The NEC SX-9: The next step forward in computational meteorology

Thomas Schoenemeyer
NEC Deutschland GmbH
HPCE Division

Thirteenth Workshop on Use of High Performance Computing in Meteorology

3 – 7 November 2008

Agenda

- Details on the architecture
- Early benchmarks
- Advantages and Benefits
- Future HPC Products

Challenges in HPC

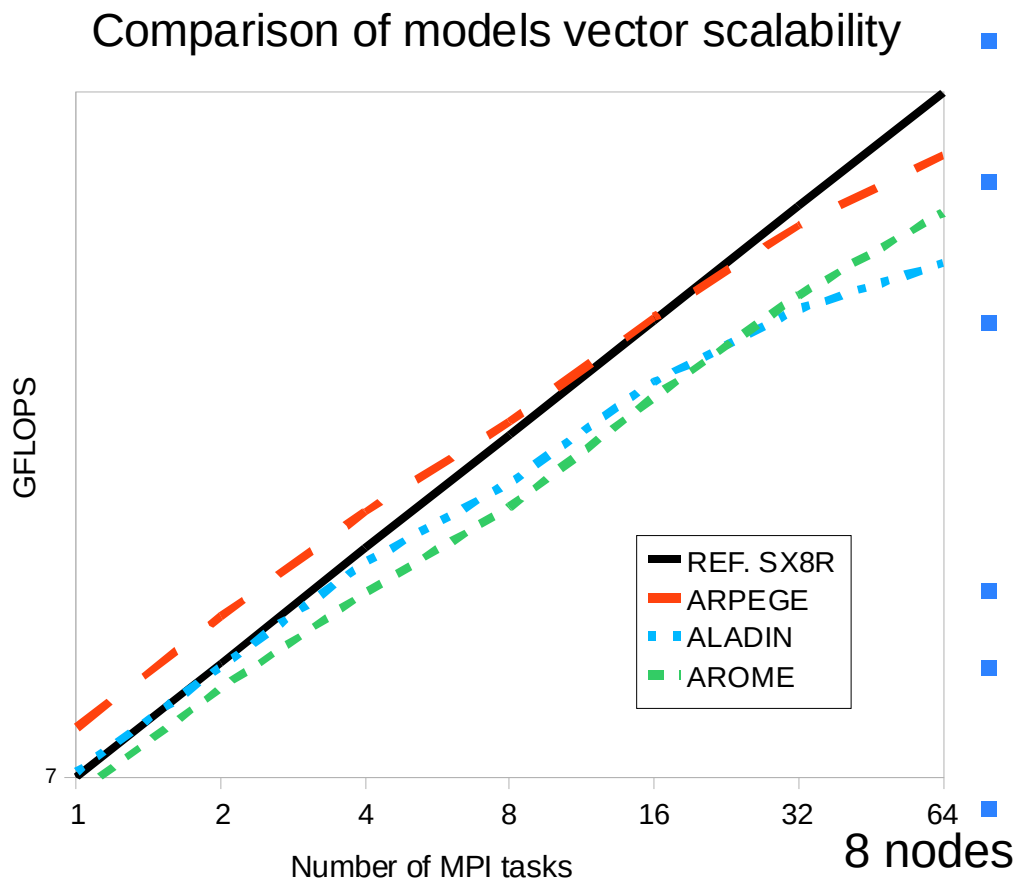
- Commodity may solve lots of computational problems, but …
- Access to memory - a big bottleneck?
- Processor Performance <-> Memory Bandwidth <-> Interconnect Bandwidth
- per TF I/O bandwidth 1? 0.5?
- Respect people skills and resources
- Preserve application software investments
- Time to solution
- Project Example →

SX-8 Vector Cluster (72 nodes, 9.2 Tflops)

Application	Area	# of nodes	Sustained Performance	Fraction of Peak
Best	CFD	72	5.8 Tflop/s	63%
CPMD	Chemistry	64	4.7 Tflop/s	57%
NEMO	Ocean	64	2.2 Tflop/s	27%
VASP	Molecular Dynamics	32	2.1 Tflop/s	41%
PARAPYR	Combustion	64	4.4 Tflop/s	53%
FRooM	Ocean	64	3.4 Tflop/s	42%
ECHAM5	Climate	64	2.4 Tflop/s	29%
URANUS	CFD	64	2.4 Tflop/s	29%
N3D	CFD	70	2.7 Tflop/s	30%
FENFLOSS	CFD	64	2.5 Tflop/s	31%
IFS	Meteorology	16	1.5 Tflop/s	37%

Achieved by the Teraflop Workbench Cooperation between HLRS and NEC

MeteoFrance



Number of processors used in operations :

ARPEGE=8 ALADIN=4 AROME=56

Relative number of gridpoints :

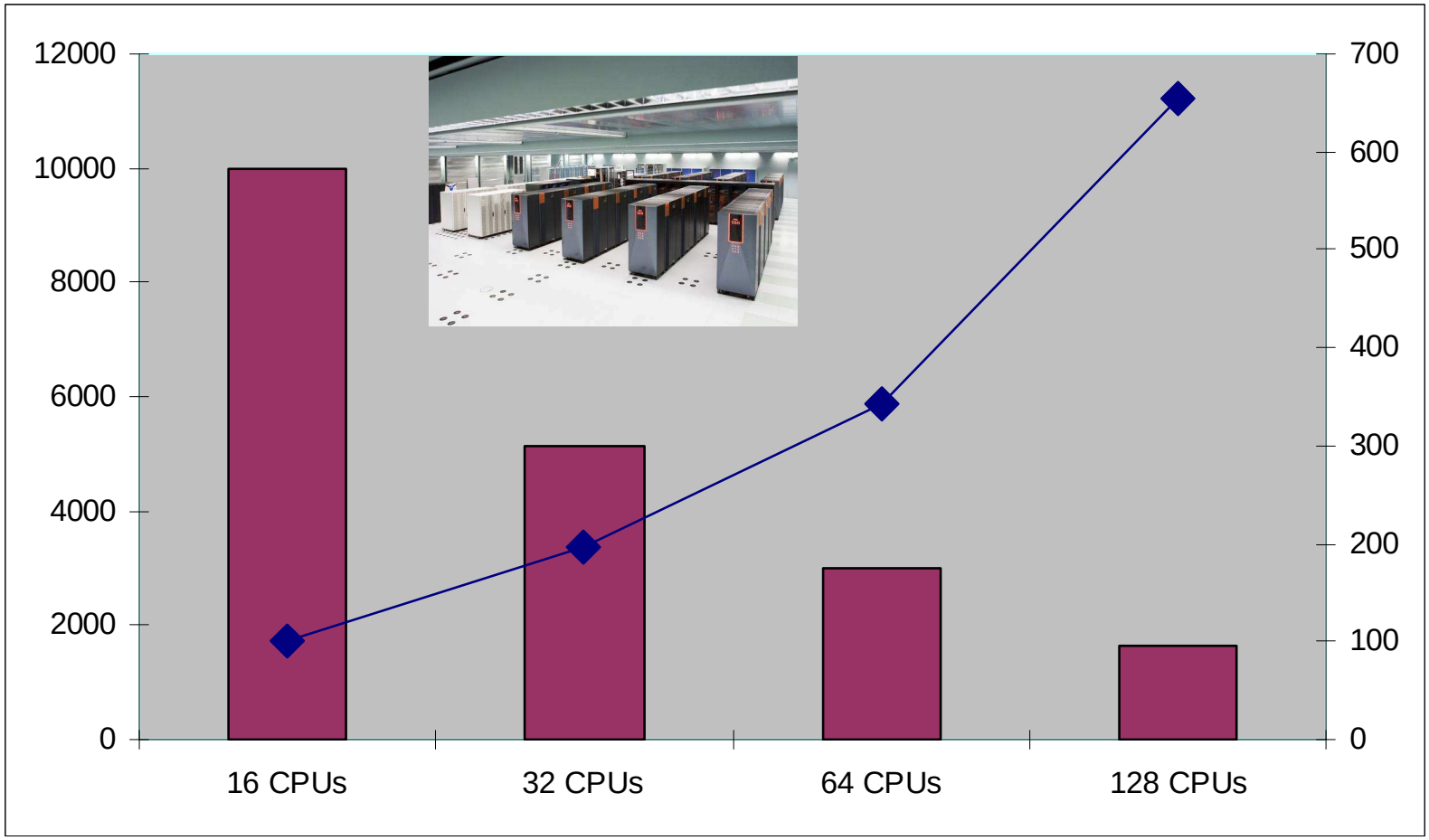
ARPEGE=4.44 ALADIN=1 AROME=2.33

- **Configurations and codes from operational suite**
- **REF-SX-8R (7 Gflops sust./CPU)**
- **Good performance of ARPEGE highly optimized matrixes multiplication routine**
- **ALADIN Cycle 33 T1**
- **AROME performance still under development**
- **But AROME keeps a good scalability : large model**

Results and Picture by courtesy of
Ryad EL KHATIB
CNRM/GMAP/ALGO
31057 Toulouse Cedex

COSMO-EU_2.8 (Benchmark for DWD)

Time [sec] Sustained Performance [Gflops]



2 nodes NEC SX-8 16 nodes

Orders SX-9

ASIA

- Japan Agency for Marine-Earth Science and Technology
- Tohoku University Cyberscience Center
- Osaka University
- National Astronomical Observatory of Japan

