EXACOMP 2011

International Workshop On Exascale Supercomputing

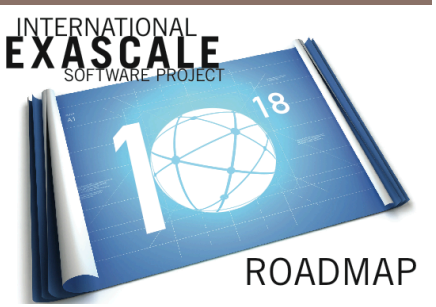
Date ; Nov. 3rd ~ 4th, 2011

Location ; COEX, Seoul, South Korea (# Room 402)

www.exascale.org

On the Future of High Performance Computing and the Importance of Software for Exascale

Jack Dongarra
University of Tennessee
Oak Ridge National Lab
University of Manchester

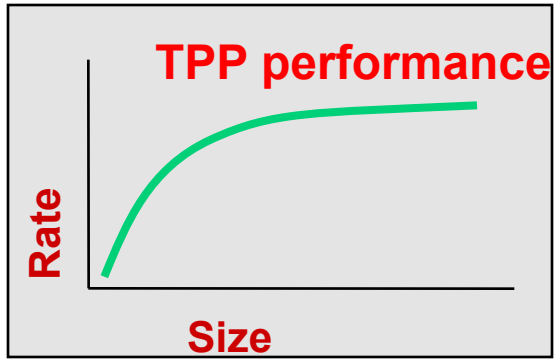


Top500 List of Supercomputers

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

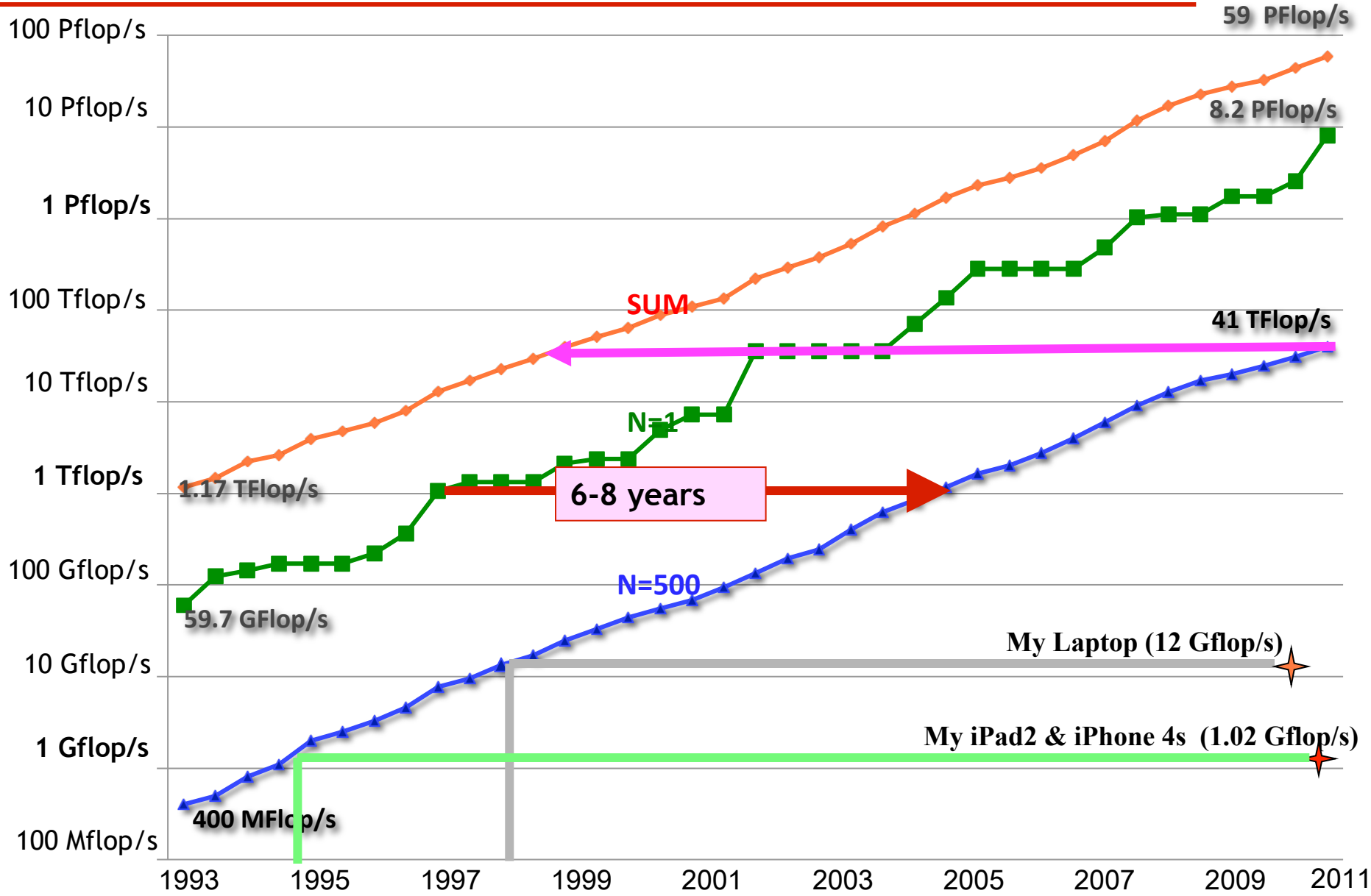
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

