



# VAMPIRE

**V**Anderbilt **M**ulti-**P**rocessor **I**ntegrated  
**R**esearch **E**ngine

Alan Tackett

Mathew Binkley

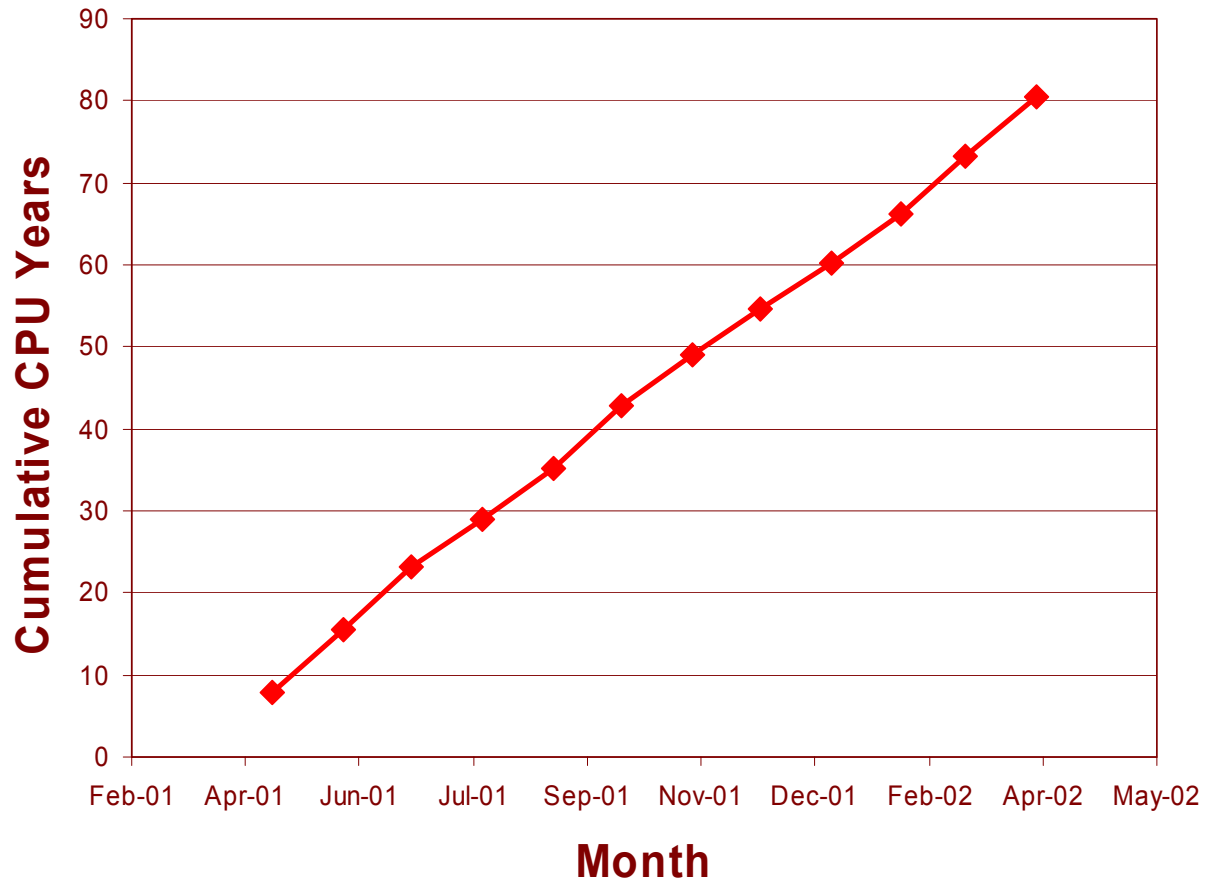
Bobby Brown

# VAMPIRE

## Vanderbilt Multi-Processor Integrated Research Engine

- VAMPIRE was conceived and created by Vanderbilt researchers in biology and physics:
  - Paul Sheldon, Physics
  - Will Johns, Physics
  - Jason Moore, Human Genetics
- It is maintained, operated, and governed by Vanderbilt faculty as a research tool responsive to the needs of their programs
- 55 compute nodes (110 processors)
  - 2U case
  - Tyan S1823DL motherboard with BX chipset
  - Dual 600MHz P3 Processors
  - 256MB memory
  - 10G disk
  - Fast ethernet
- Disk Server (700 GB)
  - 4U case with 16 hot swap IDE bays
  - SuperMicro P3TDLE motherboard with Serverworks III LE chipset
  - Dual 933MHz P3 processors
  - 1GB of memory
  - 3ware 7850 RAID5 Controller
  - Eight 100GB Maxtor DiamondMax 7200 RPM IDE drives
  - Raw sustained performance: 57MB/s for writes and 120MB/s for reads

# Cluster Usage



# VAMPIRE Users

- 70+ users
- 13 Different departments
  - Physics and Astronomy, Biostatistics, Molecular Physiology and Biophysics, Structural biology, Mechanical Engineering, Electrical Engineering and Computer Science, Biochemistry, Psychology, Pharmacology, Psychiatry, Chemistry, Microbiology
- 3 different Universities
  - Vanderbilt, Univ of Colorado Health Sciences, Meharry Medical College

# Applied Scientific Computing Class

Greg Walker (ME) and Alan Tackett (Physics)

- Purpose: Applying HPC to *actual* research projects. Not toy problems.
- Each student is working jointly with a faculty member on a current research project.
- Graduate level class with 11 students (4 A&S, 7 Engineering)
- Projects:
