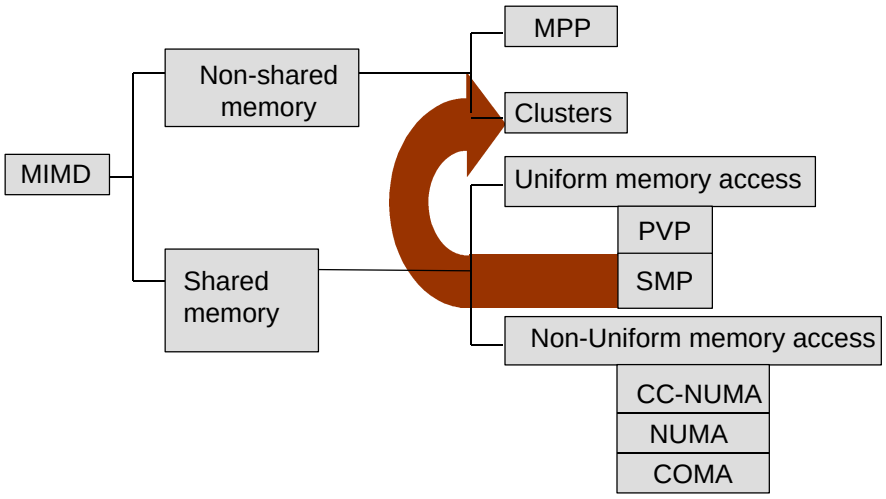
An Overview of PARAM 10000

Raise and Fall of Computer Architectures

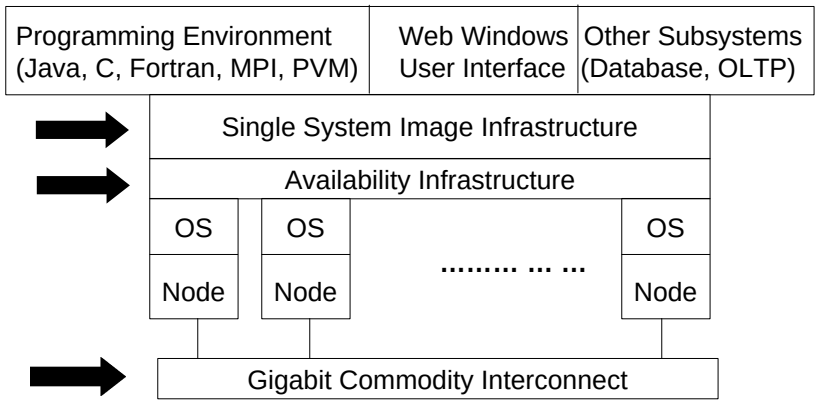
- Vector Computers (VC) – proprietary system
  - Provided the breakthrough needed for the emergence of computational science, buy they were only a partial answer.
- Massively Parallel Processors (MPP)) – proprietary systems:
  - High cost and a low performance/price ratio.
- Symmetric Multiprocessors (SMP):
  - Difficult to use and hard to extract parallel performance.
- Distributed Systems:
  - Difficult to use and hard to extract parallel performance.
- Clusters - gaining popularity:
  - High Performance Computing – Commodity Supercomputing
  - High Availability Computing – Mission Critical Applications

Architectural Models: MIMD



Cluster of Computers

Idealized Cluster Architecture



Cluster of Computers - Features

- Collection of nodes physically connected over commodity/ proprietary network
- Cluster computer is a collection of complete independent workstations or Symmetric Multi processors
- Network is a decisive factors for scalability issues (especially for fine grain applications)
- High volumes driving high performance
- Network using commodity components and proprietary architecture is becoming the trend

Cluster of Computers

(Contd…)

Clusters Features

- High-end technology to all (System Area Networks)
  - MPP interconnection technology has arrived in commodity networks
  - The “killer switch”
  - Single chip building block for scalable networks
  - Low latency and high bandwidth
  - Intelligent Network Interface

Cluster Challenges

Cluster Challenges

- Enabling Technologies
  - Building Blocks
    - Microprocessors and Workstations
  - Fast Communication
    - System Area Networks and User Level Protocols
  - Distributed/Parallel Programming Software
    - MPI/PVM/ OpenMP/Pthreads/f90/HPF/

Cluster Challenges

(Contd…)

- The users view the entire cluster as **Single system**, which has multiple processors. The user could say: "Execute my application using five processors." This is different from a distributed system.
  - Single Entry; Single File Hierarchy; Single Networking Single Input/Output; Single Point of Control; Single Memory Space ; Single Job Management System; Single User Interface; Single Process Space ; Single System Symmetry ;
- Job Management
  - Global job management; Global system management and configuration; Group based scheduling and resource allocation
  - Load Sharing Facility (LSF)
- High Availability
  - Fault tolerance and Check-pointing

