the Availability Digest

Penguin Computing Offers Beowulf Clustering on Linux

January 2007

Clustering can provide high availability and supercomputer-scalable high-performance computing at commodity prices. The original Linux clustering software was Beowulf (www.beowulf.org). Though available as open source, Beowulf clustering is offered as a supported product by Penguin Computing (www.penguincomputing.com) along with a line of servers supporting high-performance computing.

What makes Penguin unique is that the original developer of Beowulf, Donald Becker, is now Chief Technology Officer of Penguin Computing.

An Overview of Clustering

A computer cluster is a group of loosely coupled computers that work together closely so that in many respects they can be viewed as though they are a single computer.[1] The group of standalone computers are linked together by software and by high-speed networks. The primary advantages of clustering are high performance and high availability at a low cost.

The capability to achieve high performance makes clusters very suitable to high-performance computing (HPC). Since the workload is spread among the computers in the cluster, a cluster can be scaled to achieve very high performance, especially if the applications can take advantage of parallel processing. Supercomputing capabilities in the ten-gigaflop range and more were commonly achieved as early as the mid-1990s. [2] Today's clusters provide teraflops of processing power.

In addition, since clusters are highly redundant, they are suitable for applications requiring high availability (HA). It is possible to structure them to achieve very high availabilities with automatic failover and load balancing. Should any processor in the cluster fail, in principle its role can be automatically assumed by a surviving processor, providing that processor state can be preserved.

Another benefit of clusters is their inherent scalability. Processors can be added or deleted to adjust the available capacity. Coupled with an appropriate system management utility, changes in the hardware configuration of a cluster can easily be made.

High-performance computing coupled with high availability makes clusters suitable for massive computational tasks such as data mining, business intelligence, biotechnology, modeling, and simulation. High availability is extremely important in many of these applications as they can take weeks to run.

---

[1] www.wikipedia.org .

[2] See http://www.beowulf.org/overview/history.html.

Since clusters run on commodity hardware, even large clusters can be a fraction of the cost of equivalent supercomputers and mainframes.[3] As a consequence, "departmental supercomputers" are now possible.

Beowulf

Beowulf was the original Linux cluster facility. Beowulf was developed in 1993 at the NASA Goddard Space Center by Donald Becker and Dr. Thomas Sterling.

This development effort showed that commodity clusters could do the work of multimillion dollar supercomputers at a fraction of the cost.

The benefits of Beowulf include:

- running on Linux, an open source operating system that can be acquired inexpensively.
- the use of commodity hardware.
- the ability to deploy departmental HPC systems to avoid the waiting times required for many supercomputers.

Beowulf is supported by the open source community, which communicates via periodic conferences, and the Beowulf web site referenced earlier That web site also lists the companies, including Penguin Computing, that offer a supported version of Beowulf.

Additional sources of information include the white paper Breaking New Ground: The Evolution of Linux Clustering, available at http://www.scyld.com/breaking_new_ground.pdf, and the 1999 book, *How To Build a Beowulf: A Guide to the Implementation and Application of PC Clusters*, by Thomas Sterling, John Salmon, Donald Becker, and Daniel Savarese.

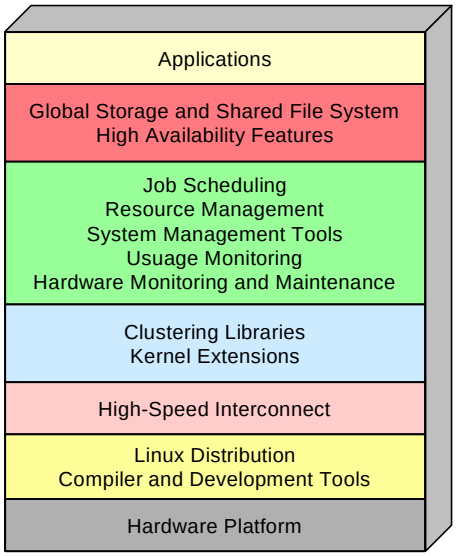
Scyld ClusterWare

Penguin Computing offers Beowulf clustering as its Scyld ClusterWare product, coupling it with the company's line of cluster servers. It obtained Scyld (pronounce "skilled") with the acquisition of Scyld Computing in 2003. Scyld Computing was founded by Donald Becker, the developer of Beowulf and the current CTO of Penguin, in 1993.

The Penguin Stack

Penguin's Scyld ClusterWare supports all of the levels of the cluster stack. With Penguin's line of cluster systems for the hardware level come the Linux operating system and the development tools needed to develop a clustering solution.

