BEOWULF CLUSTER FINAL DESIGN

UWM LSC GROUP
http://www.lsc-group.phys.uwm.edu/beowulf/medusa

B. Allen, P. Brady, D. Brown, J. Creighton, T. Creighton,
K. Flasch, A. Herbage, B. Owen, A. Wiseman, B. Wolfe

26-Jan-01

Purpose of this document

The University of Wisconsin – Milwaukee (UWM) LIGO Scientific Collaboration (LSC) Data Analysis Facility (called **MEDUSA**) will be made available for use by members of the LSC outside of UWM. For this reason, the LSC and the LIGO Laboratory have requested that a small ad-hoc committee review the final design. The purpose of this document is to present the final design for review.

History of the UWM LSC Data Analysis Facility

In November 1999, Allen, Brady and Wiseman submitted a proposal to the NSF Major Research Infrastructure program. A copy of this proposal is attached. The proposal requested funding of $415k from the NSF and $178k of matching funds from UWM for construction of an LSC Data Analysis Facility. The proposal was funded at the requested level in August 2000.

Need for review

The proposal was accompanied by letters of support from the Director of the LIGO Laboratory (Barry Barish) and the Spokesman of the LSC (Rainer Weiss). These letters requested that before purchase, a committee from the LSC and the LIGO Laboratory review the suitability of the final design.

Purpose/Philosophy/Design Goals

The primary purpose of the facility is to assist in developing the methods, algorithms and code needed to analyze LIGO data. The facility is intended for rapid, loosely scheduled prototyping on representative LIGO data sets. It is not intended to carry out complete pipelined analysis of all LIGO data. This dictates the following features of the design:

- Data are stored on disks, not tape. This permits the data to be accessed in an unscheduled random order.
- The disks are inexpensive and “unreliable”. This is because the data stored on them can be replaced from the LIGO data archive if a disk fails.
- The overall system is designed to maximize computing throughput rather than reliability. This is because system downtime will not cause permanent data loss, and unusual or significant results can be double-checked.
- The system is intended for small numbers (<20) of users, of whom only a few are expected to be working at any given time.

Initial and Final Design Summary

The NSF MRI program requires that the proposal contain a complete design with vendor quotes. The primary characteristics of the design given in the MRI proposal are listed in the table below. However, in the proposal (page C9, section B2) we state

> "The design given here is a baseline design, not the final design, because the computer hardware market changes so quickly. In fall 2000, when the funding period begins, there will be higher performance hardware available for the same cost. Specifically, the optimal point, which gives maximum performance for a given cost, will have shifted. For this reason, the first three months of the funding period will be spent in testing small two- and four-node configurations using loaner system hardware. The goal during this period will be to obtain the maximum ratio of system performance to cost. So the baseline design given here should be regarded as a *lower bound* on the performance and capabilities that will be obtained."

This testing and benchmarking phase is now over and a final design is ready. The table below compares the initial and final design parameters.

	Proposal Design 11/1999	**Final Design 1/2001**
	SYSTEM SPECS	
Number of Nodes	128	250
Node Vendor/Machine	PC Wisconsin/ASUS	Gateway Professional/M1000
CPU	Intel PIII @ 550 MHz	Intel PIII @ 1 GHz
Memory/node	512 Mbytes SDRAM	512 Mbytes SDRAM
Disks/node	2 x 37 GB ATA-66	1 x 75 GB ATA-100
Disk controller	Promise Ultra ATA-66	ATA-100 on motherboard
Aggregate Peak Gflops	70.4 Gflops	250 Gflops
Aggregate Disk Storage	9.4 TB	18.8 TB
Networking	Fully-meshed: nodes channel bonded @ 200 Mb/s	Fully-meshed: nodes @ 100Mb/s + multiple Gb/s links to backplane
Tape robot capacity	200 AIT-2 tapes @ 50 GB	30 AIT-2 tapes @ 50 GB
Power utilized	22kw (estimated)	21.5kw (measured)
Cooling needed	6 tons (estimated)	5 tons (calculated)
	SYSTEM COSTS	
Cost of CPU & Disk	$297.7k	$335k
Cost of networking	$83.7k	$46k
Cost of tape robot	$71k	$0 (existing)
Cost of control network	$13k	$0 (will be added if needed)
Cost of shelving	$2k	$2.4k
Cost of monitor	$1.5k	$1.5k
Cost of Uninteruptible Power Supplies	$1.5k	$10.8k
Network upgrade	$0	$73.2k
TOTAL COST	$470.4k	$470.4k

