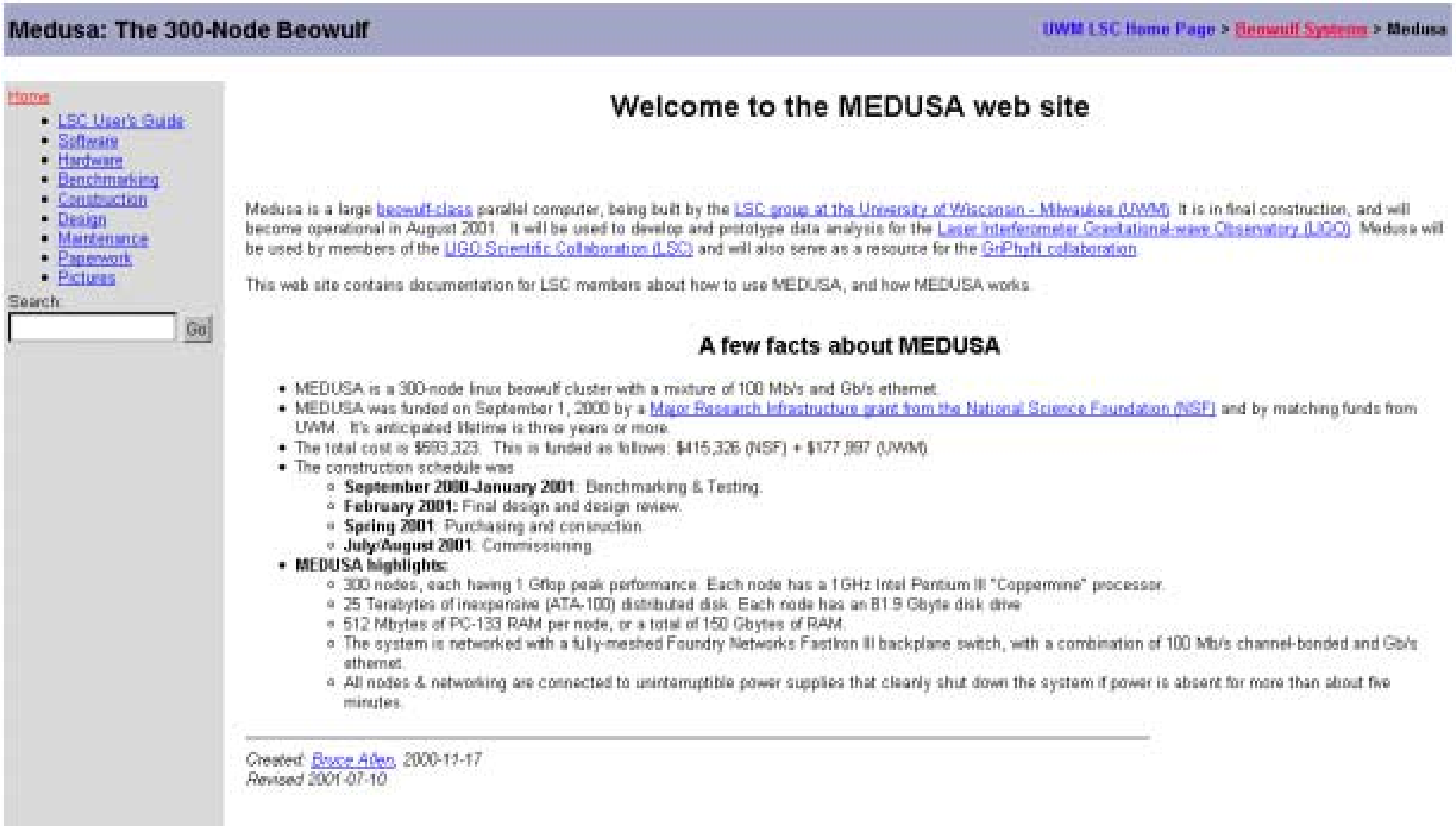
Medusa: a LSC/UWM Data Analysis Facility

University of Wisconsin - Milwaukee

LSC Meeting, August 14, 2001

LIGO-G010335-00-Z

Medusa Web Site:
www.lsc-group.phys.uwm.edu/beowulf



Medusa: The 300-Node Beowulf

UWM LSC Home Page > Beowulf Systems > Medusa

Home

- LSC User's Guide
- Software
- Hardware
- Benchmarking
- Construction
- Design
- Maintenance
- Paperwork
- Pictures

Search

Go

Welcome to the MEDUSA web site

Medusa is a large beowulf-class parallel computer, being built by the LSC group at the University of Wisconsin - Milwaukee (UWM). It is in final construction, and will become operational in August 2001. It will be used to develop and prototype data analysis for the Laser Interferometer Gravitational-wave Observatory (LIGO). Medusa will be used by members of the LIGO Scientific Collaboration (LSC) and will also serve as a resource for the GriPhyN collaboration.

