Status of Japanese Petaflops Projects and its Preparation

Taisuke Boku (taisuke@cs.tsukuba.ac.jp)
Deputy Director
Center for Computational Sciences
University of Tsukuba


Outline

- Introduction of CCS, U. Tsukuba
- Status of NGS (Next Generation Supercomputer) system development at RIKEN
- Target applications of NGS
- NGS target applications at U. Tsukuba
- Our preparation (T2K Open Supercomputer Alliance)

Center for Computational Sciences at University of Tsukuba

CCS at University of Tsukuba

- Center for Computational Sciences
- Established in 1992
  - 12 years as Center for Computational Physics
  - Reorganized as Center for Computational Sciences in 2004
- Daily collaborative researches with two kinds of researchers (about 30 in total)
  - Computational Scientists: who have NEEDS (applications)
  - Computer Scientists: who have SEEDS (system & solution)

Our main resource

PACS-CS (14.3TFLOPS)

- 2560 nodes with 2560 Intel Xeon (14.3 Tflops)
- Trunked GbE with 3-D Hyper-Crossbar network topology
- Generally used for all CCS applications, especially for QCD and RS-DFT
- Also opened to nation-wide large simulation projects

FIRST (3.5 + 35 TFLOPS)

- 256 nodes
  - 512 Intel Xeon (3.5 Tflops)
  - 256 BladeGRAPE gravity engine (35 Tflops)
- Specially designed Hybrid Cluster with gravity accelerator for Astrophysics
- Used for project to find “the first object in the Universe”

CCS at U. Tsukuba (cont'd)

- Supercomputer development
  - CP-PACS, PACS-CS, FIRST, …
- Application development and performing
  - Lattice-QCD, DFT, Hybrid Astrophyiscs, Climate, Gene-tree, …
- Contribution to Japanese next generation supercomputer project
  - System development at RIKEN: several faculties are the guest researchers in System Development Group
  - Target application development: 2 out of 21 important apps.
  - System common utilization WG
- Agreement with RIKEN for application development and system performance evaluation
  - Lattice-QCD, Real-Space DFT, FFT

Japanese Next Generation Supercomputer Development at RIKEN

(most of slides are by Dr. Watanabe @ RIKEN)

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

<Goal 1>
Discovery & Creation of Knowledge toward the future

Milky Way formation process

by RIKEN

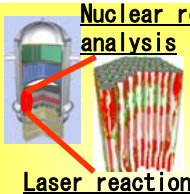
Planet formation process

by RIKEN

Aurora outbreak process

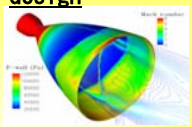
by JAMSTEC

Nuclear reactor analysis



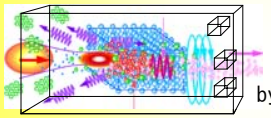
by JAEA

Rocket engine design



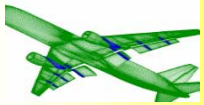
by JAXA

Laser reaction analysis



by JAEA

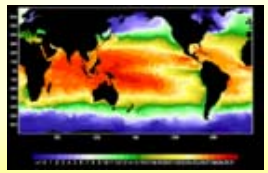
Plane development



by JAXA

< Goal 3 >
Sustainable Development - Consistent with Economy and Environment -

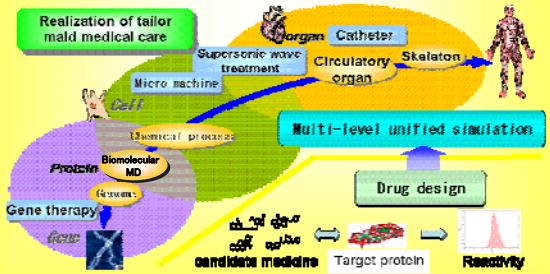
An influence prediction of El Nino phenomenon



by JAMSTEC

< Goal 5 >
Good Health over Lifetime

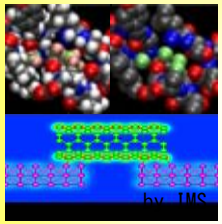
Multi-level unified simulation



by Univ. of Tokyo and RIKEN

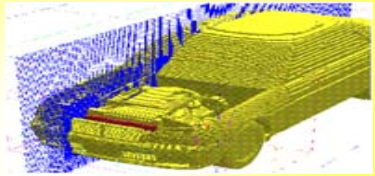
Development and Application of Next-Generation Supercomputer

Nano technology



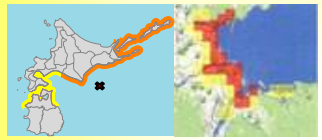
by IMS

Car development



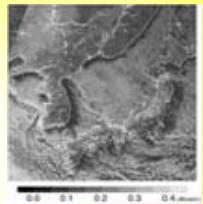
by NISSAN

Tsunami damage prediction



by Tohoku Univ.

Clouds analysis



by MRI

< Goal 2 >
Breakthroughs in Advanced Science and Technology

< Goal 4 >
Innovator Japan - Strength in Economy & Industry -

< Goal 6 >
Safe and secure Nation

Development & Application of Next-Generation Supercomputer Project by MEXT

FY2006: 3,547Million yen / FY2007: *7,736*Million yen
FY2006～FY2012 (total budget expected） about 110billion yen

1. Purpose of policy

Development and implementation of the world's most advanced and high-performance Next-Generation Supercomputer, and to develop and disseminate its usage technologies, as one of Japan's "Key Technologies of National Importance" (National Infrastructure).

2. Expected effects

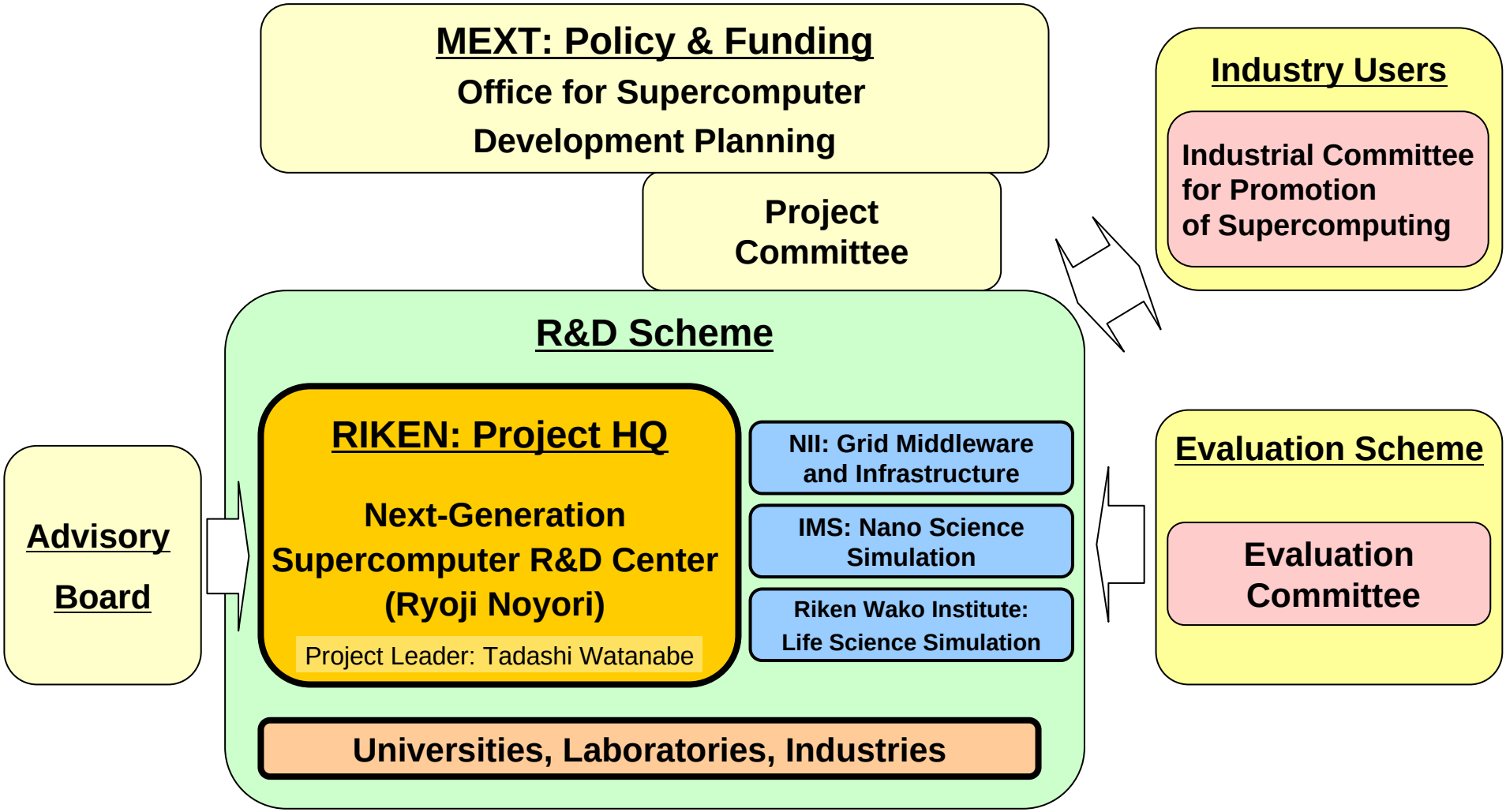
As an important tool for simulation, supercomputing needs to be developed further. This project aims to bring the Next-Generation Supercomputer to completion in 2012.
In order to maintain world-leading position in variety of areas, the following academic-industrial collaboration activities will be conducted under the initiative of MEXT.

(1) Development and implementation of the world's most advanced high-performance Next-Generation supercomputer
(2) Development and dissemination of software that makes optimum use of the supercomputer
(3) Establishment of the world's most advanced and highest standard supercomputing Center of Excellence, which includes the Next-Generation Supercomputer

3. Project Framework

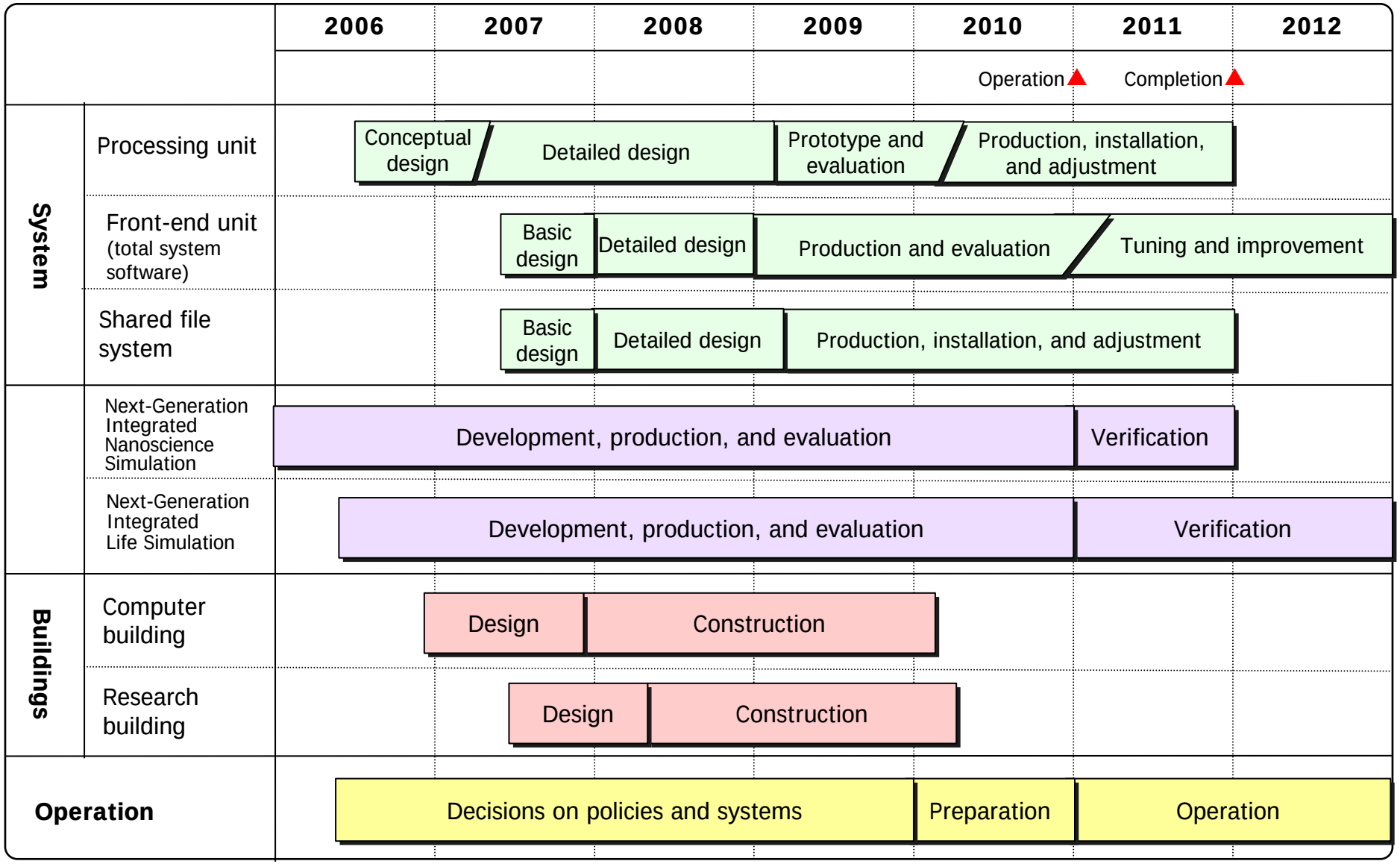
- **Integrated development of computer and software**
- **Establishment of nationwide academic-industrial collaborative structure, with RIKEN as the project headquarters**
- **A new law has been introduced for the framework of usage and administration**

