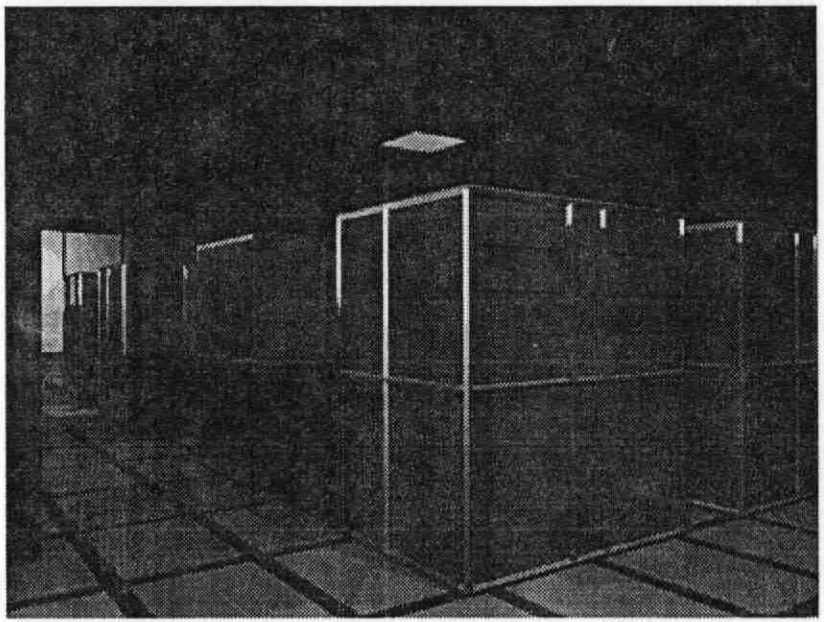
Vision of a Teraflop



Meiko Computing Surface 2.0

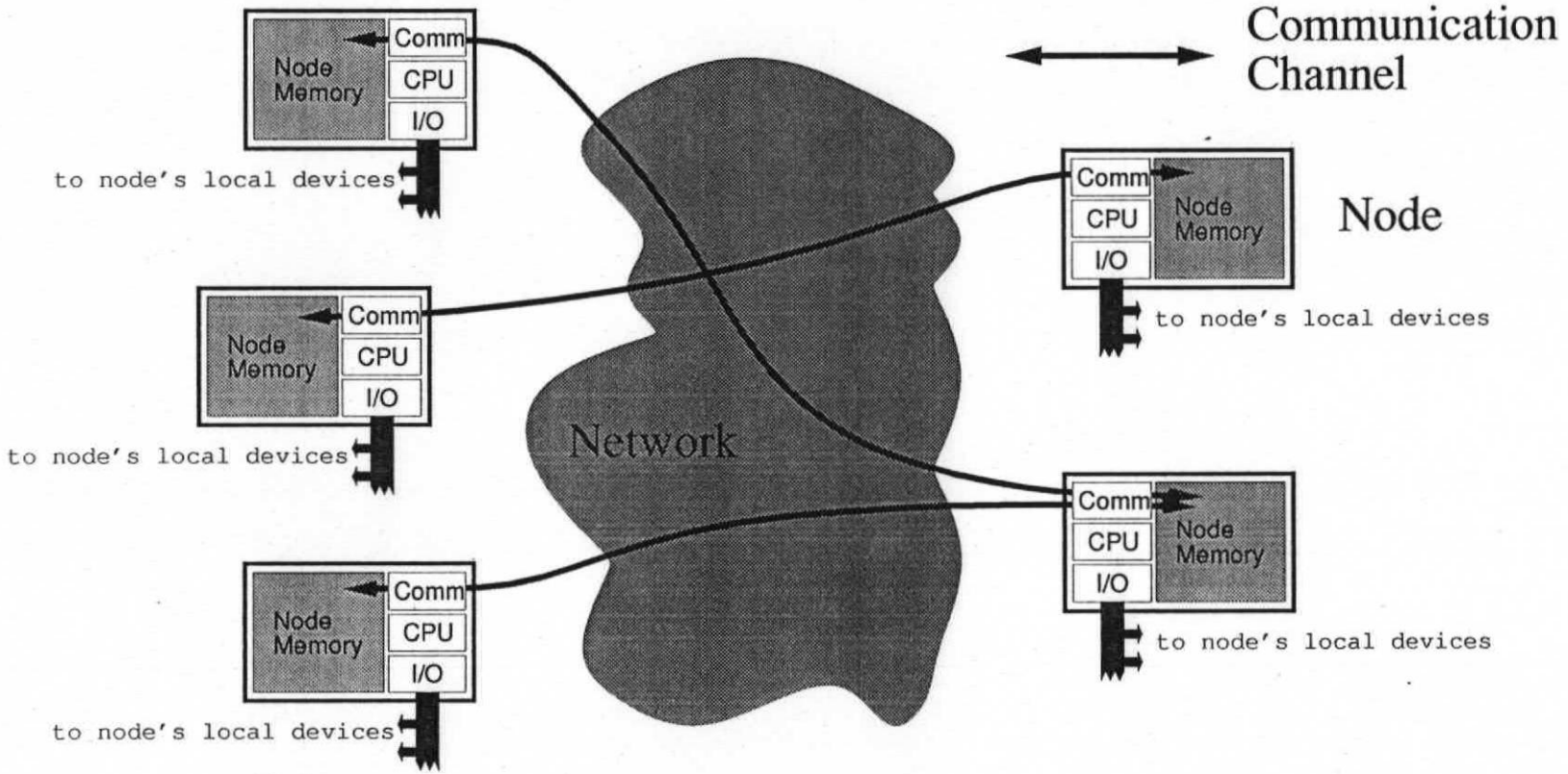
MPP vs Traditional Supercomputer

- At Least 10× Cost/Performance

Traditional Vector Now	$2500 / MFLOP
MPP Now	$1000 / MFLOP
CS 2.0 Now	$250 / MFLOP

- Outperforms Best Vector Supercomputer

Traditional Vector Now	16 D.P. GFLOPS
CS 2.0 Now	200 D.P. GFLOPS
CS 2.1 (1993)	500 D.P. GFLOPS



A Typical MPP System

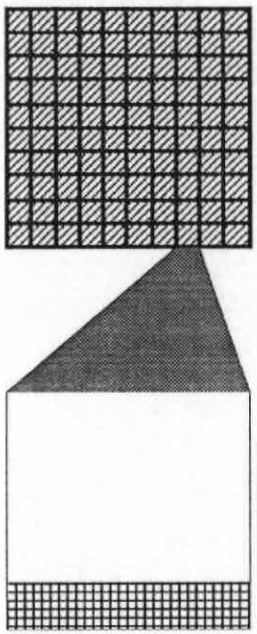
What is the CS 2.0?

