Meeting of the Advanced Scientific Computing Advisory Committee
October 17 and 18, 2002
Hilton Washington Embassy Row Hotel
2015 Massachusetts Avenue, NW, Washington, DC

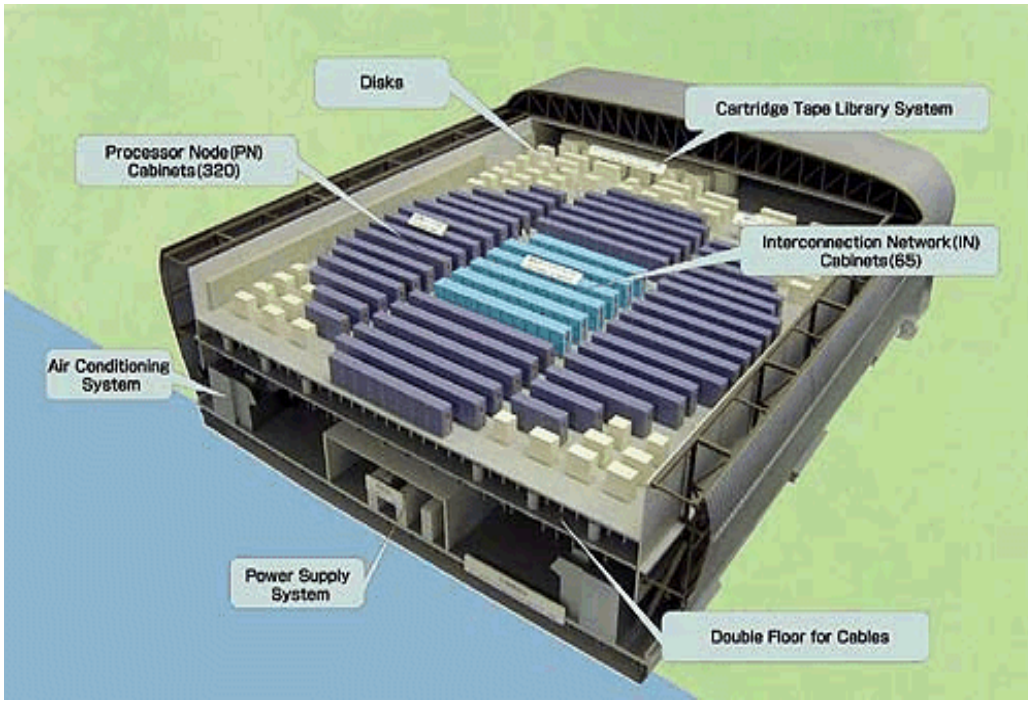
It's a Sign, Not a Race

Jack Dongarra

University of Tennessee

A Tour d'Force for a Supercomputer System

- Japanese Jaeri/Jamstec/Nasda/Riken Earth Simulator (2002)
- Target Application: CFD-Weather, Climate, Earthquakes
- 640 NEC SX/6 Nodes (mod)
  - 5120 CPUs which have vector ops
- 40TeraFlops (peak)
- 7 MWatts (ASCI White: 1.2 MW; Q: 6 MW)
  - Say 10 cent/KWhr - $16.8K/day = $6M/year!
- $250-500M for things in building
- Footprint of 4 tennis courts
- Expect to be on top of Top500 until 60-100 TFlop ASCI machine arrives
- Homogeneous, Centralized, Proprietary, Expensive!

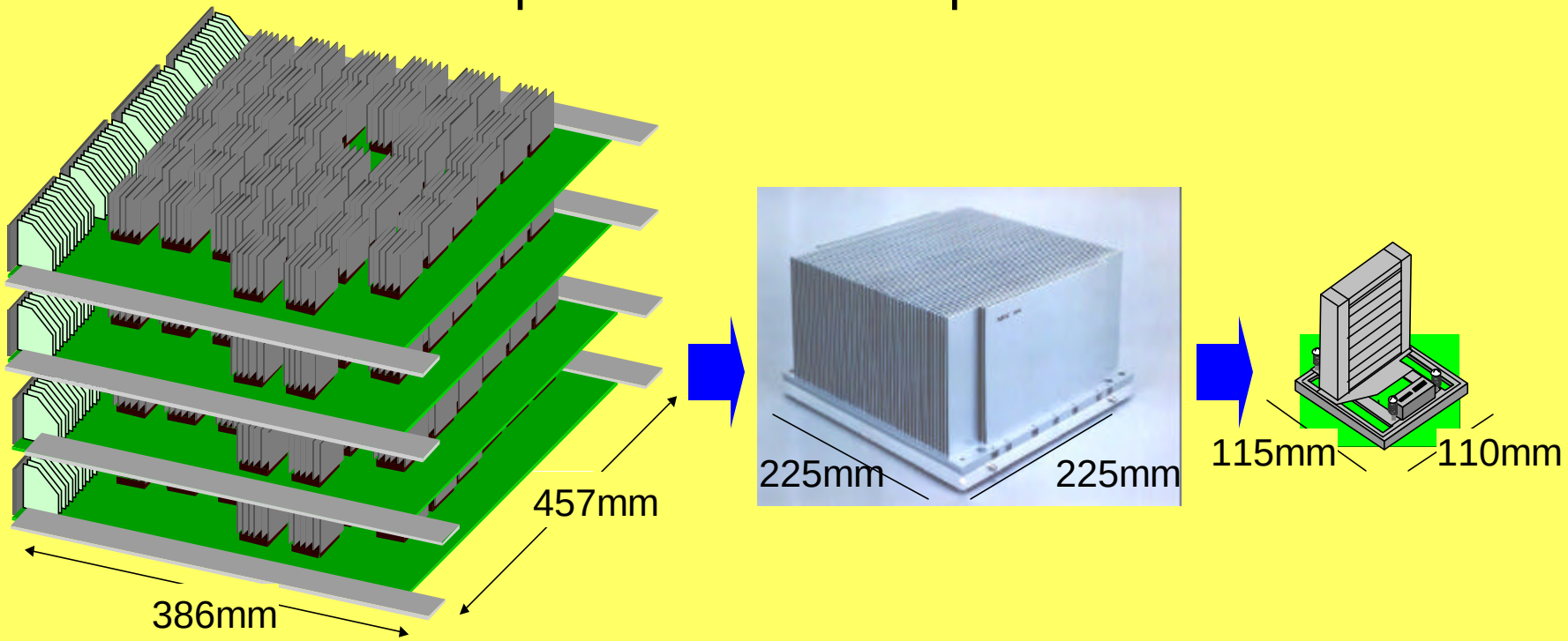


(Earth Simulator Picture from JAERI web page)

NASDA : National Space Development Agency of Japan
JAMSTEC : Japan Marine Science and Technology Center
JAERI : Japan Atomic Energy Research Institute
RIKEN : The Institute of Physical and Chemical Research

R&D results Not Revolutionary

Comparison of vector processors



SX4 (1995)	SX5 (1998)	Earth Simulator (2002)
2 GFlop/s 8 vector pipes Clock :125MHz LSI: 0.35μ m CMOS 37x4=148 LSIs	8 GFlop/s 16 vector pipes Clock :250MHz LSI: 0.25μ m CMOS; 32 LSIs	8 GFlop/s 8 vector pipes Clock :500MHz LSI: 0.15μ m CMOS 1 chip processor

Specifications of the SX-7

TOKYO, October 9, 2002 -

I. Single Node System

Central Processing Unit (CPU)	
Number of CPUs	4 ~ 32
Vector Performance	35.3 ~ 282.5GFLOPS
Vector Register	144k bytes x4 ~ 32
Scalar Register	64bits x128 x4 ~ 32
Main Memory Unit	
Memory Architecture	Shared Memory
Capacity	32 ~ 256G Bytes
Maximum Transfer Rate	1,130.2 G Bytes/Sec.
Input/Output Processor (IOP)	
Number of IOP	1 ~ 4
Maximum Channel	127 channels
Maximum Transfer Rate	8 G Bytes/Sec.