Current Project Organization



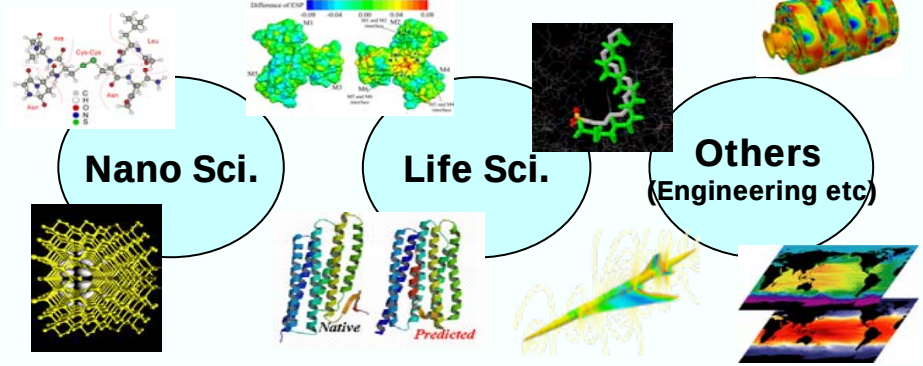
(Note) NII: National Institute of Informatics, IMS: Institute for Molecular Science

SCHEDULE


2006
2007
2008
2009
2010
2011
2012
Operation
Completion
System
Processing unit
Conceptual design
Detailed design
Prototype and evaluation
Production, installation, and adjustment
Front-end unit (total system software)
Basic design
Detailed design
Production and evaluation
Tuning and improvement
Shared file system
Basic design
Detailed design
Production, installation, and adjustment
Next-Generation Integrated Nanoscience Simulation
Development, production, and evaluation
Verification
Next-Generation Integrated Life Simulation
Development, production, and evaluation
Verification
Buildings
Computer building
Design
Construction
Research building
Design
Construction
Operation
Decisions on policies and systems
Preparation
Operation

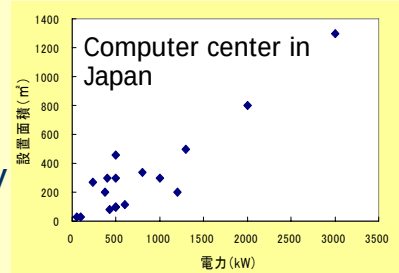
System Design

Requirement from Grand Challenges



Requirements from Computer Centers

- Power, Space
- Reliability, operability
- Cost (development, manufacturing, maintenance)



Popular numerical scheme

FMO, MO, RISM, Real space density functional method, FFT, Monte Carlo, Searching algorism, FEM, FVM, Linear Solver

Optimal system

21 Selected Target applications

- Nano Science (6)
- Life Science (6)
- Climate/ Geoscience (3)
- Engineering. (4)
- Physics/ Astro. (2)

Other Project Watch

Technology Survey

Operation & Utilization

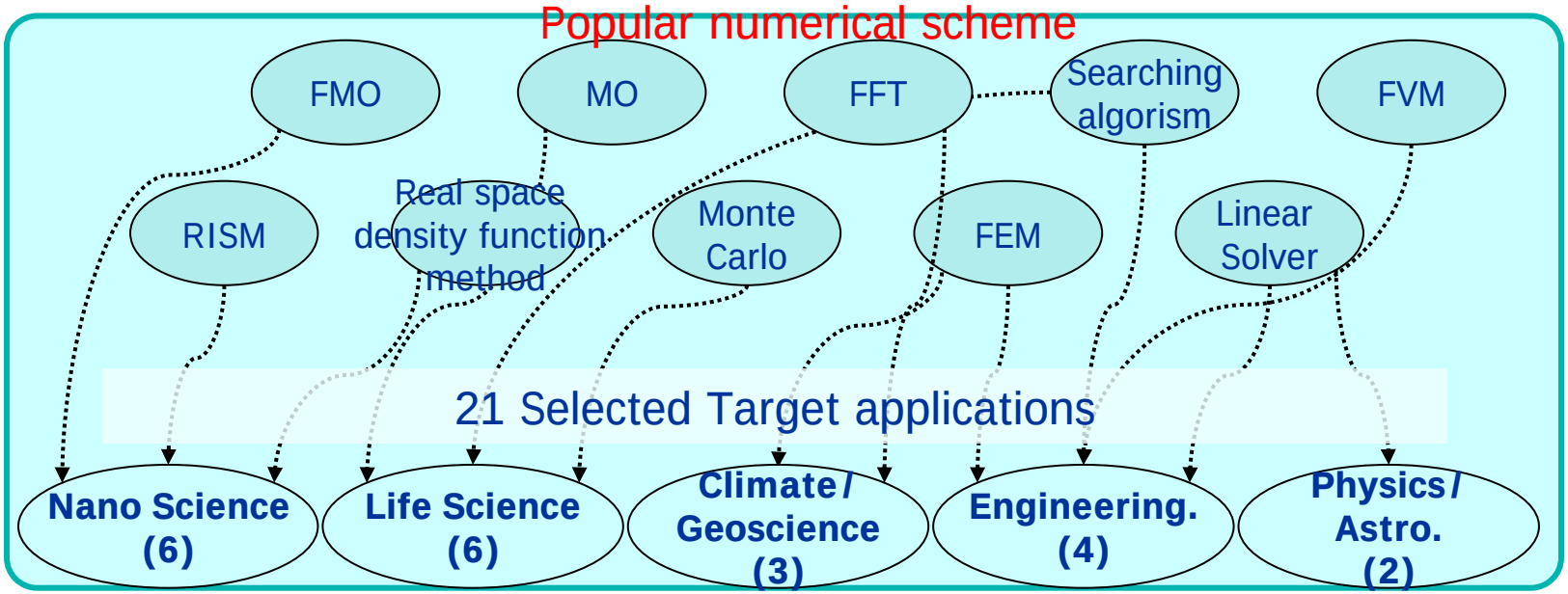
Essential Element technologies

LSI process, Power saving SOI Low-k High-k, Light transm tech., Software OS, compiler

Spin off to the consumer electronics

Technology Limit

Target applications for system performance estimation



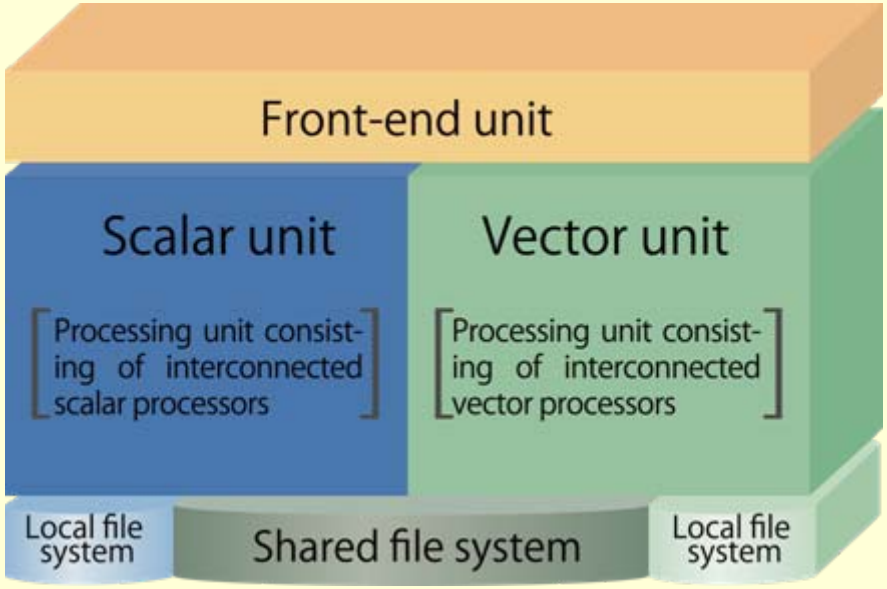
- 21 application programs have been identified as candidates to be used for system performance estimation by the application committee.
- Modified to be a benchmark test suite
  - benchmarking on the desk on each proposed system

The Next-Generation Supercomputer project

The Next-Generation Supercomputer project started in 2006 which is being carried out by RIKEN, with partners in industry, universities, and the government, under an initiative by MEXT (the Ministry of Education, Culture, Sports, Science and Technology).

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 configuration]

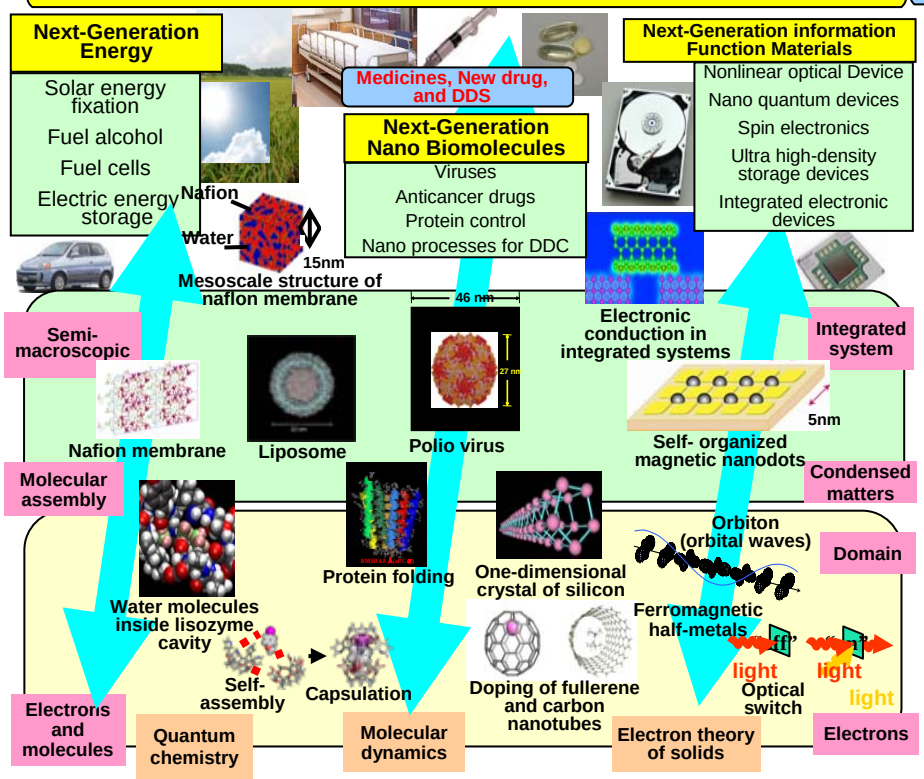


The Next-Generation Supercomputer will be hybrid general-purpose supercomputer that provides the optimum computing environment for a wide range of simulations.

- 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.

