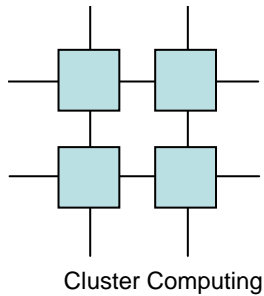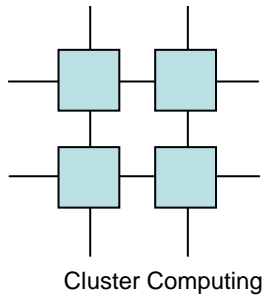Cluster Computing

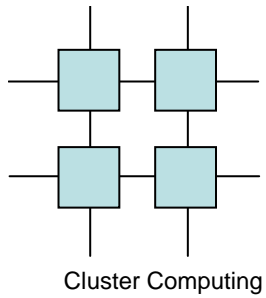# Massively Parallel Architectures

MPP Specifics

- No shared memory
- Scales to hundreds or thousinds of processors
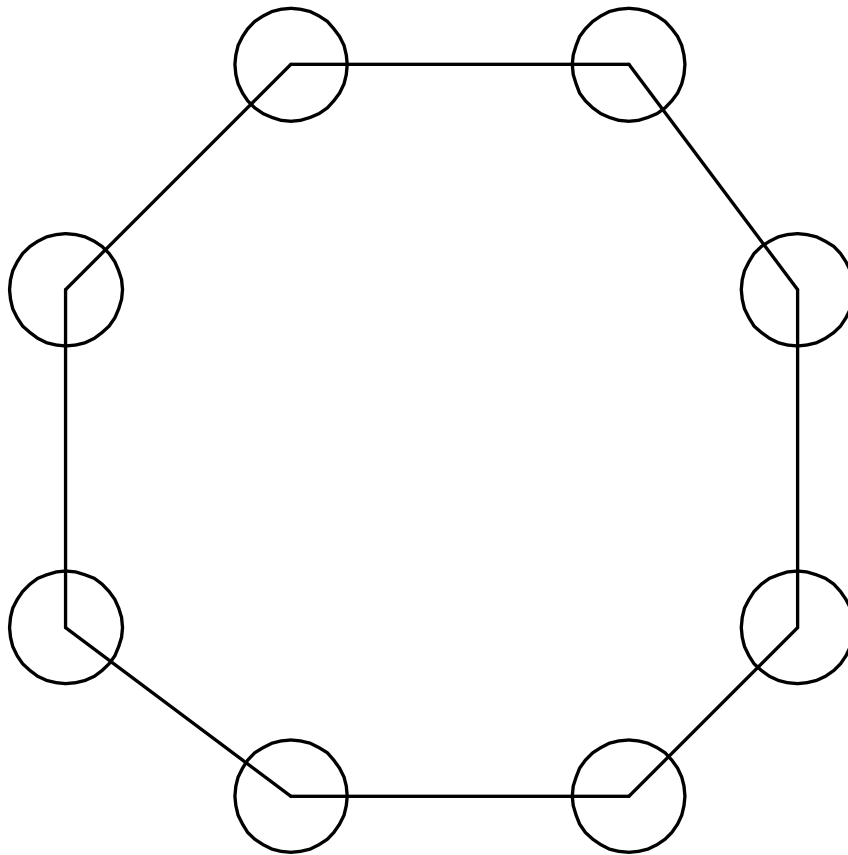- Homogeneous sub-components
- Advanced Costum Interconnects

MPP Architectures

Cluster Computing

- There are numerous approaches to interconnecting CPUs in MPP architectures:
  - Rings
  - Grids
  - Full Interconnect
  - Trees
  - Dancehalls
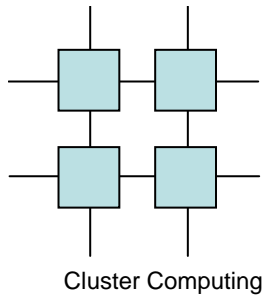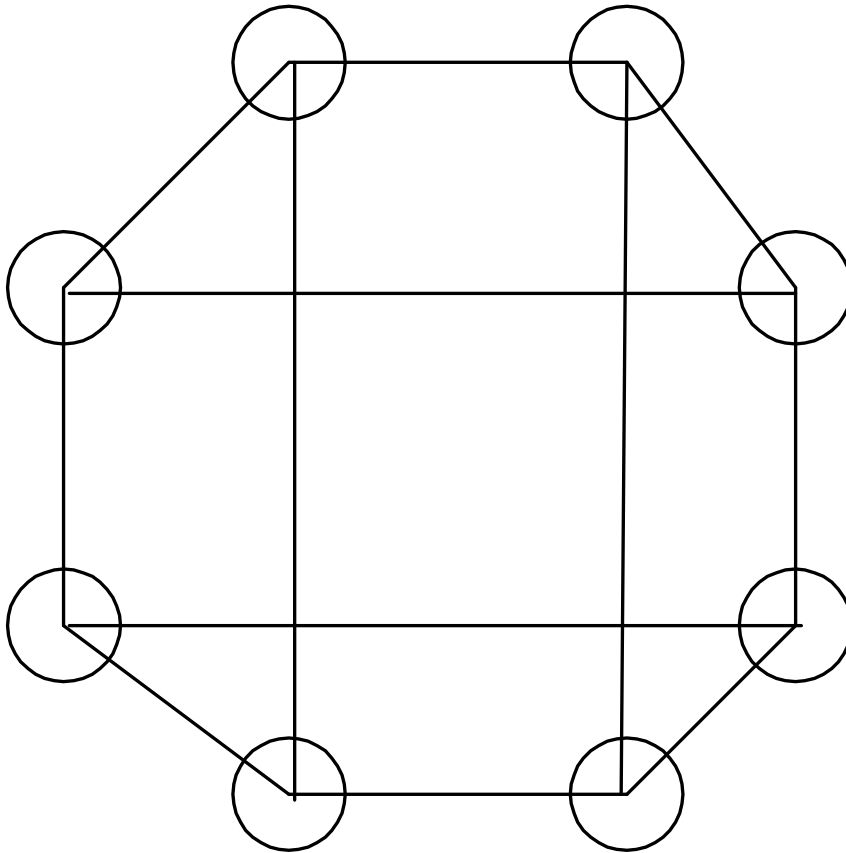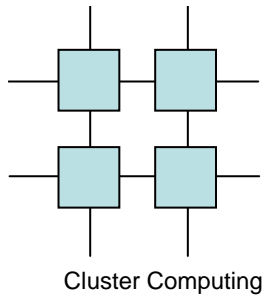 - Hybercubes

Rings

Worst case distance

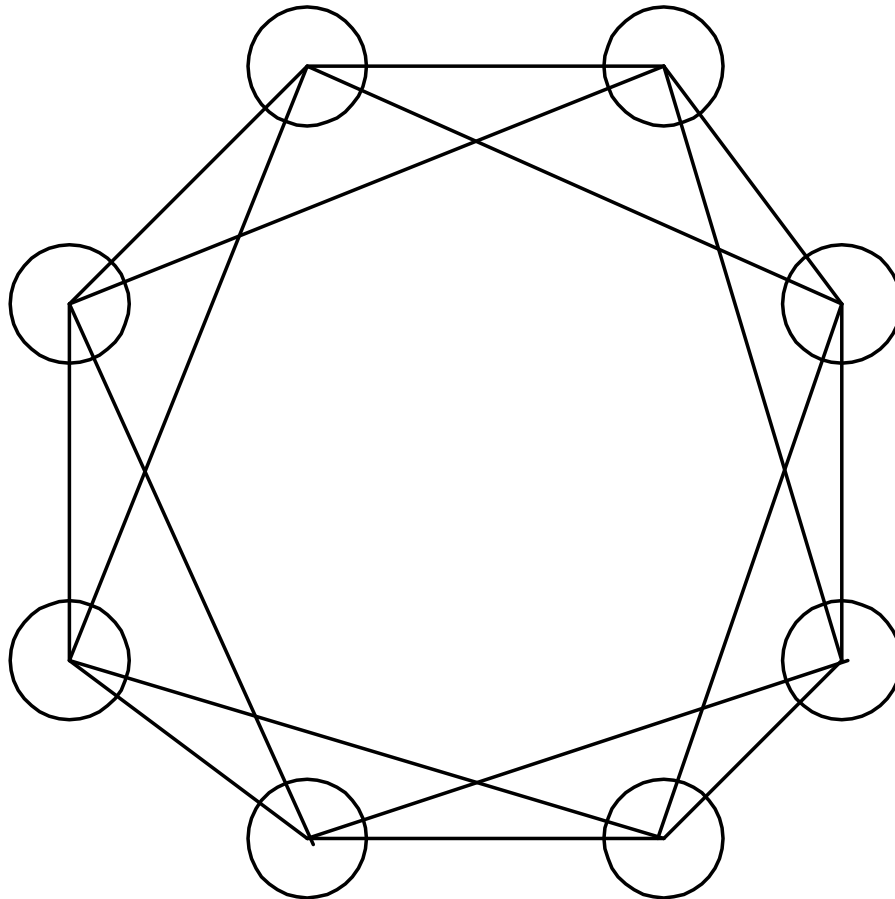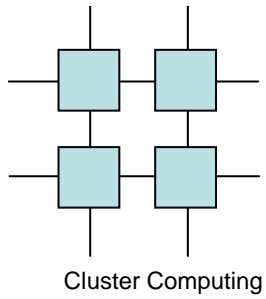$n-1$ (one ring)