  - Molecular dynamics (Cummings - CE)
  - Radiation effects on Semiconductors (Weller - EECS)
  - Nuclear Physics (Oberacker, Umar, and Ernst – Physics)
  - Gene Mining (Moore, Human Genetics)
  - Brain Activation Centers (Psychiatry)
  - Web server load balancing (Barnes – EECS)
  - Elastography (Miga - BME)
  - Ion Strike (Walker - ME)

# Diverse Applications

- Run the gamut from
  - Serial jobs. But lots of them!
    - High Energy and Nuclear Physics
  - Small/medium parallel jobs requiring 2-20 CPU's
    - Requires high-performance network
    - Amber(MD, Protein), Human Genetics apps, VASP
  - Large parallel ASCI jobs using 10-512 CPU's
    - Requires high-performance network
    - Socorro(Condensed Matter Physics)
      - 16 CPU run: 600s with Fast Ethernet vs. 4 sec with Myrinet

# Software Libraries

- Because of diverse user group there is a diverse group of software installed
  - Libraries: ATLAS/BLAS, LAPACK, FFTW, PETSc, DAKOTA, Matlab, Netsolve, IBP, MPICH, PVM
  - Compilers: Multiple gcc versions supported, Intel C/C++/F95, Absoft F77/F95
- Users not capable of building these packages. In fact they may not even know they exist!
- Most need to be compiled locally for performance

# Administration

- SystemImager (<http://systemimager.org>)
  - Propagates updates, wipes and re-installs
  - Allows us to perform a complete wipe and re-install of the entire system in 30 minutes. Scalable to 1000's of nodes.
  - Only propagates changes so minor changes take just a few minutes
- Nagios(formerly Netsaint)/Ganglia
  - <http://www.nagios.org>
  - Health monitoring with automatic service restart and SysAdmin emailing and paging if needed.
- PBS/Maui
  - Batch Scheduler and Resource Manager
- Misc Scripts
  - Pexec – Parallel execution and file copy written locally



