The Development of the Earth Simulator

Shigemune Kitawaki
Earth Simulator Center
Japan Marine Science and Technology Center
October 14, 2002

Contents

Earth Simulator Project

The Earth Simulator (ES) is an ultra high speed parallel supercomputer.

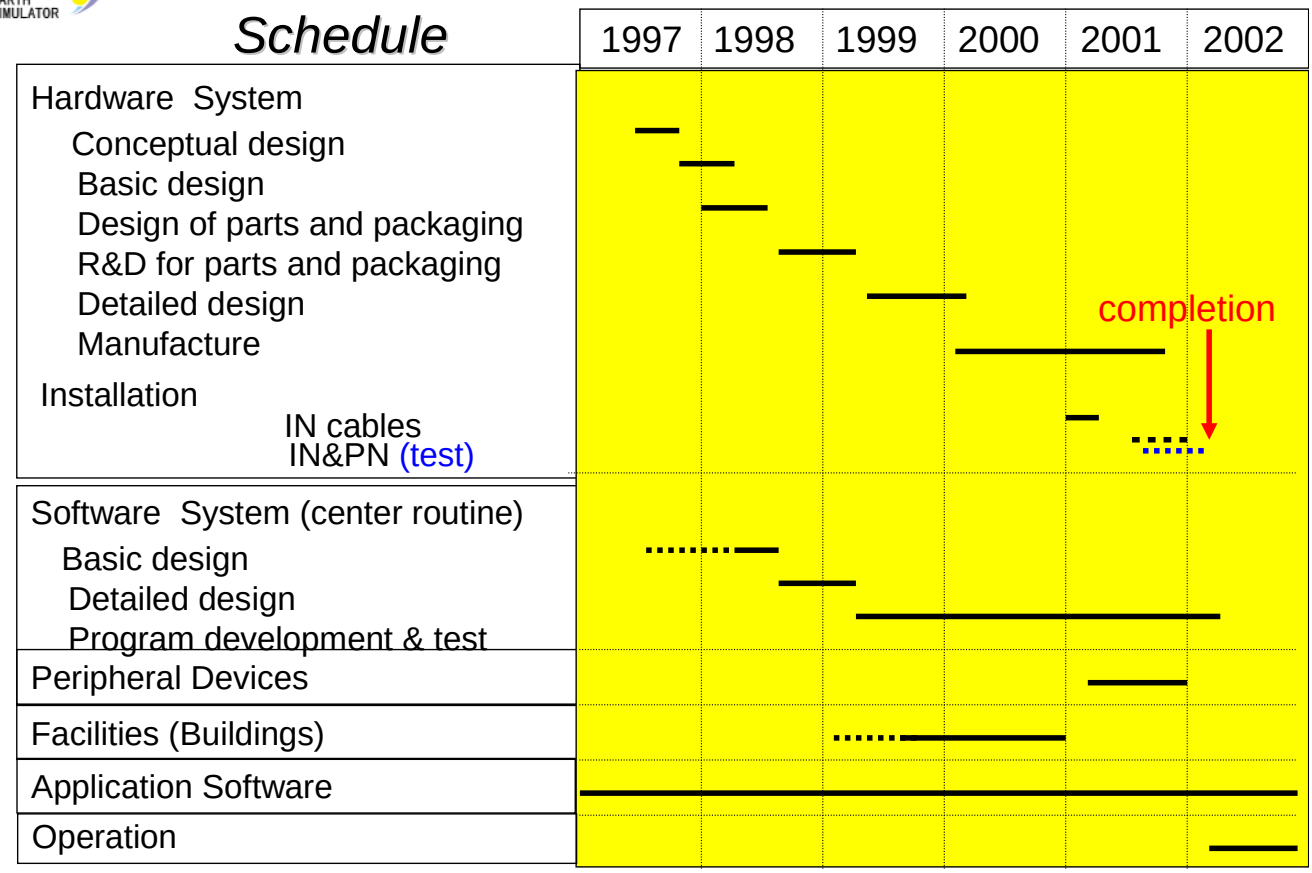
- The development of ES had started in 1997 to make an ultra high speed supercomputer for a comprehensive understanding of the global changes such as global warming, as a project of the former STA (Science and Technology Agency of Japan, now MEXT: Ministry of Education, Culture, Sports, Science and Technology) .
- It has been successfully completed achieving 40Tflops theoretical peak performance at the end of February, 2002.

Requirement & Design target

Requirements for the Earth Simulator

- Processor Type (scalar processor or vector processor)
  We selected vector type processors.
  NCAR reported CCM2 (NCAR Climate Model) shows more than 30% of peak performance on vector processor system, and less than 10% on scalar processor system.
- Total Peak Performance
  More than 32 Tflops
- Total Main memory size
  More than 8 TB
- Type of interconnection network and aggregate switching capacity
  We desired single stage crossbar network with more than 4 TB/sec of aggregate switching capacity.
  A single-stage crossbar network is superior in flexibility of allocating processor nodes to application programs and also in flexibility of executing many paradigm of applications.
- Performance of Atmospheric General Circulation Model (AGCM)
  More than 5 sustained Tflops (At least 1000 times faster than those of CRAY C90)
  We estimated the performance of AGCM of 6144x3074x255 mesh (T2047L255) as 16% of total peak performance, and also the main memory requirement as 8 TB.

Schedule



Earth simulator specifications

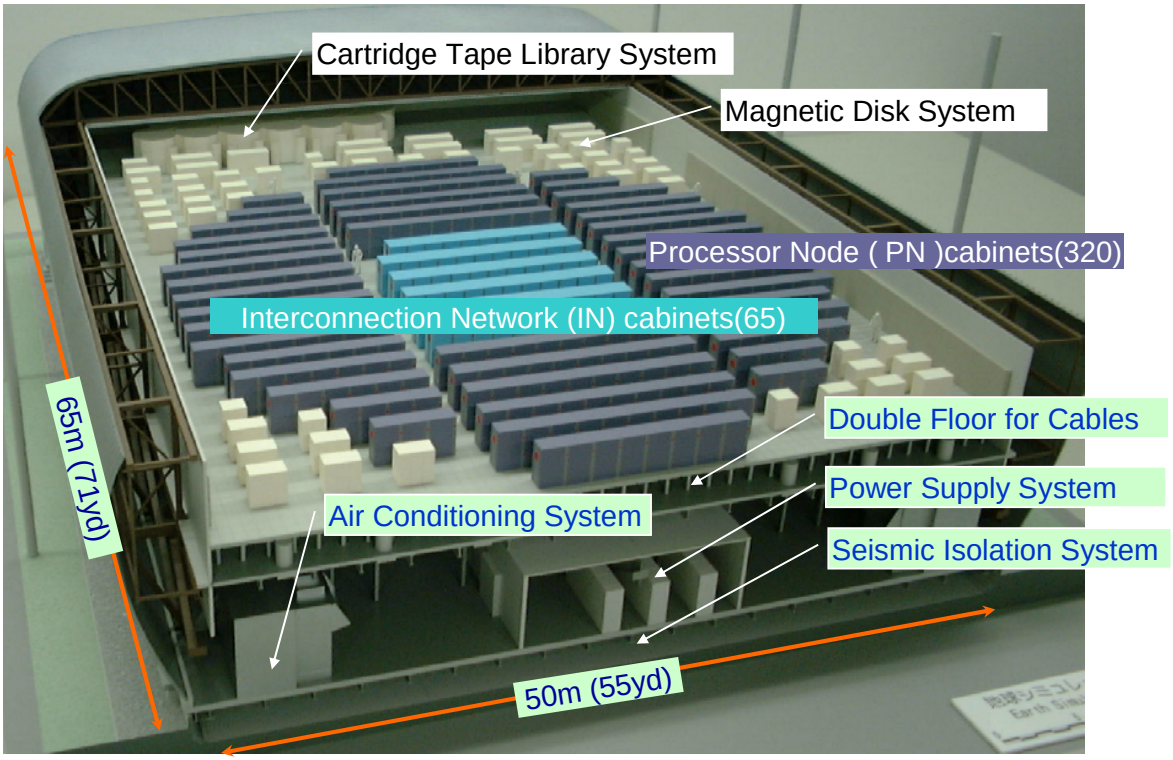
At the end of basic design stage NEC proposed specifications below, and the specifications of the Earth Simulator was decided as follows, except for the performance target of AGCM (remained as 5 Tflops).

Architecture :	A highly parallel vector processor system, consisting of 640 shared memory parallel vector processor nodes.
Total number of AP's:	5120
Total number of processor nodes :	640
Number of AP's for each PN:	8
Total peak performance:	40 Tflops
Peak performance of each PN:	64 Gflops
Peak performance of each AP:	8 Gflops
Total main memory size:	10 TB
Main memory size / PN:	16 GB
Total main memory bandwidth:	160TB/s
Main memory bandwidth / PN:	256GB/s
Main memory bandwidth / AP:	32GB/s
Interconnection network :	Single-Stage Crossbar Network

Aggregate switching capacity of interconnection network : 7872GB/s
Inter-node bandwidth / PN: 12.3GB/s (bi-sectional)
Performance estimation of AGCM code: 10.4Tflops (6144x3074x255: T2047L255)

Implementation of hardware

Scale Model of the Earth Simulator

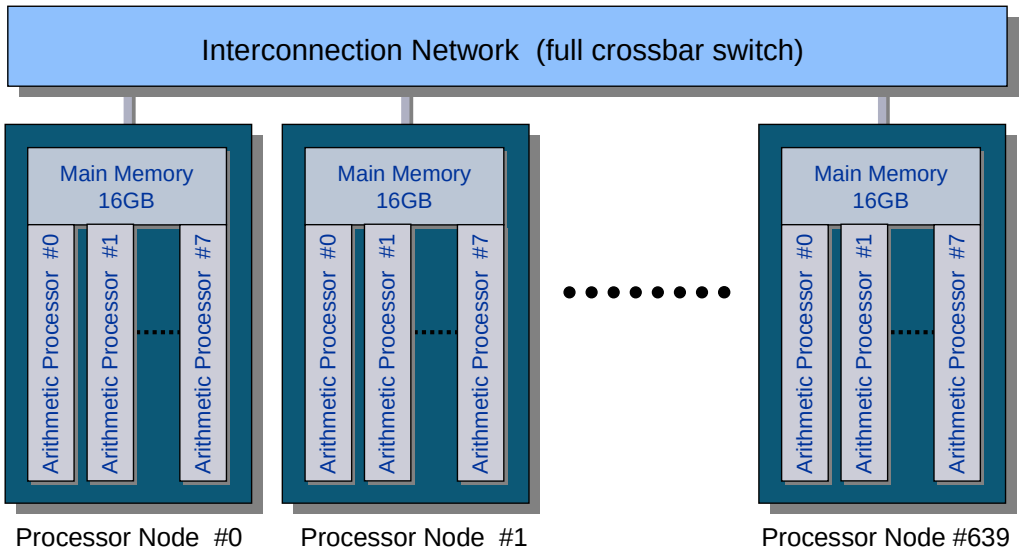


Configuration of the Earth Simulator

- Peak performance/AP : 8Gflops
- Peak performance/PN : 64Gflops
- Main memory/PN : 16GB

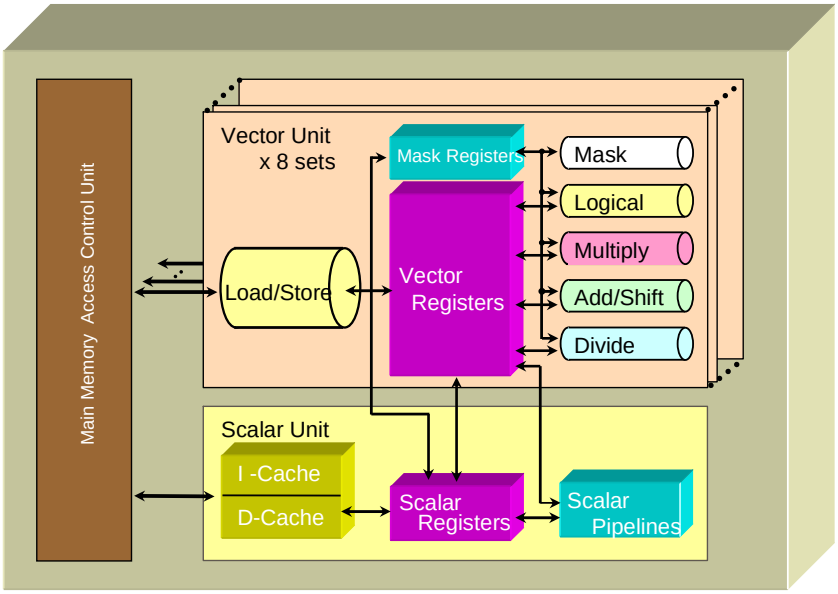
- Total number of APs : 5120
- Total number of PNs : 640

