

The Earth Simulator and its Beyond

**— Technological Considerations towards —
Sustained Peta Flops Machine**

Oct. 2, 2004

NEC

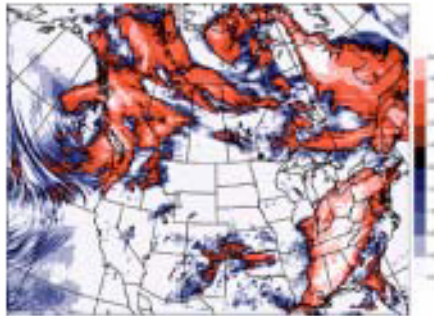
Tadashi Watanabe

(e-mail:t-watanabe@db.jp.nec.com)

Simulating “Earth” on Supercomputer

Supercomputer Simulation:

- can visualize
- can virtually experiment
- can forecast the future

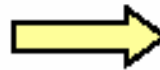


(North American 24hours Precipitation)

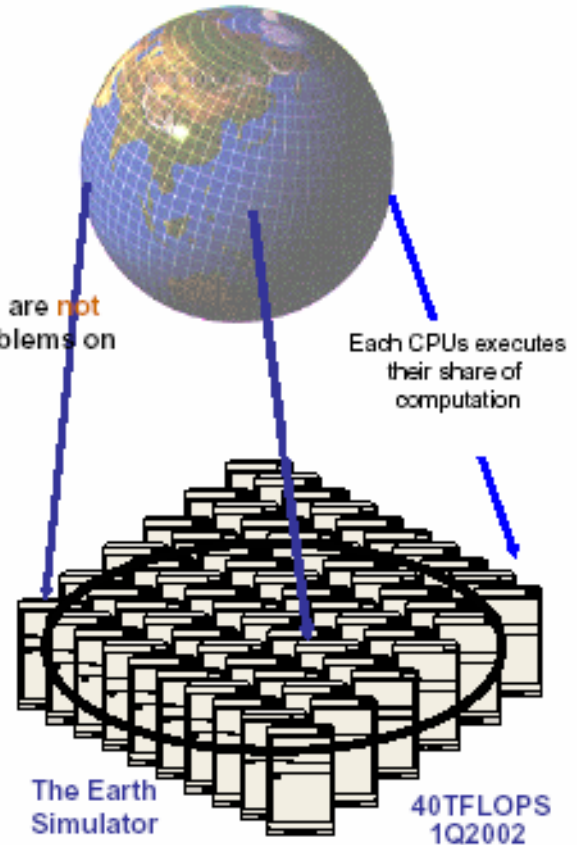


NEC SX-6/8A

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

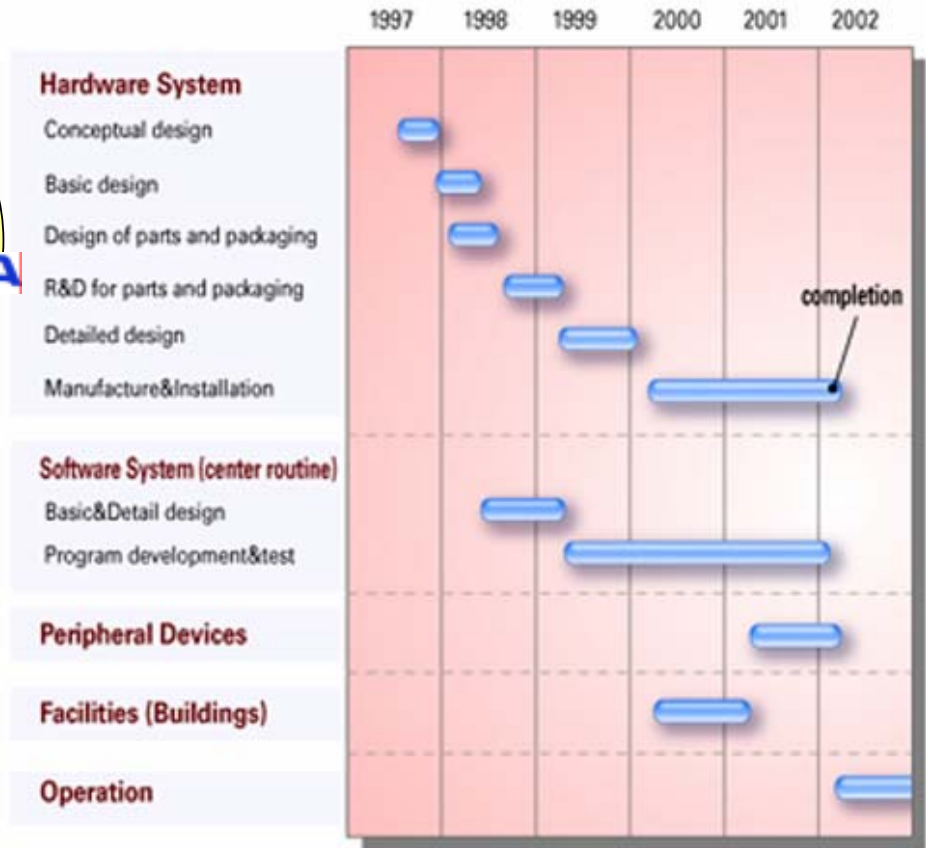
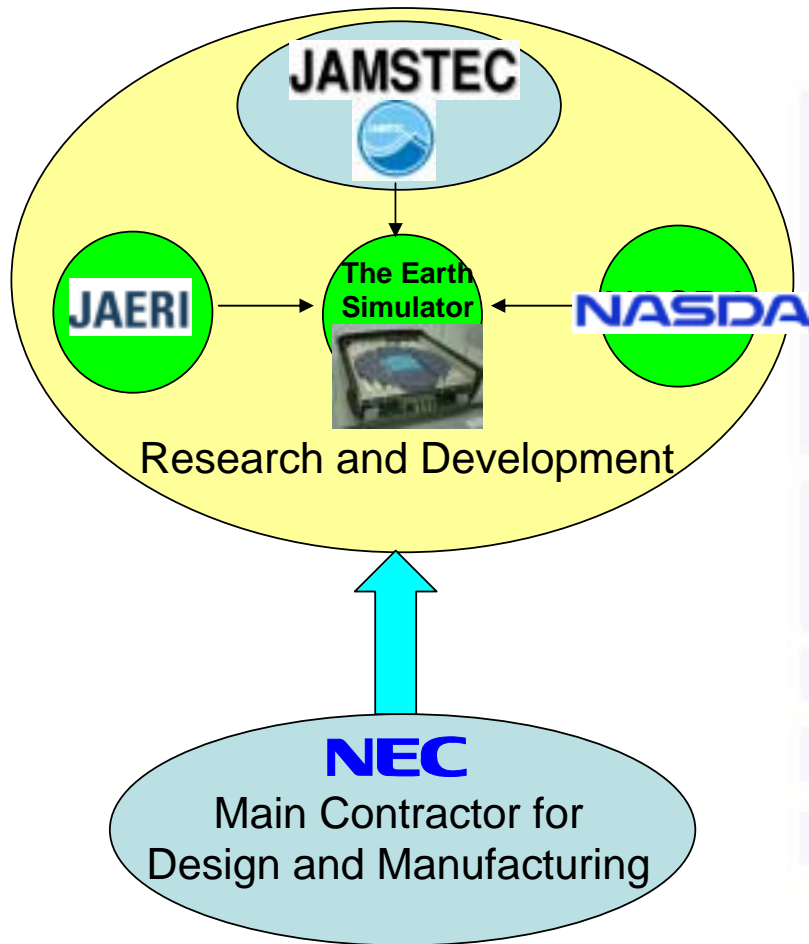


Power
x 640



Project of

Development Organization and Schedule

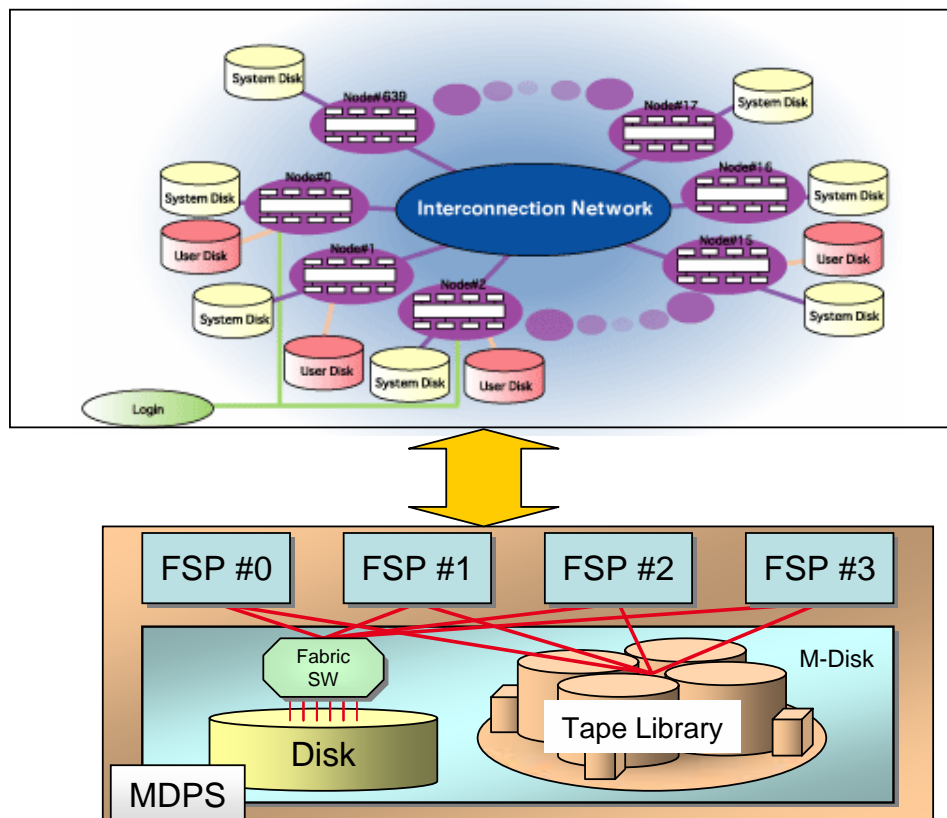


(Courtesy of JAMSTEC/Earth Simulator Center)

System and Hardware



Earth Simulator System



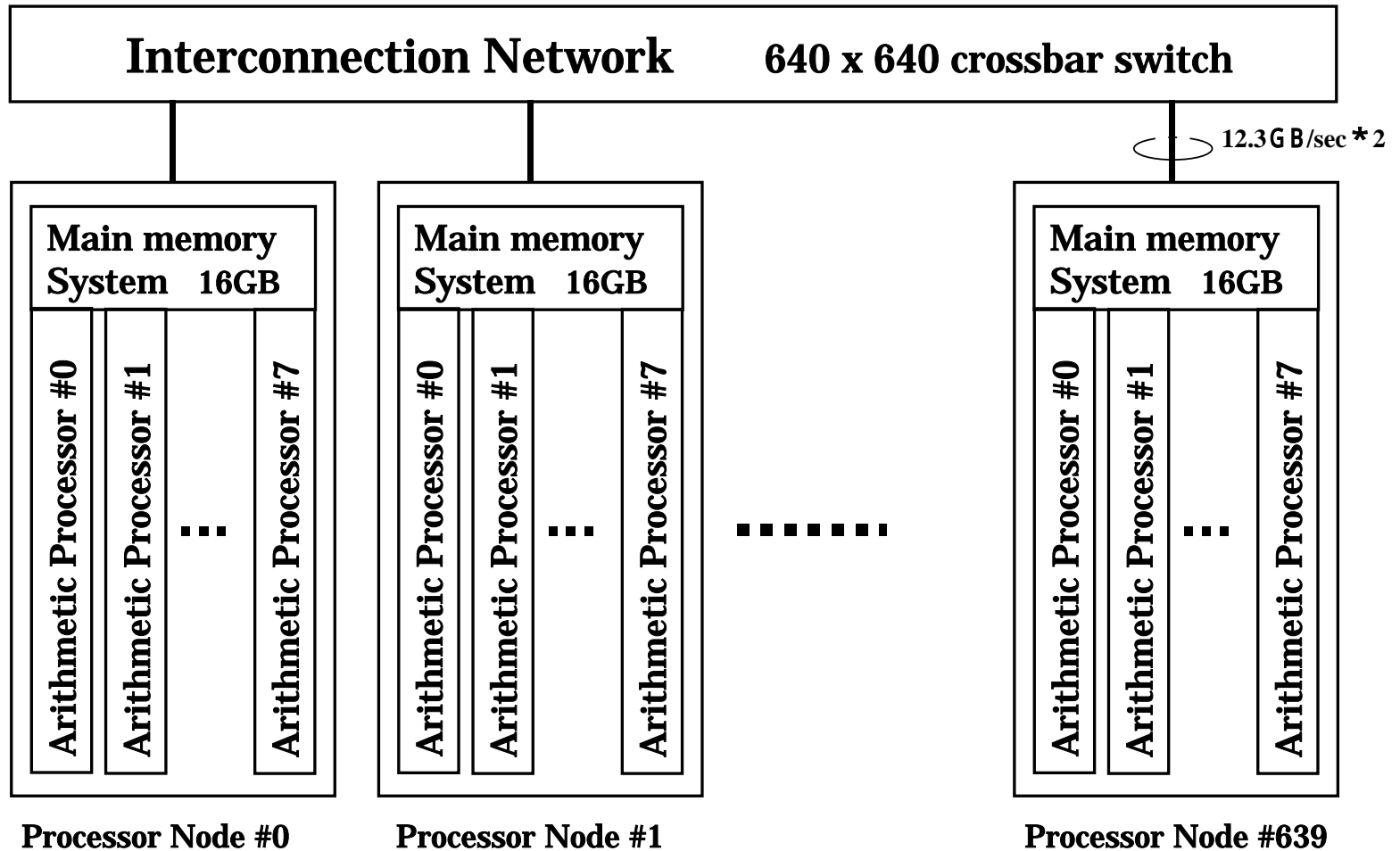
System Peak Performance
 Total No. of Arithmetic Processors (APs)
 Peak Performance/AP
 Total No. of Processor Nodes (PNs)

40TFLOPS
 5,120
 8GFLOPS
 640
 (8APs/Node: 64GFLOPS/Node)