Commercial Models of Parallel Computer

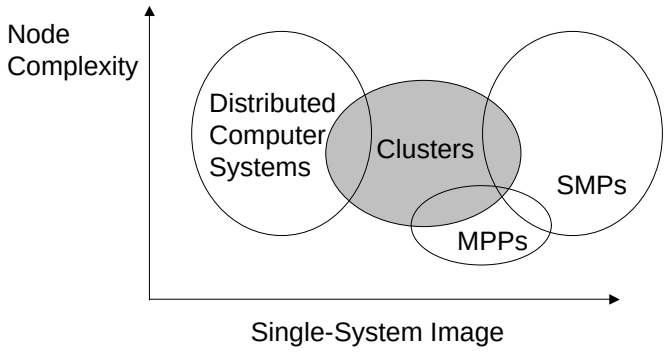
Cluster of Computers

- Digital TruCluster
- IBM SP2 (Considered as an MPP)
- Berkeley NOW project
- Cluster of SMPs – PARAM 10000

Clustering is becoming a trend in developing scalable parallel computers

Better Performance for Clusters

PARAM 10000 is based on cluster of SMPs



Overlapped design space of clusters, MPPs, SMPs, and distributed computer systems

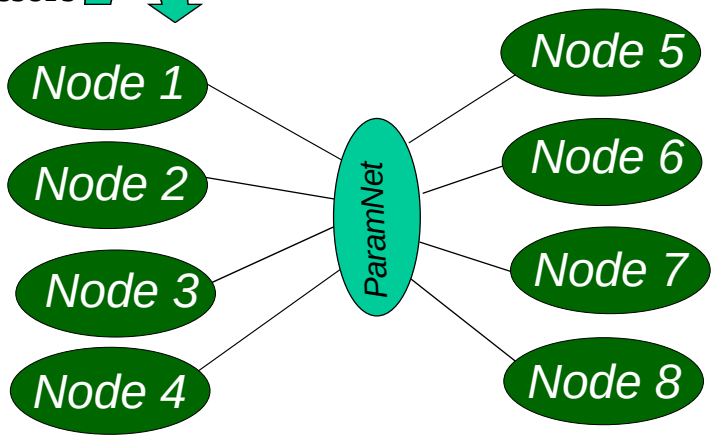
PARAM 10000 100 GF Parallel Machine

- Shared memory at Node level- 300 MHz, UltraSparc four way SMP with 2 MB external cache
- Node run replicated UNIX OS
- System Area Networks PARAMNet, Myrinet, FastEtherNet
- Message Passing : CDAC-MPI and C DAC HPCC software for Parallel Program Development
- Aggregate Main Memory : 33Gbytes
- Aggregate Storage : 800 Gbytes
- Peak Computing Power : 100 GF
- Connectivity : Intranet (Gigabit-Ethernet/ Extranet (FDDI),Internet (ERNET, VSNL)
- 100 GFlops *Peak Performance*