- Significant Step to Practical Teraflop
- Distributed Memory MIMD
  massively parallel computer
- Scalable Architecture
  - compute/memory bandwidth
  - communication bandwidth
  - network bandwidth
  - low cost I/O
- Fault Tolerant Architecture

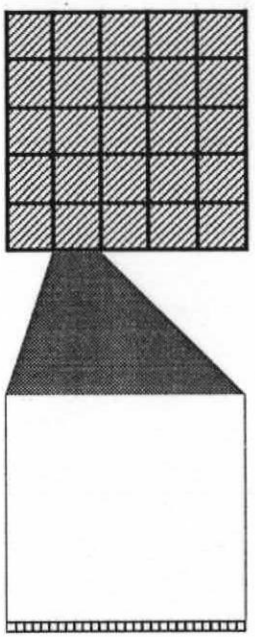
Performance of First Release Machine

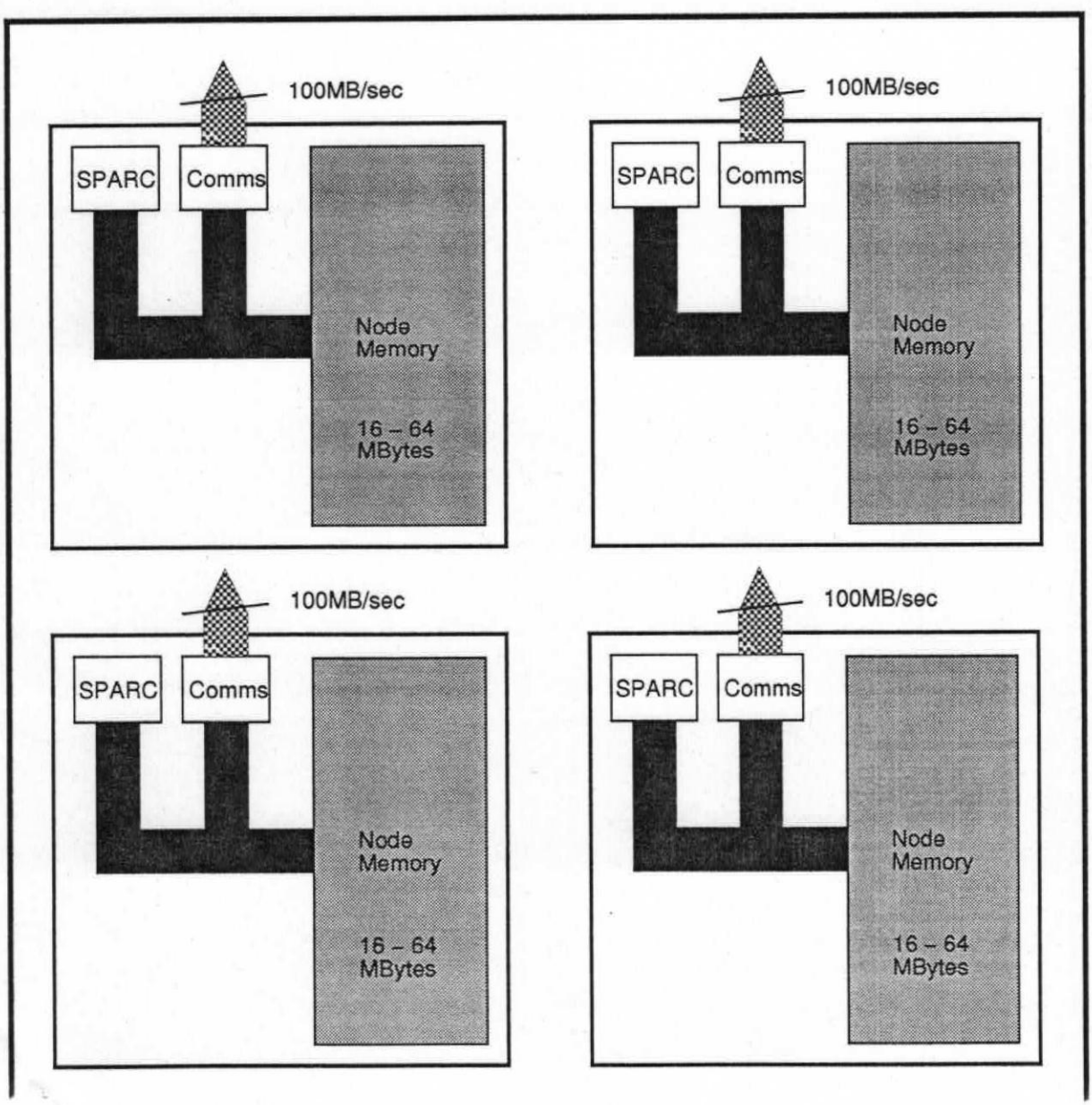
- Scalable to 1024 Nodes (200 d.p. GFLOPS)
- Node
  - 400 MFLOPS (single precision)
  - 200 MFLOPS (double precision)
  - 100 MIPS
- Comms Channel
  - 100 MBytes/sec bidirectional
- Bisectional Network Bandwidth
  - 32 nodes 3.2 GBytes/sec
  - 256 nodes 25.6 GBytes/sec
  - 1024 nodes 102.4 GBytes/sec

