System Overview and Operation of the Earth Simulator

Contents

Shigemune Kitawaki

Earth Simulator Center

Japan Agency for Marine-Earth Science and Technology

The 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.

Requirements and Design Target

- Processor Type (scalar processor or vector processor)

  Vector type processors were required >> single chip vector 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.

  Parallel Computing, Vol.21, No.10 November 1995

- Total Peak Performance

  More than 32 Tflops >> 40 Tflops

- Total Main memory size

  More than 8 TB >> 10 TB

- Type of interconnection network and aggregate switching capacity

  Single stage crossbar network with more than 4 TB/sec of aggregate switching capacity were required >> Single stage crossbar network with aggregate switching capacity: 7872GB/s

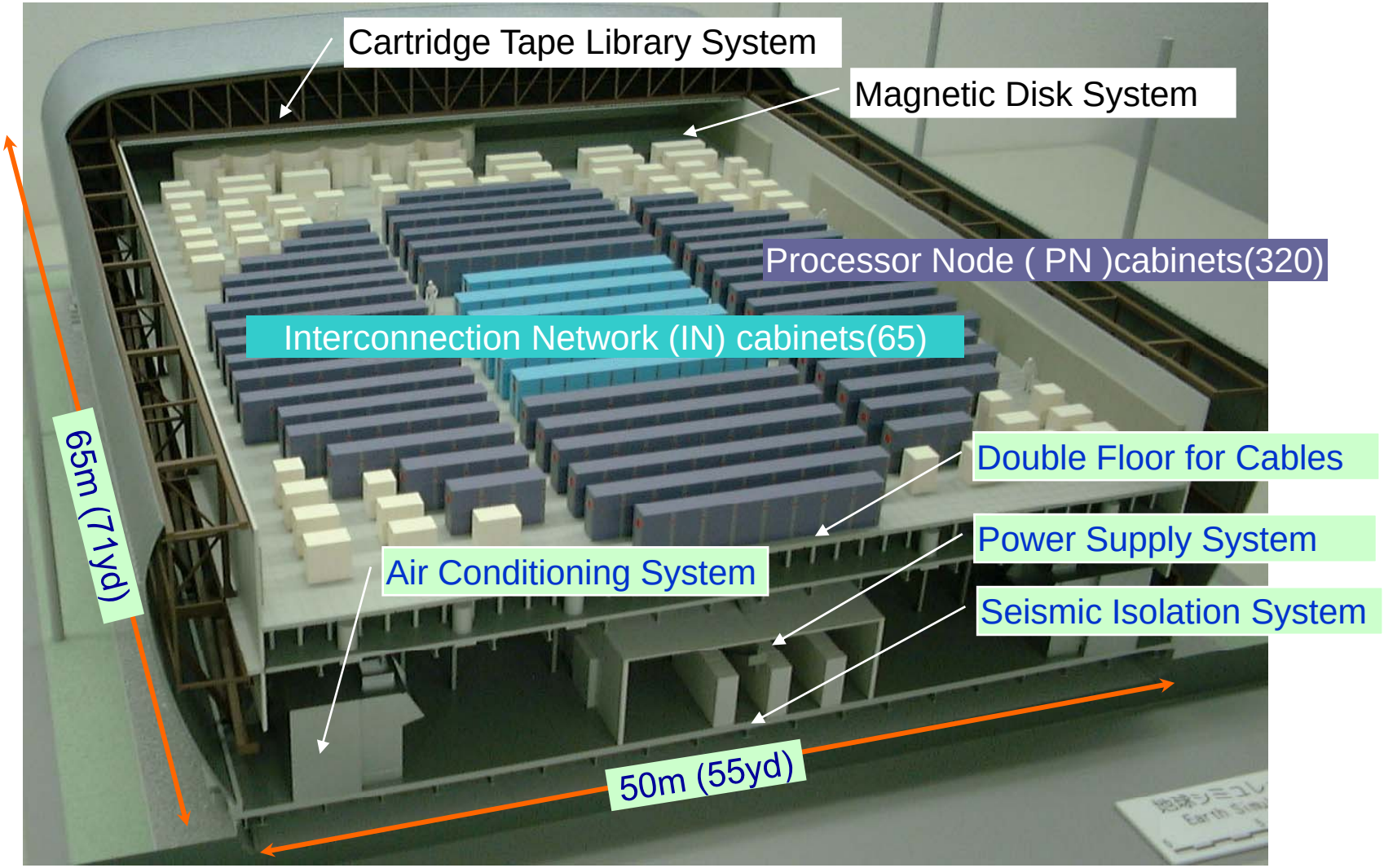
  A single stage crossbar network is superior in flexibility of allocating processor nodes to application programs and also in flexibility of executing many paradigm of applications.

- Performance of Atmospheric General Circulation Model (AGCM)

  More than 5 sustained Tflops (At least 1000 times faster than those of CRAY C90) were required >> Estimated the performance of AGCM of 6144x3074x255 mesh (T2047L255) at design stage, and evaluated the performance of AGCM of 3840x1920x96 mesh (T1279L96) at the completion of the whole system

Implementation of hardware

Scale Model of the Earth Simulator

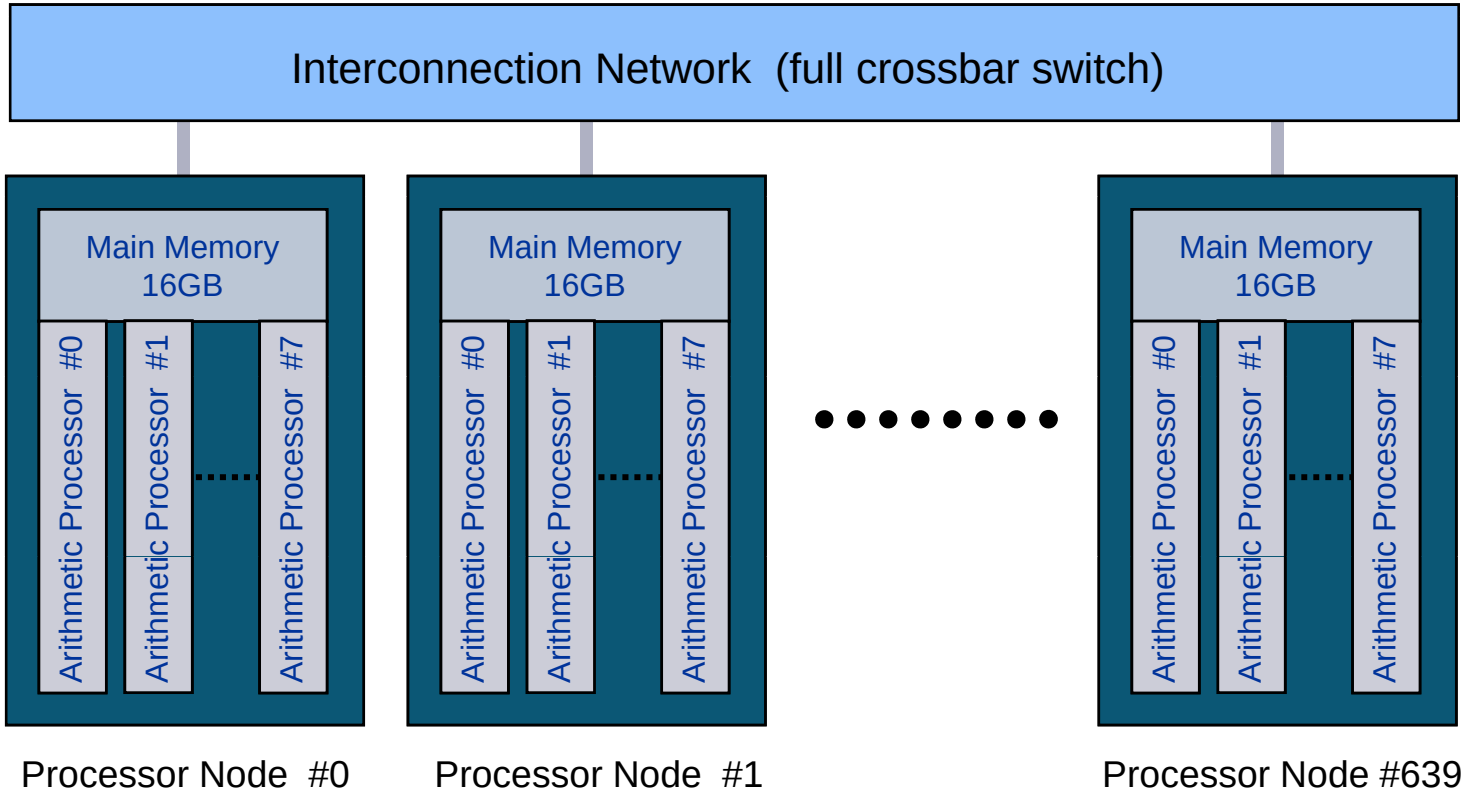


Configuration of the Earth Simulator

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

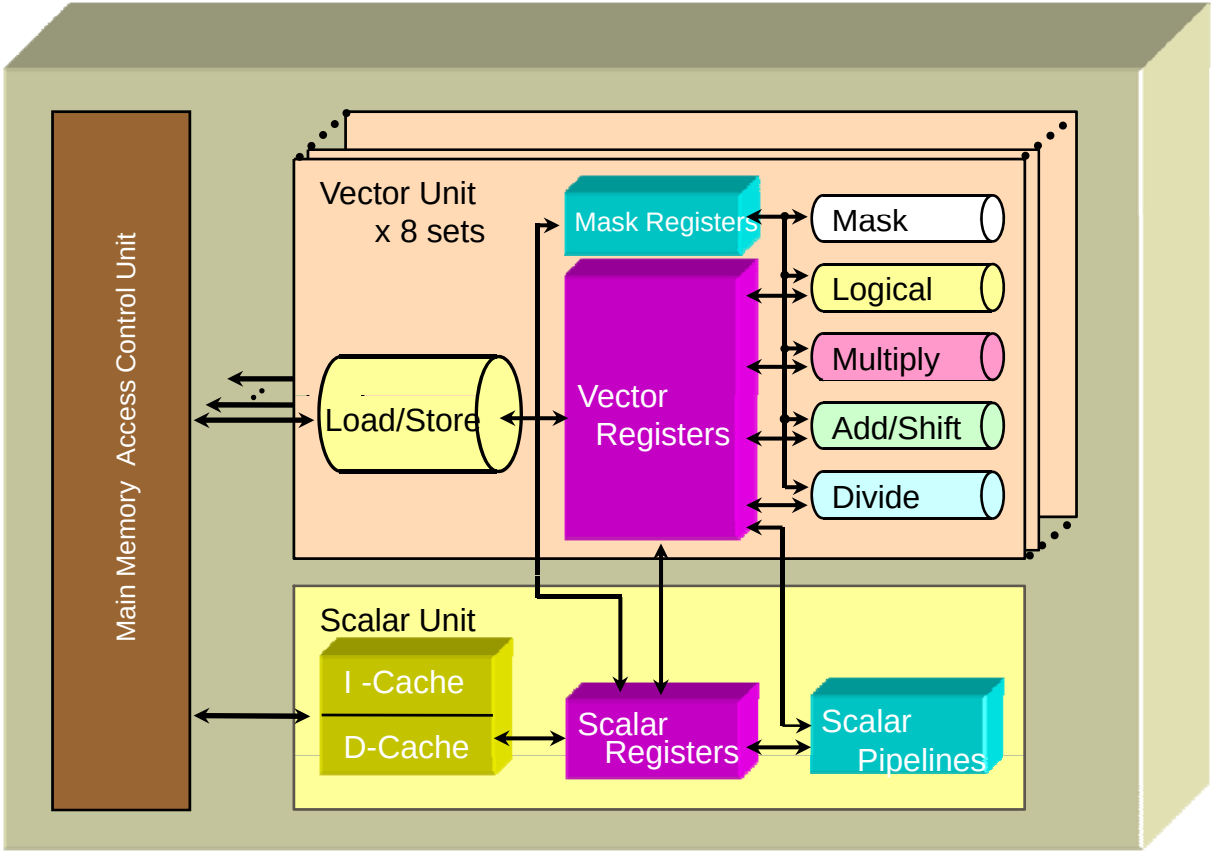
- Total number of APs : 5120
- Total number of PNs : 640

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



Arithmetic Processor configuration

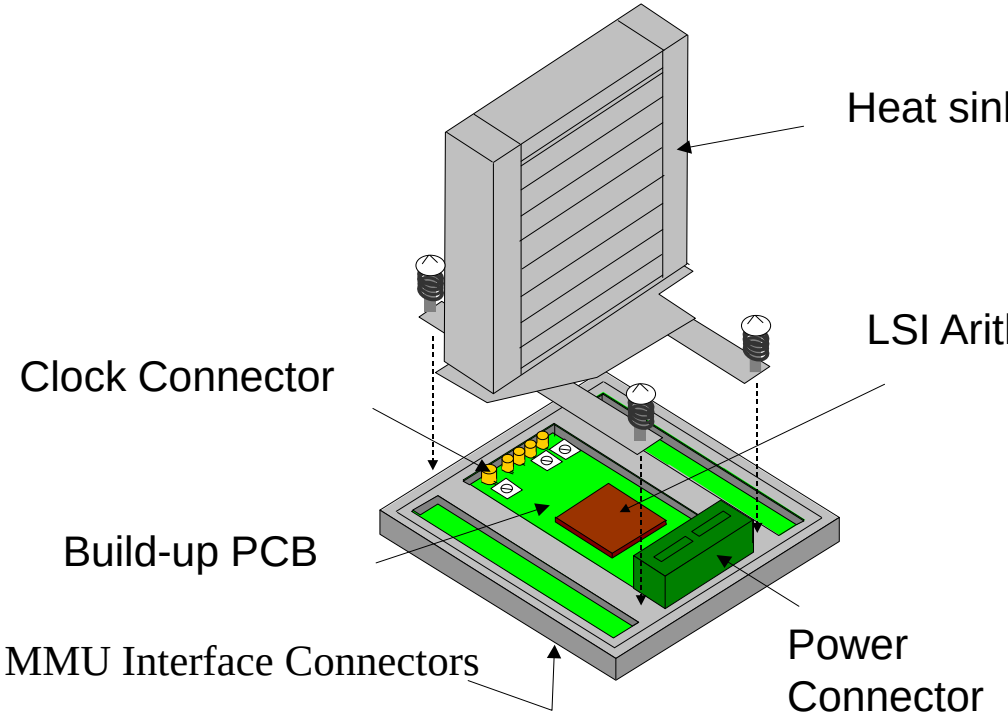
- Scalar Unit (SU)
  - 4-way superscalar
  - 128 scalar registers
  - 64KB Instruction cache
  - 64KB data cache
  - DRAM developed for ES
- 8 units of vector pipelines(VU)
  - 6 types of operation pipeline
  - 144KB vector registers
  - 256bit x 17 vector mask registers
- Main memory access contorl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



