# An Overview of High Performance Computing and Future Requirements

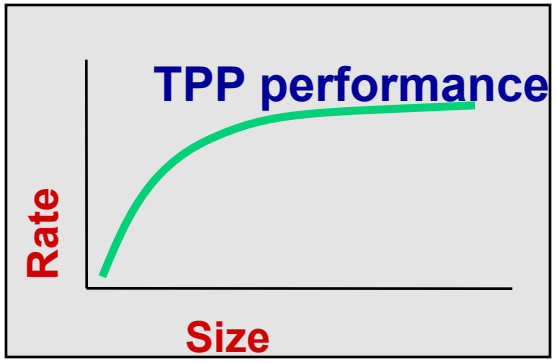

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory

**H. Meuer, H. Simon, E. Strohmaier, & JD**

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

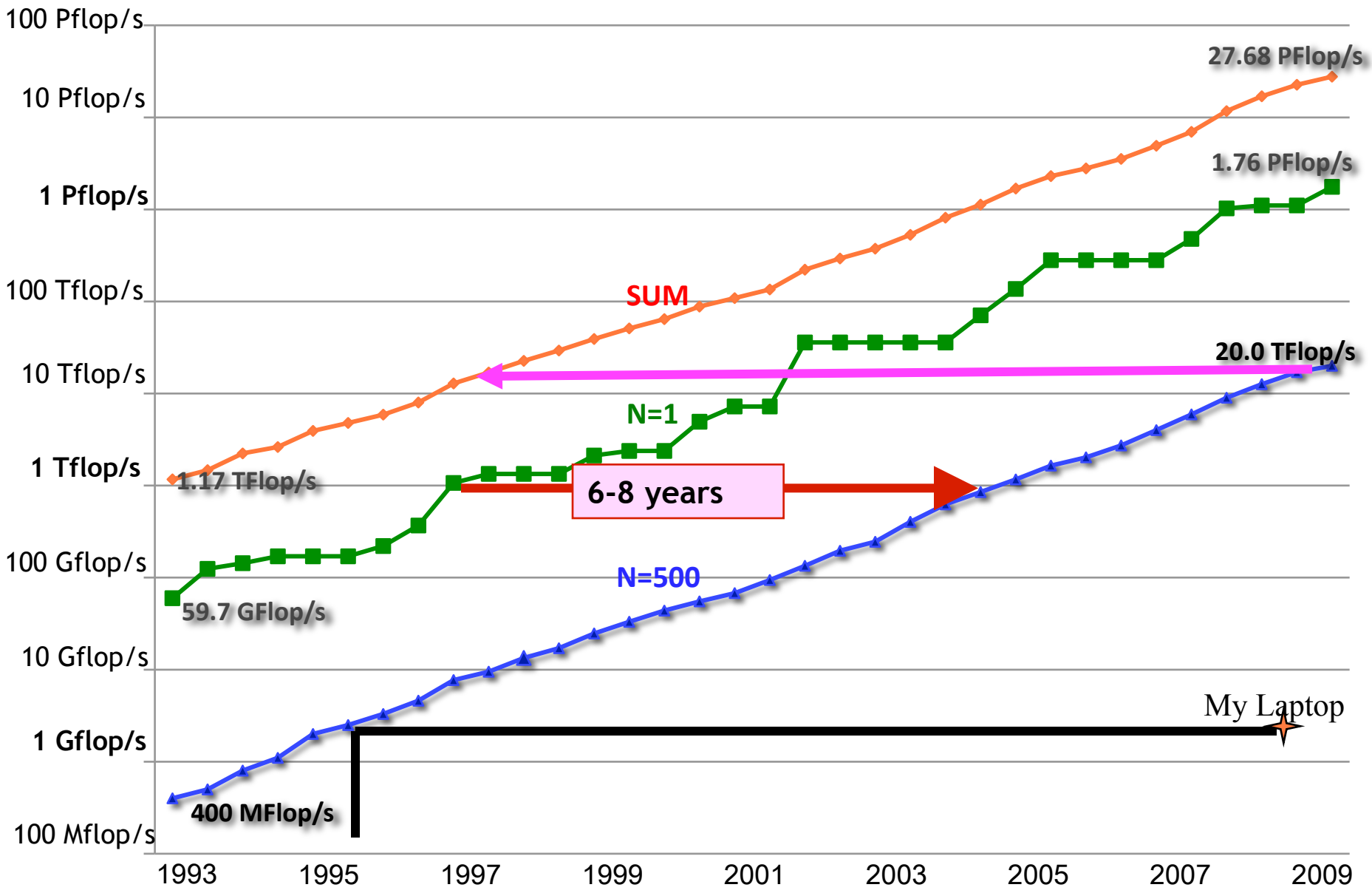
$Ax=b$, *dense problem*




- Updated twice a year
  SC‘xy in the States in November
  Meeting in Germany in June

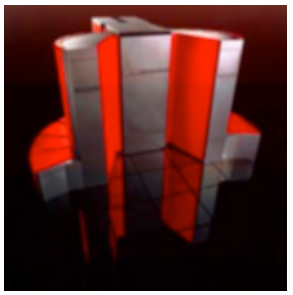
- All data available from **www.top500.org**

# Performance Development



# Looking at the Gordon Bell Prize

(Recognize outstanding achievement in high-performance computing applications and encourage development of parallel processing )

