Introduction to Shared memory architectures

Carlo Cavazzoni, HPC department, CINECA

Modern Parallel Architectures

Two basic architectural scheme:

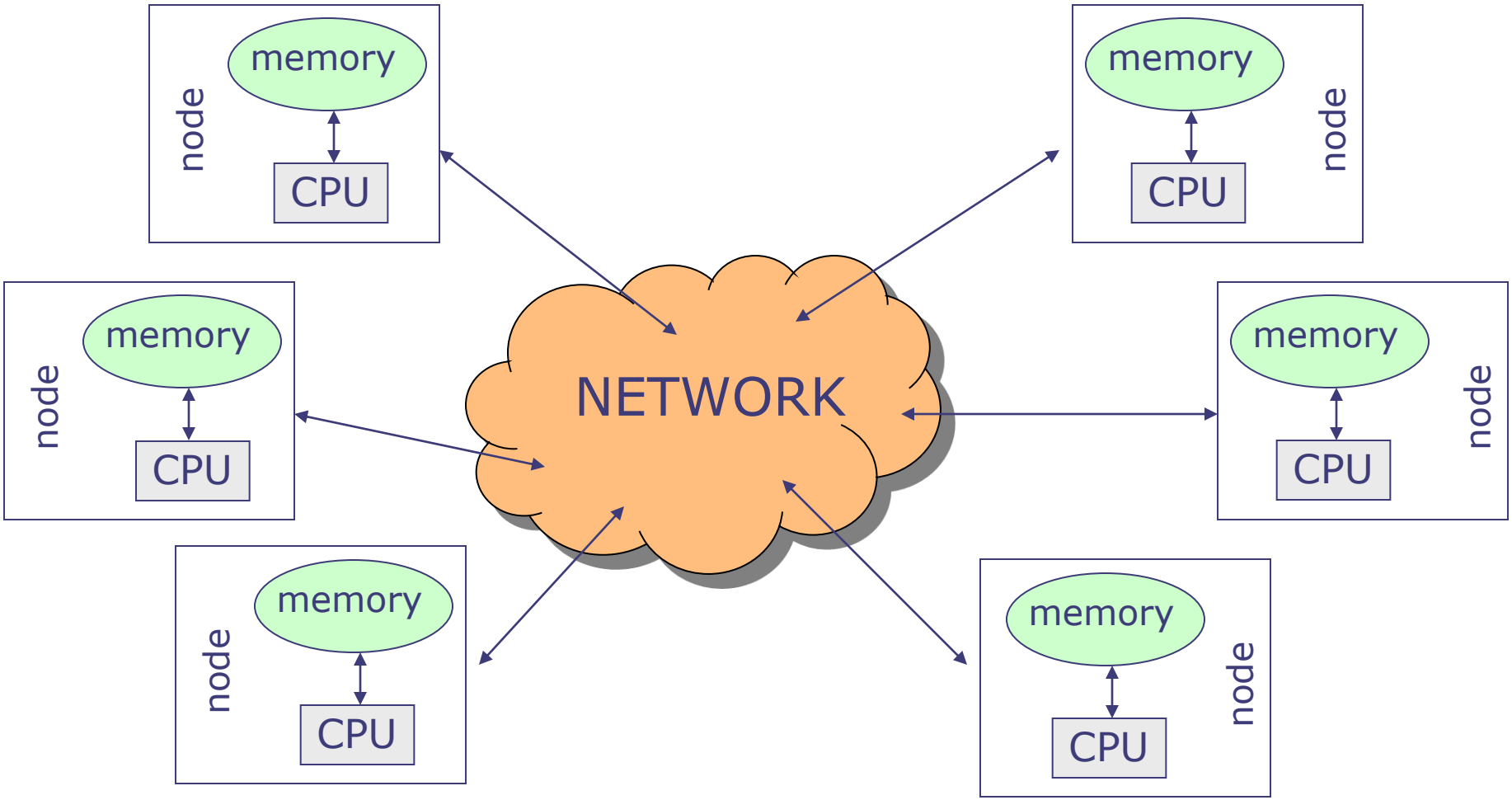
Distributed Memory

Shared Memory

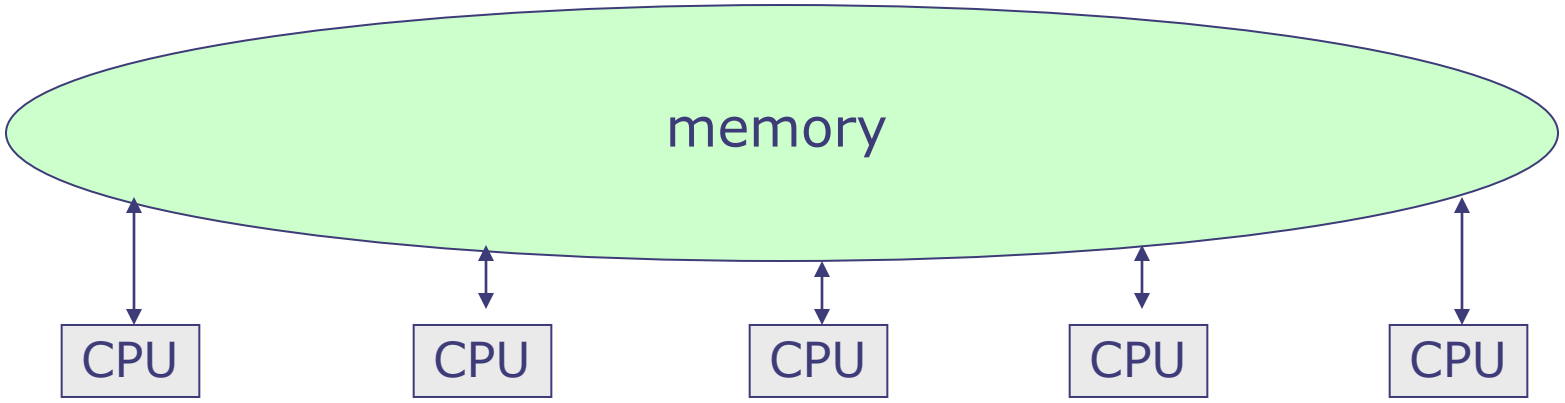
Now most computers have a mixed architecture

