Distributed [*Computing*] Systems

Interconnection Networks, Communications and the OSI Model - Part I

Isaac D. Scherson (aka The Schark)

Dept. of Computer Science (Systems)
Bren School of Information and Computer Sciences
University of California, Irvine
Irvine, CA 92697-3425

isaac@ics.uci.edu
www.ics.uci.edu/~isaac www.ics.uci.edu/~schark

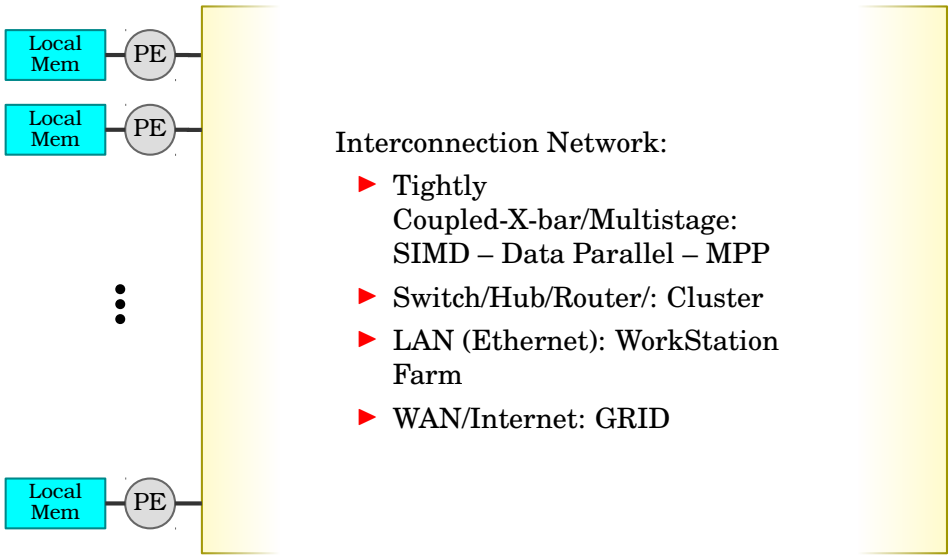
CompSci-230, Winter 2019

The Interconnection Network

Differences between Interconnection Networks

- Interconnection Network Speed (Latency and Bandwidth) differentiates between different concurrent computing architectures.
- Many network details have been oversimplified here... for simplicity's sake...

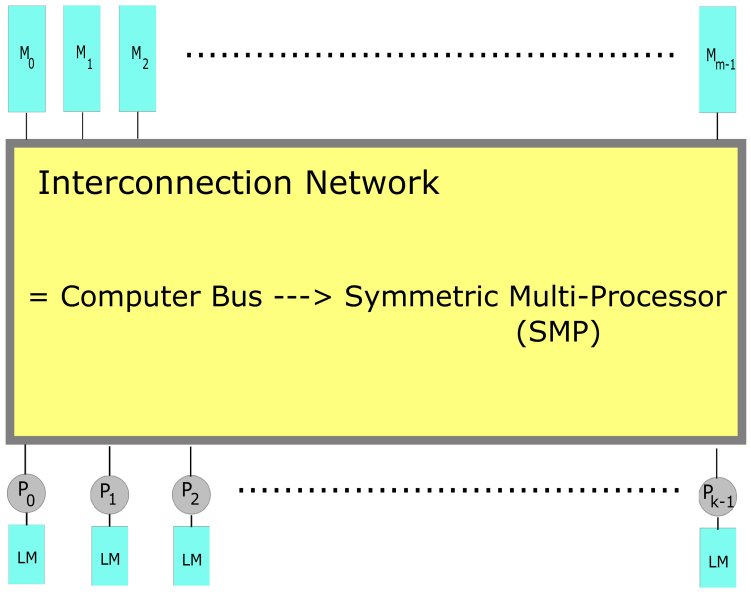
Differences between Interconnection Networks (cont"d)



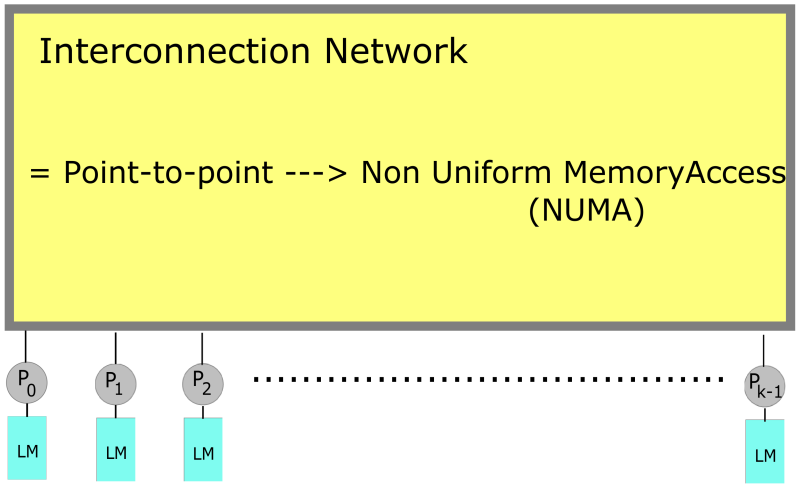
Interconnection Network:

- Tightly Coupled-X-bar/Multistage: SIMD – Data Parallel – MPP
- Switch/Hub/Router/: Cluster
- LAN (Ethernet): WorkStation Farm
- WAN/Internet: GRID

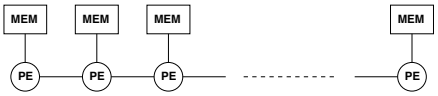
Network Characteristics → Parallel Processing Architectures



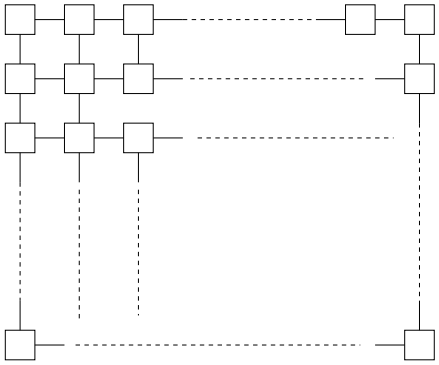
Network Characteristics → Parallel Processing Architectures



Two Examples of point-to-point Architectures



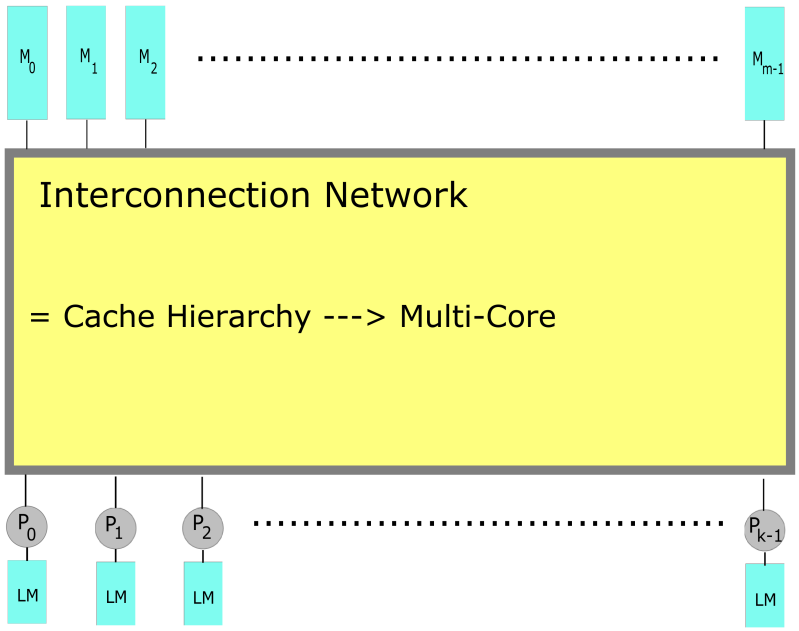
Linear Array



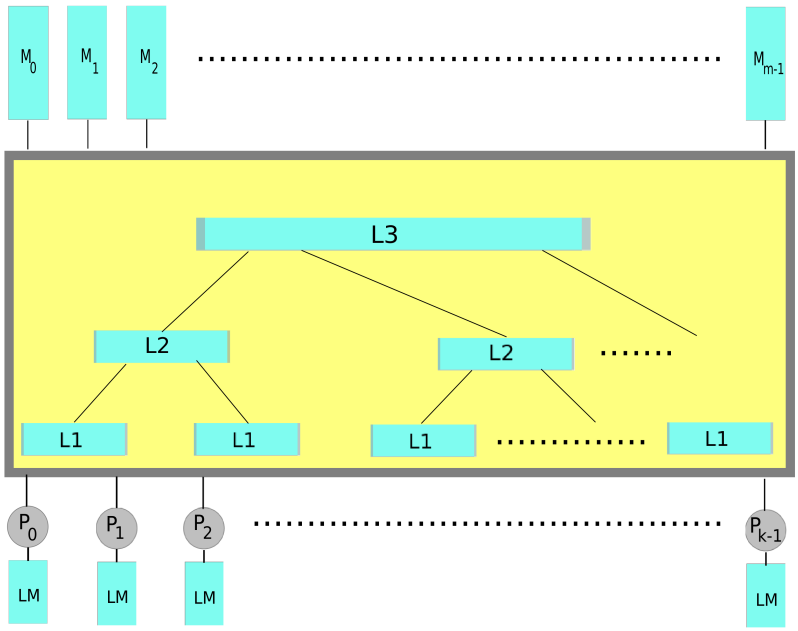
Two Dimensional Array
(Mesh-Connected Parallel Computer)