- Total peak performance: 40Tflops
- Total main memory : 10TB



Arithmetic Processor configuration

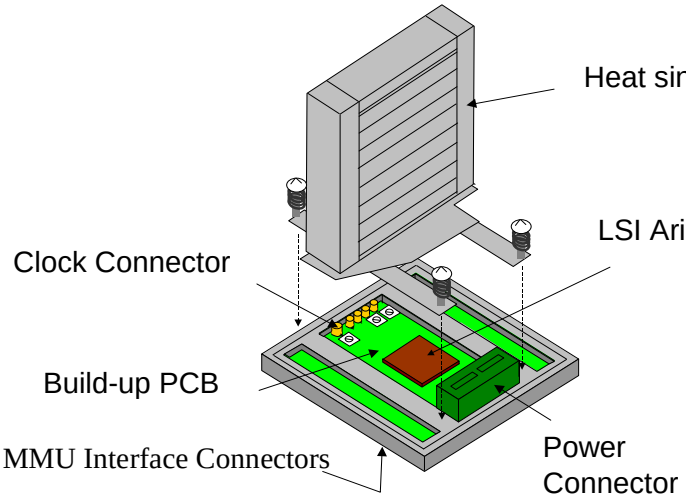
- Scalar Unit (SU)
  - 4-way superscalar
  - 128 scalar registers
  - 64KB Instruction cache
  - 64KB data cache
- 8 units of vector pipelines(VU)
  - 6 types of operation pipeline
  - 144KB vector registers
  - 256bit x 17 vector mask registers
- Main memory access contorl unit

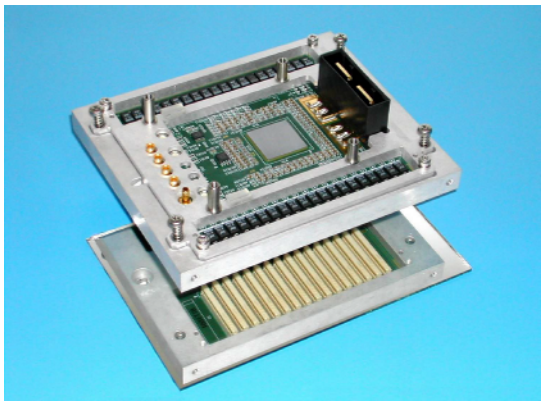


One Chip LSI: 8Gflops

- 0.15µm CMOS LSI
  + Cu interconnection
- 20.79 mm x 20.79 mm
- 60 million transistors
- More than 5000 pins
- 500MHz (partially 1GHz)

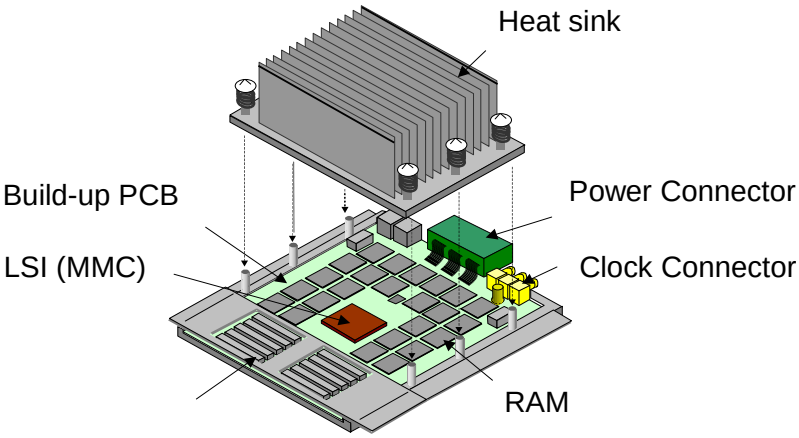
Arithmetic Processor Package

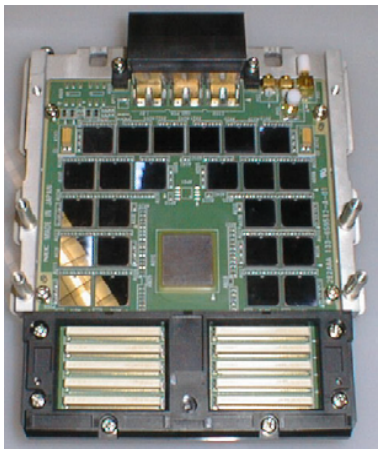




(115mm x 139mm)

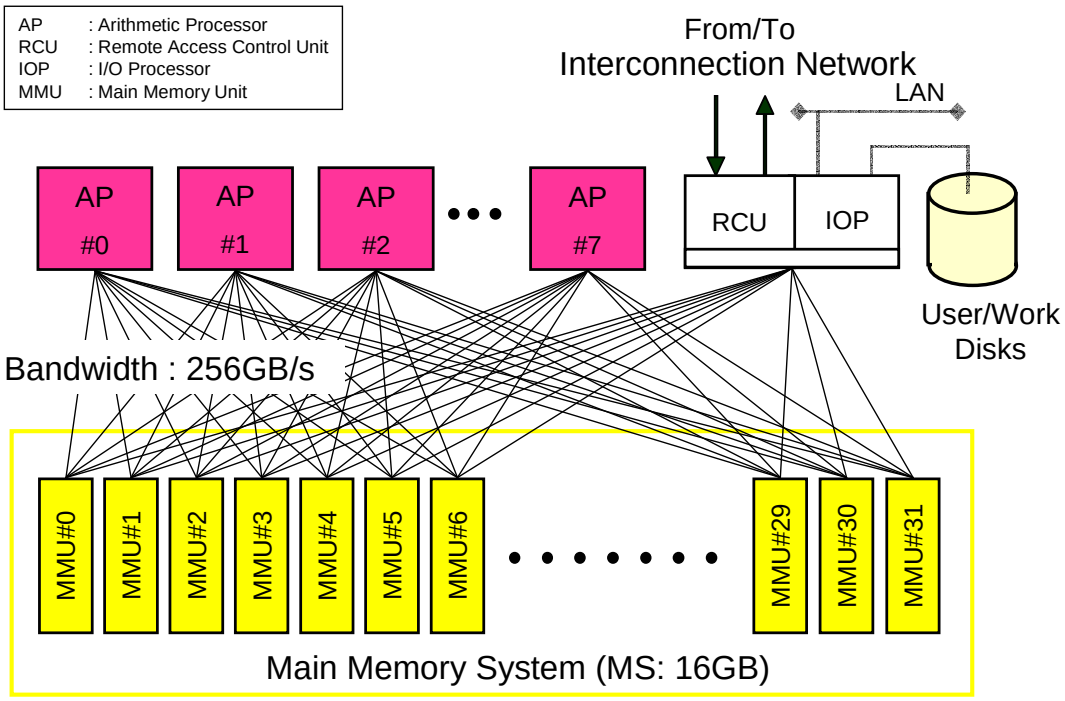
Main Memory Unit Package





(125mm x 147mm)

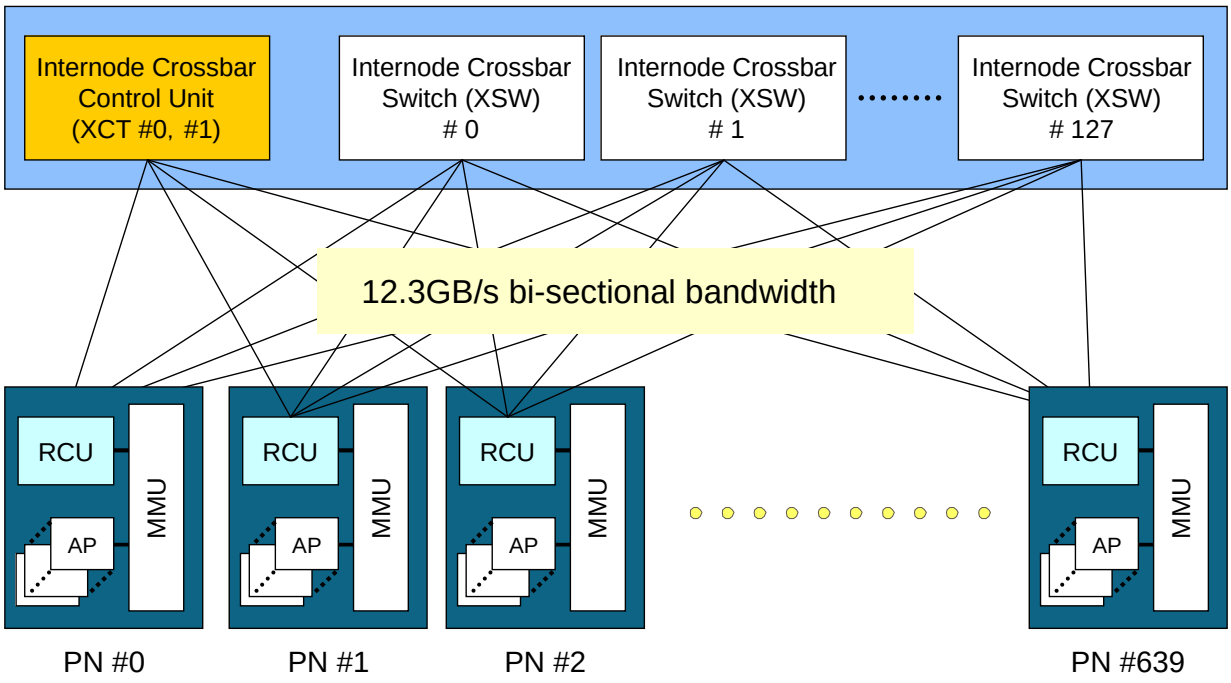
Processor Node configuration



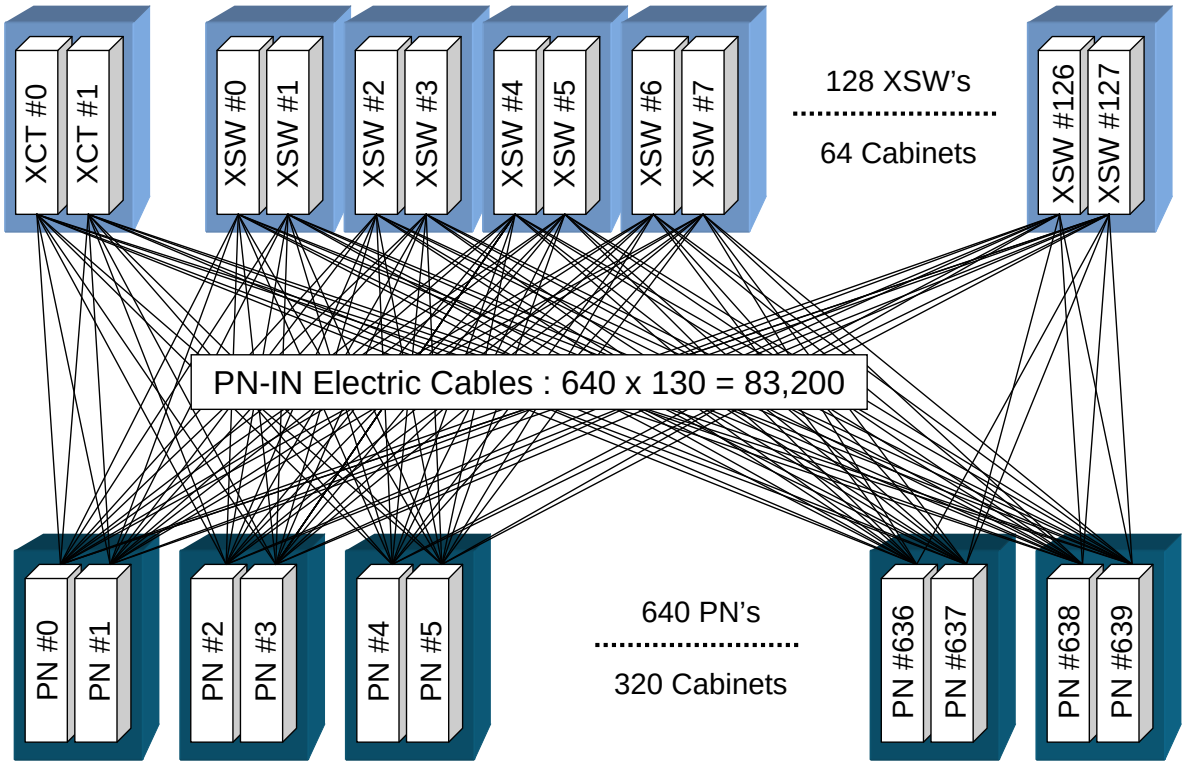
- 128M bit DRAM developed for ES (24nsec bank cycle time)
- 2048 banks

Interconnection Network (IN)

