



# ***LINPACK Benchmarking and Beyond***

---

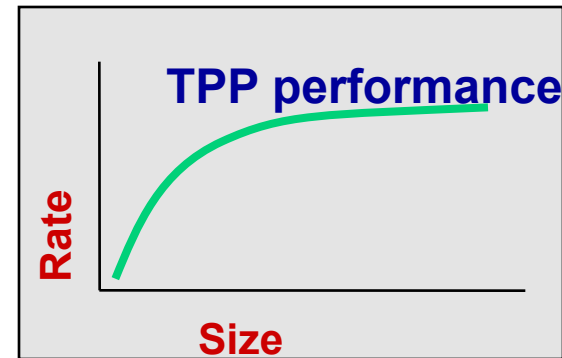
**Jack Dongarra**

**University of Tennessee  
Oak Ridge National Laboratory**

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

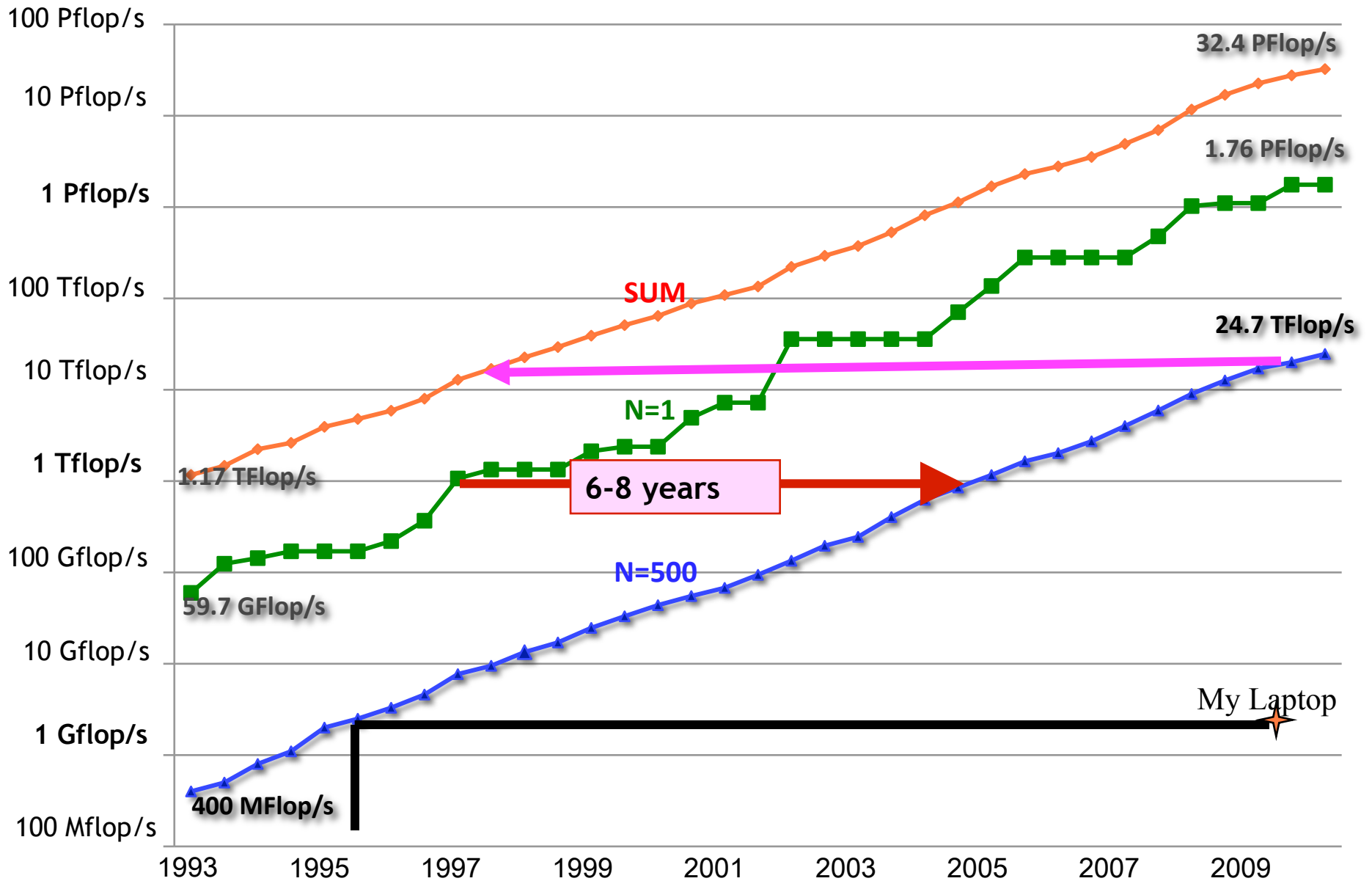
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

$$Ax=b, \text{ dense problem}$$



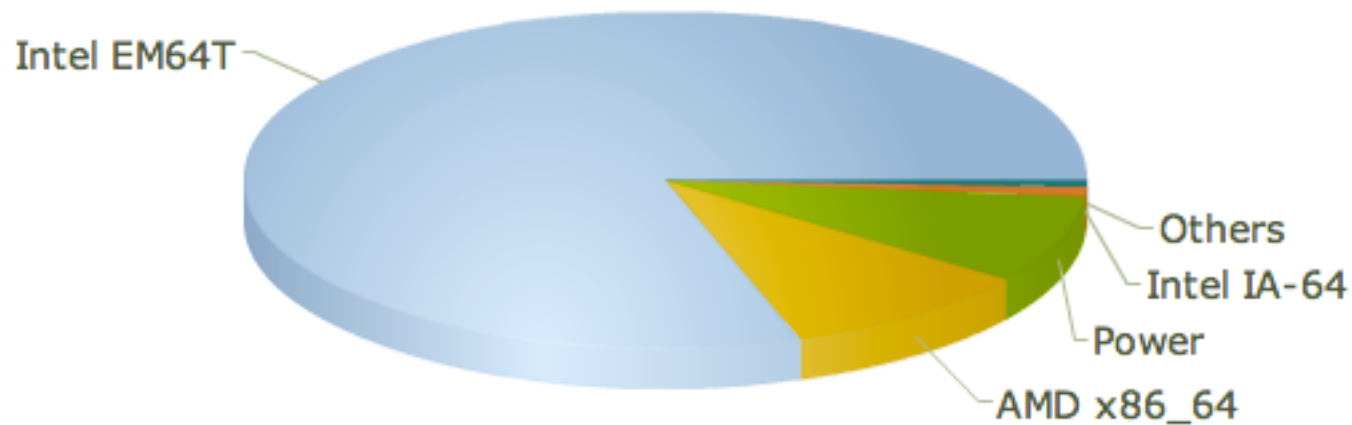
- Updated twice a year  
SC'xy in the States in November  
Meeting in Germany in June
- All data available from [www.top500.org](http://www.top500.org)

# Performance Development



# Processors Used in the Top500 Systems

---



Intel 81%  
AMD 10%  
IBM 8%

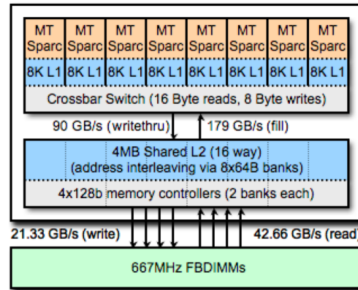




# Today's Multicores

99% of Top500 Systems Are Based on Multicore

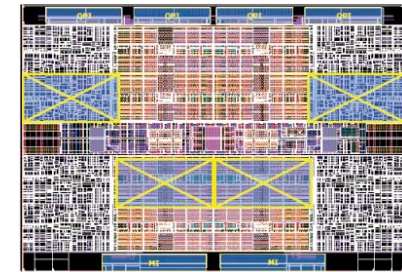
Of the Top500,  
499 are multicore.



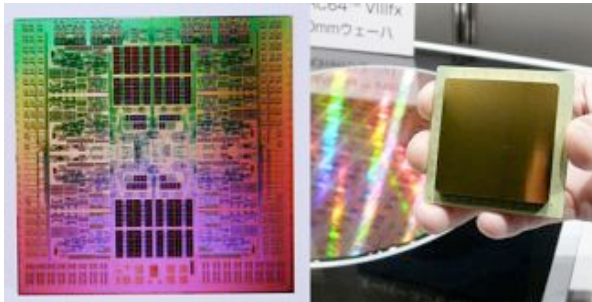
Sun Niagara2 (8 cores)



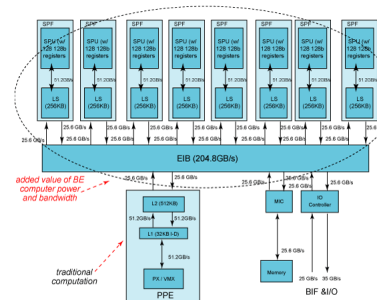
IBM Power 7 (8 cores)



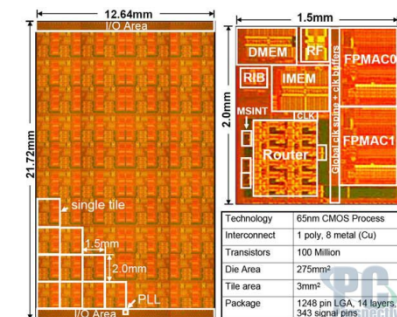
Intel Xeon(8 cores)



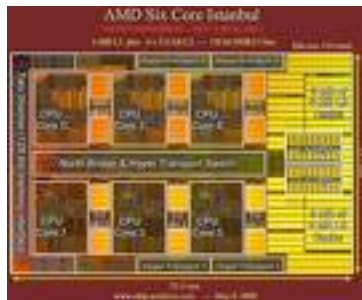
Fujitsu Venus (8 cores)



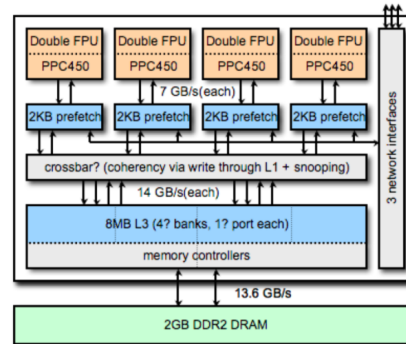
IBM Cell (9 cores)