One Chip LSI: 8Gflops

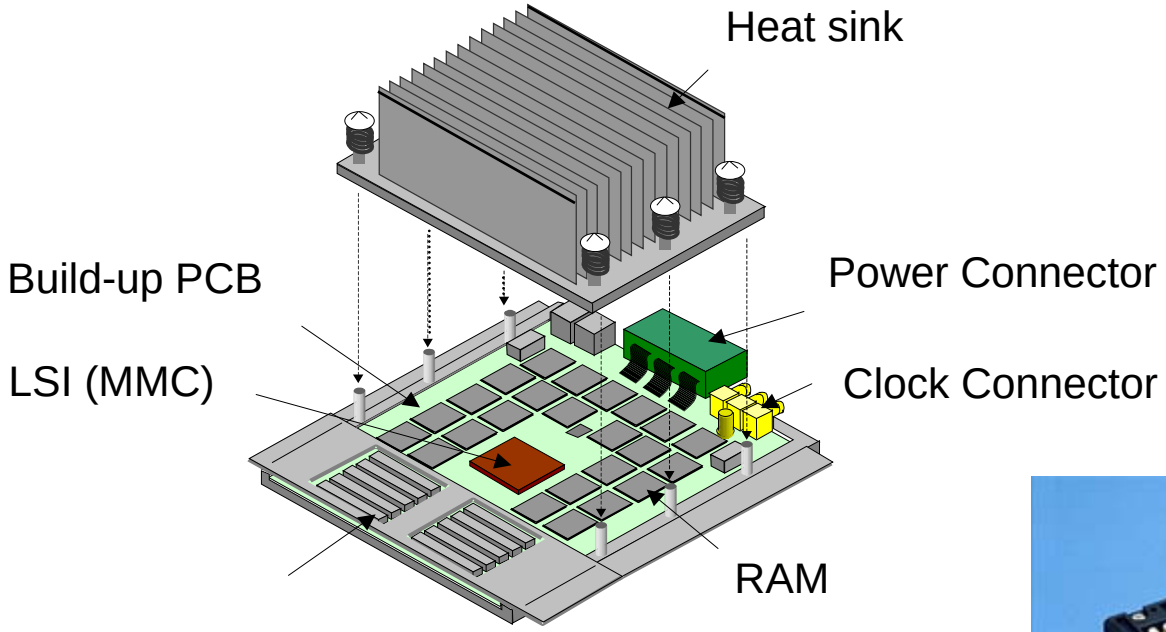
- 0.15μm CMOS LSI
  + Cu interconnection
- 20.79 mm x 20.79 mm
- 60 million transistors
- More than 5000 pins
- 500MHz

Arithmetic Processor Package

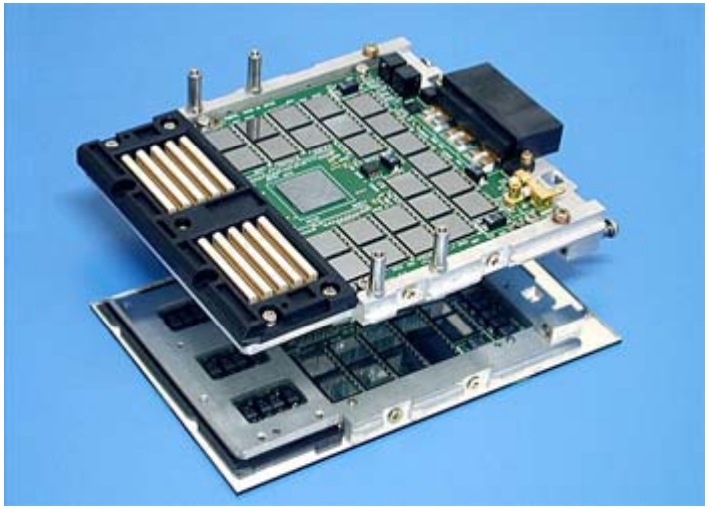


(115mm x 139mm)

Main Memory Unit Package

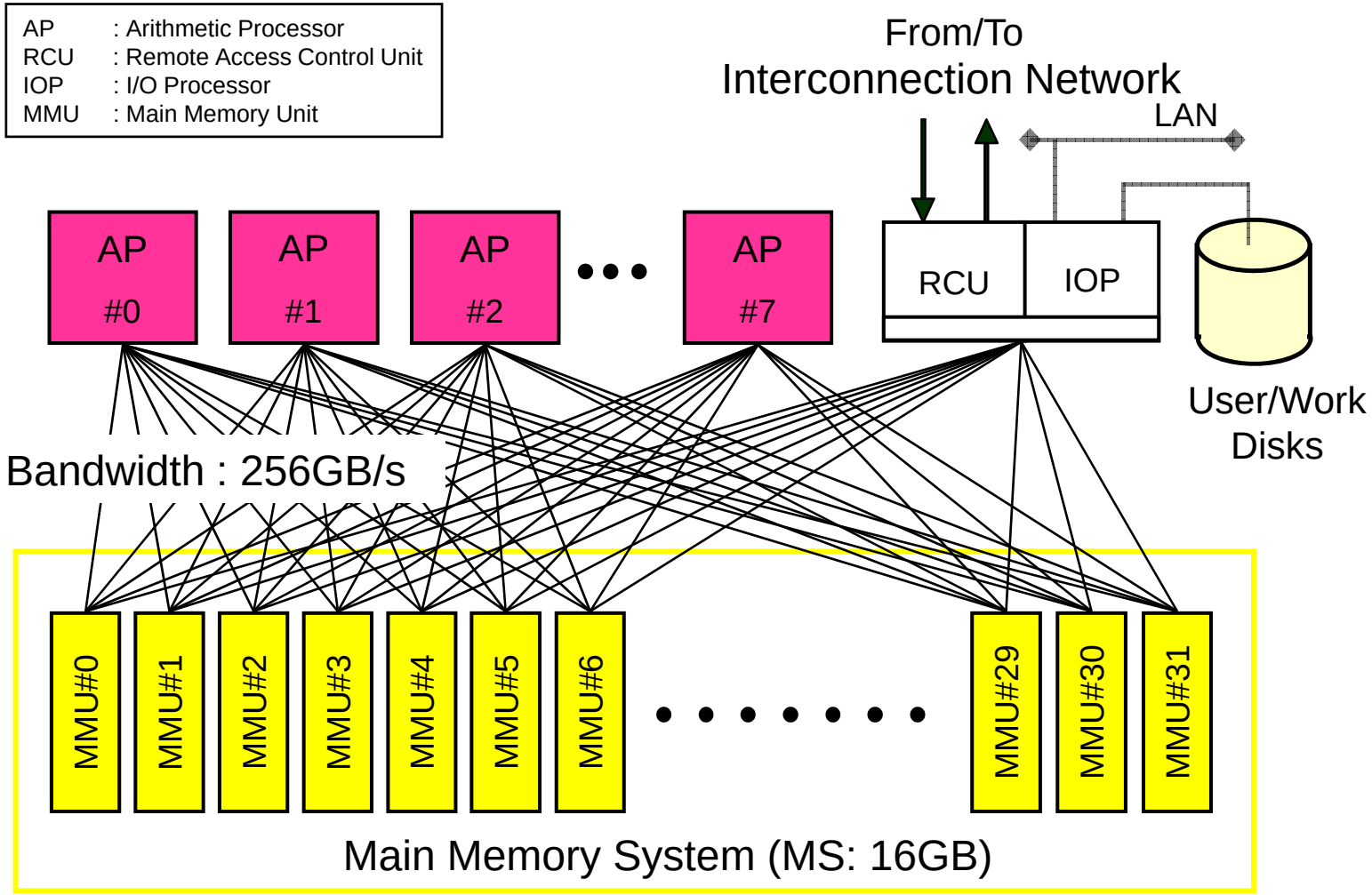


AP/RCU Interface Connectors



(125mm x 147mm)

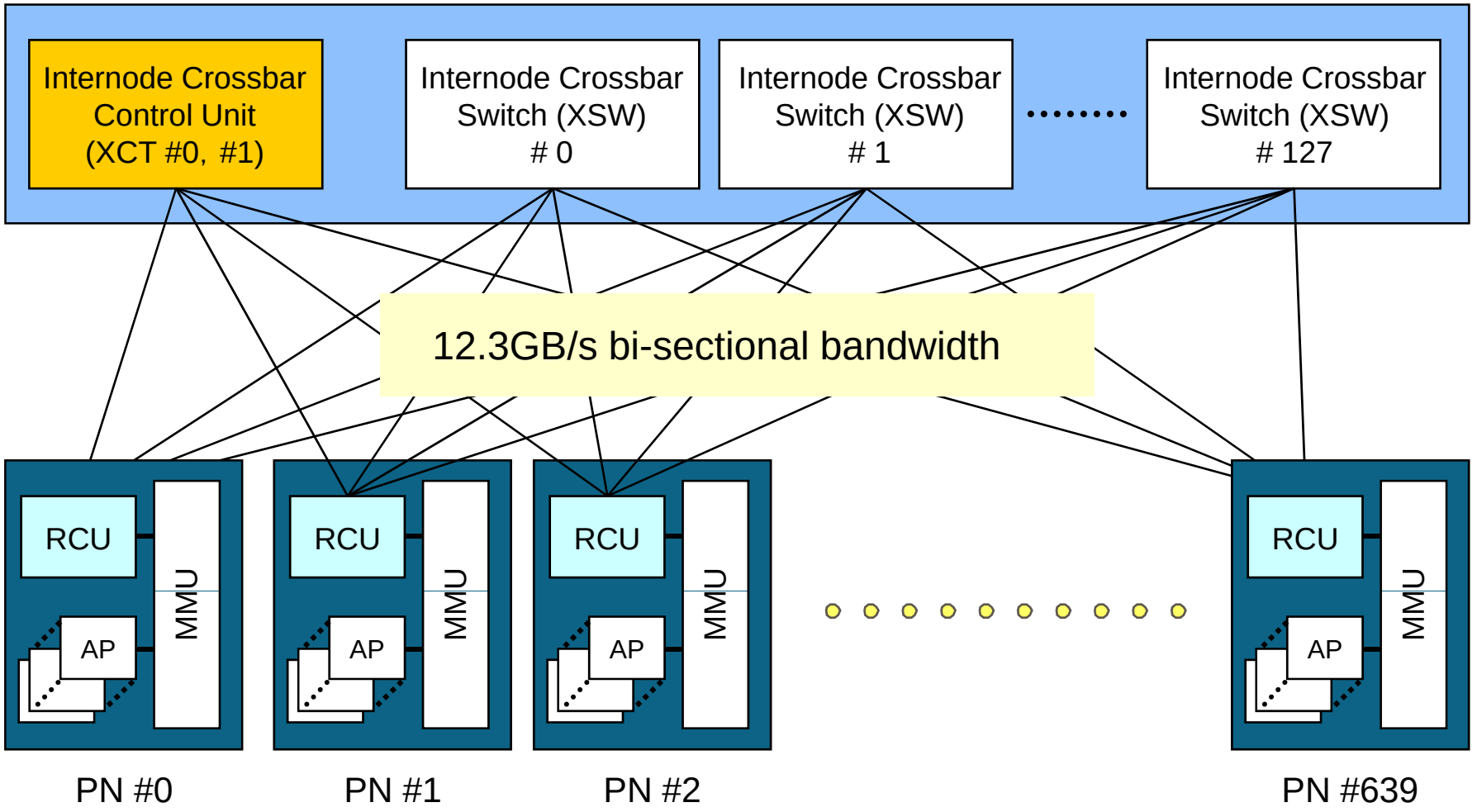
Processor Node configuration



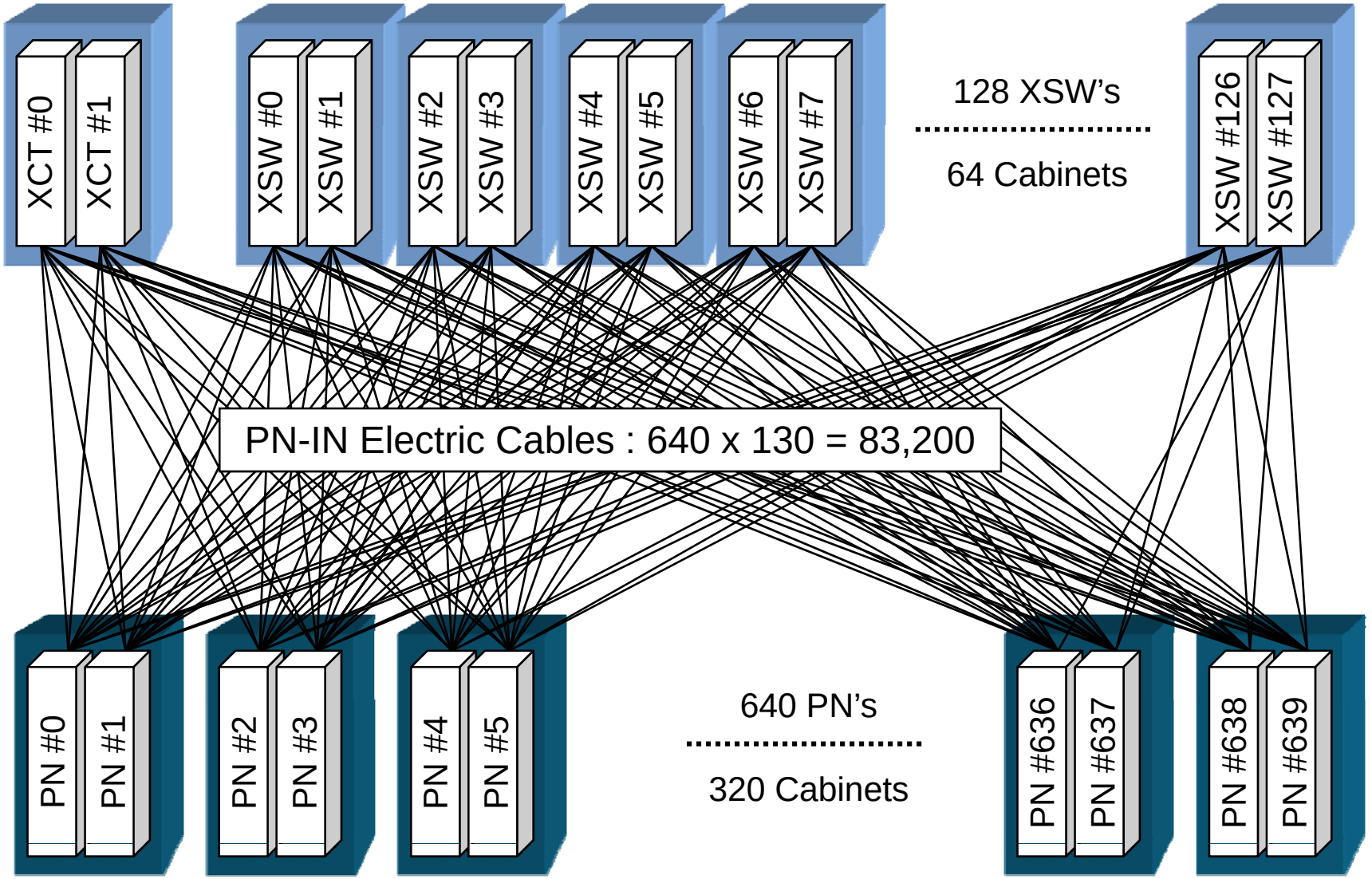
- 128M bit DRAM developed for ES (24nsec bank cycle time)
- 2048 banks