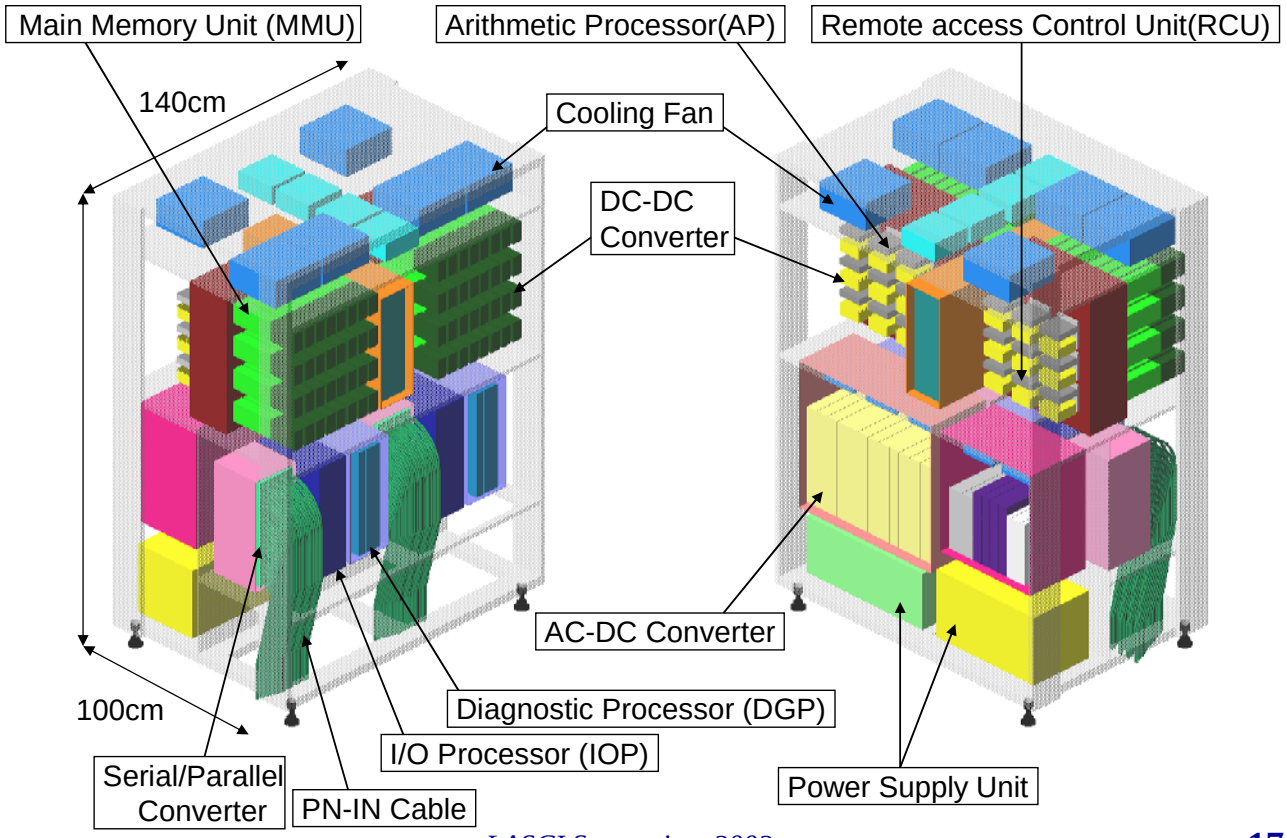
- 640 x 640 full crossbar switch
- 2 XCT's and 128 XSW's
  - XCT : Coordination of data transfer through XSW's



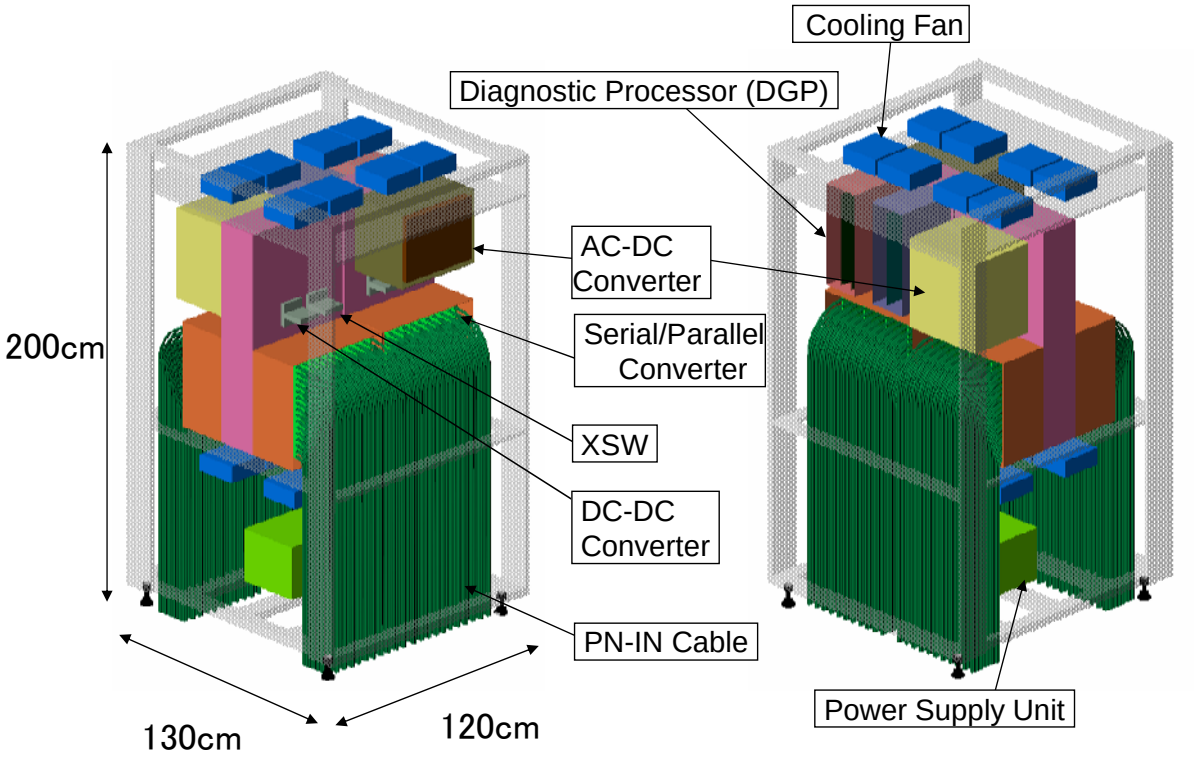
Connection between Cabinets



Processor-Node Cabinet (two nodes in a cabinet)



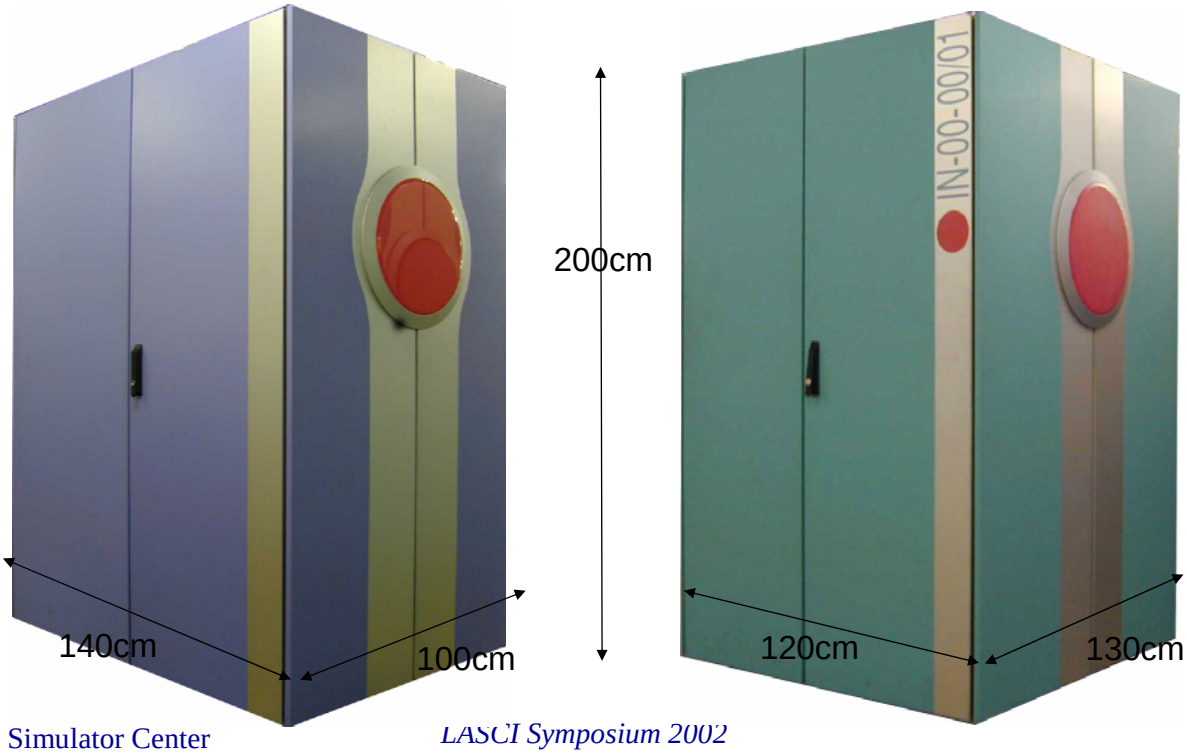
XSW Cabinet (two XSW's in a cabinet)