Total Main Memory Capacity
 Disk Storage
 Mass Storage

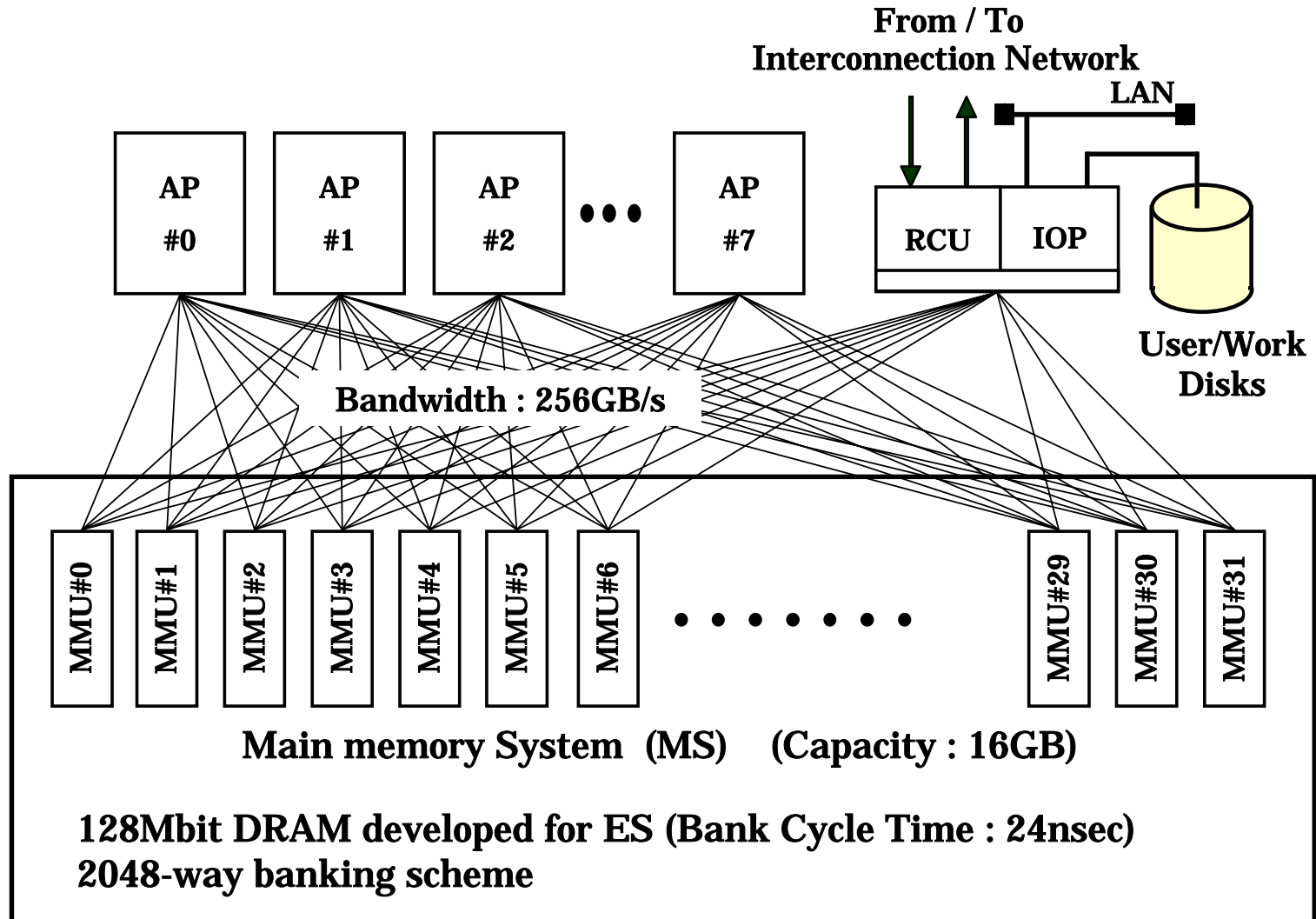
10TBytes
 940TBytes
 1.5PBytes

Central Subsystem

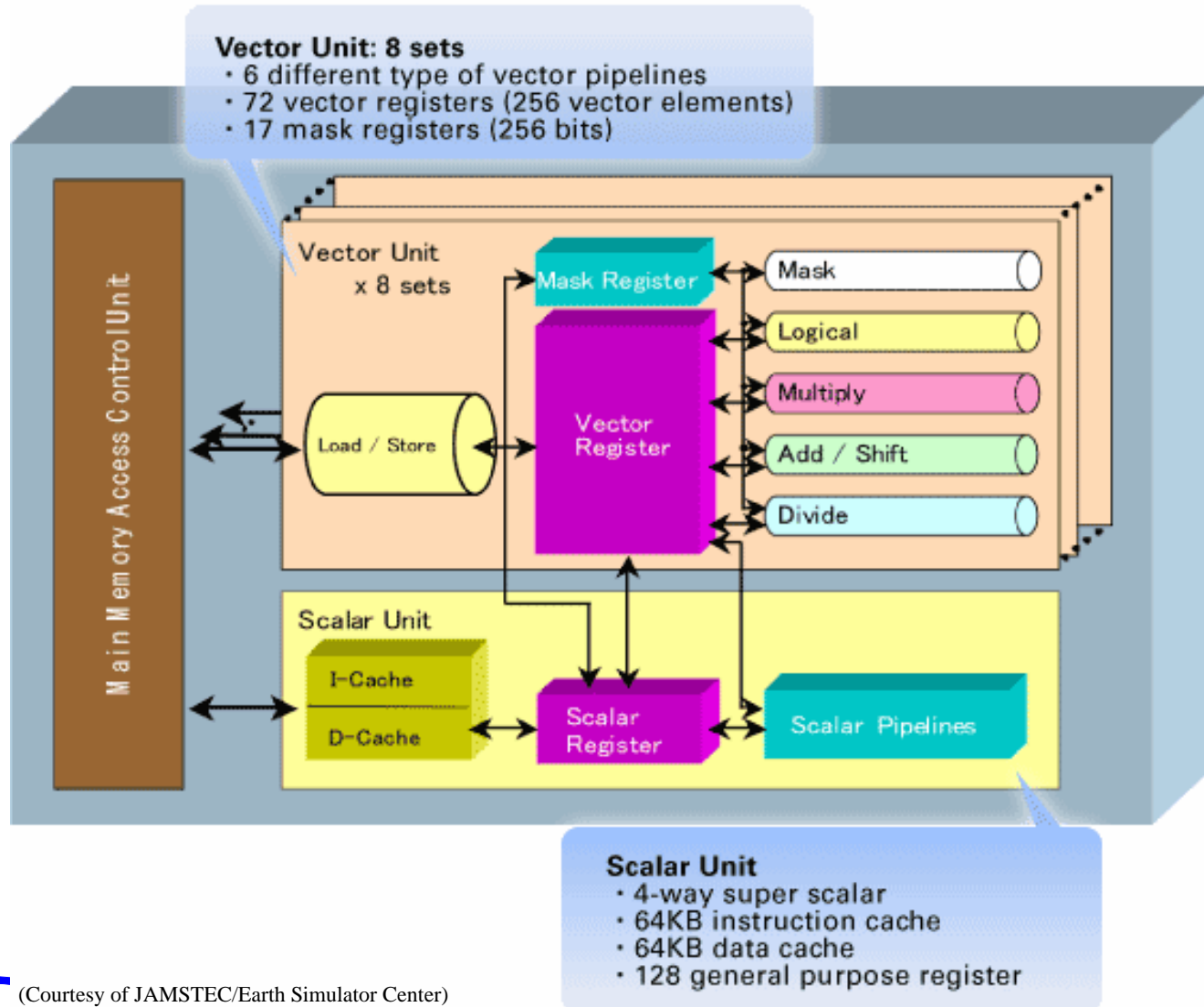


(Courtesy of JAMSTEC/Earth Simulator Center)

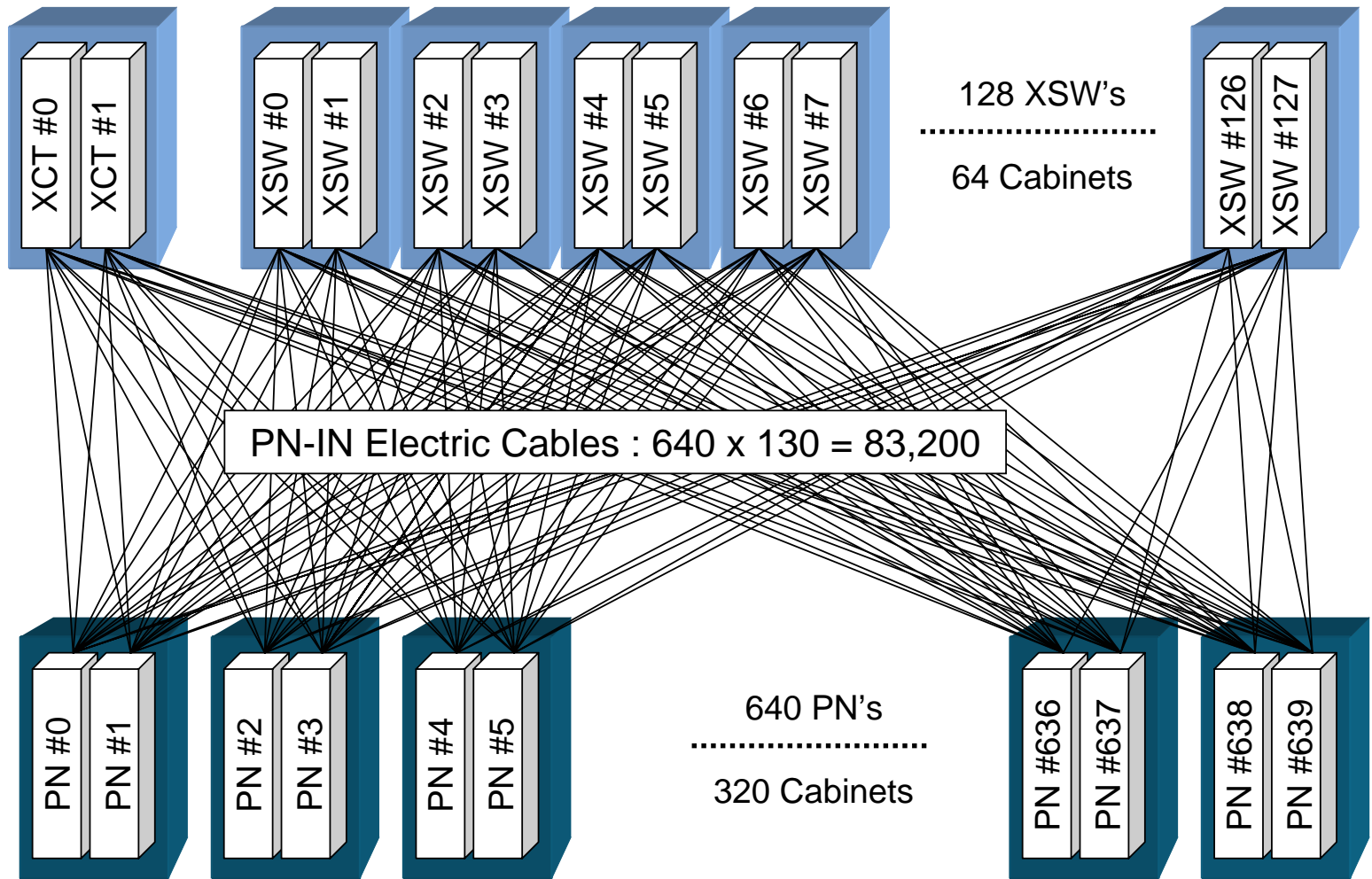
Processor Node



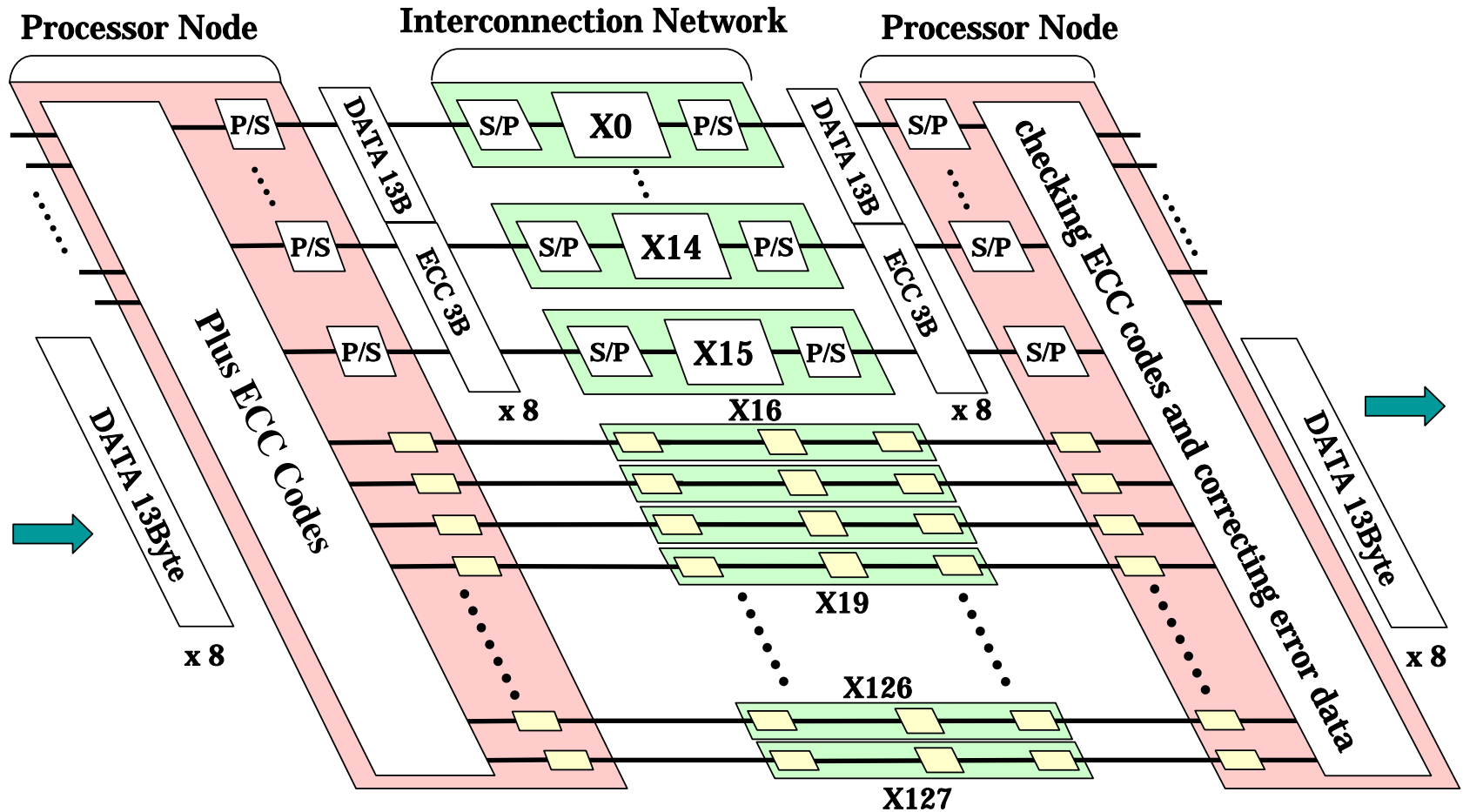
Arithmetic Processor (AP)



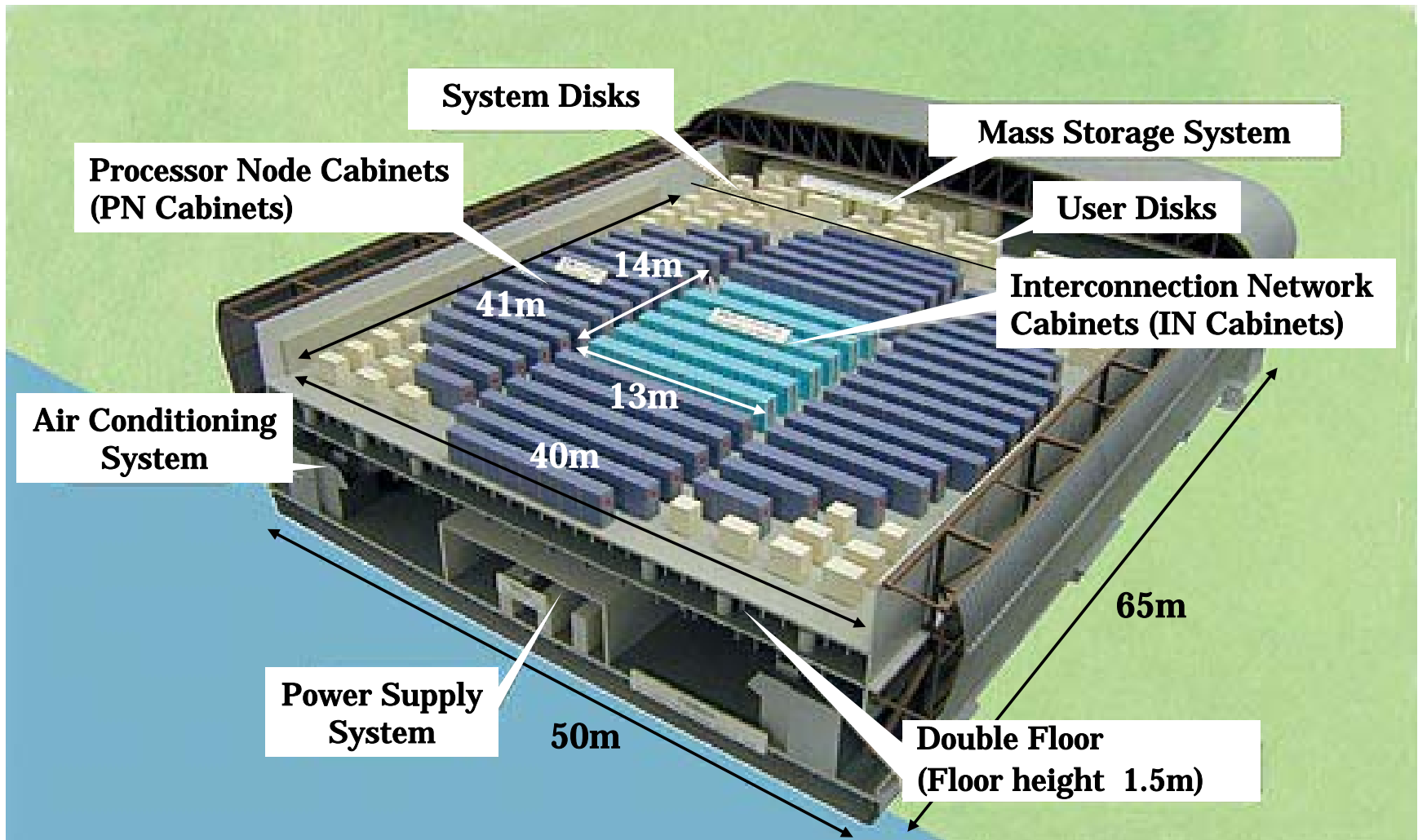
Connection between Cabinets



Data Paths in Interconnection Network(IN)