Interconnection Network (IN)

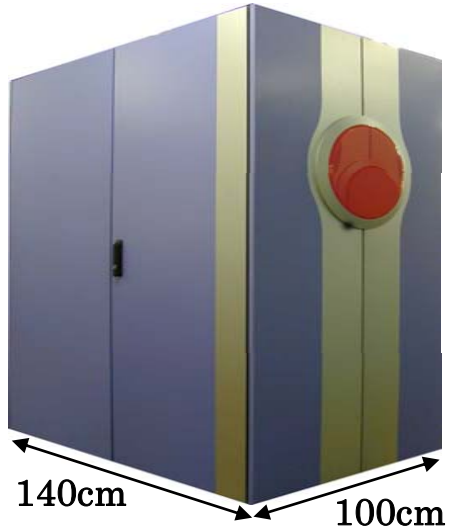
- 640 x 640 full crossbar switch
- 2 XCT's and 128 XSW's
  - XCT : Coordination of data transfer through XSW's



Connection between Cabinets

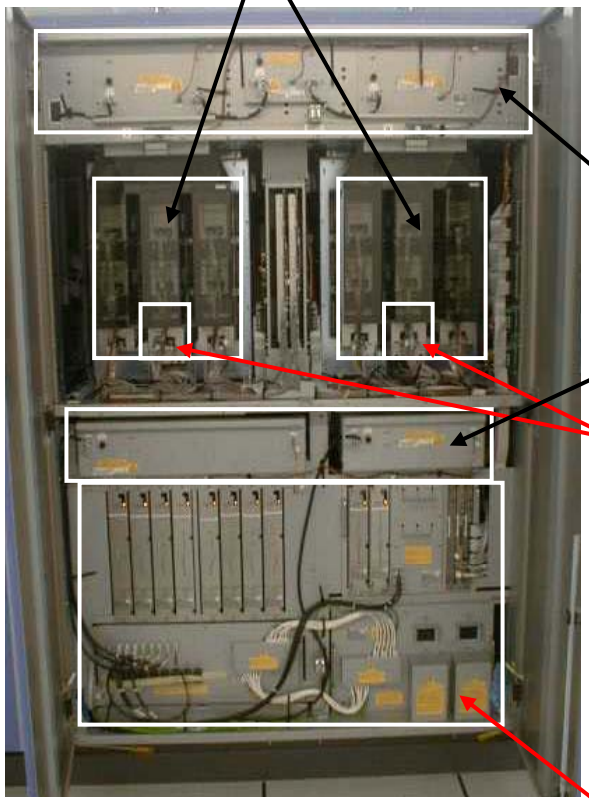


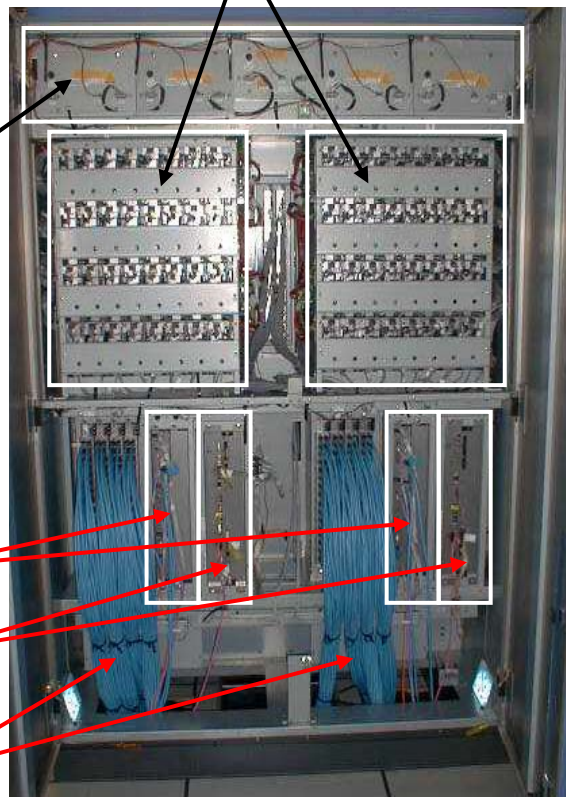
Processor-Node Cabinet (two nodes in a cabinet)



Arithmetic Processor (AP)

Main memory System (MS)





Cooling Fan

Remote access Control Unit (RCU)

Input/Output Processor (IOP)

Diagnostic Processor (DGP)

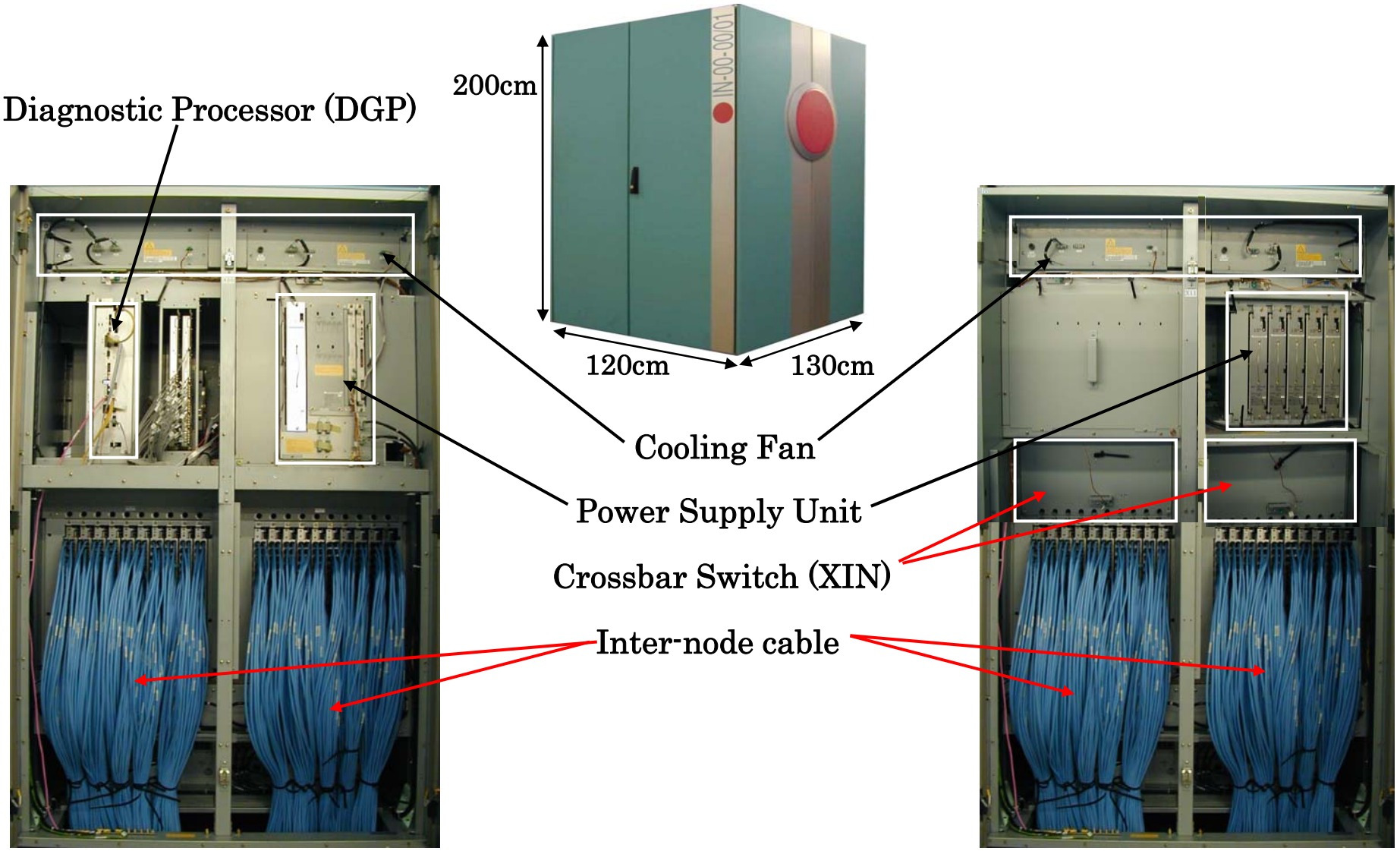
Inter-node cable

Power Supply Unit

Front

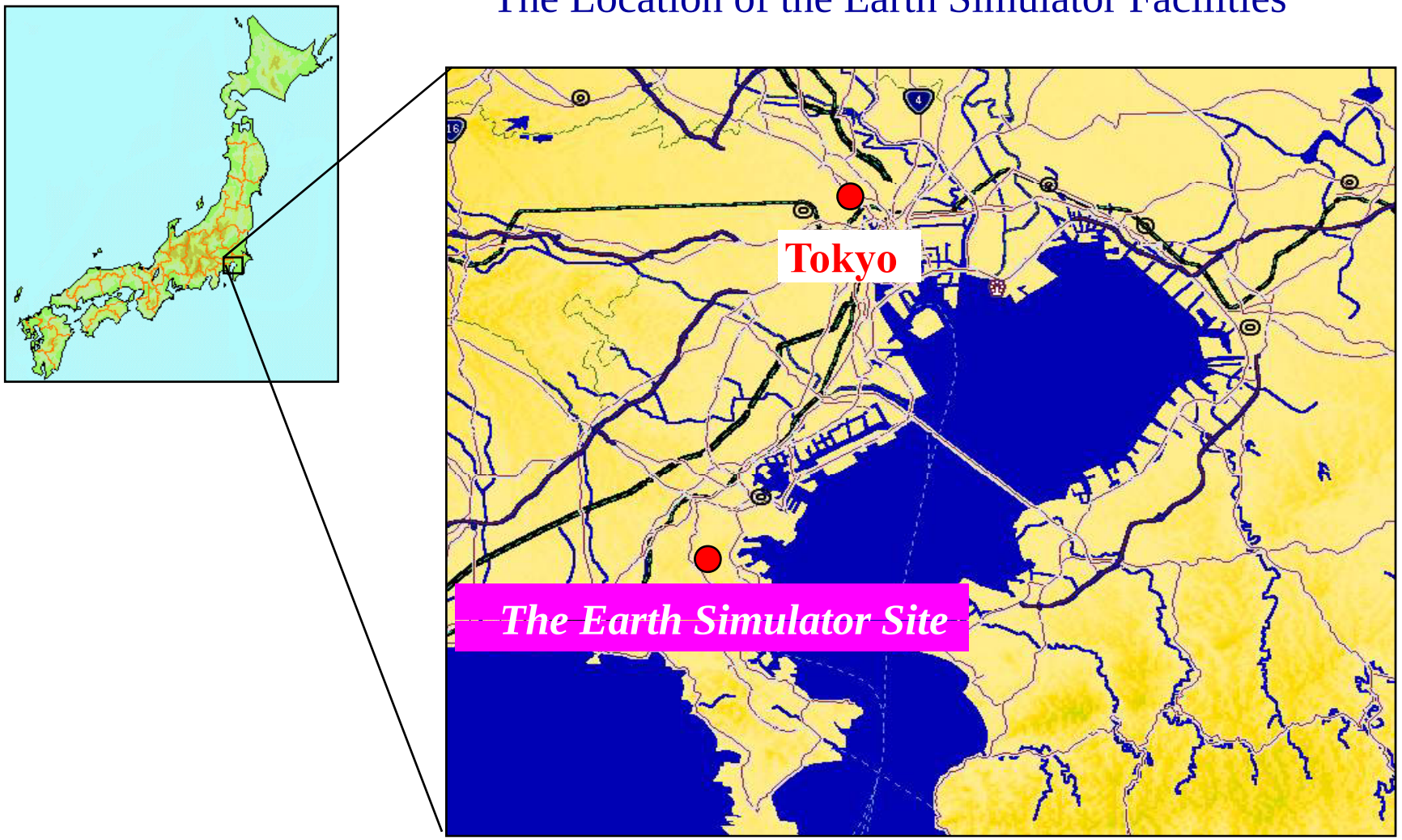
Rear

XSW Cabinet (two XSW's in a cabinet)



