Recent Trends and Projects on High Performance Computing in Japan

Mitsuhisa Sato

Center for Computational Sciences /

Graduate School of Systems and Information Engineering

University of Tsukuba

Presentation Outline

- Recent Trends of Supercomputers in Japan
  - PACS CS, Center for Computational Sciences, U. of Tsukuba
  - Supercomputing Campus Grid" Core System, Tokyo Inst. Of Tech.
- Recent Projects
  - So-called "Kei Soku Keisanki" (10PFLOPS) project
  - Status and My view
- Summary

Where are supercomputers in Japan?

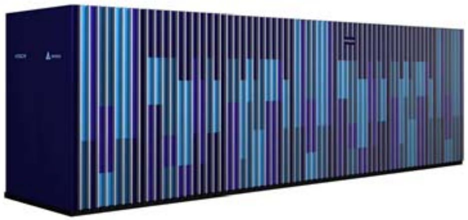
- Universities
  - 7 National University computer centers (Hokkaido, Touhoku, Tokyo, Nagoya, Kyoto, Osaka, Kyusyu) are providing services of shared computing resources (supercomputers) among universities.
  - U. of Tsukuba, Tokyo Institute of Technology (Titech), …

- Government Lab.
  - AIST, RIKEN, KEK, …
  - Meteorological Agency

- Industries
  - Automobile industries (TOYOTA, Honda, …)
  - Others, …

Recent installation of Supercomputers in Japan

- Institute of Fluid Science, Touhoku University
  - SGI Altix3700Bx 1.6TFLOPS x 4
  - NEC SX-8 128GFLOPS x 8
  - Dec., 2005
- Japan Meteorological Agency
  - Hitachi SR11000 model K1 (21.5TFLOPS)
  - Power5+ (2.1GHz), 16CPU,135GF/node, 160 nodes
  - March, 2006


SR11000 (hitachi)

- High Energy Accelerator Research Organization (KEK)
  - IBM BlueGene/L 57.3TFLOPS (10 racks)
  - SR11000model K1 2.15TFLOPS (16nodes)
  - March, 2006
  - Fastest system at this moment!

Large-scale Clusters in Japan

- AIST Supercluster
  - P-32(Opteron-dual 1072nodes, 8.6TF, Myrinet, 61TF/Linpack)
  - M-64(Itanium2x4 132nodes, 2.7TF, Myrinet, 1.6TF/Linpack)
  - F-32(Xeon-dual 268nodes, 3.3TF, GbE, 1.9TF/Linpack)
  - May, 2004
- RIKEN Combined Cluster
  - Xeon-dual 512nodes, 6.2TF x 2
  - NEC SX-4 32CPU 282GF
  - March, 2004
- PACS CS, CCS, University of Tsukuba
  - July, 2006
- Titech Campus Supercomputing Grid, Core System
  - April (?) 2006

CCS of University of Tsukuba

- Center for Computational Sciences
  - http://www.ccs.tsukuba.ac.jp/
- Established on April 2004,
  expanded and reorganized from the former organization, CCP (Center for Computational Physics)
  - Extended its research area from Computational Physics to Computational Sciences
- Collaborative researches with Computational Scientists (application) and Computer Scientists (system)
  - Needs from applications
  - Seeds from systems

Massively Parallel System CP-PACS

No.1 at TOP500 list on November 1996
614 Gflops peak perf.
368 Gflops Linpack
(The last Japanese supercomputer at No.1 before Earth Simulator)

↓

Dropped off from the list on November 2003 !!

Shutdown on Sep. 2005

Future view of resources at CCS

- Mid & long term plan
  - We will need very large-scale system replacing CP-PACS
  - System to fit the application fields of CCS, and PFLOPS system
    $\Rightarrow$ Not just an “off-the-shelf Supercomputer”

- Short term plan
  - For next few years, clusters still keep advantage on CPU performance, network performance and their balance
  - Cluster with 20-30% of efficiency is better than Vector machine with 99% efficiency (in term of cost/performance)

10 - 20 Tflops range system by PC cluster

PACS-CS

(Parallel Array Computer System for Computational Sciences)

General view of HPC clusters

- Use Intel-compatible CPU (Xeon, Opteron, Itanium2, …)
- Dual CPU SMP
  - To reduce the space and the number of network interface keeping total system peak performance
  - Lack of memory bandwidth (memory wall problem)
    (but very fast for Linpack !)
  - Low sustained performance on network bound applications
- SAN (System Area Network)
  - MyrinetXP: dual connection for 500MB/s -> 10Gbps
  - Infiniband: x4 spec. for 1GB/s
  - Gb Ethernet is still OK for non-network bound applications
    (10GbE will come soon, but still expensive)

Our view to HPC clusters

- It should provide high cost/performance ratio replacing traditional vector machine and MPP
- For high-end usage, the cost of network is getting higher while commodity CPU increases its performance (especially for thousands of CPU class)
  - Fat-Tree or Clos Network base
  - Price of NIC and Switch (dramatically changes in a few years)
- Difficult to fit to general users and applications
  - What kind of usage ?
  - CPU intensive or I/O (Network) intensive ?
  - How many CPUs / job ?
- We know more effective solution for “our users” (we can understand their applications in detail!)

Concept of PACS-CS

- "We need MPP !" (honestly saying)
  - It is difficult to develop system in CP-PACS style (needs collaborative development with vendors)
  - Not just buying it ⇒ for future HPC system research
- Yes, it is a cluster, but we keep the balance among **CPU : memory : network performances**
- To fit applications and operations in CCS

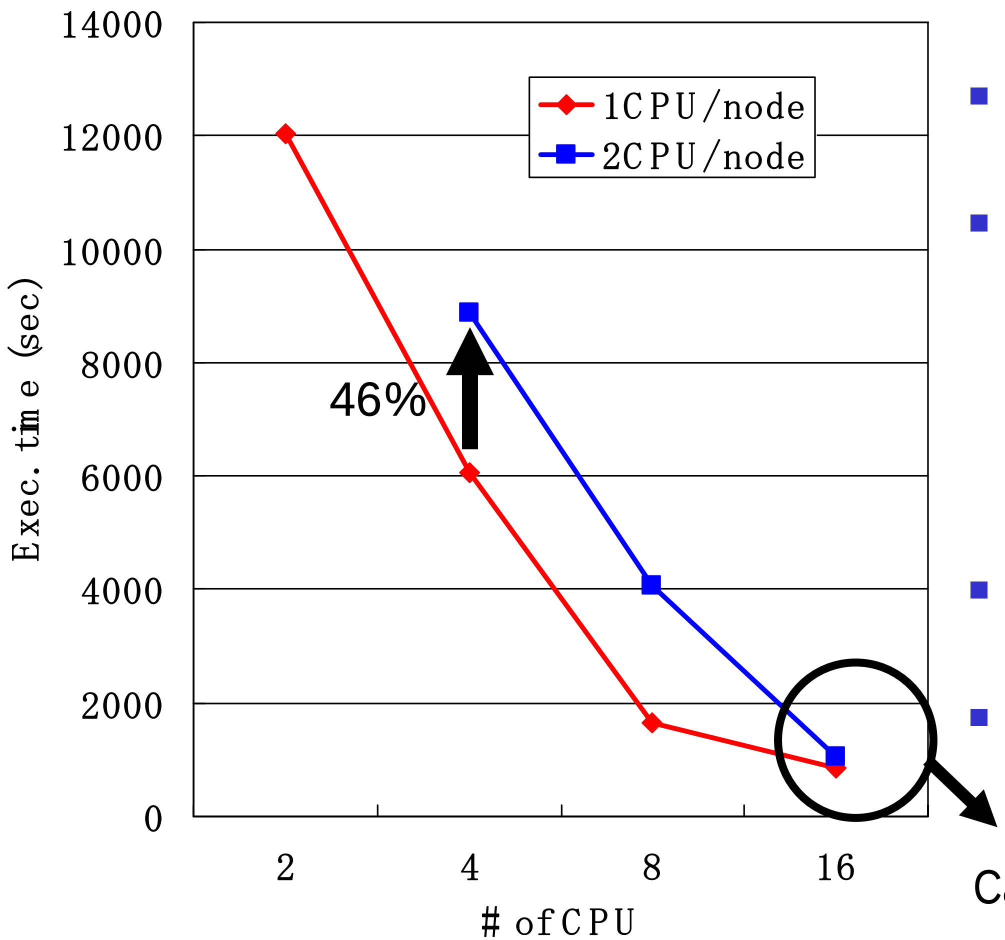
⬇

- Making MPP (-like system) with commodity technology
  - No development of LSI level
  - Commodity CPU and commodity network
  - Only Developing a mother board
  - Commodity software (+ specially customized ones)

Features of PACS CS

- Single CPU / node (not SMP!)
  - Balance between CPU and memory performance
  - Appropriate CPU speed (2.8GHz, LV Xeon, not too high)
- Hyper-Crossbar Network with trunked GbE
  - Balance between node and network performance
    - Multi-dimensional, trunking with GbE
  - "Nearest Neighbor" + "broadcast/reduction" is essential
  - Effective use of commodity technology with good cost/performance ratio
    - Many NIC ports/node, Many small switches (10-20 ports)
    - Effective solutions for large-scale systems
- Use only commodity parts
  - But, we have designed the motherboard
  - High density implementation
    - Same as traditional HPC clusters (dual CPU SMP): 2 CPU / 1 U

Why single CPU / node ?

