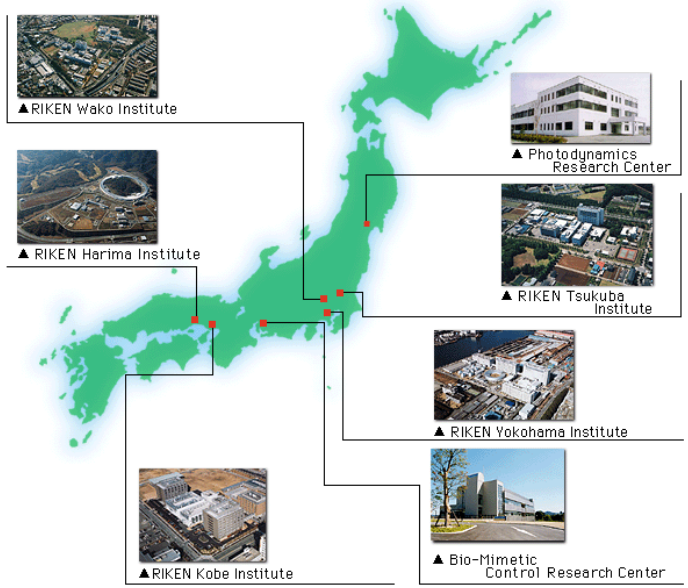
Advanced Center for Computing and Communication

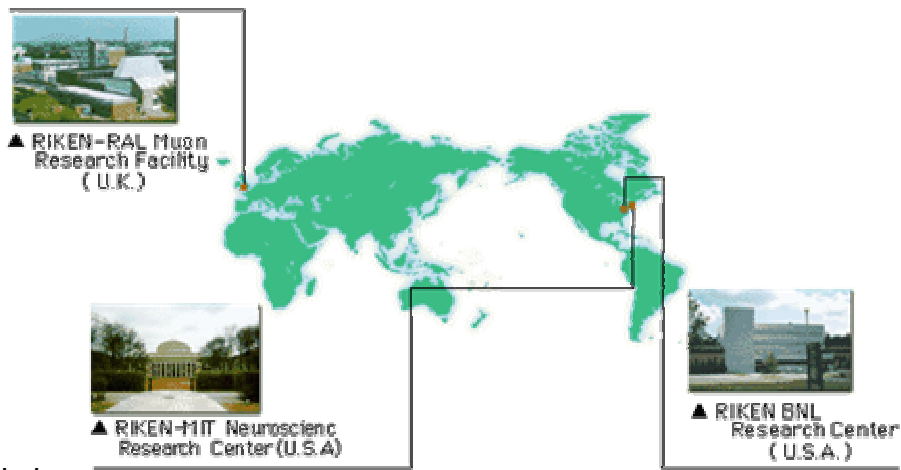
RIKEN Heterogeneous Cluster System and its Next for Multi-Scale and Multi-Physics Applications

Ryutaro Himeno
himeno@riken.jp
RIKEN

RIKEN and Advanced Center for Computing and Communication

- RIKEN
  - comprehensive research in science and technology (excluding only humanities and social sciences)
  - physics, chemistry, medical science, biology, and engineering extending from basic research to practical application
  - 6 campus in Japan, 3 outside Japan
  - about 3000 persons
  - an Independent Administrative Institution under the Ministry of Education from 2003
- Advance Center for Comp. & Com.
  - Providing researchers with computer resources and network services





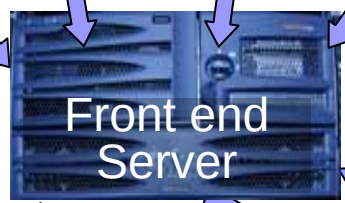
RSCC replaced VPP700

- Riken Super Combined Cluster System was introduced in Jan., 2004.
- Start operation in March, 2004.

RSCC System

Itanium2 Server: 2

User
User
User
User



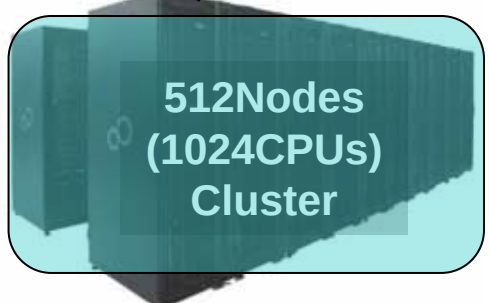
POWER4 Server: 7
Tape Drive: 4 (16)
Disk Cache: 2TB

NEC SX-7/32
282.5GFLOPS
256GB

Myrinet

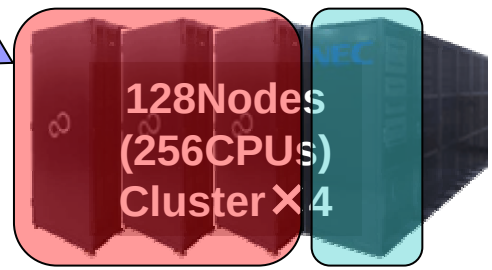
Pentium4 1.7GHz
2GB/Node,Myrinet
64GFLOPS×20

InfiniBand



Intel Xeon 3.06GHz
4GB/Node, IB
6.26TFLOPS

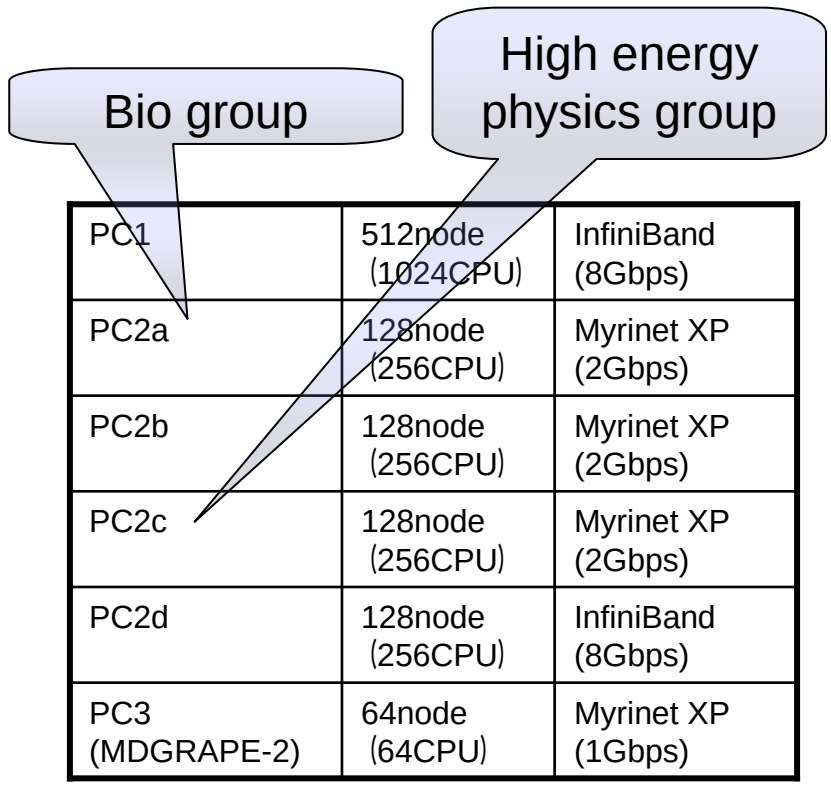
Myrinet InfiniBand



Intel Xeon 3.06GHz
2GB/Node,Myrinet/IB,
1.56TFLOPS×4

RSCC System

- Composed of sub-systems for deferent types of jobs
- 6 Linux clusters
  - Large-scale parallel jobs: 128CPU jobs in daily operation, max. 1024 CPUs
  - Execution time limit: 24 hours, max. 1 week
  - PC3 cluster has 20 MDGRAPE-2, Molecular Dynamics computers
  - User gropes may use one cluster throughout a year (high energy physics group and Bio user group are such kind of users)
- SX-7
  - Job with large memory requirement
  - Vecterized codes