$Ax=b$, *dense problem*

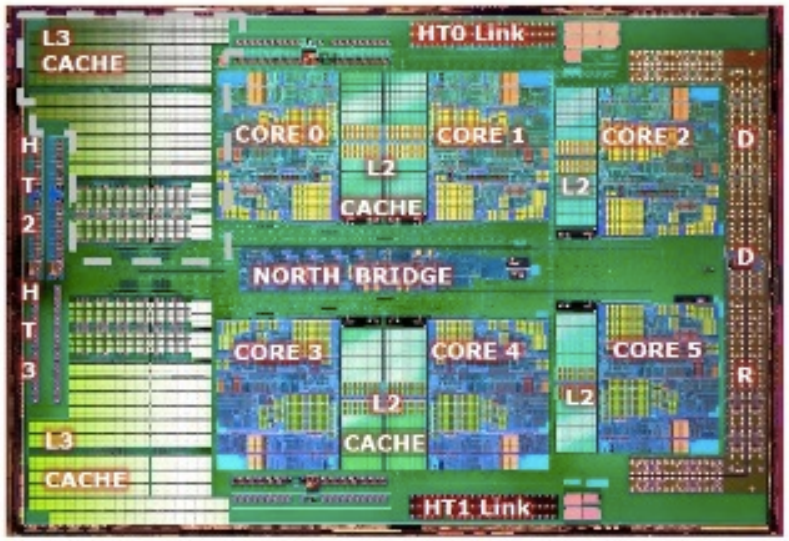


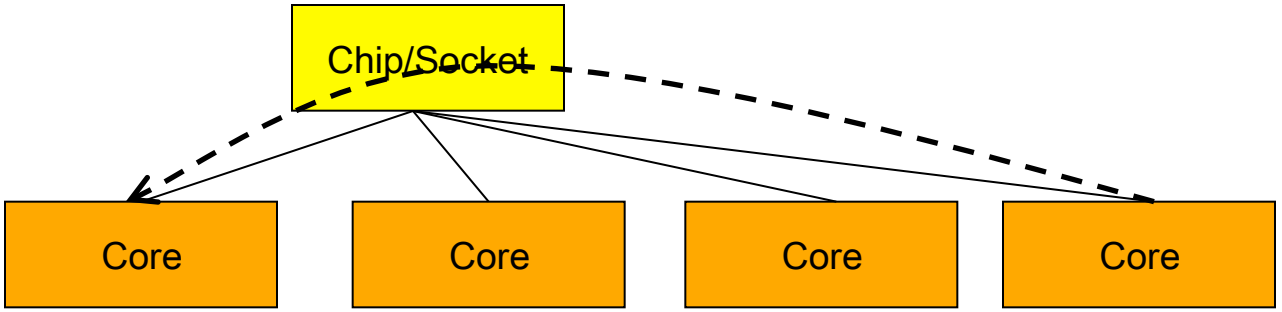
- Updated twice a year
  SC‘xy in the States in November
  Meeting in Germany in June
- All data available from **www.top500.org**