Note that the proposed final design has more than three times the floating point performance, and more than twice the on-line disk capacity of the proposed design.

Testing Methodology

The testing was done with a custom code package (DRAG) available from the UWM LSC group web site. This package tested the speed of the FFTW Fast Fourier Transform (FFT) package on the different systems. It also tested the networking speed of simultaneous (full-duplex) data transfers. In all cases the code was compiled with all known useful optimizations enabled. It was run at low run levels so that the systems/caches were not loaded. The result of the FFT testing was a table of Mflops versus data set size. . The results of the testing are shown in a graph below. Typical LIGO data set lengths for astrophysical filtering are 15-minute data segments at a 1024Hz sample rate: approximately 900k samples or 2^{20} samples. For this reason, the performance metric was chosen to be the Mflop ratings at 2^{20} samples.

Systems tested

During the testing and benchmarking phase from September – December 2000, the following systems were obtained as loaners from vendors and benchmarked:

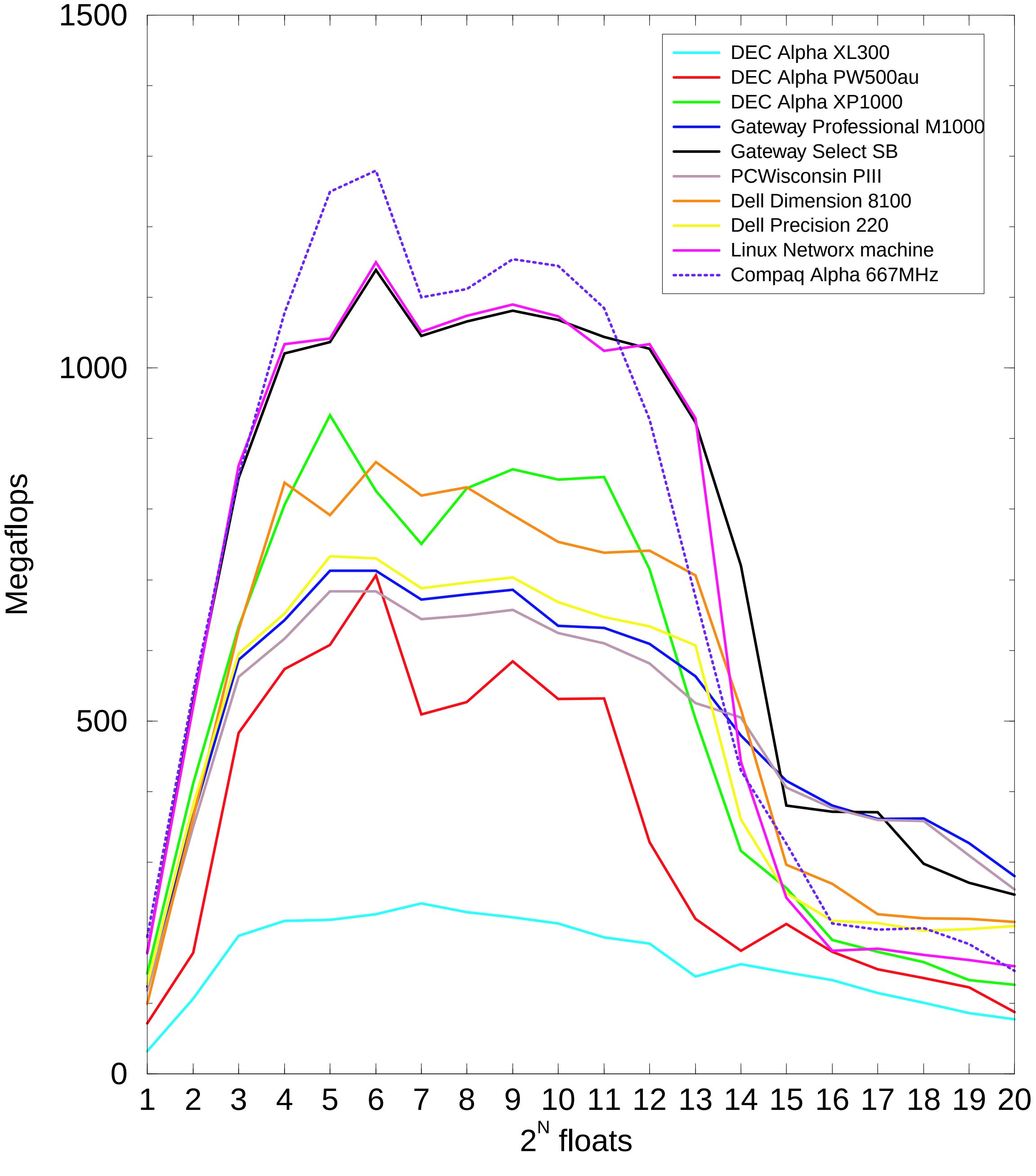
Vendor	**Board/ Machine**	**CPU/ Freq**	**Bus Clock (MHz)**	**Mflops @ 2^{20} points**	**Nodes for $335k**	**Total Gflops**
Gateway	Custom/ Professional M1000	Pentium III 1 GHz	133	280	250	70.0
PC Wisconsin	ASUS/ Custom	PIII @ 933 MHz	133	261	260	67.8
Gateway	Custom/ Select SB PC	Athlon 1.2 GHz	133	254	250	63.5
Dell	Custom/ Precision 220	Pentium III 1 GHz	133	210	178	37
Dell	Custom/ Dimension 8100	Pentium 4 @ 1.4 GHz	100	231	166	38
Linux Networks	Tysan Titan	Athlon 1.2 GHz	133	154	188	29.7
Compaq	PW500au	AXP 21164 500 MHz	?	127	No bid	
Compaq	XP1000	AXP 21264 667 MHz	83	146	No bid	