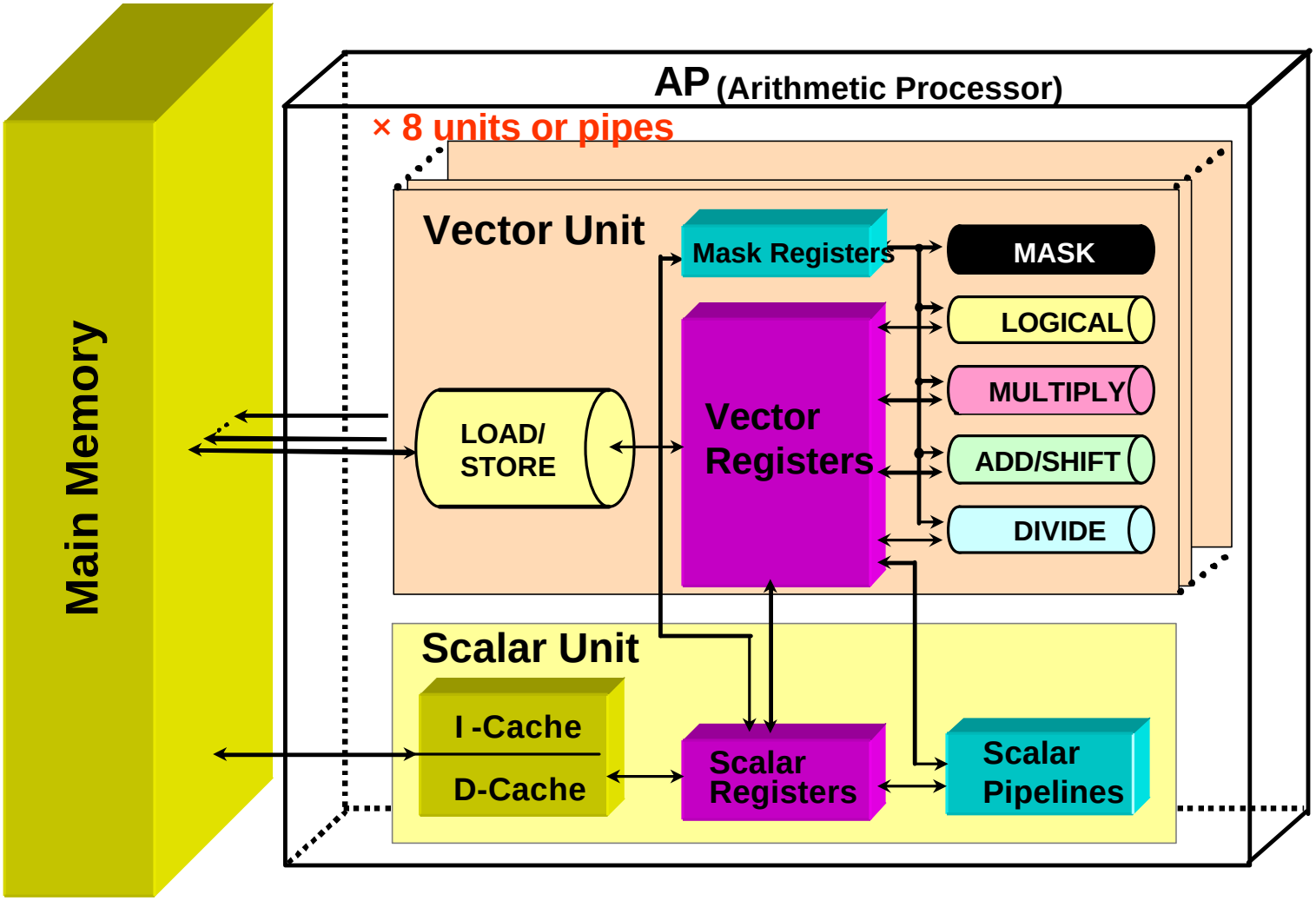
II. Multi-node System

Number of Nodes	2 ~ 64
Central Processing Unit	
Number of CPUs	16 ~ 2,048
Vector Performance	141.2G ~ 18,083GFLOPS
Vector Register	144k bytes x16 ~ 2,048
Scalar Register	64bits x128 x16 ~ 2,048
Main Memory Unit	
Memory Architecture	Shared/Distributed Memory
Capacity	128G ~ 16T Bytes
Maximum Transfer Rate	Max.72T Bytes/Sec.
Input/Output Processor (IOP)	
Number of IOP	Max.256
Maximum Channel	Max. 8,128channels
Maximum Transfer Rate	Max. 512G Bytes/Sec.
Internode Crossbar Switch (IXS)	
Maximum Transfer Rate	Max. 512G Bytes/Sec.

- 553 MHz
  - 8 vector pipes
  - 8.85 GFlop/s / proc
  - 32 proc / node
- 64 nodes
  - 18 TFlop/s / node
  - high-speed crossbar
- Ship Dec 2003
- NEC SX-8
  - on the horizon

Outline of the system

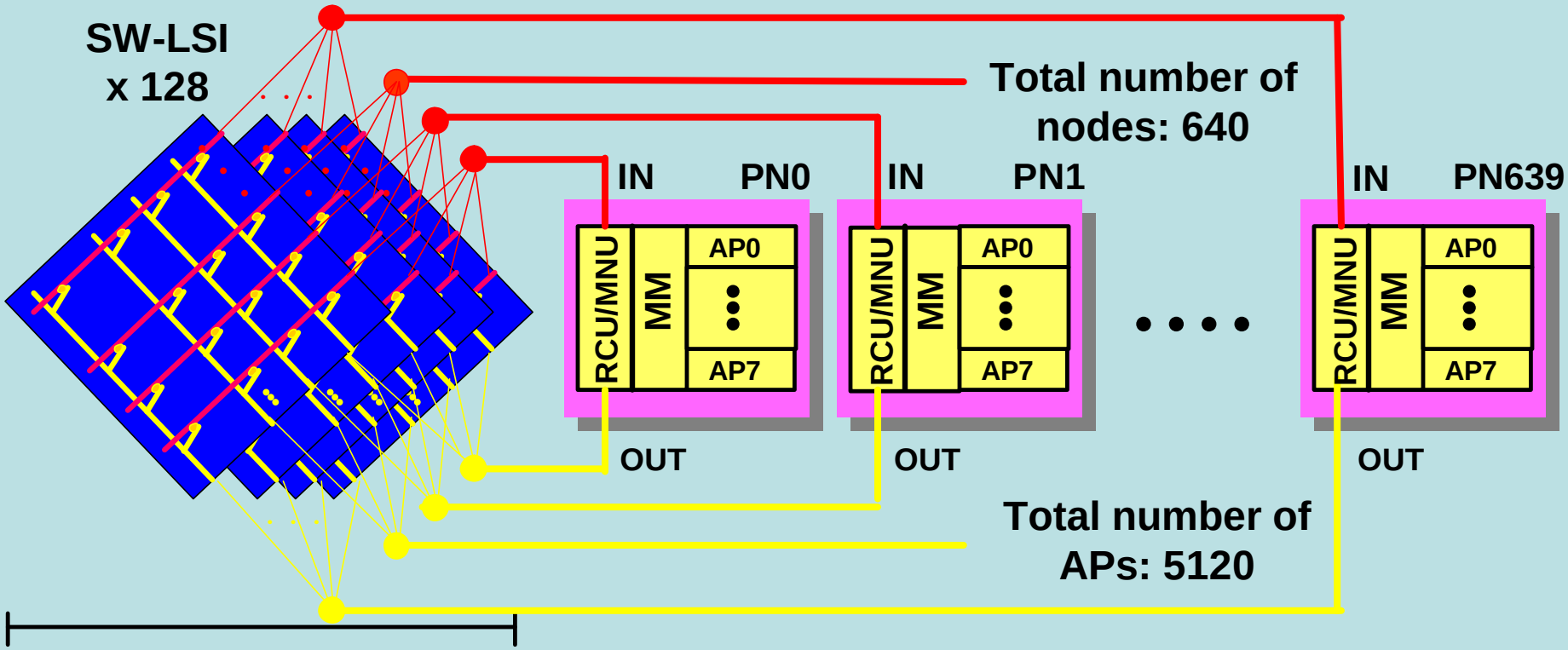
Block diagram of arithmetic processor (single chip)