$\dfrac{n}{2}$ (bi-directional ring)
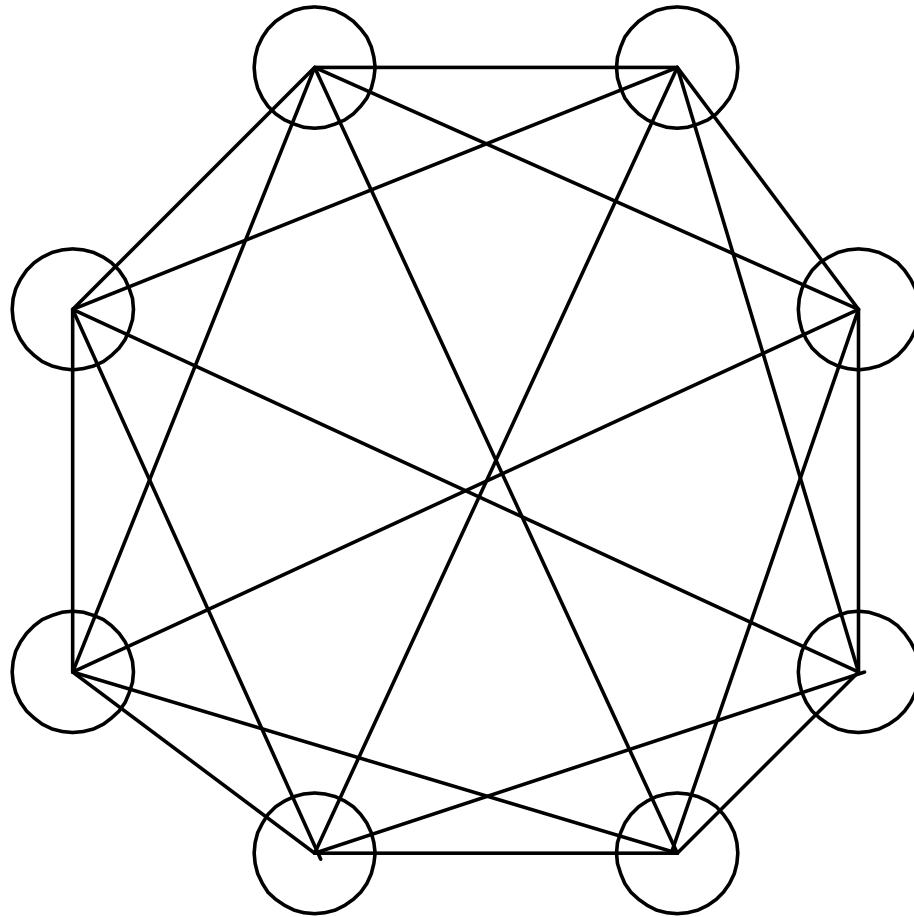
Cost n

Chordal Ring 3

Cluster Computing

Chordal Ring 4

Cluster Computing

Barrel Shifter

Cluster Computing
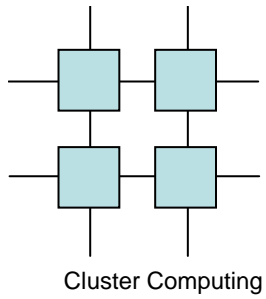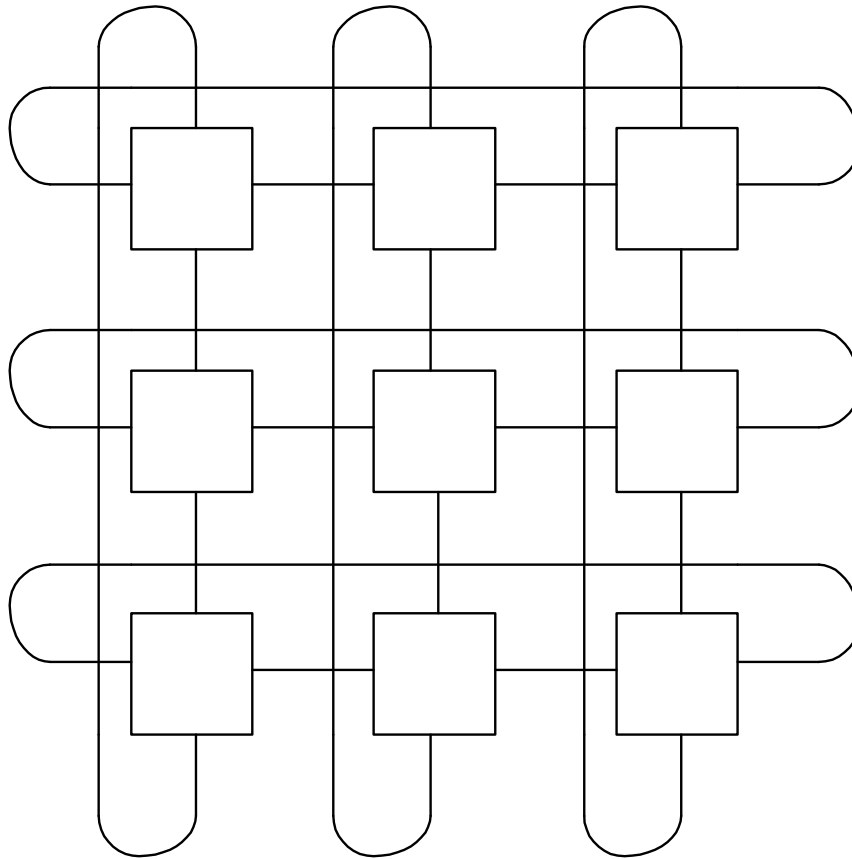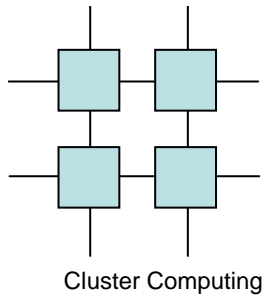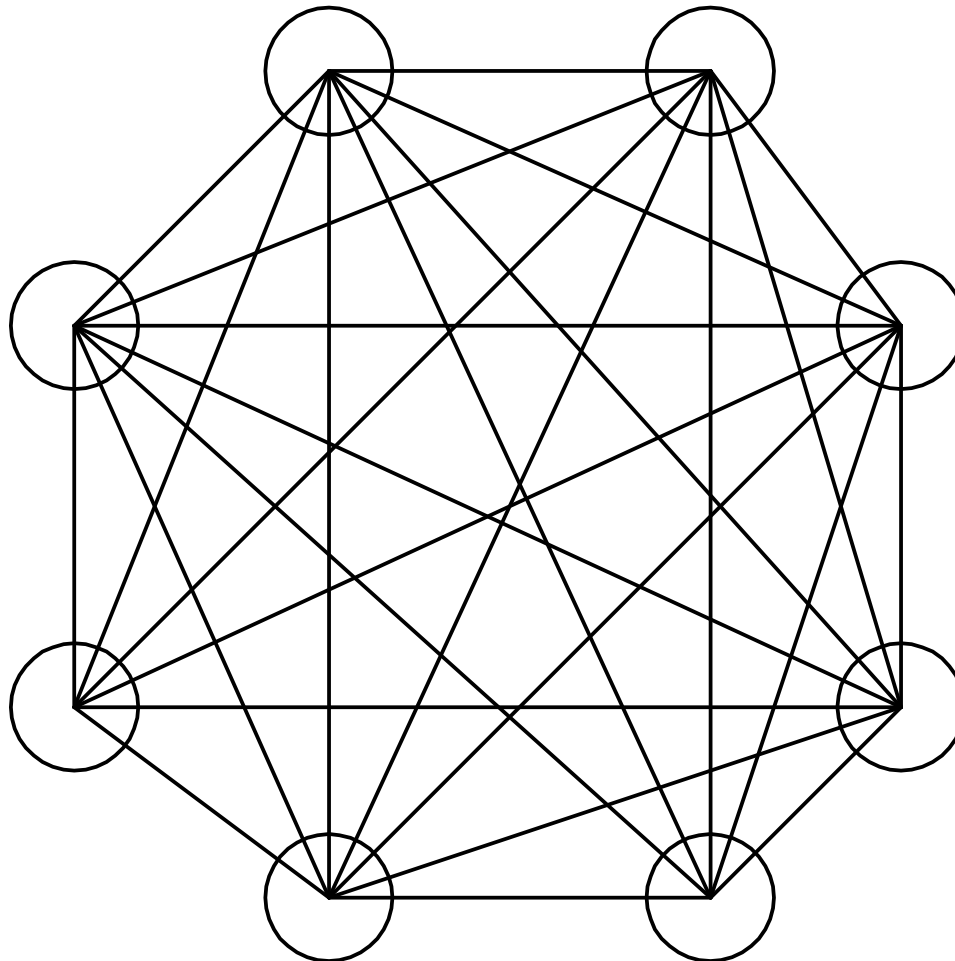
Worst case distance
n/2

