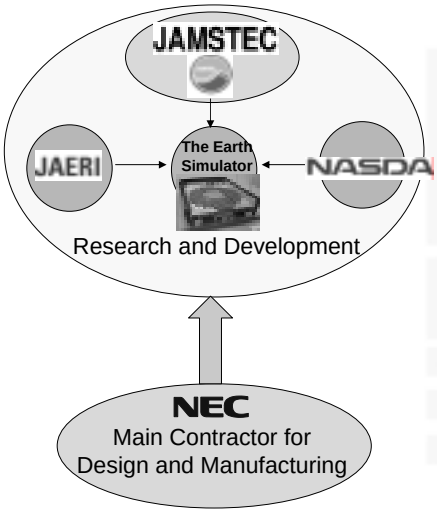
The Whole Earth Simulator

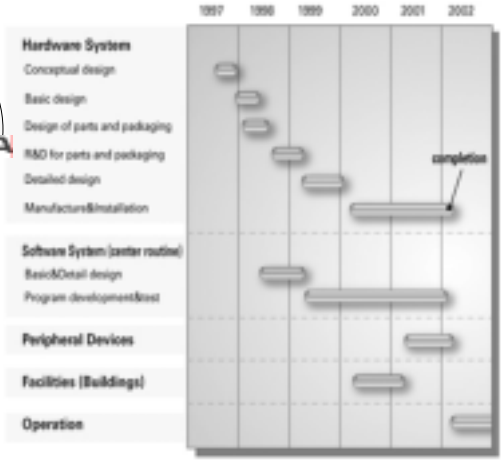
— The World's Fastest Supercomputer —

Kiyoshi Otsuka
JAMSTEC/Earth Simulator Center
(otsukak@jamstec.go.jp)

Tadashi Watanabe
NEC
(t-watanabe@db.jp.nec.com)

Development Organization and Schedule





(Courtesy of JAMSTEC/Earth Simulator Center)

Project Leader：Late Mr. Hajime Miyoshi


三好甫さんの足跡

三好甫さんを偲ぶ会

平成14年5月17日

The Earth Simulator Center

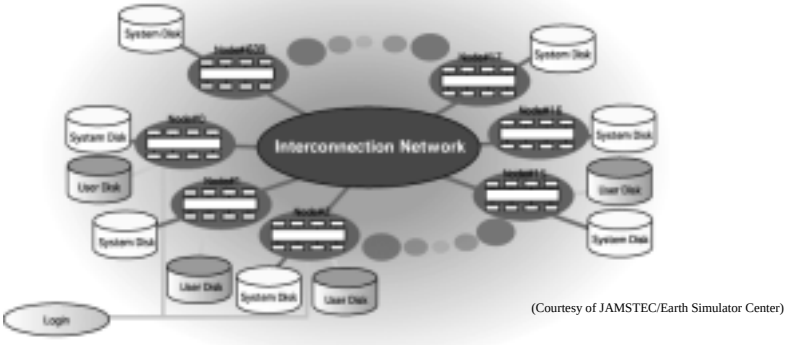
-An Organization of Japan Marine Science and Technology Center(JAMSTEC)

Basic Principles for Operating the Earth Simulator

- Quantitative Prediction and Assessment of Variations of the Atmosphere, Ocean and Solid Earth
- Production of Reliable Data to Protect Human Lives and Properties from Natural Disasters and Environmental Destruction
- Contribution to Symbiotic Relationship of Human Activities with Nature
- Promotion of Innovative and Epoch-making Simulation in any Fields such as Industry, Bioscience and Energy

System and Hardware

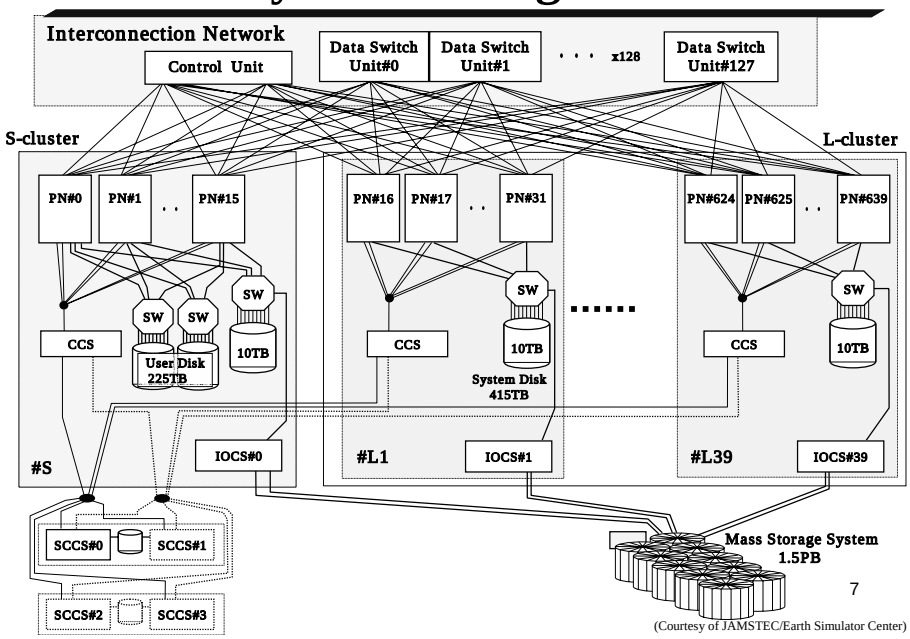
Earth Simulator System



(Courtesy of JAMSTEC/Earth Simulator Center)