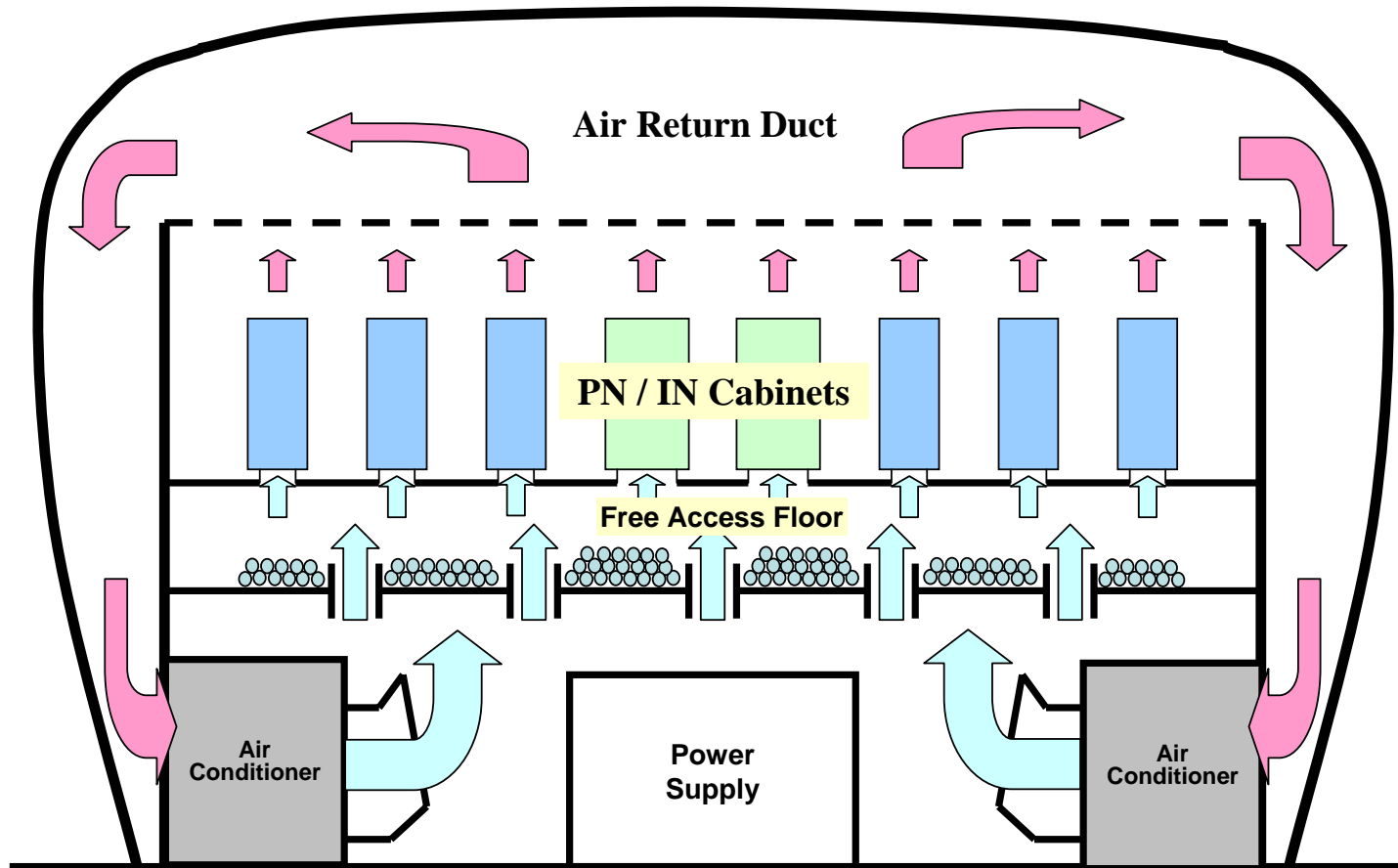
Earth Simulator Building



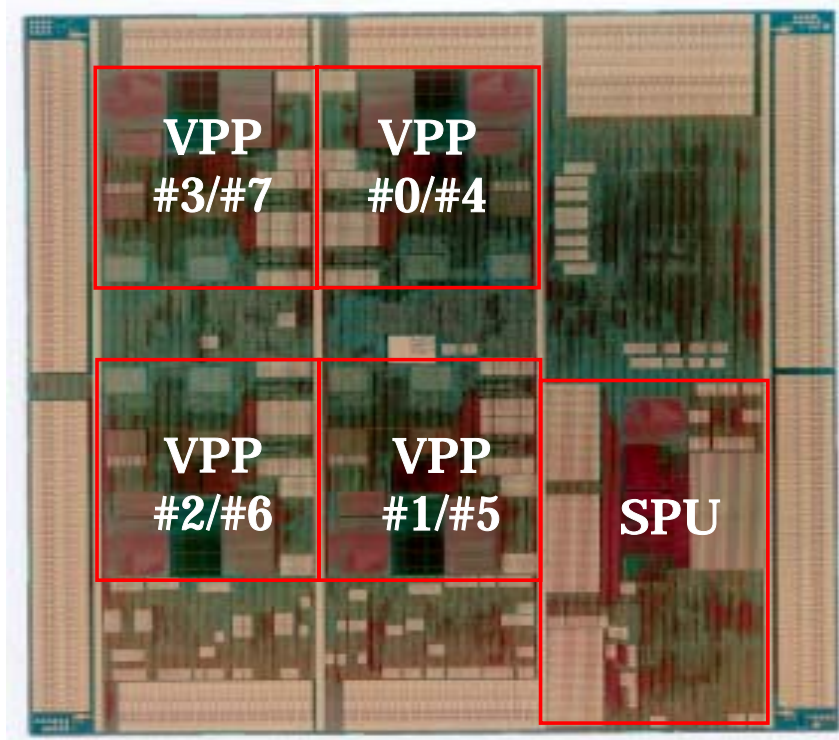
Inter-node Communication Cables



Cross-Sectional View of the Earth Simulator Building



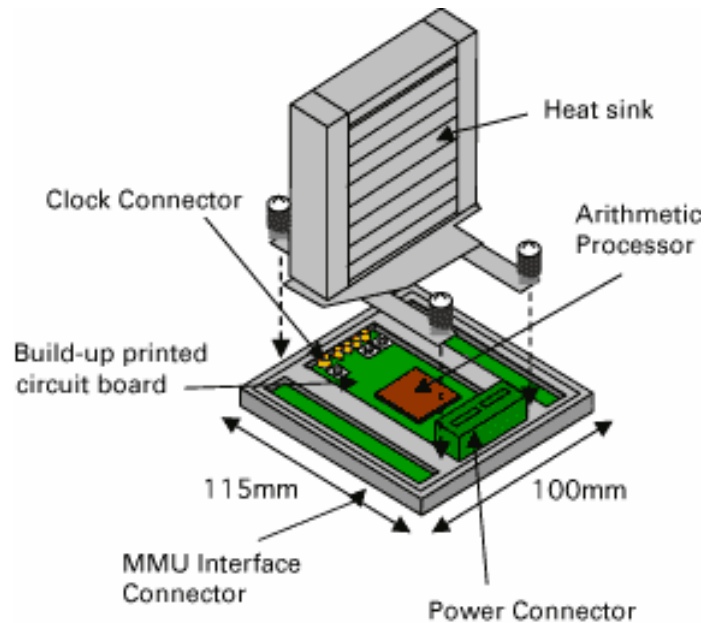
One Chip Vector Processor(AP)



(Courtesy of JAMSTEC/Earth Simulator Center)

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

AP Package



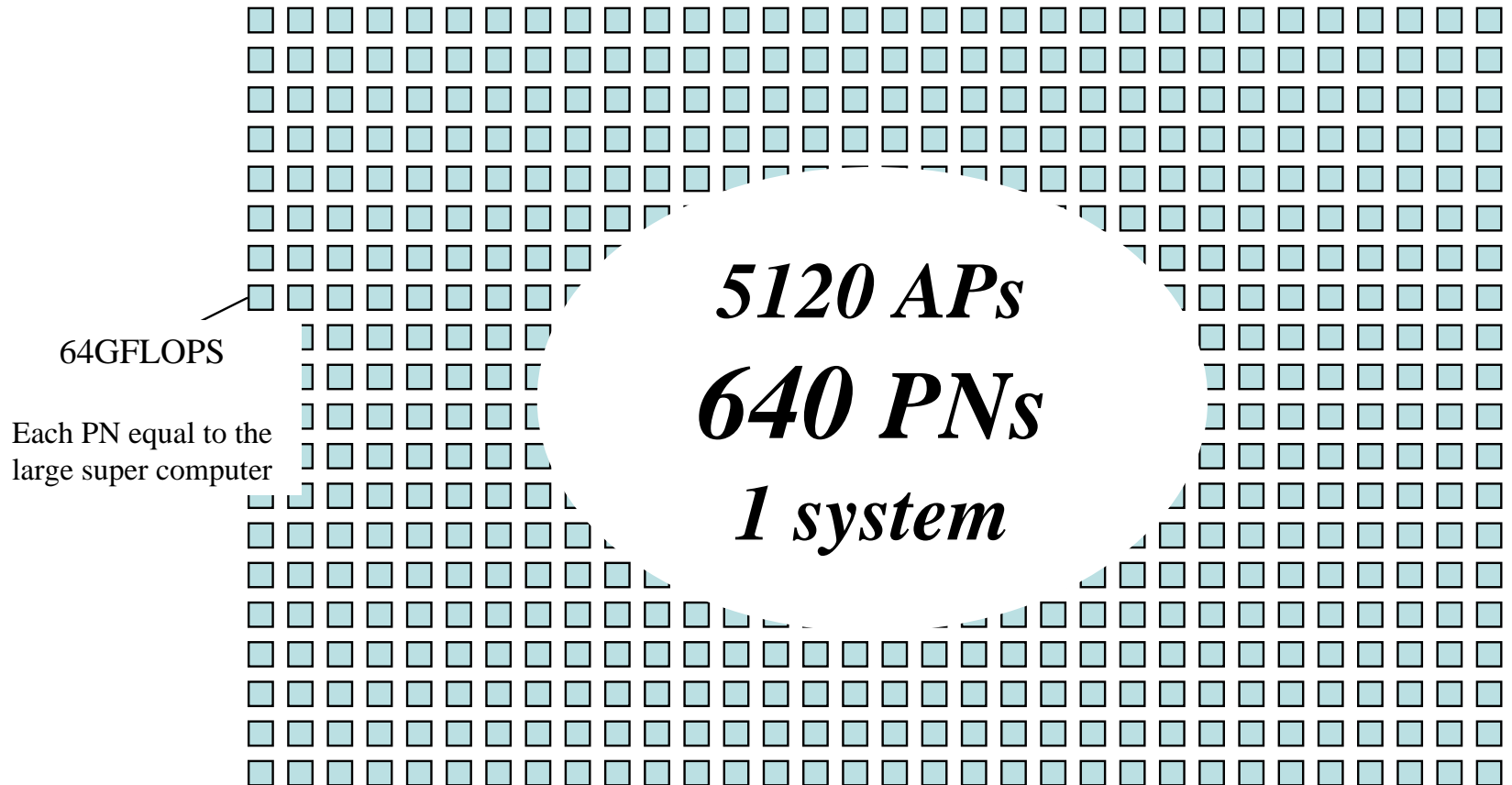
(Courtesy of JAMSTEC/Earth Simulator Center)



Software

Operation System Overview

- ✓ Operation and management system for huge distributed memory system



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 640nodes)

- ✓ 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 :

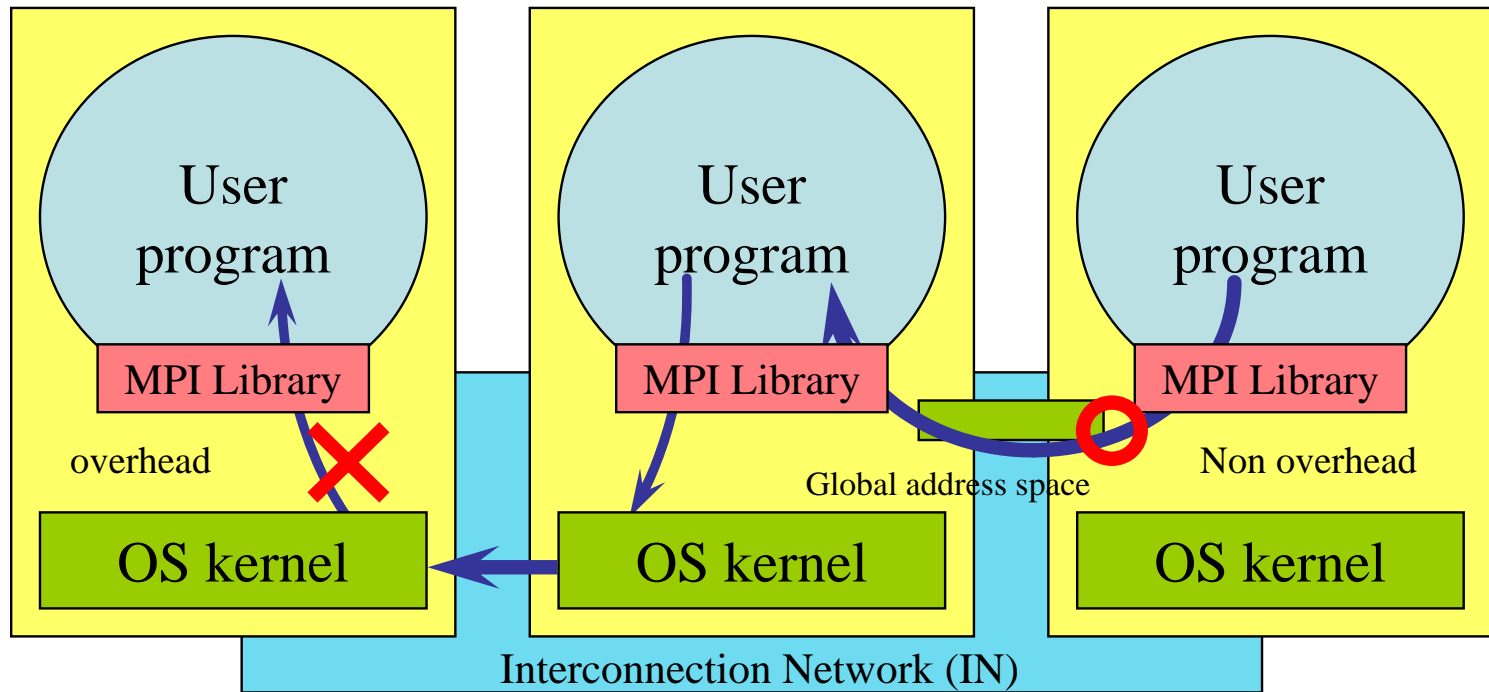
- ✓ **Super Cluster System** for system operation management
 - Two level cluster control (16nodes/cluster, 40cluster/system)
 - Resource management function of whole system (Node / IN / disk / tape)

for end users :

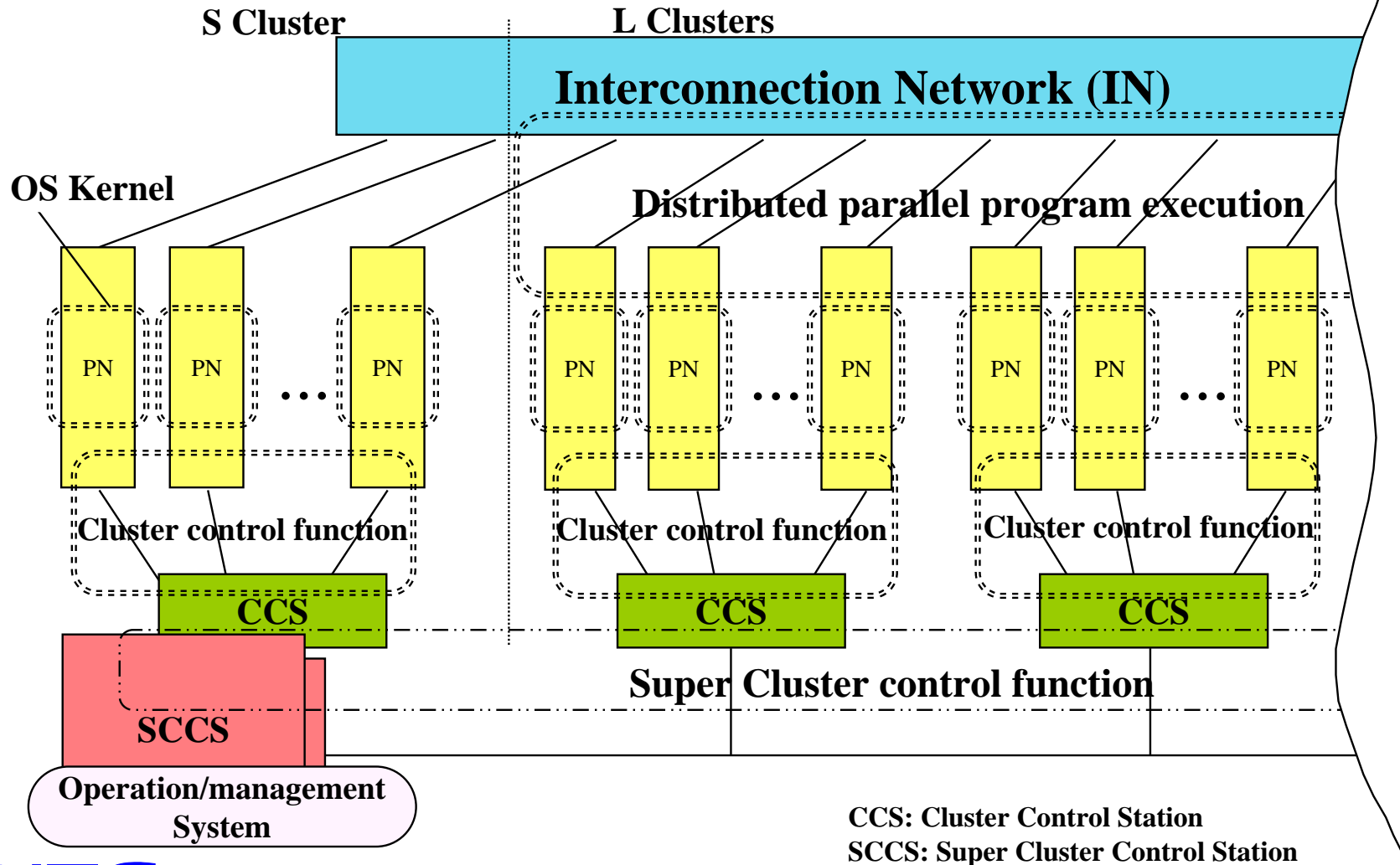
- ✓ Batch job environment for highly parallel job (NQSII,MDPS)
- ✓ Automatic file migration