External appearance of Cabinets

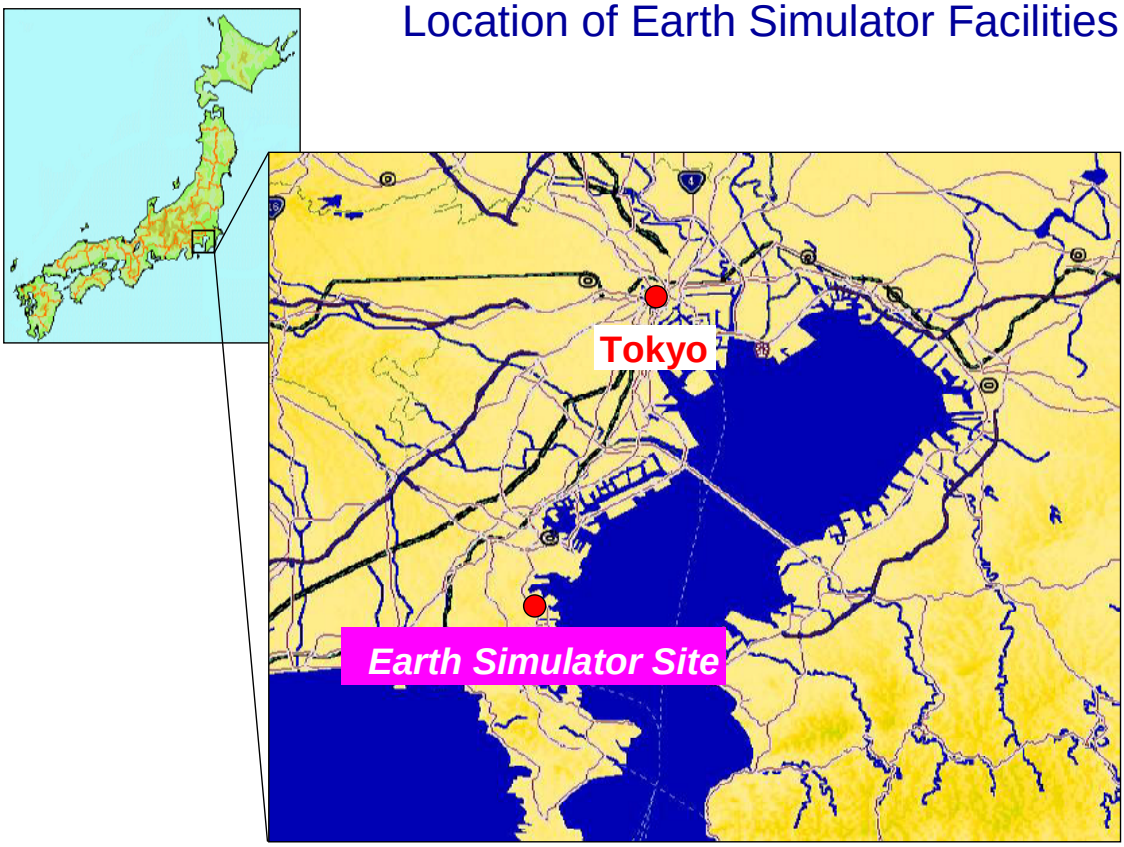
PN Cabinet

IN Cabinet

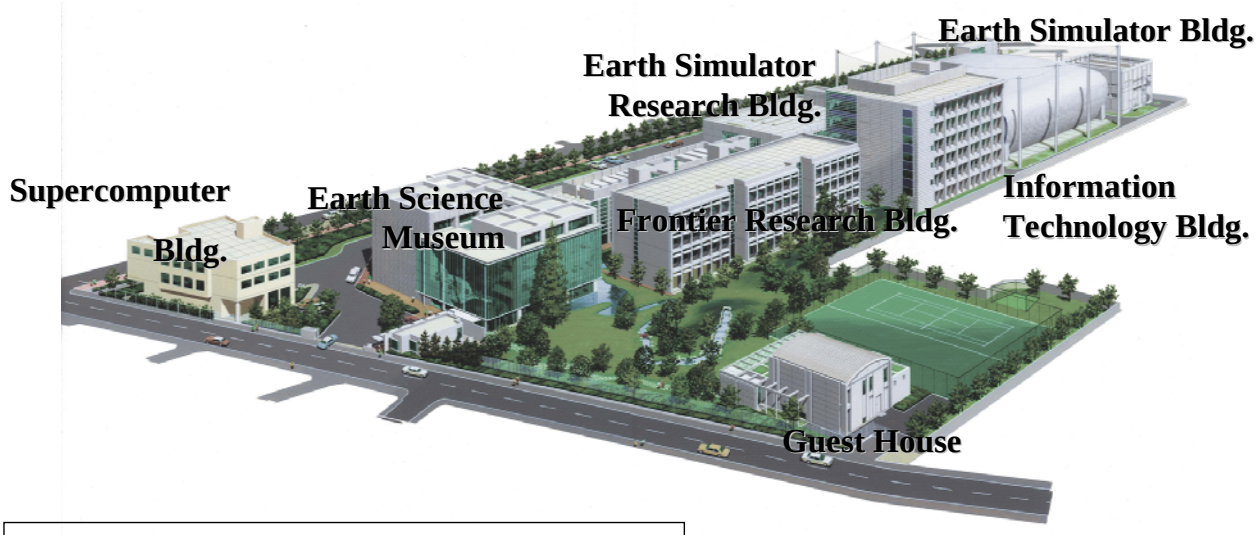


Installation of hardware

Location of Earth Simulator Facilities



the Earth Simulator Center

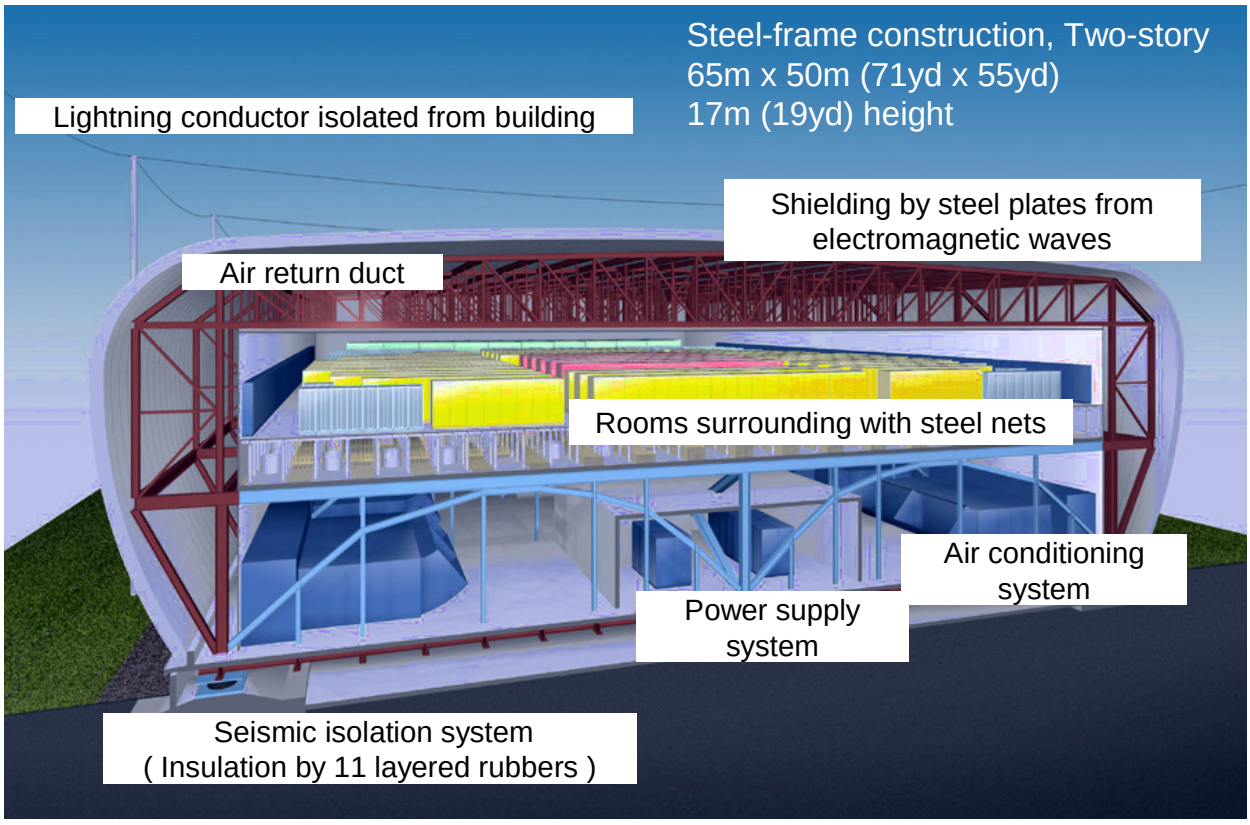


Yokohama Institute For Earth Science

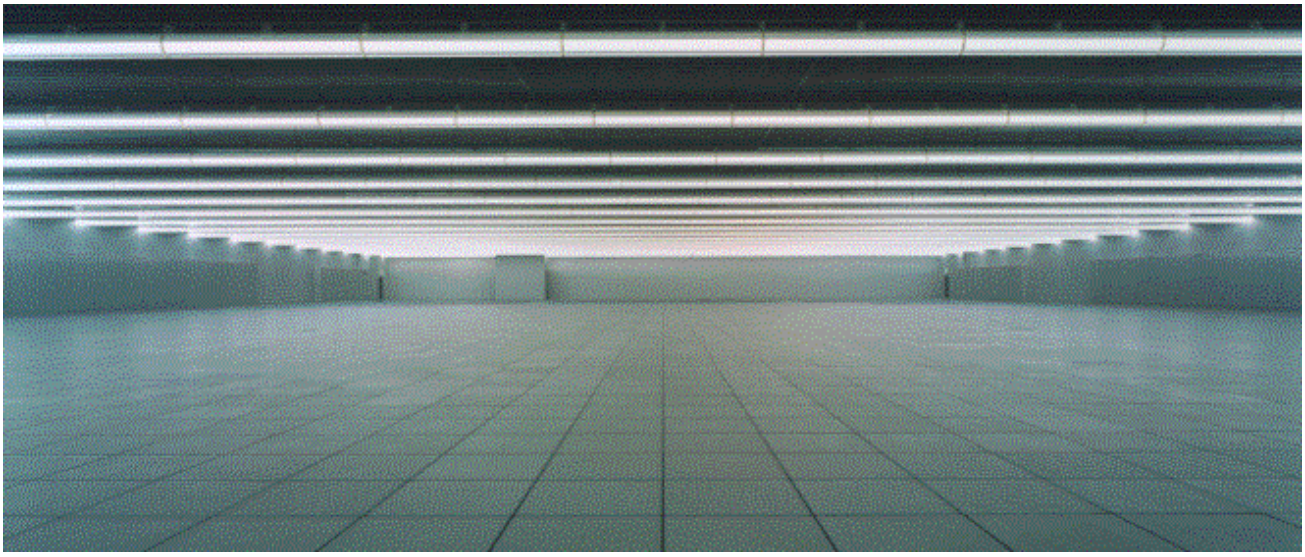
Japan Marine Science and Technology Center

Earth Simulator Building

Features of Earth Simulator Building



Lighting of Computer Room



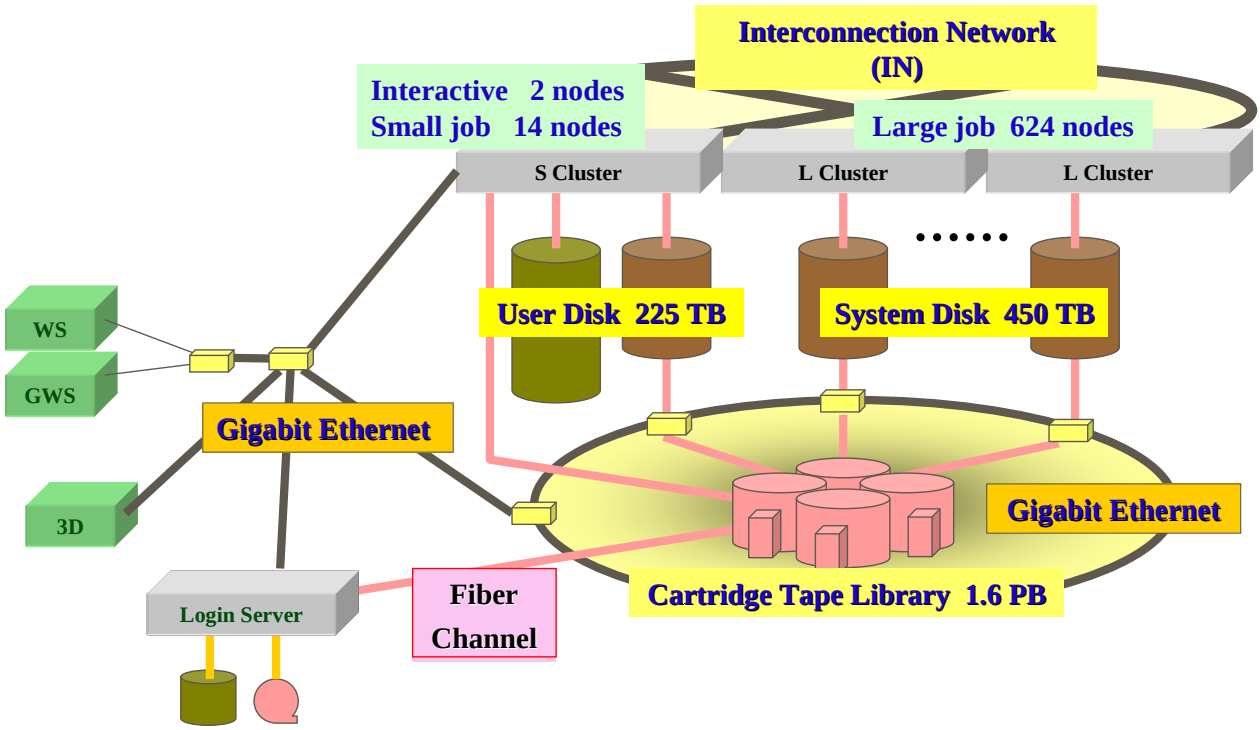
Lighting : Light propagation system inside a tube
(255mm diameter, 44m(49yd) length, 19 tubes)
Light source : halogen lamps of 1kW
Illumination : 300 lx at the floor in average

Electric Cables Connecting Cabinets

Earth Simulator at Completion

Another Photo of Earth Simulator

Connection among Peripherals



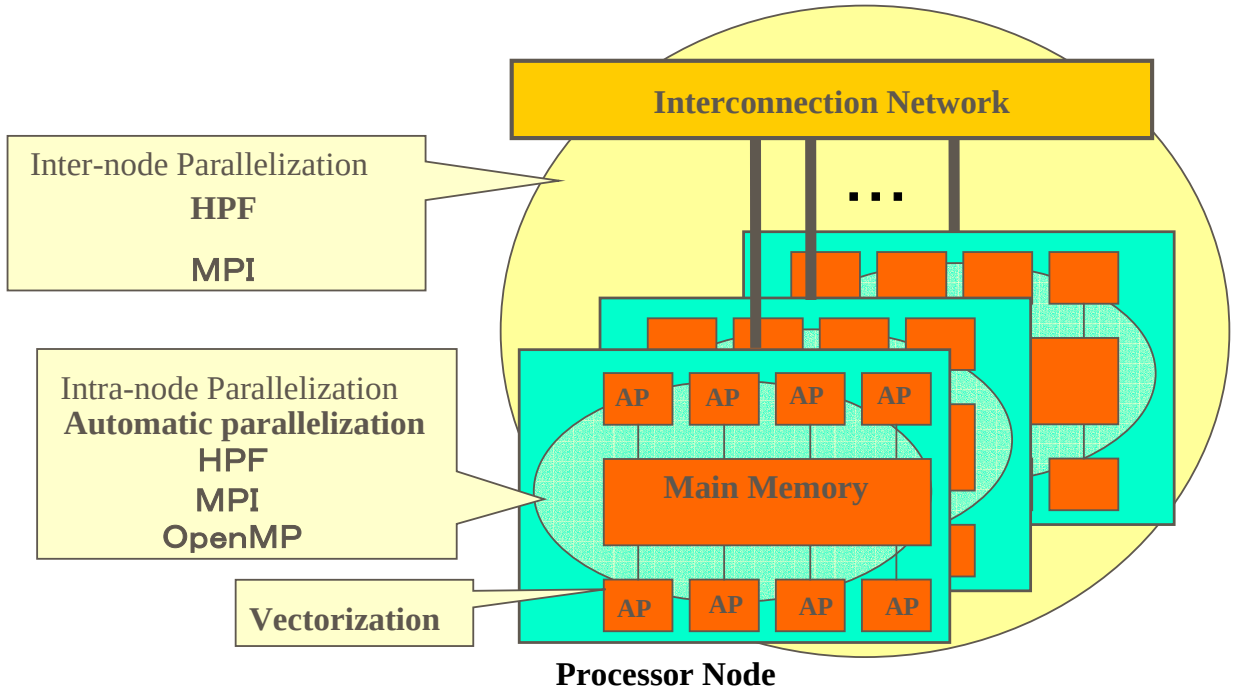
Software & Achieved Performance

Software Environment

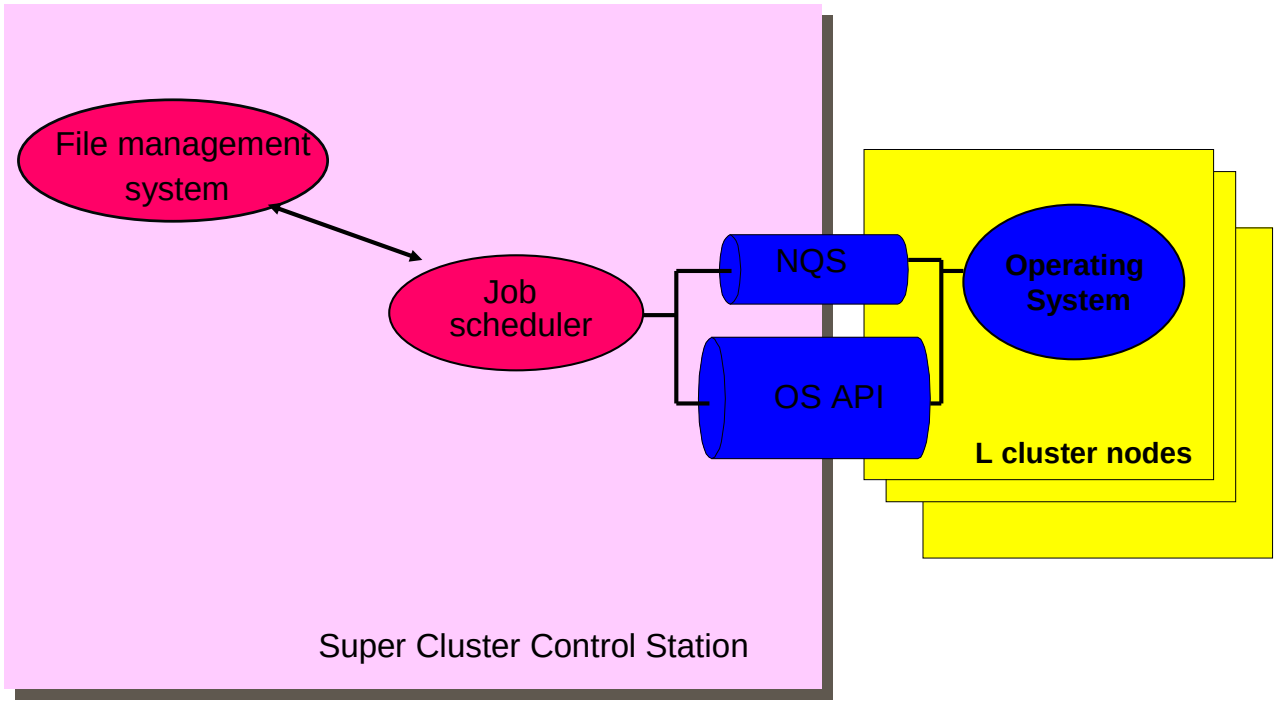
- Operating System
  - UNIX-based system (Enhanced version of NEC SUPER-UX)
  - Parallel file system (MPI-IO, HPF)
- Programming Environment
  - Parallel programming environment ({Fortran90,C,C++}+MPI, HPF)
  - Tuning tools
- Job scheduler
  - Extension of NQS
  - Running on the SCCS
  - Job assignment to PN's with file loading to appropriate system disks

These software have a good scalability up to 640 nodes.