More Power is Better

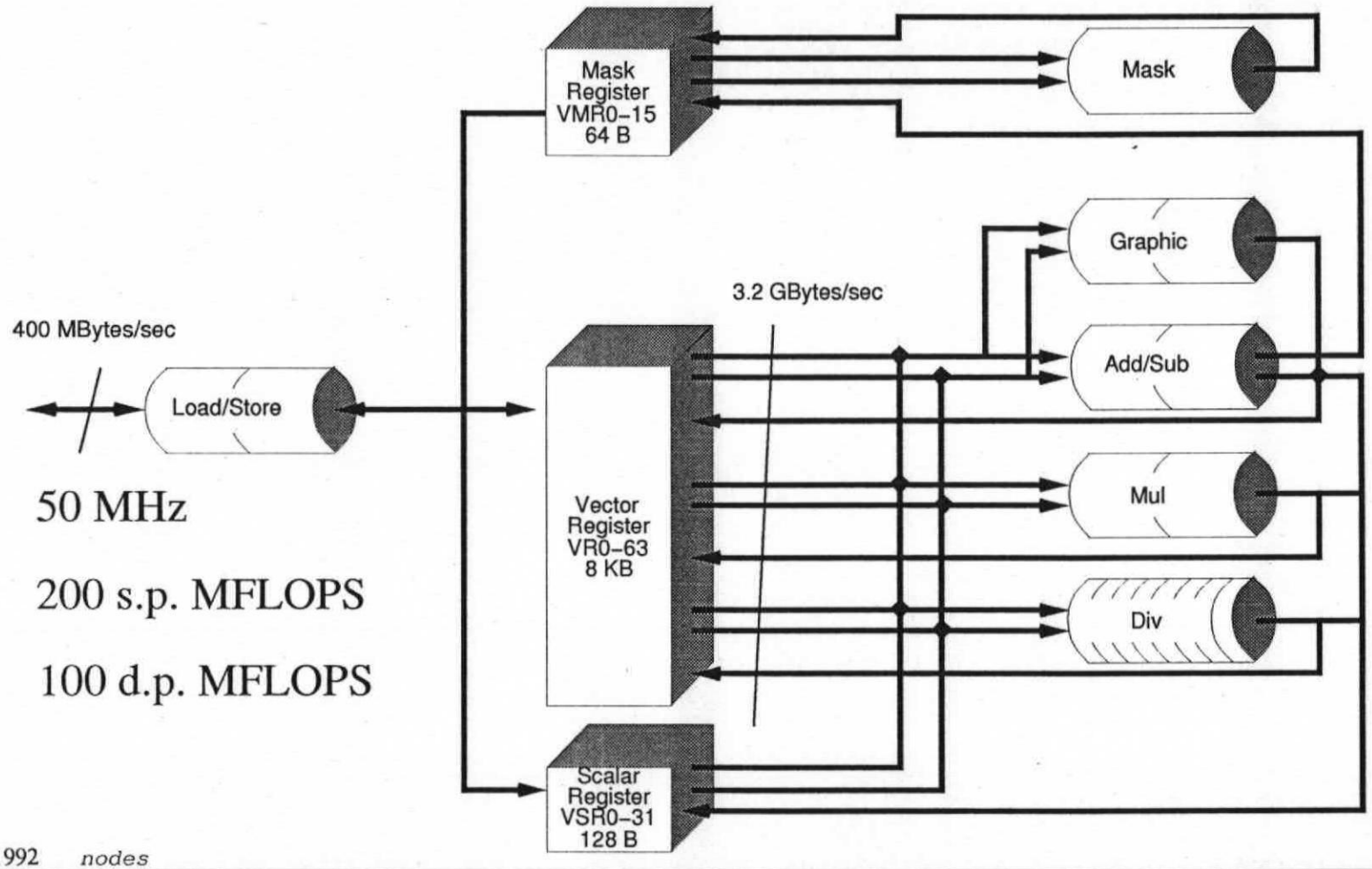


1	Compute (∝ area)	4
4	Comms (∝ perimeter)	8
1 : 4	Compute : Comms	1 : 2
	Memory Usage	
80%	Data	95%
20%	Program	5%




100MB/sec
SPARC
Comms
Node Memory
16 – 64 MBytes
100MB/sec
SPARC
Comms
Node Memory
16 – 64 MBytes
100MB/sec
SPARC
Comms
Node Memory
16 – 64 MBytes
100MB/sec
SPARC
Comms
Node Memory
16 – 64 MBytes

Internal Block Diagram of VPU



50 MHz

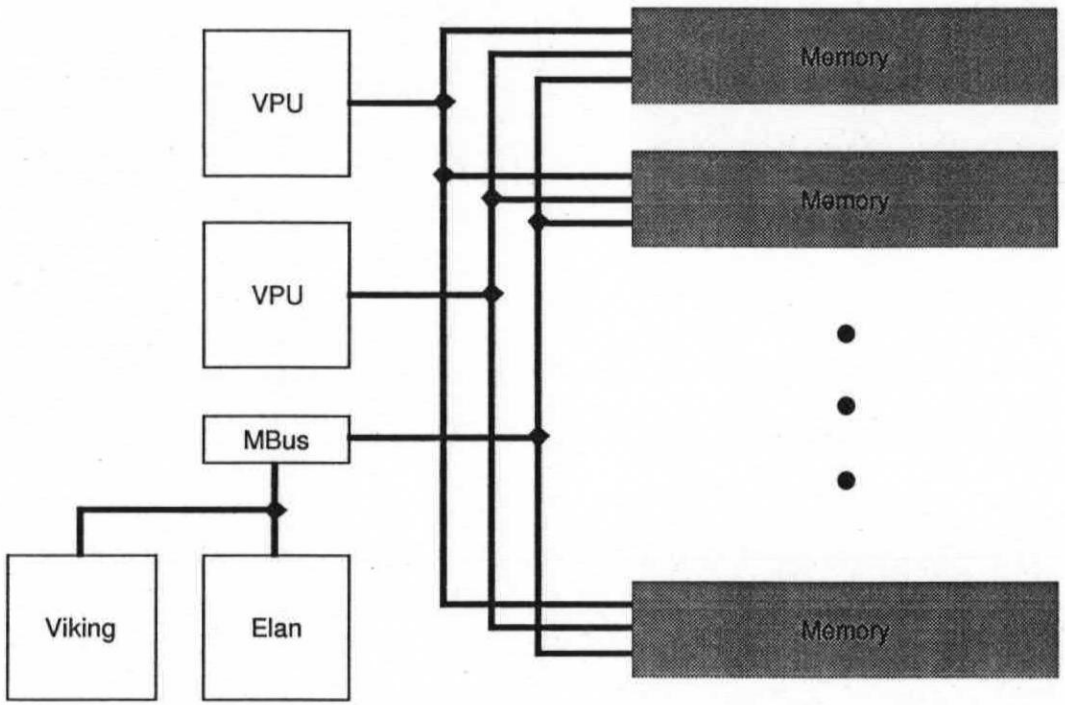
200 s.p. MFLOPS

100 d.p. MFLOPS

Viking Superscalar SPARC Performance

- 50 MHz clock
- 150 MIPS
- 50 MFLOPS
- 80 SPECmarks
- 20 KByte instruction cache
- 16 KByte data cache
- 1 MByte second level cache

