International Young Scientists Conference
"High Performance Computing and Simulation" 2012

On the Future of High Performance Computing: How to Think for Peta and Exascale Computing

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory
University of Manchester

Overview

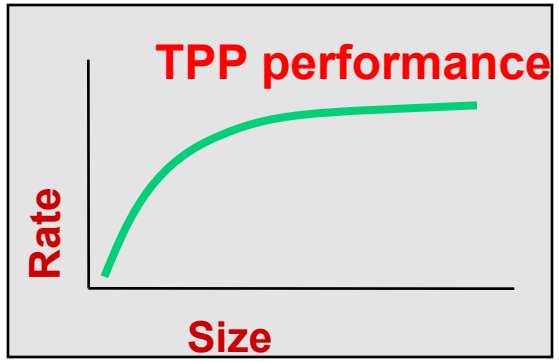
- Take a look at high performance computing
- What's driving HPC
- Future Trends

Top500 List of Supercomputers

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

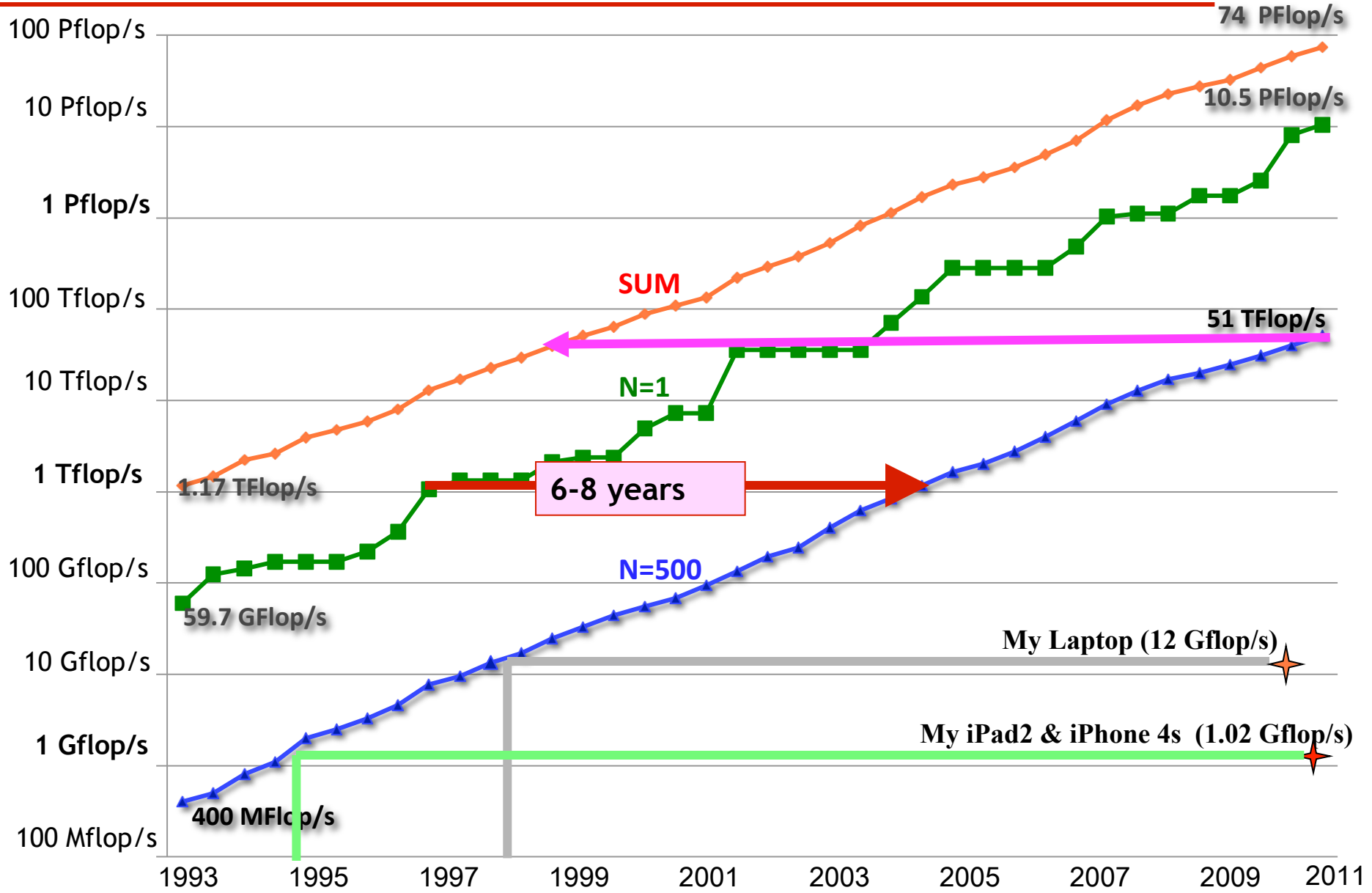
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

$Ax=b$, *dense problem*

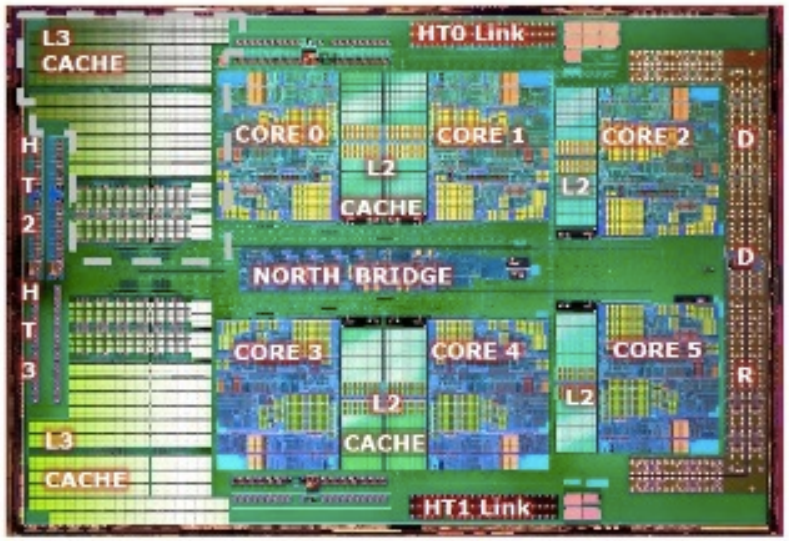


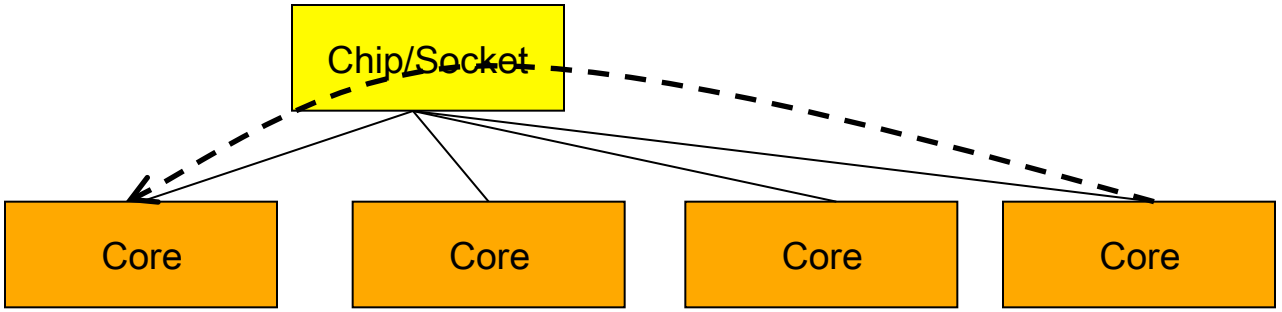
- Updated twice a year
  SC‘xy in the States in November
  Meeting in Germany in June
- All data available from **www.top500.org**

Performance Development

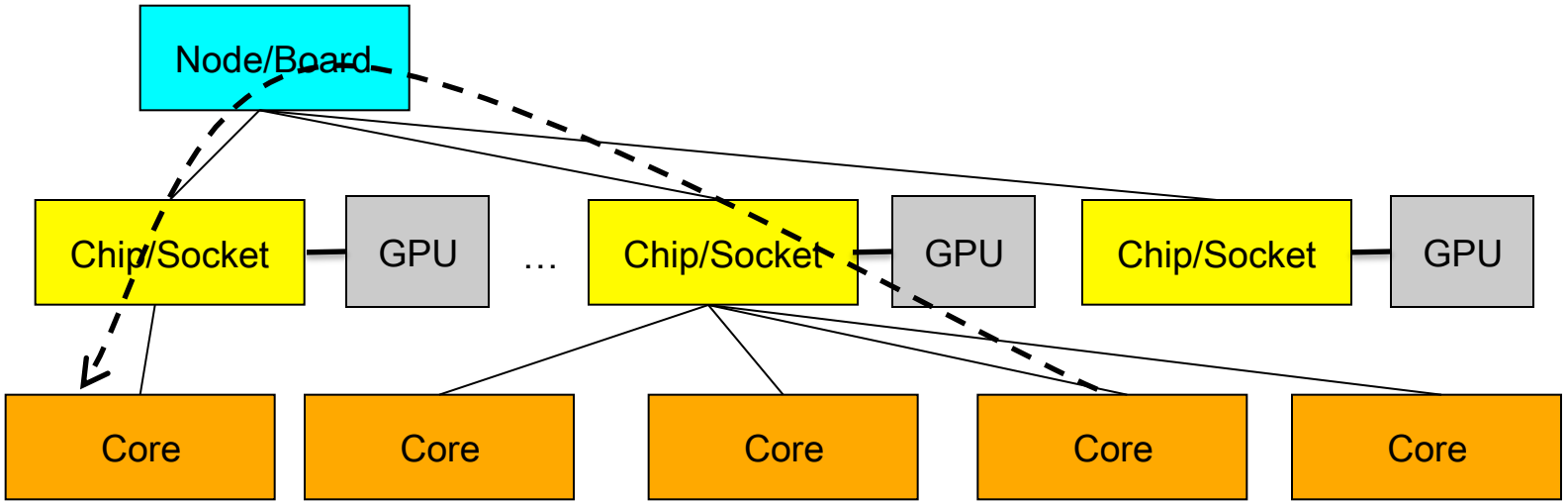


Example of typical parallel machine



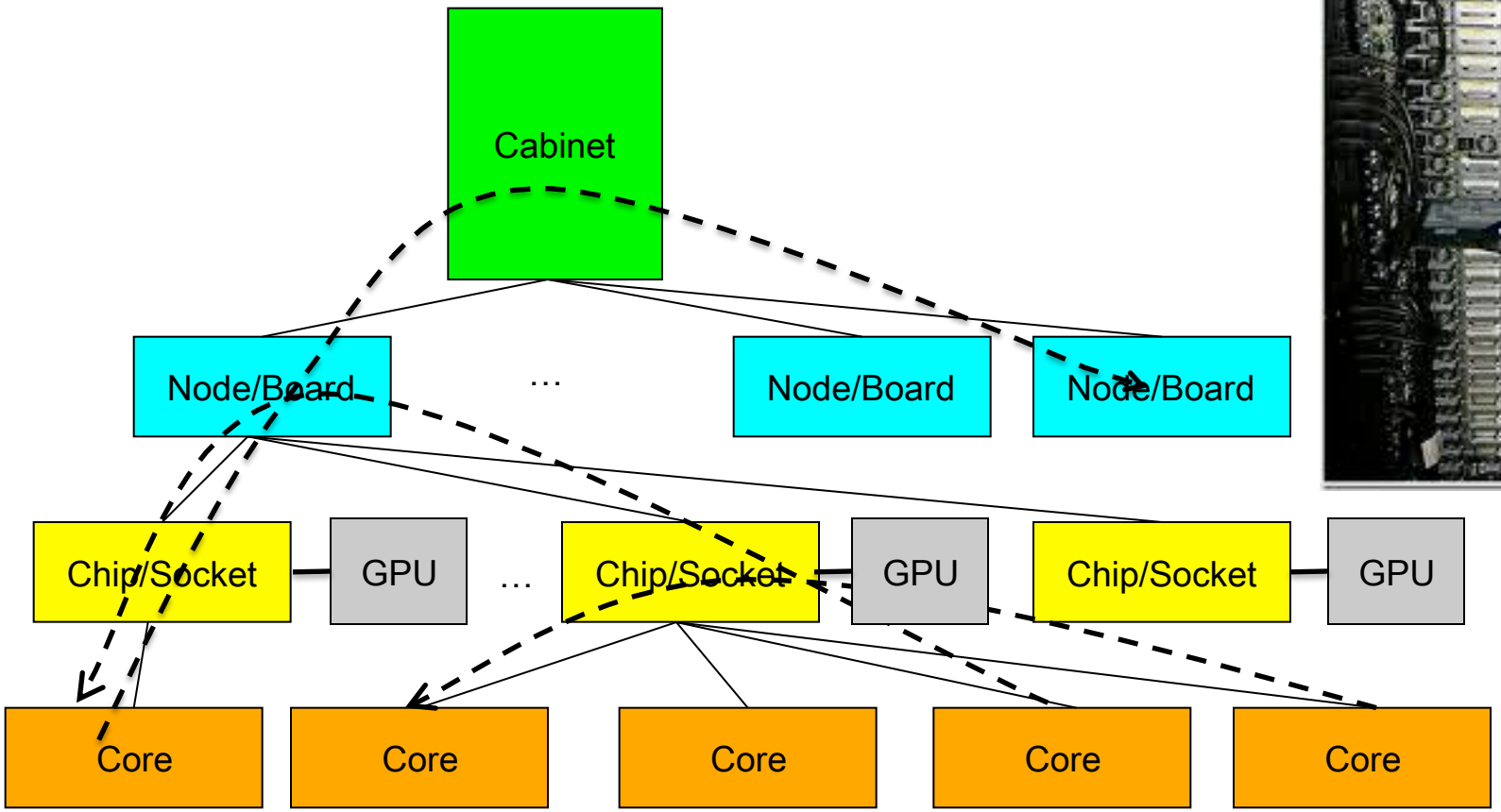


Example of typical parallel machine

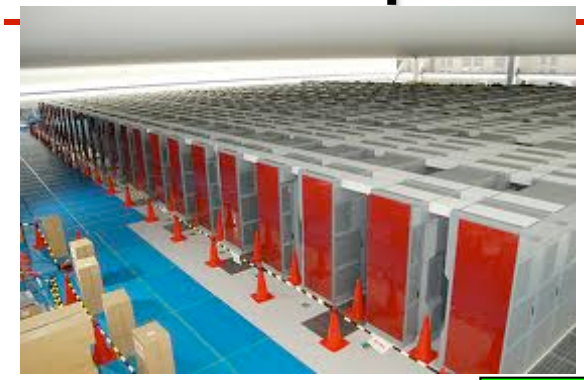


Example of typical parallel machine

Shared memory programming between processes on a board and
a combination of shared memory and distributed memory programming
between nodes and cabinets



