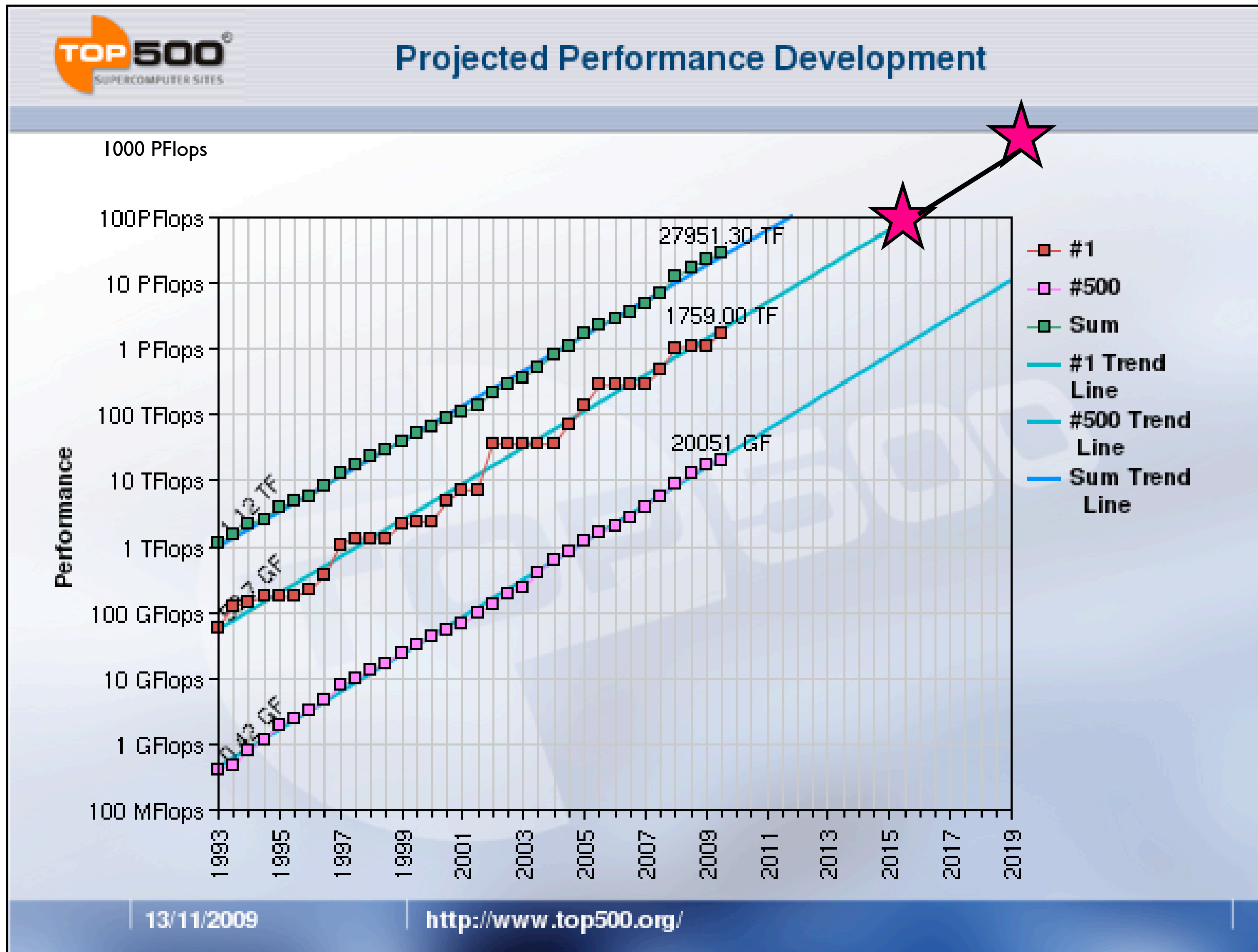
The Road to Exascale

Andreas Bechtolsheim
Arista Networks

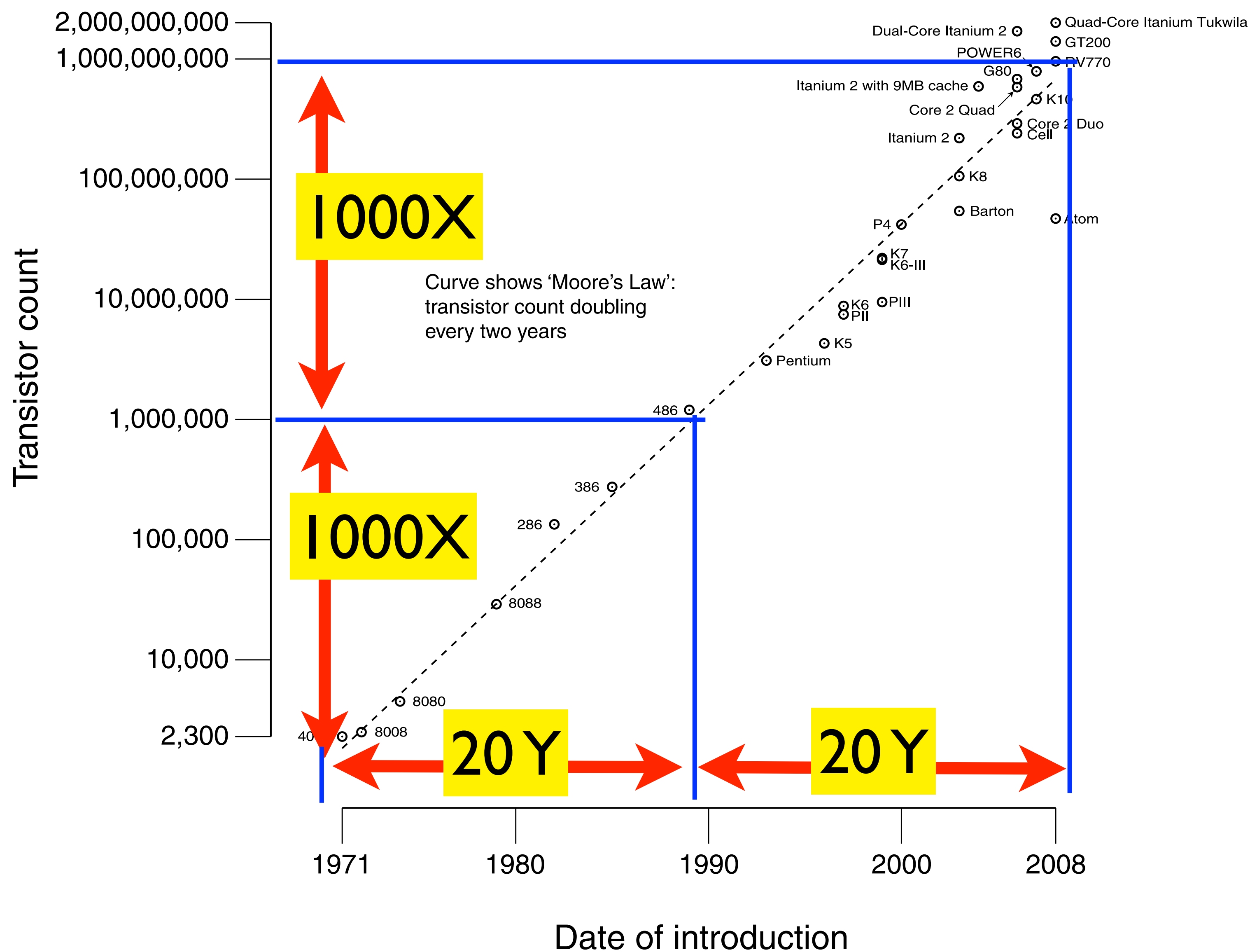
Top500 Projected Performance



Top 500 List Observations

- It took 11 years to get from 1 TF to 1 PF
- Performance doubled approximately every year
- Assuming the trend continues, 1 EF by 2020
- Question can this be achieved?
- Moore’s law predicts 2X Transistors every 2 years
- Need to double every year to achieve EF in 2020

