Technologies for Data-Intensive Computing

Andreas Bechtolsheim
Sun Microsystems Inc

October 26, 2009

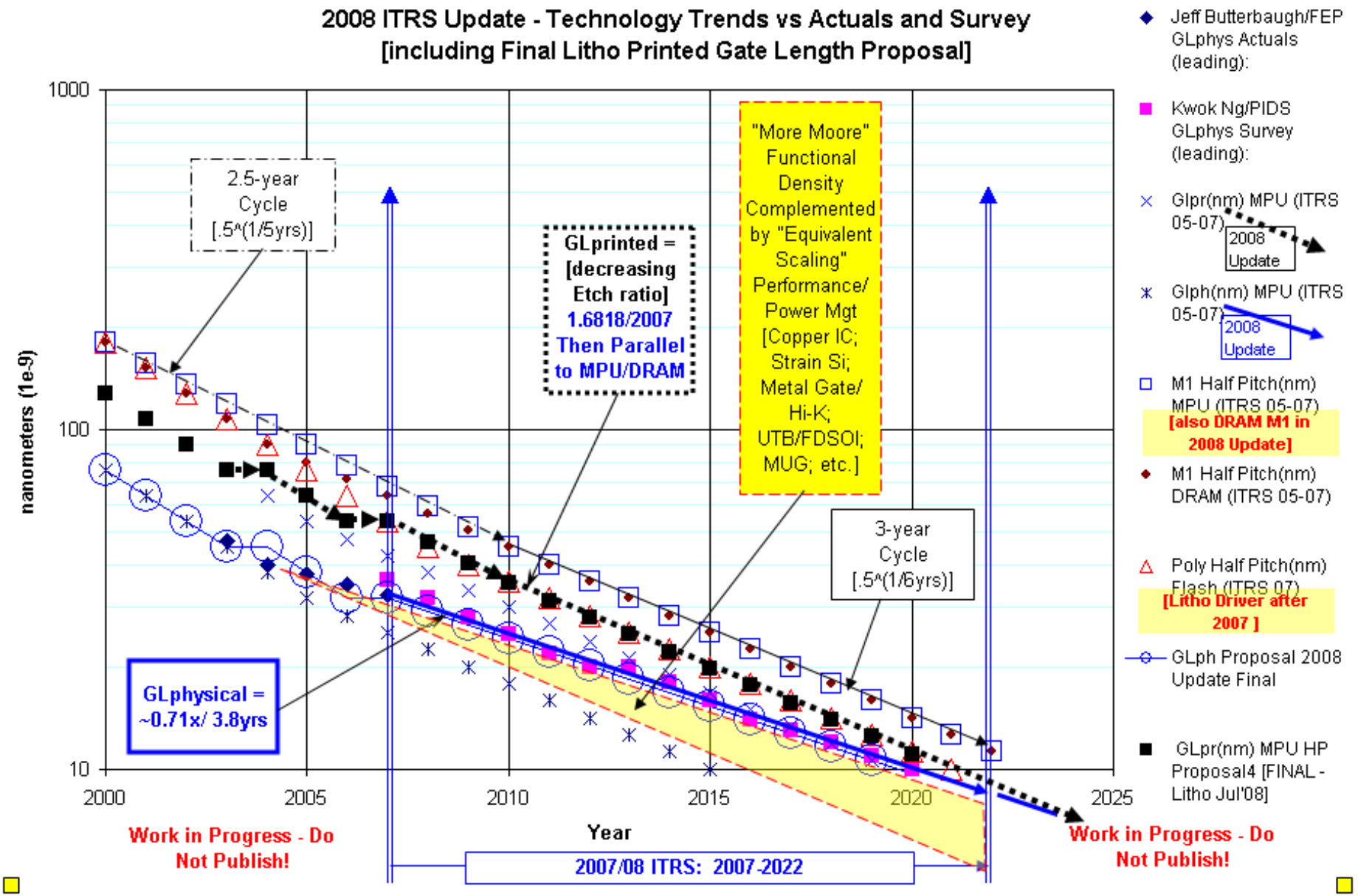
Challenges

- Semiconductor Roadmap
- CPU Roadmaps
- Memory Bottleneck
- Packaging Technology
- Power and Cooling
- Fabric Interconnect
- Exploiting Parallelism

Major Bottlenecks Ahead

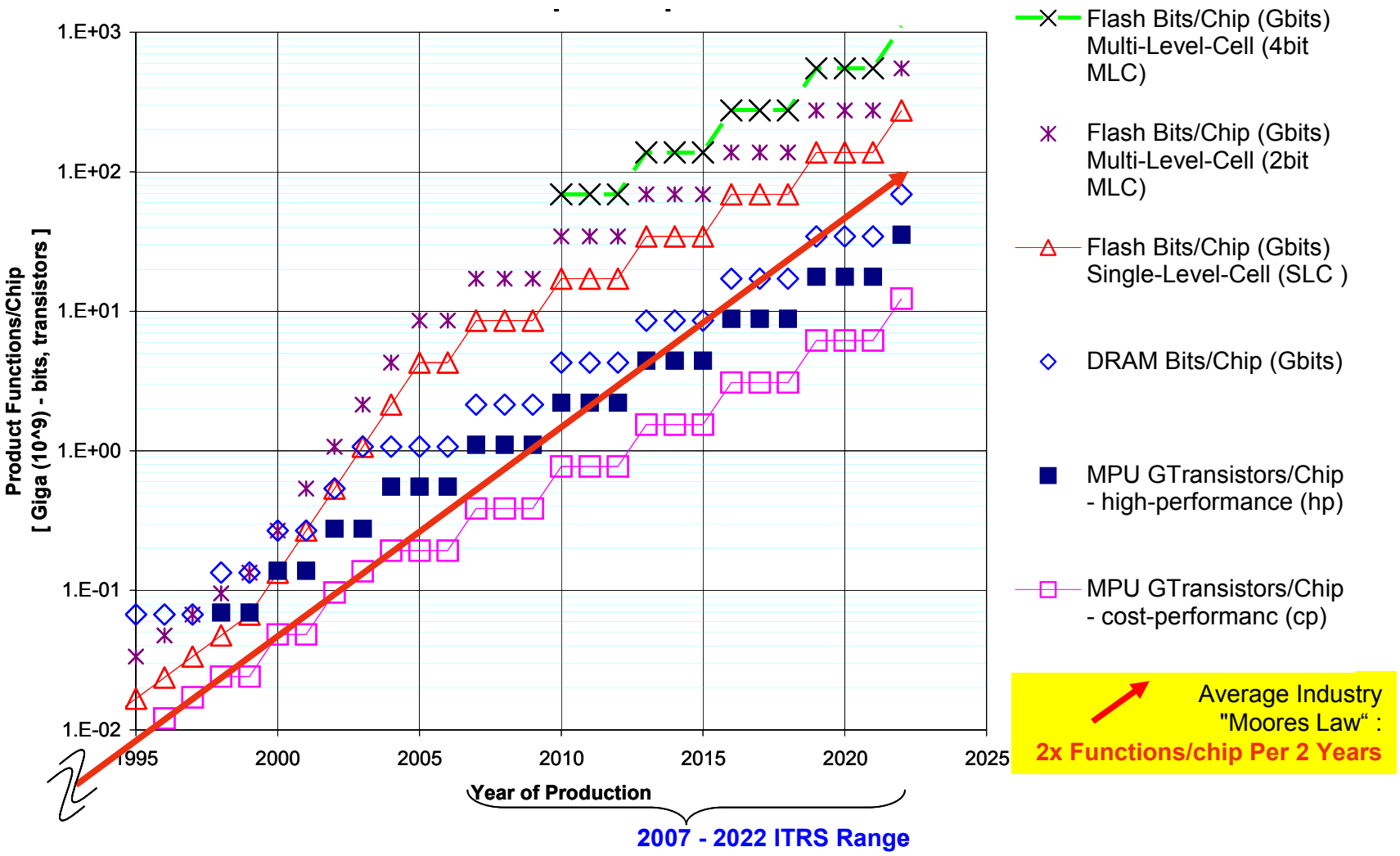
- Scaling CPU Performance
- Scaling Memory Bandwidth
- Scaling Interconnect
- Scaling Input/Output
- Managing Power