Performance Development

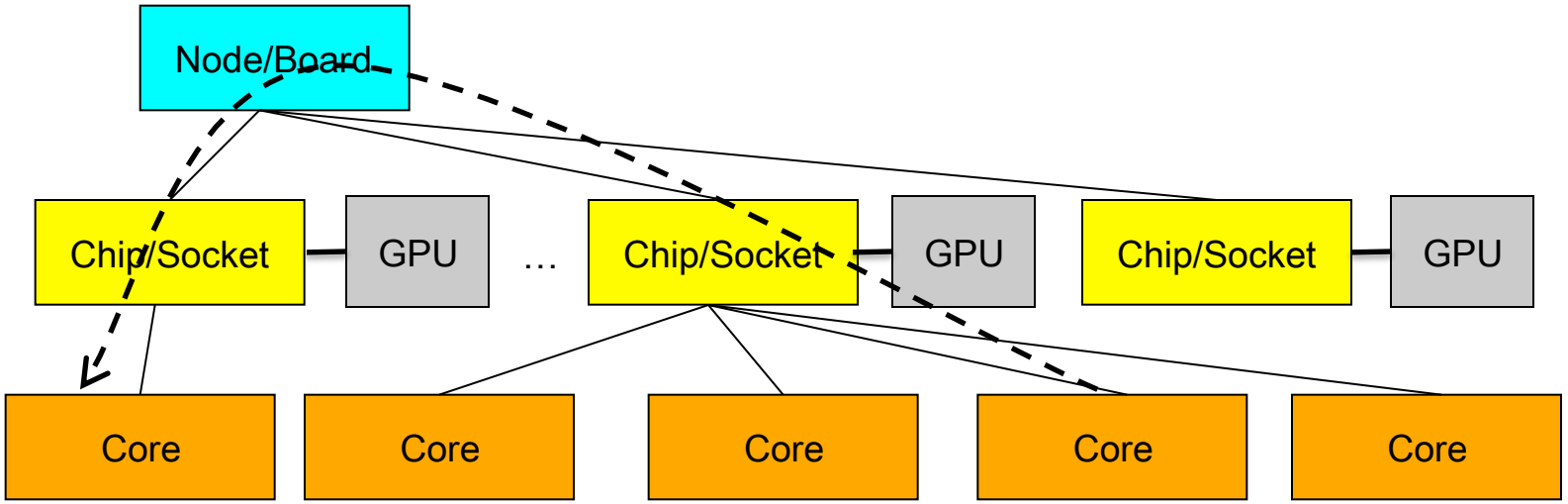


Example of typical parallel machine



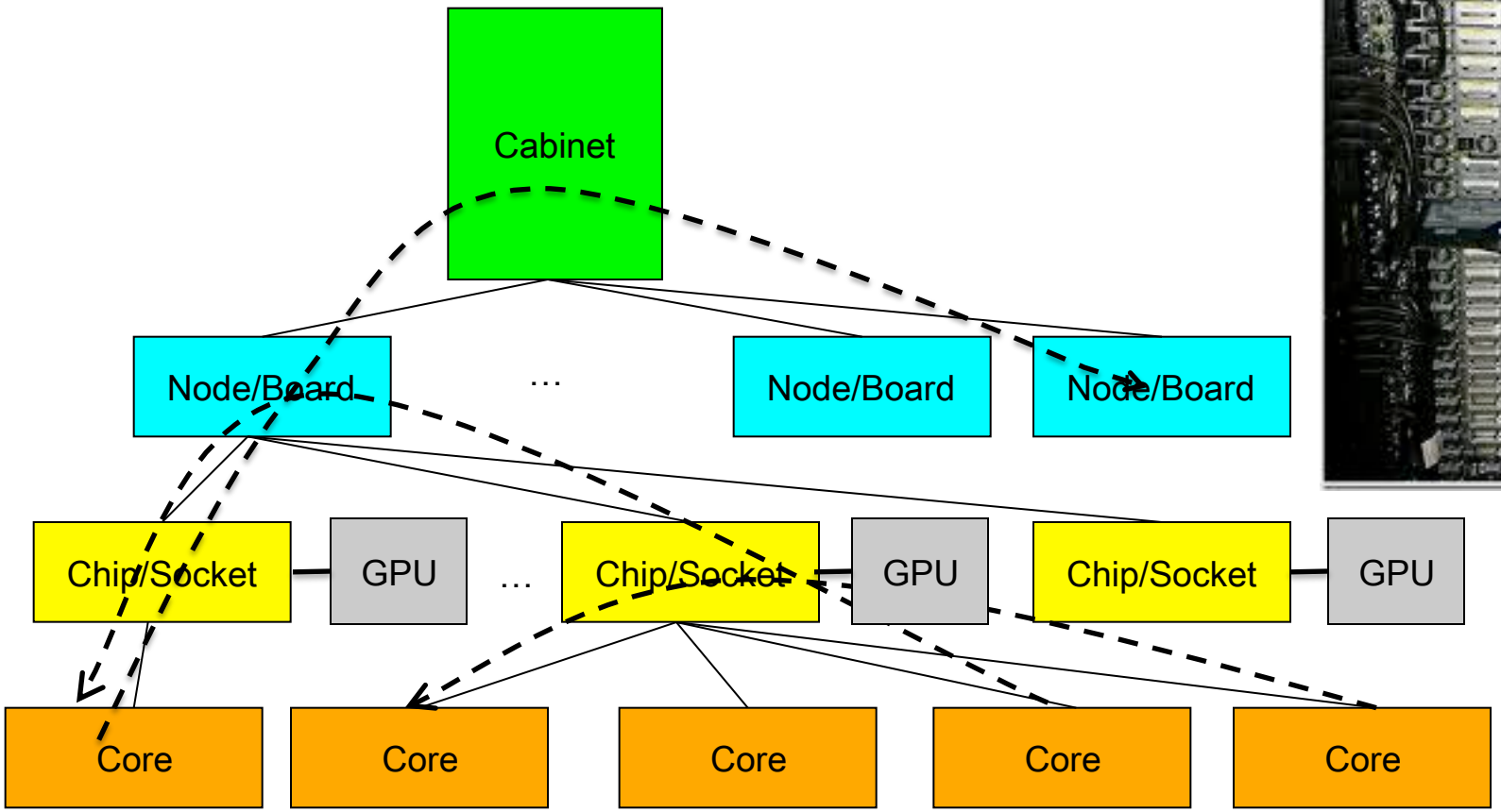


Example of typical parallel machine

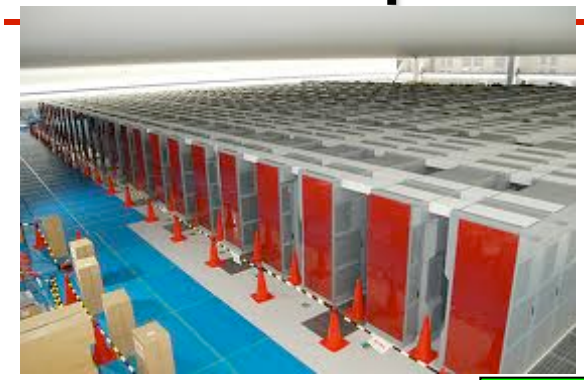


Example of typical parallel machine

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

June 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	548,352	8.16	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