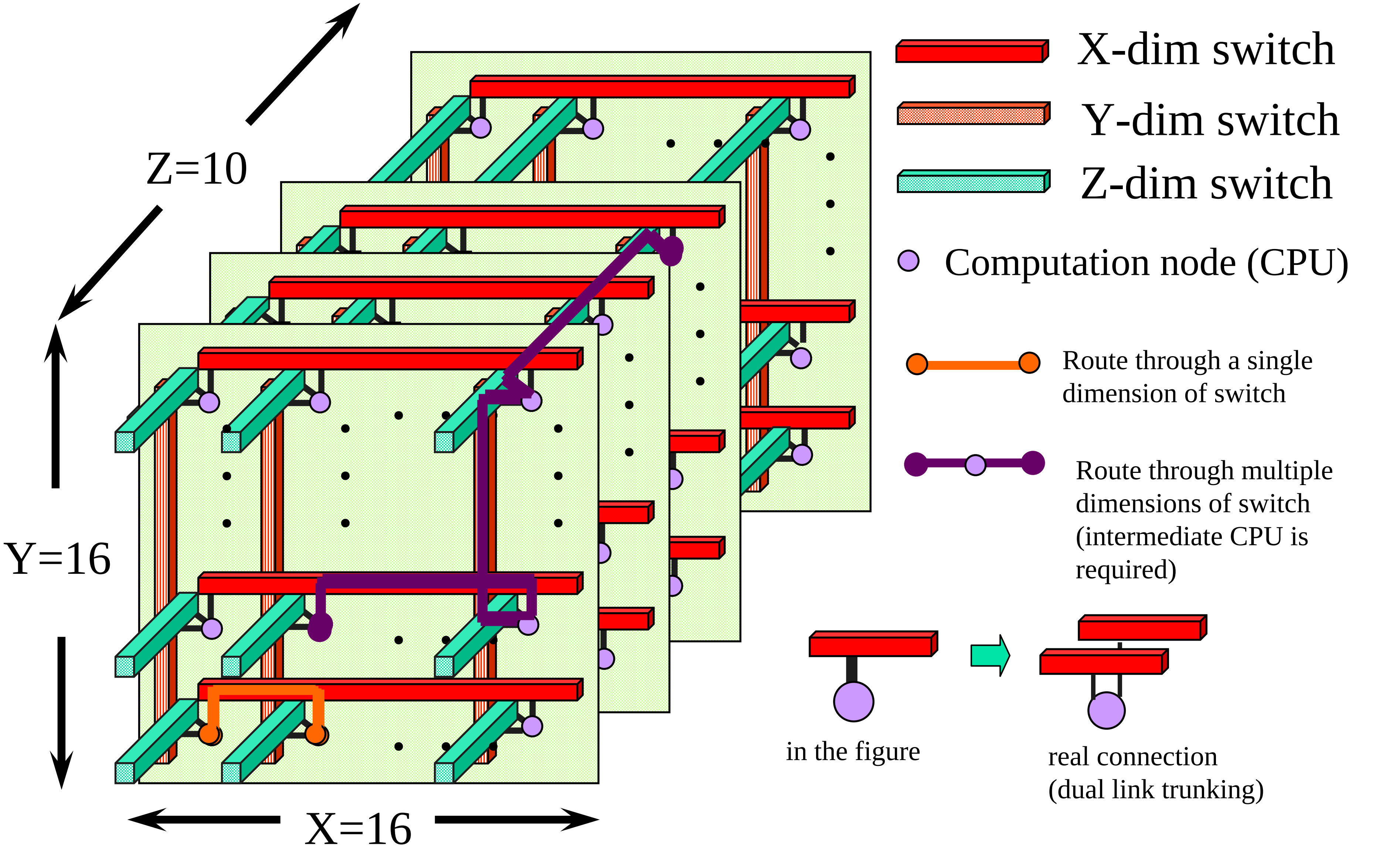


- Material Science: DFT method for Si512
- Conditions
  - CPU: Xeon 2.8GHz dual CPU SMP
  - Memory: PC2100 (4.2GB/s)
  - Network: Myrinet2000
  - Middleware： SCore 5.1 PM/Myrinet
- Varied the number of CPUs for a fixed size of problem
- Examine the occupancy of memory bandwidth by 1 or 2 CPUs

Cache all-hit -> no difference

Logical connection among nodes (CPUs)