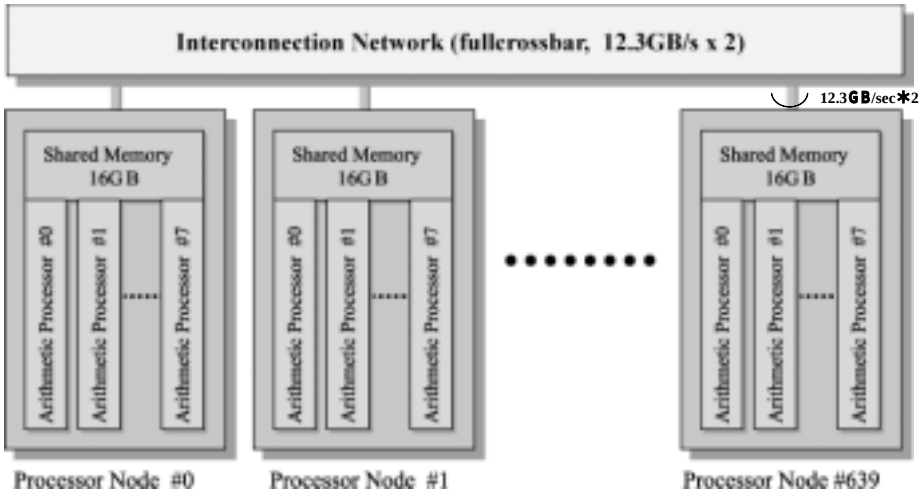
System Peak Performance	**40TFLOPS**
Total No.of Arithmetic Processors(APs)	**5,120**
Peak Performance/AP	**8GFLOPS**
Total No.of Processor Nodes(PNs)	**640**
	(8APs/Node:64GFLOPS/Node)
Total Main Memory Capacity	**10TBytes**
Disk Storage	**640TBytes**
Mass Storage	**1.5PBytes**

System Configuration



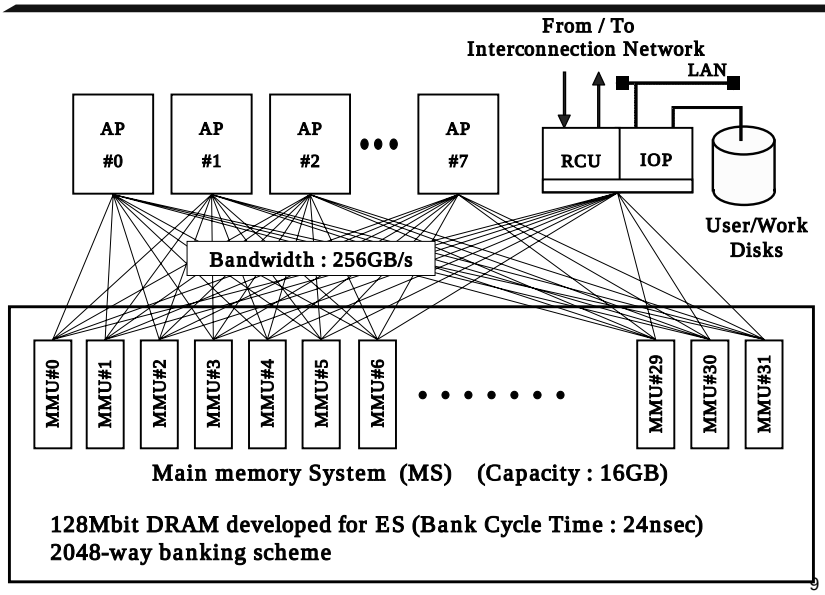
(Courtesy of JAMSTEC/Earth Simulator Center)

Central Subsystem



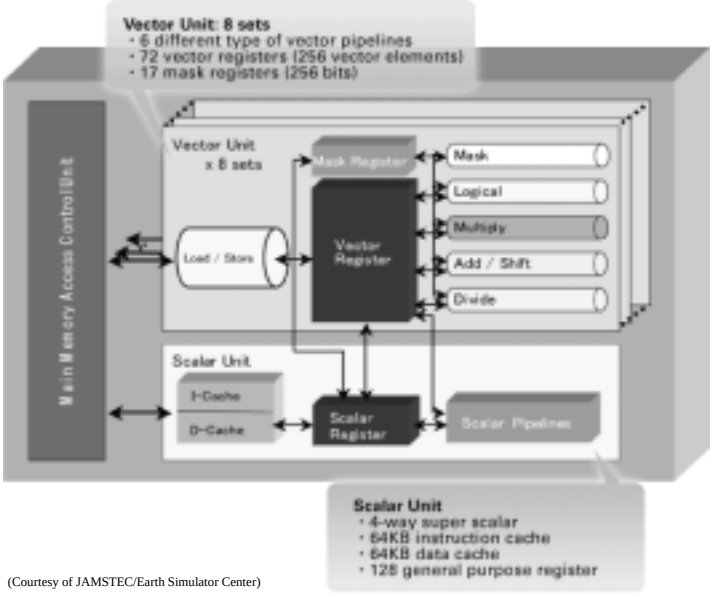
(Courtesy of JAMSTEC/Earth Simulator Center)

