Planned Developments of High End Systems Around the World

Jack Dongarra
INNOVATIVE COMPUTING LABORATORY

University of Tennessee
Oak Ridge National Laboratory
University of Manchester

Planned Development of HPC

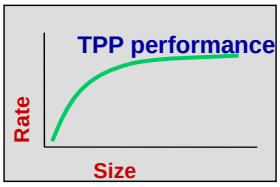
- **Quick look at current state of HPC through the "eyes" of the Top500**
- **The Japanese Efforts**
- **The European Initiatives**
- **The state of China's HPC**
- **India's machine**

TOP500 supercomputer

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

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

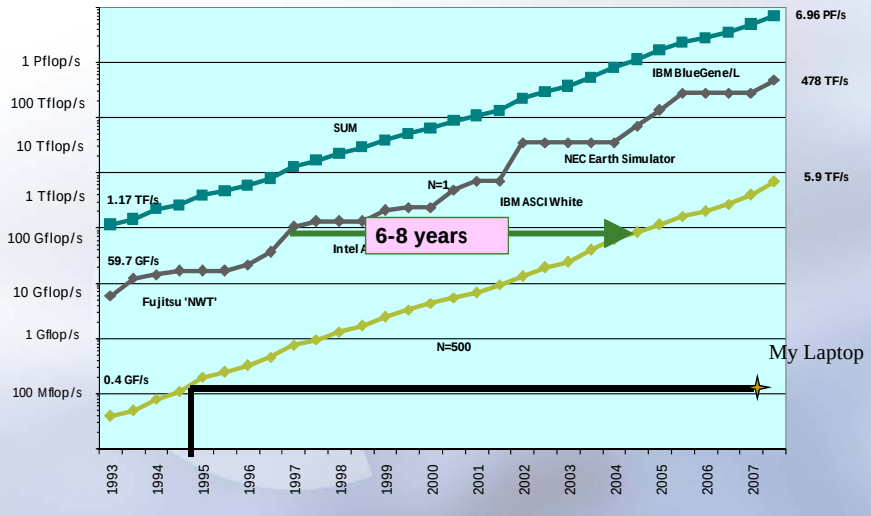
  $Ax=b$, *dense problem*



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

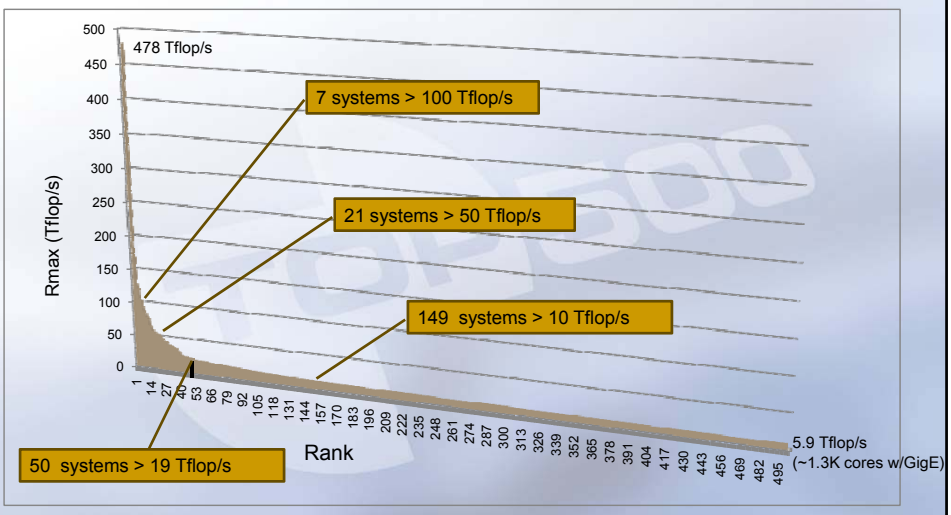
3

TOP500 Performance Development



4

Top500 Systems November 2007

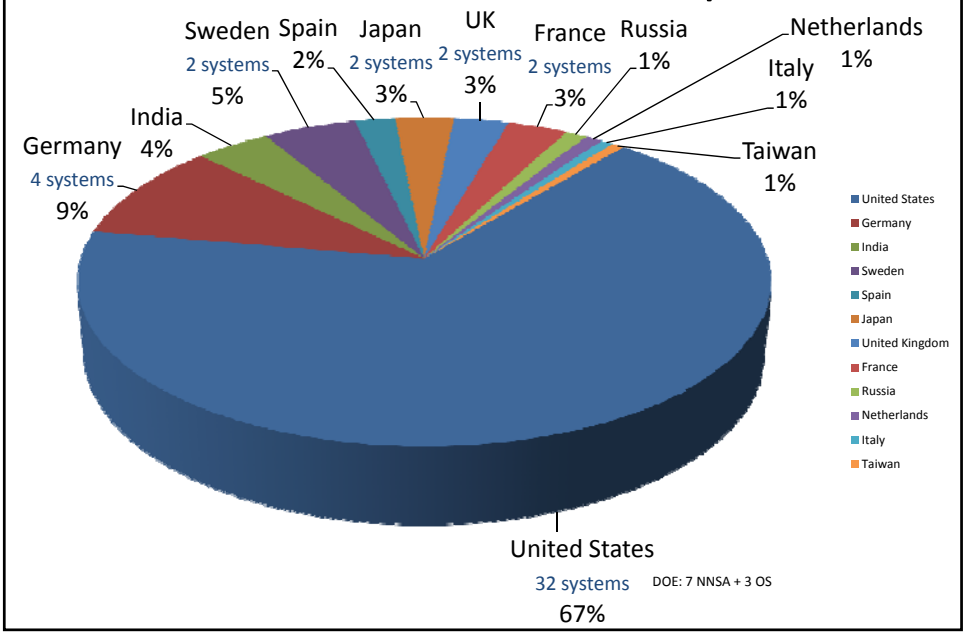


5

30th Edition: The TOP10

	Manufacturer	Computer	Rmax [TF/s]	Installation Site	Country	Year	#Cores
1	IBM	Blue Gene/L eServer Blue Gene Dual Core .7 GHz	478	DOE Lawrence Livermore Nat Lab	USA	2007 Custom	212,992
2	IBM	Blue Gene/P Quad Core .85 GHz	167	Forschungszentrum Jülich	Germany	2007 Custom	65,536
3	SGI	Altix ICE 8200 Xeon Quad Core 3 GHz	127	SGI/New Mexico Computing Applications Center	USA	2007 Hybrid	14,336
4	HP	Cluster Platform Xeon Dual Core 3 GHz	118	Computational Research Laboratories, TATA SONS	India	2007 Commod	14,240
5	HP	Cluster Platform Dual Core 2.66 GHz	102.8	Government Agency	Sweden	2007 Commod	13,728
6	Cray	Opteron Dual Core 2.4 GHz	102.2	DOE Sandia Nat Lab	USA	2007 Hybrid	26,569
7	Cray	Opteron Dual Core 2.6 GHz	101.7	DOE Oak Ridge National Lab	USA	2006 Hybrid	23,016
8	IBM	eServer Blue Gene/L Dual Core .7 GHz	91.2	IBM Thomas J. Watson Research Center	USA	2005 Custom	40,960