+ accelerators -> hybrid architectures

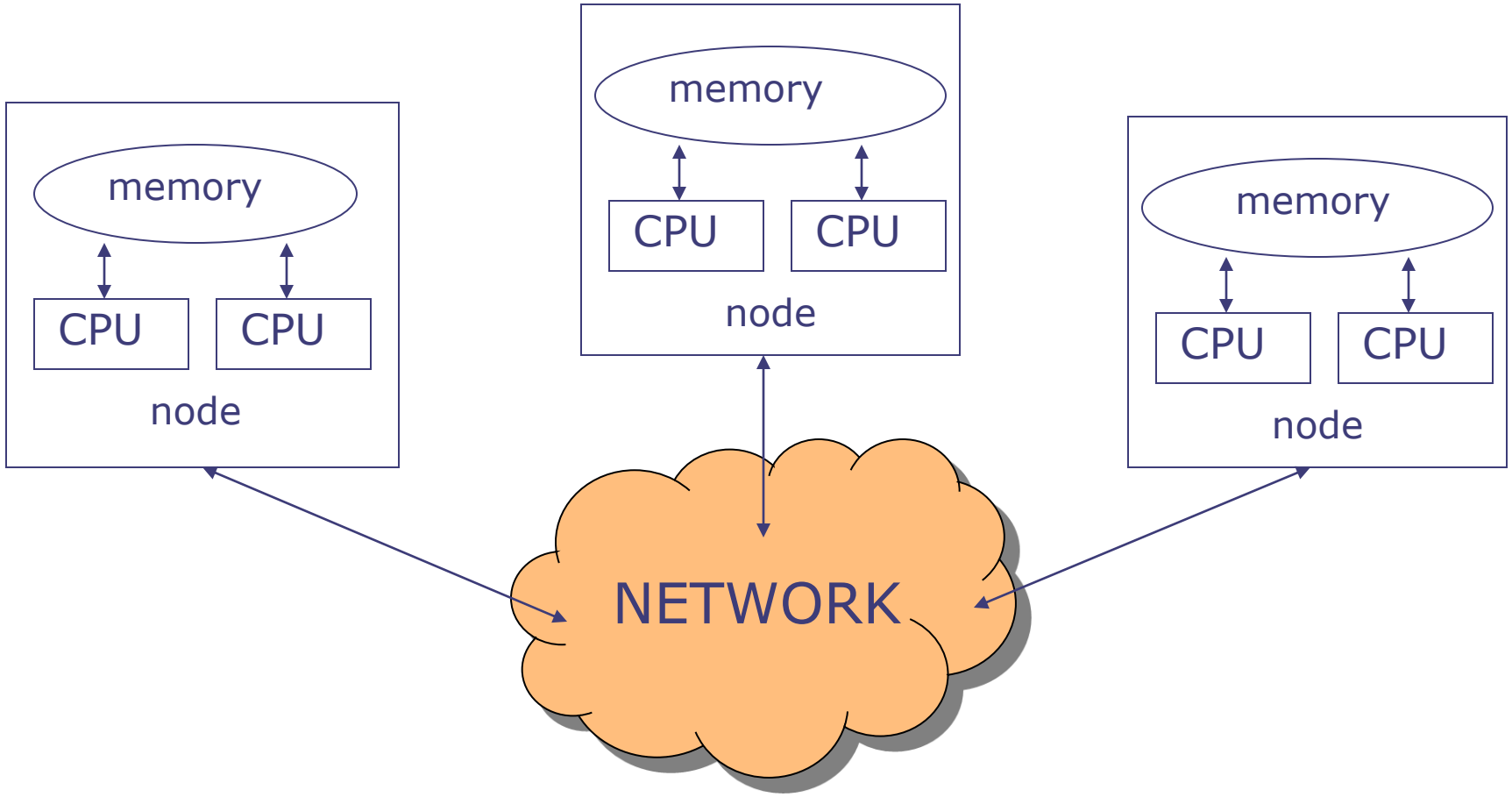
Distributed Memory



Shared Memory

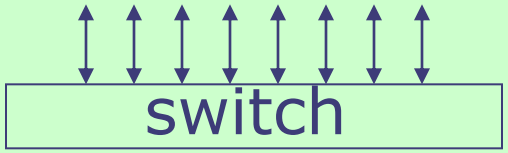


Mixed Architectures

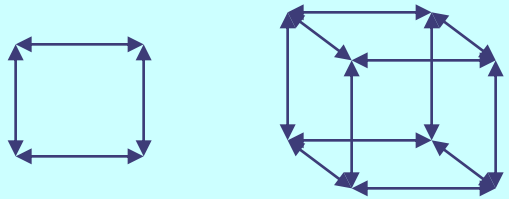


