Cluster-NFS: Simplifying Linux Clusters

Gregory R. Warnes, Ph.C.

Fred Hutchinson Cancer Research Center

May 11, 1999

The cNFS Mascot, Henry the Hydra

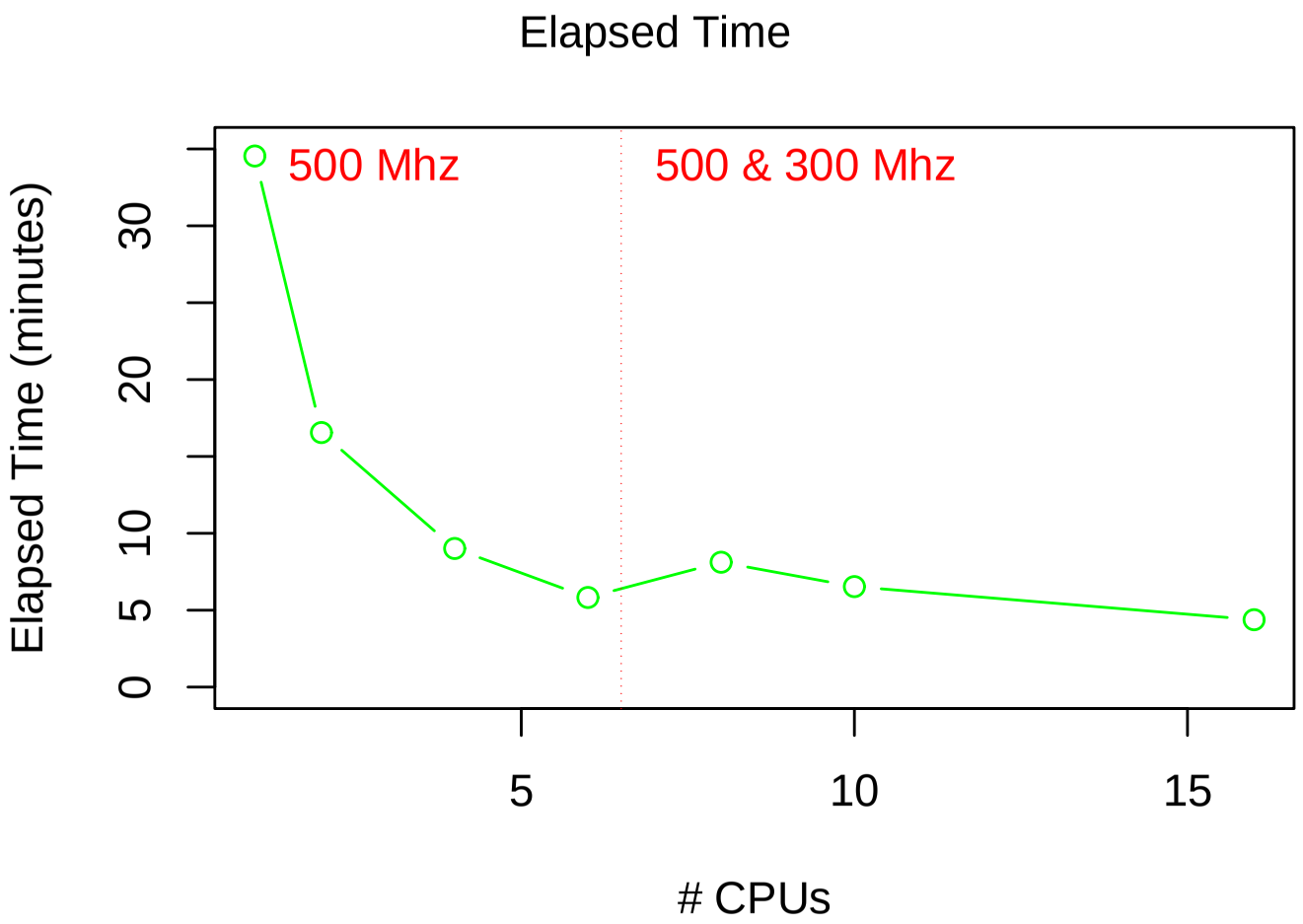
Outline

1. Why Linux Clusters?
   - High Performance + Low Cost
2. What Makes Clusters Hard?
   - Admins: Maintenance!
   - Users: Distributing Tasks
3. The Tools
   - MOSIX: Easier for Users
   - Cluster-NFS: Easier for Admins
4. MOSIX + Cluster-NFS in Action: The BioHive Cluster
5. Cluster Recipe
6. Ideas and Future Plans