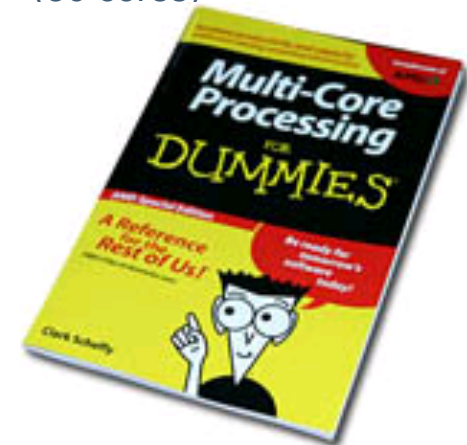
Intel Polarix [experimental]  
(80 cores)



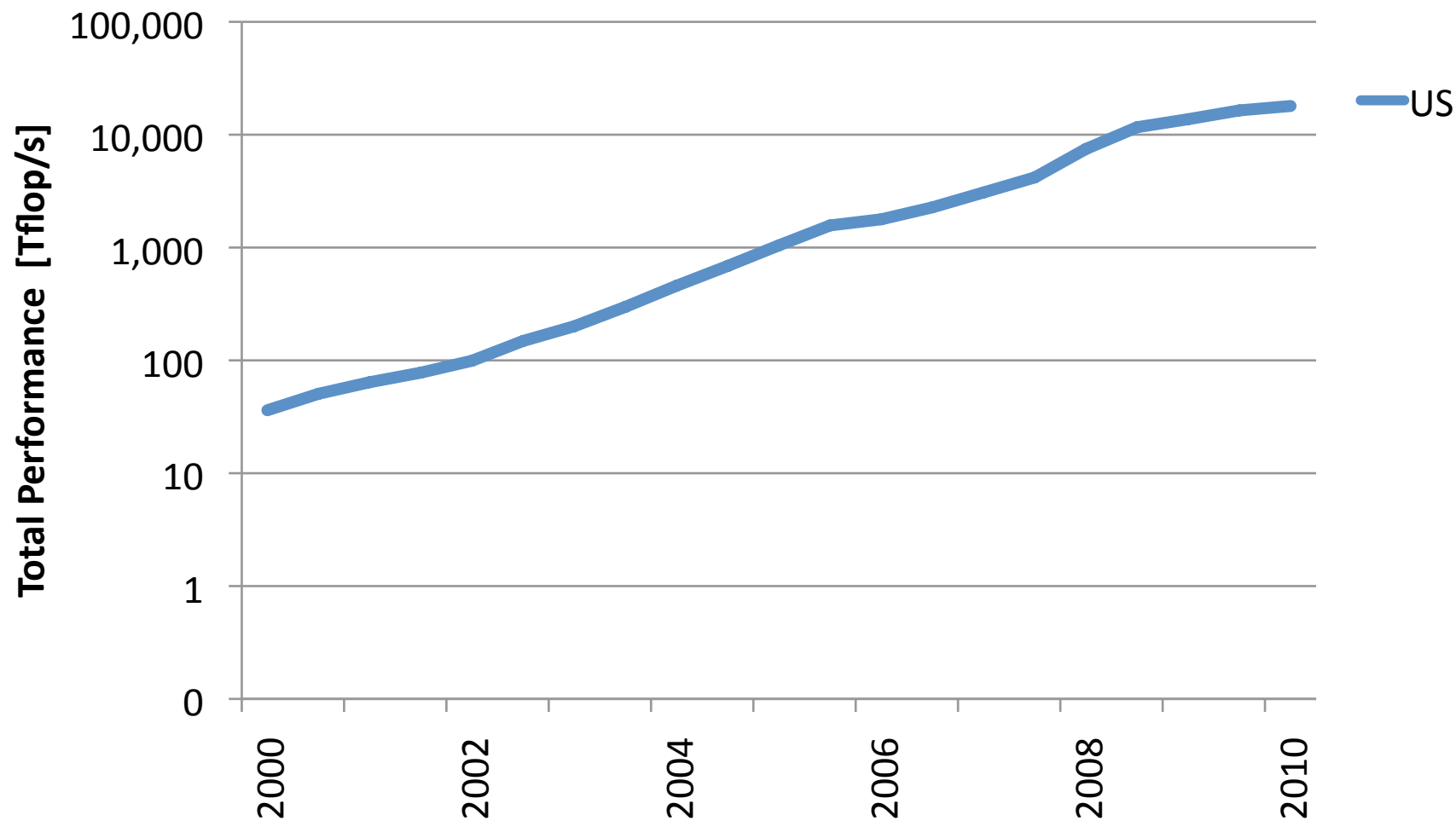
AMD Istanbul (6 cores)



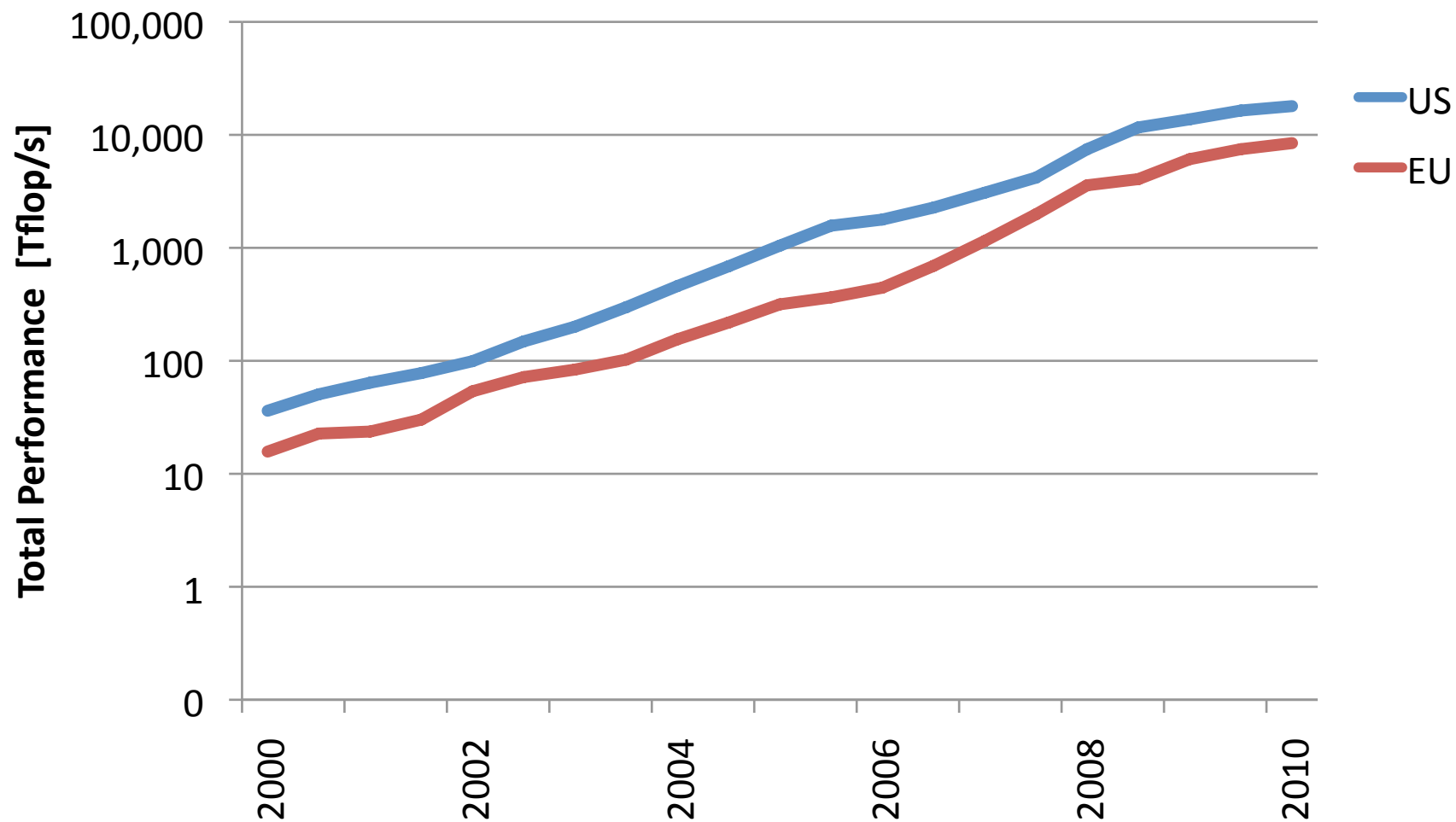
IBM BG/P (4 cores)