Grand Challenge Applications

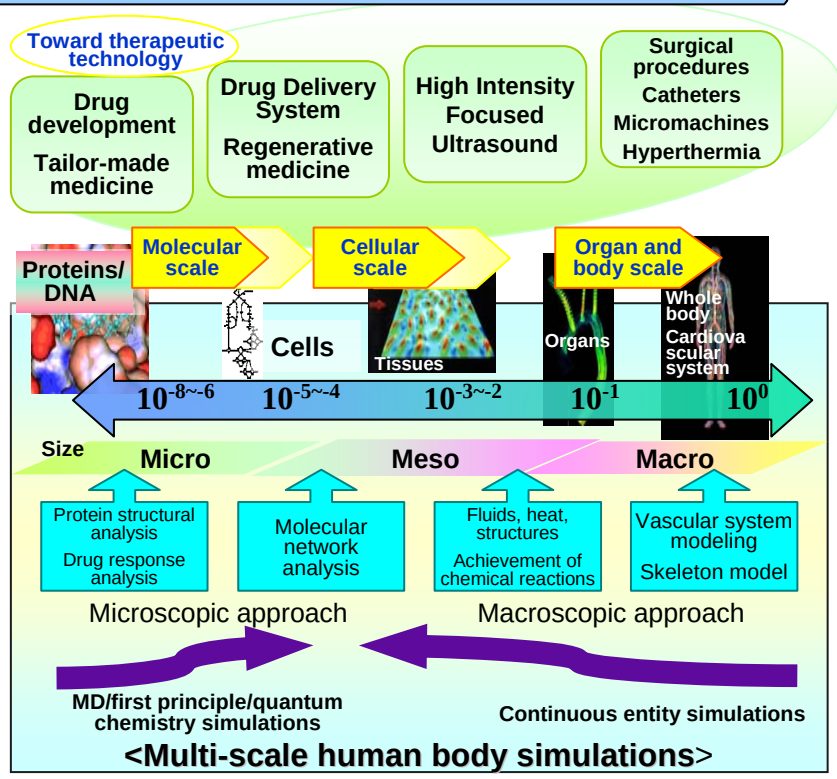
Next-Generation Integrated Nano-Science Simulation Software (2006–2010)



Base site: Institute for Molecular Science

To create next-generation nano-materials (new semiconductor materials, etc.) by integrating theories (such as quantum chemistry, statistical dynamics and solid electron theory) and simulation techniques in the fields of new-generation information functions/materials, nano-biomaterials, and energy

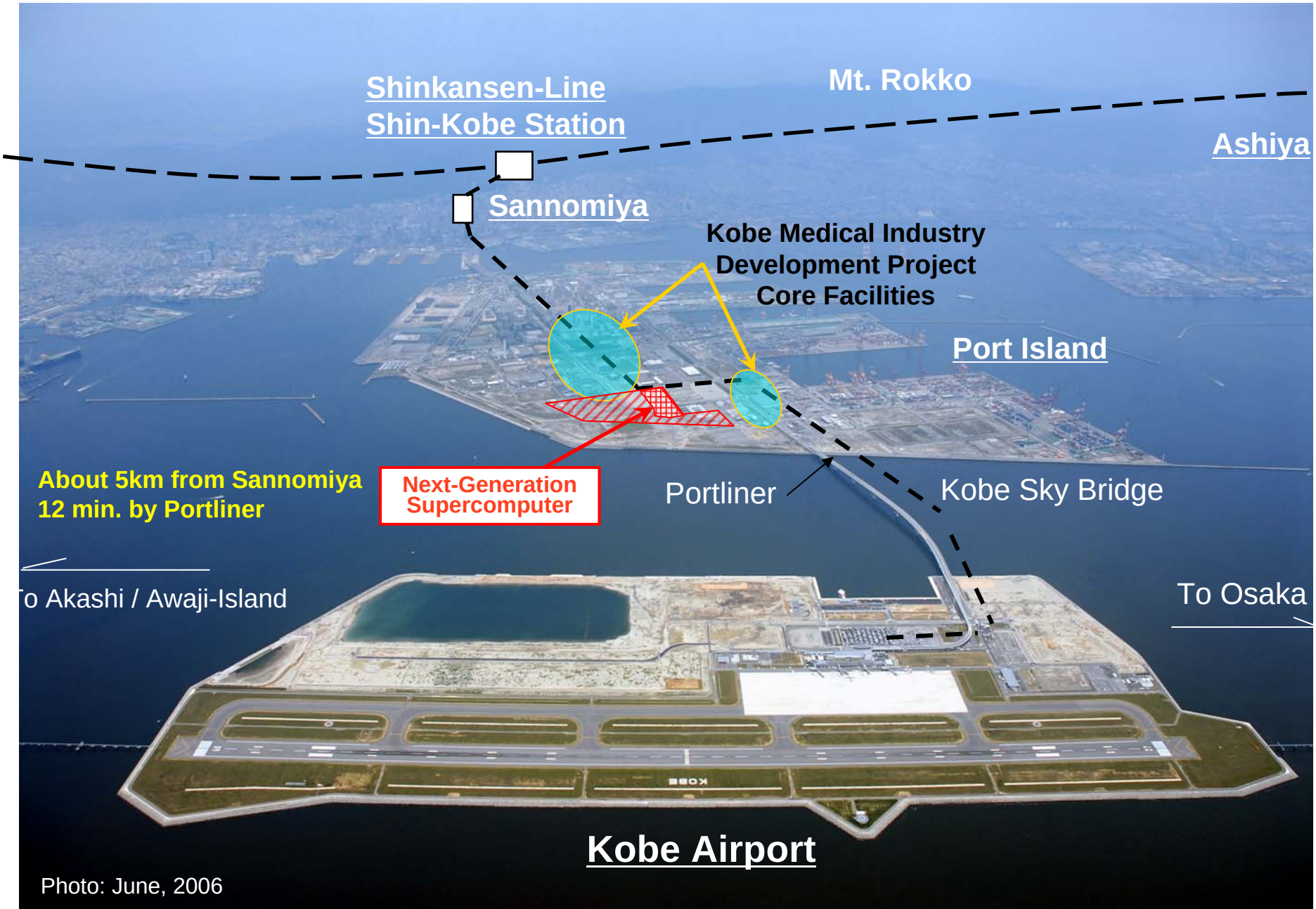
Next-Generation Integrated Life-Science Simulation Software (2006–2012)



Base site: RIKEN Wako Institute

To provide new tools for breakthroughs against various problems in life science by means of petaflops-class simulation technology, leading to comprehensive understanding of biological phenomena and the development of new drugs/medical devices and diagnostic/therapeutic methods

System Installation Venue: Kobe in Kansai area



Status of NGS development at RIKEN

- Spent about 1 year for conceptual system design(~ May 2007)
- Now, basic design is done and detailed design is undergoing for about 2 years (~ early 2009)
- RIKEN contracted with two system groups
  - Fujitsu (scalar unit)
  - NEC + Hitachi (vector unit)
- Academia's contribution on system design and evaluation
  - U. Tsukuba, U. Tokyo, Kyushu U.
- System performance target (not on real applications) with 10 Pflops-scale machine:
  - #1 on TOP500 with Linpack
  - Several #1's on HPCC benchmarks

in 2012

Target applications of NGS

21 important applications

- Nano Science
  - GAMESS/FMO
    - FMO molecular orbital calculation
  - Modylas: lower scalability
    - Massively-parallel multipurpose software for molecular dynamics calculation
  - Octa
    - Coarse-grained molecular dynamics calculation
  - PHASE
    - First-principles molecular dynamics simulation within the plane-wave pseudopotential formalism
  - RISM
    - Analysis of electron status of protein in solution with the 3D-RISM/FMO Method
  - RSDFT
    - Ab-initio molecular dynamics calculation in real space

21 important applications (cont'd)

- Life Science
  - GNISC
    - Inference of genetic networks from experimental data of gene expression
  - MC-BFlow
    - Simulation for blood-flow analysis
  - MLTest
    - Validation of statistical significance for development of individualized medicine
  - MyPresto
    - Docking simulation of protein and drug
  - ProteinDF
    - Ab initio molecular dynamics calculation in large-scale protein systems
  - SimFold
    - Prediction of protein structure

21 important applications (cont'd)

- Earth Science
  - COCO
    - Super-resolution ocean general ciculation model
  - NICAM
    - Nonhydrostatic icosahedral atmospheric model for global cloud resolving simulations
  - Seism3D
    - Simulation of seismic-wave propagation and strong ground motions
- Physics/Astronomy
  - LatticeQCD
    - Study of elementary particle and nuclear physics based on lattice QCD simulation
  - NINJA/ASURA
    - Super large-scale gravitational many-body simulation for finding the celestial origin

21 important applications (cont'd)

- Engineering
  - Cavitation
    - Computation on unsteady cavitation flow by finite difference method
  - FrontFlow/Blue
    - Unsteady flow analysis based on large eddy simulation
  - FrontSTR: large scale, communication bound
    - Structural calculation by finite element method
  - LANS
    - Computation of compressible fluid in aircraft and spacecraft analysis

Scalability problem on applications

- Relatively small number of applications which have (very) high scalability
  - NGS is designed for capability computing to support ultra-high parallelism
  - 10 Pflops scale machine consists of hundred thousands of CPU cores
  - Physics/Astronomy applications are quite highly scalable
  - Nano science & earth science include several quite highly scalable apps., but not all
  - In life science field, most of apps. have low scalability but require a large number of jobs (capacity computing)

Scalability problem (cont'd)

- In last symposium (Sept. 2007) on NGS, a large number of applications were introduced, but very few discussion on quantitative evaluation on scalability
- Computer scientists strongly warned on application's scalability
  - Capability computing on NGS is for "solving large scale problems in reasonable computation time", not for "solving small scale problems in very short computation time"
  - Most of computational scientists have a dream on the latter case
  - Capacity computing on NGS is also important, but a highly scalable interconnection network is required for that purpose
  - This situation will be reflected to the total architecture of NGS (maybe)
- Inter-disciplinary study is quite important
  - Between computational scientists and computer scientists
  - Between different fields of computational scientists

Target Applications Developed in CCS for Next Generation Supercomputer

(thanks for slide materials
Prof. Ukawa, Prof. Oshiyama & Dr. Iwata @ CCS)

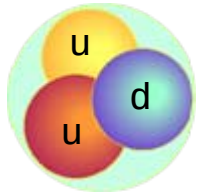

NGS target applications in CCS, U. Tsukuba

- 2 apps. out of 21
  - LatticeQCD & RS-DFT (Real-Space Density Function Theory)
  - Both apps have very high scalability with domain decomposition on large problem space
- CCS & RIKEN agreement on application development, tuning and performance evaluation on NGS
  - LatticeQCD
  - RS-DFT
  - FFT (for HPCC)
- Both codes are now free from "*vector*" machines
  - Instruction level tuning
  - Cache-aware tuning

Standard model of fundamental particles

- Particles to construct materials
  - Quark (6 types) $\begin{pmatrix} u \\ d \end{pmatrix}$ $\begin{pmatrix} s \\ c \end{pmatrix}$ $\begin{pmatrix} t \\ b \end{pmatrix}$
  - Lepton (6 types) $\begin{pmatrix} e \\ \nu_e \end{pmatrix}$ $\begin{pmatrix} \mu \\ \nu_\mu \end{pmatrix}$ $\begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix}$

Hadrons (proton, neutron, π) consist of 3 or 2 quarks



Proton=uud

- Guage particles to transfer interaction
  - Photon γ Magnetic interaction
  - Weak bozon W, Z Weak interaction
  - Glue-on g Strong interaction

Quantum Chromo-Dynamics (QCD)

Quantum Chromodynamics

Gross-Wilczek-Politzer 1973

- Basic theory on strong interaction
- Interaction among quarks and glue-ons can be described with four-dimensional (3-D space + time) “field”

$q_f(x)$ Quark field

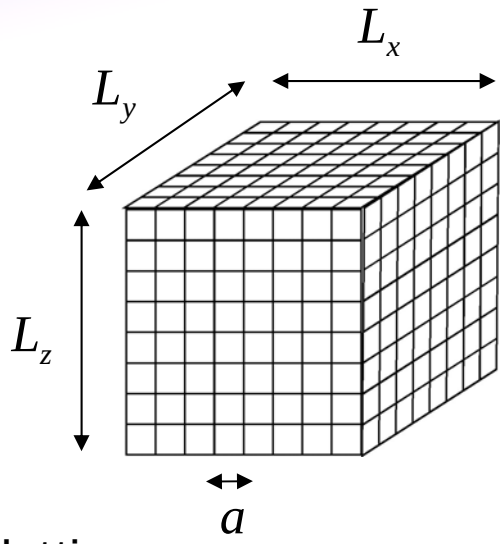
$A_\mu(x)$ Glue-on field

} Defined field on 4-dim. Space & time point X

- Quantum physical system by the strong interaction on non-linear freedom of infinite numbers