Compaq	XP1000	AXP 21264 @ 500 MHz	83	141	No bid	

Comments on testing

The testing revealed the following:

- The number of Mflops decreased rapidly with increasing data-set size, as can be seen from the figure below. This indicates that the limiting factor in the performance is not the CPU speed, but rather the memory/data bus bandwidth
- For short data segments the AMD Athlon was significantly faster than the other tested systems.
- The Alpha AXP CPU based systems were the clear winners when our group's last Beowulf was built in spring 1998. While this systems had high Mflop ratings for short data sets, once the data set moved out of L1 and then L2 cache, the performance dropped rapidly. We tested the AXP systems with both gcc and ccc compilers. The surprisingly poor performance of the FFTW package for long data set lengths prompted us to contact COMPAQ. They explained that FFTW makes poor use of the AXP architecture, and that for long data sets, the CPML routines are about a factor of 2 faster than FFTW. Since even this factor does not make AXP systems cost-effective, we did not carry out further testing.
- The "new technology" Intel P4 gave a poorer performance/price ratio than the Intel P3 and Athlon machines. These machines also provide system bus bandwidths that are better matched to CPU speed.
- The Intel P4 performed more poorly than the P3 with a slower clock speed. This may be because the size of the on-chip cache has been greatly reduced. It may also be because compiler technology has not caught up with the changes in the architecture. It may also be because changes in the instruction decoder have crippled the performance of the chip.

This graph shows measured FFTW performance in Mflops, for 2^N complex floats. For example if N=20, the data size is 8 Mbytes.