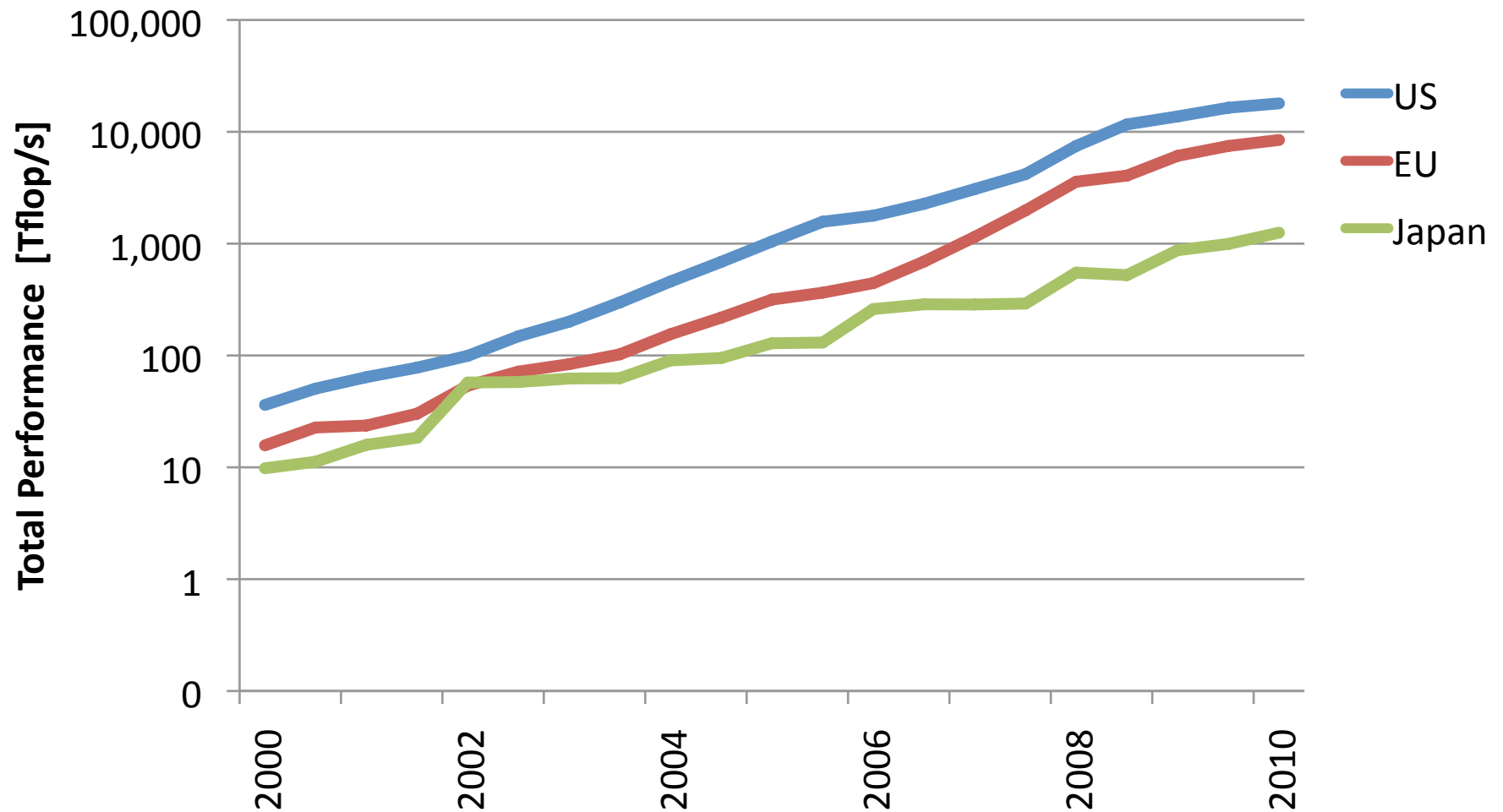
# Performance of Countries



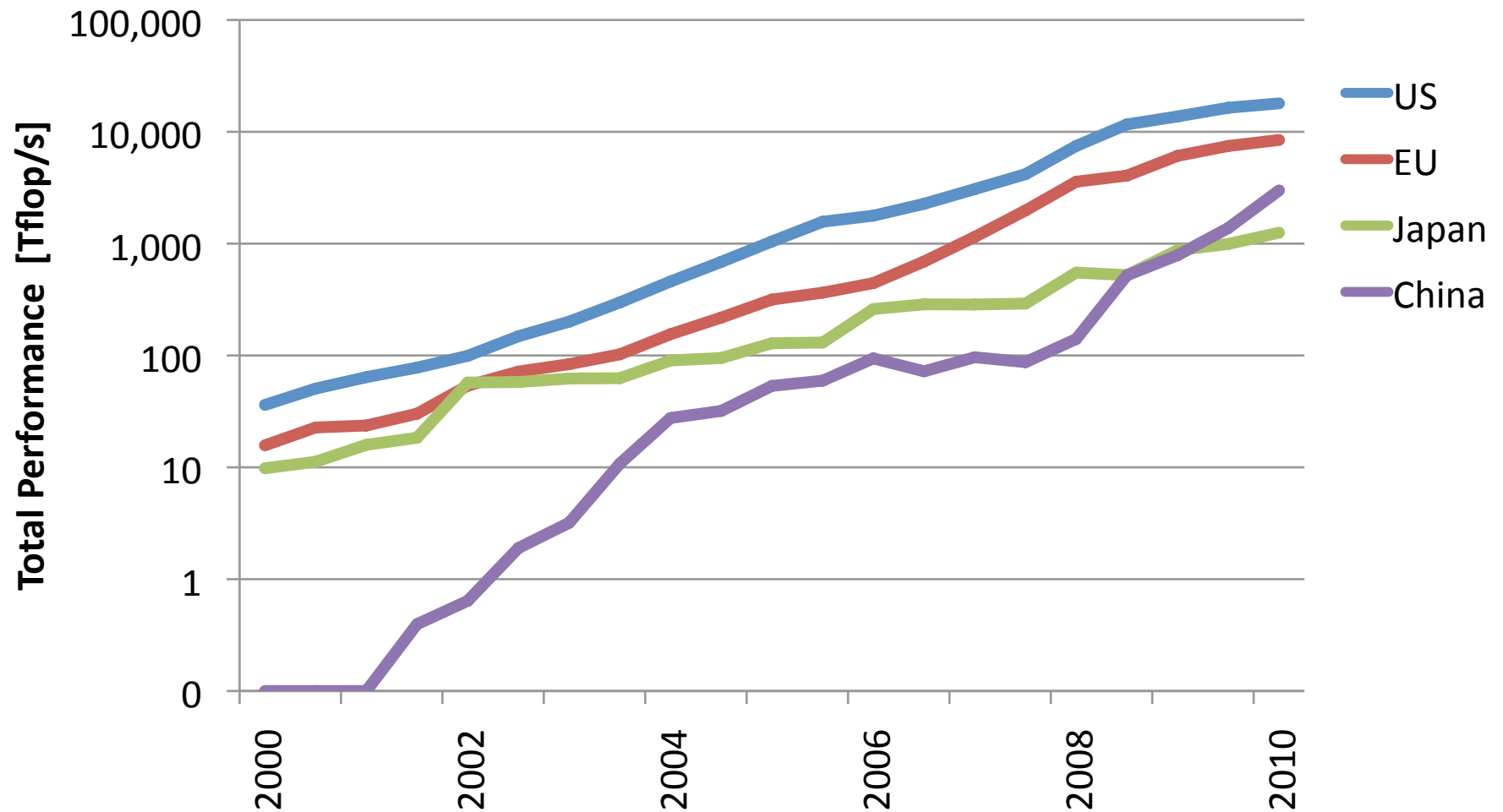
# Performance of Countries



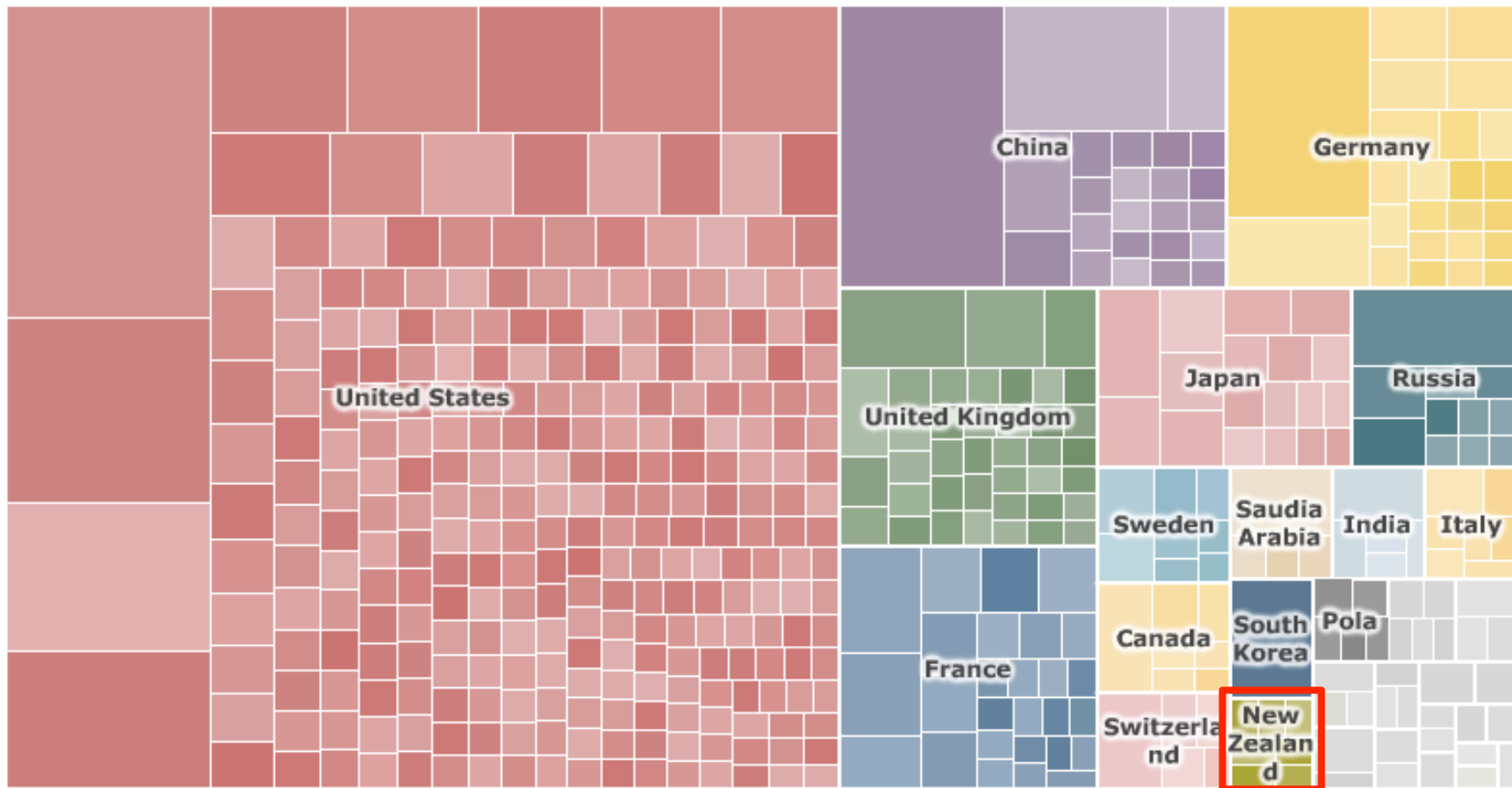
# Performance of Countries



# Performance of Countries



# Countries / System Share



# 35<sup>rd</sup> List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak
1	DOE / OS Oak Ridge Nat Lab	Jaguar / Cray Cray XT5 sixCore 2.6 GHz	USA	224,162	1.76	75
2	Nat. Supercomputer Center in Shenzhen	Nebulea / Dawning / TC3600 Blade, Intel X5650, Nvidia C2050 GPU	China	120,640	1.27	43
3	DOE / NNSA Los Alamos Nat Lab	Roadrunner / IBM BladeCenter QS22/LS21	USA	122,400	1.04	76
4	NSF / NICS / U of Tennessee	Kraken/ Cray Cray XT5 sixCore 2.6 GHz	USA	98,928	.831	81
5	Forschungszentrum Juelich (FZJ)	Jugene / IBM Blue Gene/P Solution	Germany	294,912	.825	82
6	NASA / Ames Research Center/NAS	Pleiades / SGI SGI Altix ICE 8200EX	USA	56,320	.544	82
7	National SC Center in Tianjin / NUDT	Tianhe-1 / NUDT TH-1 / IntelQC + AMD ATI Radeon 4870	China	71,680	.563	46
8	DOE / NNSA Lawrence Livermore NL	BlueGene/L IBM eServer Blue Gene Solution	USA	212,992	.478	80
9	DOE / OS Argonne Nat Lab	Intrepid / IBM Blue Gene/P Solution	USA	163,840	.458	82
10	DOE / NNSA Sandia Nat Lab	Red Sky / Sun / SunBlade 6275	USA	42,440	.433	87