9	Cray	Opteron Dual Core 2.6 GHz	85.4	DOE Lawrence Berkeley Nat Lab	USA	2006 Hybrid	19,320
10	IBM	eServer Blue Gene/L Dual Core .7 GHz	82.1	Stony Brook/BNL, NY Center for Computational Sciences	USA	2006 Custom	36,864

Performance of the Top50



DOE NNSA

- **LLNL**
  - IBM BG/L
    - Power PC
    - Cores: 212,992
    - Peak: 596 TF
    - Memory: 73.7 TB
  - IBM Purple
    - Power 5
    - Cores: 12,208
    - Peak: 92.8 TF
    - Memory: 48.8 TB
- **SNL**
  - Red Storm Cray
    - AMD Dual Core
    - Cores: 27,200
    - Peak: 127.5 TF
    - Memory: 40 TB
  - Thunderbird Dell
    - Intel Xeon
    - Cores: 9,024
    - Peak: 53 TF
    - Memory: 6 TB
- **LANL**
  - RoadRunner IBM
    - AMD Dual Core
    - Cores: 18,252
    - Peak: 81.1 TF
    - Memory: 27.6 TB
  - Q HP
    - Alpha
    - Cores: 8,192
    - Peak: 20.5 TF
    - Memory: 13 TB





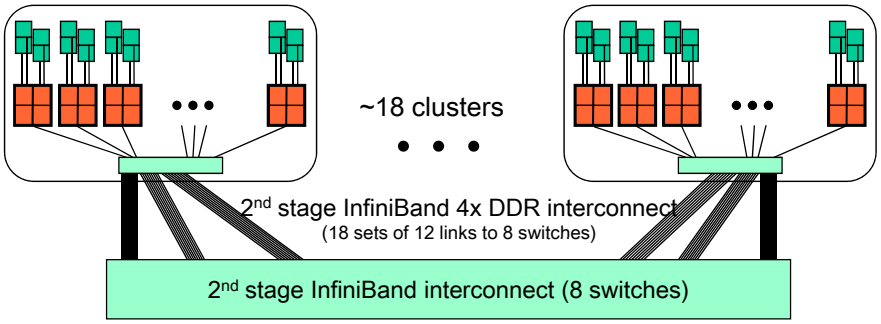

8

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



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

Approval by DOE 12/07
First CU being built today
Expect a May Pflop/s run
Full system to LANL in December 2008

DOE OS

- ORNL
  - Jaguar Cray XT
    - AMD Dual Core
    - Cores: 11,706
    - Peak: 119.4 TF
      - Upgrading 250 TF
    - Memory: 46 TB
  - Phoenix Cray X1
    - Cray Vector
    - Cores: 1,024
    - Peak: 18.3 TF
    - Memory: 2 TB

- LBNL
  - Franklin Cray XT
    - AMD Dual Core
    - Cores: 19,320
    - Peak: 100.4 TF
    - Memory: 39 TB
  - Bassi IBM
    - PowerPC
    - Cores: 976
    - Peak: 7.4 TF
    - Memory: 3.5 TB
  - Seaborg IBM
    - Power3
    - Cores:
    - Peak: 9.9 TF
    - Memory: 7.3 TB

- ANL
  - BG/P IBM
    - PowerPC
    - Cores: 131,072
    - Peak: 111 TF
    - Memory: 65.5 TB