Most Common Networks

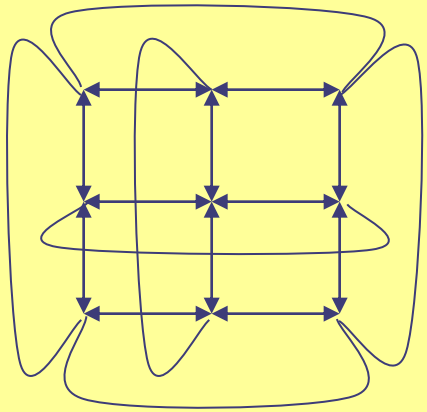
switched



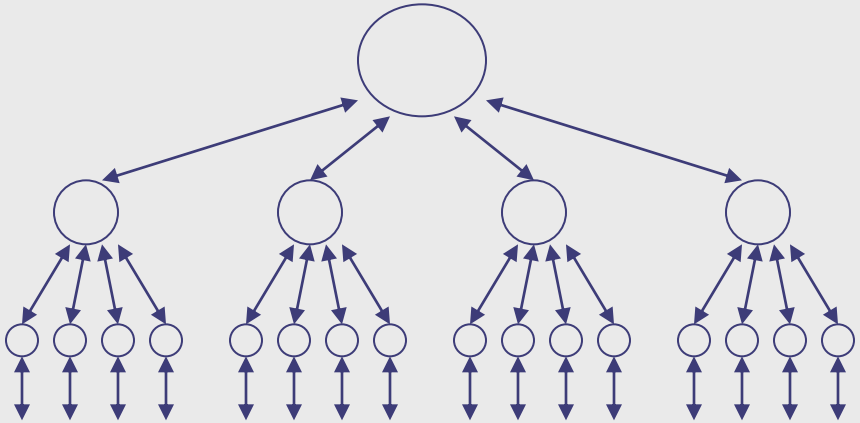
Cube, hypercube, n-cube



Torus in 1,2,...,N Dim



Fat Tree



Roadmap to Exascale
(architectural trends)