# 35<sup>rd</sup> List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	DOE / OS Oak Ridge Nat Lab	Jaguar / Cray Cray XT5 sixCore 2.6 GHz	USA	224,162	1.76	75	7.0	251
2	Nat. Supercomputer Center in Shenzhen	Nebulea / Dawning / TC3600 Blade, Intel X5650, Nvidia C2050 GPU	China	120,640	1.27	43	2.58	493
3	DOE / NNSA Los Alamos Nat Lab	Roadrunner / IBM BladeCenter QS22/LS21	USA	122,400	1.04	76	2.48	446
4	NSF / NICS / U of Tennessee	Kraken/ Cray Cray XT5 sixCore 2.6 GHz	USA	98,928	.831	81	3.09	269
5	Forschungszentrum Juelich (FZJ)	Jugene / IBM Blue Gene/P Solution	Germany	294,912	.825	82	2.26	365
6	NASA / Ames Research Center/NAS	Pleiades / SGI SGI Altix ICE 8200EX	USA	56,320	.544	82	3.1	175
7	National SC Center in Tianjin / NUDT	Tianhe-1 / NUDT TH-1 / IntelQC + AMD ATI Radeon 4870	China	71,680	.563	46	1.48	380
8	DOE / NNSA Lawrence Livermore NL	BlueGene/L IBM eServer Blue Gene Solution	USA	212,992	.478	80	2.32	206
9	DOE / OS Argonne Nat Lab	Intrepid / IBM Blue Gene/P Solution	USA	163,840	.458	82	1.26	363
10	DOE / NNSA Sandia Nat Lab	Red Sky / Sun / SunBlade 6275	USA	42,440	.433	87	2.4	180



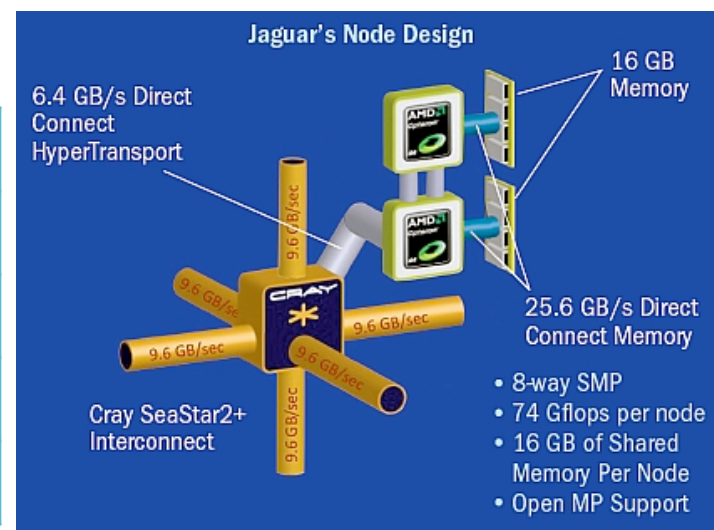


# #1 ORNL's Newest System Jaguar XT5



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 – National Supercomputer Center in Shenzhen - Dawning

---

- .. Nebulae
- .. Hybrid system, commodity + GPUs
- .. Theoretical peak 2.98 Pflop/s
- .. Linpack Benchmark at 1.27 Pflop/s
- .. 4640 nodes, each node:
  - 2 Intel 6-core Xeon5650 + Nvidia Fermi C2050 GPU (each 14 cores)
  - 120,640 cores
  - Infiniband connected
  - 500 MB/s peak per link and 8 GB/s



@supercomputing

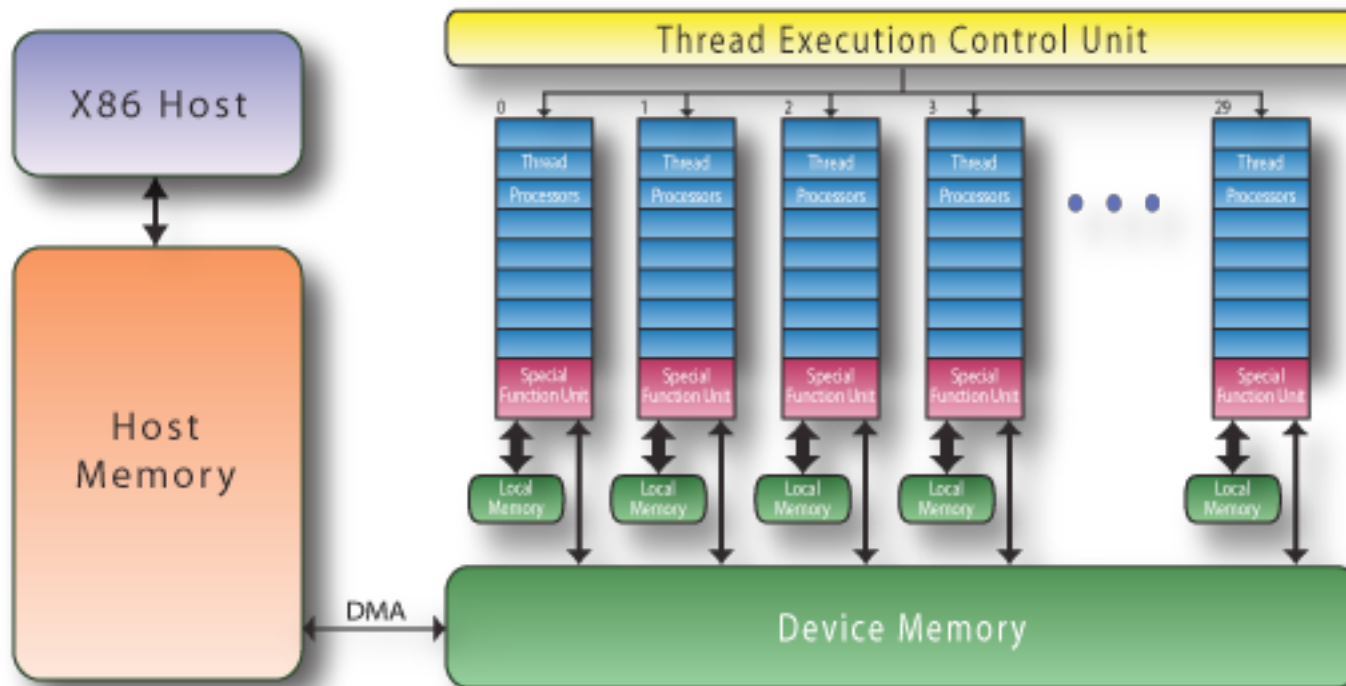
# Commodity plus Accelerators

# Commodity

Intel Xeon  
8 cores  
3 GHz  
8\*4 ops/cycle  
96 Gflop/s (DP)

## Accelerator (GPU)

Nvidia C2050 "Fermi"  
448 "Cuda cores"  
1.15 GHz  
448 ops/cycle  
515 Gflop/s (DP)



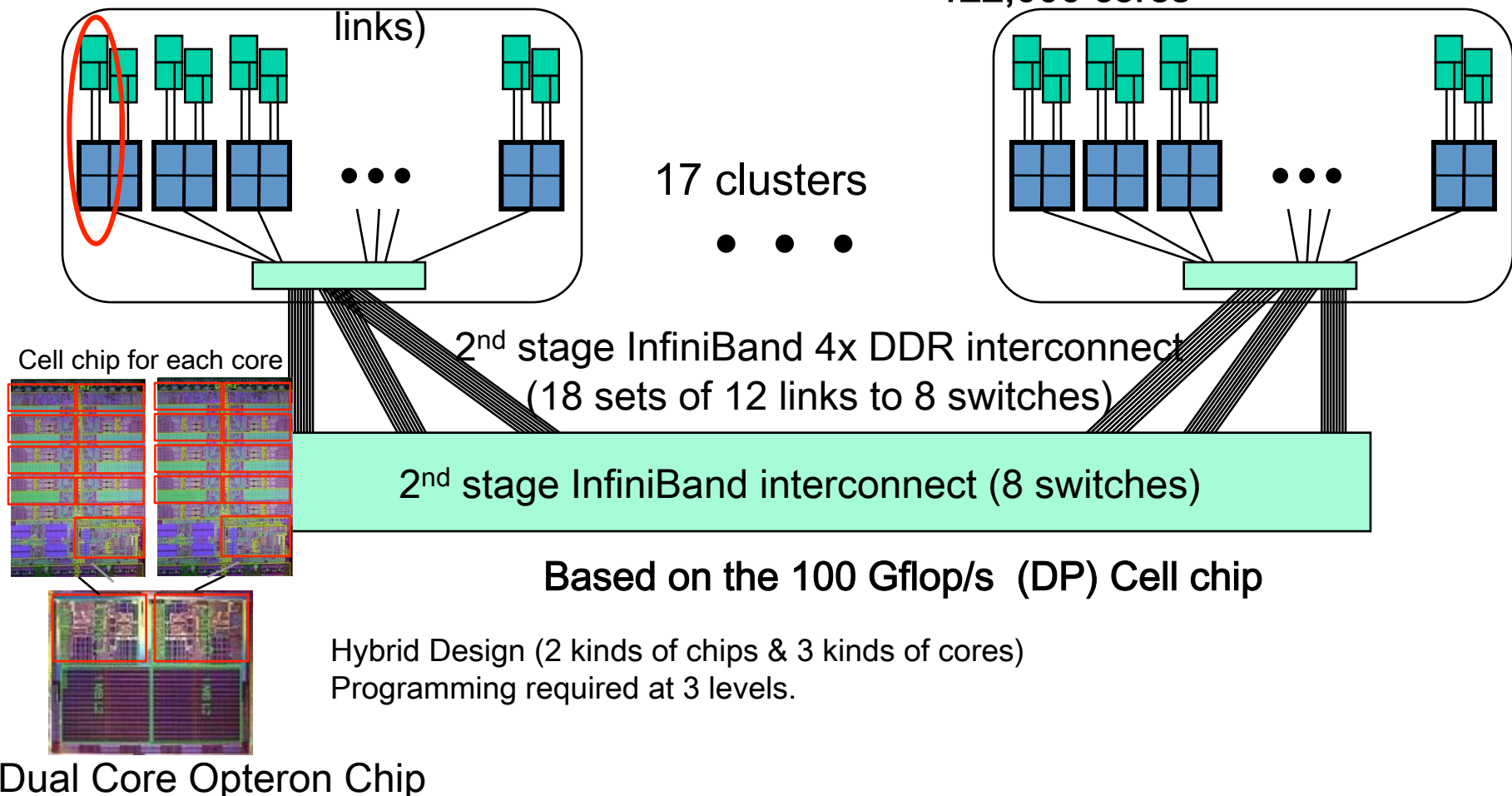
Interconnect  
PCI Express  
512 MB/s to 32GB/s  
8 MW – 512 MW

# #3 LANL Roadrunner

## A Petascale System in 2008

“Connected Unit” cluster  
192 Opteron nodes  
(180 w/ 2 dual-Cell blades  
connected w/ 4 PCIe x8

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



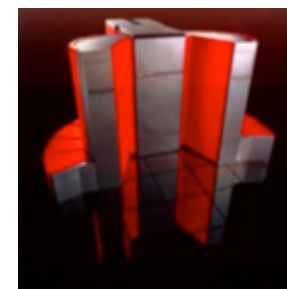


# Looking at the Gordon Bell Prize

(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



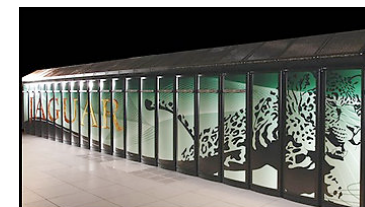
- 1 TFlop/s; 1998; Cray T3E; 1024 Processors

- ▣ Modeling of metallic magnet atoms, using a variation of the locally self-consistent multiple scattering method.