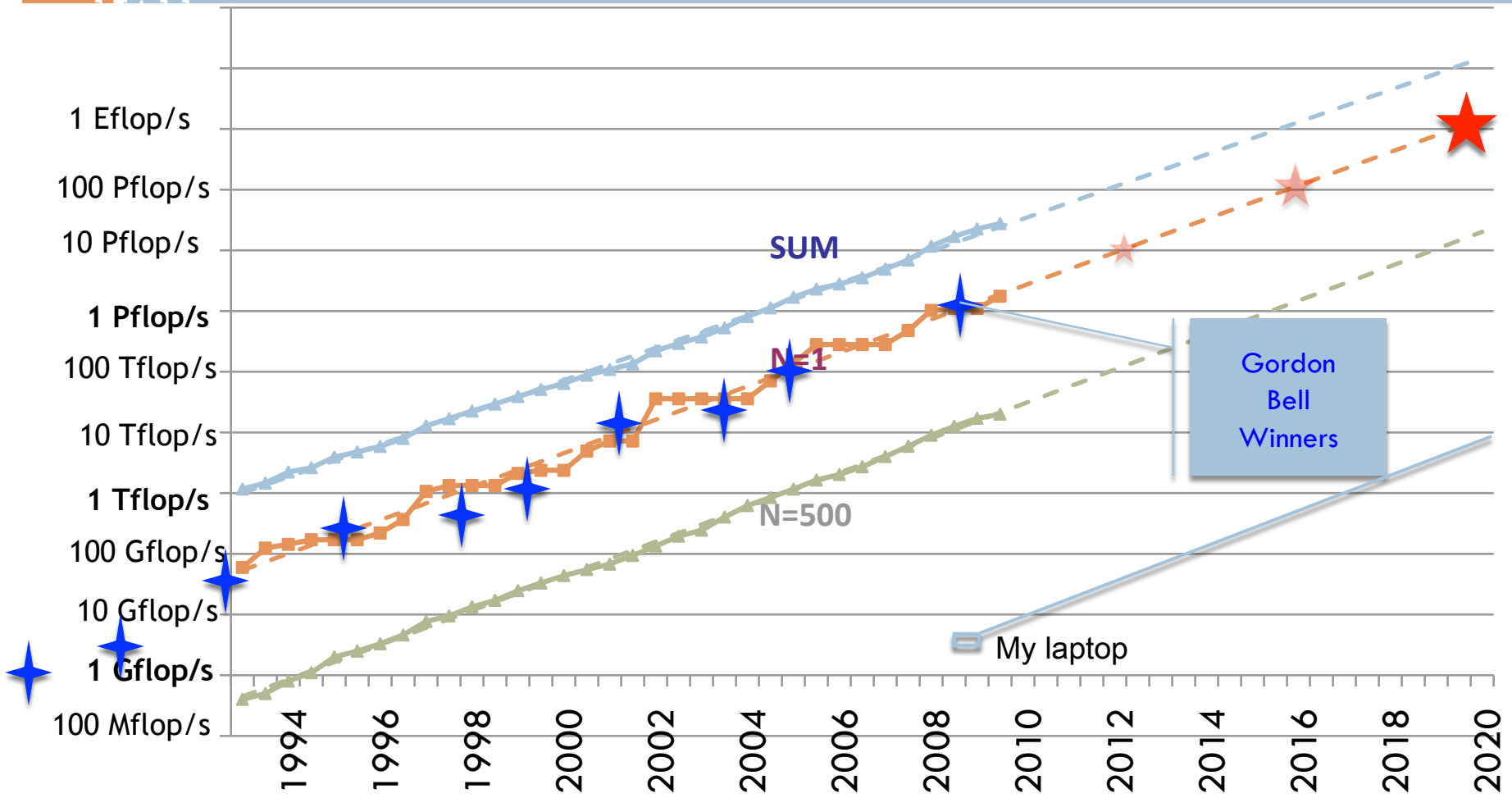


- 1 GFlop/s; 1988; Cray Y-MP; 8 Processors
  - Static finite element analysis

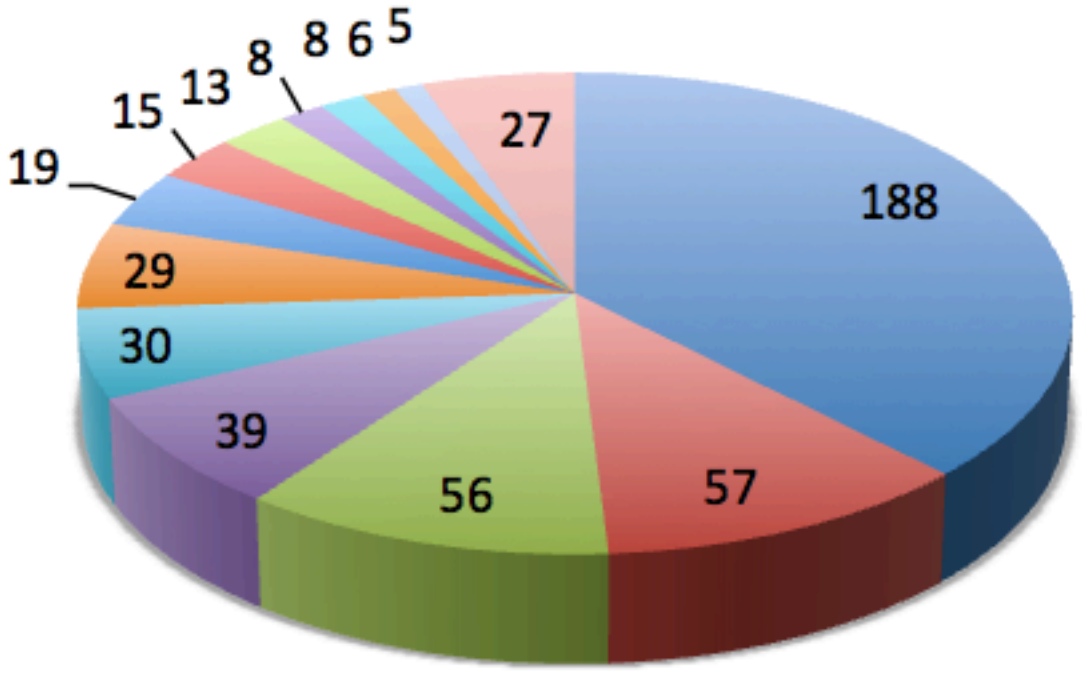
- 1 TFlop/s; 1998; Cray T3E; 1024 Processors
  - Modeling of metallic magnet atoms, using a variation of the locally self-consistent multiple scattering method.
- 1 PFlop/s; 2008; Cray XT5; $1.5 \times 10^5$ Processors
  - Superconductive materials

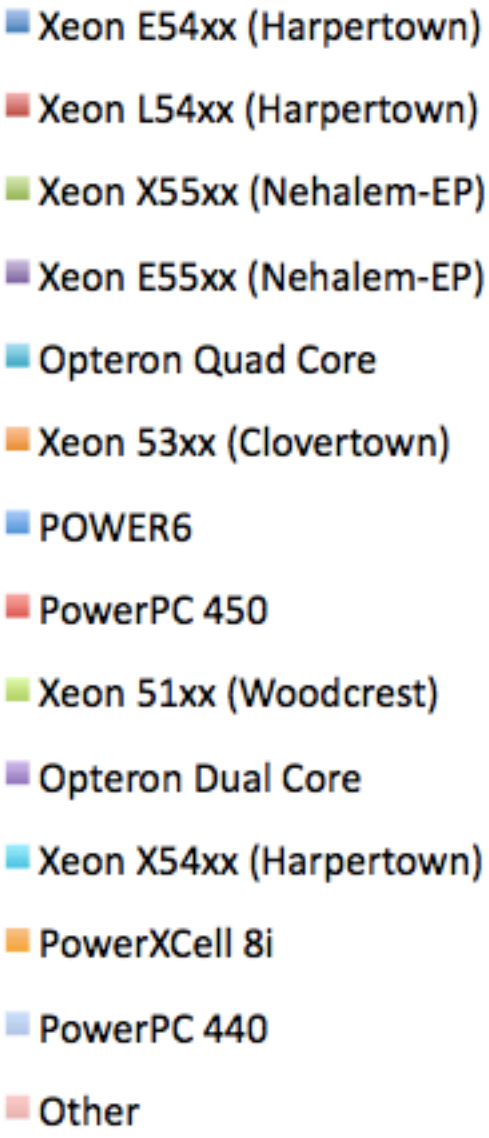

- 1 EFlop/s; ~2018; ?; $1 \times 10^7$ Processors ($10^9$ threads)