EMEA

- DWD, Germany
- CMCC, Italy
- MeteoFrance, Frankreich
- HLRS, Germany
- Research Center Karlsruhe, Germany
- Undisclosed Projects

Tohoku University

- Seismic Processes
- Direct Numerical Simulation
- FDTD-Simulation
- Simulation of Plasma Processes
- Nanotechnology
- Astrophysics

16 SX-9 nodes – 4 TB/s Interconnect Bandwidth
26.2 TF Peak
16 TB RAM

Delivery to Deutscher Wetterdienst

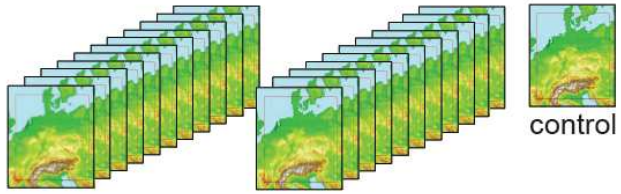
Autumn 2008

SX-9 Installation DWD

2 Compute Servers in two independent halls

Operational Forecast

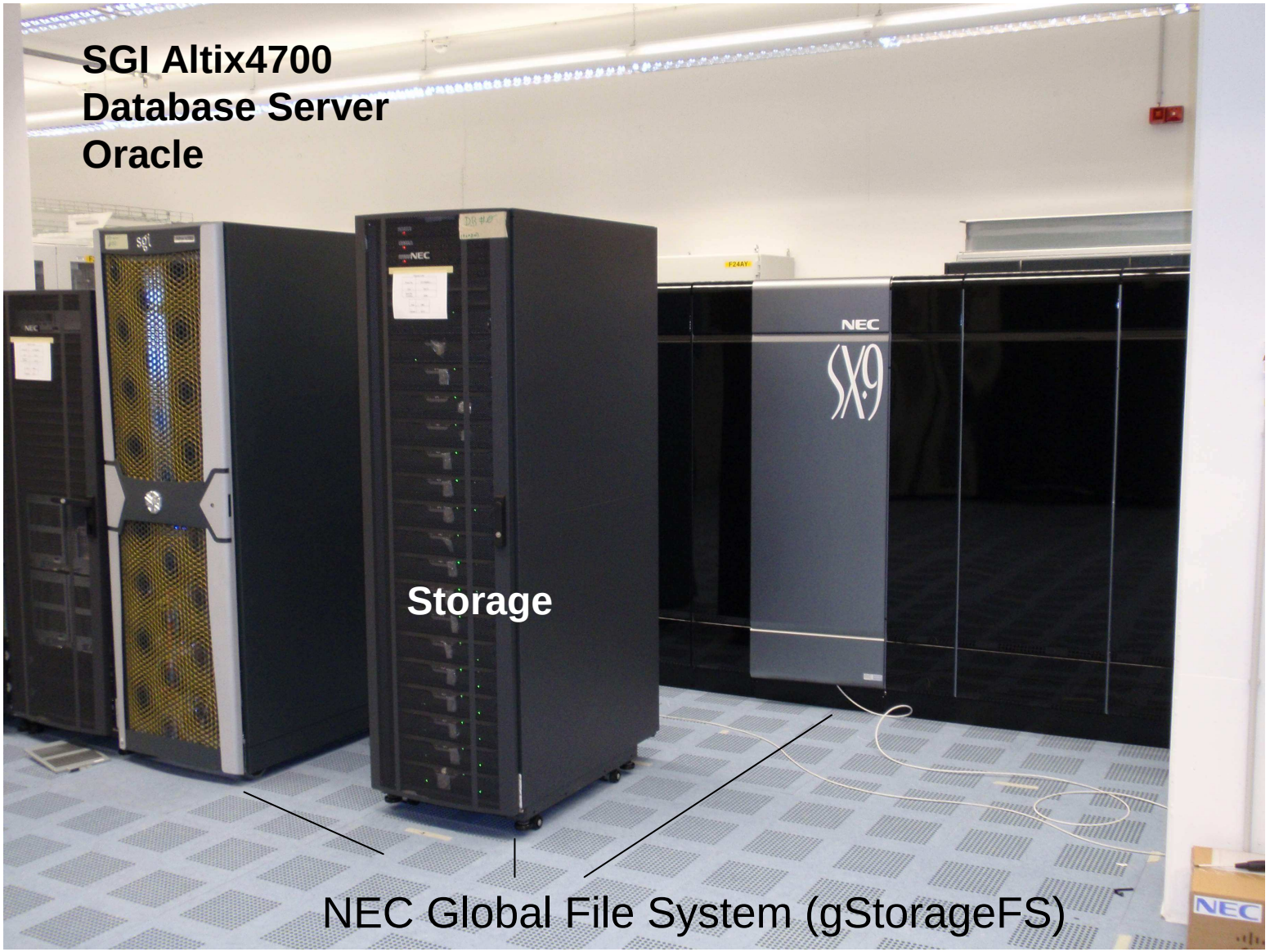
Research/Development/Backup



8 nodes per hall