Moore's Law 1970-2010



Semiconductor Technology Roadmap

Product Functions/Chip
[Giga (10^9) - bits, transistors]

1.E+03
1.E+02
1.E+01
1.E+00
1.E-01
1.E-02

1995 2000 2005 2010 2015 2020 2025

Year of Production

2007 - 2022 ITRS Range

100X

12Y

Flash Bits/Chip (Gbits) Multi-Level-Cell (4bit MLC)

Flash Bits/Chip (Gbits) Multi-Level-Cell (2bit MLC)

Flash Bits/Chip (Gbits) Single-Level-Cell (SLC)

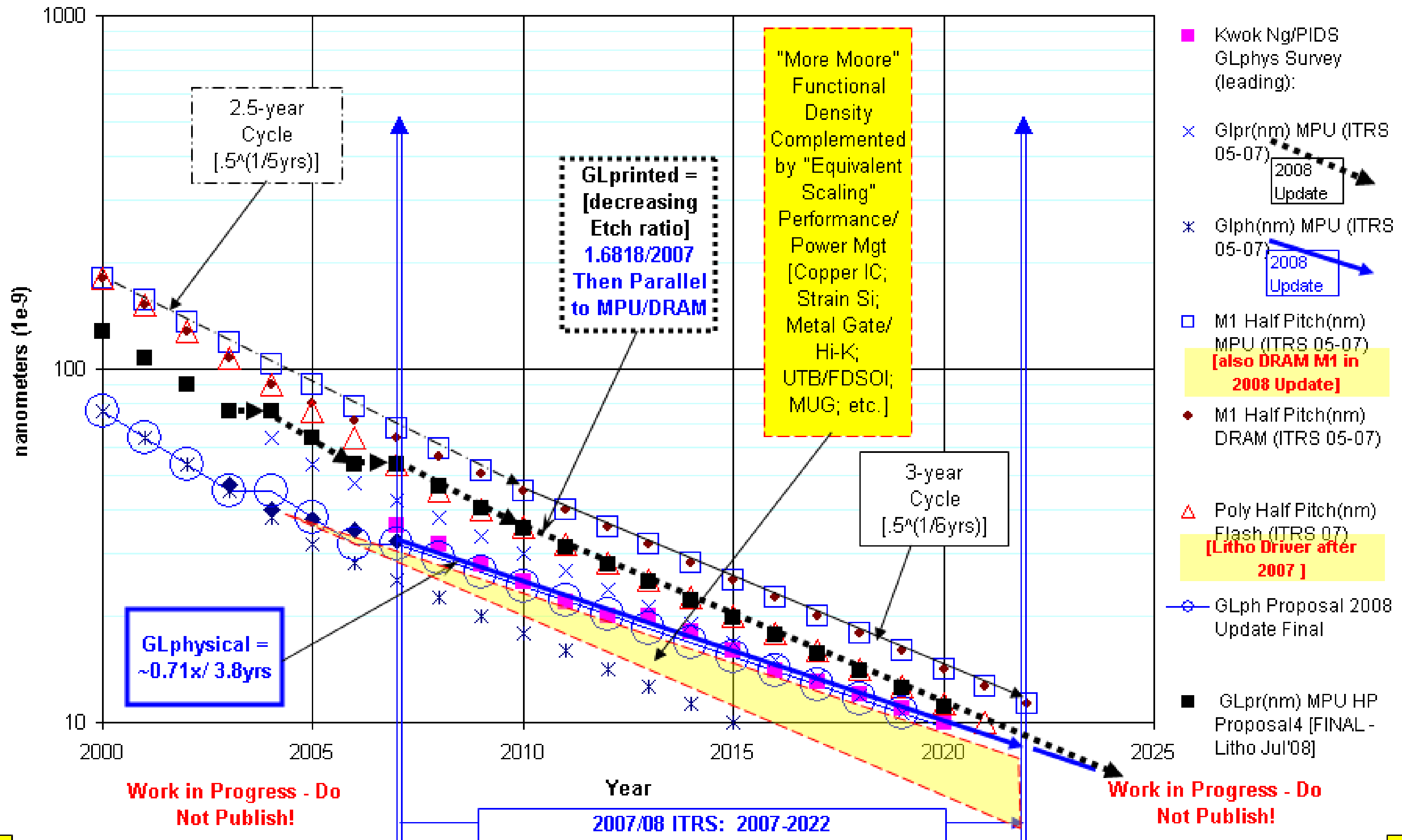
DRAM Bits/Chip (Gbits)

MPU GTransistors/Chip - high-performance (hp)

MPU GTransistors/Chip - cost-performanc (cp)