Semiconductor Technology Roadmap

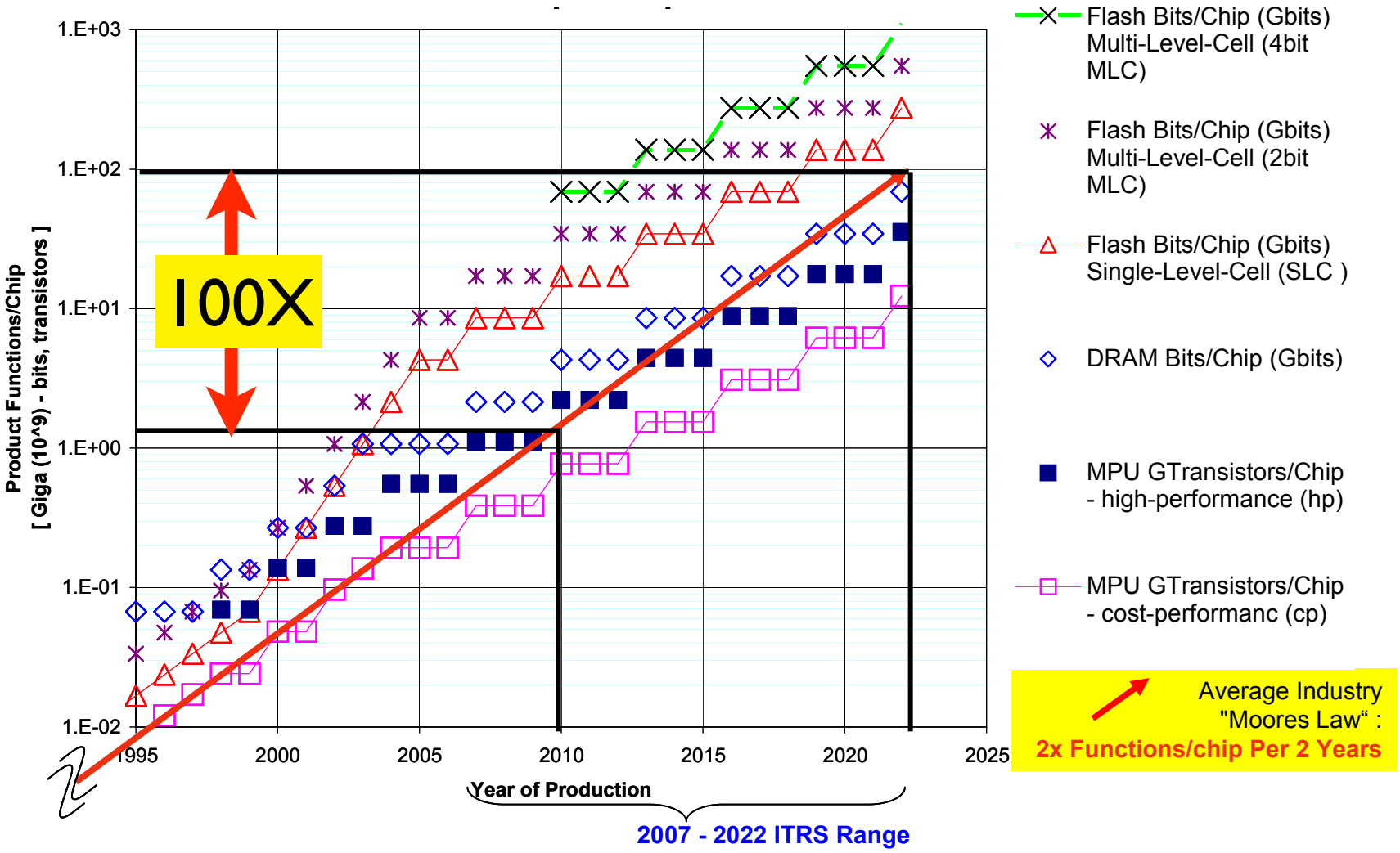


Source: ITRS 2008

Semiconductor Technology Roadmap



Semiconductor Technology Roadmap



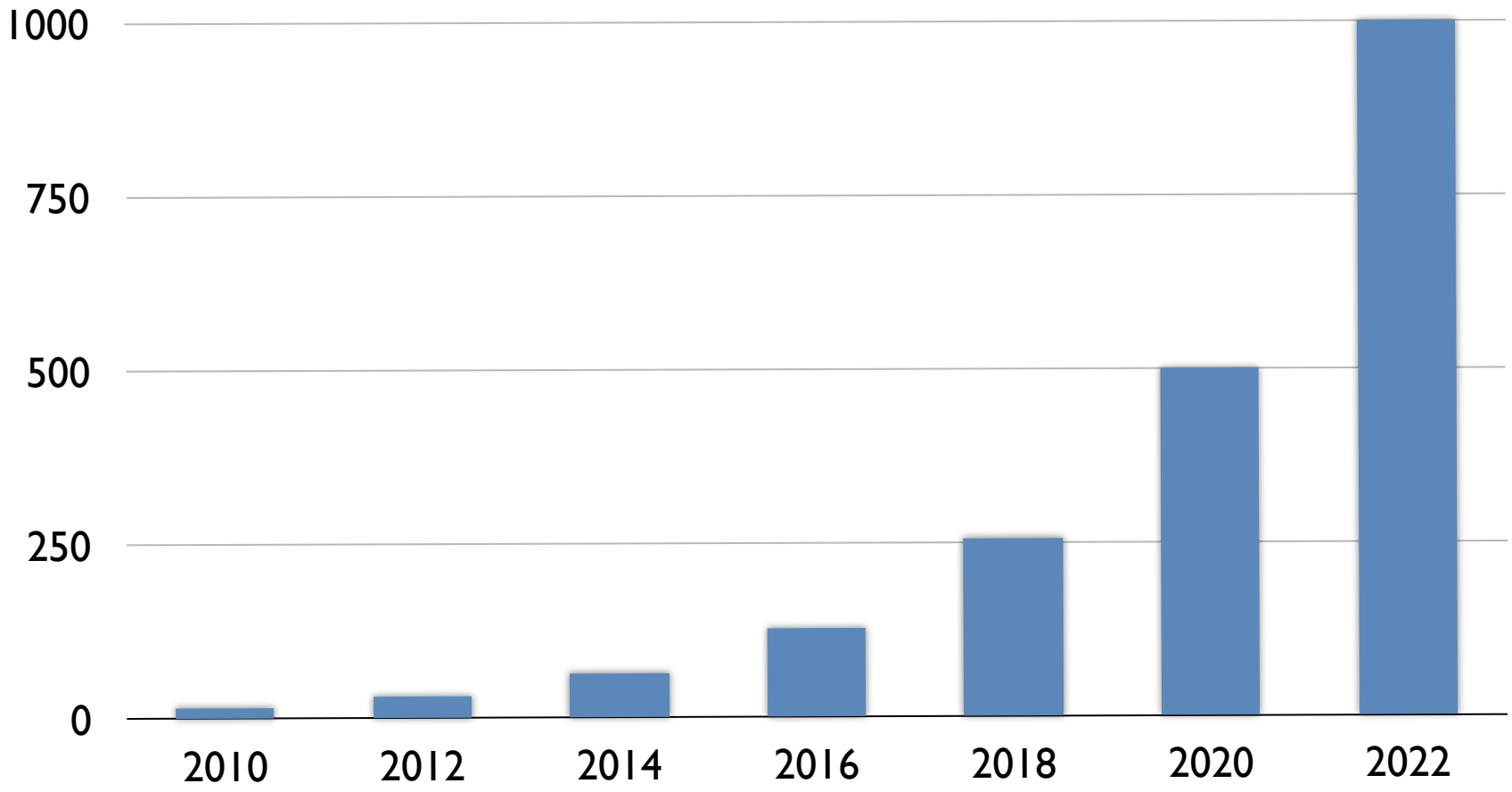
Emerging Devices

- Lots of Research on New Materials
  - Carbon Nano-Tubes (CNT)
  - Graphene Nanoribbons
  - Ferroelectric Materials
  - Phase-Change Materials
  - Nano-ionic Memories
- Challenge is to find out which are worthy
  - Long road from research lab to volume production

Silicon Roadmap Predictions

- 128X increase in transistors per chip by 2022
  - 1K Core CPUs
  - 512 Gbit DRAM
  - 8 Tbit FLASH
- What does this mean for data-intensive applications?

