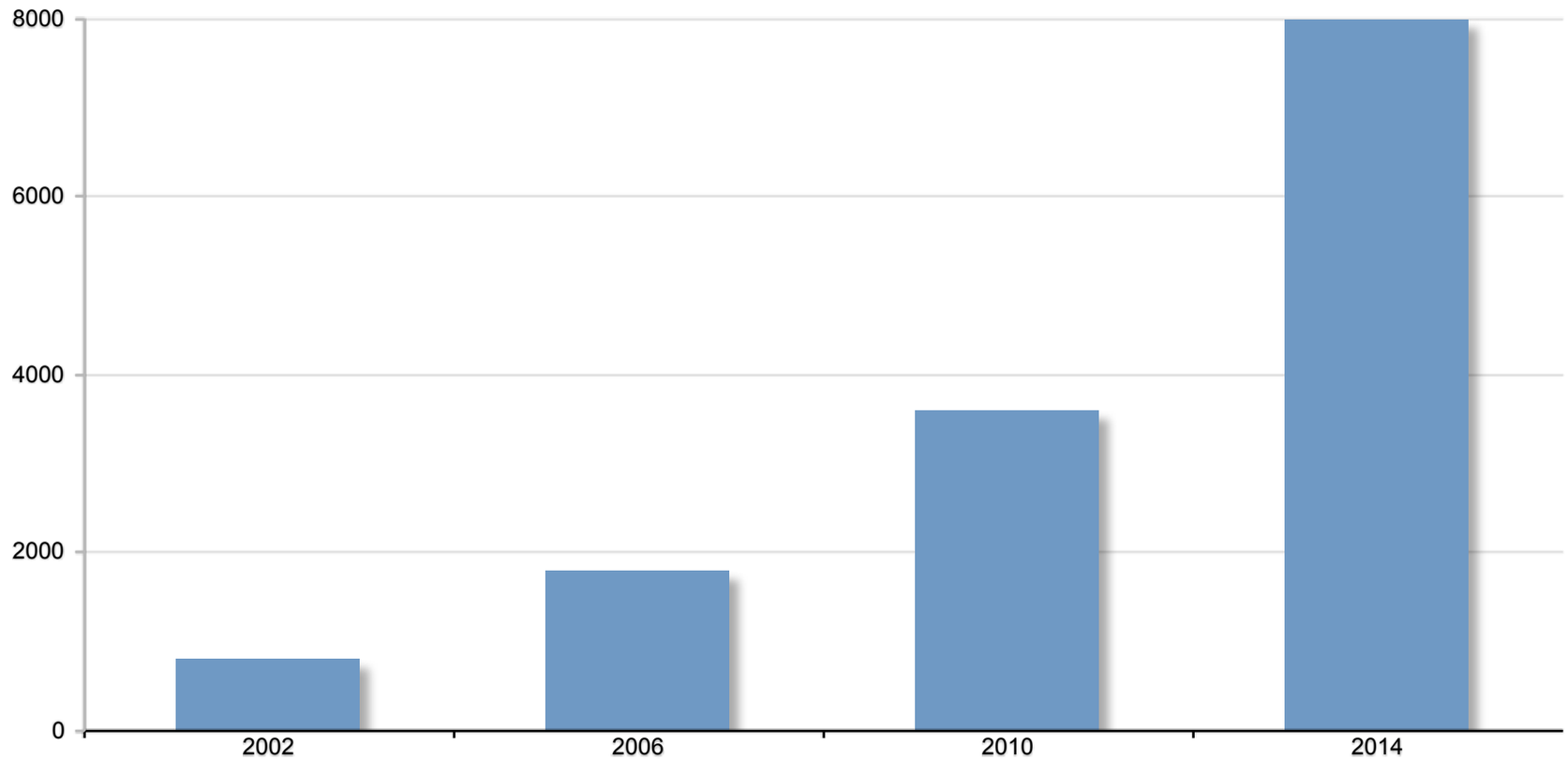
SERDES



Modest Increase in 12 Years

Maximum Throughput per Chip

Tbps



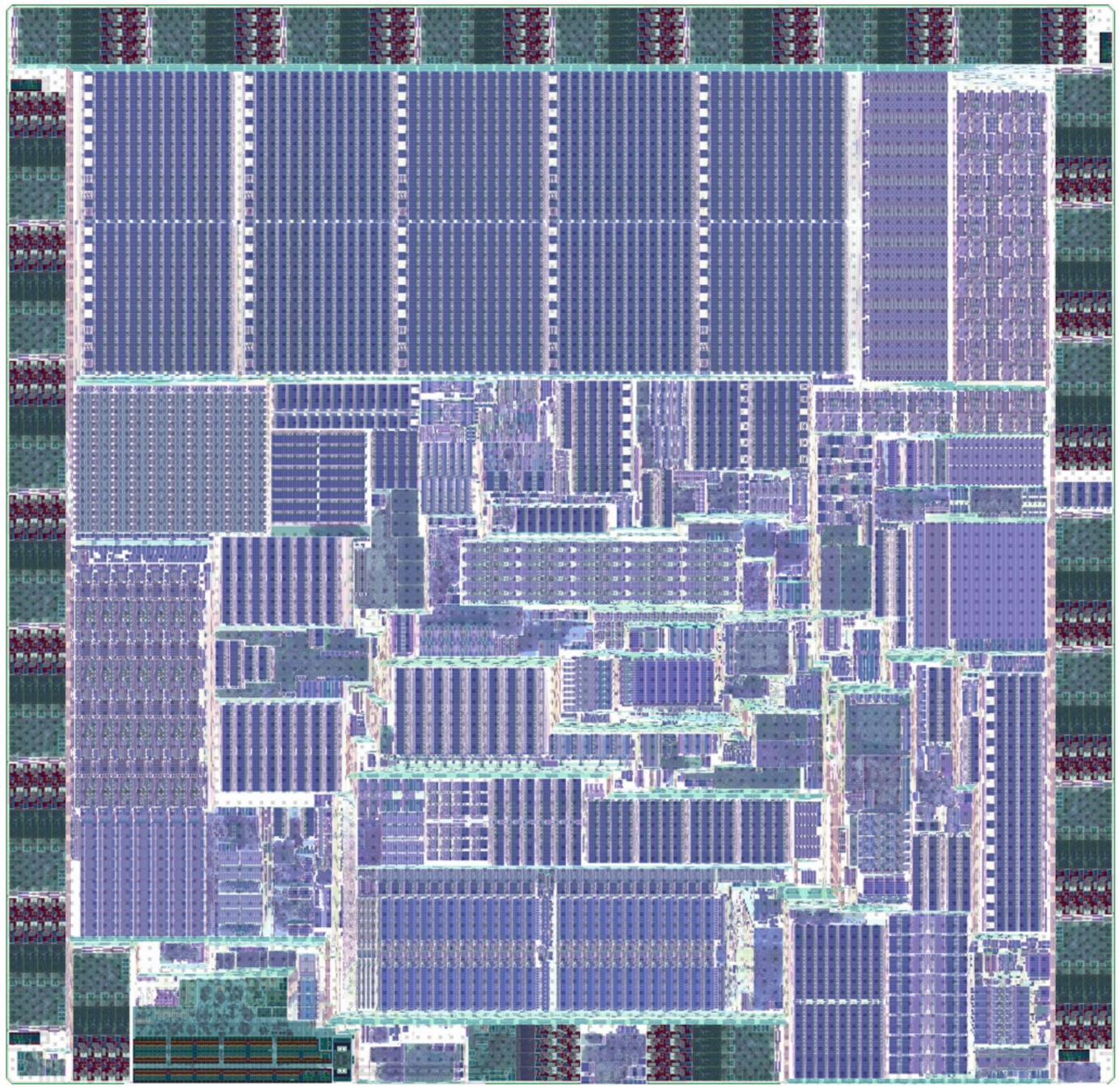
10X in 12 Years

ASIC vs Full Custom Chip Design

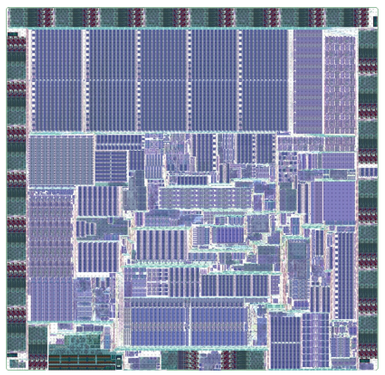
- **ASIC = Application Specific Integrated Circuit**
  - "Top-down" design, independent of layout
  - ASIC vendor does physical implementation
  - Difficult to achieve high clock rates this way
- **Full Custom Flow**
  - Chip design starts with clock rate
  - Data Paths designed to achieve clock rate
  - Only way to get to high clock rates