June 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	548,352	8.16	93	9.9	824
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	Nebulae, 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	850
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	Energy Comp	IBM Cluster, Intel + GigE	China	7,104	.041	53		

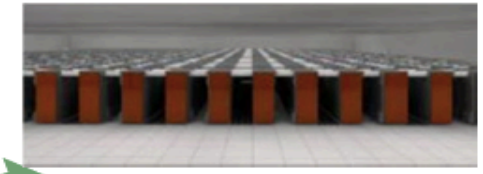
Quiz: How Many of the Top500 systems use GPUs?

Japanese K Computer

K computer Specifications

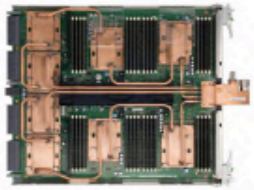
CPU (*SPARC64 VIIIfx*)	Cores/Node	8 cores (@2GHz)
	Performance	128GFlops
	Architecture	SPARC V9 + HPC extension
	Cache	L1(I/D) Cache : 32KB/32KB L2 Cache : 6MB
	Power	58W (typ. 30 C)
	Mem. bandwidth	64GB/s.
Node	Configuration	1 CPU / Node
	Memory capacity	16GB (2GB/core)
System board(SB)	No. of nodes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



System
LINPACK 10 PFlops
over 1PB mem.
800 racks
80,000 CPUs
640,000 cores

Rack
12.3 TFlops
15TB memory



System Board
512 GFlops
64 GB memory

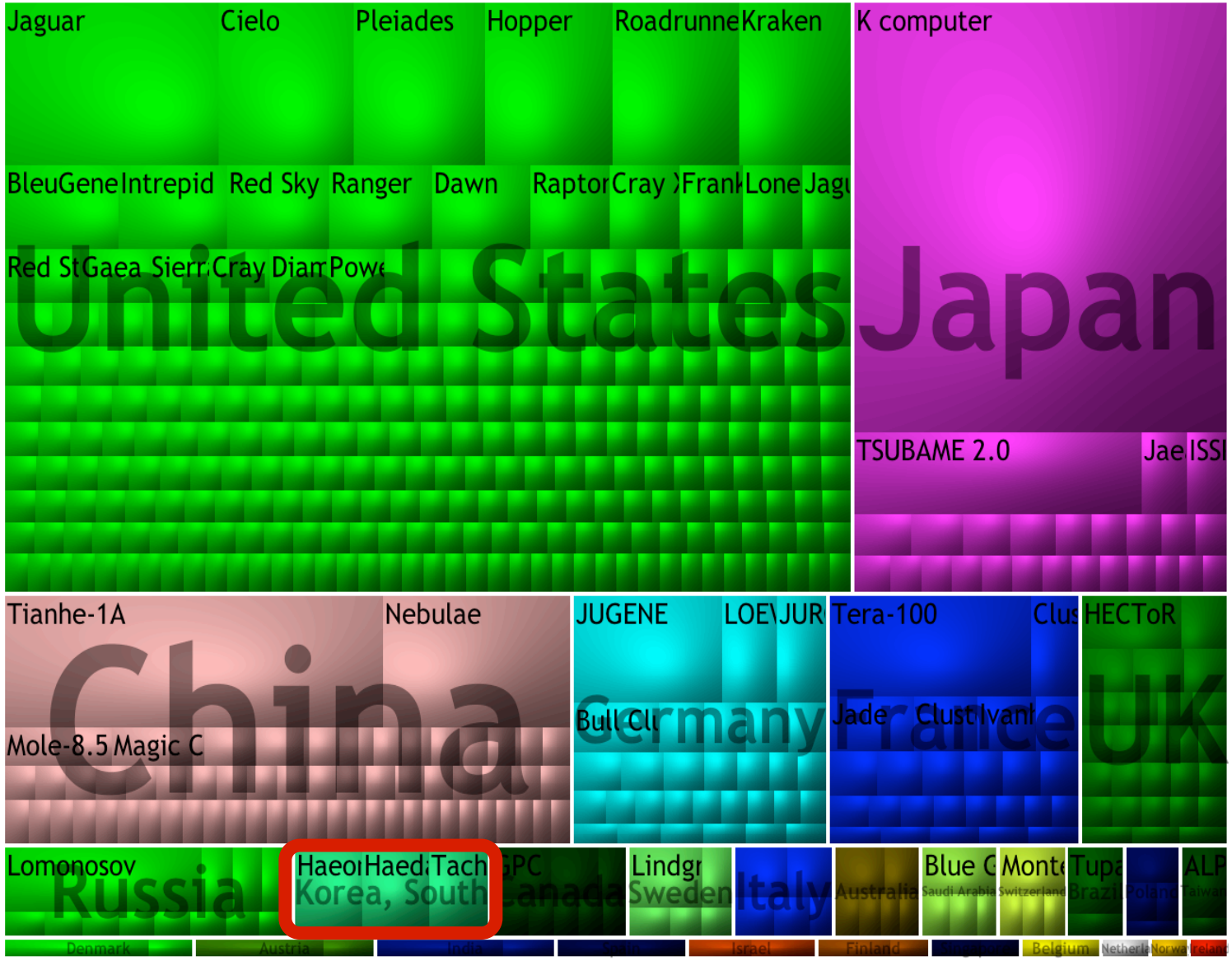
Node
128 GFlops
16GB Memory
64GB/s Memory band width