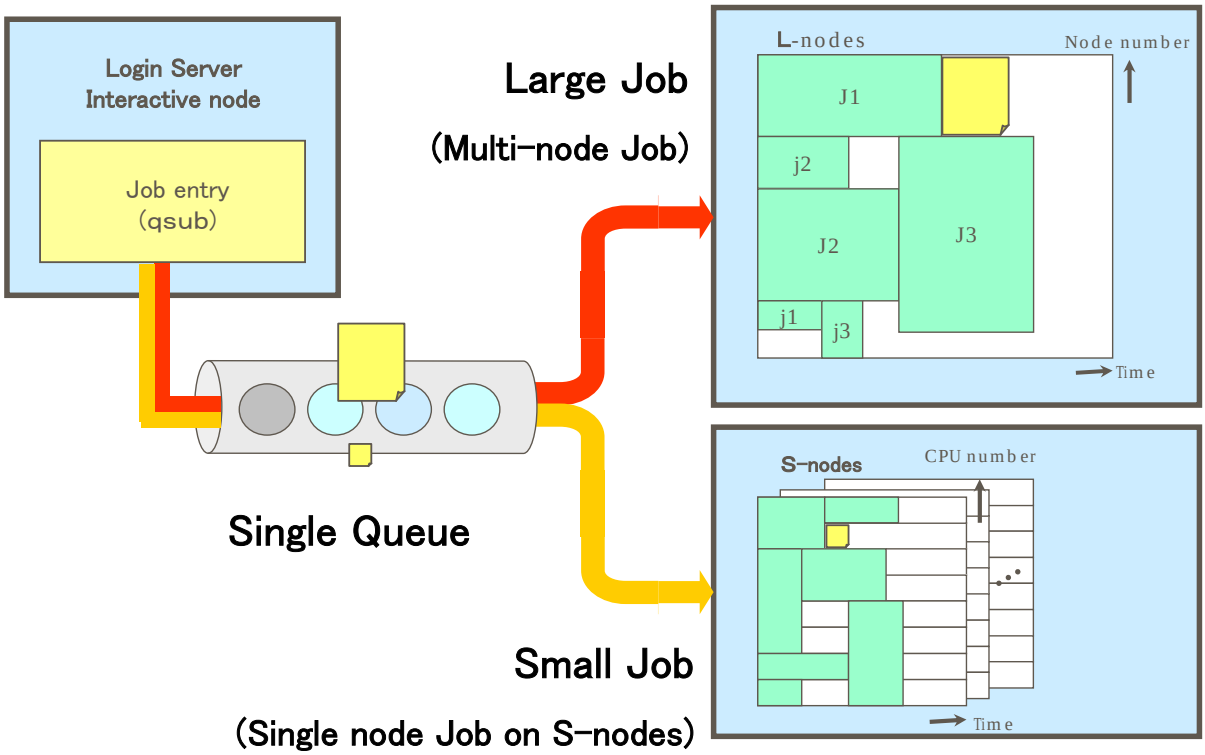
Vecrorization and Parallelization



Job Scheduler

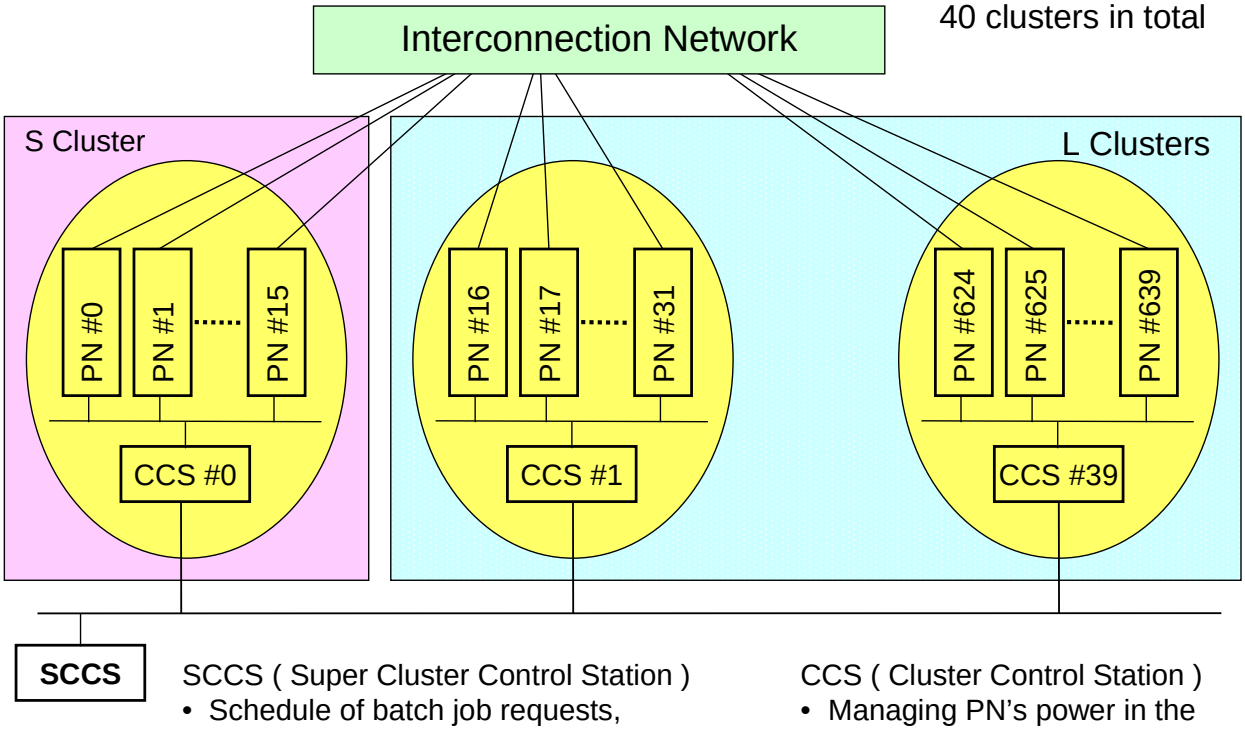


Batch Jobs



Cluster Structure in ES

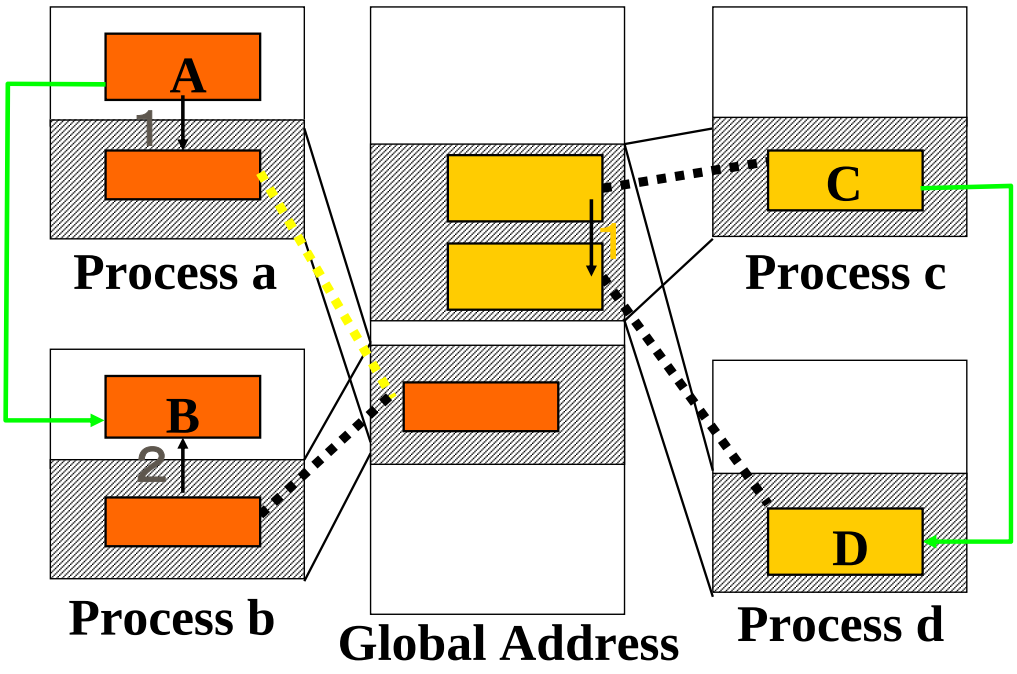
16 PNs / cluster
40 clusters in total



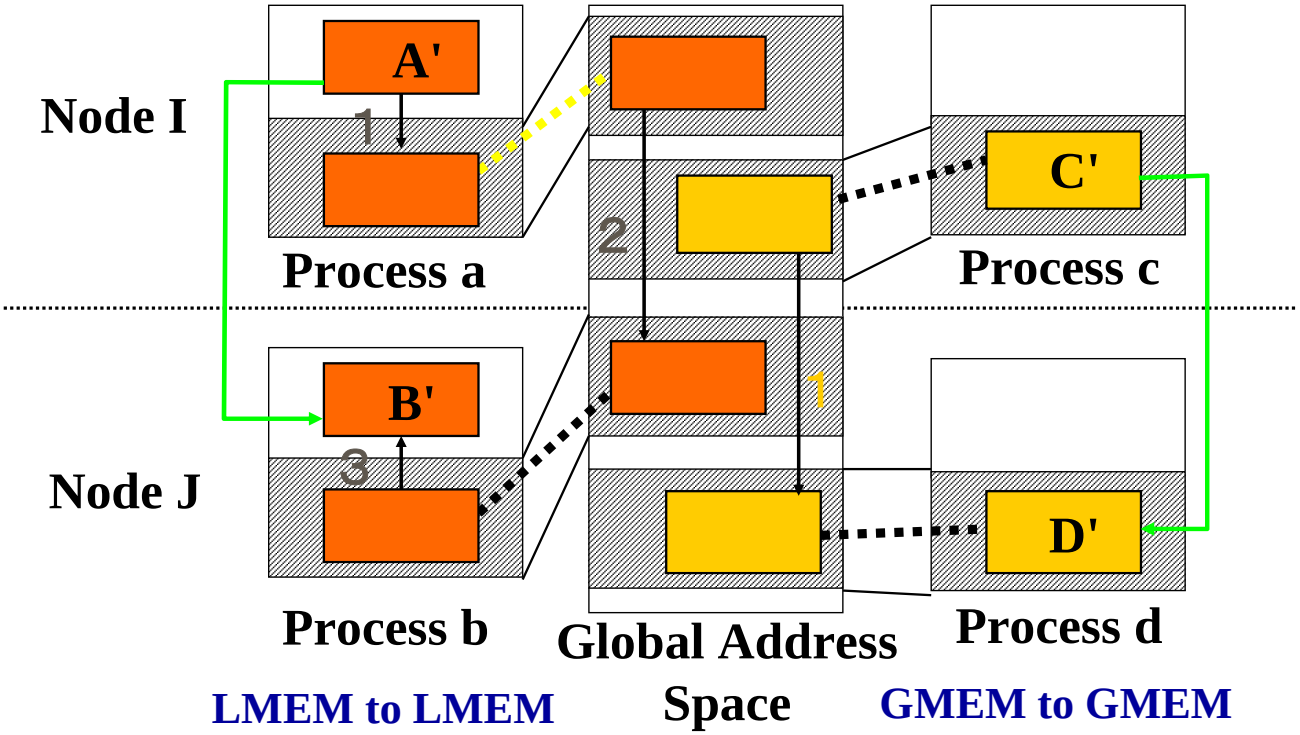
SCCS (Super Cluster Control Station)
- Schedule of batch job requests,
- Statistics for system utilization, etc.

CCS (Cluster Control Station)
- Managing PN's power in the cluster, etc.