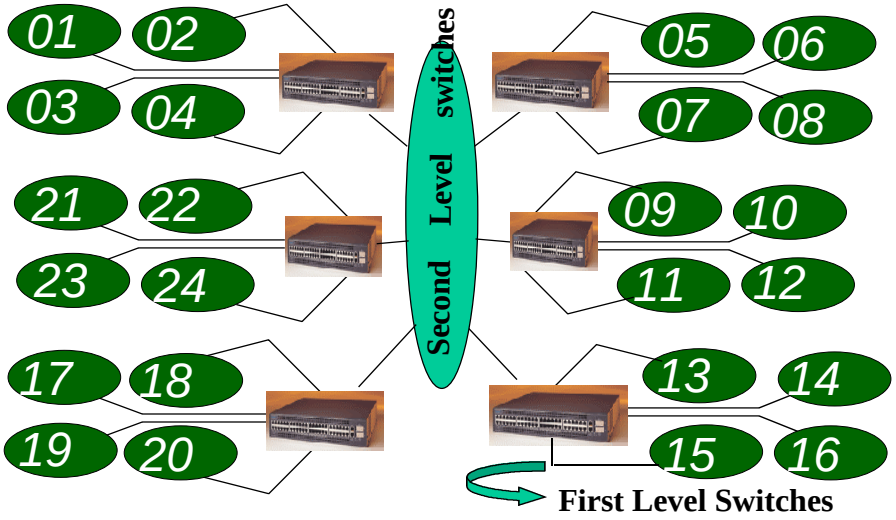
4th Milestone March 1998

A Cluster of SMPs With Single PARAMNet Switch

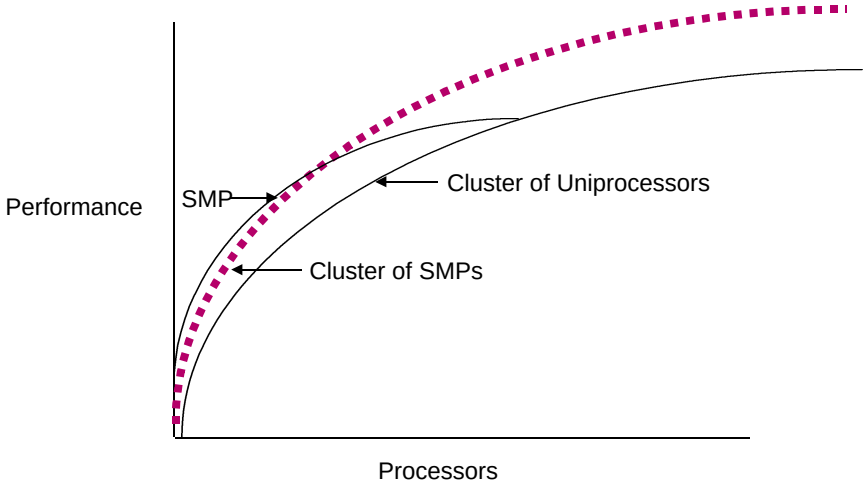
4 Processors



PARAM 10000 with Multiple PARAMNet Switches



Scalability of Scalable Parallel Processors



Symmetric Multiprocessors (SMPs)

Symmetric Multiprocessors (SMPs) characteristics

- All processors see same image of all system resources
- Equal priority for all processors (except for master or boot CPU)
- Memory coherency maintained by HW
- Multiple I/O Buses for greater Input Output
- Cost of Communication is very less

System-Area Networks (SANs) Features)

- TCP/IP is Designed for WANs
  - OS overhead for send/receive is extremely high.
  - Low bandwidth.
  - Higher error rates and high per packet processing cost.
  - Does not exploit the high reliability of SANs.
- What is required?
  - A better protocol with reduced communication overheads, high bandwidth which exploits the reliability of SANs.

System Area Networks (SANs) Features

(Contd…)

- Switched, low-latency, high speed networks Replaces the backplanes and cabinets of massively parallel processors into the traditional territory of local area networks
- Use Source-based message routing through anonymous switches
- Topologies may no longer be the static
- Topologies may no longer be well-defined and well-understood graphs such as *hypercubes*, *meshes*, *tree* etc.
- Graphs may be arbitrary and change over time

System Area Networks (SANs)

(Contd…)

PARAMNet	Myrinet	Ethernet	ATM
C-DAC's 8 Port PARAM Switch cascadable with scalable bandwidth upto 1000 nodes using multistage network	Myricom's 8/16 Port Switch	8/12 Port Fast Ethernet Switch 8 Port – Gigabit Switch	16 Port ATM Switch
400+400Mbps full duplex bandwidth per port	1.28+1.28Gbps full duplex bandwidth per port	10/100Mbps full duplex bandwidth per port	155Mbps per port & 622Mbps per port
Wormhole nonblocking with adaptive routing	Nonblocking cut-through switching	Store & Forward	Cell Switching
PCI Adapters	PCI Adapters	PCI Adapters	PCI Adapters
KSHIPRA Lightweight Communication substrate/ Streams Driver/(TCP/IP)	Low Latency Protocol Support/Streams Driver/ (TCP/IP)	TCP/IP	MultiProtocol Over ATM (MPOA) & TCP/IP

SAN : PARAMNet

PARAMNet Components

- Network Interface Cards; Optical Fibre Extender and PARAMNet Switch

Network Interface Cards

- PCI and Sbus Support
- Supports upto 4 DS-links (400 Mbits b/w each direction)
- Use of DMA for efficiency.
- Powered by C-DAC's Communication Co-processor (CCP) Chip.
- Device drivers for Solaris and Windows NT

SAN : PARAMNet

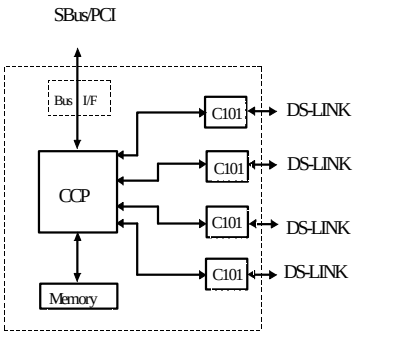
(Contd…)

CCP (Communication Co-Processor) :

Features

- Low Latency, High throughput Communication
- Latency less than 15 μsec at application level.
- Supports upto 4 DS-links.
- DMA used for both transmit and receive
- Zero copy data transfer is supported.

