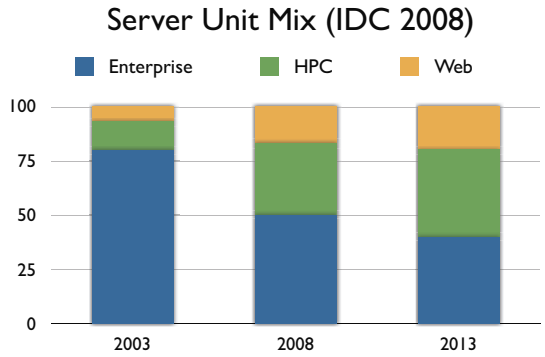
The Road from Peta to ExaFlop

Andreas Bechtolsheim

June 23, 2009

HPC Driving the Computer Business



HPC grew from 13% of units in 2003 to 33% in 2008
HPC projected to be the highest % of servers by 2013

Changes in the Computer Business

- Virtualization Consolidates Enterprise Datacenters
  - Reduces the number of servers required in Enterprise
- Cloud Computing Outsources Applications
  - Further reduces size of the traditional Enterprise server market
- High Performance Computing Gaining Share
  - Open ended demand for more performance, does not virtualize
- Moore's Law Benefits HPC Market
  - Go Faster, Faster

High-speed Fabrics Everywhere

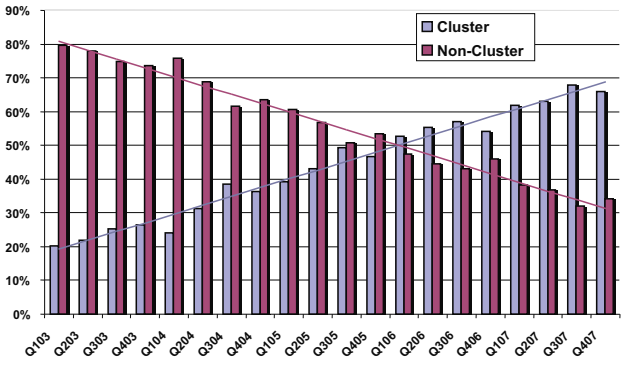
10 GigE and Infiniband shipping in volume

HPC, Database and Storage Clusters

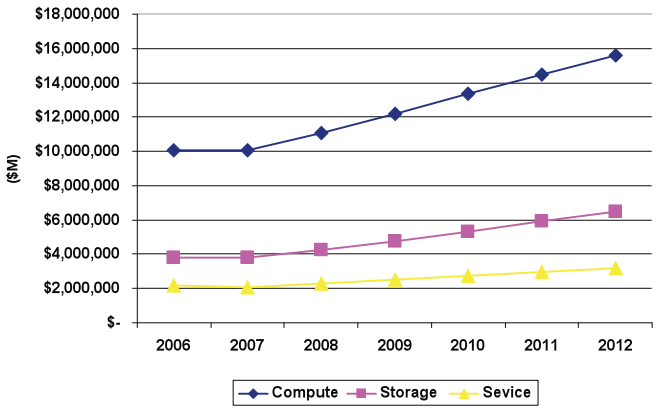
Outstanding scaling for wide range of applications