Networking

In the NSF proposal design, the networking technology was dual channel-bonded 100Mb/s Ethernet, with Gb/s concentrators, and a Gb/s central switch. At the time of the proposal, we hoped that by the time of construction, the cost of Gb/s networking would drop to the point where at least one higher speed interconnect might be affordable within our budget. For the final design proposed here, the cost of Gb/s Ethernet, Myrinet, and SCI (Dolphin) networking are all still too high.

For example, the cost of the largest available fully-meshed Gigabit switch available (Foundry Networks FastIron III with 120 copper ports) is more than $100k. By the time that the NICs are included, at $300 each, the total cost exceeds $170k.

In addition, as the table above showed, the greatest available performance is obtained by using a much larger numbers of nodes than initially proposed. For this reason, we have chosen an inexpensive and upgradable networking design. This design uses a Foundry Networks (http://www.foundrynet.com) FastIron III backplane, which has 15 card slots. 11 of these slots are populated with 24-port 100Mb/s Ethernet cards. 1 slot is populated with an 8-port 1000Mb/s copper or fiber Ethernet card. This provides a fully-meshed non-blocking switch with multiple high-bandwidth backplane connections and many potential upgrade paths. For the types of template-based searching currently envisioned, this system provides sufficient bandwidth, together with good upgrade paths, if these are needed to enable other types of distributed searches.

Our design is based on a level II rather than a level III switch. This is because we do not anticipate using separate VLANs. This might be appropriate for building several smaller beowulf systems that shared a common switch. However we intend to build a single large beowulf, and want to avoid the additional cost and complexity of level III networks. The FastIron III does permit trunking of lines on separate cards, which can be used if desired to get higher bandwidth channel-bonded Ethernet connections to specific nodes.

Other networking options were also considered. One was to use separate small switches as concentrators. Such switches, with 8 to 24 100baseT ports and 1 to 3 Gb/s copper Ethernet ports, typically cost between $700 and $1200. These switches would then be connected to a 24 to 32 port Gb/s switch. The total cost of such a configuration was comparable to the design that we have chosen, but it adds an additional level of latency in routing packets, as well as additional hardware to maintain. The chosen design is simpler, has lower latency, and has a simpler upgrade path.

We also considered networking backplane switches from HP and from Cisco. These were more expensive, and less expandable, than the design we have chosen. A Cisco switch with only 100-baseT capabilities that could handle our network is a bit less expensive than the Foundry Networks switch we have chosen, but it has no upgrade path to Gb/s and probably doesn’t have the non-blocking performance that the Foundry switch offers.

The figure below shows all of Medusa's networking connections. All of the compute nodes are connected to the FastIron III switch through symmetric 100-baseT Ethernet.