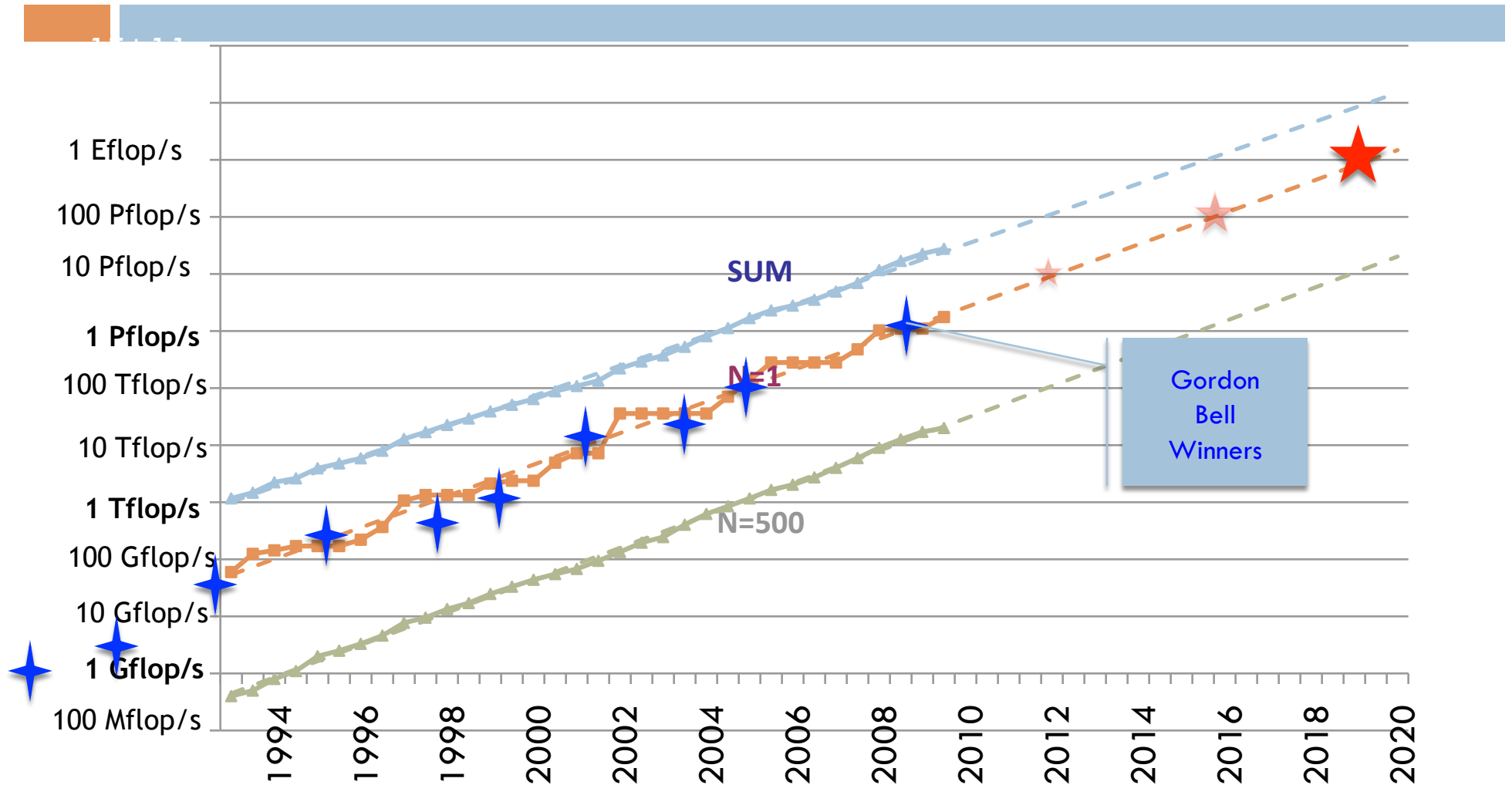
- 1 PFlop/s; 2008; Cray XT5;  $1.5 \times 10^5$  Processors

- ▣ Superconductive materials



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

# Performance Development in Top500







# Potential System Architecture

	2009	2019	Difference Today & 2019
System peak	2 Pflop/s	1 Eflop/s	O(1000)
Power	6 MW	~20 MW	
System memory	0.3 PB	32 - 64 PB [ .03 Bytes/Flop ]	O(100)
Node performance	125 GF	1,2 or 15TF	O(10) - O(100)
Node memory BW	25 GB/s	2 - 4TB/s [ .002 Bytes/Flop ]	O(100)
Node concurrency	12	O(1k) or 10k	O(100) - O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10) - O(100)
Total concurrency	225,000	O(billion) [O(10) to O(100) for latency hiding]	O(10,000)
Storage	15 PB	500-1000 PB (>10x system memory is min)	O(10) - O(100)
IO	0.2 TB	60 TB/s (how long to drain the machine)	O(100)
MTTI	days	O(1 day)	- O(10)



# Exascale ( $10^{18}$ Flop/s) Systems:

## Two possible paths

---

- **Light weight processors (think BG/P)**
  - ~1 GHz processor ( $10^9$ )
  - ~1 Kilo cores/socket ( $10^3$ )
  - ~1 Mega sockets/system ( $10^6$ )
- **Hybrid system (think GPU based)**
  - ~1 GHz processor ( $10^9$ )
  - ~10 Kilo FPUs/socket ( $10^4$ )
  - ~100 Kilo sockets/system ( $10^5$ )



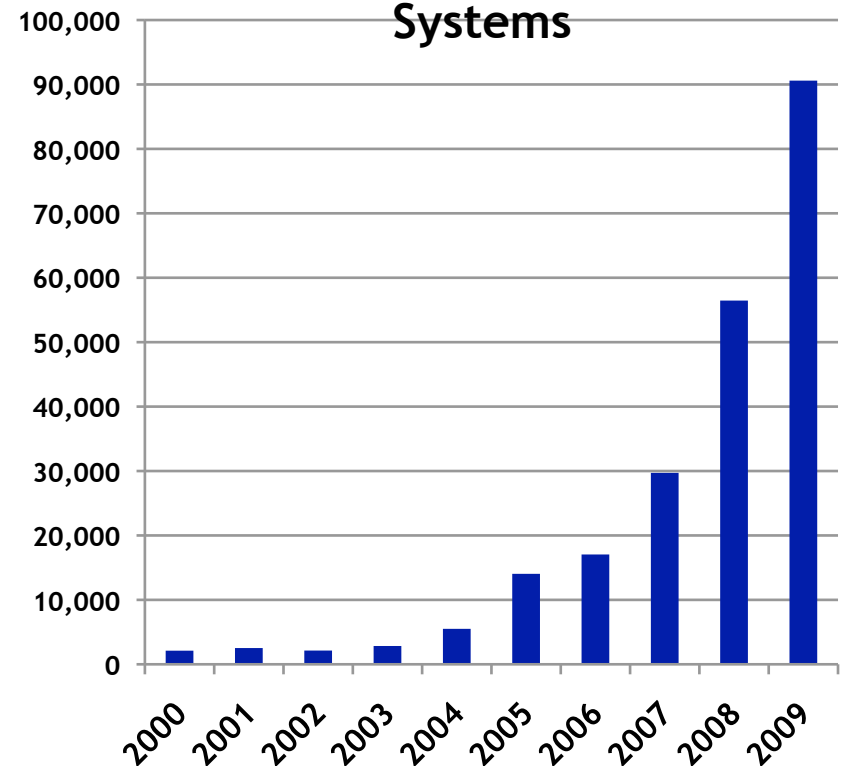


# Factors that Necessitate Redesign of Our Software

- Steepness of the ascent from terascale to petascale to exascale
- Extreme parallelism and hybrid design
  - Preparing for million/billion way parallelism
- Tightening memory/bandwidth bottleneck
  - Limits on power/clock speed implication on multicore
  - Reducing communication will become much more intense
  - Memory per core changes, byte-to-flop ratio will change
- Necessary Fault Tolerance
  - MTTF will drop
  - Checkpoint/restart has limitations

Software infrastructure does not exist today

Average Number of Cores Per Supercomputer for Top20 Systems



# Moore's Law reinterpreted

---

- Number of cores per chip will double every two years
- Clock speed will not increase (possibly decrease) because of Power

$$Power \propto Voltage^2 * Frequency$$

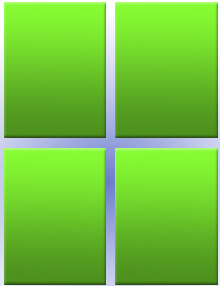
$$Voltage \propto Frequency$$

$$Power \propto Frequency^3$$

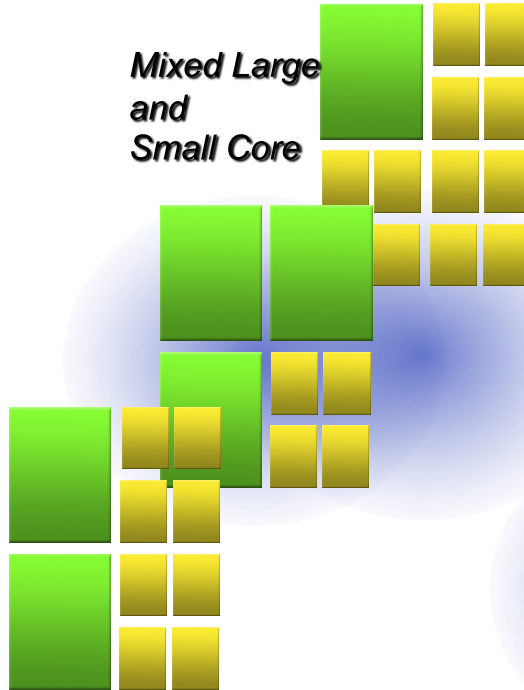
- Need to deal with systems with millions of concurrent threads
- Need to deal with inter-chip parallelism as well as intra-chip parallelism