Outline of the system

Configuration of the hardware system of the Earth Simulator

Total peak performance : 40 TFlop/s
Total main memory : 10 TB



Single-Stage Crossbar Network

: Fast Switch 1Gbps/switch
Total performance :
1 Gbps/switch x 128 swiches/8 =16 GB/s

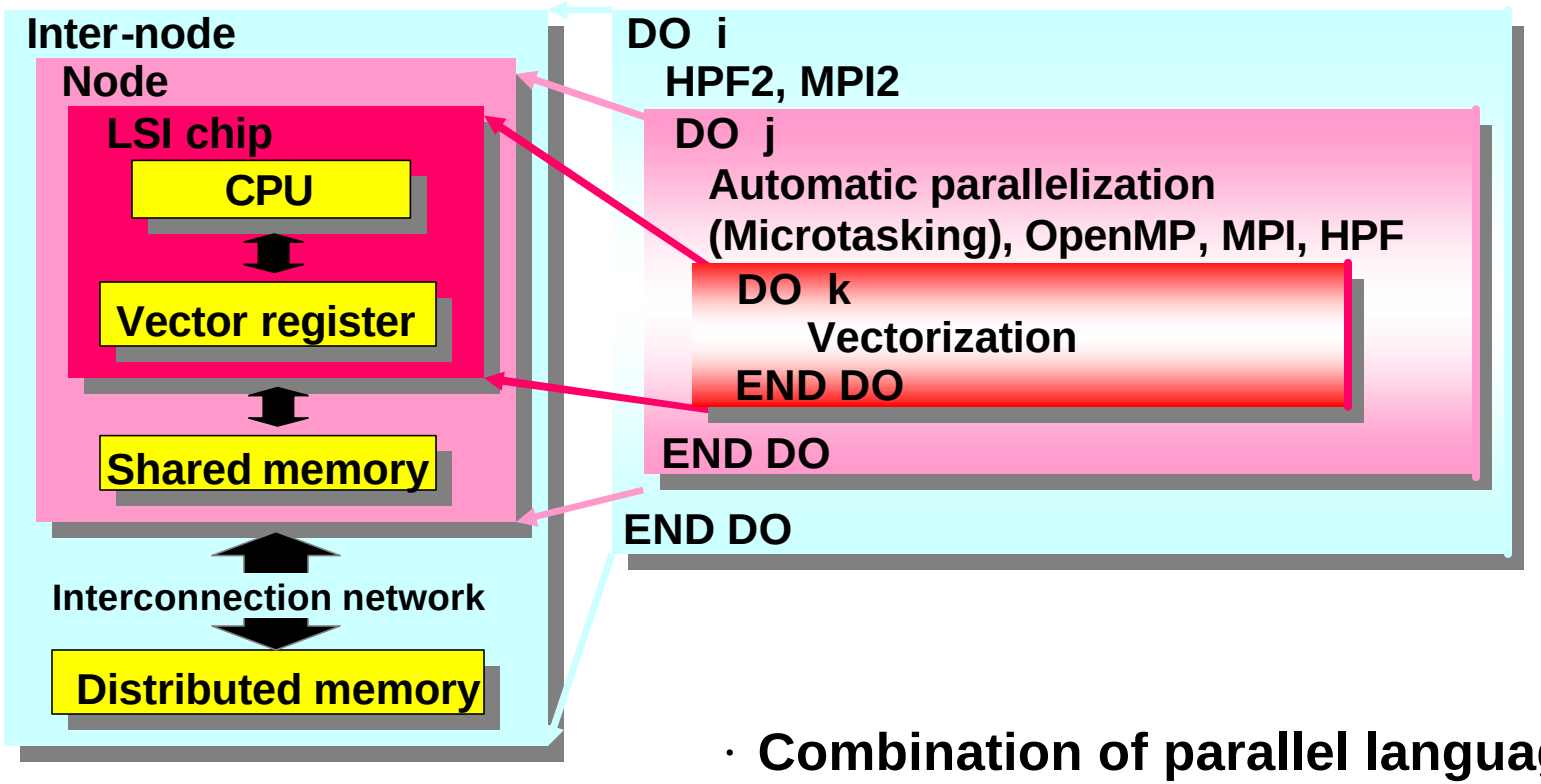
(12.3 GB/s)

PN : Processor Node (multi-processor node)
AP : Arithmetic Processor 8 GFlop/s (peak)
MM : Main Memory (sheared) 16 GB
RCU: Remote Control Unit
MNU : Memory Network Unit

Outline of the system

Parallel Programming Language

- Memory hierarchy
- Program languages and optimization



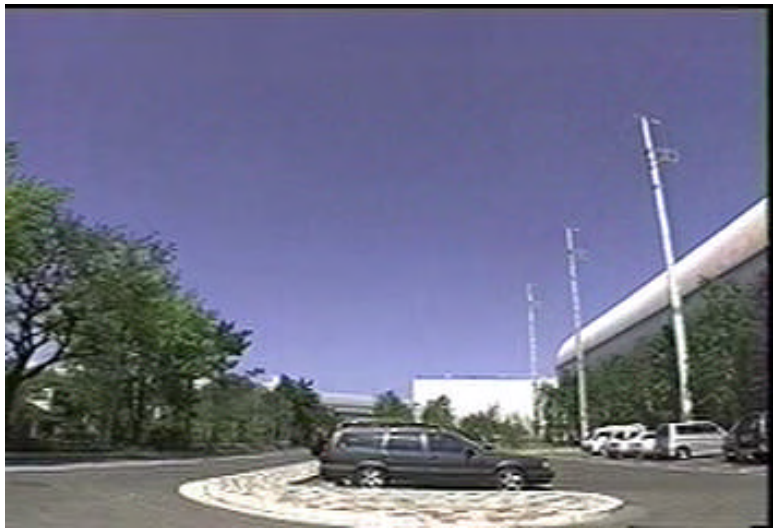
· **Combination of parallel languages**

Within node / Inter-node	Auto. Para.	OpenMP	MPI	HPF
MPI	Recommended	**Available**	Available	Not available
HPF	Recommended	Procedure call	Procedure call	Available

