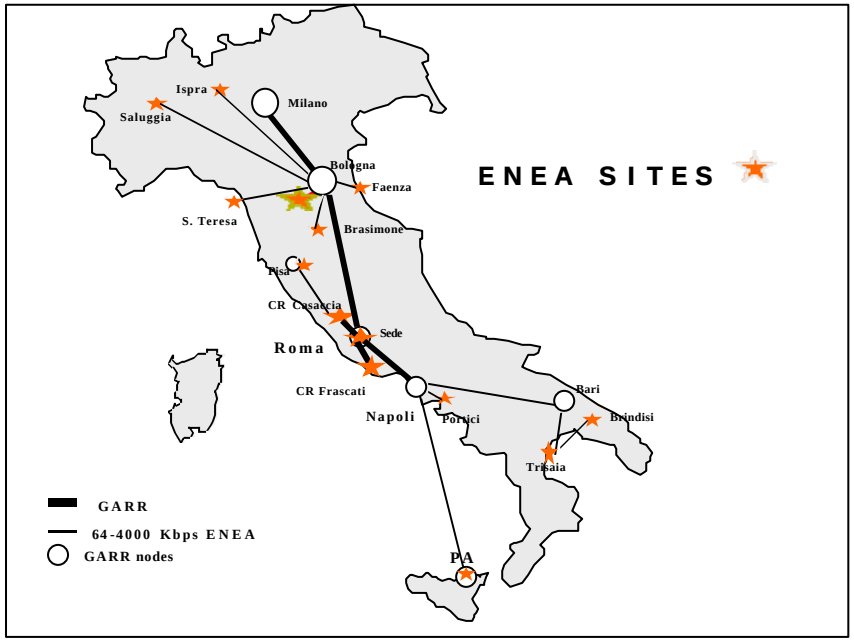
THE ENEA EXPERIENCE IN HIGH PERFORMANCE COMPUTING AND BEOWULF CLUSTERS

Massimo Celino
ENEA - Servizio Calcolo e Modellistica
C. R. Casaccia

Email: massimo .celino@casaccia.enea.it
Tel.: +39-0630483871
Fax: +39-0630484230



Hardware in the ENEA Research Centre Casaccia (2/3)

CRAY SV1

Hardware

- 16 SMP vector processors (1.2 GFlops each) with vector cache memory
- 300 Mhz of cpu clock
- 8 GBytes RAM
- 220 GBytes Disk

Software

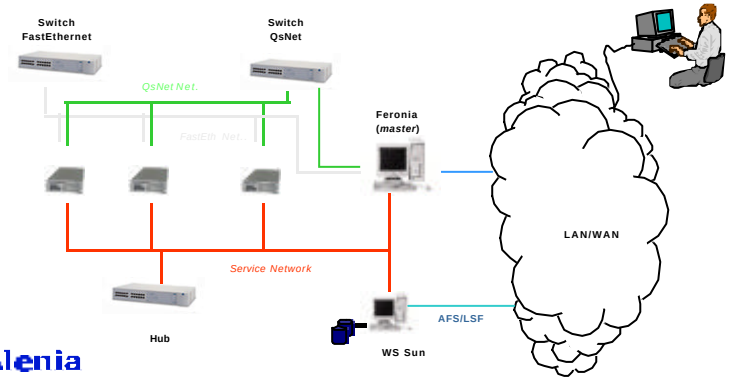
Fortran 90, C

Message Passing Interface (MPI)

Tools forvectorization and parallelization

UNICOS operative system

Hardware in the ENEA Research Centre Casaccia (3/3)



Alenia SPAZIO S.p.A. A Finmeccanica Company

www.quadrics.com

Feronia Beowulf Cluster

Node Architecture: UP2000

UP2000

- Two Alpha 21264 processors running at 667MHz, each with integrated 4Mb, L2 Cache
- 3.2GB/s L2 Cache bandwidth
- 2.65GB/s memory bandwidth
- 1GB RAM with ECC; 256-bit wide memory bus
- 6 PCI slots: Two 64-bit and four 32-bit
- One shared ISA expansion slot
- Two serial ports with modem control
- Dual USB ports
- Thermal sensor

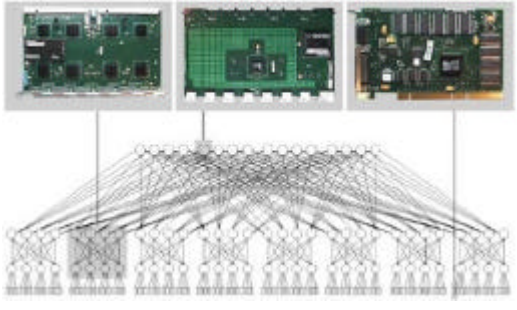
www.alpha-processor.com

Feronia Beowulf Cluster

QsNet

Network peak performance: 340 MB/S/rail
QsNet substained performace: 200 MB/S (75 %)
Mpi Latency 5 µs

Feronia computing power exploits 40 API UP2000 nodes, where each node is made up by 2 Alpha 21264 CPUs (667 Mhz), 1 GB RAM memory and 4 MB L2 Cache. System peakperformance is higher than 100 GFLOPS. In addition to the 40 nodes, Feronia has also a single CPU controller node, which acts as the system interface with respect to the external world. Feronia nodes are linked by means of 2 Fast Ethernet networks, one dedicated to general system services, while the other one is reserved for Message Passing data exchange in parallel applications. QSW proprietary interconnection technology, fat-tree QsNet is now available. Each node runs Linux Red Hat ver. 6.1 operating system with kernel 2.2.19.1qsw



QsNet Network software

Resource Management System (RMS) manages the access to the QsNet network.

RMS divides machine in multiple virtual partition

RMS manages user police of the network

RMS starts and stops the user's programs.

Pandora is an administrative tool.

It provide an high level vision of QsNet network status and functionalities. Pandora is able to provide statistics info on parallel jobs.