10

NSF HPC Systems available on TeraGrid
10/01/2007

High Performance Computing Systems

Name	Institution	System	CPUs	Peak TFlops	Memory TBytes	Disk TBytes
Abe	NCSA	Dell Intel 64 Linux Cluster	9600	89.47	9.38	100.00
Lonestar	TACC	Dell PowerEdge Linux Cluster	5840	62.16	11.60	106.50
Big Red	IU	IBM e1350	3072	30.60	6.00	266.00
BigBen	PSC	Cray XT3	4136	21.30	4.04	100.00
Blue Gene	SDSC	IBM Blue Gene	6144	17.10	1.50	19.50
Tungsten	NCSA	Dell Xeon IA-32 Linux Cluster	2560	16.38	3.75	109.00
DataStar p655	SDSC	IBM Power4 + p655	2176	14.30	5.75	115.00
TeraGrid Cluster	NCSA	IBM Itanium2 Cluster	1744	10.23	4.47	60.00
Lear	Purdue	Dell EM64T Linux Cluster	1024	6.60	2.00	28.00
Cobalt	NCSA	SGI Altix	1024	6.55	3.00	100.00
Frost	NCAR	IBM BlueGene/L	2048	5.73	0.51	6.00
TeraGrid Cluster	SDSC	IBM Itanium2 Cluster	524	3.10	1.02	48.80
Copper	NCSA	IBM Power4 p690	384	2.00	1.44	30.00
DataStar p690	SDSC	IBM Power4 + p690	192	1.30	0.88	115.00
TeraGrid Cluster	UC/ANL	IBM Itanium2 Cluster	128	0.61	0.24	4.00
NSTG	ORNL	IBM IA-32 Cluster	56	0.34	0.07	2.14
Rachel	PSC	HP Alpha SMP	128	0.31	0.50	6.00
		Total:	40780	288.08	56.15	1215.94

Does not show:
LSU: Queen Bee
TACC: Ranger
Tennessee: Cray XT/Baker

NSF - New TG systems

System	Peak TF/s	Memory (TB)	Type
LSU Queen Bee	50.7	5.3	680n 2s 4c Dell 2.33GHz Intel Xeon 8-way SMP cluster; 8GB/node; IB
UT-TACC Ranger	504	123	Sun Constellation - 3936n 4s 4c 2.0GHz AMD Barcelona - 16-way SMP cluster; 32GB/node; IB
UTK/ORNL Track 2b	164 1,000	17.8 80	Cray XT4 - 4456n 1s 4c AMD Budapest (April 2008) Cray Baker (80,000 cores) expected 2Q 09
?? Track 2c			Proposals under evaluation today
UIUC Track 1	Sustained Pflop/s		To be deployed in 2011

Japanese Efforts

- **TiTech Tsubame**
- **T2K effort**
- **Next Generation Supercomputer Effort**

13

TSUBAME as No.1 in Japan since June 2006

東京工業大学
Tokyo Institute of Technology

Originally: 85 TFlop/s Today:103 TFlop/s Peak
1.1 Pbyte (now 1.6 PB)
4 year procurement cycle, $7 mil/y

Has beaten the Earth Simulator

Has beaten all the other Univ. centers *combined*

Sun Galaxy 4 (Opteron Dual core 8-socket)
10480core/655Nodes
32-128GB
21.4TBytes
50.4TFlop/s
OS Linux (SuSE 9, 10)
NAREGI Grid MW

ClearSpeed CSX600 SIMD accelerator
~~360~~ **648 boards,**
~~35~~ **52.2TFlop/s**

1.5PB ~~1.0~~ Pbyte (Sun "Thumper")
0.1Pbyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV (over IP)
60GB/s ~~50~~GB/s aggregate I/O BW

Voltaire ISR9288 Infiniband
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)

14

Universities of Tsukuba, Tokyo, Kyoto (T2K)

- **The results of the bidding announced on December 25, 2007.**
  - **The specification requires a commodity cluster with quadcore Opteron (Barcelona).**
- **Three systems share the same architecture on each site**
  - **Based on the concept of Open Supercomputer**
    - **Open architecture, (commodity x86)**
    - **Open software, (Linux, open source)**
- **University of Tokyo: 140 Tflop/s (peak) from Hitachi**
- **University of Tsukuba: 95 Tflop/s (peak) from Cray Inc.**
- **Kyoto University: 61 Tflop/s (peak) from Fujitsu**
- **They will be installed in summer 2008.**
- **Individual procurement : Not a single big procurement for all three systems**

NEC SX-9 Peak 839 Tflop/s

- **102.4 Gflop/s per cpu**
- **16 cpu per unit**
- **512 units max.**
- **Expected ship in March 2008**

- **German Weather Service (DWD)**
  - **39 TF/s, €39 M, operational in 2010.**
- **Meteo France**
  - **sub-100 TF/s system**
- **Tohoku University, Japan**
  - **26 TF/s**