Example of typical parallel machine



Combination of shared memory and distributed memory programming

November 2011: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak
1	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx + custom	Japan	705,024	10.5	93
2	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel + Nvidia GPU + custom	China	186,368	2.57	55
3	DOE / OS Oak Ridge Nat Lab	Jaguar, Cray AMD + custom	USA	224,162	1.76	75
4	Nat. Supercomputer Center in Shenzhen	Nebulea, Dawning Intel + Nvidia GPU + IB	China	120,640	1.27	43
5	GSIC Center, Tokyo Institute of Technology	Tusbame 2.0, HP Intel + Nvidia GPU + IB	Japan	73,278	1.19	52
6	DOE / NNSA LANL & SNL	Cielo, Cray AMD + custom	USA	142,272	1.11	81
7	NASA Ames Research Center/NAS	Plelades SGI Altix ICE 8200EX/8400EX + IB	USA	111,104	1.09	83
8	DOE / OS Lawrence Berkeley Nat Lab	Hopper, Cray AMD + custom	USA	153,408	1.054	82
9	Commissariat a l'Energie Atomique (CEA)	Tera-10, Bull Intel + IB	France	138,368	1.050	84
10	DOE / NNSA Los Alamos Nat Lab	Roadrunner, IBM AMD + Cell GPU + IB	USA	122,400	1.04	76
226	RPI	IBM eServer Blue Gene Solution		32,768	.073	80

November 2011: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx + custom	Japan	705,024	10.5	93	12.7	830
2	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel + Nvidia GPU + custom	China	186,368	2.57	55	4.04	636
3	DOE / OS Oak Ridge Nat Lab	Jaguar, Cray AMD + custom	USA	224,162	1.76	75	7.0	251
4	Nat. Supercomputer Center in Shenzhen	Nebulea, Dawning Intel + Nvidia GPU + IB	China	120,640	1.27	43	2.58	493
5	GSIC Center, Tokyo Institute of Technology	Tusbame 2.0, HP Intel + Nvidia GPU + IB	Japan	73,278	1.19	52	1.40	865
6	DOE / NNSA LANL & SNL	Cielo, Cray AMD + custom	USA	142,272	1.11	81	3.98	279
7	NASA Ames Research Center/NAS	Plelades SGI Altix ICE 8200EX/8400EX + IB	USA	111,104	1.09	83	4.10	265
8	DOE / OS Lawrence Berkeley Nat Lab	Hopper, Cray AMD + custom	USA	153,408	1.054	82	2.91	362
9	Commissariat a l'Energie Atomique (CEA)	Tera-10, Bull Intel + IB	France	138,368	1.050	84	4.59	229
10	DOE / NNSA Los Alamos Nat Lab	Roadrunner, IBM AMD + Cell GPU + IB	USA	122,400	1.04	76	2.35	446
500	IT Service	IBM Cluster, Intel + GigE	USA	7,236	.051	53		

Japanese K Computer

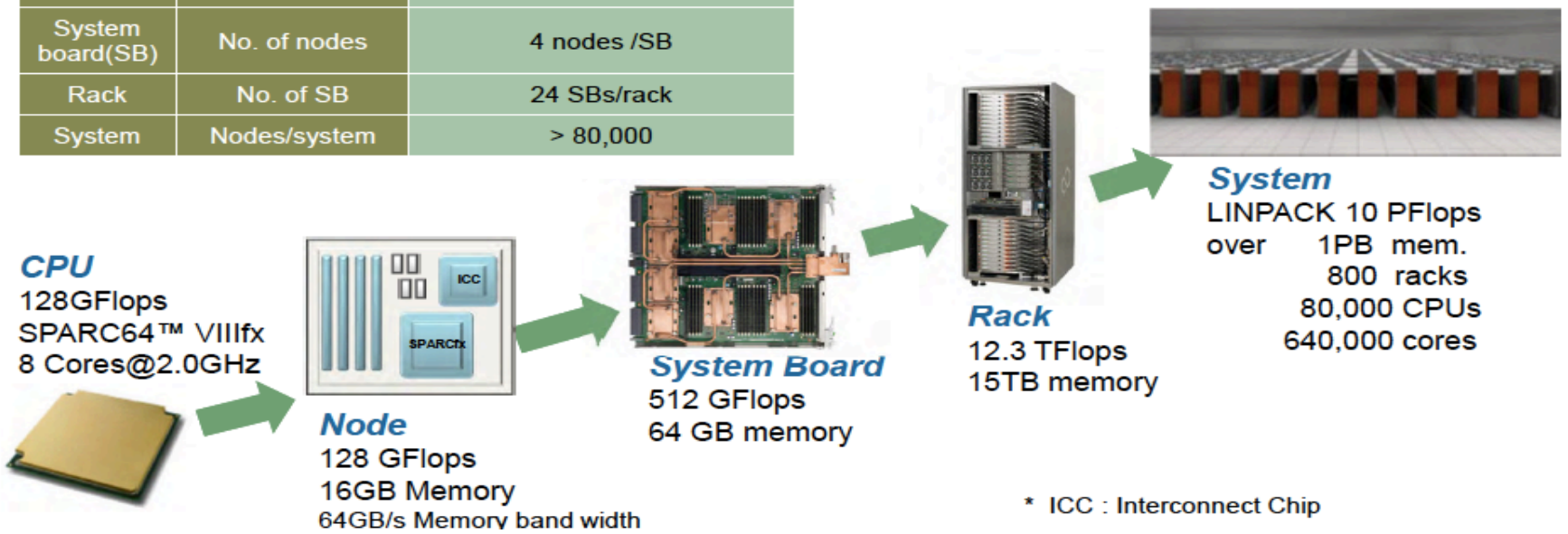
K Computer > Sum(#2 : #8)
~ 2.5X #2

K computer Specifications

FUJITSU

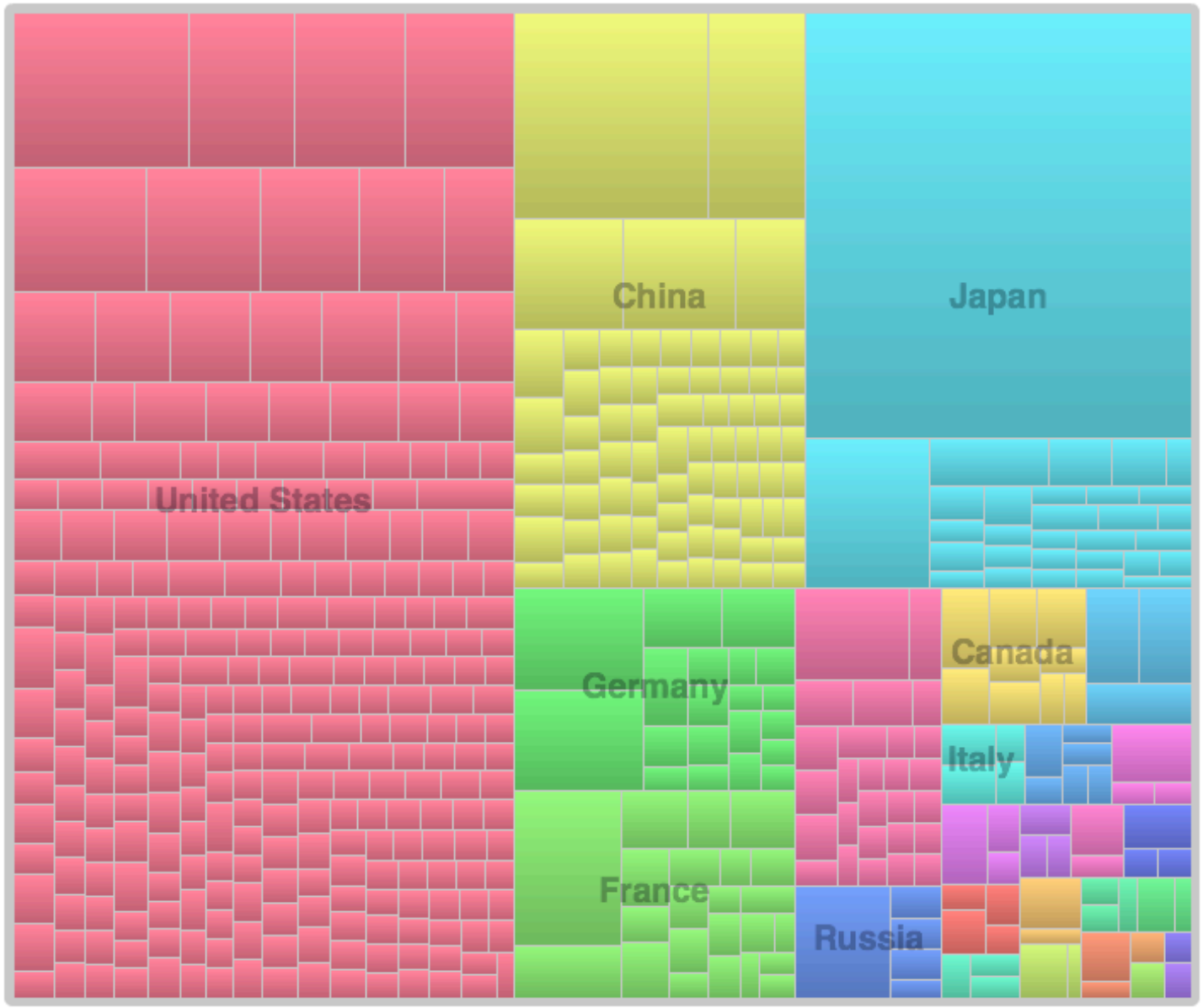
CPU (SPARC64 VIIIfx)	Cores/Node	8 cores (@2GHz)
	Performance	128GFlops
	Architecture	SPARC V9 + HPC extension
	Cache	L1(I/D) Cache : 32KB/32KB L2 Cache : 6MB
	Power	58W (typ. 30 C)
	Mem. bandwidth	64GB/s.
Node	Configuration	1 CPU / Node
	Memory capacity	16GB (2GB/core)
System board(SB)	No. of nodes	4 nodes /SB
Rack	No. of SB	24 SBs/rack
System	Nodes/system	> 80,000

Inter-connect	Topology	6D Mesh/Torus
	Performance	5GB/s. for each link
	No. of link	10 links/ node
	Additional feature	H/W barrier, reduction
	Architecture	Routing chip structure (no outside switch box)
Cooling	CPU, ICC*	Direct water cooling
	Other parts	Air cooling



07 Linpack run with 705,024 cores at 10.51 Pflop/s (88,128 CPUs), 12.7 MW; 29.5 hours
Fujitsu to have a 100 Pflop/s system in 2014