Installation of hardware

The Location of the Earth Simulator Facilities



The Earth Simulator Center

Yokohama Institute For Earth Science

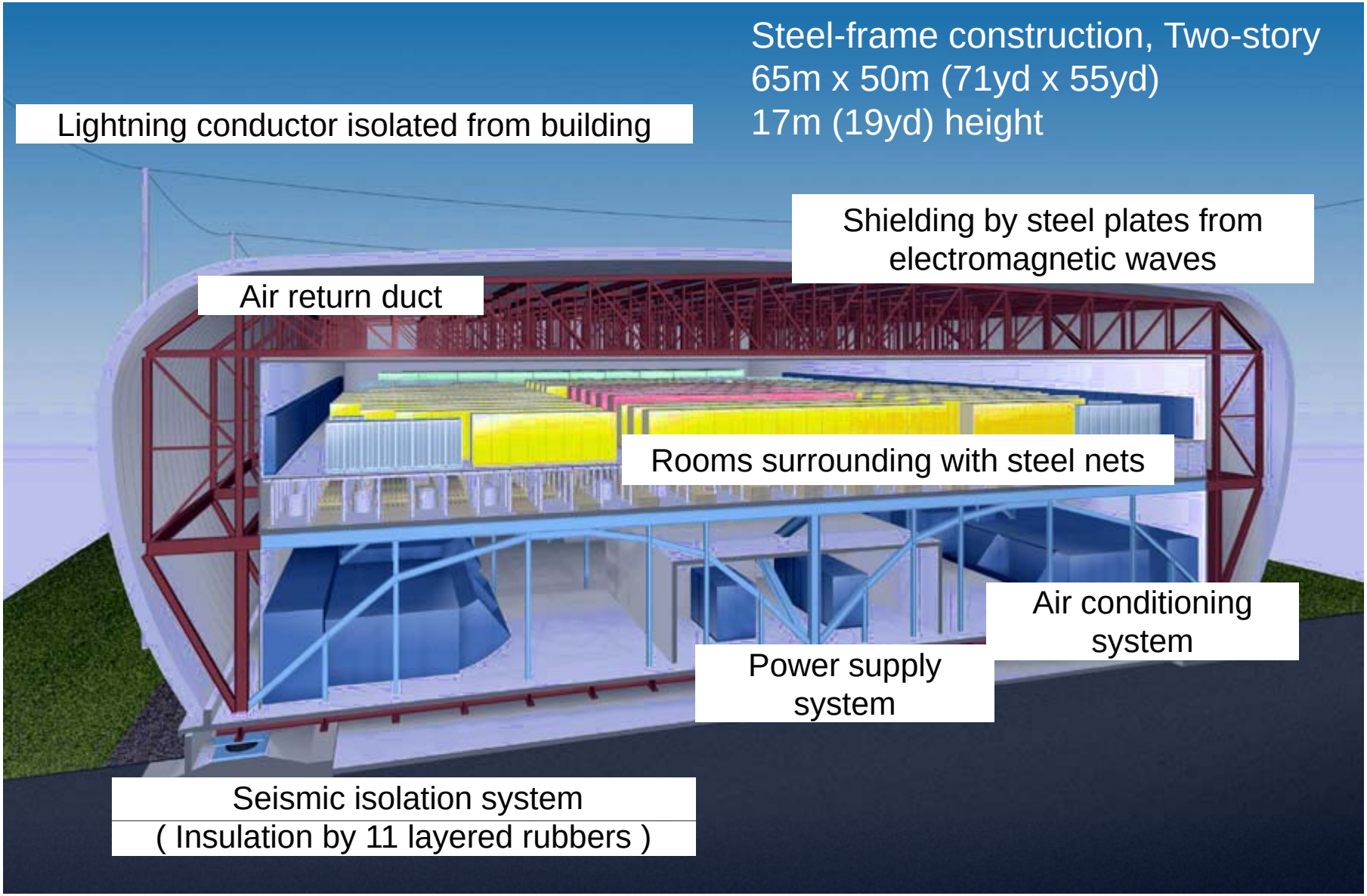
Japan Agency for Marine-Earth Science and Technology

3173-25 Showa-machi, Kanazawa-ku

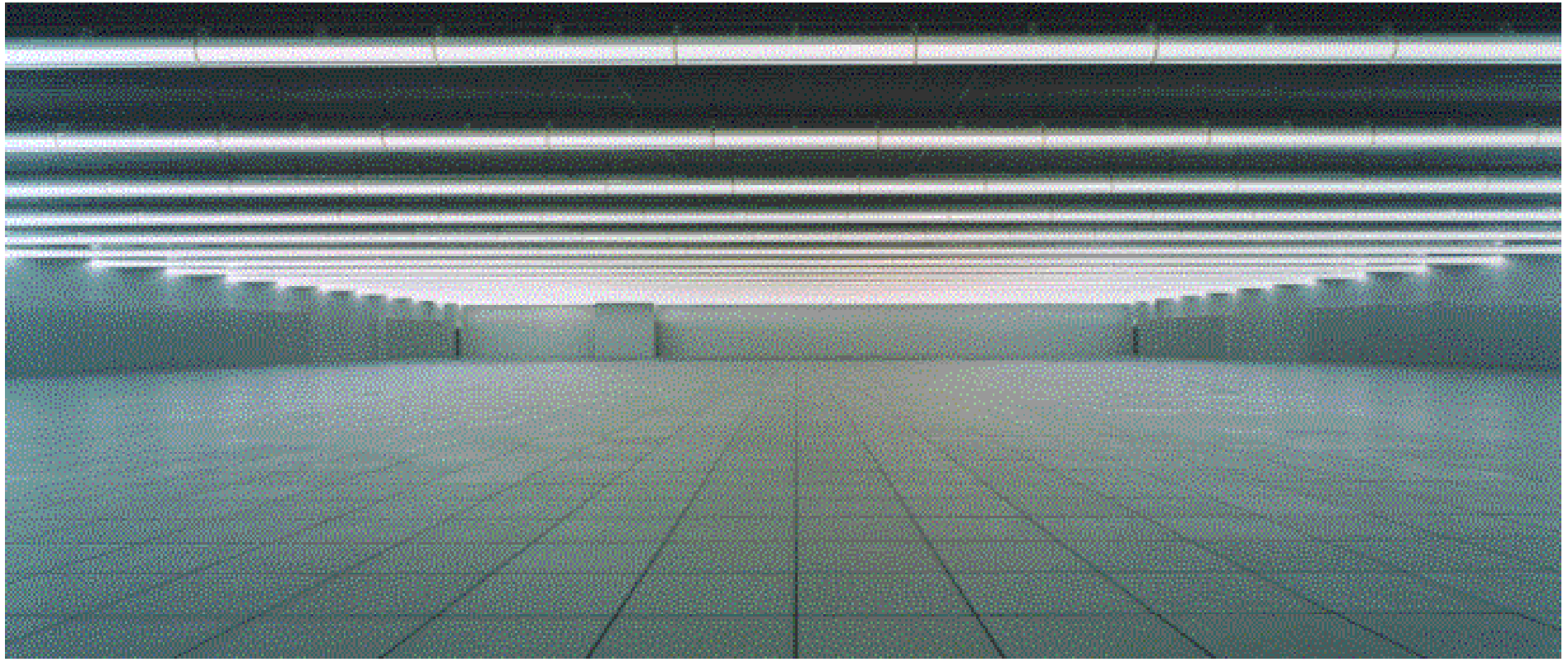
Yokohama-city, 236-0001 Japan

Earth Simulator Building

Features of the 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

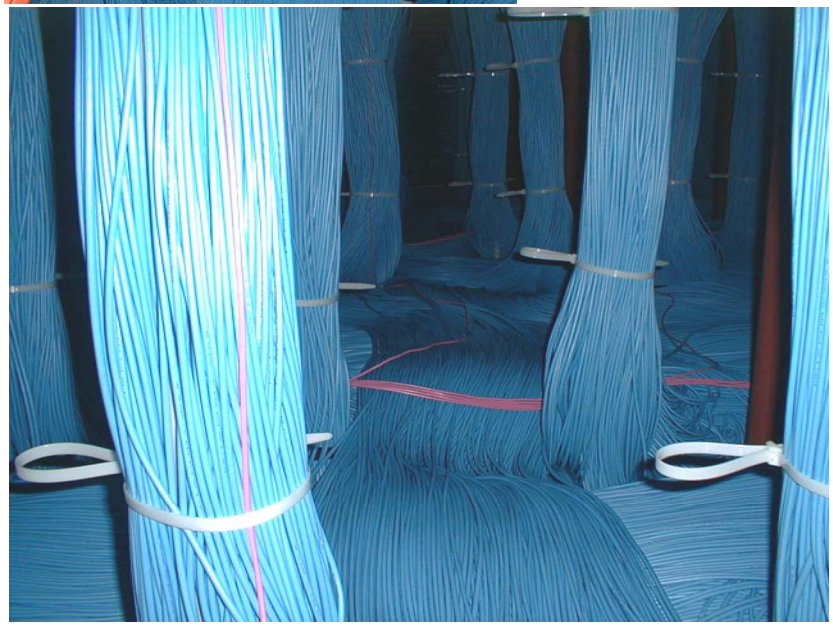
Electric Cables Connecting Cabinets

Number of cables: 83,200

Length : 10 to 40 m

Total length : 2,400km

Total weight: 140 t



The Earth Simulator at Completion

System Enhancement in FY2003

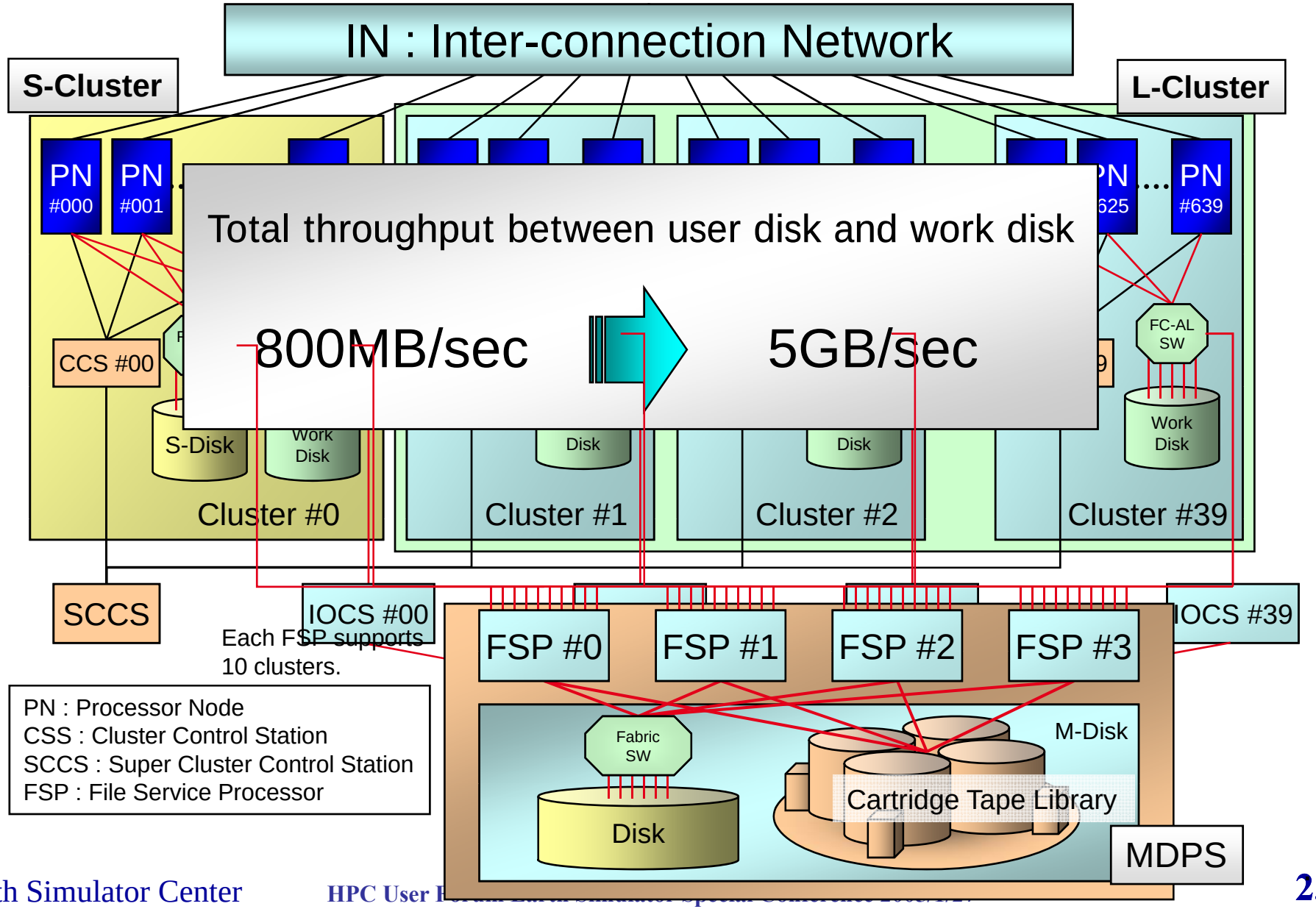
Introducing NQS II and MDPS

- Mass Data Processing System was installed.
  - The capacity of storage system has increased.
  - Accessibility of files in MT has improved.
- Job Manager was replaced from GM-NQS(JS) to NQS II
  - To utilize MDPS efficiently and
  - To improve node utilization and maintenancibility.
- This system transition was made smoothly using about 4 months without stopping ES.

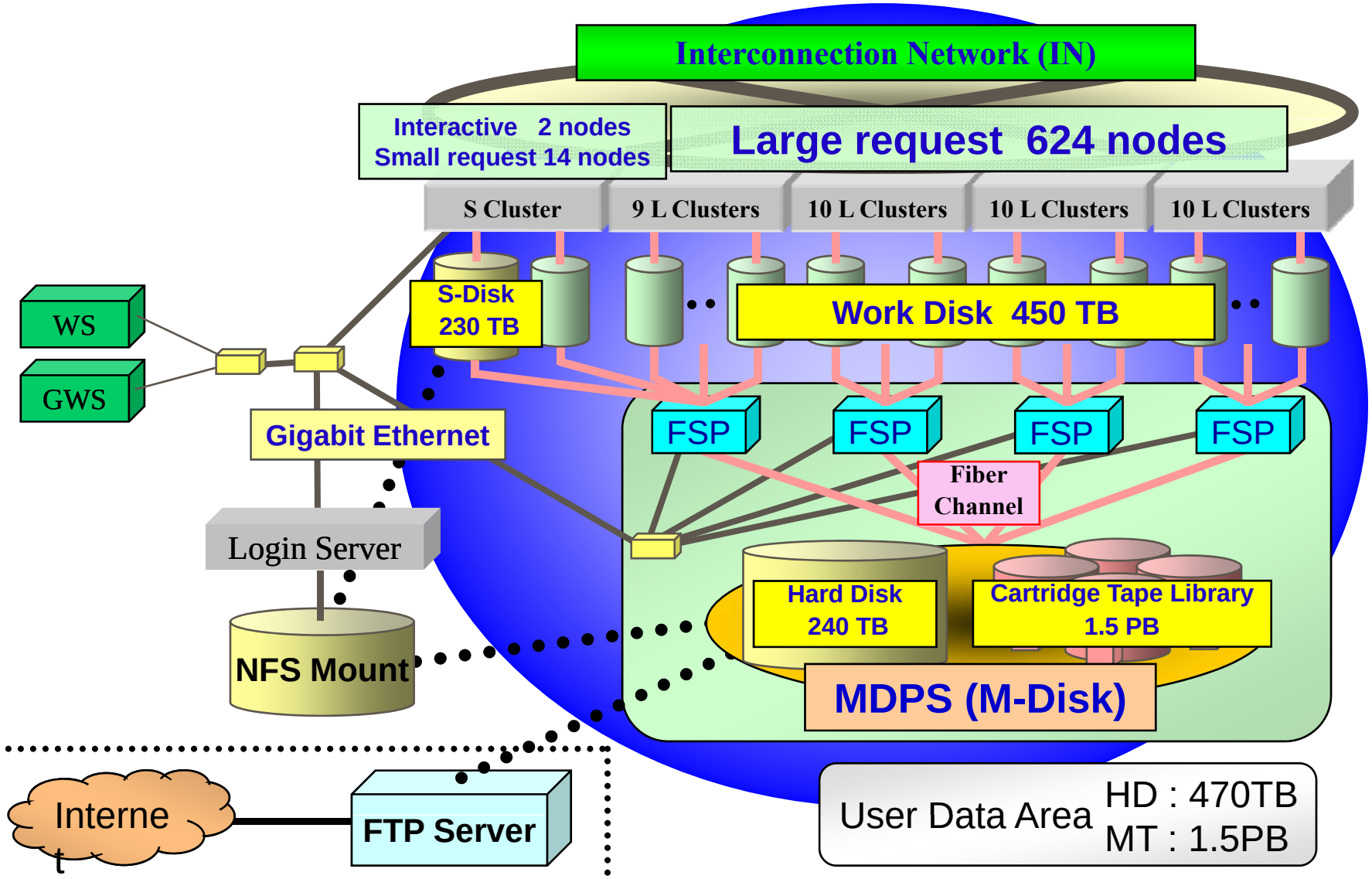
Mass Data Processing System (MDPS)

