



# Algorithmic and Software Challenges when Moving Towards Exascale

---

**Jack Dongarra**

University of Tennessee  
Oak Ridge National Laboratory  
University of Manchester

# Overview

---

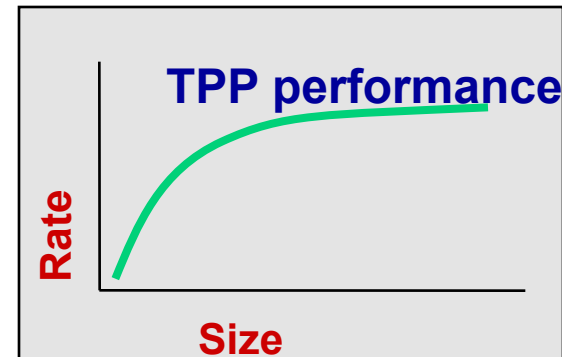
- **High Performance Computing Today**
- **The Road Ahead for HPC**
- **Challenges for Algorithms and Software Design**

---

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

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

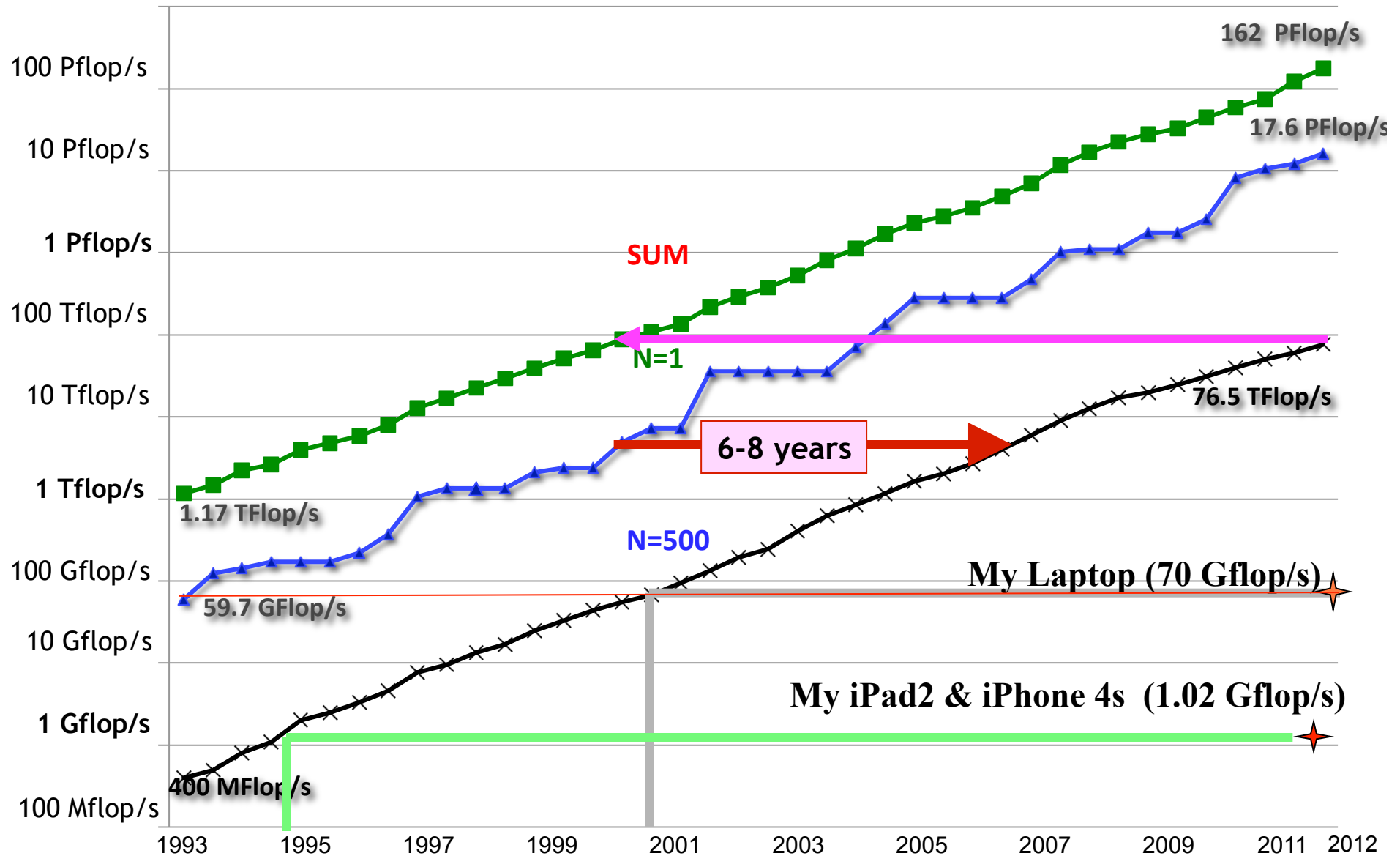
$$Ax=b, \text{ dense problem}$$



- 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 of HPC Over the Last 20 Years















# Pflop/s Club (23 systems)

Name	Pflop/s	Country	10  4  2  2  2  2  1 
Titan	17.6	US	Cray: <b>Hybrid</b> AMD/Nvidia/Custom
Sequoia	16.3	US	IBM: BG-Q/Custom
K computer	10.5	Japan	Fujitsu: Sparc/Custom
Mira	8.16	US	IBM: BG-Q/Custom
JuQUEEN	4.14	Germany	IBM: BG-Q/Custom
SuperMUC	2.90	Germany	IBM: Intel/IB
Stampede	2.66	US	Dell: <b>Hybrid</b> Intel/Intel/IB
Tianhe-1A	2.57	China	NUDT: <b>Hybrid</b> Intel/Nvidia/Custom
Fermi	1.73	Italy	IBM: BG-Q/Custom
DARPA Trial Subset	1.52	US	IBM: IBM/Custom
Curie thin nodes	1.36	France	Bull: Intel/IB
Nebulae	1.27	China	Dawning: <b>Hybrid</b> Intel/Nvidia/IB
Yellowstone	1.26	US	IBM: Intel/IB
Pleiades	1.24	US	SGI: Intel/IB
Helios	1.24	Japan	Bull: Intel/IB
Blue Joule	1.21	UK	IBM: BG-Q/Custom
TSUBAME 2.0	1.19	Japan	HP: <b>Hybrid</b> Intel/Nvidia/IB
Cielo	1.11	US	Cray: AMD/Custom
Hopper	1.05	US	Cray: AMD/Custom
Tera-100	1.05	France	Bull: Intel/IB
Oakleaf-FX	1.04	Japan	Fujitsu: Sparc/Custom
Roadrunner	1.04	US	IBM: <b>Hybrid</b> AMD/Cell/IB ( <b>First one in '08</b> )
DiRAC	1.04	UK	IBM: BG-Q/Custom