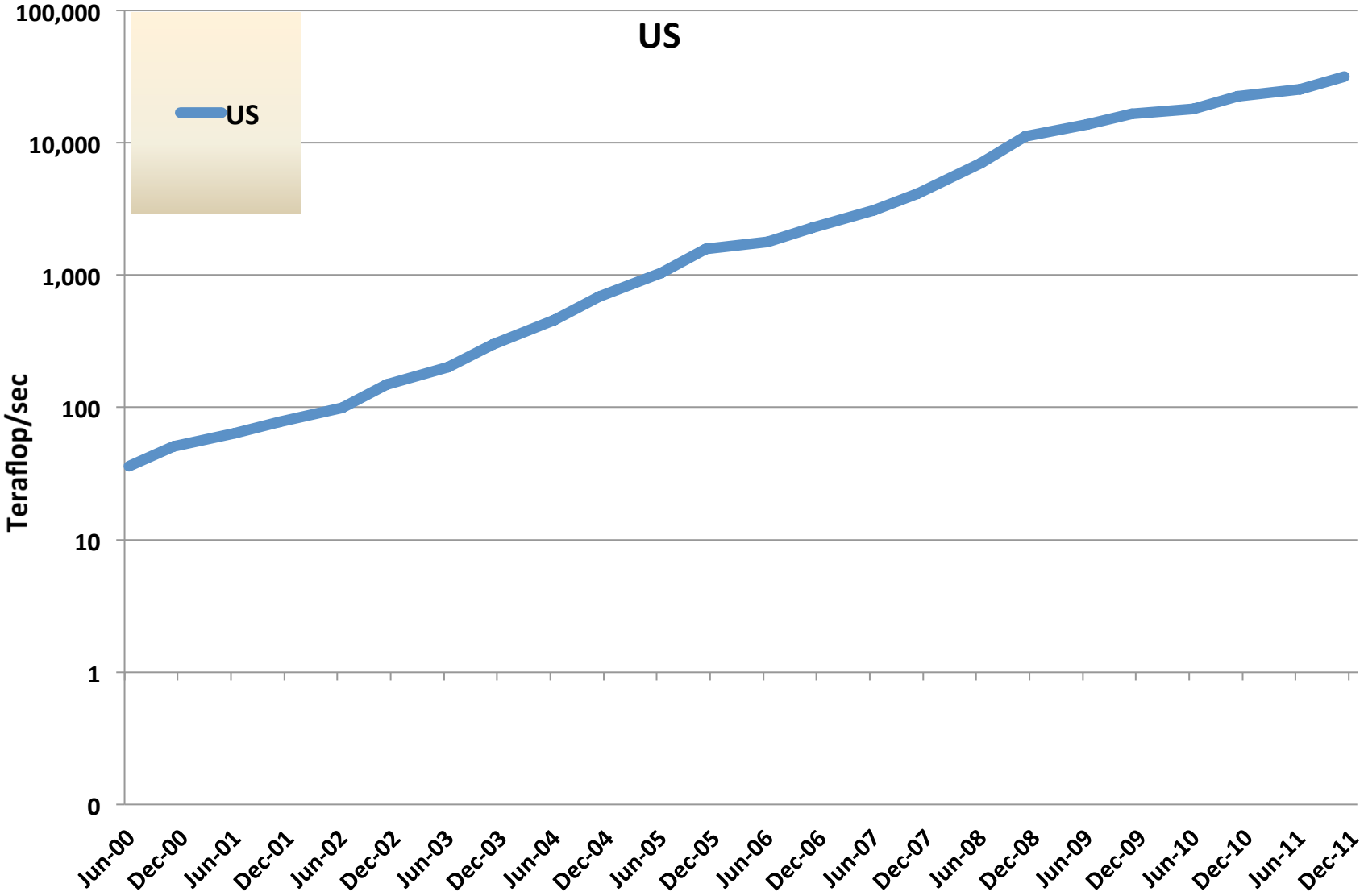
Countries Share



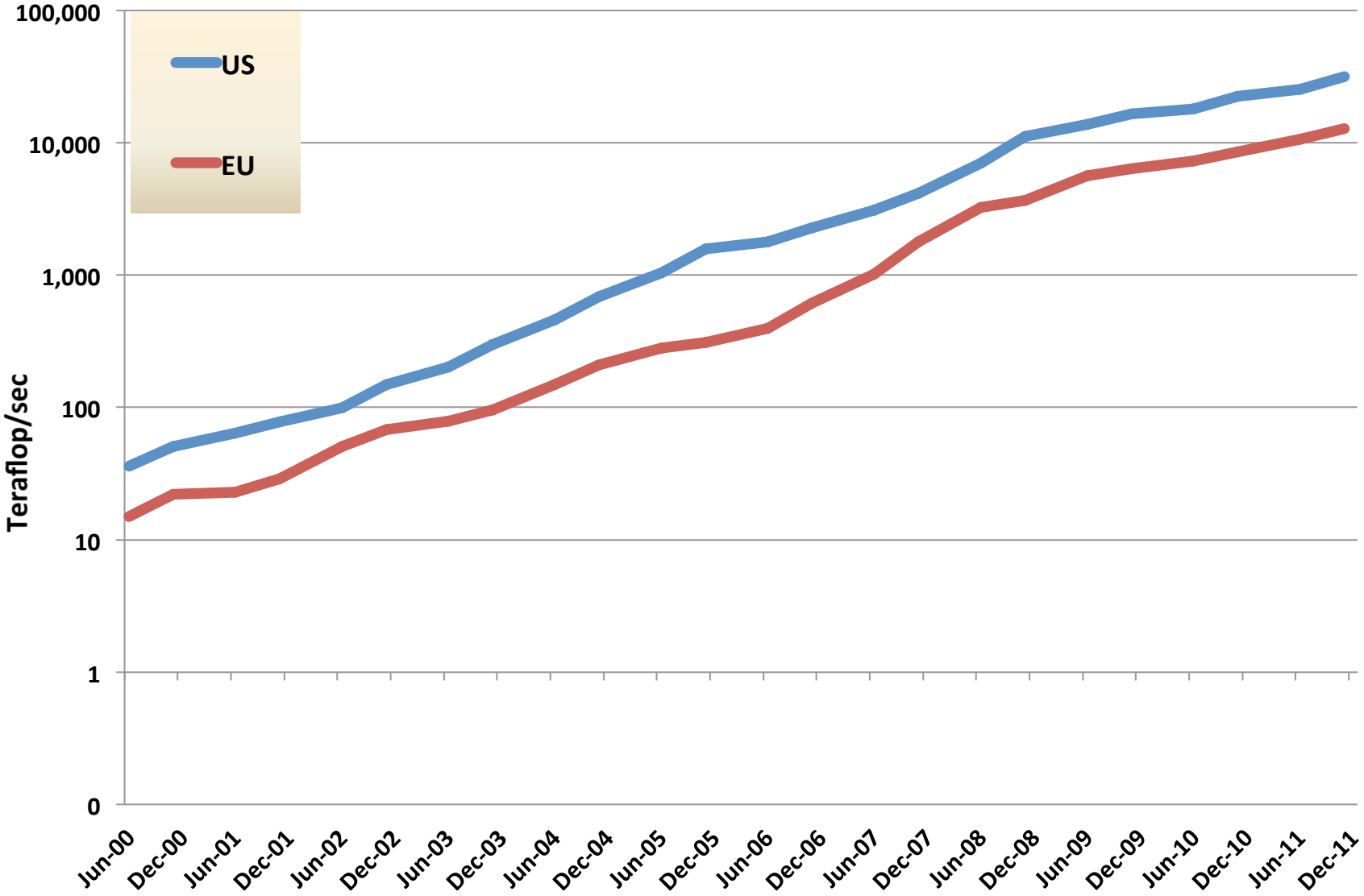
Absolute Counts

US:	263
China:	75
Japan:	30
UK:	27
France:	23
Germany:	20

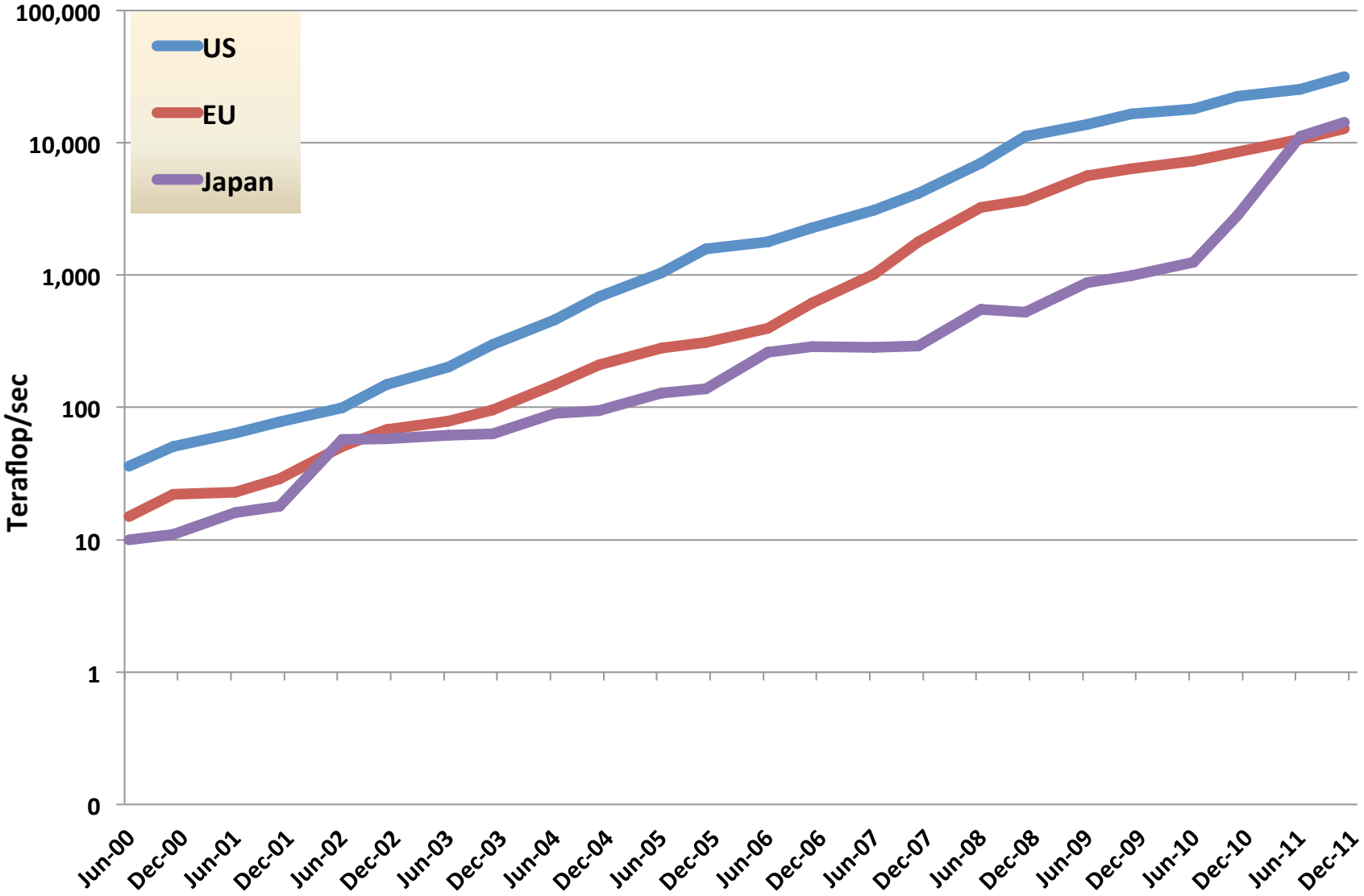
Performance of Countries



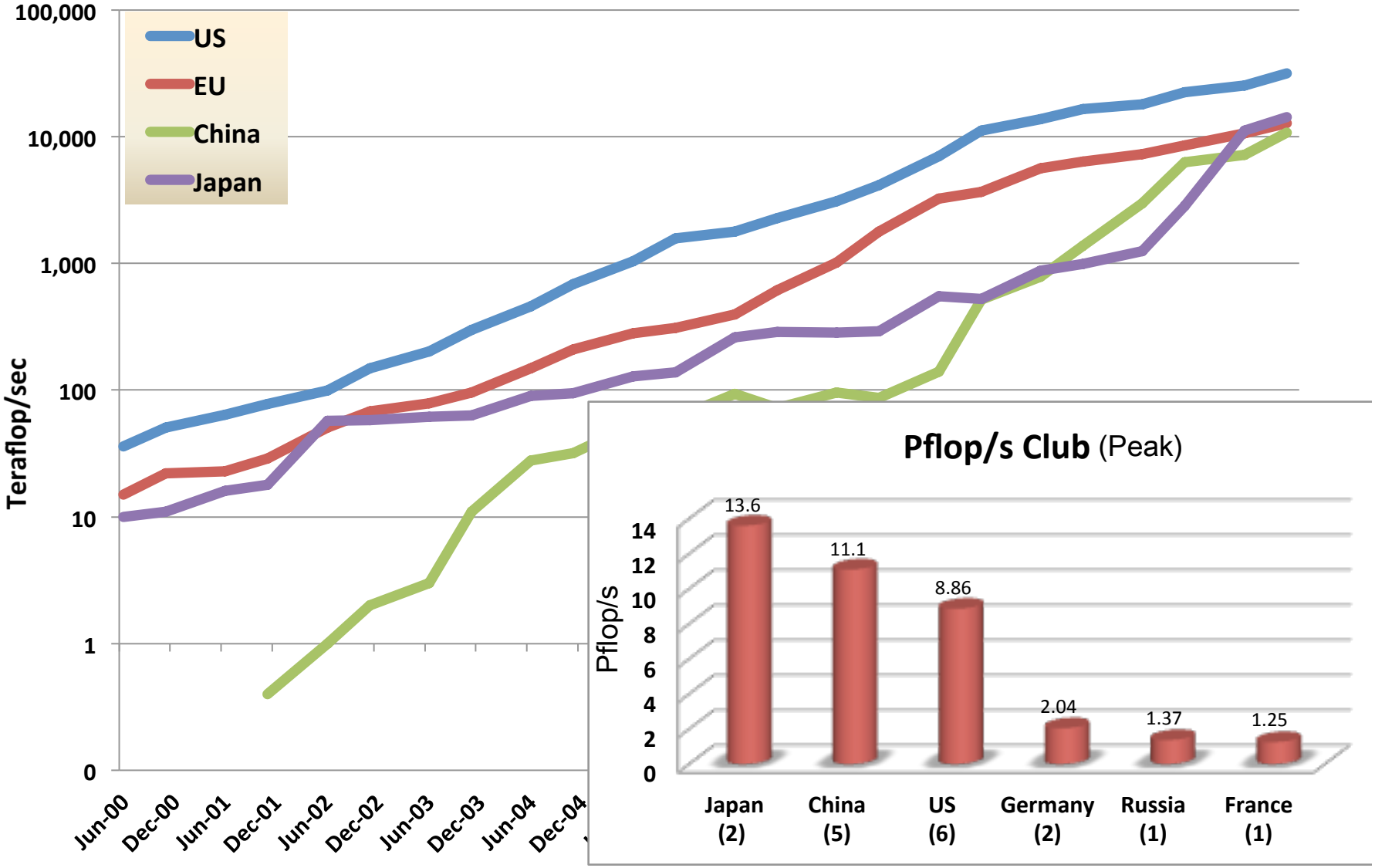
Performance of Countries



Performance of Countries



Performance of Countries



Russian Supercomputers

Rank	Name	Computer	Site	Manufacturer	Total Cores	Rmax	Rpeak
18	Lomonosov	T-Platforms T-Blade2/1.1, Xeon X5570/X5670 2.93 GHz, Nvidia 2070 GPU, Infiniband QDR	Moscow State University	T-Platforms	33072	674105	1373060
108	MVS-100K	Cluster Platform 3000 BL460c/BL2x220, Xeon 54xx 3 Ghz, Infiniband	Joint Supercomputer Center	HP	11680	107448	140160
120		Cluster Platform 3000 BL2x220, E54xx 3.0 Ghz, Infiniband	Kurchatov Institute Moscow	HP	10304	101213	123648
122	SKIF Aurora	SKIF Aurora Platform - Intel Xeon X5680, Infiniband QDR	South Ural State University	RSC SKIF	8832	100400	117000
339		HP DL160 Cluster G6, Xeon E5645 6C 2.40 GHz, Gigabit Ethernet	Web Content Provider	HP	12024	59903	115430