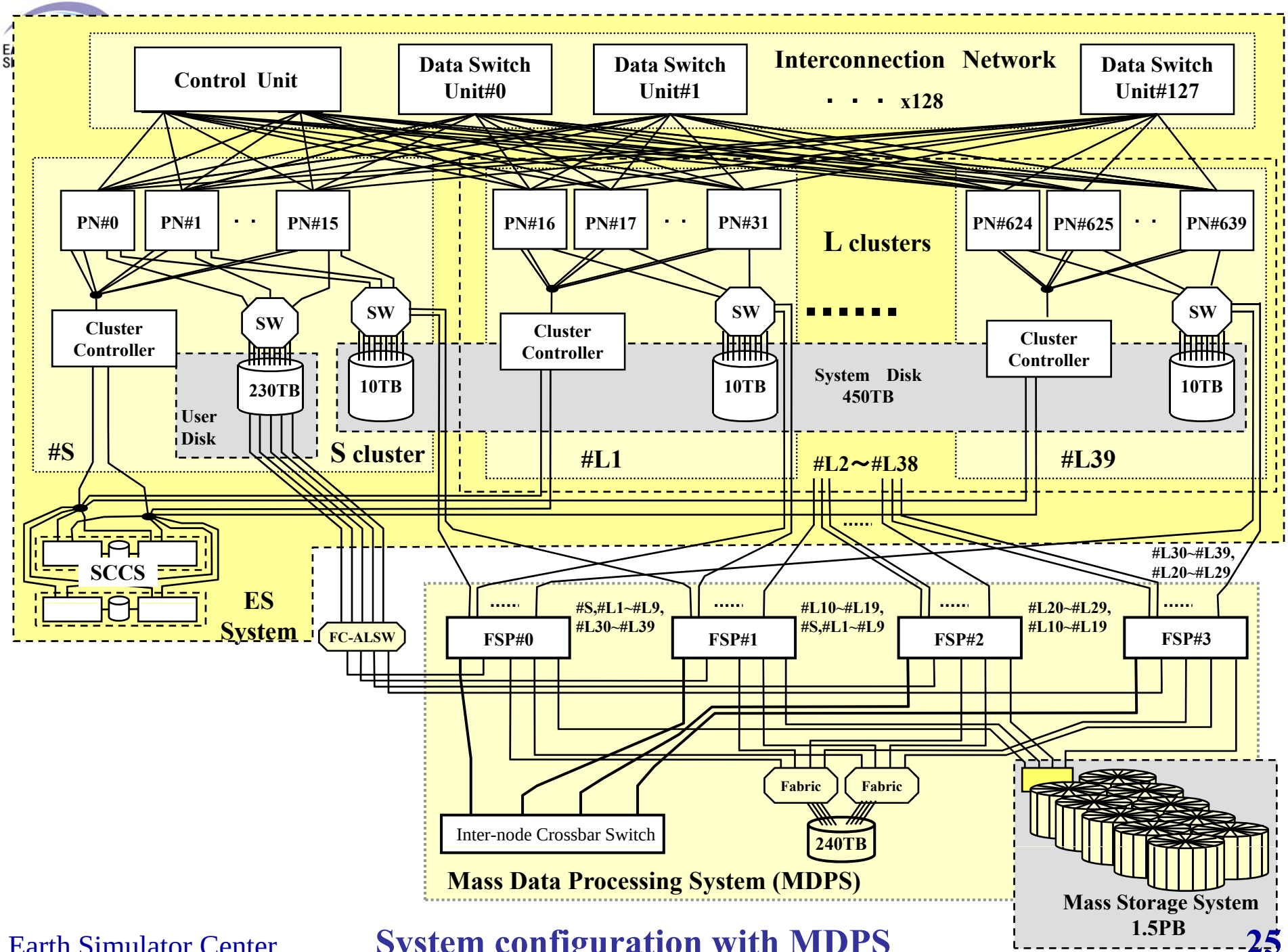
- Hierarchical Storage Management System
  - Consists of Hard Disk System and Cartridge Tape Library System
  - Capacity : HD is 240TB and MT is 1.5PB
- 4 File Service Processors manage files
  - Each FSP supports 160 nodes.
  - FSP transfers data between work disk and user virtual disk files which consist of Hard Disk System and Cartridge Tape Library System of MDPS.
  - FSP manages data size of files in Hard Disk System, and migrates them into / recalls them from Cartridge Tape Library System automatically.
- Before introducing MDPS, user should handle files in Cartridge Tape Library System directly, but MDPS enables user to access virtual disk files in MDPS as ordinal unix disk files.

Introducing Mass Data Processing System



Connection among Peripherals (After introducing MDPS system)





System configuration with MDPS

Comparison of System Characteristics

NITRD report

Selected System Characteristics

	Earth Simulator (NEC)	ASCI Q (HP ES45)	ASCI White (IBM SP3)	MCR (Dual Xeon)	Cray X1 (Cray)
Year of Introduction	2002	2003	2000	2002	2003
Node Architecture	Vector SMP	Alpha micro SMP	Power 3 micro SMP	Xeon micro SMP	Vector SMP
System Topology	NEC single-stage Crossbar	Quadrics QsNet Fat-tree	IBM Omega network	Quadrics QsNet Fat-tree	2D Torus Interconnect
Number of Nodes	640	3072 (Total)	512	1152	
Processors - per node	8	4	16	2	4
- system total	5120	12288	8192	2304	
Processor Speed	500 MHz	1.25 GHz	375 MHz	2.4 GHz	800 MHz
Peak Speed - per processor	8 Gflops	2.5 Gflops	1.5 Gflops	4.8 Gflops	12.8 Gflops
- per node	64 Gflops	10 Gflops	24 Gflops	9.6 Gflops	51.2 Gflops
- system total	40 Gflops	30 Gflops	12 Tflops	10.8 Tflops	
Memory - per node	16 GB	16 GB	16 GB	16 GB	8-64 GB
- per processor	2 GB	4 GB	1 GB	2 GB	2-16 GB
- system total	10.24 TB	48 TB	8 TB	4.6 TB	
Memory Bandwidth (peak)					
- L1 Cache	N/A	20 GB/s	5 GB/s	20 GB/s	76.8 GB/s
- L2 Cache	N/A	13 GB/s	2 GB/s	1.5 GB/s	
Main (per processor)	32 GB/s	2 GB/s	1 GB/s	2 GB/s	34.1 GB/s
Inter-node MPI					
- Latency	8.6 µsec	5 µsec	18 µsec	4.75 µsec	
- Bandwidth	11.8 GB/s	300 MB/s	500 MB/s	315 MB/s	12.8 GB/s
Bytes/flop to main memory	4	0.8	0.67	0.4	2.66
Bytes/flop interconnect	1.5	0.12	0.33	0.07	1

Most of this data is from Kerbyson, Hoisie, Wasserman; LANL; unpublished

15

http://www.krellinst.org/csgf/conf/2003/presentations/nelson.pdf

NITRD report with a slight modification

NITRD report: Selected System Characteristics (with slight modification)

National Coordination Office for Information Technology Research and Development

	Earth Simulator (NEC)	*ASCI Q* (HP ES45)	*ASCI White* (IBM SP3)	*MCR* (Dual Xeon)	*Cray X1* (Cray)
Year of Introduction	2002	2003	2000	2002	2003
Node Architecture	Vector	Alpha micro	Power 3 micro	Xeon micro	Vector
	SMP	SMP	SMP	SMP	SMP
System Toplogy	NEC	Quadrics Qsnet	IBM	Quadrics Qsnet	
	Single-stage crossbar	Fat-tree	Omega network	Fat-tree	hyper Cube+ 3D Torus
# of Node	640	3072 (Total)	512	1152	1024
Processors - per node	8	4	16	2	4
- system total	5120	12288	8192	2304	4096
Processor speed	500 MHz	1.25 GHz	375 MHz	2.4 GHz	800 MHz
Peak Speed - per processor	8 Gflops	2.5 Gflops	1.5 Gflops	4.8 Gflops	12.8 Gflops
- per node	64 Gflops	10 Gflops	24 Gflops	9.6 Gflops	51.2 Gglops
- system total	40 Tflops	30 Tflops	12 Tflops	10.8 Tflops	52.4 Tflops
Memory Bandwidth (peak)					
- L1 Cache	N/A	20 GB/s	5 GB/s	20 GB/s	76.8 GB/s
- L2 Cache	N/A	13 GB/s	2 GB/s	1.5 GB/s	
Main (per processor)	32GB/s	2GB/s	1GB/s	2GB/s	34.1GB/s
Inter-node MPI - Latency	5.6µsec	5µsec	18µsec	4.75µsec	8.6µsec
- Bandwidth	11.8GB/s	300MB/s	500MB/s	315MB/s	11.9GB/s
Intra-node MPI - Latency	1.4µsec				8.2µsec
- Bandwidth	14.8GB/s				13.9GB/s
Bytes/flop to main memory	4	0.8	0.67	0.4	2.66
Bytes/flop interconnect	1.5	0.12	0.33	0.07	1

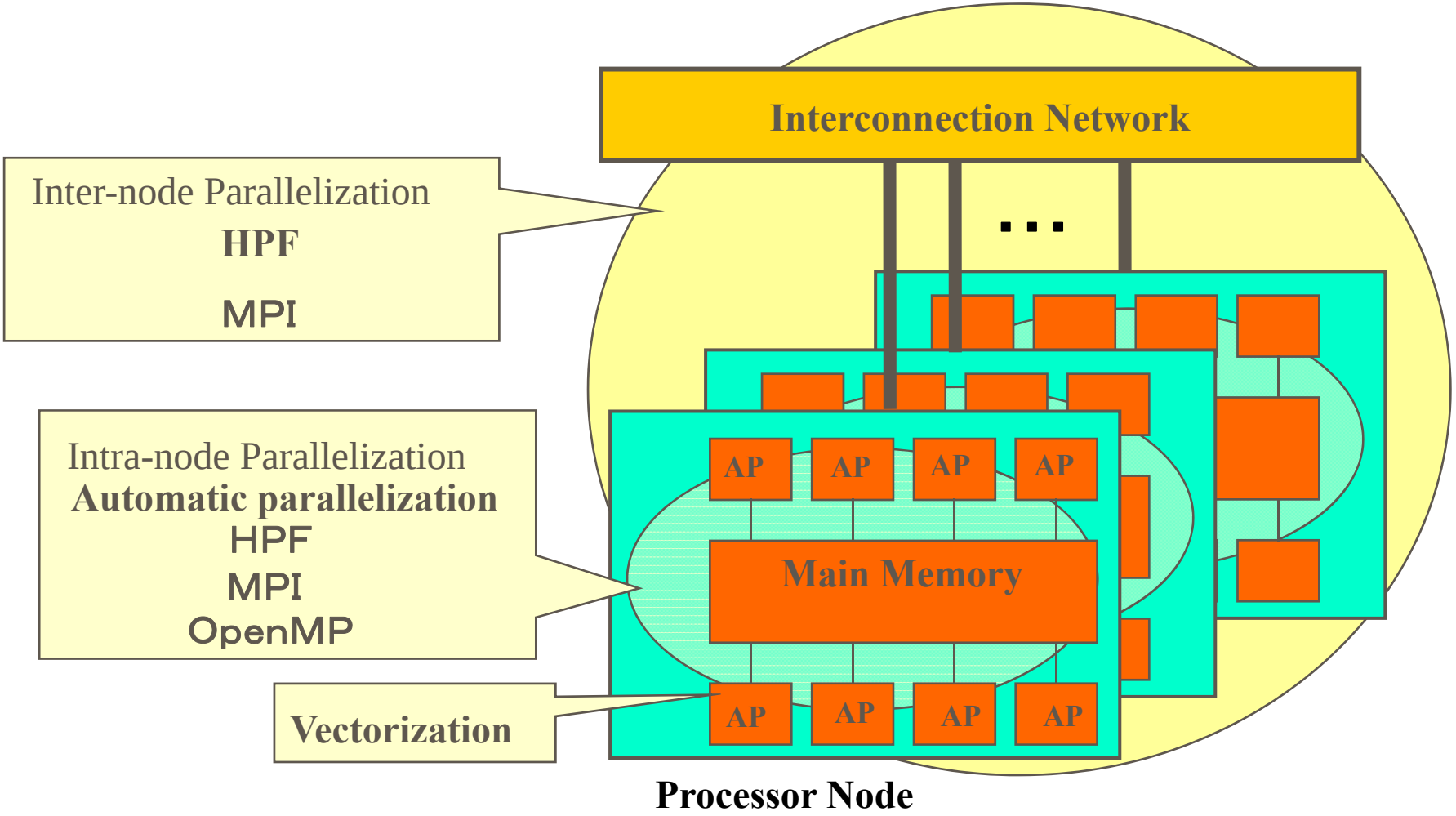
Software & Achieved Performance

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
- NQS II
  - Extension of NQS
  - Running on the SCCS
  - Allocating PN's and staging in/out the user data into/from system disk

These software have a good scalability up to 640 nodes.

