Breaking New Ground

- Evolution of Linux Clusters: Challenging Conventional Wisdom
  - Timeline of Innovation driven by upsetting the expected belief
- Fearless Forecasts for the Future
  - Conquering uncharted territory
1. Only supercomputers can do the job

- Prevailing belief that only custom designed architectures could solve complex problems
- SMP supercomputers required to meet needs of high performance computer users
- Only a small group of highly skilled programmers could write High Performance Computing (HPC) code
- Domain experts had to depend on these programmers to design the analyses and simulations

*High Performance Computing was too costly for most companies*
2. Open Source not a viable platform

- Only UNIX was considered sufficiently robust for HPC
  - Linux perceived as a “toy” system by many
- Commodity hardware too slow and primitive
- Proprietary hardware and software was required for peak performance
  - The OS vendor controlled the tools
- As recently as ’97, Windows NT even considered the only viable alternative platform given Msft’s dominance
- Attack of killer microchip anticipated

$ Million+ price tag still a huge barrier to entry for most
Disruptive Technologies Converge

- Widespread acceptance of personal computers reduces cost of commercial, off-the-shelf (COTS) components
- Higher clock rates, cheap memory and networks
- Innovation comes first on commodity platforms
- Linux and Open Source gain acceptance
  - Rebel operating system, but capable of working with broad set of commodity hardware
  - License enables coherent development without proprietary splits
Upsetting the Expected Beliefs

1. Use Networked PCs for HPC
   - Commodity hardware is now powerful enough
     - Overcome latency issues
   - Empower the domain experts to design the code

2. Use Linux for the OS
   - See potential, not a toy or enthusiast’s tool
   - Recognize networking capability of Linux
   - Build on open source vs. proprietary mindset

Birth of Beowulf Project
Beowulf Democratizes Supercomputing

- Project conceived by Becker and Sterling in ‘93 and initiated at NASA in ‘94
- Objective: show that commodity clusters could solve some of the easier problems usually handled by $million supercomputers but at a fraction of the cost
- Build a system that scaled in all dimensions
  - Networking, bandwidth, disks, main memory, processing power
- Initial prototype
  - 16 processors, Channel-bonded Ethernet, under $50K
  - Matched performance of contemporary $1M machine
- Idea spread quickly through NASA, research, academic communities

**HPC at a fraction of traditional cost**
Early Beowulf Clusters

- Unsupported
- Roll your own
- Hardware reliability issues
- Compute density required considerable floor space
- Cheap
Beowulf Pioneer Community: DIY Innovation

- Potential for a variety of applications was tremendous
- Domain expert likely to also be application architect, programmer, system administrator
- Only a subset of people had the talents, skill, and time to play all roles
- Open source meant everything was free

Mindset & practical considerations still limited who could participate
3. Roll your Own Clusters

- Sometimes the belief most in need of change is your own
  - DIY approach not perfect
- Not all domain experts had know-how, desire or time to build their own clusters, write apps, and manage system
- Commercial customers expected reliable hardware, supported apps, stability, training, and even documentation
- Financial resources were needed to advance technology further

Scyld Software founded to overcome cluster management barriers
Clustering had Inherent Scalability Problems

- While COTS hardware was cheap, the time to build your own HPC Linux cluster was not!
- Clusters required full install on each system or use of NFS (Network File System)
- Configuration assumed fixed set of machines at installation
- MPI and PVM were only interfaces for cluster programming of parallelized applications

*A commercially-viable cluster solution had to be easier than this*
Unified Cluster System Prototype: 2000

- Scyld UCS prototype - full install only on master node, netboot and compute nodes existed only to run applications
- Designed from scratch — delivers single system installation, administration, provisioning, monitoring, process space: *BeoMaster*
- Automatically, incrementally and transparently scalable, no cascading failures
  - No need to assume a fixed set of machines
- Deployment platform — standardized configuration
4. Clusters are good for scientific research and technical simulations

- PCs powerful enough to do HPC analysis for commercial applications such as MCAD/E, geoscience, bioinformatics
- Expensive supercomputers mostly reserved for government research and defense contractors
- All major hardware vendors offer Linux - recognized as
  - Stable and equally robust as UNIX
  - More scalable than Windows NT
  - More economical than other operating systems
- Key ISVs developing for distributed model
- Beowulf is an accepted approach for clusters
Mainstreaming the Movement

- Engineering teams across different industries under pressure
- Need to get products to market faster on tighter budgets
- Aging workstations are common
- Want more complex simulations earlier in design process
- Facing analysis bottlenecks
- Don’t have time to build their own clusters

*Complicated cluster management prevents broader uptake*
Linux HPC Cluster Sweet Spot

- Multi-physics Coupling (# state variables)
- Scale (# degrees of freedom)
- Dynamics (# time steps)

- Supercomputers
- Linux HPC Clusters
- Desktops
Turning it into a science not an adventure

- Scyld’s single system management makes it reasonable and cost-effective to upgrade to clusters as workstations need to be replaced.
- Scyld’s unique approach enables anyone who can administer a single Linux box to easily set up and manage a Scyld cluster up to 1000 nodes.
- Incremental scaling is possible without redesign or administrative effort.
- Combination of ease of use, power, support is ideal for commercial installations.