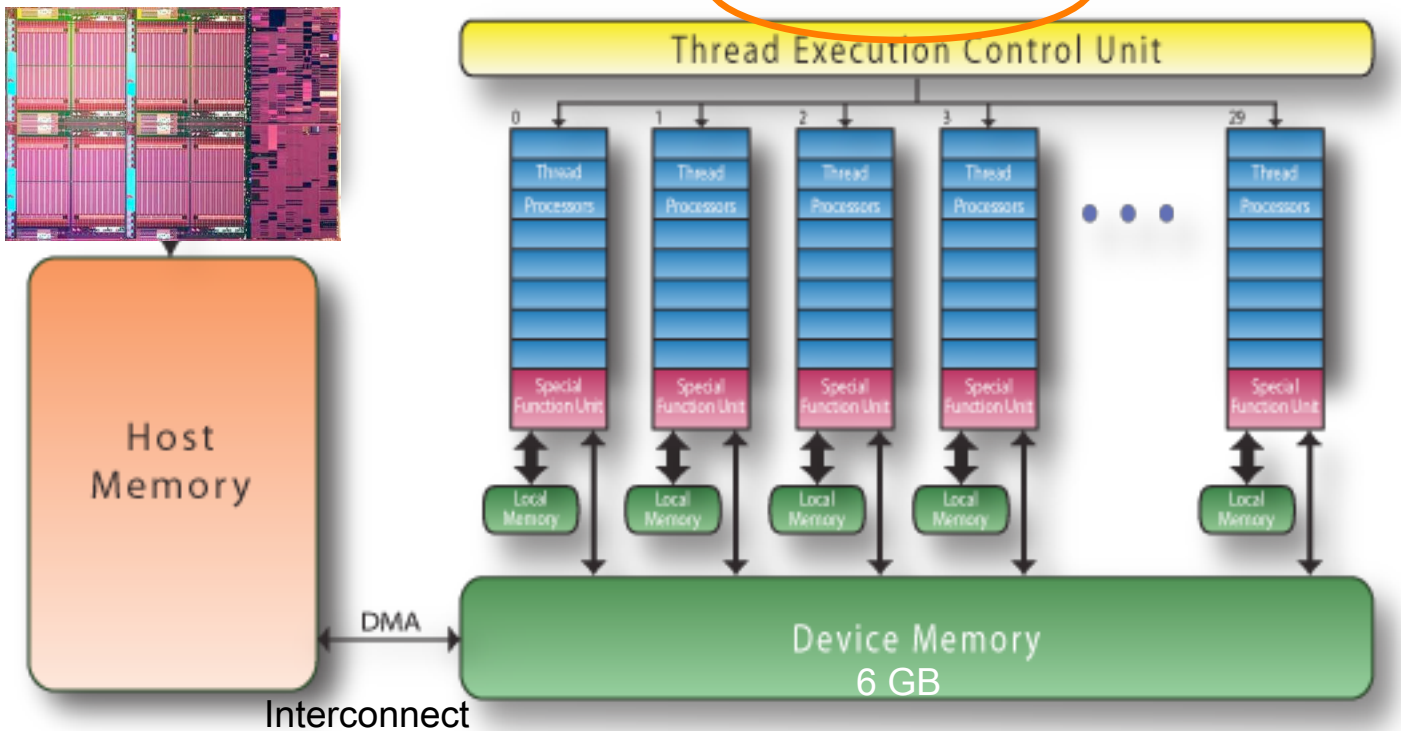
Commodity plus Accelerator

Commodity

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

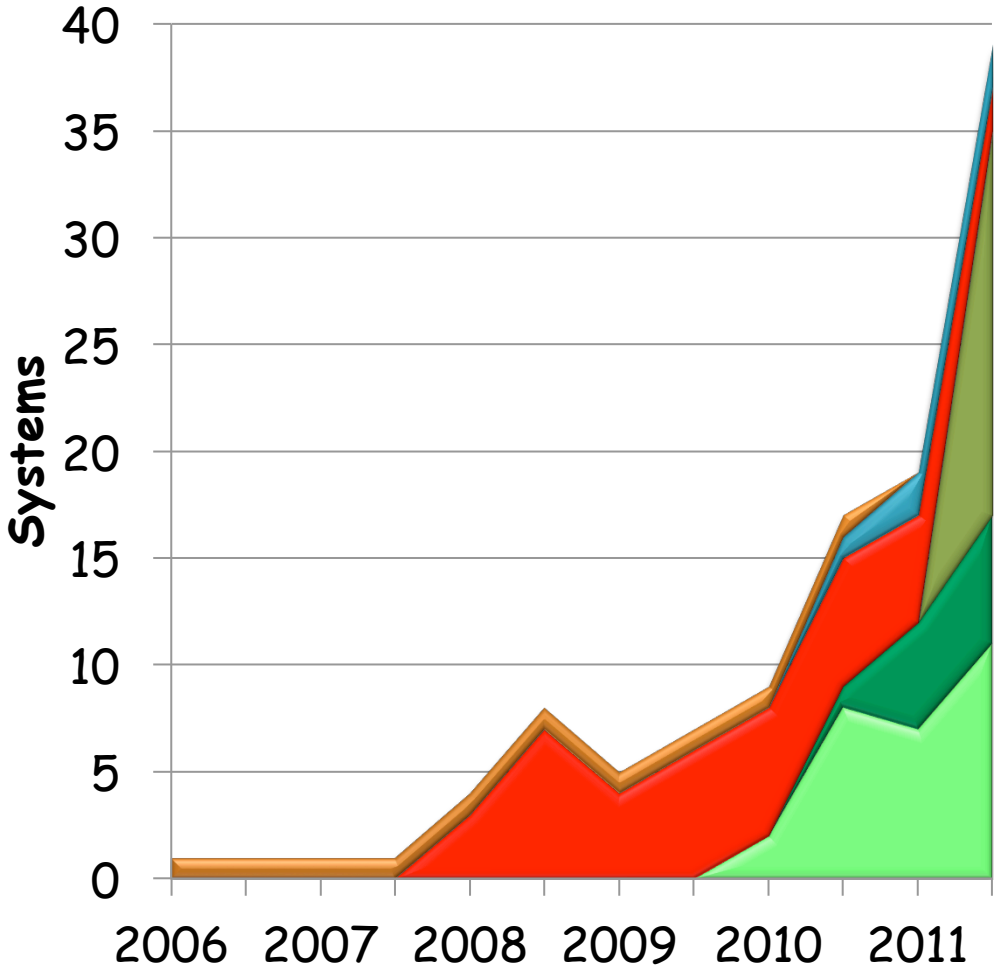
Accelerator (GPU)

Nvidia C2070 “Fermi”
448 “Cuda cores”
1.15 GHz
~~448 ops/cycle~~
515 Gflop/s (DP)



Interconnect
PCI-X 16 lane
64 Gb/s
1 GW/s

39 Accelerator Based Systems

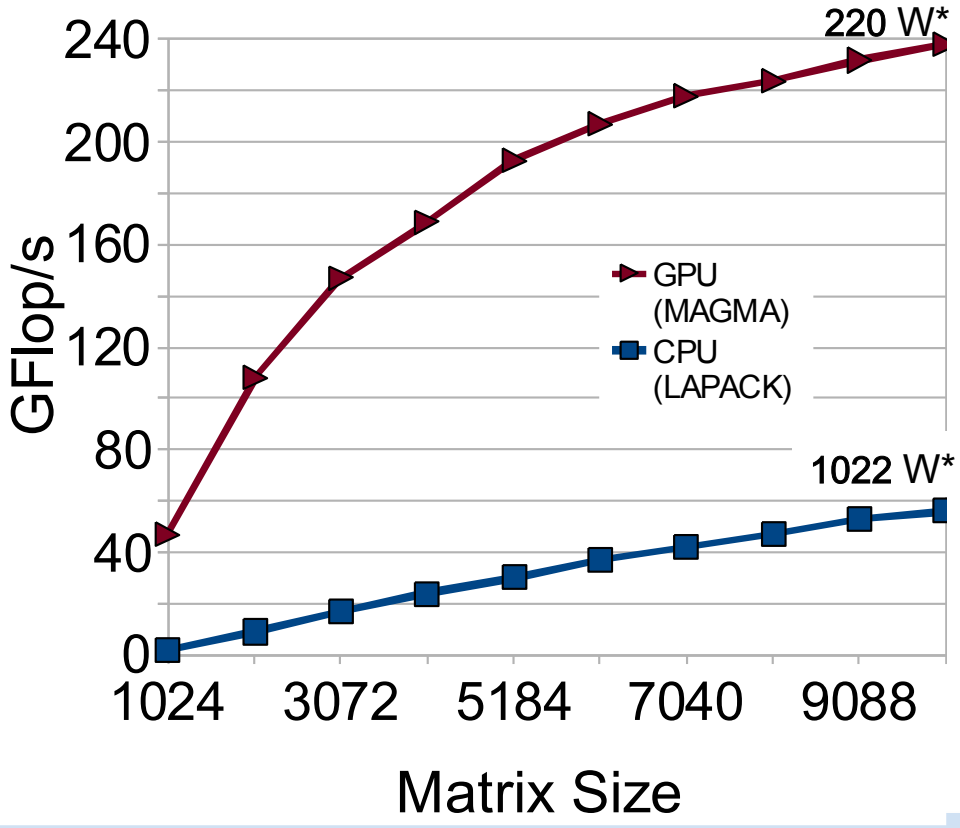


- Clearspeed CSX60022
- ATI GPU
- IBM PowerXCell 8i
- NVIDIA 2090
- NVIDIA 2070
- NVIDIA 2050

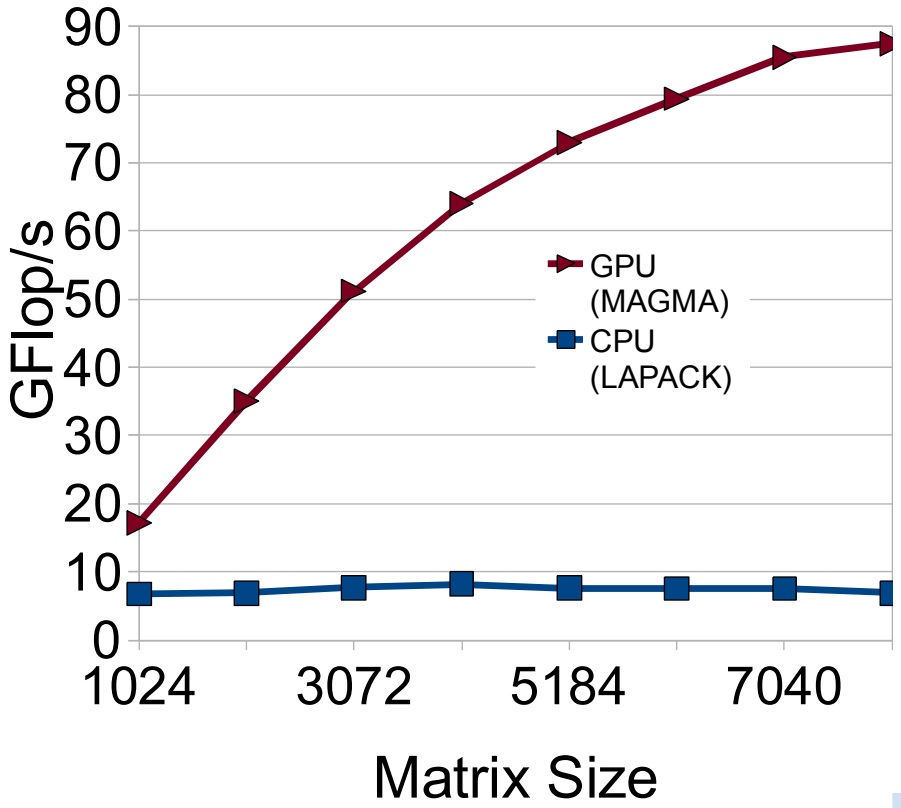
20 US	1 Italy
5 China	1 Poland
3 Japan	1 Spain
2 France	1 Switzerland
2 Germany	1 Russia
1 Australia	1 Taiwan

Accelerating Dense Linear Algebra with GPUs

LU Factorization in double precision (DP)
[for solving a dense linear system]



Hessenberg factorization in DP
[for the general eigenvalue problem]

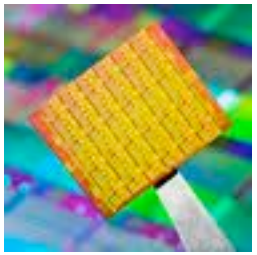


GPU Fermi C2050 [448 CUDA Cores @ 1.15 GHz] + Intel Q9300 [4 cores @ 2.50 GHz]			**CPU** AMD ISTANBUL [8 sockets x 6 cores (48 cores) @2.8GHz]
DP peak	**515 + 40** GFlop/s	~1/10 price	DP peak **538** GFlop/s
System cost	~ $**3,000**	~1/5 power	System cost ~ $**30,000**
Power *	~ **220** W	~4 X performance	Power * ~ **1,022** W

* Computation consumed power rate (total system rate minus idle rate), measured with *KILL A WATT PS, Model P430*

Future Computer Systems

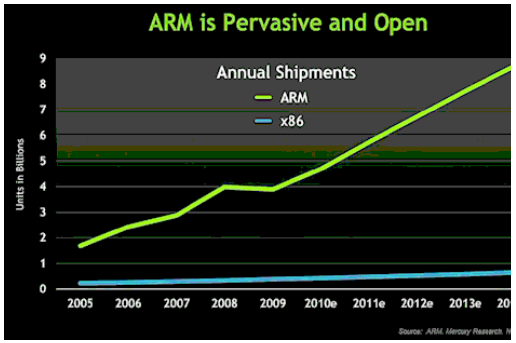
- Most likely be a hybrid design
  - Think standard multicore chips and accelerator (GPUs)
- Today accelerators are attached
- Next generation more integrated
- Intel's MIC architecture "Knights Ferry" and "Knights Corner" to come.
  - 48 x86 cores



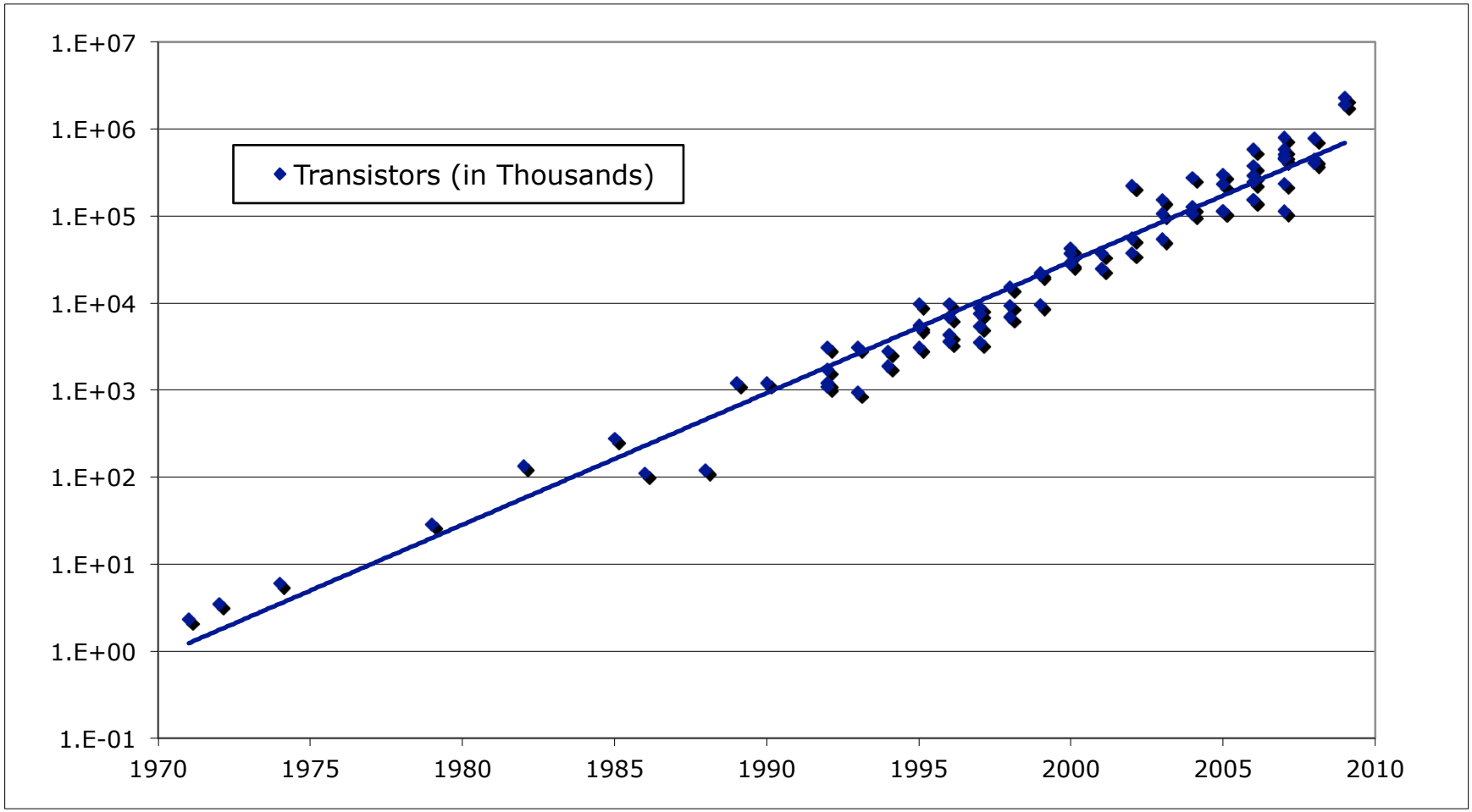
- AMD's Fusion
  - Multicore with embedded graphics ATI

AMD
The future is fusion

- Nvidia's Project Denver plans to develop an integrated chip using ARM architecture in 2013.