Typical Result: 8X Higher Density in Full Custom

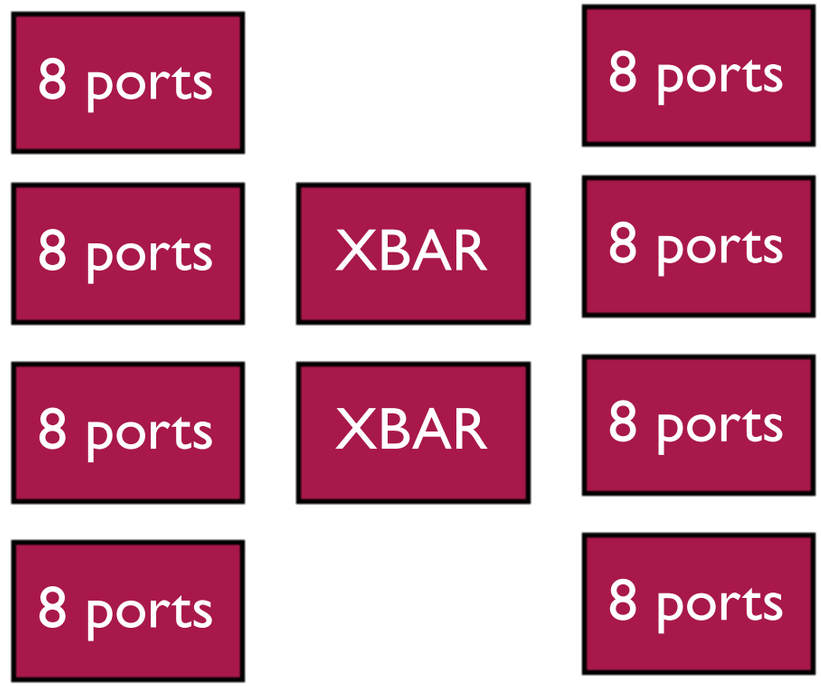
Full Custom 64 port 10G Switch Chip



64 port 10G Switch: Custom vs ASIC



Custom Design: 1 Chip



ASIC Design: 10 Chips

Advantages of Full Custom Chips

Full Custom Chips are Denser (more ports per chip), have much lower latency (due to fewer chip crossings), resulting in system designs that consume less power and are much more reliable than multi-chip designs

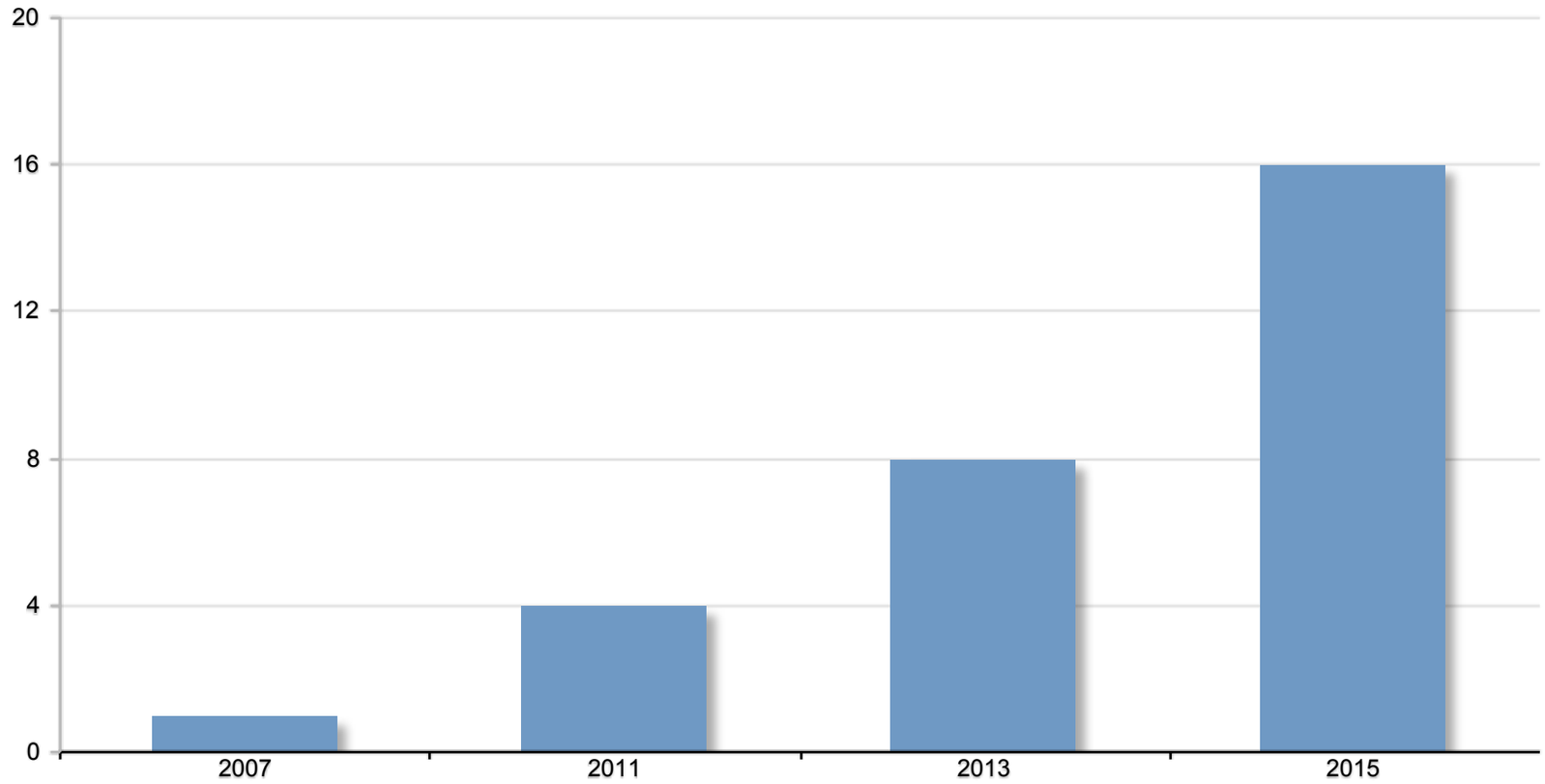
ASIC designs are not on Moore's law

Evolution of Custom Switch Silicon

Technology	130 nm	65nm	40 nm	28 nm
10G ports	24	64	128	256
Throughput	360MPPS	960MPPS	2 BPPS	4 BPPS
Buffer Size	2 MB	8 MB	16 MB	32 MB
Table Size	16K	64K	128K	256K
Port Speeds	10G	10/40	10/40/100	10/40/100
Availability	2007	2011	2013	2015
Improvement	N/A	3X/4Y	2X/2Y	2X/2X

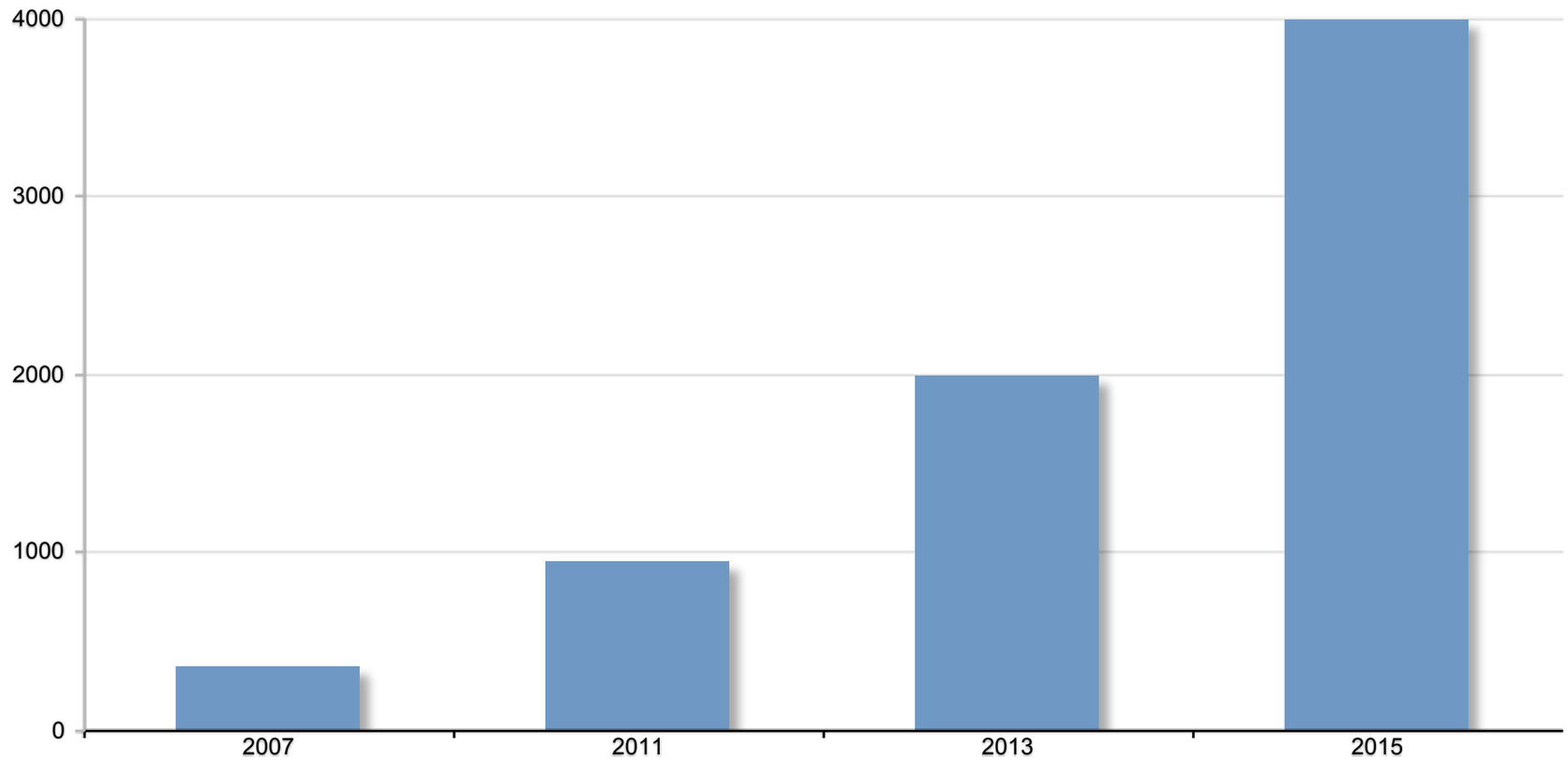
Next generation custom switch silicon is on Moore's Law!