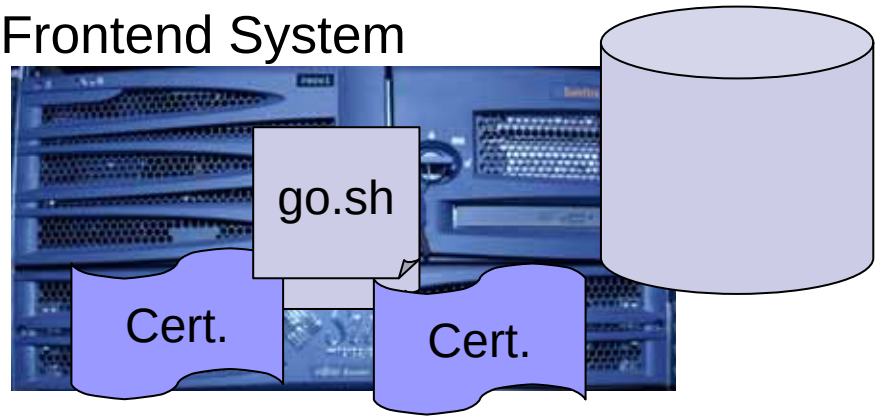
PC1	512node (1024CPU)	InfiniBand (8Gbps)
PC2a	128node (256CPU)	Myrinet XP (2Gbps)
PC2b	128node (256CPU)	Myrinet XP (2Gbps)
PC2c	128node (256CPU)	Myrinet XP (2Gbps)
PC2d	128node (256CPU)	InfiniBand (8Gbps)
PC3 (MDGRAPE-2)	64node (64CPU)	Myrinet XP (1Gbps)

Job execution

execute

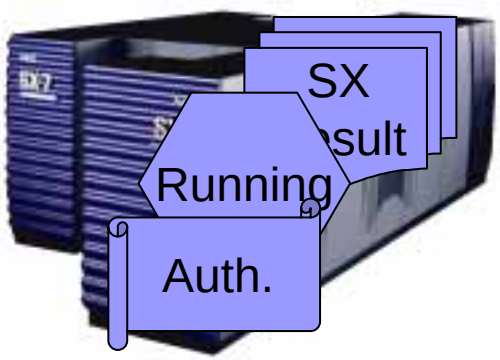
% qsub –q pc_queue go.sh

Frontend System

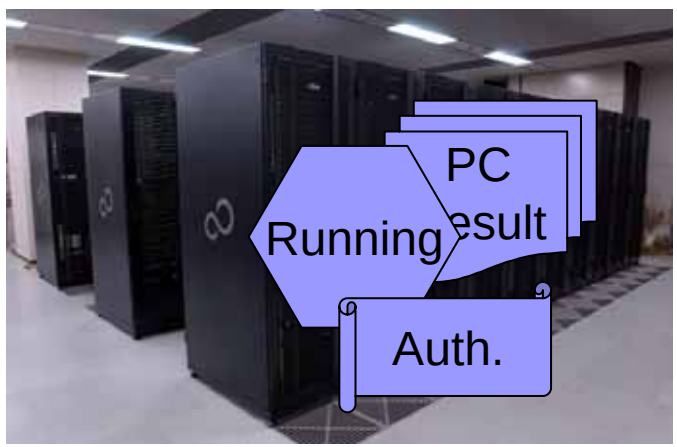


Login (sign-in)

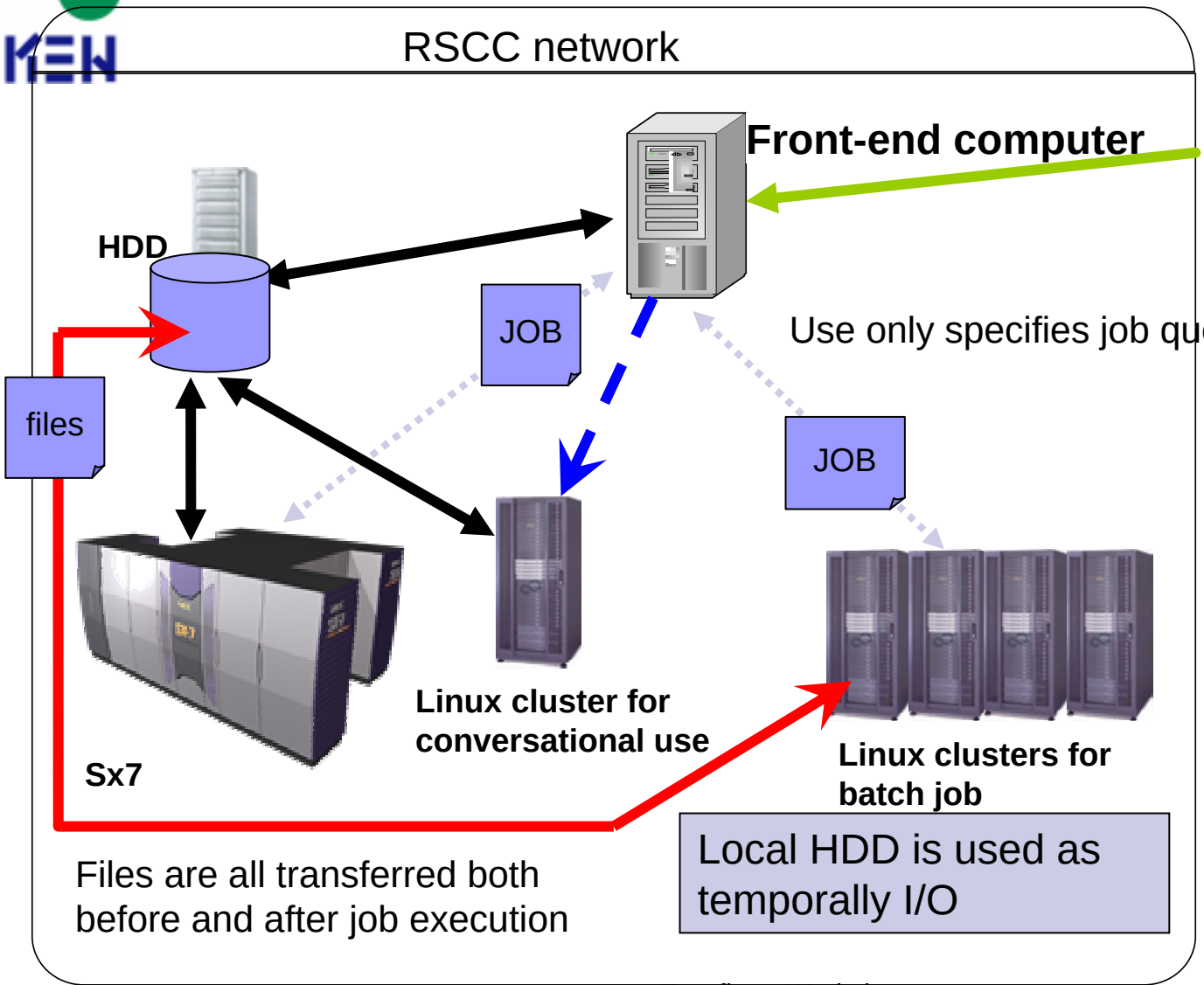
SX-7



PC Cluster



File transfer for job execution



- Login server
- Jobs from bio portal
- From ITBL

login

Batch job

File access

Other features

- **HPC and Bio portal**
  - Easy to use for beginners and experimental scientists
  - **Gene database is always up-to-date and on the local HDD for Bio user group, which is easily usable from Bio portal**
- **Real time visualization service**
- **VP FORTRAN, compatible compiler with Fujitsu VPP FORTRAN is available on Linux cluster**
- **From the cluster for high energy physics user group**
  - **HPSS high speed tape library is directly accessible**
- **IT-Based Laboratory project**
  - **Aliance of 6 organizations:** National Institute for Materials Science (NIMS), an independent administrative institution; National Research Institute for Earth Science and Disaster Prevention (NIED), an independent administrative institution; Japan Aerospace Exploration Agency(JAXA), an independent administrative institution; Institute of Physical and Chemical Research (RIKEN);Japan Atomic Energy Research Institute (JAERI); andJapan Science and Technology Agency(JST)

request

results

Web server

Small server

results

Short sequence

Here web application directly control job manager

Longer sequence

Job control and job request, status acquisition through Globus

Results

synchronized

Front-end System

Linux Cluster(256CPU)