Grid/Torus/Illiac Torus
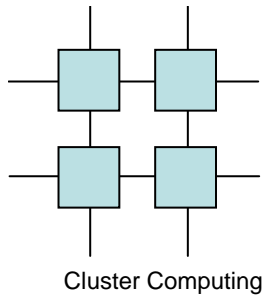
#worst case distance
$\sqrt{n}$
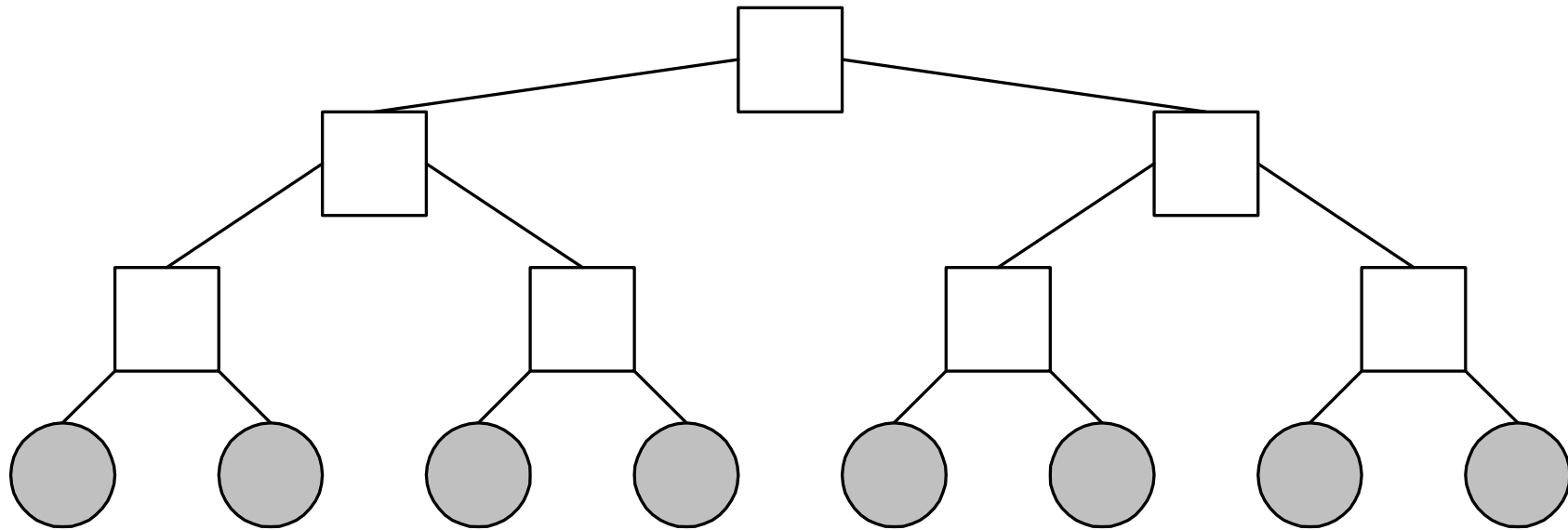
Cost n

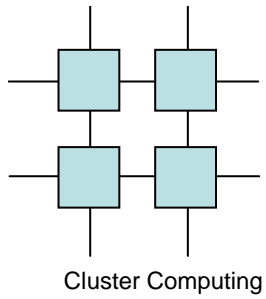# Fully interconnected

worst case distance 1

Cost $n(n-1)/2$

Trees

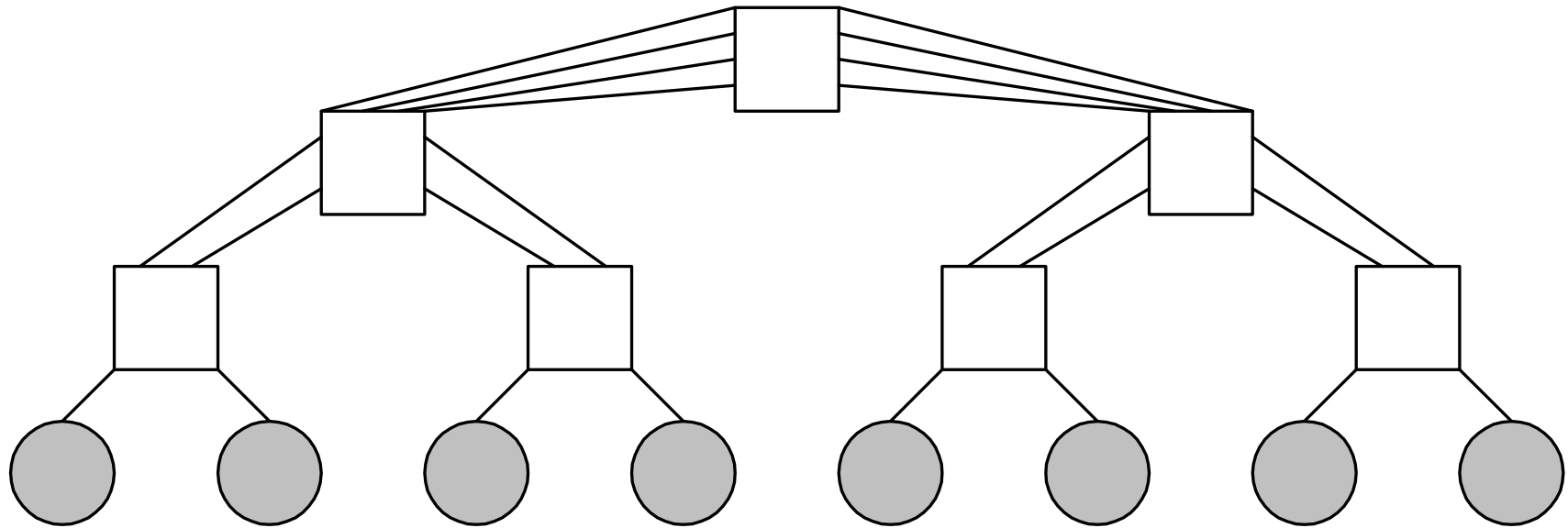worst case 2log n

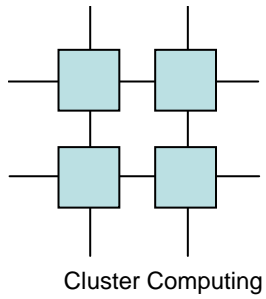Cost n log n
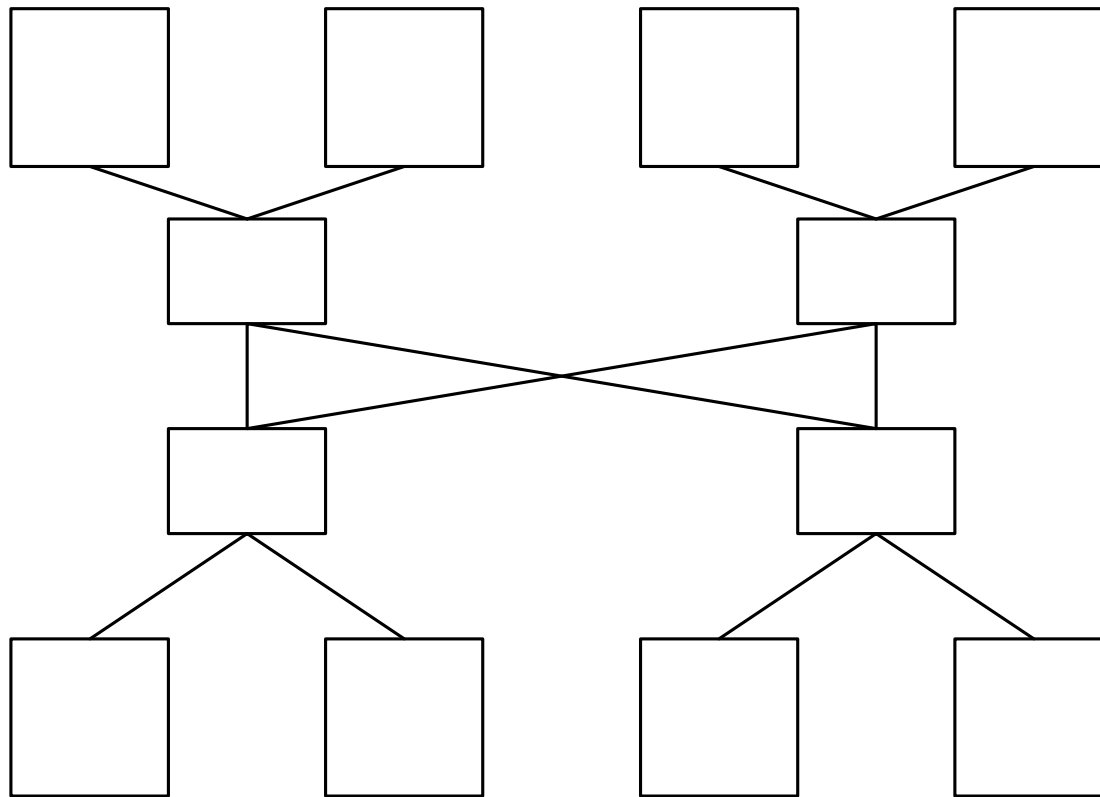
Fat Trees