Relative Device Densities



130 nm 65 nm 40 nm 28 nm

Single Chip Throughput (MPPS)

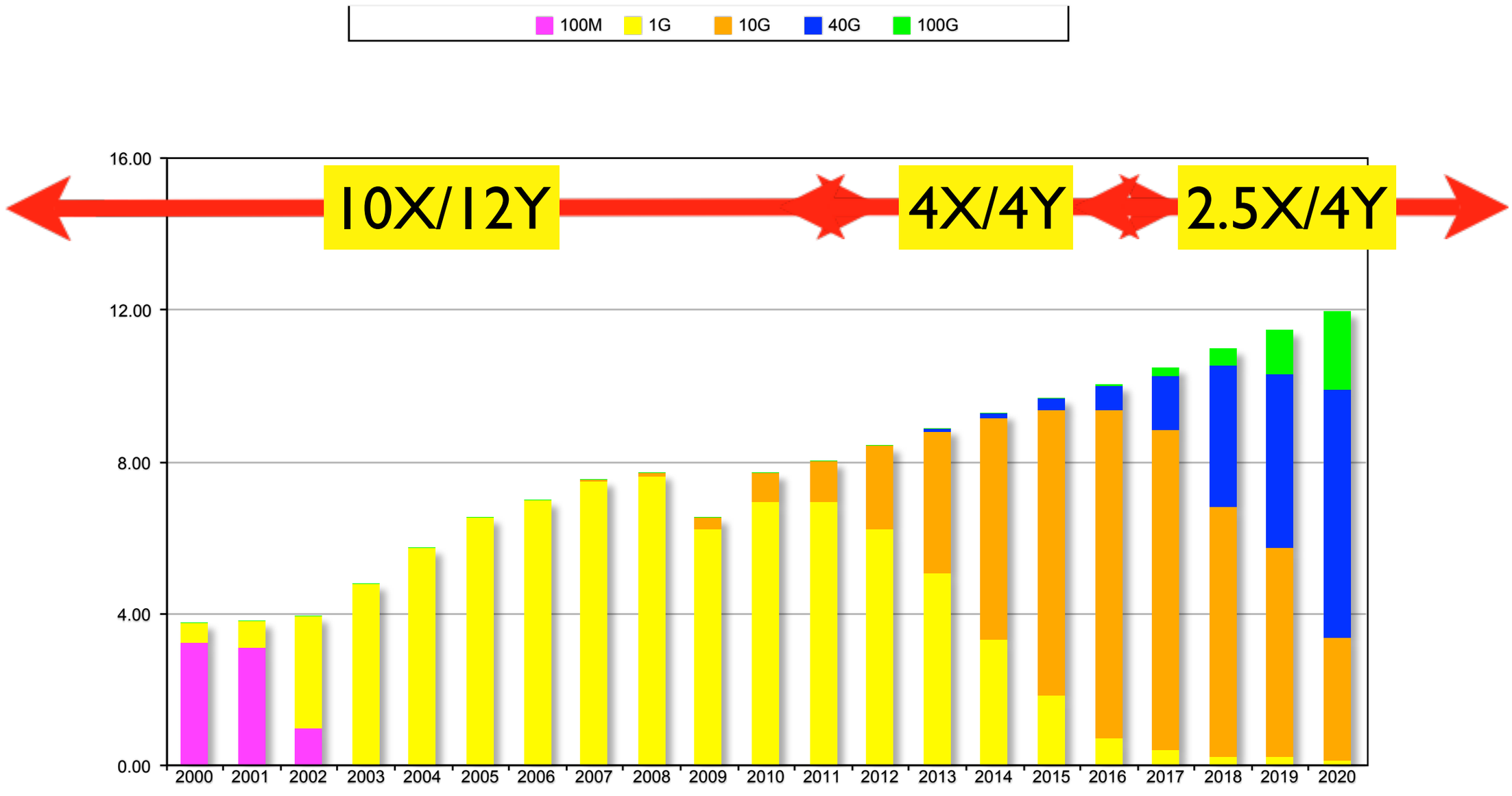


130 nm 65 nm 40 nm 28 nm

Moore's Law and Networking

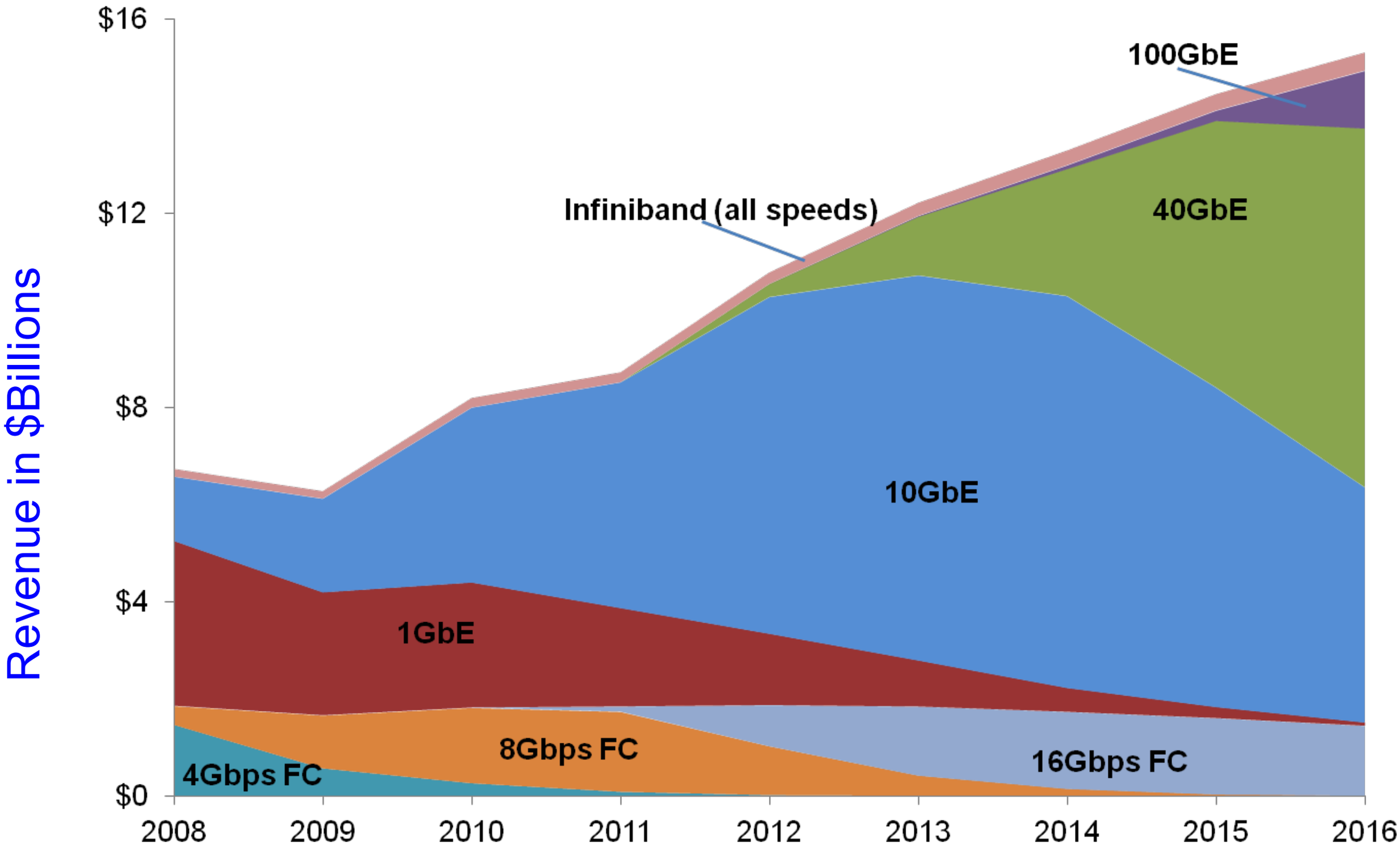
- **Next Generations scale with Moore's Law**
  - Table sizes double every process generation
  - Industry catching up on process roadmap
- **I/O Speed scales less than Moore**
  - Larger package sizes offset constraint
  - Next step is 25 Gbps SERDES in 2014
- **Full-Custom Design Flow Required**
  - ASIC design flow wastes silicon potential

Server 10/40/100G Adoption Cycle



Source: Intel LAN Group

Total Datacenter Switch Revenue by Protocol & Speed

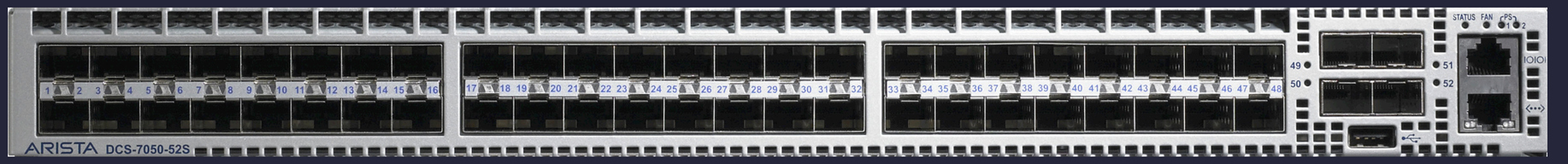


CPUs Driving Network Upgrade

- **Faster CPUs need Faster Networks**
- Sandybridge driving 10 GigE Adoption
- 50% attach rate in 2013, 80% by 2015
- **10/40/100G Market will grow quickly**
- From $4B in 2010 to $16B in 2016
- From 5M ports in 2010 to 67M ports in 2016
- **Faster End nodes need faster Backbones**
- Most Traffic going East-West, not North South
- Cluster sizes getting larger and larger