Penguin's cluster systems also come complete with a high-speed interconnect. This interconnect can either be



The Penguin Stack

[3] However, one must be cognizant of the pitfalls of "building your own" cluster. See our article in the December issue of the *Availability Digest*, Can 10,000 Chickens Replace Your Tractor?

gigabyte Ethernet or Infiniband.

The standard Linux operating system is augmented by Penguin with kernel extensions to support clustering as well as with the required set of open source clustering libraries.

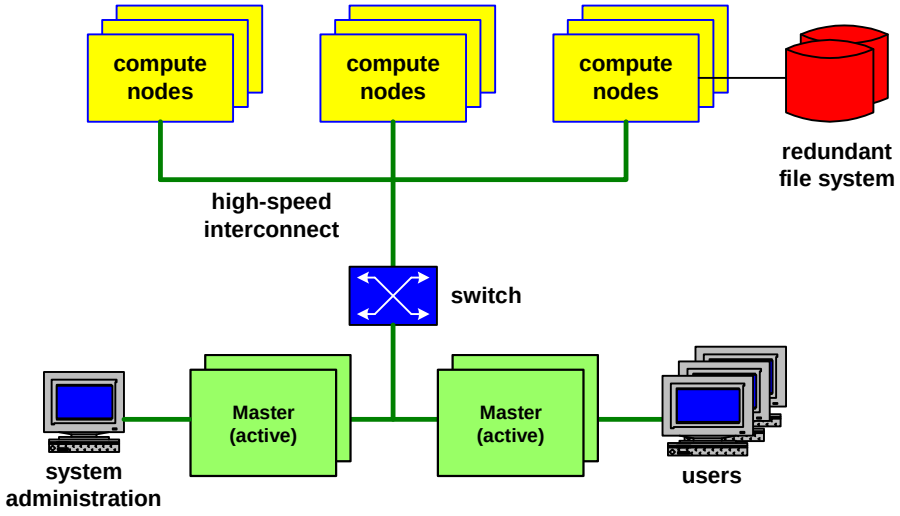
The Scyld TaskMaster Suite provides the functions of job scheduling, resource management, usage monitoring, system management tools, and hardware monitoring and maintenance.

Finally, Scyld ClusterWare includes support for a variety of file systems. Its high-availability features provide the reliability needed for these systems.

All of this is supported by services offered by Penguin.

Cluster Architecture

A Scyld ClusterWare cluster contains a group of compute nodes managed by a Master node. The Master node assigns jobs to the compute nodes based on a scheduling policy. The compute nodes are lightweight nodes with a memory-resident operating system for maximum performance.



The compute nodes and the Master node are interconnected by a high-speed, low-latency network. Also accessible through this network are the cluster file systems. These could be direct attached storage, storage area networks (SANs), or network attached storage (NAS). Redundant databases such as RAID are supported.

In addition to job scheduling, the Master node is responsible for overall cluster management.

The Master Node – The Single Point of Cluster Control

Scyld ClusterWare runs on a Master node in the cluster. The Master node acts as the single point of control for the entire cluster. Using Penguin's Scyld TaskMaster Suite, the Master node provides all of the functions of cluster management, including:

- Scheduling jobs.
- Adding and deleting compute nodes.

- Managing applications.
- Installing new applications and new versions of existing applications.
- Cluster monitoring and administration.

For systems which require high availability, a Scyld cluster can be configured with two Master nodes, one acting as the active node and the other as a backup node.

Multiple Master nodes can be configured, each with its own compute nodes. These Masters can interact, thereby coordinating their respective workloads. In effect, a cluster of clusters is created. Compute nodes can be migrated from one Master to another according to policies established by the user. The Masters can back up each other, and each Master can fail over to its backup should it fail.

Scyld ClusterWare comes bundled with the CentOS Linux distribution. Penguin also offers support for Red Hat Linux. Also included in the bundle is a toolkit that includes open source libraries such as the MPI messaging library, Ganglia web-based monitoring, cluster file systems, compilers (GCC, C, C++, Fortran), and a user interface for the monitoring of cluster status and resource utilization.

Scyld TaskMaster Suite

The Scyld TaskMaster Suite is the heart of cluster control. It virtualizes the cluster to provide a single-system image to the operator. This reduces the complexity and administration burden of clustered computing. Any person with the skills to manage a standbalone Linux system can easily manage a Penguin Beowulf cluster.