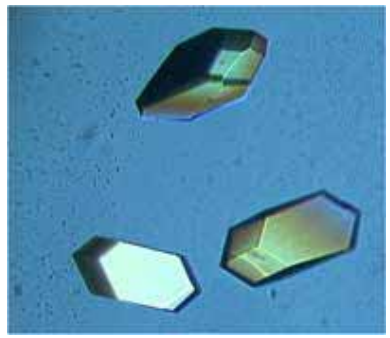
Job management, job execution request and job status requests are processed through Globus

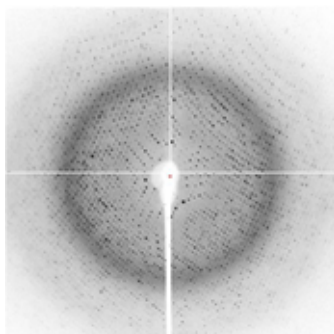
NQS server

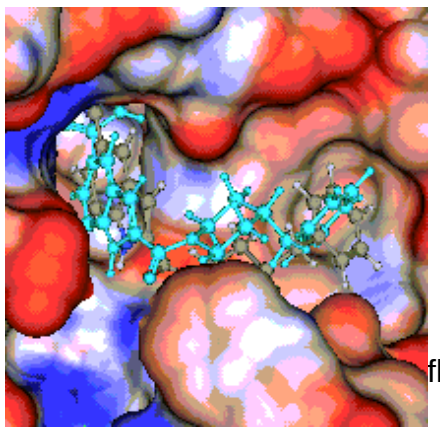
NQS selects computing nodes for execution

Focused Application(1)

- Life science
  - Bio portal for wet researchers
  - MD simulation
    - 100TFLOPS MDGrape-3 will be introduced until March, 2006.
  - Protein structure analysis