Average Industry "Moores Law" :
2x Functions/chip Per 2 Years

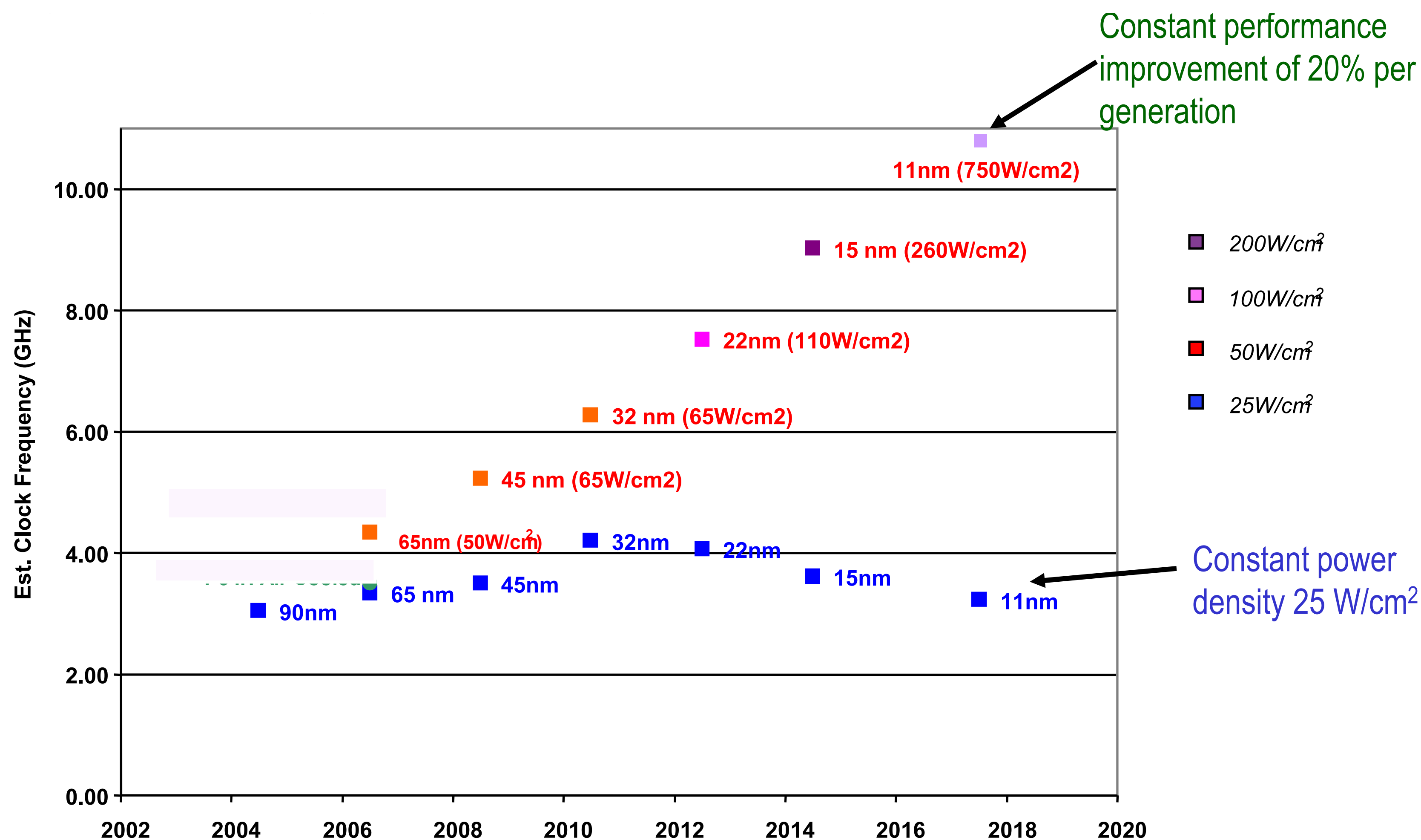
Semiconductor Technology Roadmap



Source: ITRS 2008

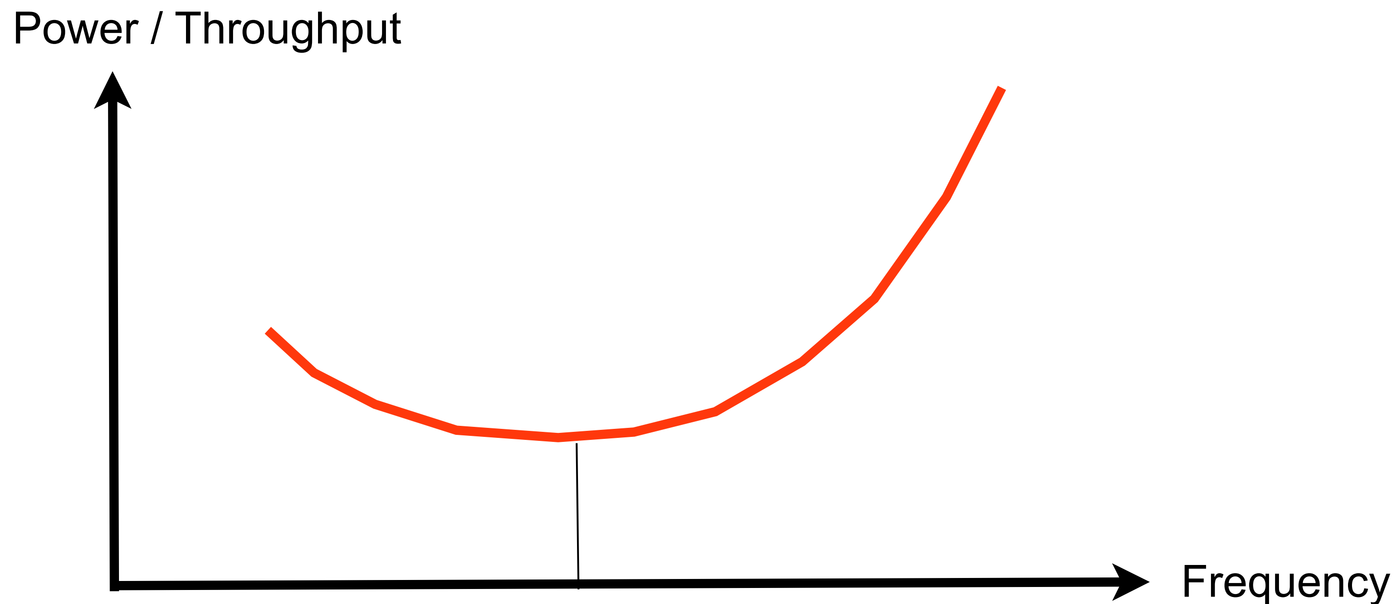
What can one do with 100 Billion Transistors/chip?

More CPU Cores!
More Floating Point Units!
More Cache!
More Memory Bandwidth!
More I/O



Source: *D. Frank, C. Tyberg,* IBM Research

Power Efficiency (Power per Throughput)