16

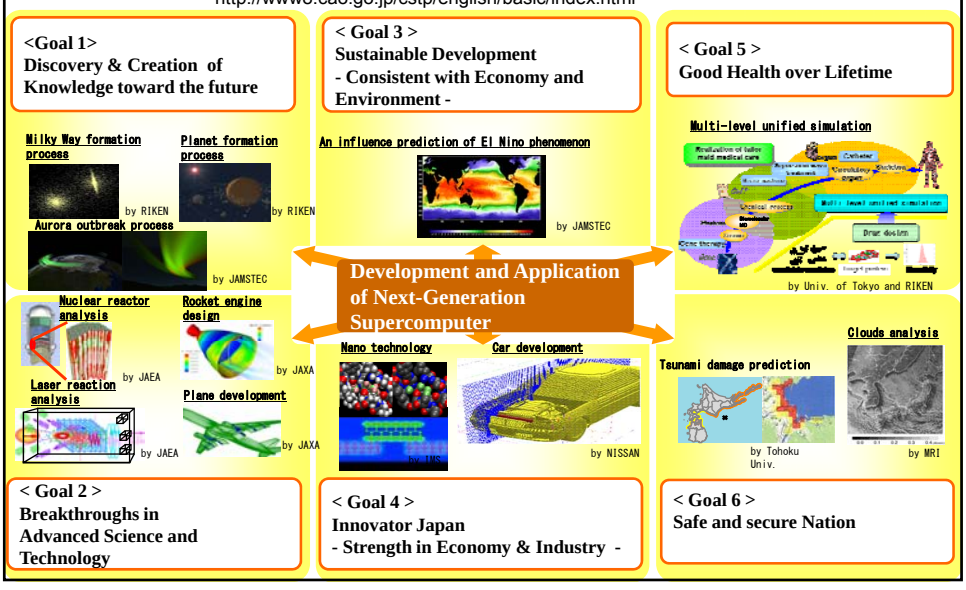
Japanese Efforts:
The Next Generation Supercomputer Project

- **Roughly every 5-10 years Japanese government puts forward a Basic Plan for S&T**
  - **http://www8.cao.go.jp/cstp/english/basic/index.html**
- **Today "3rd Science and Technology Basic Plan"**
- **The 2nd S&T Plan gave rise to the Earth Simulator**

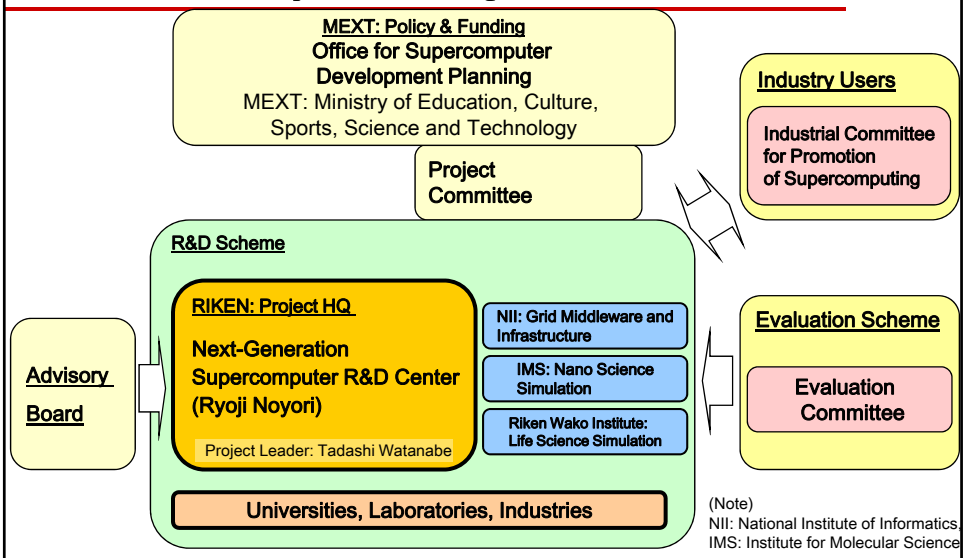
17

Six Goals of Japan's "3rd Science and Technology Basic Plan" and Next-Generation Supercomputer Project