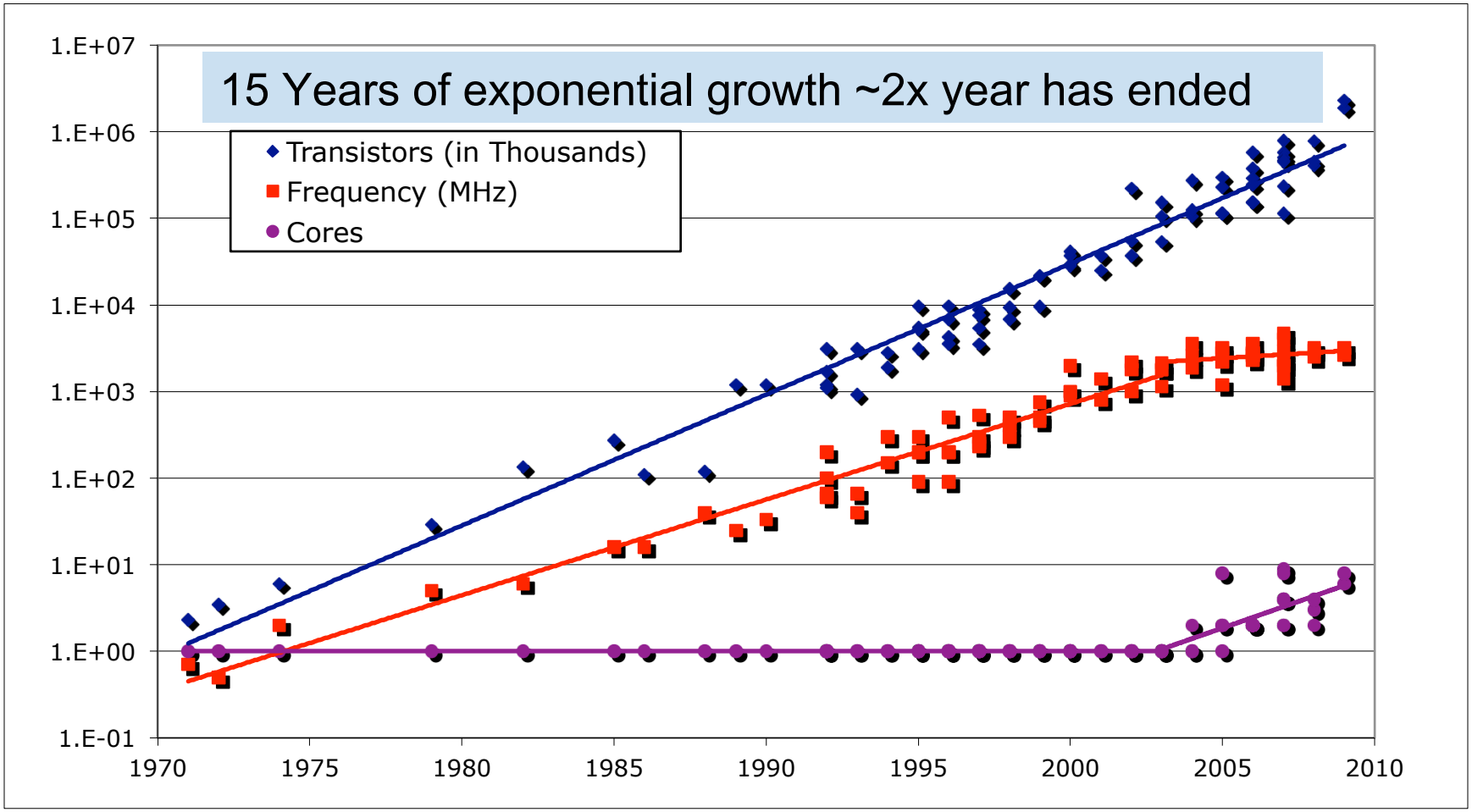


Moore's Law is Alive and Well



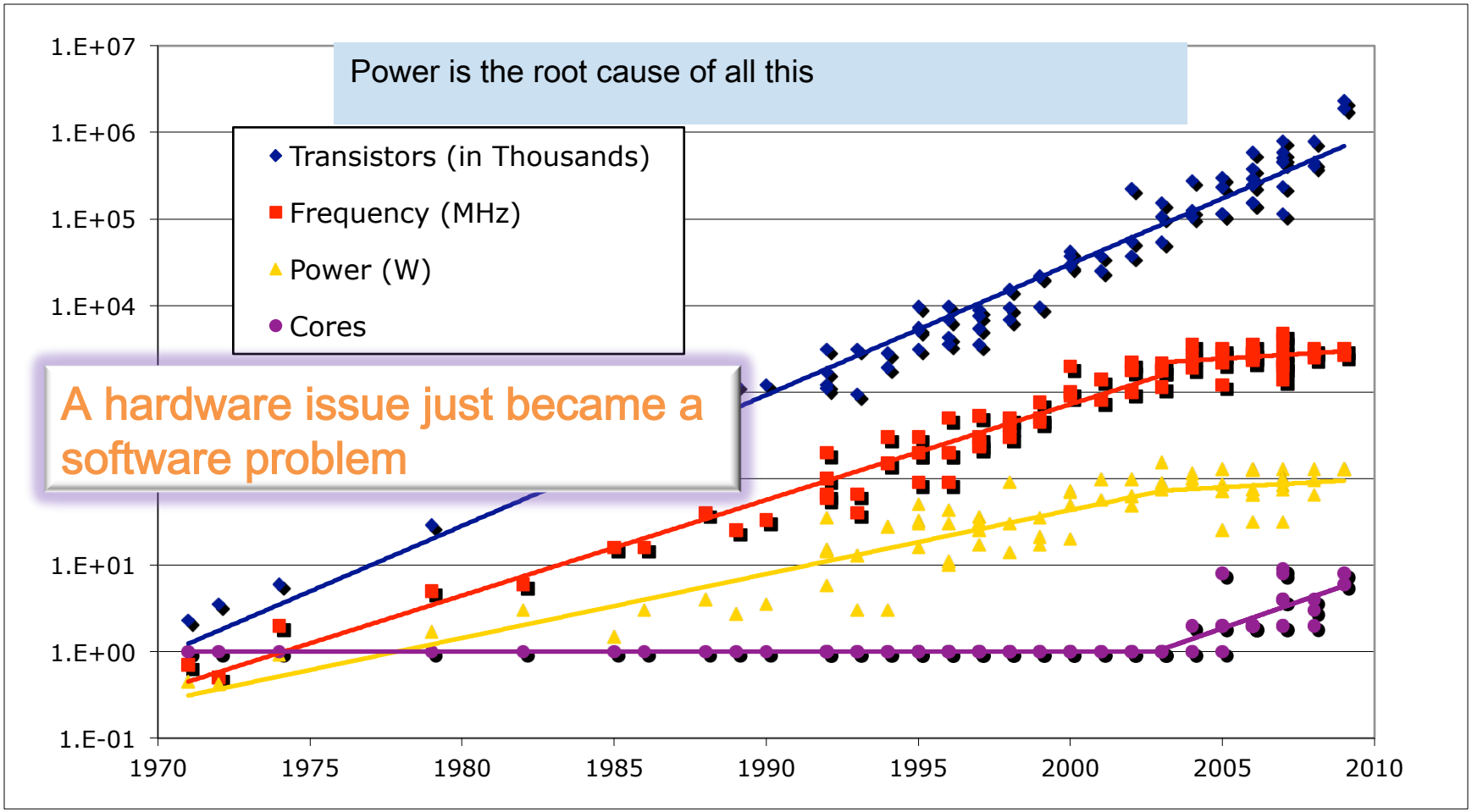
Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç
Slide from Kathy Yelick

But Clock Frequency Scaling Replaced by Scaling Cores / Chip



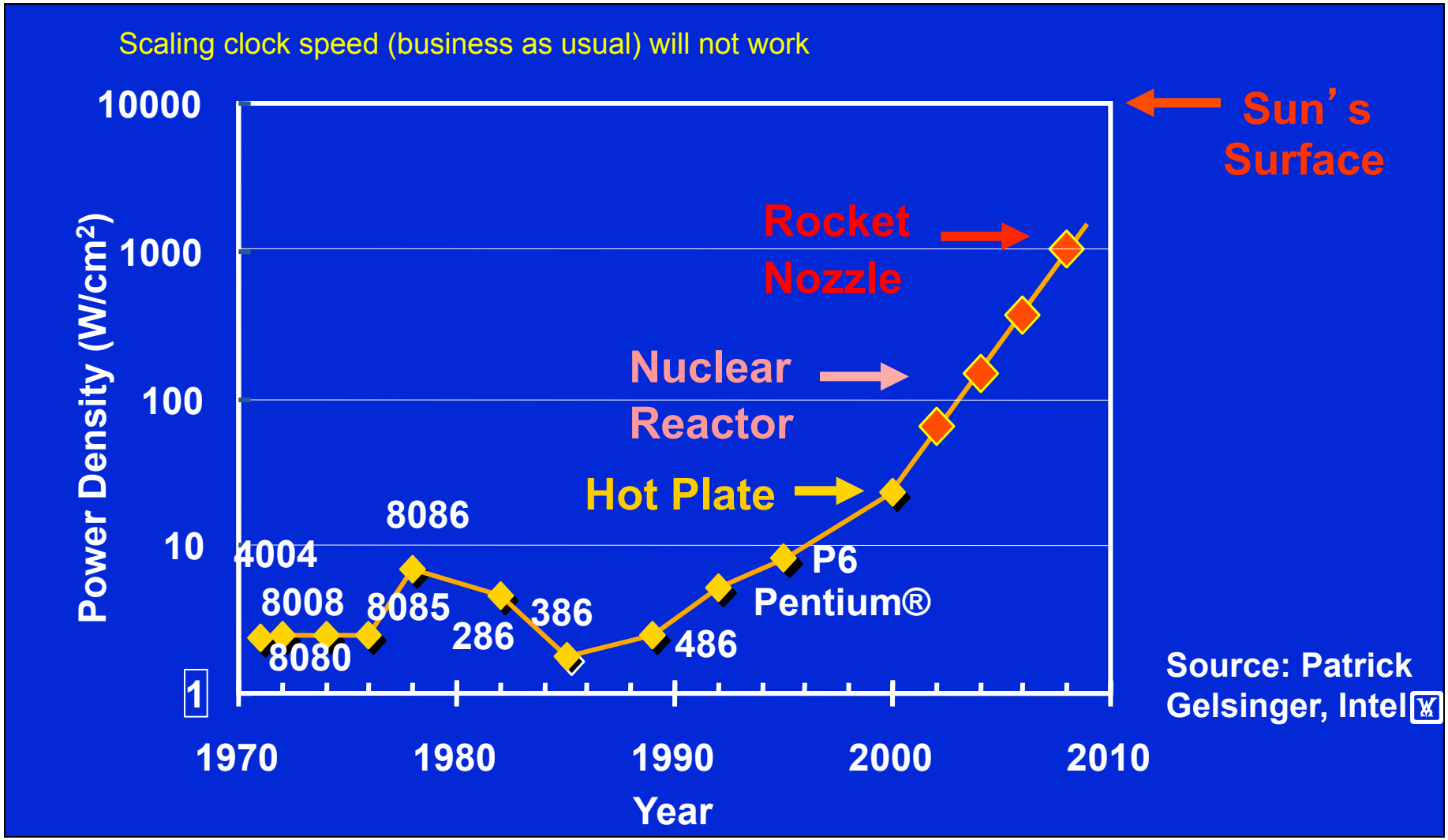
Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç
Slide from Kathy Yelick

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

Clock Scaling Hits Power Density Wall



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)

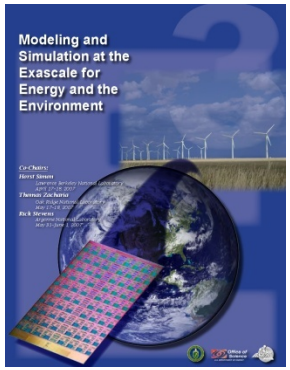
50% more performance with 20% less power

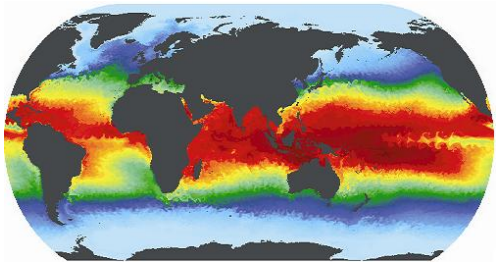
Preferable to use multiple slower devices, than one superfast device

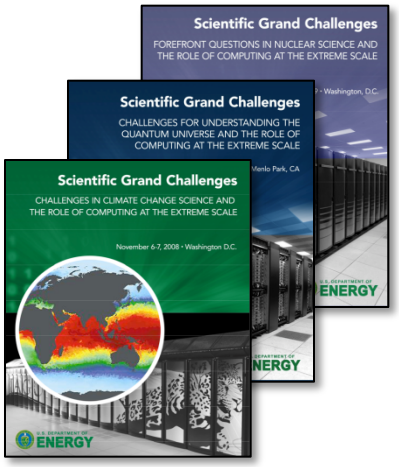
Broad Community Support and Development of the Exascale Initiative Since 2007

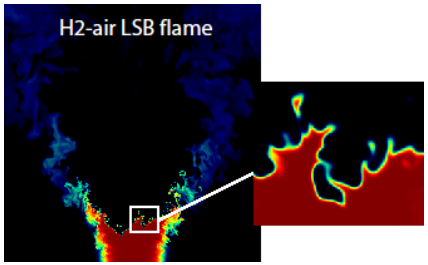
http://science.energy.gov/ascr/news-and-resources/program-documents/

- **Town Hall Meetings April-June 2007**
- **Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009**
  - Climate Science (11/08)
  - High Energy Physics (12/08)
  - Nuclear Physics (1/09)
  - Fusion Energy (3/09)
  - Nuclear Energy (5/09)
  - Biology (8/09)
  - Material Science and Chemistry (8/09)
  - National Security (10/09)
  - Cross-cutting technologies (2/10)
- **Exascale Steering Committee**
  - "Denver" vendor NDA visits (8/09)
  - SC09 vendor feedback meetings
  - Extreme Architecture and Technology Workshop (12/09)
- **International Exascale Software Project**
  - Santa Fe, NM (4/09); Paris, France (6/09); Tsukuba, Japan (10/09); Oxford (4/10); Maui (10/10); San Francisco (4/11); Cologne (10/11)