worst case 2log n   Cost n
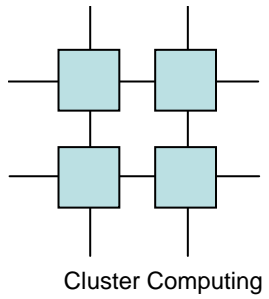
Dancehalls/Butterflys

distance

$\log n$

Cost n

Hybercubes
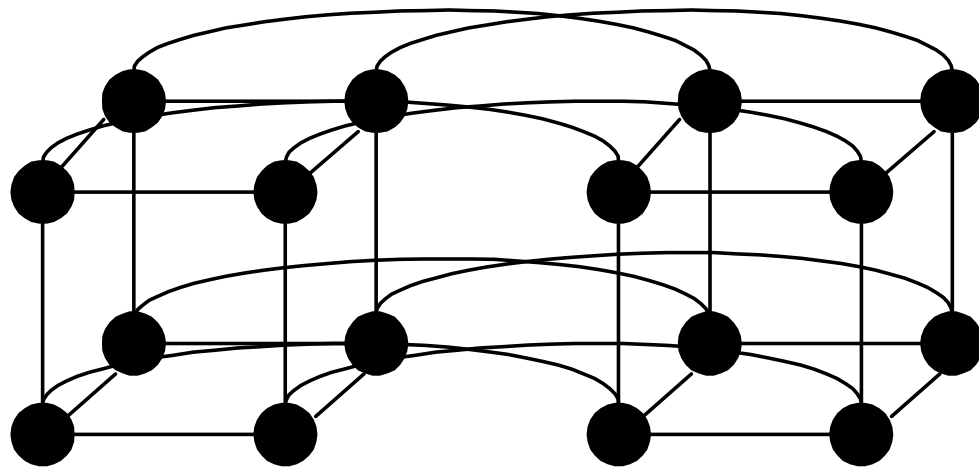
worst case distance

d

Cost d
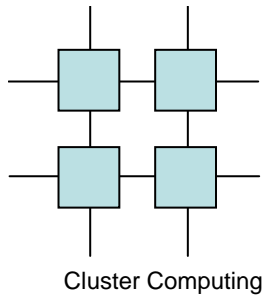
Intel Paragon

- Intel i860 based machine
- "Dual CPU" –
  - 50 MHz CPUs
  - Shares a 400 MB/sec cache coherent bus
- Grid architecture
- Mother of ASCI Red

Cluster Computing

Intel Paragon

Cluster Computing

Network

175 MB/sec

Mem

NIC

2KB FIFOs

64-bit
400 MB/s
Cache-Coherent
Mem. Bus

L1 Cache

i860XP

Compute Processor

L1 Cache

i860XP

Message Processor

DMA
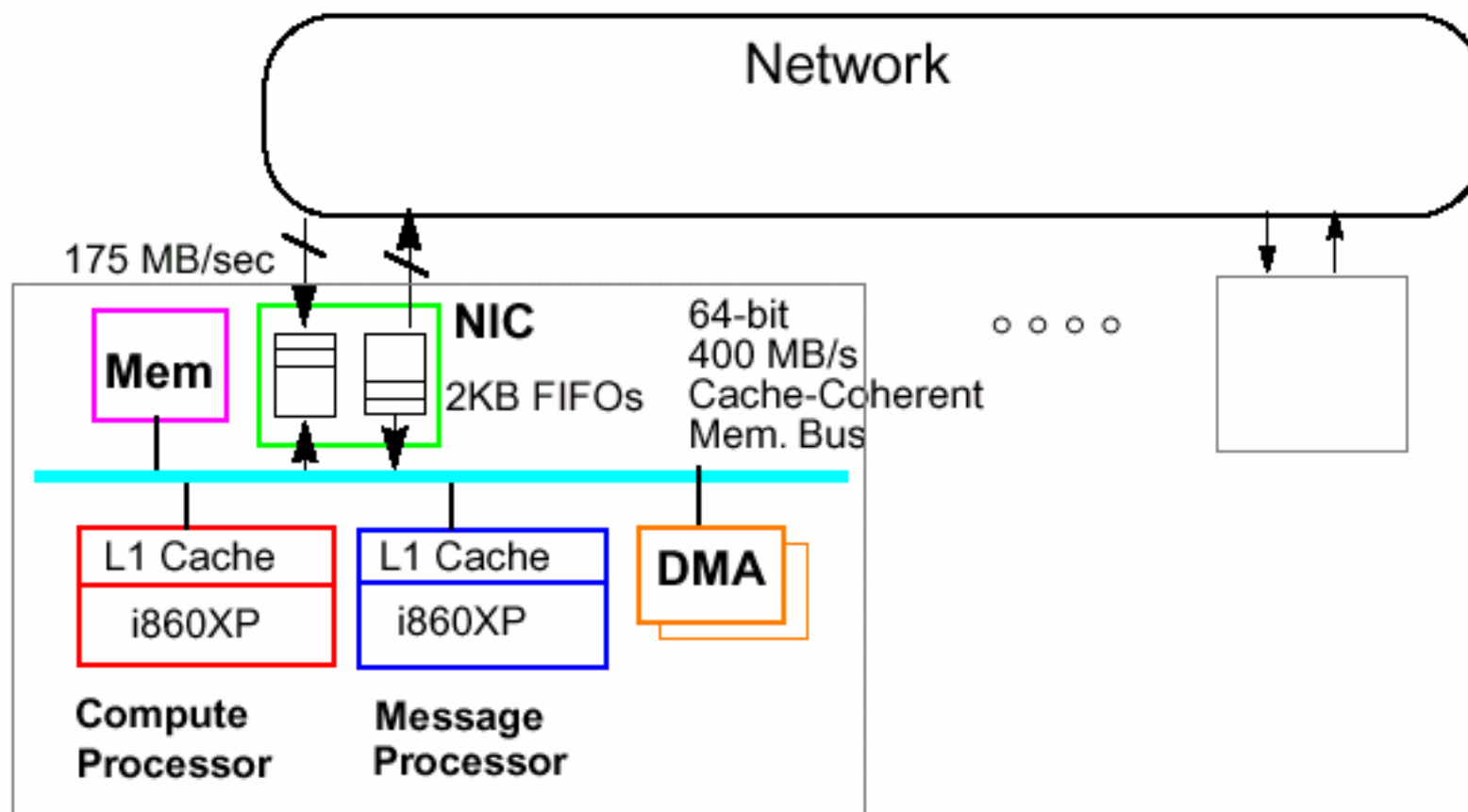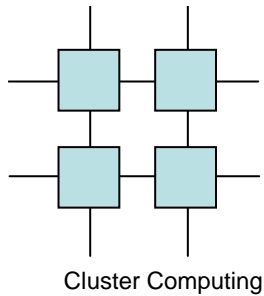
SP2

- Based in RS/6000 nodes
  - POWER2 processors
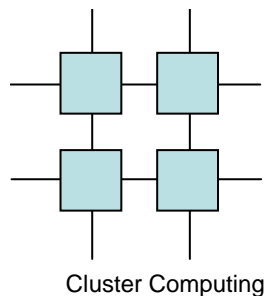- Special NIC: MSMU on the micro-channel bus
- Standard Ethernet on the micro-channel bus
- MSMUs interconnected via a HPS backplane