- These architectures have planar topologies, hence thought suitable for VLSI
- Tons of work on algorithms for linear and 2D systems
- However, a more general model won the popularity contest

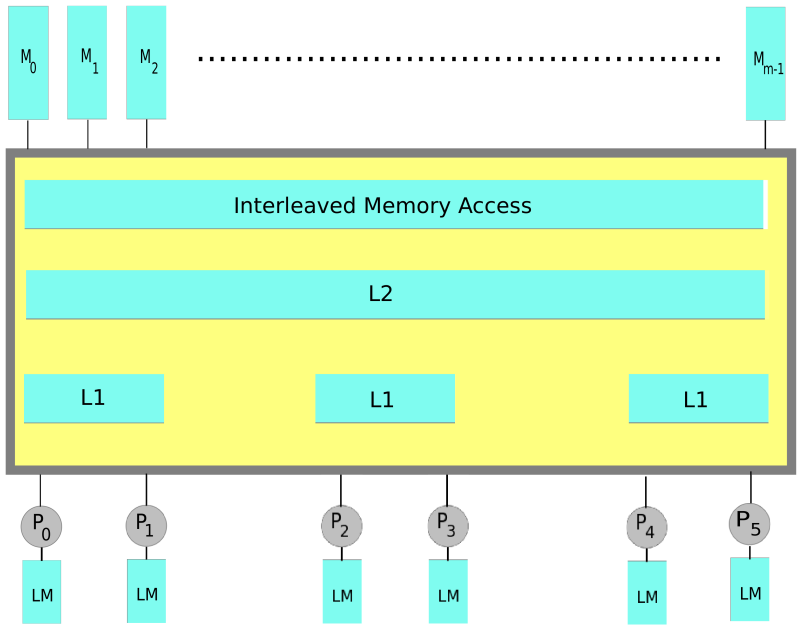
Network Characteristics → Parallel Processing Architectures



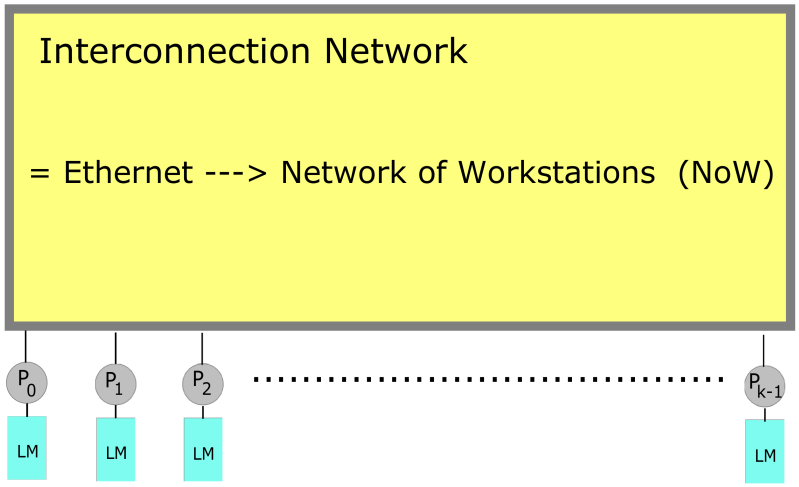
Network Characteristics → Parallel Processing Architectures



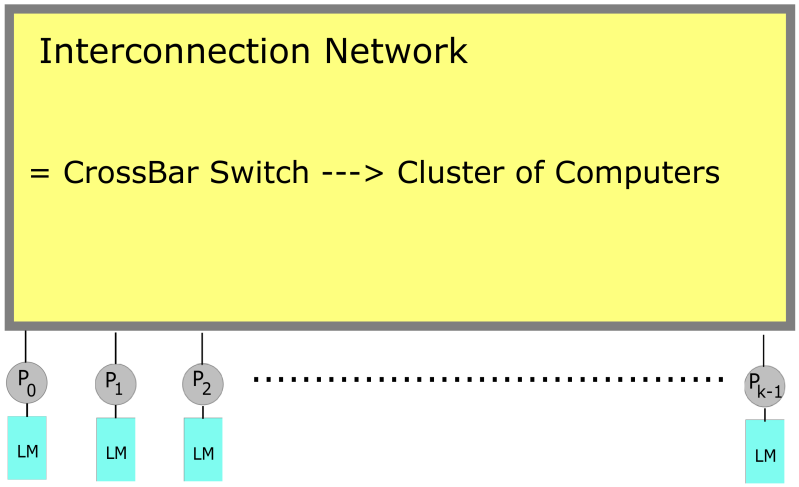
Network Characteristics → Parallel Processing Architectures