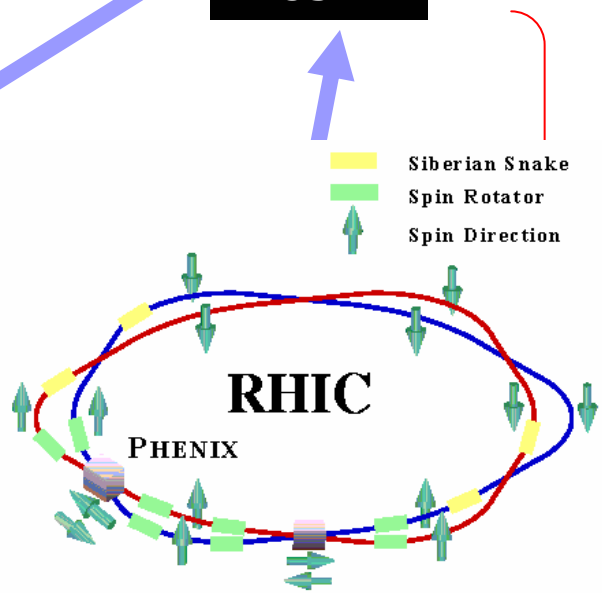
flops

Focused applications: high energy physics

HPSS high-speed tape system directly attached to cluster

RSCC

detector



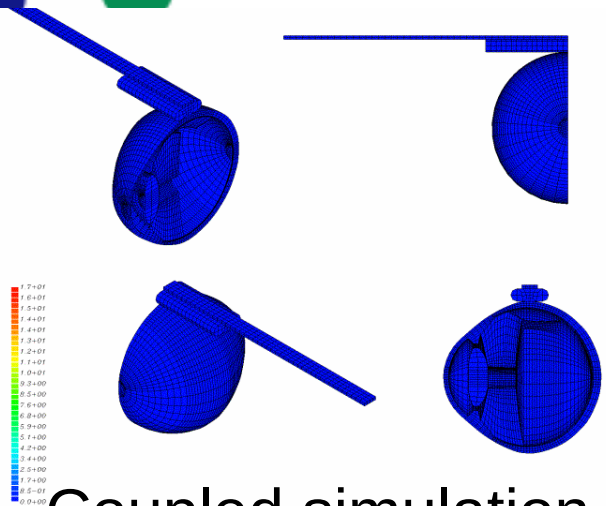
Accelerator

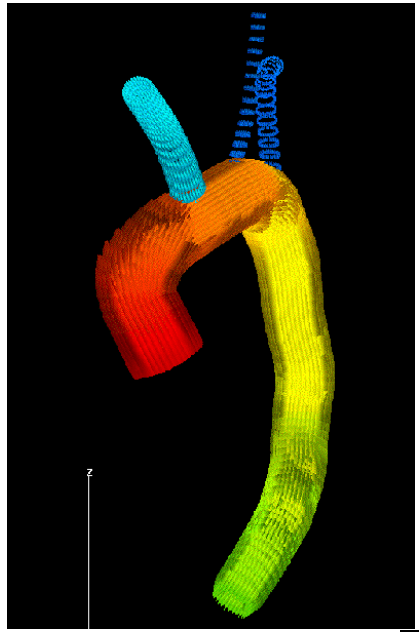
data

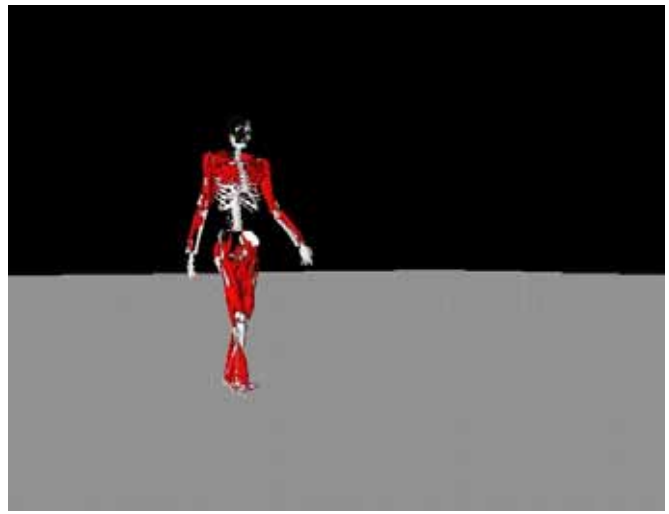
Japan

BNL in USA

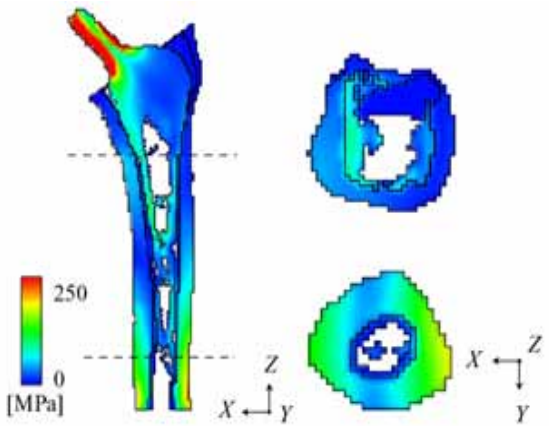
Computational Biomechanical Simulation





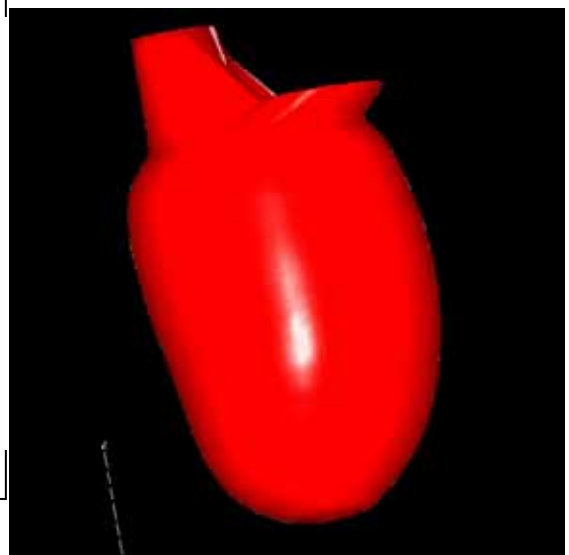


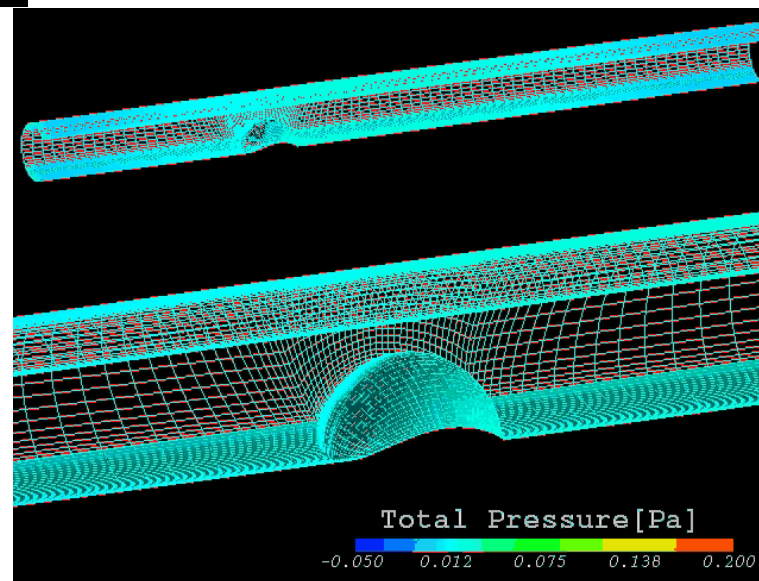
Coupled simulation with flow and structure



$X-Z$

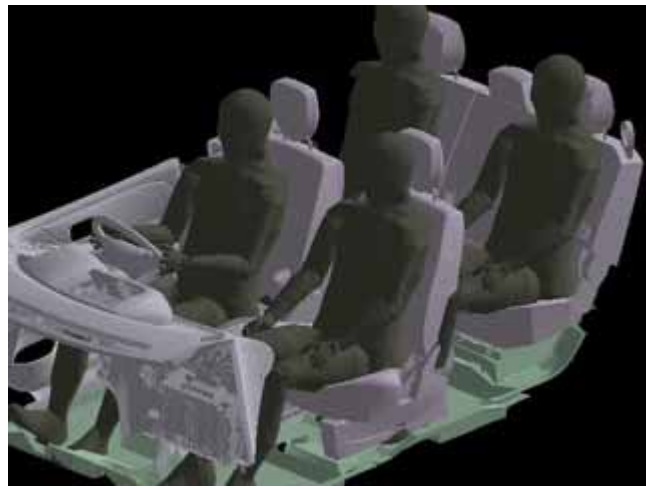
$X-Y$

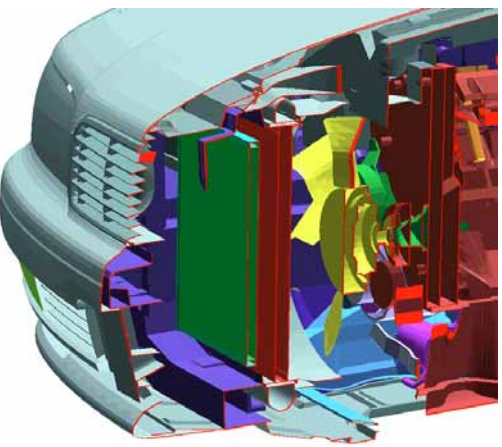


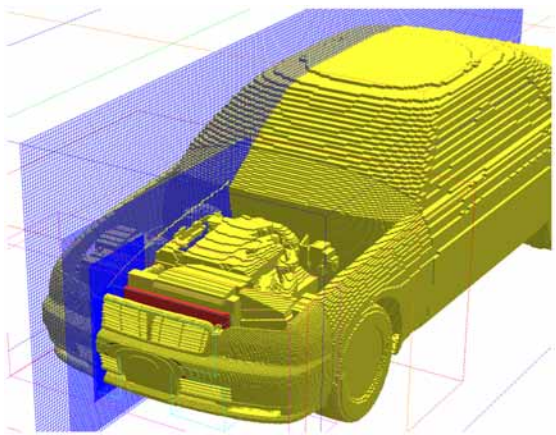


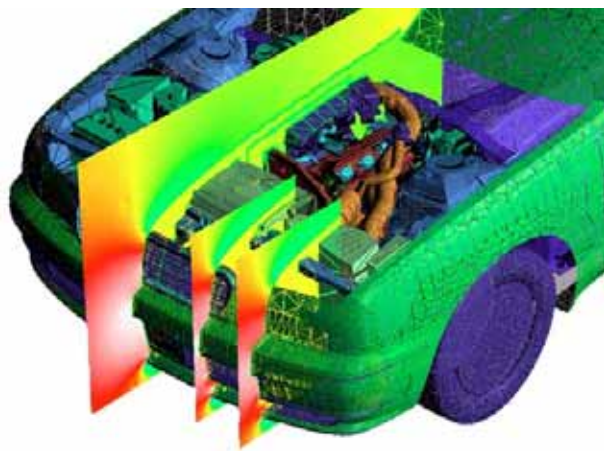
Focused applications: Digital production