Systems	2009	2011	2015	2018
System Peak Flops/s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	O(100)	O(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	O(Million)
Total Concurrency	225,000	3 Million	50 Million	O(Billion)
Storage	15 PB	30 PB	150 PB	300 PB
I/O	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	Days	Days	Days	O(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

Dennard scaling law

- new gen.
  - old gen.

$L' = L / 2$
$V' = V / 2$
$F' = F * 2$
$D' = 1 / L^2 = 4D$
$P' = P$

do not hold anymore!

The core frequency and performance do not grow following the Moore's law any longer

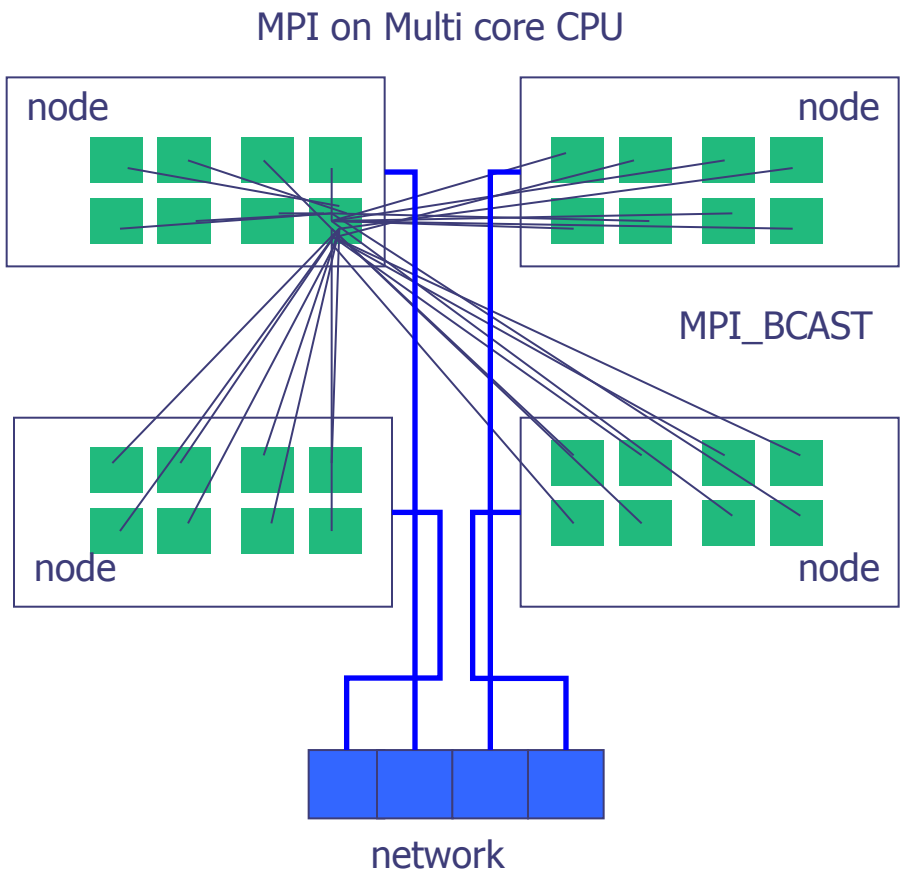
$L' = L / 2$
$V' = \sim V$
$F' = \sim F * 2$
$D' = 1 / L^2 = 4 * D$
$P' = 4 * P$

The power crisis!

Increase the number of cores to maintain the architectures evolution on the Moore's law

Programming crisis!