Processor Node



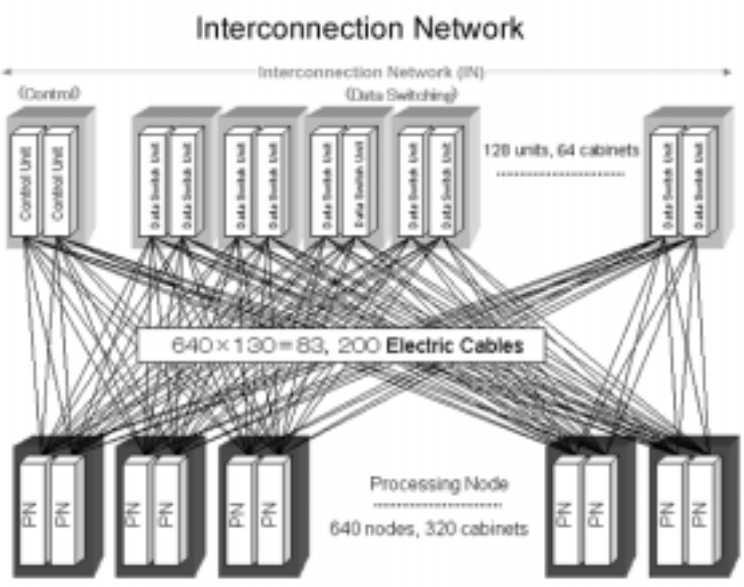
(Courtesy of JAMSTEC/Earth Simulator Center)

Arithmetic Processor (AP)



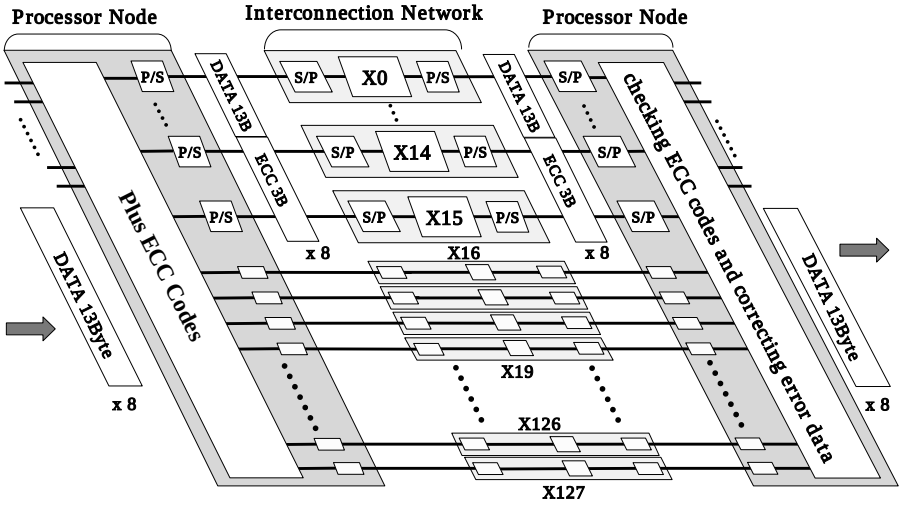
(Courtesy of JAMSTEC/Earth Simulator Center)

Interconnection Network(IN)



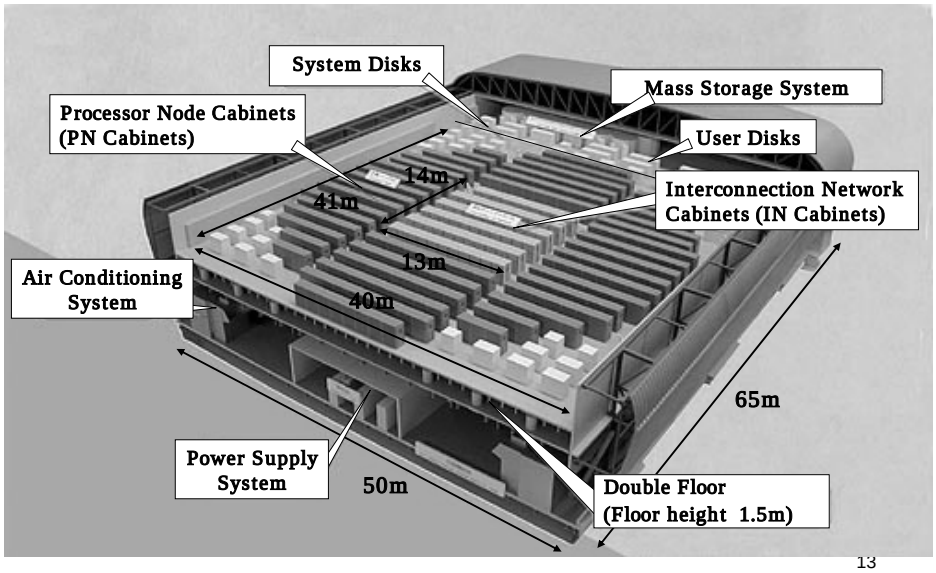
(Courtesy of JAMSTEC/Earth Simulator Center)

Data Paths in Interconnection Network(IN)



(Courtesy of JAMSTEC/Earth Simulator Center)

Earth Simulator Building

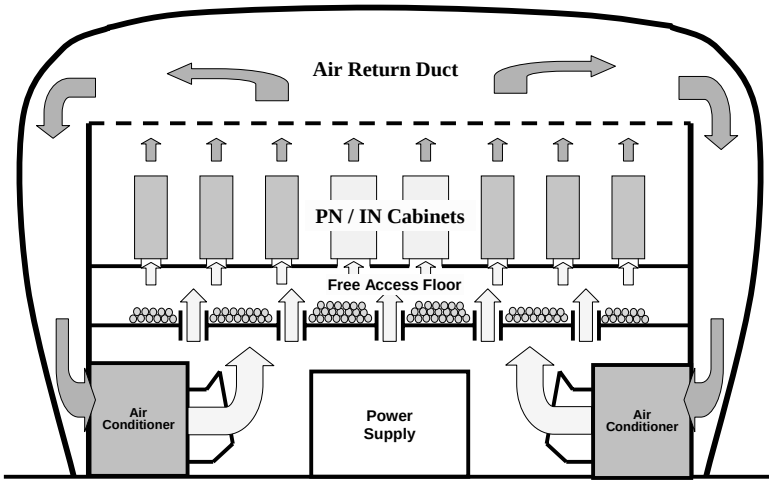


(Courtesy of JAMSTEC/Earth Simulator Center)