# Performance Development in Top500



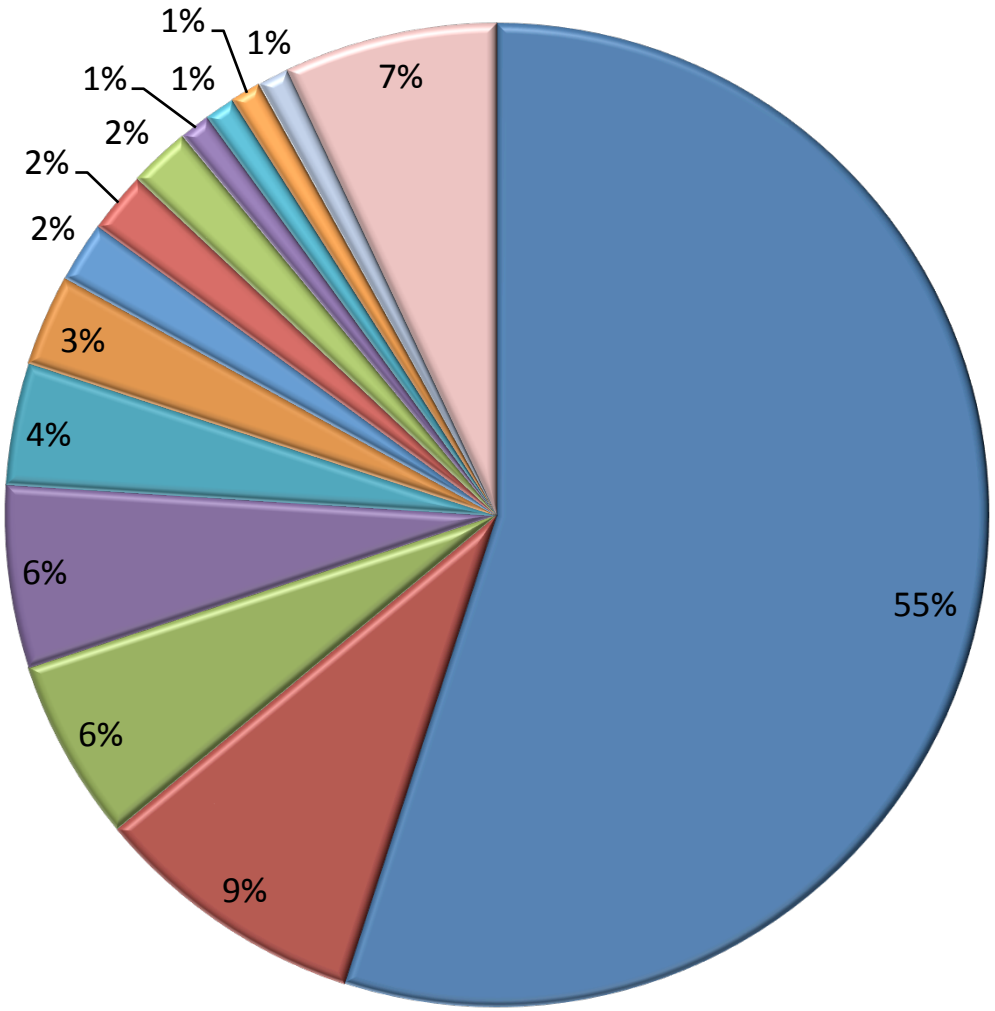
# Processors Used in the Top500 Systems



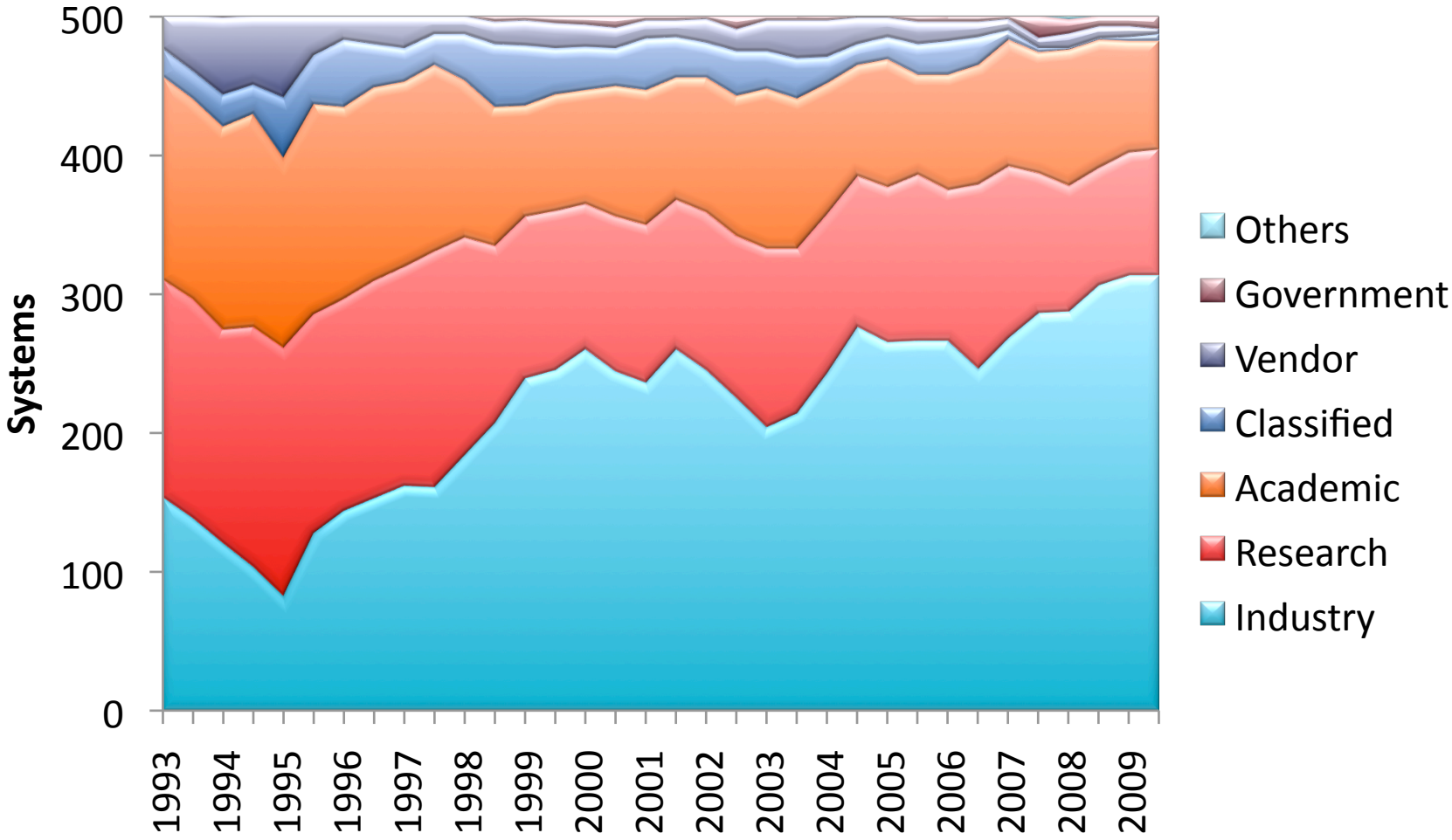

Intel 80%
IBM 10%
AMD 8.4%

# Countries / System Share




55% United States
9% United Kingdom
6% France
6% Germany
4% China
3% Japan
2% Canada
2% Austria
2% New Zealand Most flops/pop
1% Sweden
1% Russia
1% Italy
1% Spain
7% Others

# Customer Segments



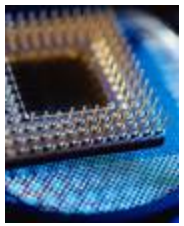
# Industrial Use of Supercomputers

- **Of the 500 Fastest Supercomputer**
  - **Worldwide, Industrial Use is > 60%**

- Aerospace
- Automotive
- Biology
- CFD
- Database
- Defense
- Digital Content Creation
- Digital Media
- Electronics
- Energy
- Environment
- Finance
- Gaming
- Geophysics
- Image Proc./Rendering
- Information Processing Service
- Information Service
- Life Science
- Media
- Medicine
- Pharmaceutics
- Research
- Retail
- Semiconductor
- Telecomm
- Weather and Climate Research
- Weather Forecasting