*Complete, commercially supported software platform for HPC clusters*
Scyld Beowulf Overview

Simplicity & Ease of Use
### Scyld Features & Benefits

**Technology leadership**

- **Single Point of Cluster Management**
  - Single system installation
  - Single system administration
  - Single system monitoring

- **Unified Process Space**
  - SMP-like environment
  - Lightweight compute nodes
  - Automatic process migration at job execution time
  - Manage processes w/ std Linux tools

**Customer benefits**

- Install once, execute everywhere
- Add or remove nodes in seconds
- More secure model
- Supports diskless nodes
- **Lower deployment, management, maintenance costs**
- Cluster invisible to end users
- Easier to submit & manage jobs
- Lower overhead for applications
- Users focused on designs, not clusters
- **Shorter design cycle**

---

**BeoMaster: Key libraries & extensions to Linux kernel for clustering**

- Single Point of Cluster Management
- Unified Process Space
### Scyld Features & Benefits

#### Technology leadership

#### Customer benefits

---

**Complete Software Platform for Linux Clustering**

<table>
<thead>
<tr>
<th>Full Linux Distribution</th>
<th>Customer benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completely standards based</td>
<td>Familiar Red Hat environment</td>
</tr>
<tr>
<td>Linux Kernel Version 2.4</td>
<td>No need to purchase additional RH licenses</td>
</tr>
<tr>
<td>Most Red Hat applications using MPI run unchanged*</td>
<td>Not proprietary, fully standards based</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Integrated &amp; Flexible HPC Toolset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bundled and pre-tested</td>
<td>Complete HPC clustering solution</td>
</tr>
<tr>
<td>Parallel libraries (MPI, PVM)</td>
<td>Integrated &amp; pre-tested</td>
</tr>
<tr>
<td>Compilers (C, C++, Fortran)</td>
<td>Flexible platform to integrate other popular HPC toolsets</td>
</tr>
<tr>
<td>Cluster file system (PVFS)</td>
<td>Works out of the box</td>
</tr>
<tr>
<td>Library interfaces to integrate other tools/workflows</td>
<td></td>
</tr>
</tbody>
</table>

* May require configuration or minor modifications to distribute across cluster
Clusters delivering on the promise

- **Hitachi Manufacturing**
  - Using CFD to study airflow in its hard drives

- **National Weather Service**
  - Weather information dissemination system
  - Relies on intensive, behind-the-scenes computation - used to issue up-to-the-minute weather updates and warnings to the public

- **University of Arizona Lunar and Planetary Lab**
  - Numerical simulations to study the formation of planet surface features & dynamics of planet atmospheres & circulation

  Scyld ‘supercluster’ has increased compute speed fifteen fold so the Lab can handle larger problems, covering a larger region of the solar system.
Scyld Future Roadmap

Platforms & Interconnect
- Xeon, Opteron
- GBEnet, Infiniband (Infinicom)
- Nacona
- Additional IB vendors

High Performance Computing
- Scyld Beowulf cz-3
- Scyld Beowulf cz-4
- Scyld Beowulf cz-5
- Scyld Beowulf Product Family

High Availability & Enterprise
- Run to Completion
- Active-active master
- Failover (stateless)
- Virtualized (stateless)

- Virtualization of services (multiple masters)

1H04 2H04 1H05 2H05 1H06
The Beliefs we challenged

1. Only supercomputers can do the job
2. Open Source not a viable platform
3. Roll your own clusters
4. Clusters are good for scientific research and technical simulations

And…

5. Grid Computing is the future of distributed computing
Fearless Forecast: Clusters Here to Stay

- Commodity hardware and Linux continue to advance

- Cluster model will be applied to enterprise uses
  - Bulk data handling, data mining
  - High Performance Throughput
  - Multiple small scale parallel jobs
  - Dynamic web applications

- All sets of machines will be managed as a cluster

Clustering is the natural evolution of the computing ecosystem
Thank you!

Booth #609

www.scyld.com

www.beowulf.org