Cores per CPU Socket over Time



CPU Module [Socket] Roadmap

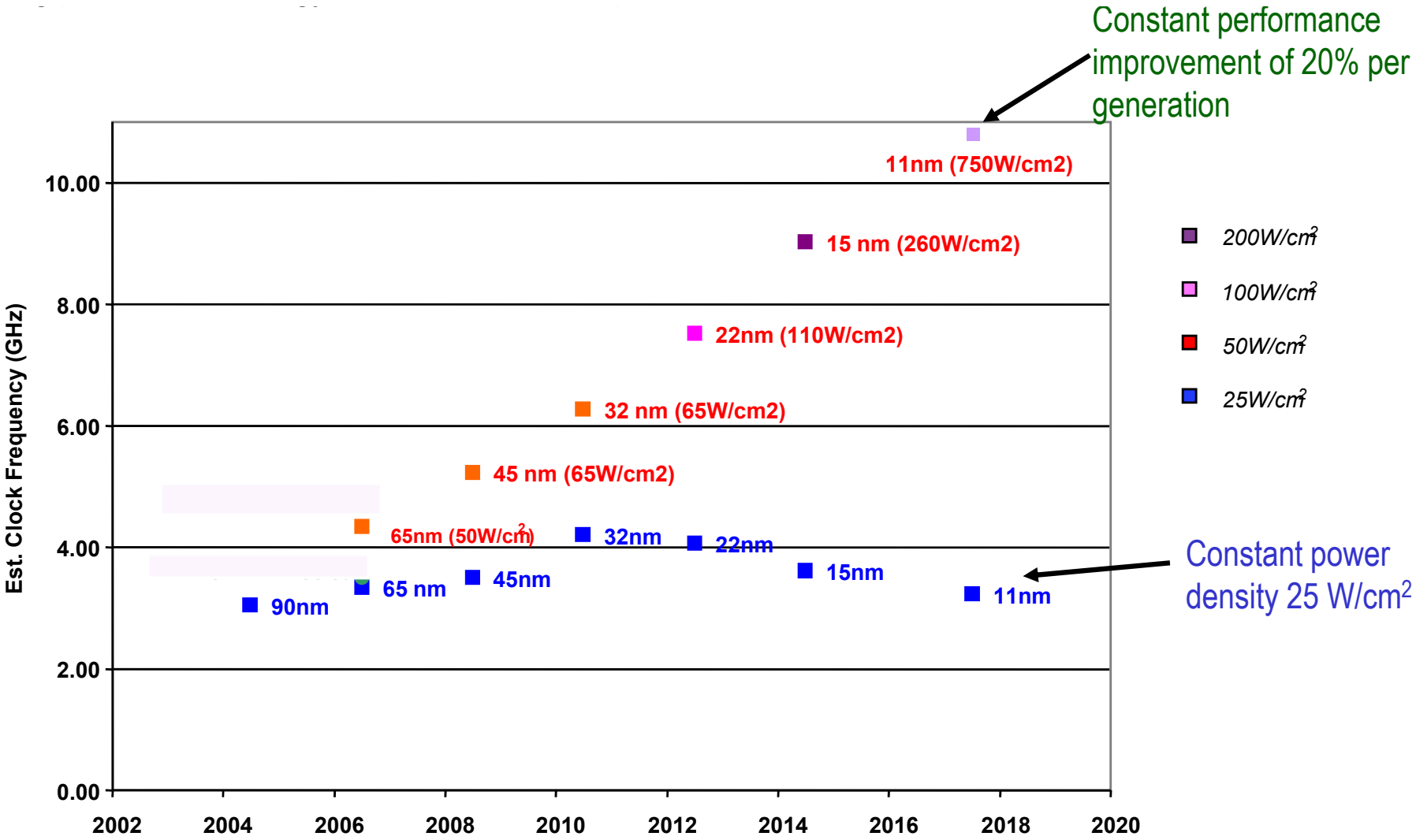
Year	2010	2022	Ratio
Clock Rate	2.5 GHz	5 GHz	2X
Cores	16	1024	64X
Core GHz	40	5120	128X
Mem Bandwidth	40 GB/s	2.5 TB/s	64X
M Bandwidth/CGHz	1	0.5	0.5X
IO Bandwidth	2 GB/s	250 GB/s	128X
IO Bandwidth/CGHz	0.1	0.05	0.5X
Power / Module	200W	500W	2X
Power Efficiency	10W/CGHz	0.1W/CGHz	100X

Amazing but this is what technology predicts

The CPU Challenge

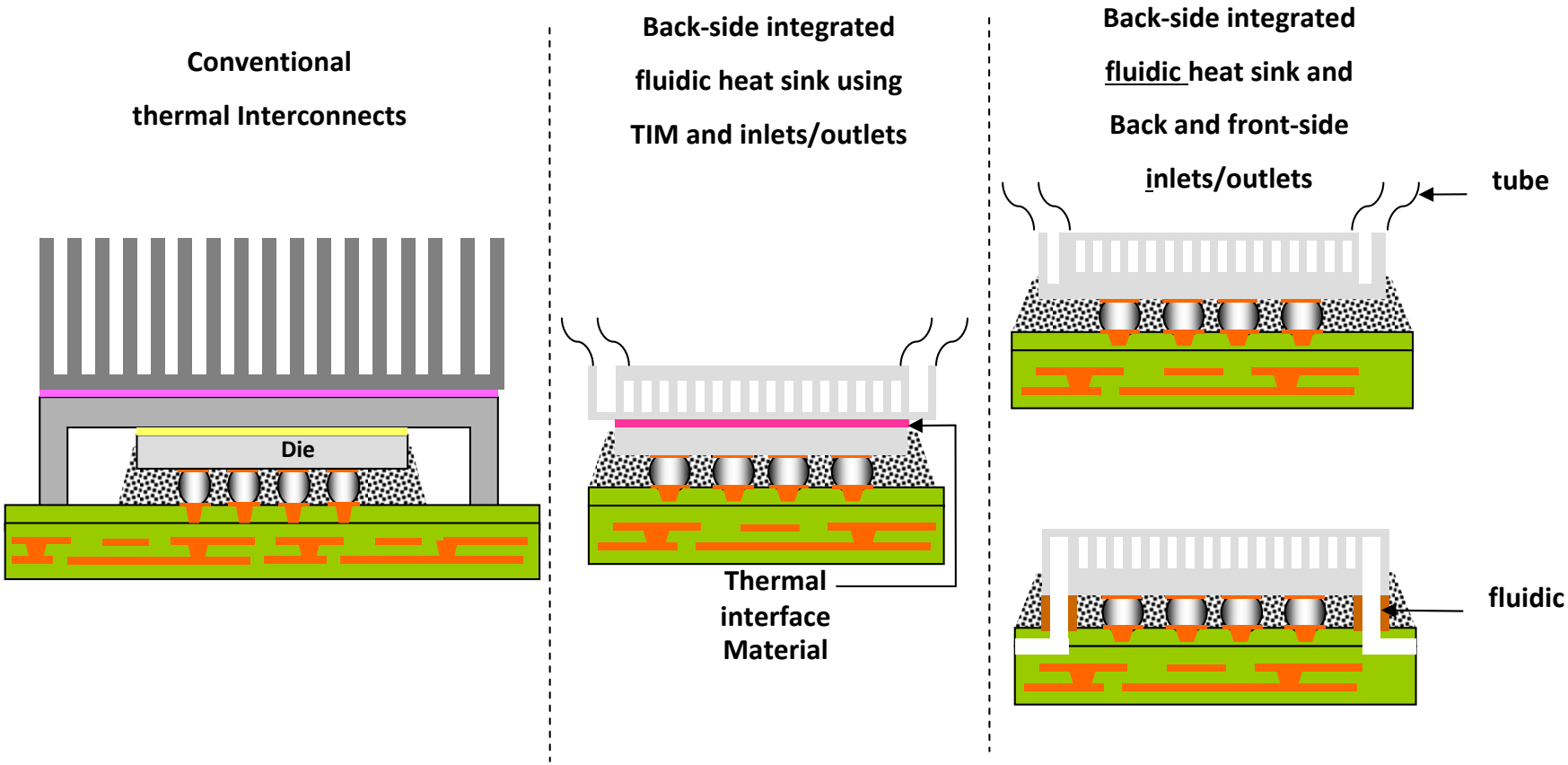
- CPU Clock Rates increasing at ~ 5%/Year
- CPU Cores doubling every other year
- Cache sizes and efficiencies also improving
- Primary constraint is power