Vecrorization and Parallelization



Clusters

- 640 nodes are divided into 40 clusters
  - Each cluster has 16 nodes
  - Cluster structure is used only for operation
- S-Cluster (1 cluster)
  - Interactive nodes : 2 nodes

    Interactive process environment (compiling, debugging)
  - S-system : 14 nodes

    Small-scale batch request environment
- L-Cluster (39 clusters)
  - L-system : 624 nodes

    Large-scale batch request environment

    Single system image

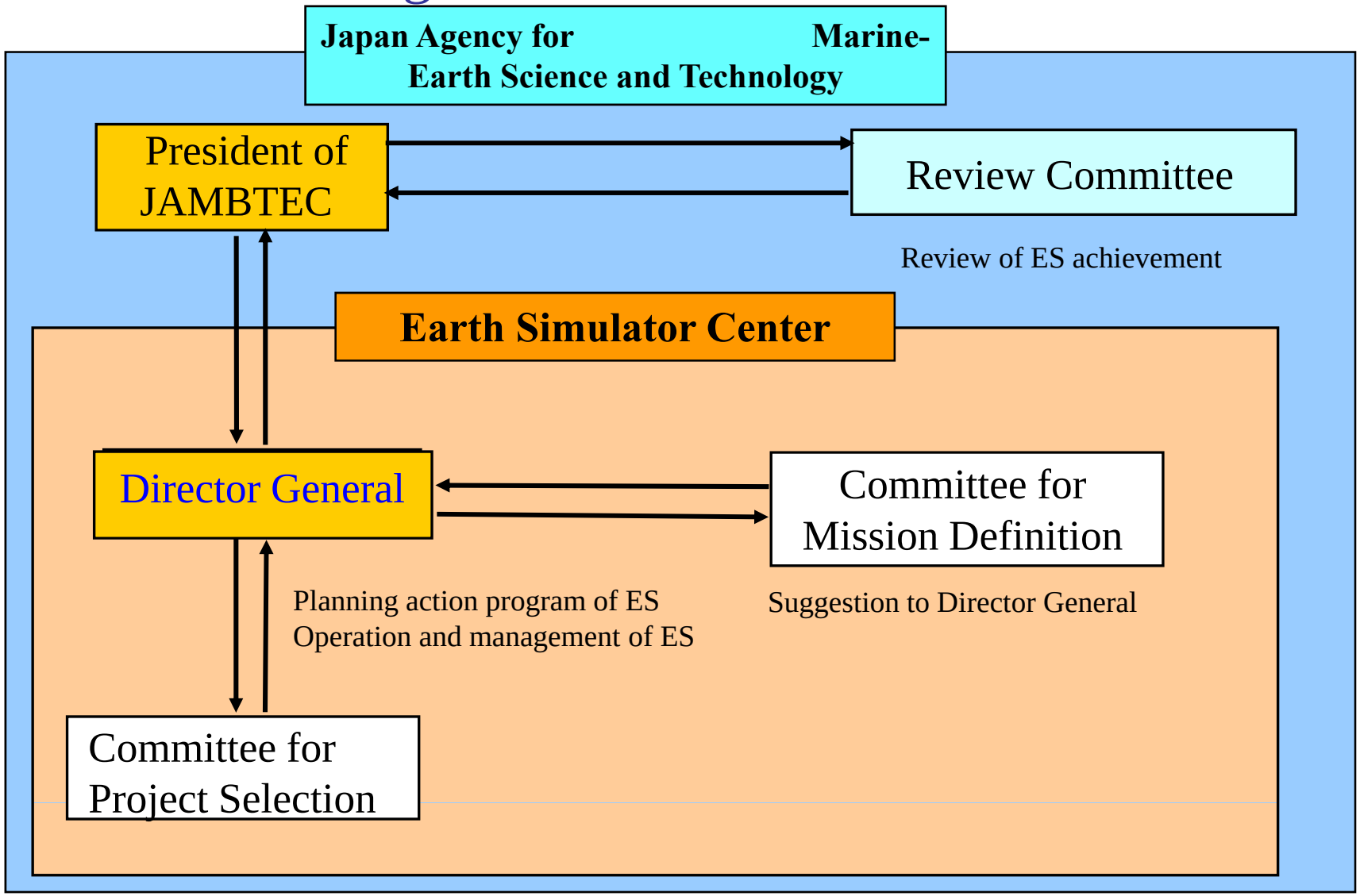
Measured Computational Performance

(in early stage in 2002)

Application	Language processors	Parallelization method	# of node used	TFLOPS	Ratio to Peak
Linpack HPC	Fortran90, Assembler	MPI, microtask	640	35.86	88%
AFES(Atmospheric general circulation model For ES)	Fortrn90, Assembler	MPI, microtask, Auto	640	26.58	65%
OFES(MOM3 optimized for ES)	Fortran90	MPI, Auto	188	2.75	23%
TRANS7(Direct Numerical Simulation of Turbulence by a Fourier Spectral Method For ES)	Fortran90	MPI, microtask	512	16.40	50%
PFES(POM parallelized with HPF/ES)	HPF	HPF	376	9.85	41%
Impact3D(Three-dimensional Fluid Simulation for Fusion Science with HPF/ES)	HPF	HPF	512	14.90	45%

Operation of the Earth Simulator

Management of the Earth Simulator



Activities of Earth Simulator Center

Annual schedule of Earth Simulator Center

- January Annual meeting for the Earth Simulator research projects
- February Public project recruitment for next fiscal year
- March Project Selection for next fiscal year
- April Starting projects for new fiscal year
- Summer (or Autumn) Earth Simulator Center Symposium

Periodicals from Earth Simulator Center

- Annual Report of the Earth Simulator Center
- Journal of the Earth Simulator ……twice per year
- Earth Simulator News (in Japanese) ……...twice per year

Scheduled Maintenance of ES system

- Switching procedure at the end of fiscal year
- Maintenance of Hardware

  1 L-cluster (16 processor nodes)

  …weekly (every Wednesday)

  whole system including S-cluster, IN, auxiliary UNIX servers, and network devices

  …bimonthly (end of Jan., March, May, etc.)
- Maintenance of Software

  … bimonthly (same timing as whole system maintenance of Hardware)
- Maintenance of all electric equipments in Yokohama campus

  … once per year

Operating Policy of the Earth Simulator

attaching Importance to the Performance for the Parallel Programs using many Processor Nodes

- Each PN allocated for a multi-PN parallel program is to be monopolized by the program.
- Input/output for a multi-PN parallel program should be local (should be directed to the work disk allocated for each PN).
- Simple Restart feature is used for the execution of a multi-PN parallel program. Restart data is to be prepared by the user program.
- Degraded Operation at the failure of AP

  For efficient execution of a multi-PN parallel program, equal capability of each PN is vary important. So a PN which any of the AP's is in failure is to be removed from the Operation of the Earth Simulator.

Condition for running multiple node program

- Through NQS (as Batch job request)
- Job amount to submit at a time for a user ≦5
- Number of PN for a job ≦ 10

（Extendable to 512 by the application ）

- Wall clock time for a job ≦ 12(in hours)
- Number of PN × wall clock time (in hours)≦1536

Condition to extend the PN number for a JOB

- Vectorization ratio≧95%

  (vector operation ratio may be used)

- Parallelization efficiency≧50%

If a program needs T1 hours with 1 node,

and Tn hours with n nodes

Parallelization efficiency = (T1/Tn)/n

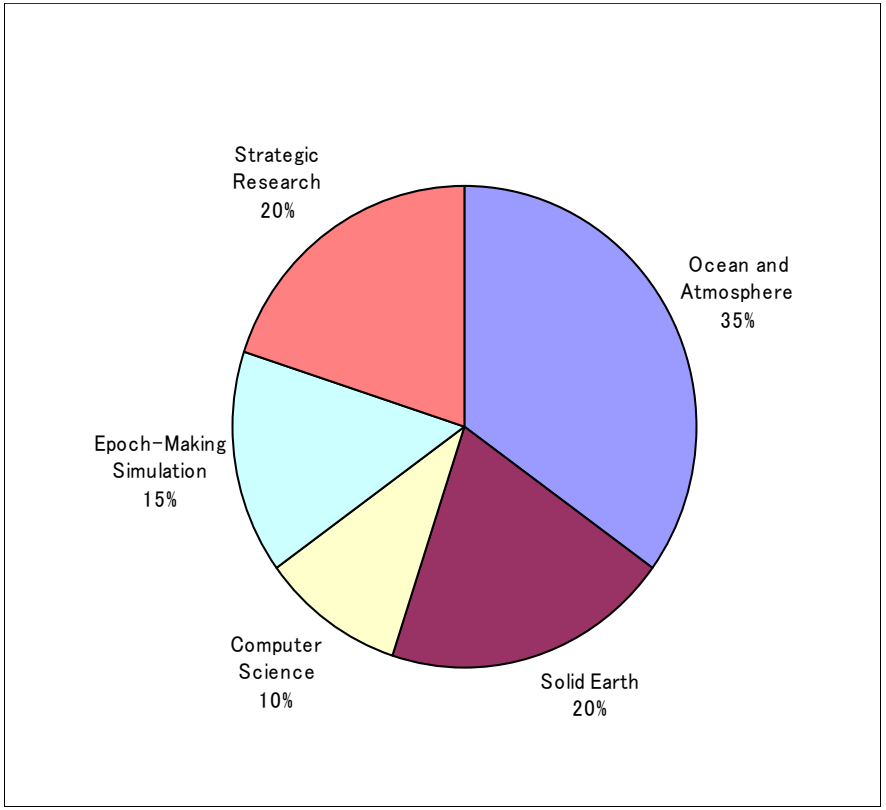
Parallelization ratio should be more than 99.9% for a program using more than 128 PN's to keep the parallelization efficiency as 50%, if the parallelization ratio is not affected by the number of PN used for the computation.

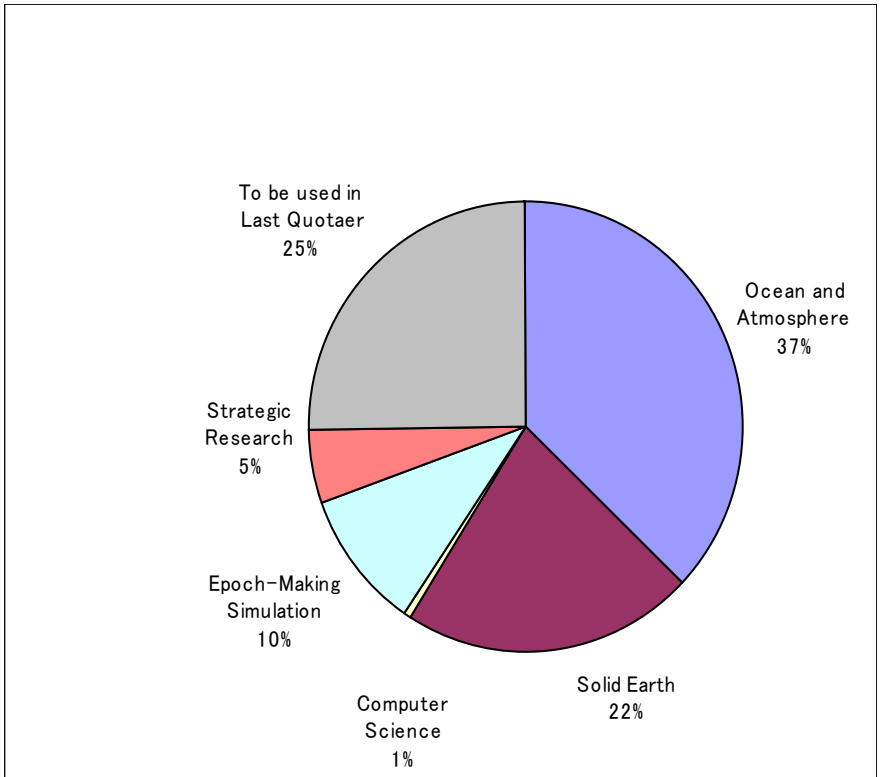
Statistics of the Operation

Selected Projects			
Research fields	# of project in FY2002	# of project in FY2003	# of project in FY2004
Ocean and Atmosphere	17	12	14
Solid Earth	8	9	9
Computer Science	4	2	2
Epoch-Making Simulation	11	11	12
Ordinary domestic prpject	40	34	37
(Strategic Research)	(3)	(7)	(11)

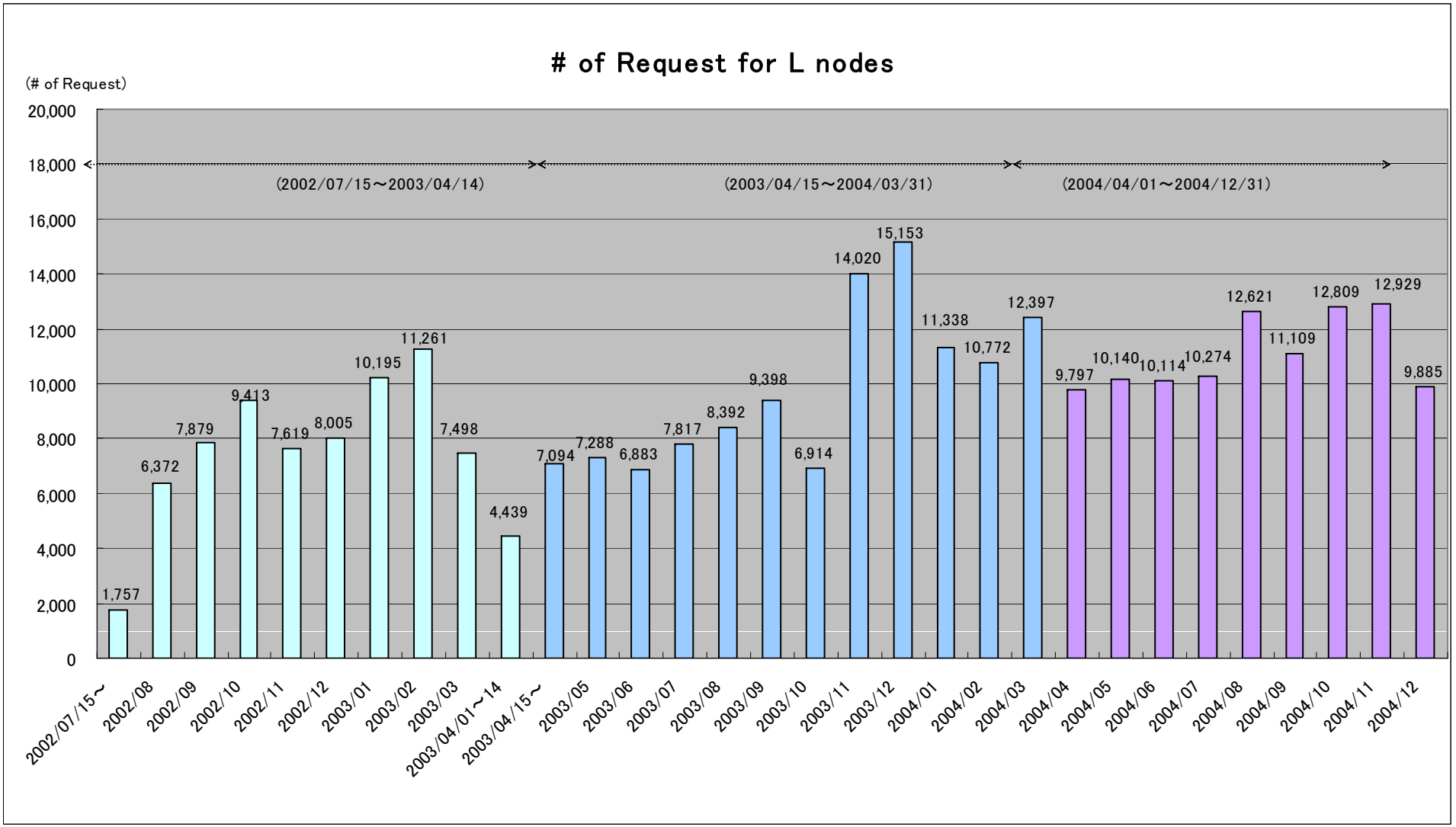
# of Users		
	# of User Organizations	**# of Users**
FY2002	93	480
FY2003	171	699
FY2004	187	758