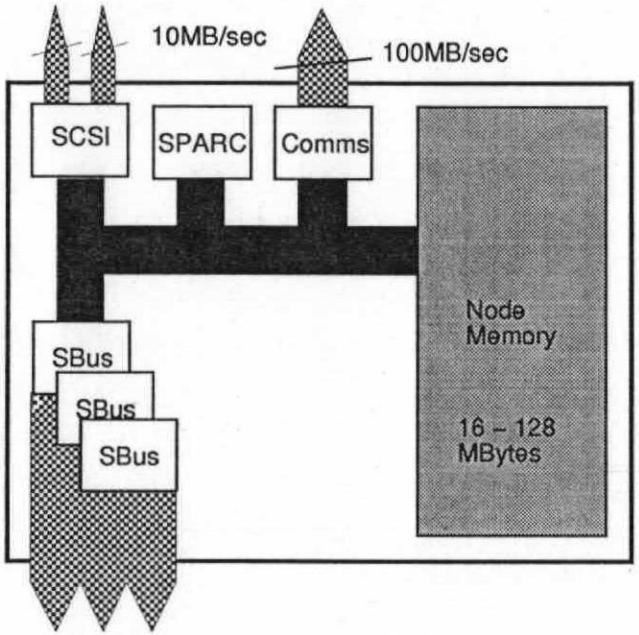
MK403 (Memory System Design)



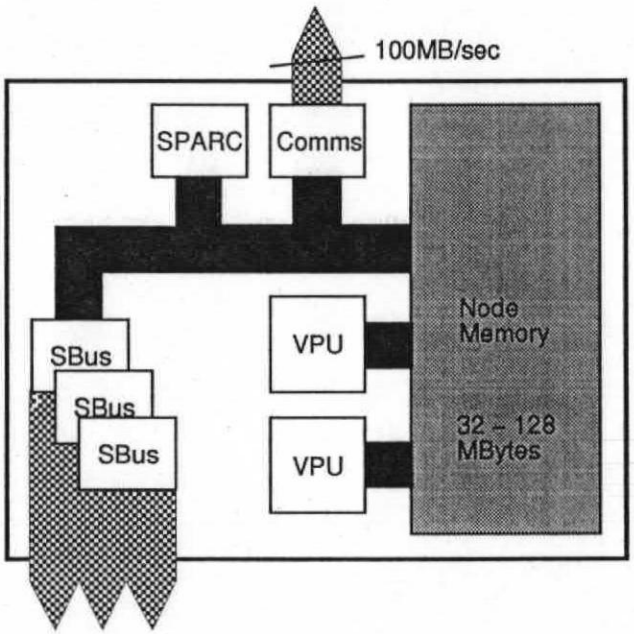
- 16 independent banks
- 32 to 128 MBytes of EDC DRAM
- Aggregate memory bandwidth 3.2 GB/sec
- Aggregate port bandwidth 1.2 GB/sec