Feronia SW Architecture

Operating System
- Linux Red Hat 6.1
- Kernel 2.2.13

Compilers
- Compiler gnu gcc / f77
- Compaq Compiler f90, C, C++
- HPF Adaptor 7.0

Libraries
- Compaq Lib CXML,
- Scalapack, Blacs, Pblas

Parallel environment
- MpiCh, MPI_QSW

Tools Management
- NFS
- Nis

Tools Monitoring
- QswMon
- Parallel debugger Totalview

Feronia is integrated in ENEA environment :

AFS integration with NFS translator

LSF jobs schedule product

Help desk
- helpdesk@casaccia.enea.it

Newsgroup
- news.casaccia.enea.it : calcolo

Web
- feronia.casaccia.enea.it

MPI Benchmarks - ***The Pallas Suite (PMB2.2)***

http://www.pallas.de/pages/pmbd.htm

It provides a concise set of benchmarks targeted at measuring the MPI functions performance

Single transfer

- *Local mode*
- No concurrency with other message passing activity
- Only run with 2 active processes

PingPong
PingPing

Parallel transfer

- *Global mode*
- In concurrency with other message passing activity

Sendrecv
Exchange

Collective...

- ... in MPI jargon
- measure the quality of the implementation

Bcast
Allgather
Allgatherv
Alltoall
Reduce
Reduce_scatter
Allreduce
Barrier

MPI Benchmarks - ***The Pallas Installation***

Makefile

include make_AlphaLinuxQsNet

include make_AlphaLinuxStd

QsNet mode

MPI_HOME = /usr/lib/mpi

FastEthernet mode

MPI_HOME = /usr/local/mpich_1.2.1/

make

OPTFLAGS: -fast -arch ev67 -tune ev67 -O5

MPI Benchmarks - *The Pallas Run*

PMB-MPI1

QsNet mode

prun -N n -n p PMB-MPI1

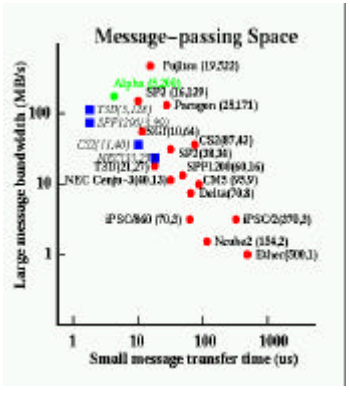
FastEthernet mode

mpirun -np p PMB-MPI1
(-machinefile filename)

With:

n	***number of nodes***
p	***number of processes***

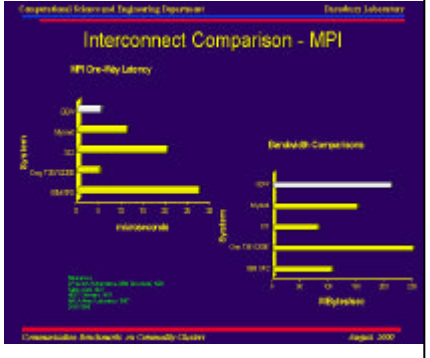
MPI Benchmarks - *Related works*



"Compaq Alpha QSNet vs. IBM SP"
http://www.ccs.ornl.gov/~dunigan/alpha/

"Communications Benchmarks on High-End and Commodity-type Computers"

http://www.dl.ac.uk/CFS/benchmarks/pmb



Pallas MPI Benchmarks Suite (PMB2.2)

PingPong

Measure the startup and throughput

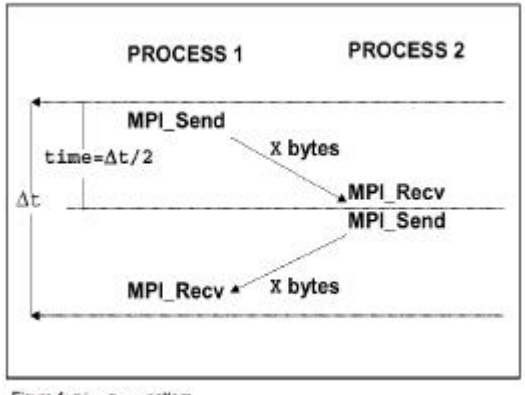
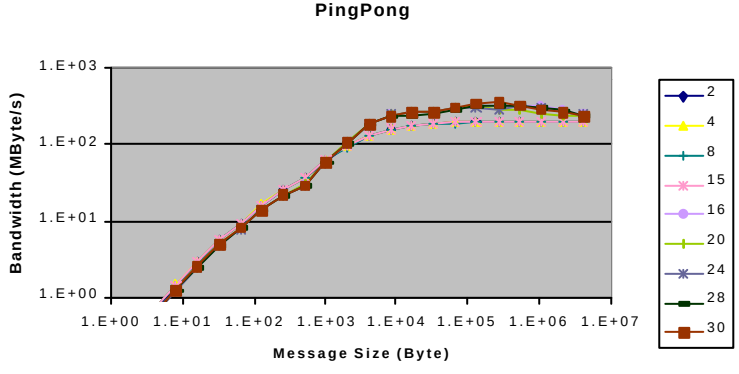