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

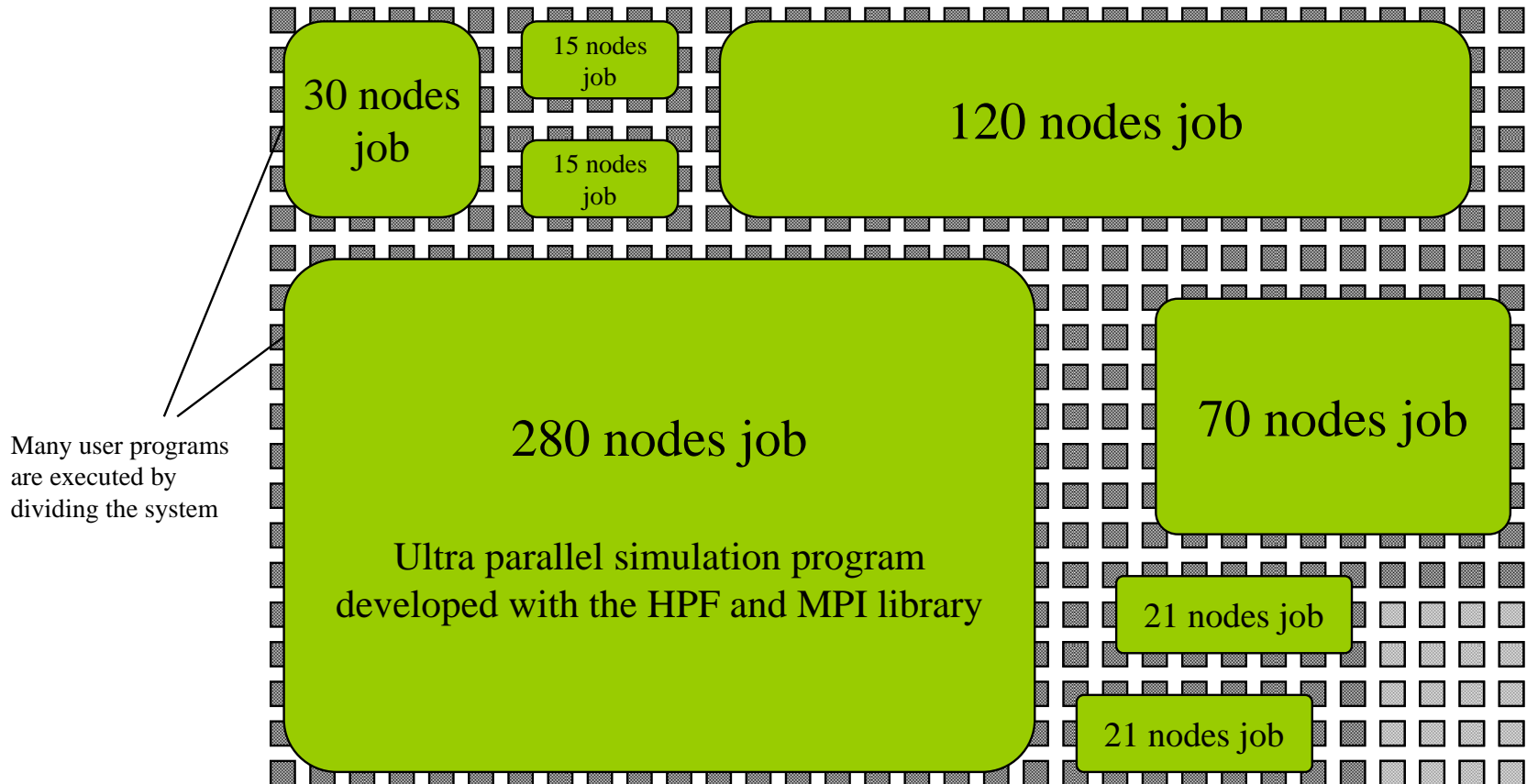


Operation management

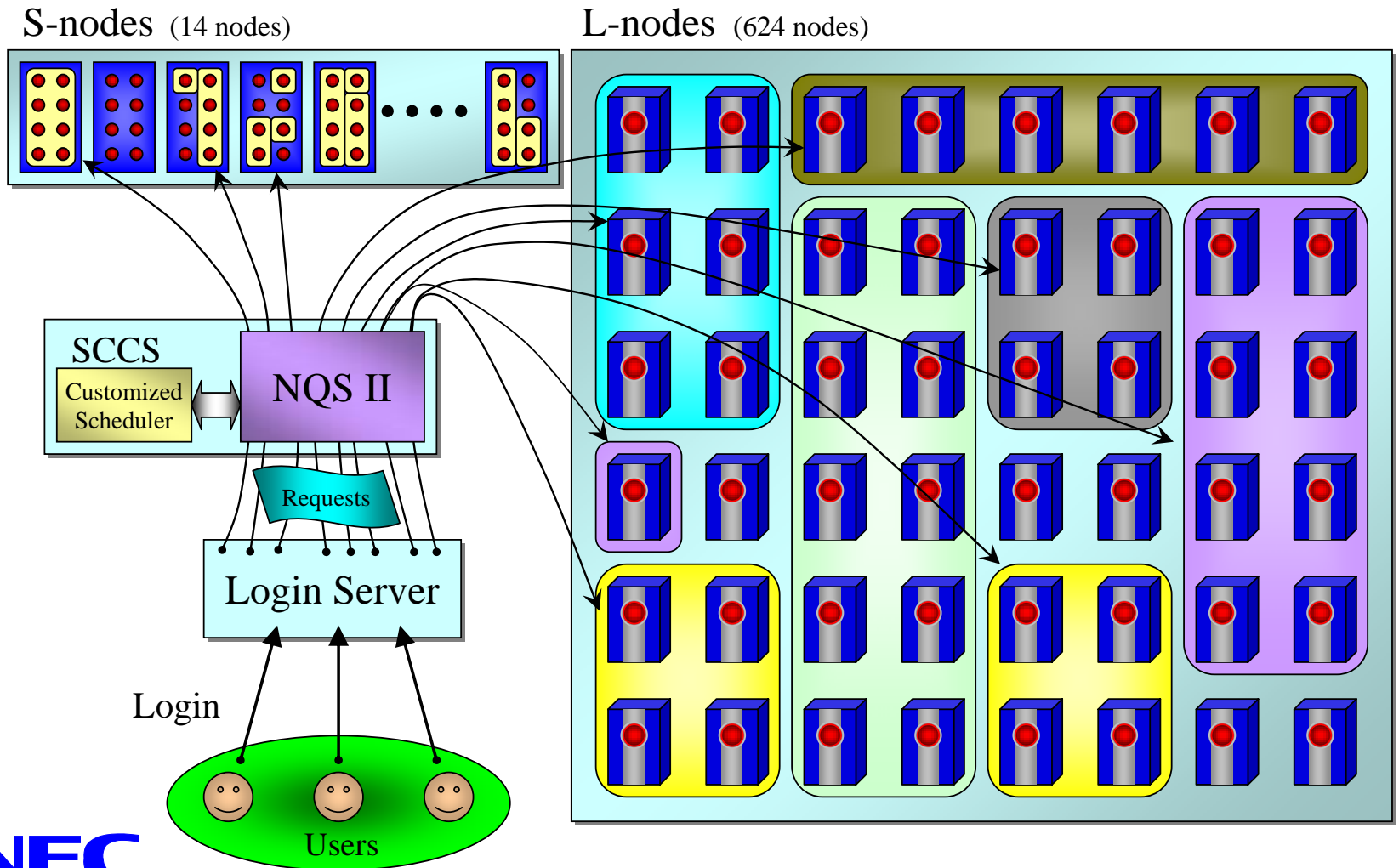


Execution of large scale job

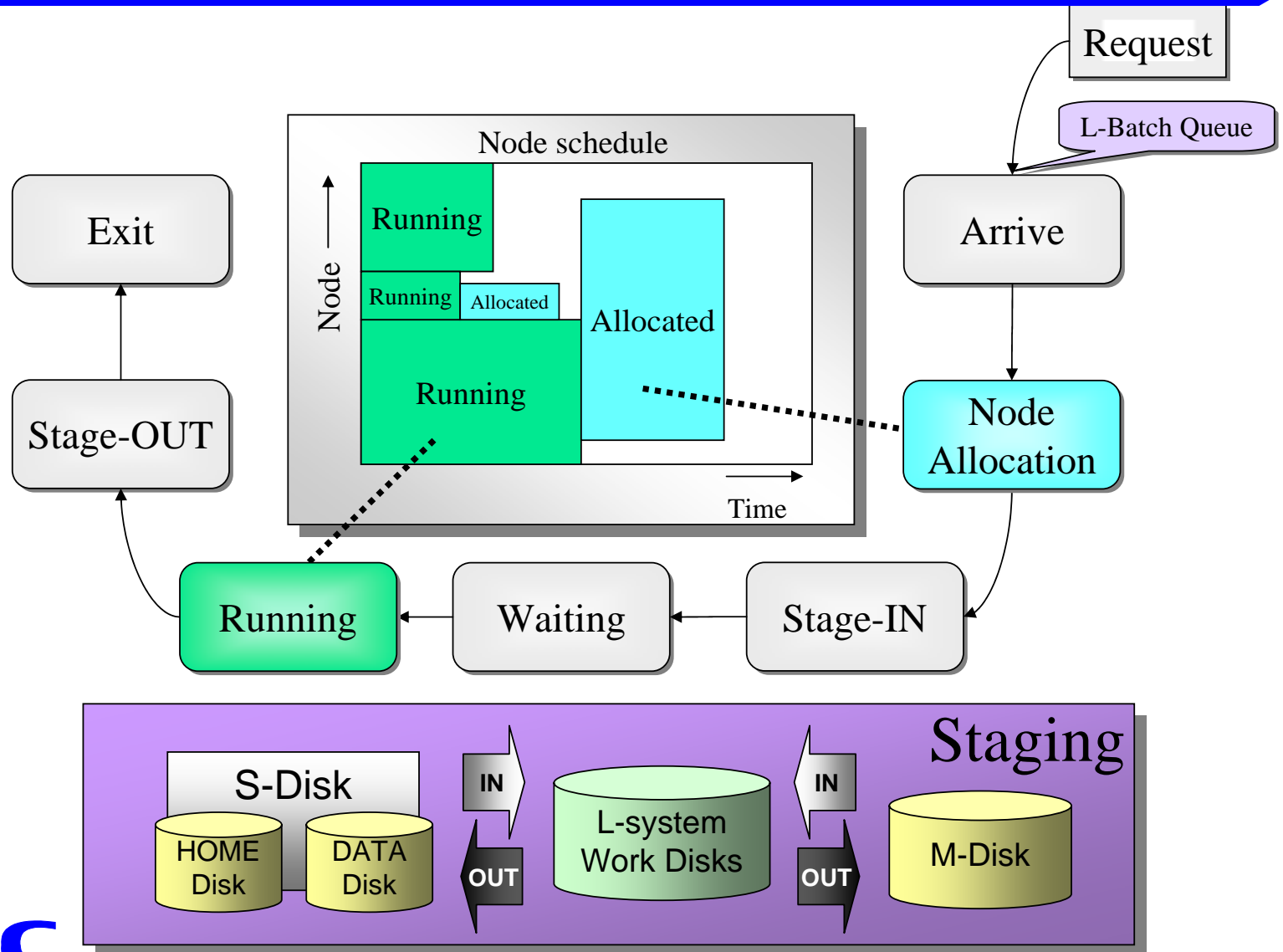
Large distributed parallel jobs



Node Allocation



Job Execution Flow



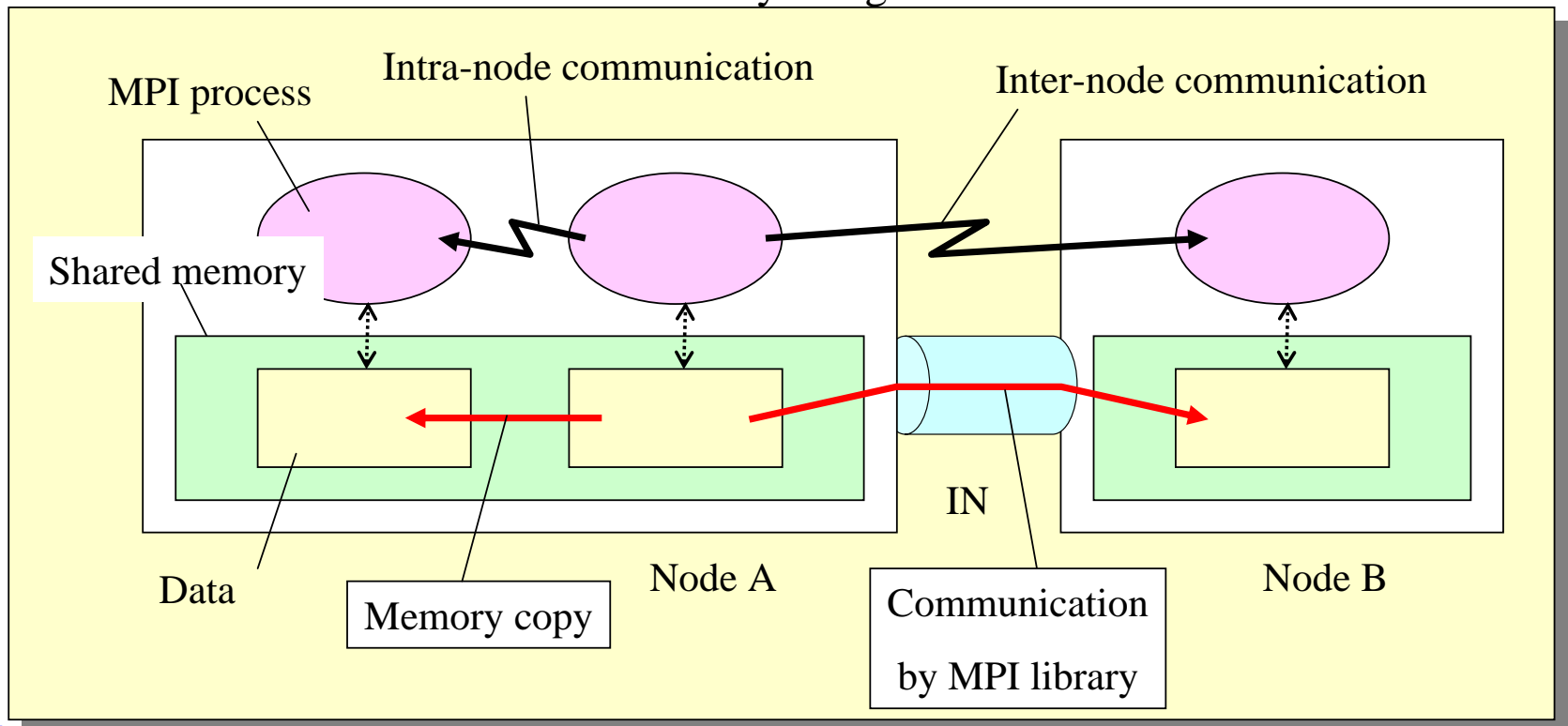
MPI (Message Passing Interface)

- ✓ 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

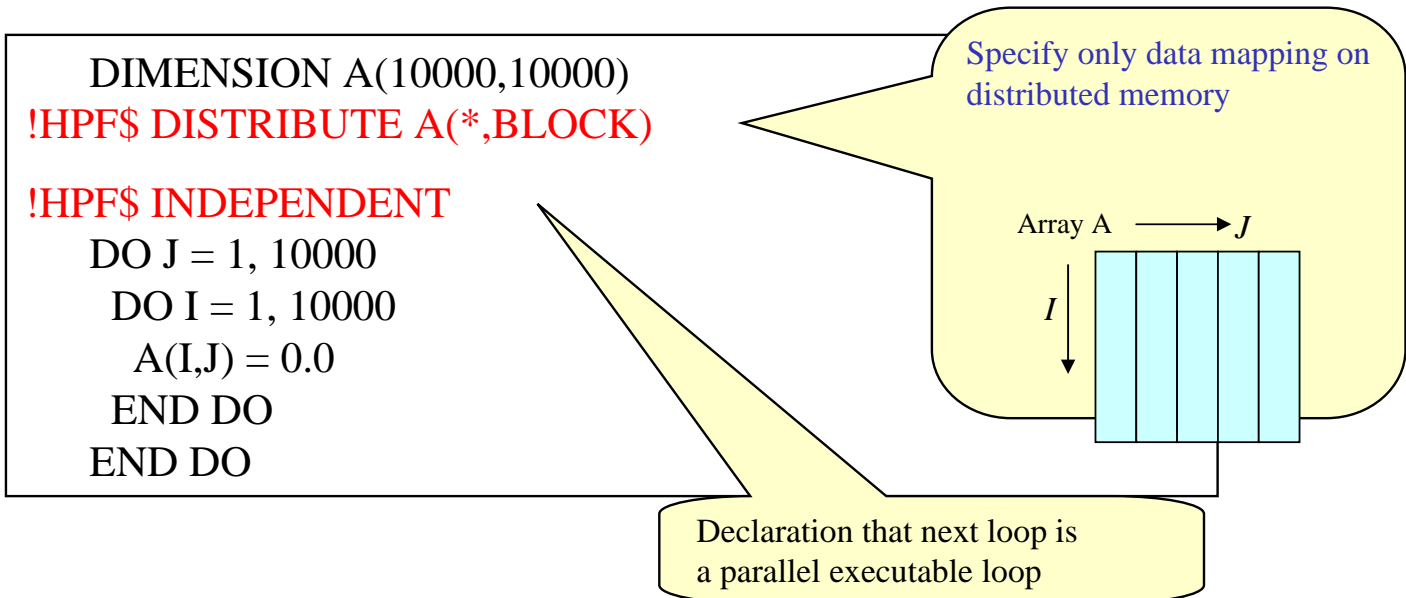
MPI library selects appropriate communication procedure

- ✓ 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)



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

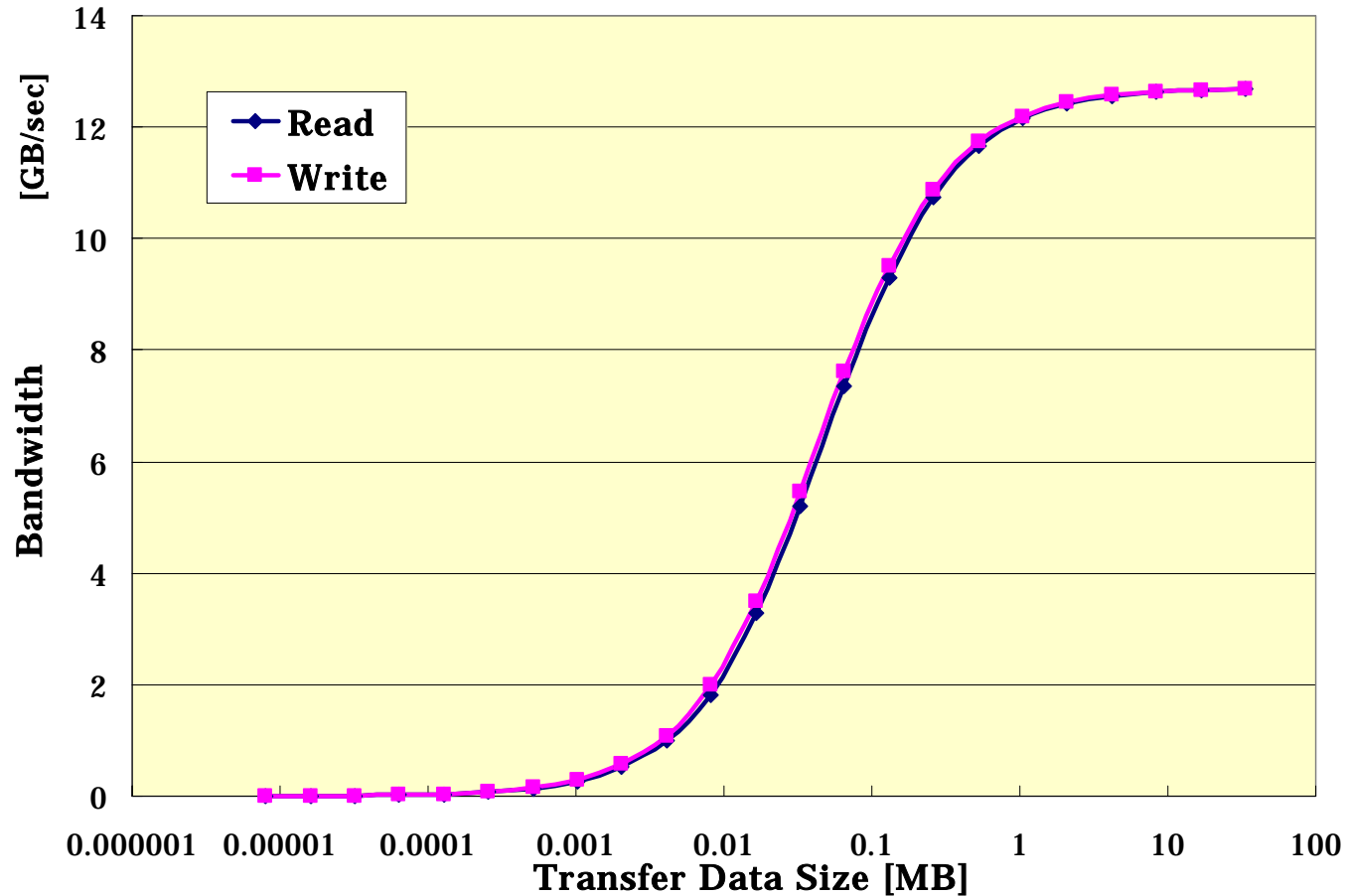
LINPACK(HPC)