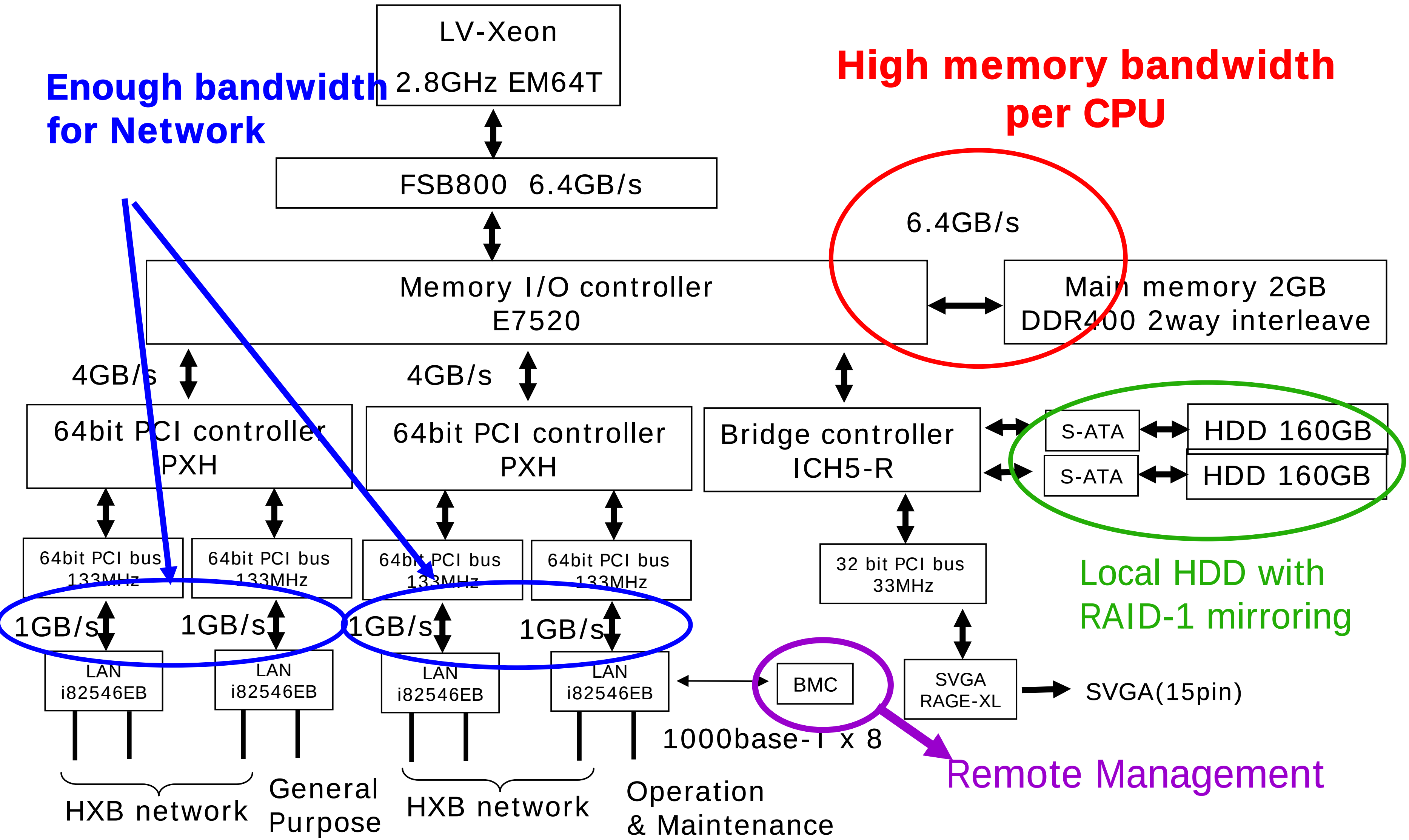
(3-D HXB network) 2560 nodes



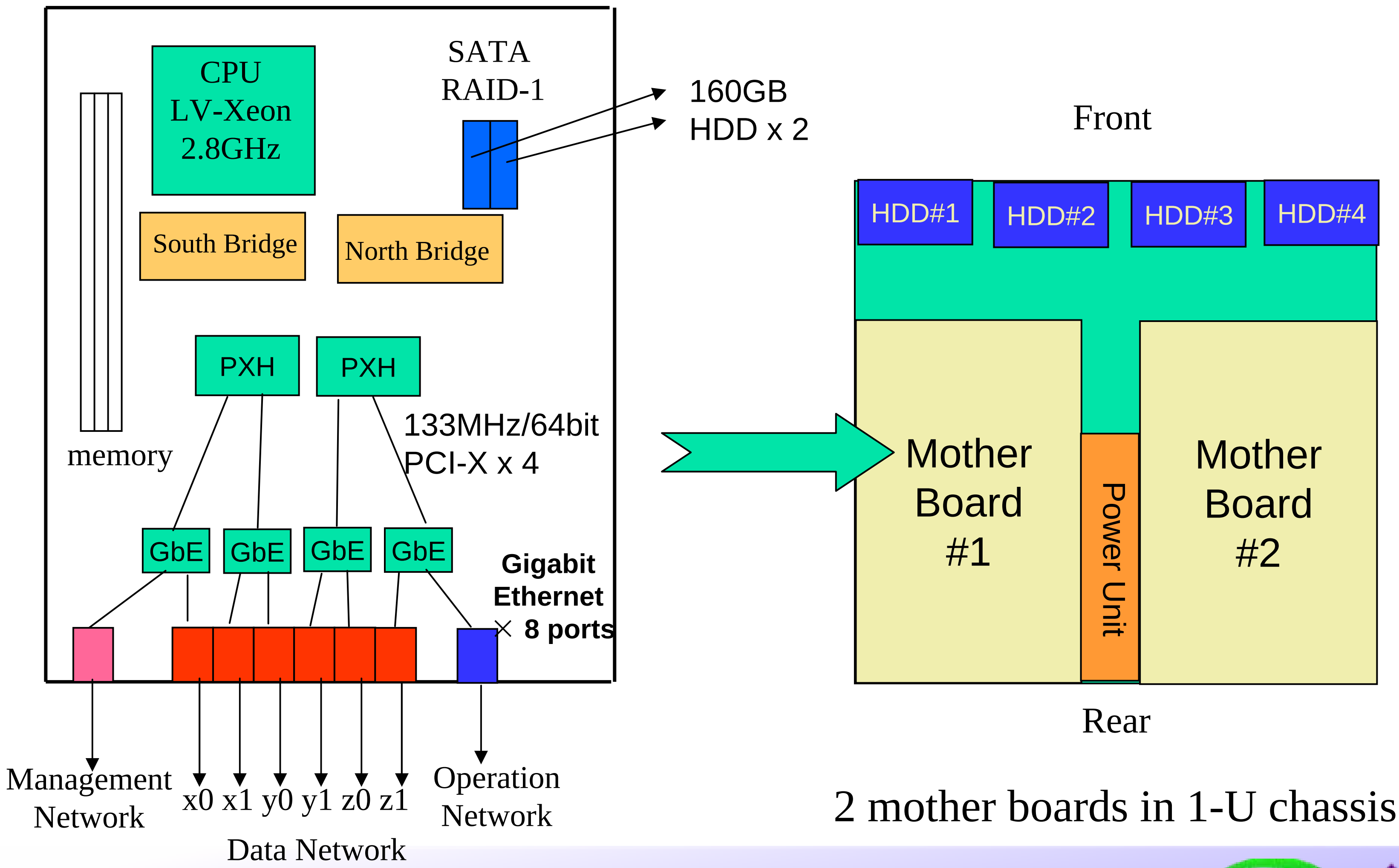
Network driver for GbE-trunk-HXB

- PM-Ethernet/HXB (by Fujitsu Lab.) enables
  - Direct inter-node communication on single dimension
    - Ex) (i, Y, Z) ⇒ (j, Y, Z), (X, k, Z) ⇒ (X, l, Z), etc.
  - Multiple GbE links are trunked to multiply bandwidth
    - Dual-link trunking doubles the bandwidth per dimension
  - Up to 3-D simultaneous sending/receiving
    - 250 MByte/sec (dual-link GbE) x 3 = 750 MByte/sec and 1.5 GByte/sec for bidirectional communication
  - Routing for a message requiring 1 or 2 hops of transfer on intermediate nodes
  - (Future plan) Fault tolerant operation for single link failure