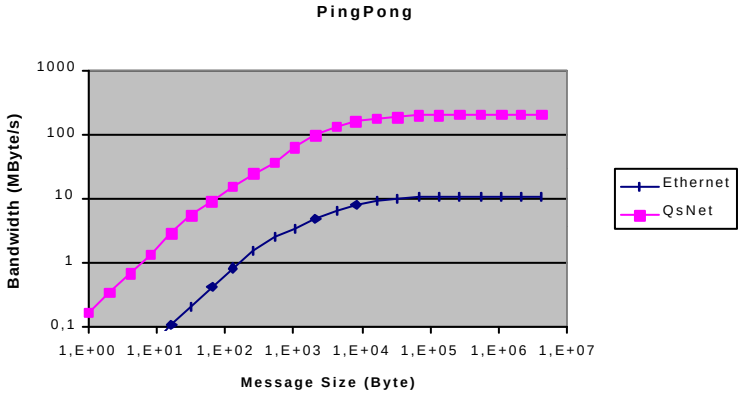
Figure 1: PingPong pattern

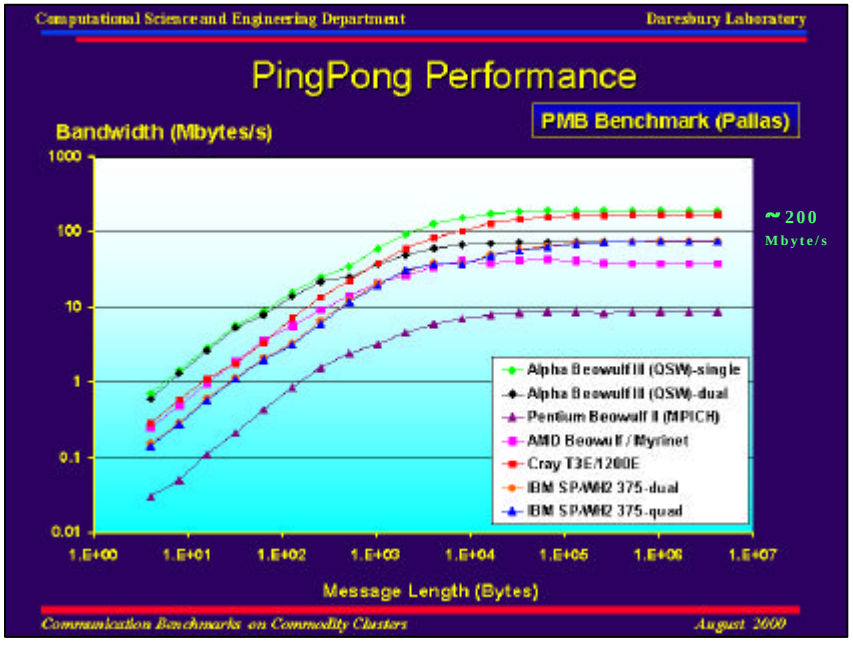
Pallas Benchmarks - *QsNet Results*



Latency:	**~ 5 µs**
Bandwidth:	**~ 200 MByte/s**

Pallas Benchmarks - *QsNet vs. FastEthernet*





Pallas Benchmark Suite (PMB2.2)

Sendrecv

The processes form a periodic communication chain.

Each process sends to the right and receives from the left neighbor in the chain

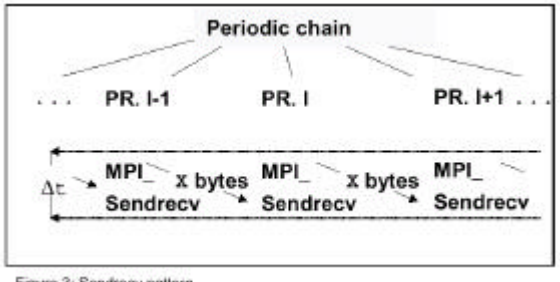
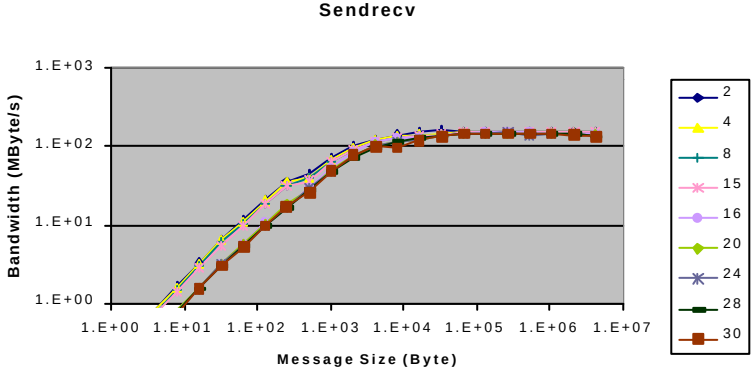
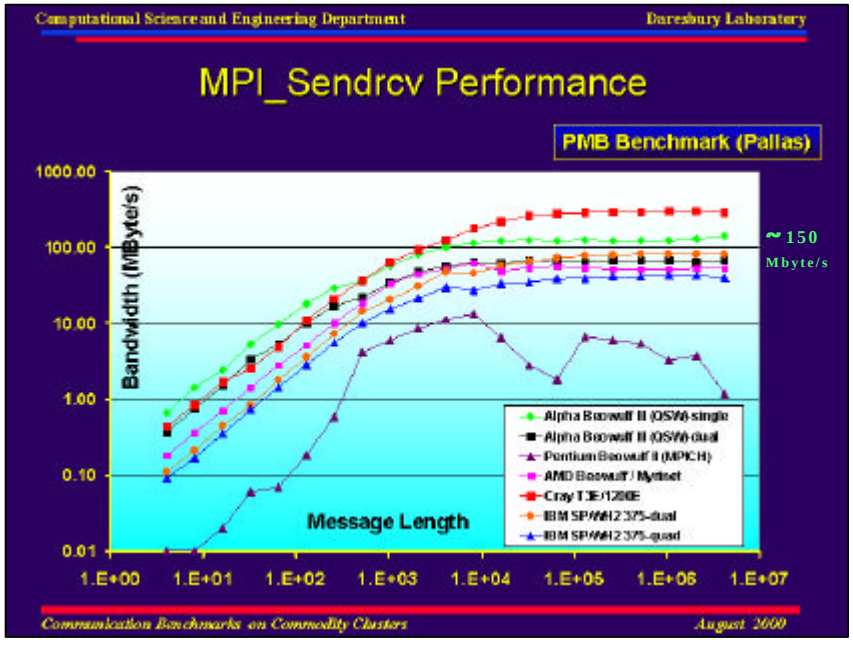
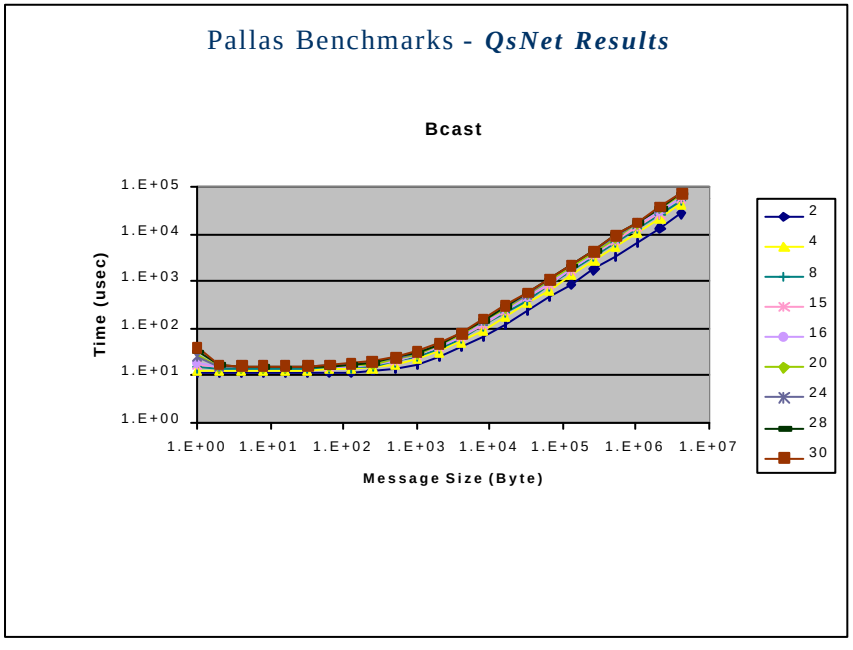