MPI inter process communications



1 MPI proces / core
Stress network
Stress OS

Many MPI codes (QE) based on ALLTOALL
Messages = processes * processes

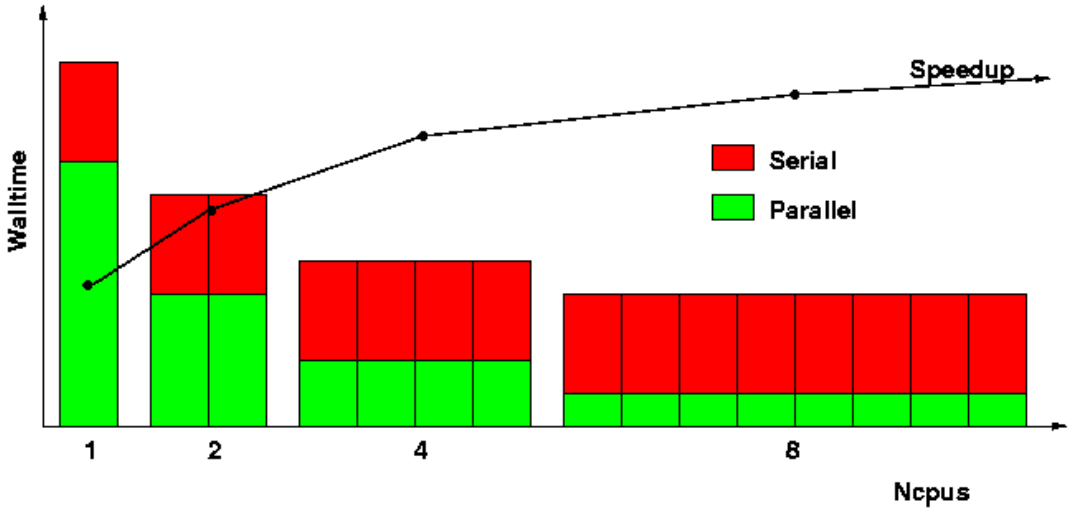
We need to exploit the hierarchy

Re-design applications

Mix message passing And multi-threading

What about Applications?

In a massively parallel context, an upper limit for the scalability of parallel applications is determined by the fraction of the overall execution time spent in non-scalable operations (Amdahl's law).



maximum speedup tends to
$1 / (1 - P)$
P= parallel fraction

1000000 core

$P = 0.999999$

serial fraction= 0.000001

FERMI (BGQ) total concurrency: 655360 (65536/rack)

Programming Models

- **Message Passing (MPI)**
- **Shared Memory (OpenMP)**
- Partitioned Global Address Space Programming (PGAS) Languages
  - UPC, Coarray Fortran, Titanium
- Next Generation Programming Languages and Models
  - Chapel, X10, Fortress
- **Languages and Paradigm for Hardware Accelerators**
  - CUDA, OpenCL
- **Hybrid: MPI + OpenMP + CUDA/OpenCL**

MPI + OpenMP

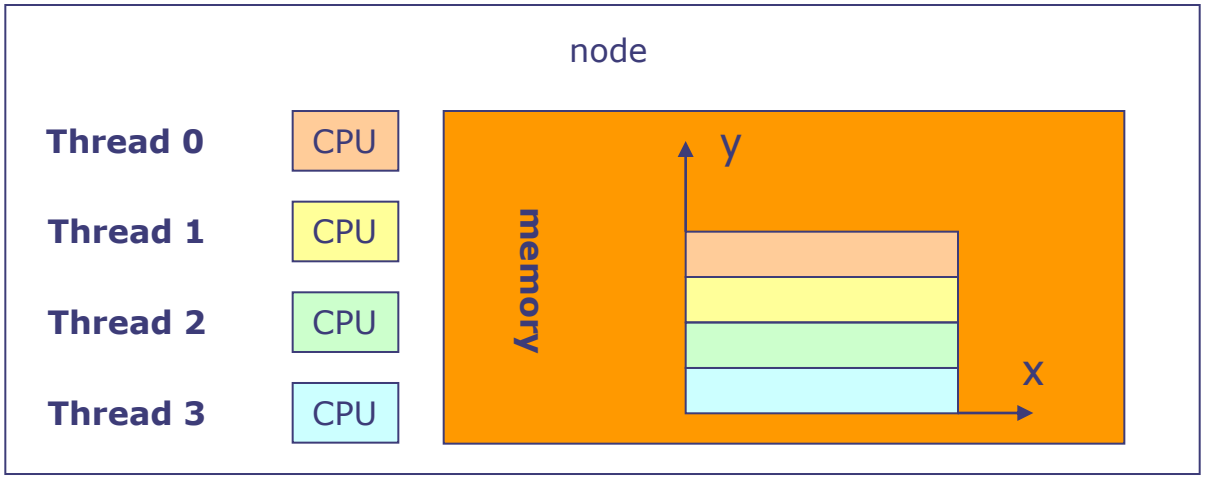
MPI (inter node)