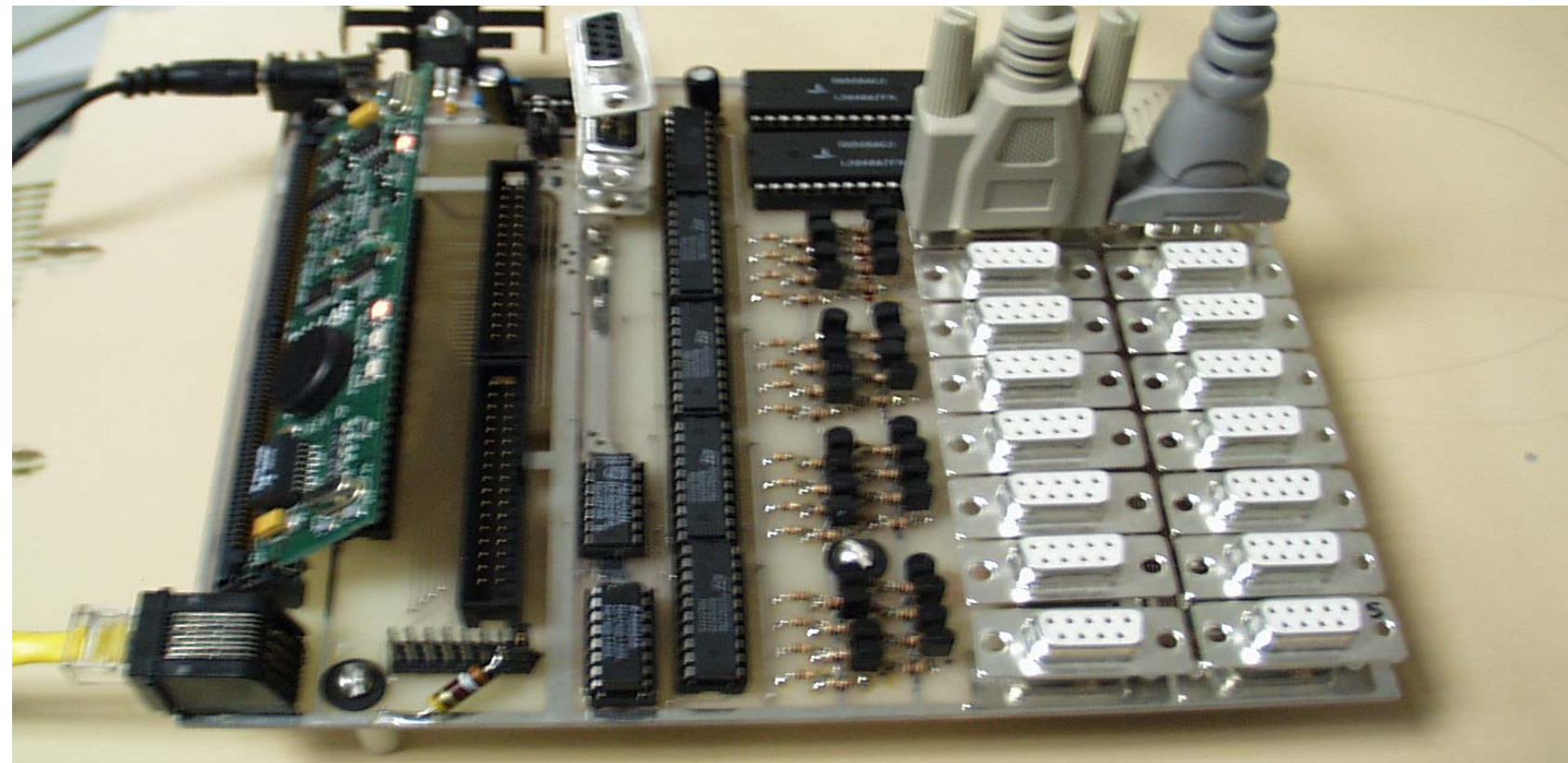
# Resource Sharing with Maui<sup>1</sup>

- Provides each group *on average* their appropriate fair share of the cluster
- Supports advanced reservations,
  - Serial and parallel jobs,
  - Node attributes for special hardware or apps
- Configurable Job priority based on
  - Group, user, account, QoS, number of CPU's, execution time, etc.
- Shortpool Queue for interactive debugging of jobs and short jobs
- Showbf command

<sup>1</sup><http://www.supercluster.org/>

# Remote Access Control

- Terminal Server designed by Will Johns
  - Cost \$10/port
- Supports Power On/Off and Reboot
- Based on TINI board
- Still need to provide finishing touches on interface