Earth Simulator Factoids

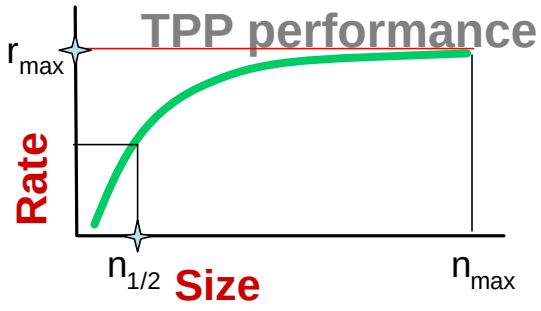
- Memory:
  - 16 GB per node or 10 TB memory in the system
- Memory bandwidth (from memory to V-regs):
  - 32 GB/s from memory to register per processor, or
    256 GB/s from memory for a node, or
    163 TB/s from memory for the system.
- Data movement through the switch:
  - 12.3 GB/s crossbar bi-directional between nodes or cross section bandwidth of 8 TB/s
  - MPI_PUT

    11.63 GB/s and 6.63 µs latency
- Disk: ~600 TB
  - 50 GB/s, .02 sec latency
- Cartage tape: 1 – 15 PB



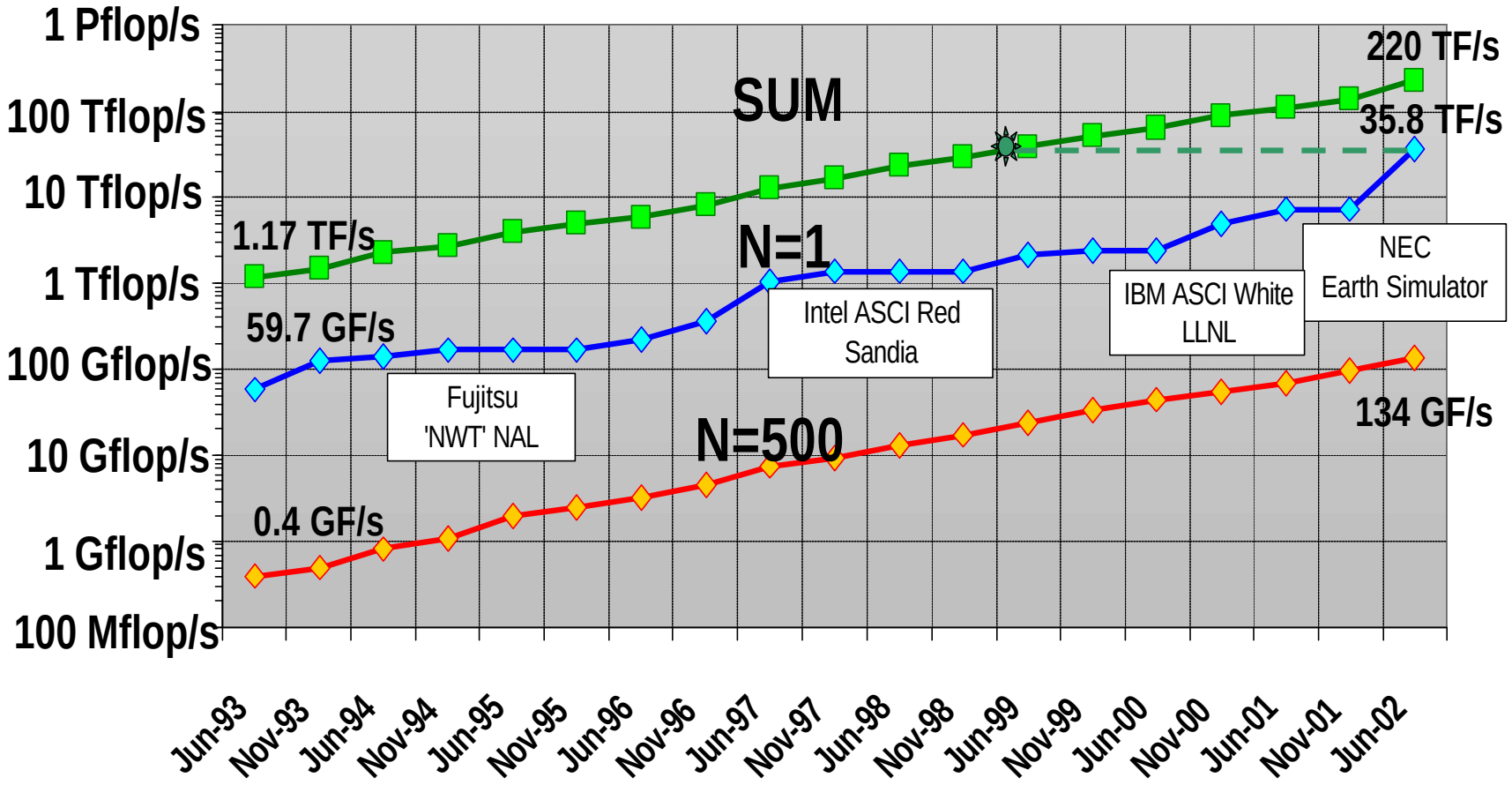
Earth Simulator Computer (ESC)

- Rmax from LINPACK MPP Benchmark ***Ax=b, dense problem***
  - Linpack Benchmark = 35.86 TFlop/s, 87.5% efficiency
  - Problem of size n = 1,041,216; (8.7 TB of memory)
  - Size at half of peak ($n_{½}$) achieved at $n_{½}$ = 265,408
  - Benchmark took 5.8 hours to run.
  - Algorithm: LU w/partial pivoting
  - Software: for the most part Fortran using MPI



- For the Top500 (June 2002)
  - S of all the DOE computers (DP + OS) = 27.5 TFlop/s
  - Performance of ESC ~ 1/6 S(Top 500 Computers)
  - Performance of ESC > S(Next Top 12 Computers)
  - Performance of ESC > S(Top 15 Computers in the US)
  - Performance of ESC > All the DOE and DOD machines (37.2 TFlop/s)
  - Performance of ESC >> the 3 NSF Center's computers (7.5 TFlop/s)