Inter-node Communication Cables

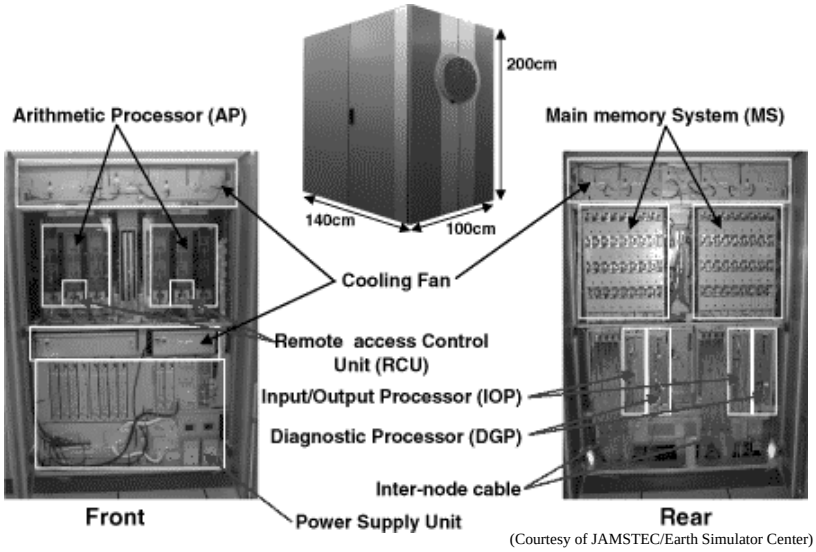
(Courtesy of JAMSTEC/Earth Simulator Center)

Cross-Sectional View of the Earth Simulator Building



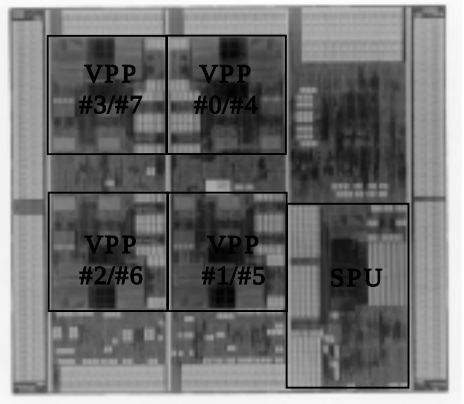
(Courtesy of JAMSTEC/Earth Simulator Center)

Processor Node(PN) Cabinet



(Courtesy of JAMSTEC/Earth Simulator Center)

One Chip Vector Processor(AP)

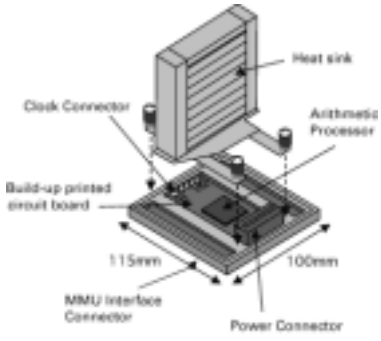


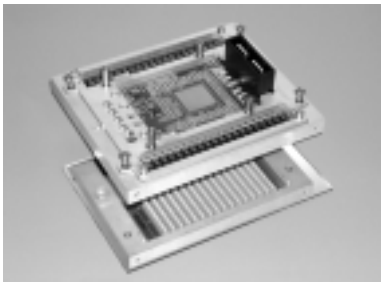
(Courtesy of JAMSTEC/Earth Simulator Center)

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

17

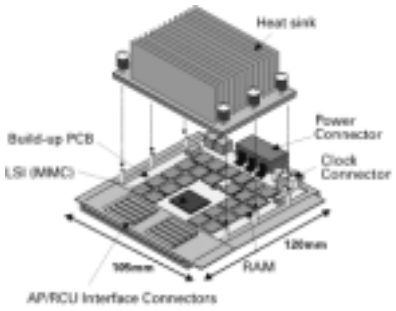
AP Package

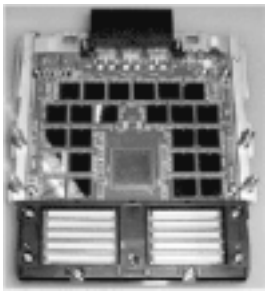




(Courtesy of JAMSTEC/Earth Simulator Center)

Memory Package





(Courtesy of JAMSTEC/Earth Simulator Center)

LSI Specifications

LSI Specifications

LSI Design	Full-custom
Design rule (μm)	0.15
Die size (mm)	20.79 x 20.79 (AP)
Number of transistors	60 million (AP)
Operating frequency (MHz)	500
Metal layer	Copper : 8
Number of I/O (Sig.)	5,185 (1,986) (AP)
I/O pitch (μm)	200
Power supply voltage (V)	1.8
Mounting	Flip-chip