Predictable roadmaps to 100 Gbps and beyond

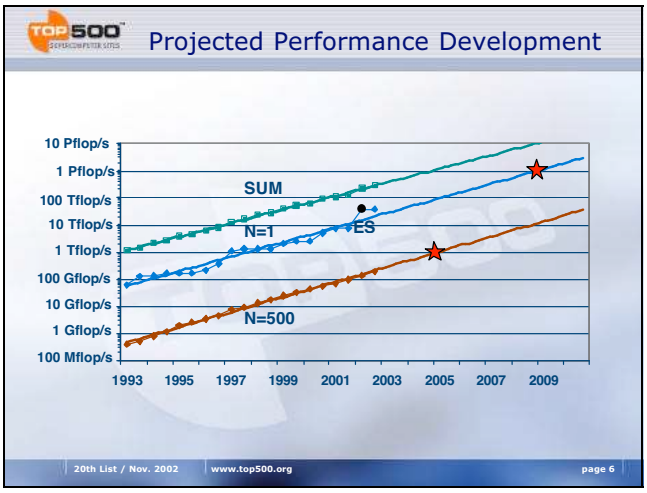
Clusters are Driving the HPC Market



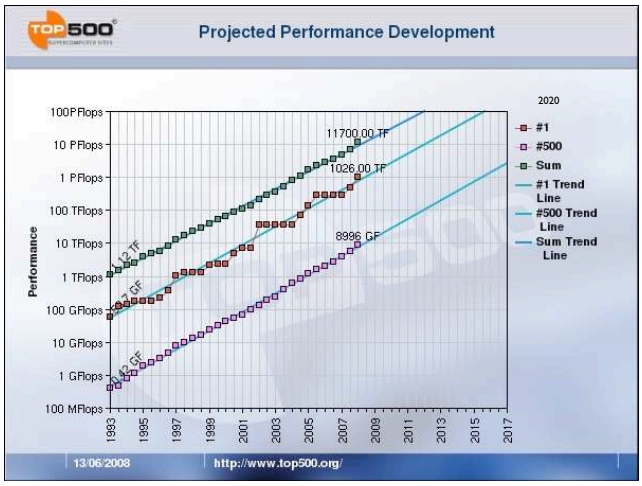
HPC Market Forecast (IDC 2008)



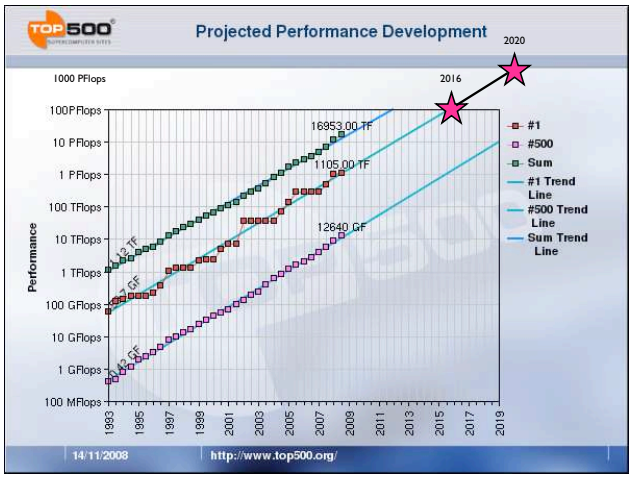
Top500 List November 2002



Top 500 June 2008: First PF Result



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

Challenges

- Semiconductor Roadmap
- Packaging Technology
- Power and Cooling
- Local Interconnect
- External Interconnect
- Storage System
- Software Scalability
- Exploiting Parallelism

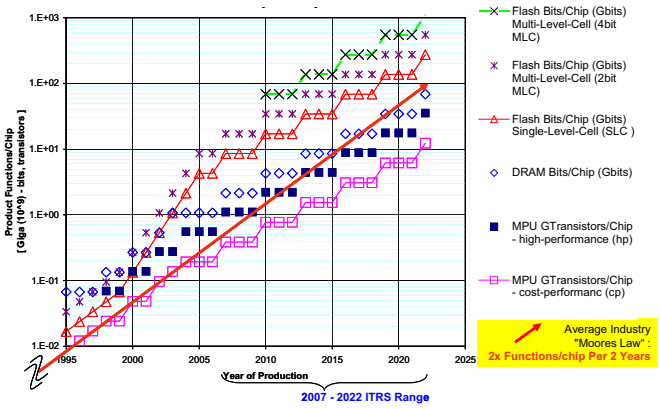
The Basic Math: "More than Moore"

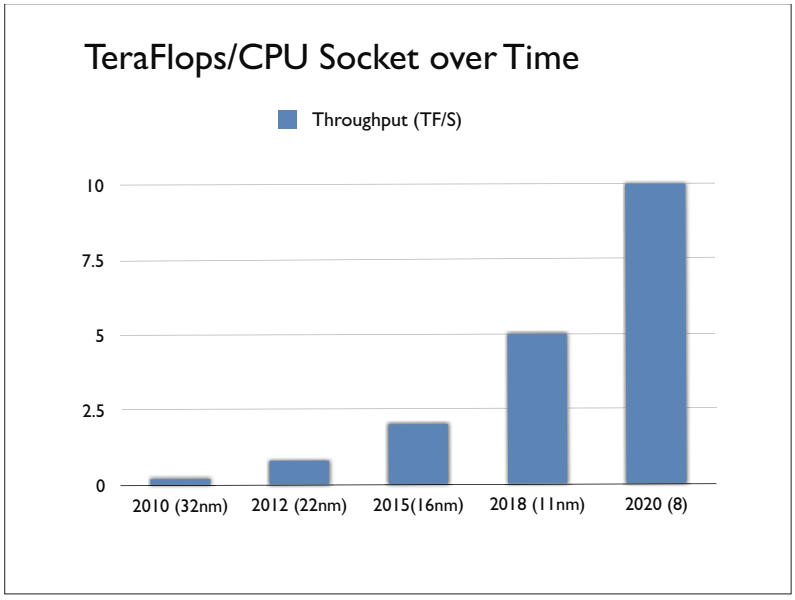
Aggregate Performance = N * C * F * I

N	Number of Modules	20% / Y	Budget, Power
C	Cores per Module	40% / Y	Technology, Power
F	Frequency	5% / Y	Technology, Power
I	Instruction Efficiency	15% / Y	Architecture, Power
	TOTAL	100% / Y	