# November 2012: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7 (16C) + <b>Nvidia Kepler GPU (14c)</b> + custom		560,640	17.6	66	8.3	2120
2	DOE / NNSA Livermore Nat Lab	Sequoia, BlueGene/Q (16c) + custom		1,572,864	16.3	81	7.9	2063
3	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8c) + custom		705,024	10.5	93	12.7	827
4	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16c) + custom		786,432	8.16	81	3.95	2066
5	Forschungszentrum Juelich	JuQUEEN, BlueGene/Q (16c) + custom		393,216	4.14	82	1.97	2102
6	Leibniz Rechenzentrum	SuperMUC, Intel (8c) + IB		147,456	2.90	90*	3.42	848
7	Texas Advanced Computing Center	Stampede, Dell Intel (8) + <b>Intel Xeon Phi (61)</b> + IB		204,900	2.66	67	3.3	806
8	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel (6c) + <b>Nvidia Fermi GPU (14c)</b> + custom		186,368	2.57	55	4.04	636
9	CINECA	Fermi, BlueGene/Q (16c) + custom		163,840	1.73	82	.822	2105
10	IBM	DARPA Trial System, Power7 (8C) + custom		63,360	1.51	78	.358	422

500 Slovak Academy Sci

IBM Power 7

Slovak Rep

3,074

.077

81

# November 2012: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7 (16C) + <b>Nvidia Kepler GPU (14c)</b> + custom	USA	560,640	17.6	66	8.3	2120
2	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16c) + custom	USA	1,572,864	16.3	81	7.9	2063
3	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8c) + custom	Japan	705,024	10.5	93	12.7	827
4	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16c) + custom	USA	786,432	8.16	81	3.95	2066
5	Forschungszentrum Juelich	JuQUEEN, BlueGene/Q (16c) + custom	Germany	393,216	4.14	82	1.97	2102
6	Leibniz Rechenzentrum	SuperMUC, Intel (8c) + IB	Germany	147,456	2.90	90*	3.42	848
7	Texas Advanced Computing Center	Stampede, Dell Intel (8) + <b>Intel Xeon Phi (61)</b> + IB	USA	204,900	2.66	67	3.3	806
8	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel (6c) + <b>Nvidia Fermi GPU (14c)</b> + custom	China	186,368	2.57	55	4.04	636
9	CINECA	Fermi, BlueGene/Q (16c) + custom	Italy	163,840	1.73	82	.822	2105
10	IBM	DARPA Trial System, Power7 (8C) + custom	USA	63,360	1.51	78	.358	422

# Top500 Systems in Mexico

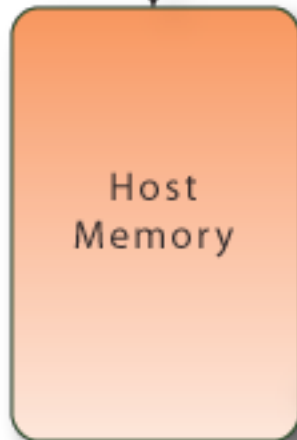
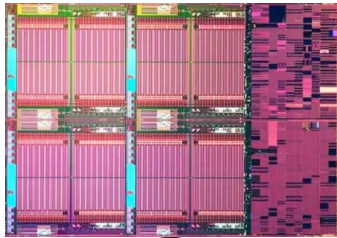
---

Rank	Computer	Site	Manufacturer	Total Cores	Rmax Tflop/s	Efficiency (%)
348	Xeon E5-2670 8C 2.6GHz, InfB	Universidad Nacional Autonoma de Mexico	HP	56,160	92	79