Sustained Performance	35.86TFLOPS(87.5%efficiency)
-----------------------	------------------------------

MPI Start-up cost

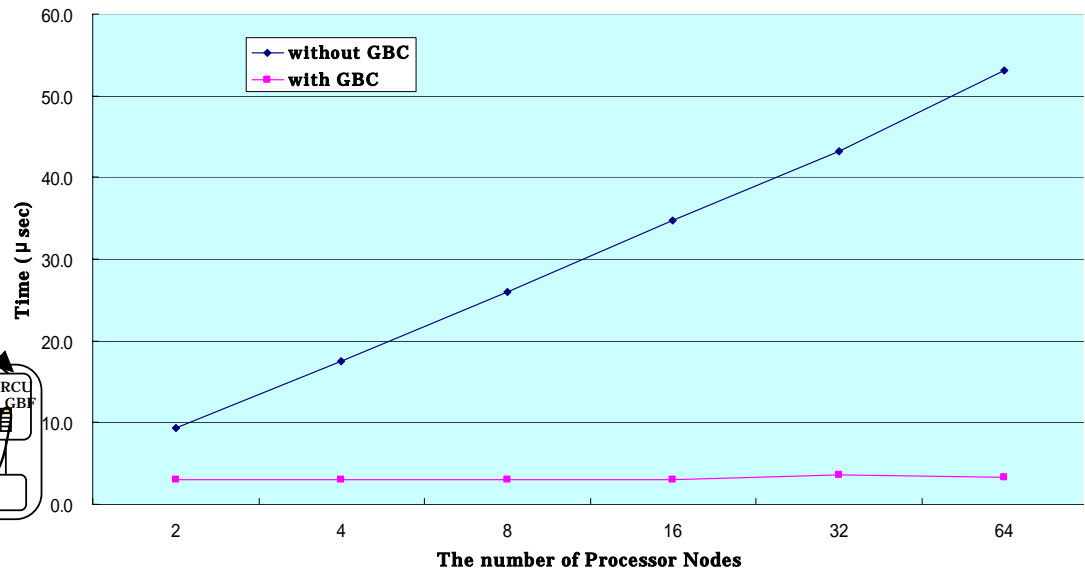
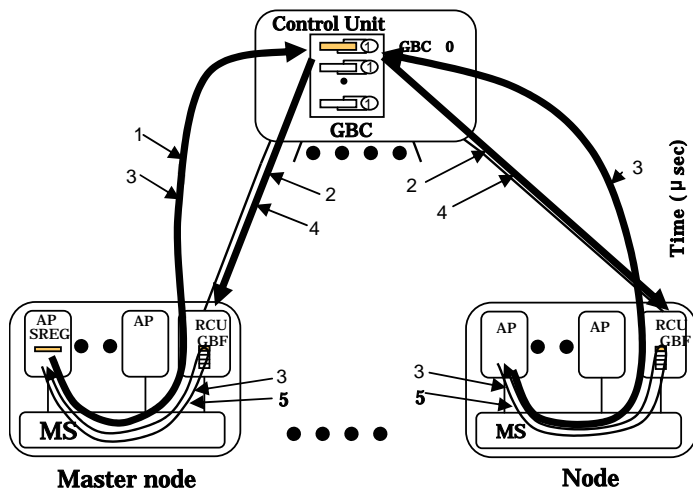
	internode	intranode
MPI_Get	6.68 μ s	1.27 μ s
MPI_Put	6.36	1.35

Internode Communication Bandwidth



(Courtesy of JAMSTEC/Earth Simulator Center)

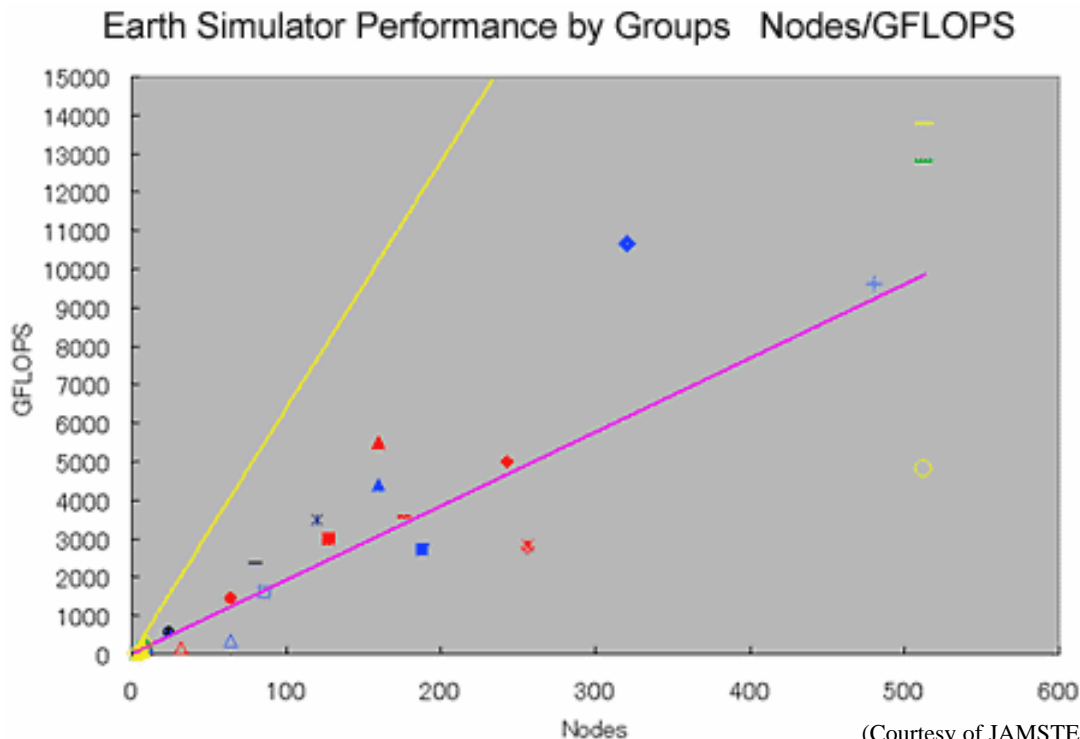
Barrier Synchronization



(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 :14.9TFLOPS(38.3%)
- for Fusion Science with HPF



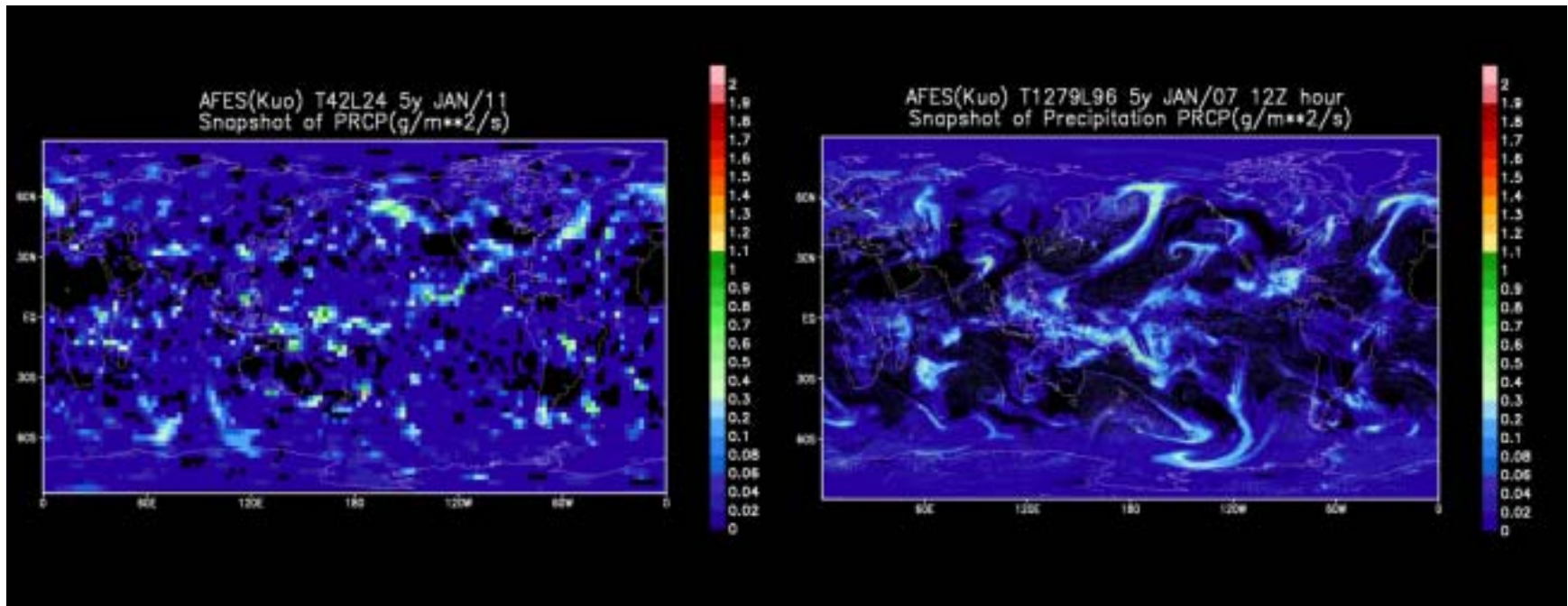
(Courtesy of JAMSTEC/Earth Simulator Center)

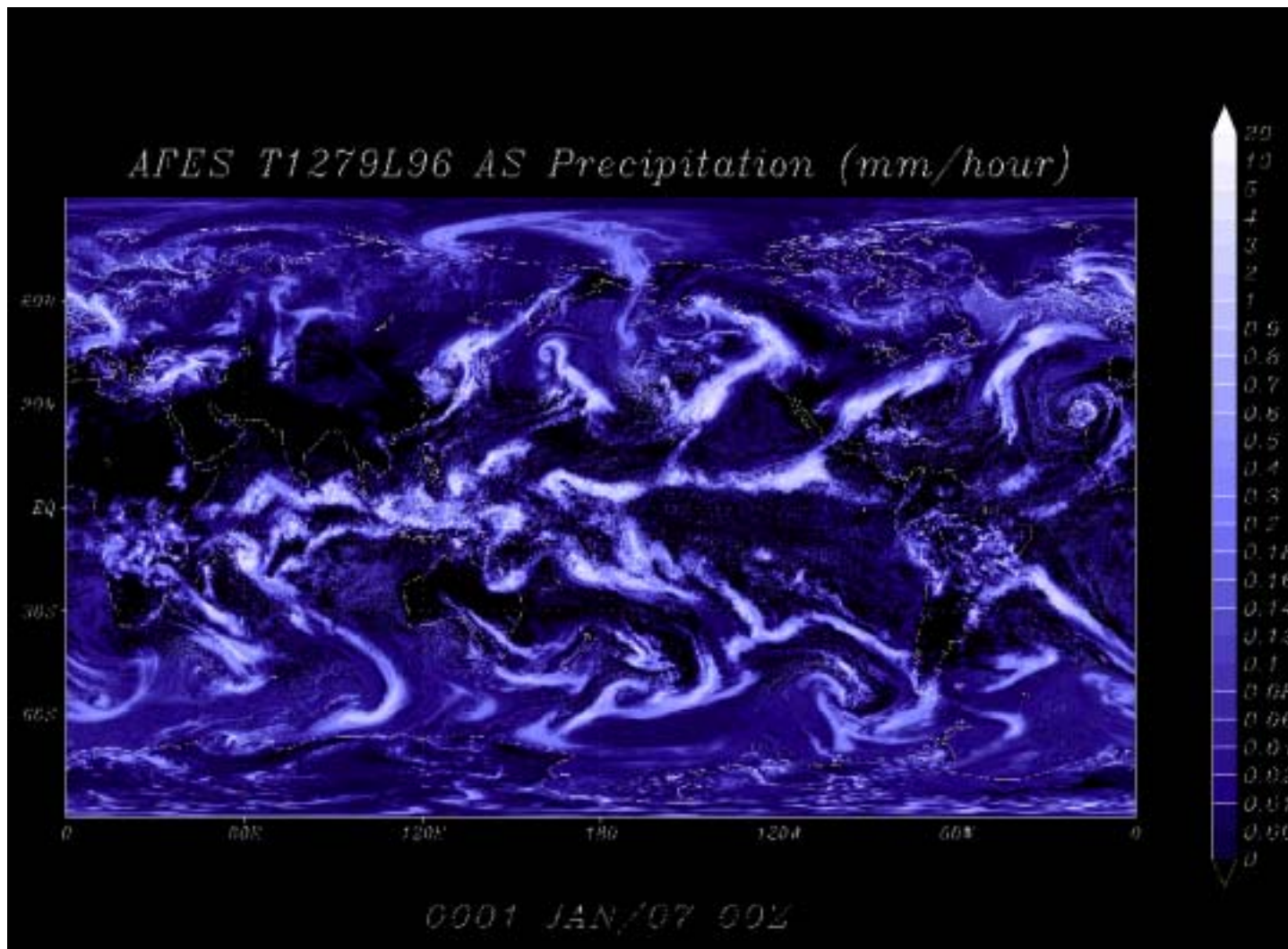
Blue: Ocean and Atmosphere Red: Solid Earth
Green: Computational Science Yellow: Epoch-Making

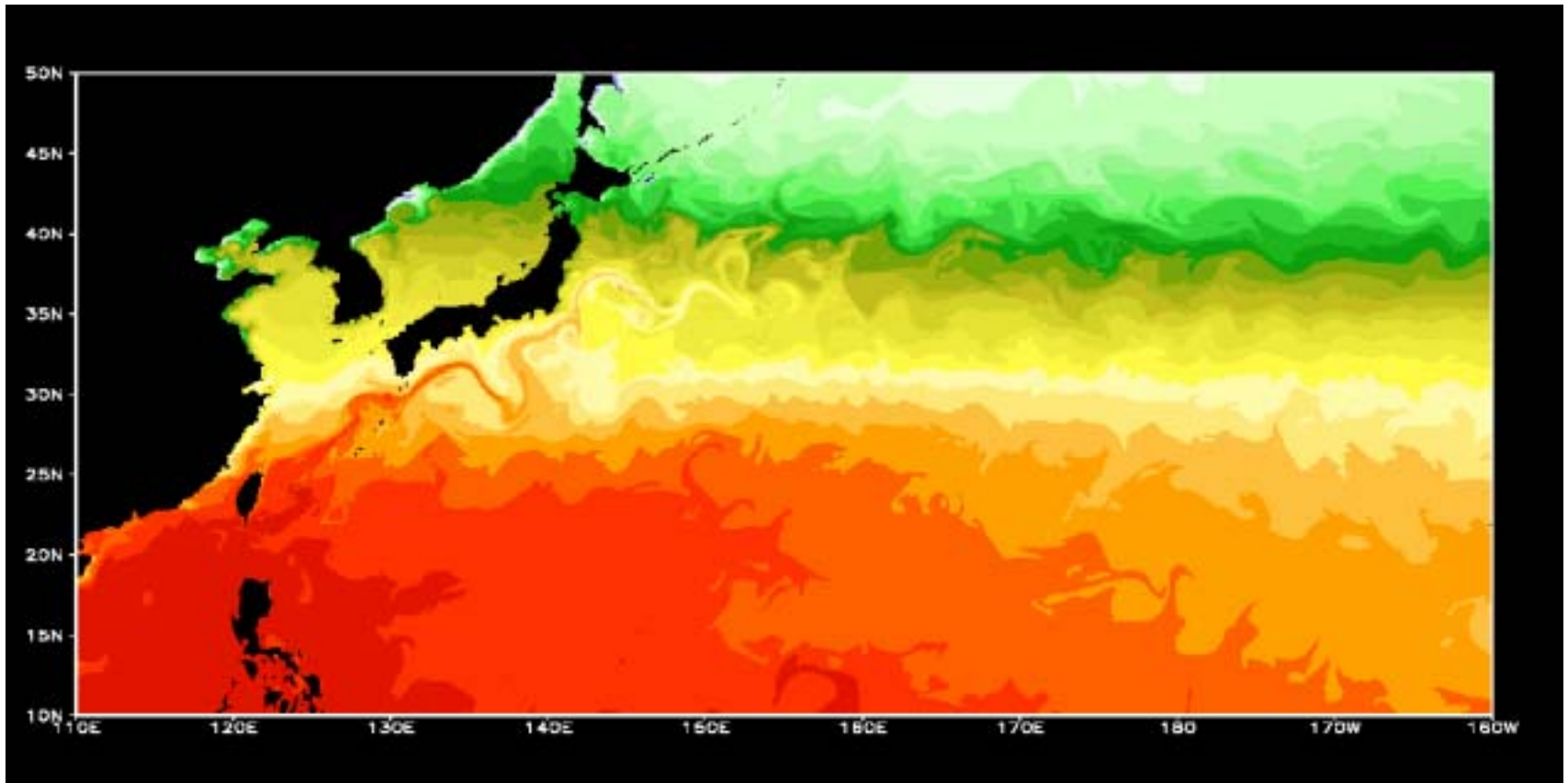


Application Results

Precipitation(312km,T42L24) Precipitation(10.4km,T1279L24)