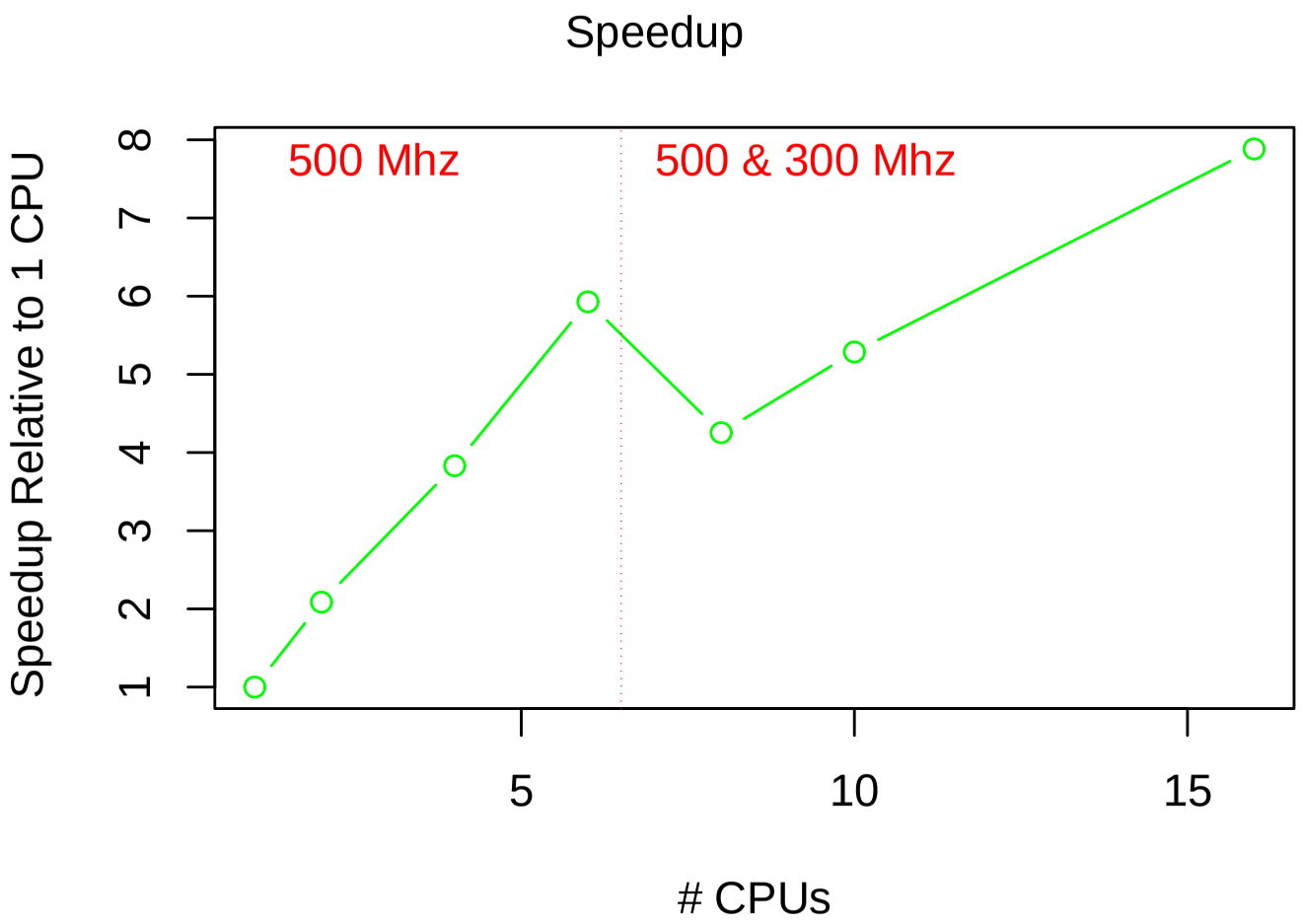
Why Use Linux Clusters?