Scaling the Cloud Network

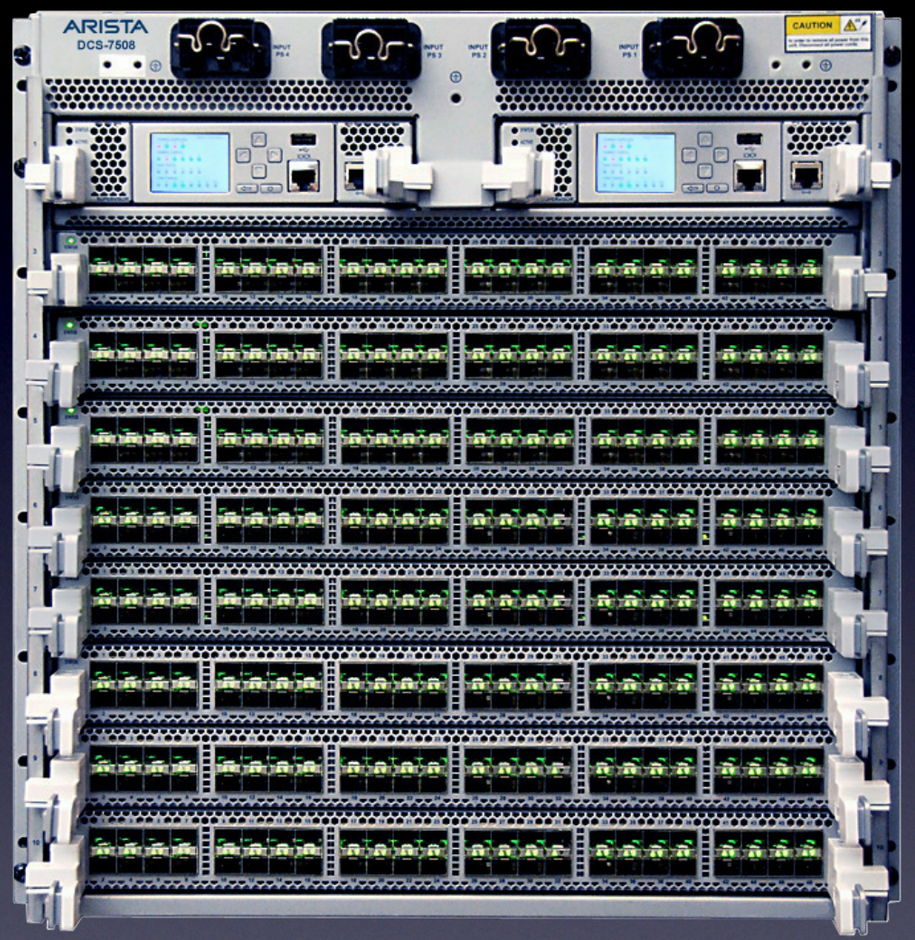
Arista 7050 Switch



64-ports 10G, 960 BPPS, 1.28 Tbps

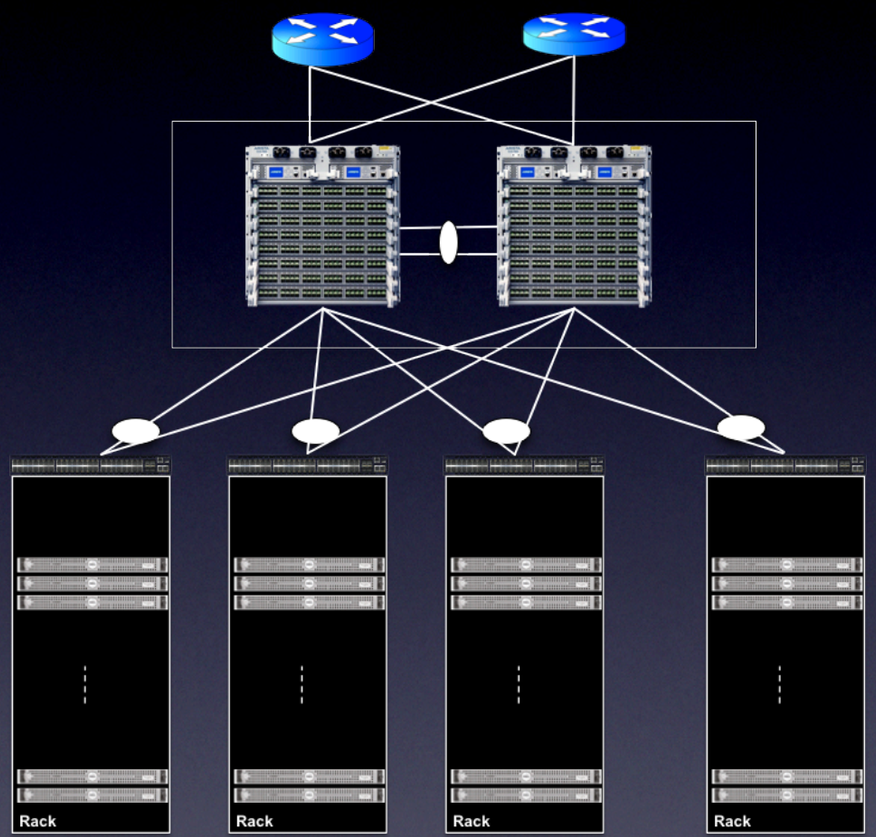
Typical Power 2 Watt/Port

Arista 7500 Switch

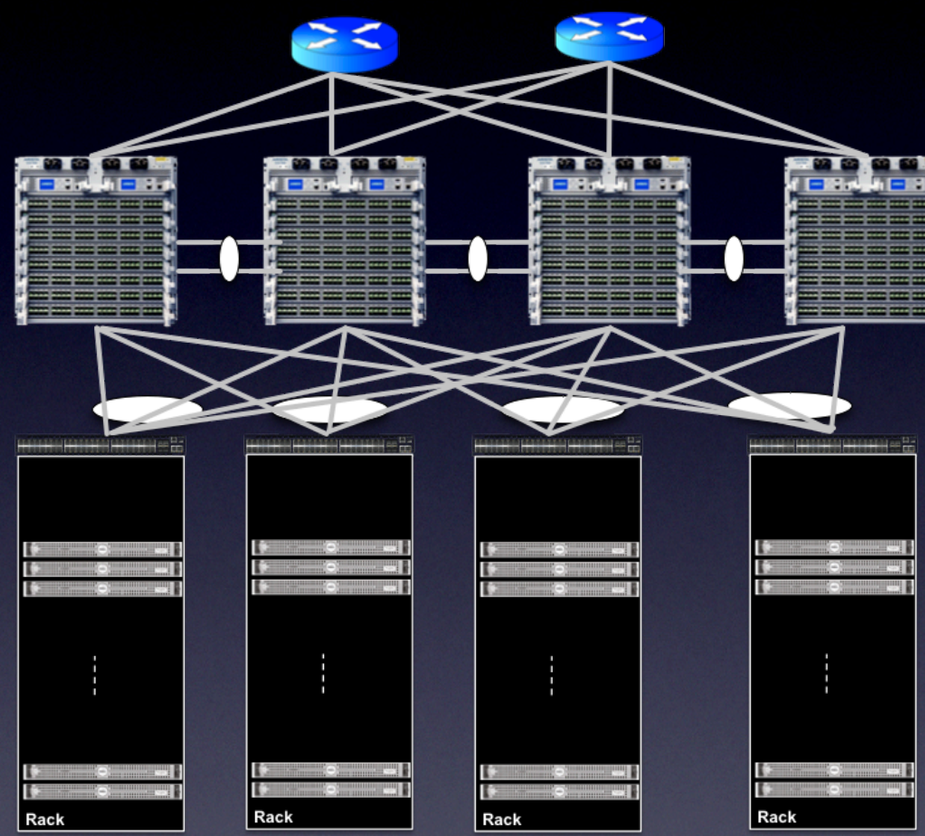


384-ports 10G, 5760 BPPS, 10 Tbps Fabric

Two ways to Scale: L2 or L3

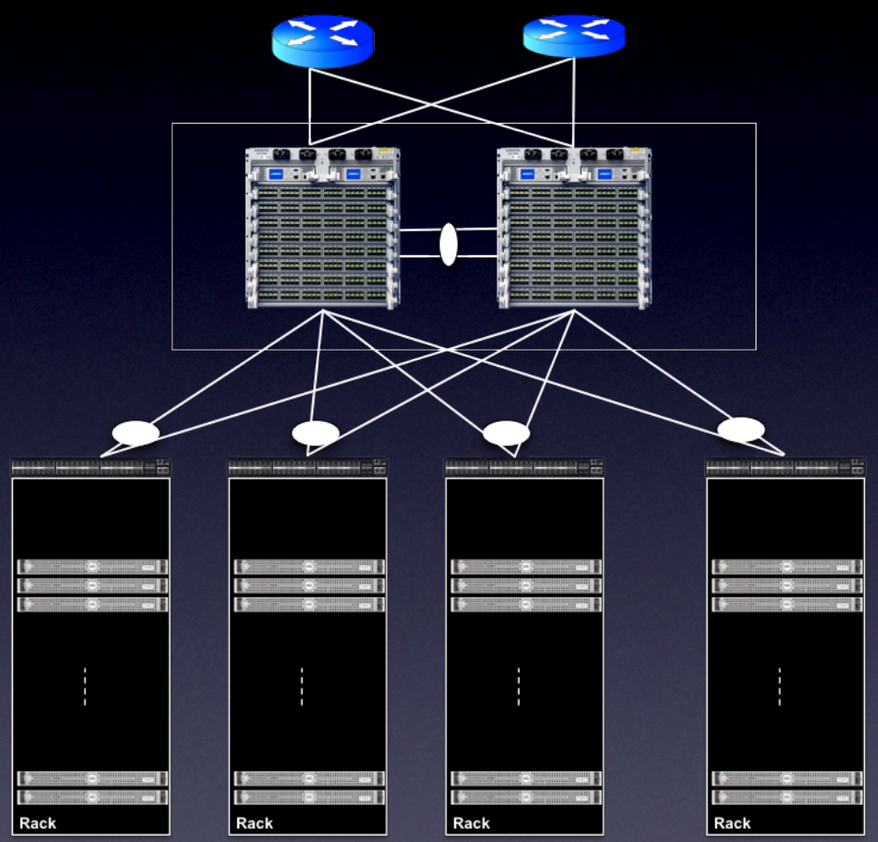


MLAG Spine (L2)