CPU
128GFlops
SPARC64™ VIIIfx
8 Cores@2.0GHz

* ICC : Interconnect Chip

New Linpack run with 705,024 cores at 10.51 Pflop/s (88,128 CPUs)

Countries Share



Absolute Counts

US:	251
China:	64
Germany:	31
UK:	28
Japan:	26
France:	25
Korea:	3

Top500 Systems in Korea

Rank	Site	Manufacturer	Computer	Cores	RMax	RPeak
20	Korea Meteorological Administration	Cray Inc.	Cray XE6 12-core 2.1 GHz	45120	316400	379008
21	Korea Meteorological Administration	Cray Inc.	Cray XE6 12-core 2.1 GHz	45120	316400	379008
26	KISTI Supercomputing Center	Oracle	Sun Blade x6048, X6275, IB QDR	26232	274800	307439

China

- First Chinese Supercomputer to use a Chinese Processor
  - Sunway BlueLight MPP
  - ShenWei SW1600 processor, 16 core, 65 nm, fabbed in China
  - 125 Gflop/s peak
  - In the Top20 with 139,364 cores & 1.07 Pflop/s Peak
  - Power Efficiency 741 Mflops/W

- Coming soon, Loongson (Godson) processor
  - 8-core, 65nm Loongson 3B processor runs at 1.05 GHz, with a peak performance of 128 Gflop/s

The New York Times

China Has Homemade Supercomputer Gain

By JOHN MARKOFF
Published: October 28, 2011

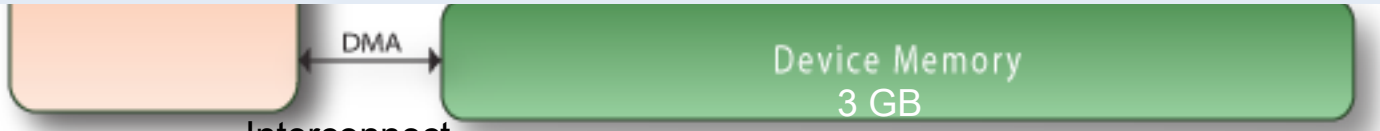
China has made its first supercomputer based on Chinese microprocessor chips, an advance that surprised high-performance computing specialists in the United States.

Commodity plus Accelerator

Commodity	Accelerator (GPU)
Intel Xeon	Nvidia C2050 "Fermi"
8 cores	448 CUDA cores
3 GHz	1.15 GHz
8*4 ops/cycle	448 ops/cycle
96 Gflop/s (DP)	515 Gflop/s (DP)

Quiz: How Many of the Top500 systems use GPUs?

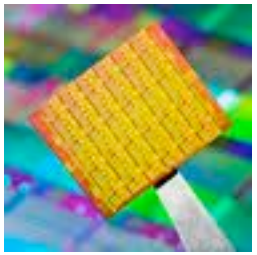
Answer:
Today only 19 systems on the TOP500 use GPUs



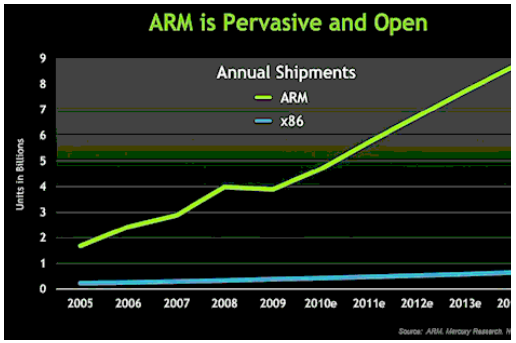
Interconnect
PCI-X 16 lane
64 Gb/s
1 GW/s