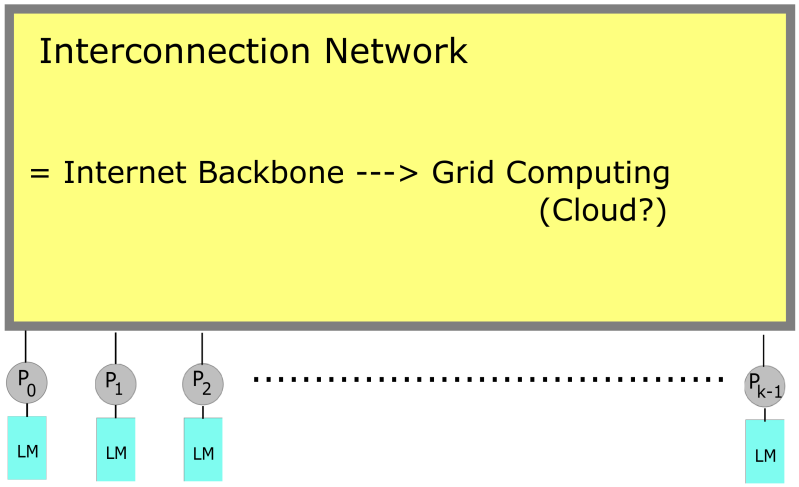
Network Characteristics → Parallel Processing Architectures



Network Characteristics → Parallel Processing Architectures

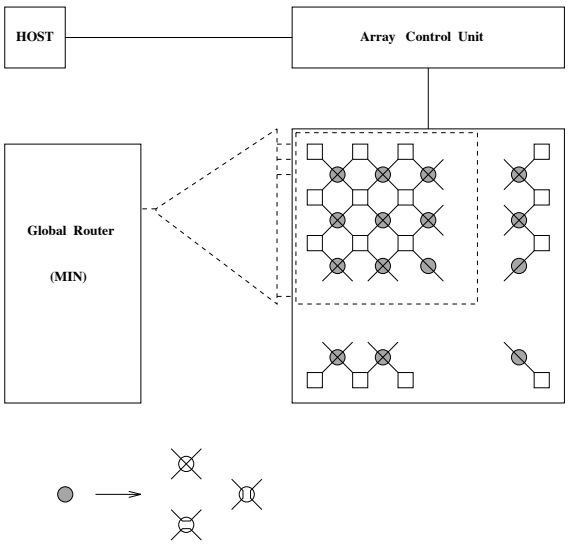


Network Characteristics → Parallel Processing Architectures



Network Characteristics → Parallel Processing Architectures

The Maspar Inc. Massively Parallel Computer Architecture:
A two-network massively parallel computer.



Network Characteristics → Parallel Processing Architectures

The Maspar Inc. Massively Parallel Computer Architecture:
A two-network massively parallel computer.

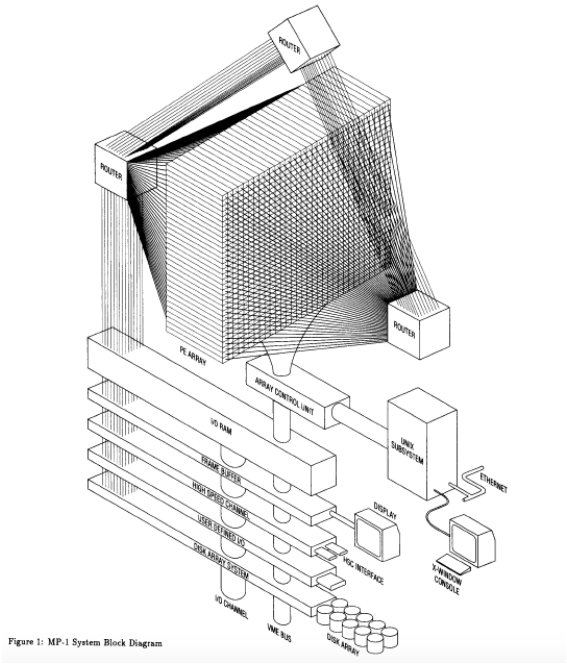
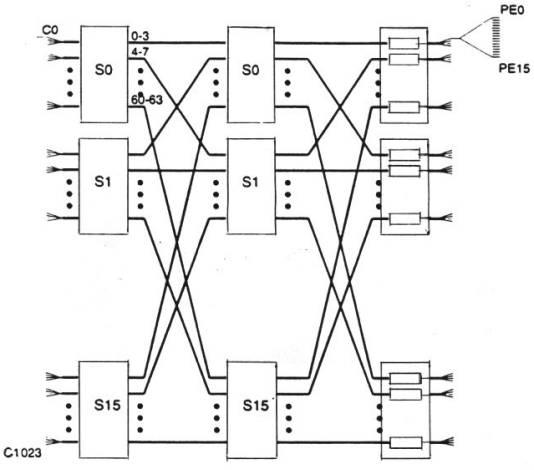