Network Interface Card
: Block Diagram

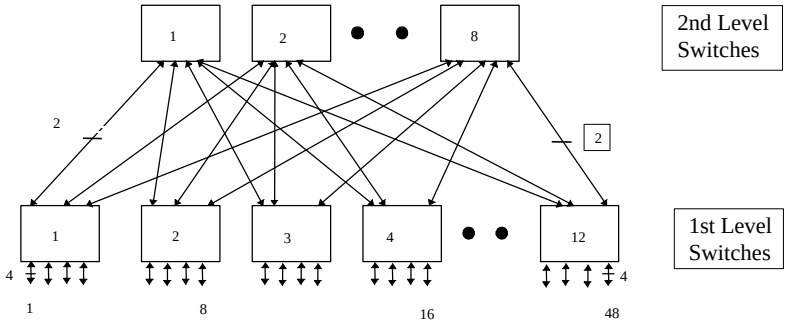


C101 : Parallel DS-LINK Adapter

SAN : PARAMNet

(Contd…)

Network Realization for 100 GF PARAM Openframe



Host Machines

SAN : PARAMNet

(Contd…)

Performance Summary of PARAMNet

- Latency
  - Low Latency < 1usec for switching
  - Low latency protocol of CCP2. (< 10 usec)
  - User level direct access to h/w to reduce kernel overheads.
  - Polling vs. Interrupt
    - Poll register is provided in main memory.
    - Interrupt is optional.

SAN : PARAMNet

(Contd…)

Performance Summary of PARAMNet

- Bandwidth
  - Multiplexing of a message on all 4 links.
  - CCP protocol to avoid unnecessary copy of message.
- Multistage n/w performance
  - Use of group adaptive routing to reduce blocking.
  - Cross sectional B/W of ((no. of nodes*400) / 2) Mbits/sec in each direction, so that all nodes can talk simultaneously to other nodes without performance degradation .

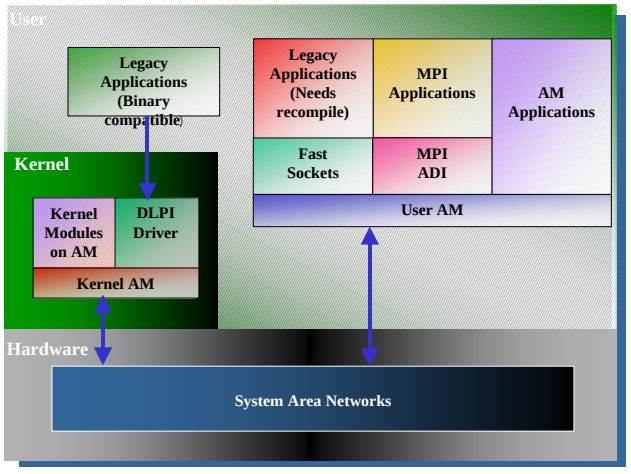
Active Messages for Fast Communication

- It is an *asynchronous communication* mechanism to realize low overhead communication
- The *objective* is to expose to the user the raw capability of the underlying communication hardware
- The basic idea is to use the control information at the header of a message as a pointer to a user-level subroutine called *message handler*
- When the message header arrives at the destination node, the message handler is invoked to extract the remaining message from the network and integrate it into the ongoing computation.

Active Messages for Fast Communication

- Active Message is a general mechanism, not required to be tied to any hardware or software problem
- It is implemented on Cluster of SMPs on PARAM 10000.
- AM is used as a low-level communication layer that can deliver a large percentage of the raw performance of the communication hardware.
- CDAC HPCC software - KSHIPRA Active Messages are designed for SANs.