Figure 3: Sendrecv pattern

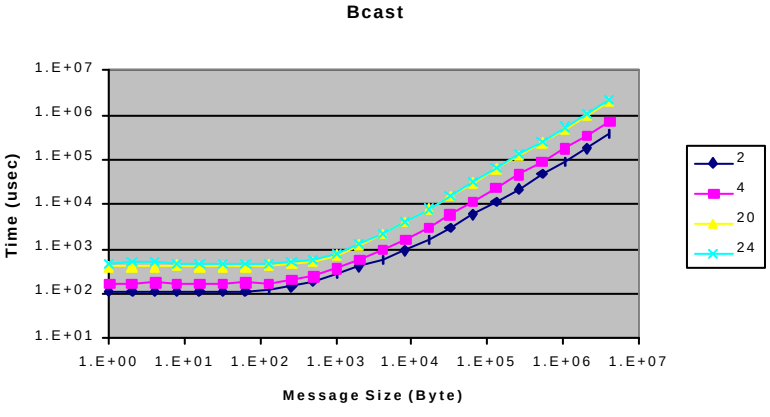
Pallas Benchmarks - *QsNet Results*




Computational Science and Engineering Department
Daresbury Laboratory
MPI_Sendrcv Performance
PMB Benchmark (Pallas)
1000.00
100.00
10.00
1.00
0.10
0.01
Bandwidth (MByte/s)
~150
Mbyte/s
Alpha Beowulf II (QSW) single
Alpha Beowulf II (QSW) dual
Pentium Beowulf II (MPICH)
AMD Beowulf / Mytinet
Cray T3E/1200E
IBM SP/WI2 375-dual
IBM SP/WI2 375-quad
Message Length
1.E+00
1.E+01
1.E+02
1.E+03
1.E+04
1.E+05
1.E+06
1.E+07
Communication Benchmarks on Commodity Clusters
August 2000

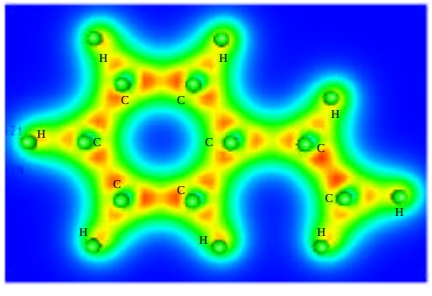

Pallas Benchmarks - QsNet Results
Bcast
Time (usec)
1.E+05
1.E+04
1.E+03
1.E+02
1.E+01
1.E+00
1.E+00
1.E+01
1.E+02
1.E+03
1.E+04
1.E+05
1.E+06
1.E+07
Message Size (Byte)
2
4
8
15
16
20
24
28
30

Pallas Benchmarks - *FastEthernet Results*



The Car-Parrinello Molecular Dynamics

The Car-Parrinello code allows to compute the evolution in time (Molecular Dynamics simulation at a given temperature) of a set of atoms (constituting for example molecules, polymers or materials) taking into account explicitly the electronic structure.



Charge density distribution of the monomer C_8H_8, building block of the phenilene-vinylene polymer, organic material used as light emitter (OLEDs) (for example in flat color displays).

The Car-Parrinello approach

CP method describes the quantum dynamic-based behavior of a system of N atoms, in the Born-Oppenheimer approximation. The CP method evaluates the system energy in the frame of the Density Functional Theory (Kohn-Sham) and takes the move from the Lagrangian

$$L = \Sigma_i \mu \int dr \, |\psi_i|^2 + 1/2 \, \Sigma_I M_I V_I^2 - E[\psi_r R_I] + \Sigma_{ij} \Lambda_{ij} \, (< \psi_i | \psi_j > - \delta_{ij})$$

where E takes into account explicitly the ion-ion, electron-electron and ion-electron interactions.

Some terms of the total energy are diagonal in real space , others in reciprocal space, thus the electronic wave functions $\psi_i(\mathbf{r})$ are developed in Fourier series:

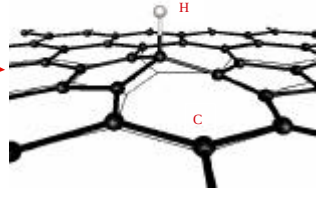
$$\psi_j(\mathbf{r}) = 1/\Omega^{1/2} \, \Sigma_{\mathbf{g}} c_j(\mathbf{g}) e^{i \mathbf{g} \cdot \mathbf{r}}$$

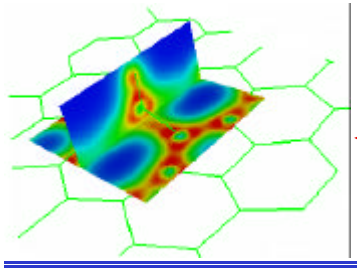
The Car-Parrinello approach

Adsorption of hydrogen in defective graphite

Application in the field of hydrogen storage in low density carbon nano-structures

Molecular dynamics evolution of an hydrogen atom on a plane of graphite





Analysis of the electronic properties of the chemical bond between the plane and the hydrogen atom

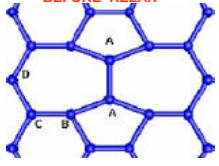
Lautrec code (courtesy of A. De Vita)

The Stone-Wales defect in graphite

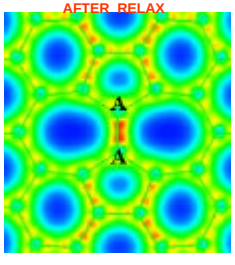
- The Stone-Wales (or 5-7) defect is introduced by switching one C-C bond by 90°, and then relaxed.
- The formation energy of the S-W defect is E_{sw}= 5.39 eV.

  (tight-binding gives E_{sw}= 5.8 eV and E_{sw}= 5.55 eV for a (6,6) nanotube).
- We observe a sizeable displacement of the C atoms up to 3rd and 4th neighbours around the switched bond (indicated with B,C,D).
- The A-A bond is contracted to d=1.31 Å, which is consistent with the formation of a double C-C bond, as corroborated by the electronic-density plot.