Figure 1: MP-1 System Block Diagram

Network Characteristics → Parallel Processing Architectures

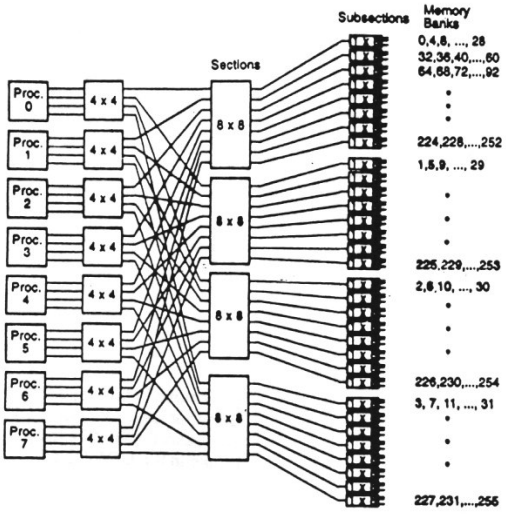
The Maspar Inc. Massively Parallel Computer Architecture:
The Global Router.



A Delta ($16^2 \times 16^2$) Network with 256 inputs
capacity of 4, clusters of 16 with shuffle at output

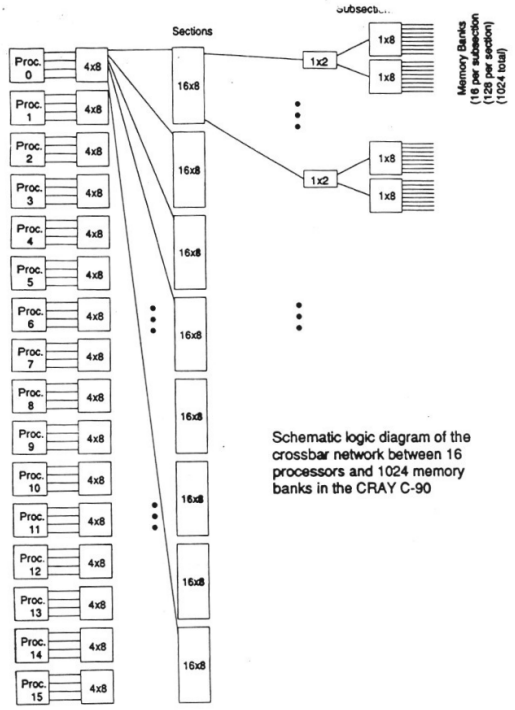
Network Characteristics → Parallel Processing Architectures

CRAY Research's C90-Y-MP Computer



Network Characteristics → Parallel Processing Architectures

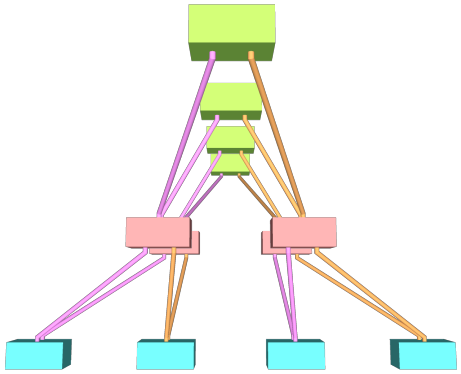
CRAY Research's C90-Y-MP Computer



Schematic logic diagram of the crossbar network between 16 processors and 1024 memory banks in the CRAY C-90