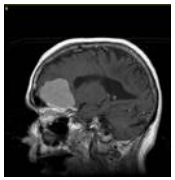




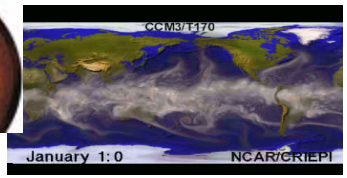









# 34rd List: The TOP10

| Rank | Site | Computer | Country | Cores | Rmax [Tflops] | % of Peak |
|---|---|---|---|---|---|---|
| 1 | DOE / OS<br>Oak Ridge Nat Lab | Jaguar / Cray<br>Cray XT5 sixCore 2.6 GHz | USA | 224,162 | 1.759 | 75 |
| 2 | DOE / NNSA<br>Los Alamos Nat Lab | Roadrunner / IBM<br>BladeCenter QS22/LS21 | USA | 122,400 | 1,042 | 76 |
| 3 | NSF / NICS /<br>U of Tennessee | Jaguar / Cray<br>Cray XT5 sixCore 2.6 GHz | USA | 98,928 | 831 | 81 |
| 4 | Forschungszentrum<br>Juelich (FZJ) | Jugene / IBM<br>Blue Gene/P Solution | Germany | 294,912 | 825 | 82 |
| 5 | National SC Center in<br>Tianjin / NUDT | Tianhe-1 / NUDT TH-1 /<br>IntelQC + AMD ATI Radeon 4870 | China | 71,680 | 563 | 46 |
| 6 | NASA / Ames Research<br>Center/NAS | Pleiades / SGI<br>SGI Altix ICE 8200EX | USA | 56,320 | 544 | 82 |
| 7 | DOE / NNSA<br>Lawrence Livermore NL | BlueGene/L IBM<br>eServer Blue Gene Solution | USA | 212,992 | 478 | 80 |
| 8 | DOE / OS<br>Argonne Nat Lab | Intrepid / IBM<br>Blue Gene/P Solution | USA | 163,840 | 458 | 82 |
| 9 | NSF<br>TACC/U. of Texas | Ranger / Sun<br>SunBlade x6420 | USA | 62,976 | 433 | 75 |
| 10 | DOE / NNSA<br>Sandia Nat Lab | Sun / SunBlade 6275 | USA | 41,616 | 424 | 87 |

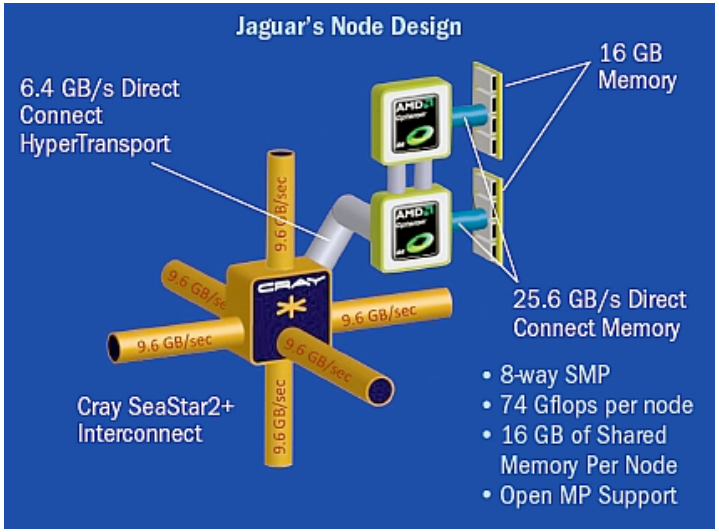
# 34rd List: The TOP10

| Rank | Site | Computer | Country | Cores | Rmax [Tflops] | % of Peak | Power [MW] | Flops/ Watt |
|---|---|---|---|---|---|---|---|---|
| 1 | DOE / OS<br>Oak Ridge Nat Lab | Jaguar / Cray<br>Cray XT5 sixCore 2.6 GHz | USA | 224,162 | 1.759 | 75 | 7.0 | 251 |
| 2 | DOE / NNSA<br>Los Alamos Nat Lab | Roadrunner / IBM<br>BladeCenter QS22/LS21 | USA | 122,400 | 1,042 | 76 | 2.48 | 446 |
| 3 | NSF / NICS /<br>U of Tennessee | Jaguar / Cray<br>Cray XT5 sixCore 2.6 GHz | USA | 98,928 | 831 | 81 | | |
| 4 | Forschungszentrum<br>Juelich (FZJ) | Jugene / IBM<br>Blue Gene/P Solution | Germany | 294,912 | 825 | 82 | 2.26 | 365 |
| 5 | National SC Center in<br>Tianjin / NUDT | Tianhe-1 / NUDT TH-1 /<br>IntelQC + AMD ATI Radeon 4870 | China | 71,680 | 563 | 46 | | |
| 6 | NASA / Ames Research<br>Center/NAS | Pleiades / SGI<br>SGI Altix ICE 8200EX | USA | 56,320 | 544 | 82 | 2.09 | 230 |
| 7 | DOE / NNSA<br>Lawrence Livermore NL | BlueGene/L IBM<br>eServer Blue Gene Solution | USA | 212,992 | 478 | 80 | 2.32 | 206 |
| 8 | DOE / OS<br>Argonne Nat Lab | Intrepid / IBM<br>Blue Gene/P Solution | USA | 163,840 | 458 | 82 | 1.26 | 363 |
| 9 | NSF<br>TACC/U. of Texas | Ranger / Sun<br>SunBlade x6420 | USA | 62,976 | 433 | 75 | 2.0 | 217 |
| 10 | DOE / NNSA<br>Sandia Nat Lab | Sun / SunBlade 6275 | USA | 41,616 | 424 | 87 | | |

# #1 ORNL's Newest System Jaguar XT5


Recently upgraded to a 2 Pflop/s system with more than 224K cores using AMD's 6 Core chip.

| | |
|---|---|
| **Peak performance** | **2.332 PF** |
| System memory | 300 TB |
| Disk space | 10 PB |
| Disk bandwidth | 240+ GB/s |
| Interconnect bandwidth | 374 TB/s |

