How do Design a Cluster

Dana Brunson

Asst. VP for Research Cyberinfrastructure

Director, High Performance Computing Center

Adjunct Assoc. Professor, CS & Math Depts.

Oklahoma State University

http://hpcc.okstate.edu

It depends.

-- Henry Neeman

What is a Cluster?

"… [W]hat a ship is … It's not just a keel and hull and a deck and sails. That's what a ship needs. But what a ship is ... is freedom."

– Captain Jack Sparrow

"Pirates of the Caribbean"

Credit: Henry Neeman

What a Cluster is…

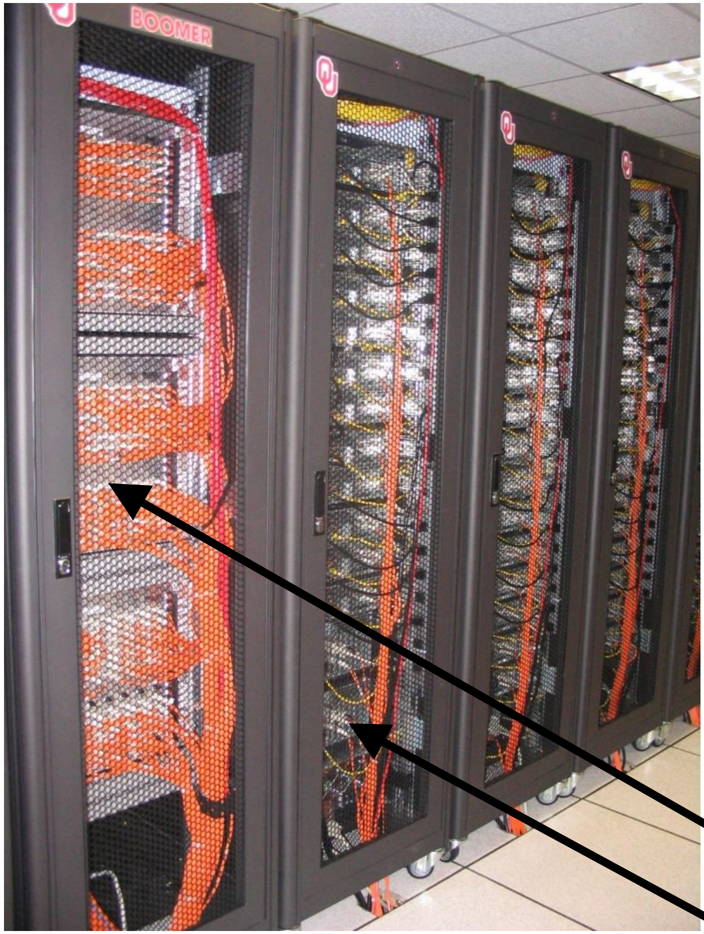
A cluster **needs** of a collection of small computers, called ***nodes***, hooked together by an ***interconnection network*** (or ***interconnect*** for short).

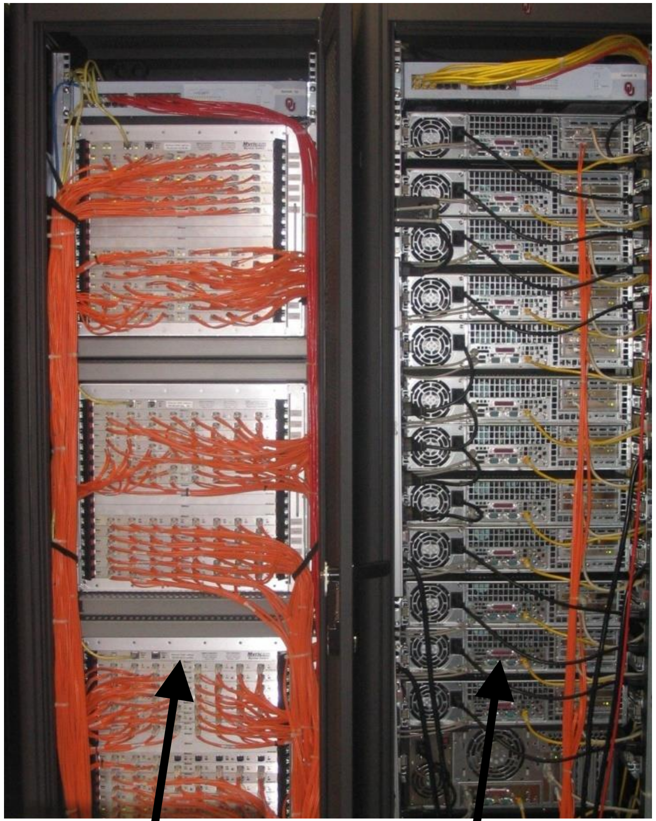
It also **needs** software that allows the nodes to communicate over the interconnect.