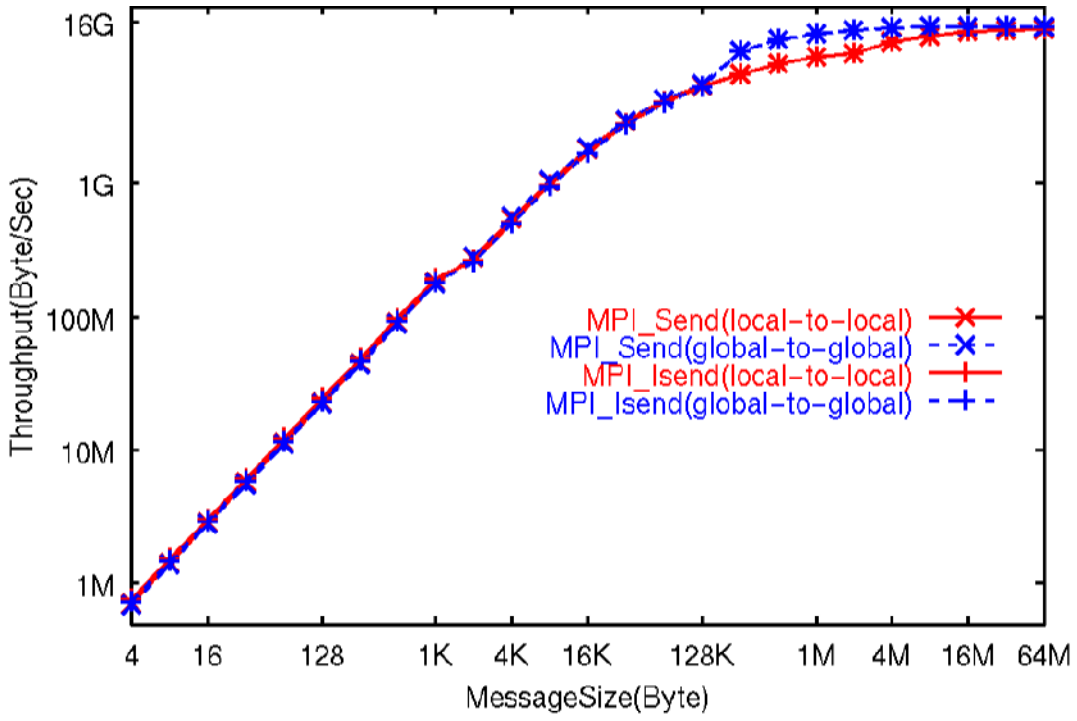
MPI implementation within PN



MPI Implementation between PN's



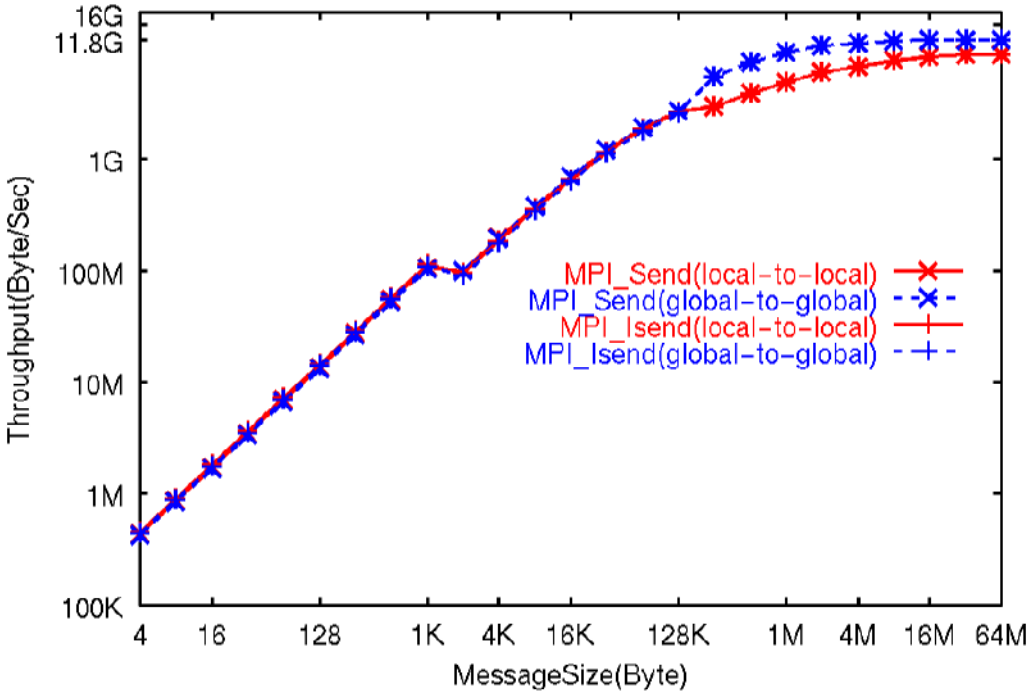
Throughput of MPI_Send Function (intra-node)



14.8GB/s

(92.5%)

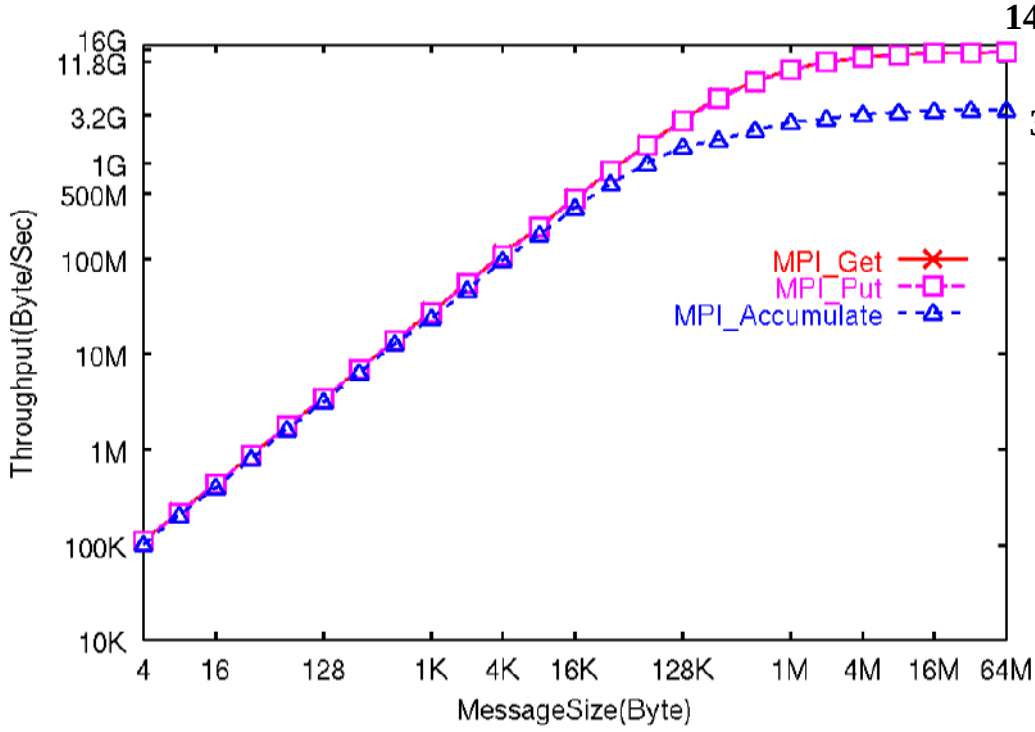
Throughput of MPI_Send Function (inter-node)



11.8GB/s

(96%)

Throughput of MPI-2 Functions (intra-node)

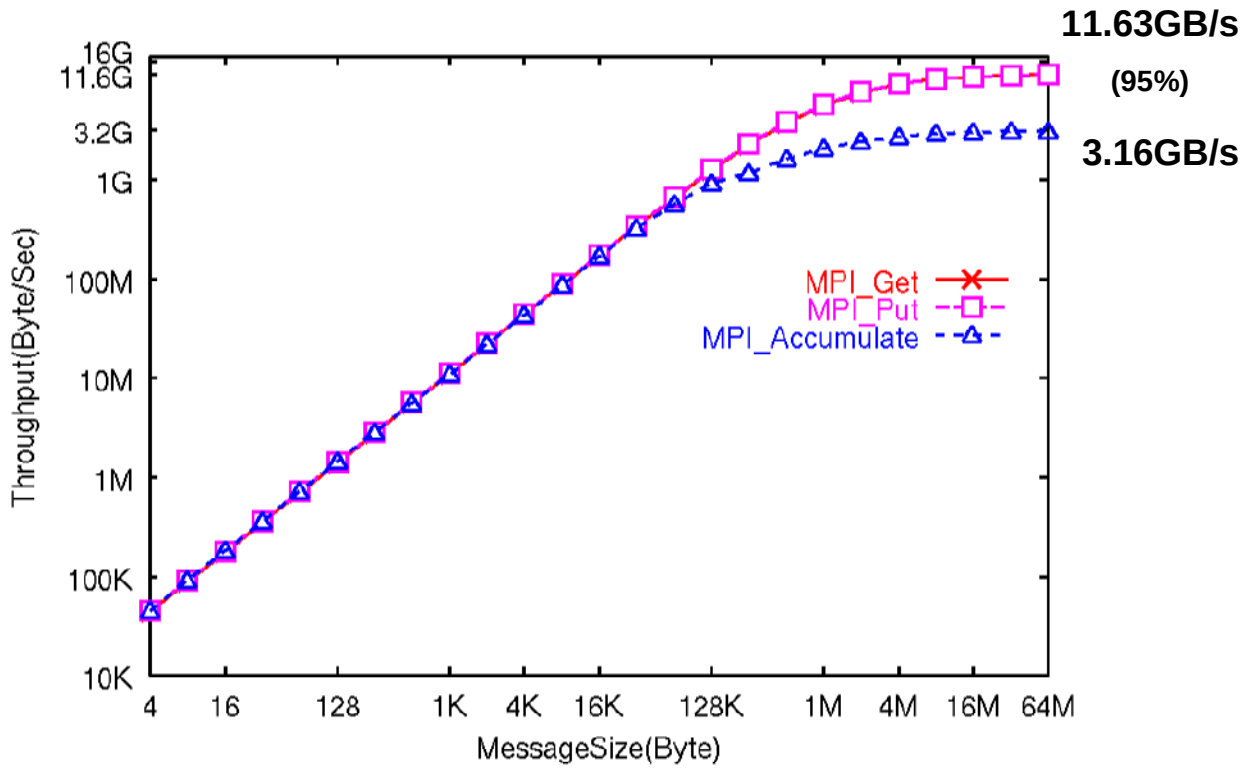


14.78GB/s

(92%)

3.64GB/s

Throughput of MPI-2 Functions (inter-node)



Latency of MPI Functions

Function name	Inter-node	Intra-node
MPI_send	5.58	1.38
MPI_Isend	5.90	1.75
MPI_Put	6.36	1.35
MPI_Get	6.68	1.27
MPI_Accumulate	7.65	3.87

(microseconds)

- The inter-node barrier synchronization uses special hardware and takes about 3.2 microsecond independent to the number of nodes.

AFES

AFES (AGCM For Earth Simulator) is an optimized code of the atmospheric general circulation model SAGCM for the Earth Simulator, which has been developed by the Earth Simulator Research and Development Center (ESRDC). The model SAGCM is based on the CCSR/NIES AGCM jointly developed by the Center for Climate System Research of the University of Tokyo and the National Institute for Environmental Studies, Japan.

AFES is a three-dimensional global hydrostatic model with spectral transform method, consists of dynamics and physics parts, and written in Fortran90.

For three years, ESRDC optimized this program for the Earth Simulator using Fortran90 automatic vectorization , micro tasking, MPI, and so on.

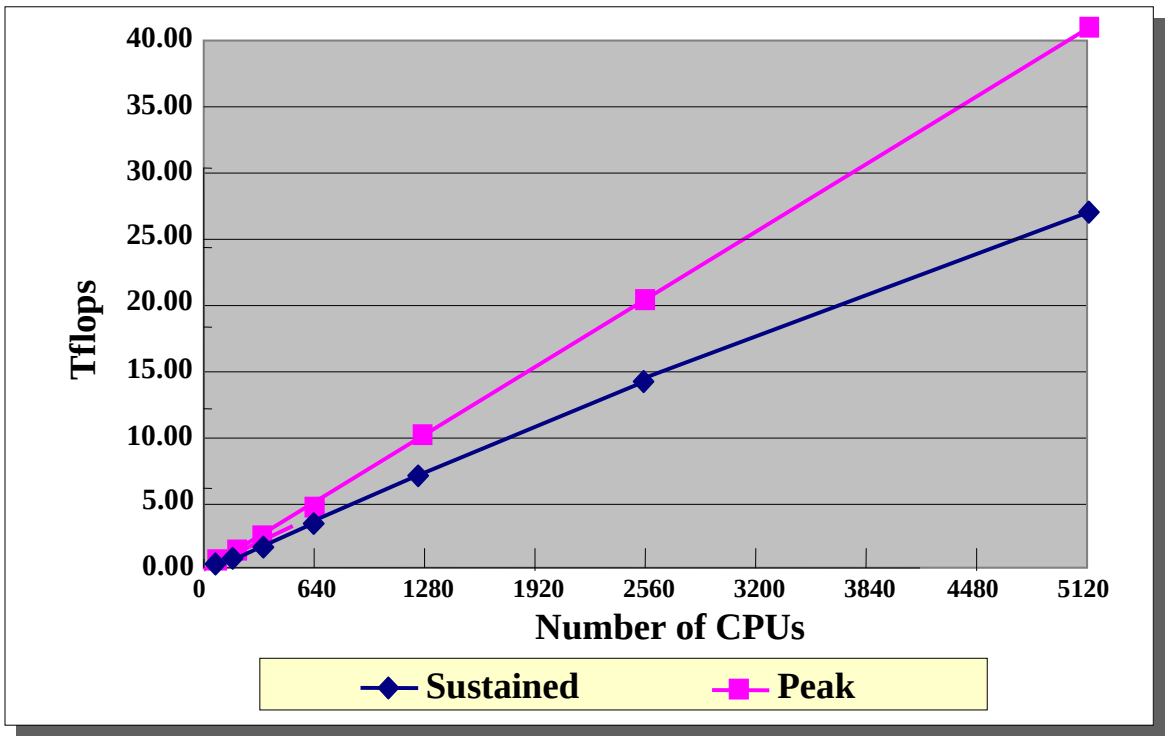
We achieved the sustained performance 26.58 Tflops for the AFES (horizontal resolution : approx. 10 km) with full exploitation of the 640 nodes configuration. The resulting computing efficiency is 65 % of the peak performance.