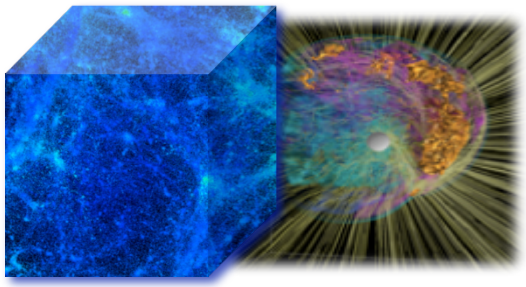






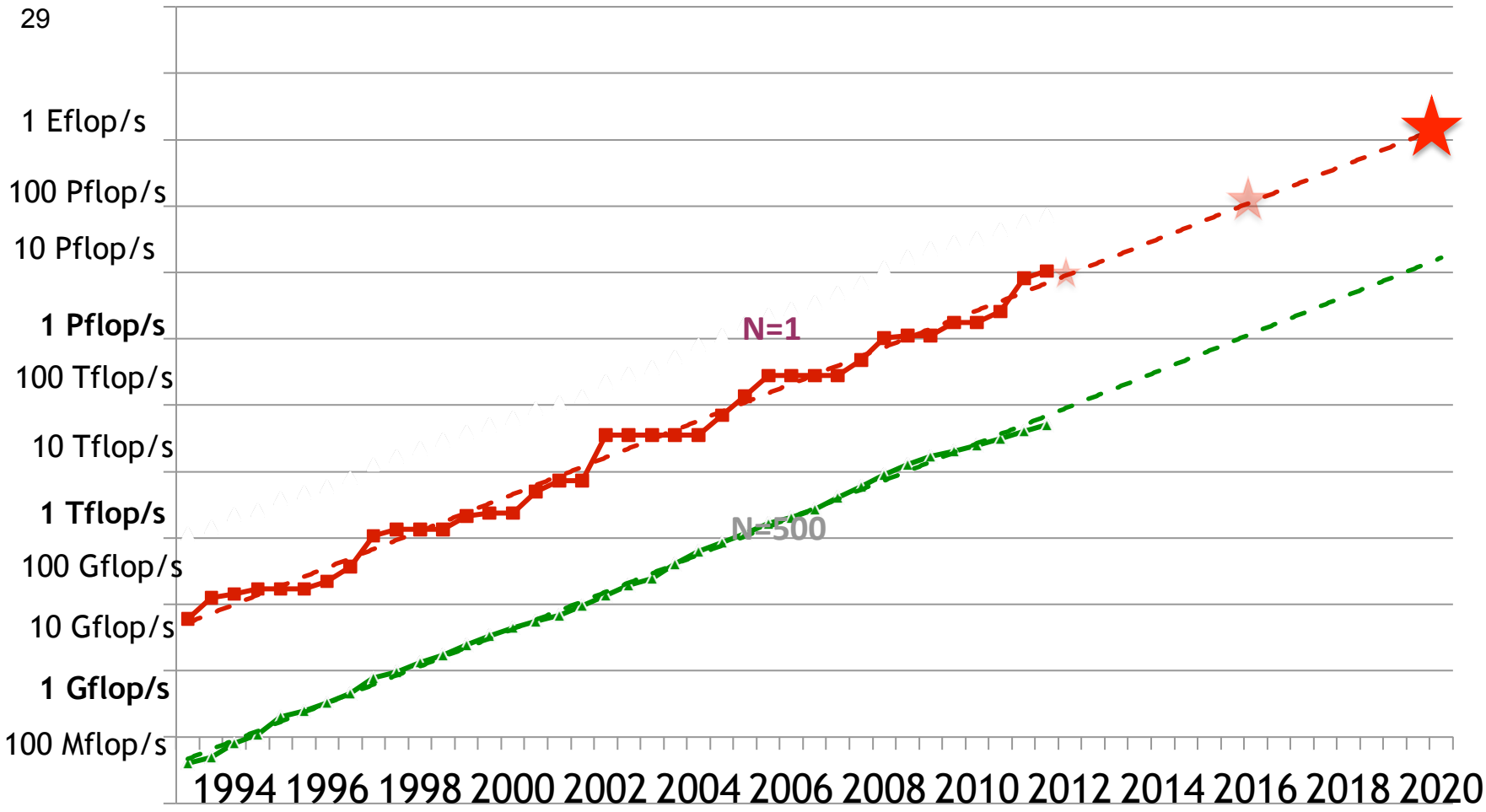


Mission Imperatives



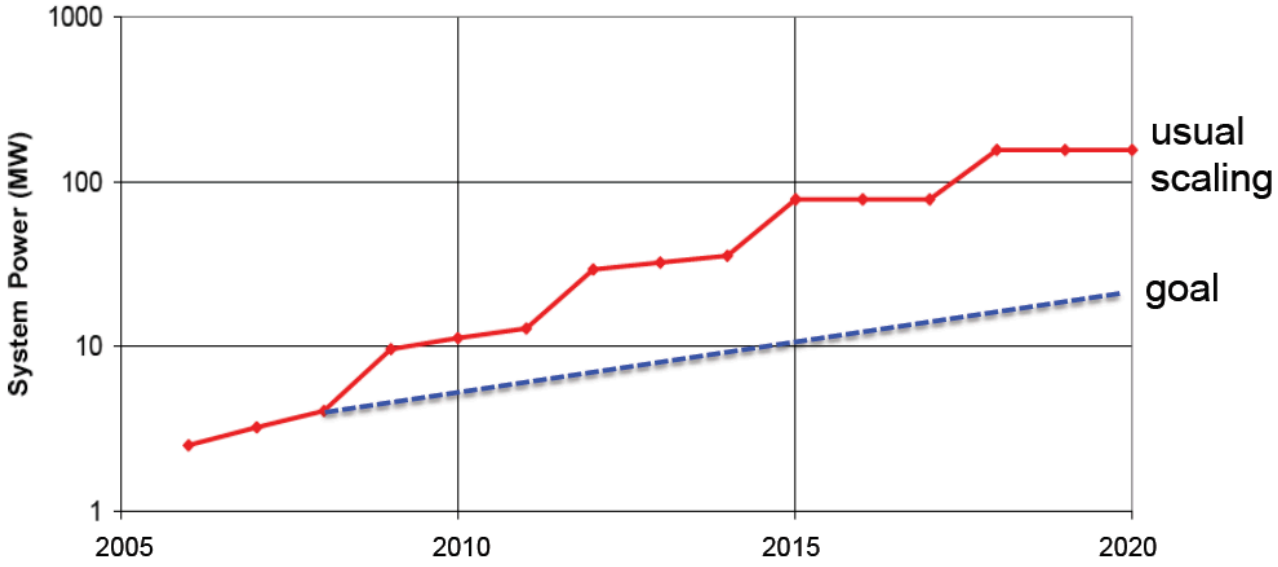
Fundamental Science

Performance Development in Top500



Energy Cost Challenge

- At ~$1M per MW energy costs are substantial
  - 10 Pflop/s in 2011 uses ~10 MWs
  - 1 Eflop/s in 2018 > 100 MWs



  - DOE Target: 1 Eflop/s in 2018 at 20

The High Cost of Data Movement

- Flop/s or percentage of peak flop/s become much less relevant

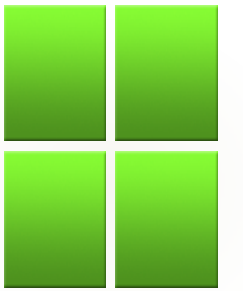
Approximate power costs (in picoJoules)

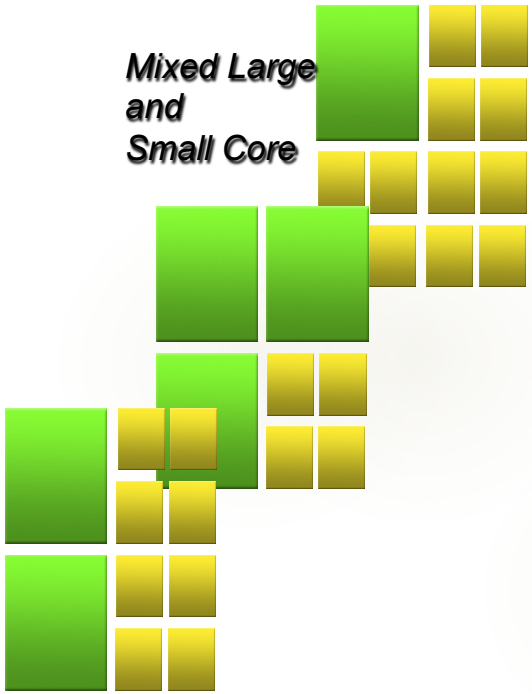
	2011
DP FMADD flop	100 pJ
DP DRAM read	4800 pJ
Local Interconnect	7500 pJ
Cross System	9000 pJ

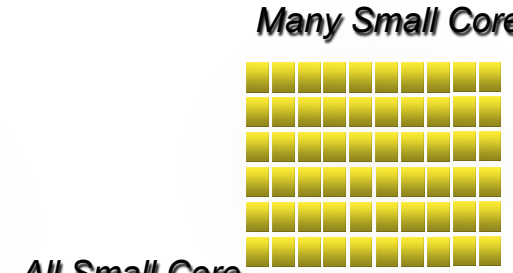
Source: John Shalf, LBNL

- Algorithms & Software: minimize data movement; perform more work per unit data movement.

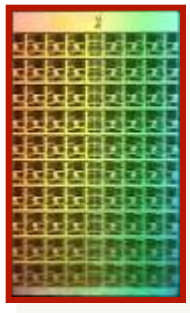
What's Next?

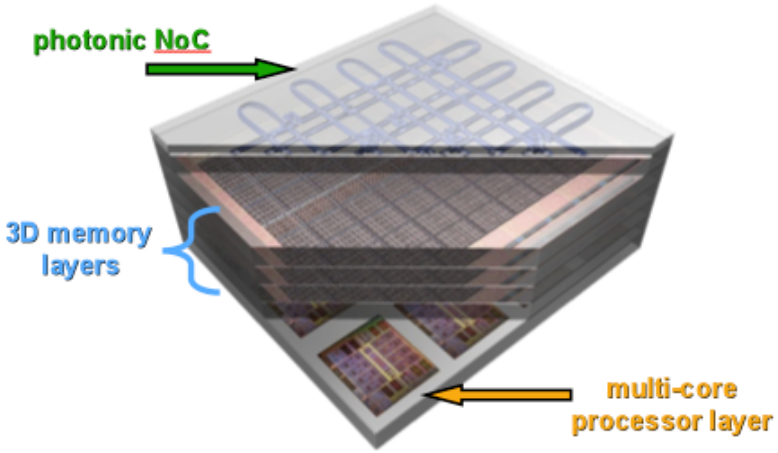






Many Floating-Point Cores





Different Classes of Chips

- Home
- Games / Graphics
- Business
- Scientific

Potential System Architecture

Systems	2011 K computer
System peak	**10.5 Pflop/s**
Power	**12.7 MW**
System memory	1.6 PB
Node performance	128 GF
Node memory BW	64 GB/s
Node concurrency	8
Total Node Interconnect BW	20 GB/s
System size (nodes)	88,124
Total concurrency	705,024
MTTI	days

Potential System Architecture with a cap of $200M and 20MW

Systems	2011 K computer	2019	Difference Today & 2019
System peak	**10.5 Pflop/s**	**1 Eflop/s**	**O(100)**
Power	**12.7 MW**	**~20 MW**	
System memory	1.6 PB	32 - 64 PB	O(10)
Node performance	128 GF	1,2 or 15TF	O(10) – O(100)
Node memory BW	64 GB/s	2 - 4TB/s	O(100)
Node concurrency	8	O(1k) or 10k	O(100) – O(1000)
Total Node Interconnect BW	20 GB/s	200-400GB/s	O(10)
System size (nodes)	88,124	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	705,024	O(billion)	O(1,000)
MTTI	days	O(1 day)	- O(10)

Major Changes to Software & Algorithms

- **Must rethink the design of our algorithms and software**
  - **Another disruptive technology**
    - Similar to what happened with cluster computing and message passing
  - **Rethink and rewrite the applications, algorithms, and software**
  - **Data movement is expense**
  - **Flop/s are cheap, so are provisioned in excess**

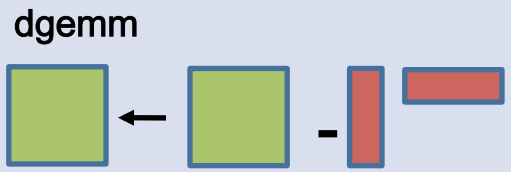
Critical Issues at Peta & Exascale for Algorithm and Software Design

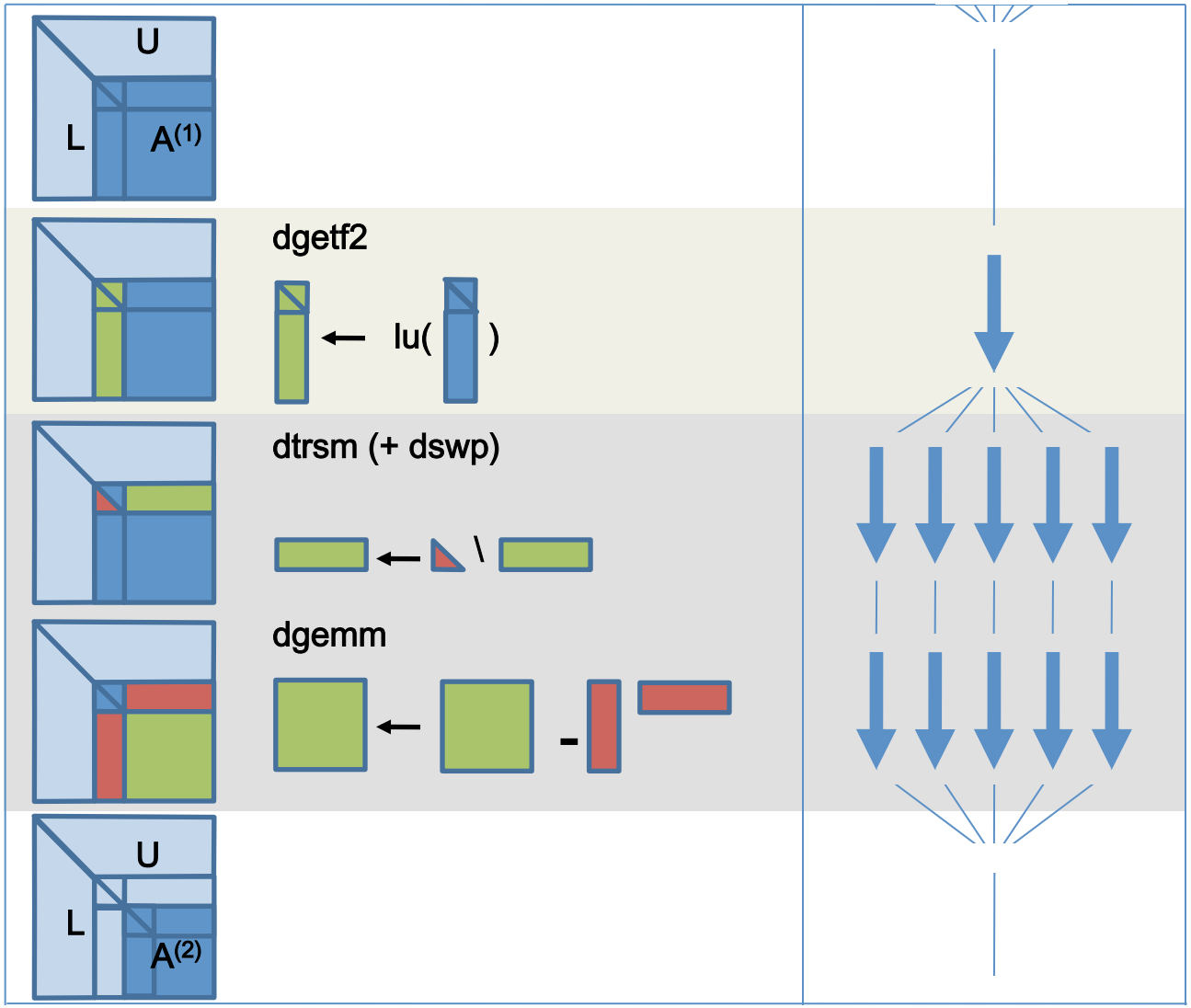
- Synchronization-reducing algorithms
  - Break Fork-Join model
- Communication-reducing algorithms
  - Use methods which have lower bound on communication
- Mixed precision methods
  - 2x speed of ops and 2x speed for data movement
- Autotuning
  - Today's machines are too complicated, build "smarts" into software to adapt to the hardware
- Fault resilient algorithms
  - Implement algorithms that can recover from failures/bit flips
- Reproducibility of results
  - Today we can't guarantee this. We understand the issues, but some of our "colleagues" have a hard time with this.