1. High performance
   - Close to 1:1 speedup (modulo CPU speed differences) for our parallel application.
   - Perfect 1:1 speedup for batches of independent simulations
2. Low Cost
   - Diskless Dual Celeron-500: $800/ea
   - Diskless Athlon-850: $1000/ea
3. Relatively Easy to Build and Maintain

Why Linux Clusters? High performance



Why Linux Clusters? High performance



Why Linux Clusters? Low Cost

Home | Company Info | Contact | News Releases | Support

HARD DRIVES NORTHWEST
"Rising High Above the Competition"

GO DIRECT! Built Exactly How You Want It! 1-888-220-8964

- View My Order!

Departments:
- Build Your Own System
- Hard Drives
- Home Computers
- Business Computers
- Notebooks

Components:
- Auctions
- Cases
- CD-Recordable Drives
- CD/DVD-ROMs
- Controllers
- CPUs
- Digital Cameras
- External HDD Enclosures
- Floppy Drives
- Keyboards
- Media
- Memory
- Mice
- Modems
- Monitors - Flat Panel
- Monitors 14" - 17"
- Monitors 19"+
- Motherboards
- Network Cards (NIC's)
- Networking Hubs
- Operating Systems
- PDA'S
- Power Protection
- Printers
- Removable Mass Media
- Scanners
- Services
- Software
- Sound Cards
- Speakers
- Tape Backup Devices
- Video Cards
- Viewsonic

Build A Computer

Description		Price
Intel Celeron 500 PPGA 66MHz Socket 370	Qty = 2	$238.00
1.44 Floppy		$14.00
Abit BP6 ATX 100MHz Dual Socket-370 Motherboard		$147.00
256mb PC-133 SDRAM	Qty = 1	$239.00
KME Jumbo-Mini Tower ATX		$42.00
Aopen SiS 6326 4mb AGP		$27.00
System Assembly		$29.00
3COM Fast Etherlink XL 3C905B-TXNM PCI 10/100 Mpbs		$53.00
	TOTAL	$789.00

Add To Order Return to Build A Computer

[Build A Computer] [Home Computers] [Business Computers] [Laptops] [Hard Drives] [Components] [View my basket] [Service] [About HDNW] [News] [Home]

©1997-2000 Hard Drives Northwest, Inc. All rights reserved.

What Makes Clusters Hard?

1. Setup - Administrator
   - Setting up a 4 node cluster by hand is pretty easy.
   - Setting up a 16 node cluster by hand is mind-numbing and prone to errors.
2. Maintenance - Administrator
   - Ever tried to update a package on every node in the cluster?
   - How about 3 configuration files?
   - How do you know if you missed one machine?
3. Running Jobs - Users
   - Running a parallel program or a set of sequential programs requires that the users figure out what hosts are available and manually assign tasks to nodes.
   - Users usually don’t want to see this much detail.

The Tools: MOSIX

Description: MOSIX is an enhancement to the Linux kernel that provides adaptive (on-line) load-balancing and memory ushering between x86 Linux machines. It uses preemptive process migration to assign and reassign the processes among the nodes to take best advantage of the available resources.

Translation: MOSIX moves processes around the cluster to balance the load, using faster machines first.

Source: Amnon Barak, CS Department of the Hebrew University of Jerusalem

URL: `http://www.mosix.cs.huji.ac.il`

The Tools: Cluster-NFS (cNFS)

Description: cNFS is a patch to the standard Universal-NFS server (uNFS) code that “parses” file requests to determine an appropriate match on the server.