Future Computer Systems

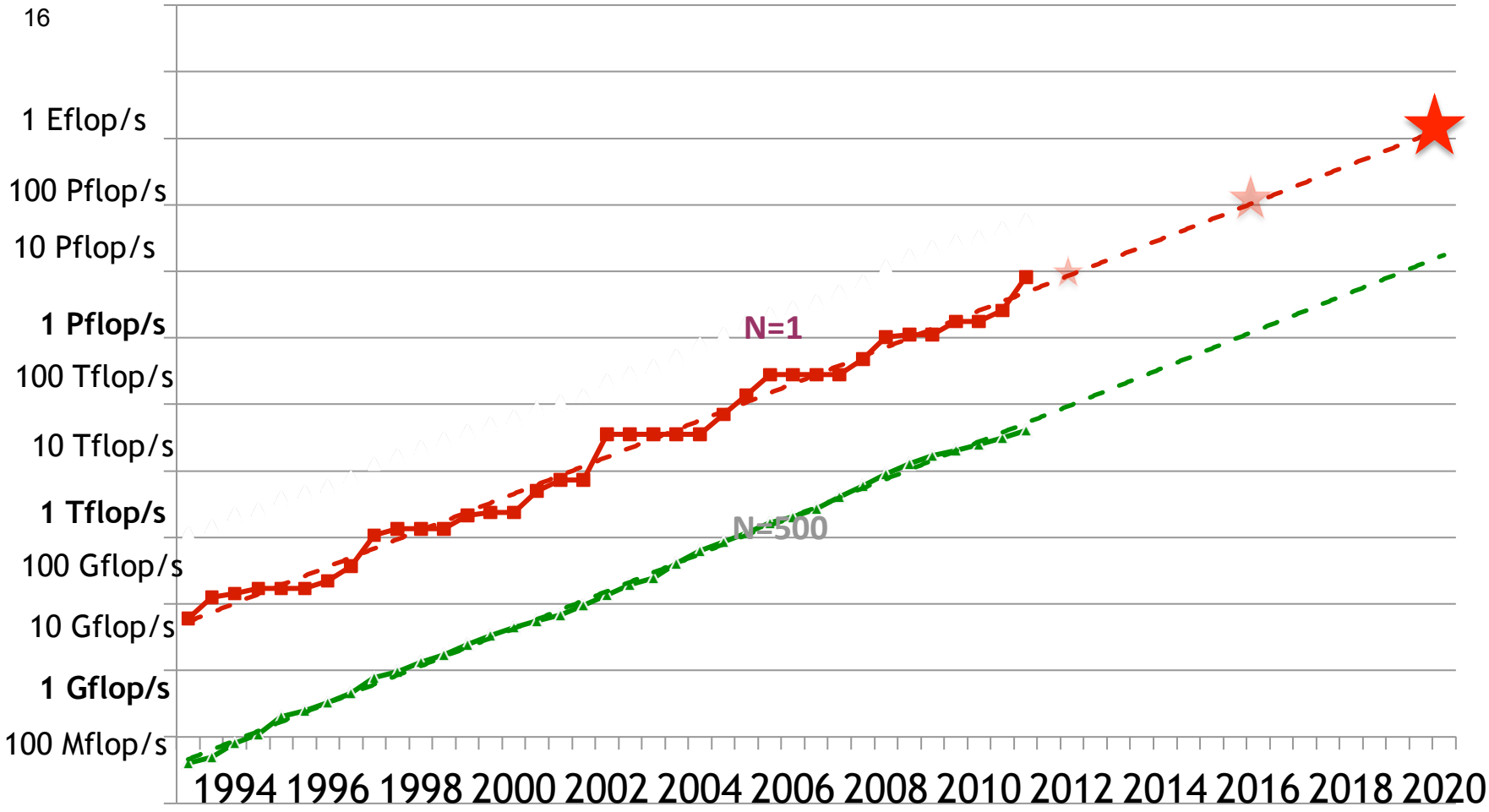
- Most likely be a hybrid design
  - Think standard multicore chips and accelerator (GPUs)
- Today accelerators are attached
- Next generation more integrated
- Intel's MIC architecture "Knights Ferry" and "Knights Corner" to come.
  - 48 x86 cores
- AMD's Fusion
  - Multicore with embedded graphics ATI
- Nvidia's Project Denver plans to develop an integrated chip using ARM architecture in 2013.



AMD
The future is fusion



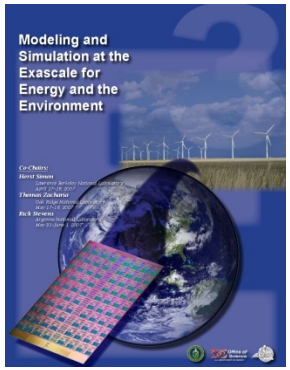
Performance Development in Top500

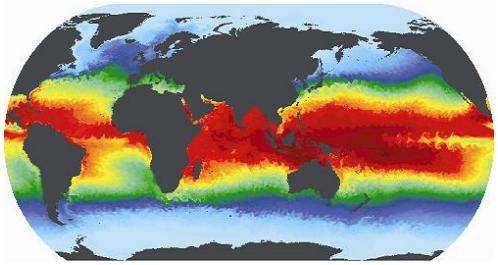


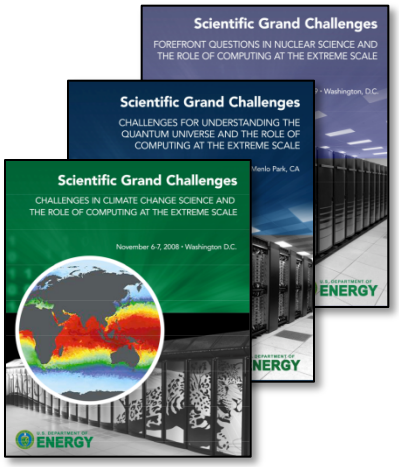
Broad Community Support and Development of the Exascale Initiative Since 2007