# What's Next?

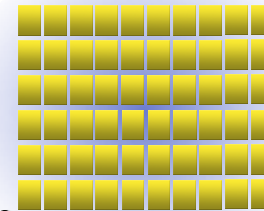
All Large Core



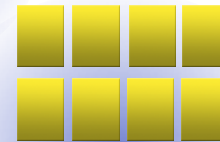
Mixed Large and Small Core



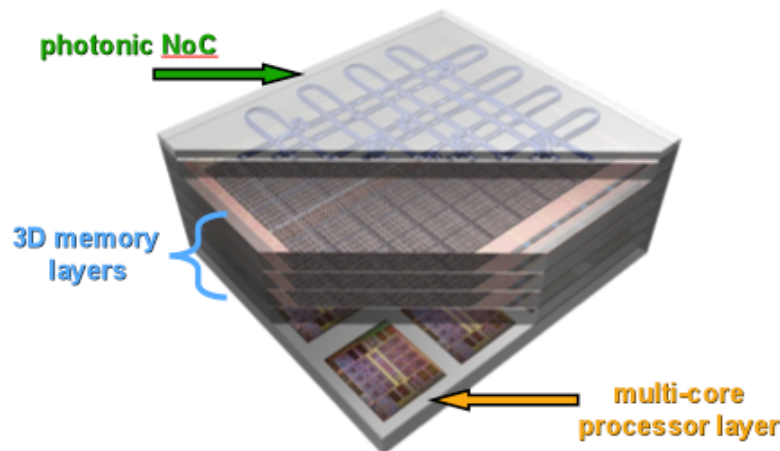
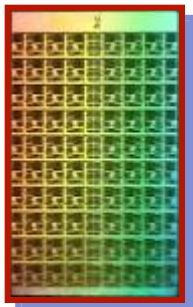
Many Small Cores



All Small Core



Many Floating-Point Cores



+ 3D Stacked Memory

Different Classes of Chips

- Home
- Games / Graphics
- Business
- Scientific



# 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**
- **Numerical libraries for example will change**
  - **For example, both LAPACK and ScaLAPACK will undergo major changes to accommodate this**

# Five Important Features to Consider When Computing at Scale

---

## 1. Effective Use of Many-Core and Hybrid architectures

- Break fork-join parallelism
- Dynamic Data Driven Execution
- Block Data Layout

## 2. Exploiting Mixed Precision in the Algorithms

- Single Precision is 2X faster than Double Precision
- With GP-GPUs 10x
- Power saving issues

## 3. Self Adapting / Auto Tuning of Software

- Too hard to do by hand

## 4. Fault Tolerant Algorithms

- With 1,000,000's of cores things will fail

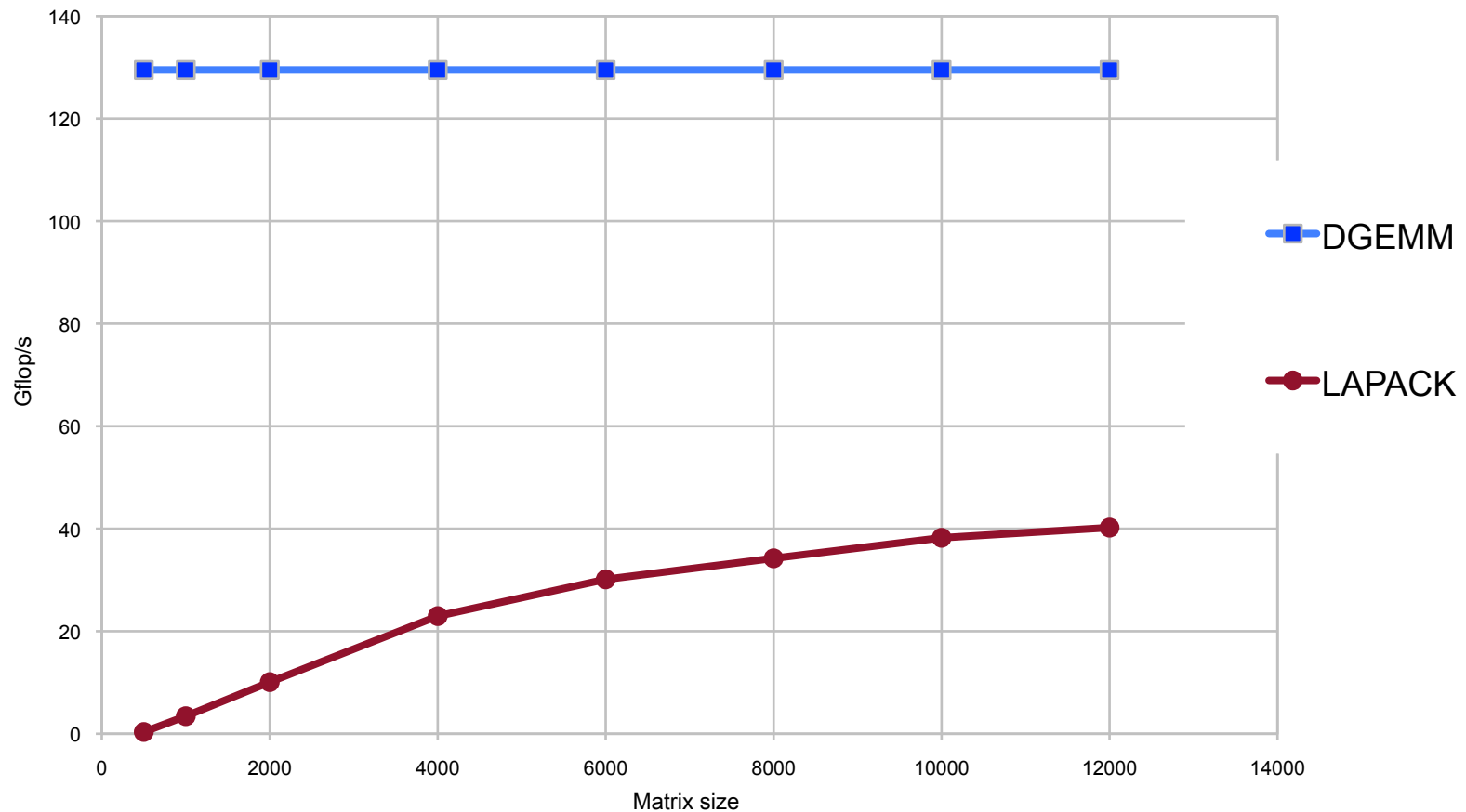
## 5. Communication Reducing Algorithms

- For dense computations from  $O(n \log p)$  to  $O(\log p)$  communications
- Asynchronous iterations
- GMRES k-step compute (  $x, Ax, A^2x, \dots A^kx$  )