12.8 TFlops Vector Peak per hall

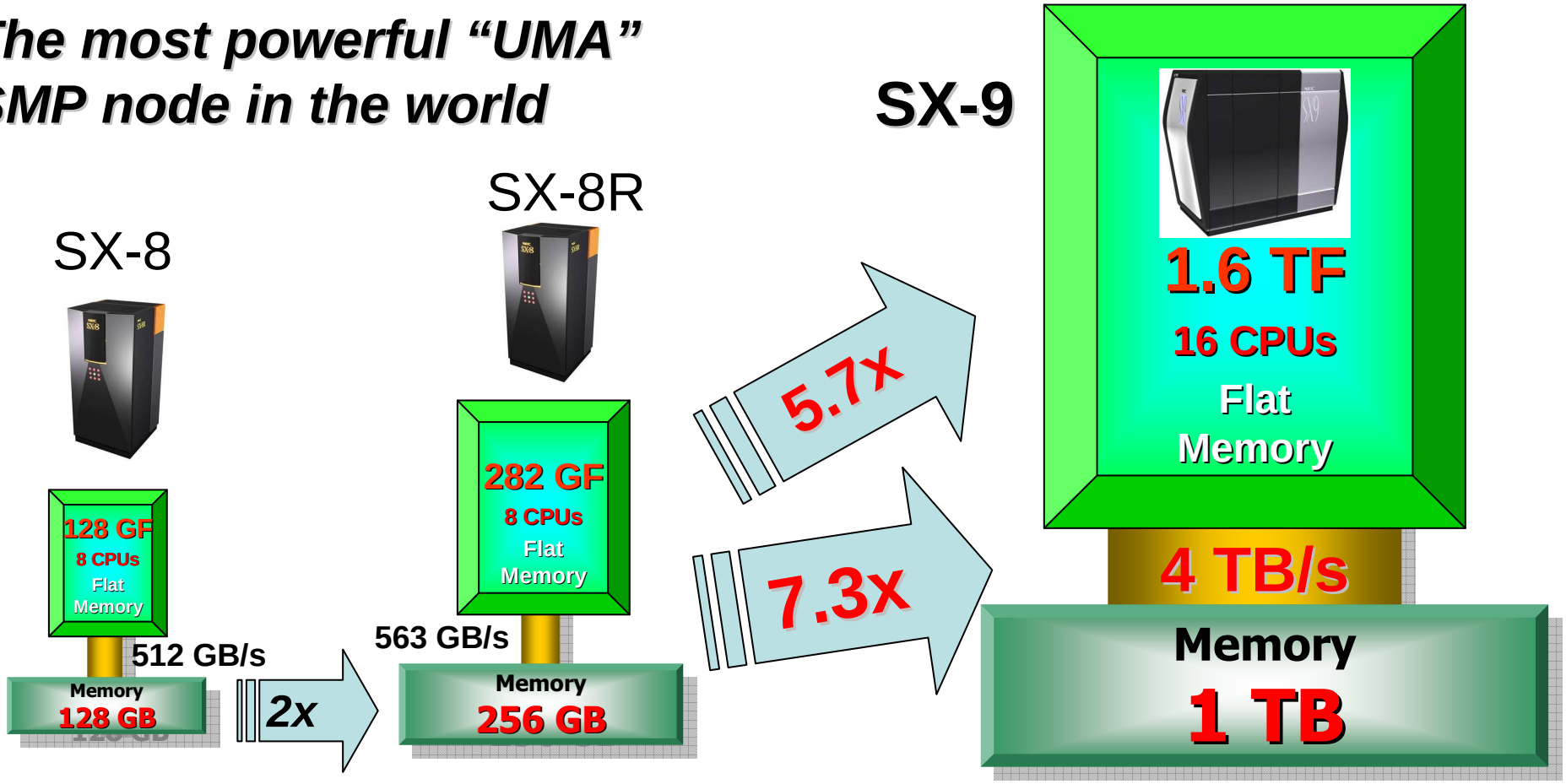
Compute & Database Server



NEC SX-9 Node: More than Evolution

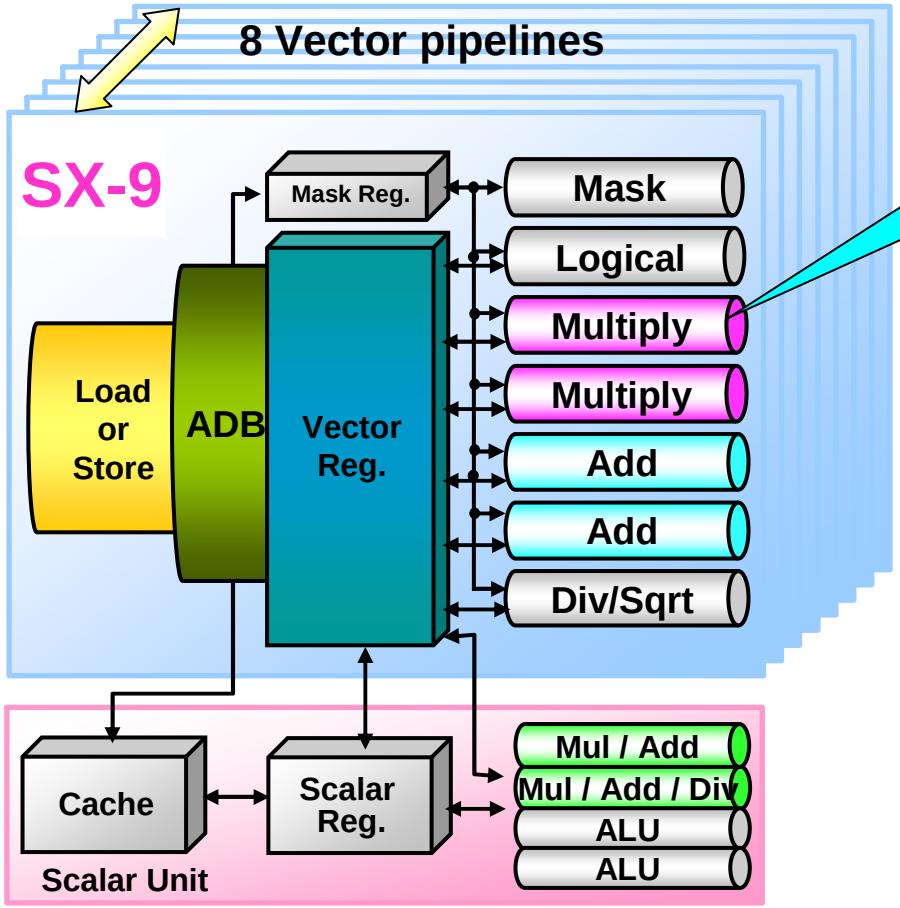
Node-Performance and UMA-Memory-Bandwidth

The most powerful "UMA" SMP node in the world



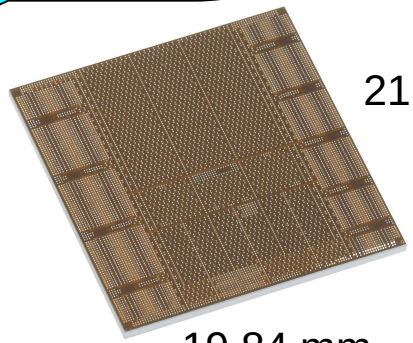
SX-9 Processor

Parallelism integrated



8 results per cycle

65nm CMOS
copper wiring
11 layers
# of pins(signal):8960(1791)
# of Tr.:350M
power:240W(max)
CLK:3.2GHz



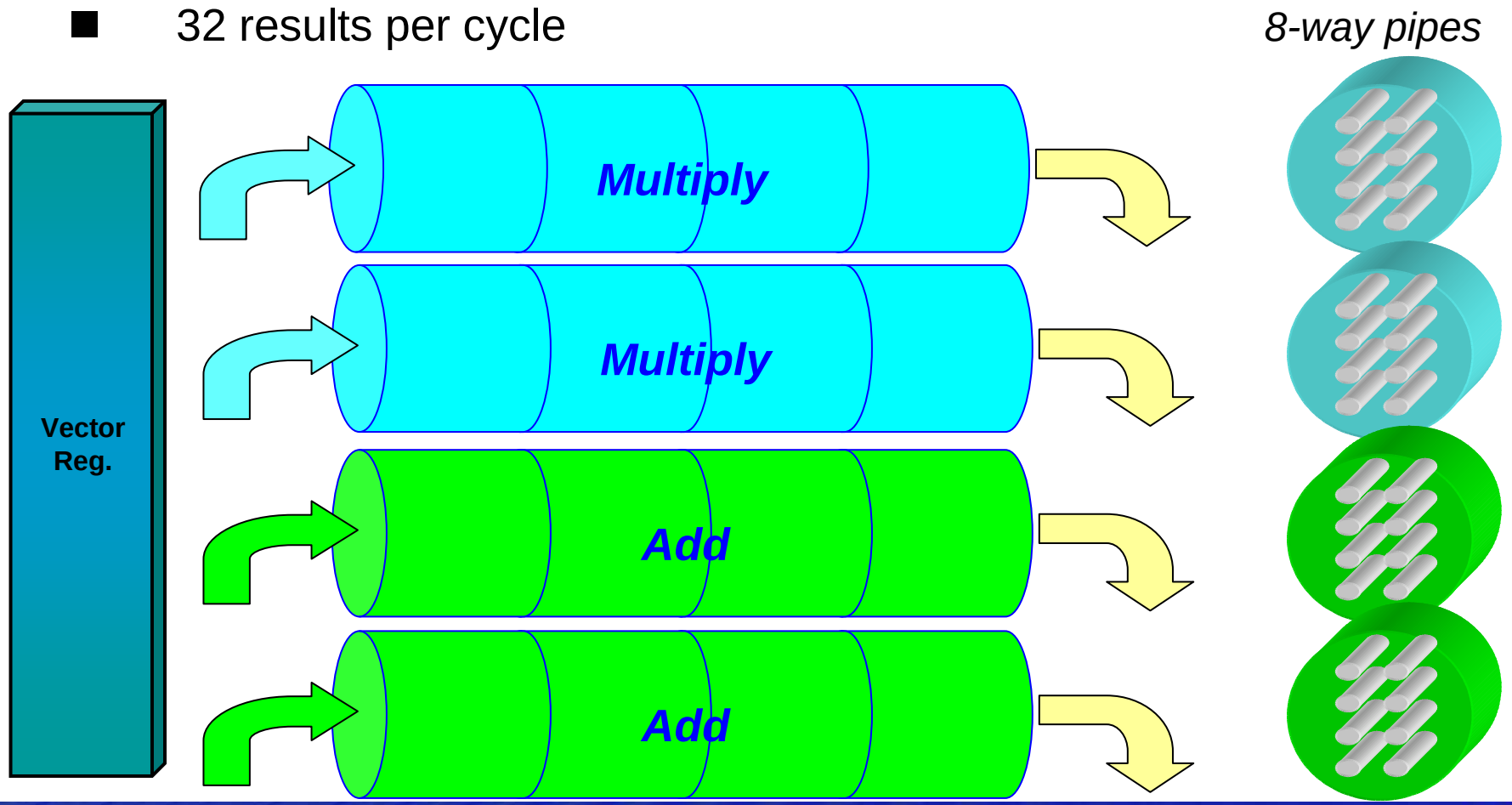
21.04 mm

19.84 mm

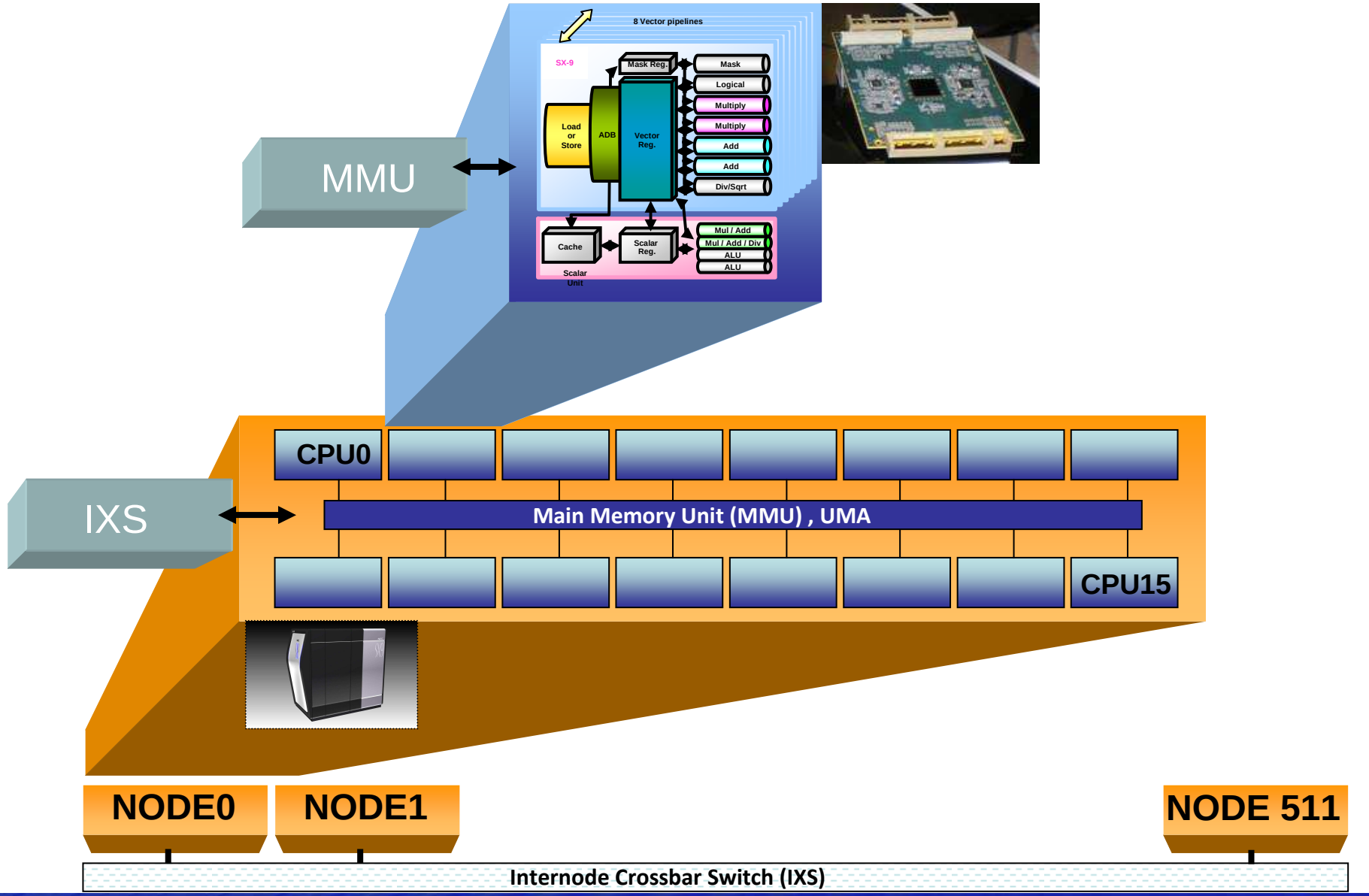
3.2 GHz
*** 8 results**
*** (2 add + 2 mult)**
→ 102.4 GFlops