Among its primary activities is job scheduling. By default, jobs are scheduled to run in the least loaded compute node. Alternatively, a job can be assigned to a particular compute node if special services are required. More generally, job scheduling is determined by a scheduling policy established by the user. If desired, reservations of system capacity can be made for jobs to be run in the future. Scyld TaskMaster provides a simulation function for “what if” testing of scheduling policies.

In addition, Scyld TaskMaster provides the following functions:

- Adding or deleting compute nodes on demand within seconds.
- Running and managing applications and ensuring that all versions are up-to-date.
- Monitoring all components in the cluster, with status reports.
- Accessing jobs, nodes, statistics, policies, and available resources (now and in the future).
- Supporting lights-out management of remote facilities through remote power management and system health monitoring.
- Providing event triggers to automate maintenance tasks, to adjust scheduling policies, and to send notifications.

A graphical interface for cluster administration provides access either locally or through the web.

A primary goal of Scyld TaskMaster is to optimize cluster utilization, with a goal of 90% to 99% utilization.

High Availability

Scyld ClusterWare and the Penguin servers provide the functions required to achieve high availability. The Master nodes can be backed up by passive standbys so that the system continues in operation should a Master fail. Furthermore, there can be multiple Master nodes backing up each other.

Should a compute node fail, its workload can be moved to another compute node.

Compute nodes and system and application processes are loosely coupled through a messaging system so that a fault in one component will not directly take down another component.

Penguin clusters can be provided with redundant power supplies, and all hardware components are hot-swappable. Component swapping requires no tools.

High-Performance Computing

Scyld ClusterWare is optimized for high-performance computing. The compute nodes exist only to run applications as specified by the Master node. Compute nodes are lightweight, having been stripped of unnecessary software and overhead. The operating environment in the compute nodes is stateless and is fully memory-resident. Libraries are automatically cached just in time (JIT) as they are needed.

Because the compute nodes are lightweight, they can be flexibly added or deleted in seconds. This allows virtually instantaneous adjustment of a cluster's capacity to meet current workloads.

The Scyld TaskMaster ensures that all operating system components and application components are always at the latest version level. There is no version skew.

Neither the compute nodes nor the Master nodes contain any unnecessary software. Therefore, they are virtually impervious to outside malicious attacks.

Penguin Systems

The Penguin Application-Ready Clusters include two configurations – the Penguin Performance Cluster and the Penguin High Density Cluster. Both are available with either Intel Xeon or AMD Opteron microprocessors. The Opteron microprocessors can be provided as either single core or dual core.

All clusters come complete with Scyld ClusterWare and all other required software installed.

Performance Cluster

A Penguin Performance Cluster includes a full range of hot-swappable SATA or SCSI disks and hardware RAID controllers. Direct attached storage, network attached storage, and storage area networks are supported. All disk units are hot-swappable without tools.

Each rack includes redundant power supplies and high-speed interconnects using gigabyte Ethernet, Infiniband, or Myrinet. All power supplies, cooling fans, and blades are hot-swappable without the need

for tools.

The cluster blades are a 2U rackmount form factor. Each blade can contain two Xeon or two Opteron (single or dual core) processors.

A cluster node can contain from 32 to 256 compute nodes, and the cluster itself can scale to thousands of compute nodes. Each node comes equipped for lights-out operation.

High Density Cluster

The High Density Cluster is similar to the Performance Cluster except for the blade configuration.. It packages 24 processors (Xeon or Opteron single or dual core) in 4U of rack space. A single 42U rack can contain up to 480 compute nodes and provides two teraflops of processing power.

Penguin Computing

Penguin, located in San Francisco, was founded in 1998. It acquired Scyld Computing in 2003.

Its Chairman and CEO is Enrico Pesatori. Donald Becker is its CTO, and Pauline Nist is Penguin's Senior Vice President of Product Development and Management.

Summary

The benefits of Scyld ClusterWare are many:

- The entire cluster is managed as a single virtual machine. There is only a single point of management.
- It requires installation only on the Master nodes.
- There is visibility into the entire cluster from the Master node.
- Automatic job scheduling is provided based on user-supplied policies.
- Compute time slots can be reserved.
- Workloads can be prioritized.
- Compute nodes can be added, deleted, or reprovisioned in seconds.
- It is highly scalable to thousands of compute nodes.
- It can be configured to be highly available.
- Because of the use of lightweight software, clusters are highly secure and virtually impervious to malicious attack.
- Accounting is provided for shared usage.