http://www8.cao.go.jp/cstp/english/basic/index.html



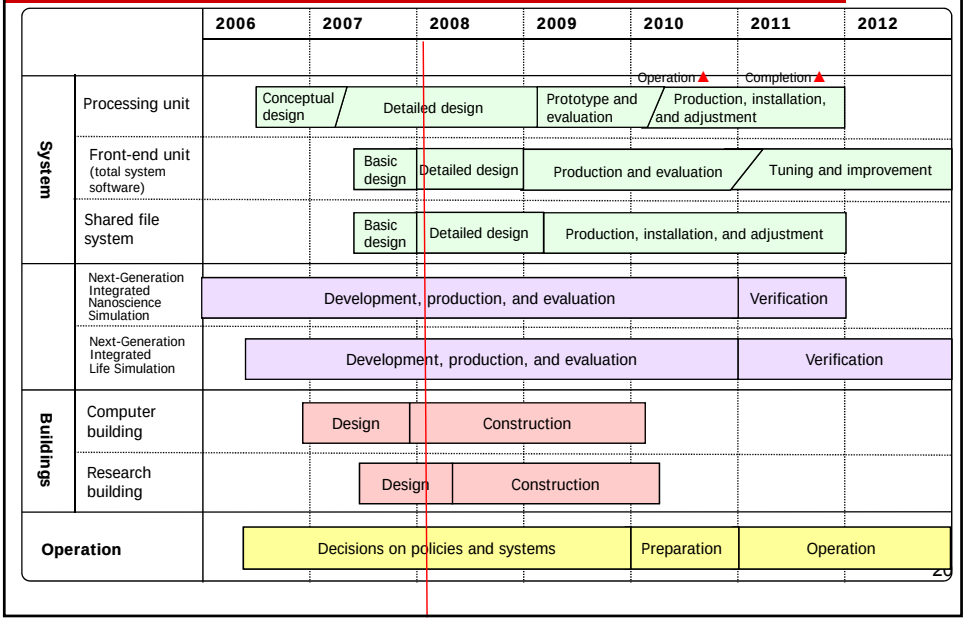
Project Organization

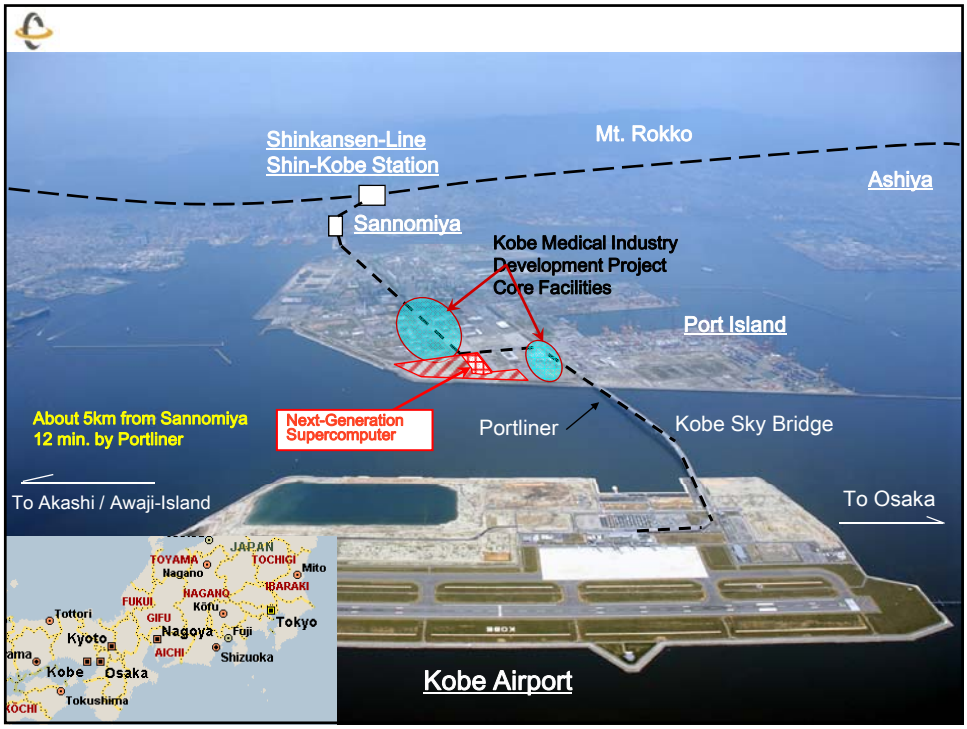


Total Budget: about 115 billion Yen (~ $ 1 billion)

19

Next Generation Supercomputer Schedule



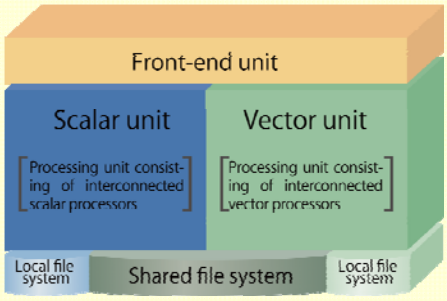


The Next-Generation Supercomputer project

Due to be ready in 2012, the peta-scale computing by the new supercomputer will ensure that Japan continues to lead the world in science and technology, academic research, industry, and medicine.

System architecture is a heterogeneous computing system with scalar and vector units connected through a front-end unit which is now being defining

[System configuration]