SX-9: Parallelism integrated

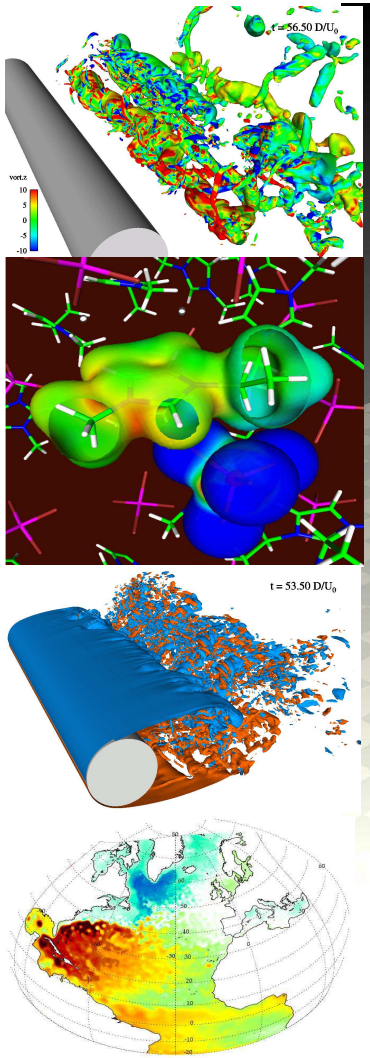
- One Instruction – Many Data
- 2 independent Pipesets for Add and 2 for Multiply:
  8-fold parallel @ 3.2GHz on SX-9
- 32 results per cycle



SX-9 System Architecture



CLUSTER OF STRONG SMP's



- Internode Bandwidth Factor 8 above SX-8
  128 GB/s per direction
- Highly Tuned MPI2
- Hybrid Parallelization with OpenMP
- Exploit all levels of Parallelism

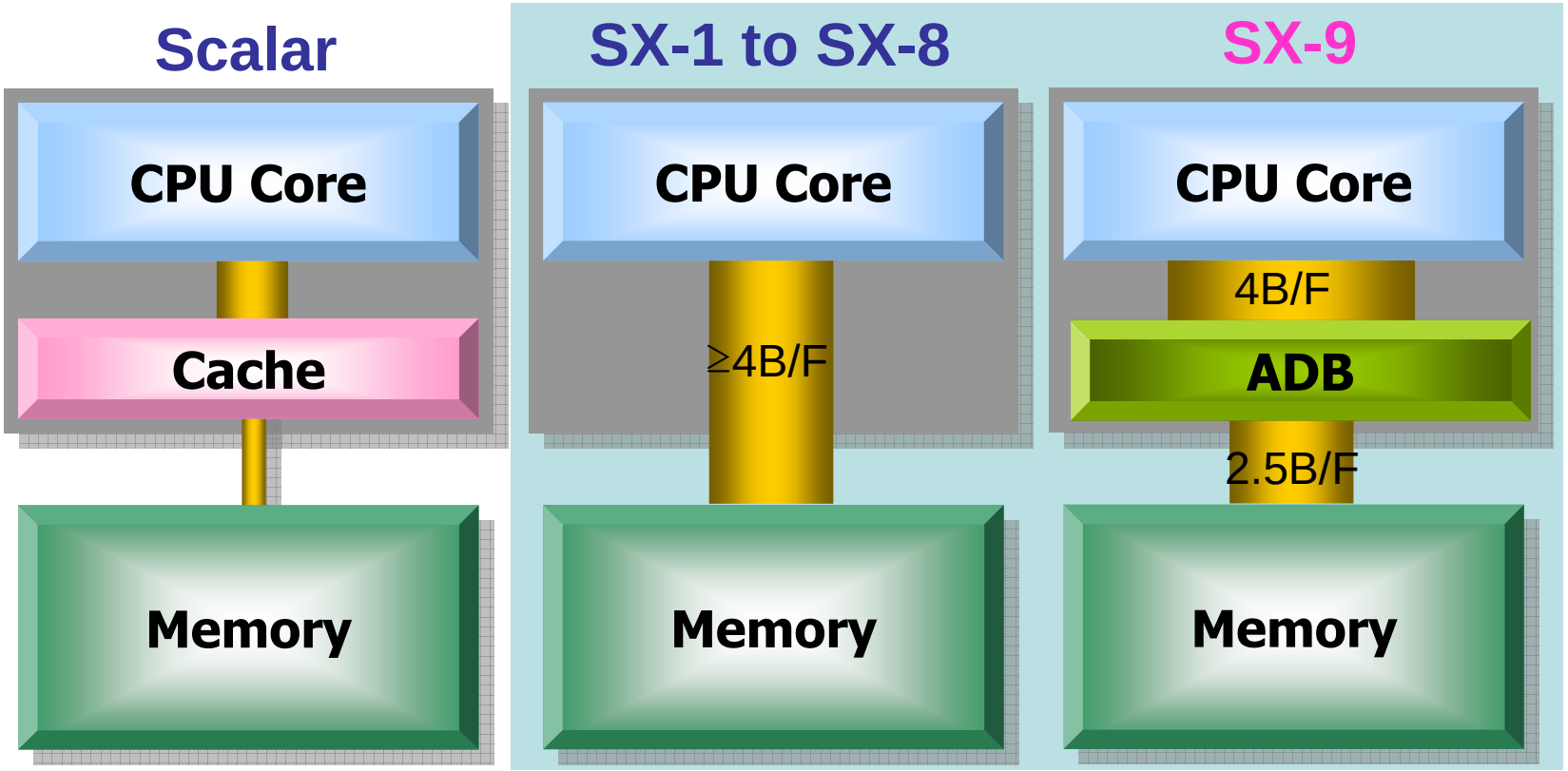
Preliminary Benchmarks SX-9 node

- **Stream ADD**
  - Stream Add: 3.14 TB/s on a node
  - Theoretical Max: 4 TB/s (256 x 16)
- **SPFLAME**
  - 355 Gflops sustained on a node

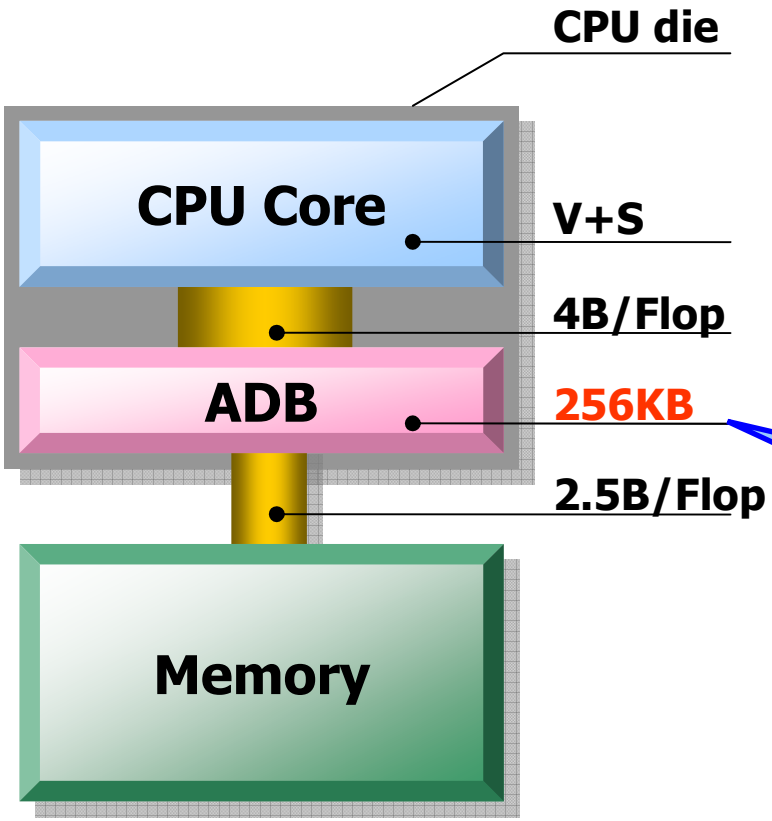
    Number of grid points 2023 x 560
- **COSMO-DE**
  - udsdx_up5_xy with more than 50 Gflops on a single CPU
- **Linpack HPL**
  - 1.43 Tflops (N=100000) on a SX-9 node
- **BLAS DGEMM**
  - 91 Gflops on a single SX-9 CPU

NEC SX-9 CPU: How to accelerate MB

"Assignable Data Buffer" (ADB)



High On-Chip Bandwidth



ADB（Assignable Data Buffer）

- Controlled by software
- Capacity :256KB
- Bandwidth
  - •Memory/ADB :256.0GB/s (2.5B/F)
  - •ADB/Core :409.6GB/s (4.0B/F)

ADB is controlled by SW in order to provide efficient ADB utilization.