Cluster Computing

SP2 MSMU

Cluster Computing

SP2 HPS

- Links are 8 bit parallel
- Contention free latency is 5 ns per stage
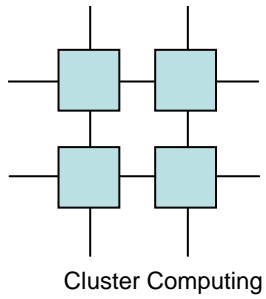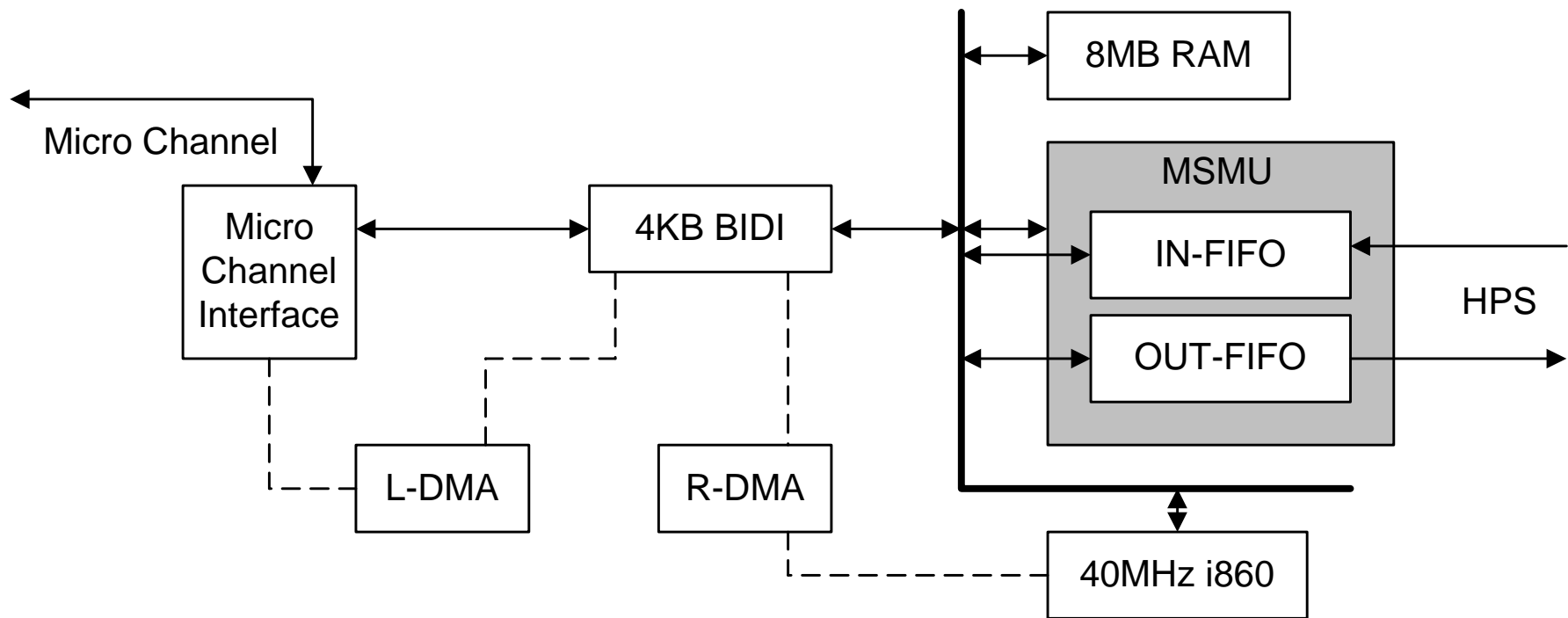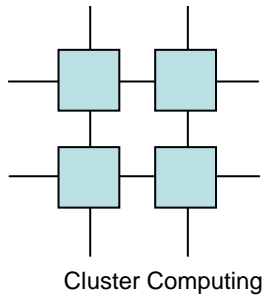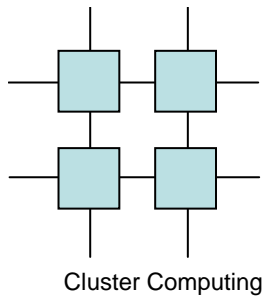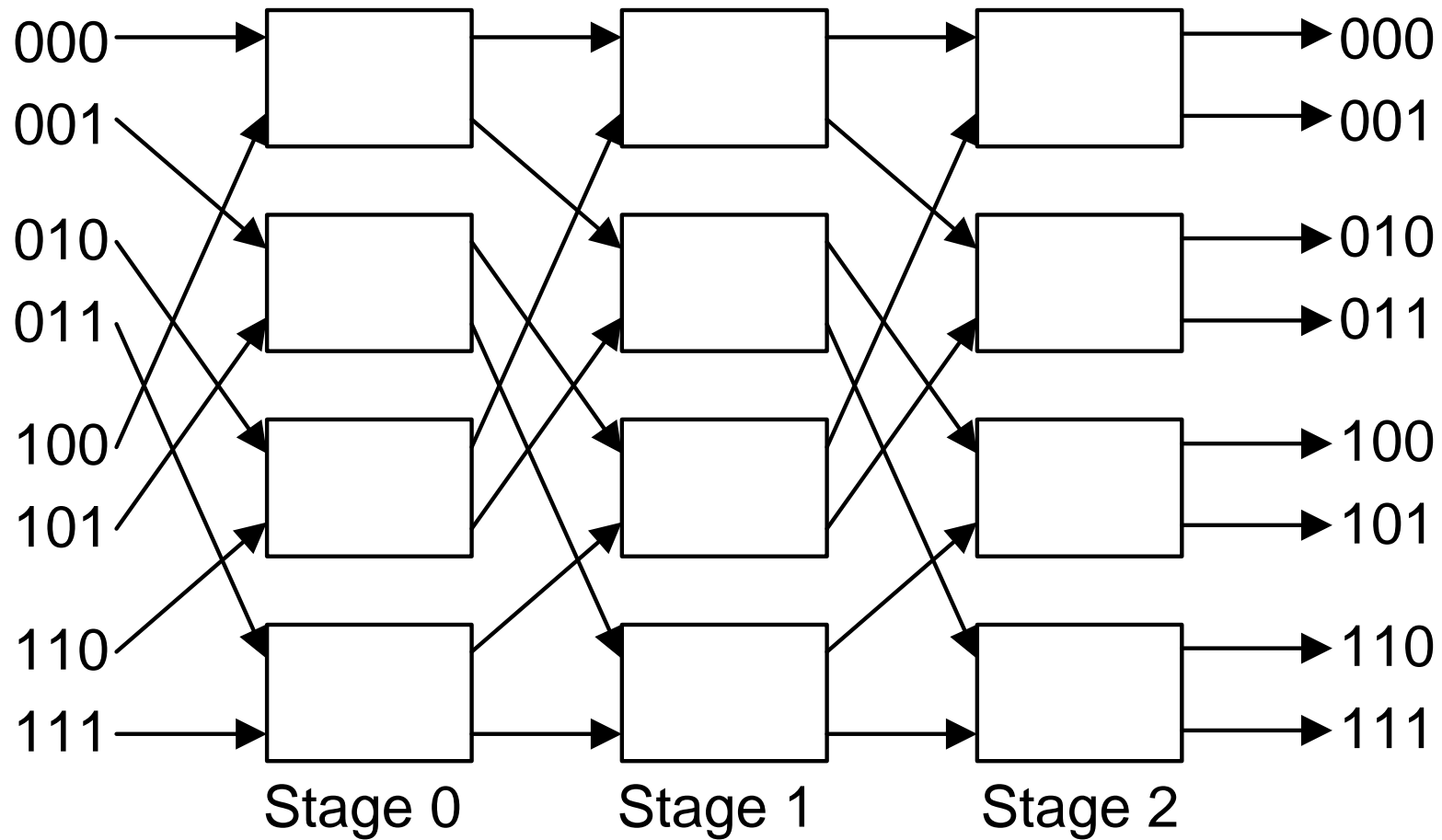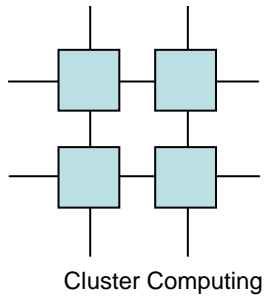 - 875 ns latency for 512 nodes