Distributed memory systems
message passing
data distribution model
Version 2.1 (09/08)
API for C/C++ and Fortran

OpenMP (intra node)

Shared memory systems
Threads creations
relaxed-consistency model
Version 3.0 (05/08)
Compiler directive C and Fortran

Shared memory



Shared Memory: OpenMP

Main Characteristic

- Compiler directives
- Medium grain
- Intra node parallelization (pthreads)
- Loop or iteration partition
- Shared memory
- Many HPC App

Open Issue

- Thread creation overhead
- Memory/core affinity
- Interface with MPI

OpenMP

```
!$omp parallel do
do i = 1 , nsl
   call 1DFFT along z ( f [ offset( threadid ) ] )
end do
!$omp end parallel do
call fw_scatter ( . . . )
!$omp parallel
do i = 1 , nzl
!$omp parallel do
  do j = 1 , Nx
    call 1DFFT along y ( f [ offset( threadid ) ] )
  end do
!$omp parallel do
  do j = 1, Ny
    call 1DFFT along x ( f [ offset( threadid ) ] )
  end do
end do
!$omp end parallel
```

PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

FERMI @ CINECA

PRACE Tier-0 System

Architecture: 10 BGQ Frame
Model: IBM-BG/Q
Processor Type: IBM PowerA2, 1.6 GHz
Computing Cores: 163840
Computing Nodes: 10240
RAM: 1GByte / core
Internal Network: 5D Torus
Disk Space: 2PByte of scratch space
Peak Performance: 2PFlop/s

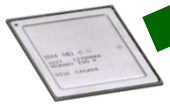
ISCRA & PRACE call for projects now open!

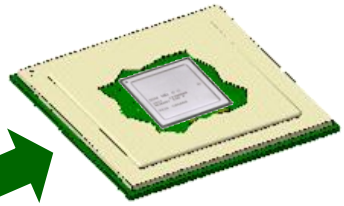
4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips, Torus