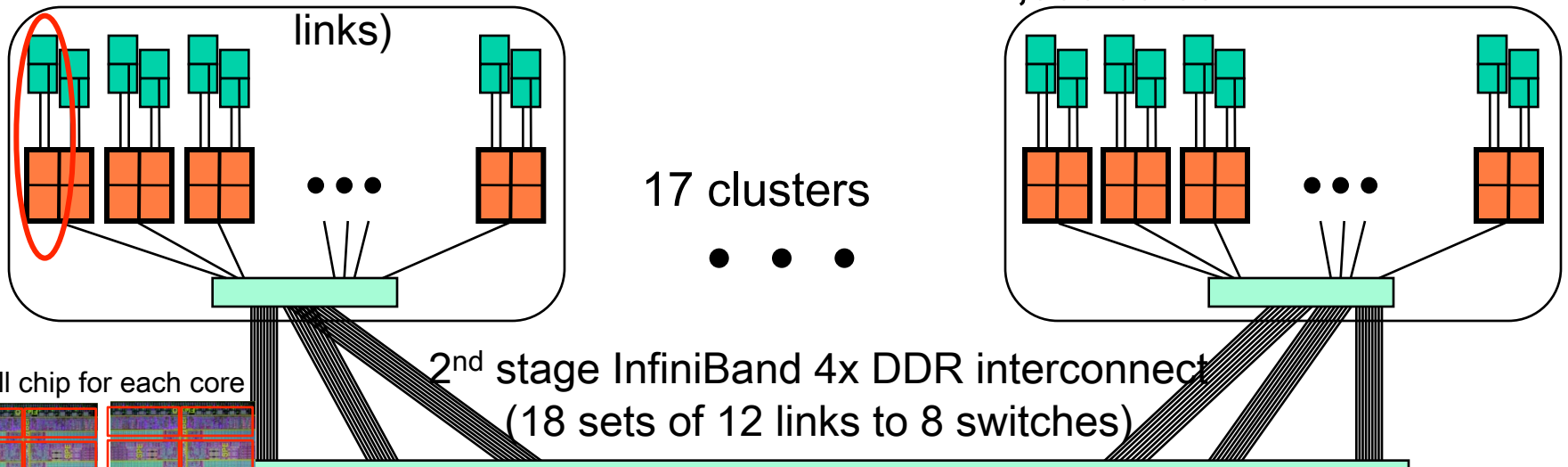



U.S. DEPARTMENT OF ENERGY | Office of Science

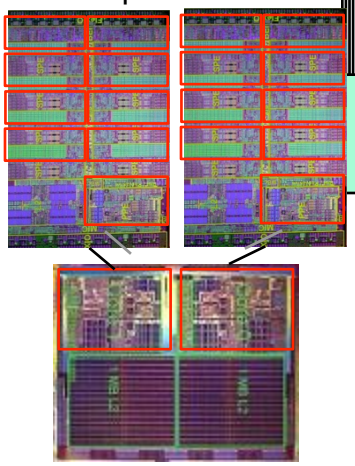
# #2 LANL Roadrunner
# A Petascale System in 2008

"Connected Unit" cluster
192 Opteron nodes
(180 w/ 2 dual-Cell blades
connected w/ 4 PCIe x8
links)

≈ 13,000 Cell HPC chips
≈ **1.33 PetaFlop/s** (from Cell)
≈ 7,000 dual-core Opterons
≈ **122,000 cores**




Cell chip for each core



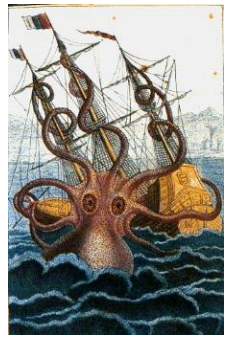
Dual Core Opteron Chip

**Based on the 100 Gflop/s (DP) Cell chip**

Hybrid Design (2 kinds of chips & 3 kinds of cores)
Programming required at 3 levels.

# #3 UT's HPC System

- University of Tennessee's National Institute for Computational Sciences
- Housed at ORNL, operated for the NSF, named Kraken

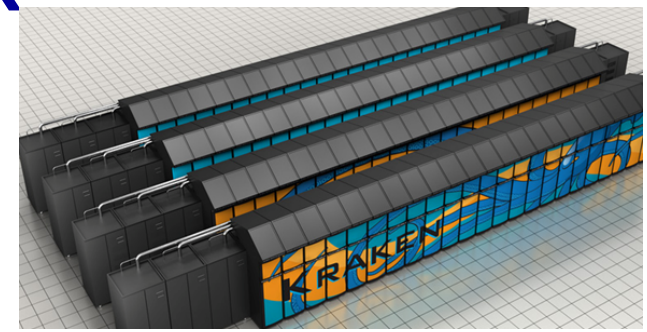


- Number 3 on the Top500

Just upgraded to 1 Pflop/s peak

99,072 cores, AMD 2.6 GHz

6 core chip, w/129 TB memory



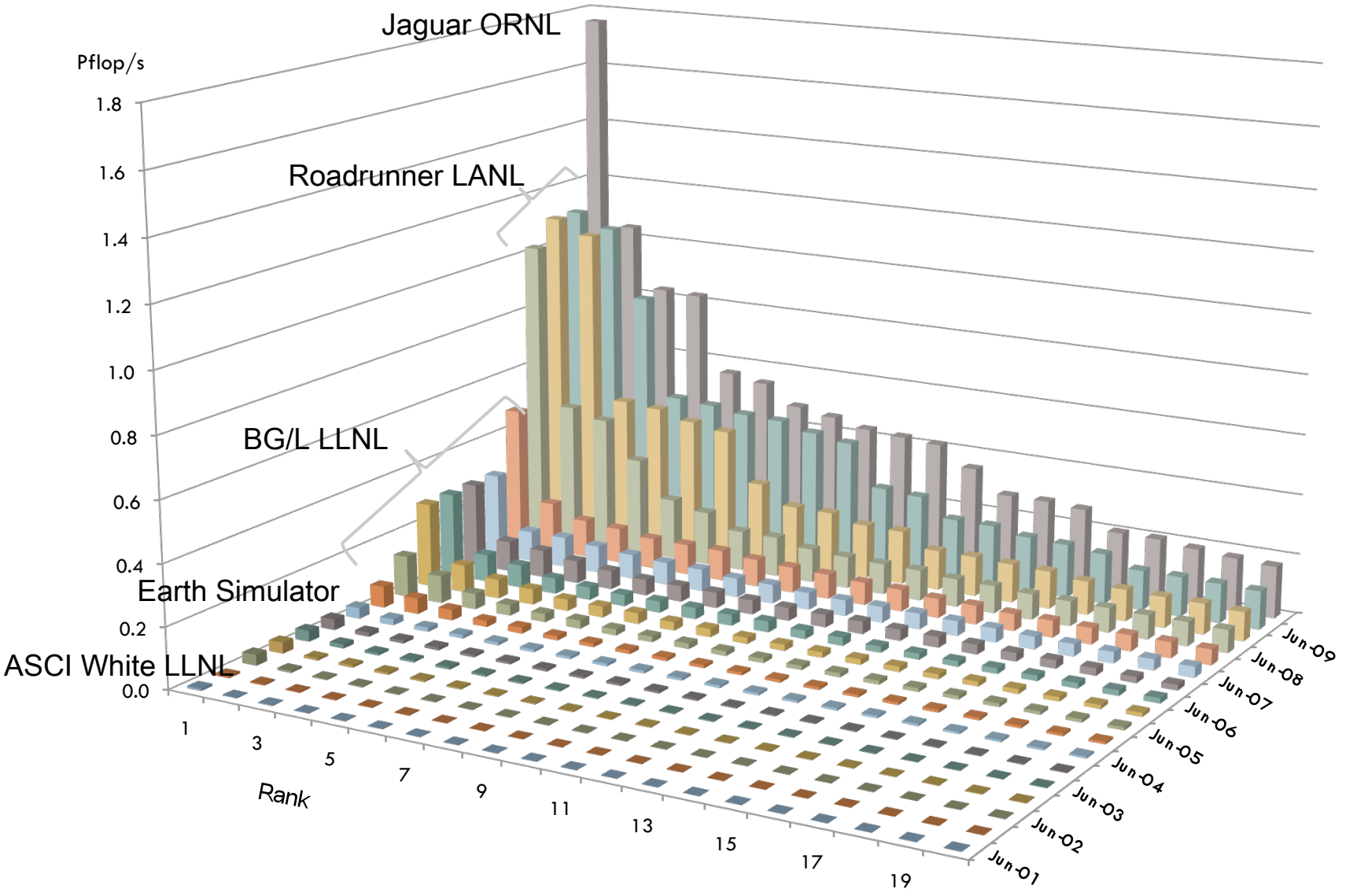
# #4 - Jugene

- IBM BG/P - 72 Racks with 32 nodecards x 32 compute nodes (total 73,728)
  - Compute node: 4-way SMP processor
  - Processor type: 32-bit PowerPC 450 core 850 MHz Processors: 294912
  - Overall peak performance: 1 Pflop/s
  - Linpack: 825.5 Tflop/s
  - Main memory: 2 Gbytes per node (aggregate 144 TB) I/O Nodes: 600 Networks: Three-dimensonal torus (compute nodes)
- Power Consumption:
  - max. 35 kW per rack