Initial CS 2.0 Nodes

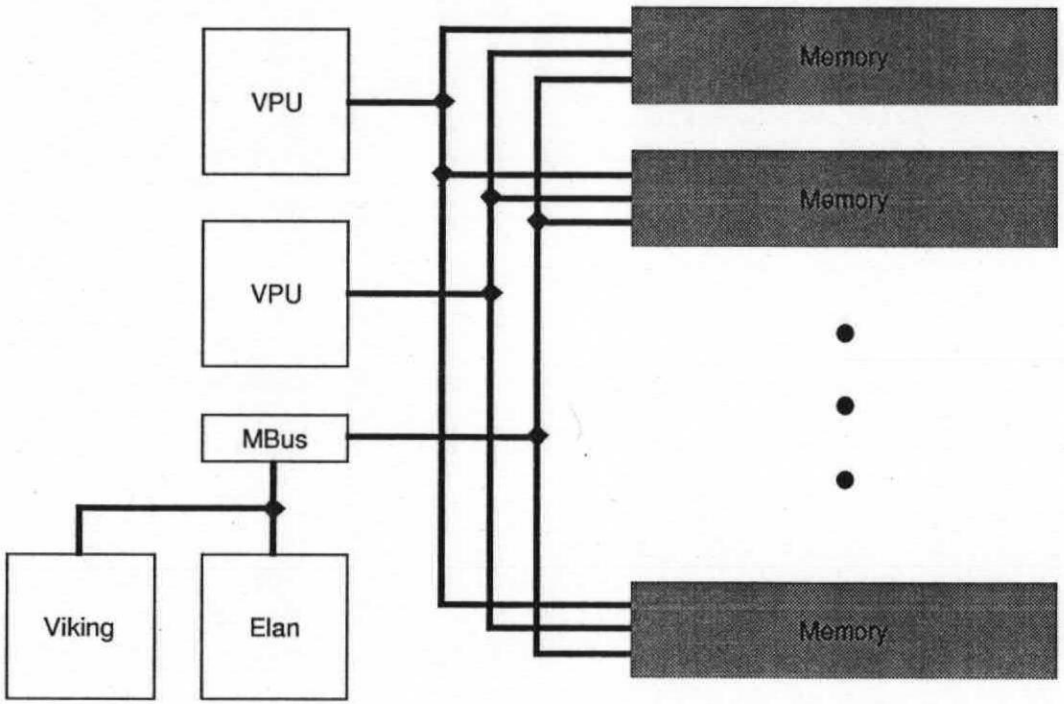
MK401



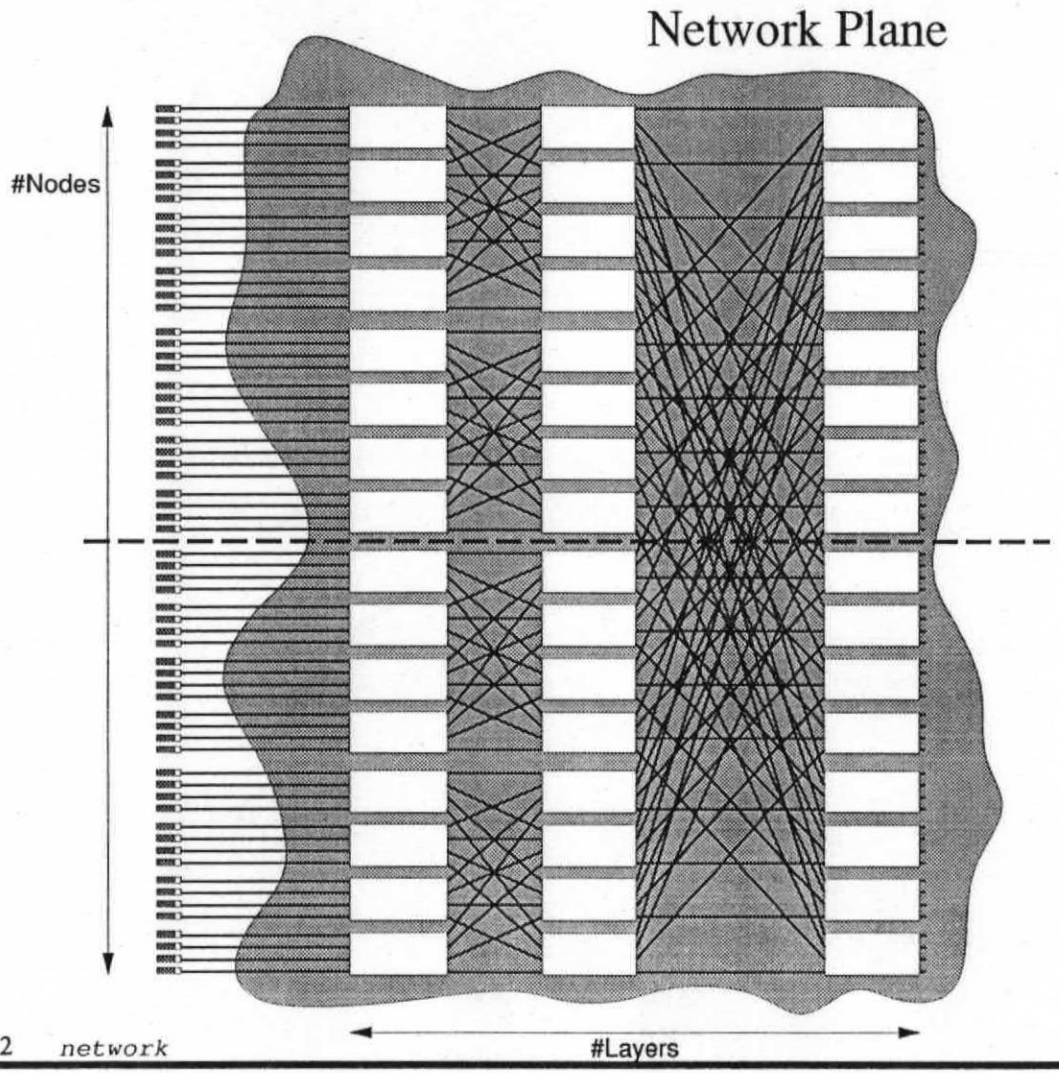
MK403



MK403 (Memory System Design)



- 16 independent banks
- 32 to 128 MBytes of EDC DRAM
- Aggregate memory bandwidth 3.2 GB/sec
- Aggregate port bandwidth 1.2 GB/sec



