



Linux Clusters

A brief introduction to Linux and Linux clusters

Beng Tan

Calyptech

© All reserved 2004.

Contents

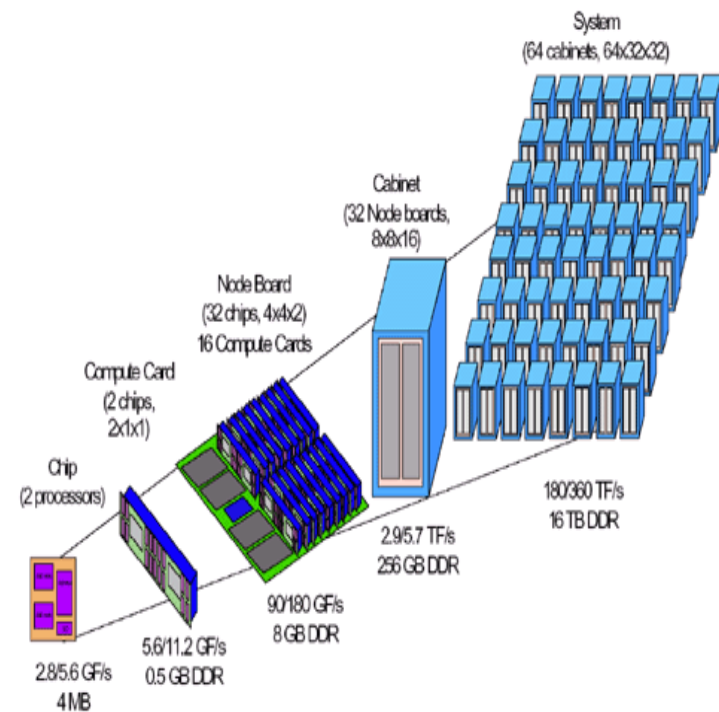
- Top Supercomputers
- Origins
- Definitions
- Applications
- Programming Model
- openMosix
- OpenSSI
- Conclusions
- References

Top Supercomputers

- Top Supercomputers List
 - List of most powerful supercomputers in the world
 - Updated twice a year
 - This year shows the ascent of Linux.
 - Previous winner was NEC's Earth Simulator, which has been relegated to 3rd.

Top Supercomputers (cont)

- Blue Gene/L (1st)
 - IBM Linux cluster
 - Not yet fully built, but already rated at 70.72 teraflops (sustained performance).
 - Final product will have 131072 700 Mhz PowerPC 400 cores, expected performance of 360+ teraflops.
 - Runs customised Linux and SuSe LES 9



Blue Gene/L Components

National Energy Research Scientific Computing Center
<http://www.nersc.gov>

Top Supercomputers (cont)

- SGI Columbia (2nd)
 - SGI Linux cluster
 - 10160 1.5Ghz Itanium 2 processors
 - 51.9 teraflops

- NEC's Earth Simulator (3rd)
 - Old style vector supercomputer
 - 35.9 tera flops
 - Was most powerful for 2.5 years.

- MareNostrum at Barcelona Supercomputing Centre (4th)
 - IBM Linux Cluster
 - 3564 IBM 2.2Ghz PowerPC 970 processors.
 - 20.5 teraflops
 - Uses conventional blade servers

- Market share
 - For top 500, IBM has 43.2%, HP has 34.6% market share
 - For entire supercomputing market, HP has 33.5%, IBM has 30.2% market share.

Origins

- The Beowulf cluster is commonly cited as the first cluster.
- Built in 1994 by Donald Becker and Thomas Sterling for Center of Excellence in Space Data and Information Sciences (CESDIS), which is part of the non-profit University Space Research Association (USRA), partially funded by NASA.
- Conceived as a commodity-based cluster system designed as a cost-effective alternative to large supercomputers.
- Consisted of 16 DX4 processors running Linux and multiple 10Mbits ethernets. A single 10 Mbit ethernet was too slow.

Origins (cont)

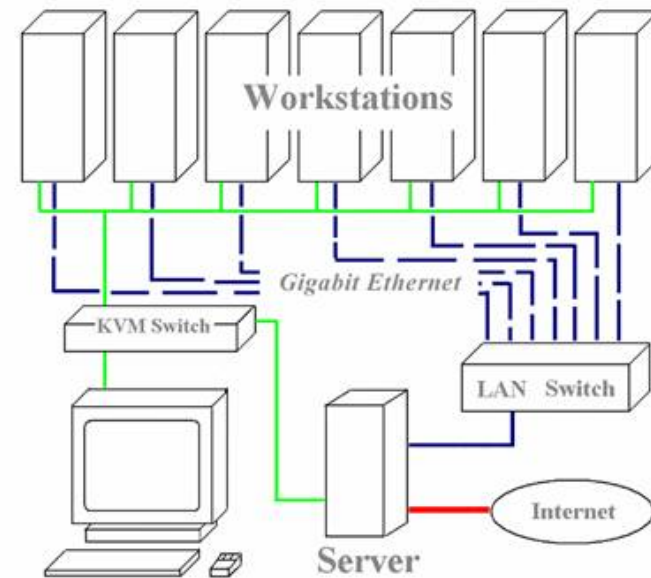
- Very successful. The concept spread into academic and research communities and similar machines were built. Such machines are now known as "Beowulf Class" clusters.
- Existence and growth due to available COTS hardware and open source software, particularly Linux, GNU and MPI/PVM.