Power = Clock * Capacitance * Vdd^2

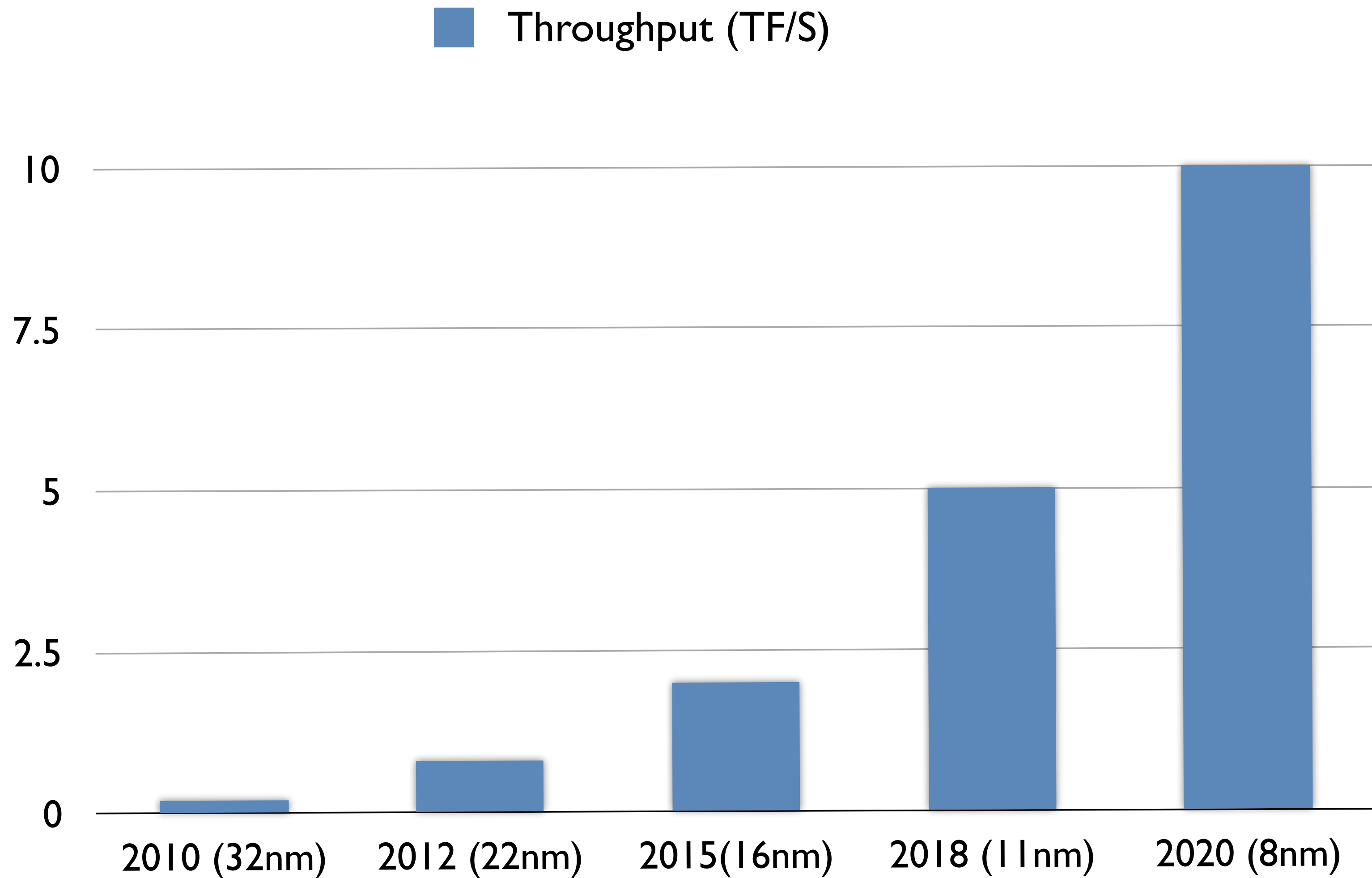
Higher-frequency designs consume much more power

The Basic Math: "More than Moore"

Aggregate Performance = C * F * I * N

	Element	Increase
C	Cores per Module	40% / Y
F	Frequency	5% / Y
I	Instruction Efficiency	15% / Y
N	Number of CPUs	20% / Y
	TOTAL	100% / Y

Primary increase is in the number of cores



TeraFlops/CPU Socket over Time
Throughput (TF/S)
10
7.5
5
2.5
0
2010 (32nm)
2012 (22nm)
2015(16nm)
2018 (11nm)
2020 (8nm)

Comments on GP-GPU

- **Technology Constraints are the same for all architectures**
  - Number of Transistors per die
  - Number of Memory channels
  - Number of I/O pins
  - Maximum Power per Chip
- **Difference is how the Transistors are Used**
  - CPUs are optimized for general purpose throughput
  - GPUs are optimized for number of floating point units
- **CPU and GPU architectures will merge going forward**
  - General purpose CPUs with vector extensions and lots of FPUs
  - GP-GPUs will add support for much larger memory

General Purpose CPU 2010 => 2020

Year	2010	2020	Ratio
Clock Rate	2.5 GHz	4 GHz	5%/Y
FLOPS/Clock	4	16	4X
FLOPS/Core	10 GF	64 GF	6.4X
Cores/Module	16	160	10X
FLOPS/Module	160 GF	10 TF	64X
Mem Bandwidth	30 GB/s	2 TB/s	64X
M Bandwidth/F	0.2 B/F	0.2 B/F	=
IO Bandwidth	3 GB/s	192 GB/s	64X
IO Bandwidth/F	0.02 B/F	0.02 B/F	=
Power / Module	250W	250W	=
Power Efficiency	0.6 GF/W	40 GF/W	64X

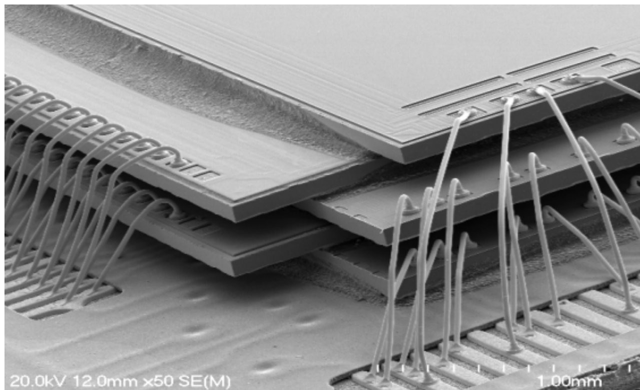
Scaling the CPU Throughput

- **Three Dimensions of Scalability**
  - Frequency, Cores, FLOPS/Core
- **Increasing Frequency is most difficult**
  - Limited by power consumption per core
- **Increasing the Number of Cores**
  - Moore’s Law predicts doubling every 2 years
- **Increasing FLOPS per Core**
  - Increase functional units, SIMD instructions
- **Throughput will Double every year**
  - Combination of number of cores and efficiency gains

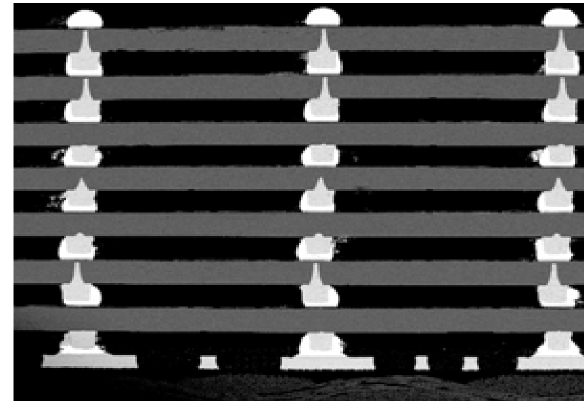
Challenge #1: Memory Bandwidth

- Memory bandwidth must grow with throughput
- 2020 CPU needs > 64X the memory bandwidth
- Traditional Package I/O pins are basically fixed
- Electrical signaling at the speed limit
- How to scale memory bandwidth?
- Solution: Multi-Chip 3D Packaging