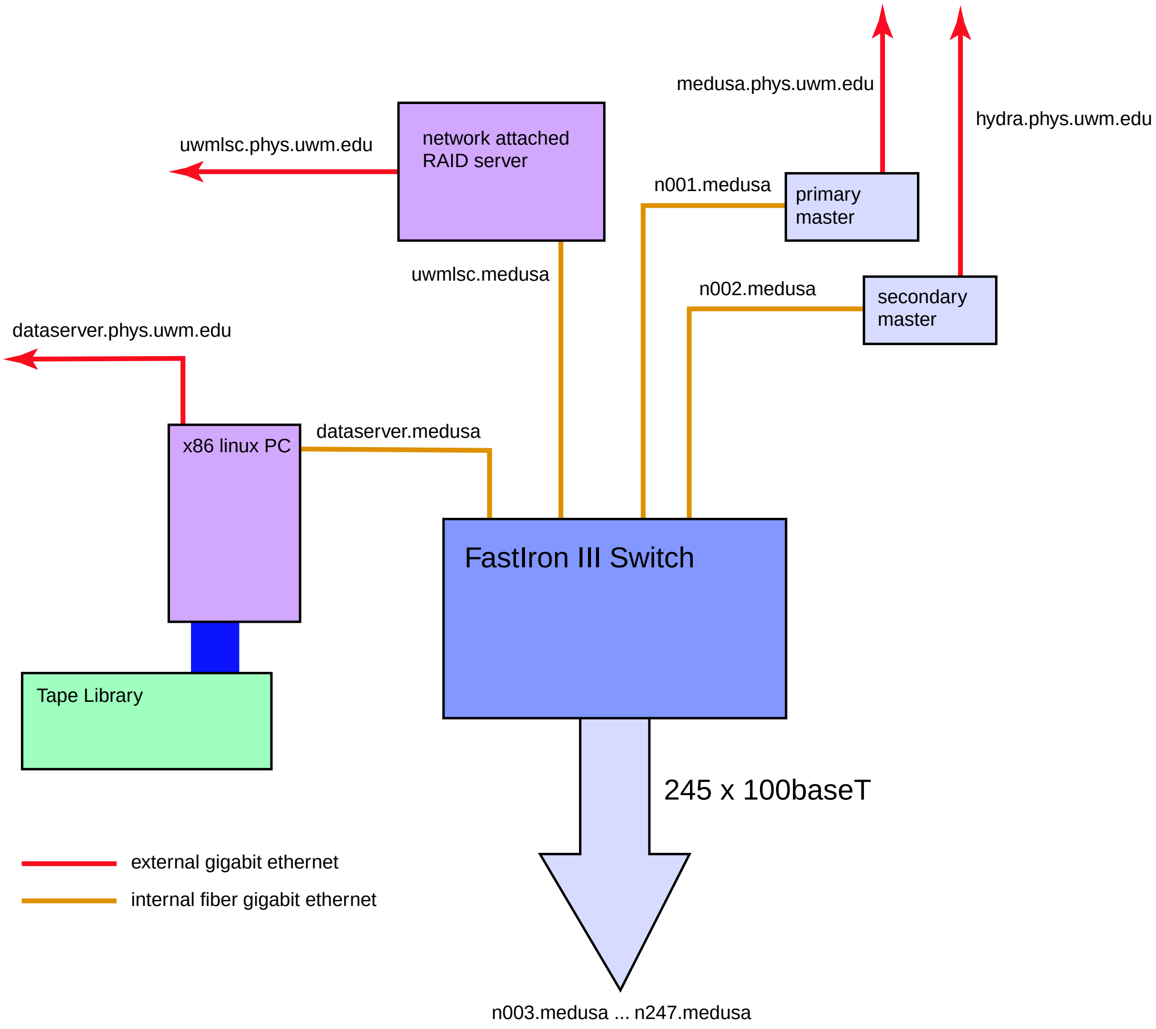
To reduce network break-in attempts, we have decided not to make the private Medusa subnet directly visible to the outside world. There is no route from the internet to the beowulf subnet across any of the machines that are connected to both. The outside world is networked to Medusa only through the master nodes.

The master nodes are connected to the Medusa switch via Gb/s fiber ethernet and to the UWM LSC group switch, via Gb/s copper ethernet. The LSC group switch is fiber connected to the UWM campus backbone.

Medusa's private subnet is also connected via Gb/s fiber ethernet links to:

- A Network Attached Storage (NAS) RAID sever
- A data server machine

This is to allow the beowulf nodes access to data, software and home directories that are stored on these machines.



Tape Robot

The NSF proposal contained funds for a 200-tape AIT2 robot. Recent testing of the network bandwidth for UWM/Caltech links has shown that it should be straightforward to transfer LIGO data over the network, eliminating the need for a tape robot. The measured bandwidth (evening, with 4 parallel FTP sessions running several hours) between UWM and Caltech is 1.3 MB/s. This is sufficient to transfer 37 GB/night of data. This will also improve with time. The current tape robot used at UWM (30 AIT-2 tapes) will be used for data ingestion/replication if needed. If this is used heavily, some contingency funds will be used to purchase an additional drive for the robot. Data ingestion, whether from tape or via the network, will be done through the dedicated data server node shown above.

