Technical University of Denmark
Informatics and Mathematical Modelling

High Performance Operating Systems

Beowulf

Authors: Bojan Pajkovski
Yavor Emilov Markov

Supervisor: Robin Sharp

Date: September 16, 2004

Introduction

- This Presentation
  - Beowulf clusters
  - Cluster classification – Beowulf
- Beowulf
  - History
  - Model
  - Evolution of the Beowulf project
  - Application domains
  - System architecture
  - Software specification
  - System communication
- Conclusion and Future work

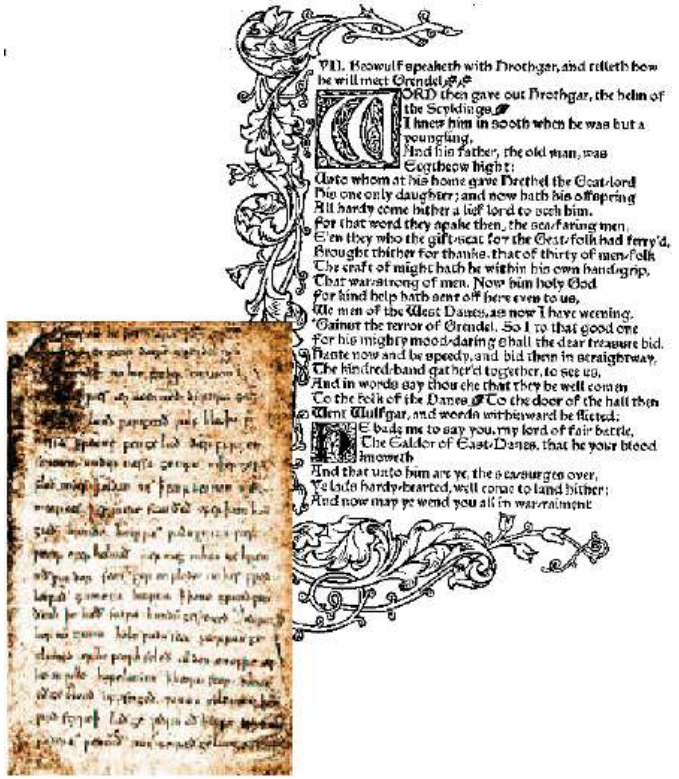
What are Beowulf clusters?

- *www.beowulf.org*
  - Beowulf clusters are scalable performance clusters based on commodity hardware, on a private system network, with open source software (Linux) infrastructure.

- Beowulf Cluster Classification
  - Application Target – High Performance (HP) cluster
  - Node Ownership – Dedicated Clusters
  - Node Hardware – CoPs / PoPs (COW / NOW)
  - Node OS – Linux
  - Node Configuration – Homogeneous cluster
  - Levels of Clustering – Group Clusters (2-99)

History

- Epic poem from the 8th century
- Beowulf – Great warrior
  - saves the Lord of the Danes and his court from the evil monster *Grendel*
- Beowulf has a 10 year anniversary this month


VII. Beowulf speaketh with Hrothgar, and telleth how he will meet Grendel.

HROD then gave out Hrothgar, the helm of the Scyldings.
I knew him in sooth when he was but a youngling.
And his father, the old man, was Ecgtheow hight:
Unto whom at his home gave Hrethel the Geat-lord His one only daughter; and now hath his offspring All hardy come hither a lief lord to seek him.
For that word they spake then, the seafaring men, E'en they who the gift-seat for the Geat-folk had ferry'd, Brought thither for thanks, that of thirty of men-folk The craft of might hath he within his own handgrip, That war-strong of men. Now him holy God For kind help hath sent off here even to us,
We men of the West Danes, as now I have weening.
'Gainst the terror of Grendel. So I to that good one For his mighty mood-daring shall the dear treasure bid.
Haste now and be speedy, and bid them in straightway, The kindred-band gather'd together, to see us,
And in words say thou eke that they be well come in To the folk of the Danes. To the door of the hall then Went Wulfgar, and word-in withinward he flitted:
HE bade me to say you, my lord of fair battle, The Galder of East-Danes, that he your blood knoweth
And that unto him are ye the o'ersurgen over, Ye lads hardy-hearted, well come to land hither;
And now may ye wend you all in war-raiment

Beowulf Model

- Low-cost supercomputing
- Ability to adopt to latest technologies
- Beowulf clusters yield same or better performance than similar MPP systems
- Basic System Environment
  - Linux OS
  - GNU development environment
  - Programs usually written in C, C++ or Fortran
  - Message passing libraries (PVM and MPI) for parallel computations

DTU

The Beowulf project (1)

- Project at Goddard Space Flight Center, NASA
  - Designed by Donald Becker
- Main goals
  - Investigate the potential of PC clusters for performing computational tasks
  - Achieve “best” overall system cost/performance ratio for the cluster
  - Achieve a 1 GFLOPS peak performance
- Initial prototype
  - Project started in 1994
  - Initially consisting of 16 Intel 486-DX4 processors
  - 10 Mbps Ethernet, up to 4.6 MFLOPS
  - 10 GB storage capacity

