Production Cluster Deployment - Then, Now, and Future Trends

Dana Brunson - Oklahoma State University

Jeff Pummill - University of Arkansas

In the old days...


UofA
Whitebox
60 Cores

Commodity PC's

ATX Cases on generic racks
Fast Ethernet
Dual socket Mobo w/Pentium III 1.0 GHz
256mb of RAM per core

Current state of affairs...


UofA
Dell Power Edge
1256 Cores

Rack Mount 1U Servers

Infiniband Interconnects
Dual socket Quad Core 64-bit
2Gb of RAM per core
Lustre Parallel File System

Larger Examples...


TACC
Sun Constellation
62,976 Cores


JAGUAR
NCCS
Cray XT5
224,256 Cores

Future Trends...

Accelerators

Increasing Core Counts

Multi-function Chips

Virtualization

Alternate Programming Languages

Do you really need to deploy your own supercomputer?

Considerations before "taking the plunge"...

- Deployment takes time from doing science
  - Use Teragrid, college or state shared resources instead

Words to live by

"It depends..."

--H. Neeman

TeraGrid

- Many systems of various architectures
- Wide variety of applications installed
- Technical expertise second to none

http://teragrid.org

If you must deploy your own...

- Simple cluster takes a few hours to deploy.
- Adding infiniband and parallel filesystem can add months to deployment.
- Find out if your applications benefit from added complexity and how much.
- Ask app user community and HPC community.

Things to consider

Power Usage Examples

Machine	# of racks	Theortical (Gflops)	Power (kilowatts)
Cimarron (OSU-mini)	0.5	896	5.75
Pistol Pete (OSU)	3	5,448	33.44
Star (UARK)	8	13,364	73.37
Roadrunner	296	1,375,780	2,345.50

Relevant Links

http://www.dell.com/calc

http://www.sun.com/servers/x64/x2250/calc/index.jsp

http://www-03.ibm.com/systems/bladecenter/resources/powerconfig/index.html

Intended Use

- R & D Cluster
- Production Cluster

Quote

"In theory, there is no difference between theory and practice.
But in practice, there is!"

-- anonymous

Hardware Choices

- Whiteboxes
- Commodity Servers
- True Supercomputers
- Hybrid Hardware

Cluster Software Stacks

- Free: ROCKS, OSCAR, CAOS, or xCAT
- Commercial: OCS or Rocks+
- Alternately: Roll-Yer-Own

Relevant Links

http://www.rocksclusters.org
http://svn.oscar.openclustergroup.org/trac/oscar
http://www.caoslinux.org/index.html
http://xcat.sourceforge.net/
http://my.platform.com/products/platform-ocs
http://clustercorp.com/rocksplus/index.html
http://www.platform.com/Products/platform-cluster-manager
http://debianclusters.cs.uni.edu/index.php/Main_Page

Filesystem Choice

- NFS
- PVFS2
- Lustre
- Panasas
- GPFS

HPC File System Articles

HPC File Systems by Jeff Layton

http://www.linux-mag.com/id/4169 (part 1)

http://www.linux-mag.com/id/4181 (part 2)

http://www.linux-mag.com/id/4358 (part 3)

Relevant Links

http://nfs.sourceforge.net/nfs-howto

http://www.pvfs.org/

http://wiki.lustre.org

http://www.panasas.com/

http://www-03.ibm.com/systems/clusters/software/gpfs/index.html

Quote

We stand at a crossroads. One path leads to despair, the other to destruction. Let's hope we make the right choice.

--Woody Allen

Interconnect Options

- Gigabit Ethernet
- Infiniband / Myrinet
- 10Gig Ethernet

Latency & Bandwidth comparison

Interconnect	Latency (microseconds)	Bandwidth (Gbps)
GigE	~29-120	~1
10 GigE	9.6	~6.9
DDR Infiniband	1.2-3.5	20
QDR Infiniband	1.2-3.5	40

bandwidth calculator: http://web.forret.com/tools/bandwidth.asp?speed=125&uni

HPC Network Articles

HPC Networks by Jeff Layton

http://www.linux-mag.com/id/3507 (part 1)

http://www.linux-mag.com/id/4146 (part 2)

Relevant Links

http://en.wikipedia.org/wiki/Gigabit_ethernet
http://en.wikipedia.org/wiki/InfiniBand
http://www.qlogic.com/Products/HPC_products_landingpage.aspx
http://www.myri.com/
http://en.wikipedia.org/wiki/10_Gigabit_Ethernet

Schedulers / Resource Managers

- Honor System
- Free: Torque / SGE / Slurm
- Commercial: LSF / MOAB

Relevant Links

http://www.clusterresources.com/pages/products/torque-resource-manager.php
http://www.sun.com/software/gridware/
https://computing.llnl.gov/linux/slurm/
http://www.platform.com/Products/platform-lsf
http://www.clusterresources.com/pages/products/moab-cluster-suite.php

Software Applications

- Serial vs Parallel
- Open Source vs Commercial

Cluster Tools

- Modules
- IPMI
- Commercial Remote Access Tools

Data Management

- Backups
- User Quotas
- Aging Scripts

Quote

I think I might believe what I just said!

-- Bill Camp