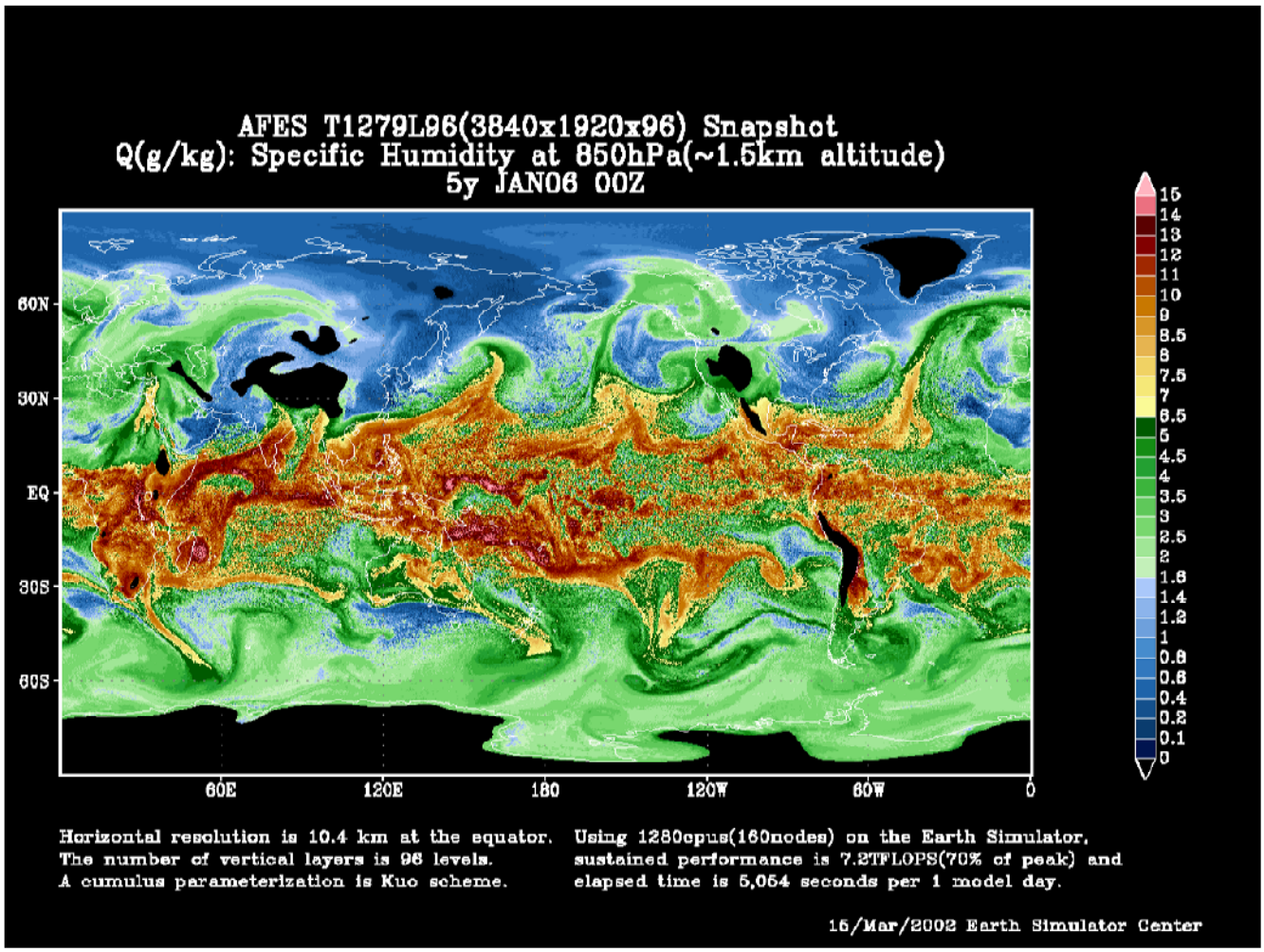
Performance Scalability of AFES (T1279L96)

Total CPU	Node	CPU/Node	ELAPSE (sec)	Tflops		Ratio (%)
				Sustained	Peak	
80	80	1	238.04	0.52	0.64	81.1
160	160	1	119.26	1.04	1.28	81.0
320	320	1	60.52	2.04	2.56	79.8
640	80	8	32.06	3.86	5.12	75.3
1280	160	8	16.24	7.61	10.24	74.3
2560	320	8	8.52	14.50	20.48	70.8
5120	640	8	4.65	26.58	40.96	64.9

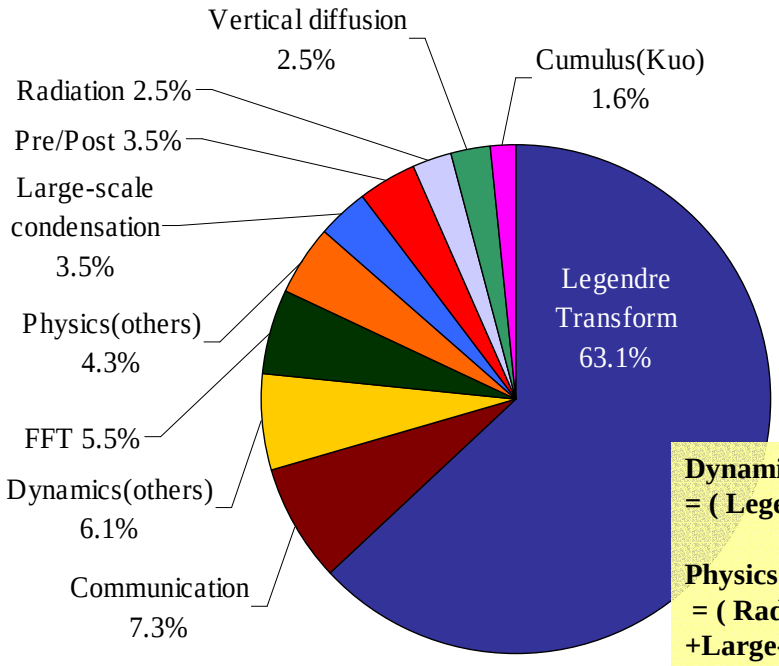
Number of time integration steps : 10

Performance Scalability of AFES (T1279L96)





Cost profile of AFES (T1279L96) for a 1-day simulation on the ES (160 nodes * 8 CPUs per node =1280 CPUs)



1 model day integration
7.20 Tflops(70.3%)

Process	Elapse Time (sec)	Ratio (%)
Dynamics	3779.2	74.7%
Physics	729.8	14.4%
Communication	370.2	7.3%
Pre/Post	178.2	3.5%
Total	5057.5	100%

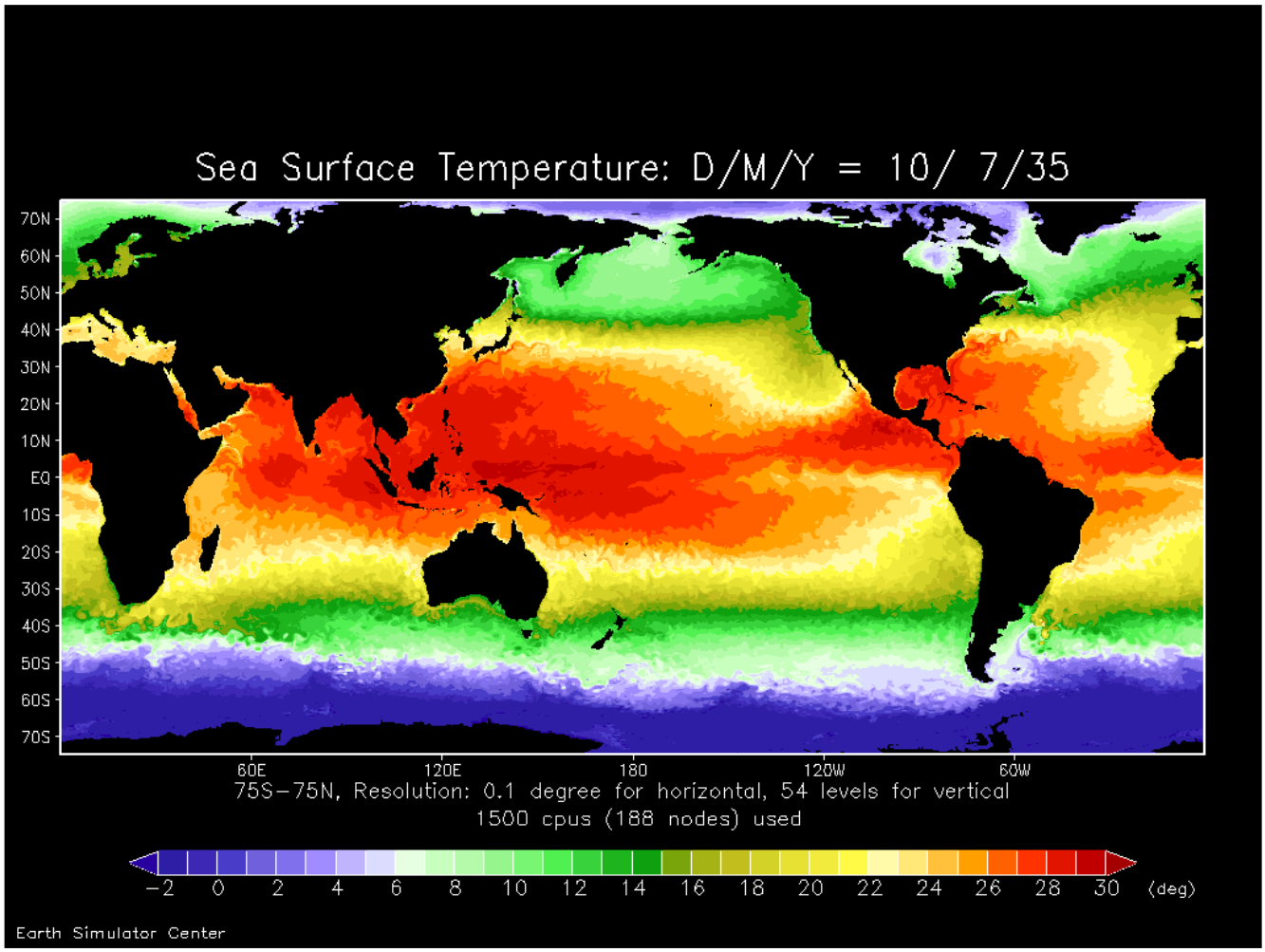
(Root process profile)

Dynamics Process
= (Legendre Transform + FFT + others)

Physics Process
= (Radiation
+Large-scale condensation
+ Vertical diffusion
+ Cumulus convection + others)

MOM3

MOM3 :Oceanic General Circulation Model developed in GFDL, USA is optimized and parallelized for the Earth Simulator, by the Earth Simulator Research and Development Center (ESRDC). The test run using 175 nodes with eddy-resolving resolution (0.1 degree) shows the Kuroshio Current and Gulf Stream and achieved 2.5Tflops (22% of the peak performance).



EARTH SIMULATOR

Achievements of the High Performance Fortran (HPF) for the Earth Simulator

- PFES (Oceanic General Circulation Model based on Princeton Ocean Model) achieved 9.85TFLOPS with 376 nodes (41% of the peak performance).
- Impact3D (Plasma fluid code using TVD scheme) achieved 14.9 TFLOPS with 512 nodes (45% of the peak performance).

Linpack (Highly Parallel Computing) Benchmark

We achieved 35.86 Tflops for the Linpack benchmark suite with full node configuration. Nmax and N1/2 were 1075200 and 266240, respectively.

