The Earth Simulator and its Beyond
— Technological Considerations towards —
Sustained Peta Flops Machine

Oct. 2, 2004

Tadashi Watanabe
(e-mail:t-watanabe@db.jp.nec.com)
Simulating “Earth” on Supercomputer

Supercomputer Simulation:
- can visualize
- can virtually experiment
- can forecast the future

However, current supercomputers are not enough for further analysis of problems on Planet Earth.

(North American 24-hour Precipitation)

NEC SX-6/8A

Power x 640

The Earth Simulator

40TFLOPS 1Q2002

Each CPUs executes their share of computation

Project of...
Development Organization and Schedule

- **JAMSTEC**
  - The Earth Simulator
  - Research and Development
- **NASDA**
- **JAERI**
  - Main Contractor for Design and Manufacturing
- **NEC**

(Courtesy of JAMSTEC/Earth Simulator Center)
System and Hardware

The Earth Simulator Center

NEC
Earth Simulator System

<table>
<thead>
<tr>
<th>System Peak Performance</th>
<th>40TFLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total No.of Arithmetic Processors(APs)</td>
<td>5,120</td>
</tr>
<tr>
<td>Peak Performance/AP</td>
<td>8GFLOPS</td>
</tr>
<tr>
<td>Total No.of Processor Nodes(PNs)</td>
<td>640</td>
</tr>
<tr>
<td>(8APs/Node:64GFLOPS/Node)</td>
<td></td>
</tr>
<tr>
<td>Total Main Memory Capacity</td>
<td>10TBytes</td>
</tr>
<tr>
<td>Disk Storage</td>
<td>940TBytes</td>
</tr>
<tr>
<td>Mass Storage</td>
<td>1.5PBytes</td>
</tr>
</tbody>
</table>
Central Subsystem

Interconnection Network

640 x 640 crossbar switch

Main memory System 16GB

Arithmetic Processor #0
Arithmetic Processor #1
Arithmetic Processor #7

Processor Node #0

... — 12.3 GB/sec * 2

Processor Node #1

Main memory System 16GB

Arithmetic Processor #0
Arithmetic Processor #1
Arithmetic Processor #7

Processor Node #639

Main memory System 16GB

Arithmetic Processor #0
Arithmetic Processor #1
Arithmetic Processor #7

(Courtesy of JAMSTEC/Earth Simulator Center)
Processor Node

Fig 3  Processor Node Configuration

Main memory System (MS)  (Capacity : 16GB)

From / To Interconnection Network
LAN

MMU#0
MMU#1
MMU#2
MMU#3
MMU#4
MMU#5
MMU#6

MMU#7

Bandwidth : 256GB/s

AP #0
AP #1
AP #2
AP #7

RCU IOP

User/Work Disks

128Mbit DRAM developed for ES (Bank Cycle Time : 24nsec)
2048-way banking scheme

(Courtesy of JAMSTEC/Earth Simulator Center)
Arithmetic Processor (AP)

Vector Unit: 8 sets
- 6 different types of vector pipelines
- 72 vector registers (256 vector elements)
- 17 mask registers (256 bits)

Scalar Unit
- 4-way super scalar
- 64KB instruction cache
- 64KB data cache
- 128 general purpose register

(Courtesy of JAMSTEC/Earth Simulator Center)
Connection between Cabinets

128 XSW’s
64 Cabinets

PN-IN Electric Cables: 640 x 130 = 83,200

640 PN’s
320 Cabinets
Data Paths in Interconnection Network (IN)

(Courtesy of JAMSTEC/Earth Simulator Center)
Earth Simulator Building

- Processor Node Cabinets (PN Cabinets)
- Mass Storage System
- System Disks
- User Disks
- Interconnection Network Cabinets (IN Cabinets)
- Air Conditioning System
- Power Supply System
- Double Floor (Floor height 1.5m)

(Courtesy of JAMSTEC/Earth Simulator Center)
Inter-node Communication Cables

(Courtesy of JAMSTEC/Earth Simulator Center)
Cross-Sectional View of the Earth Simulator Building

(Courtesy of JAMSTEC/Earth Simulator Center)
One Chip Vector Processor (AP)

- 0.15 µm CMOS
- 8 layers copper interconnection
- 20.79mm × 20.79mm
- 60 million Tr
- 5185 pins
- Clock Frequency: 500 MHz (1 GHz)
- Power Consumption: 140 W (typ.)