# Commodity plus Accelerator Today

## Commodity

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

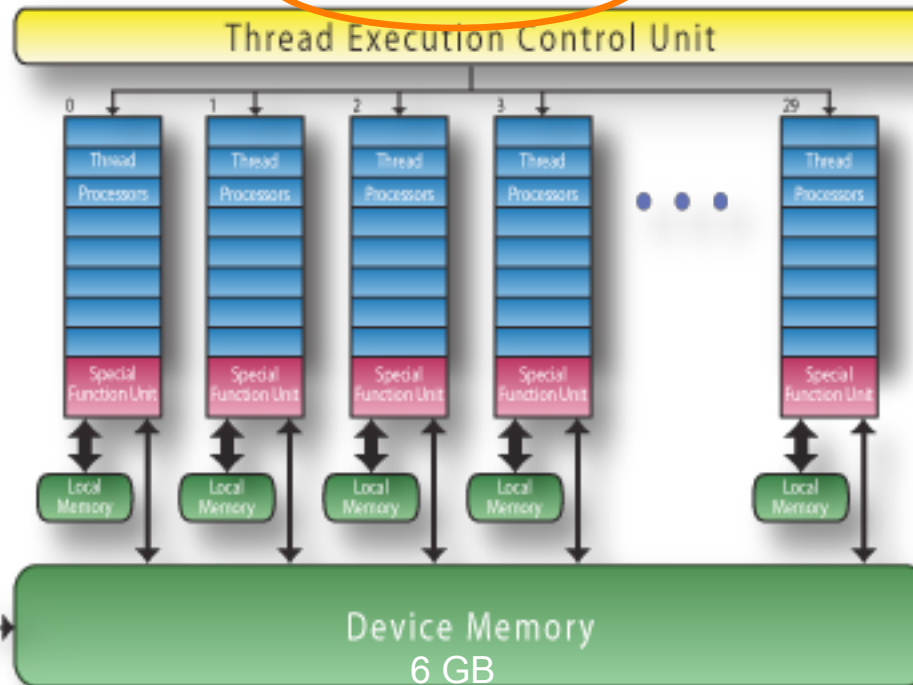


Interconnect  
PCI-X 16 lane  
64 Gb/s (8 GB/s)  
1 GW/s

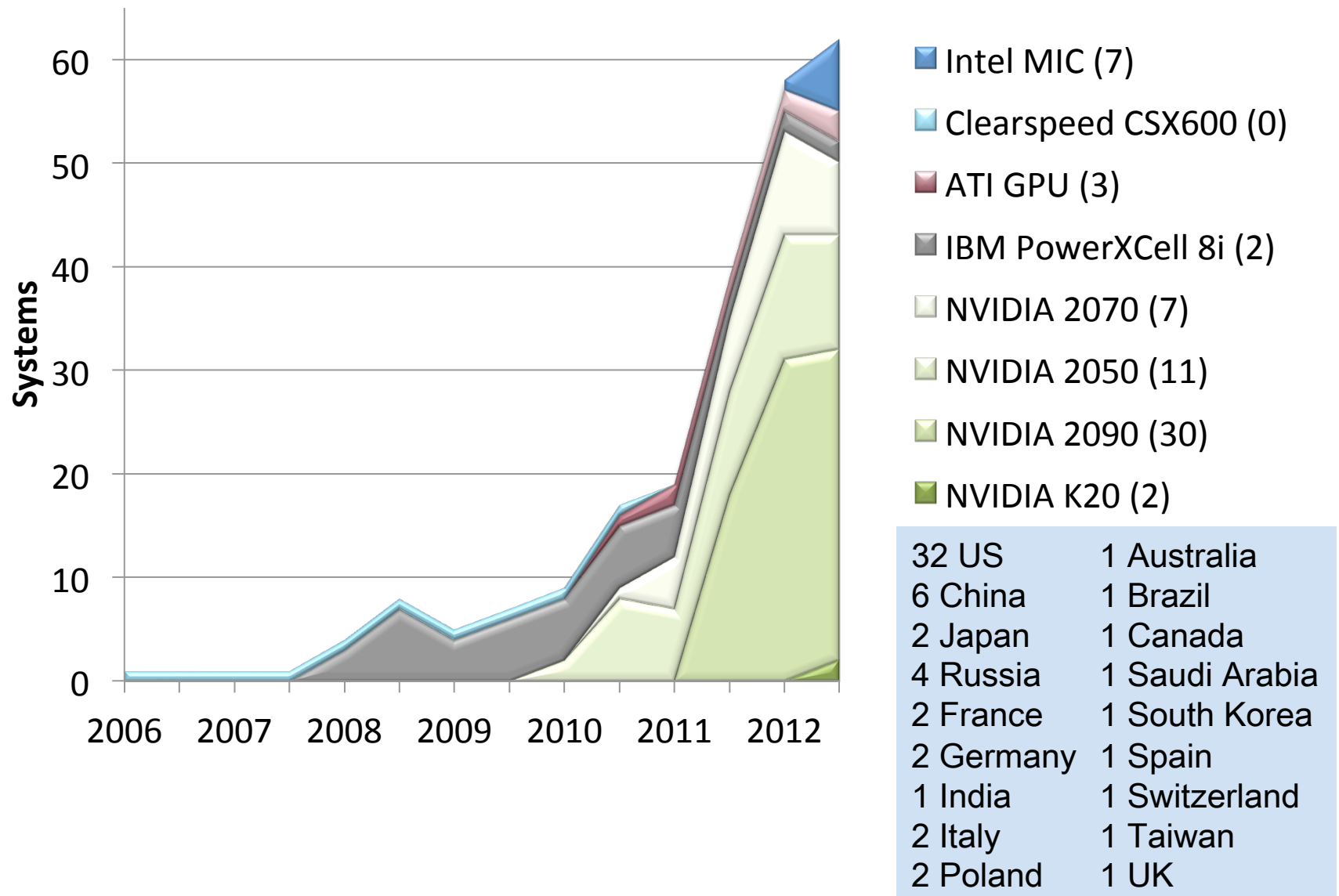
## Accelerator (GPU)

Nvidia K20X "Kepler"  
2688 "Cuda cores"  
.732 GHz  
2688\*2/3 ops/cycle  
1.31 Tflop/s (DP)

192 Cuda cores/SMX  
2688 "Cuda cores"



# Accelerators (62 systems)





# We Have Seen This Before

- Floating Point Systems FPS-164/MAX Supercomputer (1976)
- Intel Math Co-processor (1980)
- Weitek Math Co-processor (1981)



**1976**

**THREE HUNDRED FORTY ONE MILLION FLOATING POINT OPERATIONS PER SECOND. THE FPS-164/MAX.**

Today's scientific and engineering problems increasingly call for supercomputers that can handle the most demanding tasks in the shortest time. The FPS-164/MAX is a supercomputer with the speed and accuracy of a supercomputer, but at a fraction of the cost.

**The FPS-164/MAX is fast.** With peak performance of over 340 million floating point operations per second, depending on configuration, and up to 700 million if all four processors are available to the user, the FPS-164/MAX gives you all the speed and accuracy you need to solve those most demanding engineering problems.

**The FPS-164/MAX is cost-effective.** The FPS-164/MAX is designed for the demanding requirements of Floating Point Systems. With 22 built-in service units, full memory diagnostic capabilities, and a record of proven quality and reliability, the FPS-164/MAX will be up, running, and ready to solve your problems today.

For complete information and applications, call toll free 1-800-567-1415.

Model	Peak Speed (MFLOPS)	Number of Processors	Memory (MB)	Weight (lb)
FPS-164/MAX-1	340	4	16	100
FPS-164/MAX-2	700	8	32	200

Circle Number 318 on Reader Service Card

**The Intel® Math CoProcessor™ is for crunching numbers faster.**

There's one for every machine.

**80387™ Family, for 386™ based machines.**

**80287™ Family, for 286™ based machines.**

**80187™ Family, for 8086™ and 8088™ based machines.**

**80487™ Family, for 486™ based machines.**

**It's FAST!**  
The Intel Math CoProcessor dramatically speeds up the number crunching parts of the work you do every day: budgeting, statistical analysis, financial analysis, CAD and other engineering analysis. In fact, the Math CoProcessor is supported by more than 100 commonly used software packages including Lotus 1-2-3, dBase IV, AutoCAD, and most languages and statistical packages.

**It's EASY!**  
Intel makes a variety of math co-processors. Every PC has a built-in socket. Just plug it in and go.

**It's SAFE!**  
Made by Intel, the same people who designed your PC's microprocessor, each and every Math CoProcessor is backed by an industry leading the way warranty and full technical support. You are assured the highest degree of quality, compatibility, reliability and support for your investment.