ECMP Spine (BGP)

Scaling with MLAG (L2)



MLAG Spine (L2)

MLAG provides active-active load-sharing redundancy

Max Throughput: 20 Tbps with current Arista 7500

Maximum Scale: 360 Racks with current Arista 7500

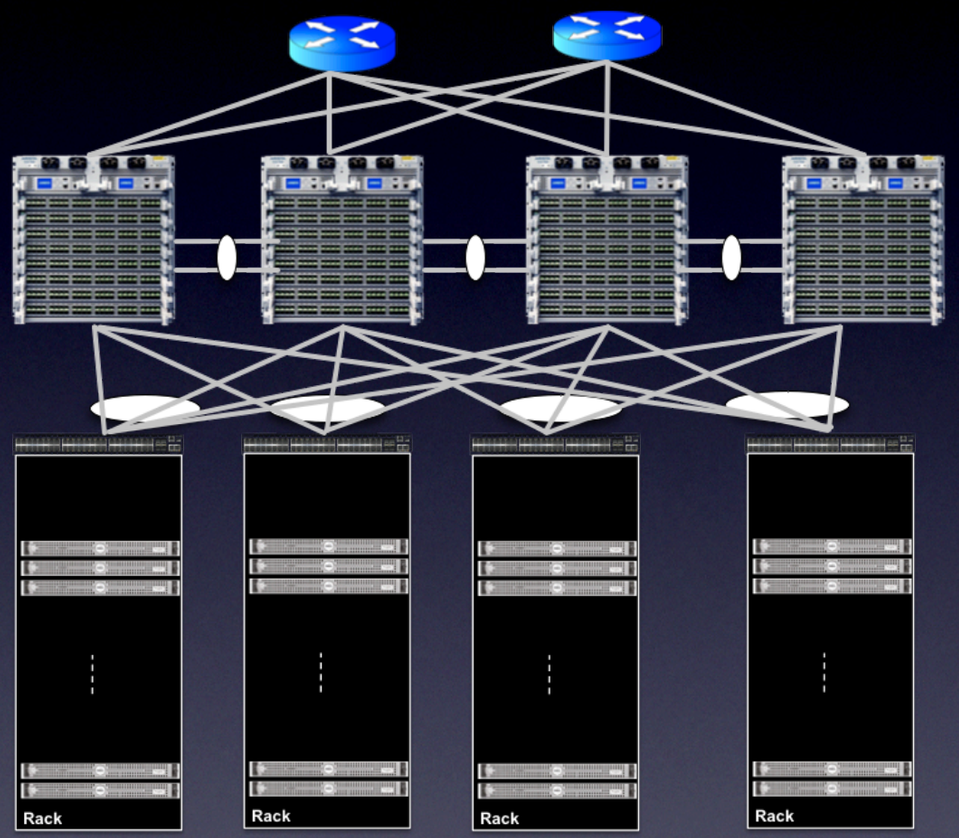
No proprietary Fabric Required

Scaling with ECMP (L3)

ECMP provides scalable active-active load-sharing

Max Throughput: 320 Tbps with current Arista 7500

Maximum Scale: 360 Racks using current Arista 7500



ECMP Spine (BGP)

No proprietary Fabric Required

ECMP Scale

ECMP Spine (OSPF/BGP)

ECMP	Spine Capacity	Cluster Size	Oversubscription
4-way	40Tb	23000	10:1
8-way	80Tb	21000	5:1
12-way	120Tb	19000	3:1
16-way	160Tb	18000	2.5:1
32-way	320Tb	36000	1.25:1

Planning Guide

1. Decide pod size and bandwidth per server
   => determines total cluster bandwidth

2. Select ECMP Redundancy level (4-32 way)
   => determines bandwidth per spine switch

3. Size Spine switch to match servers / rack
   and ECMP Fanout Factor

Optimize cost of bandwidth per server

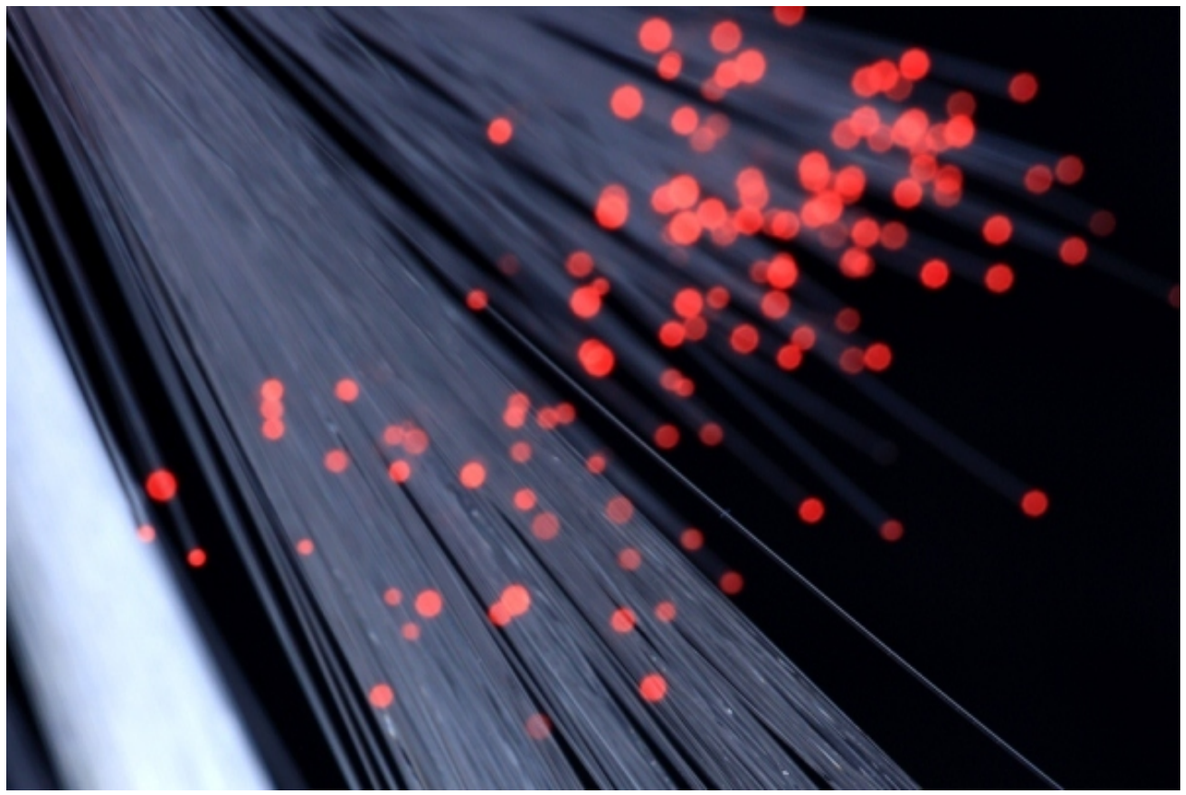
Network Utility Function

The value of a network is not the cost per port, but the cost per bandwidth delivered to servers, including the cost of leaf switches, spine switches, cost of optics, fiber cabling and power over time.

Higher interface speeds only improve utility if they improve $/Gbps cost-performance, i.e. one 100G port costs < 10*10G ports

Status of 40 GigE and 100 GigE

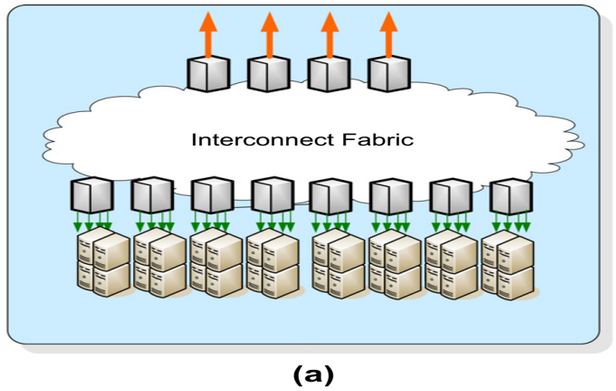
- **IEEE Standards completed years ago**
  - 40G and 100G products shipping
- **Issue is cost-performance utility**
  - 40 GigE > 4X Cost of 10 GigE
  - 100 GigE >>> 10X Cost of 10 GigE
- **Biggest problem is optics cost**
  - 100 GigE optics are extremely expensive
  - Even 40G optics are > 4X 10G Optics
- **Volume Adoption requires Cheaper Optics**



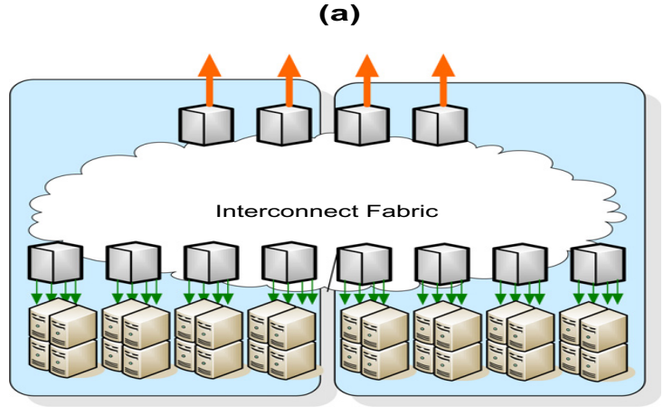
10/40/100G Physical Layers for large-scale Datacenters

Leaf-Spine Cluster Configuration

Fiber Technology 17 (2011) 363–367

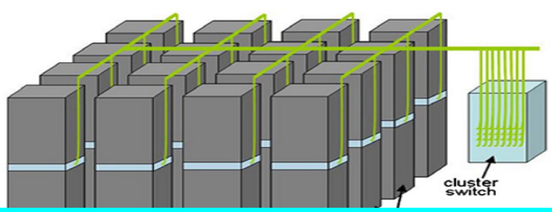


(a)



(b)

Fig. 2. Hierarchies of intra-datacenter cluster-switching interconnect fabrics (a) within a single building (b) across multiple buildings.