Block diagram of a node

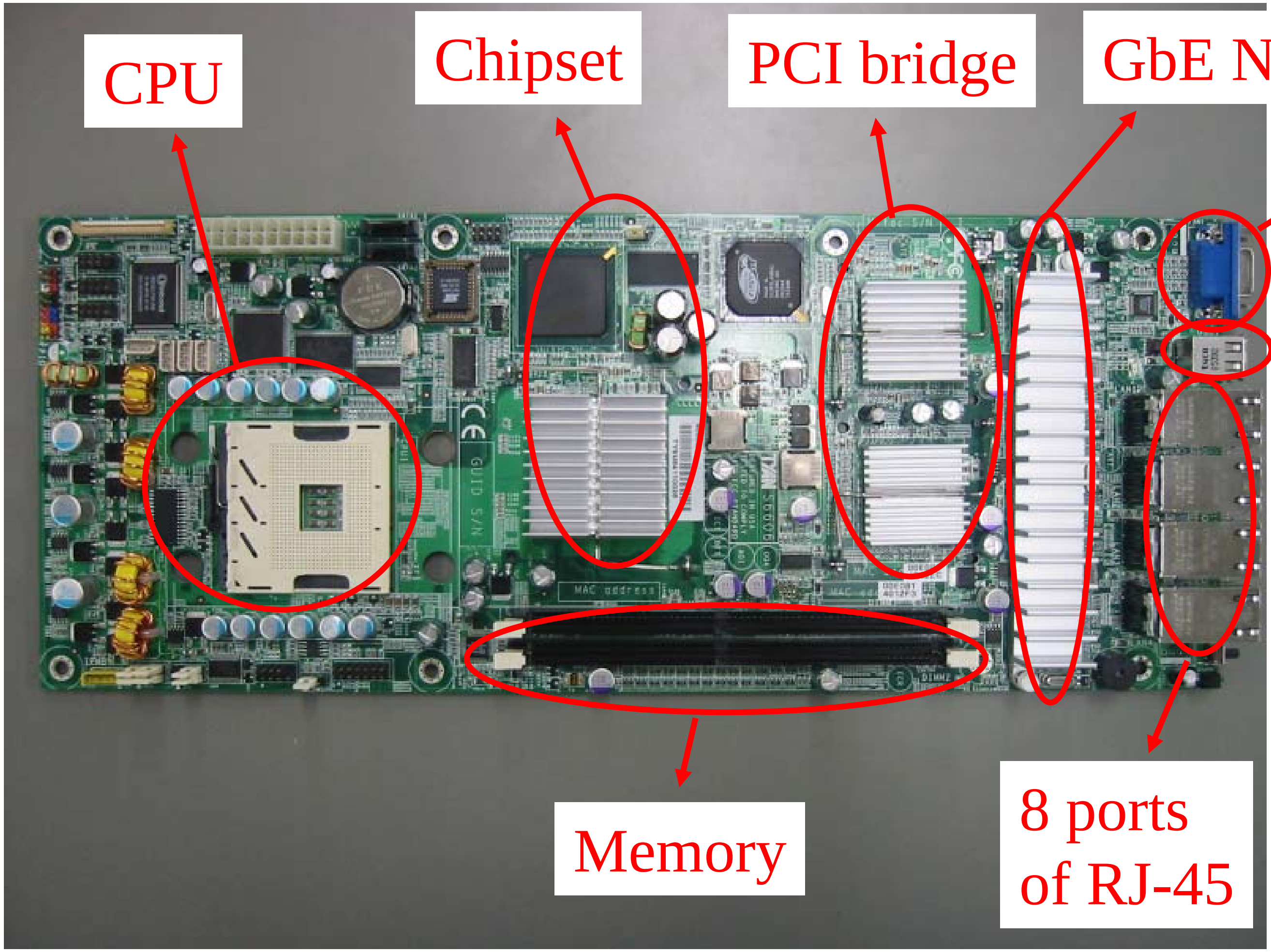


Mother board & Chasis



2 mother boards in 1-U chassis

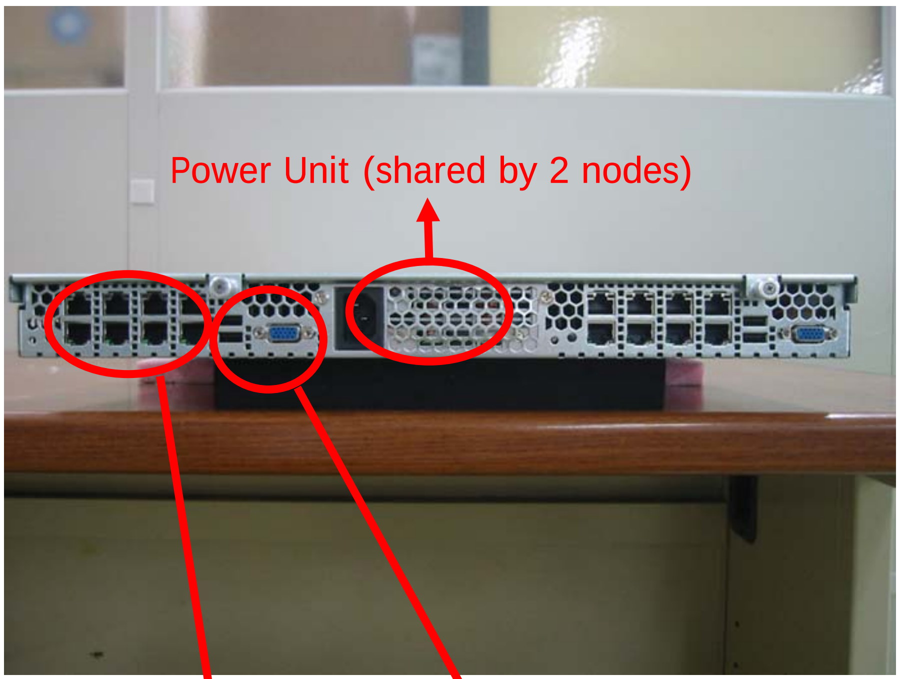
1st cut prototype motherboard



Front & Back view (1st cut prototype)

Front View

Rear View



8 ports of GbE (RJ-45)   USB x 2 + VGA

Chassis (1st cut prototype)

3.5 inch HDD x 4 drive = RAID-1 x 2

4 categories of interconnection network

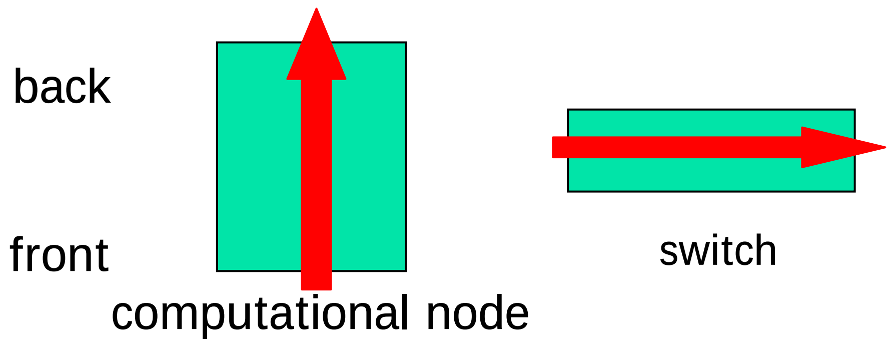
- Parallel Processing Network (for Data)
  - 3-D HXB network based on dual-GbE trunking
  - For high speed parallel processing on applications
- General Purpose Network
  - Generic tree network with link aggregation (LACP)
  - Generic UNIX network processing : NFS, NIS, DNS, rsh, ...
- Operation & Maintenance Network
  - Tree network
  - Remote operation (power on/off, reboot, individual/broadcast console access) to each/all node
- Surveillance Network
  - Watching a large number of (about 500) switches
  - All switches are managed and monitored by SNMP

Summary of PACS-CS spec.

# of nodes	2560 (16 x 16 x 10)
peak performance	14.3 Tflops
node configuration	single CPU / node
CPU	Intel LV Xeon EM64T, 2.8GHz, 1MB L2 cache
memory	2GB/node (5.12 TB/system), PC-3200 interleaved
network for parallel processing	3-dimensional Hyper-Crossbar Network
link bandwidth	one-sided: 250MB/s/dim. one-sided: 750MB/s (3-D simultaneous trans.)
local HDD	160 GB/node（RAID-1）
total system size	59 rack
power consumption	550 kW

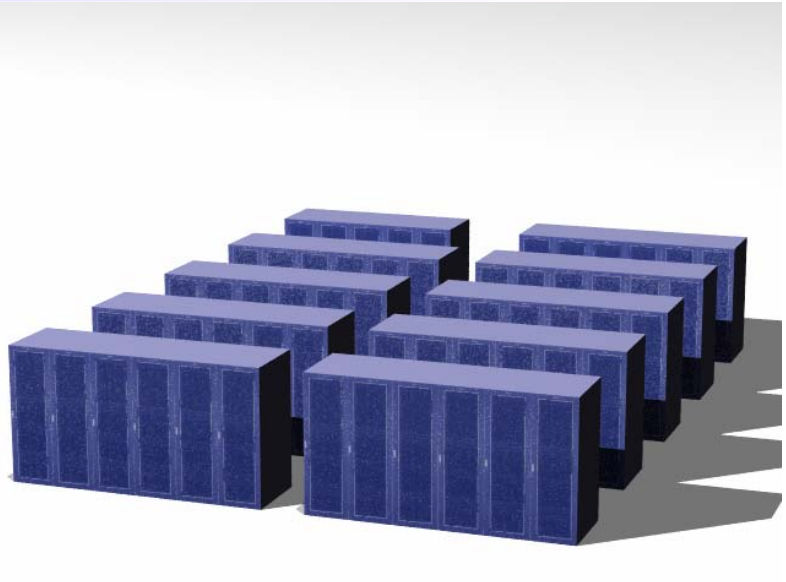
System implementation

- 1U rack mountable mother board and switch
- Separated racks for computational node-only and switch-only
  - node-switch mixture rack
    ⇒ air-flow problem