This web site contains documentation for LSC members about how to use MEDUSA, and how MEDUSA works.

A few facts about MEDUSA

- MEDUSA is a 300-node linux beowulf cluster with a mixture of 100 Mb/s and Gb/s ethernet.
- MEDUSA was funded on September 1, 2000 by a Major Research Infrastructure grant from the National Science Foundation (NSF) and by matching funds from UWM. It's anticipated lifetime is three years or more.
- The total cost is $593,323. This is funded as follows: $415,326 (NSF) + $177,997 (UWM).
- The construction schedule was:
  - **September 2000-January 2001**: Benchmarking & Testing.
  - **February 2001:** Final design and design review.
  - **Spring 2001**: Purchasing and consruction.
  - **July/August 2001**: Commissioning.
- **MEDUSA highlights:**
  - 300 nodes, each having 1 Gflop peak performance. Each node has a 1GHz Intel Pentium III "Coppermine" processor.
  - 25 Terabytes of inexpensive (ATA-100) distributed disk. Each node has an 81.9 Gbyte disk drive.
  - 512 Mbytes of PC-133 RAM per node, or a total of 150 Gbytes of RAM.
  - The system is networked with a fully-meshed Foundry Networks FastIron III backplane switch, with a combination of 100 Mb/s channel-bonded and Gb/s ethernet.
  - All nodes & networking are connected to uninterruptible power supplies that cleanly shut down the system if power is absent for more than about five minutes.

Created: Bruce Allen, 2000-11-17
Revised 2001-07-10

Medusa Overview

Beowulf cluster

- •296 Gflops peak
- •150 Gbytes RAM
- •23 TBytes disk storage
- •30 Tape AIT-2 robot
- •Fully-meshed switch
- •UPS power