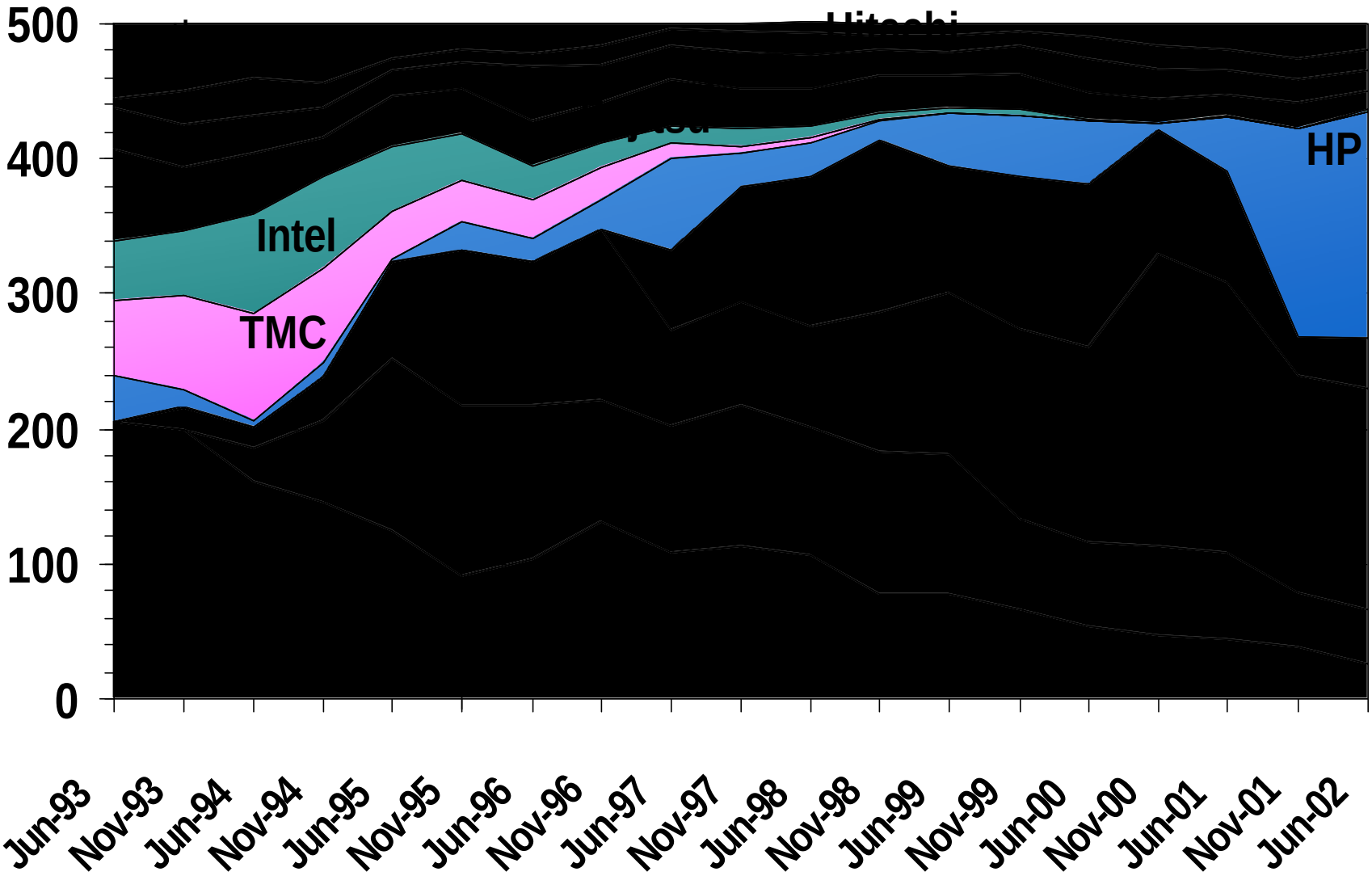
TOP500 - Performance



Machines at the Top of the List

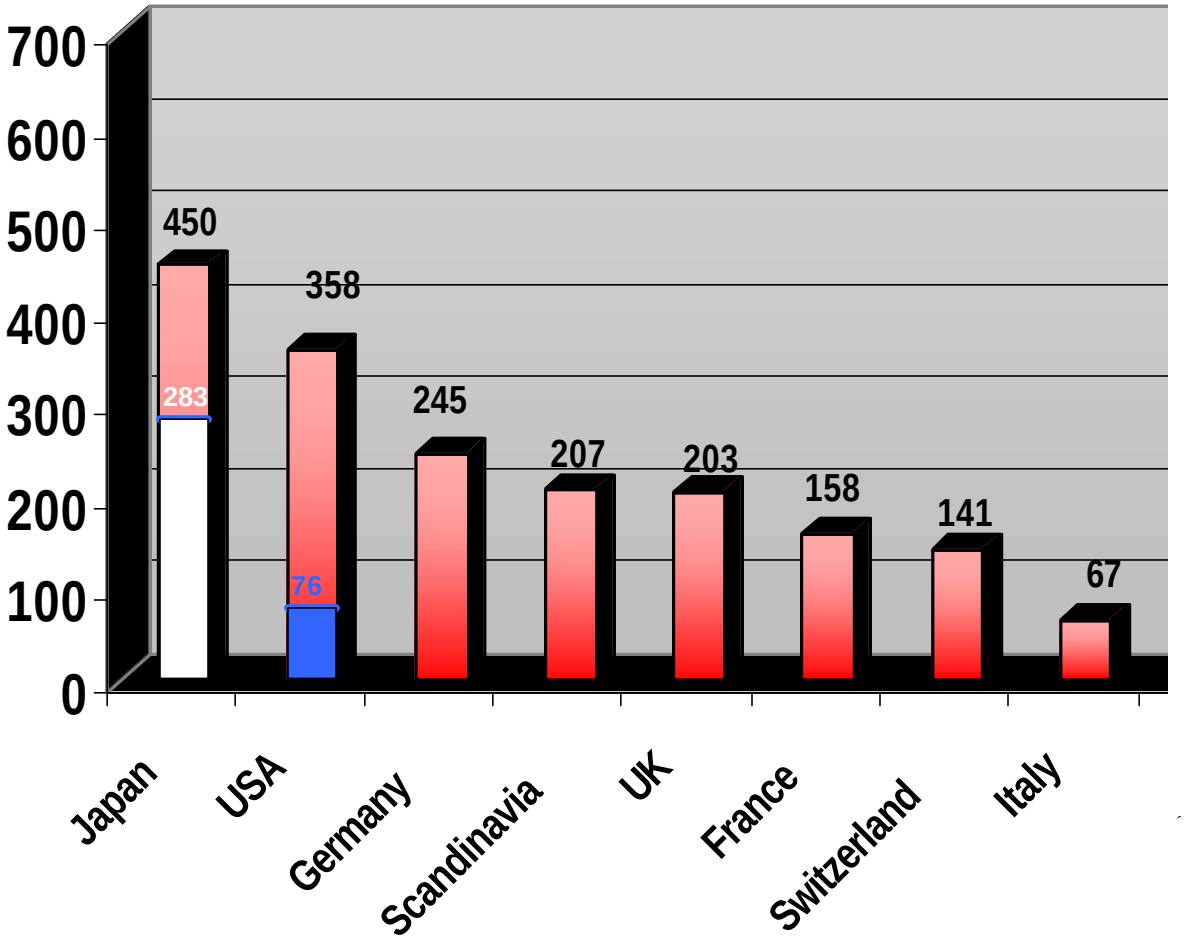
Year	Computer	Measured Gflop/s	Factor ? from Pervious Year	Theoretical Peak Gflop/s	Factor ? from Pervious Year	Number of Processors	Efficiency
2002	Earth Simulator Computer, NEC	35860	5.0	40832	3.7	5104	87%
2001	ASCI White-Pacific, IBM SP Power 3	7226	1.5	11136	1.0	7424	65%
2000	ASCI White-Pacific, IBM SP Power 3	4938	2.1	11136	3.5	7424	44%
1999	ASCI Red Intel Pentium II Xeon core	2379	1.1	3207	0.8	9632	74%
1998	ASCI Blue-Pacific SST, IBM SP 604E	2144	1.6	3868	2.1	5808	55%
1997	Intel ASCI Option Red (200 MHz Pentium Pro)	1338	3.6	1830	3.0	9152	73%
1996	Hitachi CP-PACS	368.2	1.3	614	1.8	2048	60%
1995	Intel Paragon XP/S MP	281.1	1	338	1.0	6768	83%
1994	Intel Paragon XP/S MP	281.1	2.3	338	1.4	6768	83%
1993	Fujitsu NWT	124.5		236		140	53%

Manufacturers



HP 168 (12 < 100), IBM 164 (47 < 100)

Kflops per Inhabitant



White is ES contribution and Blue is ASCI contribution

Introduction

Slide from Keiji Tani from ESC

- Necessary type of computer for E.S.