# Proposed Scientific Computing Center (SCC)

- Centralized computational Cluster (1000 nodes) for Large-scale applications
- Smaller specialized clusters located in faculty research labs
- 50TB disk array using a parallel file system
- Expandable backup facility capable of handling Petabytes of data

# Which CPU? P4 or Athlon

- Athlon 10-20% faster on some apps
  - Best for apps compiled for P3, integer math, or compiled using g77
- P4 2x faster on apps making heavy use of double precision BLAS
  - SSE2 provides SIMD instructions for double precision
- Chipset also makes a difference

# Application Benchmarks

<b>Amber</b>	<b>Secs</b>		<b>Memory Bandwidth (MB/s)</b>	<b>L1</b>	<b>L2</b>	<b>Main</b>
P3-1GHz/LE / g77-2.96	2364		P4-1.8/GC	14727	12564	1049
300MHz O2k	2118		P4-1.8/E7500	14720	12552	1114
P4-1.8/E7500 / g77-2.96	2063		Athlon-1.8	10992	3648	864
P4-1.8/E7500 / g77-3.2	1772		P3-1GHz / 440BX	9770	4429	309
P4-1.8/GC / g77-3.2	1743					
Athlon-1.4 / g77-2.96	1556					
Athlon-1.8 / g77-2.96	1210					
P4-1.8/E7500 / ifc6	805					
P4-1.8/GC / ifc6	773					
<b>Socorro</b>	<b>Secs</b>					
P4-1.8/E7500 / ifc6	1689					
P4-1.8/GC / ifc6	1209					

# Disk Benchmarks

## Disk Server Configuration

- Dual P4 system
- 16 Maxtor 160GB IDE
- Dual 3ware 7850 controllers
- Each controller is configured as RAID5 with software Striping
- 2.25TB of disk space
- XFS File system

## Bonnie++

Version 1.02c

-----Sequential Output-----

--Sequential Input- --Random-

-Per Chr- --Block-- -Rewrite-

-Per Chr- --Block-- --Seeks--

Machine Size K/sec %CP K/sec %CP K/sec %CP

K/sec %CP K/sec %CP /sec %CP

Inner Mount 4G 20851 88 **83647** 27 82919 37

23294 97 **246939** 52 270.5 1



# Immediate Expansion Plans

Grass roots effort with all money from individual researchers

- 110TB Backup Server (just arrived)
  - Quantum ATL P7000 Tape Library
  - SDLT 320 Drives
  - Expandable to over 500TB
  - Supporters from Psychiatry, Physics, Ingram Cancer, Human Genetics, Struct. Biology, Medical Imaging Center




- ~200 additional VAMPIRE compute nodes (IBM x335)
  - Serverworks GC-LE Chipset (>25% faster than Intel E7500)
  - Dual 2.0GHz P4 Xeon
  - 1G of ECC DDR-266 memory
  - Myrinet 2000
  - Dual Gigabit Ethernet
  - Supporters from EECS, BME, CE, ME, and Physics

Intel processor-based servers

**xSeries** 335 overview

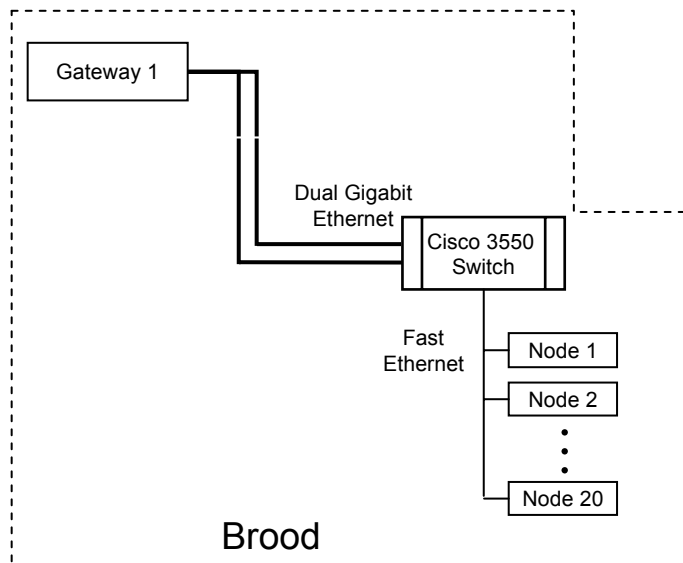


- 2 small compute clusters
  - M. Miga (BME) and J. Moore (Human Genetics)