The need for backups will be reduced because user home directories and shared system files will be stored on a 1 TB RAID-5 disk array. This system is a dedicated high-reliability hot-swap redundant system that permits on-the-fly reconstruction of data from any single failed disk

Uninteruptible Power Supplies (UPS)

Our experience with our existing Beowulf has taught us that to minimize maintenance, it is desirable to shield ourselves from failures in the electrical power grid. These are more prevalent in the summer than in the winter, but occur sporadically all year around. Most outages are of short duration. For the longer outages, our goal is to shut down the systems automatically with consistent file systems and without file system damage.

We considered a single dedicated central UPS system. Quotes were obtained from two vendors: Kramer Datapower and American Power Conversion. A 40 kW central system would weigh approximately 4000 lbs., and cost approximately $30k.

In the spirit of the rest of our design, we have instead chosen a commodity, less expensive solution. This is to use multiple small-system UPS units, which cost approximately $400 each. 24 such systems are sufficient to power the entire system for between 10 and 20 minutes, at a cost of about $10k. The UPS systems provide serial port connections, which can be used to notify the machines of power failure and to shut them down cleanly. The batteries in the UPS systems should last at least three years, after which they can be replaced for about $50 each.

Physical Design

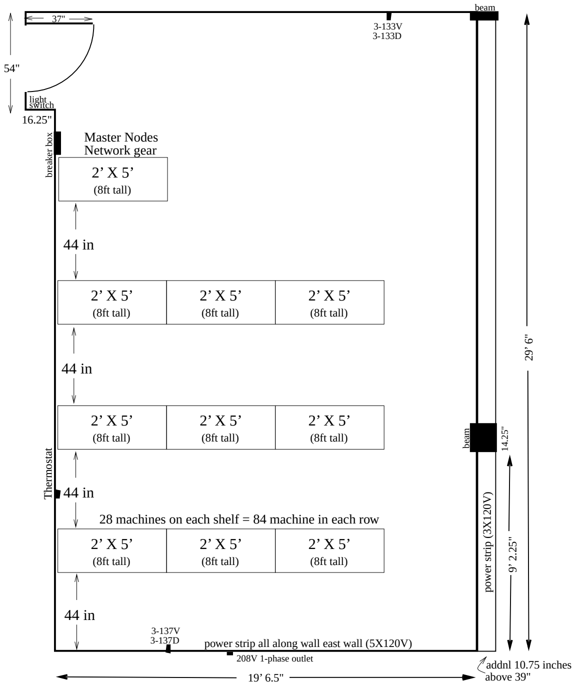
Starting in September 2000, a pair of rooms (Physics 331/333) in the UWM Physics Department have been remodeled to house the facility. The resulting room is shown in the figure below. It has the following characteristics:

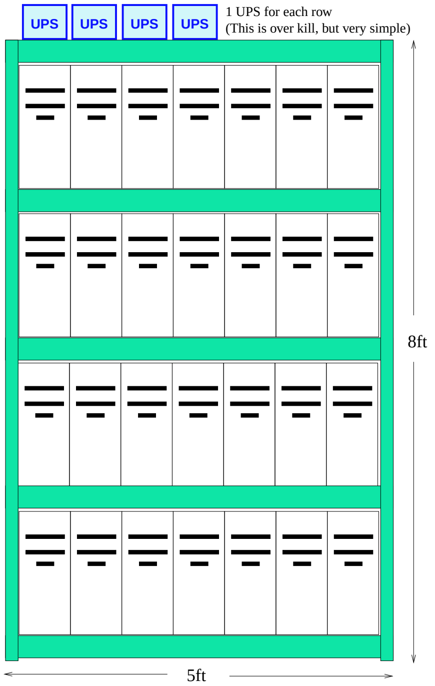
Dimensions: 19.5 feet x 29.5 feet, ceiling height: 10.1 feet.
Electrical power: 208 volts three-phase at 100 amps (or equivalently 300 amps single phase at 120 volts).

