Clusters

Rob Kunz and Justin Watson
Penn State Applied Research Laboratory
rfk102@psu.edu
Contents

- Beowulf Cluster History
- Hardware Elements
- Networking
- Software
- Performance & Scalability
- Infrastructure
- Build topics
- Summary
Beowulf Cluster History

- The name ‘Beowulf’ comes from the Old English epic poem of the same name. The character Beowulf is a hero whose task is to destroy the monster Grendel who is bullying the general populous.

- In this way the name is significant. The Beowulf Cluster can be viewed as the hero that took the supercomputing monopoly away from the giant institutional research centers and created the real possibility of highly parallel computation for the masses.
Beowulf Cluster History

- 1993: Perfect conditions for the first Beowulf Cluster
- PC Advancements
  - Faster CPUs 286->386
  - RAM of reasonable costs and densities (8MB)
  - Disk Drives of several hundred MegaBytes available
  - Ethernet (10 Mbps) and cheaper hubs
  - Linux improved rapidly and was in a usable state
  - PVM accepted cross-platform message passing model
- Clustering had already been done with commercial UNIX but the cost was high
- The 1st Beowulf cluster built by NASA in 1994 - budget of $50K - 1Gflop performance
Clusters have exploded in popularity in 10 years:

- Low cost per node (\$/FLOP)
- They are everywhere – broad knowledge base (e.g., http://www.beowulf.org)
- Extremely easy and cheap to expand
- Builder control over system design (hardware+software)
  - Hardware selection for CFD application system would be different than FEM application system
- The systems are very stable and robust
- Many turnkey systems are becoming available – i.e. vendor delivers a cluster to your doorstep ready to rip
- Easy to keep up with current technology – systems are virtually throw away!!
Hardware Elements – Commodity Components

- Can be constructed using any PC based technology from low cost desktops to high end servers
- Can be single, dual, or quad processor machines
- The performance per processor of the dual and quad processor systems is lower than single processor systems but the price per processor is lower. You also get more processors in the same amount of space.
- They use Intel or AMD processors 32 bit, 64 bit, or 32 bit extended
Networking

- Fast Ethernet and Gigabit Ethernet are the commodity interconnects.
  - Standard add on to any computer
  - Standard communication protocol (TCP)
  - Multiple devices can be ganged together in parallel to improve performance (Channel bonding)
  - Relatively cheap
- GigaNet and Myrinet are proprietary interconnects.
  - Proprietary communication protocol (VIA and GM respectively)
  - Expensive but faster and lower latency than Ethernet.
- Future interconnects 10 Gigabit, InfiniBand, ISCSI
The Operating system is typically Linux based but can be BSD or Windows. Windows based clusters are not typically referred to as Beowulf.

Linux is typically used

- It is free or low cost distributions can be used
- Open source (i.e. highly customizable)
- Familiar UNIX like OS (popular gnu utilities have been ported to it)
- Has proven to be stable reliable OS
- Full suite of optimized libraries and programs are available
Software – Compilers

- Clusters are distributed memory systems:
  - Copy of the code is running on each processor as if it were a stand alone system
  - Accordingly no special compiler technology is required on cluster systems
  - gcc, pgcc, pgf77, lf90, etc…
Communications model for clusters, as with all distributed memory parallel systems is based on message passing.

Some data is passed in chunks in between processors at user defined locations during execution.

Model is well suited to “domain decomposition”.

Open source message passing libraries/standards (MPI, PVM):
- User instruments their application software to perform necessary message (data) passing using high level calls (e.g., MPI_Send, MPI_Bcast, …)
Software – Job Control Software

- As with most HPC systems, clusters have many users running calculations simultaneously.

- Queuing software is required to prioritize jobs:
  - Control when, where and how the calculation will run
  - PBSPro or OpenPBS (Open Source) are typically used on Beowulf clusters
  - Highly customizable to suit the users needs
  - Can compile usage statistics towards optimizing system utilization
Software – Job Control Software

OpenPBS accounting file – script plots data:

- 1988 Jobs
- 28280 Processor-Hours
- 54.5 % System Usage
Performance and Scalability

- The performance and scalability of a system depends on:
  - Type of job; applications range from:
    - “Embarrassingly” parallel jobs which have little or no inter-process communication
    - Jobs with large amount of communication
    - Relative speed of processors/network
  - System design should reflect the type of problems the system will primarily be used to solve
Performance and Scalability

Lung model, 415,357 cells, single-phase, turbulent, non-AMG PETSc

32 processor model
Performance and Scalability

High lift pod concept

- Ideal
- NPHASE, 1-phase, 2.3x10^6 cells, PETSc
- NPHASE, 2-phase, 3.3x10^6 cells, PETSc
- NPHASE, 2-phase, 3.3x10^6 cells, Jacobi

48 processor model
Performance and Scalability

NPB (NAS Parallel Benchmarks): LU Decomposition

- Differences between 1 processor/node and 2 processors/node caused by resource sharing on the dual processor nodes
  - Memory Bandwidth
  - Disk IO
  - Network IO
Infrastructure

- To improve the performance of processors more transistors are added to the chip.
  - As the number of transistors increases so does the power consumption of the chip.
  - Increases the overall power demand of the system
  - Increases the heat produced by the system
- These factors play a role in power supply to the system, UPS, and cooling of the system
- The system footprint and processor density also is a factor. The more processors you have in an unit area the harder time you have cooling the system.
- Manufacturers are increasing the number and speed of fans in the systems to solve the cooling problem but this leads to a power consumption and noise increases
Infrastructure

- Two of our clusters (72 processor + 23 processor) sit in a small room surrounded by acoustical tile and with two huge air conditioners blasting on them since we have engineers sitting in next room!

- **I386** – 1 watt
- **I486** – 2 watts
- **Pentium** – 14 watts
- **Pentium Pro** – 30 watts
- **Pentium II** – 35 watts
- **Pentium III** – 35 watts
- **Pentium 4** – 75 watts
- **Itanium** – 130 watts

- By 2010, > $10^9$ transistors dissipating over 1 kW/cm$^2$ power
Build Topics

- Short Assembly Time
  - Sadie: 48 processor system; hardware delivery to 1st production calcs < 2 weeks (1 man)

- Extremely easy and cheap to expand

- Builder is responsible for designing the system
  - You have control over which hardware and software is used.

- Easy to keep up with current technology

- The systems are very stable and robust

- Relatively easy to reorganize node topology to suit the application
Build Topics

PC boxes on a shelf approach  rack mount
Summary

- Cluster computing is here and here to stay
- Virtually all scientists and engineers performing large scale computations are or soon will be using these systems
- They are cheap, powerful, configurable, fairly small, easy to maintain
- They have truly brought supercomputing to the masses