HPCC - KSHIPRA Design



HPCC - KSHIPRA Design

(Contd…)

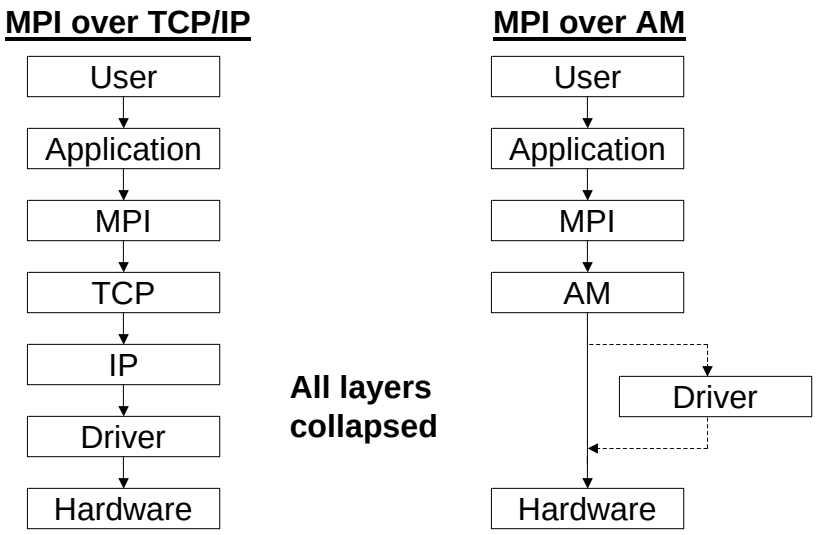
- Designed for clusters
  - KSHIPRA Active Messages are designed for SANs.
- Protected User Level Primitives
  - No OS overhead for send/receive.
- Success oriented protocols
  - Exploits high reliability of SANs.
  - Minimal per packet processing cost.

HPCC - KSHIPRA Design

(Contd…)

- Designed for clusters
  - KSHIPRA Active Messages are designed for SANs.
- Protected User Level Primitives
  - No OS overhead for send/receive.
- Success oriented protocols
  - Exploits high reliability of SANs.
  - Minimal per packet processing cost.

C-DAC- HPCC software - Active Messages



PARAM - Main Software Components (contd…)

- Parallel Computing Training
- Vis 5D, VolVis, VolPack, Khoros, IRIS Explorer/AVS Express
- MesaGL, OpenGL, Open Inventor
- Dedicated slot booking
- Job Accounting
- Batch jobs (PBS)
- GUI for Job execution
- Sun Workshop 6.0
- ScaLAPACK, PetSc, LINPACK
- ORACLE, Database Tools
- C-DAC F90 Tools Source browser, Profiler, Debugger
- SPARC Tools Source browser, Profiler, Debugger
- Multithread Tool LoopTools, Thread Analyzer, Debugger
- HPCC software tools for parallel programs (MPI)
- C-DAC/Apogee F90
- SPARC F90
- SPARC F77, MF77
- SPARC C/MC
- SPARC C++
- JAVA

C-DAC HPCC Software : DIViA

- DIViA (Debugger with Integrated Visualizer and Analyzer) is an advanced portable and flexible parallel debugging environment.
- It consists of a coherent set of tools that help programmer in both correctness and performance debugging.
- Correctness Debuggers:-
  - Multiprocess Debugger (MPD)
  - Message Debugger
  - Visual Debugger (ParViD)
  - Execution Monitor (ExMon)
- Performance Debuggers:-
  - Automatic Communication Bottleneck Detector (ABND)
  - Profile Visualizer (ProfViz)

Sun Enterprise 250 : Tools

About Sun Workshop Tools

- **FORTE:** Sun High Performance Computing (HPC) (A set of graphical tools that allow you to create and maintain your FORTRAN 77, Fortran 95, and C applications)
- **FORTE C++:** (A set of graphical tools that allow you to create and maintain your C++ and C applications)
  - Performance, Debugging and File Management Tools
  - Sun Workshop Visual /Team Ware
  - Multithreaded Development Tools
  - Sun Workshop Compilers/Debuggers

Sun Enterprise 250 : Sun Performance Libraries

❖ **Sun Performance Libraries**

- LAPACK version 3.0 For Solving algebra equations
- BLAS 1, 2 3 (Basic Linear Algebra Subprograms) - For performing vector-vector operations, matrix-vector operations and matrix-matrix operations
- FFTPACK - version 4 - For performing the fast Fourier transform
- VFFTPACK - version 2.1 - A vectroized version of FFTPACK for performing the fast Fourier transform
- LINPACK - For solving linear algebra problems in legacy applications containing routines that have not been upgraded to LAPACK 3.0

Explicit Parallelism

Explicit parallel models on PARAM 10000

Three dominant parallel programming models are :

- Data-parallel model (f90/HPF)
- Message-passing model (MPI/PVM)
- Shared-variable Model (OpenMP, Pthreads)
- Combination of Message Passing and Shared-variable Model (MPI-OpenMP and MPI-Pthreads)

Conclusions

Summary

- Features of PARAM 10000 – A cluster of SMPs are covered
- Importance of SAN in clusters are explained

Thank you