3. Compute card:
One chip module,
16 GB DDR3 Memory,

2. Single Chip Module

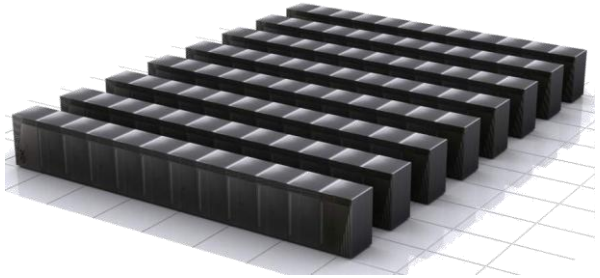
1. Chip:
16 λP cores





5b. IO drawer:
8 IO cards w/16 GB
8 PCIe Gen2 x8 slots

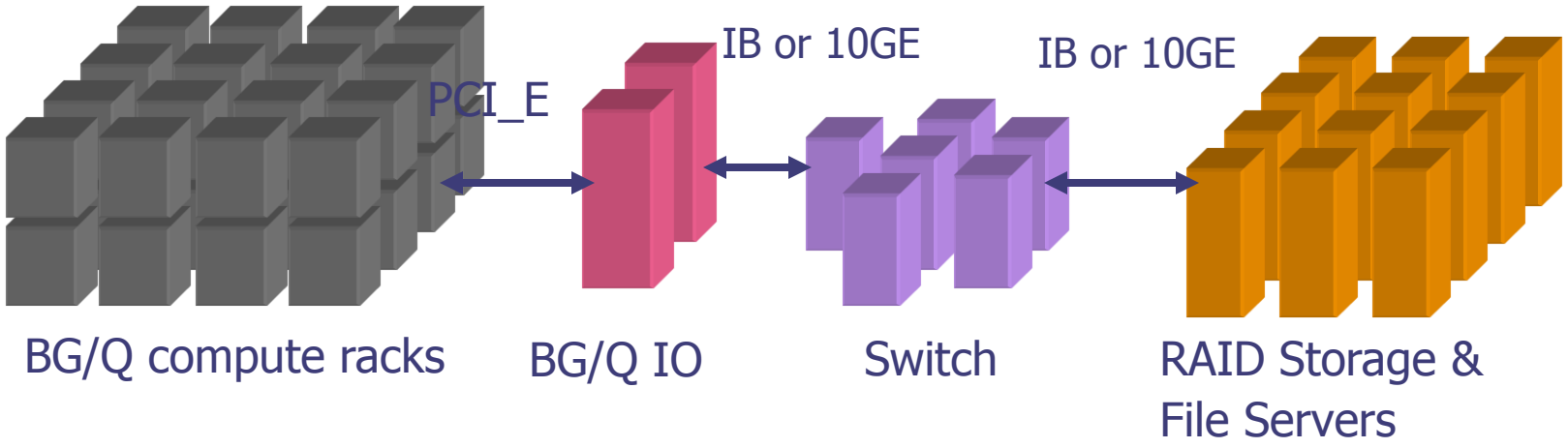
7. System:
20PF/s



5a. Midplane:
16 Node Cards

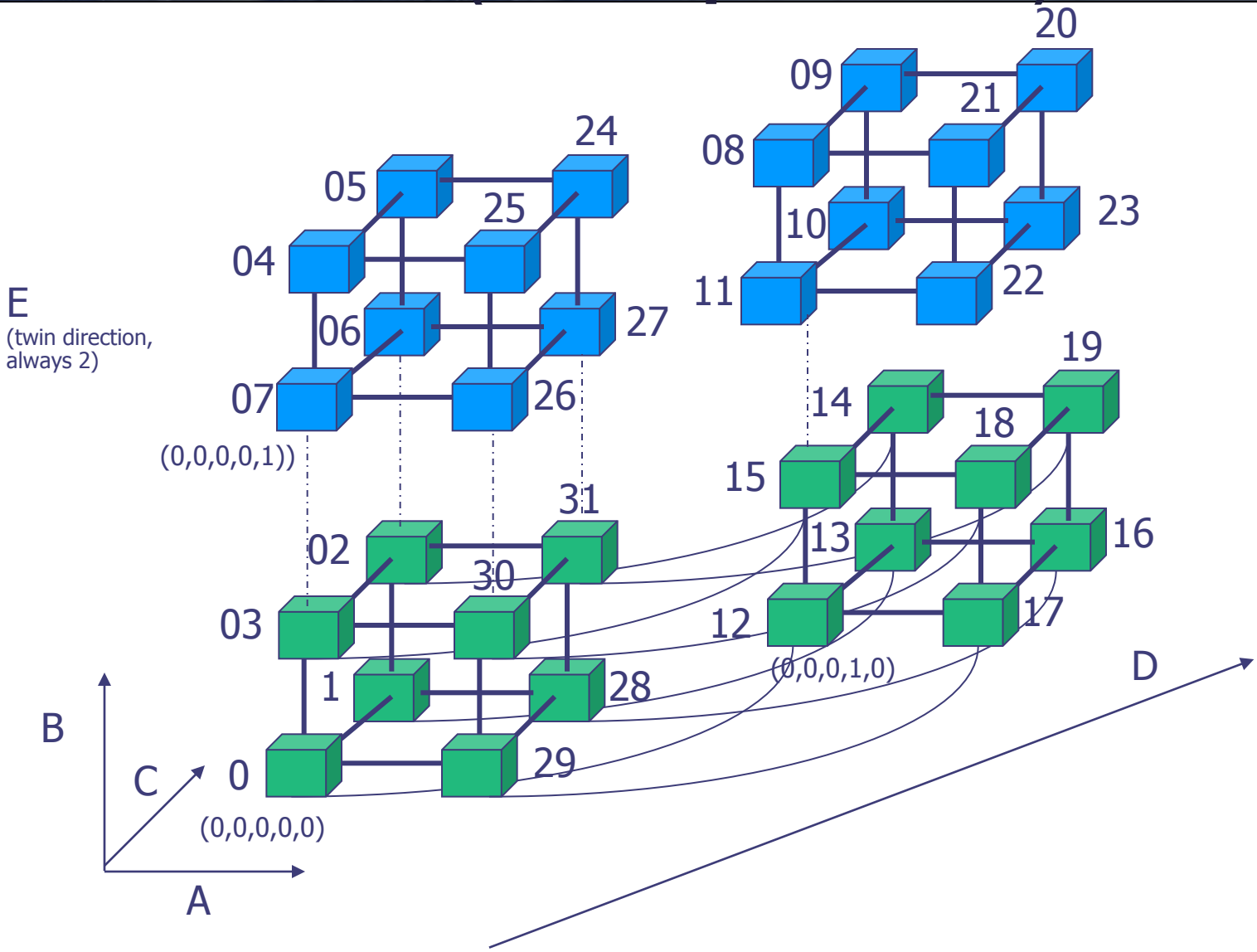
6. Rack: 2 Midplanes

BG/Q I/O architecture



- BlueGene Classic I/O with GPFS clients on the logical I/O nodes
- Similar to BG/L and BG/P
- Uses InfiniBand switch
- Uses DDN RAID controllers and File Servers
- BG/Q I/O Nodes are not shared between compute partitions
  - **IO Nodes are bridge data from function-shipped I/O calls to parallel file system client**
- Components balanced to allow a specified minimum compute partition size to saturate entire storage array I/O bandwidth

