

Linux Clusters: Details and Case Studies

Jim Phillips and Tim Skirvin



What is your situation?

- Who are the users?
- What application(s) will they run?
- Faster turnaround or higher throughput?
- How much am I willing to spend?
- Where can I put the machines?
- Who can administer the machines?

User Rules of Thumb

- 1-4 users:
 - Yes, you still want a queueing system.
 - Plan ahead to avoid idle time and conflicts.
- 5-20 users:
 - Put one person in charge of running things.
 - Work out a fair-share or reservation system.
- > 20 users:
 - User documentation and examples are essential.
 - Decide who makes resource allocation decisions.

Application Rules of Thumb

- 1-2 programs:
 - Don't pay for anything you won't use.
 - Benchmark, benchmark, benchmark!
 - Be sure to use your typical data.
 - Try different compilers and compiler options.
- > 2 programs:
 - Select the most standard OS environment.
 - Benchmark those that will run the most.
 - Consider a specialized cluster for dominant apps only.

Parallelization Rules of Thumb

- Throughput is easy...app runs as is.
- Turnaround is not:
 - Parallel speedup is limited by:
 - Time spent in non-parallel code.
 - Time spent waiting for data from the network.
 - Improve serial performance first:
 - Profile to find most time-consuming functions.
 - Try new algorithms, libraries, hand tuning.

Budget Rules of Thumb

- \$2K to \$20K: Desktop PCs on shelves, 24-port gigabit switch, \$700/CPU.
- \$20K to \$50K: Dual-CPU rackmount, 24-port gigabit switch, \$1000/CPU.
- > \$50K:
 - Single large gigabit cluster for throughput.
 - Myrinet or Infiniband for turnaround, \$2000/CPU.
 - Consider multiple 24-node gigabit clusters.

Environment Rules of Thumb

- 12 CPUs per 20A 110V circuit
 - In a rack, 24 CPUs per 20A 208V circuit
- 20 CPUs per ton of air conditioning
- Buy a Kill-A-Watt for \$30 and measure!
 - $\text{Watts} = \text{Amps} * \text{Volts} * \text{Power Factor}$
 - Run something intense like “cpuburn”:
 - Nodes draw 50% more current under load.
 - Poorly cooled or unstable machines crash!

SysAdmin Rules of Thumb

- Automate everything you can:
 - Small differences are a pain to debug.
 - Use install/setup scripts for the head node too.
- Limit root access:
 - A little knowledge is a dangerous thing.
 - Have one or two trusted backup admins.
 - In a medical emergency you call 911 and follow instructions until the paramedics arrive, right?
 - Post numbers call in an emergency. A call at 3am is better than finding a broken cluster at 9am.

Some Details Matter More

- What limiting factor do you hit first?
 - Budget?
 - Space, power, and cooling?
 - Network speed?
 - Memory speed?
 - Processor speed?
 - Expertise?

Limited by Budget

- Don't waste money solving problems you can't afford to have right now:
 - Regular PCs on shelves (rolling carts)
 - Gigabit networking and multiple jobs
- Benchmark performance per dollar.
 - The last dollar you spend should be on whatever improves your performance.
- Ask for equipment funds in proposals!

Limited by Space

- Benchmark performance per rack
- Consider all combinations of:
 - Rackmount nodes
 - More expensive but no performance loss
 - Dual-processor nodes
 - Less memory bandwidth per processor
 - Dual-core processors
 - Less memory bandwidth per core

Extreme Space Issues

- Blade servers
 - Proprietary and expensive
- Cooling integrated into the rack
 - CSE Turing cluster uses these
 - Expensive
 - Requires plumbing
 - Makes me nervous

Limited by Power/Cooling

- Benchmark performance per Watt
- Consider:
 - Opteron or PowerPC rather than Xeon
 - Dual-processor nodes
 - Dual-core processors

Extreme Power Issues

- Orion Multisystems deskside cluster
 - Proprietary and expensive
 - 96 Transmeta processors draw 15A
 - 1/8 the power of a normal CPU
 - 1/3 the performance of normal CPU
 - Must scale better to run at same speed
 - Same performance per Watt from dual core?
 - www.orionmulti.com

Limited by Network Speed

- Benchmark your code at NCSA.
 - 10,000 CPU-hours is easy to get.
 - Try running one process per node.
 - If that works, buy single-processor nodes.
 - Try Myrinet.
 - If that works, can you run at NCSA?
 - Can you run more, smaller jobs?