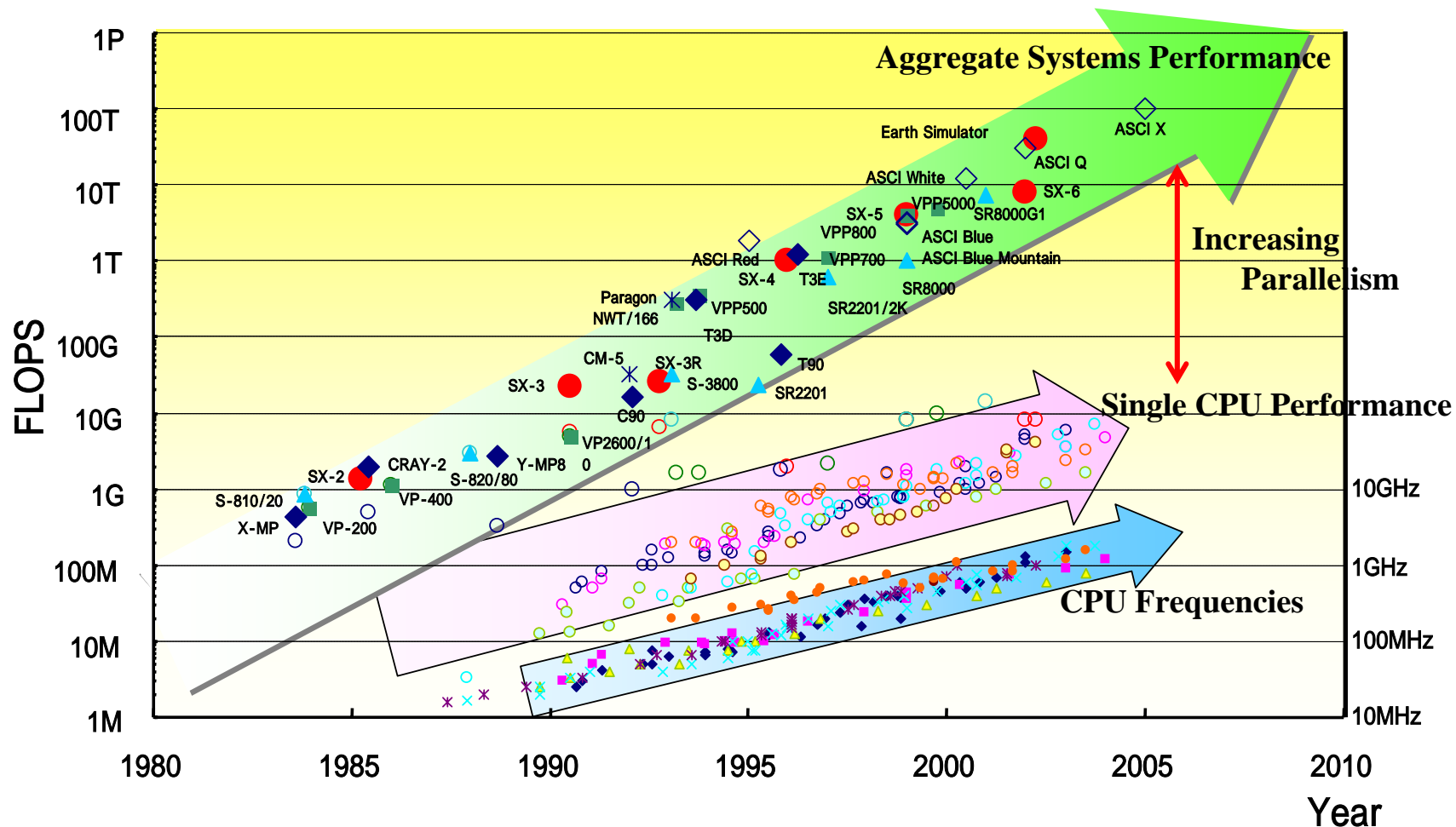


Copyright :JAMSTEC/Earth Simulator Center



Future Technological Challenges for Peta Flops Computing

History of High Performance Computers



Faster the Speed, More the Parallel

The Largest configuration in SX-3



22GFlops/4Cpu

1990

The Earth Simulator

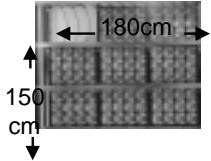

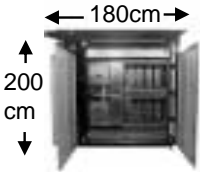

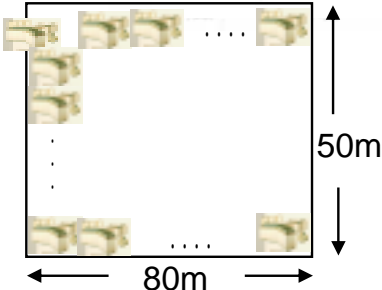



40TFlops/5120Cpu

2002



Evolution of SX Series for 20 years

	<u>'83</u>	<u>'03</u>	<u>Magnification</u>
CPU Performance	1.3GFLOPS	8 GFLOPS	x 6
System Performance	1.3GFLOPS	40TFLOPS (Earth Simulator)	x $3 * 10^4$
#of CPUs	1	5120 (Earth Simulator)	x 5,120
Total Memory Capacity	256MBytes	10Tera Bytes (Earth Simulator)	x $4 * 10^4$
CPU Size			x 1/6,750
# of chips per cpu	2,250chips	1 Chip	x 1/2,250
Memory Size		 2Carriers	x 1/4,000
System Size	 SX-2 x 49Cabinets	 64GFLOPS/8CPU	x 1/4,000

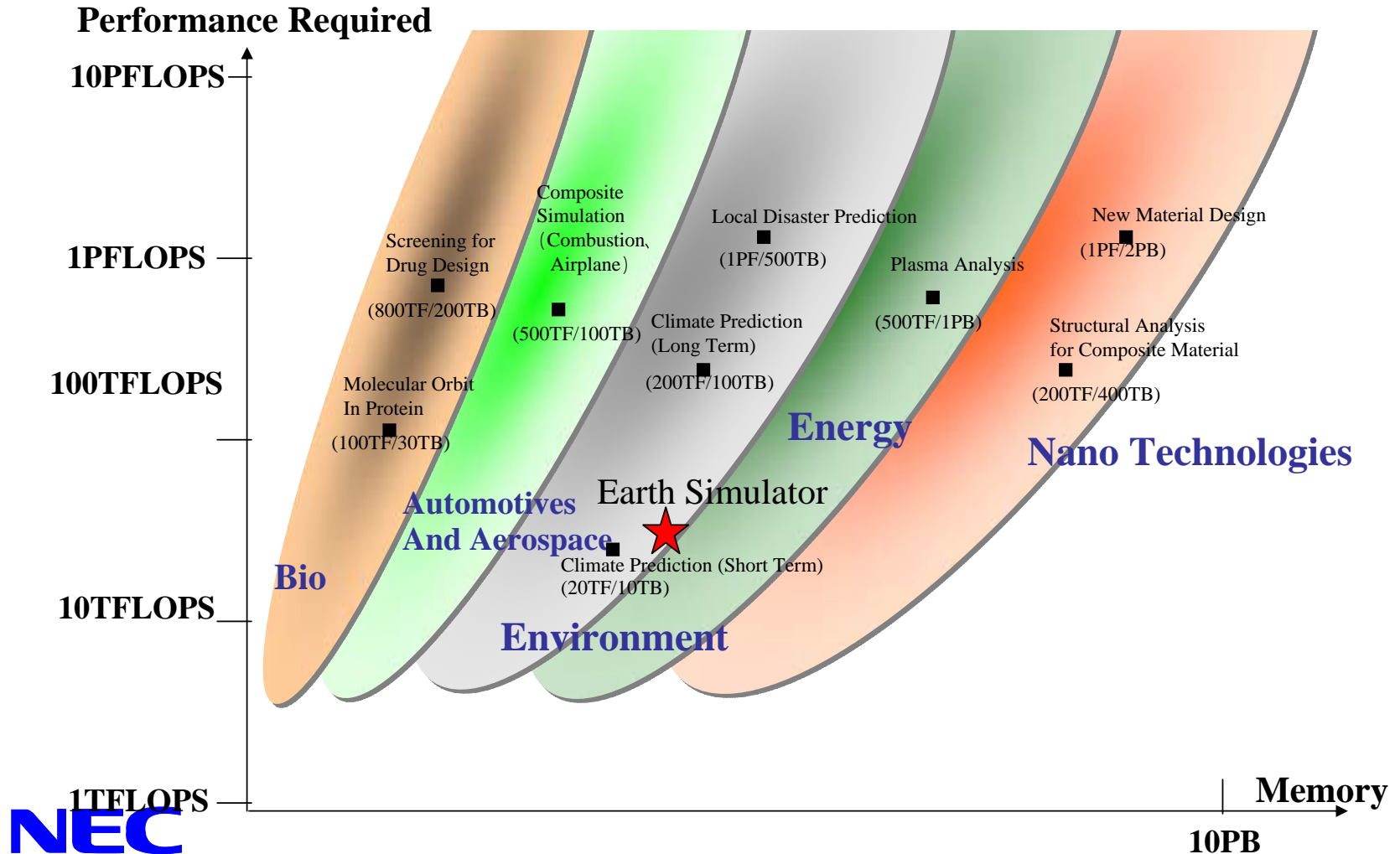
Will this technological evolution continue?

Are there any problems or difficulties to overcome?

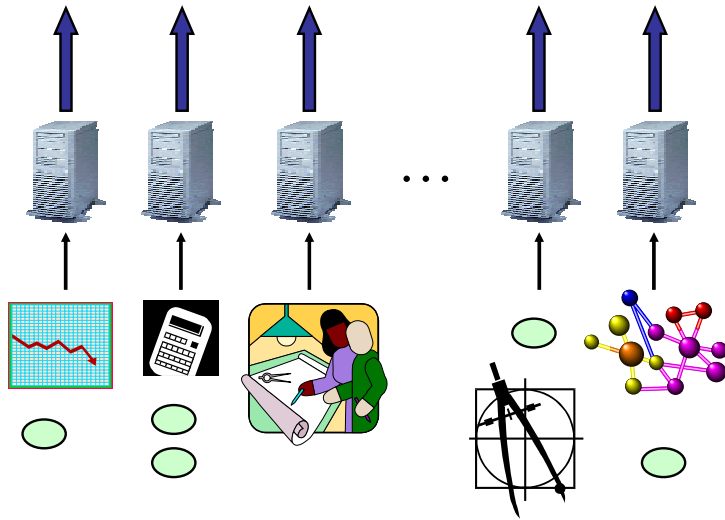
If so, what are they?

Do We Need a Peta Flops Computer?

Application Areas and Required Performance



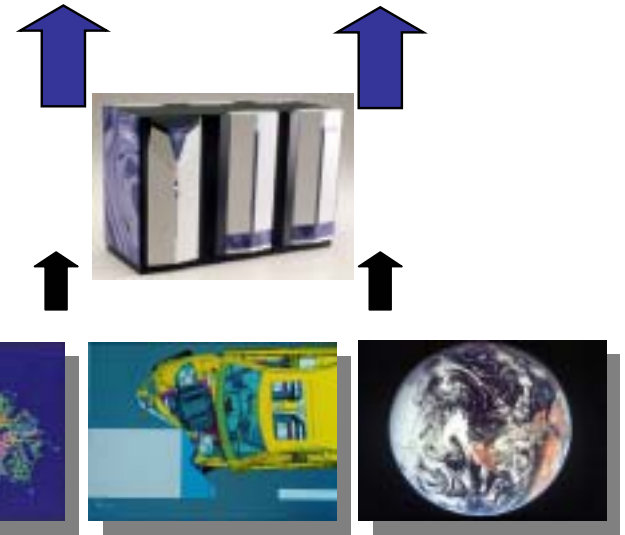
Capacity Computing and Capability Computing



PC Cluster / Blade Server

Capacity Computing

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

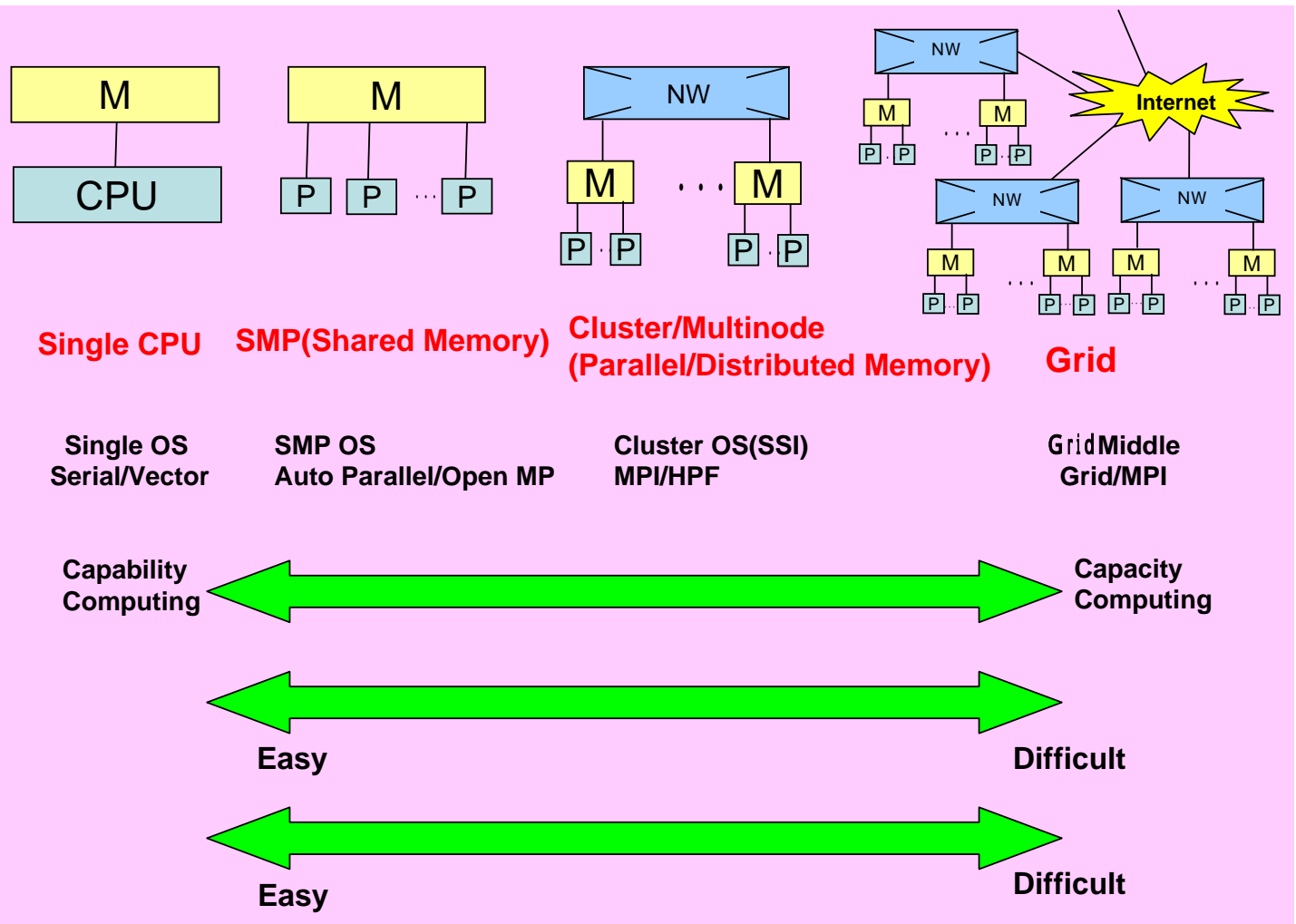


Vector / SX

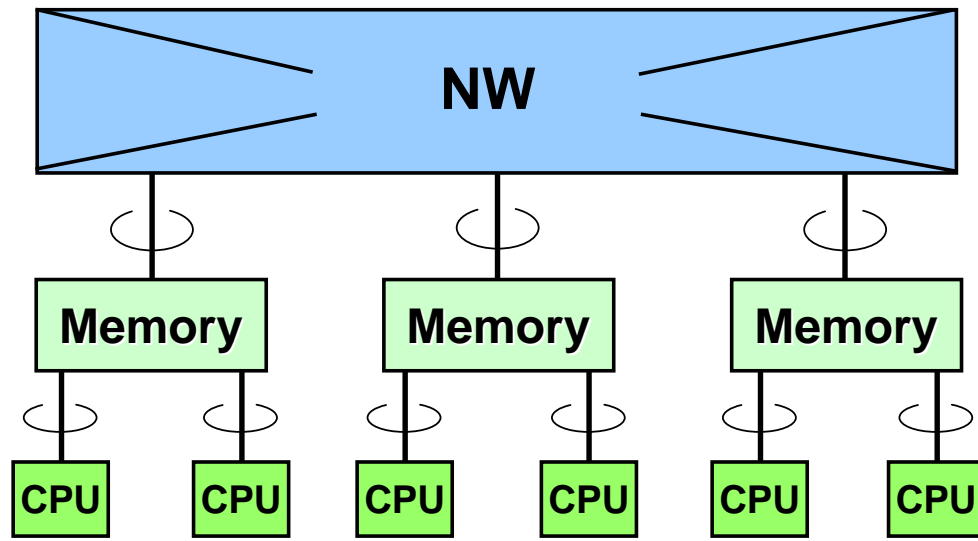
Capability Computing

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

System Configuration and User View



Highly Efficient Capability Computing



- Low Latency
- High Throughput

High Bandwidth Interface



Powerful CPU



Capability Computing ~ To Increase Sustained Performance ~

Technological Issues

- **Amdahl's Law**

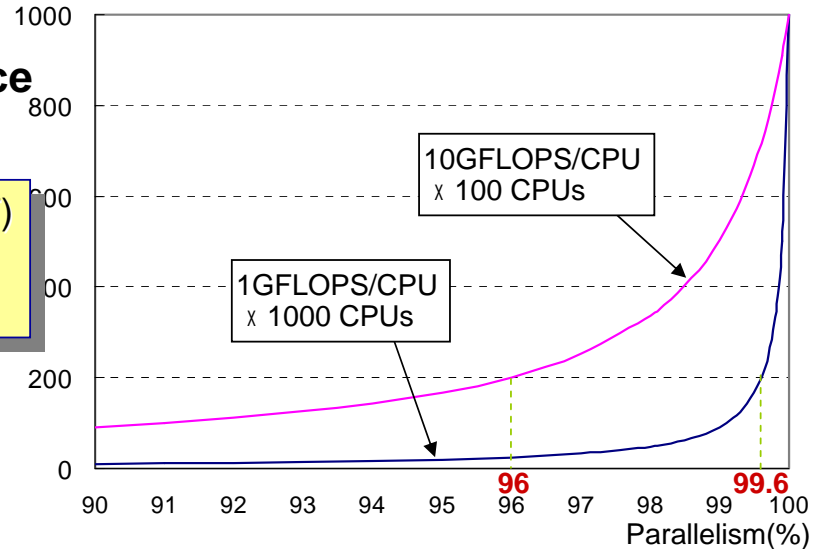
Importance of Single CPU's Performance



(Ex.) To achieve 20% efficiency of Peak Speed (1TF)