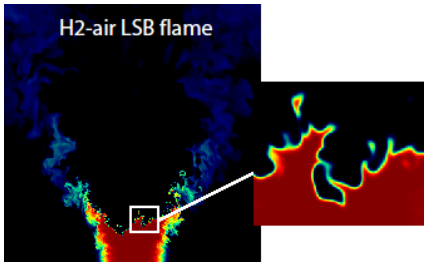
http://science.energy.gov/ascr/news-and-resources/program-documents/

- **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)

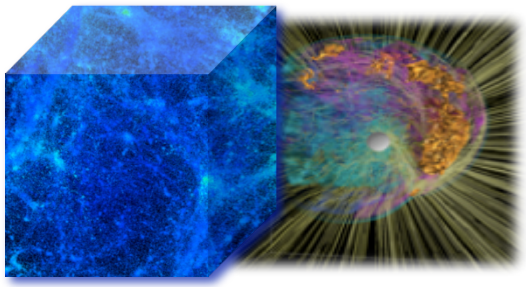








Mission Imperatives



Fundamental Science

Potential System Architecture

Systems	2011 K Computer
System peak	**8.7 Pflop/s**
Power	**10 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)	68,544
Total concurrency	548,352
MTTI	days

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

Systems	2011 K Computer	2019	Difference Today & 2019
System peak	**8.7 Pflop/s**	**1 Eflop/s**	**O(100)**
Power	**10 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)	68,544	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	548,352	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.

International Exascale Software Project

- Overall goal:
  - To develop a plan for producing a software infrastructure capable of supporting exascale applications
- Meetings so far:
  1. Santa Fe, NM, US, April 2009
  2. Paris, France, June 2009
  3. Tsukuba, Japan, October 2009
  4. Oxford, UK, April 2010
  5. Maui, HI, US, October 2010
  6. San Francisco, US, April 2011
  7. Cologne, Germany, October 2011
  8. Kobe, Japan April 2012
- SC08 (Austin), SC09 (Portland), ISC10 (Hamburg), SC10 (New Orleans), ISC11 (Hamburg), SC11 (Seattle)

IESP Makeup

- Attendees from universities, research institutes, government, funding agencies, research councils, hardware and software vendors, industry
- 65 - 85 participants per workshop
- Rough distribution per meeting is rather constant:
  - 70% universities/research institutes
  - 15% vendors/industry
  - 15% government/funding agencies

- Steering Committee
  - Jack Dongarra, U of Tennessee/Oak Ridge National Lab, US
  - Pete Beckman, Argonne Nat. Lab, US
  - Franck Cappello, INRIA, FR
  - Thom Dunning, NCSA, US
  - Thomas Lippert, Jülich Supercomputing Centre, DE
  - Satoshi Matsuoka, Tokyo Inst. of Tech, JP
  - Paul Messina, Argonne Nat. Lab, US
  - Patrick Aerts, Netherlands Organization for Scientific Research, NL
  - Anne Trefethen, Oxford, UK
  - Mateo Valero, Barcelona Supercomptuing Ceneter, Spain

Initial Objectives

- The IESP software roadmap is a planning instrument designed to enable the international HPC community to improve, coordinate and leverage their collective investments and development efforts.
- To develop a plan for producing a software infrastructure capable of supporting exascale applications
  - Thorough assessment of needs, issues and strategies
  - Develop a coordinated software roadmap
  - Provide a framework for organizing the software research community
  - Engage vendors to coordinate on how to deal with anticipated scale
  - Encourage and facilitate collaboration in education and training

International Community Effort

- 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 coordinated with software development

IESP Discussion Topics

- Many topics have been discussed and brought forward during the IESP workshops
  - Science drivers
  - Applications
  - Software stack
  - Open source
  - Vendor involvement
  - Standardization activities
  - Co-design
  - Exascale software centers
  - Initiatives
  - Funding and governance models

IESP Roadmap

- **IESP has created version 1.0 and 1.1 of a software roadmap:**
  - Science and technology trends
  - Software stack components (with cross-cutting aspects)
    - Not only the usual suspects (MPI, compilers, OpenMP, …)
    - But also new challenges, like resiliency, power, bit-wise reproducibility, …
  - Application involvement
    - Algorithms
    - Data Analysis, Visualization & Management
    - Co-design vehicles
  - Vendor involvement
  - Organization and governance

Roadmap Components

4.1 Systems Software

- 4.1.1 Operating systems
- 4.1.2 Runtime Systems
- 4.1.3 I/O systems
- 4.1.4 Systems Management
- 4.1.5 External Environments

4.2 Development Environments

- 4.2.1 Programming Models
- 4.2.2 Frameworks
- 4.2.3 Compilers
- 4.2.4 Numerical Libraries
- 4.2.5 Debugging Tools

4.3 Applications

- 4.3.1 Application Element: Algorithms
- 4.3.2 Application Support: Data Analysis and Visualization
- 4.3.3 Application Support: Scientific Data Management

4.4 Crosscutting Dimensions

- 4.4.1 Resilience
- 4.4.2 Power Management
- 4.4.3 Performance Optimization
- 4.4.4 Programmability