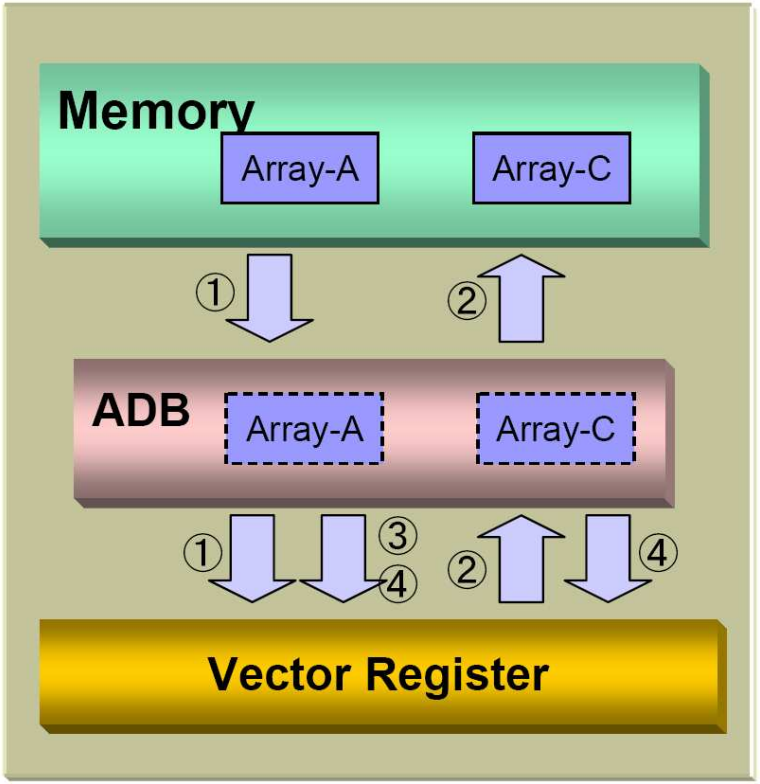
Advantages of ADB

- Enhancing memory bandwidth
- Reducing memory access latency
- Avoiding memory bank conflict

Buffering Vector Data in ADB

```
REAL, DIMENSION(1:300):: A,B,C
…
…

!CDIR ON_ADB(A)
!CDIR ON_ADB(C)

    DO I = 1,N
 ①   … = A(I) + B(I)
 ②  C(I)= …
 ③   … = A(I) + …
    END DO
   …
!CDIR ON_ADB(A)
!CDIR ON_ADB(C)

    DO I = 1,N
    …
 ④  … = A(I) + C(I)
    …
   END DO
```

Directives

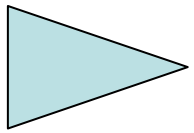


① Load A from Memory
② Store C in ADB
③ Load A from ADB
④ Load A & C from ADB
→ Save 3 Loads from Memory

How I/O Bandwidth can benefit from SX-9

*„A supercomputer is a device for turning compute-bound problems into I/O-bound problems“ **

- *NAS* ☹
- *SAN* 😐 *or* ☺ *(e.g. gStorageFS)*
- *Local Disk* ☺ *(4 Gb FC)*

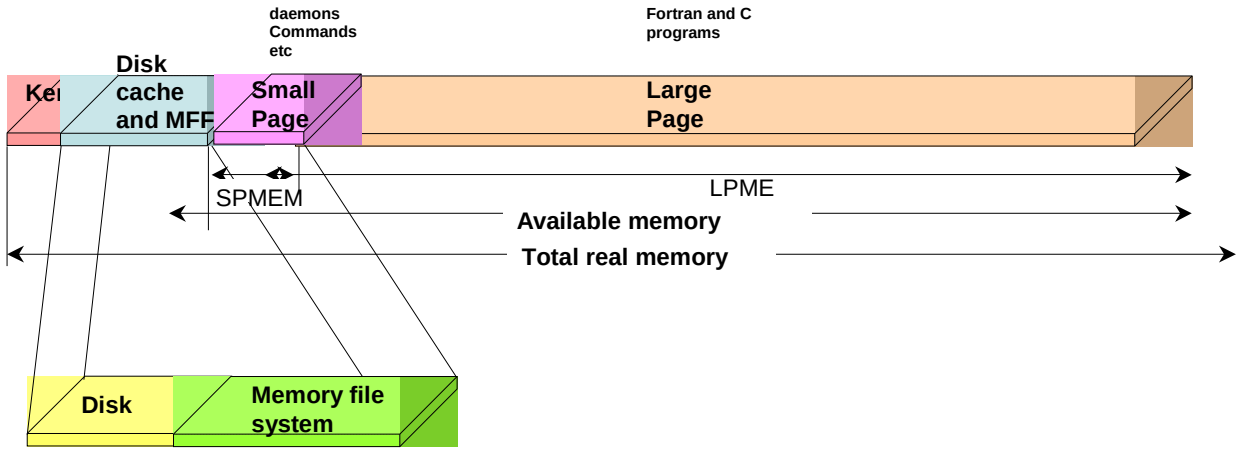


- SUPER-UX supports the concept of memory file system (MFF)
- Allows to perform IO operations with the minimum latency and the highest bandwidth
- Move the output afterwards into Shared File System

**Seymour Cray or Ken Batcher?*

MFF: A special tool for SX

- MFF (=Assignable Disk Cache for I/O)
- Very Fast
- 20% = 200 GB
- Example



512 GB
1024 GB

NEC MFF versus NEC gStorageFS

	/mff (SFS)		(8 way GFS)	
I-O size	**READ**	**WRITE**	**READ**	**WRITE**
1Mbyte	7837	6859	116	47
2Mbyte	9976	8577	229	94
4Mbyte	11585	9889	359	168
8Mbyte	12577	10675	466	238
16Mbyte	13258	11169	591	349
32Mbyte	13487	11451	599	467

Measured on SX-8 Cluster at HLRS, 4DVAR T159R

Priority of Jobs on NEC SX

- Operational Jobs must run with reproducable time need

  No disturbance by any other jobs (with lower priority)

- Emergency jobs shall start immediately run after submission

  The elapsed time for the emergency run should not exceed the time on a dedicated system

SX-9 uses an enhanced version of NQSII and Jobmanipulator developed for SX-8R

Even large and huge jobs can be suspended

Example for Priority Scheduling

Job	COSMO-DE	COSMO-EU 2.8
Grid size	421 × 461 × 50	1500 × 1500 × 50
Resolution (ca.)	2.8 km	2.8 km
Time step	30 s	30s
Forecast time	21 h (2520 steps)	12 h (1440 steps)
Disk space for Input	1 × 207.8 MB 22 × 442.1 MB	1 × 2.50 GB 13 × 2.04 GB
Disk space for Output	1 × 24.1 MB 22 × 442.1 MB	1 × 0.28 GB 13 × 2.83 GB
Memory needed (ca.)	20 GB (128 MPI tasks)	156 GB (256 MPI tasks)