→ Impossible to solve analystically (with “pencil and paper”), and a large scale simulation is the only method

Basic variables and operations



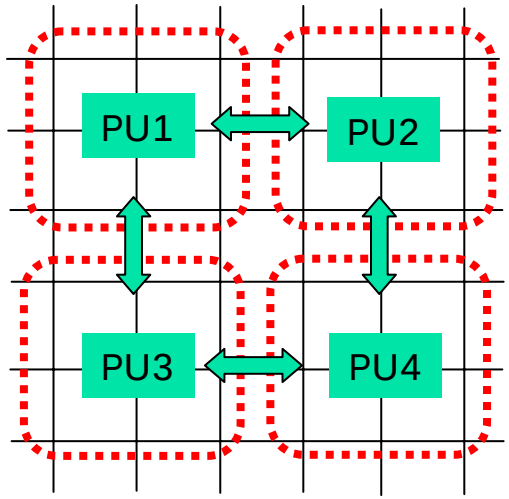
- 4-dim. simple cubic
  - # of lattice
  - Lattice distance

$$\frac{V}{a} = L_x \times L_y \times L_z \times L_t$$

- Basic variables
  - Glue-on field U_n^{ab} 3x3x4xV vector on between lattice
  - Quark field $q_n^{\alpha a}$ 3x3x4xV vector on lattice

- Parallelization
  - Decomposing the physical lattice into logical ones and mapping on processor array
  - Interactions of basic variables are only between nearest neighboring lattice

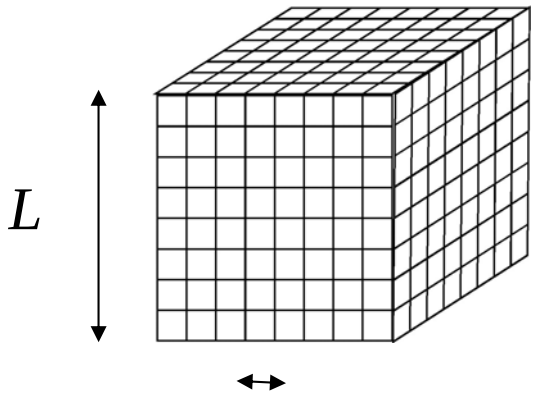
    → local communication only



Highly scalable

Scaling on floating-point operations

- Basic parameters:
  - Quark mass m_π / m_ρ
  - Lattice size $L\,(fm)$
  - Lattice distance $a\,(fm)$



$1\,fm = 10^{-15}\,m$

- # of operations in traditional HMC algorithm

$$\#FLOP's = C \cdot \left[\frac{\#conf}{1000}\right] \cdot \left[\frac{m_\pi / m_\rho}{0.6}\right]^{-6} \cdot \left[\frac{L}{3\,fm}\right]^{5} \cdot \left[\frac{a}{0.1\,fm}\right]^{-7} Tflops \cdot year$$

$$C \approx 2.8$$

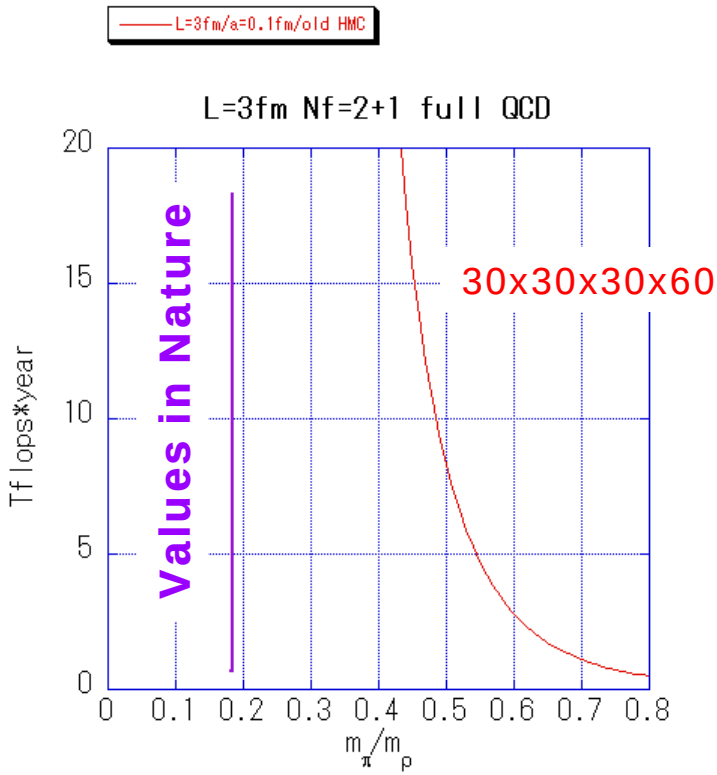
- Acceleration by domain decomposition & multi time-step method

$$\#FLOP's = C \cdot \left[\frac{\#conf}{1000}\right] \cdot \left[\frac{m_\pi / m_\rho}{0.6}\right]^{-4} \cdot \left[\frac{L}{3\,fm}\right]^{5} \cdot \left[\frac{a}{0.1\,fm}\right]^{-7} Tflops \cdot year$$

$$C \approx 0.11$$

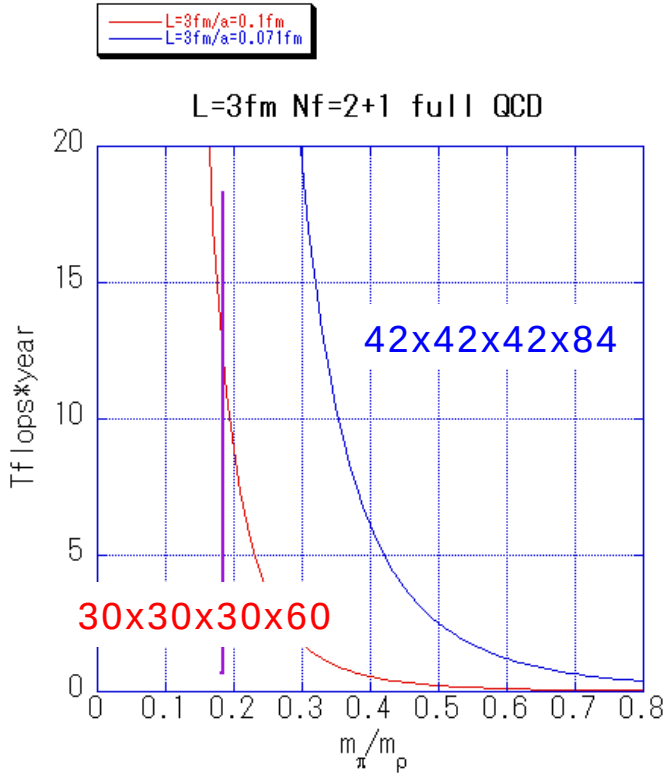
Current Lattice QCD # of operations

Traditional HMC algorithm



Quark mass

Accelerated HMC algorithm

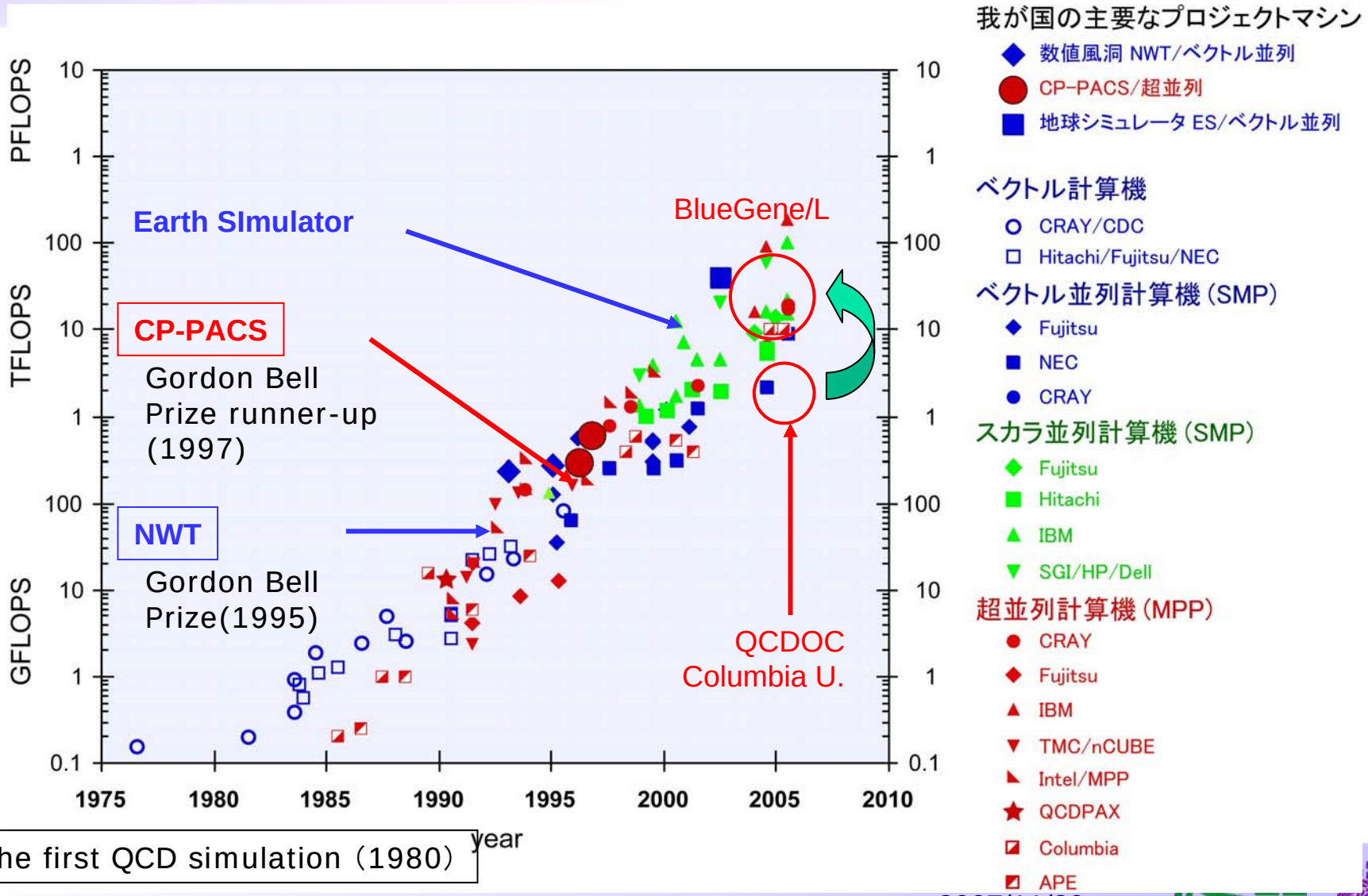


Quark mass

Characteristics of Lattice QCD

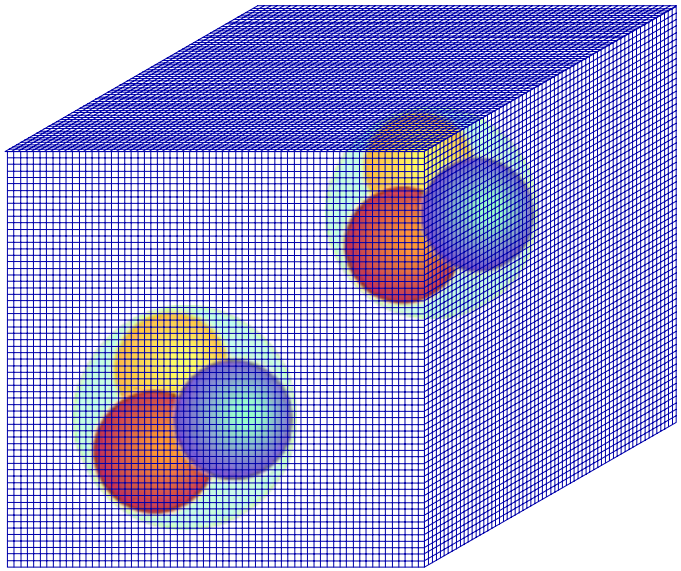
- Simple computing construction
  - 4-dimensional, but simple cubic
  - Single scale problem (not multi-scale)
- Heavy load both on computation and communication
  - Complex number calculation
  - Nearest neighboring communication
- A large amount of floating-point operations
  - Hard scaling toward parameters close to natural feature
    - Condition number of quark-matrix is proportional to 1 / (quark mass)
    - Larger physical size is better
    - Smaller lattice distance is better
  - Precise computation is required
    - Considering both statistical error and structural error, then make it less than few percents