Power per Core

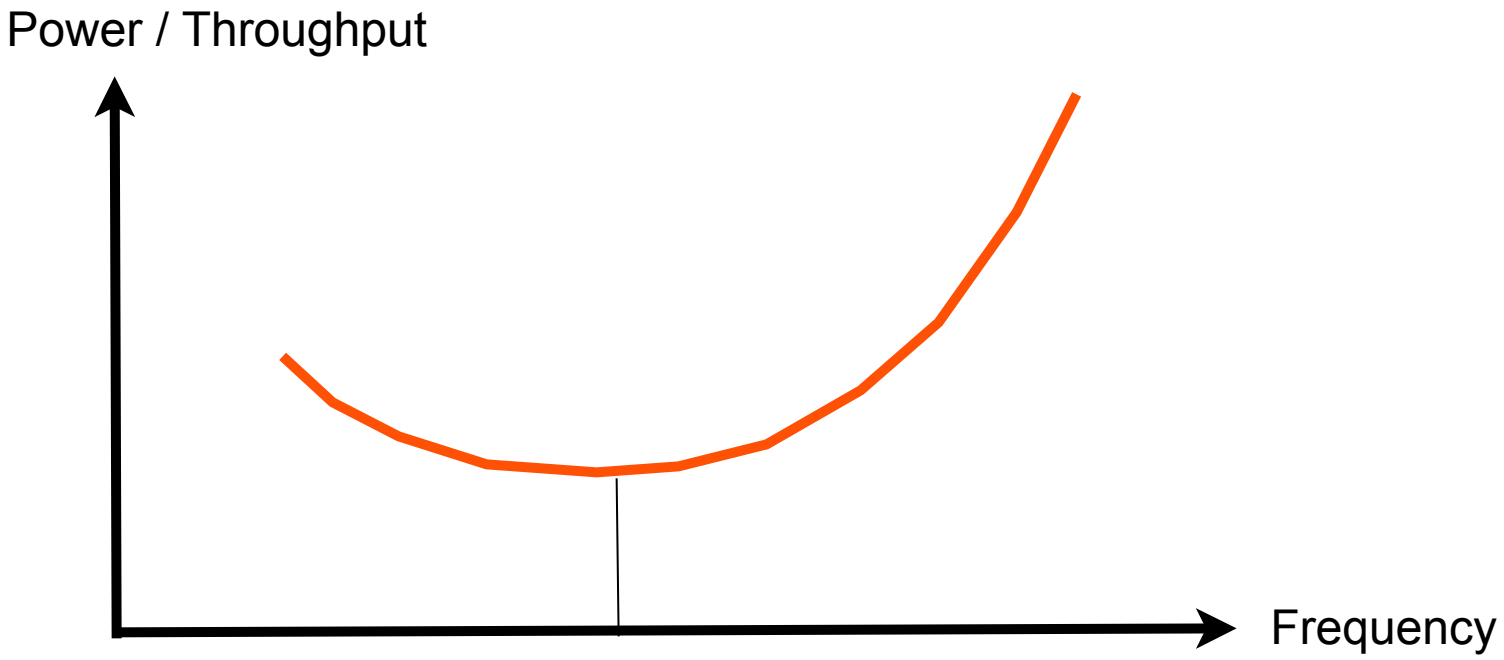


Source: *D. Frank, C. Tyberg,* IBM Research

Microchannel Fluidic Heatsinks



Power Efficiency (Power per Throughput)



Power = Clock * Capacitance * Vdd^2

Higher-frequency designs consume much more power

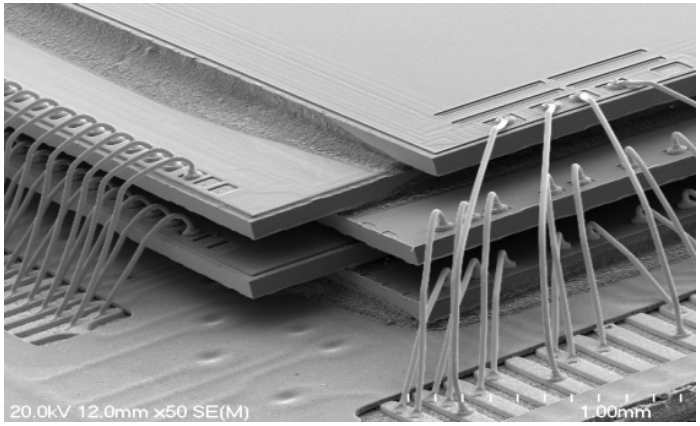
CPU Power Strategy

- With fixed power density, clock rates look flat
  - Increasing power densities is very challenging
  - Best solution appears to be liquid cooling at device level
- High clock rates are less power efficient
  - Higher frequency CPUs require higher voltage levels
  - Power increases quadratically with voltage
- To reduce power, simplify CPU architecture
  - Lower memory latency simplifies pipelines
  - New Memory Interfaces and integrated I/O subsystems
- Most savings are from better packaging

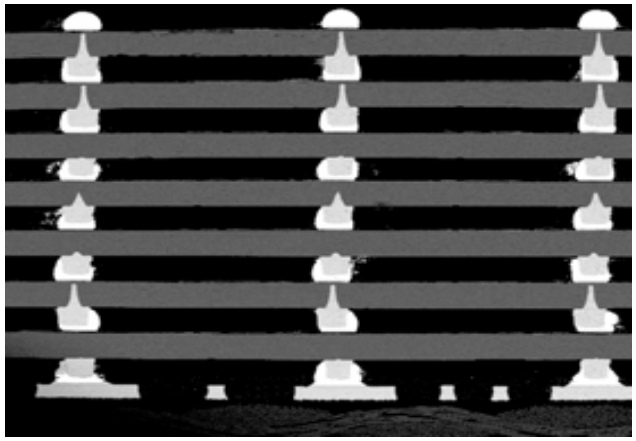
The Memory Bandwidth Challenge

- Memory bandwidth must grow with throughput
- 2022 CPU needs > 100X the memory bandwidth
- Traditional Package I/O pins are basically fixed
- Electrical signaling hitting speed limits
- How to scale memory bandwidth?
- Solution: Multi-Chip 3D Packaging