But what a cluster **is** … is all of these components working together as if they're one big computer ... a **super** computer.

Credit: Henry Neeman

An Actual Cluster





Interconnect

Nodes

Also named Boomer, in service 2002-5.

Credit: Henry Neeman

Considerations

- Your budget
- Power, space, and cooling
- Your researchers’ workload
- Your funding cycle
- Your staff
- What else?

OSU’s considerations

- Budget ~ $1.3M
- Building out new power & cooling
- Workload is mix of single core, shared memory and jobs up to ~512 cores
- Funding cycle is one big purchase every 4-ish years (thanks NSF!)
- Staff = 1 dedicated person for combined sysadmin, user support, application installation.

Components overview

- Compute Nodes (standard, large memory, accelerated)
- Storage (slow & fast)
- Interconnects (slow & fast)
- Login nodes
- Management nodes
- Other? (Data transfer node, web interfaces, etc.)

Compute nodes

- “Standard” compute nodes
  - Processor type
  - RAM (speed, # channels to fill)
  - Cheap as possible and still make users happy
- Special compute nodes
  - Large memory
  - Accelerators

Choices depends on your users’ needs and the sweet spot in pricing.

OSU's latest basic outline: Compute nodes

Compute Nodes: (depends on sweet spot of pricing)

- Standard: processor 2620 or better, 32-64 GB RAM
- Large memory: One 1 TB RAM, 4x 256 GB
- GPU nodes (specs will come from the researcher wanting them.)
- We already have Xeon Phi from recent purchase

Storage

- Scratch: Do you need a parallel filesystem?
  - Needs staff and/or great support (expensive)
  - Size depends on workload and purge policy
- Home -- small and simple and is often backed up.
- Work – big, not too slow.
- Archive – PetaStore (what do others do?)

OSU's latest basic outline: Storage

- Home: ~20TB storage appliance x3 (we want this redundant-ish)
- Scratch: 100TB appliance with 800 number on it (unclear if we can get it this small.)
- Work: 1 PB (servers full of disks, NFS, RAID6, cheap and simple.)

Interconnects

- Gigabit Ethernet – workhorse
- Infiniband/Omnipath
  - Depends on workload (oversubscription can save money if your workload doesn't have many large parallel jobs.)
- IPMI network for out-of-band mgmt
  - Worth the small expense

OSU's latest basic outline: Interconnects

- Infiniband/Omnipath:
  - as cheap as we can get it
  - Highly oversubscribed – all our parallel jobs fit within a single switch
- GigE – top of rack, uplinked to central 10G
- IPMI

Login & Management

- Login nodes
  - Get enough to handle all your users
  - >1 can give high availability
  - Round robin DNS
- Management nodes:
  - Much diversity in how this is done
  - Where you can run all the cluster-wide services
  - Depends on size of cluster, services needed
- Very small clusters often have a single server for both login and management…

OSU's latest basic outline: Management

- 2x login nodes (same proc as compute)
- 3x mgmt nodes

- We also want Vendor installation and support (remember 1 FTE dedicated to the technical stuff.)

Other optional bits

- Data Transfer node (see ESNet for specs)
- Web interfaces/science gateways, etc.
- What else?

- OSU – yes to DTN, we already have sufficient webby stuff (aka virtual server pool)

Strategy

- We gave this list to any vendor that wanted to talk to us.
- We told them the budget
- The resulting quotes were very informative – go over them carefully!
- We have plenty of time because we’re waiting for the new power and cooling to be installed.
- We will go out for bid

Watch out for:

- Enterprise vendors have a completely different mindset: uptime, redundancy, etc.
- The level of redundancy on a cluster is the compute node.
- Get references from someone who bought from that vendor a similarly sized cluster who has a similarly sized staff
- Compare notes with as many people as possible

Be warned:

No matter what you do, the minute you send out the PO you'll think of something you should've done differently.

And it's okay, we all feel that way.

Topics from Etherpad

- Advertising specs
  - casual or RFP?
- Scheduler & policies
  - what meets your researchers needs?
- Provisioning mgmt. system
  - Rocks, xcat, openhpc, razor, puppet, vendor supplied, etc.
- Replacement schedule
  - What’s your funding cycle like?

Topics from etherpad

- Single system vs two?
  - we keep our old system around a while to ease transition
- Liquid cooling?
  - $$$$
- Voltage
  - Depends on how much and what you have available
  - UPS > 100KW are usually (only?) 3 phase 408V
- Rack standards
  - Space available? Density of cooling available?

Acquisition start to finish (was part 2 plan)

- Get money
- Spec system based on what’s needed
- Get bids (informally or formally)
- Buy
- Don’t sign acceptance before you test everything with at least some of your workload

Thanks!

Questions?

Dana Brunson

dana.brunson@okstate.edu

OSU High Performance Computing Center

http://hpcc.okstate.edu