- Voxel based simulation
- Multi physics:
  Flow/heat/structure /noise

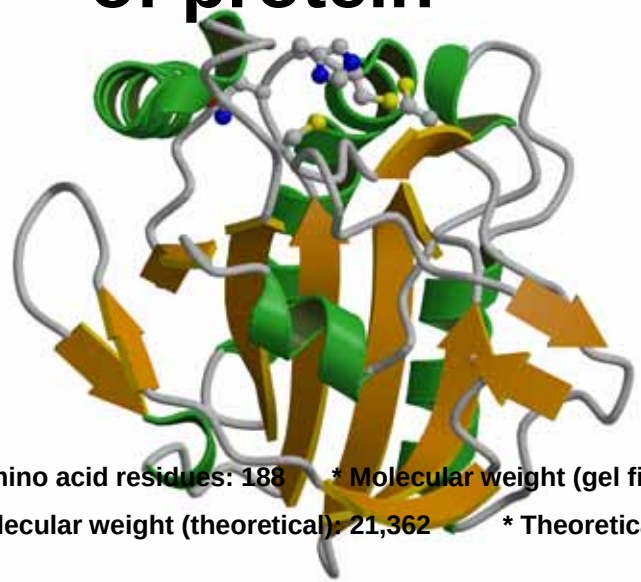






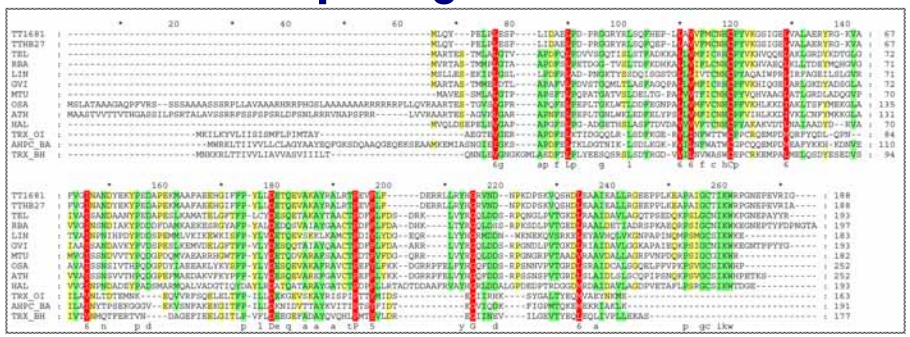


Future orientation: Analysis of unknown function of protein

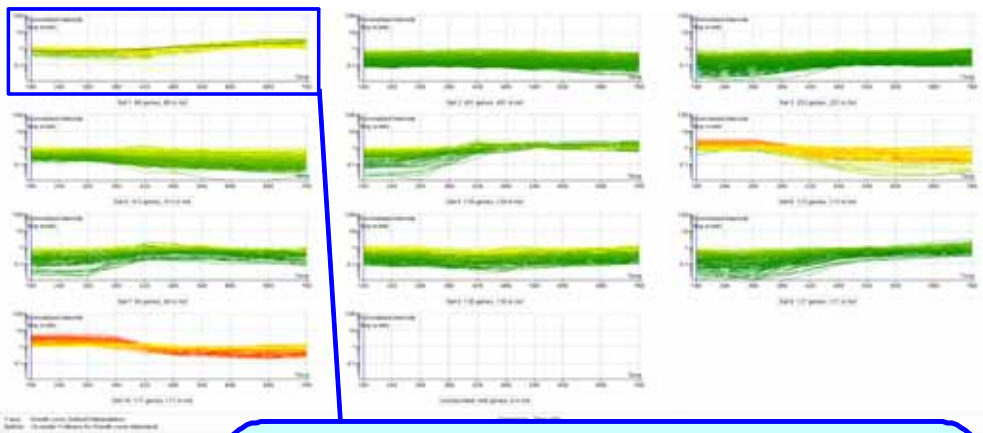


*** Amino acid residues: 188** *** Molecular weight (gel filtration): 20k**

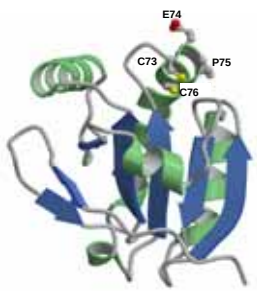
***Molecular weight (theoretical): 21,362** *** Theoretical p*I*: 5.1**



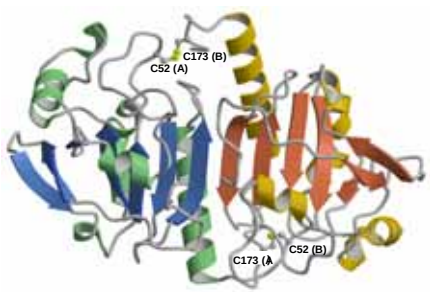
Hypothetical proteins TTHB27, *Thermus thermophilus* HB27; TEL, *Thermosynechococcus elongatus*; RBA, *Rhodopirellula baltica*; LIN, *Leptospira interrogans*; GVI, *Gloeobacter violaceus*; MTU, *Mycobacterium tuberculosis*; OSA, *Oryza sativa*; ATH, *Arabidopsis thaliana*; HAL, *Halobacterium sp.*; TRX_OI, thioredoxin [*Oceanobacillus iheyensis*]; AHPC_BA, AhpC/TSA family protein [*Bacillus anthracis*]; TRX_BH, thioredoxin [*Bacillus halodurans*]



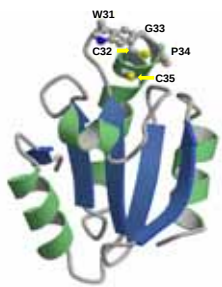
Structural similarity to a thioredoxin-fold (by DALI)



Thiol-disulfide oxidoreductase ResA (PDB ID 1ST9)



Thioredoxin peroxidase 2 (PDB ID 1QQ2)



Thioredoxin (PDB ID 1ERV)

iron-sulfur cluster biosynthesis protein IscA
iron-sulfur cluster biosynthesis protein IscU
probable *c*-type cytochrome
cytochrome c_{552} precursor (C552)
cytochrome caa_3 oxidase subunit I (polypeptide I + III)
cytochrome caa_3 oxidase subunit IIc
probable *c*-type cytochrome
thiol:disulfide interchange protein
ferric uptake regulatory protein

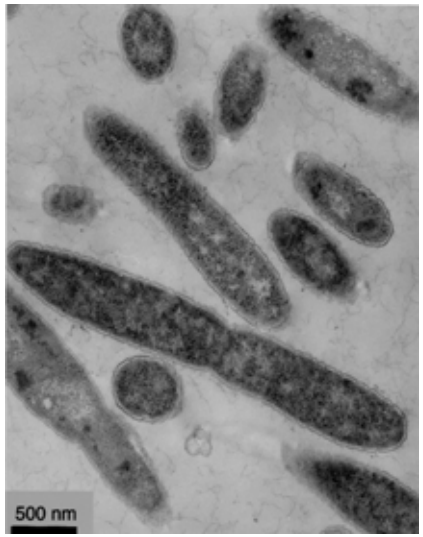
Ebihara et al

Interaction analysis of Protein

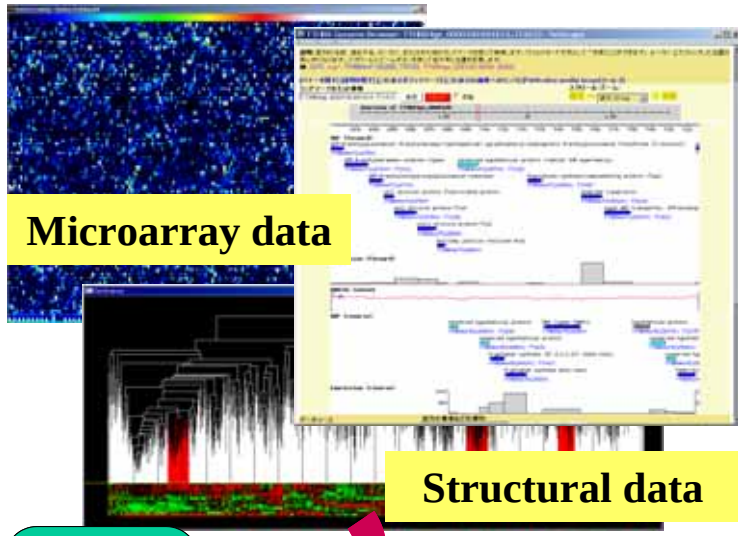
- Quantum mechanics:
  - Precise but high cost
  - Less no. of atoms
- MD simulation
  - Low cost : custom chip MD-grape
  - Large no. of atoms
- Coupled simulation
  - Near field: quantum simulation
  - Other field: MD

Future orientation: simulation of life

1) **Structure and function of each molecule : today's target**
2) **When, where, how does reaction occur?**
3) **Simulation of whole phenomena of life**