Network Characteristics → Parallel Processing Architectures

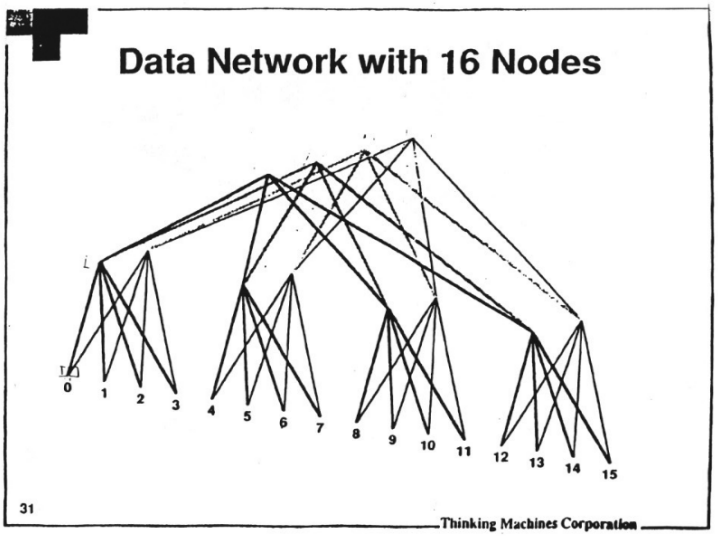
- Dan Hillis at MIT introduced the Connection Machine as part of his PhD thesis.
- Based on the idea, Thinking Machines Corporation was funded.
- CM1 and CM2 evolved into CM5, one of the most popular supercomputers of the time.
- CM5 had a Data and a Synchronization tree-like networks.



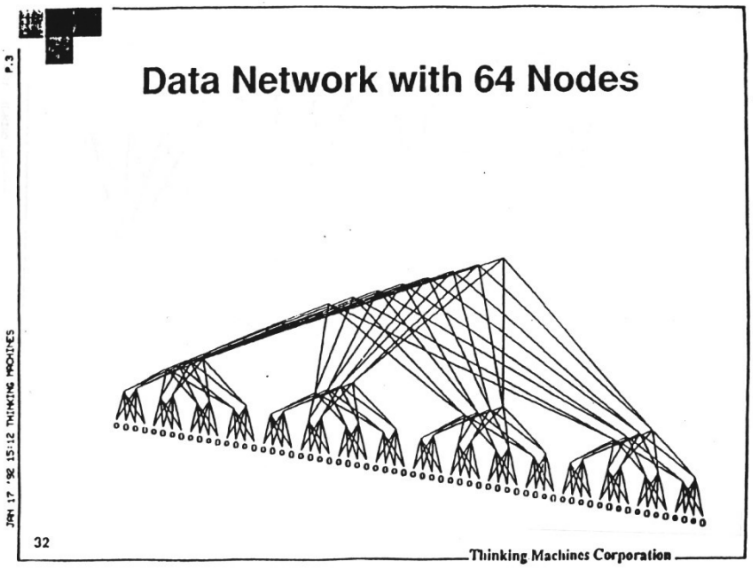
CM5 used a Network called a Fat Tree

Network Characteristics → Parallel Processing Architectures

The CM5 sales pitch transparencies
The fat tree network.



The CM5 sales pitch transparencies
The fat tree network.



Lipovski and Malek
The Texas Reconfigurable Array Computer (TRAC).

Parallel Computing
Theory and Comparisons

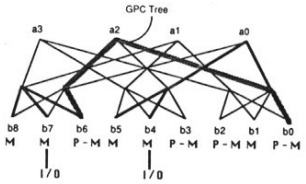
G. Jack Lipovski

Miroslaw Malek

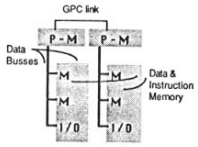
Department of Electrical Engineering and Computer Science
University of Texas at Austin

A Wiley-Interscience Publication
JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore
1987

Lipovski and Malek
The Texas Reconfigurable Array Computer (TRAC).



a) Tree structures



b) Programmer's view

Figure 8.12. (Virtual) SIMD using the GPC tree.

In preparation for the discussion of packet switching, consider a circuit switched data tree (possibly augmented by active chains of several shuttle trees), the data is transferred in a memory cycle. The memory address is sent, and later data is transferred to or from memory, in fixed parts of the cycle. There will be parts of the cycle where the memory management unit or the memory is busy. During such a part, a packet can be moved through an arc in that data tree. When the data pins are being used in circuit switched mode, the arbitration and control of packet movement can be evaluated so the packet can move during the next time they are not used for circuit

Lipovski and Malek
The Texas Reconfigurable Array Computer (TRAC).

are synergetically shared to provide both circuit and packet switching. Note that a data tree and active chains attached to it must use the same clock (which can be tailored to the length of the data tree if an asynchronous bus protocol is used), but the packets can move into and out of trees with different clocks, as long as each arc is controlled for the time slots it has available in it. Also, arcs that are not part of a tree can be used for packet switching at any time.

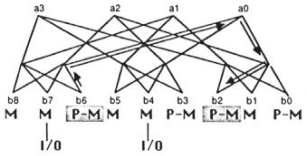


Figure 8.13. Packet switching.