BEFORE RELAX



AFTER RELAX



The Car-Parrinello approach

Data are stored in large 3D and 1D arrays describing physical quantities on both grids, this implies a large use of :

FFT routines, large scalar products,
matrix moltiplications, matrix diagonalizations

N=**120 Carbon atoms** ($1s^2\ 2s^2\ 2p^2$)
Cell dim = 13Å x 12.5Å x 6.65Å
FFT grid = 100 x 96 x 32 points
E cut-off = 40 Rydberg (15600 plane waves)

N=**32 H_2O molecules** (H=1s, O=$1s^2\ 2s^2 2p^4$)
Cell dim = 9.88Å x 9.88Å x 9.88Å
FFT grid = 120 x 120 x 120 points
E cut-off = 90 Rydberg (46700 plane waves)

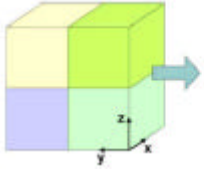
→ In these simple cases some Gigabytes of memory are used and minutes of simulations are needed for one time step

Production runs are memory and cpu bounded !!

→ Parallel computer with large memory and very fast inter-node network

The parallelization strategy

Because the most time consuming routine are the
FFT and the orthogonalization (essentially solvable by
scalar products of wave functions) ones, and
because there is the need to have access to great quantity of memory:



The parallelization is obtained distributing among the
PE grid all arrays (domain decomposition) storing quantities tha t are
function of real and reciprocal space coordinates **r** and **g** (wave
functions, charge density and all their related quantities).

Positions and forces on ions are not distributed and are present on all
the PE.

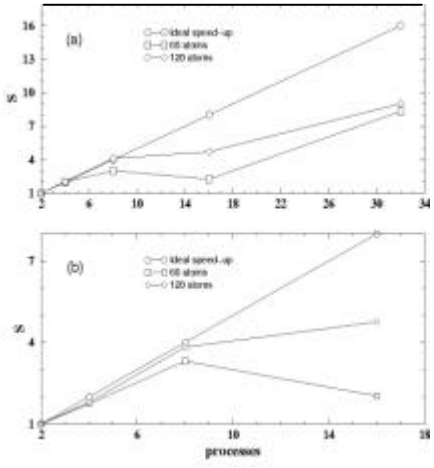
The parallelization strategy

In this way scalar products within orbitals and all the real space quantities are efficiently computed, since they correspond to integrals whose domain is distributed among processors: only the integration subtotals have to be communicated between computing nodes.

At the same time, the three-dimensional FFT is implemented by using optimal communication routines "dedicated" to the specific problem
(for example, FFT is not computed on columns of elements equal to zero).
The data matrices required by the algorithm were also distributed (above all, for memory reasons), and specific communication routines were coded to perform the necessary distributed matrix algebra.

Finally, the electronic orbitals can be chosen to be *real* functions, so that two of them can be packed into a complex-to-complex FFT, while only half of the memory allocation which would be necessary for complex orbitals is actually needed. This technique does not introduce any extra communications between computing nodes, if the data distribution is properly handled. Code Performance

Performance tests



Speedup S

In (a) two processes per node

In (b) one process per node

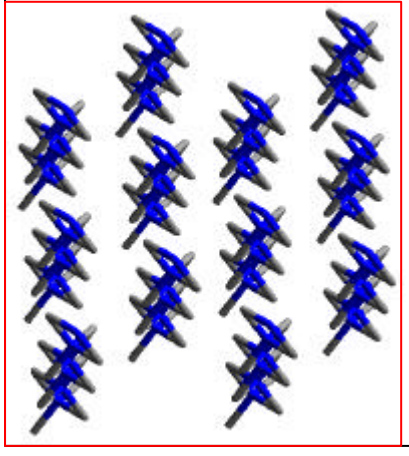
$$S = \frac{T_{cpu}(N_p = 2)}{T_{cpu}(N_p = 2^p)}$$

In (a): at 16 processes the speedup seems lower Because the data grid used is not optimally distributed

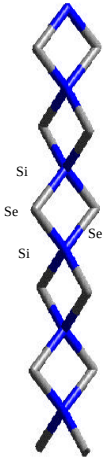
In (b) the lower of the speedup at 16 processes Is due to the same reason as in Fig.(a).

Lautrec code (courtesy of A. De Vita)

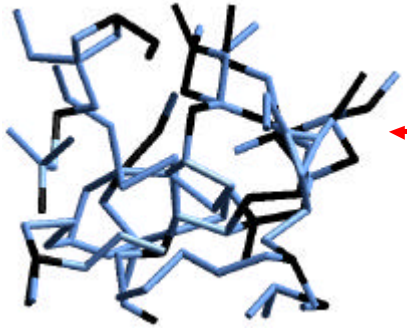
Crystalline $SiSe_2$



Structure characterized by parallel chains of distorted edge-sharing tetrahedra.

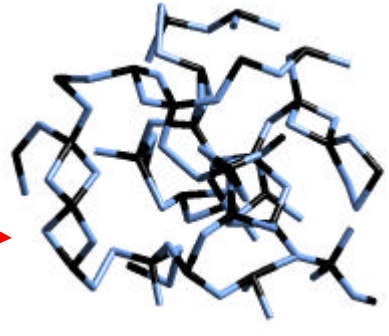


Applications in the field of batteries because when used as matrices for solid electrolytes they show high ionic mobility (Ag, Li and Na)



Starting configuration:

- homopolar bonds both Si-Si and Se-Se
- no Si tetrahedrally coordinated

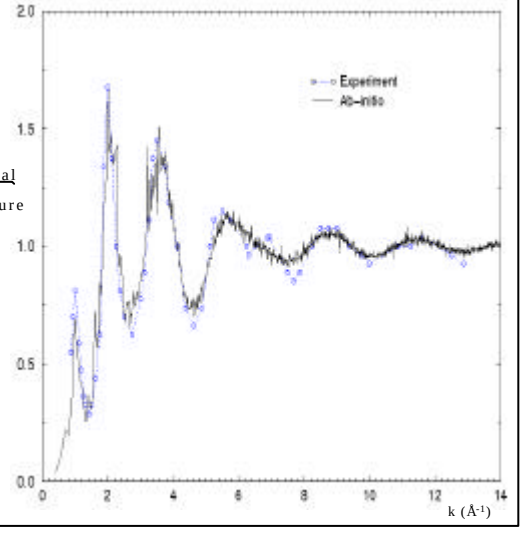


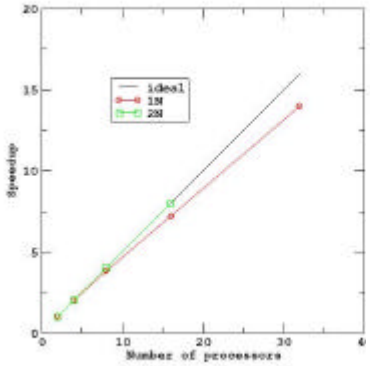
Equilibrated configuration:

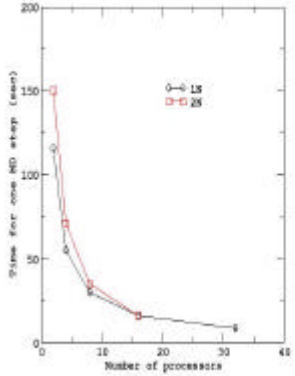
- a few Si-Si homopolar bonds
- most Si atoms tetrahedrally coordinated

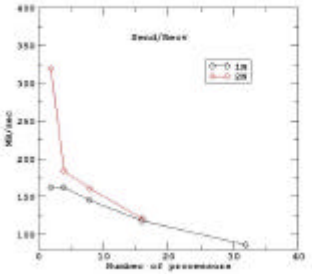
Total structure factor S(k)

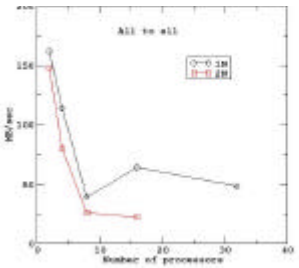
Comparison between experimental S(k) of a-$SiSe_2$ at room temperature and ab-initio S(k) of l-$SiSe_2$ at T = 1000 K.

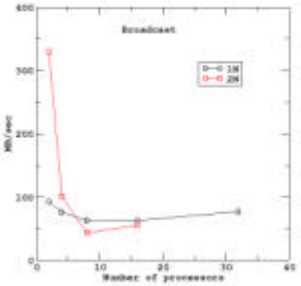



Speedup
Number of processors
ideal
1M
2M

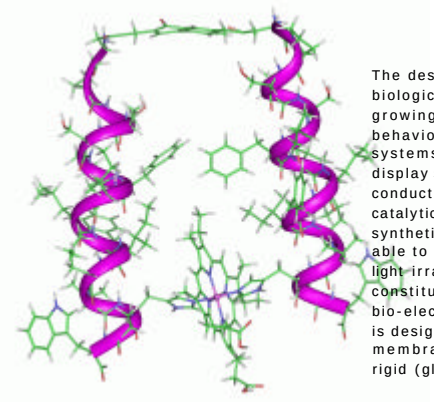

Time for one MD step (sec)
Number of processors
1M
2M


Send/Recv
MB/sec
Number of processors
1M
2M


All to all
MB/sec
Number of processors
1M
2M


Broadcast
MB/sec
Number of processors
1M
2M

Building a molecular bio-diode: design and modeling



The design and the synthesis of organic-biological structures have recently received a growing interest. These structures mimic behaviour and properties of complex biological systems and can be purposely designed to display specific properties (e.g. large conductivity, photoluminescence, specific catalytic properties etc.). Our aim is to build a synthetic system based on a proteic scaffold able to accomplish a fast electron transfer after light-irradiation. This molecule can thus constitute the basic element of a miniaturized bio-electronical device: a bio-diode. The system is designed to be immobilized on a lipid membrane that, in turn, can be mounted on a rigid (glass) substrate.

GROMACS

GROMACS is a general-purpose molecular dynamics computer simulation package for the study of biomolecular systems. Its purpose is threefold:

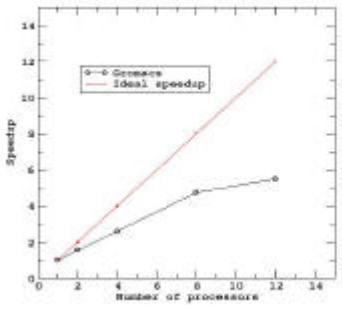
- Simulation of arbitrary molecules in solution or crystalline state by the method of molecular dynamics (MD), stochastic dynamics (SD) or the path-integral method.
- Energy minimisation of arbitrary molecules.
- Analysis of conformations obtained by experiment or by computer simulation.

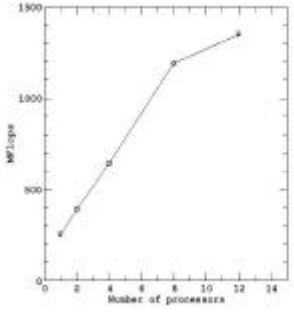
Berendsen, H.J.C., van der Spoel, D. and van Drunen, R., GROMACS: A message-passing parallel molecular dynamics implementation, Comp. Phys. Comm. 91 (1995), 43-56.

Lindahl, E., Hess, B. and van der Spoel, D., GROMACS 3.0: A package for molecular simulation and trajectory analysis J. Mol. Mod. 7 (2001) 306-317.

GROMACS BENCHMARK

A phospholipid membrane, consisting of 1024 DPPC lipids in a bilayer configuration with 23 water molecules per lipid, for a total of 121,856 atoms. It was simulated with a twin-range group based cut-off of 1.8 nm for electrostatics and 1.0 nm for Van der Waals interactions. The long-range Coulomb forces between 1.0 nm and 1.8 nm were updated every tenth integration step during neighborlist generation. The force field described by Berger et al (1997) was used for the lipids while the water was simulated with the SPC model.

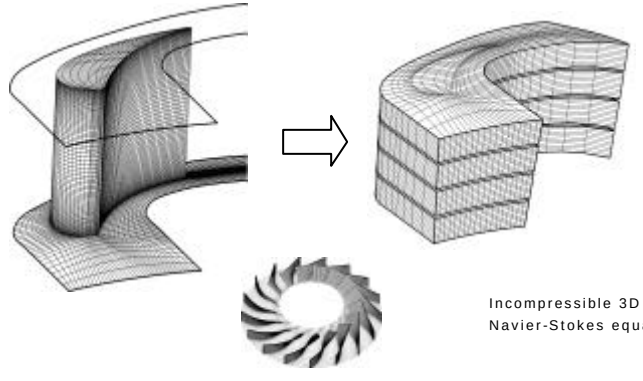




Parallel Computations for Turbomachinery

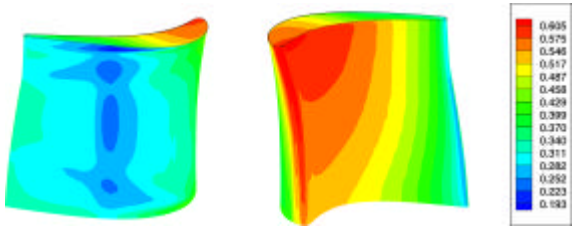
Paolo Giangiacomo, Vittorio Michelassi
University of Roma Tre, Department of Mechanical and Industrial Engineering
michelas@uniroma3.it