Multi-Chip 3D Packaging



Wire bonded stacked die



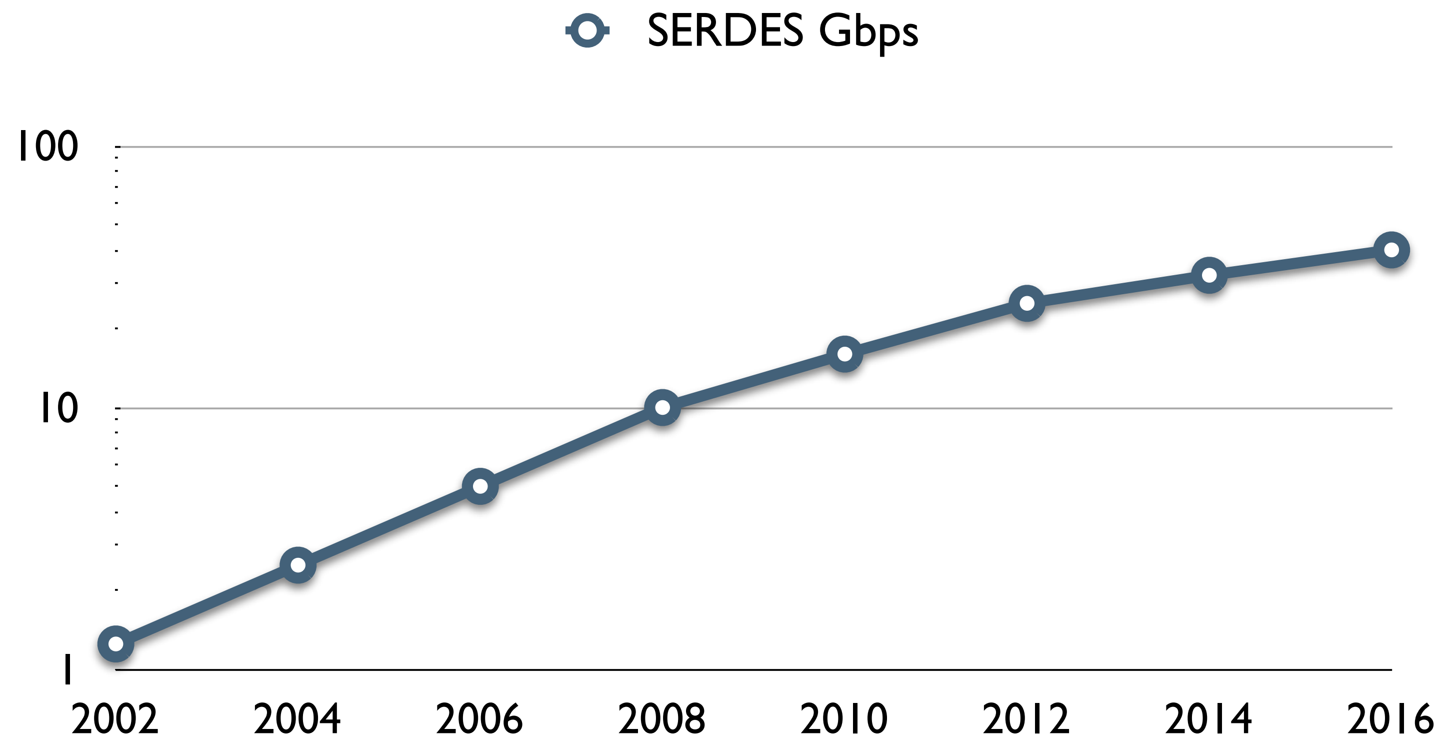
Thru-Si via Stacking

Need to combine CPU + Memory on one Module with lots and lots of memory channels

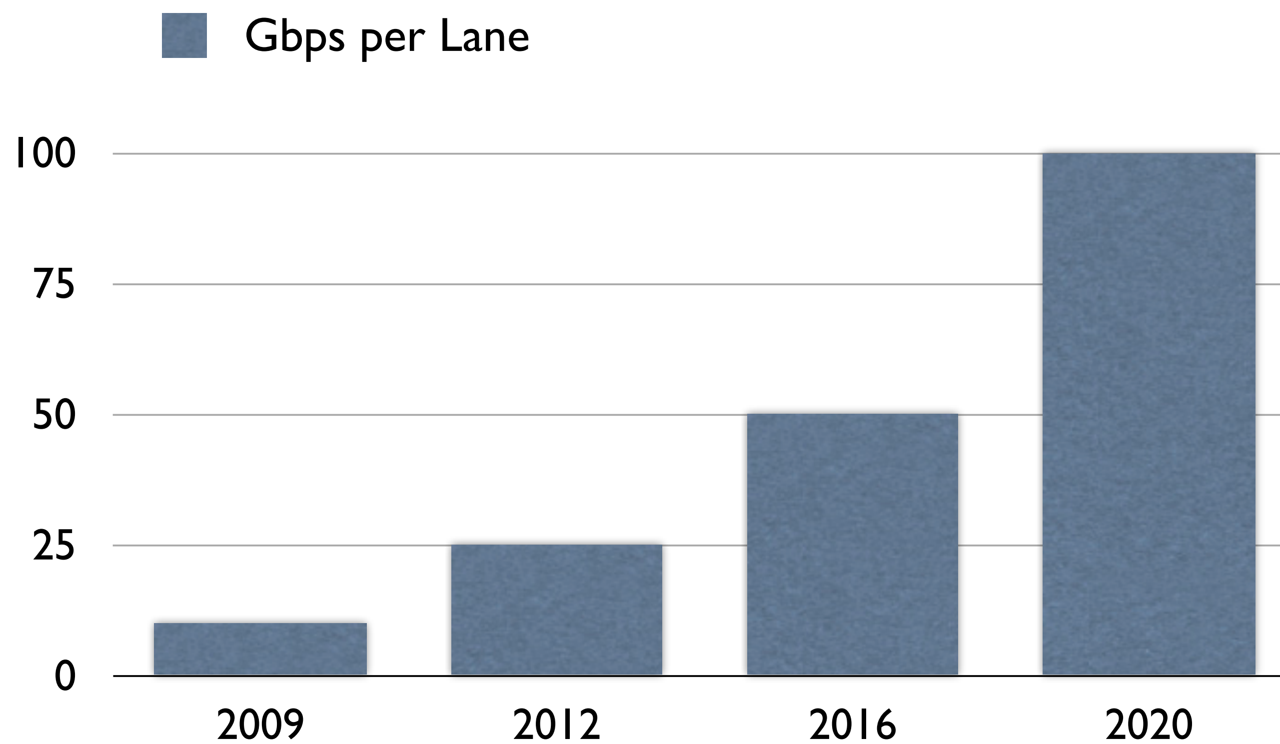
Challenge #2: I/O Bandwidth

- I/O bandwidth must grow with throughput
- 2020 CPU needs > 64X the I/O bandwidth
- Electrical signaling at the speed limit
- How to scale I/O bandwidth?
- Solution: Integrate NIC in CPU
- High-speed SERDES with MCM Optics

