

3 System Hardware Configuration

3.1 DAISy

The DAISy Cluster is dual homed network of 16 Intel Pentium 90MHz workstations and very inexpensive UNIX compatible software. DAISy is a homogeneous research prototype used for scientific parallel distributed computing, and, a model for a minimum cost fast distributed computational system. The motherboards support 3 PCI (Peripheral Control Interface) [14] bus cards and 4 ISA (Industrial Standard Architecture) bus cards. Each node has 256Kbytes of 2nd level cache and 64Mbytes of random access memory (RAM). The PCI is noteworthy in that it is the first high performance, asynchronous I/O bus available for commodity PC architectures. Disk I/O functionality is handled by a bus-mastering PCI SCSI-II controller.

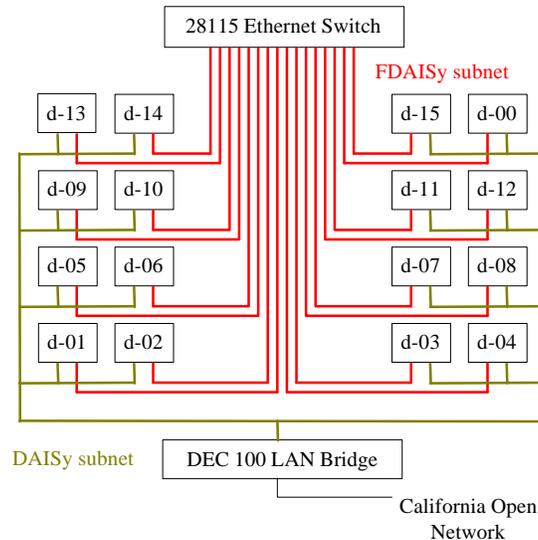


Figure 10. DAISy & FDAISy Subnet. Individual CPUs are labeled d-00 through d-15.

These systems are connected with standard 10Mb/s Ethernet (10BASE-2) and ISA bus 10BASE-2 Ethernet network interface cards(NICs) in a bus broadcasting network topology. This interface is used for client node NFS mounts, and any client node interactive work users find necessary. A second network facility is also being used, a high speed frame switch (28115 Fast Ethernet Switch by Bay Networks) with PCI bus Fast Ethernet (10/100BaseT) NICs connected in a point-to-point star topology. The architecture of the 100BASE-TX network is designed to ensure contention free, high performance network communications. DAISy is a subnet on Sandia's Internet backbone via a DEC 100 bridge.

As Figure 10 shows, DAISy is accessible via traditional ethernet through the DEC LAN bridge. The bridge is used to separate the cluster from Sandia's open network and treat it as a true subnet. Only data packets that need to cross the bridge do so. This helps reduce the amount of network collisions that the cluster will see. Also notice that the switched (FDAISy) subnet is also isolated. Since the network media is point-to-point though the switch the FDAISy subnet is also a true subnet. daisy-00 (d-00) is the

main system on the cluster, but access to the DAISy cluster can be from any system d-00 through d-15.

Access to FDAISy can only be accomplished after a user has established a connection on any one of the DAISy nodes.

Each node in the model consists of:

Intel Pentium based workstation (3 PCI, 4 ISA slots), 90MHz

Motherboard: Intel Premier 90MHz w/Neptune chipset, P54C-PCI w/256k cache
 CPU: Intel Pentium P54C 90MHz
 RAM: 64MB (2@8x36) 60ns 72 pin SIMMs, w/parity
 SCSI Controller: PCI fast SCSI II NCR53810 controller
 Ethernet: 3COM 3C509 Etherlink III Combo, EISA
 SMC EtherPower 10/100, PCI
 Hard Drive: Quantum PD1080S, 1GB fast SCSI II 9.5ms internal
 Floppy: Teac, 1.44MB 3.5"
 Video: SVGA 512k, 1024 x 768
 Case & PS: medium tower case w/250W power supply

D-00 also includes the following:

CDROM Drive: fast SCSI 3x speed NEC 3x1 CDR-510
 Tape Drive: 8-16Gb Wangdat internal Dat
 Video: ATI Mach 64 Win Turbo, 2Mb VRAM
 Hard Drive: IBM fast SCSI II 4GB internal, <10ms

3.2 Memory

Figure 11 shows a more detailed view of how DAISy can be viewed as a generalized parallel processor with private memory. As mentioned earlier, each node contains 64M bytes of RAM and 256k external cache. The file system in each node contains 100M bytes of swap space, and 3.6G bytes of user space (/) in d-00 and 600M bytes of user space in d-01 through d-15. A 300M byte partition is left on each node for researching other operating systems.

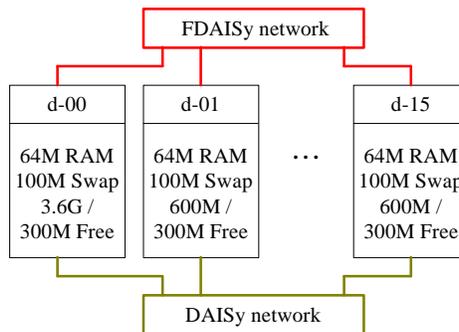


Figure 11. Detailed description of DAISy as a generalized parallel processor

3.3 The P54C Pentium™ 90 MHz Processor

The defacto workstation cluster consisted of a network of RISC UNIX systems as described in the introduction. With free UNIX compatible software already available on architectures based on Intel's i386™ and i486™ CPUs the Distributed Computing Research Group at SNL,CA decided to construct a network

of CISC UNIX systems consisting of commodity components. The intent in purchasing the hardware for DAISy was to obtain the most technologically advanced PC hardware available that UNIX compatible software could run on. Since the P54C Pentium™ Processor is backward-compatible with the i386™ and i486™ CPUs and freeware software, such as Linux, NetBSD, and FreeBSD, run on these CPUs the P54C Pentium™ Processor was the logical choice. At the time there were two choices of motherboards. The Intel Premier 90MHz w/ neptune chipset and the SuperMicro 90MHz w/ Opti chipset. The Intel Premier 90MHz w/neptune chipset was chosen. Performance benchmarks were run by Russell Carter (contractor at Sandia) showing the Intel board with the NCR SCSI controller was the faster of the two motherboards.