Resource Allocation in FY2004 (Plan and Result)

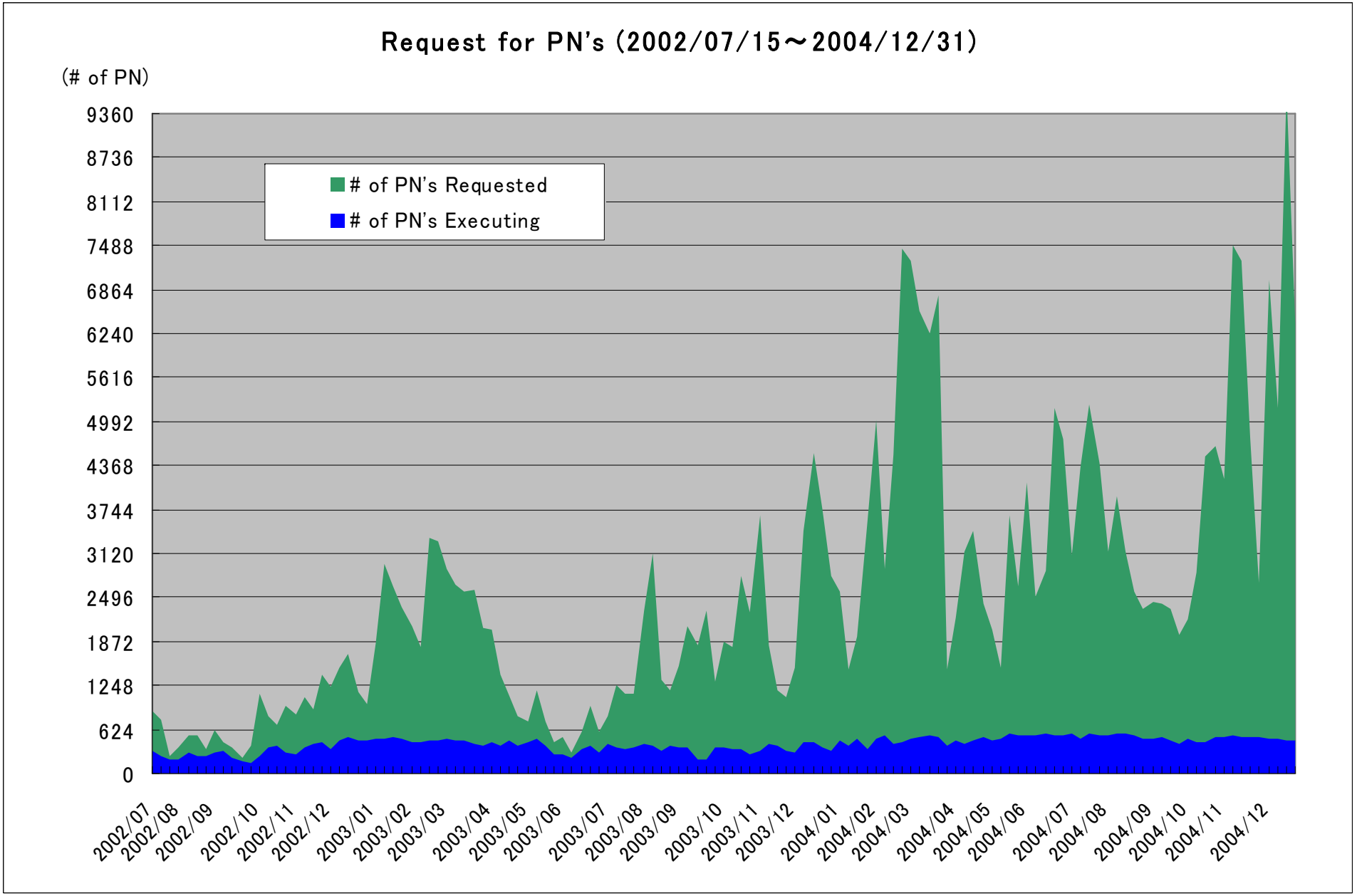




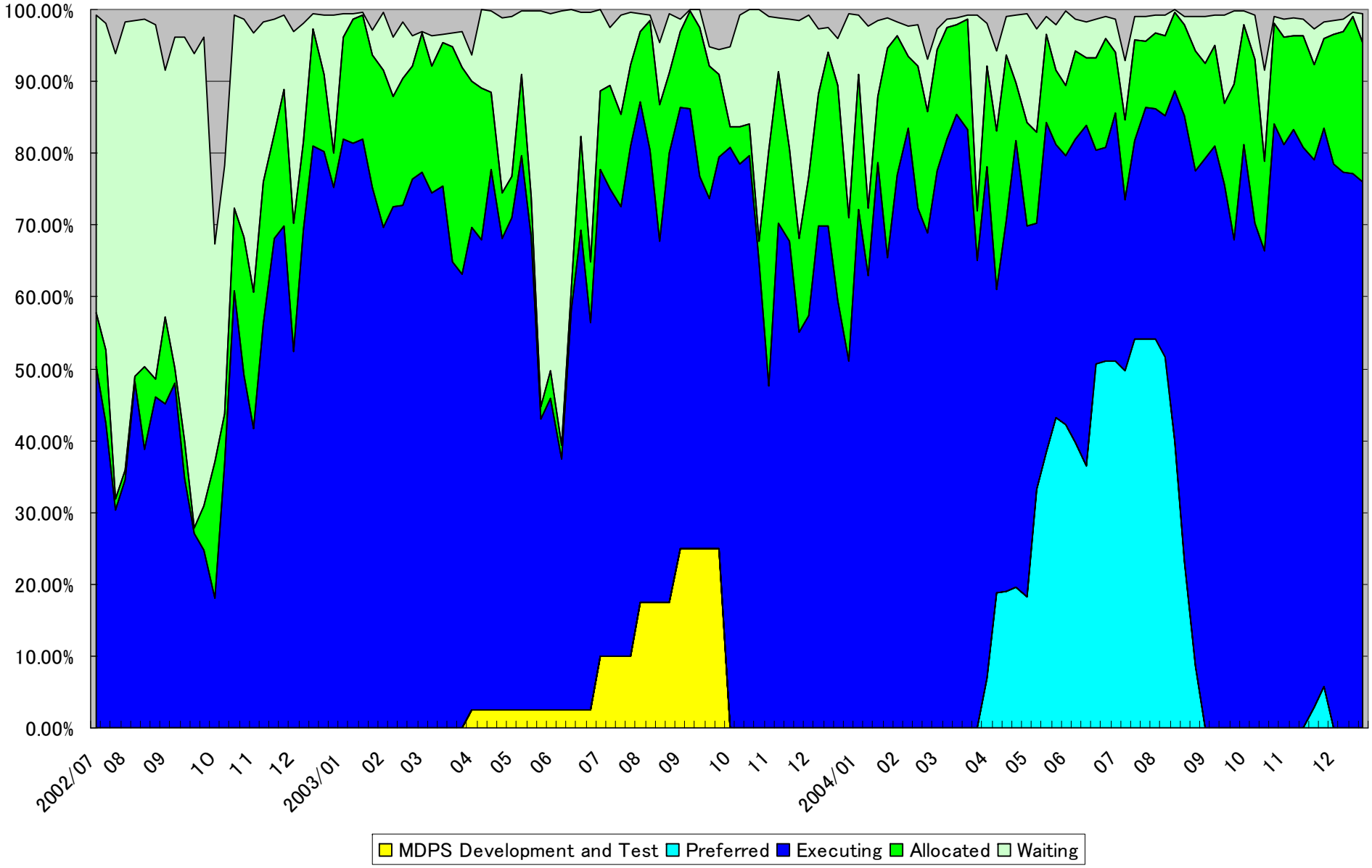
Number of Multi-node programs executed



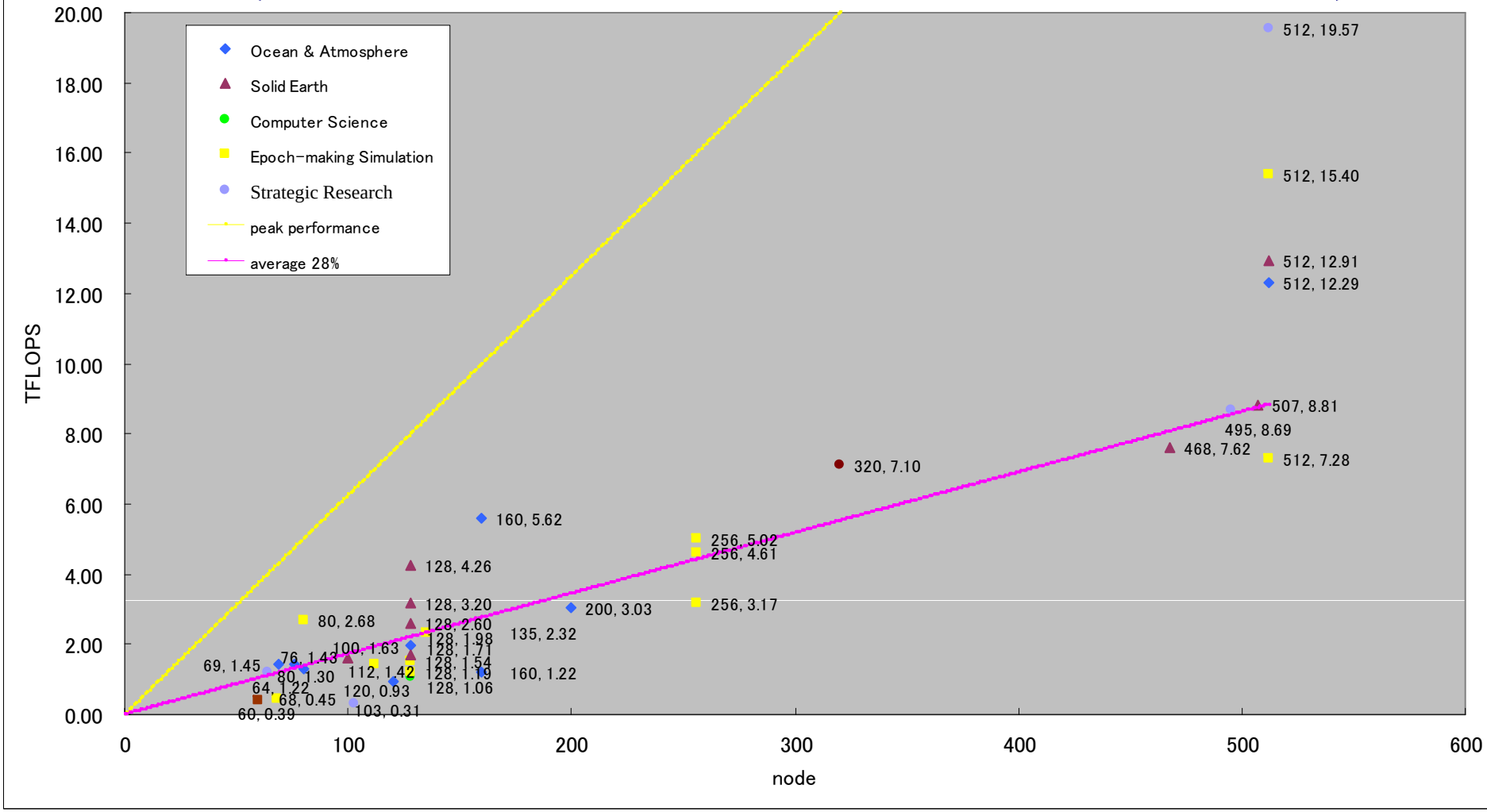
PN Request by Multi-node parallel programs