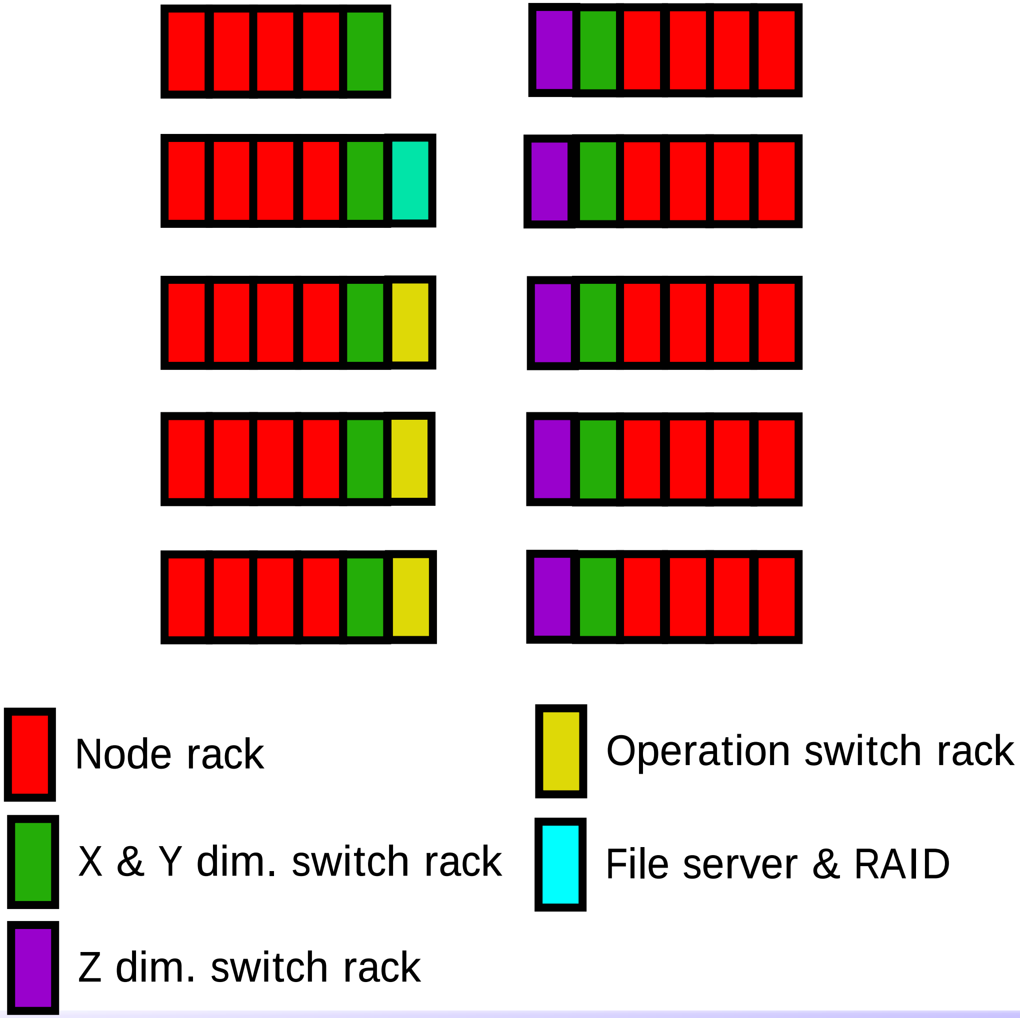


- Node rack：1280U
- Resources for network construction
  - for parallel processing network: 48 port switch divided into three VLANs with 16 port for each ⇒ space efficiency
  - GbE link (1000base-T): 8 ports / node x 2560 = **20480 links (= cables)**
  - total number of switches （48 port）：351
- Compact and Air-Flow-Aware cabling is required
- Hardware manufacturer: Hitachi

Whole system image and rack floor plan



Whole system image



Node racks and network switch racks are separated

Software

- Linux + SCore
  - PM/Ethernet-HXB driver : developed by Fujitsu
  - Partitioning, Monitoring
- Batch/Queue （PBS, SGE, ...） : not complicated
- Parallel programming in MPI
- Languages: Fortran90, C, C++
- Math Libraries

Main applications of PACS CS

- Particle Physics: QCD (Quantum-Chromo Dynamics)
  32x32x32x64 full QCD
- Material Physics: Nano-material “first-principal” simulation
  Simulation based on Real-Space DFT (Dense Function Theory) for 10,000 atoms

- Both applications require high bandwidth on memory
- Communication
  - nearest neighbor & broadcast & reduction (for both applications)
- Allocate 512～2048 CPUs per run for these large scale simulations for several days to several weeks

Tokyo Institute of Technology (TITech) and Global Scientific Information and Computing Center (GSIC)

- Japan's premier Technical Institute (University) in science and technology, over 800 faculty members.
- GSIC established in April 2001, reincarnated from traditional-style supercomputing center
- Responsible for R&D and deployment of advanced supercomputing infrastructure, also leading center for Grid computing in Japan

- The first full-scale campus-wide Grid "Titech Grid", seeding the entire campus (15 sites) with over 1300 processors, starting on April 2002, of high-performance PC blade servers, in addition to the existing supercomputers, and interconnecting them with a campus Gigabit backbone …

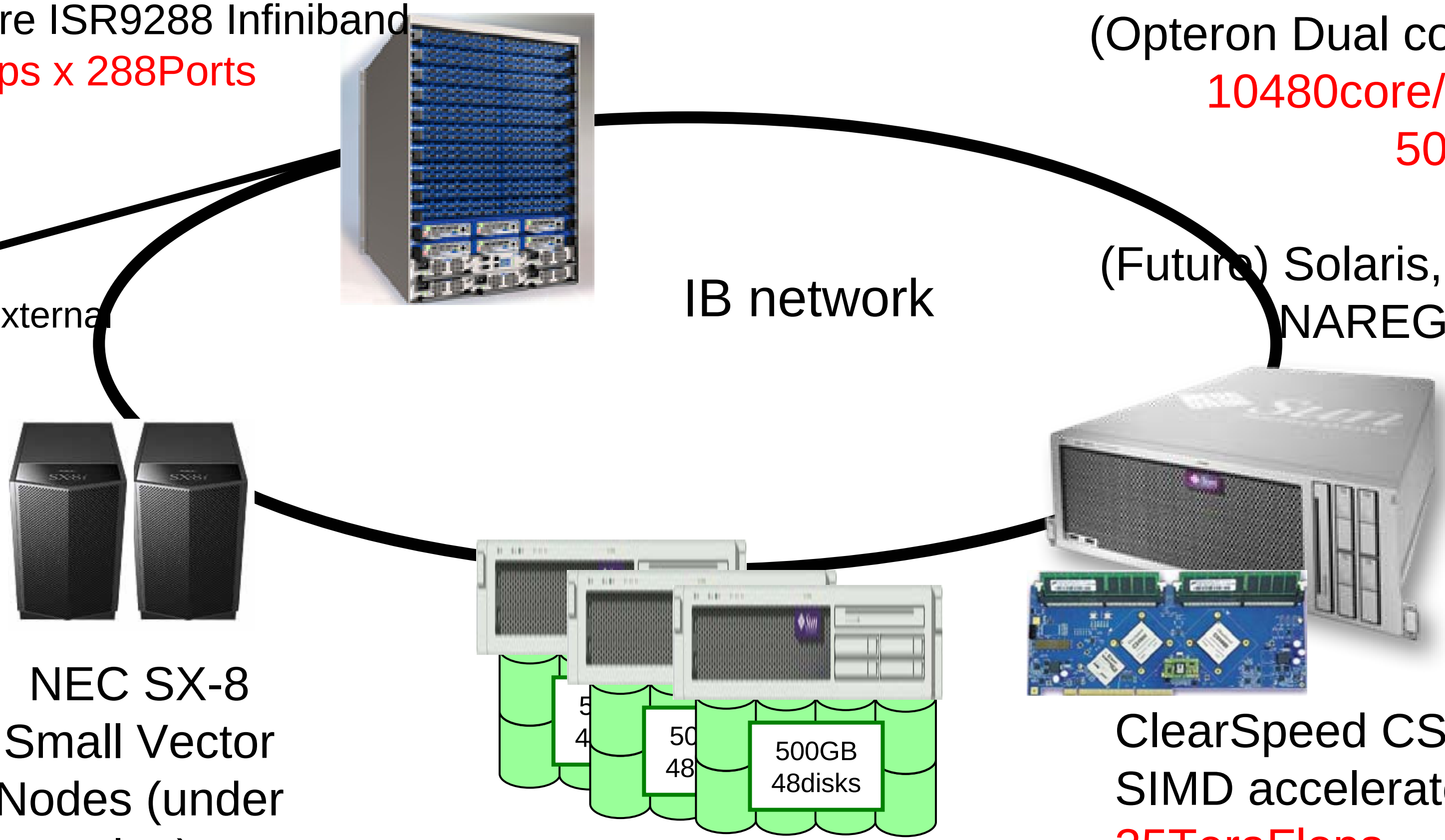
- **The New "Supercomputing Campus Grid" Core System, Spring 2006, which have performance**

The New “Supercomputing Campus Grid” Core System, Spring 2006

Voltaire ISR9288 Infiniband
10Gbps x 288Ports

10Gbps+External Network

Sun Galaxy 4
(Opteron Dual core 8-Way)
10480core/655Nodes
50TeraFlops
OS Linux
(Future) Solaris, Windows
NAREGI Grid MW