At the time of this writing, 166 MHz Pentiums and 200MHz Pentium Pros are standard, leading to CPU performance increases of a factor of 2-4. Advances in cache memory and main memory technology have increased performance of these critical components by over 50%. The performance of the P54C-90 is used as the basis for the computation of price/performance.

3.4 PCI (Peripheral Control Interface)

The PCI bus was the appropriate choice for workstation like performance for DAISy, as long as hardware and software drivers were available. The potential bandwidth [15] of the PCI bus is about 60MB/sec

as well as allowing shared hardware interrupts and posted writes. The following is a comparison of different busses available on PCs and is borrowed information from the SCSI-HOWTO by Drew Eckhardt from "Dr. Linux" as part of the Linux Documentation Project [15].

- ISA** The original PC expansion bus extended for the IBM AT and clone computers. At 8MHz it should have a bandwidth of about 5MB/sec and edge triggered tri-state interrupts effectively prevent interrupt sharing.
- VESA** The Video Electronics Standard Association local bus is basically an extension of the processors bus which was designed for high speed video applications. The transfer speed of approximately 30MB/sec makes it useful for other applications.
- MCA** IBM's attempt to corner the PC market with a proprietary bus which they would license to others for a stiff fee. MCA bus machines are no longer widely available and have little real support under Linux.
- EISA** The extended industry standard architecture bus is a consortium of PC clone makers' answer to IBM's MCA. It is backwards compatible with the ISA bus and supports some interrupt sharing as well as a 30MB/sec transfer rate. Since the EISA bus was designed with bus-mastering in mind, it usually performs slightly better than the VESA bus for SCSI disk access.
- PCI** Peripheral Control Interface Intel's answer to the VESA local bus, the PCI bus is the latest addition to the bus wars. It combines the best of the other busses with a 60MB/sec transfer rate. Both master and slave bus-mastering are supported as well as shared interrupts making PCI appear to be the bus of the future. If current trends continue, the PCI buss will be ubiquitous within a few years; however, in the PC world, the future is never assured.

3.5 Frame Switched 100BASE-TX Fast Ethernet

The 100BASE-TX Fast Ethernet network uses a Synoptics 28115 frame switch [16] to reduce latency and increase aggregate bandwidth to message passing functions in user programs on the DAISy cluster. Frame switching is primarily used to enhance network performance by increasing the total amount of available aggregate bandwidth and decreases the overall communication latency. In the case of DAISy, an increased bandwidth of a factor of ten increases the number of parallel applications suitable for the cluster. Bandwidth is increased because contention is eliminated and multiple transmissions are allowed. For instance, ordinary shared media broadcast through a type of pipe communication. That is, all nodes connected to that pipe can see the broadcast, therefore; (1) all nodes look at the broadcast frame, (2) decide if the frame belongs to them, and, (3) act accordingly if the frame was addressed to them, otherwise (4) the nodes just continue to monitor the pipe. The advantage of a frame switch is that frames are unicast only to the port attached to the destination, much like the crossbar interconnect network seen in multiprocessor machines. Because the frame is only transmitted on a single port, other ports are available for other simultaneous transmissions. That allows a 16-port frame switch such as the LattisSwitch to receive frames on eight ports while it is transmitting on the other eight ports. Figure 12 shows how the 28115 frame switch can be modeled after a multiprocessor crossbar switch. The red dots indicate a closed switch. A total throughput of x8 can be achieved if eight ports are configured as inputs and 8 ports configured as outputs.

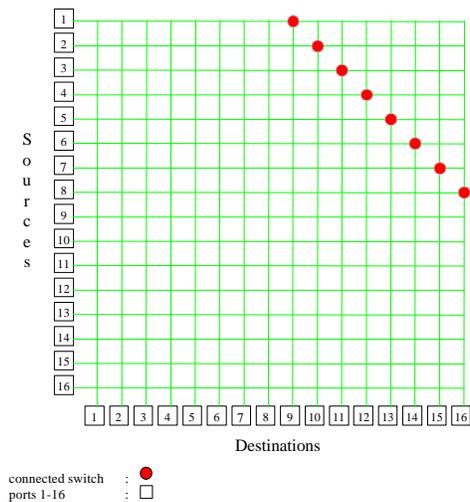


Figure 12. Frame switch modeled after multiprocessor crossbar interconnect network

The following measurement uses a modified version of the *lmbench* [3] TCP latency micro-benchmark. Specifically, all function calls were replaced with macros in order to minimize overhead. The TCP latency of DAISy's 28115 LattisSwitch was measured using this code as follows. First, two nodes were connected to the frame switch and a request response latency was measured. The test was repeated, this time connecting both nodes together directly using a crossover cable. The difference between the two latencies is attributed to the overhead incurred by sending packets through the switch. The results show a switch latency of 13.74 micro seconds.

<i>LAT_TCP</i>	
<i>571.95 us</i>	<i>w/switch</i>
<i>558.21 us</i>	<i>point-to-point</i>
<i>13.74 us</i>	<i>latency through switch</i>