Whenever a client requests the file `filename`, the server check for the existence of one of the following files, returning the first match:

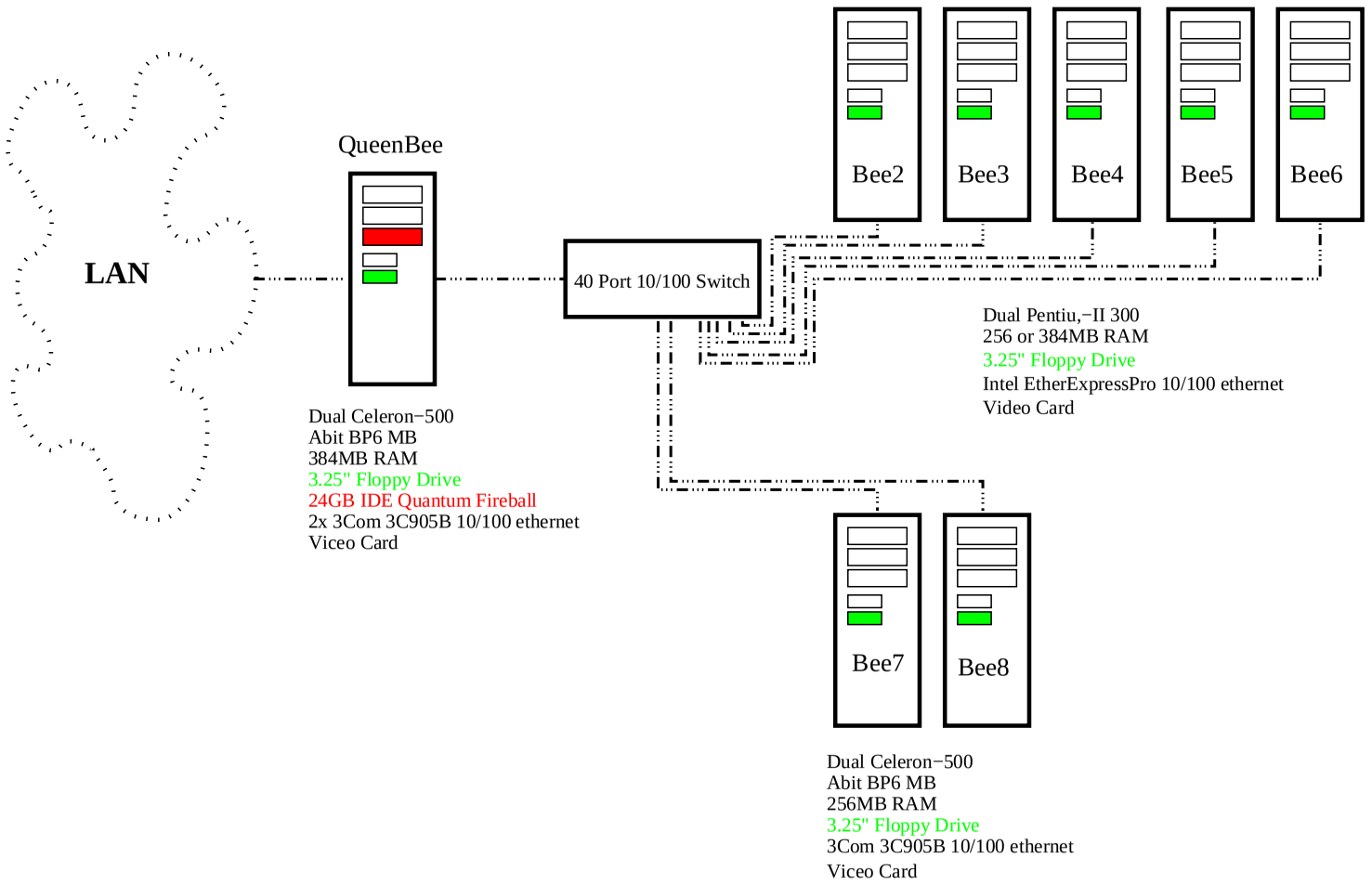
`filename$$UID=xxxx$$`	user’s id
`filename$$GID=xxxx$$`	user’s group id
`filename$$HOSTNAME=ssss$$`	client hostname
`filename$$IP=xxx.xxx.xxx.xxx$$`	client ip number
`filename$$CLIENT$$`	always matches
`filename`	default

Example: When client machine `bee3` asks for file `/etc/hostname` it gets the contents of `/etc/hostname$$HOST=bee3$$`.