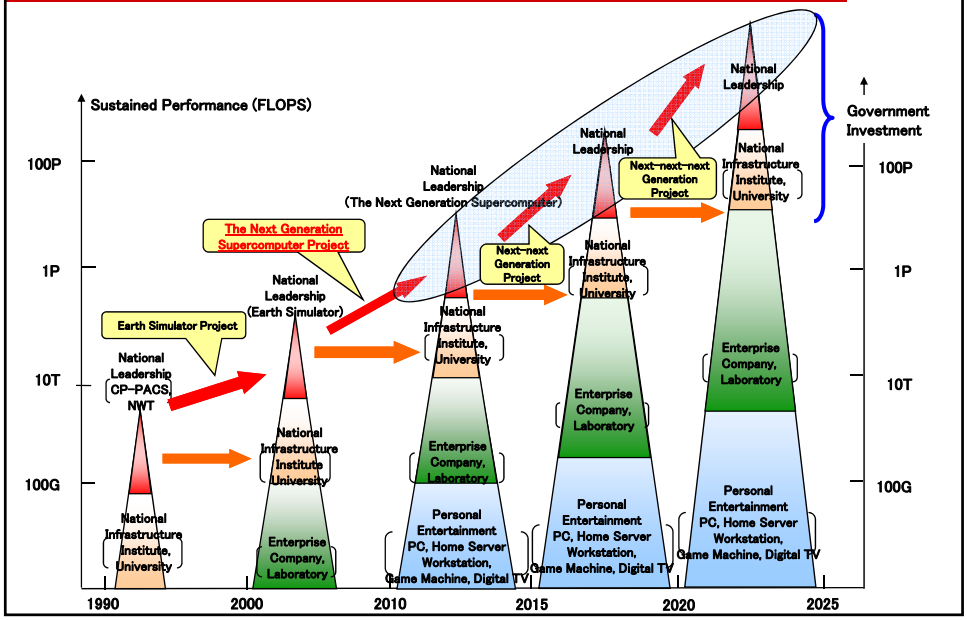


The Next-Generation Supercomputer will be hybrid general-purpose supercomputer that provides the optimum computing environment for a wide range of simulations.

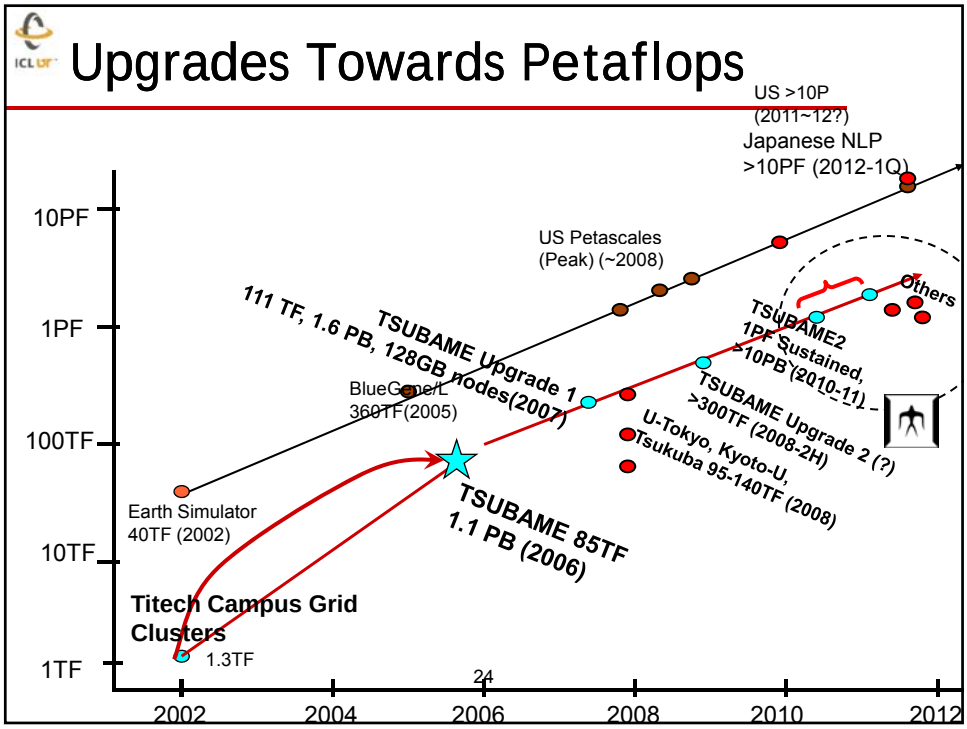
- Calculations will be performed in processing units that are suitable for the particular simulation.
- Parallel processing in a hybrid configuration of scalar and vector units will make larger and more complex simulations possible.

22

MEXT's Vision for Continuous Development of Supercomputers



Upgrades Towards Petaflops



European Systems

- **France: 2 machines in the Top50 (CEA)**
  - **CEA has 2 systems from Bull**
    - **Itanium, Quadrics, 9968 cores, 53 Tflops/s peak in 2006**
    - **Itanium, Infiniband, 7680 cores, 42 Tflop/s peak in 2007**
    - **Expected to acquire a Pflop/s system in 2010.**
  - **CNRS – IDRIS (Institut du Développement et des Ressources en Informatique Scientifique)**
    - **IBM BG/P (10 rack) 139 Tflop/s peak**
    - **IBM Power6 68 Tflop/s peak**
    - **1/08 Installed, full operation 3/08**
  - **EDF**
    - **IBM BG/P (8 rack) 111 Tflop/s peak**
    - **1/08 installed, full operation 6/08**
  - **CINES (Montpellier)**
    - **Center funded by the ministry of research**
    - **RFP for a 50 Tflop/s system**

European Systems (continued)

- **England: 2 machines Top50 (Edinburgh #17 & AWE #35)**
  - **U of Edinburgh's HECToR 63.4 Tflop/s Cray XT4 system today, going to 250 Tflop/s in 2009, £113M**
  - **ECMWF, 2 IBM POWER6 systems to be installed total 290 Tflop/s in 2008**
- **Netherland: 1 machine in the Top50 (Groningen #37)**
  - **SARA (Stichting Academisch Rekencentrum) to upgrade from 14 to 60 Tflop/s (Power6) in May 2008.**
- **Spain: 1 machine in the Top50 (Barcelona #13)**
  - **Barcelona, PowerPC w/Myrinet, 10K processors, 94 Tflop/s peak since 2006**
- **Finland: No machines in the Top50**
  - **CSC has a "new" 70 Tflop/s Cray XT and a 10 Tflop/s HP cluster**

26

European Systems (continued)