296 nodes, each with

- •1 GHz Pentium III
- •512 Mbytes memory
- •100baseT Ethernet
- •80 Gbyte disk

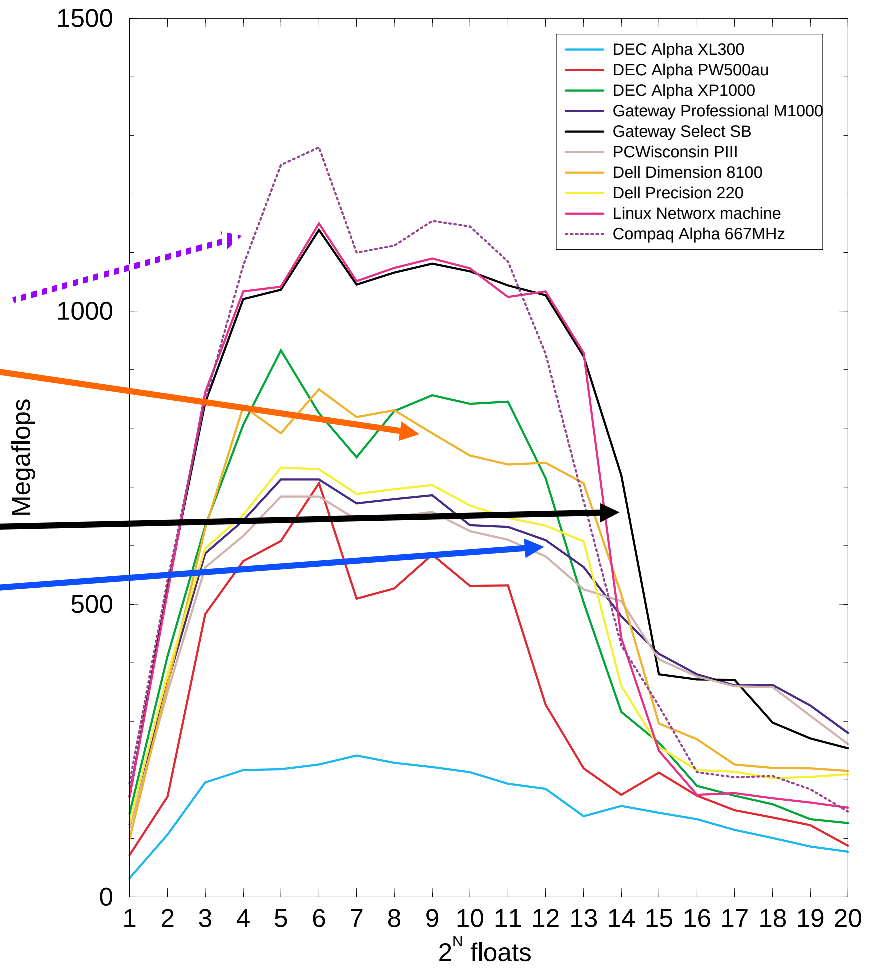
LIGO-G010335-00-Z

Medusa Design Goals

- Intended for fast, flexible data analysis prototyping, quick turn-around work, and dedicated analysis.
- Data replaceable (from LIGO archive): use inexpensive distributed disks.
- Store representative data on disk: use internet or a small tape robot to transfer it from LIGO.
- Analysis is unscheduled and flexible, since data on disks. Easy to repeat (parts of) analysis runs.
- System crashes are annoying, but not catastrophic: analysis codes can be experimental
- Opportunity to try different software environments
- Hardware reliability target: 1 month uptime

Some design details...

- Choice of processors determined by performance on FFT benchmark code
  - » AXP 21264 (expensive, slow FFTS)
  - » Pentium IV (expensive, slower than PIII on our benchmarks)
  - » Athlon Thunderbird (fast, but concerns about heat/reliability)
  - » Pentium III (fast, cheap, reliable)
- Dual CPU systems slow
- Also concerned about power budget, $$$ budget, and reliability



LIGO-G010335-00-Z

No Rackmounts

- •Saves about $250/box
- •Entirely commodity components
- •Space for extra disks, networking upgrade
- •Boxes swapable in a minute

LIGO-G010335-00-Z

Some design details...

Motherboard is an Intel D815EFV. This is a low-cost high-volume "consumer" grade system

- Real-time monitoring of CPU temperature and motherboard temperature
- Real-time monitoring of CPU fan speed and case fan speed
- Real time monitoring of 6 voltages including CPU core voltage
- Ethenet "Wake on LAN" for remote power-up of systems
- Used micro-ATX form-factor rather than ATX (3 PCI slots rather than 5) for smaller boxes.
- Lots of fans!

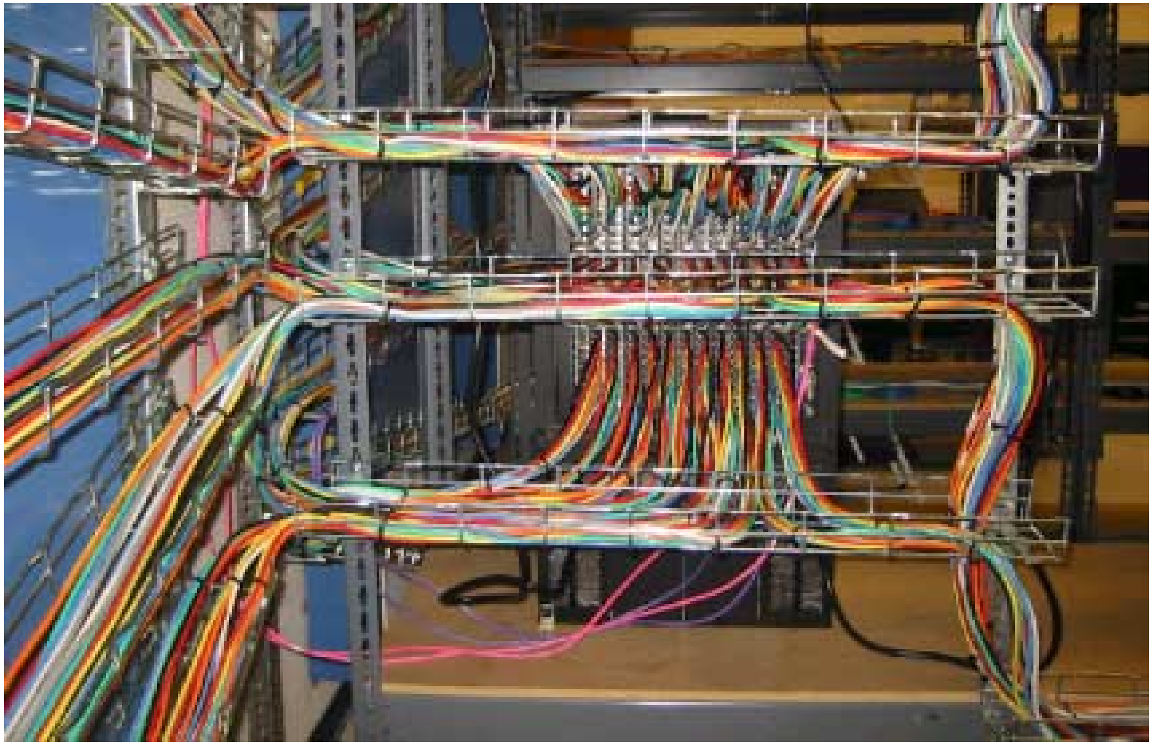
Systems are well balanced:

- memory bus transfers data at 133 MHz x 8 bytes = 1.07 GB/sec
- disks about 30 MB/sec in block mode
- ethernet about 10 MB/sec

Some design details...