Axial turbine stator



Incompressible 3D Euler and Navier-Stokes equations

Surface Pressure Distributions

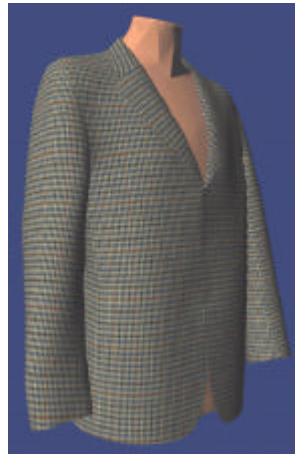


Details of the tip flow

Distributed implementation of a particle-based cloth simulator.

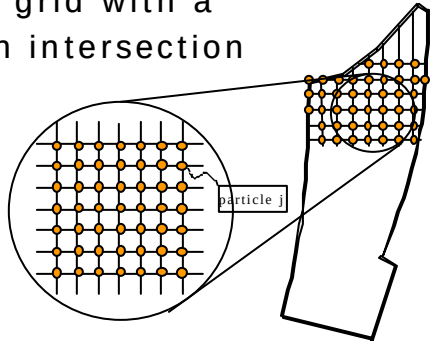
A. Galimberti

KAEMaRT Group
Industrial Engineering
Department
Parma University



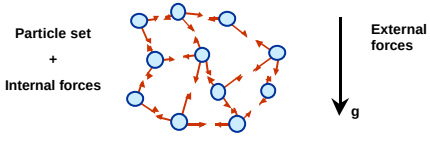
Particles

- A piece of fabric is modeled with a squared grid with a particle at each intersection
- The grid is aligned with warp and weft directions



Non rigid bodies

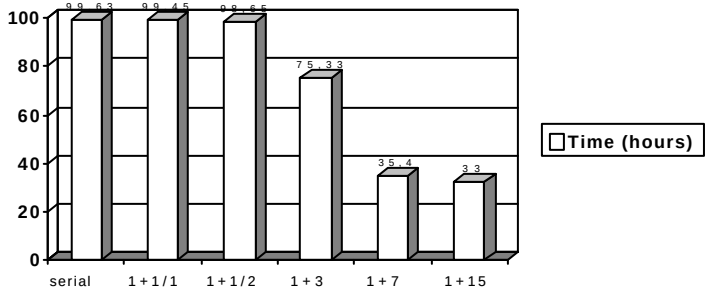
- A non rigid body is a set of particles connected by forces (at least springs) with external forces acting on it



- Simulation is performed by an ODE solver and is based on Newton's laws

Execution on Feronia cluster

- 1 second of simulation
  - 1 master + *n* slaves running on *m* different nodes



The Distributed Memory MM5 on QSW Alpha Linux Beowulf @ ENEA

B. Tomassetti [1,2], G. Visconti [1], F.Valentinotti [3], G. Giuliani [2], and L. Bernardini [2]

(1) CETEMPS, University of L'Aquila (Italy)

(2) Scientific and Technology Park of Abruzzo (Italy)

(3) Quadrics Supercomputers World Ltd.

Barbara.Tomassetti@aquila.infn.it
franco.valentinotti@roma.quadrics.com

The Meteorological MM5 Model

Fifth-Generation NCAR/Penn State Mesoscale Model

- A 3D primitive equations (equations of momentum, mass continuity, and energy conservation) model
- Prognostic variables: wind, temperature, specific humidity, and pressure ((U, V), T, Q, P)
- non-hydrostatic dynamics
- finite difference technique in time and space
- a multiple-nest capability (one or two way)
- more physics options
- several features important for climate regional prediction

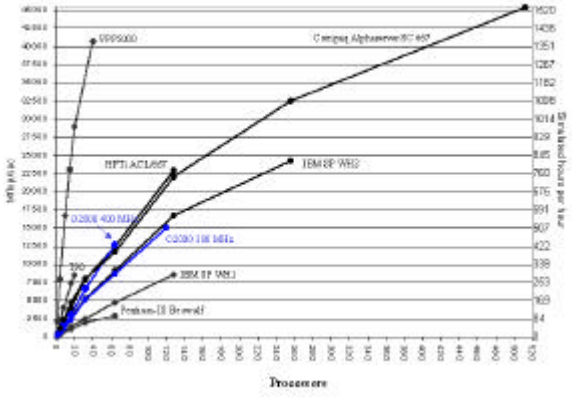
The Climatological run on the Fucino lake

Very high grid resolution: in order to "see" the presence of the lake (150 km^2)

Multiple nested domains (3 domains: 27 km, 9 km, and 3 km)

Very long simulation time, a **season,** to be repeat **twice** in order to isolate the differences between the current vs. old (lake) situation.

MM5 Performance: NCAR/t3a benchmark



t3a **benchmark features**

Res.	**36 km**
Dt	**91 s**
NX	**136 cells**
NY	**112 cells**
Nz	**33 layers**
Total cells	**503000**
Radiat.	**1/22 step**
Gflop	**2.398**

National Centre for Atmospheric Research (NCAR)
http://www.mmm.ucar.edu/mm5/mm5-home.html

IAPSO - IABO

HYDRAULIC REGIME WITHIN THE STRAIT OF GIBRALTAR: 3D NUMERICAL SIMULATION

Sannino G., A. Bargagli and V. Artale

gianmaria.sannino@casaccia.enea.it

Divisione Ambiente Globale e Mediterraneo
Sezione Clima

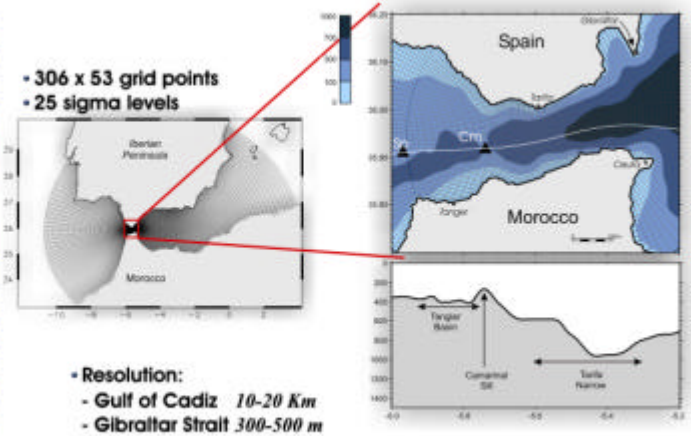
ENTE PER LE NUOVE TECOLOGIE, L'ENERGIA E L'AMBIENTE

GRID & BATHYMETRY

ENTE PER LE NUOVE TECNOLOGIE L'ENERGIA E L'AMBIENTE

- Outline
- Geography
- Background
- 3D-Model
- Description
- Grid & Bathym.
- Init. & B. Cond.
- Advec. Scheme
- Results
- Hydraulic Contr.
- Future Work