  Two Types of parallel computer

  1) Parallel system with commodity microprocessors,
  2) Parallel system with vector type processors.

  Investigation of the execution speed of CCM2 (Community Climate Model Version 2) with present computer systems ;

  - parallel computers with microprocessors : efficiency 1 - 7%,
  - parallel computers with vector processors : efficiency about 30%,

  where the efficiency = (execution performance) / (theor. peak performance)

  We employ a parallel system with vector type processors.

- Necessary type of interconnection network for E.S.

  For flexibility of parallelism for many different types of applications,

  We employ a single-stage crossbar network.

  Because of technological limitations,
  total number of nodes < 1,000

It's All About Moving Data

- NEC SX-6 From memory, 1 word per cycle per vector pipe.
- With 8 vector pipes per processor there are 64 B/Cycle or 32 GB/s or 4 GWord/s memory bandwidth.
  - With a peak of 8 GFlop/s, data reuse in registers is a must for getting close to peak.
  - Scales with the 8 vector processor/node
    - (256 GB/s or 32 GWord/s ⇒ 64 GFlop/s)

- IBM Power4-P690 HPC there are 16 processors share 12 GB/s (1.5 GWord/s) memory bandwidth each processor with a peak of 5.2 GFlop/s (total 83.2 GFlop/s).

What About the Scalar Side?

NEC SX-6

- The clock rate is 500 MHz, theoretical peak 8 Gflop/s per processor, 4 way super scalar.
- There is a 64 Kbytes data cache and a 64 KB instruction cache.

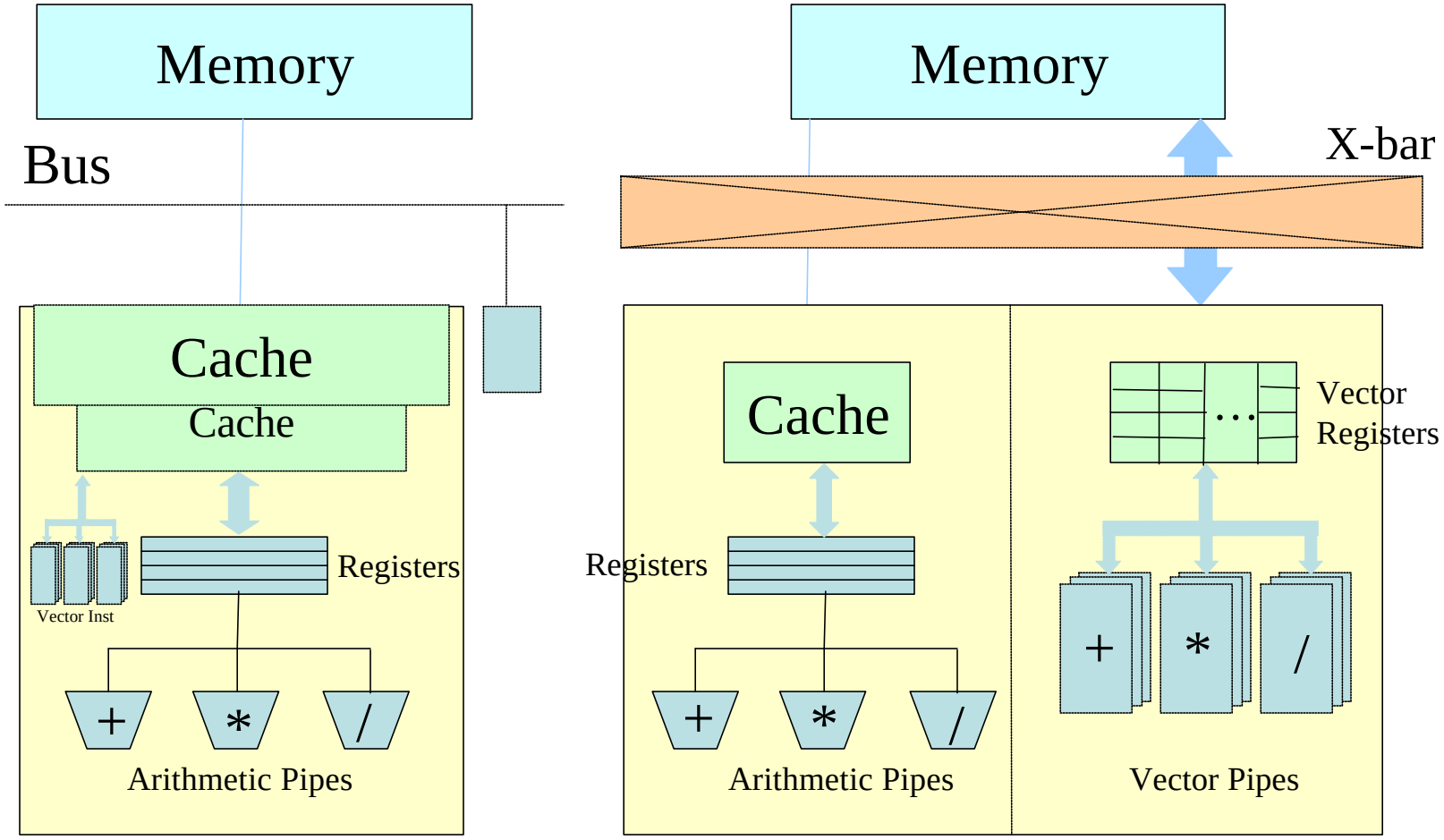
- On the scalar side it's only 4 GB/s data movement
  - (Pentium 4: 4.264 GB/s)

- The scalar performance of the SX/6 is relatively poor, 1 Gflop/s peak.

Processor	cycle time GHz	100x100 GFlop/s	1000x1000 GFlop/s	Theoretical Peak GFlop/s
Intel Pentium 4	**2.53**	**1.19**	**2.36**	**5.1**
NEC SX-6/1	**.50**	**1.16**	**7.58**	**8.0**

- For the 100x100 case the vector length is relatively short, scalar performance is more dominant.

Scalar vs Vector



μ-Processor

Vector Processor

Pentium 4 and AMD w/SSE2

What About Follow On Systems in Japan?

1. Fujitsu announced on August 22, 2002 the PRIMEPOWER HPC2500 Supercomputer
   ~ 85 TFlop/s (available Jan 2003)

2. Developing large Cluster system
   ~ 20 TFlop/s in 2003

3. Proposing the deployment of Grid computing in Japan
   ~300 TFlop/s 2003-2007

See: http://pr.fujitsu.com/en/news/2002/08/22.html

[PRESS RELEASE]

2002-0200E
Fujitsu Limited

Fujitsu Launches PRIMEPOWER HPC2500 Supercomputer

Massively Parallel Scalar System Boasts Record Performance, Scalability

Tokyo, August 22, 2002 - Fujitsu Limited today announced the release of a new massively parallel scalar supercomputer, the PRIMEPOWER HPC2500, which boasts the world's highest theoretical peak performance (up to 85.1 teraflops[*1]) and scalability (maximum of 128 x 128-processor nodes[*2], or 16,384 processors). Targeted at the High-Performance Computing (HPC) market, the new supercomputer becomes the top-of-line model in Fujitsu's flagship PRIMEPOWER servers for the UNIX market.