(Courtesy of JAMSTEC/Earth Simulator Center)
AP Package

(Courtesy of JAMSTEC/Earth Simulator Center)
Operation System Overview

✓ Operation and management system for huge distributed memory system

- 5120 APs
- 640 PNs
- 1 system

Each PN equal to the large super computer

64GFLOPS
Operating System Overview

SUPER-UX
Operating System and Language for SX series
  - Vector processing
  - Parallel processing for Shared memory
  - Parallel processing for Distributed memory
  - Batch system (NQS)
  - High performance I/O
  - Cluster management

ES Operating System
  Extend scalability (up to 640 nodes)
  - Processors performance
  - I/O performance
  - Specification limits

Add the function for the Earth Simulator
  - Efficient execution environment for highly parallel job
  - Single system image (SSI)
    - Operation management
    - Batch job environment for highly parallel program
Operating System Overview

Characteristics of the ES Operating System

Efficient execution environment for highly parallel programs

✓ High speed inter-node communication function utilizing IN
✓ Global address space between PNs using IN
✓ HPF compiler, MPI library

Single System Image (SSI)

for system administrator :

✓ **Super Cluster System** for system operation management
  ➢ Two level cluster control (16nodes/cluster, 40cluster/system)
  ➢ Resource management function of whole system (Node / IN / disk / tape)

for end users :

✓ Batch job environment for highly parallel job (NQSII,MDPS)
✓ Automatic file migration
Multi-node parallel program execution environment

- OS provides the global address space between PNs (memory protection proof)
- MPI library transfers data directly using IN data transfer instructions, without systemcall
Operation management

S Cluster

Interconnection Network (IN)

L Clusters

OS Kernel

Distributed parallel program execution

Cluster control function

CCS: Cluster Control Station
SCCS: Super Cluster Control Station

Operation/management System

SCCS
Execution of large scale job

Large distributed parallel jobs

Many user programs are executed by dividing the system

Ultra parallel simulation program developed with the HPF and MPI library
Node Allocation

S-nodes (14 nodes)  L-nodes (624 nodes)

SCCS
Customized Scheduler

NQS II

Login Server

Requests

Login

Users

NEC
Job Execution Flow

Node schedule

- Running
- Allocated

Node allocation

Request

L-Batch Queue

Exit

Stage-OUT

Running

Waiting

Stage-IN

Running

S-Disk

HOME Disk

DATA Disk

L-system Work Disks

M-Disk

Staging

IN

OUT

IN

OUT

IN

OUT
MPI (Message Passing Interface)

- Standard specification of message passing library for parallel processing
- Common API specification (platform-independent)
- Library procedure interface which can be called from C, C++, Fortran programs
- May, 1995 MPI-1.1 specification release
- July, 1997 MPI-1.2 and MPI-2 specification release
- ES supports full MPI (MPI-2) specification
MPI data transfer

MPI library selects appropriate communication procedure

- Intra-node: memory copy using vector load and vector store instructions
- Inter-node: data transfers directly using IN data transfer instructions

![Diagram showing MPI process, shared memory, intra-node communication, inter-node communication, memory copy, and data transfer using MPI library.]
HPF (High Performance Fortran)

- Extension of Fortran language for distributed-memory parallel computer system
- Defacto standard
- Easy to write, high portability (Fortran + directives)

```
DIMENSION A(10000,10000)
!HPF$ DISTRIBUTE A(*,BLOCK)

!HPF$ INDEPENDENT
DO J = 1, 10000
  DO I = 1, 10000
    A(I,J) = 0.0
  END DO
END DO
```

Specify only data mapping on distributed memory

Declaration that next loop is a parallel executable loop
HPF (High Performance Fortran)