Fork-Join Parallelization of LU and QR.

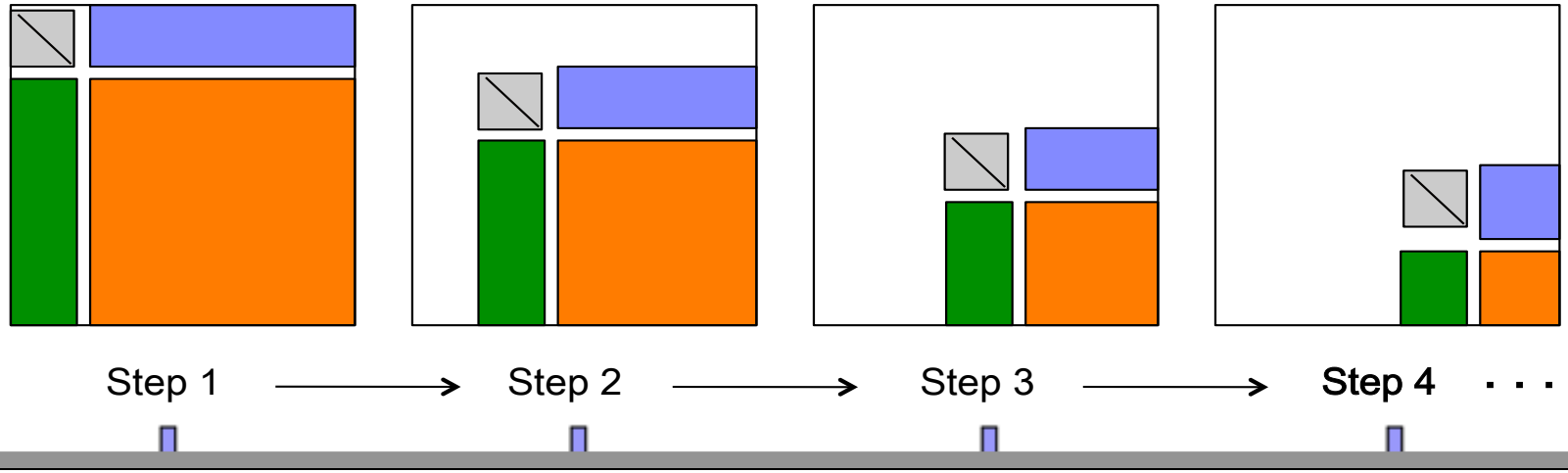
Parallelize the update:

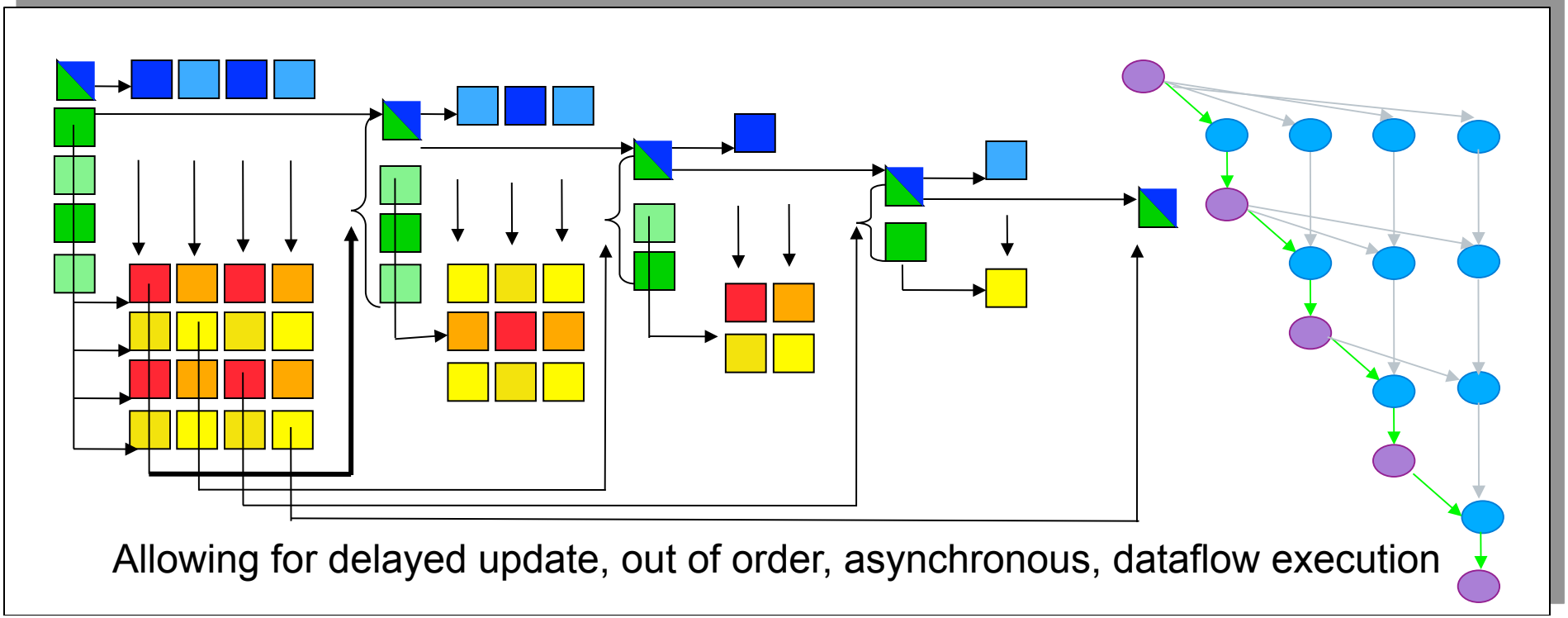
- Easy and done in any reasonable software.
- This is the $2/3n^3$ term in the FLOPs count.
- Can be done efficiently with LAPACK+multithreaded BLAS





1. Synchronization (in LAPACK LU)

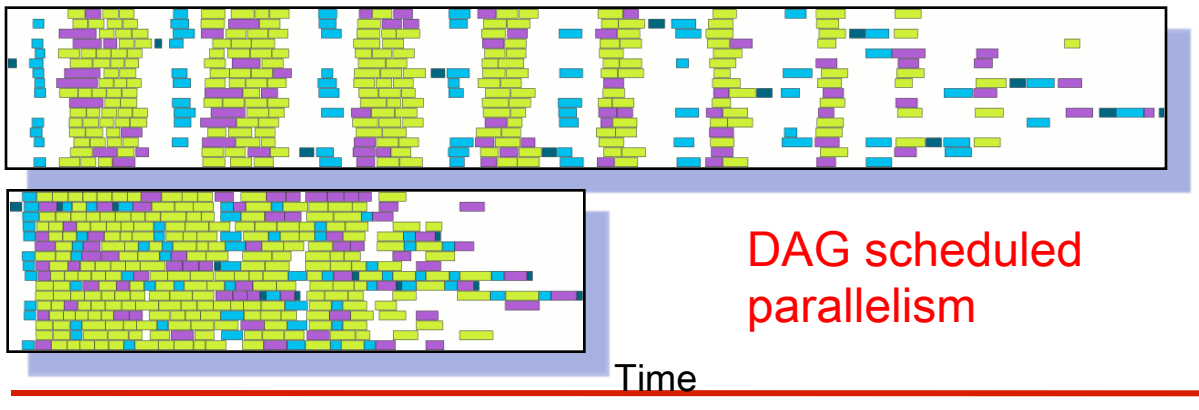


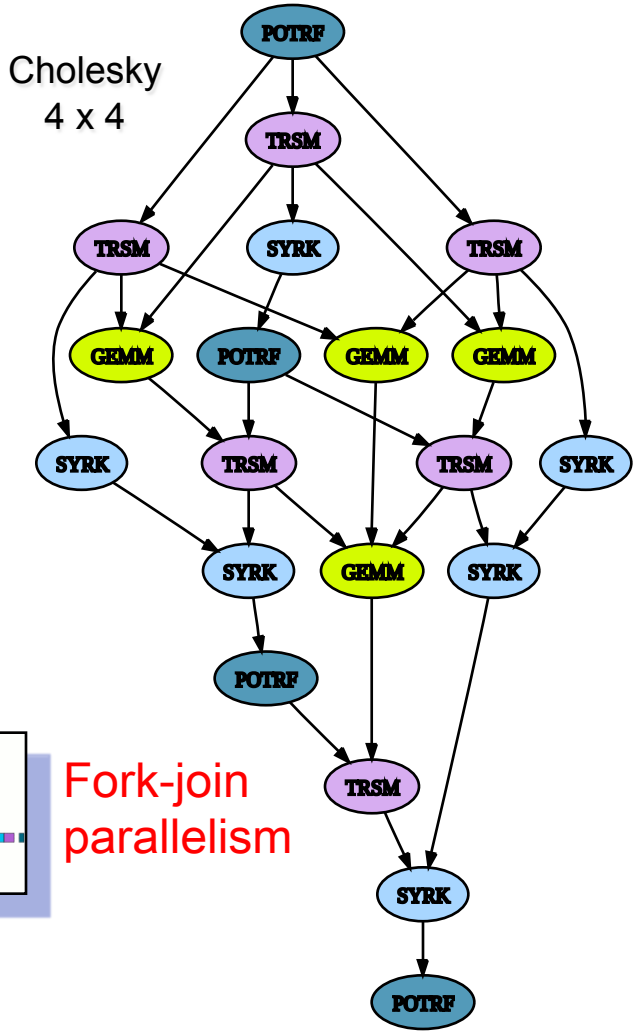


Allowing for delayed update, out of order, asynchronous, dataflow execution

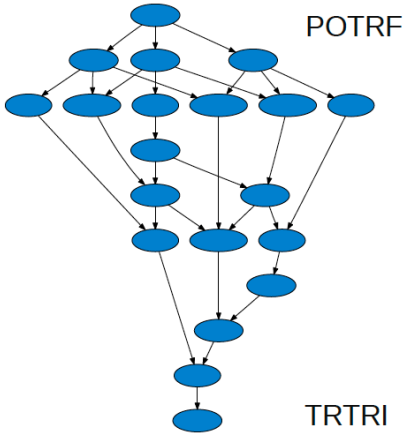
PLASMA/MAGMA: Parallel Linear Algebra s/w for Multicore/Hybrid Architectures

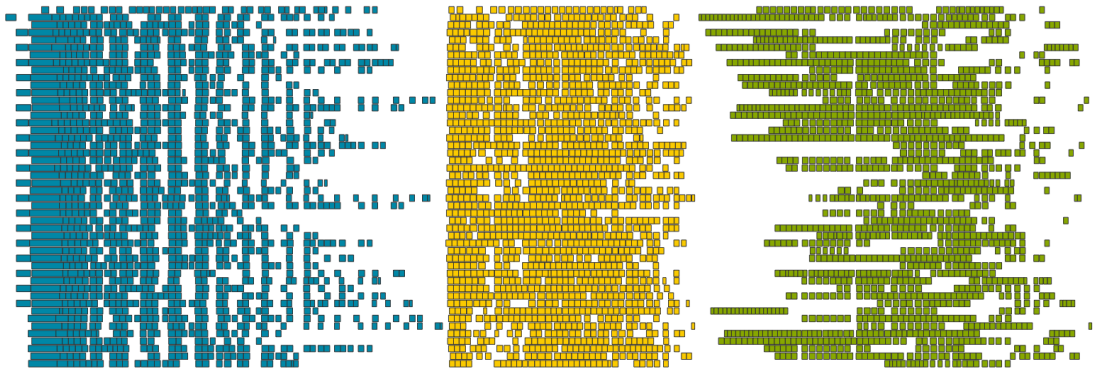
- **Objectives**
  - High utilization of each core
  - Scaling to large number of cores
  - Synchronization reducing algorithms
- **Methodology**
  - Dynamic DAG scheduling (QUARK)
  - Explicit parallelism
  - Implicit communication
  - Fine granularity / block data layout
- **Arbitrary DAG with dynamic scheduling**

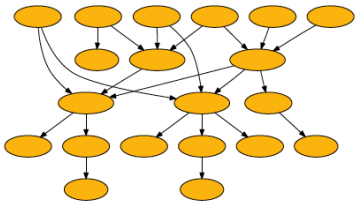


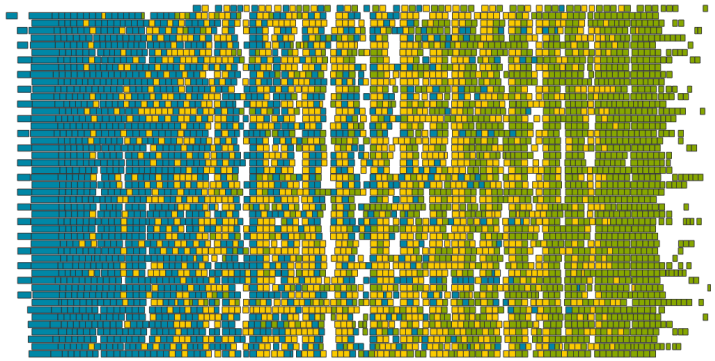


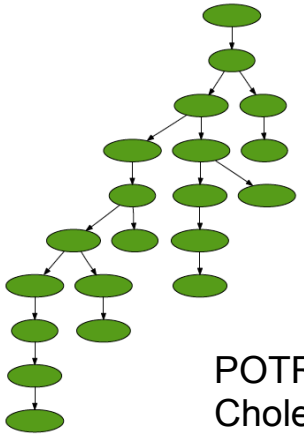
Pipelining: Cholesky Inversion
3 Steps: Factor, Invert L, Multiply L's

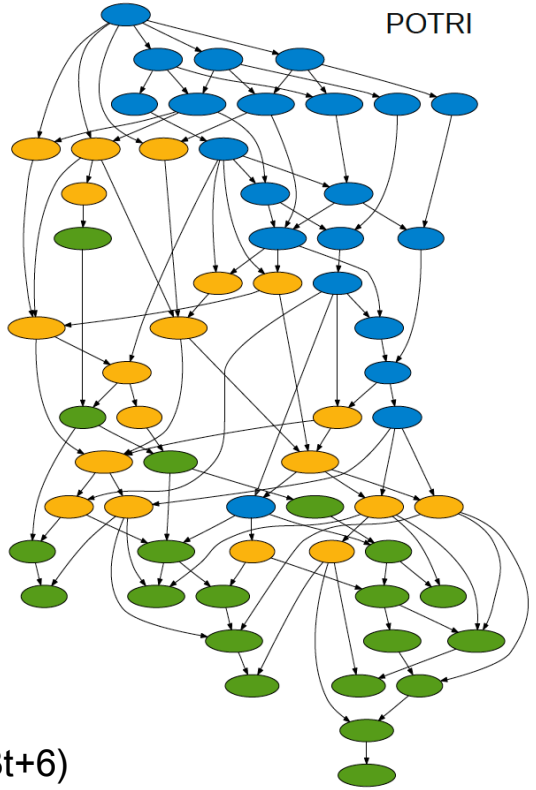












48 cores
POTRF, TRTRI and LAUUM.
The matrix is 4000 x 4000,tile size is 200 x 200,