For more information, or technical support call:  
(800) 538-3373 in the US and Canada  
(510) 638-7584 for International

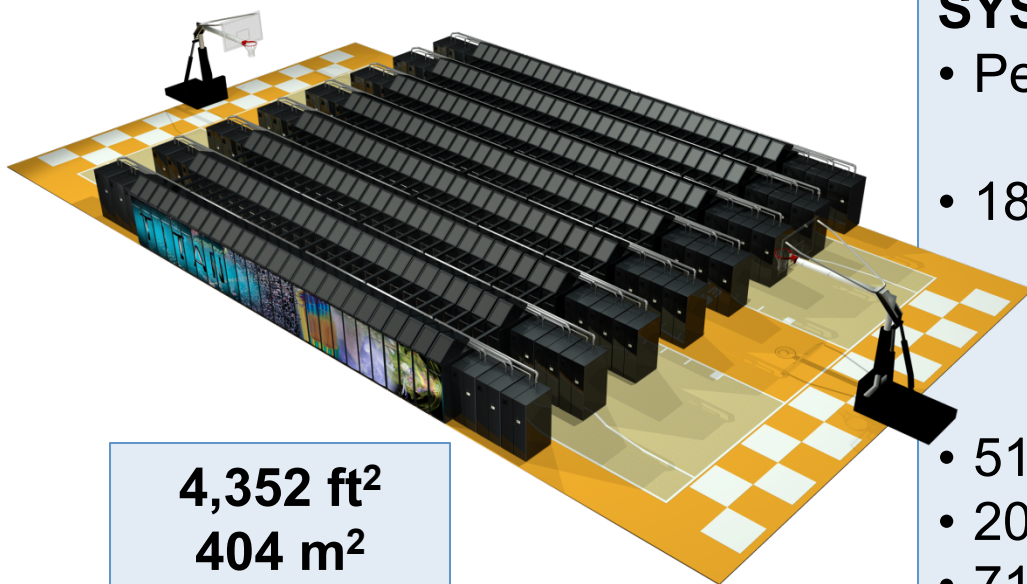
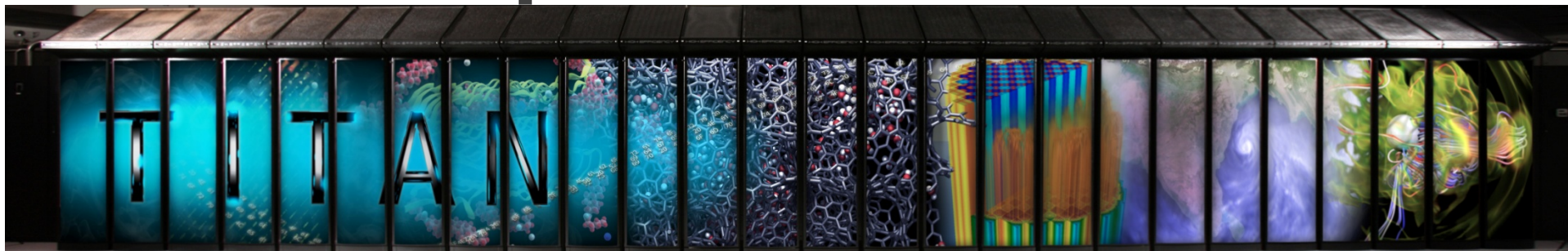
Intel Math CoProcessors are 100% VLSI and 100% tested and proven under 100% burn-in.  
Intel and the Intel logo are trademarks of Intel Corporation.

**intel**

Personal Computer Enhancement

**1980**

# ORNL's "Titan" Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla processors



**4,352 ft<sup>2</sup>  
404 m<sup>2</sup>**

## SYSTEM SPECIFICATIONS:

- Peak performance of 27 PF
  - 24.5 Pflop/s GPU + 2.6 Pflop/s AMD
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - 14-Core NVIDIA Tesla "K20x" GPU
  - 32 GB + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 9 MW peak power



# Cray XK7 Compute Node

## XK7 Compute Node Characteristics

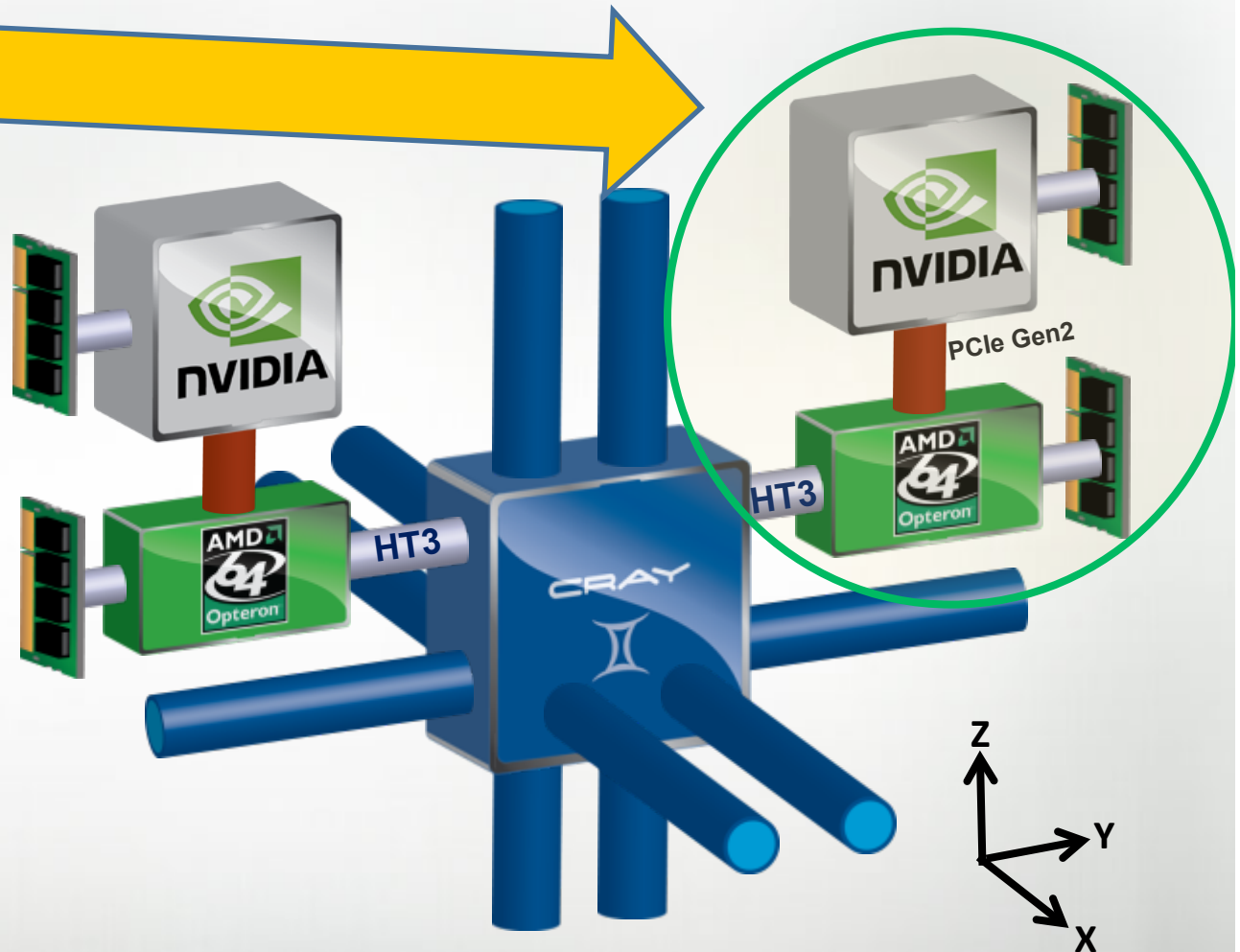
AMD Opteron 6274 Interlagos  
16 core processor