Multi-Chip 3D Packaging



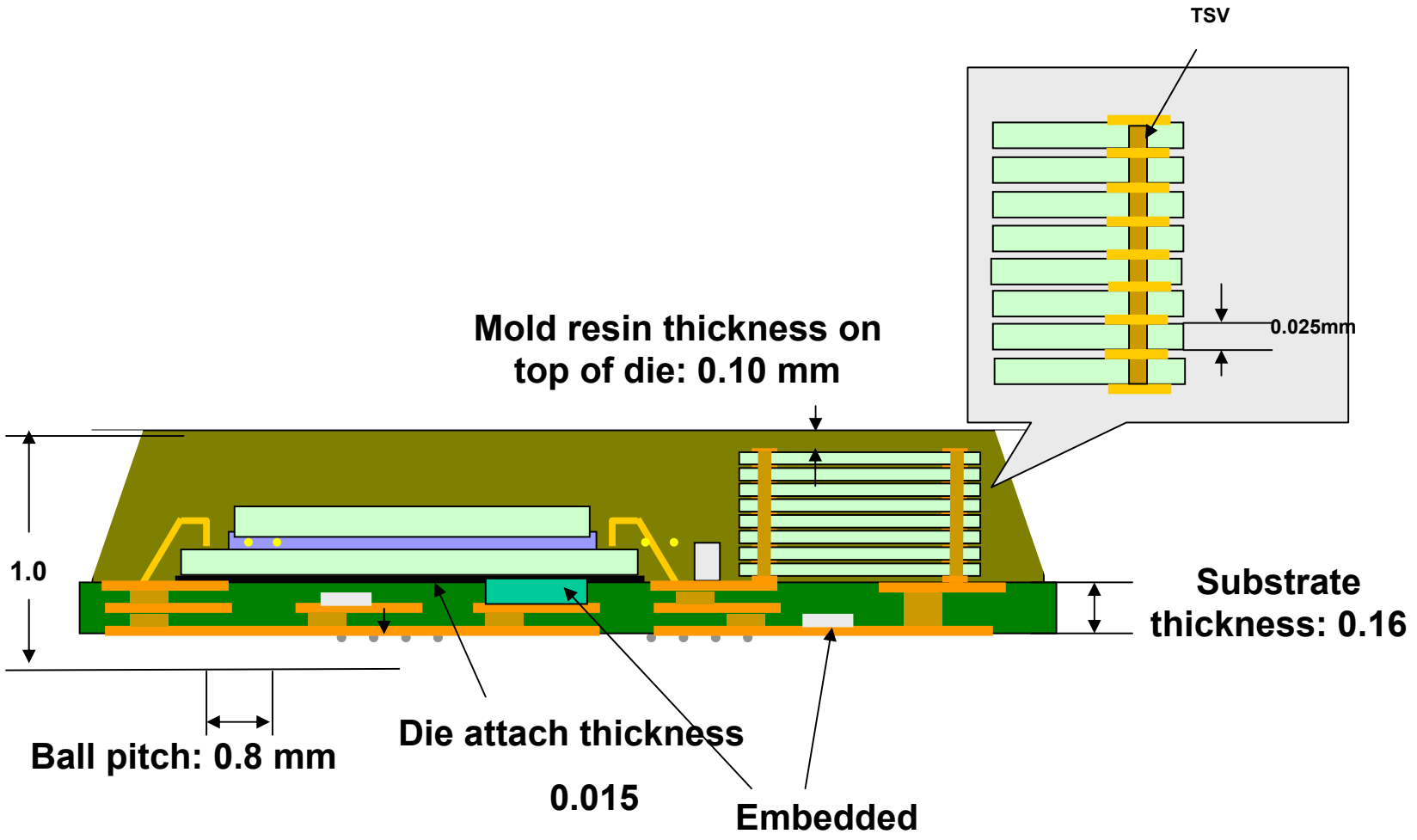
Wire bonded stacked die



Thru-Si via Stacking

Need to combine CPU + Memory on one Module

High-density 3D Multi-Chip Module (MCM)



Benefits of MCM Packaging

- Enables much higher memory bandwidth
- More channels, wider interfaces, faster I/O
- Greatly reduces memory I/O power
- Memory signals are local to MCM
- Reduces system size and power

Challenges with MCM Packaging

- Total Memory Size is limited to ~ 64 devices
- With 64 GB/device, that is 4 TBytes
- Assuming 1K cores, that is 4 GByte per core
- Consistent with today's systems but no better
- Applications must fit this profile

MCM Enables Fabric I/O Integration

- 2010: 1*4X QDR (32 Gbps / direction)
- 2020: 6*12X XDR (1.72 Tbps / direction)
- Mesh or Higher Radix Fabric Topologies
- 12X Copper for Module-Module Traces
- 12X Optical for Board-Board, Rack-Rack
- Support for global memory addressing

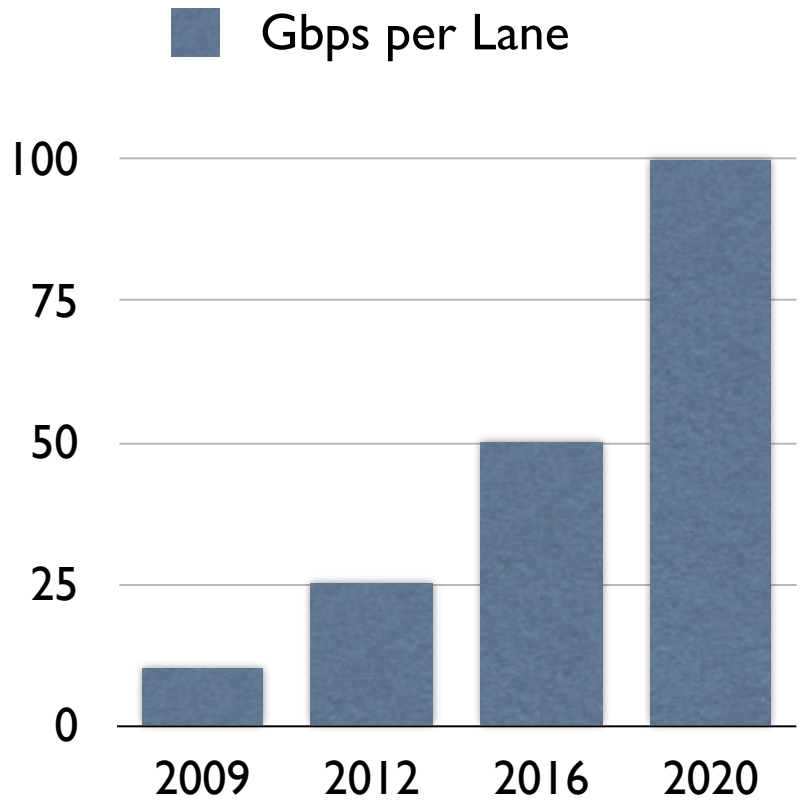
Benefits of integrating Router with CPU

- Best way to get highest message rate
- Match Injection and Link Bandwidth
- No congestion on receive
- Avoids intermediate bus conversions
- Eliminates half of the I/O pins and power
- Lowest cost and lowest power design
- Separate router chips are I/O Bound

What is the Best Fabric for Exascale?