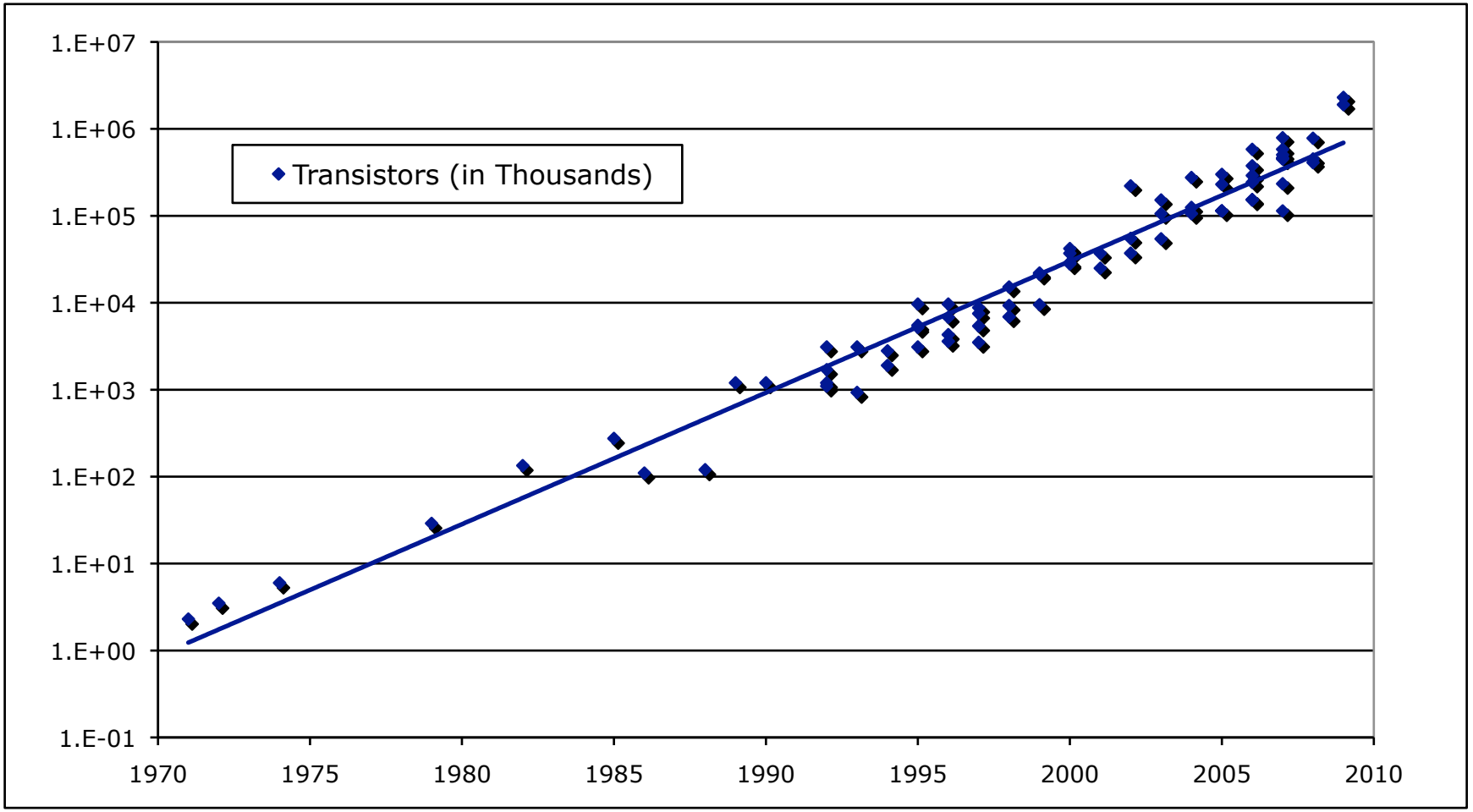
# #5 - National University of Defense Technology (NUDT)

- Tianhe-1
- Hybrid system, commodity + GPUs
- Theoretical peak 1.21 Pflop/s
- Linpack Benchmark at 563.1 Tflop/s
- 2560 nodes, each node: 2 Intel Quadcore Xeon5500 + 5,120 AMD ATI 4780 GPUs (each 10 cores)
  - 71,680 cores
  - Infiniband connected