Tesla K20x @ 1311 GF

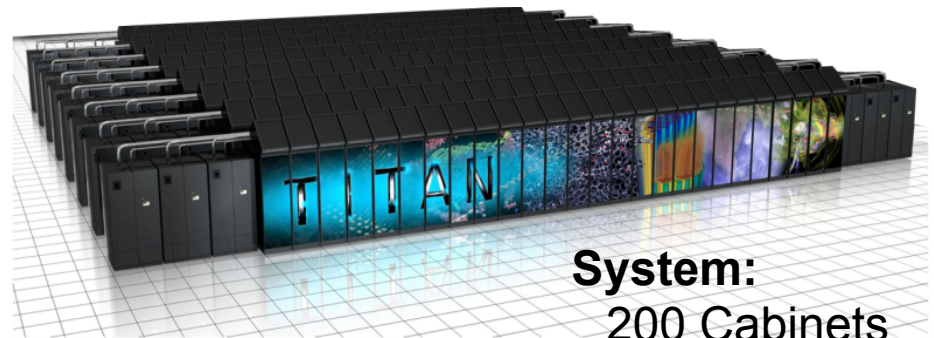
Host Memory  
32GB  
1600 MHz DDR3

Tesla K20x Memory  
6GB GDDR5

Gemini High Speed Interconnect

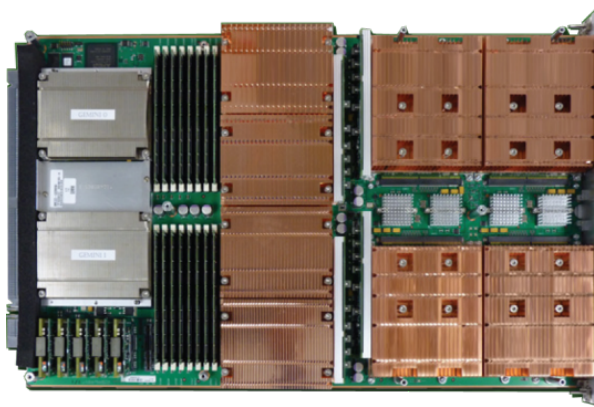


# Titan: Cray XK7 System



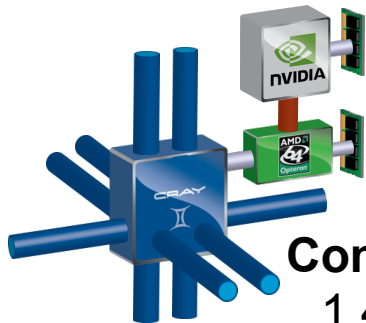
## System:

200 Cabinets  
18,688 Nodes  
27 PF  
710 TB



## Board:

4 Compute Nodes  
5.8 TF  
152 GB



## Compute Node:

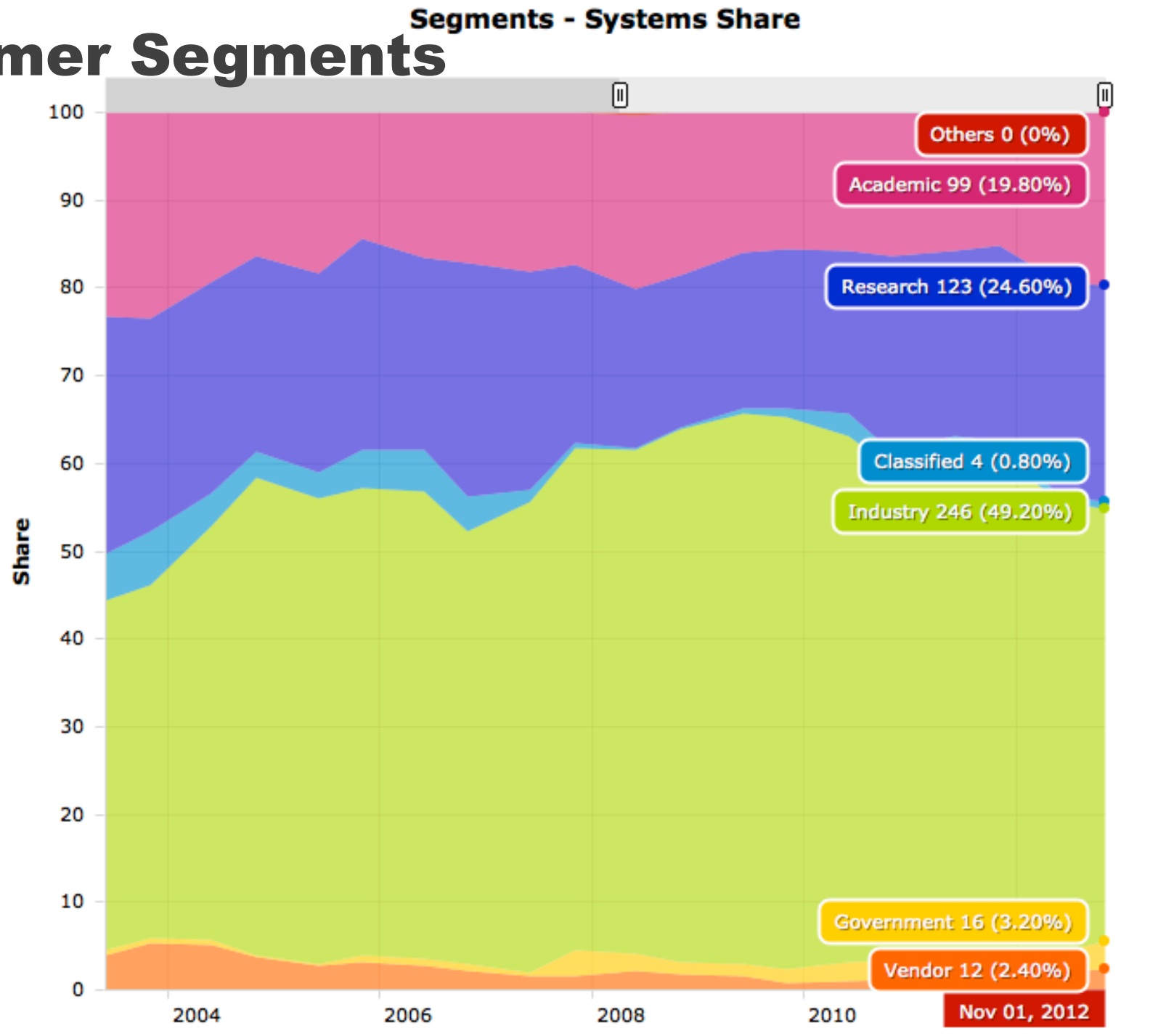
1.45 TF  
38 GB



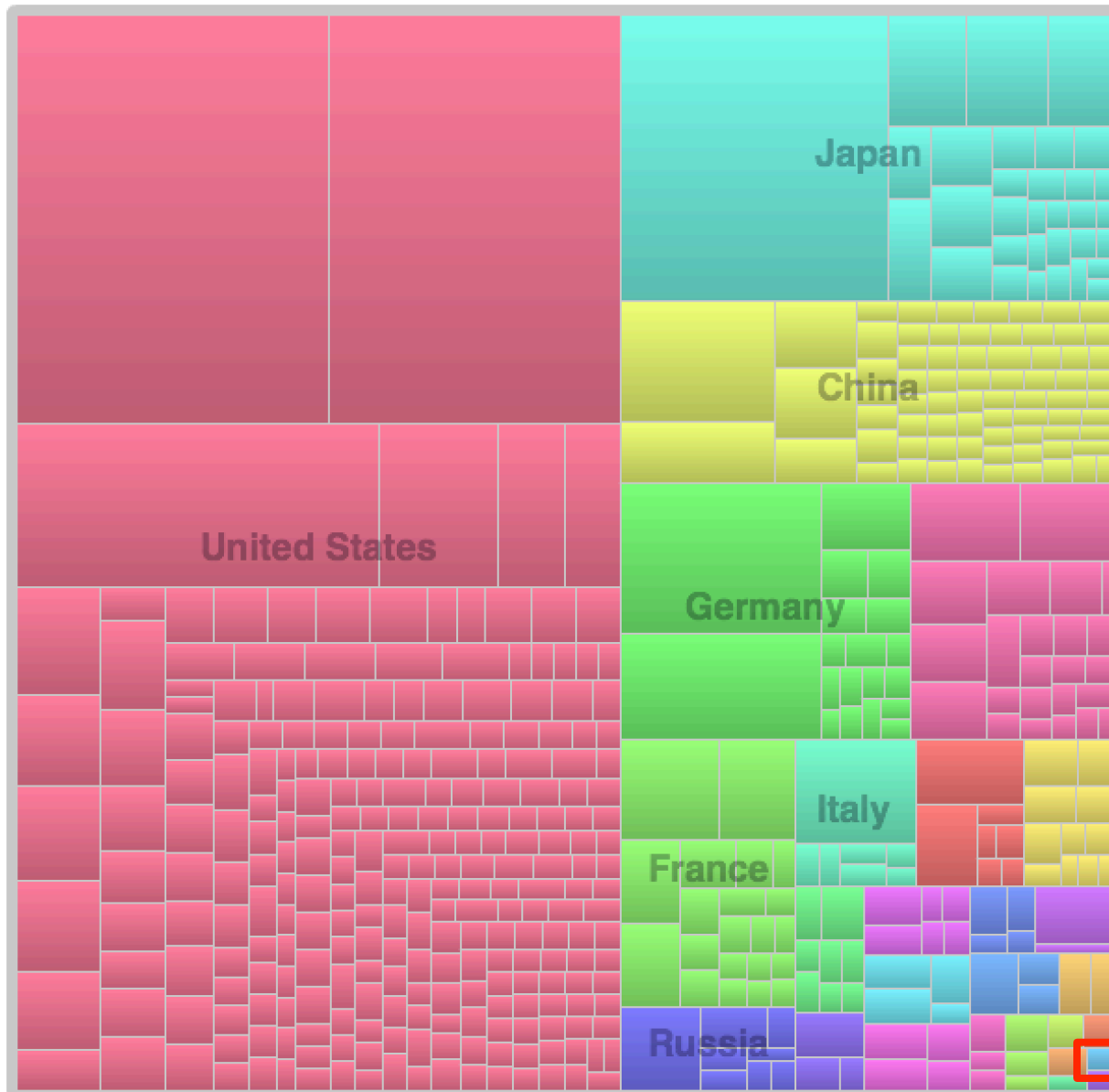
## Cabinet:

24 Boards  
96 Nodes  
139 TF  
3.6 TB

# Customer Segments



# Countries Share

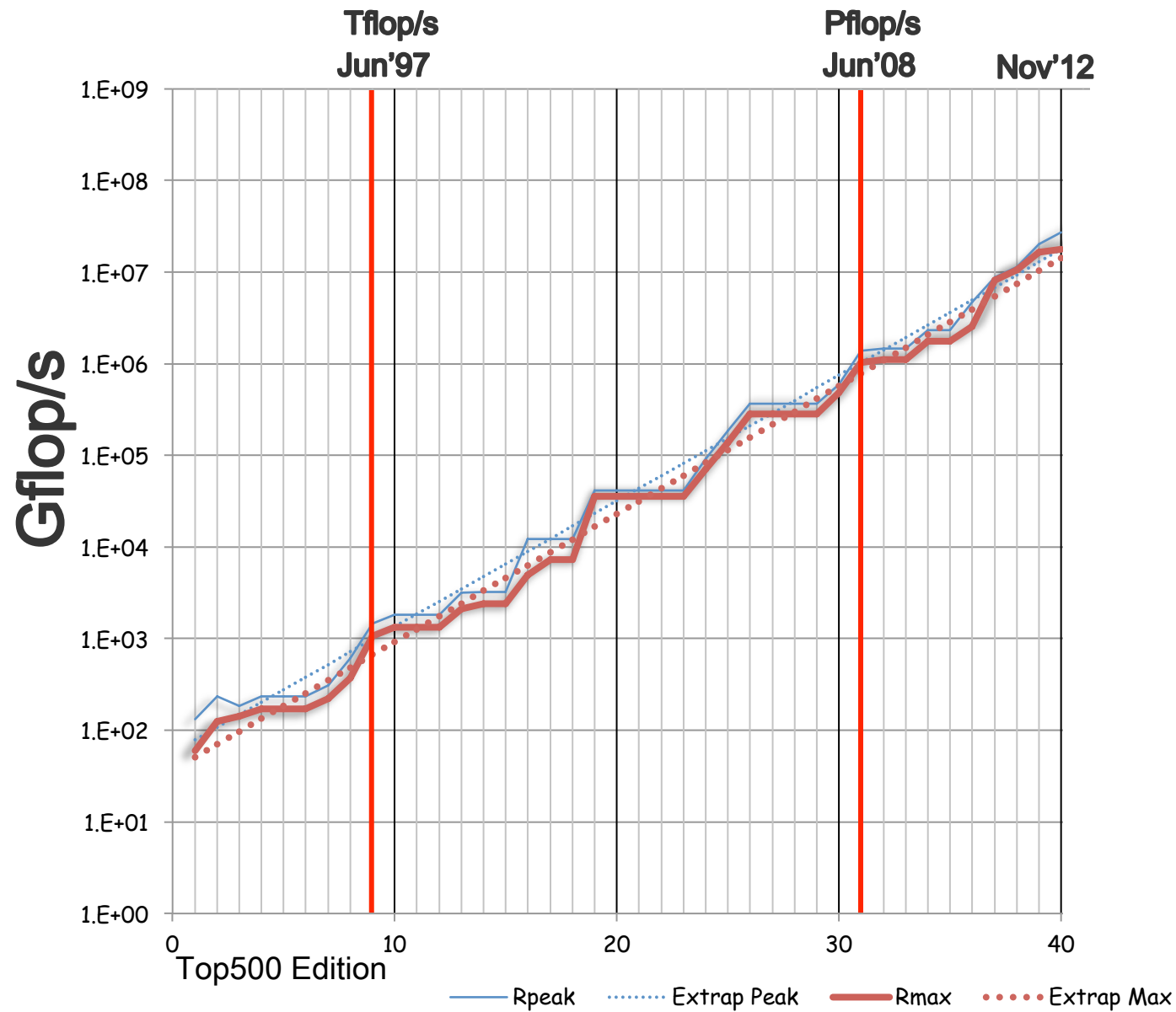


## Absolute Counts

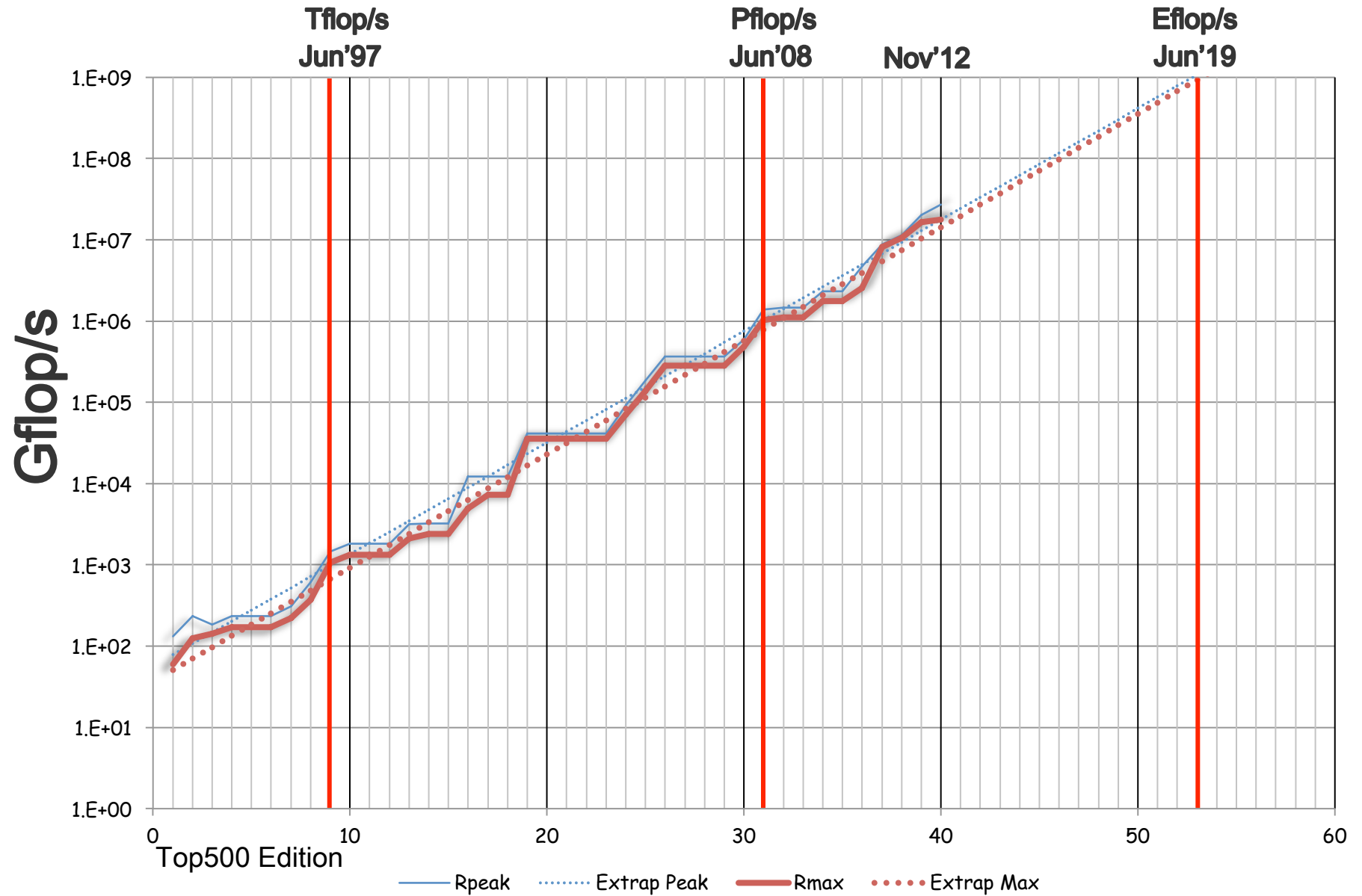
US:	251
China:	72
Japan:	31
UK:	24
France:	21
Germany:	20

Mexico

# TOP500 Editions (40 so far, 20 years)



# TOP500 Editions (53 edition, 26 years)



# The High Cost of Data Movement

---

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

Approximate power costs (in picoJoules)