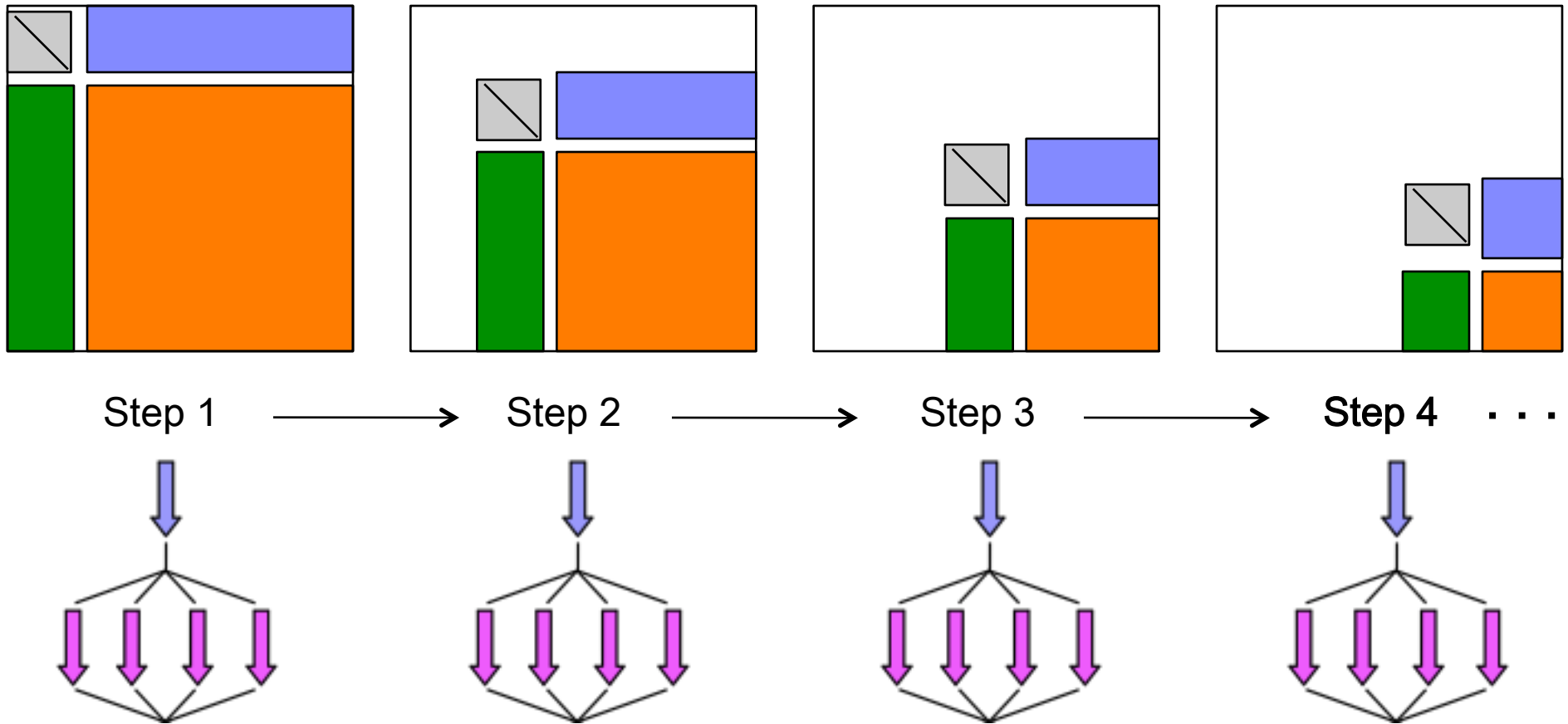


# LAPACK LU - Intel64 - 16 cores

DGETRF - Intel64 Xeon quad-socket quad-core (16 cores) - th. peak 153.6 Gflop/s

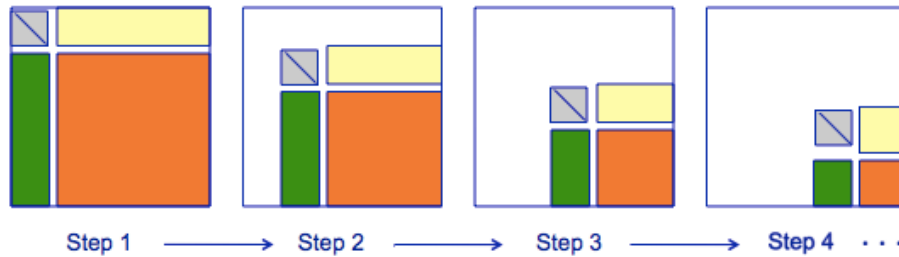


# LAPACK LU

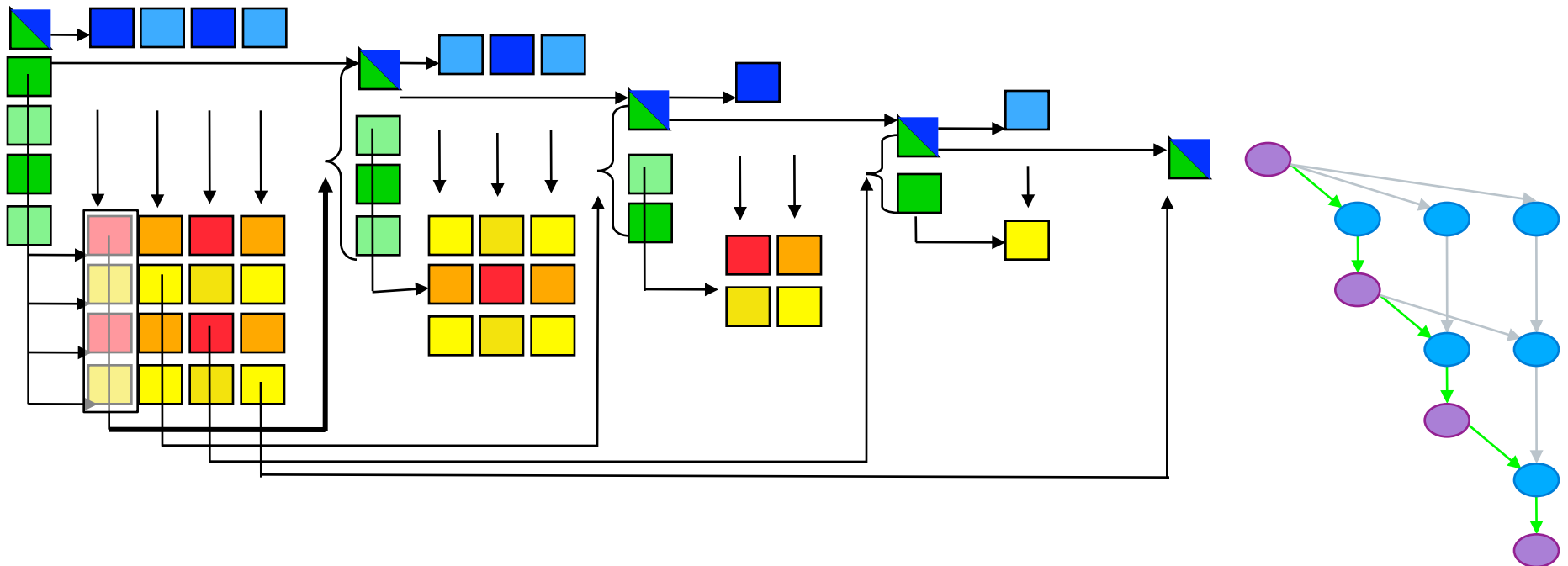


- Fork-join, bulk synchronous processing

# Parallel Tasks in LU



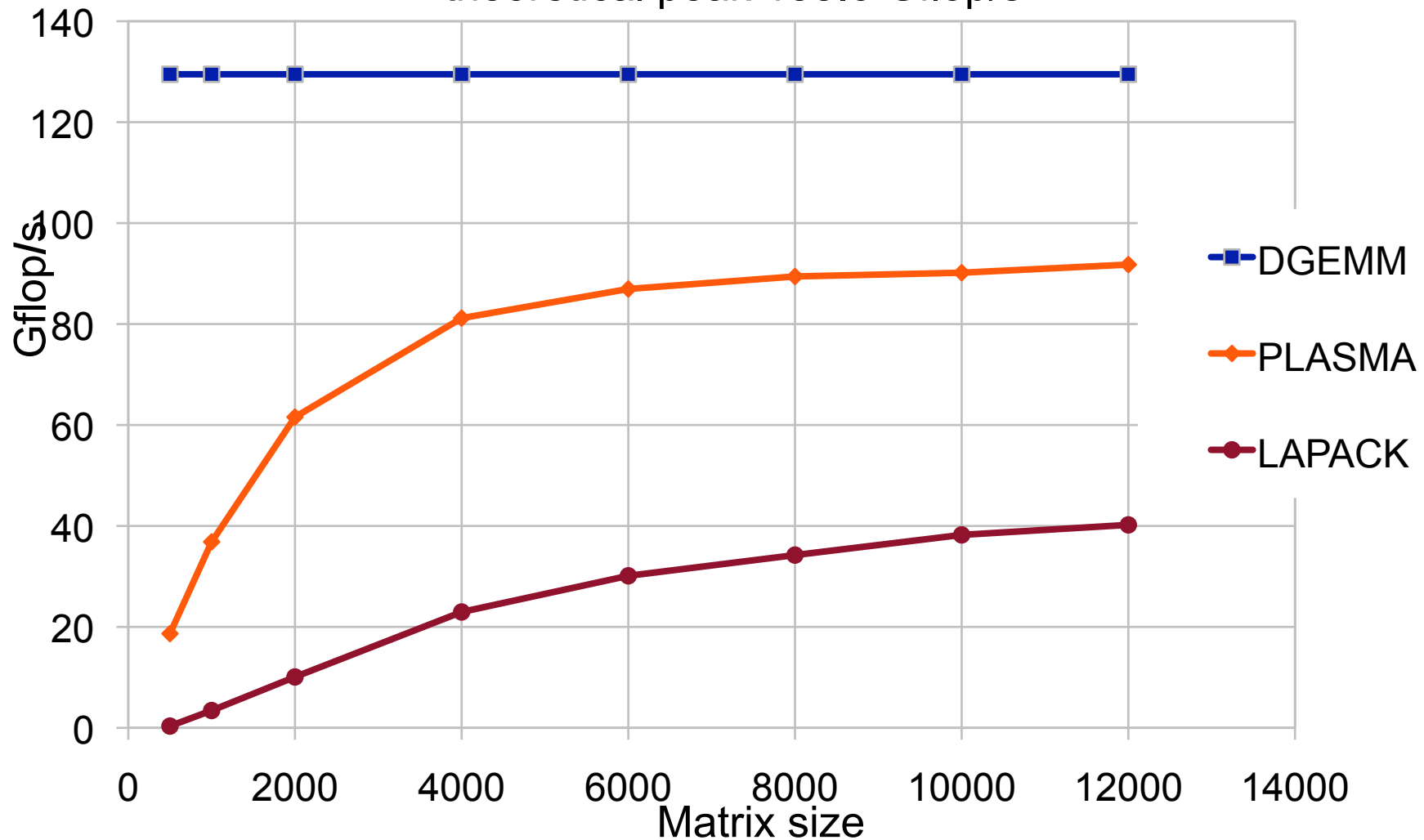
- Break into smaller tasks and remove dependencies