Table 3: **Highly Parallel Computing**

Computer (Full Precision)	Number of Processors	R_{max} Gflop/s	N_{max} order	$N_{1/2}$ order	R_{peak} Gflop/s
Earth Simulator ****	5120	35860	1075200	266240	40960
ASCI White-Pacific, IBM SP Power 3(375 MHz)	8000	7226	518096	179000	12000
Compaq AlphaServer SC ES45/EV68 1GHz	3016	4463	280000	85000	6032
Compaq AlphaServer SC ES45/EV68 1GHz	3024	4059	525000	105000	6048
Compaq AlphaServer SC ES45/EV68 1GHz	2560	3980	360000	85000	5120
IBM SP Power3 208 nodes 375 MHz	3328	3052.	371712		4992
Compaq Alphaserver SC ES45/EV68 1GHz	2048	2916	272000		4096
IBM SP Power3 158 nodes 375 MHz	2528	2526.	371712	102400	3792
ASCI Red Intel Pentium II Xeon core 333MHz	9632	2379.6	362880	75400	3207
IBM p690 cluster, Power 4 1.3 GHz	864	2310	275000	62000	4493
ASCI Blue-Pacific SST, IBM SP 604E(332 MHz)	5808	2144.	431344	432344	3868
ASCI Red Intel Pentium II Xeon core 333MHz	9472	2121.3	251904	66000	3154
Compaq Alphaserver SC ES45/EV68 1GHz	1520	2096	390000	71000	3040
IBM p690 cluster, Power 4 1.3 GHz	768	2002	252000		3994
IBM SP 112 nodes (375 MHz POWER3 High)	1792	1791	275000	275000	2688
HITACHI SR8000/MPP/1152(450MHz)	1152	1709.1	141000	16000	2074
HITACHI SR8000-F1/168(375MHz)	168	1653.	160000	19560	2016
ASCI Red Intel Pentium II Xeon core 333Mhz	6720	1633.3	306720	52500	2238
SGI ASCI Blue Mountain	5040	1608.	374400	138000	2520
IBM SP 328 nodes (375 MHz POWER3 Thin)	1312	1417.	374000	374000	1968
Intel ASCI Option Red (200 MHz Pentium Pro)	9152	1338.	235000	63000	1830
NEC SX-5/128M8(3.2ns)	128	1192.0	129536	10240	1280

Table 3: Highly Parallel Computing

Computer	Number of Processors	Rmax Gflop/s	Nmax order	$N_{1/2}$ order	Rpeak Gflop/s	Rmax/Rpeak
Earth Simulator ** **	**5120**	**35860**	**1075200**	**266240**	**40960**	**87.5%**
ASCI White-Pacific, IBM SP Power 3(375 MHz)	**8000**	**7226**	**518096**	**179000**	**12000**	**60.2%**
Compaq Alpha Server SC ES45/EV68 1GHz	**3016**	**4463**	**280000**	**85000**	**6032**	**74.0%**
Compaq Alpha Server SC ES45/EV68 1GHz	**3024**	**4059**	**525000**	**105000**	**6048**	**67.1%**
Compaq Alpha Server SC ES45/EV68 1GHz	**2560**	**3980**	**360000**	**85000**	**5120**	**77.7%**
IBM SP Power3 208 nodes 375 MHz	**3328**	**3052**	**371712**		**4992**	**61.1%**
Compaq Alpha server SC ES45/EV68 1GHz	**2048**	**2916**	**272000**		**4096**	**71.2%**
IBM SP Power3 158 nodes 375 MHz	**2528**	**2526**	**371712**	**102400**	**3792**	**66.6%**
ASCI Red Intel Pentium II Xeon core 333MHz	**9632**	**2379.6**	**362880**	**75400**	**3207**	**74.2%**
IBM p690 cluster, Power 4 1.3 GHz	**864**	**2310**	**275000**	**62000**	**4493**	**51.4%**

****** The Earth Simulator is not a commercial product, it is a computer of the Earth Simulator Center, the arm of the Japan Marine Science and Technology Center. It is based on vector processors that are manufactured by NEC.**

The columns in Table 3 are defined as follows:

Rmax the performance in Gflop/s for the largest problem run on a machine.

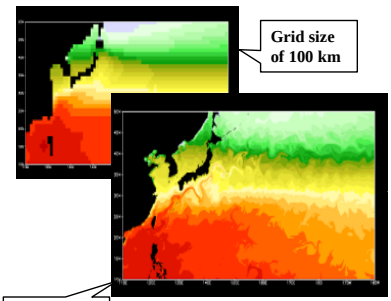
Nmax the size of the largest problem run on a machine.

$N_{1/2}$ the size where half the Rmax execution rate is achieved.

Rpeak the theoretical peak performance in Gflop/s for the machine.

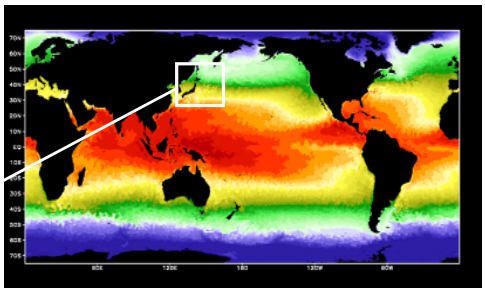
Activities of Earth Simulator Center

Comparison **between Two Simulation Results of Low (100 km) and High (10 km) Resolution Grid Size. Close-up snapshot of sea surface temperature near Japan.**

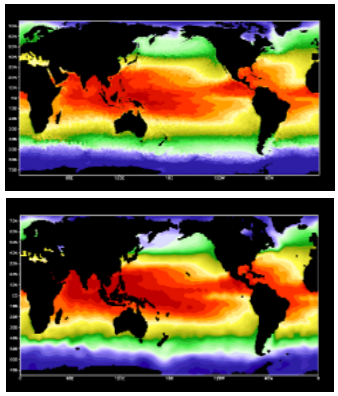


Oceanic Global Simulation

A result of the oceanic global simulation. Colored snapshot of sea surface temperature on one summer day. Higher temperature is shown in red, and lower temperature in purple.



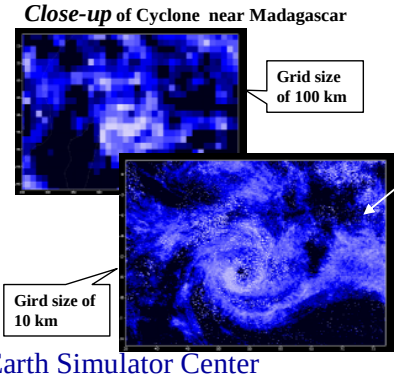
Comparison **between Earth Simulator's Result (upper) and Observation (lower)**

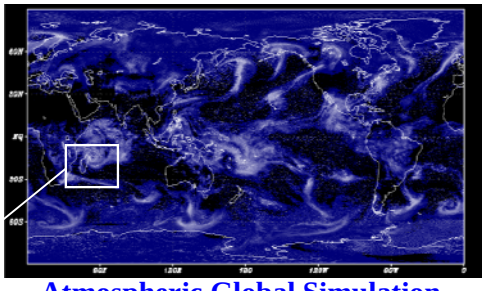


Starting Up the Earth Simulator

Operation of the Earth Simulator began in March 2002.
The Earth Simulator Center has already produced promising results through the ocean and atmospheric global simulations with an extremely high resolution of 10 km horizontal distance, which would place our hopes on reliable prediction of climate changes.

Close-up **of Cyclone near Madagascar**

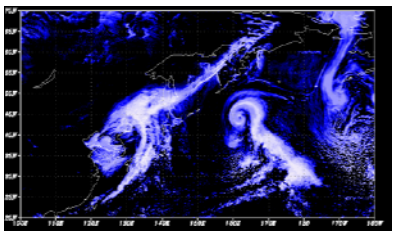




Atmospheric Global Simulation

Snapshot of the atmospheric global simulation in winter. Higher precipitation is shown in white.

Close-up **of Precipitation near Japan**



Message from Director-General Dr. Sato (extracts)

How will our mass consumption, such as automobiles, airplanes, electric device and chemical products which we have created, influence the future of global environment? Although we have been faced with various forms of highly unpredictable events such as earthquakes or global warming phenomena, we had never a measure to predict accurately what to happen in the future. With Earth Simulator, we have come to a point where we can predict the future. I wish to contribute to ensure people's lives and properties from natural disasters and environmental destructions, and bring forth a harmonized relationship with our mother Earth.

Mission of the Earth Simulator

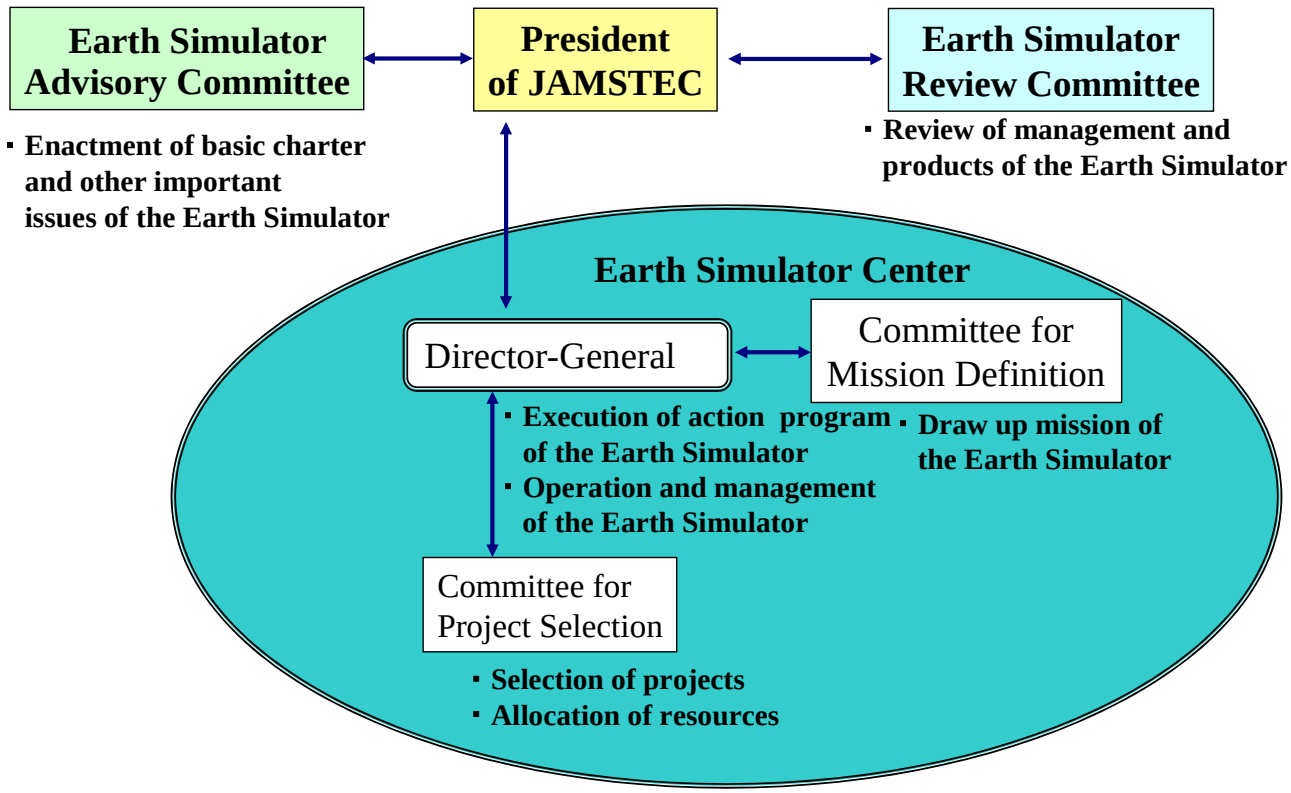
I. **Quantitative prediction and assessment of variations of the atmosphere, ocean and solid earth.**

II. **Production of reliable data to protect human lives and properties from natural disasters and environmental destructions**

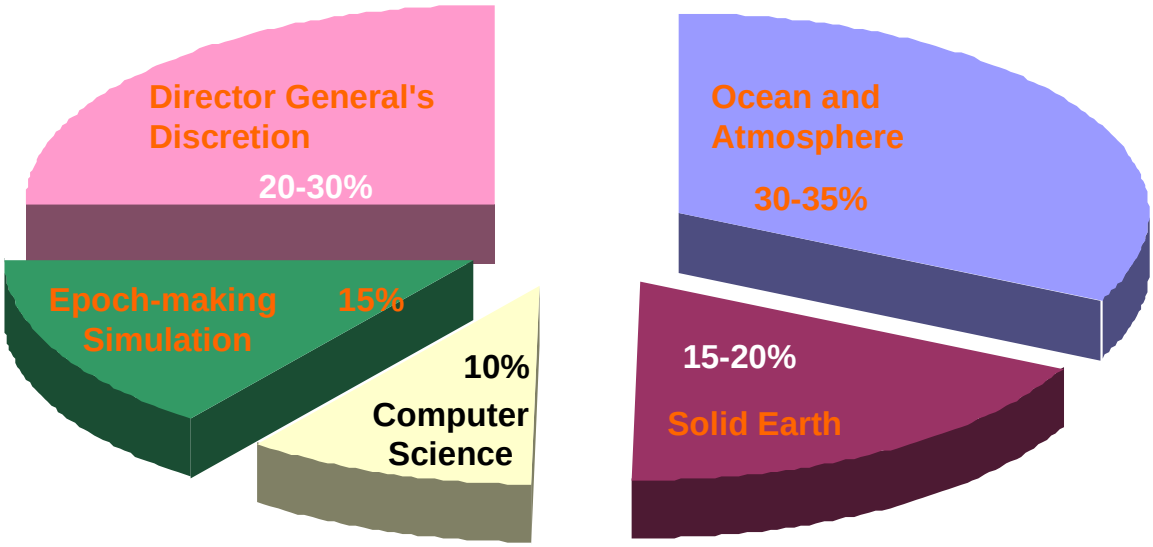
III. **Contribution to symbiotic relationship of human activities with nature.**

IV. **Promotion of innovative and epoch-making simulation in any fields such as industry, bioscience, energy science and so on.**

Managing System of the Earth Simulator



Allocation of Computer Resources in Year 2002



Selected Projects in 2002: 28

Ocean and Atmosphere	15	Solid Earth	8
Computer Science	3	Epoch-Making Simulation	2

Summary

- The world's fastest supercomputer, Earth Simulator, is successfully completed showing 40 Tflops theoretical peak performance.
- 35.86 Tflops is obtained in Linpack (Highly Parallel Computing).
- Application programs prepared for Earth Simulator were executed to evaluate the hardware performance. These have shown excellent results.
- Earth Simulator has already been in operation at the Earth Simulator Center (ESC), a branch of Japan Marine Science and Technology Center (JAMSTEC) .

Thank you for your attention

Shigemune Kitawaki

Earth Simulator Center
Japan Marine Science and Technology Center