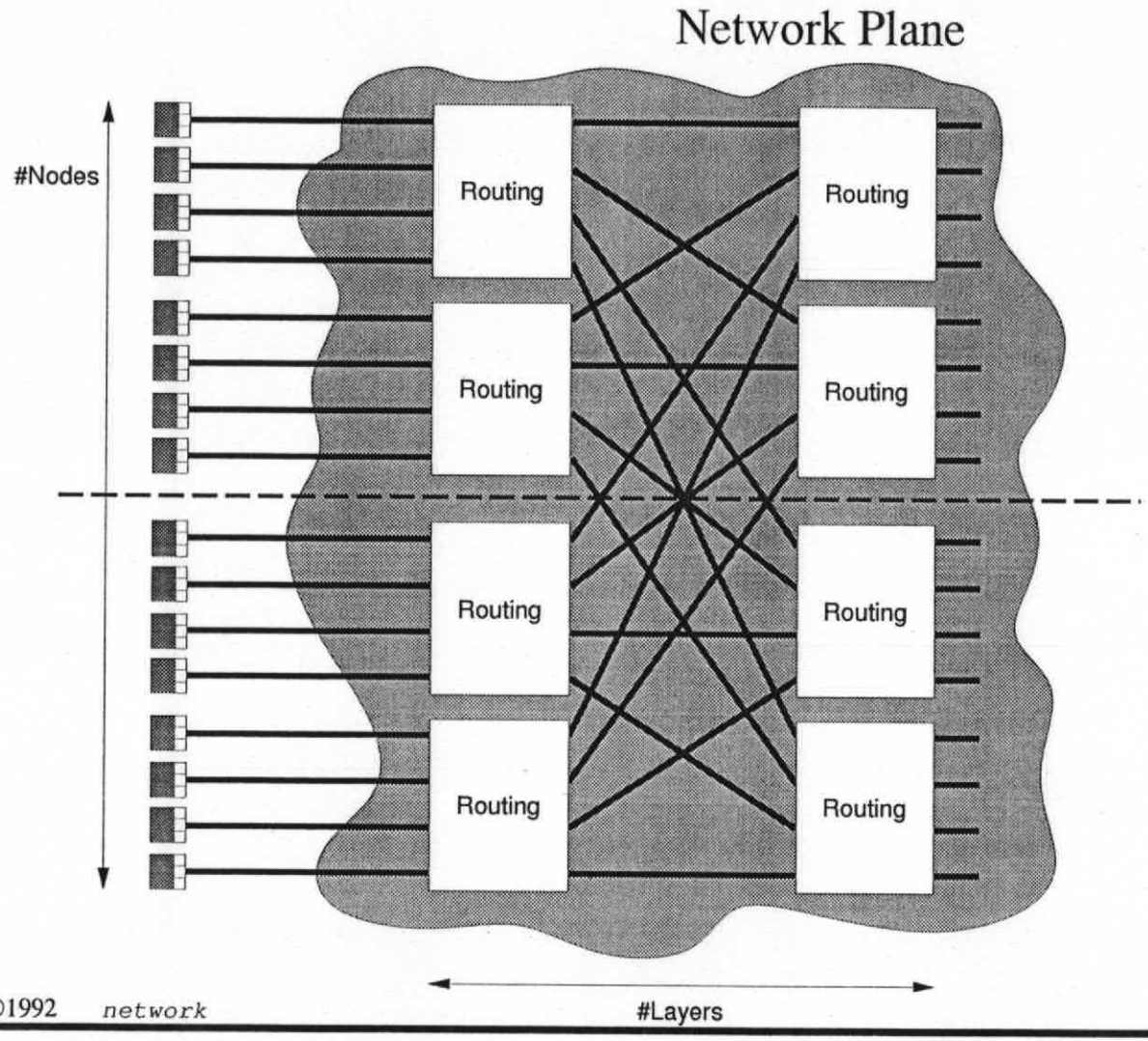
- 64 processors
- 3 switch layers

∴ Bisectional bandwidth

= ½ × 64 × 100

= 3.2 GB/sec

meiko

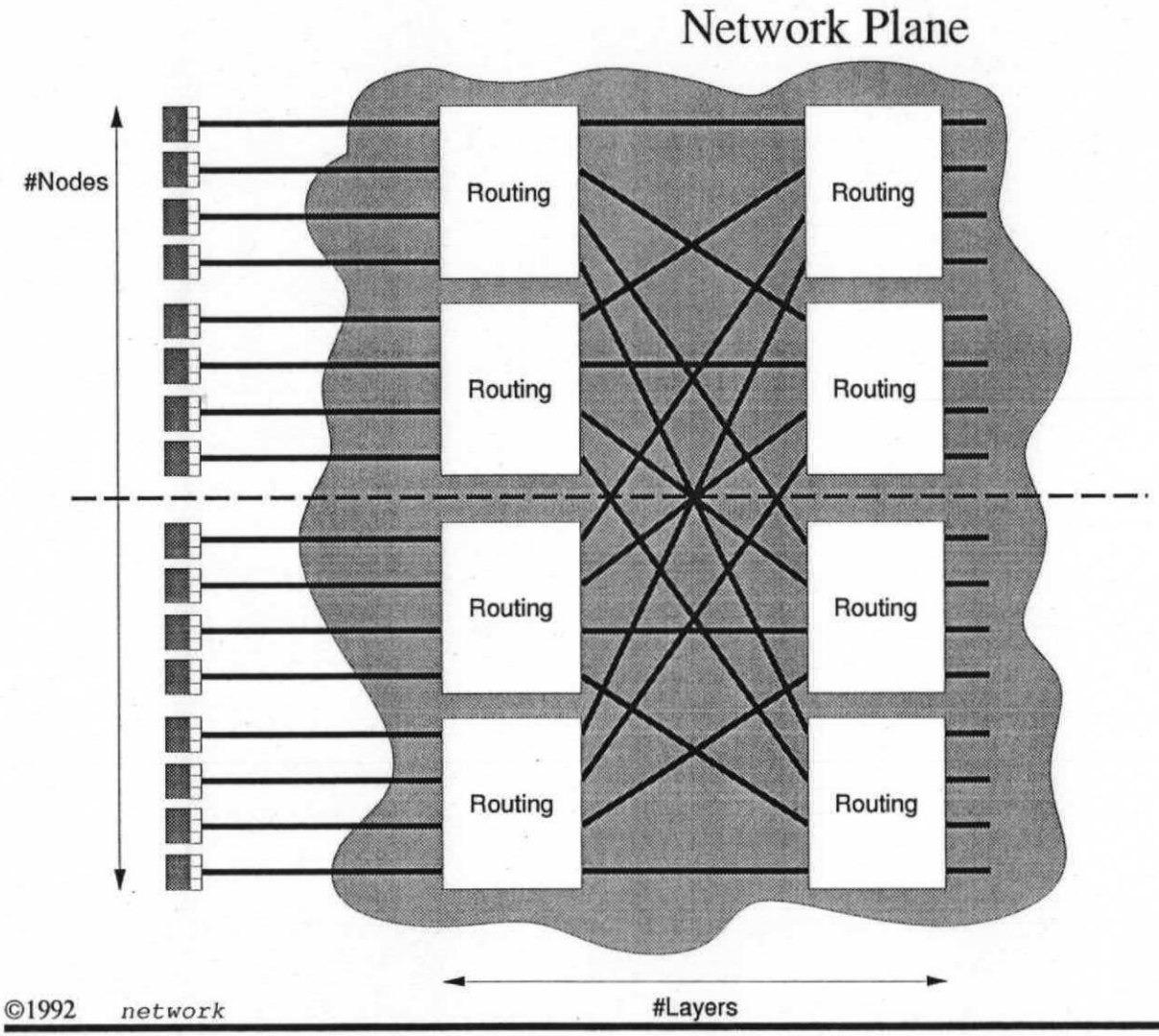


- 16 processors
- 2 switch layers

∴ Bisectional bandwidth

= ½ × 16 × 100

= 800 MB/sec



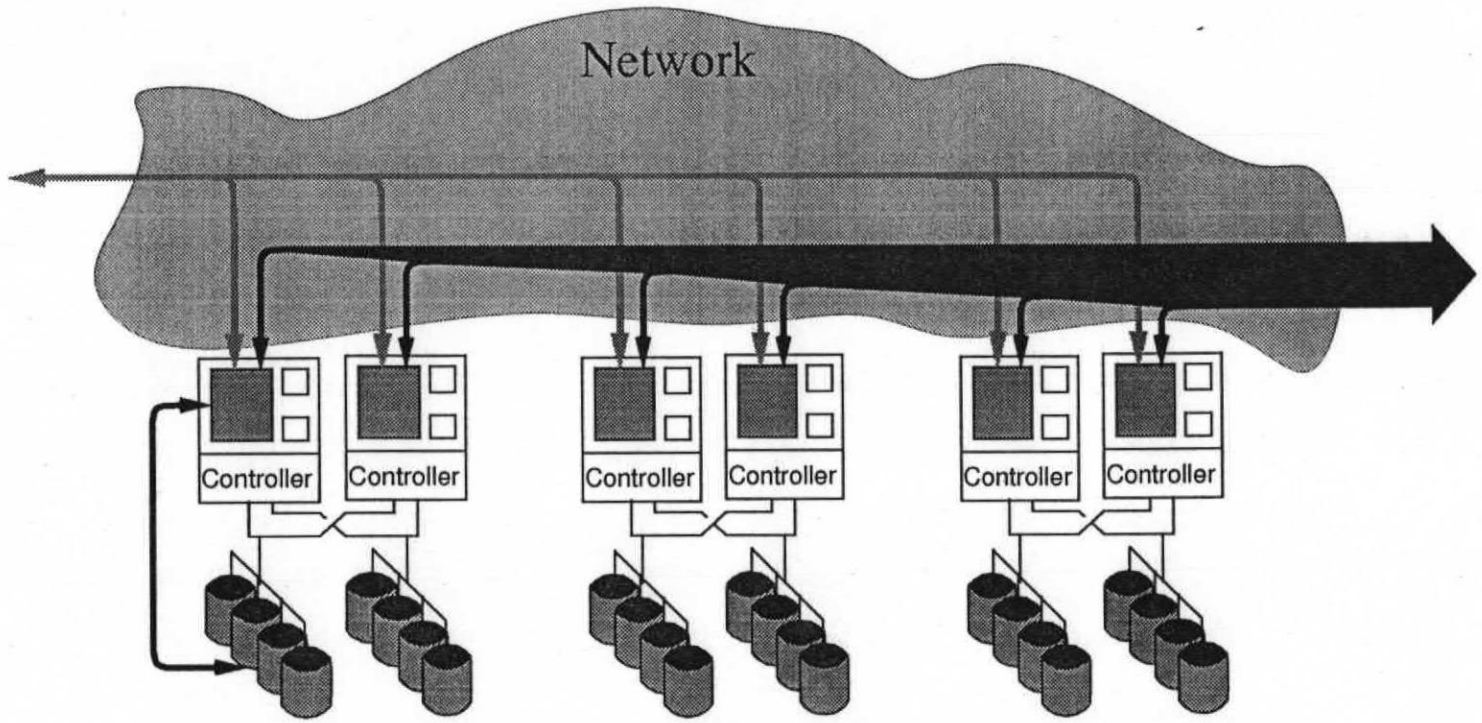
- 16 processors
- 2 switch layers

∴ Bisectional bandwidth

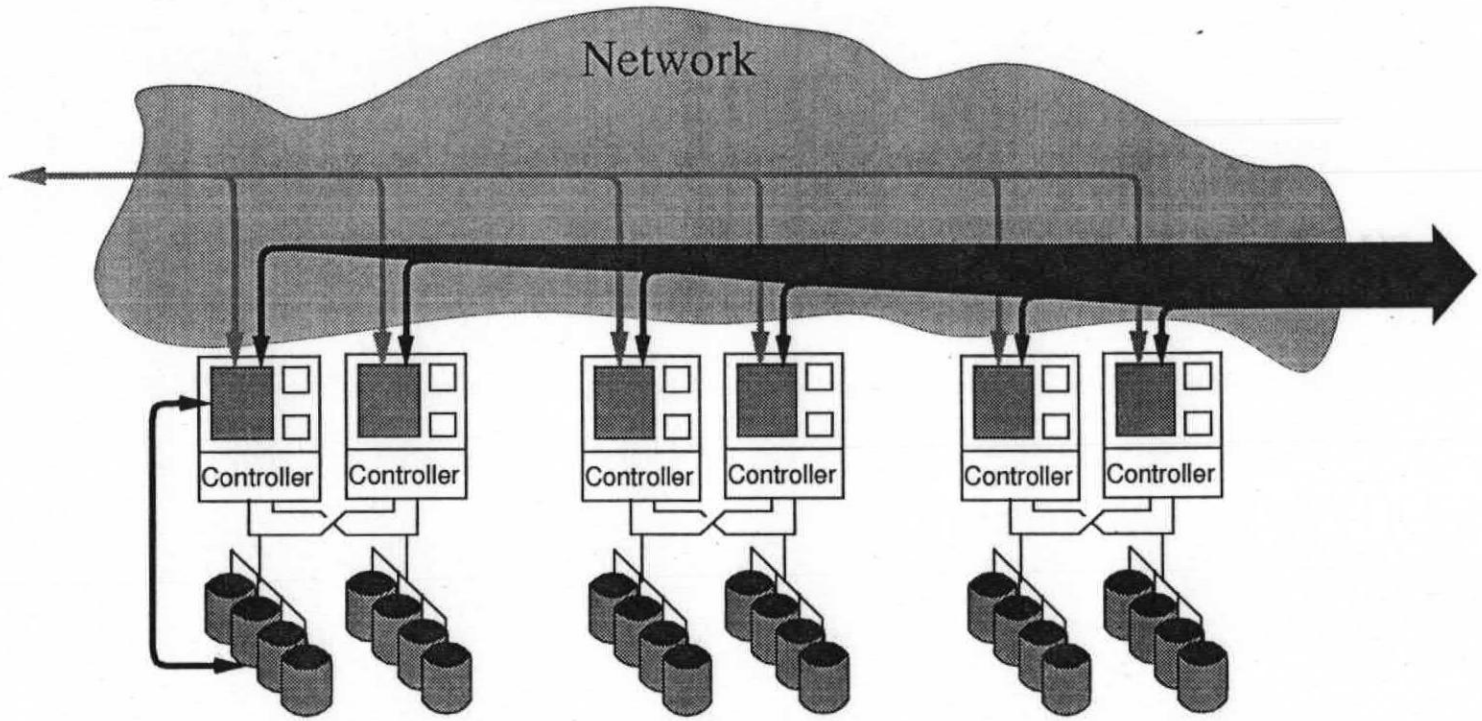
= ½ × 16 × 100

= 800 MB/sec

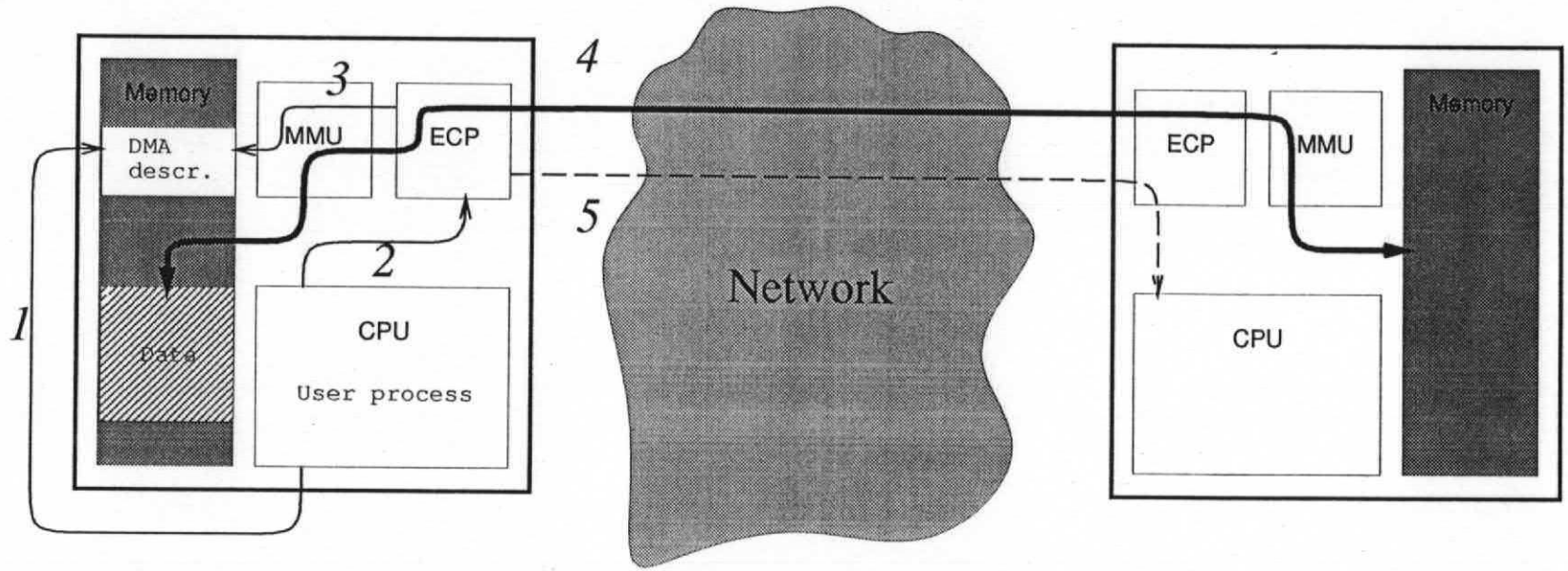
Parallel Disc Architecture



Parallel Disc Architecture



Elan Operation



meiko

Fault Tolerance

All memory has error detection and correction

All communications are CRC error checked, errors corrected by re-transmission or re-route

Network supports multiple routes

"n+1" redundancy with "hot spares" on line

Live removal and insertion of independently powered and cooled modules

Application Software

Uni-processor Math Library

- BLAS 1, 2, 3
- FFT
- NAG, IMSL, SAS

Parallel Math Library

- Dense Matrix solvers (both in and out of core)
- Interactive Sparse Matrix solvers
- 2 and 3 Dimensional FFT's

Explicit Parallel Libraries

CS-Tools, common to all generations of Computing Surface

- parallel application generator
- communications libraries