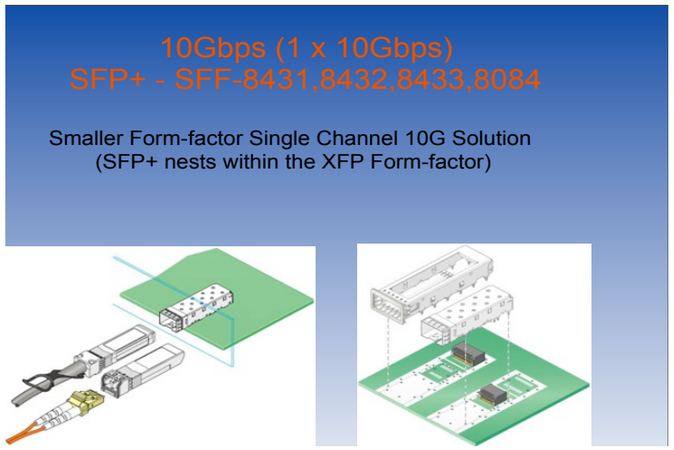


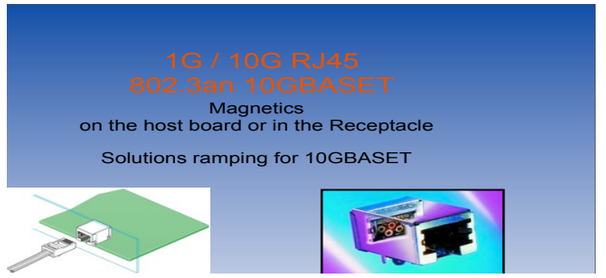
Reach from leaf-switch to spine switch: 100-300m

Cloud Optics Requirements

- **100-300m Reach, in some cases up to 1km**
- Rack-top to spine switch to core router
- Support of 40G and 100Gbps Ethernet
- Ideally over the same fiber infrastructure
- **Minimize total solution cost**
- Switch Port + Laser + Fiber + Power

10G Today: 10G-SFP+ and 10GBASE-T
48 Ports per 1U Front Panel





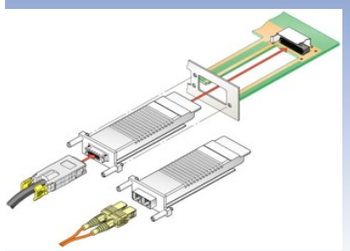
SFP+ supports laser and twin-ax copper cables

RJ45 supports 10GBASE-T + 1000BASE-T interoperable

10 Year Struggle for 10G to get here: XENPAK, XPAK, X2, XFP, SFP+

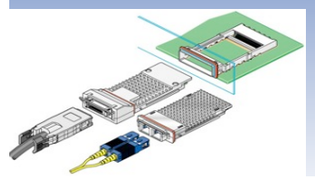
10Gbps (4 x 2.5Gbps)
10GBASECX4/LX4 – Xenpak

eneration, Four Channel, Pluggable Form-factor
e" Form-factor compared to GBIC or SFP
ired slots in host board
al & Copper Applications



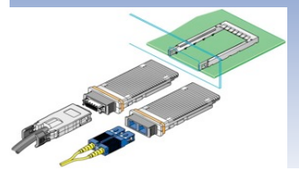
10Gbps (4 x 2.5Gbps)
0GBASECX4/LX4 – Xpak

r Second generation Xenpak
sized – less real estate, less beachfront
for slots in host board eliminated
cations for Servers/PCI Adapter Cards
shipping for legacy products



10Gbps (4 x 2.5Gbps)
0GBASECX4/LX4 – X2

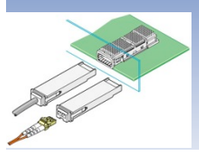
d generation Xenpak
nsized – less real estate, less beachfront
d for slots in host board eliminated
ch Applications
shipping for legacy products



x 10Gbps)
MSA

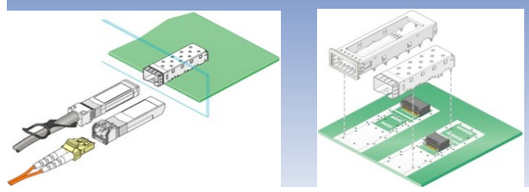
el10G Pluggable Solution

narily due to cost
more volumes



10Gbps (1 x 10Gbps)
SFP+ - SFF-8431,8432,8433,8084

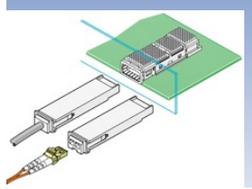
Smaller Form-factor Single Channel 10G Solution
(SFP+ nests within the XFP Form-factor)

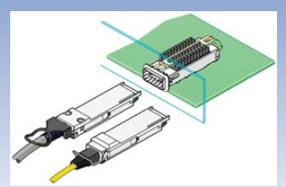


40G Today: QSFP
32-36 Ports per 1U Front Panel

10Gbps (4 x 2.5 & 5Gbps)
QSFP MSA
SFF-8436, IBTA Spec

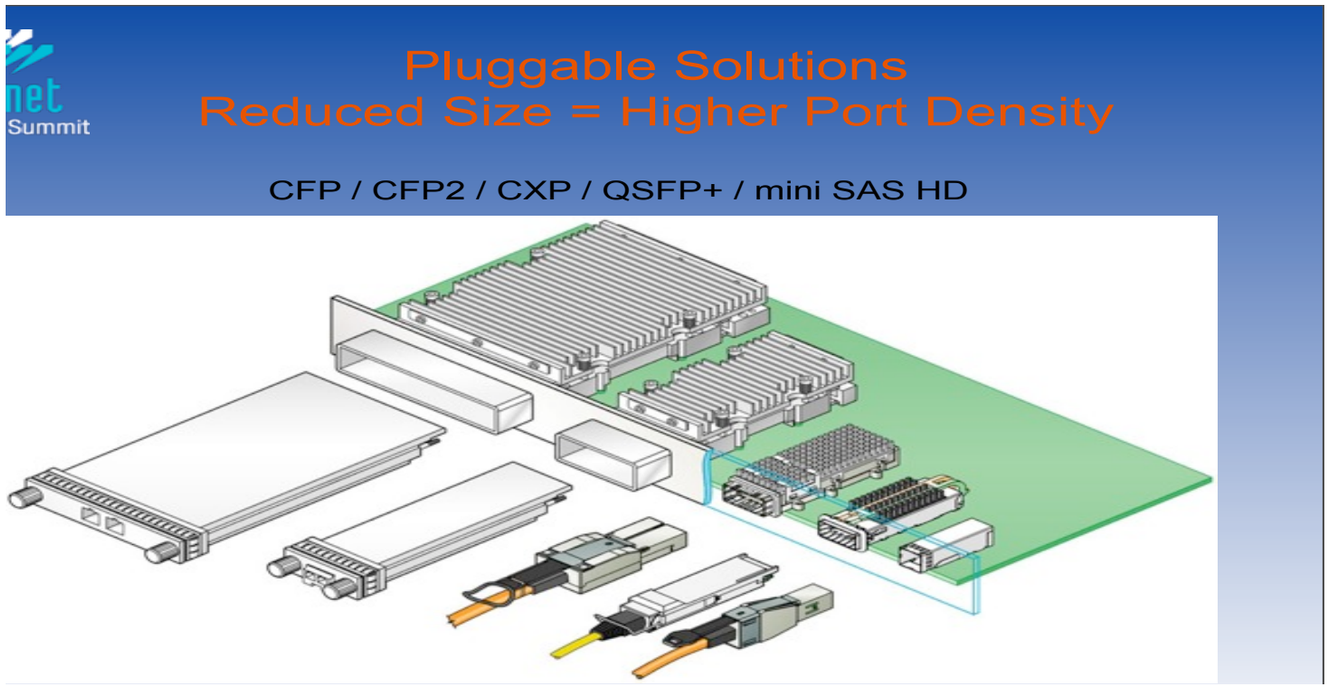
Started out as just a shorter XFP Form-factor that was 4 channels
Moved from 30ckt to 38ckt host board connector
- SFP style connectors are high speed upper row only
- Single to quad channel required more contacts





40G-QSFP supports 40G-LR4, 40G-SR4, twin-ax copper and active optical cables

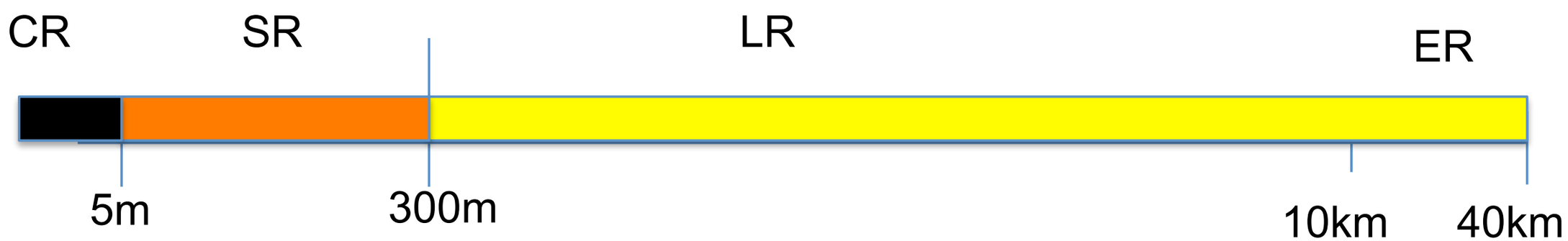
100 GigE PHY MSA Confusion: CFP, CFP2, CFP4, CXP, QSFP+