Must increase system size, cores per module, and Instruction Efficiency to double every year

Semiconductor Technology Roadmap




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

CPU Module [Socket] (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	500W	2X
Power Efficiency	0.6 GF/W	20 GF/W	32X

Challenging but doable

System Performance (2010 => 2020)

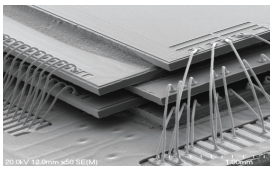
Year	2010	2020	Ratio
FLOPS/Module	160 GF	10 TF	64X
Modules/System	16,000	100,000	6X
FLOPS/System	2.5 PF	1 EF	320X
Cores/Module	16	160	10X
Cores/System	256,000	16M	64X
Memory/Module	30 GB/s	2 TB/s	64X
Memory/System	0.2 B/F	0.2 B/F	=
IO Bandwidth	3 GB/s	192 GB/s	64X
IO Bandwidth/F	0.02 B/F	0.02 B/F	=
Power / Module	250W	500W	2X
Power / System	4 MW	50 MW	12X

Challenging but doable

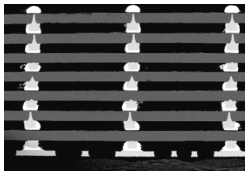
The Biggest Challenge: Memory Bandwidth

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

Multi-Chip 3D Packaging



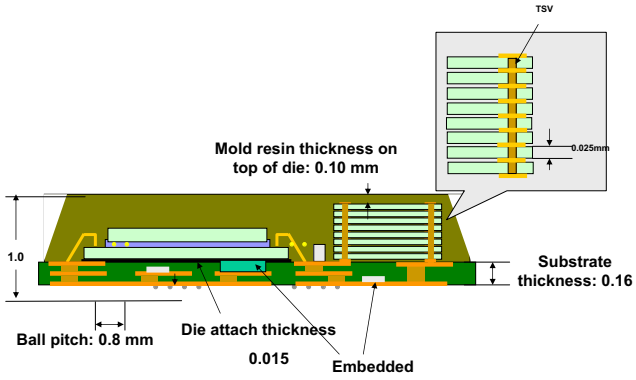
Wire bonded stacked die



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

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
- Very high message rates (Several Billion/sec)
- Support for global memory addressing

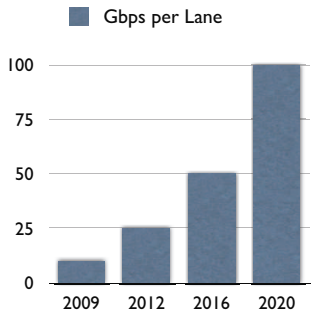
Benefits of integrating Router with CPU

- 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

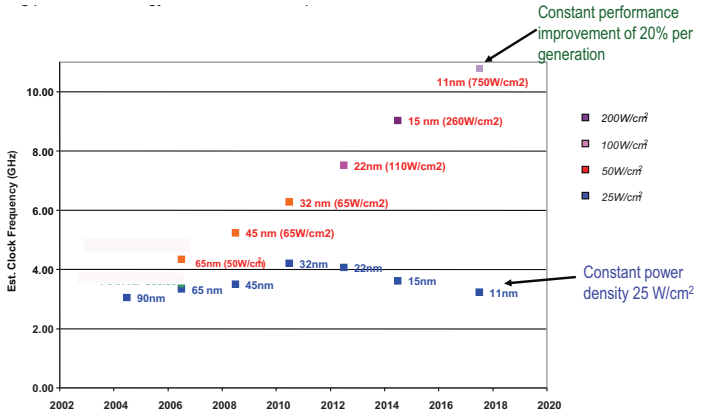
50 Gbps in 2016

Higher speeds will require integrated optics interface

Next Challenge: Power

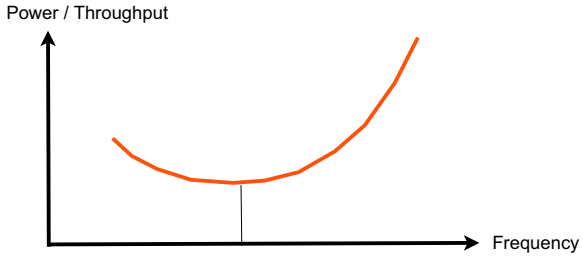
- Power Efficiency is critical
- Design for greatest power efficiency
- Clock frequency versus power
- Minimize interface and I/O power
- Optimize CapEx and OpEx