Networking: 9 Cat5e connections to the UWM backbone, currently at 100Mb/s. These can be upgraded to Gb/s if desired. Gb/s Fiber can also be added if needed.

The physical design uses approximately 2/3 of the floor space to house the compute nodes on industrial shelving units. Each shelving unit is 5 feet long, 2 feet deep and 8 feet high. The unit has five shelves, separated by 2 feet vertically, and has approximately a 1-ton load capacity. Each shelving unit will hold 7 nodes per shelf on four shelves. The top shelf will hold four UPSs, one for each of the four sets of 7 nodes. The systems are arranged as shown below.

Networking cables will be run in 8 trays located on the south wall. Each tray will carry cables to the 21 nodes located on a given “level” of a run of three shelving units. The networking cables will be velcroed together in sets of 7 and color-coded. The networking cables will be purchased from a “high quality” vendor, to custom lengths, and will be individually tested with a cable tester.





The physical design has the following characteristics:

- A 44-inch or wider corridor can be used to access all machines, front and back. This is large enough for a rolling cart, which can be used for maintenance.
- All electrical power is brought down from the ceiling to the top of the equipment shelving. This prevents floor clutter.
- The networking switch is centrally located. This reduces the length of the Cat-5e network cabling, and should simplify a potential future Gb/s upgrade.

Electrical Power

The power consumption of several individual nodes has been monitored using a digital true RMS AC ammeter. Typical idle current consumption of a node configured with disk and memory is 0.48A rms at 102VAC. Maximum observed power consumption with disk grinding and CPU pegged is 0.60A rms.

The total current draw of the system is thus expected to be approximately 150 amps rms. Since the (20% de-rated) capacity of the room is 240A rms, this is more than acceptable.

Two 15A 120V circuits will be provided to each equipment shelving unit, for a total of 18 15A 120V circuits. These circuits will "drop" to the top of each shelving unit. Each 15A circuit will power two UPSs. Each UPS will be connected to a single four-foot power strip on the appropriate shelf, and will provide power for 7 nodes.

Additional 20A 120V circuits will be provided for the other equipment areas, and for the networking switches.

Heat & Cooling

The power consumption above corresponds to 18kW. We anticipate no more than 3kW additional for the networking equipment and additional items. This corresponds to 21kW or approximately 72,000 BTU/hour of heat generation.

The room already has approximately 1-ton (12,000 BTU/hour) of cooling capacity. We intend to add an additional roof-mounted 5-ton air conditioning unit. This will be controlled by a single room-mounted thermostat, and will be connected to a separate dedicated source of electrical power.

Within the room, a single evaporator coil and fan will be centrally located high on the ceiling. Four air ducts, 44" x 6" will bring the cool air down to floor level, as shown in the diagram above.

Issues for the Review Committee

The UWM LSC group is confident that our final design is a good one. Nevertheless, there are a few issues on which we would appreciate advice from the review committee.

- **CPU Architecture**: With no change in cost, or in the number of nodes, we could shift from Intel P3 processors to AMD Athlon processors. While the P3 gives approximately 7% better performance on long FFTs, the Athlon is approximately a factor of 2 faster on short FFTs. This is because it has a larger cache, and more floating-point units than the P3. The AMD may be a better general-purpose floating-point engine. Does the committee believe that shifting to AMD processors would be wise?
- **Non ECC memory**
- **Tape Library**: As previously described, we have eliminated the large tape robot. We believe that simple scripts to transfer data at night from Caltech will be sufficient.,

and that with time, data transfer rates and reliability will increase. Does the committee agree that this is a satisfactory solution for obtaining data?

- **Private Network**: Is the committee aware of other networking solutions that might offer a higher-bandwidth or lower-latency connection within our design budget? Does the committee agree that it makes sense to postpone a network upgrade, if needed, for the future?