SP2 HPS

Cluster Computing

000   000
001   001
010   010
011   011
100   100
101   101
110   110
111   111
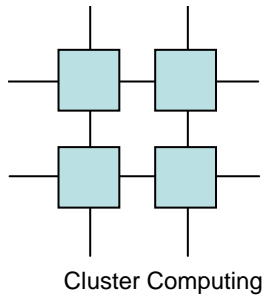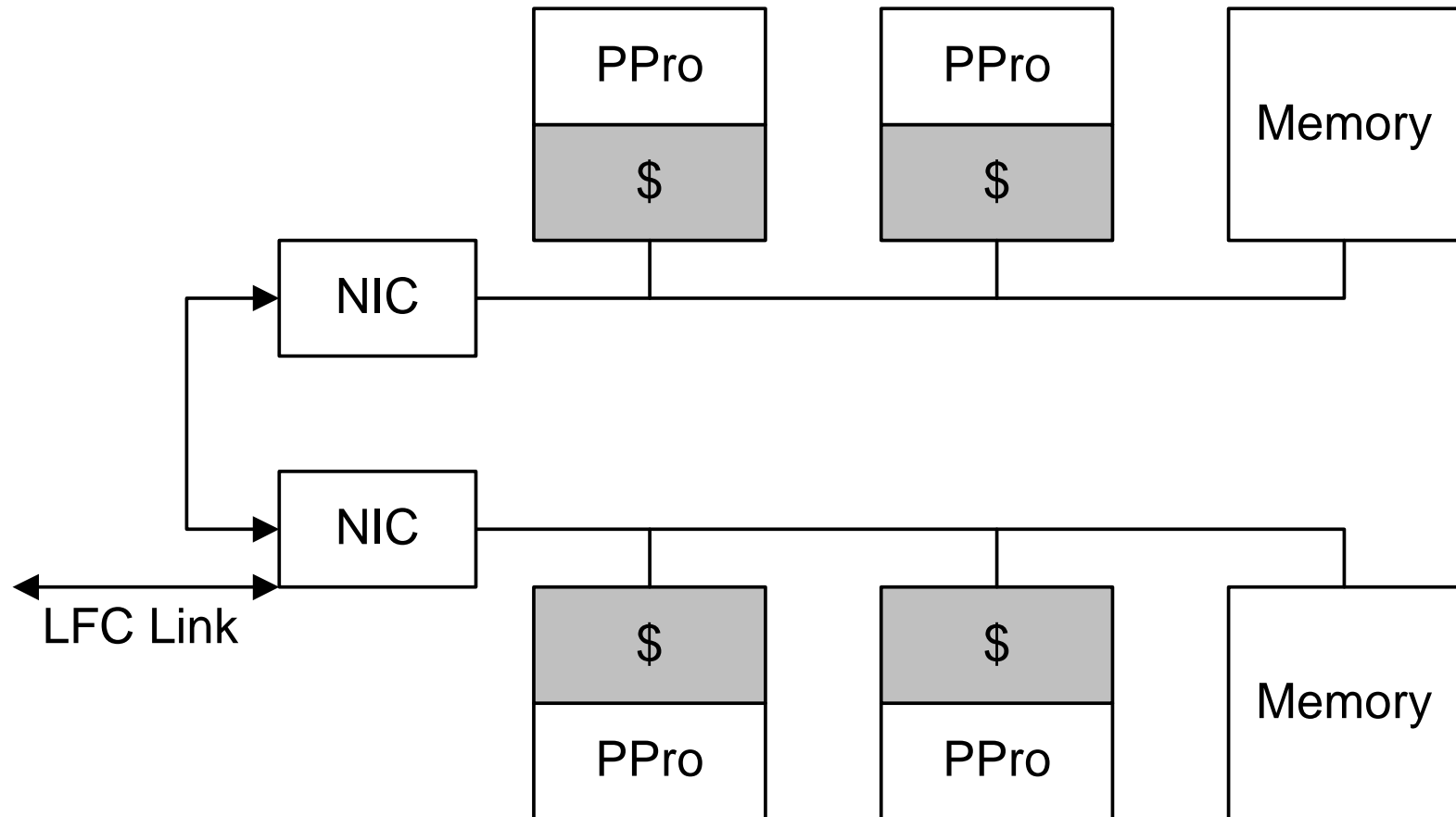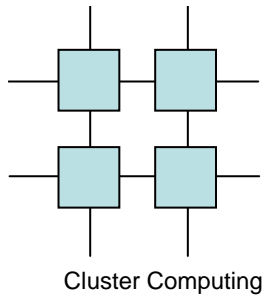
Stage 0   Stage 1   Stage 2

ASCI Red

Cluster Computing

- Build by Intel for the Department of "Energy"
- Consist of almost 5000 dual PPro boards with a special adaptation for user-level message-passing
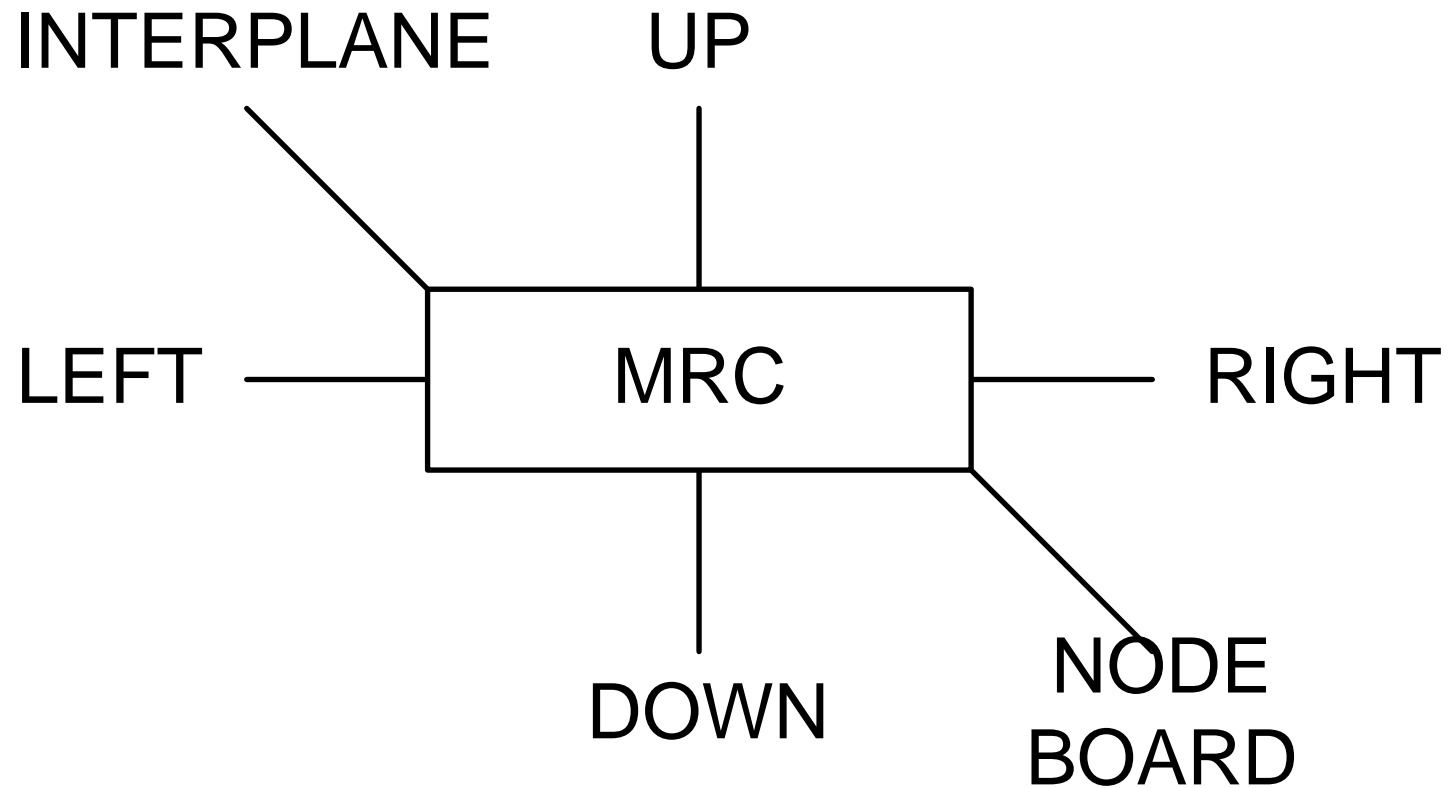- Special support for internal 'firewalls'