Memory Device Specifications

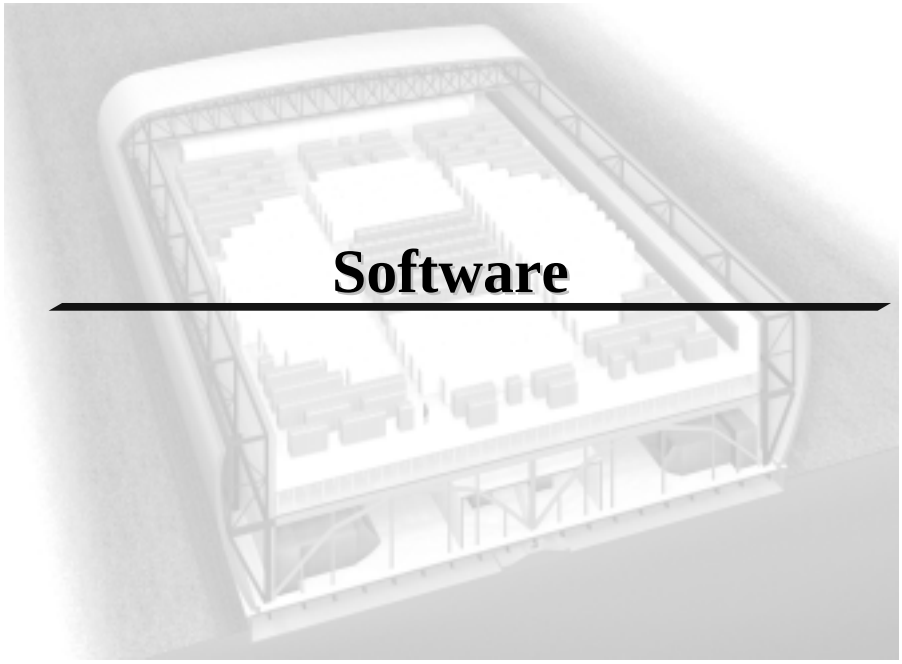
Memory Device Specifications

Capacity (Mb)	128
Number of banks	8
Clock frequency (MHz)	133
Random cycle time (ns)	21.6
Access time (ns)	30
Supply voltage (V)	2.55 (I/O 1.8)
Package	100pin μ-BGA
Power dissipation (W)	1

AP&MMU Packages

Specifications of AP & MMU packages

	AP	MMU
Substrate	Build-up printed circuit board	
Size (mm)	100×115	120×105
Thickness (mm)	1.57	
Number of layer	4 build-up layers on both sides 6 core layers	
Line/Space (μm)	25 / 25	
Via/Land (μm)	50 / 75	
Wiring length (m)	175	120
Device	CPU LSI x1 (Flip-chip)	MMC LSI x1 (Flip-chip) 128Mb-FPLRAM x48 (μ-BGA)
Number of I/O terminal (Sig.)	3,960 (1,980)	1,200 (600)
I/O terminal pitch(mm)	0.5	
Power dissipation(W)	140	60



Software

Operation System Overview

- ✓Operation and management system for huge distributed memory system

64GFLOPS

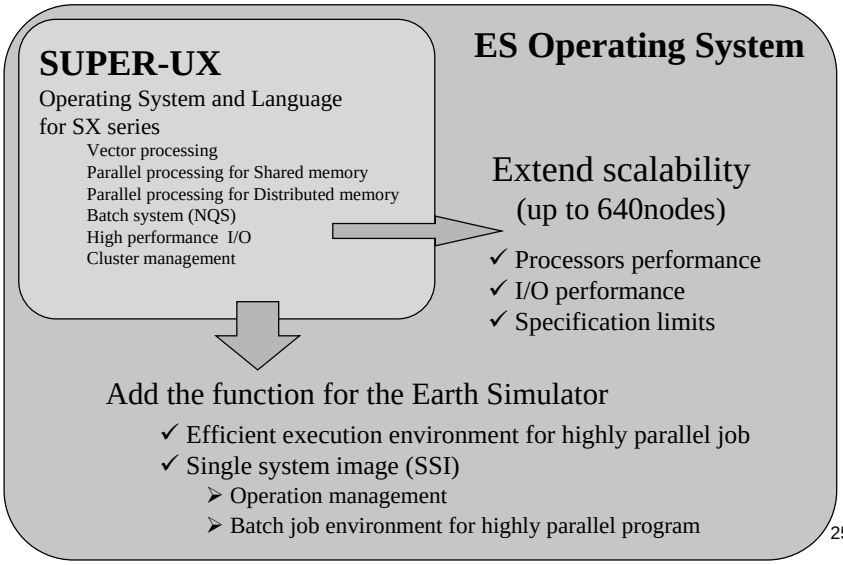
Each PN equal to the large super computer

5120 APs

640 PNs

1 system

Operating System Overview



Operating System Overview

Characteristic functions of the ES Operating

Efficient execution environment for highly parallel programs

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

Single System Image (SSI)

for system administrator :

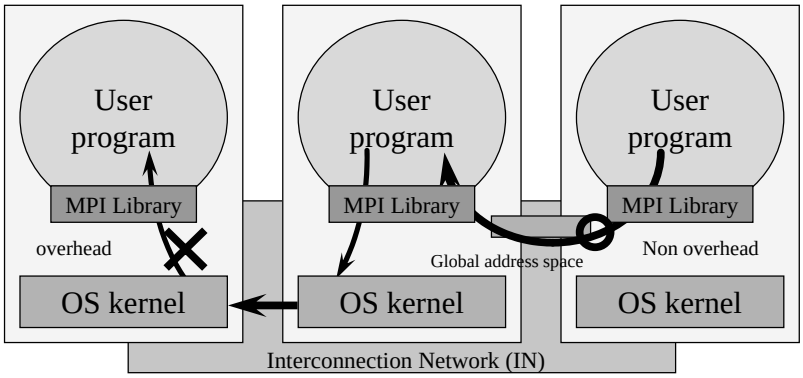
- ✓ **Super Cluster System** for system operation management
  - ➢ Two level cluster control (16nodes/cluster, 40cluster/system)
  - ➢ Resource management function of whole system (Node / IN / disk / tape)

for end users :