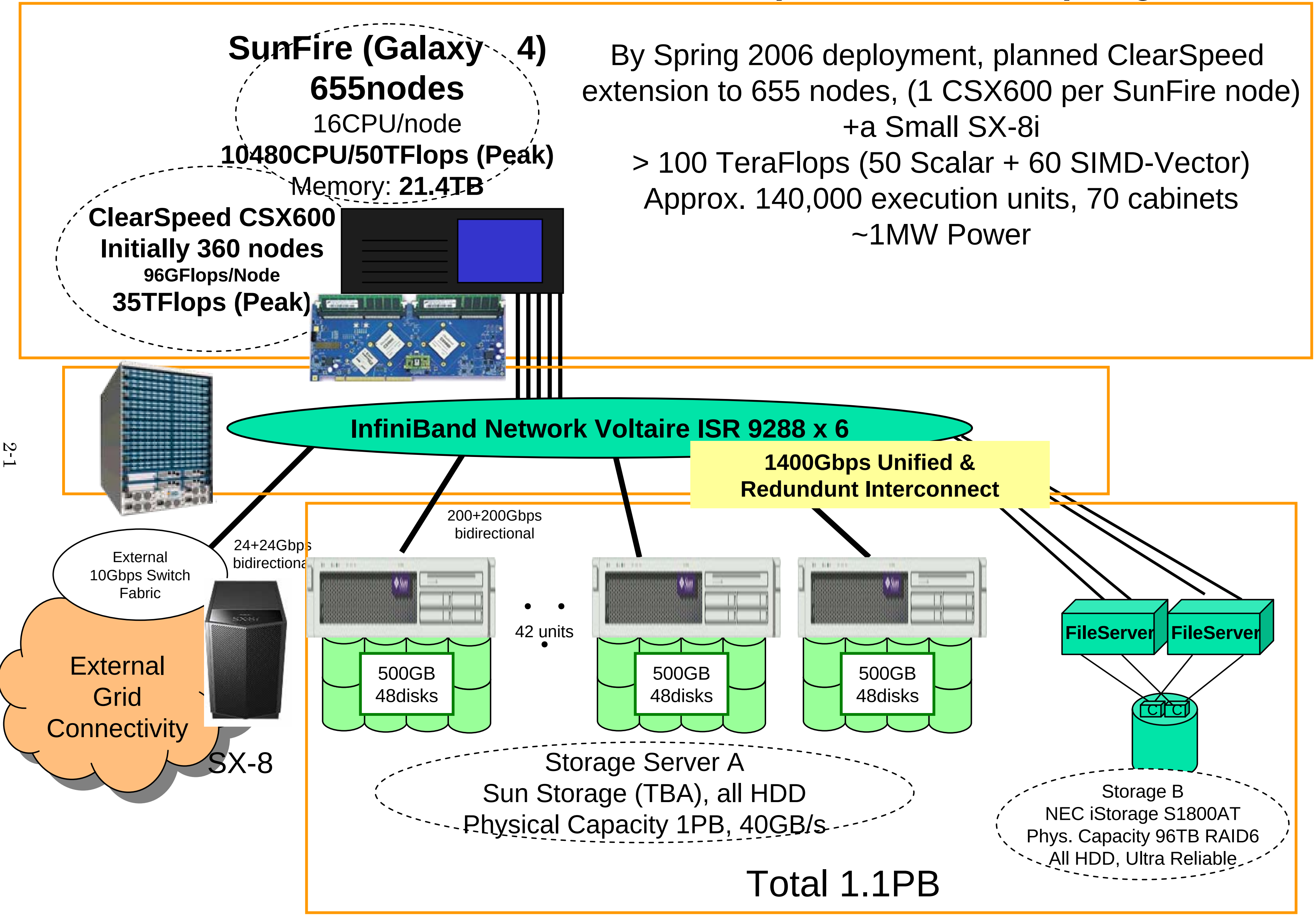


NEC SX-8
Small Vector
Nodes (under
plan)

Storage
1 Petabyte (Sun “Thumper”)
0.1Petabyte (NEC iStore)
Lustre FS, NFS (v4?)

ClearSpeed CSX600
SIMD accelerator
35TeraFlops ➔
60TeraFlops (1 board
per node)

NEC/Sun Campus Supercomputing Grid: Core Supercomputer Infrastructure @ Titech GSIC - to become operational late Spring 2006 -

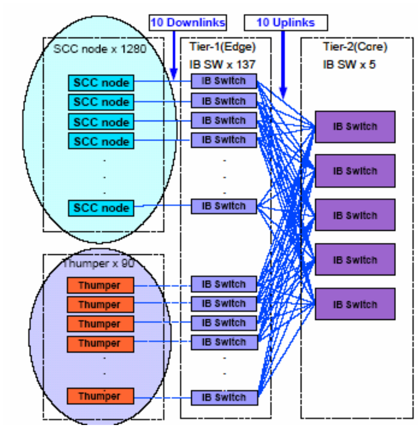


2-1

Titech Supercomputing Grid as No.1 in Japan

TITECH GRID

東京工業大学
Tokyo Institute of Technology



>50+50 TeraFlops, 1100 Terabytes, 4 year procurement cycle

Will beat the Earth Simulator

Will beat all the other Univ. centers combined

GSIC
Global Scientific Information and Computing Center

>>

Over 20 times C/P

All University National Centers

Total 41 TeraFlops, 300 Terabytes
Total $110 million/year, 6 year procurement cycle

Recent Project: “Kei Soku Keisanki”

Notice

- Source:
  - Web page, http://www.mext.go.jp/
  - Public announcement
  - Open symposium
  - (almost material only in Japanese)

The “Keisoku Keisan-ki” project

京 速 計算機

Kei
（１０^{１６}）

Soku
(speed, fast)

Keisan-ki
(computer)

数値(Value)	記号(Sign)	単位(Unit)	英数詞(English)	漢数詞(Japanese)
10^{24}	Y	Yota-(?)		杼 (じょ:jo)
10^{21}	Z	Zeta-(?)	sexillion(US)	=sextillion;=10 垓 がい:gai)
10^{20}				垓 がい:gai)
10^{18}	E	Exa-	quintillion	百京(=10 億ギガ)
10^{16}	E	10 Peta-	10 quadrillion	京(=10 億ギガ)
10^{15}	P	Peta-	quadrillion	千兆
10^{12}	T	Tera-	trillion	兆 ちょう:chyō)
10^{09}	G	Giga-	billion	十億
10^{08}	G	100 Mega-	100 million	億
10^{06}	M	Mega-	million	百万
10^{04}	M	10 Kib-	10 thousand	万
10^{03}	K	Kib-	thousand	千
10^{02}	h	hecto-	hundred	百
10^{01}	da	deca-	ten	十
10^{00}		mono	one	一

京（Kei) = １０^{１６}

10PFLOPS!!