ASCI Red Node

ASCI Red MRC

Cluster Computing

INTERPLANE   UP

LEFT —— MRC —— RIGHT

DOWN   NODE BOARD
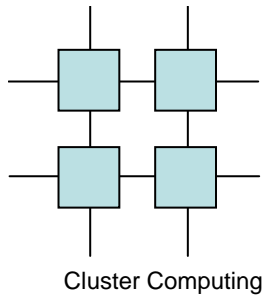
ASCI Red Grid

Scali

Cluster Computing
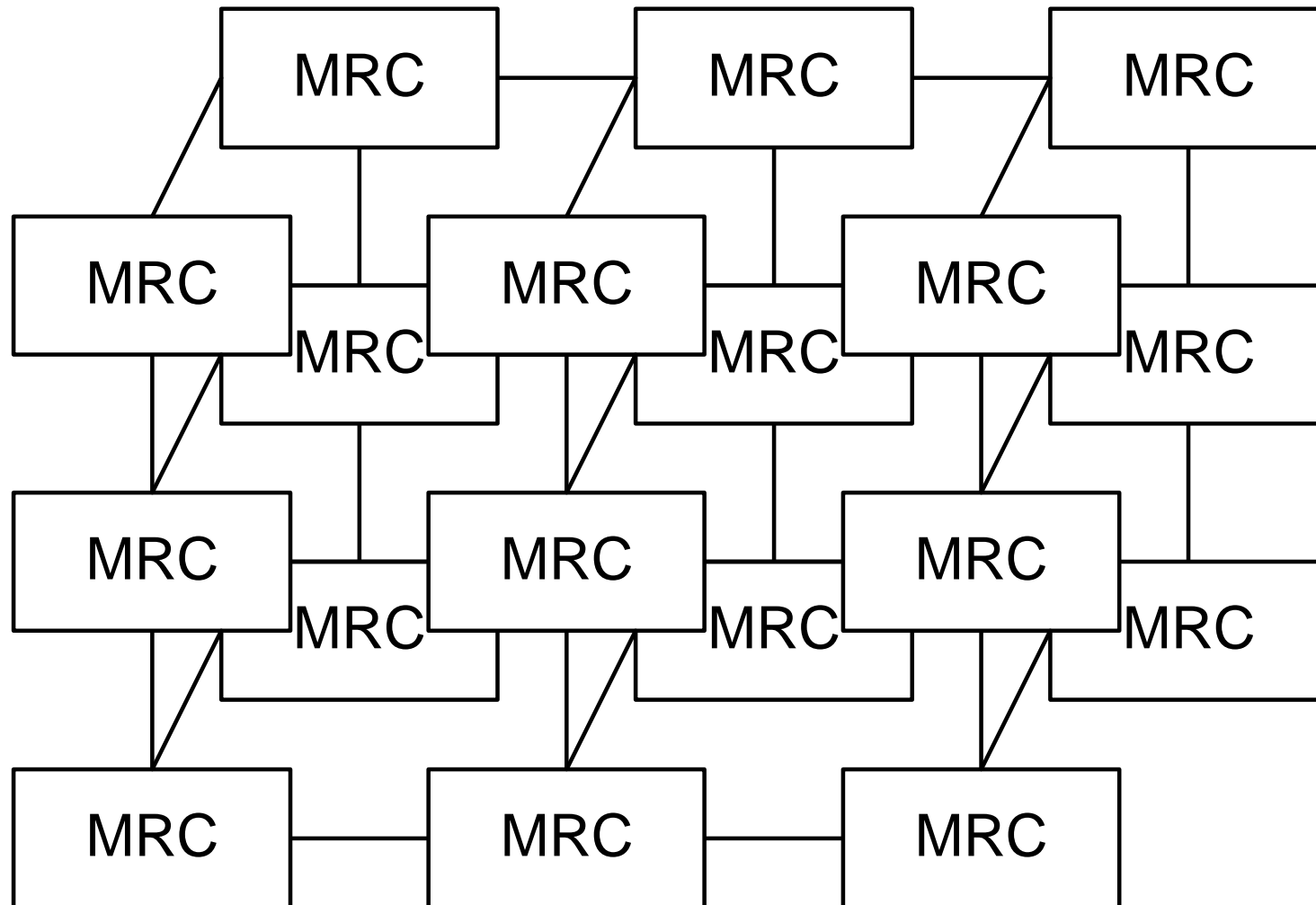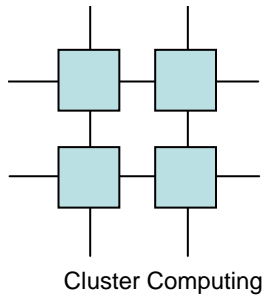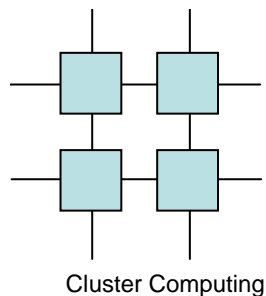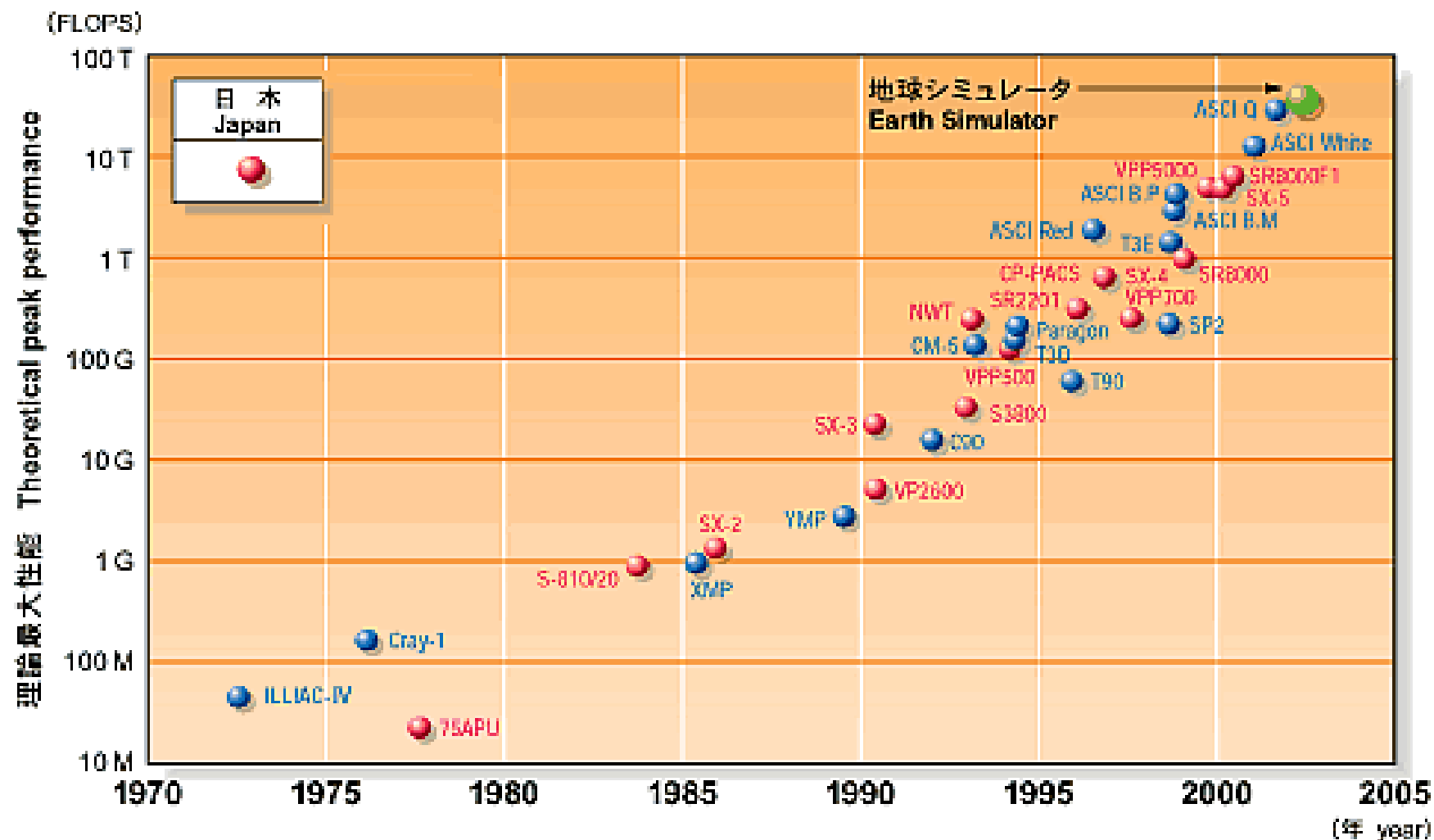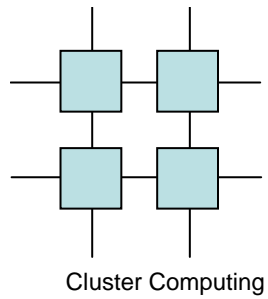
- Based on Intel or Sparc based nodes
- Nodes are connected by a Dolphin SCI interface, using a grid of rings
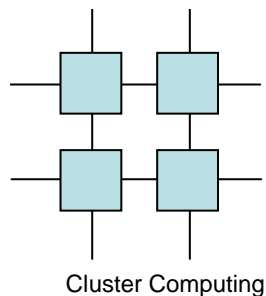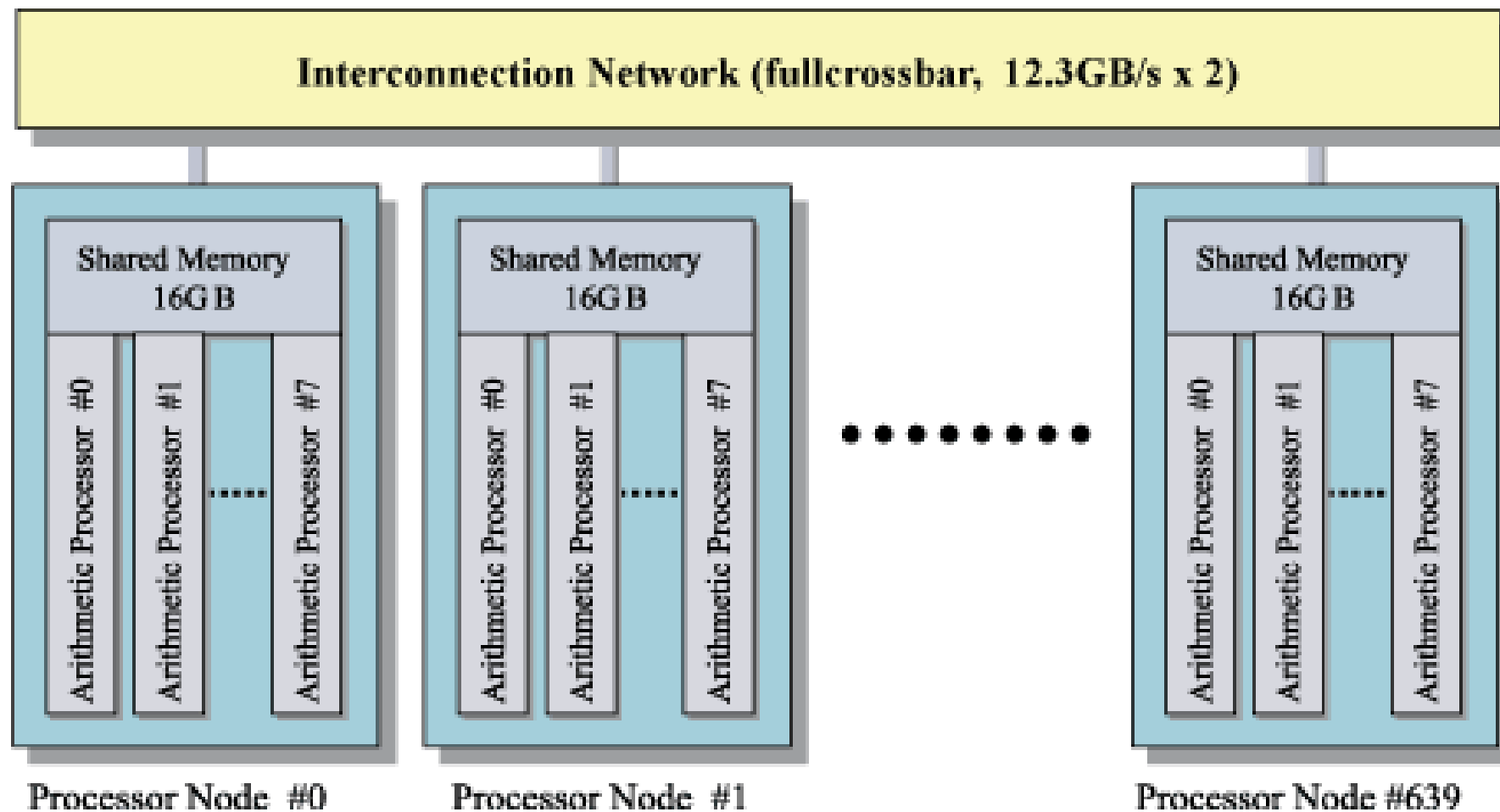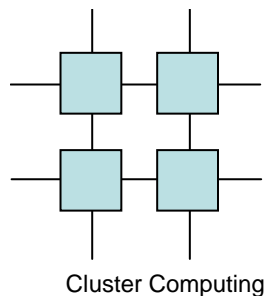- Very high performance MPI and support for commodity operating systems
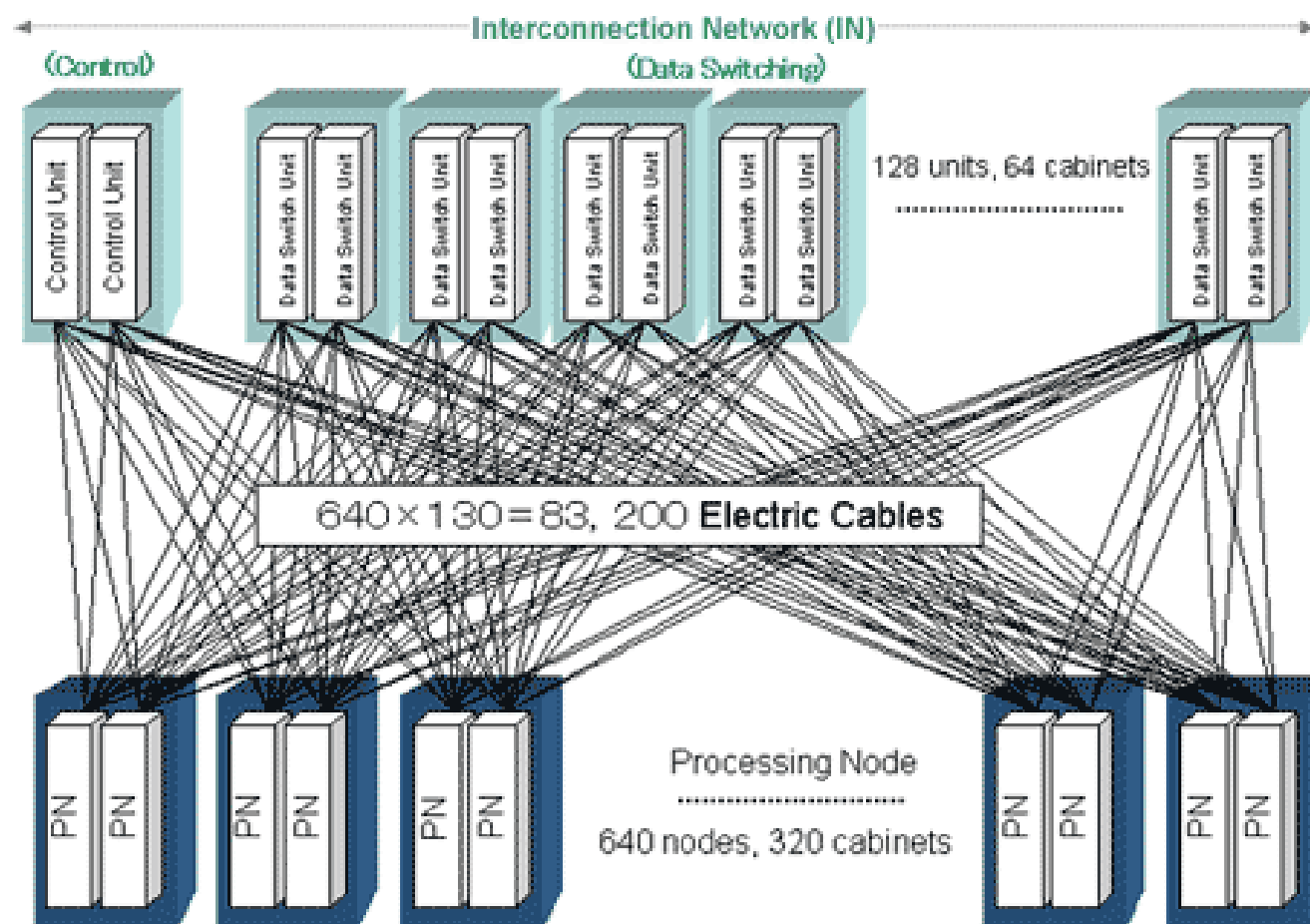
Performance???

Earth Simulator

New Earth Simulator Facilities

Power plant

Building for computer system

Building for operation and research

ES

Cluster Computing

Interconnection Network (fullcrossbar, 12.3GB/s x 2)

Processor Node #0	Processor Node #1	· · · · · · · · ·	Processor Node #639
Shared Memory 16G B	Shared Memory 16G B		Shared Memory 16G B
Arithmetic Processor #0 — Arithmetic Processor #1 · · · · · Arithmetic Processor #7	Arithmetic Processor #0 — Arithmetic Processor #1 · · · · · Arithmetic Processor #7		Arithmetic Processor #0 — Arithmetic Processor #1 · · · · · Arithmetic Processor #7

ES

Interconnection Network

Interconnection Network (IN)

(Control)   (Data Switching)

128 units, 64 cabinets

$640 \times 130 = 83,200$ **Electric Cables**

Processing Node

640 nodes, 320 cabinets

ES

Cluster Computing

ES

Cluster Computing