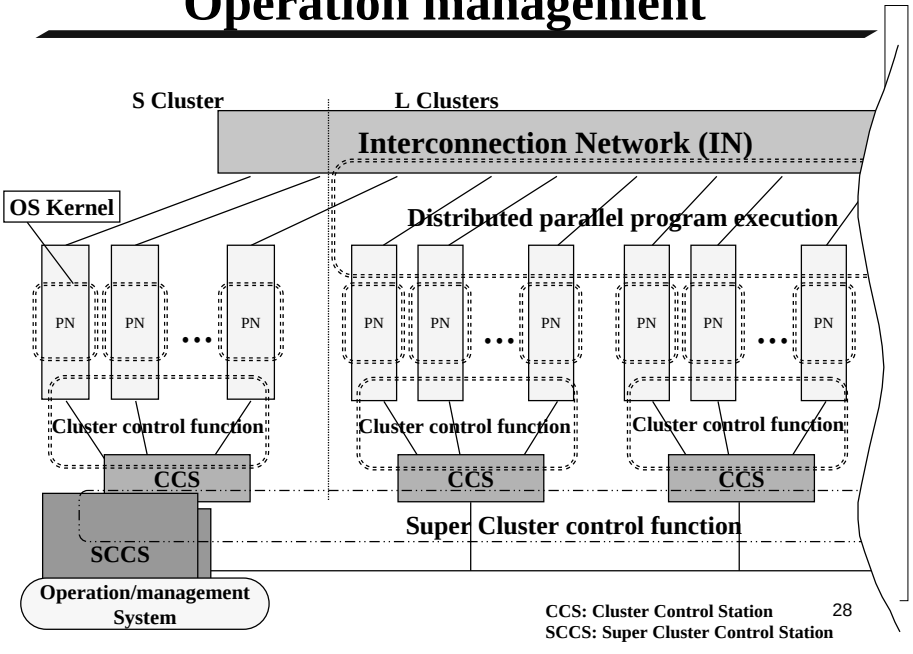
- ✓ Batch job environment for highly parallel job (Job Scheduler,NQS)
- ✓ Automated file migration and recall