The 3 Phases of parallel program development:
(a) Data partitioning/allocation to the parallel processor
(b) Computation divide/scheduling to the parallel processor
(c) Insert the communication code
HPF automates (b), (c) phases

<table>
<thead>
<tr>
<th></th>
<th>MPI</th>
<th>HPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Data mapping/allocation</td>
<td>manual</td>
<td>manual</td>
</tr>
<tr>
<td>(b) Computation divide/scheduling</td>
<td>manual</td>
<td>automatic</td>
</tr>
<tr>
<td>(c) Insert the communication process</td>
<td>manual</td>
<td>automatic</td>
</tr>
</tbody>
</table>

The case of typical isotopic simulation:

<table>
<thead>
<tr>
<th></th>
<th>MPI</th>
<th>HPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallelization</td>
<td>Modify whole program</td>
<td>Add directives (about 5%)</td>
</tr>
<tr>
<td>Performance</td>
<td>100%</td>
<td>About 70-80%</td>
</tr>
</tbody>
</table>
Performance
Basic Performance Data

Peak Performance

- System Performance: 40TFLOPS
- Per Node (8APs): 64GFLOPS
- Per Processor: 8GFLOPS

Bandwidth

- Memory to Processor: 32GB/sec
- Per Node (8 SMP): 256GB/sec
- Inter-node Per node: 12.3GB/sec * 2

LINPACK (HPC)

- Sustained Performance: 35.86TFLOPS (87.5% efficiency)

MPI Start-up cost

<table>
<thead>
<tr>
<th>Function</th>
<th>internode</th>
<th>intranode</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPI_Get</td>
<td>6.68 µs</td>
<td>1.27 µs</td>
</tr>
<tr>
<td>MPI_Put</td>
<td>6.36</td>
<td>1.35</td>
</tr>
</tbody>
</table>
Internode Communication Bandwidth

Fig. 28 Internode Communication Bandwidth

(Courtesy of JAMSTEC/Earth Simulator Center)
Barrier Synchronization

![Diagram of Barrier Synchronization](image)

(Courtesy of JAMSTEC/Earth Simulator Center)
Application Performance

- Global Atmospheric Simulation : 26.58 TFLOPS (66.5%)
- Direct Numerical Simulation of Turbulence : 16.4 TFLOPS (41.0%)
- Three-dimensional Fluid Simulation : 14.9 TFLOPS (38.3%)

for Fusion Science with HPF

(Courtesy of JAMSTEC/Earth Simulator Center)
Application Results
Precipitation(312km,T42L24)  Precipitation(10.4km,T1279L24)
Future Technological Challenges for Peta Flops Computing
History of High Performance Computers

Single CPU Performance

CPU Frequencies

Aggregate Systems Performance

Increasing Parallelism

CPU Frequencies

Single CPU Performance

Increasing Parallelism

History of High Performance Computers

Increasing Parallelism
Faster the Speed, More the Parallel

The Largest configuration in SX-3

The Earth Simulator

1990
22GFlops/4Cpu

10Years

2002
40TFlops/5120Cpu
Evolution of SX Series for 20 years

<table>
<thead>
<tr>
<th>Year</th>
<th>Magnification</th>
</tr>
</thead>
<tbody>
<tr>
<td>'83</td>
<td>6</td>
</tr>
<tr>
<td>'03</td>
<td>$3 \times 10^4$</td>
</tr>
</tbody>
</table>

- **CPU Performance**
  - '83: 1.3 GFLOPS
  - '03: 8 GFLOPS
  - 40 TFLOPS (Earth Simulator)

- **System Performance**
  - '83: 1.3 GFLOPS
  - '03: 40 TFLOPS (Earth Simulator)
  - 5120 (Earth Simulator)

- **# of CPUs**
  - '83: 1
  - '03: 5120

- **Total Memory Capacity**
  - '83: 256 MBytes
  - '03: 10 Tera Bytes

- **CPU Size**
  - '83: 180 cm x 2 cm
  - '03: 2 cm

- **# of chips per cpu**
  - '83: 2,250 chips
  - '03: 1 Chip

- **Memory Size**
  - '83: 200 cm

- **System Size**
  - '83: 80 m x 50 m
  - '03: 49 Cabinets
  - 64 GFLOPS/8 CPU

- **Other**
  - '03: 1 Chip
  - 1/2,250 x 1/4,000

---

**Note:** The diagrams show the physical sizes and arrangements of the CPUs and memory units for the '83 and '03 models, illustrating their respective scales and configurations.
Will this technological evolution continue?