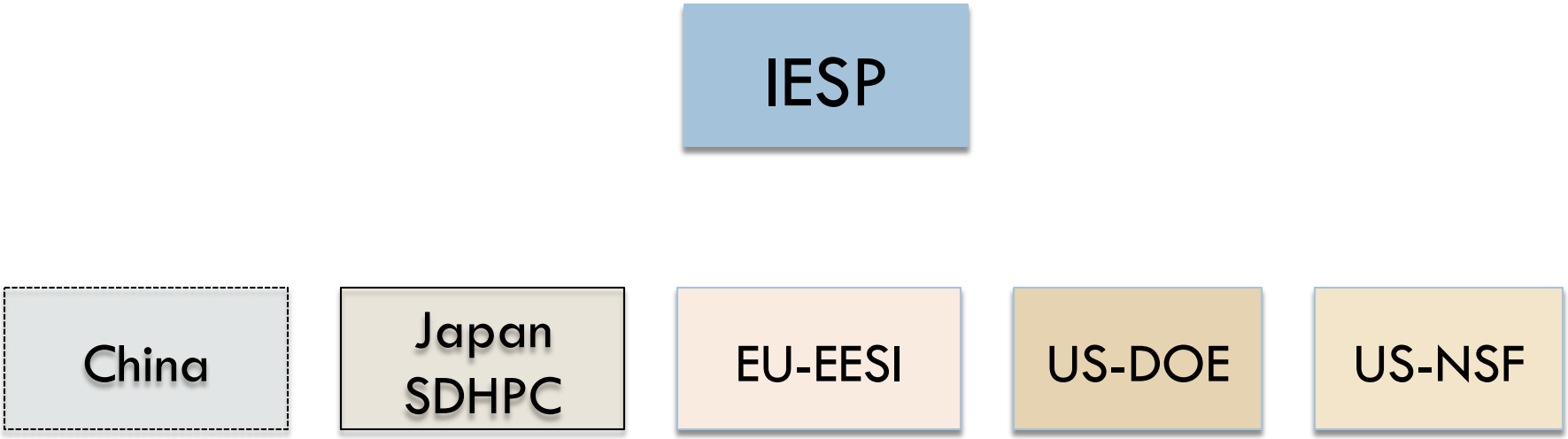
Each component subdivided in capabilities
Each with a level of uniqueness and criticality for exascale

see IJHPCA, Feb 2011, http://hpc.sagepub.com/content/25/1/3

Output from IESP

- Broad international community involvement
- Software Roadmap
- Application inventory from the HPC centers
- Software stack in use at major HPC centers
- Framework for international software development
- Survey of computational science and engineering educational programs (just started)

Example Organizational Structure:



- IESP provides coordination internationally, while regional groups have well managed R&D plans and milestones

What Next? (1/3)

Moving from "What to Build" to "How to Build"

- Technology
  - Defining and developing the roadmap for software and algorithms on extreme-scale systems
  - Setting a prioritized list of software components for Exascale computing as outlined in the Roadmap
  - Assessing the short-term, medium-term and long-term software and algorithm needs of applications for peta/exascale systems

What Next? (2/3)

Moving from "What to Build" to "How to Build"

- Organization
  - Exploring ways for funding agencies to coordinate their support of IESP-related R&D so that they complement each other
  - Exploring how laboratories, universities, and vendors can work together on coordinated HPC software
  - Creating a plan for working closely with HW vendors and application teams to co-design future architectures

What Next? (3/3)

Moving from "What to Build" to "How to Build"

- Execution
  - Developing a strategic plan for moving forward with the Roadmap
  - Creating a realistic timeline for constructing key organizational structures and achieving initial goals
  - Exploring community development techniques and risk plans to ensure key components are delivered on time
  - Exploring key components of any needed Intellectual Property agreements

Next Steps

- Revise and extend initial draft
- Build collaboration plans
- Work with funding agencies to plan research activities
- Next workshop
  - April 12-13, 2012
  - Kobe Japan
- Roadmap available at:
  - www.exascale.org

SPONSORS

INTERNATIONAL EXASCALE SOFTWARE PROJECT ROADMAP

Jack Dongarra, Pete Beckman, Terry Moore, Patrick Aerts, Giovanni Aloiso, Jean-Claude Andre, Jean-Yves Berthou, Taisuke Boku, Bertrand Braunschweig, Franck Cappello, Barbara Chapman, Xuebin Chi, Alok Choudhary, Sudip Dosanjh, Thom Dunning, Sandro Fiore, 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, Alain Lichnewsky, Thomas Lippert, Bob Lucas, Barney Maccabe, Satoshi Matsuoka, Paul Messina, Peter Michielse, Bernd Mohr, Matthias Mueller, Wolfgang Nagel, Hiroshi Nakashima, Michael E. Papka, Dan Reed, Mitsuhisa Sato, Ed Seidel, John Shalf, David Skinner, Marc Snir, Thomas Sterling, Rick Stevens, Fred Streitz, Bob Sugar, Shinji Sumimoto, William Tang, John Taylor, Rajeev Thakur, Anne Trefethen, Mateo Valero, Aad van der Steen, Jeffrey Vetter, Peg Williams, Robert Wisniewski, Kathy Yelick