PRIMEPOWER HPC2500 uses a massively parallel scalar design based on 1.3 GHz SPARC64™ V processors to achieve the world's highest theoretical peak performance. With this blazing processing performance, for example, the time required for carrying out automotive crash analysis is greatly reduced. Fujitsu's own previous HPC server, the vector-type VPP5000 Series, took roughly 8.2 hours to complete the required calculations, while the new HPC2500 took only 1.1 hours, making it a solid seven times faster[*3]. Another notable potential area of use is in protein folding calculations for sequenced human genome research, which could be reduced eight-fold, from one year to about one and a half months[4].

In developing the new supercomputer, Fujitsu capitalized on its established track record and strengths in vector-based supercomputers, including cutting-edge processor technology, parallel execution of calculation instructions and ultrafast interconnect technology, and combined these features with shared memory and memory access control technologies developed for its PRIMEPOWER line of servers. As a result, Fujitsu was able to achieve a theoretical peak performance of 665.6 gigaflops5 per 128-processor node and, when 128 of these nodes (16,384 processors total) are clustered together using its high-speed optical interconnect, an astounding 85.1 teraflops overall. Both of these figures represent record-breaking theoretical peak performance levels. Compared with the VPP5000 Series, the new PRIMEPOWER HPC2500 delivers approximately 70 times higher performance per node, and 17 times higher overall (maximum system configuration of 16,384 processors).

To ensure a smooth transition for customers moving from the VPP Series to the new PRIMEPOWER, Fujitsu is providing a new Fortran compiler (XPFortran) that encapsulates the VPP Fortran specification. The company is also providing its new compiler, which is designed to get the most out of the SPARC64 V processors through cache optimization and other techniques, and to boost performance by recompiling existing programs. Fujitsu also assists customers in migrating to PRIMEPOWER systems with a variety of support services.

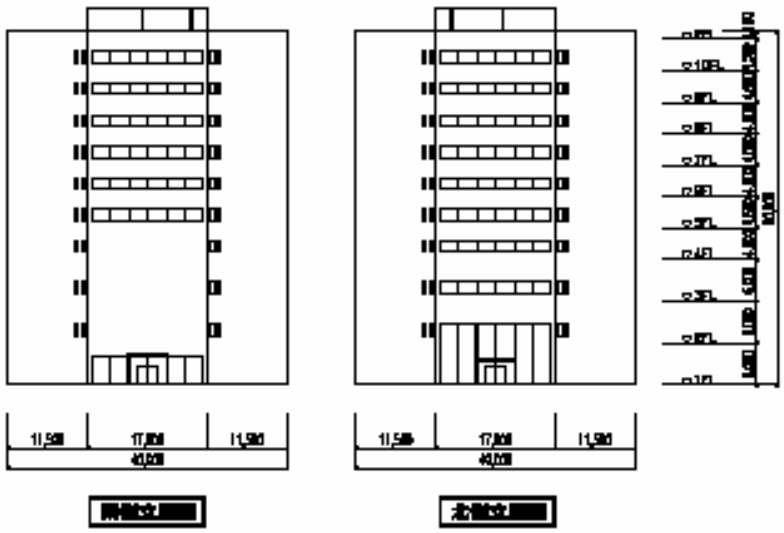
PRIMEPOWER HPC2500 Principal Specifications

CPU	Processor	SPARC64 V
	Clock speed; theoretical peak performance	1.3 GHz; 5.2 GFLOPS
Node	Number of processors	8-128 CPUs per node
	Main memory	max 512 GB per node
System	High-speed optical interconnect	2 channels of max.16 GB/sec (in, out) per node
	Maximum system configuration (CPUs; theoretical peak performance)	128 nodes (16,384 CPUs; 85.1 TFLOPS)

Delivery: From January 2003

AIST SuperCluster (2003)
(AIST GTRC, PI Satoshi Sekiguchi)

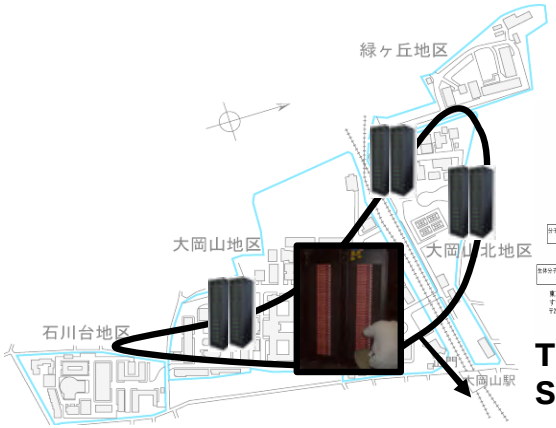
- Very large commodity cluster
  - CPU: Prescott, Opteron, Madison?
  - Interconnect: Myrinet, Quadrix, Infininband?
- 20 TFlop/s (Linpack)
  - < a MWatt Power
  - ~40 TFlop/s Peak > ES
  - #Procs 5000-8000?, #Nodes?
- To be housed in the new AIST Grid Technology Research Center Bldg.
  - Director: Satoshi Sekiguchi
  - To be used as a part of National Research Grid Infrastructure

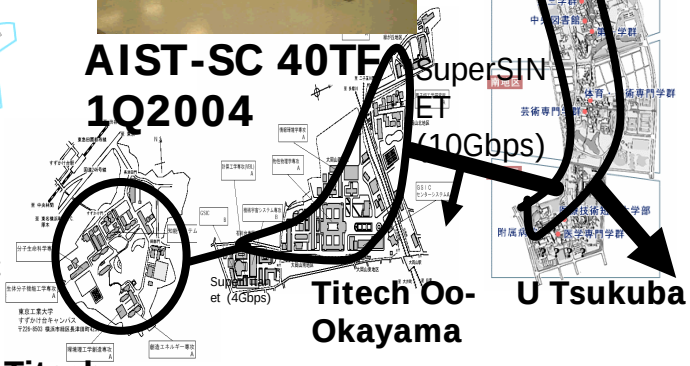


GTRC Bldg: 3000m^2/floor, 11 stories

Onto Nationwide: National Campus Grid Infrastructure

- National HEC Infrastructure for the 21st Century





AIST-SC 40TF
1Q2004

SuperSINET
(10Gbps)

Titech Oo-Okayama

U Tsukuba

Titech Suzukakedai

2002-3
Titech Campus Grid+lab

~2000CPUs, 3-4TFlop/s
130 TBytes
16 sites on Campus
SuperTITANET (1-4Gbps)

2003-7
National Research Grid Infrastructure (proposed)

20K-40K CPUs, 300TFlop/s
N PBytes
6 campuses, SuperSINET (10-40Gbps)

2007-
National Grid Infrastructure (speculated)

> 100,000 CPUs, >PFlop
30 PBytes
30 Campuses Nationwide
(Next Generation SuperSINET)

Lax Report

- Report on the panel on Large Scale Computing in Science and Engineering
  - December 26, 1982
- Four basic recommendations:
  - Increase access for the S&E research community through high bandwidth networks
  - Increase research in computational math, software and algorithms necessary to the effective and efficient use of SC systems
  - Training of personnel in S&E computing
  - R&D basic to the design and implementation of new SC systems of substantially increased capability and capacity, beyond that likely to arise from commercial requirements alone.