POTRF+TRTRI+LAUUM: 25 (7t-3)
Cholesky Factorization alone: 3t-2

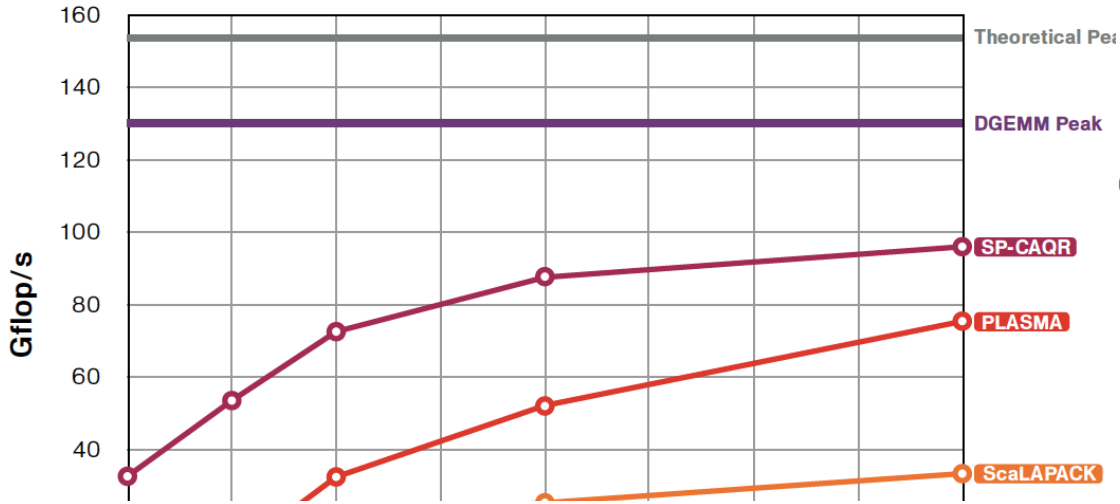
Pipelined: 18 (3t+6)

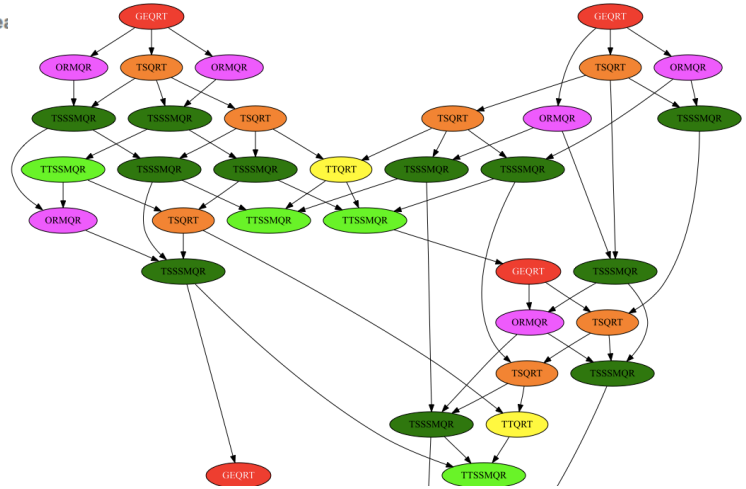
Communication Avoiding Algorithms

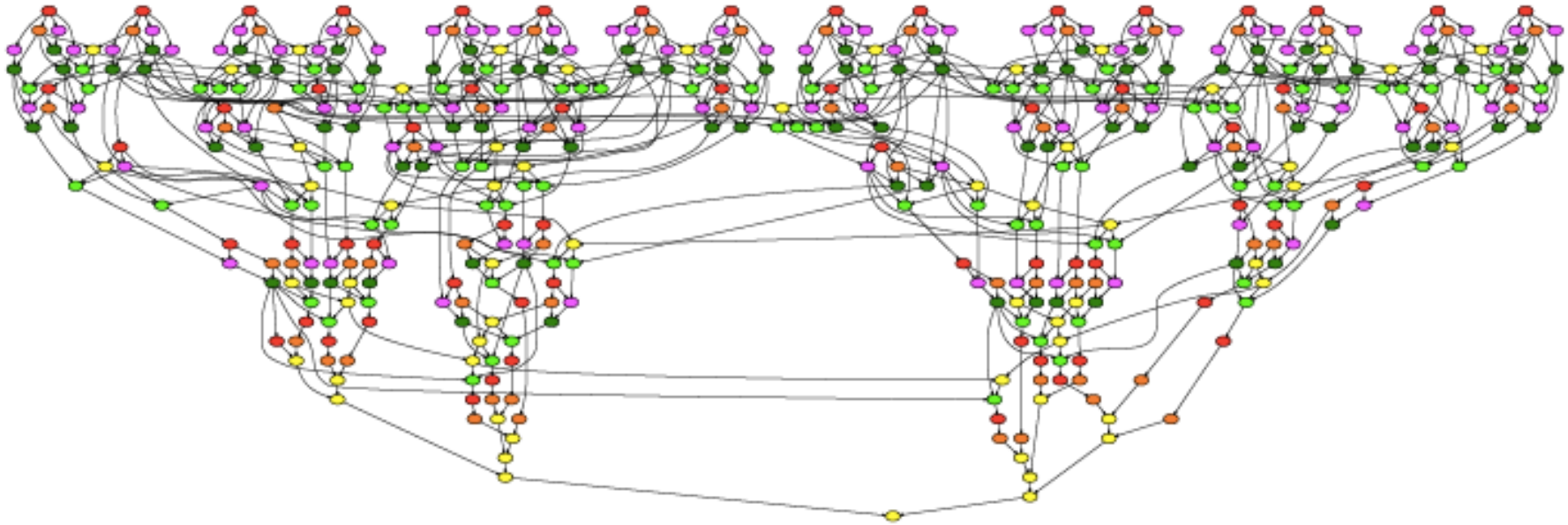
- Goal: Algorithms that communicate as little as possible
- Jim Demmel and company have been working on algorithms that obtain a provable minimum communication. (M. Anderson yesterday)
- Direct methods (BLAS, LU, QR, SVD, other decompositions)
  - Communication lower bounds for *all* these problems
  - Algorithms that attain them (*all* dense linear algebra, some sparse)
- Iterative methods - Krylov subspace methods for Ax=b, Ax=λx
  - Communication lower bounds, and algorithms that attain them (depending on sparsity structure)
- For QR Factorization they can show:

	Lower bound
# flops	$\Theta(mn^2)$
# words	$\Theta(\frac{mn^2}{\sqrt{W}})$
# messages	$\Theta(\frac{mn^2}{W^{3/2}})$

Communication Reducing QR Factorization





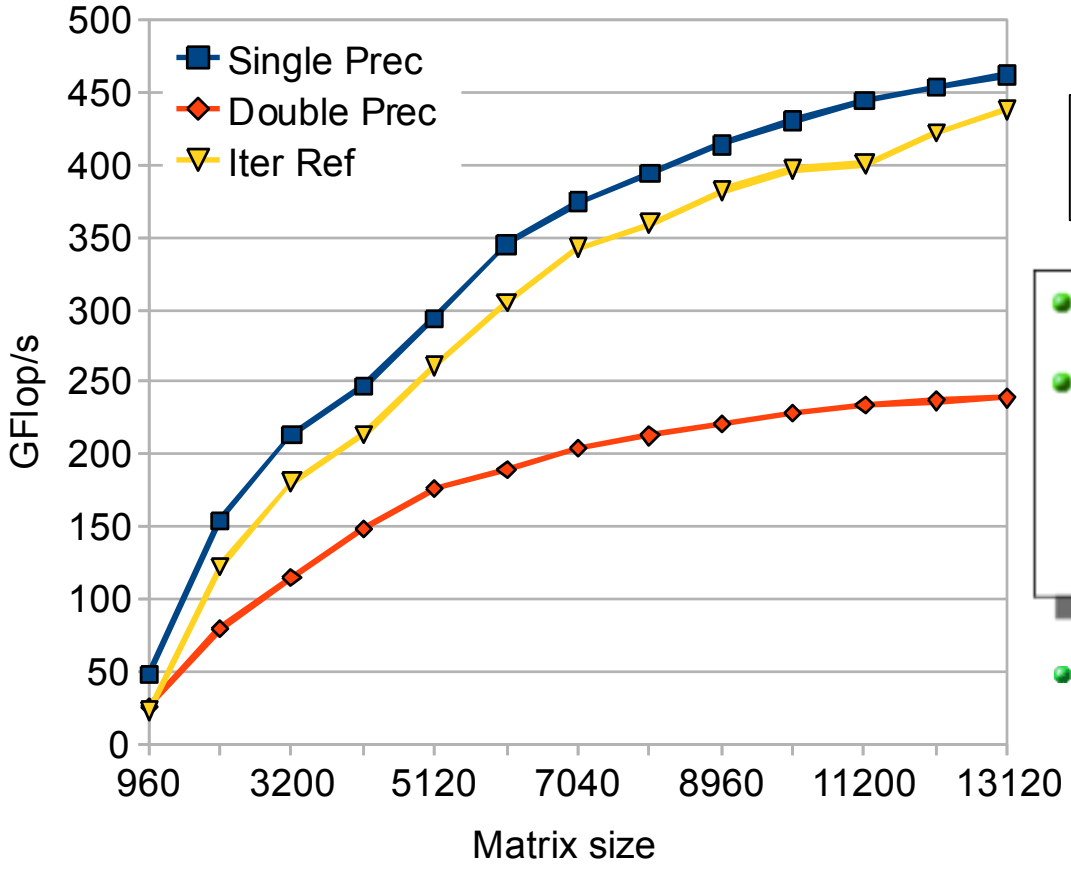


Mixed Precision Methods

- **Mixed precision, use the lowest precision required to achieve a given accuracy outcome**
  - **Improves runtime, reduce power consumption, lower data movement**
  - **Reformulate to find correction to solution, rather than solution [Δx rather than x].**

Mixed Precision Solvers

MAGMA LU-based solvers on Fermi (C2050)



FERMI Tesla C2050: 448 CUDA cores @ 1.15GHz
SP/DP peak is 1030 / 515 GFlop/s

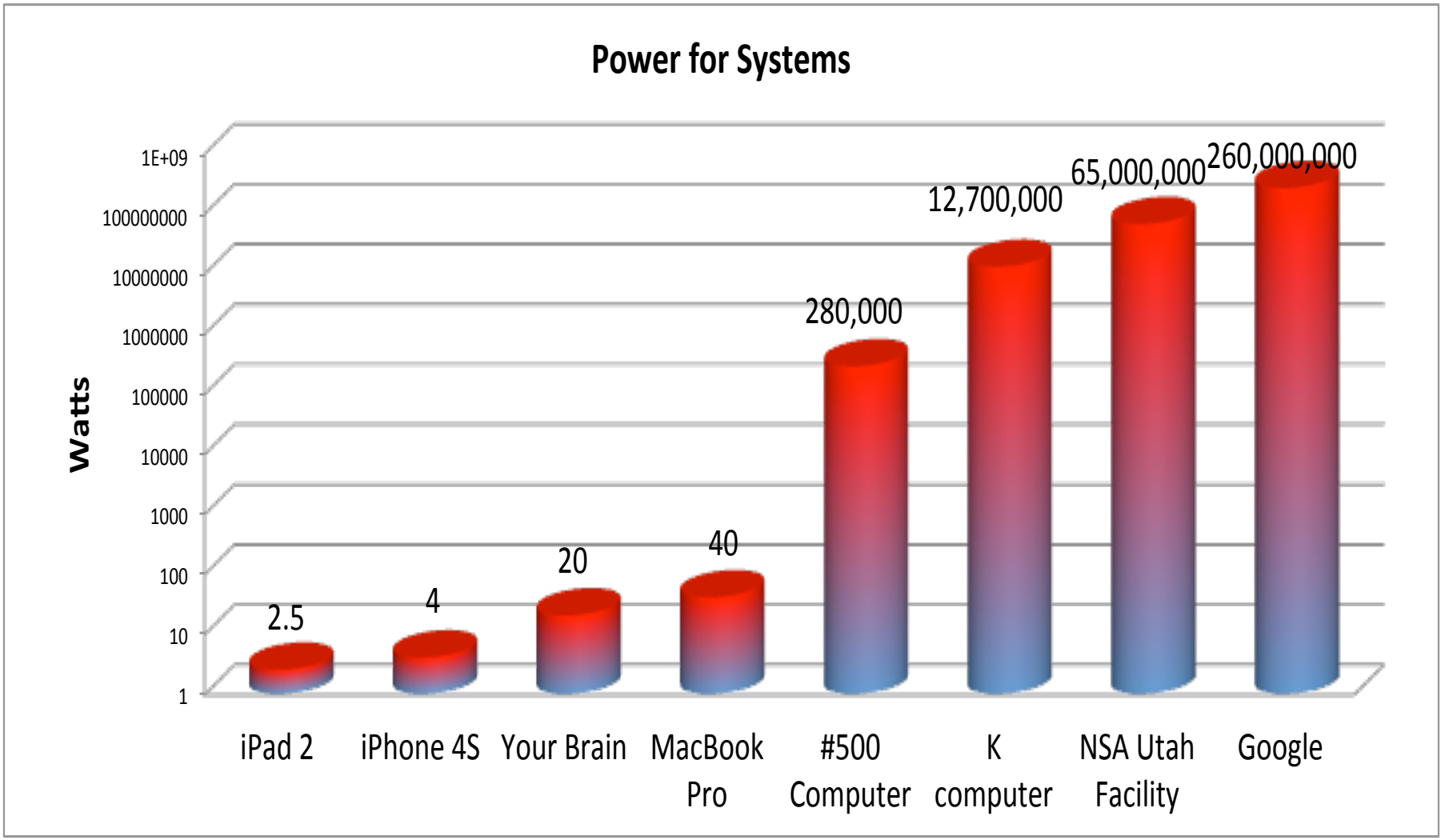
- **Direct solvers**
  - Factor and solve in working precision
- **Mixed Precision Iterative Refinement**
  - Factor in single (i.e. the bulk of the computation in fast arithmetic) and use it as preconditioner in simple double precision iteration, e.g.

$$x_{i+1} = x_i + (LU_{SP})^{-1} P (b - A x_i)$$

- Similar results for Cholesky & QR

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.
- High Performance Ecosystem out of balance
  - Hardware, OS, Compilers, Software, Algorithms, Applications
    - No Moore's Law for software, algorithms and applications


Power for Systems
Watts
1E+09
100000000
10000000
1000000
100000
10000
1000
100
10
1
2.5
4
20
40
280,000
12,700,000
65,000,000
260,000,000
iPad 2
iPhone 4S
Your Brain
MacBook Pro
#500 Computer
K computer
NSA Utah Facility
Google