Before we proceed with a routing scheme, we discuss the base-base path in a one-sided SW-banyan. A base-base path in a one-sided banyan is any path connecting a pair of bases (b_i, b_j) such that every node is traversed exactly once. It is uniquely defined by a triplet (b_i,b_j,c_k), where b_i is a source, b_j is a destination and c_k is a connection node at node level ℓ reachable from b_i and from b_j by a path of length $\ell = D(\{b_i,b_j\})$. Self-routing is based on the labeling scheme and uses an extension of a binary tree coding method discussed in [GOKE 1] for the networks built of 2x2 switches. In an (s,f,l) SW-banyan, bases are numbered using an l-digit number in base f and from any base node, a connection node at level i can be located by an i digit number in base s. Suppose b_i is to send a packet to b_j. The packet will turn around at a node c_k at level $D(\{b_i,b_j\})$. This node can be selected by an address , or else can be randomly chosen, by choosing the arcs in the upward path using a round-robin priority circuit. Random selection of c_k is used whenever the path need not be fixed, as this will give better throughput. The downward path is chosen by the address of the destination b_j, which is in the packet. Alternatively, a flip-flop can be set in some arcs to form a tree whose leaves are a set of resources to which a packet is to be broadcast; to broadcast

Lipovski and Malek
The Texas Reconfigurable Array Computer (TRAC).

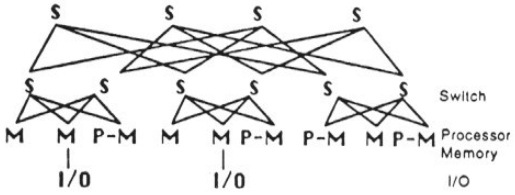
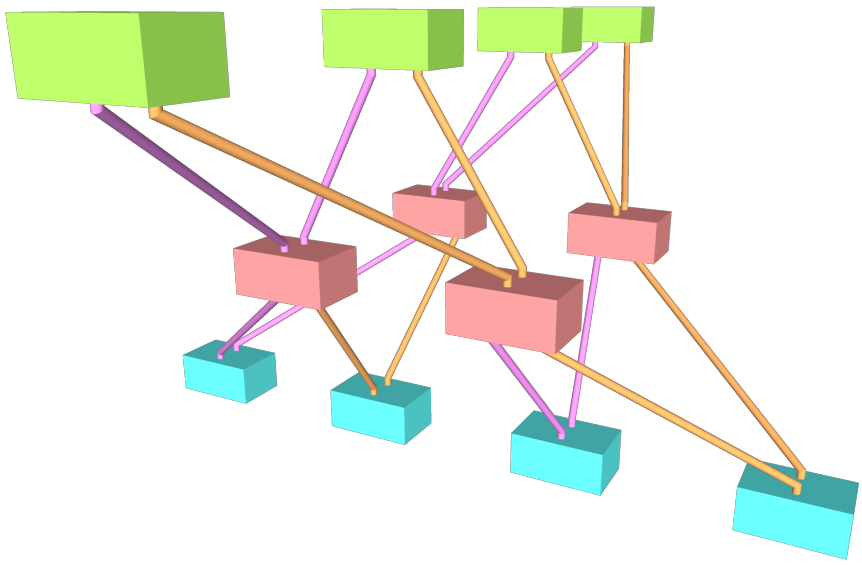


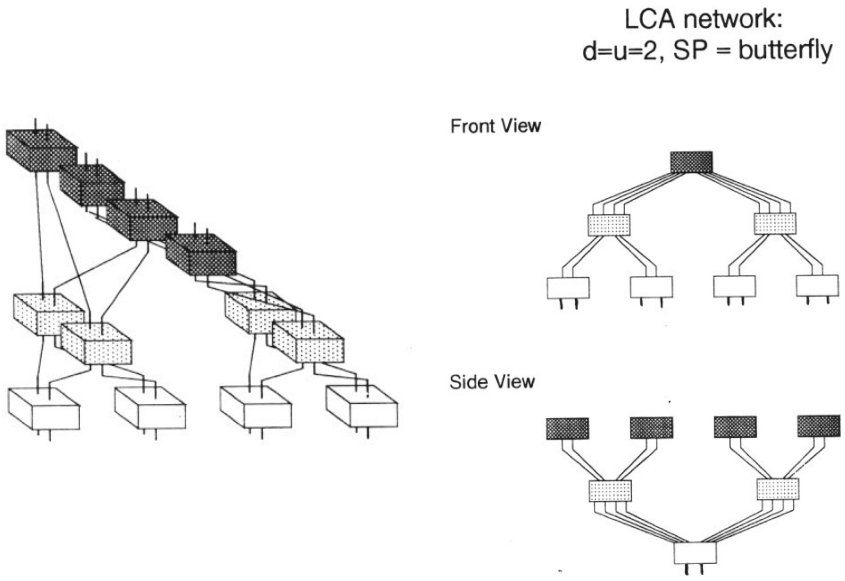
Figure 8.7. Structure of TRAC 2.0.