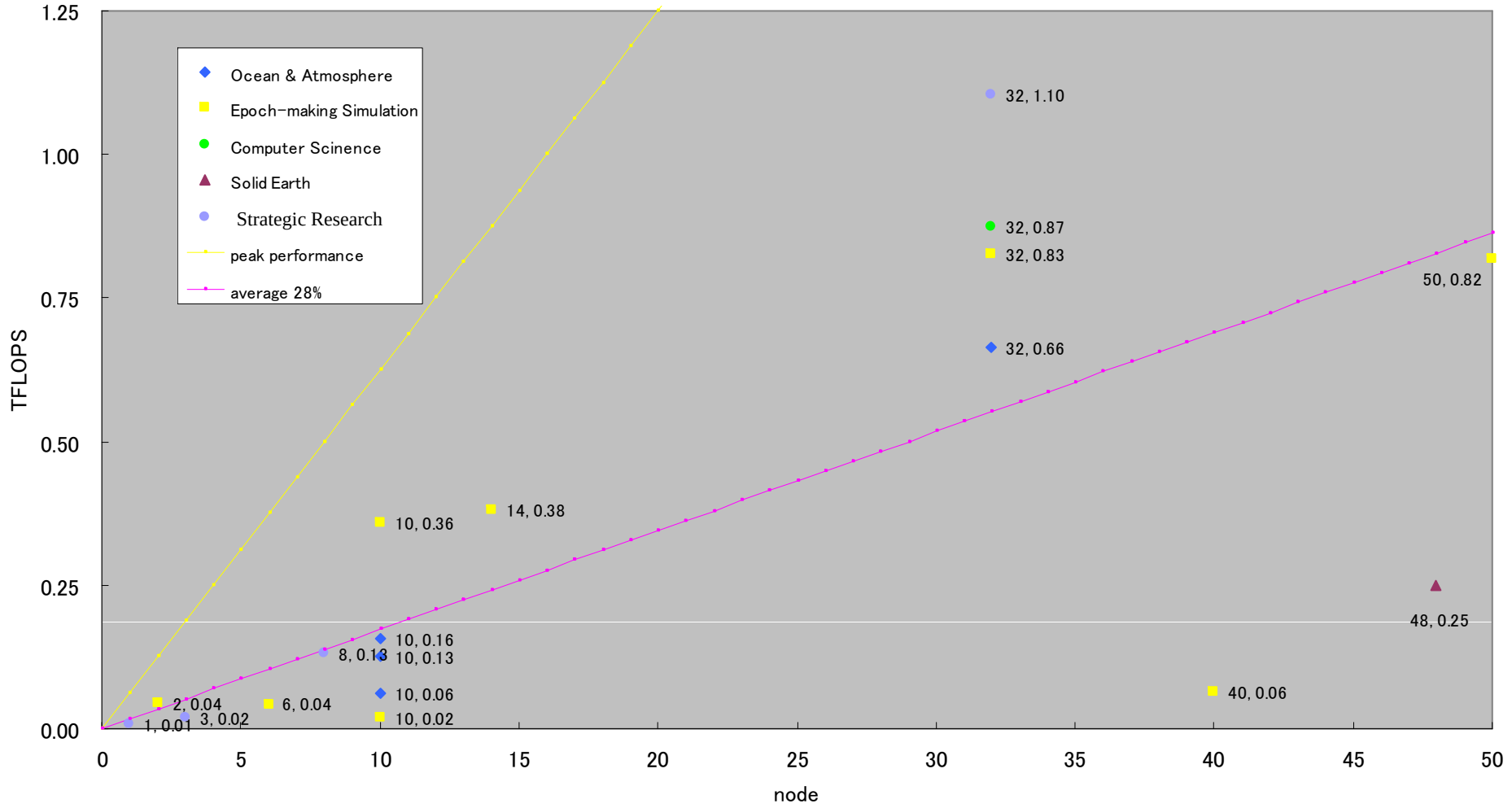
Processor Node availability for multi-node parallel programs



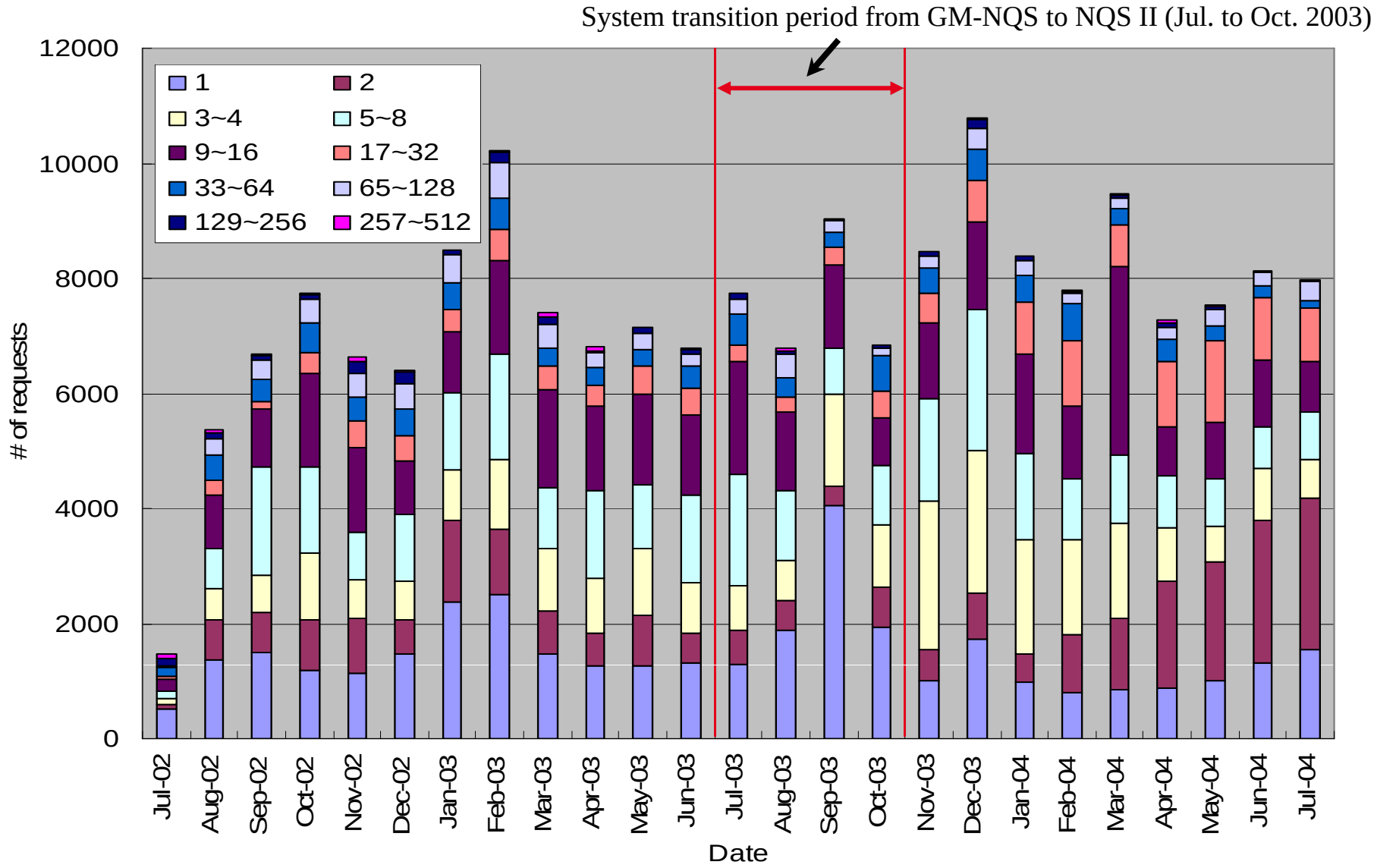
Highest Performance Program of Each project
of PN/TFLOPS
(2004/04/01～2004/12/31 More than 50 PNs)



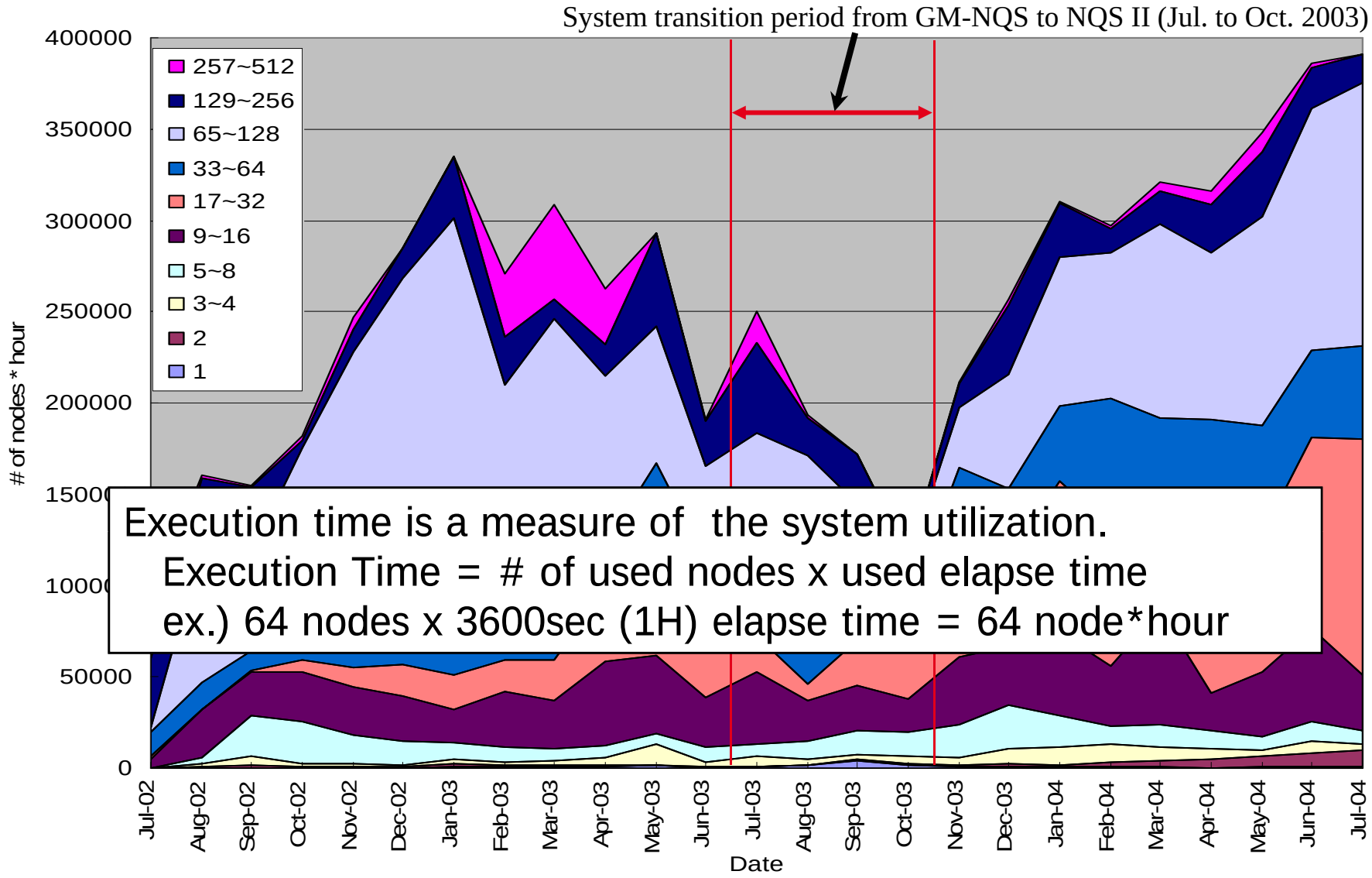
Highest Performance Program of Each project
of PN/TFLOPS
(2004/04/01～2004/12/31 Less than 51 PNs)



Number of Requests executed in ES



Execution Time

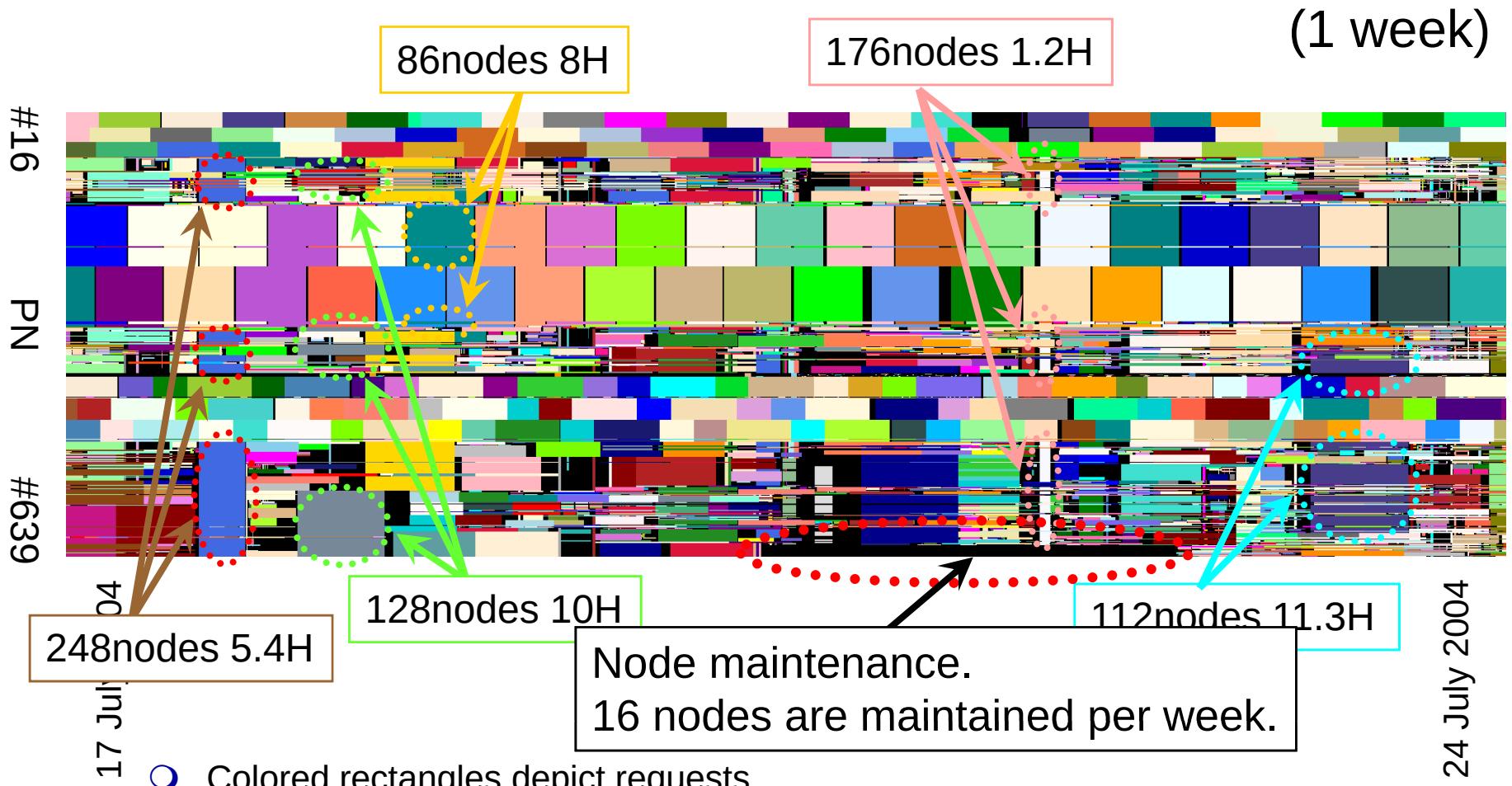


Execution time is a measure of the system utilization.

Execution Time = # of used nodes x used elapse time

ex.) 64 nodes x 3600sec (1H) elapse time = 64 node*hour

Number and size of requests run



- Colored rectangles depict requests.
- Black area means node maintenance or waiting execution.
- Requests of many sizes are executed simultaneously in ES.

EARTH SIMULATOR

Thank you for your attention