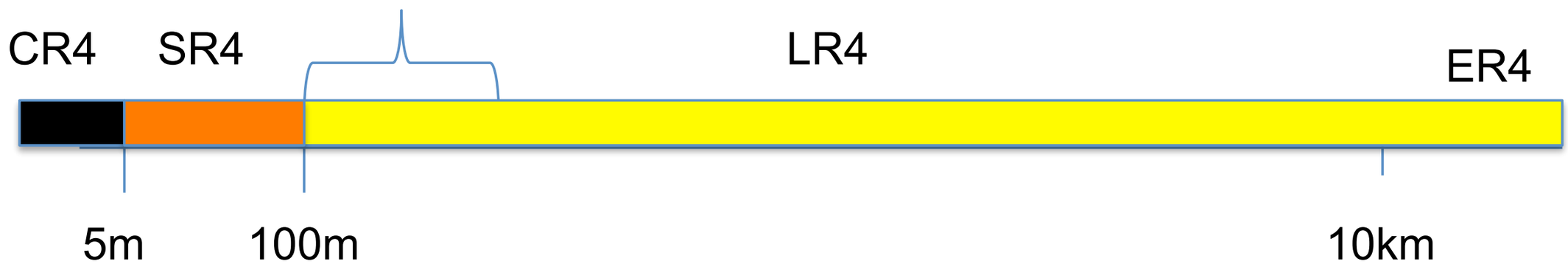
More choices than original 10G Ethernet

The 10G to 100G MMF Reach GAP

10G-SR 300m meets most customer requirements



Cost optimized 100-500m
solution is critical to success of 100G



100-SR4 Reach is limited to 100m maximum

Current State of 100G PHYs

- **Highest Demand is for Leaf-Spine Links**
- Distances of 100-300m in the Cloud
- In some cases up to 1km
- **100G-SR4 over OM4 is limited to 100m**
- Dispersion limit of 25 Gbps in OM4
- No easy way to increase reach
- 100G-LR4 can do 10km over duplex SMF
- However 100G-LR4 is not cost-effective
- No easy way to make it size or power efficient

What to do???

Existing 100G Optics Standards missed the Web/Cloud Datacenter

- **No cost-effective solution for 100-500m Reach**
- SR4 limited to 100m
- LR4 not cost-effective
- **100G-CFP MSA does not help**
- Very large, power hungry, and expensive
- Even CFP2 is way too large
- Many Standards Meetings, limited Progress
- Existing vendors protecting their turf

A cost-effective 100G Solution for the Cloud Datacenter is Needed

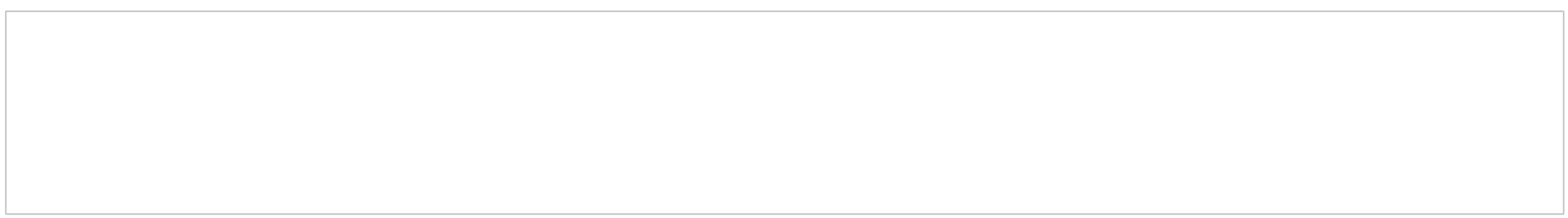
- **Goal is to minimize overall system cost**
- Total cost = Laser + Fiber + Power
- **Maximize 100G port density**
- Allow 48 ports 100G per 1U
- Minimum Reach 300m
- Able to support 500m up to 1km

Existing IEEE Standards have not addressed this

Solution: SiliconPhotonics over parallel Single Mode Fiber (pSMF)

- **Lowest overall system cost**
- Lowest cost fiber
- Lowest cost transceiver
- Lowest power transceiver
- **Highest 100G port density**
- Allows more than 48 ports 100G per 1U
- Supports 10m - 1km reach
- One solution can handle all requirements

Parallel 24F Fiber Cable



LEONI Fiber Optics GmbH

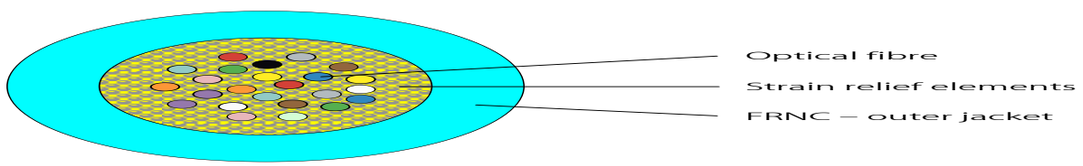
LEONI

Technisches Datenblatt – Technical Data Sheet –Technisches Datenblatt – Technical Data Sheet –Technisches Datenblatt – Technical Data Sheet

LEONI Part No.: **8407112FI655ZSYLX**

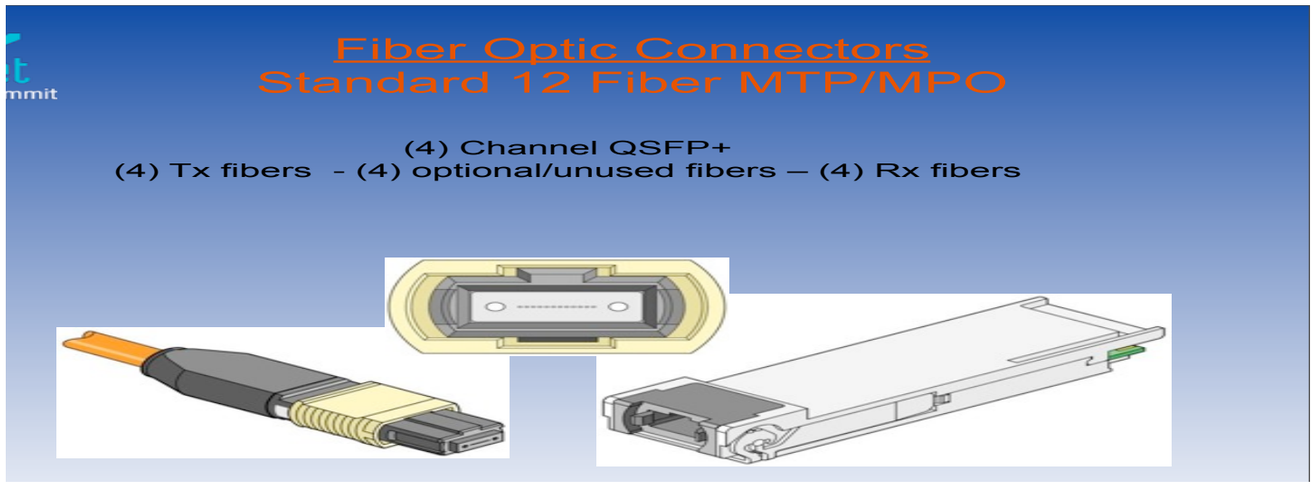
I-F(ZN)H 24G50/125 OM3 4.5 mm

Profile view:



12 duplex channels in 4.5mm, 12X denser than Cat-5e
Much lower cost than individual duplex fiber cables

MTP/MPO Multi-Fiber Connector



Invented by NTT in Japan in 1980's for Telecom
This has become the standard for multi-fiber termination in large-scale data centers

MTP/MPO Multi-fiber Connector

Maintaining Polarity In Cassette-Based Systems

PANDUIT TECHNICAL REFERENCE TR39

Purpose

Optical fiber links typically require two fibers to make a complete circuit. Optical transceivers have a transmit side and receive side, and typically employee a duplex fiber connector as the interface. In any installation, it is important to ensure that the optical transmitter at one end is connected to the optical receiver at the other. This matching of the transmit signal (Tx) to the receive equipment (Rx) at both ends of the fiber optic link is referred to as polarity. For traditional cabling systems using single fiber connectors, such as LC or SC, maintaining polarity is as simple as insuring that the A side of one connector pair matches to the B side of the other connector pair in any patch cord or permanent link. This procedure is well documented in the TIA/EIA-568-B.1 standard.

Pre-terminated, high-density cabling systems based on MTP*/MPO array connectivity require a new set of design rules and have their own requirements for maintaining proper polarity. In this document, three different methods for maintaining polarity in pre-terminated MTP* systems are reviewed. These three methods are defined by TIA/EIA-568-B.1-7. The methods define installation and polarity management practices, and provide guidance in the deployment of these types of fiber array links. Once a method is chosen, these practices must be put into place to insure proper signaling throughout the installation.

MTP*/MPO Array Connectors

As a single fiber connector terminates 1 fiber per connector, array connectors terminate multiple fibers in a single high-density interface. 12-fiber array connectors are the most common, though 4-, 6- and 8-fiber connectors are also available. Array connectors are employed in high-density permanent link installations and can be found in pre-terminated cassettes, trunk and hydra cable assemblies used extensively in data centers. Cassettes and hydra cable assemblies transition the high-density cabling on the permanent link of the installation to the single fiber connectors required by the transceivers in the switches.