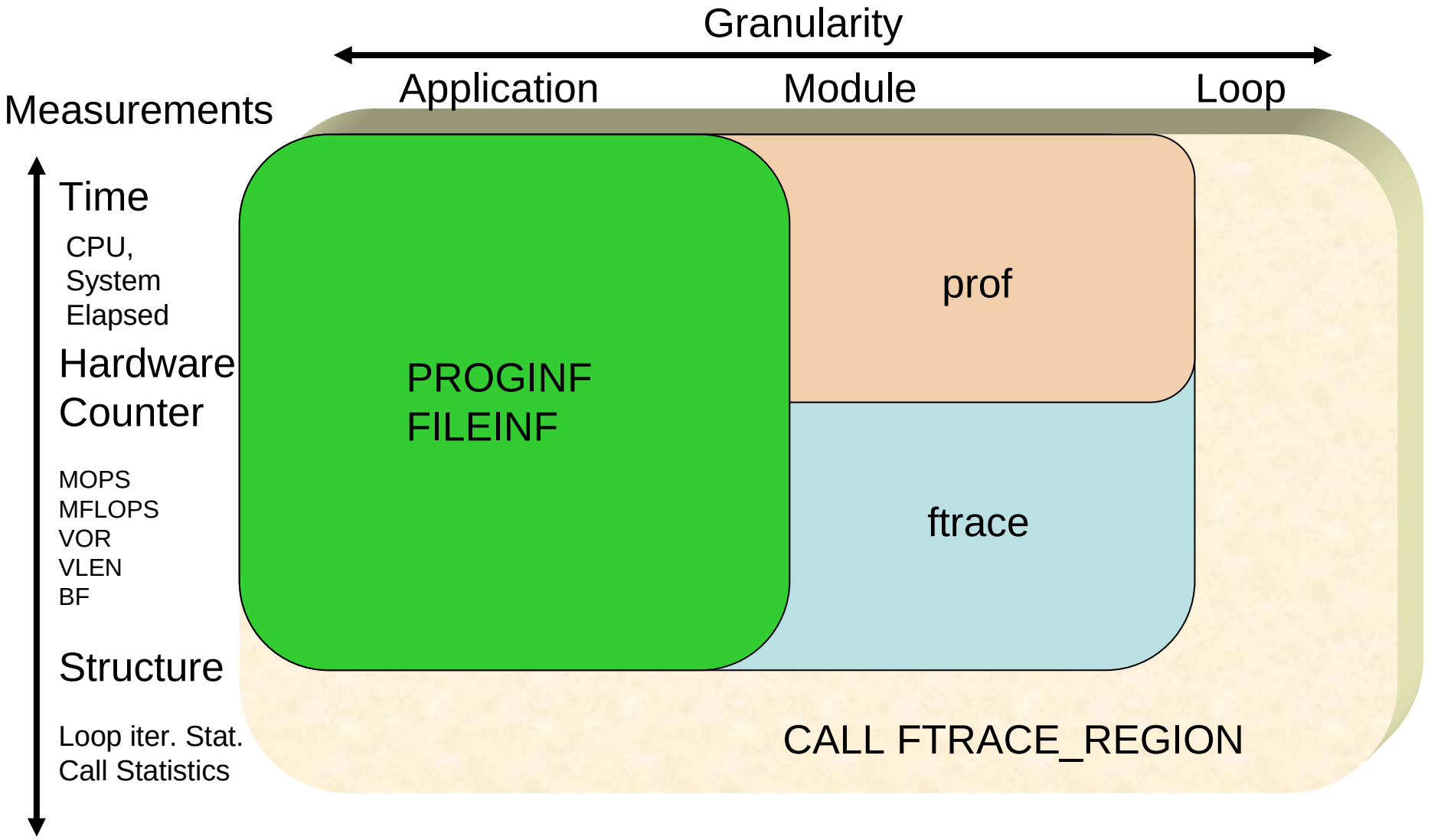
1 TB of RAM can be filled

With following results:

Priority counts – NEC NQSII

Job	Submission	Start	End	Elapsed Time	Difference to dedicated run
COSMO-EU_2.8	7:14:40	7:15:00	9:24:01	7741 s	+3027 s
COSMO-DE (Copy 1)	7:24:40	7:24:58	20:14:06	2948 s	-9 s
COSMO-DE (Copy 2)	7:24:40	7:24:58	20:14:12	2954 s	-3 s
COSMO-DE (Copy 3)	7:24:40	7:24:58	20:14:08	2949 s	-10 s
COSMO-DE (Copy 4)	7:24:40	7:24:59	20:14:13	2954 s	-3 s

How to use the NEC SX efficiently



PROGINF

```
                                  UT   =
                                  IC   =
SIC = IC -VIC                     VIC  =
SVC = SIC+VEC                     VEC  =
                                  FC   =
                        SVC / UT  =
                        FC  / UT  =
                        VEC / VIC =
                        VEC / SVC =
```

```
     ******  Program Information  ******
Real   Time (sec)          :       274.312165
User   Time (sec)          :       272.141663
System Time (sec)          :         1.473929
Vector Time (sec)          :       257.934811
Instruction Count          :     38211119460
Vector Instruction Count   :      8688304757
Vector Element Count       :   2057078964016
FLOP Count                 :    890893604266
MOPS                       :      7667.336768
MFLOPS                     :      3273.639157
Average Vector Length      :       236.764135
Vector Op. Ratio (%)       :        98.585125
Memory size used (MB)      :      1497.01942
MIPS                       :       140.408929
Inst. Cache miss (sec)     :         0.229672
Operand  Cache miss (sec):           2.678918
Bank  Conflict Time (sec):           9.299637

 Start Time (date)  :  2004/08/12 05:25:08
 End   Time (date)  :  2004/08/12 05:29:46
```

FTRACE

```
*---------------------------*
  FLOW TRACE ANALYSIS LIST
*---------------------------*
Execution : Fri Sep 20 16:43:38 2002
Total CPU : 0:21'56"273

PROG.UNIT  FREQUENCY  EXCLUSIVE       AVER.TIME    MOPS MFLOPS V.OP  AVER. I-CACHE O-CACHE    BANK
                      TIME[sec](  % )    [msec]                RATIO V.LEN  MISS    MISS      CONF
cft_3          27409   759.312( 57.7)    27.703  7555.4 3144.8 98.37 173.1  0.9885  2.5659 17.4761
s_1psi         24624   159.697( 12.1)     6.485 14709.7 7314.3 99.80 255.4  0.0197  0.0127  0.4511
ccgdiagg          11   138.952( 10.6) 12631.989 11546.6 6189.9 99.70 255.3  0.0542  0.0404  1.3057
cinitcgg          11    93.879(  7.1)  8534.443  5094.7 2230.6 98.90 194.3  0.0018  0.0016  0.0171
add_1vuspsi    13123    78.866(  6.0)     6.010 11951.1 7919.3 99.53 254.5  0.0383  0.0155  0.2382
h_1psi         13123    28.129(  2.1)     2.144  3978.4 1018.6 99.36 255.8  0.0378  0.0189  6.1214
xc           8748000    10.866(  0.8)     0.001   188.0   49.5  0.00   0.0  0.0000  0.0901  0.0000
-------------------------------------------------------------------------------------------------
total       19462138  1316.273(100.0)     0.068  8732.9 4109.4 98.95 202.8  1.2628  3.9578 26.5560
```

- Just compile&link with -ftrace
- This is the starting point of program optimisation
- ftrace produces overhead, remove after optimisation !