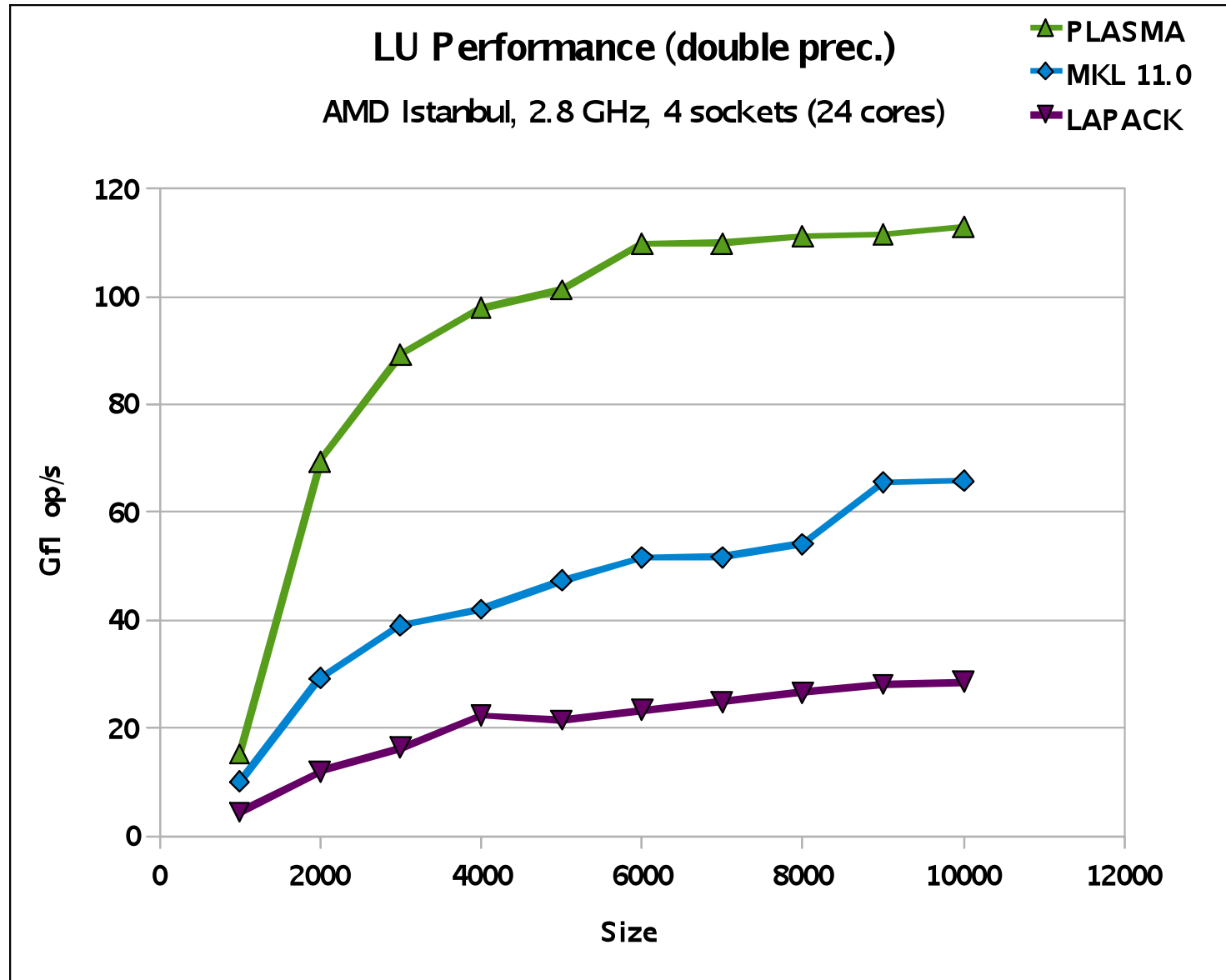


# LU - Intel64 - 16 cores

DGETRF - Intel64 Xeon quad-socket quad-core (16 cores)  
theoretical peak 153.6 Gflop/s

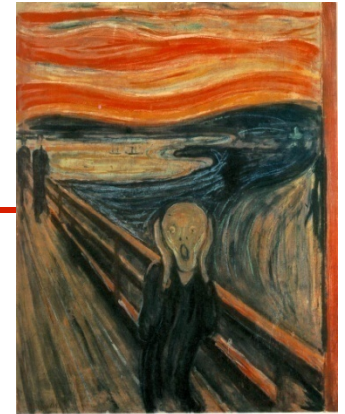


# Performance - 24 cores, LU



# Fault Tolerance

- Trends in HPC:
  - High end systems with millions of processors.
- Increased probability of a system failure
  - Most nodes today are robust, 3 year life.
  - Mean Time to Failure is growing shorter as systems grow, devices shrink, parts are powered on and off...
- MPI widely accepted in scientific computing.
  - Process faults not tolerated in MPI model.

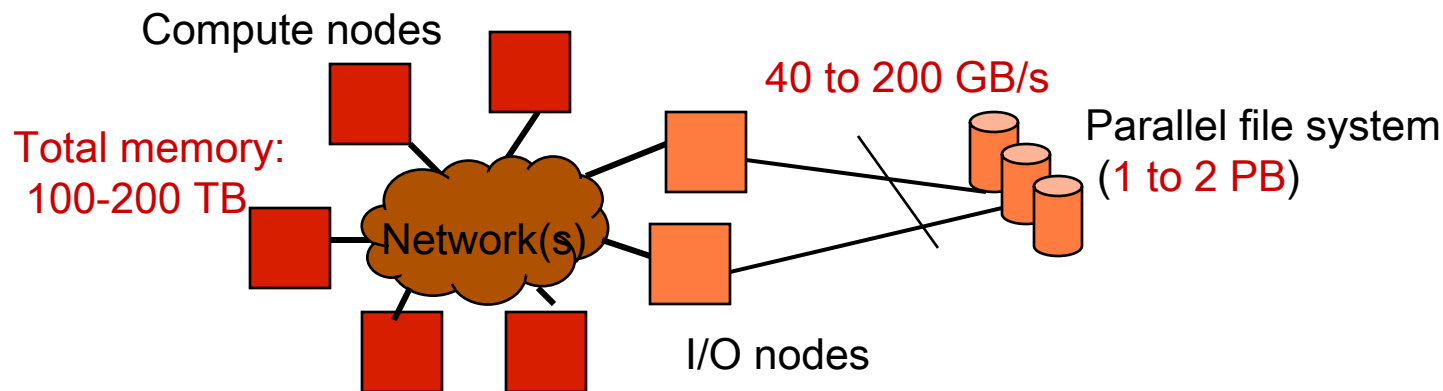


## Interesting studies:

- The computer failure data repository (CFDR) <http://cfdrr.usenix.org/>
- LANL Study: A Large-scale Study Of Failures In High-performance Computing Systems, B. Schroeder , & G. Gibson, *International Symposium on Dependable Systems and Networks (DSN 2006)*.

# Classic approach for FT: Checkpoint-Restart

## Typical “Balanced Architecture” for PetaScale Computers



1000 sec. < Ckpt < 2500 sec.

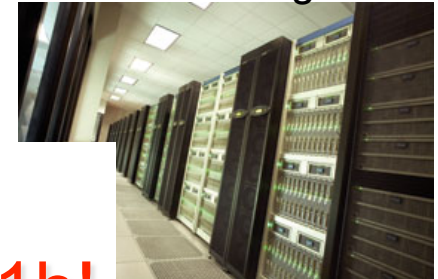
Systems	Perf.	Ckpt time	Source
Roadrunner			
LLNL BG/L			
Argonne BG/L			
Total SGI Altix			
IDRIS BG/P	100 TF	30 min.	IDRIS

Without optimization,  
Checkpoint-Restart needs about 1h!  
(~30 minutes each)

RoadRunner



TACC Ranger



LLNL BG/L



# A Call to Action



- 33
- Hardware has changed dramatically while software ecosystem has remained stagnant
  - Need to exploit new hardware trends (e.g., manycore, heterogeneity) that cannot be handled by existing software stack, memory per socket trends
  - Emerging software technologies exist, but have not been fully integrated with system software, e.g., UPC, Cilk, CUDA, HPCS
  - Community codes unprepared for sea change in architectures
  - No global evaluation of key missing components



# International Exascale Software Program

---



Improve the world's simulation and modeling capability by improving the coordination and development of the HPC software environment

Workshops:

**Build an international plan for coordinating research for the next generation open source software for scientific high-performance computing**

# International Community Effort

---

35

- **We believe this needs to be an international collaboration for various reasons including:**
  - **The scale of investment**
  - **The need for international input on requirements**
  - **US, Europeans, Asians, and others are working on their own software that should be part of a larger vision for HPC.**
  - **No global evaluation of key missing components**
  - **Hardware features are uncoordinated with software development**

# Where We Are Today:



36

- ☐ SC08 (Austin TX) meeting to generate interest
- ☐ Funding from DOE's Office of Science & NSF Office of Cyberinfrastructure and sponsorship by Europeans and Asians
- ☐ US meeting (Santa Fe, NM) April 6-8, 2009
  - ☐ 65 people
- ☐ European meeting (Paris, France) June 28-29, 2009
  - ☐ 70 people
  - ☐ Outline Report
- ☐ Asian meeting (Tsukuba Japan) October 18-20, 2009
  - ☐ Draft roadmap
  - ☐ Refine Report
- ☐ SC09 (Portland OR) BOF to inform others
  - ☐ Public Comment
  - ☐ Draft Report presented
- ☐ European meeting (Oxford, UK) April 13-14, 2010
  - ☐ Refine and prioritize roadmap
  - ☐ Explore governance structure and management models for IESP

Apr 2009

Jun 2009

Oct 2009

Nov 2009

Apr 2010



# Roadmap Components

[www.exascale.org](http://www.exascale.org)

<b>4.1 Systems Software.....</b>	
4.1.1 Operating systems .....	
4.1.2 Runtime Systems .....	
4.1.2 I/O systems .....	
4.1.3 External Environments .....	
4.1.4 Systems Management.....	
<b>4.2 Development Environments.....</b>	
4.2.1 Programming Models .....	
4.2.2 Frameworks .....	
4.2.3 Compilers.....	
4.2.4 Numerical Libraries.....	
4.2.5 Debugging tools.....	
<b>4.3 Applications.....</b>	
4.3.1 Application Element: Algorithms.....	
4.3.2 Application Support: Data Analysis and Visualization .....	
4.3.3 Application Support: Scientific Data Management .....	
<b>4.4 Crosscutting Dimensions .....</b>	
4.4.1 Resilience.....	
4.4.2 Power Management .....	
4.4.3 Performance Optimization .....	
4.4.4 Programmability.....	

# INTERNATIONAL EXASCALE SOFTWARE PROJECT

38

- [www.exascale.org](http://www.exascale.org)



## ROADMAP

Jack Dongarra  
Pete Beckman  
Terry Moore  
Jean-Claude Andre  
Jean-Yves Berthou  
Taisuke Boku  
Franck Cappello  
Barbara Chapman  
Xuebin Chi

Alok Choudhary  
Sudip Dosanjh  
Al Geist  
Bill Gropp  
Robert Harrison  
Mark Hereld  
Michael Heroux  
Adolfy Hoisie  
Koh Hotta

Yutaka Ishikawa  
Fred Johnson  
Sanjay Kale  
Richard Kenway  
David Keyes  
Bill Kramer  
Jesus Labarta  
Bob Lucas  
Barney Maccabe

Satoshi Matsuoka  
Paul Messina  
Bernd Mohr  
Matthias Mueller  
Wolfgang Nagel  
Hiroshi Nakashima  
Michael E. Papka  
Dan Reed  
Mitsuhsa Sato

Ed Seidel  
John Shalf  
David Skinner  
Thomas Sterling  
Rick Stevens  
William Tang  
John Taylor  
Rajeev Thakur  
Anne Trefethen

Marc Snir  
Aad van der Steen  
Fred Streitz  
Bob Sugar  
Shinji Sumimoto  
Jeffrey Vetter  
Robert Wisniewski  
Kathy Yelick

[www.exascale.org](http://www.exascale.org)

### SPONSORS





# If you are wondering what's beyond ExaFlops

---

## Mega, Giga, Tera, Peta, Exa, Zetta ...

$10^3$	kilo
$10^6$	mega
$10^9$	giga
$10^{12}$	tera
$10^{15}$	peta
$10^{18}$	exa
$10^{21}$	zetta

$10^{24}$	yotta
$10^{27}$	xona
$10^{30}$	weka
$10^{33}$	vunda
$10^{36}$	uda
$10^{39}$	treda
$10^{42}$	sorta
$10^{45}$	rinta
$10^{48}$	quexa
$10^{51}$	pepta
$10^{54}$	ocha
$10^{57}$	nena
$10^{60}$	minga
$10^{63}$	luma