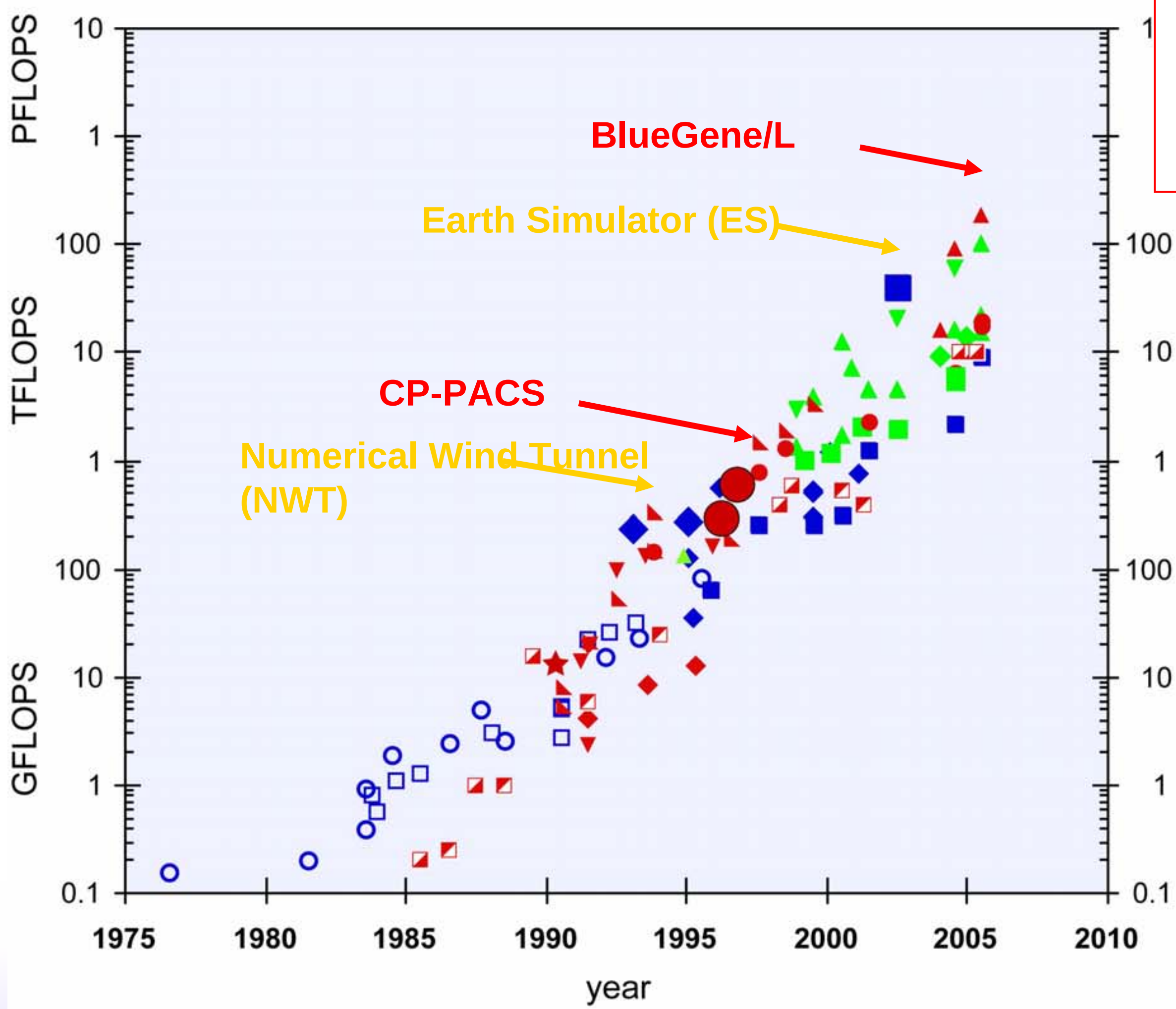
Table of Number Unit

Proposed project

- Formal title: 「最先端・高性能汎用スーパーコンピュータの開発利用」（Development and Utilization of cutting-edge high performance general-purpose supercomputer)
  - 7 years (2006-2012), total budget 110 billion yen (800M euro)
  - ＭＥＸＴ（Ministry of education, culture, sports, science and technology)
- Objective: taking the leadership in development of cutting-edge supercomputers and strengthening in high-performance computing technology.
  - Note that 10PFLOPS is just a target, not mentioned formally(?)
- Research Topics
  - Development of software (OS, middleware, application) to make use of supercomputers.
  - Development of a cutting-edge high-performance general-purpose national supercomputer systems
  - Establishment of COE (Center Of Excellence) for research and education of HPC

Progress of Supercomputers

1 million times faster in the last 30 years (1976～2005)

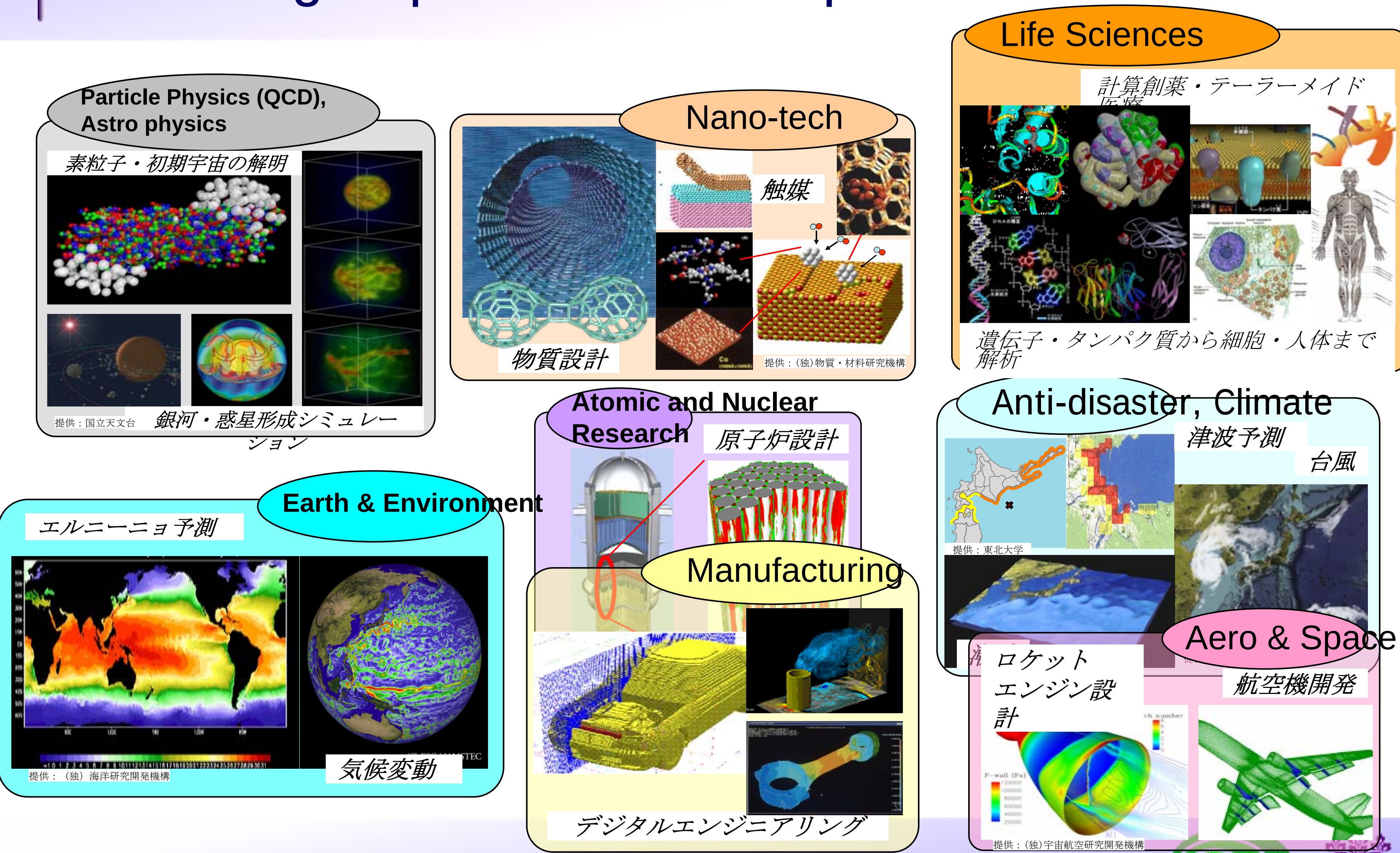


Vector SMP

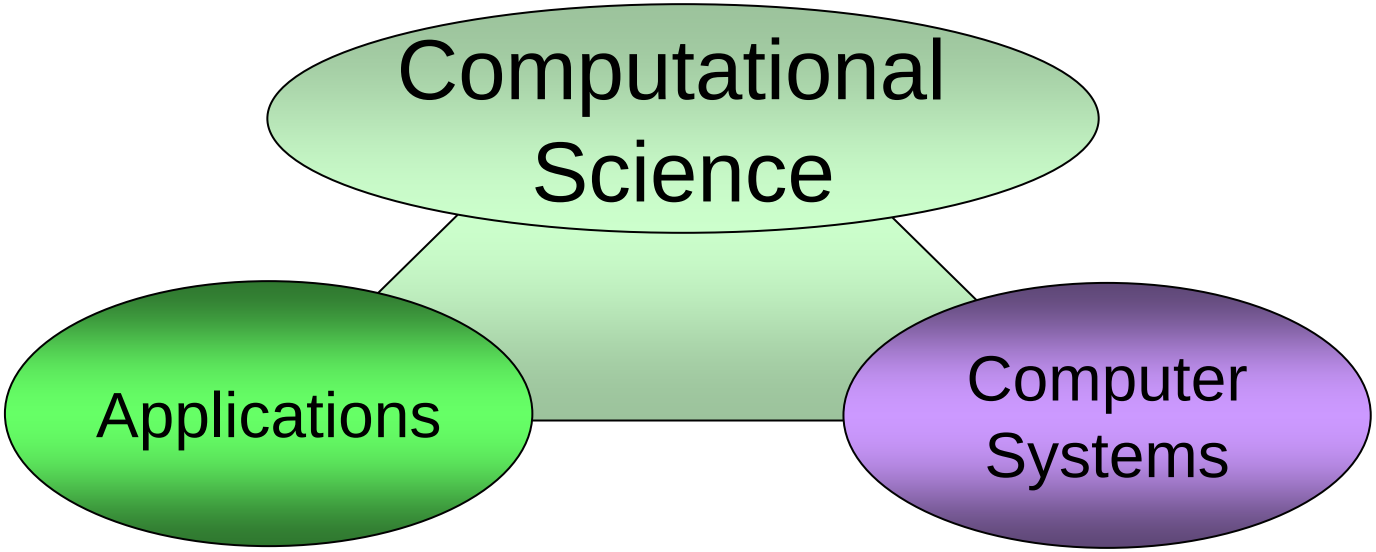
- **Epoch-making supercomputers in Japan**
  - **Numerical Wind Tunnel (NWT), NAL**
    - **1993, Nov 1st position in Top500**
    - **Vector-parallel**
  - **CP-PACS, CCS, University of Tsukuba**
    - **１９９６, Nov 1st position in Top500**
    - **MPP**
  - **Earth Simulator**
    - **2002, June 1st position in Top500**
    - **Vector-parallel**

No roadmaps in Japanese supercomputer development!!