Source: Gregory Warnes, Fred Hutchinson Cancer Research Center

URL: `http://queenbee.fhcrc.org/ClusterNFS/`

MOSIX + ClusterNFS in Action: the BioHive Cluster



http://queenbee.fhcrc.org

Making Clusters Easy : MOSIX + Cluster-NFS

1. Setup - Administrator
   - Setup server
   - Compile rootNFS kernel. Make floppies.
   - Plug in switch
   - Plug in nodes. Insert Floppy.
   - Boot.
2. Maintenance - Administrator
   - Changes made to server immediately take effect on all clients.
   - Adding a node requires changing or copying 8 files and making a bootdisk.
3. Running Jobs - Users
   - Users log into a “master” node, MOSIX distributes tasks.

Making Clusters Easy for Users: MOSIX

MOSIX (`http://www.mosix.cs.huji.ac.il`) is a dynamic load-balancing system that transparently migrates tasks between machines.

1. Users log into “master” node
2. Jobs started on the master node *automagically* migrate to fastest / least loaded machine.
   - Parallel jobs need not specify nodes
   - Sequential jobs started as if on SMP
3. Job Control (ps, top, kill) occurs as if whole cluster is one system

Users never need to know details of cluster configuration.

Diskless Servers: Traditional Method

Server:

- BOOTP server
- NFS server
- Separate root directory for each client

Client:

- BOOTP to obtain IP
- TFTP or boot floppy to load kernel
- rootNFS to load root file system

Diskless Servers: Traditional Method

This method requires a separate root directory structure for each node.

Hard to Set Up

- Lots of directories with *slightly* different contents.
- Even with symlinks this gets messy fast.

Difficult to Maintain

- Changes must be propogated to each directory.
- No easy way to see what differs between directories.

Diskless Servers: Cluster-NFS Method

Server:

- BOOTP server
- **Cluster-NFS server**
- **Single root directory for server and clients**

Client:

- BOOTP to obtain IP
- TFTP or boot floppy to load kernel
- rootNFS to load root file system

Diskless Servers: Cluster-NFS Method

Cluster-NFS allows all machines (including server) to share the root filesystem

- All files are shared by default.
- Files for all clients are named `filename$$CLIENT$$`
- Files for specific clients are named `filename$$IP=xxx.xxx.xxx.xxx$$` or `filename$$HOST=host.subdomain.domain$$.`

Diskless Servers: Cluster-NFS Method

- Easy to set up

  Just copy/create the files that need to be different.

- Easy to maintain
  - Changes to shared files are global.
  - Easy to make customizations.
  - Easy to look for customizations: `find / -name "*\$\$*\$\$"`
  - Easy to add nodes, add node to 4 server files and create 7 machine-specific files.

Cluster-NFS Recipe

On the Server

1. Install and configure Debian Linux
2. Install Cluster-NFS
3. Download and Compile MOSIX and Kernel, enabling BOOTP and RootNFS.
4. Copy the Kernel to Floppies
5. Add entries for each client to
   - `/etc/hosts`,
   - `/etc/mosix.map`,
   - `/etc/bootptab`,
   - `/etc/exports`, and
   - `/etc/hosts.allow`.
6. Create files that are the same for all clients, `filename$$CLIENT$$`.

7. Create files that are specific to individual clients

   `filename$$IP=xxx.xxx.xxx.xxx$$`

8. reboot server to restart all services

Cluster-NFS Recipe

On the Client

1. Insert boot floppy
2. Boot
3. Record Ethernet MAC address displayed by kernel
4. Add to `\etc\bootptab` on server
5. Reboot

Plans and Ideas

- Need to write this up as a paper.

Wanted: Volonteer to do the writing in exchange for co-authorship.

- Instant Cluster:

Pile of Windows Desktops

Pile of Boot Floppies

\+ 1 Linux Server running ClusterNFS

Instant Linux Cluster

No configuration on the client!

Plans and Ideas

- Use of DHCP instead of BOOTP?
- Auto-configuration: Unrecognized MAC address causes server to
  1. Assign a new IP number and hostname
  2. Add appropriate entry to `/etc/bootptab`
  3. Create new machine specific files using template scripts, say `filename$$TEMPLATE-SCRIPT$$`,
- ??