Node Board (32 Compute Nodes): 2x2x2x2x2

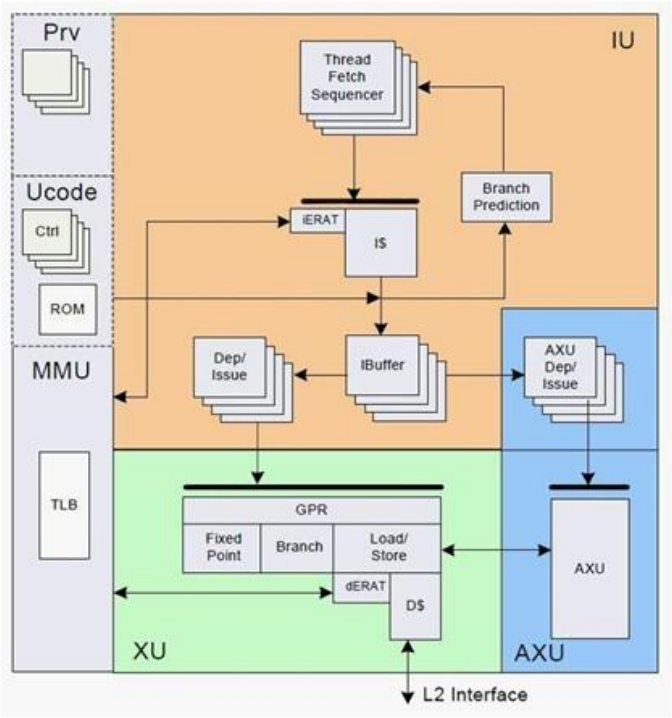


PowerA2 chip, basic info

16 cores + 1 + 1
1.6GHz
32MByte cache
system-on-a-chip design
16GByte of RAM at 1.33GHz
Peak Perf 204.8 gigaflops
power draw of 55 watts
45 nanometer copper/SOI process (same as Power7)
Water Cooled

PowerA2 core

4 FPU
4 way SMT
64-bit instruction set
in-order dispatch, execution, and completion
16KB of L1 data cache
16KB of L1 instructions cache



PowerA2 FPU

Each FPU on each core has four pipelines
execute scalar floating point instructions
Quad pumped
four-wide SIMD instructions
two-wide complex arithmetic SIMD inst.
six-stage pipeline
permute instructions
maximum of eight concurrent
floating point operations
per clock plus a load and a store.