Power per Core



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

Power Efficiency (Power per Throughput)



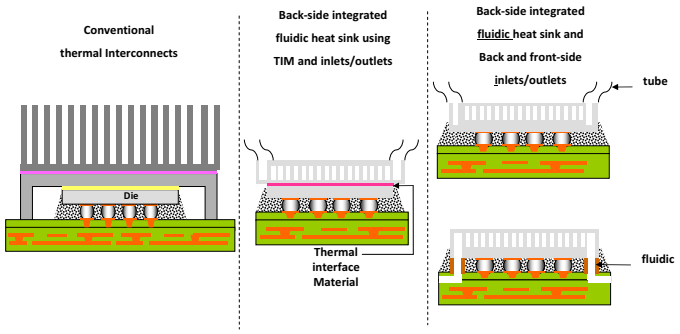
Power = Clock * Capacitance * Vdd^2

High-frequency designs consume much more power

Power Efficiency Strategy

- Reduce I/O power as much as possible
  - Requires MCM packaging, lower voltage interface levels
  - Saves more than 50% compared to power today
- Minimize Leakage Current
  - Lower temperature/liquid cooling helps
  - Optimize transistor designs and materials
- Simplify CPU Architecture
  - Lower memory latency simplifies pipelines
  - Integrate NIC and I/O subsystems
- Most savings from better packaging

Microchannel Fluidic Heatsinks



Liquid Cooling is Essential

- **Reduces Power**
  - Reduces power required to drive I/O
  - Reduces leakage currents
  - Avoids wasting power on moving air
- **Increases Rack Density**
  - Reduces Number of Racks
  - Reduces Weight and Structural Costs
  - Reduces Cabling
- **Improves Heat Removal per Socket**
  - Liquid Cooling required to increase system density
  - Physics of air cooling are not changing
  - Liquid Cooling with Microchannels look promising

Packaging Technology Summary

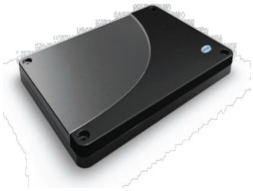
- **Main new development is Multi-Chip Modules**
  - Many more signals available on-module than off-module
  - Increases memory bandwidth while reducing power
  - Decouples memory bandwidth from I/O Pins
- **Fabric Interface**
  - Integrated NIC reduces power and improves performance
  - Integrated Router supports Mesh and Fat-Tree topologies
  - SERDES support copper and optical (fiber) interfaces
- **Power and Cooling**
  - 480VAC to each Rack
  - Liquid Cooling to reduce power and increase density
  - Dramatic Increase in Throughput (and power) per rack

ExaScale Storage

- **Storage Bandwidth Requirements**
  - 0.1 GB/s per TF
  - 100 GB/sec per PF
  - 100 TB/sec per EF
- **Forget Hard Disks**
  - Disks are not going any faster
  - Useful as a tape replacement
  - At 100 MB/sec per disk, 100 TB/sec would require 1M disks
- **Solid State Storage**
  - Arriving just in time
  - Rapid Performance Improvements
  - Rapid Cost-reduction expected

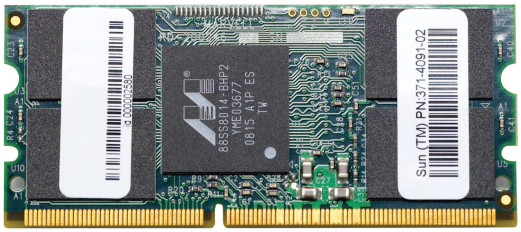
Today's SSD vs HDD

- **Conventional 2.5" HDD**
  - 15K RPM, 146 GB
  - 180 Write IOPS
  - 320 Read IOPS
  - $1 per IOPS