# Performance of Top20 Over 10 Years

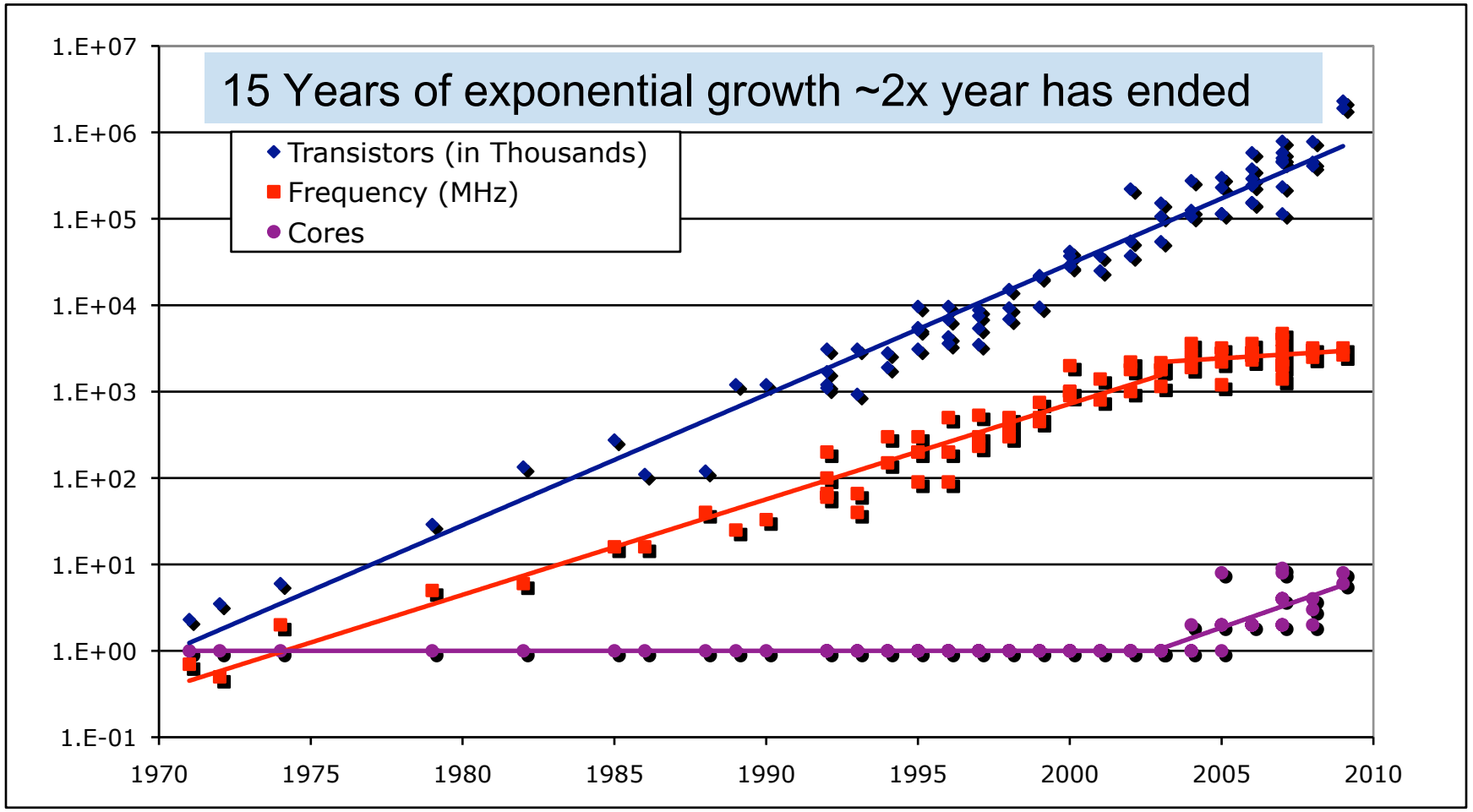


# Moore's Law is Alive and Well



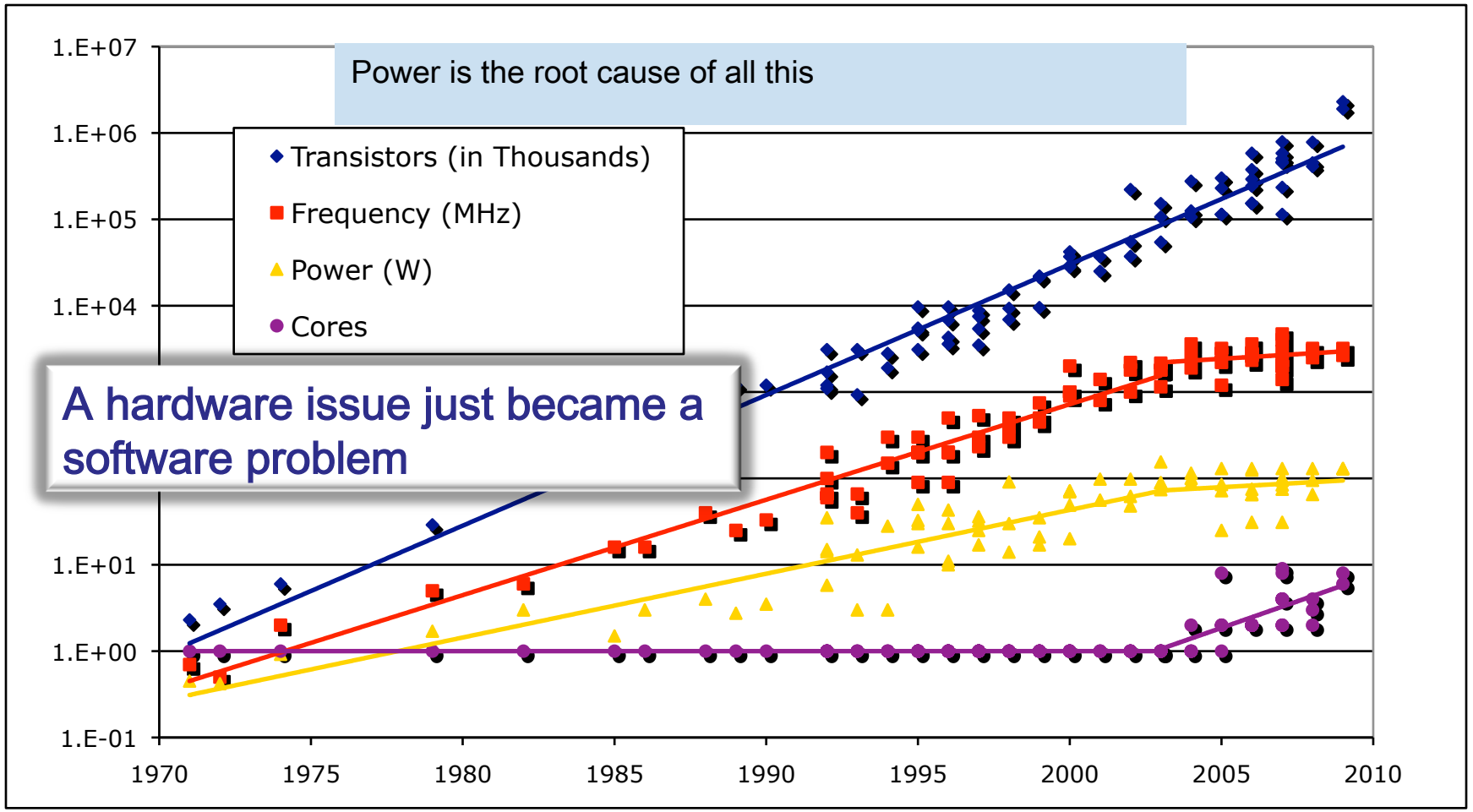

Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç
Slide from Kathy Yelick

# But Clock Frequency Scaling Replaced by Scaling Cores / Chip




Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç
Slide from Kathy Yelick

# Performance Has Also Slowed, Along with Power




Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç
Slide from Kathy Yelick

# Power Cost of Frequency

- Power ∝ Voltage$^2$ x Frequency (V$^2$F)
- Frequency ∝ Voltage
- Power ∝ Frequency$^3$

| | Cores | V | Freq | Perf | Power | PE (Bops/watt) |
|---|---|---|---|---|---|---|
| Superscalar | 1 | 1 | 1 | 1 | 1 | 1 |
| "New" Superscalar | 1X | 1.5X | 1.5X | 1.5X | 3.3X | 0.45X |

# Power Cost of Frequency

- Power ∝ Voltage$^2$ x Frequency (V$^2$F)
- Frequency ∝ Voltage
- Power ∝ Frequency$^3$

| | Cores | V | Freq | Perf | Power | PE (Bops/watt) |
|---|---|---|---|---|---|---|
| Superscalar | 1 | 1 | 1 | 1 | 1 | 1 |
| "New" Superscalar | 1X | 1.5X | 1.5X | 1.5X | 3.3X | 0.45X |
| Multicore | 2X | 0.75X | 0.75X | 1.5X | 0.8X | 1.88X |

(Bigger # is better)