The Beowulf project (2)

- 2nd Beowulf
  - Pentium 100 MHz
  - 100 Mbps Ethernet
  - 280 MFLOPS
- 3rd Beowulf
  - Built at NASA and other research labs
  - Pentium Pro processors
  - Performance over 2 GFLOPS
- Beowulf clusters among the top 500 performing supercomputers

Application Domains

- Scientific and Engineering problems
  - Simulations
  - Biotechnology
  - Petro-clusters
  - Financial market modelling
  - Data mining
  - Stream processing
- Web servers and databases
  - Internet servers for audio and games

System Architecture (1)

- No fixed system architecture
- Processor
  - Initially the most important but not anymore
  - Beowulf works on dedicated processors
  - Work on Intel processors
- Memory
  - Today very important together with bus speed
  - Distributed Shared Memory (DSM)

System Architecture (2)

- Network
  - TCP/IP communication between different processors
  - Bandwidth of the network was the bottleneck
  - Fast Ethernet (100 Mbps) ?
  - Gigabit Ethernet
    - Latency high – applications must be restructured for latency tolerance
    - Better interprocessor communications performance
  - Future: Fiber optic networks by 2010 ?
- Secondary storage systems
  - No longer a limiting factor
  - For a cluster of 100 Nodes, over 1 TB storage capacity

Software Specifications

- No standard software package
- Open source software
- Linux OS
  - Support for multiprocessor thread scheduling
  - Custom kernel configuration and compilation
- GNU compilers (C, C++, Fortran)
  - Deliver very good performance (compile-time analysis, code gen.)
- PVM / MPI libraries
  - Applications should be programmed for parallel execution, communicating using PVM or MPI

Grendel

- Collection of software tools within the Beowulf project
  - Resource management
  - Support for distributed applications
- Includes programming environments and development libraries
  - Message passing libraries (PVM, MPI, BSP)
  - SYS V-style IPC, POSIX threads interface

System Communication

- TCP/IP over the Ethernet internal to cluster
- Limited performance
  - Ethernet performance characteristics
  - System software managing message passing
- Multiple Ethernet networks to work in parallel
  - Every Beowulf workstation has user access to multiple parallel Ethernet networks

DTU

Conclusion

- Low cost alternative to supercomputing
- Composed of COTS (Commercial Off the Shelf Components)
  - Open source operating system (Linux)
  - GNU development environment
  - PVM and MPI
- No fixed system architecture
- High Performance cluster

Future Work

- Beowulf in the 21st century
  - Processing Nodes
  - Storage
  - System Area Networks
  - The $1M TFLOPS Beowulf
  - Barriers

Scyld Beowulf

- Scyld Beowulf
  - Developed by Donald Becker and part of the original Beowulf team
  - Based on RedHat Linux 6.2 distribution with special software to aid in cluster installation, maintenance, and performance
  - Allows for diskless installation of nodes “out of the box”
  - Includes standard MPICH package

Extended Beowulf clusters

- Scyld Cluster OS by Donald Becker
  - http://www.scyld.com
- ROCKS from NPACI
  - http://www.rocksclusters.org
- OSCAR from Open Cluster Group
  - http://oscar.sourceforge.net
- OpenSCE from HPCNC
  - http://www.opensce.org

DTU

More Information

- Beowulf.org
  - http://www.beowulf.org
- IEEE Computing Research Repository on Cluster Computing
  - http://www.ieeetfcc.org/ClusterArchive.html
- Building a Beowulf system
  - http://www.cacr.caltech.edu/beowulf/tutorial/building.html
- PVM / MPI
  - PVM 3.3 http://www.netlib.org/pvm3/index.html
  - MPI http://www.hpclab.niu.edu/mpi/
- Management software
  - KCAP http://smile.cpe.ku.ac.th/research/kcap2/
  - bWatch http://www.sci.usq.edu.au/staff/jacek/bWatch

Questions and Discussion

Thank you for your attention!

Backup slides

Clusters Specification

- Based on Node OS Type
  - Linux Clusters (Beowulf)
  - Solaris Clusters (Berkeley NOW)
  - NT Clusters (HPVM)
  - AIX Clusters (IBM SP2)
  - SCO/Compaq Clusters (Unixware)
  - Digital VMS Clusters, HP clusters, etc.

Comparison of Cluster Systems

Project	Platform	Communications	OS	Other
Beowulf	PCs	Multiple Ethernet with TCP/IP with TCP/IP	Linux and Grendel	MPI/PVM, Sockets and HPF
Bereley NOW	Solaris-based PCs and workstations	Myrinet and Active Messages	Solaris + GLUunix + XFs	AM, PVM, MPI, HPF, Split-C
HPVM	PCs	Myrinet with Fast Messages	NT or Linux connection and global resource manager + LSF	Java-frontend, FM, Sockets, Global Arrays, SHMEM and MPI
Solaris MC	Solaris-based PCs and workstations	Solaris-supported	Solaris + Globalization layer	C++ and CORBA