Are there any problems or difficulties to overcome?

If so, what are they?
Do We Need a Peta Flops Computer?
Application Areas and Required Performance

Performance Required

Nano Technologies

Environment

Bio

Automotives And Aerospace

Energy

New Material Design

Screening for Drug Design

Composite Simulation (Combustion, Airplane)

Molecular Orbit In Protein

Local Disaster Prediction (1PF/500TB)

Climate Prediction (Long Term) (500TF/1PB)

Climate Prediction (Short Term) (20TF/10TB)

Earth Simulator (200TF/100TB)

Plasma Analysis (500TF/1PB)

Structural Analysis for Composite Material (200TF/400TB)

Composite Simulation

Molecular Orbit

Bio

Automotives And Aerospace

Energy

Nano Technologies

Memory

Application Areas

10TFLOPS

100TFLOPS

1PFLOPS

10PFLOPS

10PB
Capacity Computing and Capability Computing

**Capacity Computing**
- Goals: Workload and Throughput Many Jobs per time
- Many Small Problems
- Parallel or Cluster Machine based on Microprocessor

**PC Cluster / Blade Server**

**Capability Computing**
- Goals: High Speed Execution of Single Job
- Large and Critical Problem – Grand Challenge
- Powerful Processor and Highbandwidth Network

**Vector / SX**
System Configuration and User View
Highly Efficient Capability Computing

- Low Latency
- High Throughput

High Bandwidth Interface

Powerful CPU

NEC
Capability Computing ~To Increase Sustained Performance~

**Technological Issues**

- **Amdahl’s Law**
  Importance of Single CPU’s Performance

(Ex.) To achieve 20% efficiency of Peak Speed (1 TF)
- 10 GFLOPS/CPU x100 CPUs ➞ 96% of Parallelism
- 1 GFLOPS/CPU x1000 CPUs ➞ 99.6% of Parallelism

- **To Increase CPU Performance**
  - Device (LSI) Technology
  - Memory Performance (Bandwidth)

- **To Increase Performance of Parallel Processing**
  - High Scalability and High Efficiency by High Speed CPU
  - Small Scale Parallel Processing: High Bandwidth SMP
  - Large Scale Parallel Processing: High Performance Communication (MPI)
    High Speed Synchronization Mechanism

NEC
Road Map of Semiconductors

Roadmap for Semiconductors (ITRS2003)
Device Technology

Road Map of LSI CMOS Process

Technological Issues
- Higher Density
- Ultra-finer Elements

Problems to Overcome
- Lower Power → High- k
- High Speed → Low- k
- Lower Voltage → Finer Process

International Technology Roadmap for Semiconductors (ITRS) 2003
Power Reduction of LSI

Increase of Power Consumption

Power

Increase of Power Consumption

- 90nm
- 65nm
- 45nm

Leak power
Operating Power

Power= \[ C \cdot Vdd^2 \cdot f + (I_{\text{off leak}} + I_{\text{gate leak}}) \cdot Vdd \]

Increase of operating Power due to Speed Increase and Finer Process

- Performance
- Pattern Pitch

\[ \rightarrow \text{Frequency} \ (f) \ \rightarrow \text{Capacitance} \ (C) \]

Increase of Leak Current due to Finer Process

- Gate Length
- Gate Insulation

\[ \rightarrow I_{\text{off leak}} \ \rightarrow I_{\text{gate leak}} \]

Counter Plan

- Vt dynamic control (Stand by Vt)
- Low material Capacitance by Low material(C)