Increase Access For The S&E Research Community Through High Bandwidth Networks

- Significant improvements
- Change the way business is being done.
- Internet 2, Abilene, Vbns, ESNet
- But its not yet in our home

Lax Report

- Report on the panel on Large Scale Computing in Science and Engineering
  - December 26, 1982
- Four basic recommendations:
  - Increase access for the S&E research community through high bandwidth networks
  - Increase research in computational math, software and algorithms necessary to the effective and efficient use of SC systems
  - Training of personnel in S&E computing
  - R&D basic to the design and implementation of new SC systems of substantially increased capability and capacity, beyond that likely to arise from commercial requirements alone.

Increase Research In Computational Math, Software And Algorithms Necessary To The Effective And Efficient Use Of SC Systems

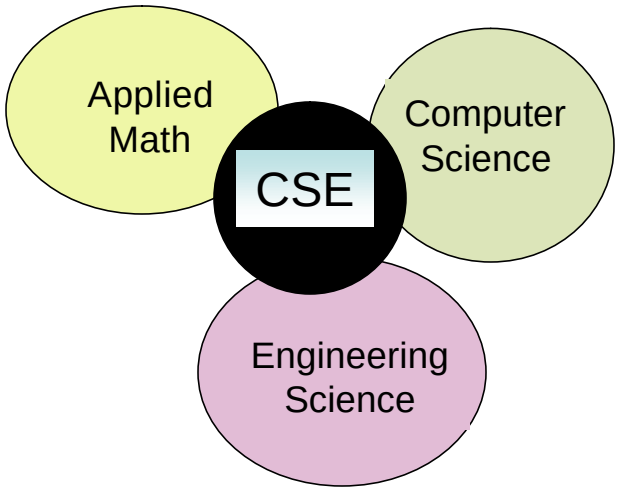
- SciDAC is a excellent start of lab cooperation
  - ~$60M/year
  - The original program called for an investment of $250M/year.
- PITAC (2/1999):
  - **Finding: The Nation is underinvesting in fundamental software research.**
    Over the past decade the U.S. has underinvested in research to create fundamentally new software technologies. Most of today's software technology is based on research performed 15 or more years ago.
  - **Major Recommendation: Make fundamental software research an absolute priority**

Lax Report

- Report on the panel on Large Scale Computing in Science and Engineering
  - December 26, 1982
- Four basic recommendations:
  - Increase access for the S&E research community through high bandwidth networks
  - Increase research in computational math, software and algorithms necessary to the effective and efficient use of SC systems
  - Training of personnel in S&E computing
  - R&D basic to the design and implementation of new SC systems of substantially increased capability and capacity, beyond that likely to arise from commercial requirements alone.

Training Of Personnel In Computational Science & Engineering

- About 50 Universities offer a graduate Computational Science & Engineering Program
- Grad Ed in CSE, SIAM Review, Volume 43, Number 1, pp. 163-177.
  - http://epubs.siam.org/sam-bin/getfile/SIREV/articles/37974.pdf
  - Eg. Stanford, U of IL, U of TX, ETH, KTH



- Undergraduate not at the right level today.

Lax Report

- Report on the panel on Large Scale Computing in Science and Engineering
  - December 26, 1982
- Four basic recommendations:
  - Increase access for the S&E research community through high bandwidth networks
  - Increase research in computational math, software and algorithms necessary to the effective and efficient use of SC systems
  - Training of personnel in S&E computing
  - R&D basic to the design and implementation of new SC systems of substantially increased capability and capacity, beyond that likely to arise from commercial requirements alone.

R&D Basic To The Design And Implementation Of New SC Systems, Beyond That Likely To Arise From Commercial Requirements Alone.

- PITAC: No well-established U.S. company remains with its primary focus on technical, high performance computing.
  - **Finding: The high-end computing capability available to the science and engineering community is falling dangerously behind the state of the art.**
- Yes there are projects in quantum computing and perhaps in 30 years…
- DARPA High Productivity Computing
  - Provide a focused research and development program, creating new generations of high end programming environments, software tools, architectures, and hardware components in order to realize a new vision of high end computing, *high productivity computing systems (HPCS).*
- Today focused on small decreasing number of vendors, research needs to come from the academic community in large numbers.
  - Need to invest in developing a cadre of researchers in this area. We've lost this.
- Vendors are not producing system that are specifically tuned to the needs of the scientific market.

Lax Report

- Report on the panel on Large Scale Computing in Science and Engineering
  - December 26, 1982
- Four basic recommendations:
  - Increase access for the S&E research community through high bandwidth networks
  - Increase research in computational math, software and algorithms necessary to the effective and efficient use of SC systems
  - Training of personnel in S&E computing
  - R&D basic to the design and implementation of new SC systems of substantially increased capability and capacity, beyond that likely to arise from commercial requirements alone.

It Not A Race

- But what's the program?
- Long term planning matters
  - Need a 10 year plan strategic plan
  - Strategic, rather than tactical objectives
  - Multiple generations of implementation
    - lessons learned influence next generations
  - 5 year project with a new focus every year …
    - SciDAC relabeled as ES response?
- It not just about hardware…
  - Funding for research in architecture, algorithms, software, systems, and hardware

- ESC path based on tying together vector processors
  - Problems with applications that are scalar orientated.
- US path based on tying together scalar commodity processors.
  - Creates performance problems with applications that could benefit from vector operations.
- For some problems vectors are not the solution.

- Need a combination of different architectures to handle a wider range of applications.

- Different paths are aggressively being followed in Japan
  - NEC on vector (SX-8 > 2 X performance of SX-6)
  - Fujitsu to focus on scalar commodity processors
  - Cluster computing
  - Grid computing as well

- Need to direct computer industry to develop equipment that can effectively solve our most challenging problems.
  - There's a big difference between web server and scientific computer.
  - Can't expect to do our most challenging problems on a large bunch of web servers.
  - There's not a market at the high end, need to provide incentives
  - Need to attract and sustain a cadre of talented scientists and engineers in development of these systems
- Cooperation of the OS labs for a concerted effort.
  - Need to circle the wagons and shoot out
  - Now is the time to make serious progress on pulling together a common story and plan.

- 20 years ago Lax Report made some recommendations
- 2 ½ years ago the PITAC reinforced those recommendations and recommended a roadmap to a Petaflop for scientific computing by 2010, but no agency picked up that roadmap.

It’s a Sign

- The current scalable, parallel, high-end computing systems are not well suited to many nationally important, strategic applications.
- Additional funding should be focused on innovative architectures, hardware technologies, and software strategies, and algorithm development that overcome the limitations of today's systems.
  - DOE has to take on the role of stewardship of research in computers design for computational science.
  - Must change the relationship of HW vendors and Computational Scientist
    - Must be integral in the development team

Lax Report

- Report on the panel on Large Scale Computing in Science and Engineering
  - December 26, 1982
- Four basic recommendations:
  - Increase access for the S&E research community through high bandwidth networks
  - Increase research in computational math, software and algorithms necessary to the effective and efficient use of SC systems
  - Training of personnel in S&E computing
  - R&D basic to the design and implementation of new SC systems of substantially increased capability and capacity, beyond that likely to arise from commercial requirements alone.