SERDES Speed / Channel

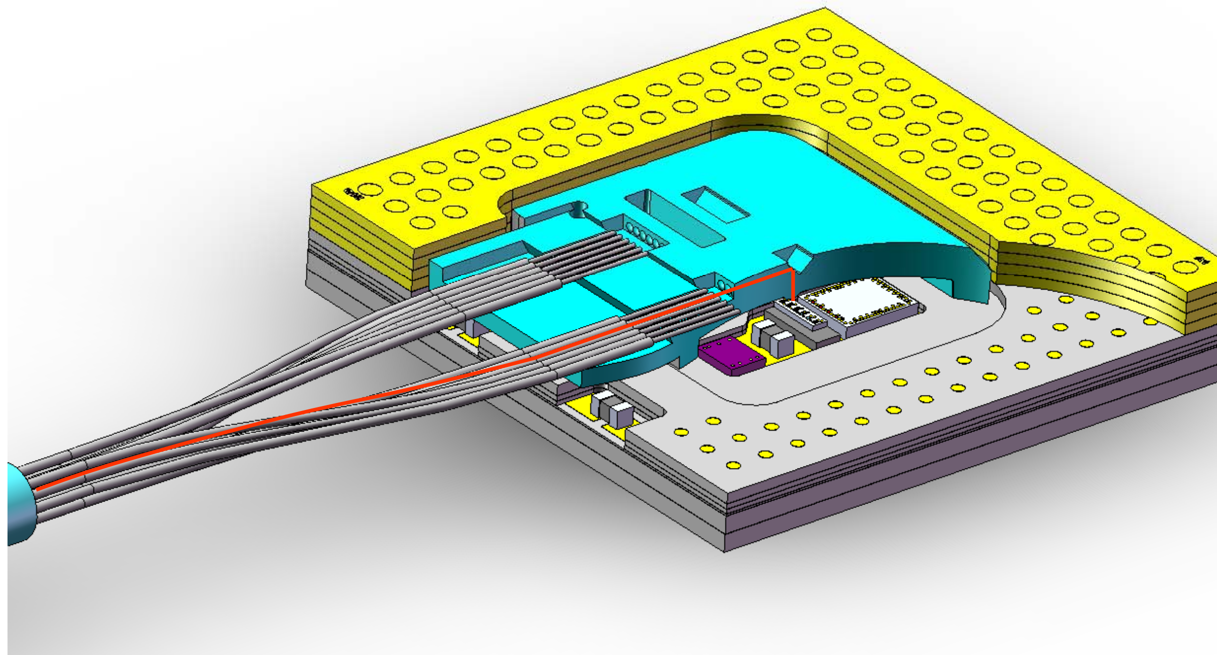


Expected Serdes Data Rate per Channel

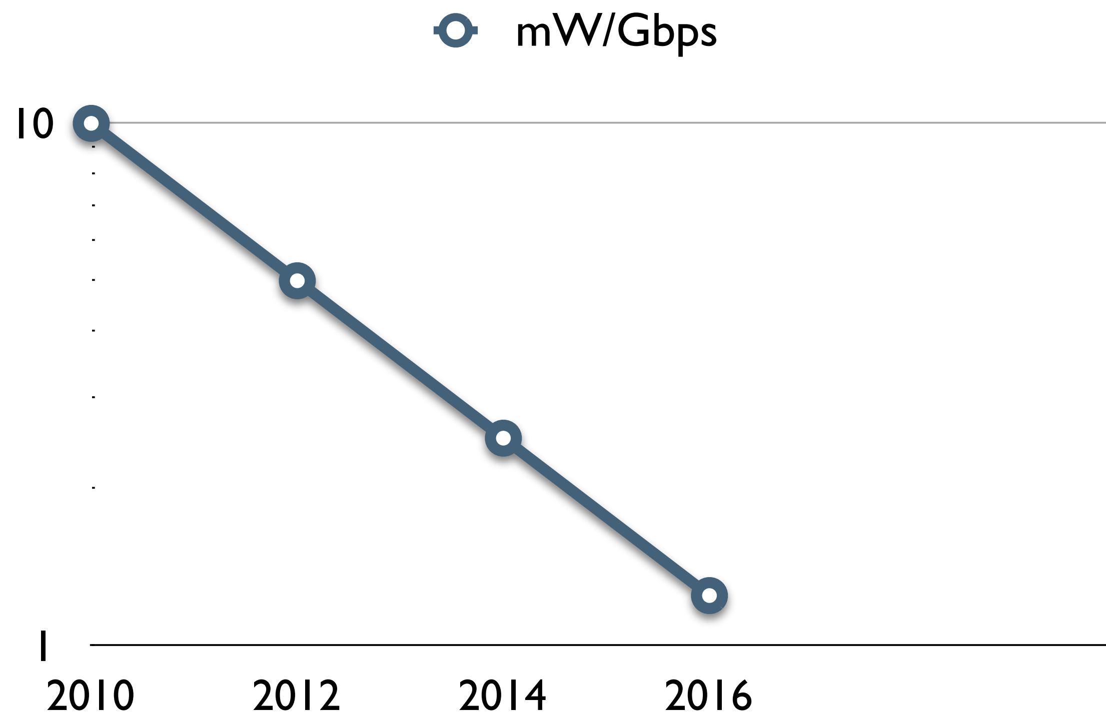


Improving with transistor speed, not Moore's law

Parallel Optical Interface with fiber pig-tail



Parallel Optics Power Outlook



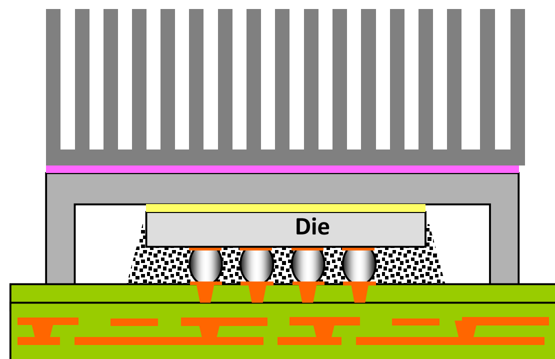
Challenge #3: Power

- Integrating memory and I/O on one MCM is very power efficient, but nevertheless *increases* power per MCM module
- How to cool?
- Solution: Microchannel Fluid Heatsinks

Microchannel Fluidic Heatsinks

Conventional

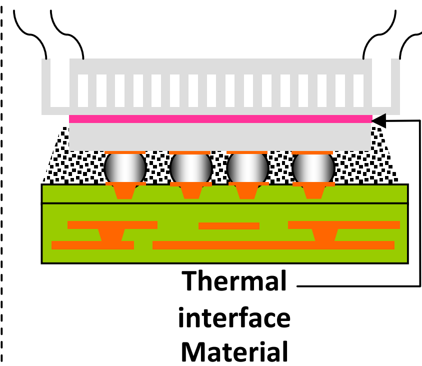
thermal Interconnects



Back-side integrated

fluidic heat sink using

TIM and inlets/outlets

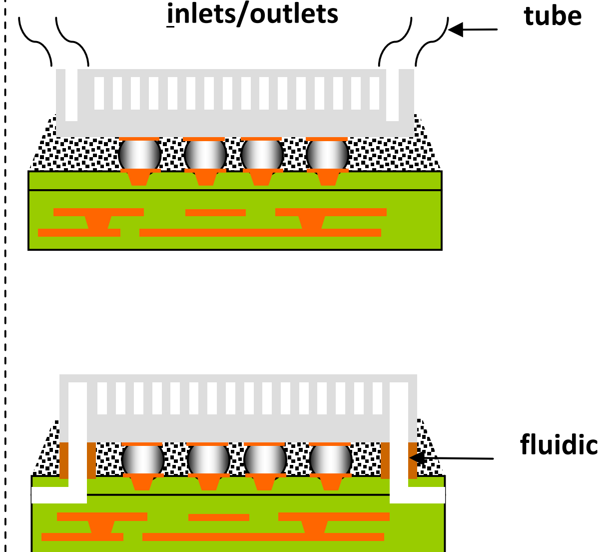


Back-side integrated

fluidic heat sink and

Back and front-side

inlets/outlets



Benefits of MCM Packaging

- Only way to achieve memory and I/O bandwidth
- Greatly reduces overall power consumption
- Enables denser packaging and better cooling