Nasa HYPRWULF Cluster
Source: hapb-www.larc.nasa.gov

Definitions

- Cluster
 - A collection of computers that provide common resources to solve a common problem.
 - Usually the nodes in a cluster are identical, homogenous.
 - Typically concerned only with pooling computational resources
 - Usually in close physical location. Lower interconnect latency.
 - Details and possibilities are wide and diverse. Many different variations of implementation.



Definitions (cont)

- Grid
 - Grids consist of heterogenous resources - different vendors, different OS.
 - Includes storage and networking as well as computational resources.
 - Geographically distributed
 - Dynamic. Resources can be added or removed on an ongoing basis.

- Clusters and grids are complementary technologies.
 - Grids may contain clusters.

- Symmetric Multiprocessing (SMP)
 - Multiple identical processors within a single computer (commonly 2 for consumer products).
 - Effectively a single computer.

Applications

- Three main cited applications of clustering, with some overlap.
 - High Availability
 - Load Balancing
 - High Performance Computing
- High Availability / Fail Over
 - Purpose is to ensure high availability of services should an individual node fail.
 - Possible scenario: Two hosts with a heartbeat connection to monitor each other. When a service on a computer fails, the other tries to take over.

Applications (cont)

- Load Sharing / Balancing
 - Purpose is to balance computational load on many nodes.
 - Incoming service requests are routed to the least busy machine ie. web server farms.
 - Load balancing cluster may also have HA characteristics

- High Performance Computing
 - Purpose is to aggregate computational resources for heavy processing tasks - supercomputing
 - Beowulf is an example of this.
 - Computations are farmed out, remotely executed, and then retrieved and synchronised.
 - Tend to have some load balancing features

Programming Model

- Two types
 - Beowulf Cluster
 - Single System Image (SSI) Cluster

- Beowulf Cluster
 - Clustering done at application level
 - Applications are specially written using clustering software libraries such as Message Passing Interface (MPI) or Parallel Virtual Machine (PVM).
 - Fully customisable, but rewrite required.
 - Custom code is usual for academic/research community.
 - Arbitrary computational granularity. True parallel execution.
 - High performance.
 - MPI / PVM are open APIs. Easily ported to future environments.

Programming Model (cont)

- Single System Image (SSI) Cluster
 - SSI - Presents a view of a single computer, administerable from any node.
 - Clustering done at kernel level, dynamically migrates processes to balance load.
 - Existing programs run unmodified.
 - Granularity is only at the process level.
 - Work in progress to migrate threads.
 - Clustering file system.
 - openMosix and OpenSSI

- Beowulf and SSI technologies are complementary.

openMosix

- History

- Mosix started in early 1980's on PDP/11. Ported to BSD/OS environments. Linux implementations surfaced around 1997.
- Mosix adopted closed license in 2002. openMosix was forked from it and licensed under GPL.
- openMosix no longer has any Mosix code.

openMosix (cont)

- Implementation
 - Kernel patch. Easy to set up.
 - User space tools available to administer cluster.
 - ClusterKnoppix distribution.
 - Adaptive load balancing algorithm is based on market economic theory which considers comparative prices for each resource and cost of migration.

- Support
 - Over 3000 installations.
 - Average number of nodes is 25. Reportedly scales to 2000 nodes.
 - 97% of Mosix clusters moved to openMosix.
 - Very active and supportive community. Lots of information and tutorials.
 - Well established and mature.

OpenSSI

- Recently conceived project.
- Attempts to be a clustering framework incorporating required technologies.
- Possibly technically superior to openMosix, but still in early development.

Conclusions

- Linux clusters are an established technology.
- Good performance at 1/10 price of supercomputers.
- Requires sysadmins, not specialists.
- Linux is the only OS to receive attention for HPC efforts.
- US\$500 million venture capital for clustering in 2002

- Disrupts traditional supercomputers

- Evaluation
 - openMosix is straightforward to implement and evaluate.
 - MPI is more specialised and requires more effort.

References

Top 500 Supercomputer Sites
<http://www.top500.org>

Blue Gene, Linux top supercomputing list
http://news.com.com/Blue+Gene,+Linux+top+supercomputing+list/2100-7337_3-5443764.html

Linux, x86 clusters take over top 500 supercomputer ranking
http://www.cbronline.com/article_news.asp?guid=A6E54915-E012-4B62-B7AE-382A4F670154

Beowulf History
<http://www.beowulf.org/overview/history.html>

Perspectives on grid: Grid computing -- next-generation distributed computing
<http://www-106.ibm.com/developerworks/grid/library/gr-heritage/>

Advantages of openMosix on IBM xSeries, Part 1
<http://www-128.ibm.com/developerworks/eserver/articles/openmosix.html>

The Secrets of openMosix
<http://www.samag.com/documents/s=8817/sam0313a/0313a.htm>

ClusterKnoppix
<http://bofh.be/clusterknoppix/>

Introducing openMosix
<http://www.linuxdevcenter.com/pub/a/linux/2004/02/19/openmosix.html>

openMosix
<http://openmosix.sourceforge.net/>

Linux Clusters State of the Art
http://www.cineca.it/streaming/openmosix/slides_moshe.php?radice=Diapositiva&last=32

OpenSSI
<http://www.openssi.org>