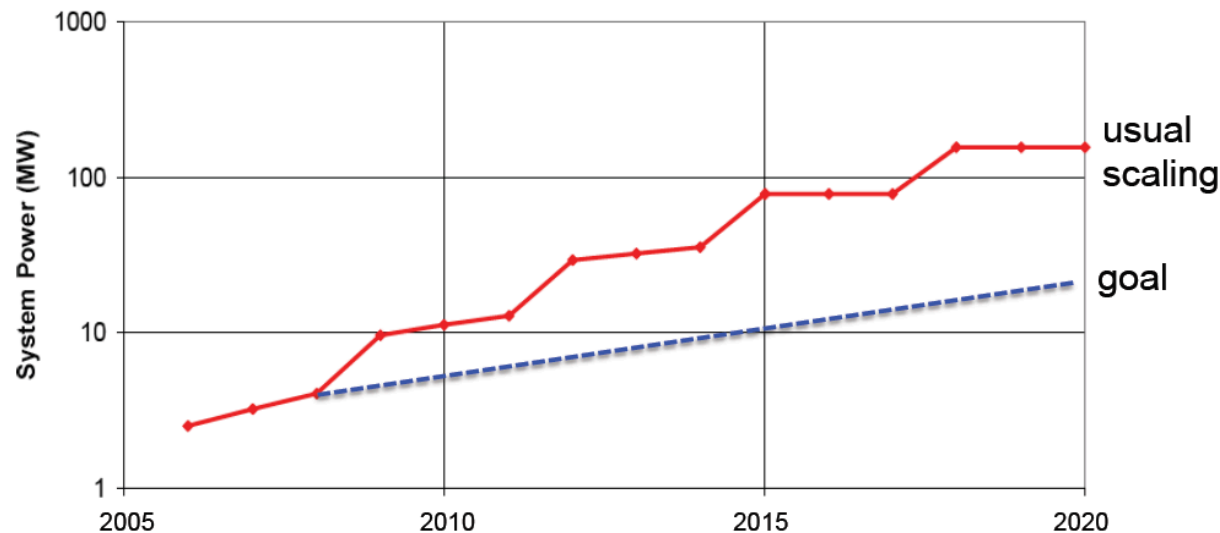
	2011
DP FMADD flop	100 pJ
DP DRAM read	4800 pJ
Local Interconnect	7500 pJ
Cross System	9000 pJ

Source: John Shalf, LBNL

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

# 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 around 2020-2022 at 20 MWs





# Potential System Architecture

Systems	2013 Titan Computer
System peak	27 Pflop/s
Power	8.3 MW (2 Gflops/W)
System memory	710 TB (38*18688)
Node performance	1,452 GF/s (1311+141)
Node memory BW	232 GB/s (52+180)
Node concurrency	16 cores CPU 2688 CUDA cores
Total Node Interconnect BW	8 GB/s
System size (nodes)	18,688
Total concurrency	50 M
MTTF	?? unknown



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

Systems	2013 Titan Computer	2020	Difference Today & 2020
System peak	27 Pflop/s	1 Eflop/s	O(100)
Power	8.3 MW (2 Gflops/W)	~20 MW (50 Gflops/W)	O(10)
System memory	710 TB (38*18688)	32 - 64 PB	O(100)
Node performance	1,452 GF/s (1311+141)	1.2 or 15TF/s	O(10)
Node memory BW	232 GB/s (52+180)	2 - 4TB/s	O(10)
Node concurrency	16 cores CPU 2688 CUDA cores	O(1k) or 10k	O(100) - O(10)
Total Node Interconnect BW	8 GB/s	200-400GB/s	O(100)
System size (nodes)	18,688	O(100,000) or O(1M)	O(10) - O(100)
Total concurrency	50 M	O(billion)	O(100)
MTTF	?? unknown	O(<1 day)	O(?)



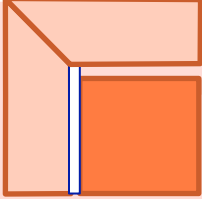

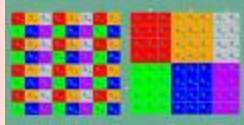
# Critical Issues at Peta & Exascale for Algorithm and Software Design

---

- **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.

# A New Generation of DLA Software

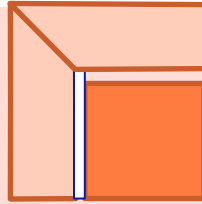
## Software/Algorithms follow hardware evolution in time

LINPACK (70's) (Vector operations)		Rely on - Level-1 BLAS operations
LAPACK (80's) (Blocking, cache friendly)		Rely on - Level-3 BLAS operations
ScaLAPACK (90's) (Distributed Memory)		Rely on - PBLAS Mess Passing

# A New Generation of DLA Software

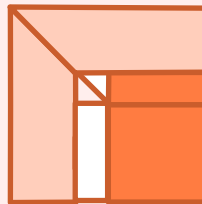
Software/Algorithms follow hardware evolution in time

LINPACK (70's)  
(Vector operations)



Rely on  
- Level-1 BLAS operations

LAPACK (80's)  
(Blocking, cache friendly)



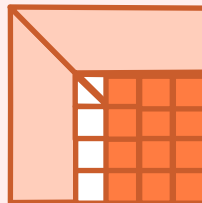
Rely on  
- Level-3 BLAS operations

ScaLAPACK (90's)  
(Distributed Memory)



Rely on  
- PBLAS Mess Passing

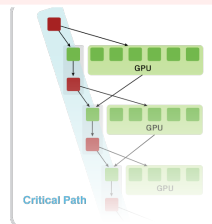
PLASMA  
New Algorithms  
(many-core friendly)



Rely on  
- a DAG/scheduler  
- block data layout  
- some extra kernels

**MAGMA**

Hybrid Algorithms  
(heterogeneity friendly)



# Summary

---

- **Major Challenges are ahead for extreme computing**
  - **Parallelism  $O(10^9)$** 
    - Programming issues
  - **Hybrid**
    - Peak and HPL may be very misleading
    - No where near close to peak for most apps
  - **Fault Tolerance**
    - Today Sequoia BG/Q node failure rate is 1.25 failures/day
  - **Power**
    - 50 Gflops/w (today at 2 Gflops/w)
- **We will need completely new approaches and technologies to reach the Exascale level**



# Collaborators / Software / Support

---

- **PLASMA**  
<http://icl.cs.utk.edu/plasma/>
- **MAGMA**  
<http://icl.cs.utk.edu/magma/>
- **Quark (RT for Shared Memory)**  
<http://icl.cs.utk.edu/quark/>
- **PaRSEC**(Parallel Runtime Scheduling and Execution Control)  
<http://icl.cs.utk.edu/parsec/>



- Collaborating partners  
University of Tennessee, Knoxville  
University of California, Berkeley  
University of Colorado, Denver  
  
INRIA, France  
KAUST, Saudi Arabia

These tools are being applied to a range of applications beyond dense LA:  
Sparse direct, Sparse iterations methods and Fast Multipole Methods