Future CPUs will look very different than todays

The Fabric

Three HPC Fabrics

Ethernet

Infiniband

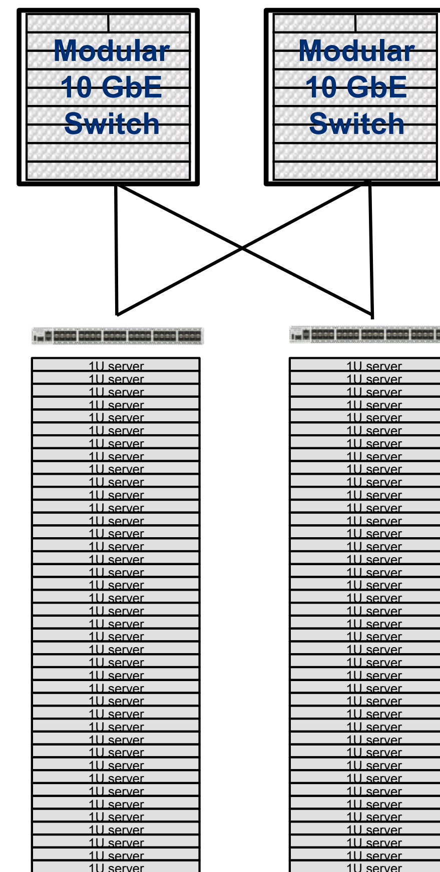
Proprietary

Ethernet

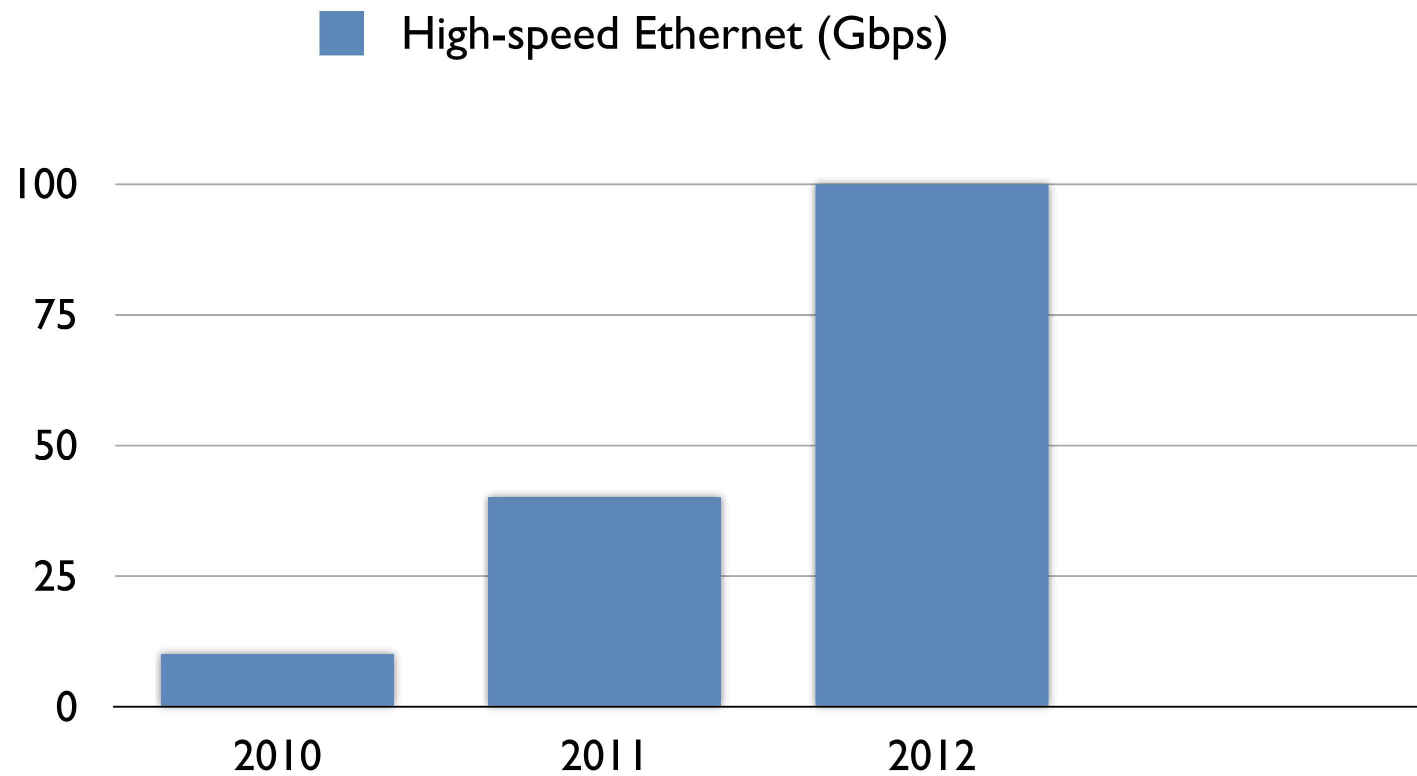
- Quite popular in HPC
  - 50% of Top 500 List
  - In particular Oil&Gas
- Advantages
  - Very easy to use
  - Low cost (Gigabit Ethernet)
- Disadvantages
  - 10 Gbps has been expensive (so far)
  - Limited Switch Scalability (until now)

Large Flat Fabric Design

- Core Switch
  - Hundred of 10G ports
  - Wire-speed architecture
- Leaf Switch
  - 48 1 or 10G ports
  - 4 or more 10G uplinks
- Overall Capacity
  - >10,000 ports
  - >10 Tbps throughput