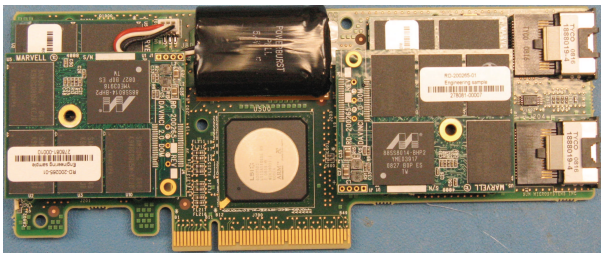
- **Solid State 2.5" SSD**
  - 0 RPM, 64 GB
  - 8K Write IOPS
  - 35K Read IOPS
  - $0.10 per IOPS

Sun Flash DIMMs



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

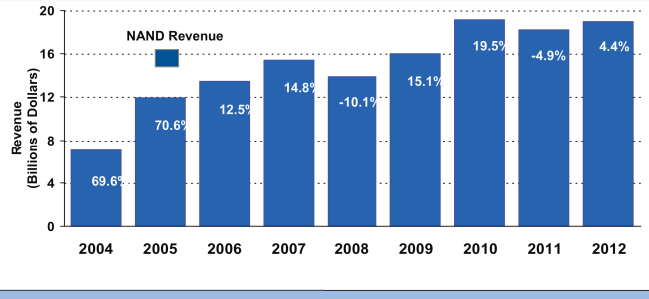
PCI Flash Storage



100,000+ IOPS, Up to 1 TB Capacity
Low Profile PCI Card Slot

Gartner Flash Forecast (August 2008)

										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%	



										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

Solid State Storage Summary

- Density Doubling Each Year
- Cost Falling by 50% Per Year
- Access times improving quickly
- Throughput improving quickly
- Phase-Change Technology looks promising
- Interface moving from SATA to PCI Express
- Multi-GB/sec per PCI Controller
- 100 TB/sec suddenly looks possible

Scaling to ExaScale: CPU Throughput

- **Three Dimensions of Scalability**
  - Frequency, Cores, FLOPS/Core
- **Increasing Frequency is most difficult**
  - Expect modest increases going forward
  - Problem is power consumption per core
- **Increasing the Number of Cores per Chip**
  - Proportional to Technology Improvements
  - Moore's Law predicts doubling every 2 years
- **Increasing FLOPS per Core**
  - Increase instruction set parallelism
  - More Functional Units and SIMD instructions

Scaling Throughput per Core

GF/Core	10	16	32	64
1 PF	100K	64K	32K	16K
10 PF	1M	640K	320K	160K
100 PF	10M	6.4M	3.2M	1.6M
1000 PF	100M	64M	32M	16M

Critical to increase throughput per core

Scaling Throughput per Power

GF/W	0.64	3	10	20
1 PF	1.5M	300K	100K	50K
10 PF	15M	3M	1M	500K
100 PF	150M	30M	10M	5M
1000 PF	1500M	300M	100M	50M

Critical to improve power efficiency to reduce OPEX

Scaling Throughput per CPU Module

GF/M	160	640	2500	10000
1 PF	6.4K	1.6K	400	100
10 PF	64K	16K	4K	1K
100 PF	640K	160K	40K	10K
1000 PF	6.4M	1.6M	400K	100K

Number of CPU Modules drives system size and cost

Technology Summary

- **Moore's Law will continue for at least 10 Years**
  - Transistors will double approximately every 2 year
  - Not enough to double performance every year
  - "More than Moore" required for 1 EF by 2020
- **Frequency Gains are very difficult**
  - Power increases super-linear with clock rate
  - Must exploit parallelism with more cores
- **Need to increase FLOPS/Core**
  - Predictable way to increase performance
  - Mul-add, multiple FPUs, SIMD extensions
- **Need to increase memory and I/O bandwidth**
  - Need to scale with throughput
  - Need a factor of 64X by 2020

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 (SMP Threads)
  - Data parallelism across nodes (MPI Tasks)
- **Ultimate question is application parallelism**
  - Will require re-architecting of applications
  - Not all applications will scale to Exaflop size
- **RAS Related Challenges**
  - MTBI declines with system size
  - Needs high-speed checkpoint restart

System Conclusions

- **Expect Throughput to Double every year**
  - Combination of Moore's law and efficiency gains
  - First 1 EF System likely by 2020
  - Smallest Top 500 System likely 10 PF in 2020
- **PetaFlop Systems will be very small and affordable**
  - One PetaFlop per rack by 2016
  - Personal PetaFlop computer by 2020
- **Greatly Broadens the HPC market**
  - HPC market growth will continue for a long time
  - Big opportunity for small, medium, and large systems
- **HPC is a major driver to advance server technology**
  - CPUs, Memory, Fabric, Storage, Software
  - Ripple-down effect from the largest to the smallest systems