Another view of a Fat Tree

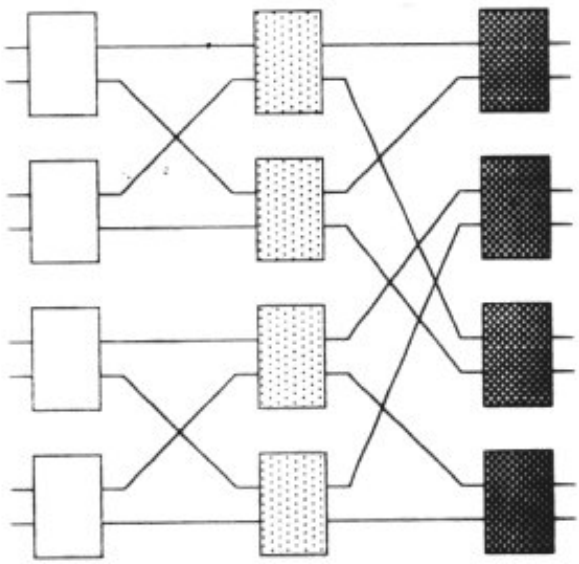
Network Characteristics → Parallel Processing Architectures

Another view of a Fat Tree.



Network Characteristics → Parallel Processing Architectures

Another view of a Fat Tree.



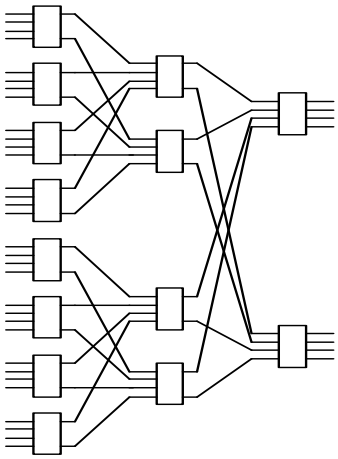
Network Characteristics → Parallel Processing Architectures

A 3D representation of the Fat Tree.

Play Video

Network Characteristics → Parallel Processing Architectures

A network we shall look at derived from Fat Trees.
These will be called Least Comon Ancestor Networks (LCANs)



Clusters

Problems with Conventional Supercomputers

- Very costly
  - Specialized processors not cheaply available
  - Specialized interconnects to support bandwidth needed
- Harder to program
  - Uncommon processors
  - Lack of standard programming model and interface
  - Lack of standard tools
- Shorter life span
  - Harder to upgrade
  - Scalability a problem for many

Enablers for Clusters

- Individual machines are becoming very powerful, no need for specialized processors to achieve required speed at each node
- Faster network technology reduces the need for specialized, proprietary interconnects between processors
- Incremental scalability – add nodes as needed
- Use of common-off-the-shelf (COTS) components implies lower cost and ready availability
- Development tools are more mature
- Standardized programming interfaces like PVM, MPI etc. makes programs portable

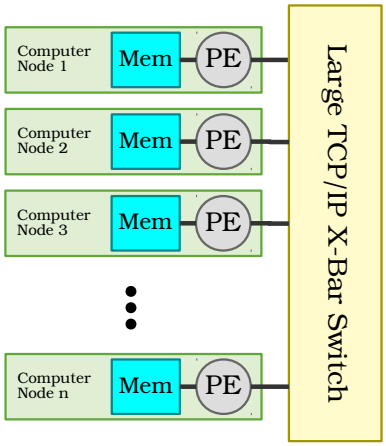
Popularity of Clusters in HPC

- `www.top500.org` - list of the top 500 supercomputers in the world, updated twice per year
- Ranked according to their performance on the standard Linpack benchmark
- 294 of them in the list of Nov. 2004 are clusters!
- Highest rank of a cluster – 2 (approx. 51 teraflops)

Cluster Components

- A typical cluster
  - Stand alone machines
  - A fast network connecting them
  - Low latency communication protocols
  - Software to give Single System Image
  - Programming Tools
- Additional components
  - Network RAM
  - Parallel I/O

Typical Cluster Architecture



Several interconnected stand-alone machines

Example: Berkeley NOW

- 100+ SUN UltraSparc machines (Ultra 170)
- 200 disks
- Myrinet interconnection within cluster– 160 MB/s
- Switched Ethernet to ATM backbone for external communication
- GLUnix – global OS over Solaris for process management
- AM (Active Message) communication protocol
- MPI for programming

Cluster Classification

- Target applications
  - High-Performance Clusters – for scientific apps.
  - High-Availability Clusters – for critical apps.
- Node ownership
- Node Hardware
- Node OS
- Node Configuration
- Clustering Levels