ftrace_region : Analyser on loop level

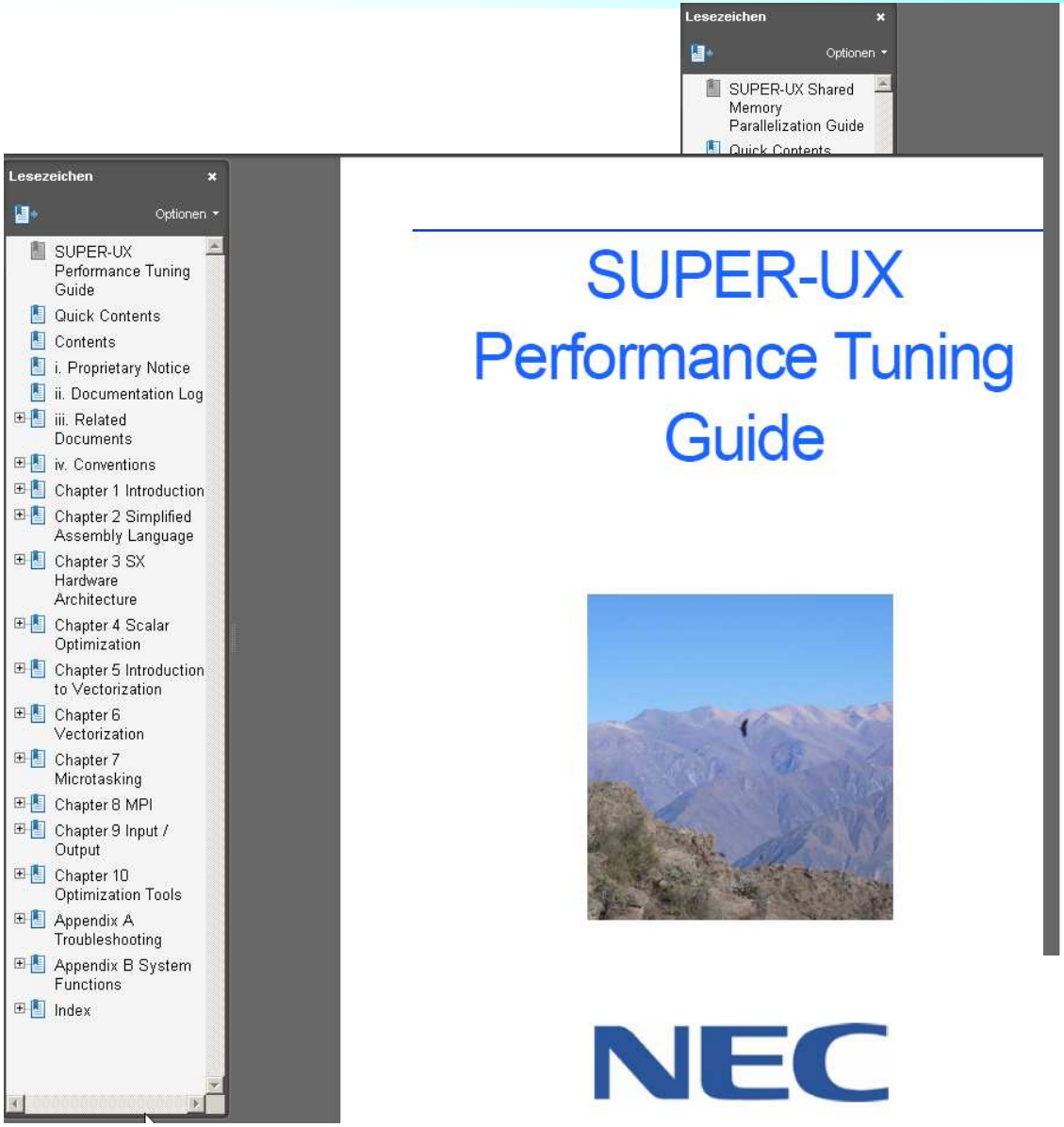
Instrument code as follows

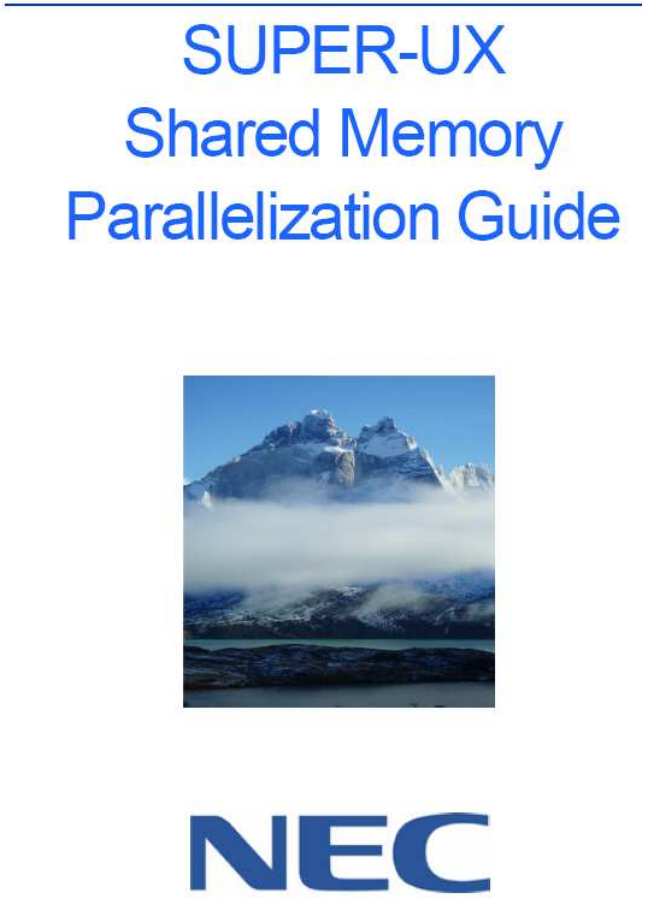
```
...
CALL FTRACE_REGION_BEGIN (“REGION A“)

      DO …
      …
      END DO

CALL FTRACE_REGION_END (“REGION A“)
...
```

```
*---------------------------*
  FLOW TRACE ANALYSIS LIST
*---------------------------*
Execution : Mon Jan 29 23:38:35 2001
Total CPU : 0:00'00"004
PROG.UNIT  FREQUENCY  EXCLUSIVE       AVER.TIME   MOPS MFLOPS V.OP  AVER.   VECTOR I-CACHE O-CACHE    BANK
                      TIME[sec](  % )    [msec]               RATIO V.LEN    TIME   MISS    MISS      CONF
sub               1     0.003( 63.7)     2.800   35.7    0.0  0.00   0.0    0.000  0.0000  0.0000  0.0000
main              1     0.002( 36.3)     1.598 1307.7    0.0 97.34 255.9    0.001  0.0002  0.0001  0.0000
-------------------------------------------------------------------------------------------------------
total             2     0.004(100.0)     2.199  497.9    0.0 92.90 255.9    0.001  0.0002  0.0001  0.0000
REGION_A          1     0.001( 22.7)     0.998 2043.2    0.0 98.08 256.0    0.000  0.0000  0.0000  0.0000
-------------------------------------------------------------------------------------------------------
```

PROG.UNIT	ELAPSE [sec]	COMM.TIME [sec]	COMM.TIME / ELAPSE	IDLE TIME [sec]	IDLE TIME / ELAPSE	AVER.LEN [byte]	COUNT	TOTAL LEN [byte]
c_do	652.523	652.520		0.160		16.0	21	336.0
0.0	652.523	652.520	1.000	0.160	0.000	16.0	21	336.0
gdrmpi	191.514	191.215		167.422		29.6M	900	26.0G
0.1	153.499	153.294	0.999	152.892	0.996	26.5M	200	5.2G
0.2	153.842	153.542	0.998	152.929	0.994	40.1M	200	7.8G
0.3	191.514	191.215	0.998	166.736	0.871	26.8M	300	7.8G
0.4	168.054	167.853	0.999	167.422	0.996	26.5M	200	5.2G
flroe3nnl	65.953	0.000		0.000		0.0	0	0.0
0.1	65.679	0.000	0.000	0.000	0.000	0.0	0	0.0
0.2	65.907	0.000	0.000	0.000	0.000	0.0	0	0.0
0.3	65.828	0.000	0.000	0.000	0.000	0.0	0	0.0
0.4	65.953	0.000	0.000	0.000	0.000	0.0	0	0.0
c_do	56.467	56.464		0.166		16.0	84	1.3K
0.1	56.467	56.464	1.000	0.166	0.003	16.0	21	336.0
0.2	40.468	40.465	1.000	0.135	0.003	16.0	21	336.0
0.3	5.613	5.610	0.999	0.115	0.020	16.0	21	336.0
0.4	35.911	35.907	1.000	0.117	0.003	16.0	21	336.0
invlui3fm1	41.324	0.000		0.000		0.0	0	0.0
0.1	41.324	0.000	0.000	0.000	0.000	0.0	0	0.0
0.2	40.688	0.000	0.000	0.000	0.000	0.0	0	0.0
0.3	40.387	0.000	0.000	0.000	0.000	0.0	0	0.0
0.4	40.746	0.000	0.000	0.000	0.000	0.0	0	0.0
invlus3fm1	38.015	0.000		0.000		0.0	0	0.0
0.1	37.640	0.000	0.000	0.000	0.000	0.0	0	0.0
0.2	37.967	0.000	0.000	0.000	0.000	0.0	0	0.0
0.3	37.853	0.000	0.000	0.000	0.000	0.0	0	0.0
0.4	38.015	0.000	0.000	0.000	0.000	0.0	0	0.0
slpnl3	33.511	0.000		0.000		0.0	0	0.0
0.1	33.473	0.000	0.000	0.000	0.000	0.0	0	0.0
0.2	33.511	0.000	0.000	0.000	0.000	0.0	0	0.0
0.3	33.189	0.000	0.000	0.000	0.000	0.0	0	0.0
0.4	33.460	0.000	0.000	0.000	0.000	0.0	0	0.0