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

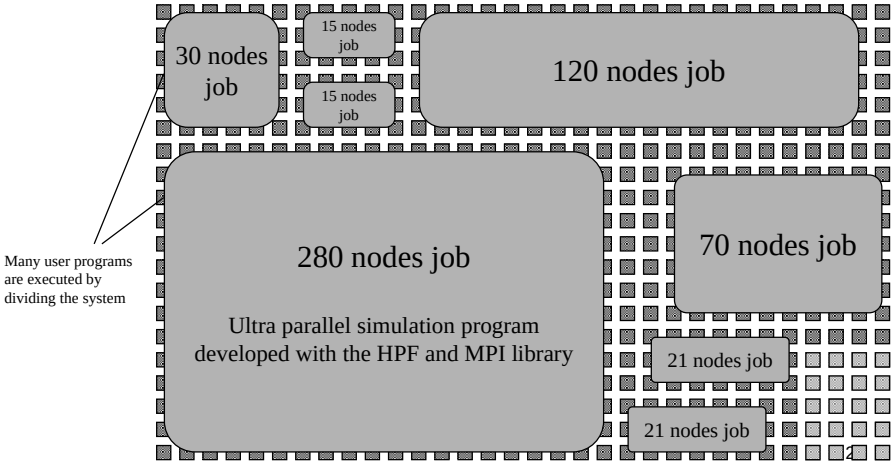


Operation management

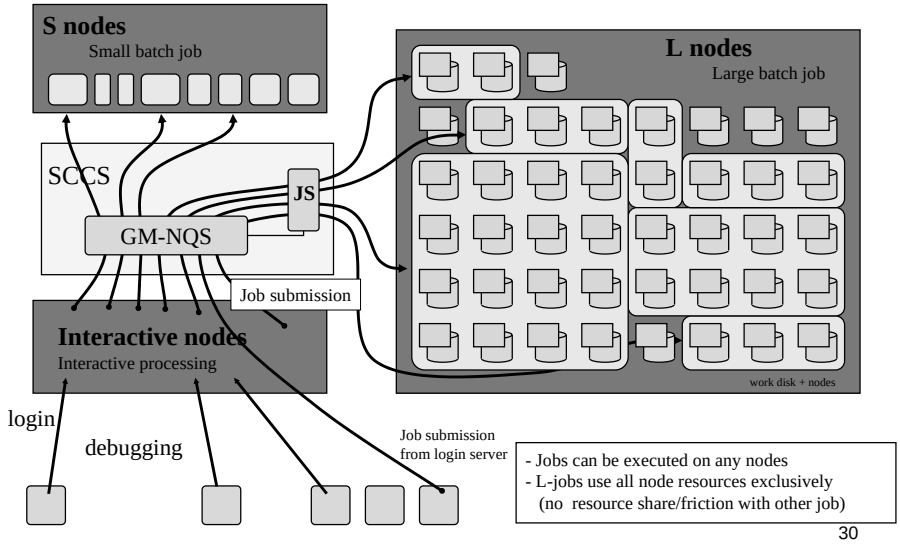


Execution of large scale job

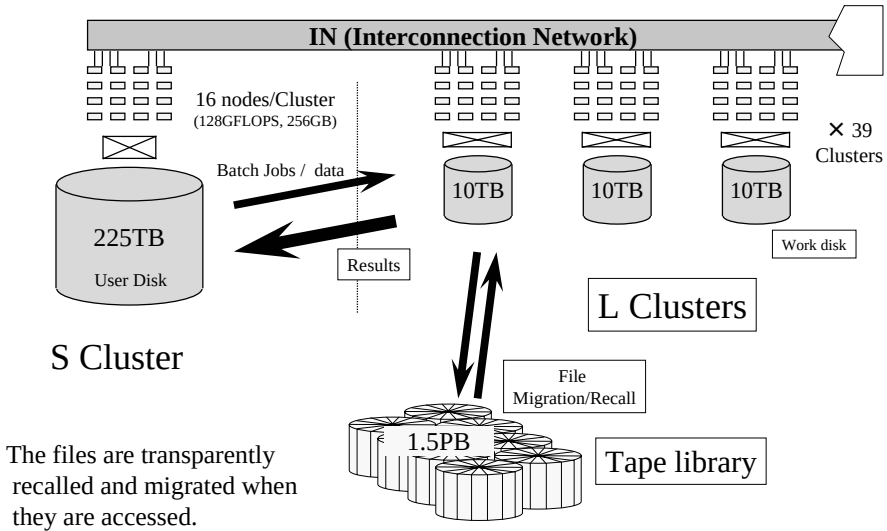
Large distributed parallel jobs



Job Execution



Automated file recall and migration



The files are transparently recalled and migrated when they are accessed.

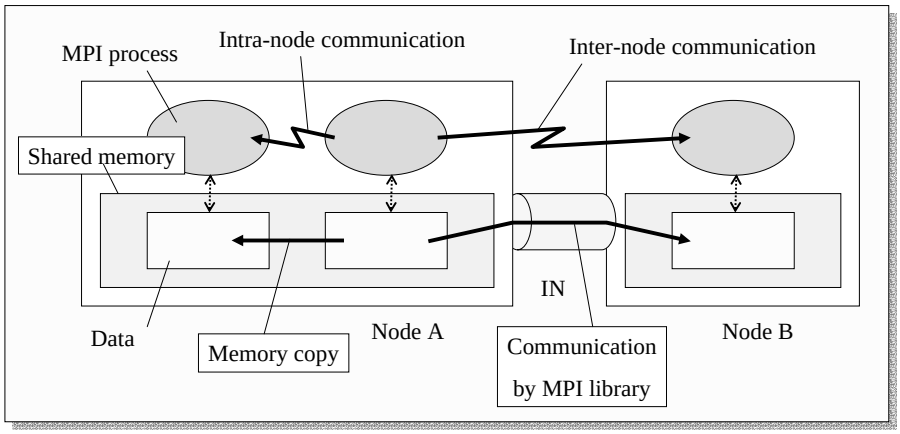
MPI (Message Passing Interface)

- ✓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



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

Array A → J

I

Declaration that next loop is a parallel executable loop

HPF (High Performance Fortran)

The 3 Phases of parallel program development:

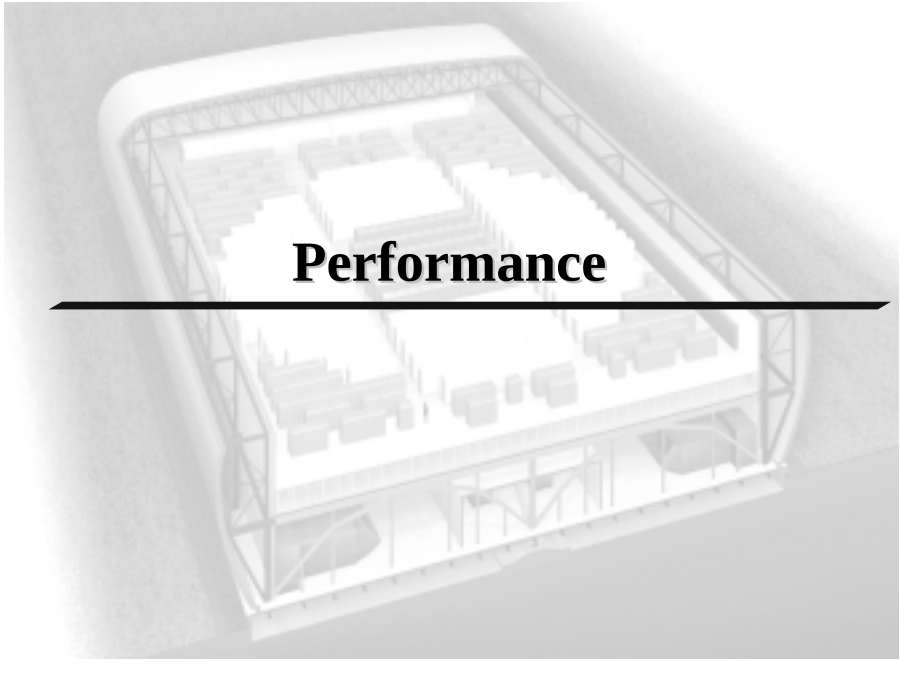
(a) Data partitioning/allocation to the parallel processor

(b) Computation divide/scheduling to the parallel processor

(c) insert the communication code

HPF automates (b), (c) phases

	MPI	**HPF**
(a) Data mapping/allocation	manual	manual
(b) Computation divide/scheduling	manual	**automatic**
(c) Insert the communication process	manual	**automatic**
The case of typical isotopic simulation :		
Parallelization	Modify whole program	Add directives (about 5%)
Performance	100%	About 70-80%