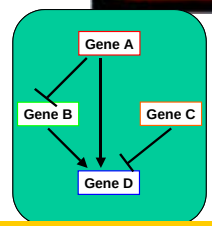


Thermus thermophilus HB8

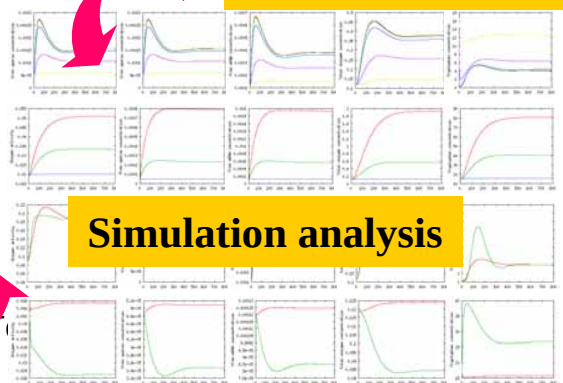


Microarray data

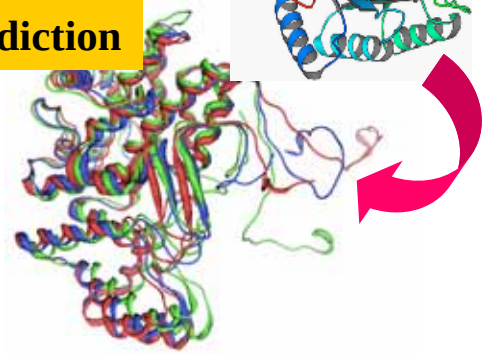
Structural data



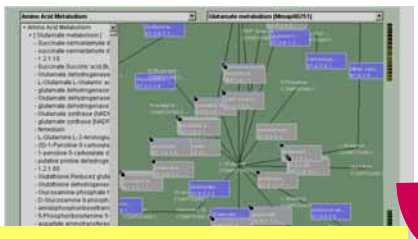
Gene network prediction



Simulation analysis



Molecular simulation



Metabolic pathway map

Future orientation: Multi-scale human simulation

From micro size

From macro size

MD, quantum physics

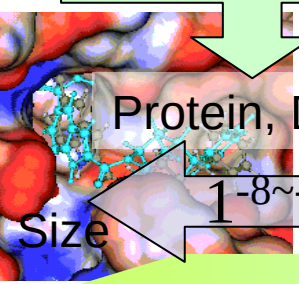
Fluid/structure analysis

Protein rigand reaction

System biology

Flow/Heat/Structure/ Chemical Reaction

Circulatory System Musculoskeletal Model



Protein, DNA

cell

tissue

organ

Size

$1^{-8\sim-6}$

$1^{-5\sim-4}$

$1^{-3\sim-2}$

1^{-1}

1^{0}

Micro

Meso

Macro

Drag design

DDS

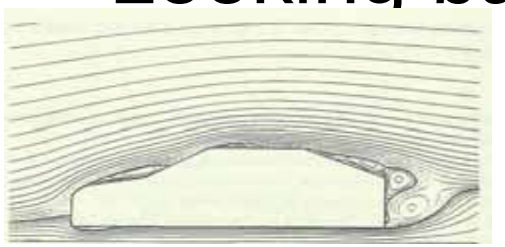
HIFU

Surgical Operation

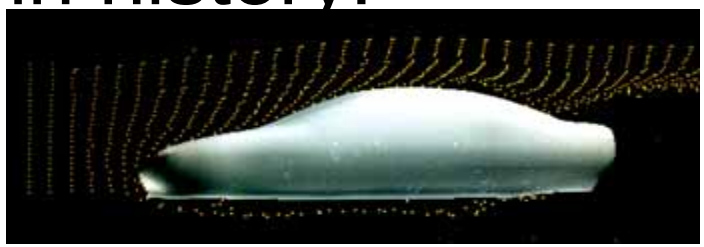
Treatment Technology

Do you believe it?

- Looking back in history:

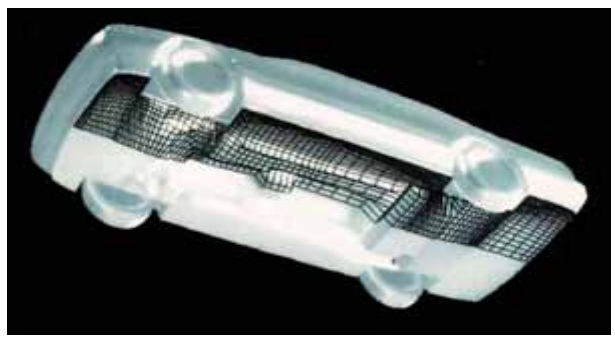


1985

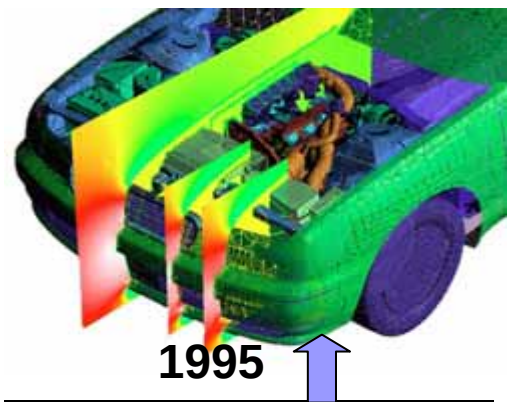


1987

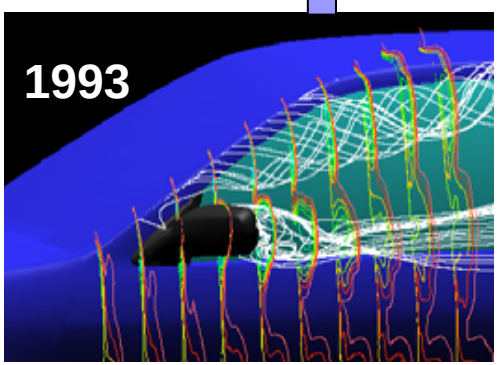
1988

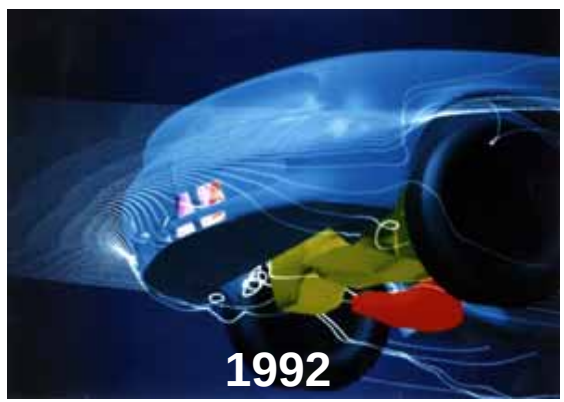


1990



1995





Progress in accuracy

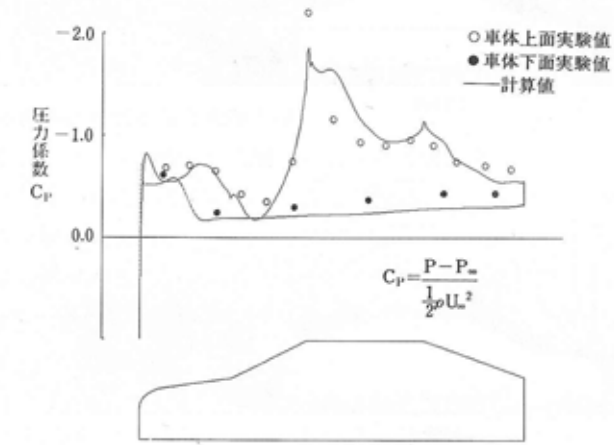
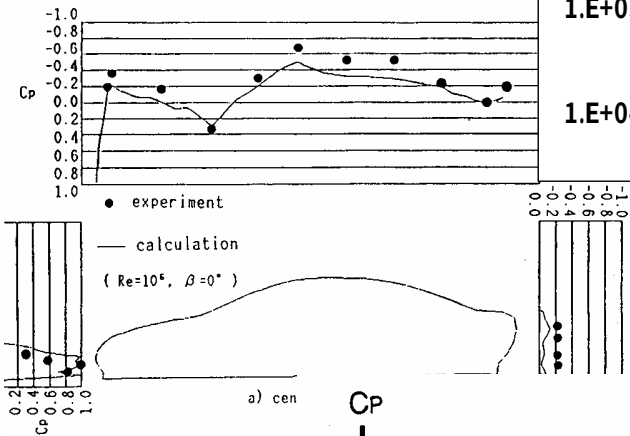
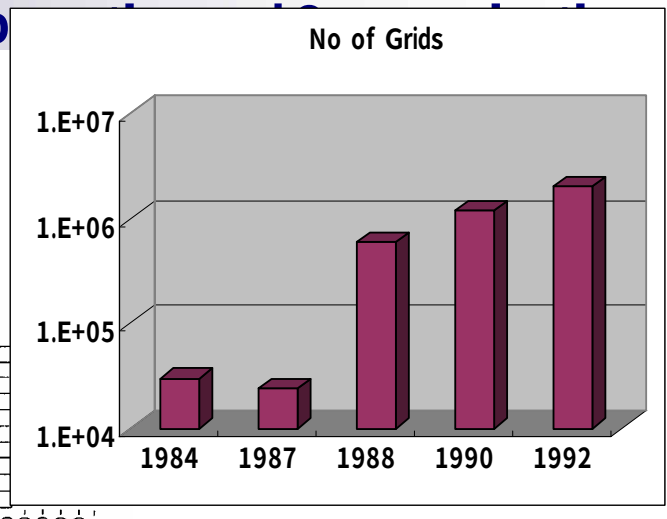
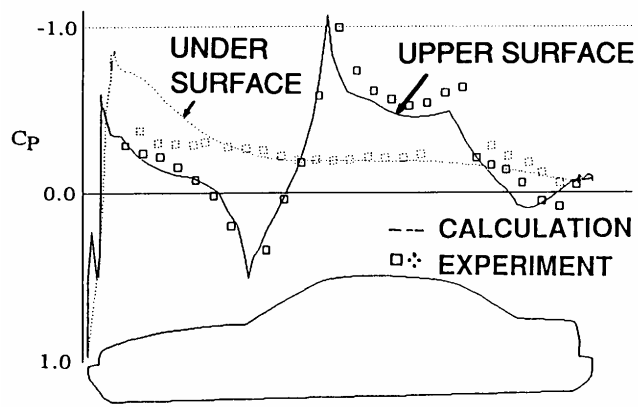