- **Sweden: 2 machines in Top50 (#'s 5 & 23)**
  - **The National Defense Radio Establishment**
    - **HP Cluster, 146 Tflop/s peak**
  - **Computer Center, Linköping University**
    - **HP Cluster, 60 Tflop/s peak**
- **Italy: 1 machine in Top50 (#48)**
  - **CINECA**
    - **IBM Cluster, 61 Tflop/s peak**
- **Russia: 1 machine in Top50 (#33)**
  - **Joint Supercomputer Center**
    - **HP Cluster, 45 Tflop/s peak**

27

European Systems (continued)

- **Germany: 4 machines in the top50 (#'s 2, 15, 28 and 40)**
  - **2 BG/P and a BG/L (FZJ and MPI) also SGI Altix (LRZ Munich)**
  - **HLRN (6 North German States) SGI Altix**
    - **70 Tflop/s system (split between Berlin and Hannover) in Q2-2008**
    - **312 Tflop/s system in 2009**
    - **30 M € total**

  - **German Climate Computing Centre (DKRZ)**
    - **Planning a new IBM (Power6) with a peak speed of 140 Tflop/s in 2008**
  - **FZ Jülich**
    - **General purpose cluster > 200 Tflop/s (Intel w/Quadrics) in 2008**
    - **A Pflop/s system in 2009**
  - **HLRS University of Stuttgart**
    - **Planning for 1-2 Pflop/s in 2011**

28

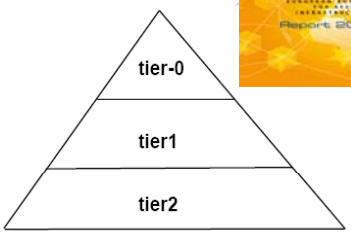
HPC now in European Research Infrastructures Roadmap

The European HPC infrastructure need was recognized in the ESFRI Roadmap (2006)

- Estimated construction cost of 200-400 M€
- Indicative running cost of 100–200 M€ / year
- High end should be renewed every 2-3 years
- Close links to national/regional centers to establish a European HPC ecosystem

European High Performance Computing Service includes:

- Capability Computing
- Grid architectures
- Software
- Data management and curation



There is a need for a combination of centralized, distributed, and networked aspects, based on a pyramid-like organization, including a few very high-end centres at the top

- HPC Provisioning pyramid
  - Tier-0: 3-5 European HPC-facilities
  - Tier-1: National HPC-facilities
  - Tier-2: Regional / University Centres

EUROPEAN ROADMAP FOR RESEARCH INFRASTRUCTURES
Report 2006

The Partnership for Advanced Computing in Europe (PRACE) Initiative

- The PRACE MoU has been signed by the representatives of 14 European countries
- The goals:
  - Prepare an European structure funding and operating a permanent Tier 0 HPC Infrastructure
  - Provide a smooth insertion in the European HPC Ecosystem of national and topical centres, networking incl. GEANT and DEISA, user groups and communities.
  - Joint endeavours, incl. a FP7 « Preparatory Phase ».
  - Promote the most effective use of Numerical Simulation at the leading edge
  - Promote European presence and competitiveness in HPC

PRACE Partnership for Advanced Computing in Europe

What is going to happen with PRACE?

- **Project will start 1.1.2008**
- **Consortium partners (14 countries)**
  - Austria, Finland, France, Germany, Greece, Italy, Norway, Poland, Portugal, Spain, Sweden, Switzerland, The Netherlands, United Kingdom
- **Two years, 10+10 MEUR volume**
- **Prototypes for petaflop computing during 2008-2009**
- **Target to have the first center operational in 2009-2010**
- **Open issues to be solved during the preparatory phase:**
  - Which companies to prototype and where to place them?
  - Who will host the petaflop centers?
  - Who will pay for construction?
  - Who can use the resources and under which conditions?
  - How to link with other projects, for example DEISA?

CSC

PRACE Cost Sharing – EU and Host Country

- **The host country will be determined by which government will invest the majority of cost (and also have access to majority of cycles).**
- **Primary partners (= willing to host) appear to be:**
  - **Germany, UK, France, Spain and the Netherlands.**

China

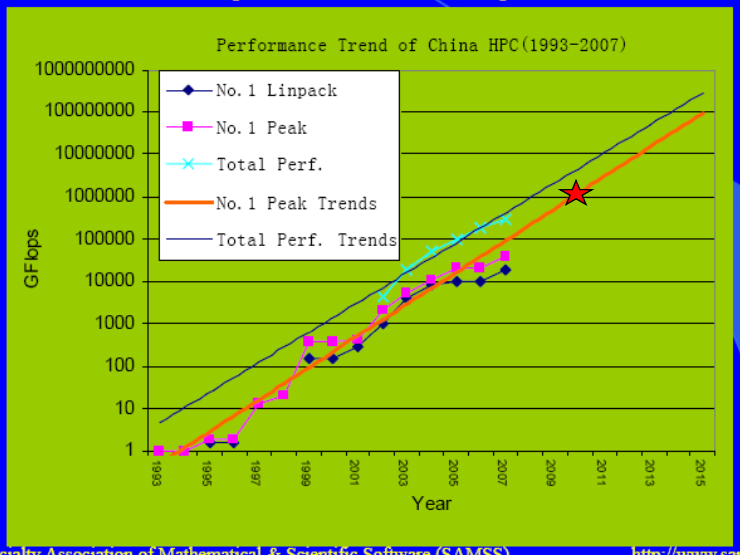
- A dozen national HPC centers at major universities (each a few TF) connected by gigabit level network
  - Research at universities is weak but improving
  - But ample numbers of CS graduates
- HPC Technical Committee to direct national priorities
- HPC Standardization Committee to coordinate and create Chinese standards (i.e., for blades, cluster OS, security, etc) with vendor participation