Array connectors, shown in Figure 1, are pin and socket connectors -- requiring a male side and a female side. Cassettes and hydra cable assemblies are typically manufactured with a Male (pinned) connector. Trunk cable assemblies typically support a Female (unpinned) connector. The connectors are also keyed to ensure that proper endface orientation occurs during the mating process. Generally, when looking at the endface of the connector with the key is in the "up" position, Fiber 1 is the far left fiber on the same side as the white dot on the connector, shown in Figure 2.

Figure 1 — 12-Fiber MTP* Male and Female Connectors

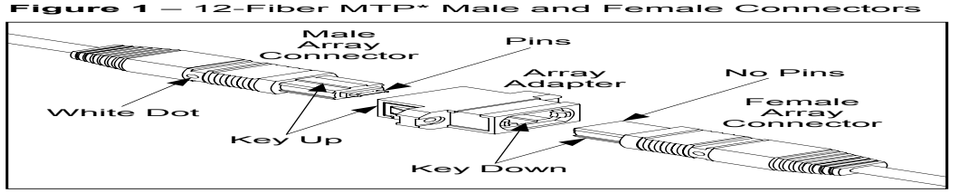
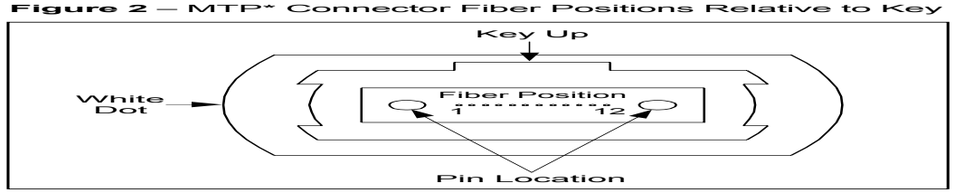
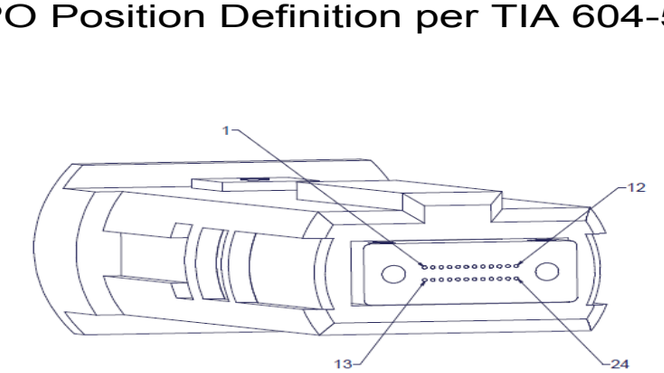


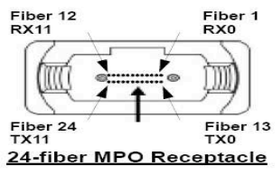
Figure 2 — MTP* Connector Fiber Positions Relative to Key



Supports 12 fibers per row, 24 per 2 rows, etc
Highest density fiber connector on the market

24 Fiber MPO Connector

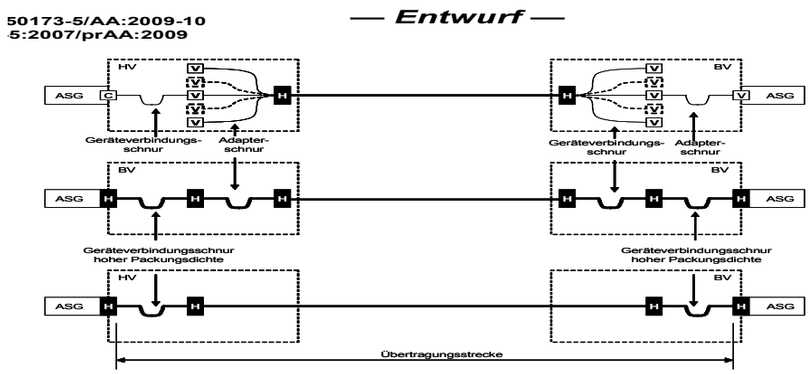




24-fiber MPO Receptacle

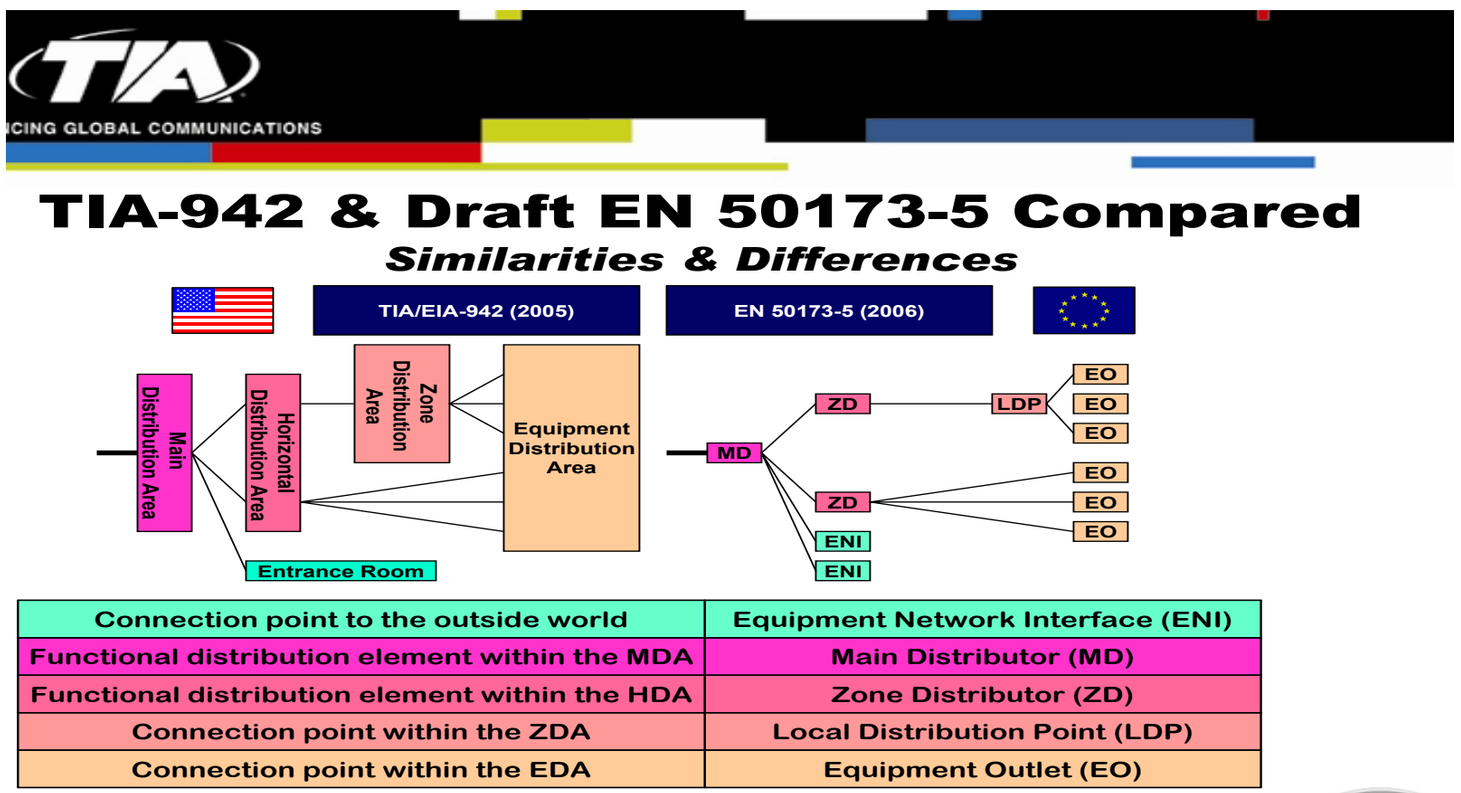
24F MTP Connector can handle 3x40/100G or 12 10G Ethernet channels

EN 50173-5 (2007) Standard



Only two fiber connectors in EN standard: LC for duplex
MPO connector for parallel fiber structured cabling

TIA-942 and EN 50173-5 Datacenter Fiber Standards



Different terminology, same basic idea

Fiber Cable Cost Comparison

Fiber Cable	$/8F 300m	$/2F 300m	Relative Cost
2F OM4	$720	$180	540%
24F OM4	$566.67	$141.67	425%
2F SMF	$266	$66.66	200%
24F SMF	$133	$33.33	100%

Parallel SMF cable is by far the lowest cost solution

100G Ports Total Cost Comparison

Element	Current Choice	Best Choice	Cost Reduction	Comments
Fiber	pMMF	pSMF	75-80%	Parallel SMF is 1/4 the cost of pMMF
Optics	VCSEL	SiPh	TBD	Silicon Photonics is lower cost than VCSEL
Reliability	Good	Highest	TBD	Significant life cycle Cost Reduction
Power	2W	1W	50%	Power Reduction is key for density
Total				

Total Cost = Equipment Laser + Fiber + Power (3Y)

Datacenter Optics Conclusions

- **Silicon Photonics is good**
- Lowest cost, lowest power, highest reliability
- Supports 100m-300m reach requirement
- **Parallel SMF Cable is good**
- Saves 75% in cost over OM4
- However most installed cable is MMF
- **Fewer Fiber Connectors is good**
- Reduces installation costs
- Fewer things that can go wrong

Summary

- **Datacenter Switching back on Moore's Law**
- Rapid cost-performance improvements ahead
- Expect 2X improvement every 2 Years
- **40G and 100G Adoption limited by costs**
- What matters is cost of bandwidth
- Particular problem is optics costs
- **Silicon Photonics with pSMF look promising**
- Lowest known optics and fiber cost
- A lot less cables and connectors