Lattice QCD and top-class machines

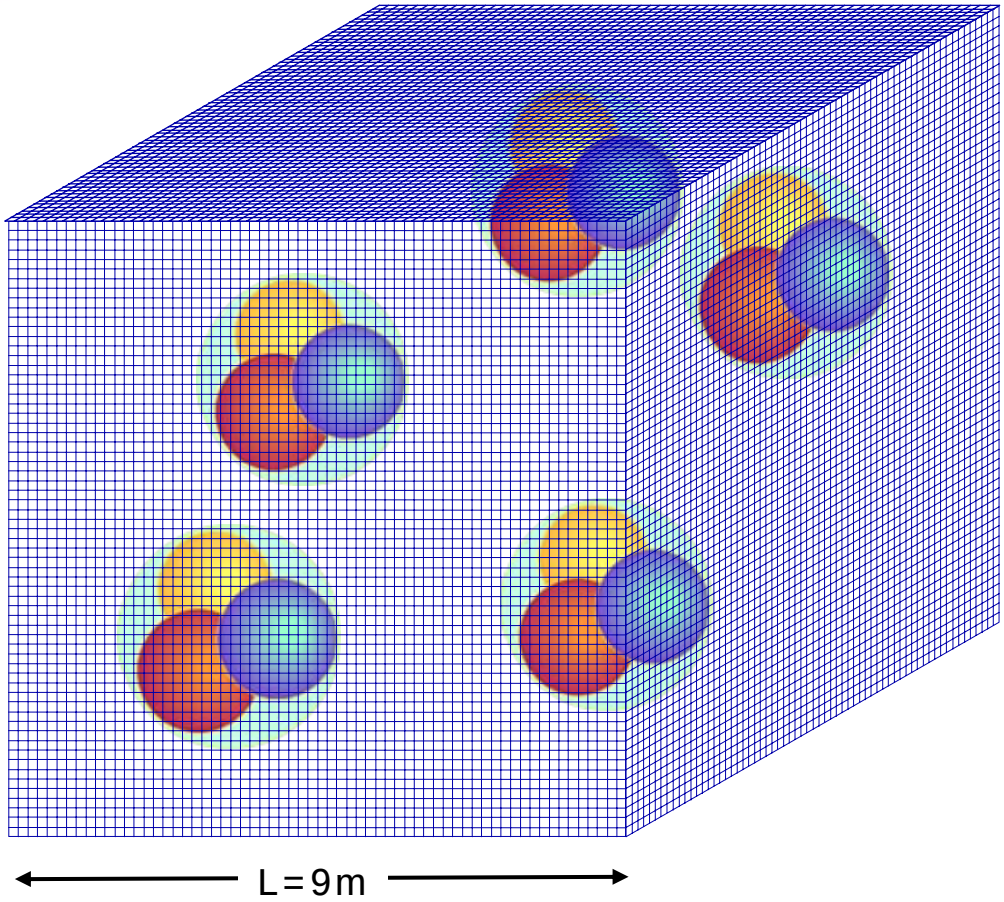


The first QCD simulation（1980）

Lattice QCD on NGS: from particle to nuclear & Universe

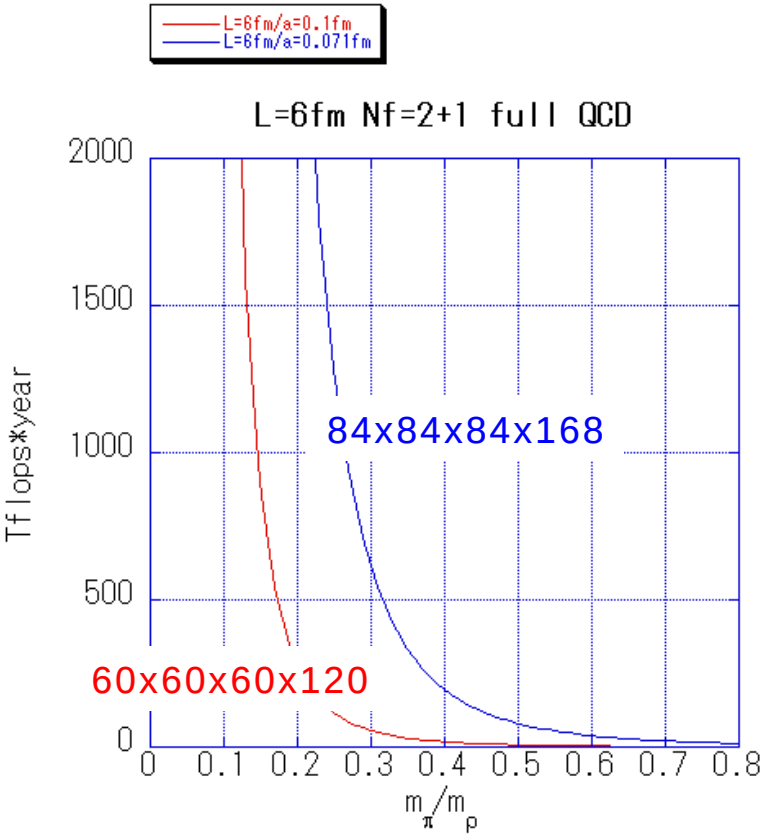


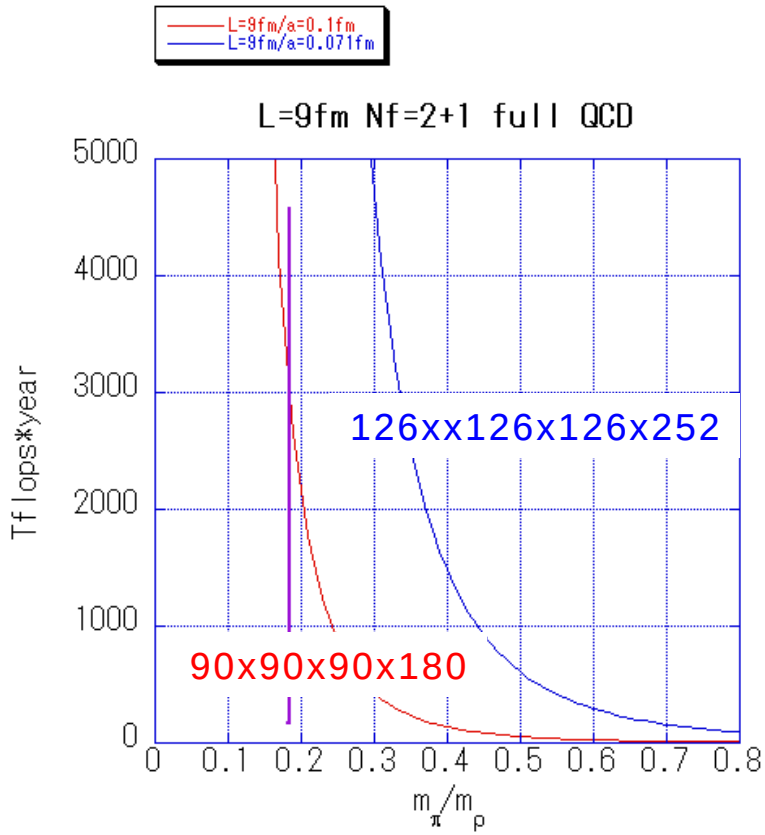
Hadron-hadron interaction



Hadron gas ⇔
Quark/Glue-on plasma phase transition

Increasing of floating-point operations on NGS

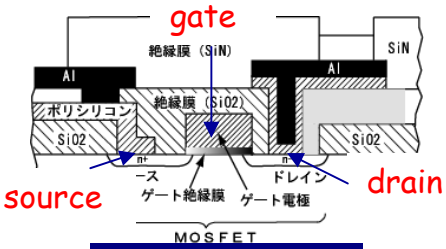




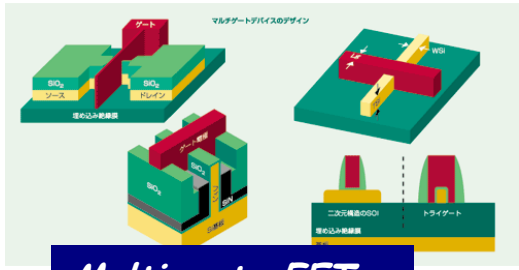
O(1) Pflops * year for problem solving

Nano science for post-scaling technology

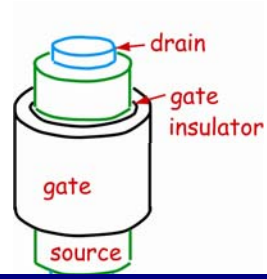
Ex) Field Effect Transistor (More gates to increase channels)



Planer FET

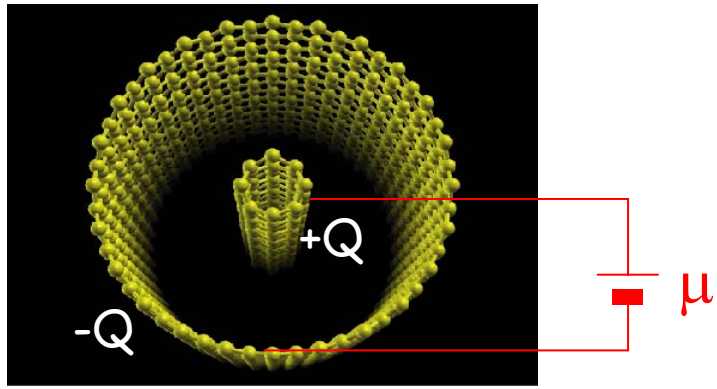


Multi-gate FETs



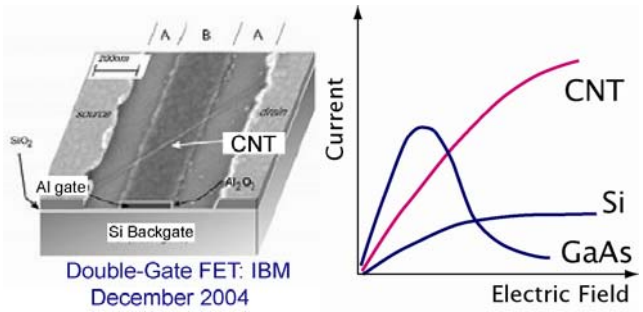
Surrounding-gate FET

Nano = quantum
⇒
Computational Quantum Physics
⇒
Device process simulation is shifting to Quantum predictor construction instead of Classical Theory



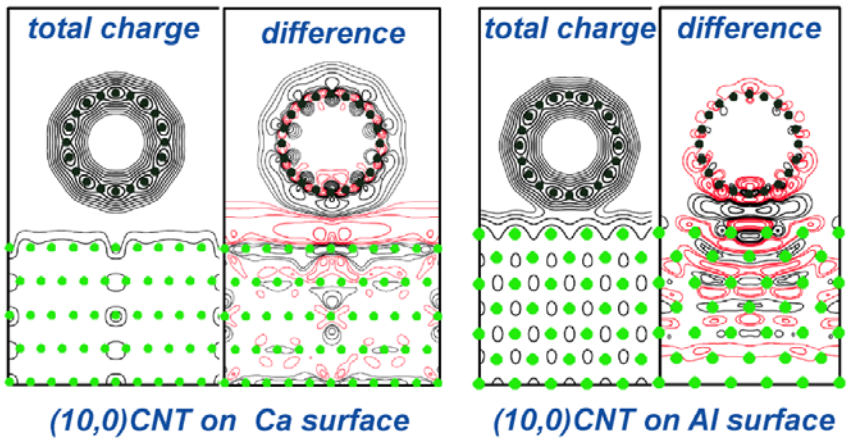
Carbon-Nano-Tube:
The ultimate cylinder structure