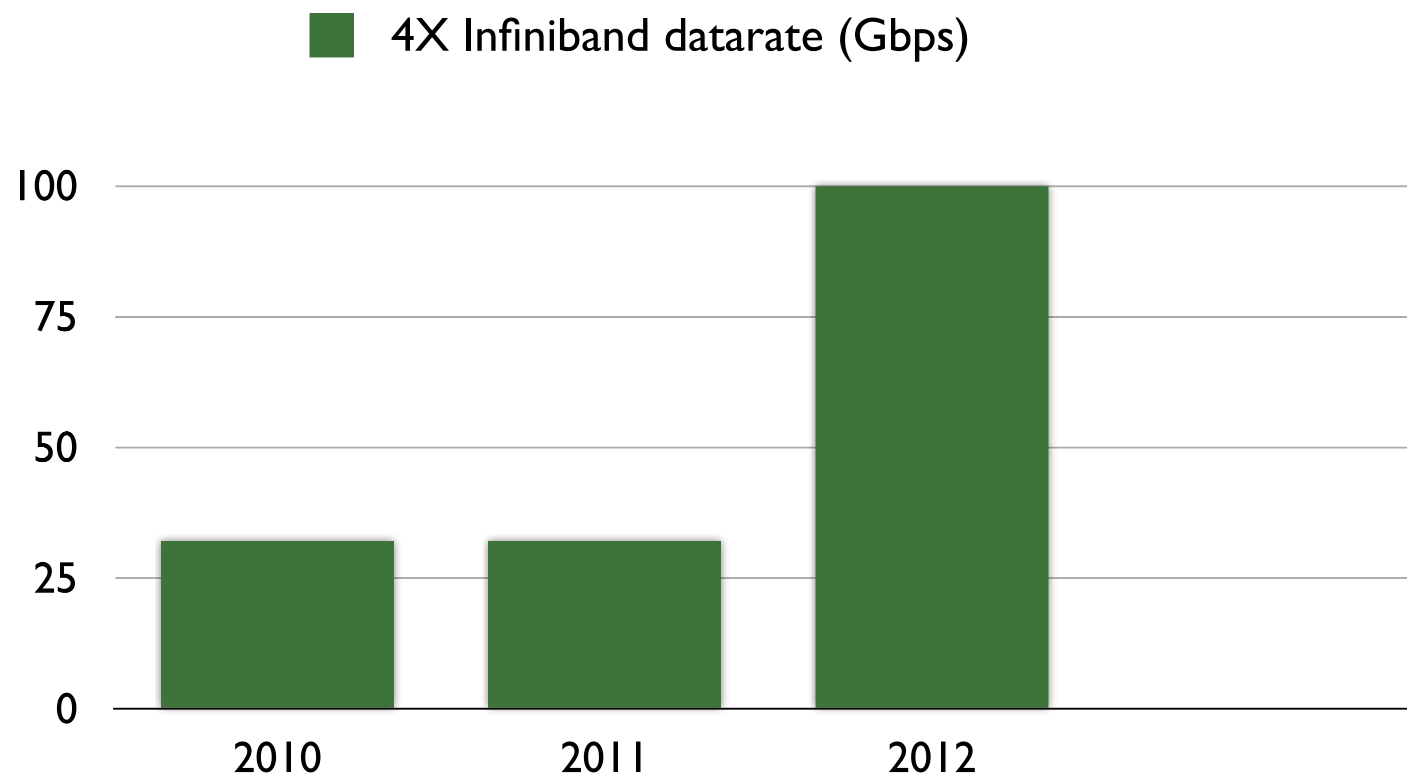
Ethernet Roadmap



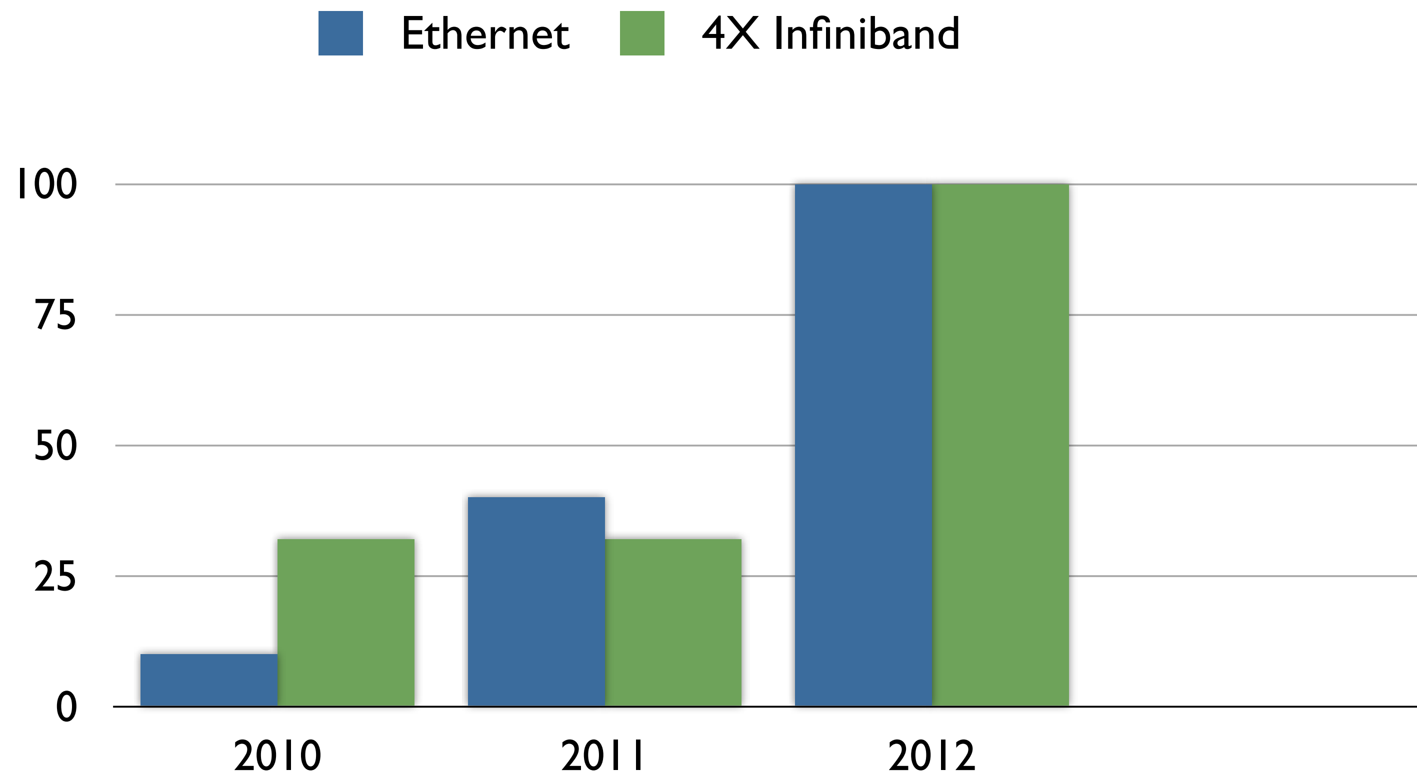
Infiniband

- Quite popular in HPC
  - 30% of Top 500 List
  - Particularly good for MPI
- Advantages
  - Latency (1- 2 usec)
  - Scalability (1000s of nodes)
- Disadvantages
  - Difficult to manage
  - Multi-stage CLOS Fabric effects

Infiniband Roadmap



E/IB Speed Convergence



Ethernet / IB Convergence

- Speed is converging
  - Because both will use same physical layer
- HCA is converging
  - Same Mellanox HCA will support both
- Protocols are converging
  - "Infiniband over Ethernet"

Remaining delta is packet header
and switch architecture / design

Proprietary Fabrics

- Quite popular in Top 50
  - Highest performance
  - Support shared memory
- Advantages
  - Not constrained by existing standards
  - Lower latency / more bandwidth
- Disadvantages
  - Custom design, not standards based
  - Needs to be tightly coupled to CPU

Fabric Summary

- **Choice of Fabric Depends on Application**
  - Proprietary Fabrics at the very high-end
  - Infiniband for high I/O intensive workloads
  - Ethernet for all other workloads
- **Fabric Protocols are Converging**
  - RDMA over Ethernet
  - “Infiniband over Ethernet”
- **Physical Layer is Converging**
  - All fabrics will use same fiber optics and fiber cabling
  - Speed and cost of physical layer will become same

Road to Exascale Summary

- **Cluster Throughput Doubling every year**
  - Increasing Nodes, Cores, Frequency and FLOPS/Core
- **Memory and I/O Bandwidth challenging**
  - Requires MCM multi-chip packaging
- **Fabric scaling quite challenging**
  - More bandwidth, lower latency, larger switches needed
- **ExaFlop looks feasible by 2020**
  - Key issue is power efficiency and system size
- **Writing software is most difficult**
  - O (10M) Parallelism