Counter Plan

- Gate length
- Operating current
Cooling Technology

- Micro Channel
- Micro Channel Liquid Cooling
- Three Dimensional Cooling
- Vapor Chamber
- Air Cooling

Component Technology:
- LSI
- Micro Channel Process
- Long Life Pump
- Anti-Erosion
- Liquid Cooled Board
- Liquid Cooled Package
- Micro Fin Heat Exchanger

Cooling Performance (W/Chip)
High Density Optical Interconnection by Multi-Layer Wave Guide

Optical Cross Interconnection
Internal Chip Configurations

• PIM(Processor in Memory)
  - Insufficient Memory for Numerical Intensive Applications
    \[ M \sim (P)^{3/4} \sim GB/GFLOPS \]
  - Commodity Product such as Media Processor, Home/Industry Equipment

• P Core + Special Engines
  - Special Engines: Graphics/Video/DSP/Image/FFT
  - Commodity Products such as Mobile Phone, Home/Industry Equipment, and Cars

• P Core + Vector Engine/Multiple P Cores
  - HPC for Scientific/Engineering Use
  - High-end/Affordable HPC
    ( P Core: VLIW/Superscalar/Multithread)
Challenges in Software

- Operation and Resource Management
- Huge Volume of Data Management
- Reliability, Availability and Serviceability
- Support of Development Environment (Compiler and Tools) for Ultra Large Scale Parallel Processing System

>10,000CPUs

>1 Peta Bytes Storage

NEC
Post Silicon
Top View of an EJ-MOSFET

lower gate  upper gate

field  n+ region

200nm  L_{LG}  2mm

8nm < L_{LG} < 100 nm
Tox=5nm

SEM image
Post Silicon & Post Switch ???

- **new structure device**
  - High mobility
  - FinFET

- **new material**
  - Low-k/Cu, High-k
  - strained Si
  - ultra-thin SOI
  - partial SOI
  - PD+BST

- **BST-SOI circuit technology**
  - High speed, Low power

- **Gate length (nm)**
  - 500, 100, 50, 10

- **Clock frequency**

- **Post Si**
  - CNT-Tr.
  - Nano-SW

- **Post SW**
  - Q-comp.
Carbon-Nanotube Field-Effect Transistors

- Possible application: low-cost, low-power LSI, rf drivers
- Position-controllable on-wafer growth (catalyst CVD)
- Extremely high transconductance:
  \[ g_m = 8.7 \, \text{mS/tube} \]
  \[ (5800 \, \mu\text{S/mm}) \]

Si nFET: 1000~1200 \( \mu\text{S/\mu m} \)
pFET: 400~600 \( \mu\text{S/\mu m} \)
Quantum Entangled State in a Solid State Device

**Sample**

Capacitive coupling

Superconductor-Based Device

Quantum Beat from Entangled Two Qubits, as predicted (published in Nature, Feb.20,’03)

**Next Step**

**Fundamental 2 Bit Logic Gate Operation (C-NOT)**

Provides Universal Gate, combined with 1 Bit Gate

<table>
<thead>
<tr>
<th>Control Bit</th>
<th>Target Bit</th>
<th>C-NOT Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>A’</td>
<td>B’</td>
<td></td>
</tr>
<tr>
<td>in</td>
<td>out</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Control Bit</th>
<th>Target Bit</th>
<th>C-NOT Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>01</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>00</td>
</tr>
</tbody>
</table>

Target bit is flipped only if control bit is 1

<table>
<thead>
<tr>
<th>Truth Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>
Post Silicon Technology

- Silicon technology miniaturization
  Possible down to 5nm
  1/30 of the Earth Simulator Technology

- Not easy to get high performance and low power
  Key technology: Parallel architecture
  Post scaling solution

- Post Silicon
  CNT Tr., Atomic switch
  Quantum computing
What will be the Future?

I Believe the Evolitional Development in these 10 years.

The More Parallism, the More Difficulties will Increase in HW Volume, Operations and Programming
Lesson 9

- The Success or failure of any new supercomputer development is finally going to rest on the ability and willingness of users to adapt to the strange world of parallel processing, and the consequent need to restructure algorithm, if not total processes.