Hybrid material: the material of future



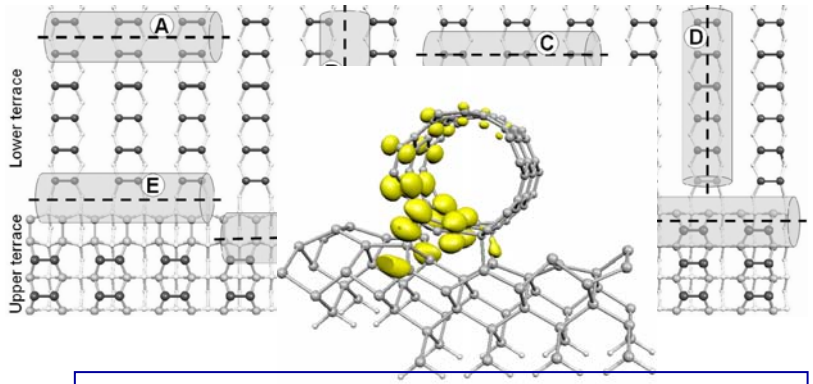
For future hybrid material simulation, nano-scale surface analysis of atomic scale is necessary

Carbon-Nano-Tube FET



Electron movement on the border of metal/carbon
→ Quantum theory is necessary

Okada & Oshiyama: Physical Review Letters (2005)



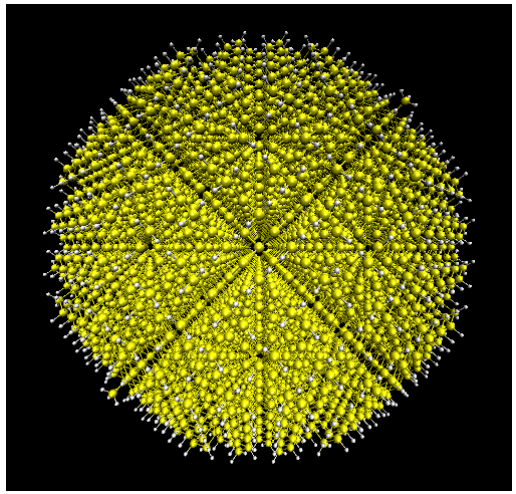
Wiring of Carbon-Nano-Tube on Silicon surface
→ Fabrication based on Quantum simulation

Berber & Oshiyama: Physical Review Letters (2007)

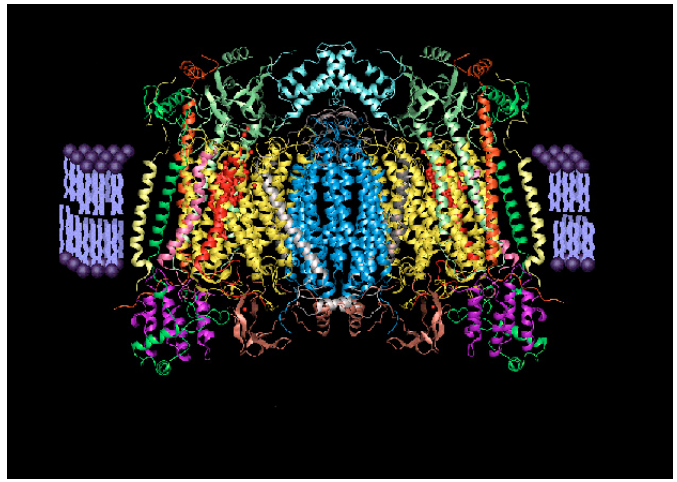
Real-Space Density Functional Theory

～100 atoms ・・・ ordinary system size for current DFT study

3nm size Si Cluster (～7000atoms)



Cytchrome c Oxidase (～30000 atoms)



Real-Space Method is suitable for parallel computations compared to the conventional plane-wave basis method. So we have developed RSDFT in order to study extremely large systems by using massively parallel computers.

Real-Space Finite-Difference Method and Parallel Computation

KS equation is solved as finite-difference equation.

KS equation (finite-difference eq.)

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + v[\rho](\mathbf{r})\right)\psi_n(\mathbf{r}) = \varepsilon_n\psi_n(\mathbf{r})$$

Higher-order finite difference

$$\frac{\partial^2}{\partial x^2}\psi_n(x,y,z) \approx \sum_{m=-6}^{6} C_m\psi_n(x+m\Delta x, y, z)$$

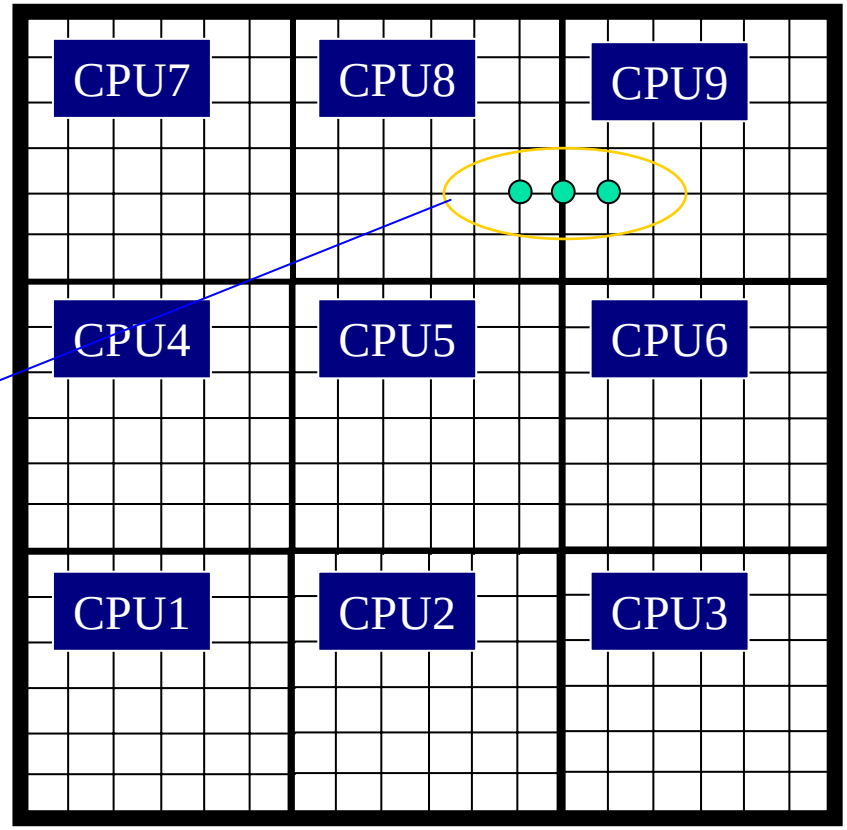
MPI_ISEND, MPI_IRECV

Integration

$$\int v(\mathbf{r})\psi(\mathbf{r})d\mathbf{r} \approx \sum_{i=1}^{Mesh} v(\mathbf{r}_i)\psi(\mathbf{r}_i)\Delta x\Delta y\Delta z$$

MPI_ALLREDUCE

3D grid is divided by some regions for parallel computation.



Flow chart

Number of electrons N
Number of grid points M ($\propto$ N)

Conjugate-Gradient Method
(or Residual-Minimization Method)

↓

Gram-Schmidt orthonormalization

↓

Density, Potentials update

↓

Convergence check $\left\| v_{new} - v_{old} \right\| < \varepsilon$?

↓

Subspace Diagonalization

Computational Cost
~O(N³)

Conjugate-Gradient Method

$$\frac{\langle \psi_n | H_{KS} | \psi_n \rangle}{\langle \psi_n | \psi_n \rangle} \rightarrow \text{minimize}$$

Residual-Minimization Method

$$\left\| H_{KS}\psi_n - \varepsilon_n \psi_n \right\| \rightarrow \text{minimize}$$

Gram-Schmidt

$$\psi'_n = \psi_n - \sum_{m=1}^{n-1} \psi_m \langle \psi_m | \psi_n \rangle \quad \cdots \mathbf{O(MN^2)}$$

Subspace Diagonalization

$$H_{m,n} = \langle \psi_m | H_{KS} | \psi_n \rangle \quad \mathbf{O(MN^2)}$$

$$\begin{pmatrix} & \mathbf{H}_{N\times N} & \end{pmatrix} \begin{pmatrix} \\ \vec{c}_n \\ \\ \end{pmatrix} = \varepsilon_n \begin{pmatrix} \\ \vec{c}_n \\ \\ \end{pmatrix} \quad \mathbf{O(N^3)}$$

$$\psi'_n(\mathbf{r}) = \sum_{m=1}^{N} c_{n,m}\psi_m(\mathbf{r}) \quad \mathbf{O(MN^2)}$$

Collaboration with Computer and Computational Scientist

- Optimization of inter-node communications
- Utilize SCALAPACK(divide-and-conquer method) in subspace diag. routine
- Active use of Level 3 BLAS in O(N³) computation (BLAS=Basic Linear Algebra Subprograms)

e.g.) Gram-Schmidt Orthogonalization • • • O(N³)

$$
\begin{aligned}
\psi_1' &= \psi_1 \\
\psi_2' &= \psi_2 - \psi_1' \langle \psi_1' | \psi_2 \rangle \\
\psi_3' &= \psi_3 - \psi_1' \langle \psi_1' | \psi_3 \rangle - \psi_2' \langle \psi_2' | \psi_3 \rangle \\
\psi_4' &= \psi_4 - \psi_1' \langle \psi_1' | \psi_4 \rangle - \psi_2' \langle \psi_2' | \psi_4 \rangle - \psi_3' \langle \psi_3' | \psi_4 \rangle \\
\psi_5' &= \psi_5 - \psi_1' \langle \psi_1' | \psi_5 \rangle - \psi_2' \langle \psi_2' | \psi_5 \rangle - \psi_3' \langle \psi_3' | \psi_5 \rangle - \psi_4' \langle \psi_4' | \psi_5 \rangle \\
\psi_6' &= \psi_6 - \psi_1' \langle \psi_1' | \psi_6 \rangle - \psi_2' \langle \psi_2' | \psi_6 \rangle - \psi_3' \langle \psi_3' | \psi_6 \rangle - \psi_4' \langle \psi_4' | \psi_6 \rangle - \psi_5' \langle \psi_5' | \psi_6 \rangle
\end{aligned}
$$

Part of the calculations can be performed as Matrix × Matrix operation!

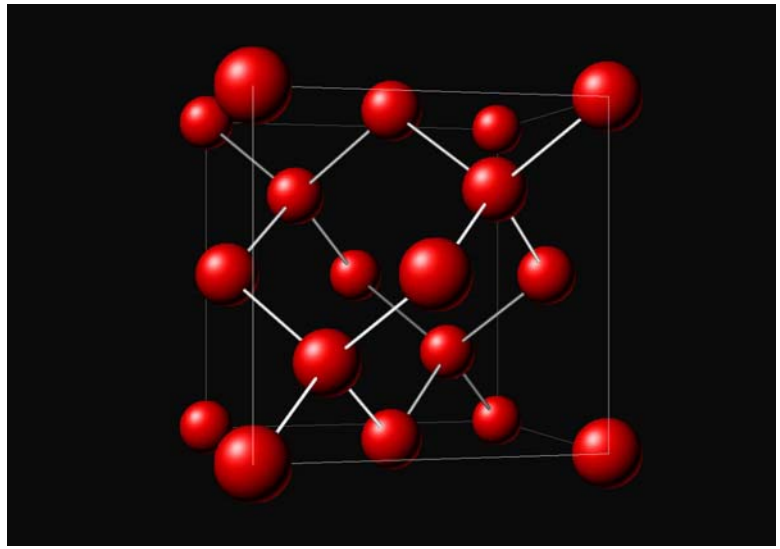
Performance of Gram-Schmidt routine

Theoretical Peak	Operation	Operation & Communication
5.6 GFLOPS/cpu	4.3 GFLOPS/cpu	3.5 GFLOPS/cpu

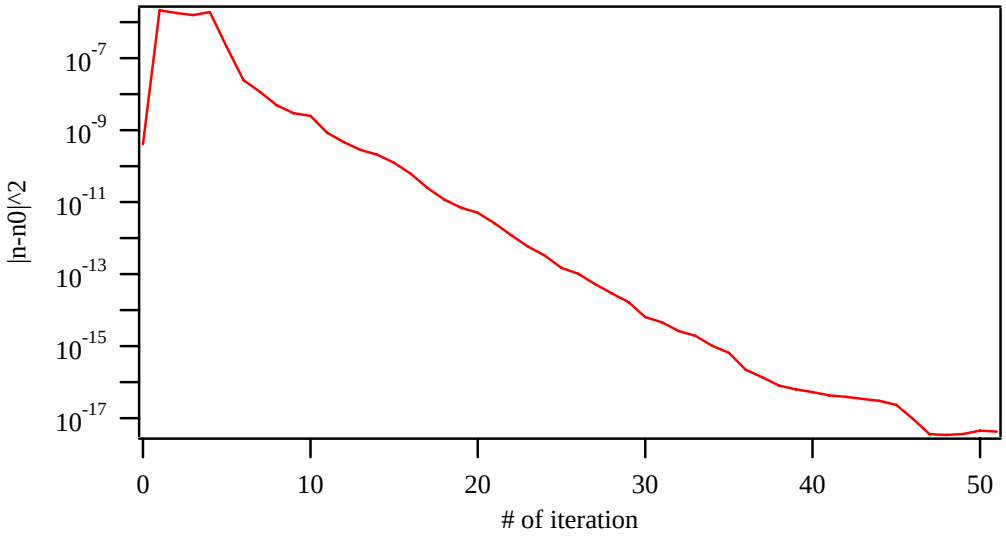
O(N³) part can be computed at 80% of the theoretical peak performance!