Extreme Network Issues

- Three main choices:
 - Myrinet...proprietary but well established
 - Infiniband...multi-vendor but new
 - 10 Gigabit Ethernet...very new
- Consider
 - Fewer nodes with more CPUs/cores
 - The opinions of those with experience

Limited by Serial Performance

- Is it memory performance? Try:
 - Single-core Opterons
 - Single-processor nodes
 - Larger cache CPUs
 - Lower clock speed CPUs
- Is it really the processor itself? Try:
 - Higher clock speed CPUs
 - Dual-core CPUs

Limited by Expertise

- There is no substitute for a local expert.
- Qualifications:
 - Comfortable with the Unix command line.
 - Comfortable with Linux administration.
 - Cluster experience if you can get it.

Install It Yourself

- Don't use the vendor's pre-loaded OS.
 - They would love to sell you 100 licenses.
 - What happens when you have to reinstall?
 - Do you like talking to tech support?
 - Are those flashy graphics really useful?
 - How many security holes are there?

Purchasing Tips: Before You Begin

- Get your budget
- Work out the space, power, and cooling capacities of the room.
- Start talking to vendors early
 - But don't commit!
- Don't fall in love with any one vendor until you've looked at them all.

Purchasing Tips: Design Notes

- Make sure to order some spare nodes
 - Serial nodes and hot-swap spares
 - Keep them running to make sure they work.
- If possible, install HDs only in head node
 - State law and UIUC policy requires all hard drives to be wiped before disposal
 - It doesn't matter if the drive never stored anything!
 - Each drive will take 8-10 hours to wipe.
 - Save yourself a world of pain in a few years...
 - ...or just give your machines to some other campus group, and make them worry about it.

Purchasing Tips: Get Local Service

- If a node dies, do you want to ship it?
- Two choices:
 - Local business (Champaign Computer)
 - Major vendor (Sun)
- Ask others about responsiveness.
- Design your cluster so that you can still run jobs if a couple of nodes are down.

Purchasing Tips: Dealing with Purchasing

- You will want to put the cluster order on a Purchase Order (PO)
 - Do not pay for the cluster until it entirely works.
- Prepare a ten-point letter
 - Necessary for all purchases >\$25k.
 - Examples are available with your business office (or bug us for our examples).
 - These aren't difficult to write, but will probably be necessary.

Purchasing Tips: The Bid Process

- Any purchase >\$28k must go up for bid
 - Exception: sole-source vendors
 - Number grows every year
 - Adds a month or so to the purchase time
 - If you can keep the numbers below the magic \$28k, do it!
 - The bid limit may be leverage for vendors to drop their prices just below the limit; plan accordingly.
- You will get lots of junk bids
 - Be very specific about your requirements to keep them away!

Purchasing Tips: Working the Bid Process

- Use sole-source vendors where possible.
 - This is a major reason why we buy from Sun.
 - Check with your purchasing people.
 - This won't help you get around the month time loss, as the item still has to be posted.
- Purchase your clusters in small chunks
 - Only works if you're looking at a relatively small cluster.
 - Again, you may be able to use this as leverage with your vendor to lower their prices.

Purchasing Tips: Receiving Your Equipment

- Let Receiving know that the machines are coming.
 - It will take up a lot of space on the loading dock.
 - Working with them to save space will earn you good will (and faster turnaround).
 - Take your machines out of their space as soon as reasonably possible.

Purchasing Tips: Consolidated Inventory

- Try to convince your Inventory workers to tag each cluster, and not each machine
 - It's really going to be running as a cluster anyway (right?).
 - This will make life easier on you.
 - Repairs are easier when you don't have to worry about inventory stickers
 - This will make life easier for them.
 - 3 items to track instead of 72

Purchasing Tips: Assembly

- Get extra help for assembly
 - It's reasonably fun work
 - ...as long as the assembly line goes fast.
 - Demand pizza.
- Test the assembly instructions before you begin
 - Nothing is more annoying than having to realign all of the rails after they're all screwed in.

Purchasing Tips: Testing and Benchmarking

- Test the cluster before you put it into production!
 - Sample jobs + cpuburn
 - Look at power consumption
 - Test for dead nodes
- Remember: vendors lie!
 - Even their demo applications may not work; check for yourself.

Security Tips

- Restrict physical access to the cluster, if possible.
 - Make sure you're involved in all tours, to make sure nobody touches anything.
- If you're on campus, put your clusters into the Fully Closed network group
 - Might cause some limitations if you're trying to submit from off-site
 - Will cause problems with GLOBUS
 - The built-in firewall is your friend!