Performance

Basic Performance Data

Peak Performance

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

Bandwidth

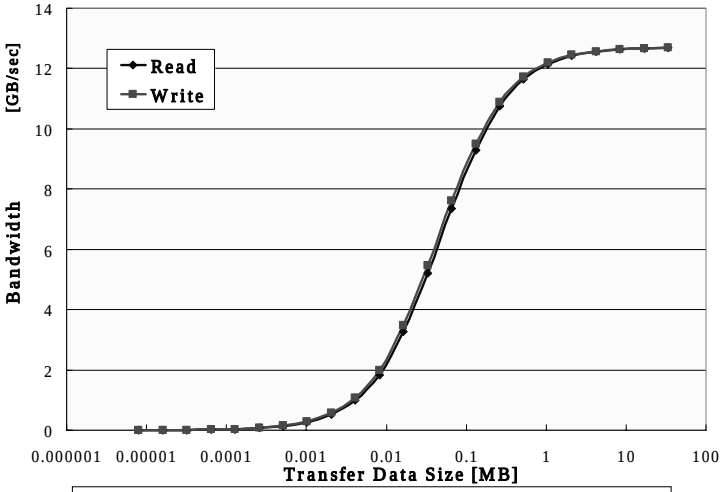
Memory to Processor	32GB/sec
Per Node(8 SMP)	256GB/sec
Inter-node Per node	12.3GB/sec∗2

LINPACK(HPC)

Sustained Performance	35.86TFLOPS(87.5%efficiency)

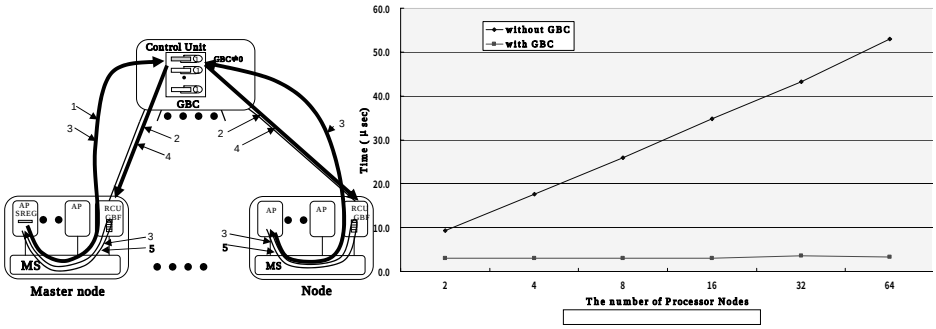
MPI Start-up cost	internode	intranode
MPI_Get	6.68μs	1.27μs
MPI_Put	6.36	1.35

Internode Communication Bandwidth



(Courtesy of JAMSTEC/Earth Simulator Center)

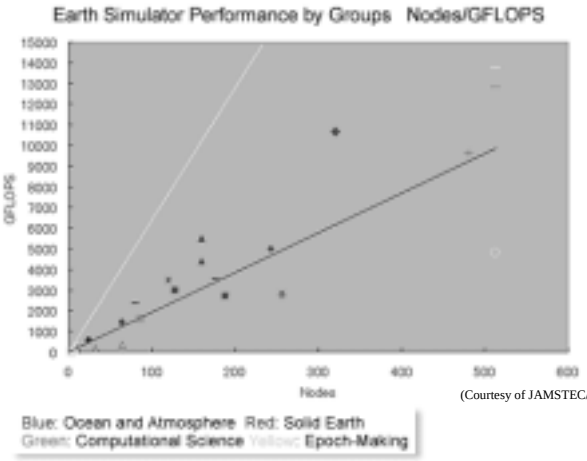
Barrier Synchronization



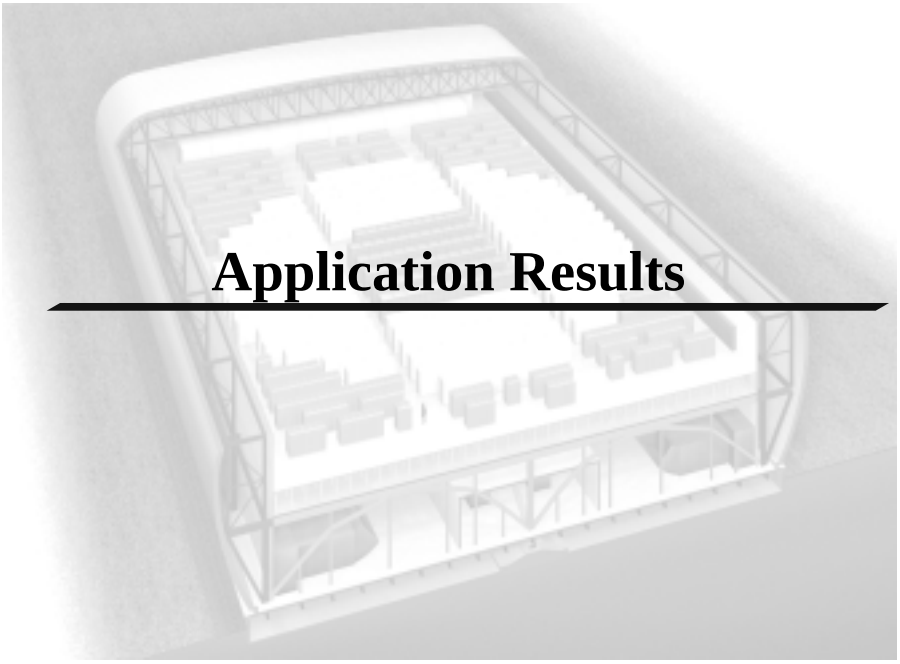
(Courtesy of JAMSTEC/Earth Simulator Center)

Application Performance

- **Global Atmospheric Simulation :26.58TFLOPS(66.5%)**
- **Direct Numerical Simulation of Turbulence :16.4TFLOPS(41.0%)**
- **Three-dimensional Fluid Simulation for Fusion Science with HPF :14.9TFLOPS(38.3%)**



(Courtesy of JAMSTEC/Earth Simulator Center)



Application Results