Si4096 atom

of atoms : 4096
lattice : 96^3=884736
# of bands : 8196
# of CPUs : 256



密度の収束性



1-SCF

	Time (sec)	MFLOPS
Part. Diagonalization O(N^3)	1046	～4000
Gram-Schmidt Orthogonalkization O(N^3)	178.8	～2600
Conjugate-Gradient O(N^2)	100～350	～70
TOTAL	1400～1661	900～1700

Very high efficiency even on ordinary scalar processor (Intel Xeon)

RS-DFT summary

- RS-DFT is highly scalable application without FFT including high-cost communication
- Very coarse-grained computation and nearest neighboring (6 or 26 neighbors on 3-D) communication is essential
  → applicable to ultra-large scale MPP
- For N atoms, O(N^3) computation and O(N^2) memory capacity is required
- On NGS, both vector and scalar units must utilize spatial locality to save memory bandwidth, and our orthogonalization algorithm based on Level-3 BLAS takes important role
- Currently, 10,000 Si computation is undergoing with 1024 – 2048 CPUs on PACS-CS
- We will perform 10,000 x n Si simulation on T2K platform, toward the final goal of 100,000 Si simulation on NGS

Code development

- Current code is well-tuned for CPU utilization
  - QCD: SSE3 intrinsic level coding to achieve 30% of peak performance
  - RS-DFT: Level3-BLAS is available for orthogonalization with more than 50% of peak performance
- Codes (not just our 2, but all 21 apps.) are just written by simple MPI without hybrid manner (shared memory model)
- Each node of NGS will be multi-core, and the compiler capability is strongly required for efficient CPU utilization on multi-core
- We need to develop hybrid code (for example, OpenMP + MPI), or to find other way to exploit the best performance on multi-core SMP and large scale interconnection network
- T2K platform will be a good development environment
- We are also researching a new paradigm of hybrid programming based on OpenMP-style coding (OpenMPD)

T2K Open Supercomputer Alliance

T2K Open Supercomputer Alliance

Who Allied ?

Kyoto U.
Academic Center for Computing & Media Researches

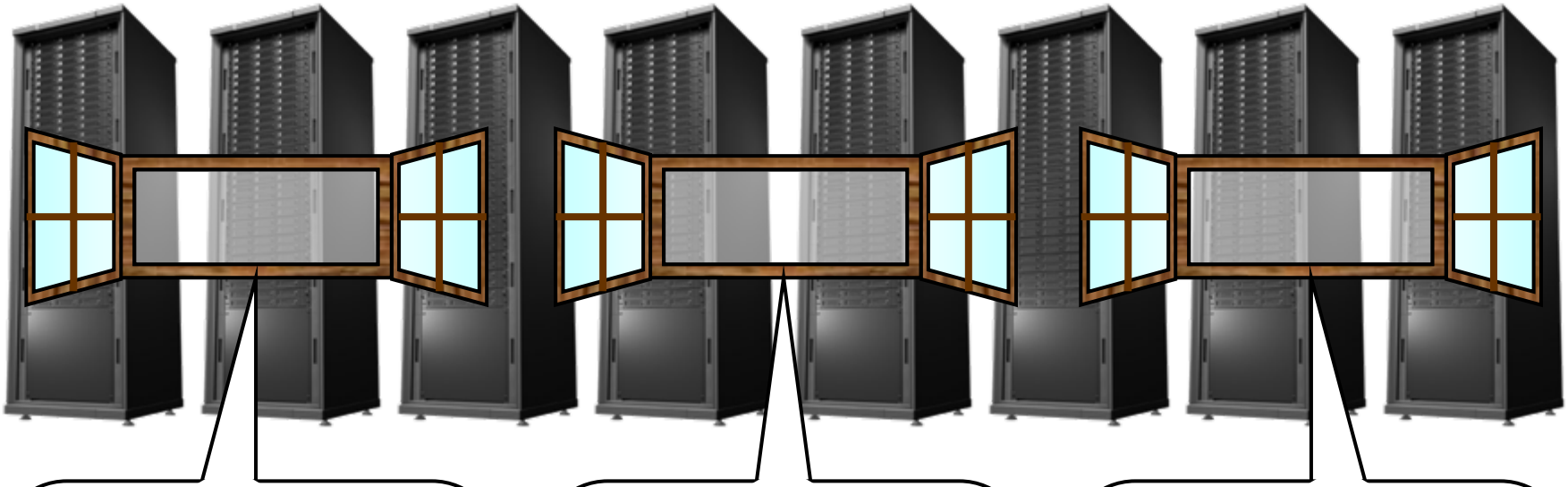
U. Tsukuba
Center for Computational Science

U. Tokyo
Information Technology Center

Why Allied ?

- Change procurement style from
  - passive & vender initiative
  - choice from product market
- to new scheme with
  - active & university initiative
  - creation from technology market
- aiming at high-performance solution
  - with most advanced technologies
  - for wide-spectrum of university users

How Opened ?



Open Hardware Arch

- commodity device base e.g. x86, IB/Myri-10G
- ➔ most efficient/advanced components from current IT market
- no special/dedicated components for HPC

Open Software Stack

- open source & standard OS/middleware e.g. Linux, MPI, Globus
- ➔ bases for also open source HPC middleware, libraries and tools

Open to User's Needs

- in addition to FP users in computational science
- attract INT users who are using PC clusters for searching/mining, natural lang. processing, genomic informatics, ...

What Specified ?

- Common Requirements
  - Hardware
    - shared memory node of 16+ x86 cores and 32+GB ECC memory with 40+GB/sec (aggr.)
    - bundle (even #) of inter-node links of 5+GB/sec (aggr.)
    - on-node 250+GB RAID-1 disk (optional) and IPMI2.0
  - Software
    - Red Hat or SuSE Linux
    - Fortran, C and C++ with OpenMP and auto-parallelizer
    - Java with JIT compiler
    - MPI of 4+GB/sec and 8.5-μsec RT latency
    - BLAS, LAPACK and ScaLAPACK
  - Benchmarks (not required performance numbers)
    - SPEC CPU2006, SPEC OMP2001, HPC Challenge (part)
    - our own for memory, MPI and storage performance

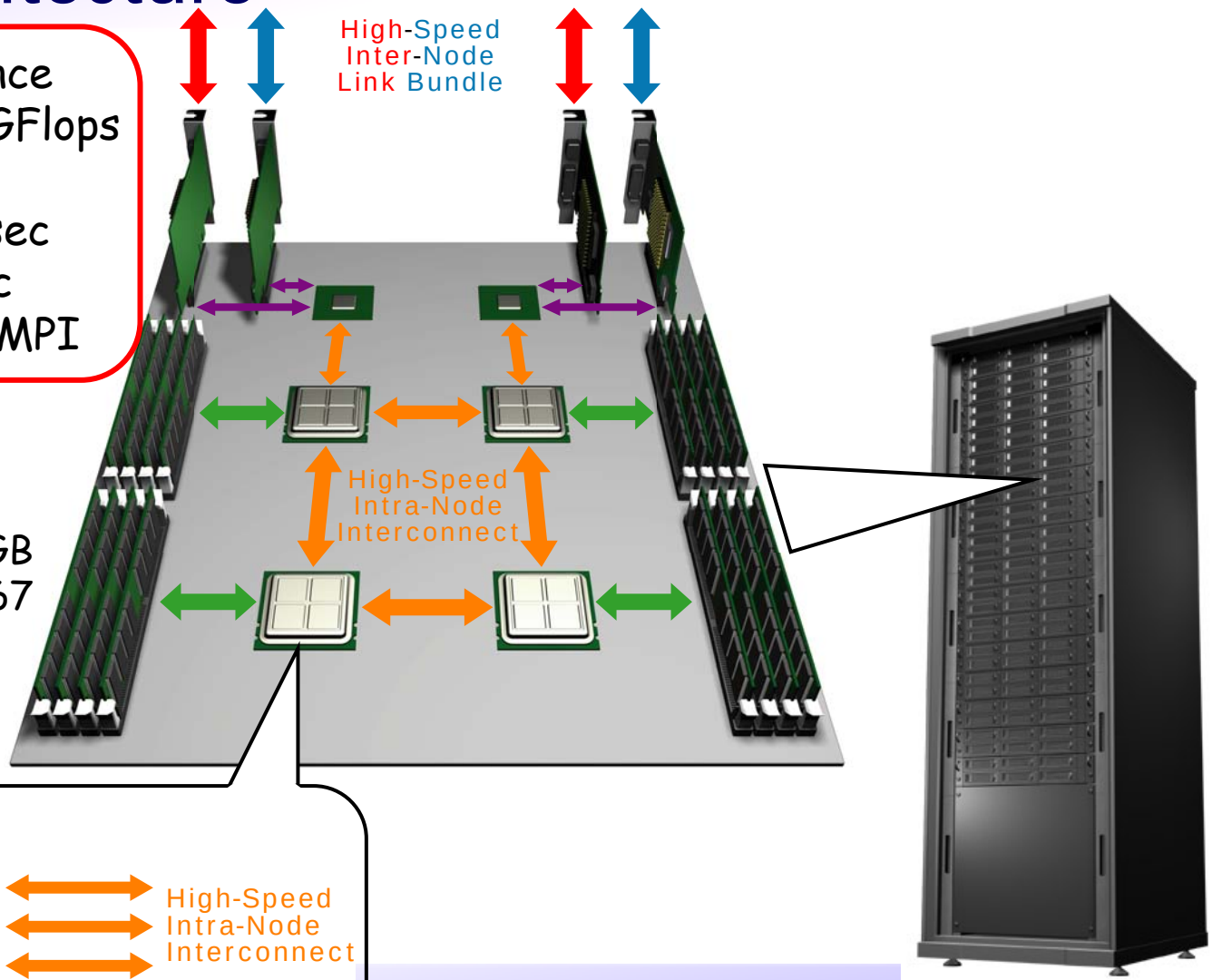
What Not Specified ?