- 306 x 53 grid points
- 25 sigma levels



- Resolution:
  - Gulf of Cadiz *10-20 Km*
  - Gibraltar Strait *300-500 m*
  - Alboran Sea *8-15 Km*

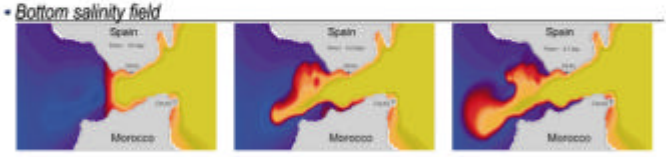
SPINUP PHASE

ENTE PER LE NUOVE TECNOLOGIE L'ENERGIA E L'AMBIENTE

- Outline
- Geography
- Background
- 3D-Model
- Results
- SpinUp Phase
- Transports
- Obser. VS Model
- Salinity Sections
- Hydraulic Contr.
- Future Work

- *Surface salinity field*

- *Bottom salinity field*



▸Two superimposed density currents begin to form and propagate, with a frontal speed of 57 *cm/sec* and 69 *cm/sec* for the Atlantic and Mediterranean waters, respectively.

HPCN in Internet:

http://www.enea.it/hpcn

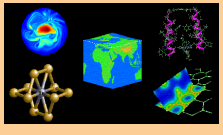
High Performance Computing and Networking

HPCN

Interdepartmental Project

HPCN at ENEA | >> Welcome to the ENEA-HPCN Website << | Related links

General Info
Research
HPC Resources
People
Publications
Opportunities



Events
Other Websites

ENEA Website

Current News

"Calcolo e Reti ad Alte Prestazioni in ENEA", Internal Note, 25.09.2000 *(in Italian)*

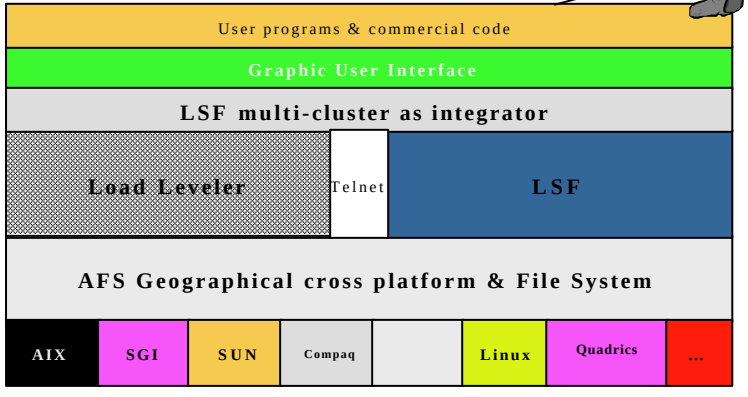
ENEA-HPCN
Lungotevere G.A. Thaon di Revel, 76
00196 ROMA - *(Italy)*
Phone: +39 -063627-2570 Fax: +39-063627-2683

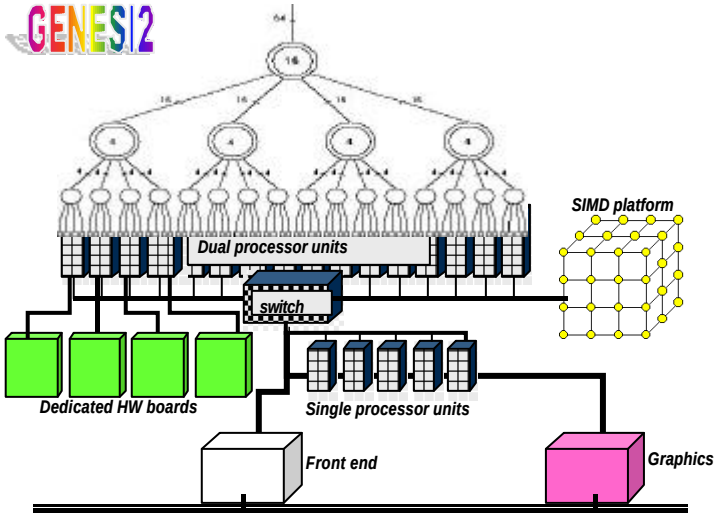
Disclaimer

All comments, error reports and requests for alterations etc., should go to Paolo Novelli or Massimo Celino

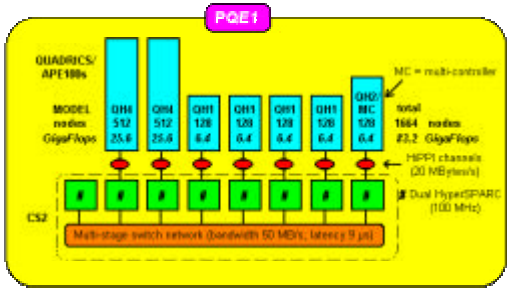
ENEA

Software Conceptual Structure





Hardware in the ENEA Research Centre Casaccia (1/3)



PQE1 is a 'heterogeneous' parallel system composed by a general purpose MIMD platform (Meiko/QSW CS-2) coupled to 7 SISAMD (single instruction single address multiple data) platforms (APE100/Quadrics).

The **APE100/Quadrics SIMD** section has 1664 nodes, 83.2 Gigaflops of aggregate computational speed, 20.8 Gigabytes/sec of bandwidth and 6.5 Gigabytes of RAM.

The **CS**-2 MIMD section has 8 twin nodes, 1Gigaflops of peak speed, 1 Gigabyte of RAM and 800 Megabytes/sec of aggregate bandwidth. The SIMD systems communicate through 7 HiPPI channels with the MIMD section, so the communication bandwidth between the two systems is 140 Megabytes/sec.