2007 China TOP100(1-10)

Rank	Vendor	System	Installation Site	Installation Year	Application Area	Processor Cores	Linpack (Gflops)	Linpack Source	Peak (Gflops)	Efficiency
1	IBM	IBM BladeCenter HS21 Cluster, Intel Xeon Woodcrest Dual Core 2.33 GHz/Giga-E	China Petroleum & Chemical Corporation	2007	Energy/Geophysics	4096	18600.00	T	38223.90	0.487
2	IBM	IBM SP eServer pSeries 655 (1.7 GHz Power4+) /Federation	China Meteorological Administration	2005	Meteorology	3200	10310.00	T	21760.00	0.474
3	DAWNING	DAWNING /640x4 Opteron 2.2GHz/ Myrinet	ShangHai Supercomputing Center	2004	Scientific Computing	2560	8061.00	Q/C	11264.00	0.716
4	SGI	Altix 4700/ Itanium2 Montecito 1.6GHz/ NUMALink4+Infiniband	China Meteorological Administration	2007	Meteorology	1280	7127	C	8192	0.870
5	HP	Blade Cluster BL-20P, Pentium4 Xeon 3.2GHz/ Giga-E	Gaming Company B-Shanghai 1	2007	Industry/ Game	1950	6976	C	12480	0.559
6	HP	Blade Cluster BL-20P, Pentium4 Xeon 3.2GHz/ Giga-E	Gaming Company B-Chengdu	2007	Industry/ Game	1950	6976	C	12480	0.559
7	HP	Blade Cluster BL-20P, Pentium4 Xeon 3.2GHz/ Giga-E	Gaming CompanyB-Shanghai 2	2007	Industry/ Game	1950	6976	C	12480	0.559
8	HP	Blade Cluster BL-20P, Pentium4 Xeon 3.2GHz/ Giga-E	Gaming Company B-Shanghai 3	2007	Industry/ Game	1950	6976	C	12480	0.559
9	HP	Blade Cluster BL-20P, Pentium4 Xeon 3.2GHz/ Giga-E	Gaming Company B-Beijing	2007	Industry/ Game	1950	6976	C	12480	0.559
10	HP	Blade Cluster BL-20P, Pentium4 Xeon 3.2GHz/ Giga-E	Game CompanyB-Xi'an	2007	Industry/ Game	1950	6976	C	12480	0.559

The Specialty Association of Mathematical & Scientific Software (SAMSS) http://www.samss.org.cn

Trend of China HPC Performance (1993-2007)



The Specialty Association of Mathematical & Scientific Software (SAMSS) http://www.samss.org.cn

Trends and Predictions for China HPC

- **Strong government commitment**
- **2008: 100 Tflop/s peak system will be in use**
- **2008 – 2009: Total Performance in China will be at 1 Pflop/s**
- **2010 – 2011: 1 Pflop/s peak machine will be in use.**

37

INDIA

India's #4 in the Top500 notwithstanding
China leads India in all aspects of HPC
- Infrastructure & facilities
- Diffusion into industry
- Local vendors
- Research output and quality
- Government commitment

India

- CRL (Computational Res Labs)
  - Pune facility, funded by Tata & Sons Inc
    - Tata: ~4% of India's GDP
    - History of long term investment in strategic national facilities.
      - Tata Inst of Science ➔ Indian Inst of Science (IISc) (100yrs)
      - Tata Inst of Fundamental Research (TIFR)
  - US$30M for large blade system from HP
    - #4 on Top500 (Nov 2007) 120TF Linpack (200TF peak)
    - Purchased and installed quickly in 3Q-4Q2007

India

- Universities & Govt labs
  - Weak HPC presence
    - Few large systems (IISc, TIFR have some HPC presence)
    - Researchers are not driven to push their problems to large HPC environments
    - Little credible HPC research
      - Few CS PhDs
      - Emphasis on searching technologies (i.e., for Google, Yahoo!, etc)
  - HiPC is best HPC meeting in the country. Most recent Dec 2007, found few HPC research achievements from Indian universities

Summary

- US dominates in the use of HPC
  - US dominates producing the components (processors, interconnects, and software) for HPC
- Japan will have a 10 Pflop/s system in 2010-2011
- Coordinated European effort will place a Pflop/s system soon
- India system is a one off, no national effort

41

Thanks

- Buddy Bland, ORNL
- David Kahaner, ATIP
- Kimmo Koski, CSC, Finland
- Thomas Lippert, Jülich, Germany
- Satoshi Matsuoka, TiTECH, Japan
- Hans Meuer, Mannheim, Germany
- Gerard Meurant, CEA, France
- JiaChang Sun, CAS, China
- Aad van der Steen, SARA, Netherlands
- Tadashi Watanabe, Riken, Japan

42