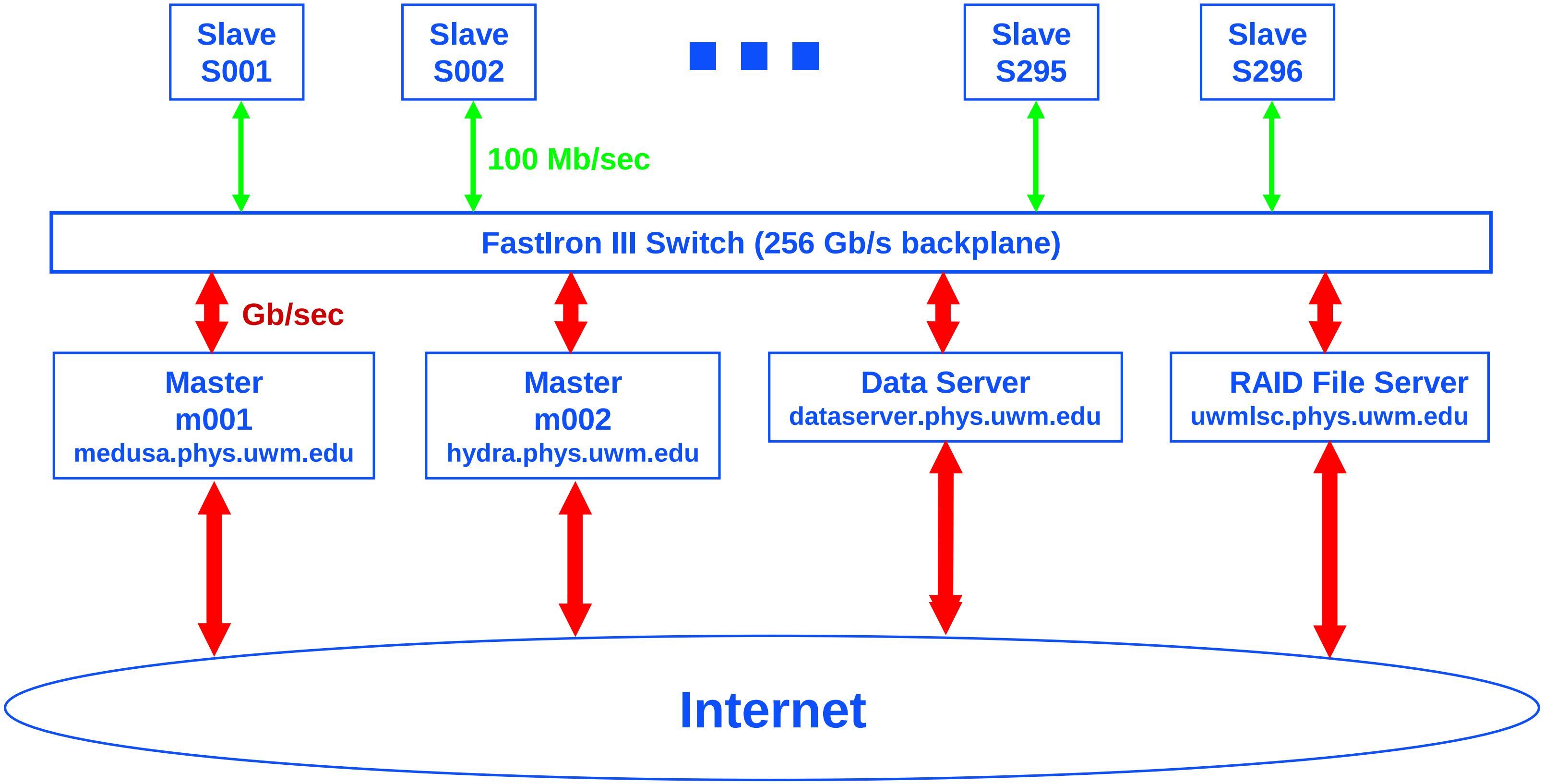
“Private” Network Switch: Foundry Networks FastIron III