図 4.133　時間平均した車体表面の圧力分布と実験値

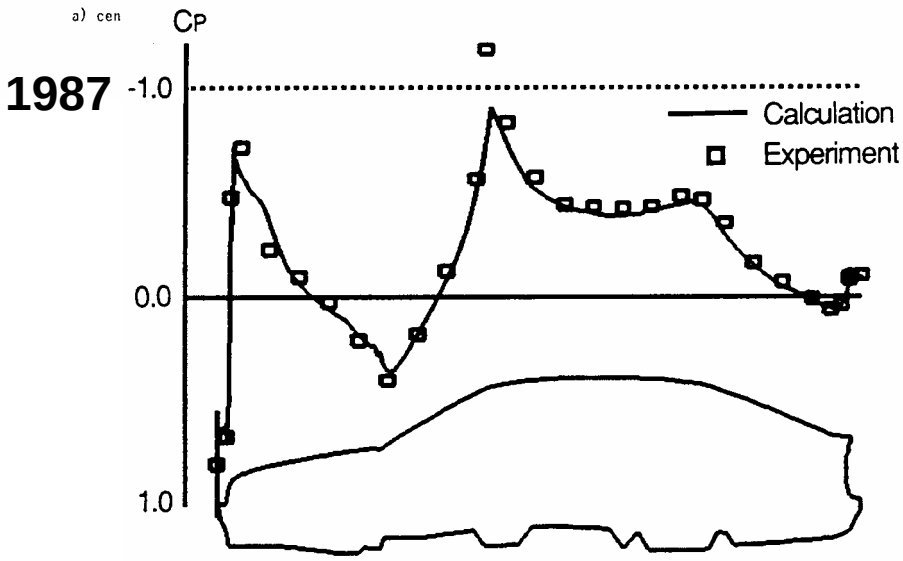
1985







1990



1987

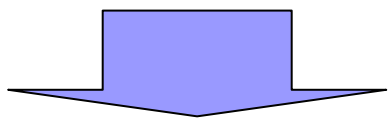
1993

From history

- Progress in 10 years is beyond our imagination
- Needs (or dreams) drive the progress
- Speed-up by hardware is less than speed-up by computation scheme

Future orientation

- Multi-physics
  - Flow, structure, heat, sound, chemical reaction,
  - Multiple governing equations
- Multi-scale
  - From atom, molecule size to human size
  - Multiple governing equations



no single architecture can fit them

Power consumption is key of Peta scale computer

- Estimation of effective 1Peta FLOPS in 2010
  - Vector: efficiency 1/3, 3 Peta in theoretical peak
    - 63 Giga FLOPS/ CPU
    - 16 CPU/node,1 Tera Flops/node, 3,072 nodes in total
    - electric power: 47.1 MW (x 7.85 Earth Simulator)
  - Scalar: efficiency 1/10, 10 Peta in peak
    - 30 Giga FLOPS/CPU
    - 32 CPU/node, 1 Tera Flops/node, 10,240 nodes in total
    - electric power : 40 MW (x 6.7 of Earth Simulator)
  - MD-Grape4: efficiency ½, 2 Peta in peak
    - Electric power: 0.3-0.5 MW (x 0.07 of Earth simulator)
    - Very few limited applications

Efficient special purpose computer is one of key!!

Prospect to Peta FLOPS

1. Blue Gene type approach: Low-power first
   - high-density packaging
   - Lower clock, small memory, fewer function unit, low latency communication, but scalable up to 100,000 for some applications
   - Limited application
2. Mass pipeline approach
   - Many pipelined function units for Matrix and vector operation : CELL, MD-Grape, Grape-DR
   - Unknown efficiency in actual application
3. Choose a set of different architecture to fit more application

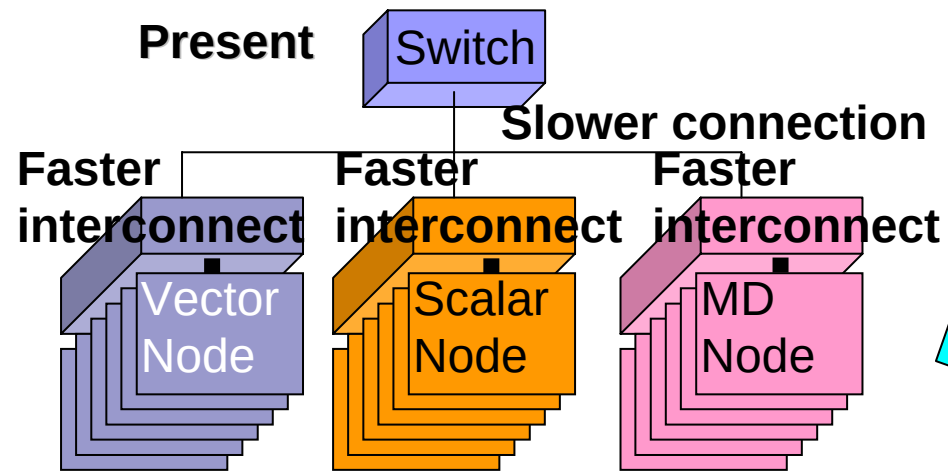
Today's problem and its solution

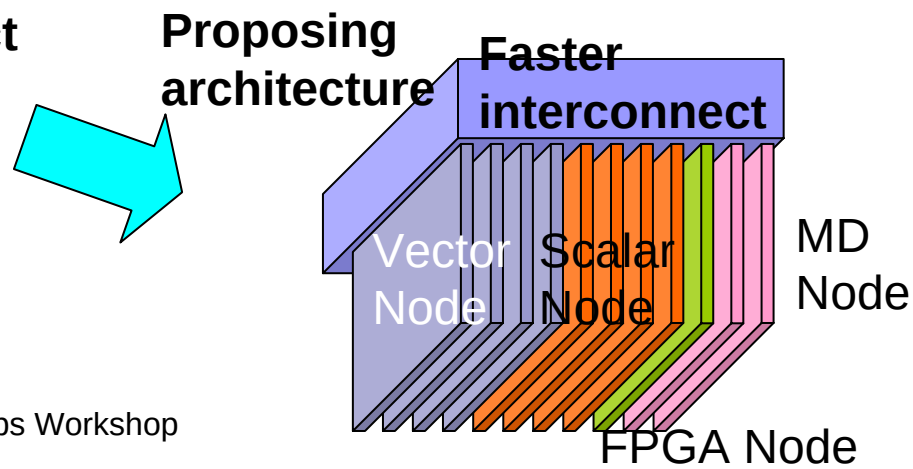
Needs of Multi-scale Multi-physic simulation