50% more performance with 20% less power

Preferable to use multiple slower devices, than one superfast device

# Moore's Law Reinterpreted

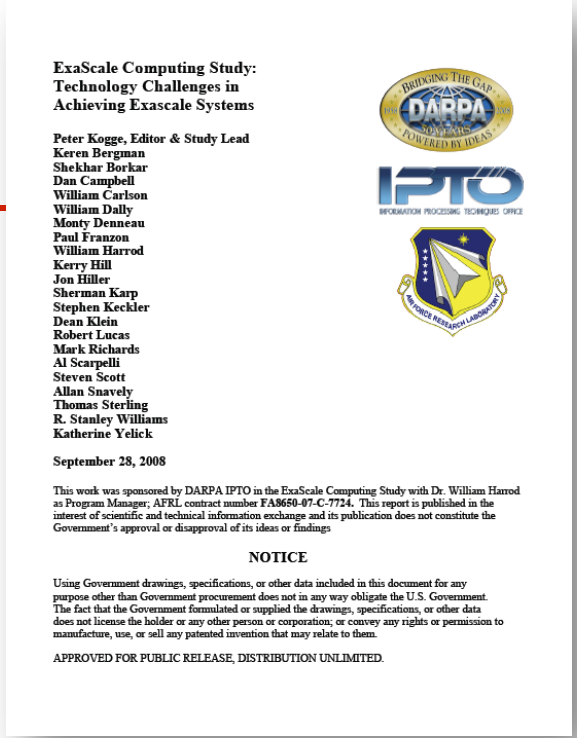
- Number of cores per chip doubles every 2 year, while clock speed decreases (not increases).
  - Need to deal with systems with millions of concurrent threads
    - Future generation will have billions of threads!
  - Need to be able to easily replace inter-chip parallelism with intro-chip parallelism
- Number of threads of execution doubles every 2 year

# Potential System Architectures

| Systems | 2015 | 2018-2020 |
|---|---|---|
| System peak | 100-200 Pflop/s | 1 Eflop/s |
| System memory | 5 PB | 10 PB |
| Node performance | 200-400 Gflop/s | 1-10 Tflop/s |
| Node memory bandwidth | 100 GB/s | 200-400 GB/s |
| Node concurrency | O(100) | O(1000) |
| Interconnect bandwidth | 25 GB/s | 50 GB/s |
| System size (nodes) | O(100,000) | 100,000-1,000,000 |
| Total concurrency | O(50,000,000) | O(1,000,000,000) |
| Storage | 150 PB | 300 PB |
| IO | 10 TB/s | 20 TB/s |
| MTTI | days | O(1 day) |
| Power | ~10 MW | ~20 MW |

# Exascale Computing

- Exascale systems are likely feasible by 2017±2
- 10-100 Million processing elements (cores or mini-cores) with chips perhaps as dense as 1,000 cores per socket, clock rates will grow more slowly
- 3D packaging likely
- Large-scale optics based interconnects
- 10-100 PB of aggregate memory
- Hardware and software based fault management
- Heterogeneous cores
- Performance per watt — stretch goal 100 GF/watt of sustained performance ⇒ >> 10 - 100 MW Exascale system
- Power, area and capital costs will be significantly higher than for today's fastest systems



**ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems**


**Peter Kogge, Editor & Study Lead**
**Keren Bergman**
**Shekhar Borkar**
**Dan Campbell**
**William Carlson**
**William Dally**
**Monty Denneau**
**Paul Franzon**
**William Harrod**
**Kerry Hill**
**Jon Hiller**
**Sherman Karp**
**Stephen Keckler**
**Dean Klein**
**Robert Lucas**
**Mark Richards**
**Al Scarpelli**
**Steven Scott**
**Allan Snavely**
**Thomas Sterling**
**R. Stanley Williams**
**Katherine Yelick**


**September 28, 2008**


This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number **FA8650-07-C-7724.** This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings


**NOTICE**



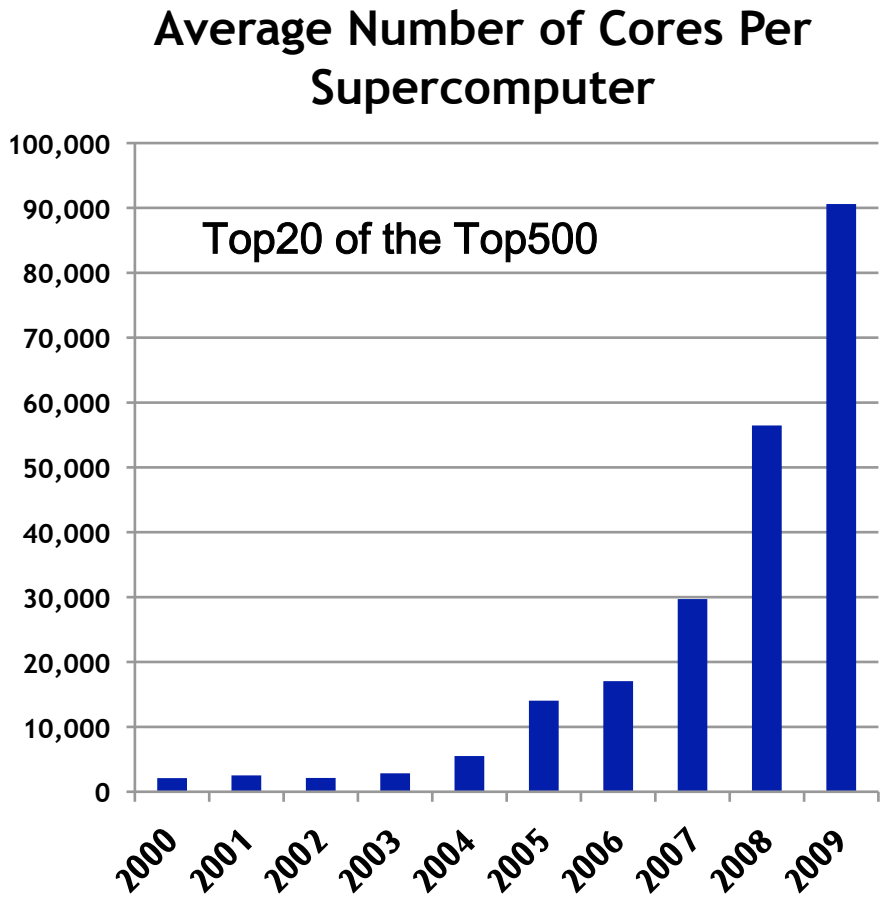
Google: exascale computing study

# Major Changes to Software

- **Must rethink the design of our software**
  - **Another disruptive technology**
    - Similar to what happened with cluster computing and message passing
  - **Rethink and rewrite the applications, algorithms, and software**

# Hardware and System Software Scalability

- Barriers
  - Fundamental assumptions of system software architecture did not anticipate exponential growth in parallelism
  - Number of components and MTBF changes the game
- Technical Focus Areas
  - System Hardware Scalability
  - System Software Scalability
  - Applications Scalability
- Technical Gap
  - 1000x improvement in system software scaling
  - 100x improvement in system software reliability



# Conclusions

- For the last decade or more, the research investment strategy has been overwhelmingly biased in favor of hardware.
- This strategy needs to be rebalanced - barriers to progress are increasingly on the software side.
- Moreover, the return on investment is more favorable to software.
  - Hardware has a half-life measured in years, while software has a half-life measured in decades.
- High Performance Ecosystem out of balance
  - Hardware, OS, Compilers, Software, Algorithms, Applications
    - No Moore's Law for software, algorithms and applications

# Collaborators / Support

Employment opportunities for post-docs in the ICL group at Tennessee

- **Top500**
  - **Hans Meuer, Prometeus**
  - **Erich Strohmaier, LBNL/NERSC**
  - **Horst Simon, LBNL/NERSC**