- performance analysis tools
- parallel debugger

Portable libraries

- PARMACS (Argonne Lab.)
- PVM (Oak Ridge Lab.)
- other vendor specific libraries

meiko

Compiler Technology

State of art optimizing C and FORTRAN for SPARC nodes

Vectorizing FORTRAN and C for Vector Nodes

High Performance Fortran for Vector and Scalar nodes

Software Technology

Every Node runs Solaris from SunSoft
- conforms to SPARC ABI
- conforms to SVR4, X/Open, POSIX

Multi-user, Multi-domain operating modes

Support for multiple parallel programming paradigms
- distributed multi-processing
- explicit parallel programming
- data parallel programming

Machine I/O

File I/O
- distributed parallel filesystem
- RAID sub-systems

Network I/O (scalable)
- Ethernet
- FDDI
- HIPPI

Bus Interfaces
- SBus
- VME (by bridge)

Elite Network Switch

Proprietary ASIC, 8 x 8 crosspoint switch

Hardware Broadcast

Fair arbitration of routing contention

Route checking of transaction CRC

Multi-stage
- packet switched network with wormhole routing

Multi-plane
- network redundancy
- improved bi-sectional bandwidth

Network Considerations

1. Congestion caused by concurrent communications between processors

2. Scalable Bi-sectional bandwidth

3. Fault tolerance

Elan Communication Processor

Proprietary ASIC, RISC CPU core, uCoded DMA engine

Packetized network transactions
- hardware context checking
- deterministic or random network routing

Byte wide bi-directional links
- 1.26 Gbit/sec line rate
- 100 Mbyte/sec bi-directional data rate

User process latency
- < 10uS between any two nodes in 1024 node machine
- operating system protection in hardware

Communications Considerations

1. Time to move data from processor A to B
   - latency
   - bandwidth

2. Anywhere to Anywhere connectivity
   - supports arbitrary and variable communication topologies
   - hardware broadcast

3. Communications Models
   - message passing
   - shared memory

4. Fault tolerance

Processing Nodes

Heterogeneous Architecture
- rapidly track high performance commodity CPU's
- provides better fit to different classes of application
- avoids proprietary CPU R&D burden

Industry standard Operating System
- Solaris 2.0 from SunSoft
- provides ABI for sequential applications

Design at optimum price performance point

Key attributes of MPP

- Node performance
- Communications channels
- Internal Network bandwidth
- I/O bandwidth

As you increase the number of nodes the bandwidth to communicate must increase proportionately.