IBM server xSeries

# Brood

## Fundamental Building Block



- Brood Configuration

- Gateway
- Switch
- 20 or more compute nodes

- Gateway responsible for

- Health monitoring
- Updates and Installs
- Compute Nodes DHCP service
- Exporting of /usr/local to nodes

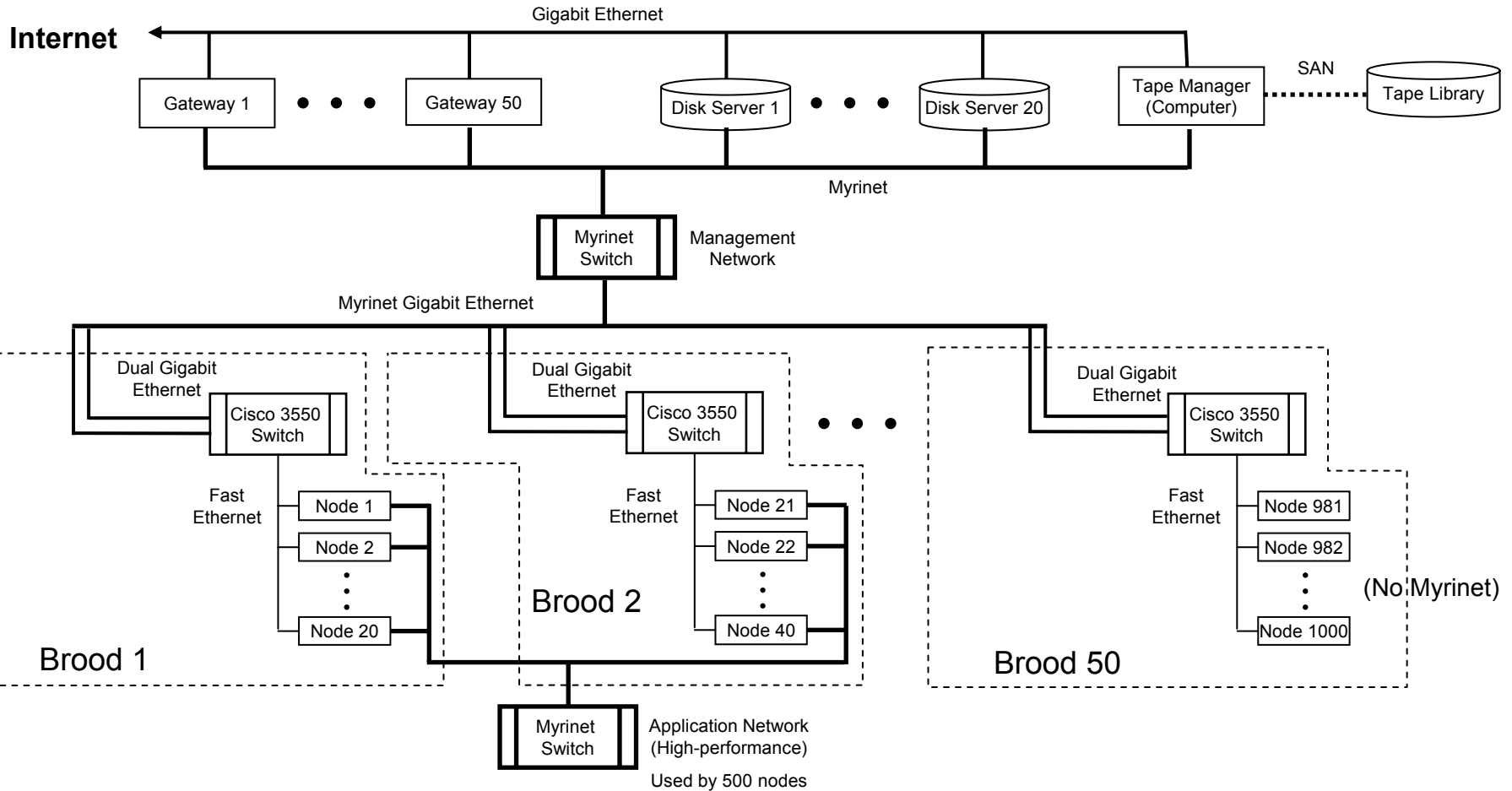
- Brood Flexibility

- Complete Mini-Cluster
- Can be segregated from the main cluster for users specialized needs.
  - Testing special hardware, kernels, different OS's, apps
- Easily reintegrated with larger cluster using SystemImager