- Optimal solution depends on economics
  - Cost of NIC, Router, Optical Interconnect
- Combination of mesh and tree look promising
  - Good global and local bandwidth
- Higher radix meshes significantly reduce hop-count
  - Pure 3D Torus for Exascale system is too large
- Robust Dynamic Routing desirable
  - Needed for load balancing and to recover from hardware failures

Expected Link Data Rate



10 Gbps shipping today

25 Gbps expected 2012

50 Gbps in 2016

100 Gbps in 2020

Higher speeds will require integrated optics interface

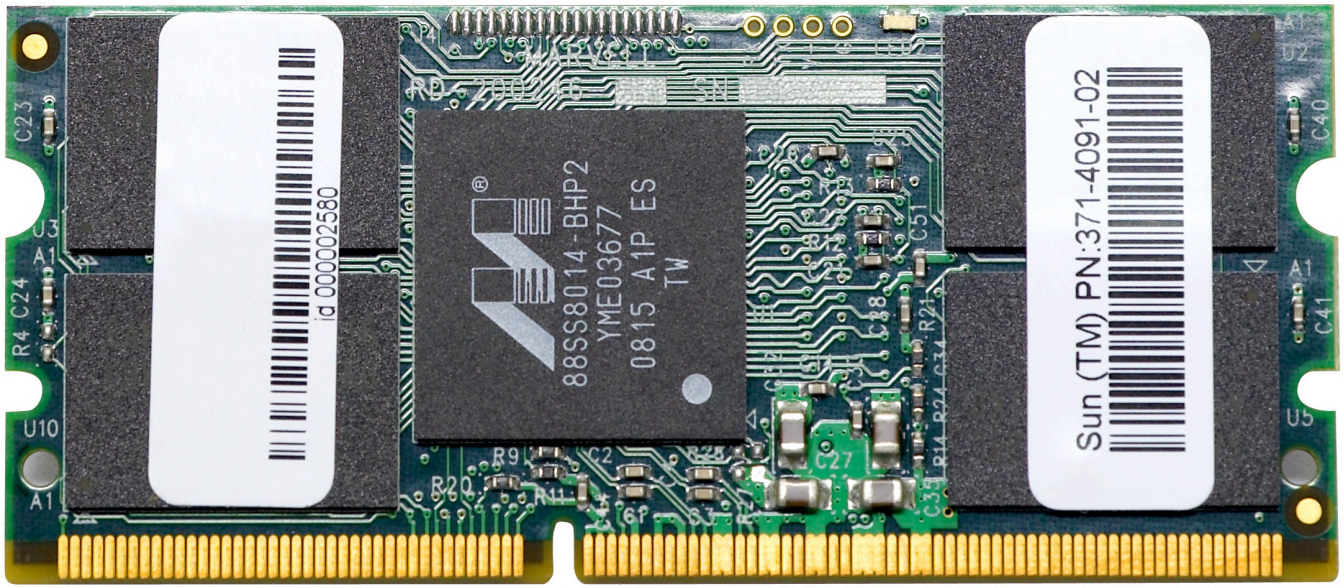
ExaScale Storage

- **Forget Hard Disks**
  - Disks are not going any faster
  - Useful as a tape replacement
  - At 100 MB/sec per disk, 1 TB/sec would require 10,000 disks
- **Solid State Storage**
  - Arriving just in time
  - Rapid Performance Improvements
  - Rapid Cost-reduction expected

Today's SSD vs HDD

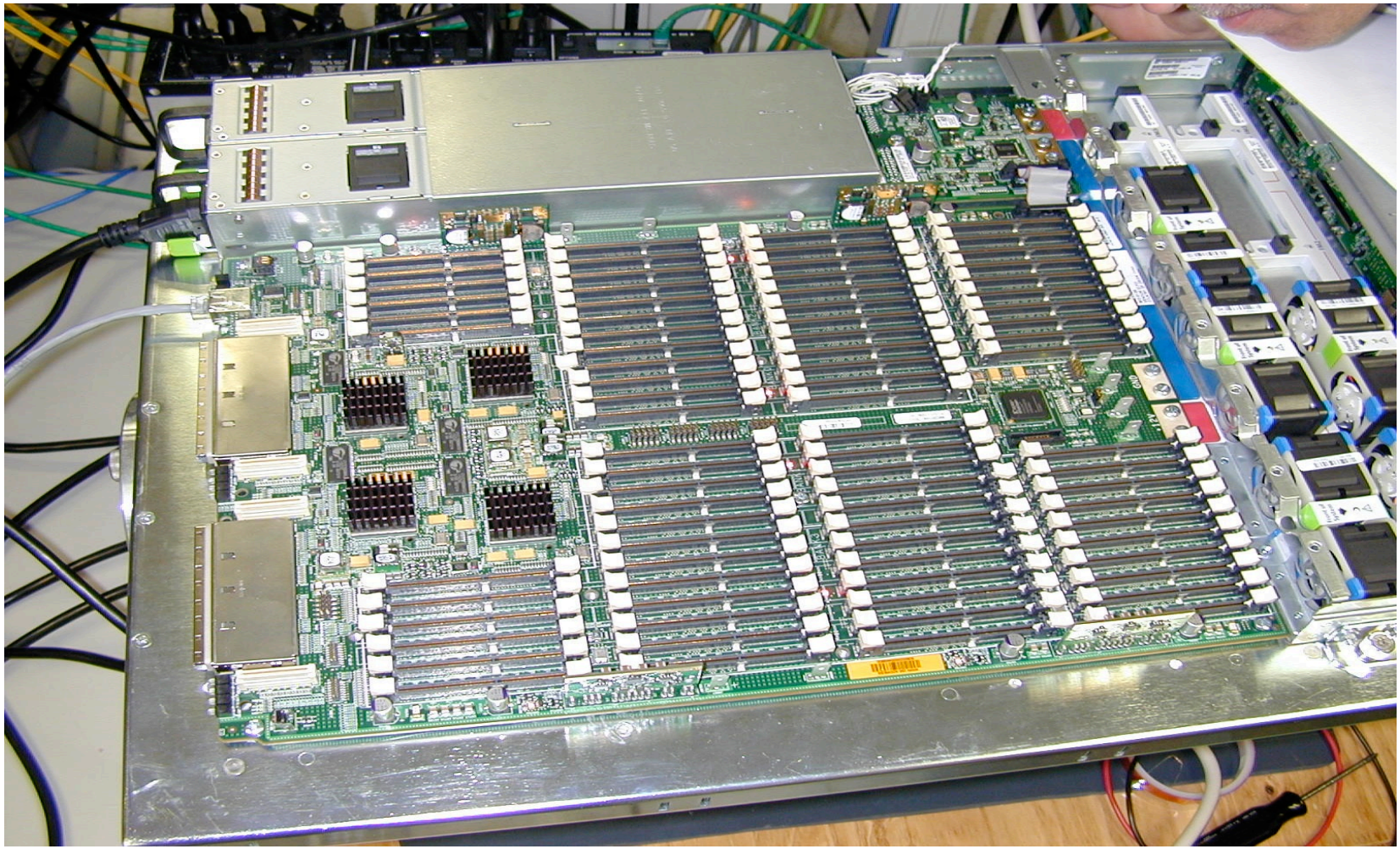
- **Conventional 2.5" HDD**
  - 15K RPM, 146 GB
  - 180 Write IOPS
  - 320 Read IOPS
  - $1 per IOPS
- **Solid State 2.5" SSD**
  - 0 RPM, 64 GB
  - 8K Write IOPS
  - 35K Read IOPS
  - $0.10 per IOPS