- Fully-meshed
- Accomodates up to 15 blades, each of which is either 24 100TX or 8 1000TX ports
- Will also accomodate 10 Gb/s blades
- All cabling is CAT5e for potential gigabit upgrade
- 1800 W



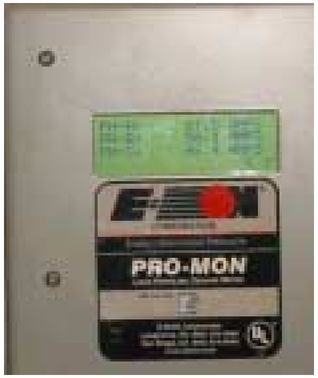
LIGO-G010335-00-Z

Networking Topology


Slave S001
Slave S002
Slave S295
Slave S296
100 Mb/sec
FastIron III Switch (256 Gb/s backplane)
Gb/sec
Master m001 medusa.phys.uwm.edu
Master m002 hydra.phys.uwm.edu
Data Server dataserver.phys.uwm.edu
RAID File Server uwmlsc.phys.uwm.edu
Internet

Cooling & Electrical

- Dedicated 5 ton air conditioner
- Dedicated 40 kVA UPS would have cost about $30k
- Instead used commodity 2250 VA UPS's for $10k
- System uses about 50 Watts/node, 18 kW total
- Three-phase power, 150 amps



System Software

- Linux 2.4.5 kernel, RH 6.2 file structure
- All software resides in a UWM CVS repository
  - » Base OS
  - » Cloning from CD & over network
  - » Nodes “interchangeable” - get identity from dhcp server on master
- Installed tools include LDAS, Condor, MPICH, LAM
- Log into any machine from any other (for example)

```
rsh s120
```

- Disks of *all* nodes automounted from all others

```
ls /net/s120/etc
cp /netdata/s290/file1 /netdata/s290/file2
```

simplifies data access, system maintenance

Memory Soft Error Rates

Cosmic rays produce random soft memory errors. Is ECC (Error Checking & Correction) memory needed? System has 9500 memory chips ~ 10^{13} transistors

- Modern SDRAM is less sensitive to cosmic-ray induced errors - so only a one inexpensive chipset (VIA 694) supports ECC, but performance hit significant (20%).
- Soft errors arising from cosmic rays well-studied, error rates measured:
  - Stacked capacitor SDRAM (95% of market) worst-case error rates ~ 2/day
  - **T**rench **I**nternal **C**harge capacitor SDRAM (5% of market) worst-case error rates 10/year, expected rates ~ 2/year
- Purchased systems with **TIC** SDRAM, no ECC

Procurement

- Used 3-week sealed bid with detailed written specification for all parts.
- Systems delivered with OS, “ready to go”.
- Nodes have a 3-year vendor warranty, with back-up manufacturers warranties on disks, CPUs, motherboards and memory.
- Spare parts closet at UWM maintained by vendor.
- 8 bids, ranging from $729/box to $1200/box
- Bid process was time-consuming, but has protected us.

Overall Hardware Budget

Nodes	$222 k
Networking switch	$ 60 k
Air conditioning	$ 30 k
Tape library	$ 15 k
RAID file server	$ 15 k
UPS’s	$ 12 k
Test machines, samples	$ 10 k
Electrical work	$ 10 k
Shelving, cabling, miscellaneous	$ 10 k
TOTAL	**$ 384k**

Remaining funds contingency: networking upgrade, larger tape robot, more powerful front-end machines?

Proposed versus Delivered

PROPOSED

- 128 nodes @ 550 MHz
  70 Gflops aggregate
- 9.4 TBytes disk
- 200 tape robot
- Two-level mix of 100baseT & gigabit

DELIVERED

- 296 nodes @ 1 GHz
  296 Gflops aggregate
- 23.7 TBytes disk
- 30 tape robot
- Single-level backplane switch with 100baseT and gigabit
- UPS systems for clean shutdown if power fails

What’s next?

- System currently in “shakedown” phase
- Some hardware delivered with dead fans, dead disks, wrong type of memory, etc. This is being corrected.
- Two UPS’s need repair.
- By the end of the month, expect system to pass burn in test (several hundred cycles of gcc **`make bootstrap`**).
- Then...start hunting in engineering data!

LSC Involvement

- MRI proposal was supported by the LIGO Lab Director and LSC Spokesman
- LIGO/LSC committee reviewed final design before purchasing/procurement phase
- In addition to UWM users, system will be available to other LSC members
  - » Support one “external” LSC user for each “internal” user
  - » Chosen 3 times/year by committee of Allen, Brady, LIGO Lab director, LSC spokesman, software coordinator
  - » If you’d like to use this system, please send me a short proposal.