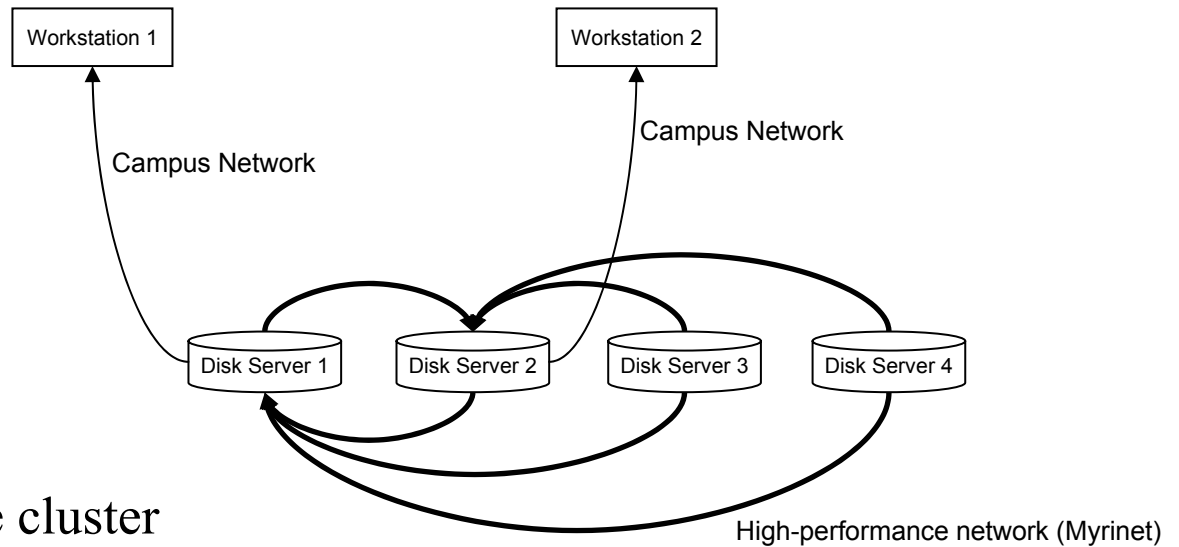
# Scientific Computing Center using Myrinet + GigE

## Gateways and Disk Servers also function as Management Nodes

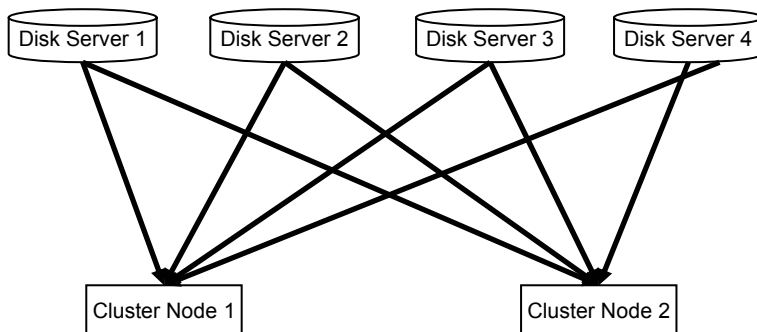


# Parallel Filesystem Access

Parallel I/O outside the cluster using load-balancing



Parallel I/O within the cluster



# Summary

- VAMPIRE is a multidisciplinary cluster
- Currently adding 110TB tape backup and ~200 compute nodes
- Future expansion next year to add an additional 1000 nodes and 50TB of disk space.