Sun Flash DIMMs

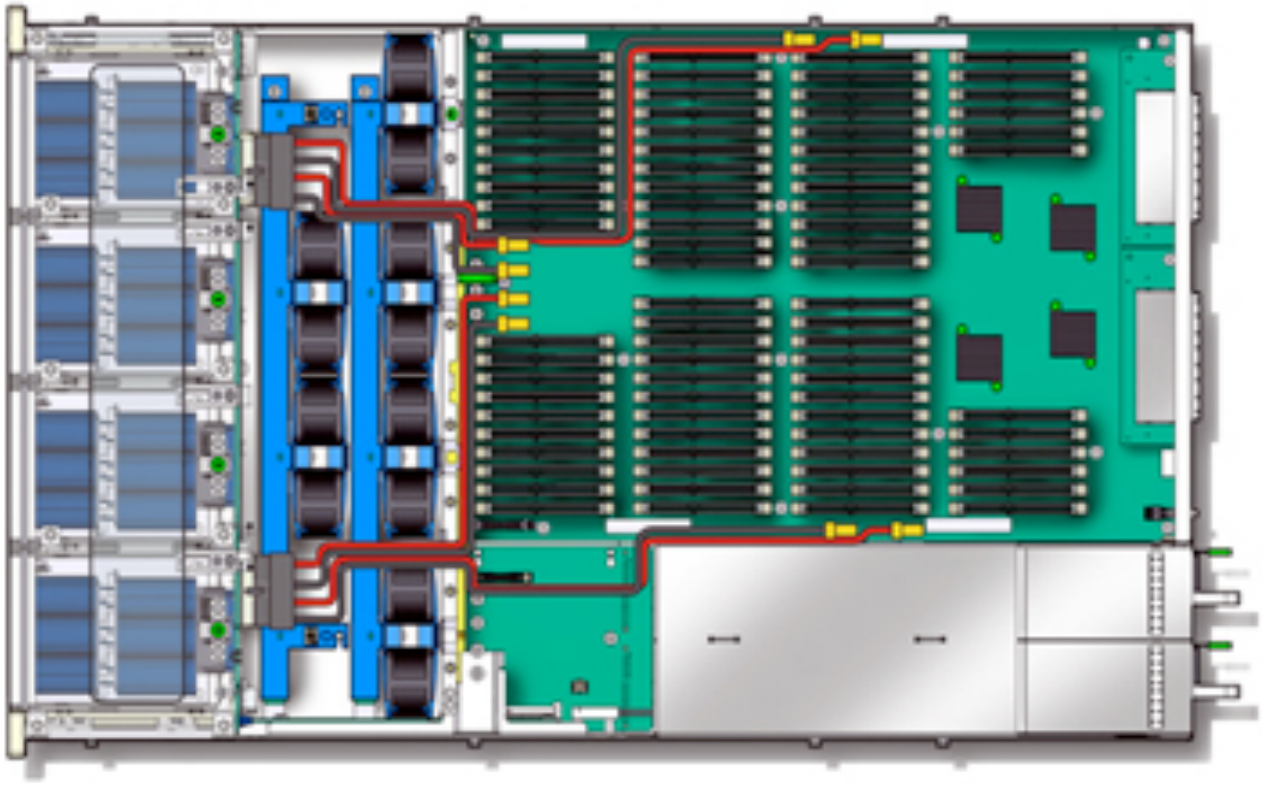


30,000 Read IOPS, 10,000 Write IOPS
Single-Level Flash SATA Interface

Sun F5100 Flash Storage Array

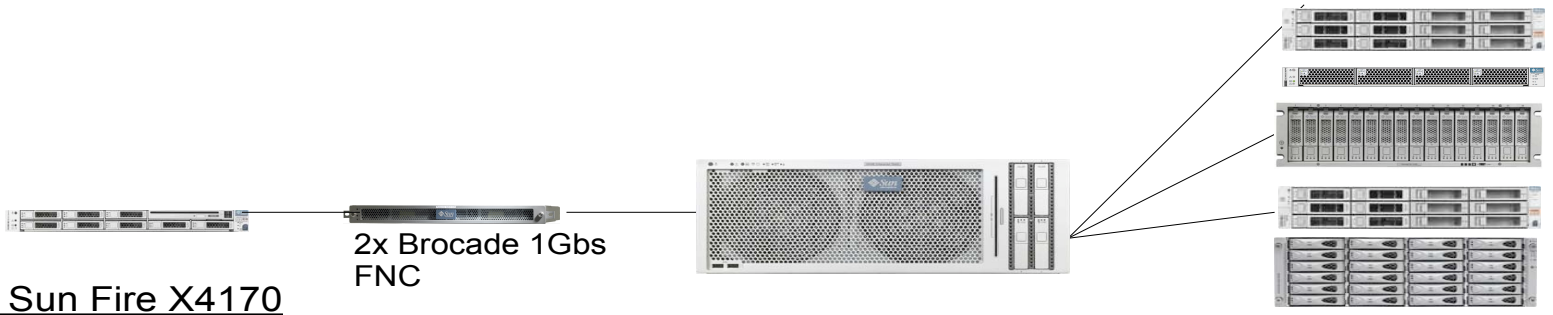


Sun F5100 Flash Array



1U Chassis with 80 Flash DIMM
64 SAS 2.0 Channels
> 1M IOPS

Oracle/Sun TPC-C World Record: 7,717,510 tpm-C with 4800 Flash DIMMs



2x Brocade 1Gbs FNC

24 Sun Fire X4170
2 2.53GHz Intel Xeon E5540 QC
48GB Memory
300GB SAS disk

12 Sun SPARC Enterprise T5440
4 1.6GHz UltraSPARC T2 Plus
512GB Memory
2 300GB 10K rpm SAS disks
3 8Gbs FC dual port
2 4Gbs FC dual port
4RU High

60 X4275 COMSTAR
5 1TB SATA disks
Sun F5100 Flash Array
1.92TB

X4275 COMSTAR
6 J4400 Arrays w/
24 1TB SATA disks ea.

24 ST6140
16 300GB 15K rpm disks
4GB write cache

Source: www.tpc.org

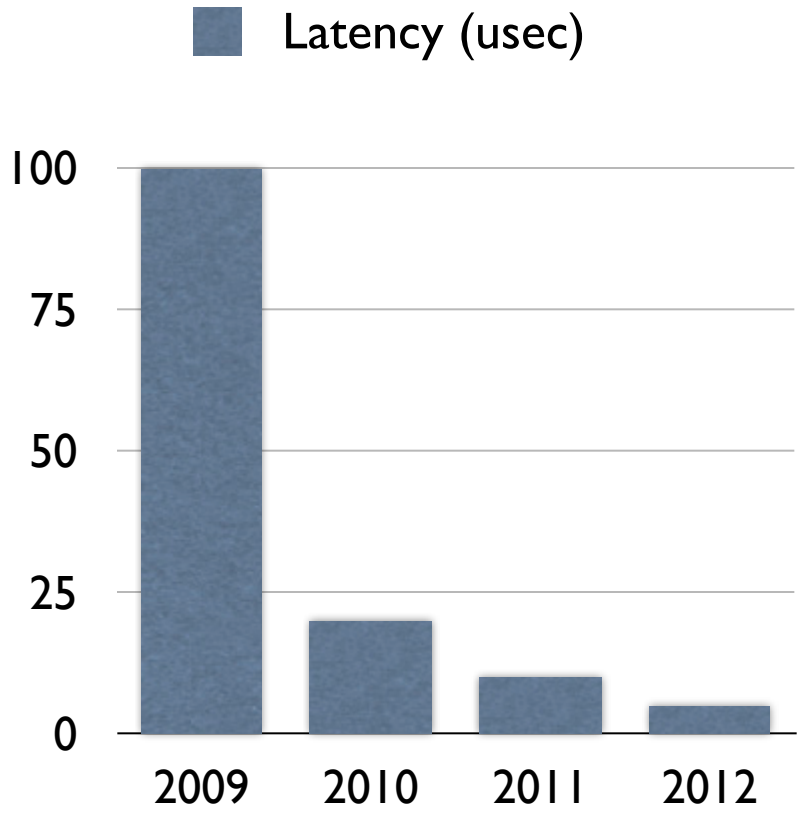
Flash Experience So Far

- **Not that easy to get to Millions of IOPS**
  - Limitation is the I/O controllers, not the FLASH
  - Needs lots of controllers and Flash channels
- **Writes are a problem**
  - Wear leveling algorithms far from perfect
  - Write performance degrades over time
- **SAS/SATA interface is not optimal**
  - Significant command processing times
  - SATA/SAS HBAs not designed for high I/O rates
- **Direct PCI-Express interface looks more promising**
  - Can support many more Flash channels per controller
  - Lower latency and more throughput