New Training Material available



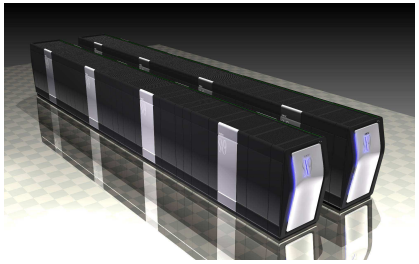


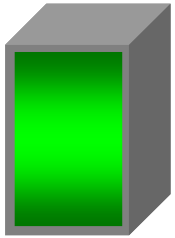
NEC Roadmap

High performance
Low power
High density

SX-8
128 GF/node

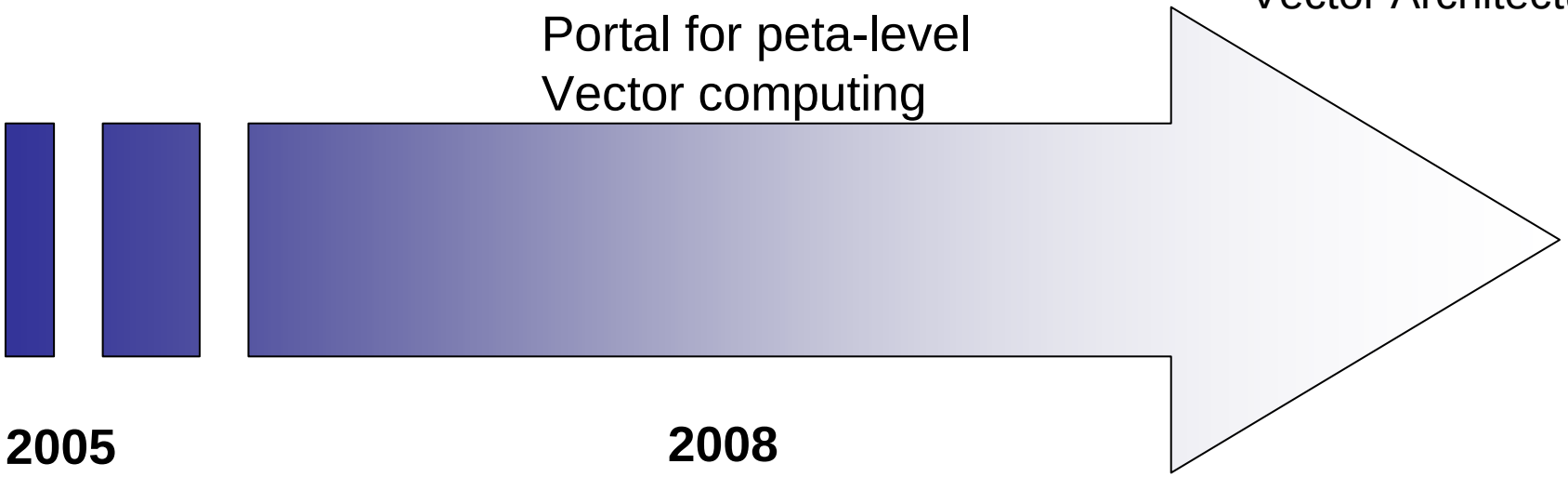
SX-9
1.6TF/node





Next Generation
Vector with new
Vector Architecture

Portal for peta-level
Vector computing



2005

2008

Future HPC Product Vision

Result of the internal ongoing investigation

- **Performance / Price → Performance / TCO**
  - Ingredients to TCO: Power consumption, Floor space, *Human Resources*
- **Strong single cpu**
  - enhanced vector architecture
  - Utilisation of data locality (enhanced ADB)
  - Multi-core vector
- **Simplified memory architecture**
  - High memory bandwidth, short latencies
  - Reduce power consumption!
- **Cost-effective for smaller configurations**
  - Smaller upgrade-increment
- **Hybrid parallelisation will be a key to scalability**

Subject to change without notice!






PRACE PROTOTYPES

PRACE, Partnership for Advanced Computing in Europe, has selected a broad coverage of promising architectures for Petaflop/s-class systems to be deployed in 2009/2010. Prototypes will be installed at six partner sites starting in 2008. The prototypes will be installed to the following sites:

BSC (Barcelona Supercomputing Center, Spain), installs a hybrid prototype combining IBM Cell and Power6 processors. The Cell processors are used for computation, and the Power6 processors for service.

CEA (French Atomic Energy Commission, France) and **FZJ** (Forschungszentrum Jülich, Germany) jointly use Intel Nehalem/Xeon processors in their systems. Two shared-memory multiprocessors (thin node clusters) will be distributed over the two sites; a prototype produced by BULL at CEA and a larger system of the same architecture at FZJ.

CSC (The IT Center for Science, Finland) and **CSCS** (Swiss National Supercomputing Centre, Switzerland) jointly evaluate the Cray XT5 architecture. This Massively Parallel Processing (MPP) prototype will be installed at CSC's facilities.

FZJ provides its already installed IBM BlueGene/P system, as a Massively Parallel Processing prototype.

HLRS (High Performance Computing Center Stuttgart, Germany) offers a NEC SX-9 and an x86 based cluster as a hybrid prototype.

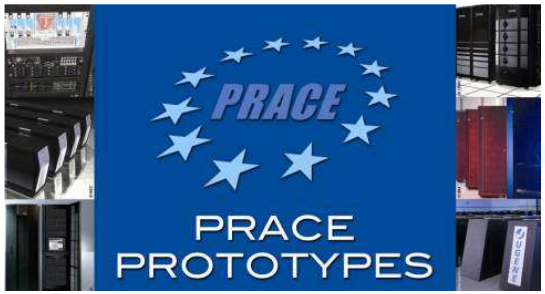
NCF (Netherlands Computing Facilities Foundation, The Netherlands) evaluates the IBM Power6 architecture, a shared-memory multiprocessor (fat node cluster). This prototype will be installed in SARA Computing and Networking Services facilities in Amsterdam.

www.prace-project.eu

e-infrastructure

CAPACITIES

The PRACE project receives funding from the EU's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° RI-211528.



NEC Hybrid System
More Revolution than Evolution

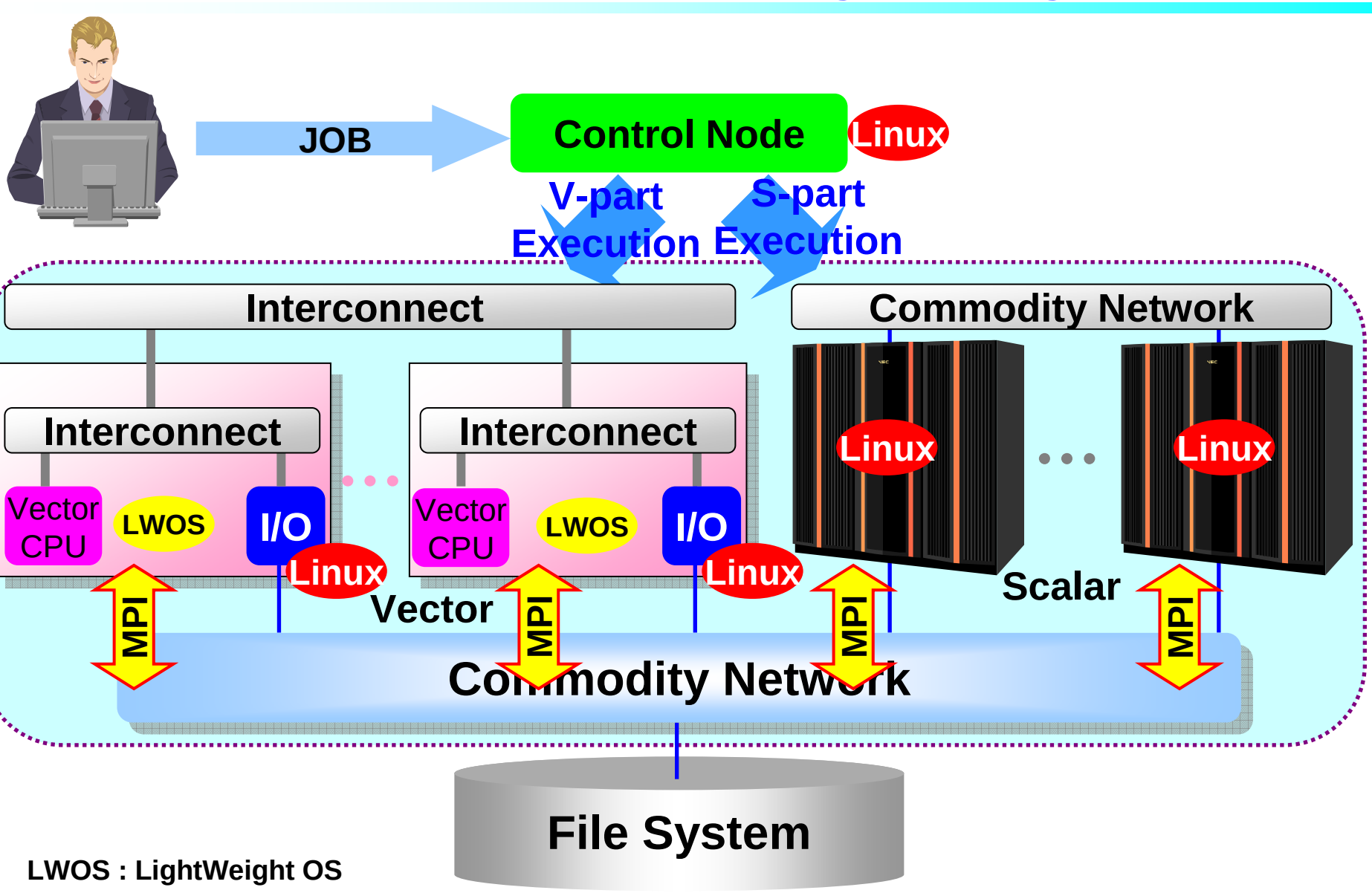
NEC SX-9 Vector
4-8 nodes*
64-128 cores*
6.5-13 TF*

NEC HPC Linux Cluster Dual Socket Intel Nehalem EX*
64-512 nodes*
512-4096 cores*
6.1-50 TF*

- **HLRS**
  - Specific Features
    - Unique „System of Systems“ concept
      - Multi-physics / multi-scale apps on optimized hardware
      - Hybrid configuration with SX-9 and HPC Cluster
      - Highly innovative configuration
      - Expandable (e.g. with GPU, FPGA, …)
      - Shared file system and heterogenous network
    - Concept enables industry-related network
- **Contribution to the PRACE project**
  - New Programming models and methods
  - Close collaboration between customer & vendor (joint Linux OS porting)
  - Necessary intermediate step towards new hybrid systems
  - Specific I/O and network challenges will be investigated
- **Availability March 2009**

*Current estimate, subject to ongoing negotiation

Future Concept – The Hybrid System



LWOS : LightWeight OS

Future Vector Product Vision

Preliminary target specifications compared to SX-9

- **Tflops per floor space : 7-10 x**
- **Tflops per kWatt : 7-10 x**
- **Tflops per €, $, ¥ : 7-10 x**

compared to SX-9

Subject to change without notice!

Empowered by Innovation NEC

NEC HPC Technical Workshop

Wednesday, November 19, 2008 3:30 pm to 5:30pm

Ballroom E on 4th Floor at Hilton Austin

Hybrid Supercomputing

Use of GPU's in HPC-Clusters

NEC SX-9 Vector Supercomputer

For more information and to register
Please access the workshop registration page

www.necam.com/sc08workshop

Use the registration code: nec1319

Thank you for your attention!