Needs of multiple computation components

Integration of multiple architecture

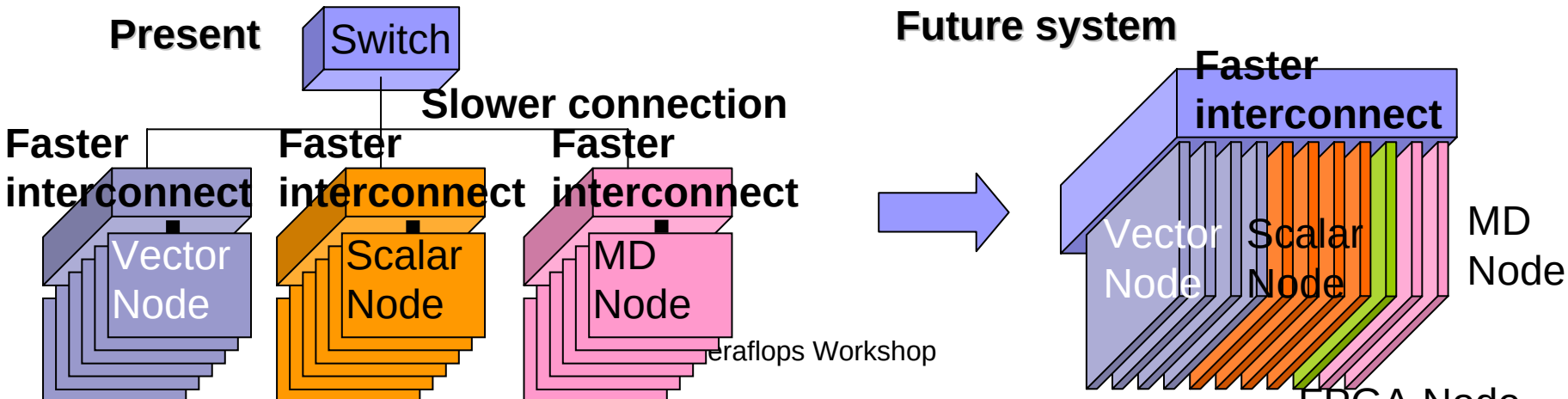
Tightly-coupled heterogeneous computer





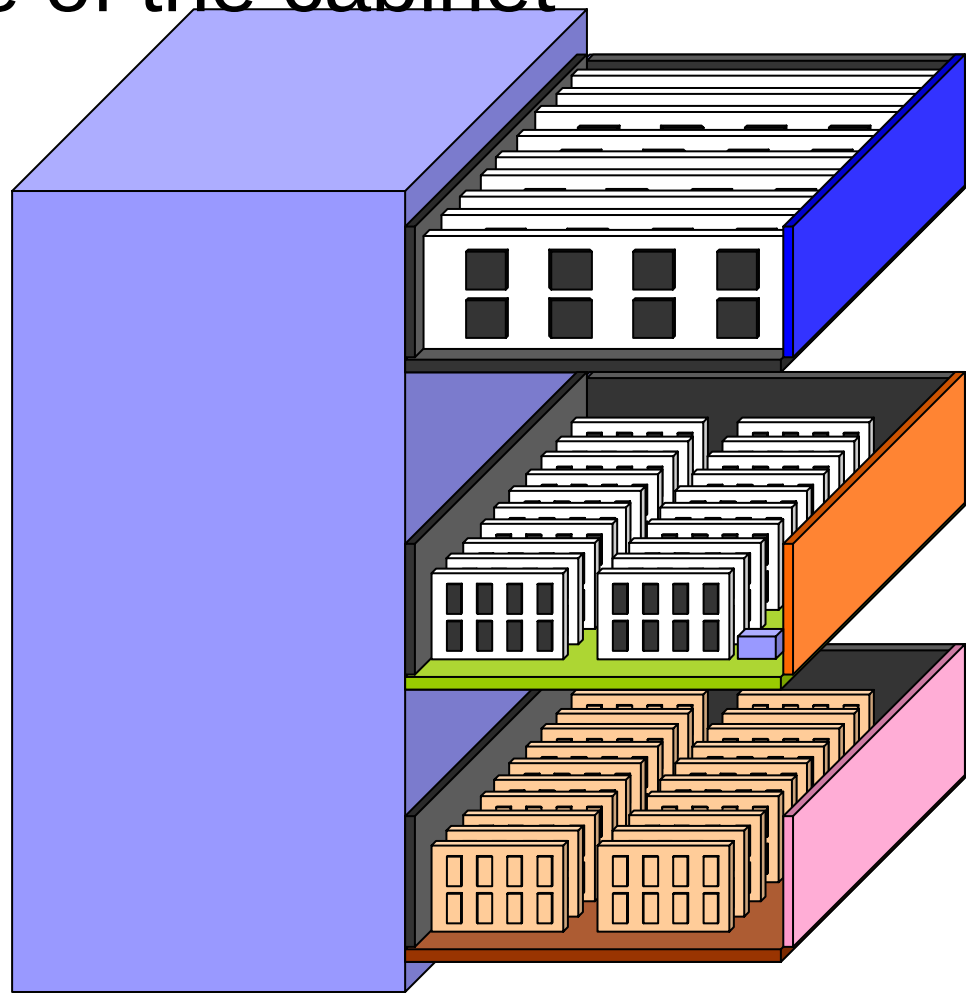
Tightly-Coupled Heterogeneous System

- Scalable, fits any computer center
  - Size, cost, ratio of components
- Easy and low-cost to develop new component
- Scale merit of components
- From application side, natural speed up of grid computing program



In detail

- Image of the cabinet



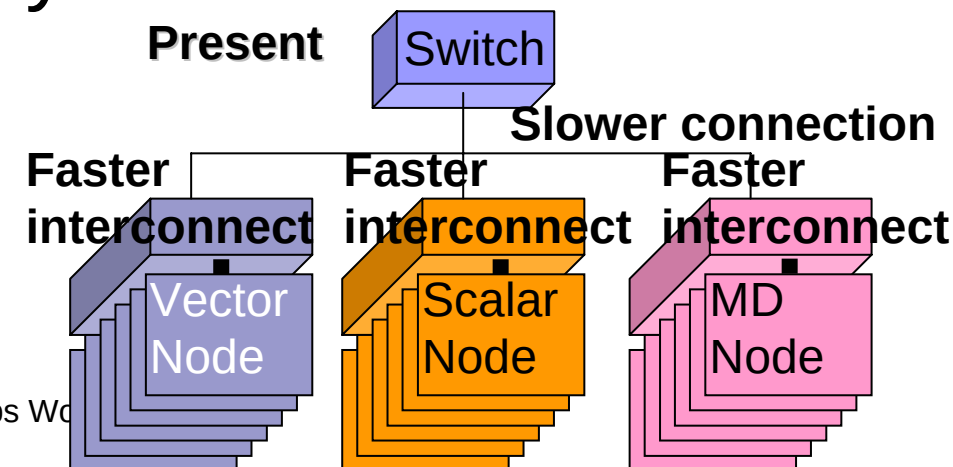
So much traffic between comp.?

- Scalar←→MD: high
- Vector←→MD: high
- Vector←→Vector: highest
- Scalar←→Vector: not so high
- Scalar←→Scalar: high
- Scalar ←→I/O:high
- Vector ←→I/O:high

The bandwidth will be determined by scalar/vector–MD traffic

Why tightly coupled?

- Communication between different sizes or different physics may be small
- However, synchronous CPU allocation is difficult in usual operation and the total bandwidth between vector-scalar, scalar-MD, vector-IO, scalar-IO is beyond commodity interconnect.



Reaction of vendors

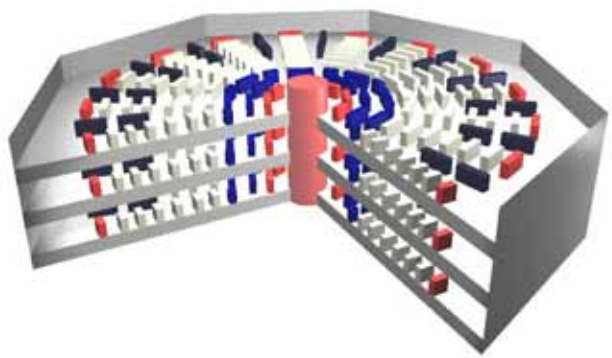
- NEC, Fujitsu and Hitachi accepted the proposed architecture
- We will start feasibility study and design work in more detail
- We also discuss possibility of visualization component with SGI Japan

Need your comment and supports!!

Our target

Start mass production in 2010

- Start operation in March, 2011
- Theoretical peak performance, a few Peta FLOPS: Vector + Scalar, 20+ Peta FLOPS: MD+
- Scalar : vector = 3 : 1
- Memory ¼ of vector + scalar FLOPS
- Nano & molecular biology simulation: >10 Petaflops
- Other multi-scale, multi-physics simulation: >0.5 peta
- 45 nm processing technology
- Optical interconnect and optical switc
- Power consumption
  - 30 MW in peak
  - 60 MW including air-conditioning



Other information

- 2nd machine will ship in 2011
- We are asking government support to Ministry of education, science and technology (sponsor of the Earth Simulator)

Benchmark test program

- For design work, we will start gathering major application programs and make up a benchmark suit.
- Please submit your request and/or programs
  - Our interests are in Nano and bio, multi-scale, multi-physics application