Increasing Importance of Computational Sciences



Two important factors to enable computational sciences

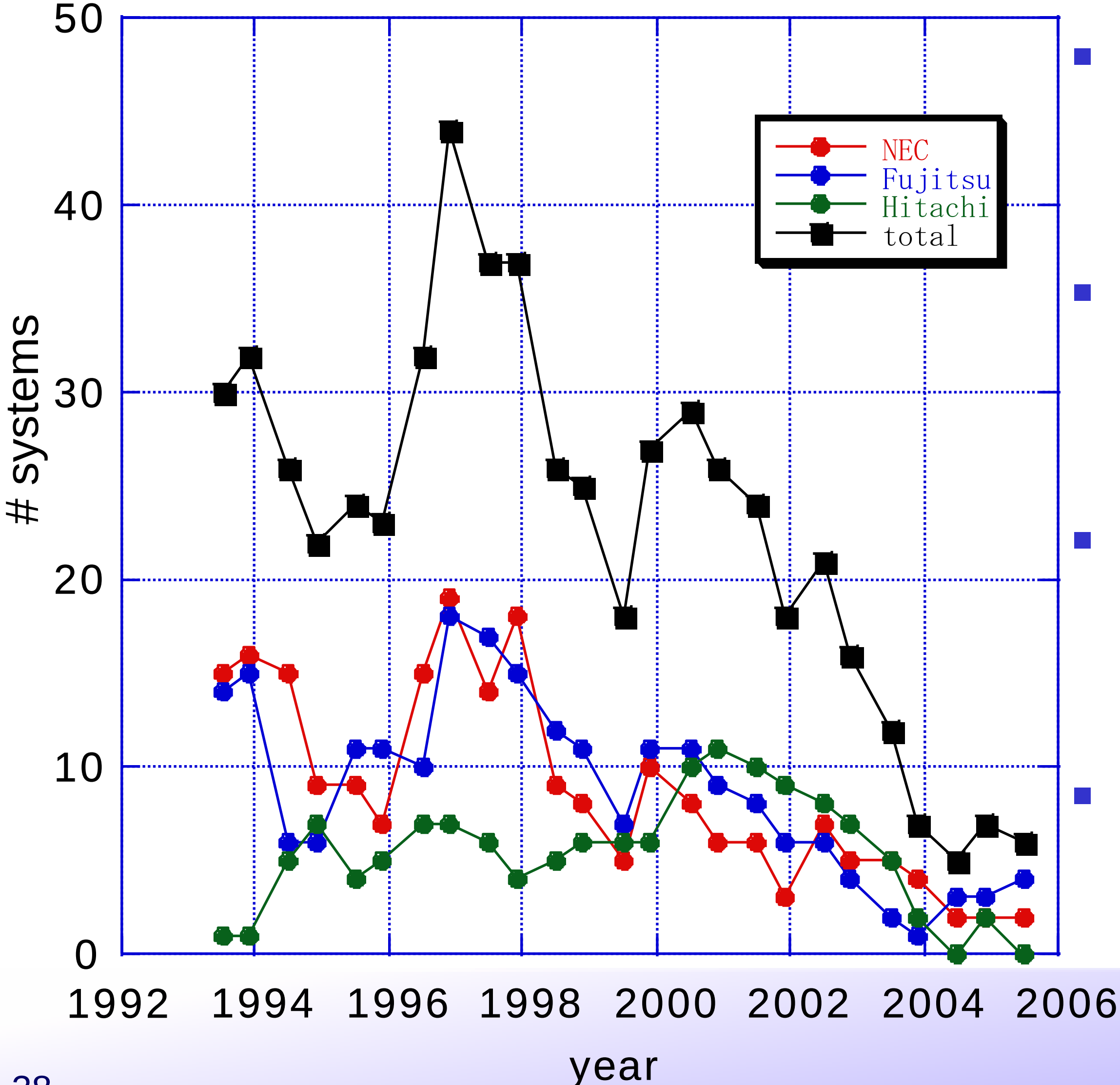


Potentials of research in computational sciences using supercomputers

Potentials to develop/make supercomputer hardware & systems

We have a big concern for the current status of HPC …

Trends: # of Japanese Supercomputers within top100 (TOP500 list)



- The number of Japanese supercomputers in Top100 was dramatically decreasing since 1997's peak!
- The same trends in each Japanese Vender and in vectors, scalar SMP.
- Because …
  - Killer Micro
  - Clusters
- We worry about our weaken competitive positions in world HPC markets.

Status (1)

- The committee on IT in MEXT have been discussing the policy and plan to promote the computational science since Aug. 2004.
  - The working group was organized to discuss on promotion of computational science and technology.
- Leading projects on "Elemental Supercomputing technology" and "Innovative applications" have been launched in 2005 (2005-2007)
- Informal announcement (July 25, 2005) on development of 10PF supercomputers -> so-called "Kei Soku Keisanki"
- Politicians (inc. ex-minister of MEXT) organized a party to promote projects of the development of supercomputer.
  - Big science needs "political" supports.

Leading projects

- Leading projects on ‘Elemental Supercomputing technology” in “R&D to build future IT infrastructure” have been launched in 2005
  - 4 projects were accepted.
    - R&D on Optical Interconnections for Ultra High-Speed Computers (NEC & Titech)
    - R&D on Low-power Device, Circuit and Processor Architecture (Hitachi, U. of Tokyo, **U of Tsukuba**)
    - R&D on Peta-Scale Interconnect (Kyushu-U, Fujistu, Titech)
    - R&D on IP internal interconnection (GRAPE-DR, U. of Tokyo)
  - 3 years (2005-2007), total budget 30M euro/year
  - The elemental technologies are expected to be used in future supercomputers (esp. in “Kei soku”)
- Leading Projects on “Innovative Applications” also started.

Status (2)

- The project plan was submitted from MEXT to the government for 2006 budget plan.
- Evaluation by CSTP. The final evaluation report was published by CSTP.
- The project plan was accepted (Jan. 2006)
- The projects will be launched April 2006 !!!

- Dr. Tadashi Watanabe (the former ES development leader of NEC) took up a Program Manager of this project in MEXT since Jan. 2006.
- RIKEN (Research Agency of MEXT) was selected as an organization to take leadership in development and operation.
  - The Next-Generation Supercomputer R&D Center (NSC) was organized by RIKEN on January 1, 2006
- NARGI (National Research Grid Initiative) will be merged into “Keisoku” projects
  - NAREGI Grid software is expected to be used also for the next generation supercomputers.

My view on “Kei Soku Keisanki”

The points to be considered

- The pursuit of effective performance

Effective performance
＝ （peak performance of node）
× （the number of nodes）
× （Efficiency）

- Feasibility
  - Realistic space for installation (less 10MW, must be equal to ES)
  - Realistic electric power (less than 2000m^3, must be equal to ES)
  - Cost for Development

ES (2000m^2, 7MW)

Scaling from PC !?

PC at 2006 ⇒
10GFLOPS/100W

3 Approaches to 10PFLOPS system

- Fat-node parallel system
  - 1TFLOPS/node x 10,000
  - Vector parallel system, big-SMP clusters
  - High-efficiency and wide-applications area are expected, but serious problem on power
  - Japanese computer venders prefer to this approach

- MPP system with low-power elements
  - 100GFLOPS/node x 100,000 or 10GFLOPS/node x 1,000,000
  - Similar approach as BlueGene/L
  - Need extensive low-power technologies for processor and networking …

- High FLOPS by special-purpose accelerators
  - ClearSpeed, GRAPE, (Cell?) …
  - There is a very strong group on accelerators such as GRAPE (for Gravity cal) in Japan
  - Possibly Limited applications

- Do you think whether 10PFLOPS system will be possible in 2010-2012?

Other Issues

- "Capacity" computing vs. "Capability" Computing
  - Capacity : the system should accept a wide class of applications?
  - Capability: the system should achieve some "break-through" in a few (or more) ground-challenge applications?
  - These two aspects may sometime conflict.

- Establish "Eco-system" of supercomputer technologies
  - "Supercomputer" tech. cannot survive only in their market!
  - "Vertical" effects
    - Smaller systems or subset of the system should be used in university computer center or lab.
  - "Horizontal" effects
    - The system should share the same or similar technology of other IT devices or systems such as embedded, or game system.

Summary

- Clusters still keep advantage on CPU performance, network performance.
    - PACS CS, Center for Computational Sciences, U. of Tsukuba
    - Supercomputing Campus Grid" Core System, Tokyo Inst. Of Tech.
  - "Cluster with 20-30% of efficiency is better than Vector machine with 99% efficiency in term of cost/performance"
  - But, Vector machines still survive to accept wide range of applications, especially in supercomputer centers.

- Japanese "Kei soku Kisanki" Project.
  - Target will be 10 PFLOPS in time range between 2010-2012.
  - The project is just being launched on April 2006.
  - The basic design will be decided within 2006, and the system will be installed in 2010-2011.