Case Studies

- The best way to illustrate cluster design is to look at how somebody else has done it.
 - The TCB Group has designed four separate Linux clusters in the last six years

2001 Case Study

- Users:
 - Many researchers with MD simulations
 - Need to supplement time on supercomputers
- Application:
 - Not memory-bound, runs well on IA32
 - Scales to 32 CPUs with 100Mbps Ethernet
 - Scales to 100+ CPUs with Myrinet

2001 Case Study 2

- Budget:
 - Initially \$20K, eventually grew to \$100K
- Environment:
 - Full machine room, slowly clear out space
 - Under-utilized 12kVA UPS, staff electrician
 - 3 ton chilled water air conditioner (Liebert)

2001 Case Study 3

- Hardware:
 - Fastest AMD Athon CPUs available (1333 MHz).
 - Fast CL2 SDRAM, but not DDR.
 - Switched 100Mbps Ethernet, Intel EEPro cards.
 - Small 40 GB hard drives and CD-ROMs.
- System Software:
 - Scyld clusters of 32 machines, 1 job/cluster.
 - Existing DQS, NIS, NFS, etc. infrastructure.

2003 Case Study

- What changed since 2001:
 - 50% increase in processor speed
 - 50% increase in NAMD serial performance
 - Improved stability of SMP Linux kernel
 - Inexpensive gigabit cards and 24-port switches
 - Nearly full machine room and power supply
 - Popularity of compact form factor cases
 - Emphasis on interactive MD of small systems

2003 Case Study 2

- Budget:
 - Initially \$65K, eventually grew to ~\$100K
- Environment:
 - Same general machine room environment
 - Additional machine room space is available in server room
 - Just switched to using rack-mount equipment
 - Still using the old clusters; don't want to get rid of them entirely
 - Need to be more space-conscious

2003 Case Study 3

- Option #1:
 - Single processor, small form factor nodes.
 - Hyperthreaded Pentium 4 processors.
 - 32 bit 33 MHz gigabit network cards.
 - 24 port gigabit switch (24-processor clusters).
- Problems:
 - No ECC memory.
 - Limited network performance.
 - Too small for next-generation video cards.

2003 Case Study 4

- Final decision:
 - Dual Athlon MP 2600+ in normal cases.
 - No hard drives or CD-ROMs.
 - 64 bit 66 MHz gigabit network cards.
 - 24 port gigabit switch (48-proc clusters).
 - Clustermatic OS, boot slaves off of floppy.
 - Floppies have proven very unreliable, especially when left in the drives.
- Benefits:
 - Server class hardware w/ ECC memory.
 - Maximum processor count for large simulations.
 - Maximum network bandwidth for small simulations.



2003 Case Study 5

- Athlon clusters from 2001 recycled:
 - 36 nodes outfitted as desktops
 - Added video cards, hard drives, extra RAM
 - Cost: ~\$300/machine
 - Now dead or in 16-node Condor test cluster
 - 32 nodes donated to another group
 - Remaining nodes move to server room
 - 16-node Clustermatic cluster (used by guests)
 - 12 spares and build/test boxes for developers

2004 Case Study

- What changed since 2003:
 - Technologically, not much!
 - Space is more of an issue.
 - A new machine room has been built for us.
 - Vendors are desperate to sell systems at any price.

2004 Case Study 2

- Budget:
 - Initially ~\$130K, eventually grew to ~\$180K
- Environment:
 - New machine room will store the new clusters.
 - Two five-ton Liebert air conditioners have been installed.
 - There is minimal floor space, enough for four racks of equipment.

2004 Case Study 3

- Final decision:
 - 72x Sun V60x rack-mount servers.
 - Dual 3.06GHz Intel processors – only slightly faster
 - 2GB RAM, Dual 36GB HDs, DVD-ROM included in deal
 - Network-bootable gigabit ethernet built in
 - Significantly more stable than any old cluster machine
 - 3x 24 port gigabit switch (3x 48-processor clusters)
 - 6x serial nodes (identical to above, also serve as spares)
 - Sun Rack 900-38
 - 26 systems per rack, plus switch and UPS for head nodes
 - Clustermatic 4 on RedHat 9

2004 Case Study 4

- Benefits:
 - Improved stability over old clusters.
 - Management is significantly easier with Sun servers than PC whiteboxes.
 - Network booting of slaves allows lights-off management.
 - Systems use up minimal floor space.
 - Similar performance to 2003 allows all 6 clusters (3 old + 3 new) to take jobs from a single queue.
 - Less likely to run out of memory when running an “express queue” job.
 - Complete machines easily retasked.