- 10GF/CPU x 100CPUs → 96% of Parallelism
- 1GF/CPU x 1000CPUs → 99.6% of Parallelism

Performance (GFLOPS)



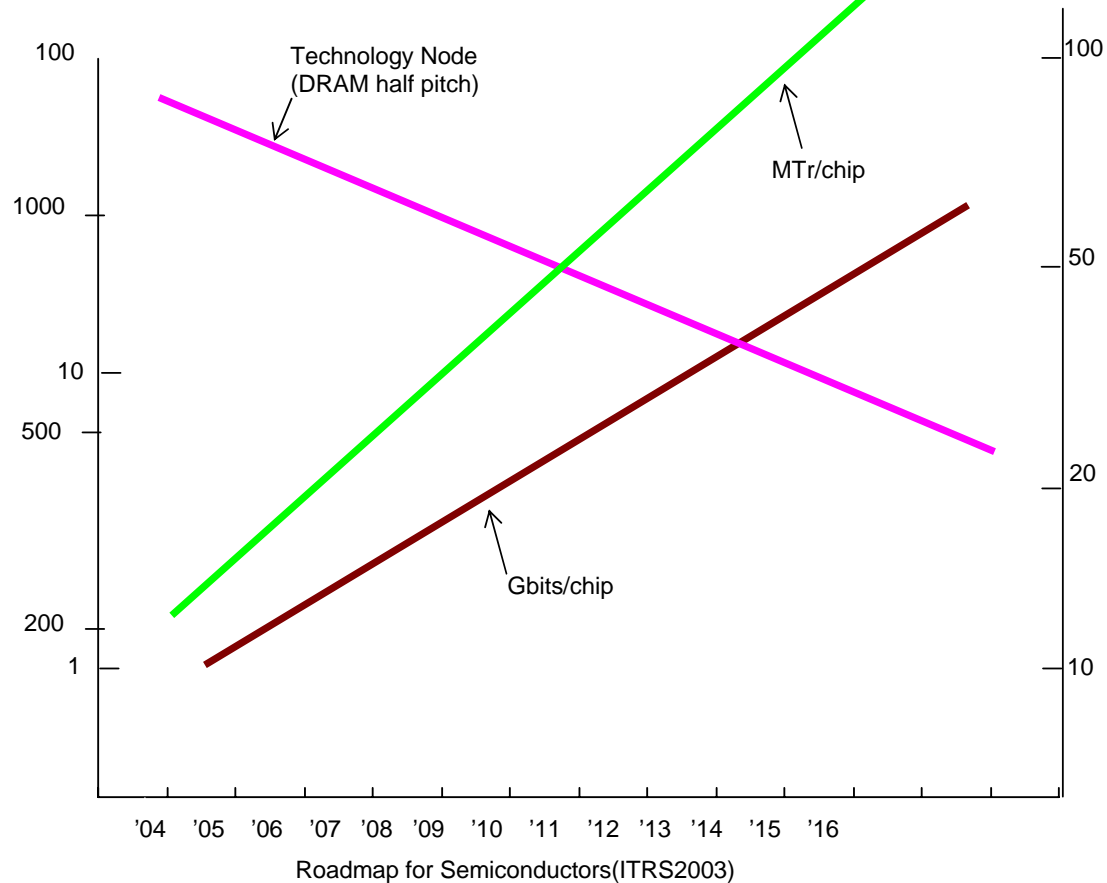
- **To Increase CPU Performance**

- Device (LSI) Technology
- Memory Performance (Bandwidth)

- **To Increase Performance of Parallel Processing**

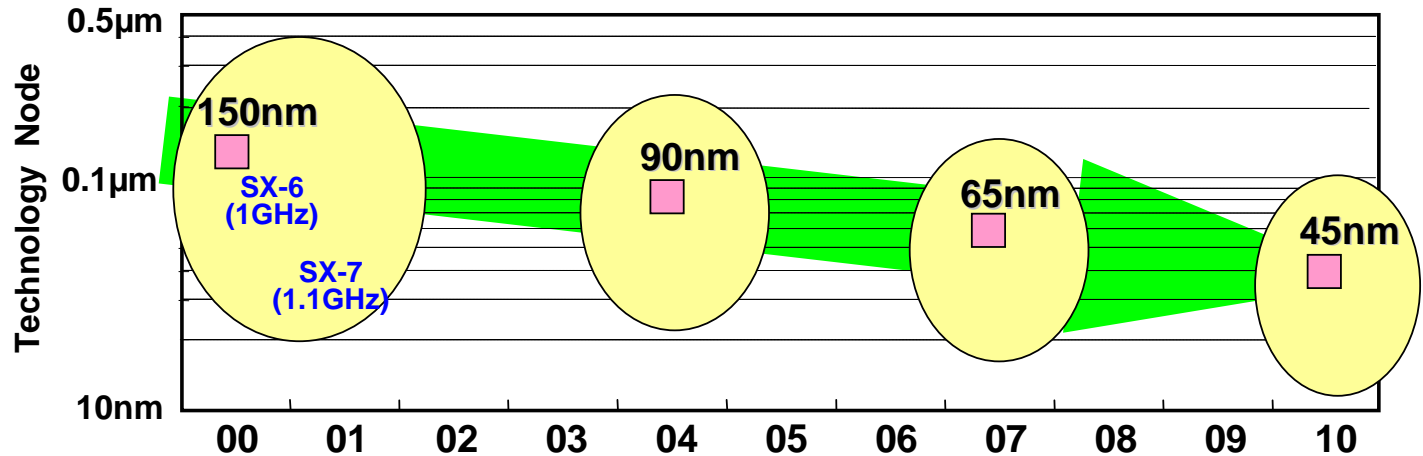
- High Scalability and High Efficiency by High Speed CPU
- Small Scale Parallel Processing: High Bandwidth SMP
- Large Scale Parallel Processing: High Performance Communication (MPI)
High Speed Synchronization Mechanism

Road Map of Semiconductors



Device Technology

Road Map of LSI CMOS Process



International Technology Roadmap
for Semiconductors (ITRS) 2003

Technological Issues

Higher Density
Ultra-finer Elements

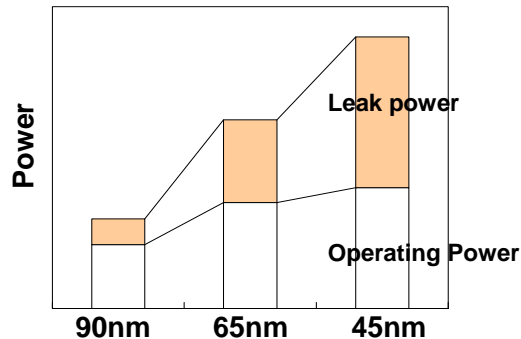


Problems to Overcome

- Lower Power → High-k
- High Speed → Low-k
- Lower Voltage → Finer Process

Power Reduction of LSI

Increase of Power Consumption



$$\text{Power} = \underbrace{C \cdot V_{dd}^2 \cdot f}_{\text{On Power}} + \underbrace{(I_{\text{off leak}} + I_{\text{gate leak}}) \cdot V_{dd}}_{\text{Off Power}}$$

Increase of operating Power due to Speed Increase and Finer Process

- Performance → Frequency (f)
- Pattern Pitch → Capacitance (C)

《Counter Plan》

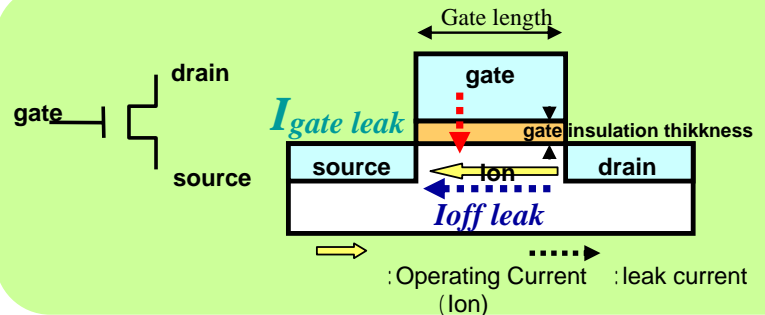
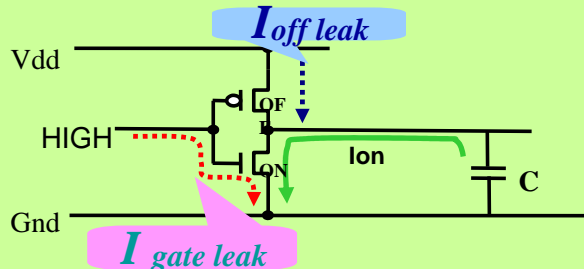
- Operating ratio by clock gating ()
- Low material Capacitance by Low material(C)

Increase of Leak Current due to Finer Process

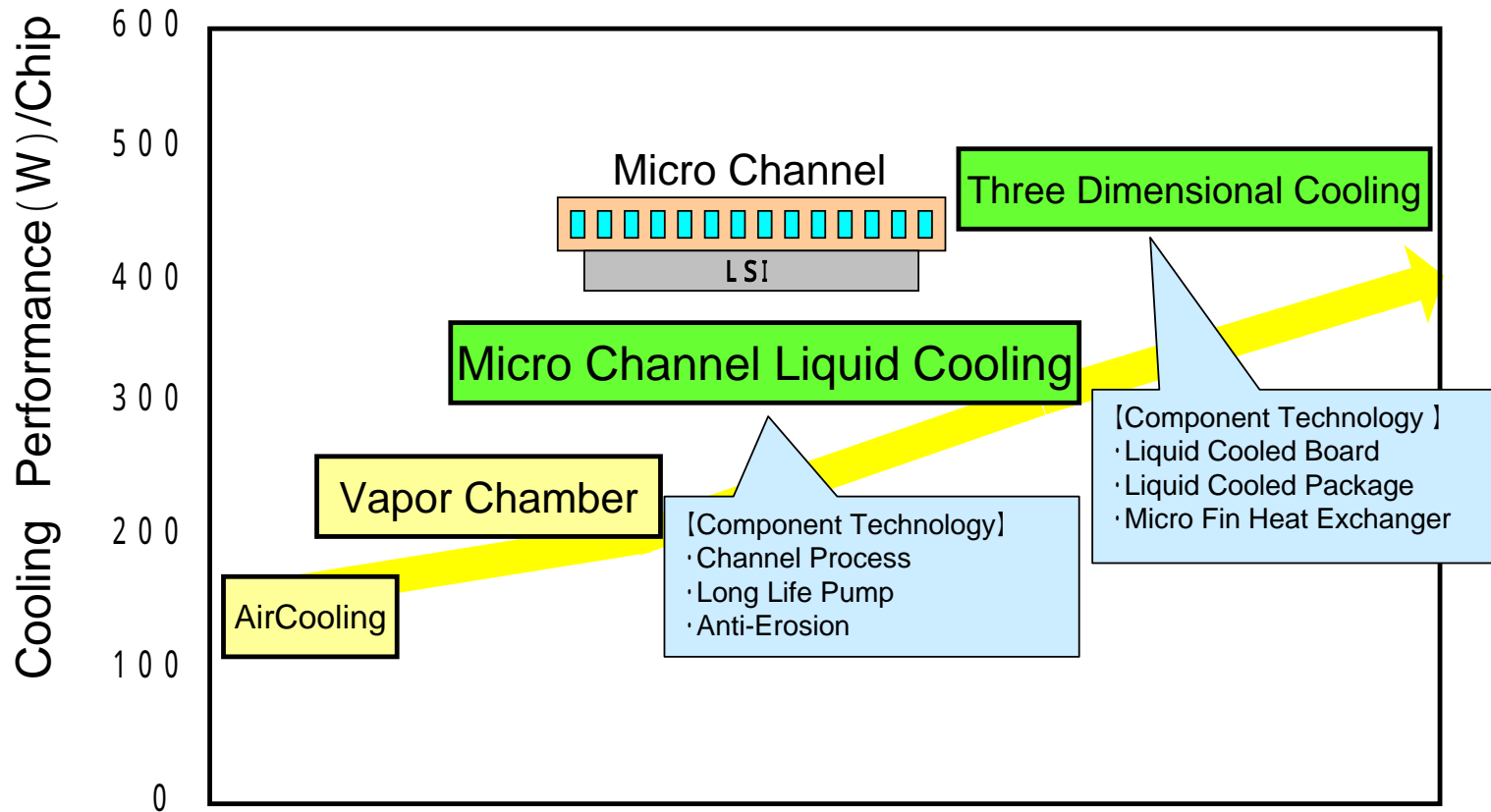
- Gate Length → $I_{\text{off leak}}$
- Gate Insulation → $I_{\text{gate leak}}$

《Counter Plan》

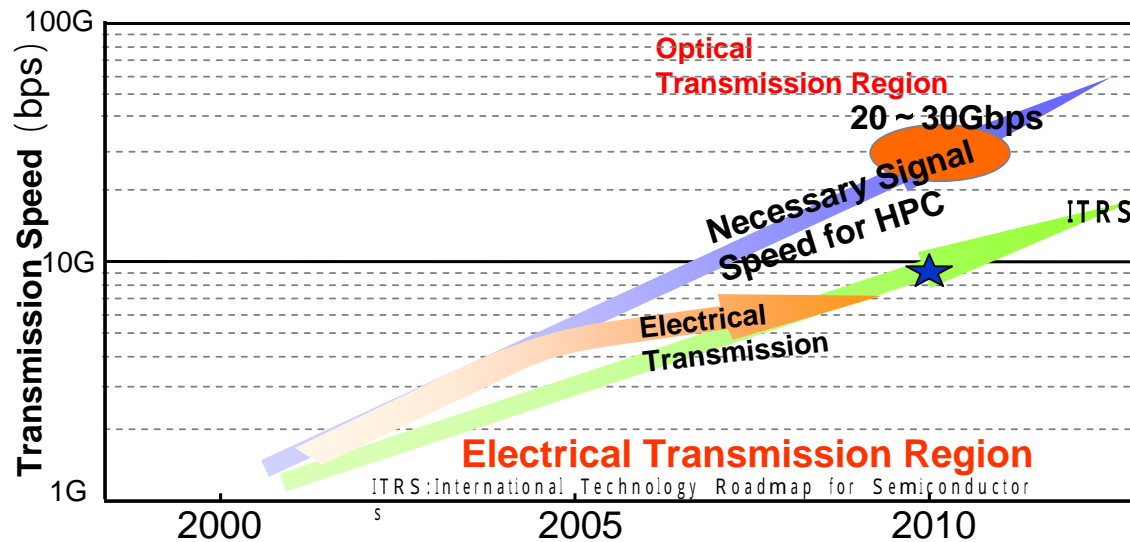
- Vt dynamic control (Stand by Vt) : $I_{\text{off leak}}$
- High gate insulation material : $I_{\text{gate leak}}$



Cooling Technology



Signal Transmission



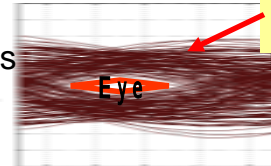
Electrical Signal *Transmission Signal of 20 Gbps*