- Site Matters
  - Hardware
    - core/socket/node performance
    - # of nodes and total performance
    - inter-node connection and its performance
    - storage system capacity and its performance
    - power consumption
  - Software
    - batch job scheduler, fault tolerance, system management
    - commercial library and applications
  - Benchmarks
    - performance numbers of common benchmarks
    - site-specific application benchmarks

T2K Open Supercomputer

Node Architecture

Per-Node Peformance
- Rpeak = 147-160 GFlops
- Memory = 32GB
  - 40GB/sec
- Links = 5-8 GB/sec
  - 4GB/sec@MPI

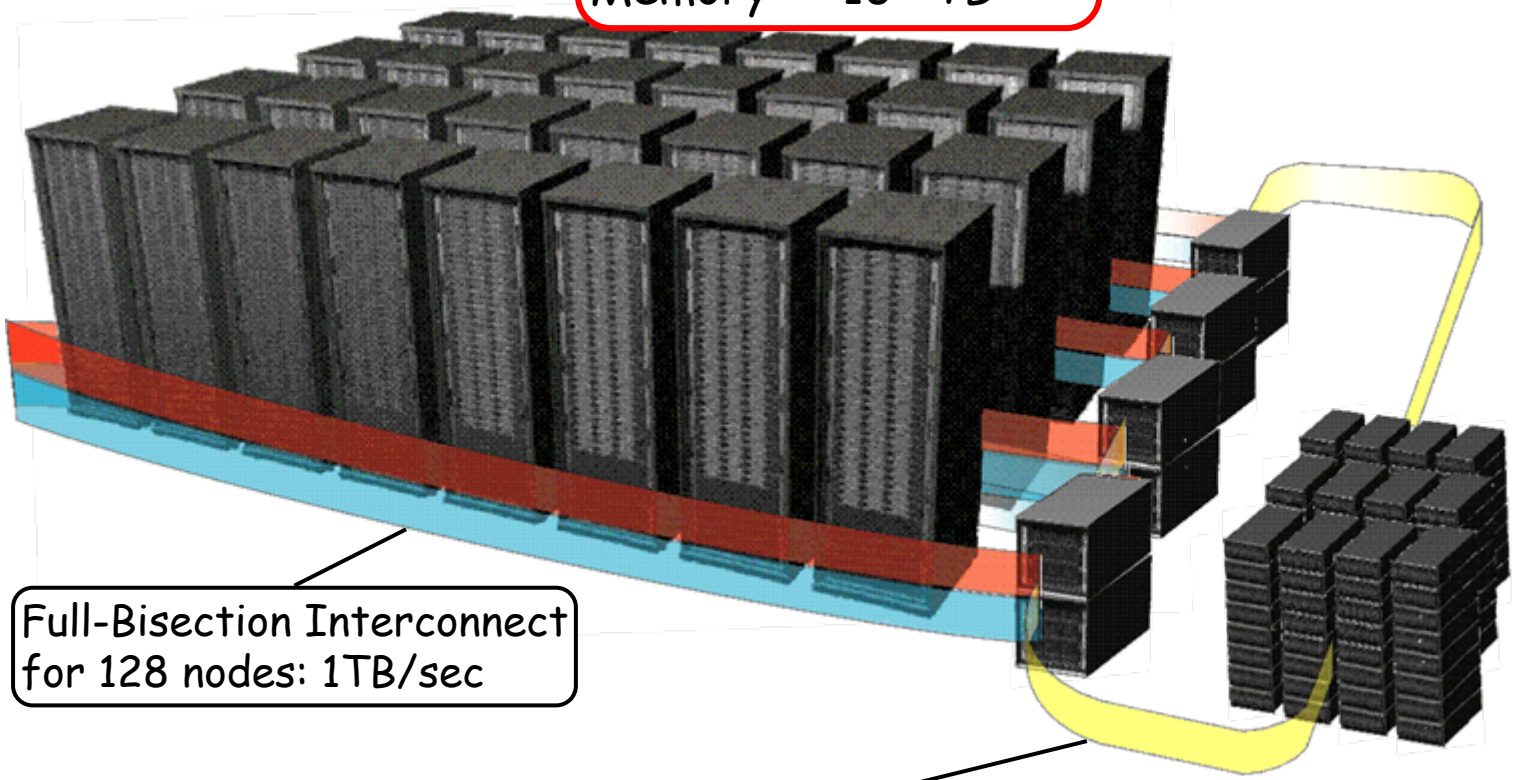


T2K Open Supercomputer

Configuration of U. Tsukuba

95 Tflops ?

# nodes = 512+
Rpeak = 75+ TFlops
Memory = 16+ TB



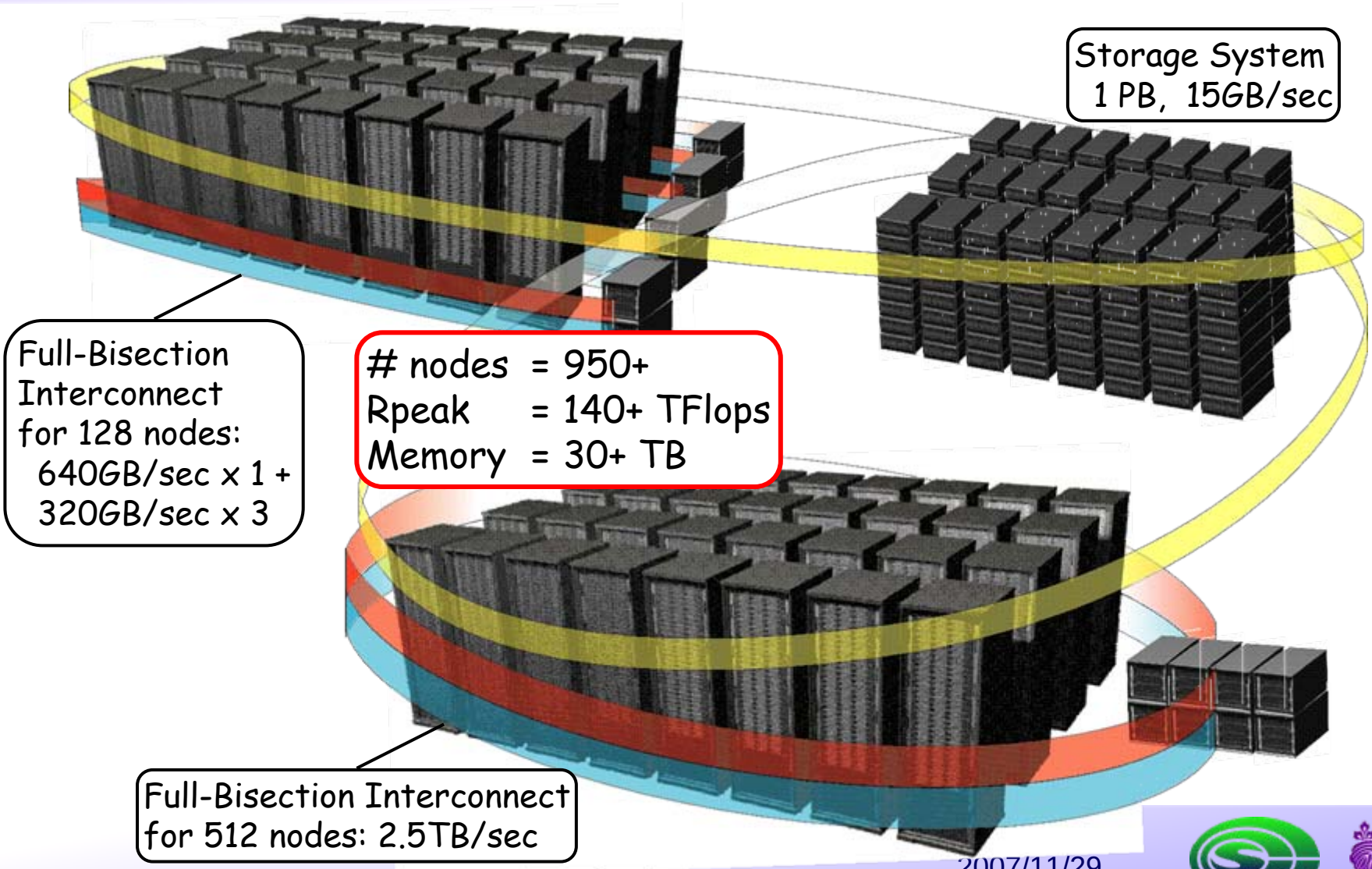
Full-Bisection Interconnect
for 128 nodes: 1TB/sec

1/8-1/1 x Full-Bisection
Interconnect
0.5-4TB/sec

Storage System
400 TB, 10GB/sec

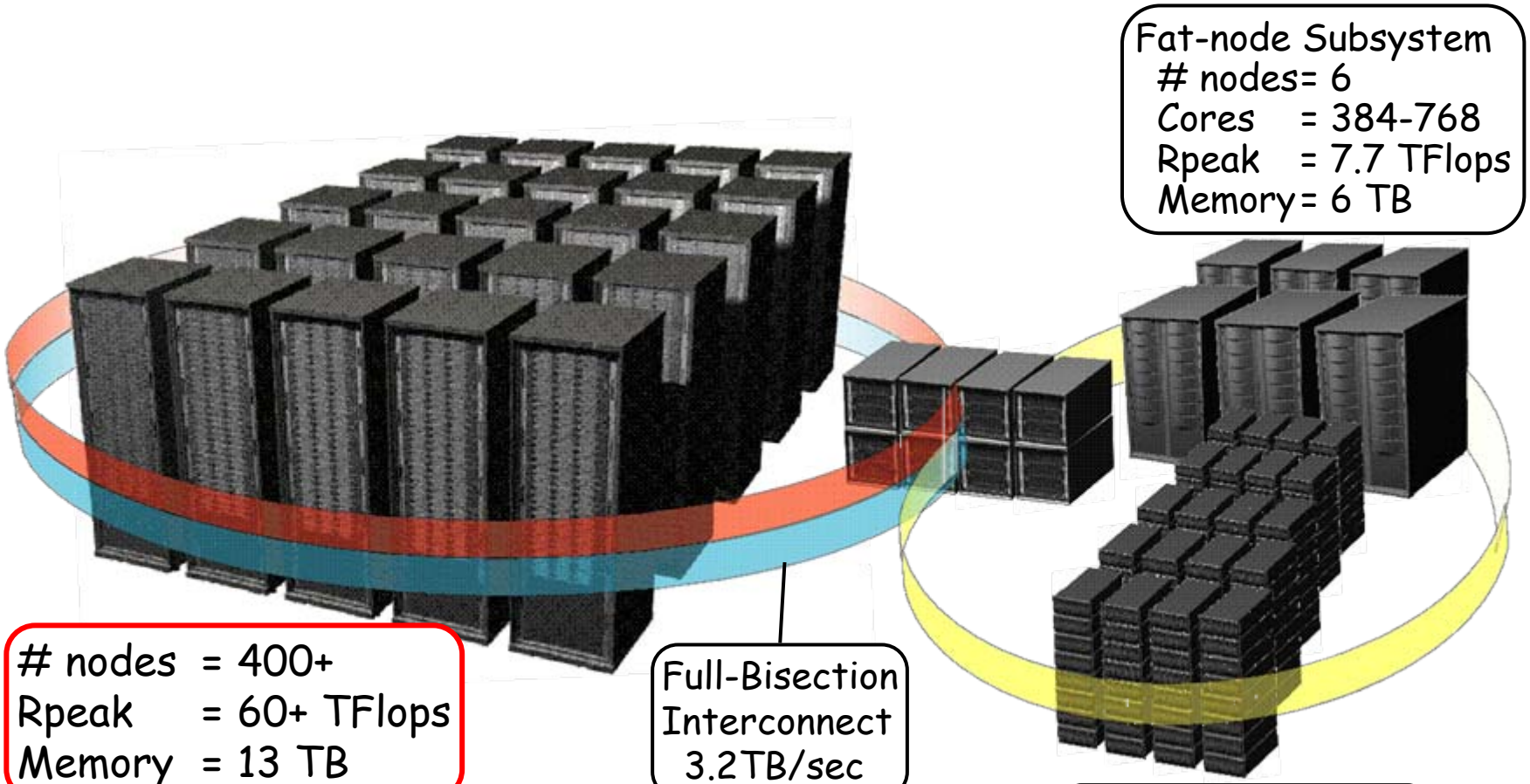
T2K Open Supercomputer

Configuration of U. Tokyo



T2K Open Supercomputer

Configuration of Kyoto U.



T2K and NGS (at U. Tsukuba)

- In CCS, U. Tsukuba will use T2K system to develop and evaluate our two NGS applications (LatticeQCD, RS-DFT)
- For development of Pflops-class apps., we need a computational facility to support 10 Tflops/job easily
- For performance estimation of Pflops-class apps., we need ~50 Tflops/job
- Ordinary cluster system is not sufficient because of insufficient memory and network bandwidth
  → T2K system has high memory bandwidth and strong interconnection capability compared with CPU performance
- We are ready to open T2K facilities for other NGS application groups to support large-scale scientific codes

Summary

- Japanese Next Generation Supercomputer system development has just started its detailed design, and fixing the system design will take more time
- RIKEN is responsible for system development, and 21 important target applications are under development (or tuning) on various institutes and universities
- Scalability of many applications are not sufficient, but NGS will be used both for capability computing and capacity computing
- CCS, U. Tsukuba is developing two important apps., and also involved to system design, performance tuning and evaluation under agreement with RIKEN
- T2K Open Supercomputer Alliance and its machines will contribute to the development of NGS applications

Acknowledgment

- Dr. Watanabe, Project leader of NGS at RIKEN
- Prof. Ukawa @ U. Tsukuba (QCD)
- Prof. Oshiyama & Dr. Iwata @ U. Tokyo (RS-DFT)
- Prof. Nakashima@Kyoto U.
  Prof. Ishikawa@Tokyo U.
  Prof. Sato@U. Tsukuba
  (all for T2K)