Flash in the Memory Hierarchy

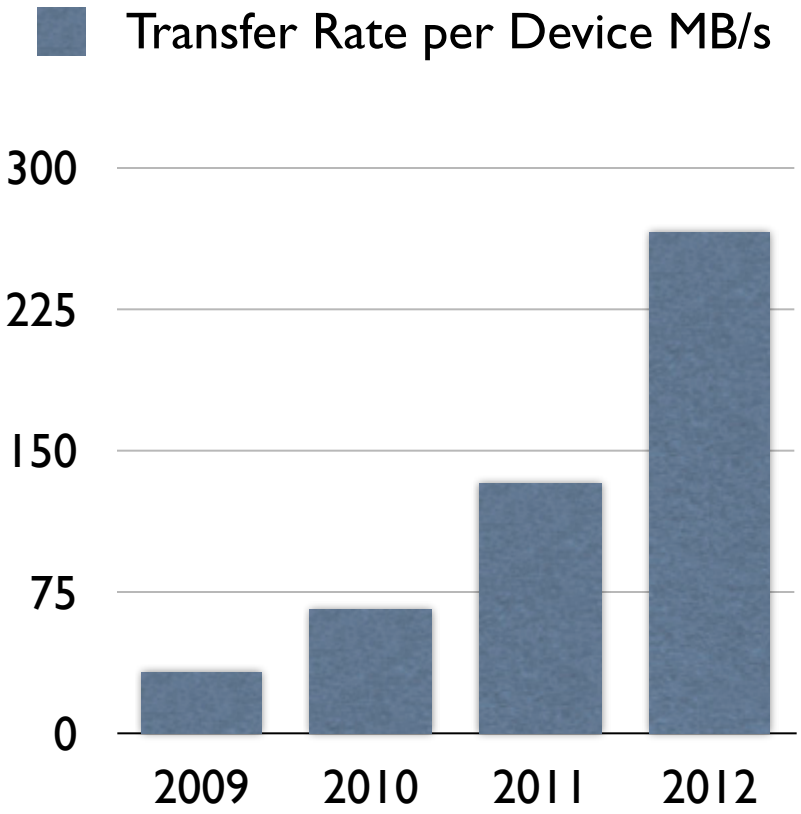
- **Flash is not random-access memory**
  - Block access oriented, not random access
  - Almost 1000X longer read latency than DRAM
- **Flash can be used as stable storage**
  - Writes are committed writes
  - Supercap magic behind the scenes
- **Tremendous Throughput and Size**
  - Terabytes of capacity cost-effective short term
  - Gigabytes/sec throughput
- **Today's Limitation is the Controller**
  - SAS/SATA has high overhead
  - Direct PCIe looks more promising

Flash Access Times Roadmap



Flash latency projected to be cut 50% per year for the next several years

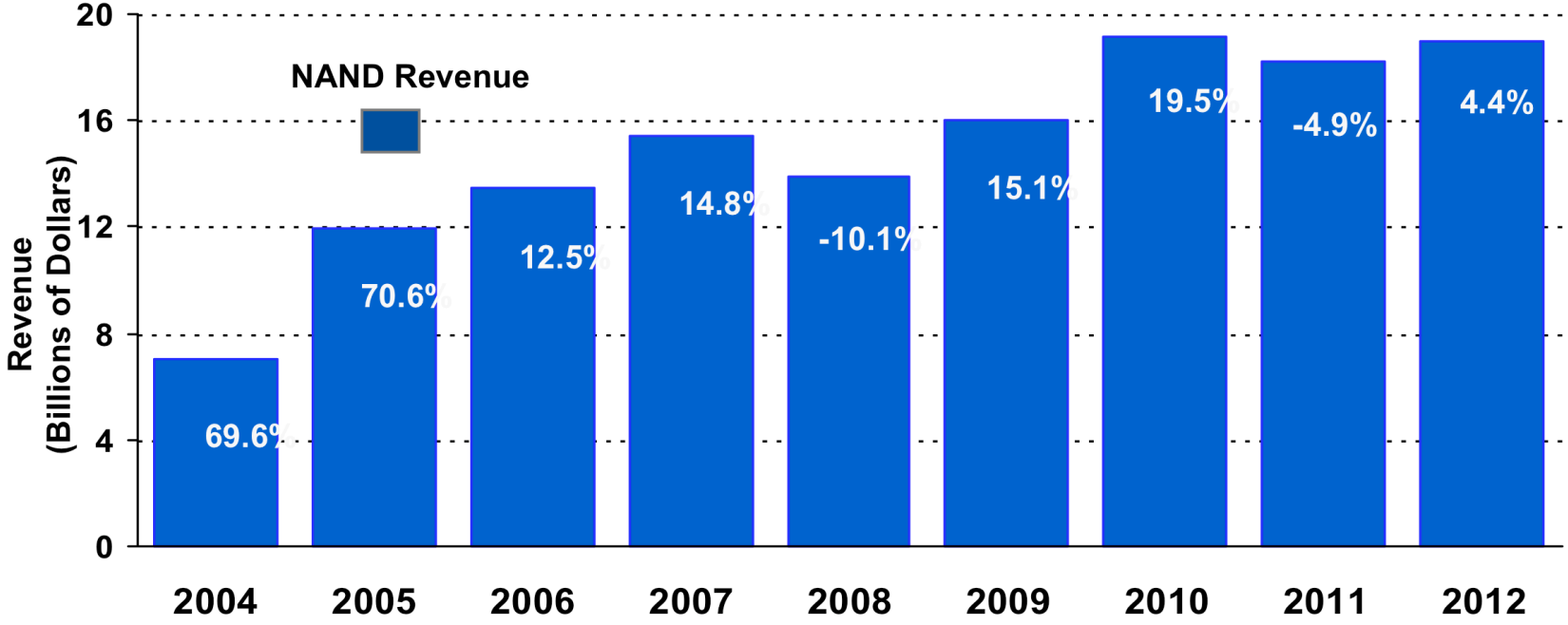
Flash Throughput Roadmap



Flash transfer rate will double each year for the next several years

Gartner Flash Forecast (August 2008)

	2004	2005	2006	2007	2008	2009	2010	2011	2012	CAGR 07-12
Gigabytes (M)	64.1	229.3	703.1	1,945	4,684	10,415	22,074	42,022	75,101	107.7%
Bit Growth	222%	258%	207%	177%	141%	122%	112%	90%	79%	



	2004	2005	2006	2007	2008	2009	2010	2011	2012	CAGR 07-12
ASP 1GB eqv.	109.4	52.1	19.1	7.94	2.97	1.54	0.87	0.43	0.25	-49.8%
ASP Change	-47.4%	-52.3%	-63.3%	-58.5%	-62.7%	-48.2%	-43.6%	-50.0%	-41.6%	

Source: Gartner, August 2008

Flash Summary

- Density doubling each year
- Costs falling by 50% per year
- Access times falling by 50% per Year
- Throughput doubling every year
- Controllers improving rapidly
- Interface moving from SATA to PCI Express
- Multi-GB/sec per PCI Controller
- Very large-scale I/O looks feasible

Technology Summary

- **Moore's Law will continue for at least 10 Years**
  - Transistors per area will double ~ every 2 year
  - 128X increase in density by 2022
- **Frequency Gains are more difficult**
  - Power increases super-linear with clock rate
  - Must exploit parallelism with more cores
- **Need to increase memory and I/O bandwidth**
  - Need to scale with throughput
  - Need a factor of 128X by 2020
- **Most promising technology is memory stacks and Flash**
  - Supports lots of channels to scale bandwidth
  - Very high bandwidth and transaction rates appears feasible

The Software Challenge

- **The limits of application parallelism**
  - Instruction set parallelism
  - Number of cores per CPU Module
  - Number of CPU modules per system
- **Need to exploit parallelism at all levels**
  - Quality of compiler code generation
  - Functional parallelism within node
  - Data parallelism across nodes
- **Ultimate question is application parallelism**
  - Will require re-architecting of applications
  - Not all applications will scale to Exascale