20Gbps
1 m



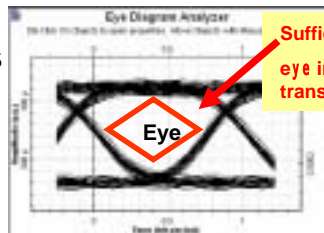
No eye in electrical transmission

20Gbps
5 cm



Optical Signal

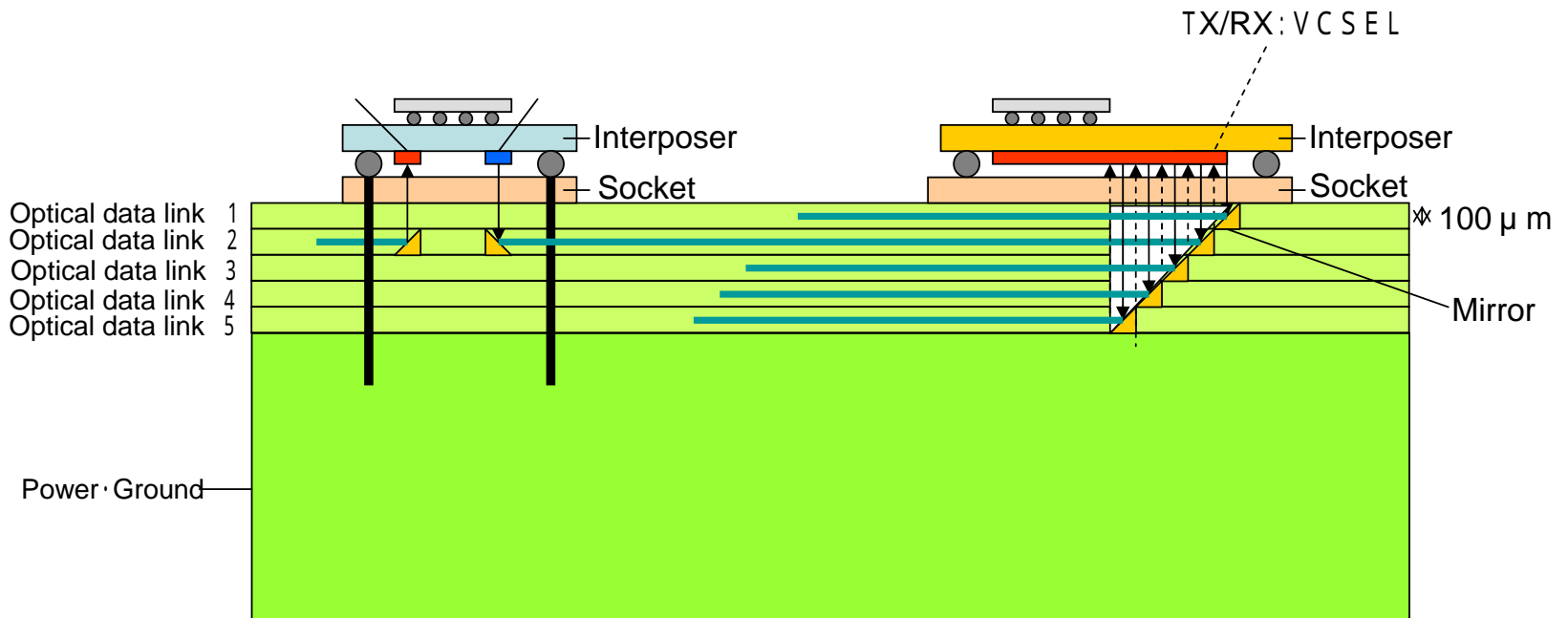
20Gbps
1 m



Sufficient eye in optical transmission

Optical Interconnection

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

$$M \quad (P)^{3/4} \quad \simeq \quad \text{GB/GFLOPS}$$

- Commodity Product such as Media Processor, Home/Industry Equipment

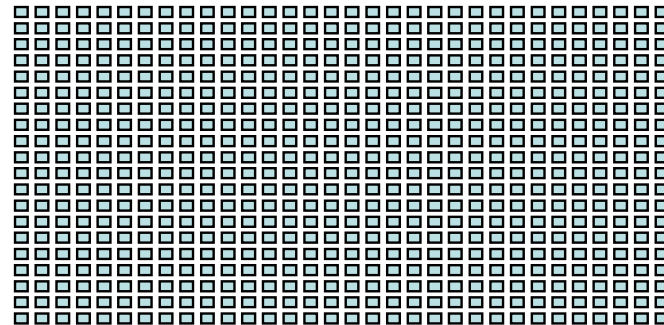
- **μ -P Core +Special Engines**

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

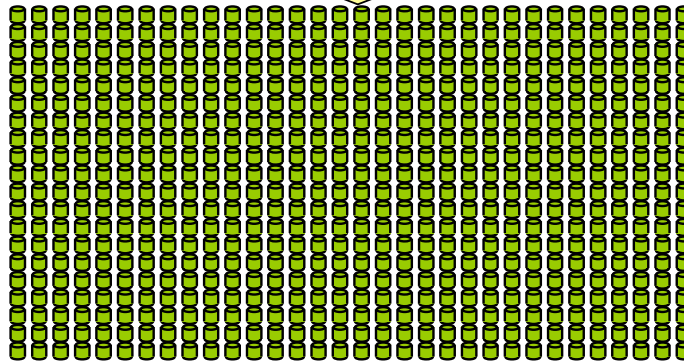
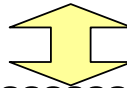
- **μ -P Core+Vector Engine/Multiple μ -P Cores**

- HPC for Scientific/Engineering Use
 - High-end/Affordable HPC
- (μ -P Core : VLIW/Superscalar/Multithread)

Challenges in Software



>10,000 CPUs

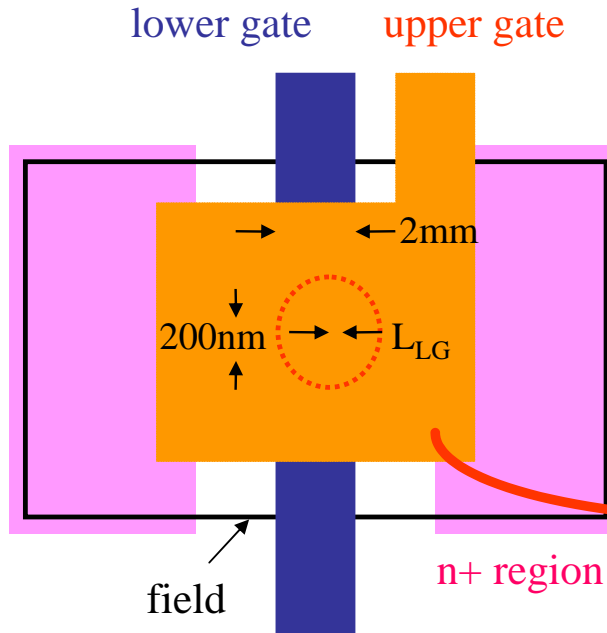


>1 Peta Bytes Storage

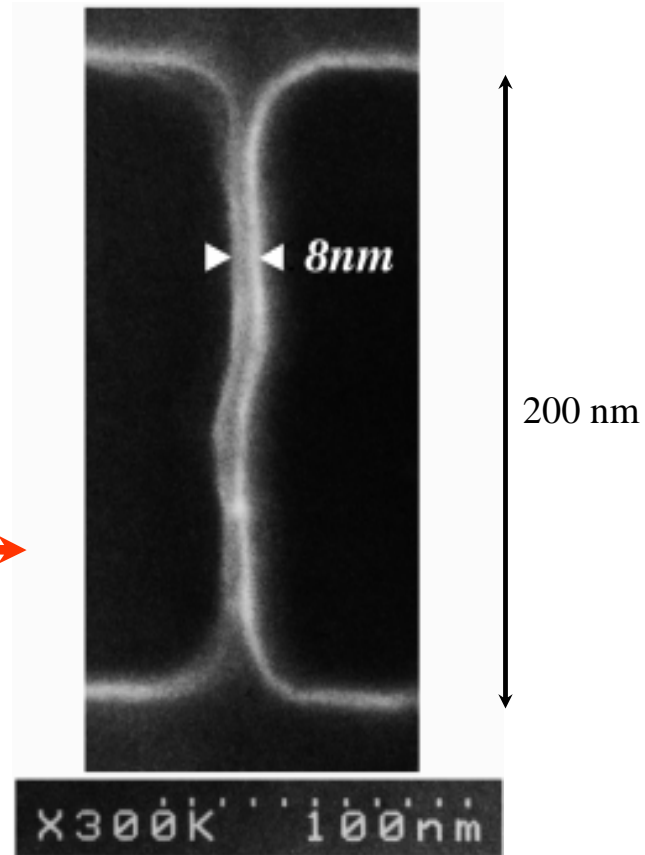
- Operation and Resource Management
- Huge Volume of Data Management
- Reliability, Availability and Serviceability
- Support of Development Environment(Compiler and Tools)
for Ultra Large Scale Parallel Processing System

Post Silicon

Top View of an EJ-MOSFET

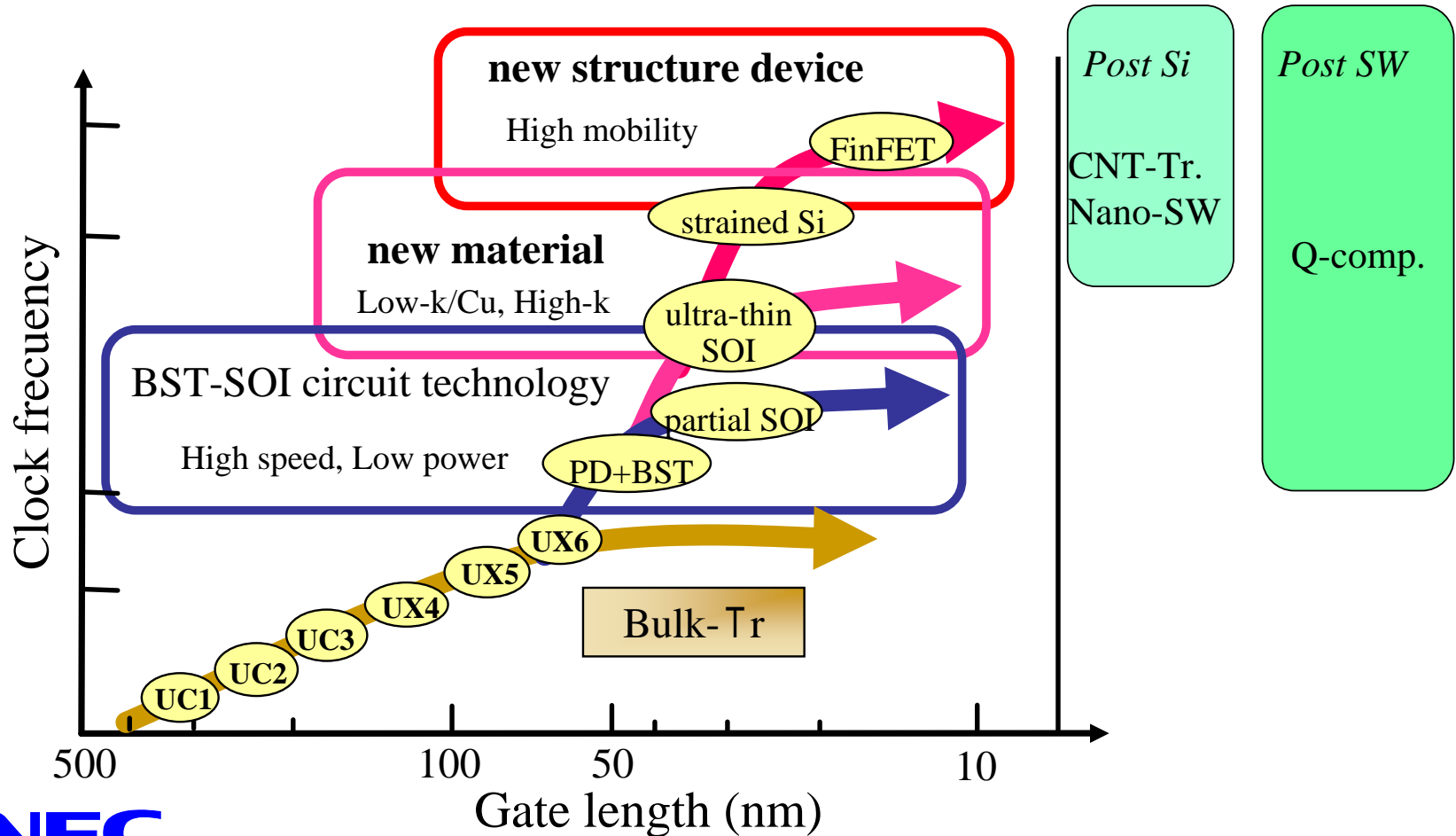


$$8\text{nm} < L_{LG} < 100\text{ nm}$$
$$T_{ox}=5\text{nm}$$



SEM image

Post Silicon & Post Switch ???



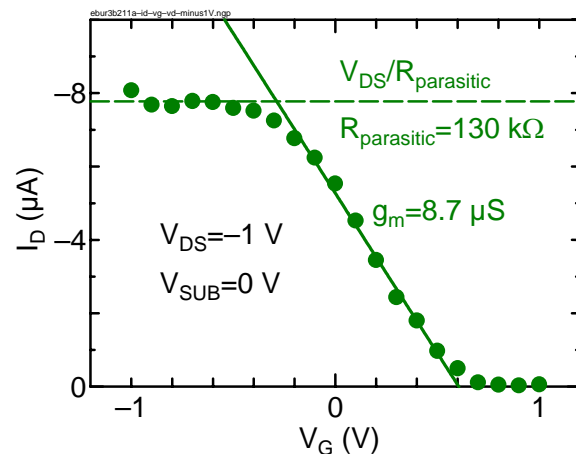
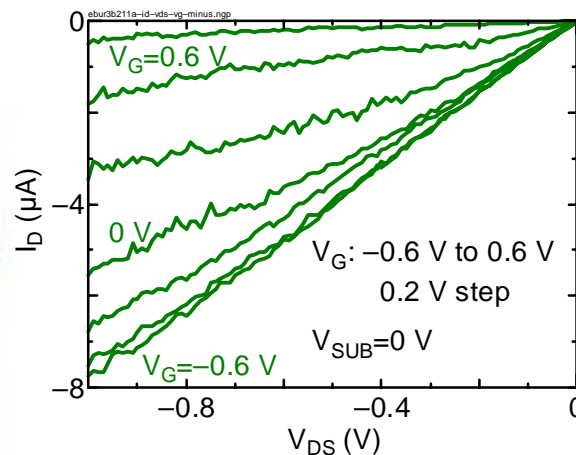
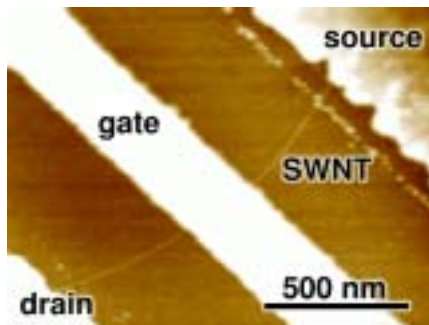
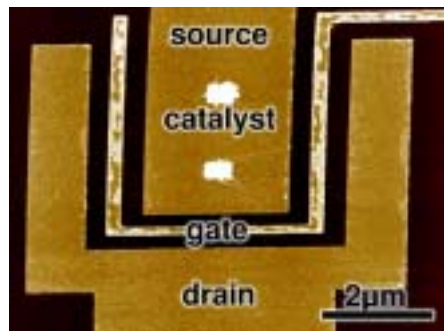
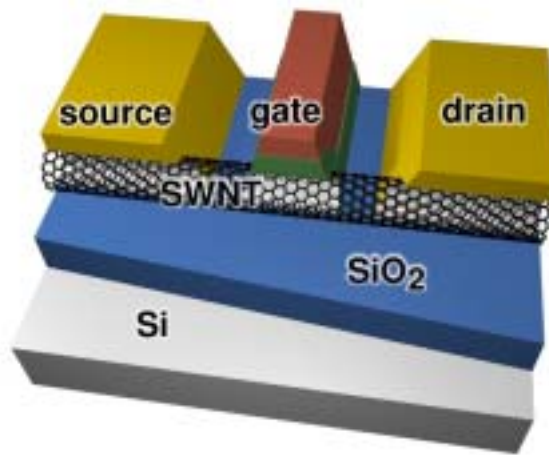
Carbon-Nanotube Field-Effect Transistors

- Possible application: low-cost, low-power LSI, rf drivers
- Position-controllable on-wafer growth (catalyst CVD)
- Extremely high transconductance:

$$g_m = 8.7 \text{ mS/tube}$$
$$(5800 \text{ } \mu\text{S/mm})$$

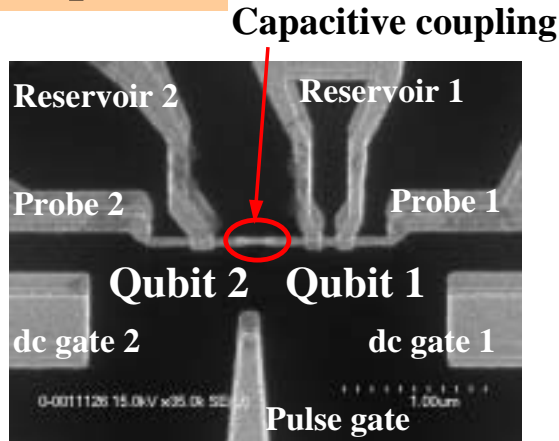
Si nFET: 1000~1200 $\mu\text{S}/\mu\text{m}$

pFET: 400~600 $\mu\text{S}/\mu\text{m}$



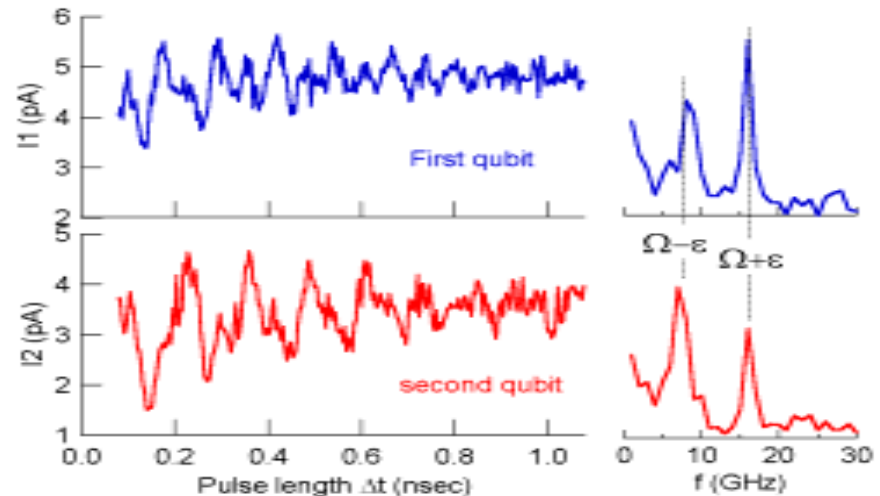
Quantum Entangled State in a Solid State Device

Sample



Superconductor-Based Device

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

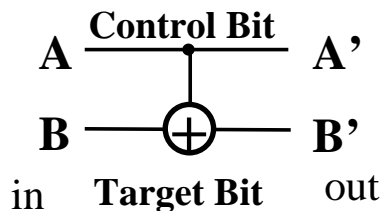


Next Step

Fundamental 2 Bit Logic Gate Operation (C-NOT)

Provides Universal Gate, combined with 1 Bit Gate

2 Bit Logic Gate (Controlled-NOT)



A	B	A'	B'
0	0	0	0
0	1	0	1
1	0	1	1
1	1	1	0

Truth Table

Control Bit Target Bit C-NOT Operation

$$\begin{aligned} |10\rangle &\longrightarrow |11\rangle \\ |00\rangle &\longrightarrow |00\rangle \end{aligned}$$

Target bit is flipped only if control bit is 1

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

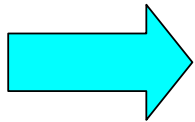
Post Silicon

CNT Tr., Atomic switch

Quantum computing

What will be the Future?

**I Believe the Evolutional Development
in these 10 years.**



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

Conclusion

—**Nine Lessons Learned in the Design of CDC6600(N.R.Lincoln)**—

**It's Really not as much Fun Building a Supercomputer as it is
Simply Inventing One**

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.**