Scyld ClusterWare :
An Innovative Architecture For Maximizing Return On Investment In Linux Clustering

By Donald Becker and Bob Monkman

Abstract

Concepts such as virtualization of computing resources, clustered and grid computing, and service oriented architecture are all gaining greater mindshare as enterprises struggle to manage large pools of servers. Enterprises require commercial grade high performance computing (HPC) that scales on-demand in order to adapt to ever changing workload requirements and provide optimal system utilization. These needs in turn have driven many useful innovations.

However, these improvements were all based on the fundamental assumption that a cluster or grid configuration must be provisioned as a static, disk-based, full operating system installation on every single server. This assumption leads to the use of extensive scripting and middleware in the belief that the complexity can be abstracted out of managing 100s or 1,000s servers. In reality, this outdated approach only masks the complexity from view by adding a second layer of code without removing the underlying problem. This actually magnifies the operating costs of managing and maintaining large pools of servers.

Many organizations assumed these constraints were inherent within HPC and chose to either live with less than optimal return on investment (ROI) or avoided HPC altogether. Instead, rethinking these fundamental concepts can yield surprising results that can eliminate the very complexities many software "solutions" strive to merely camouflage.

Scyld ClusterWare, for example, turns this flawed assumption on its head. It re-architects the foundation using several well-recognized concepts in a unique combination that delivers virtualized cluster systems that make large pools of servers appear and act like a single, consistent virtual machine. A properly architected solution that leverages disk-less provisioning, a lightweight compute operating environment and simulates the appearance and end user experience of a single virtual system has a tremendous ripple effect on rapid provisioning, manageability, scalability, security and reliability within the cluster.

The result is an elegantly simple and powerful new paradigm of virtualized clustered computing. This new paradigm eliminates the need for multiple levels of cost and support. It also dramatically increases efficiency and reduces operating costs while delivering a dependable HPC service to organizations in a highly competitive business environment.

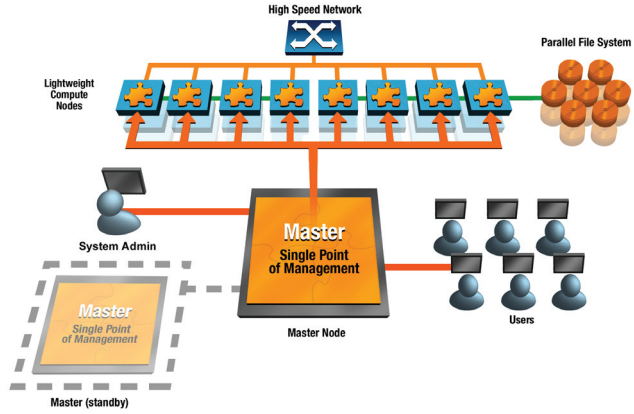
Donald Becker is the co-inventor of Linux Clustering and CTO, Penguin Computing
Bob Monkman is Director of Product Management for Scyld Software at Penguin Computing

Scyld ClusterWare Overview

Scyld ClusterWare is the latest evolution of innovative Linux Cluster Virtualization software from Scyld Software, which makes large pools of Linux servers appear and act like a single virtual system. Through a unique "Master/Slave" architecture, employing a single point of command and control (or Master) hundreds of servers, up to 1000 or more servers, can be managed as if they were a single, consistent, virtual system, dramatically simplifying deployment and management and significantly improving server performance and data center resource utilization

The compute nodes exist only to run applications specified by the Master, and are, in effect, extensions of the Master. Because the compute nodes run a lightweight, in-memory distribution, they can be provisioned rapidly so users can flexibly add or delete nodes on demand in seconds, making the cluster extraordinarily scalable and resilient. The lightweight compute node environment is stripped of any unnecessary system services and associated vulnerabilities, making the cluster nearly impossible to attack, thus inherently more secure. Finally, with single point provisioning and consistency control mechanisms, there are no version skews so the system is intrinsically more reliable. Compared to legacy technologies, Scyld ClusterWare offers a more efficient, secure way to manage servers and storage, while delivering a space-maximizing infrastructure that delivers the productivity, reliability, scalability and lower total cost of ownership (TCO) that high-performance business computing demands. Figure 1 below depicts the architecture of Scyld ClusterWare.

Figure 1. Scyld ClusterWare Unique Master/Slave Architecture



In the following sections, the unique architecture and other elements of the Scyld solution are examined in greater detail regarding their capabilities and contribution to a greater return on investment (ROI) compared to legacy approaches to cluster management.

Installation, Configuration and Updating:
Cutting Time to Productivity From Days to Hours

The biggest hurdle in getting to productive, value-add return on you cluster investment is getting through the initial installation and configuration of the software. If the system software ever needs to be completely upgraded this process must be repeated. Unfortunately, with the vast majority of cluster management approaches, this initial installation and initial configuration of the cluster can easily take 5-10 days to complete.

Scyld ClusterWare installation will generally be ready to run inside of a couple of hours. The reasons why Scyld is so much faster are related to the unique architecture of Scyld ClusterWare compared to the state of most of the other offerings available today. Another major factor is the pre-integrated and architected to work seamlessly, its stateless provisioning.
The vast majority of alternatives to Scyld ClusterWare are pure open source projects that have little or no documentation, no technical support, no integration with the various components, no intuitive graphical user interface (GUI), etc. The operating system (OS) and the cluster tools install separately with little coordination and, in many cases, the post-installation configuration of the servers to act together in a cluster is left as a manual exercise for the user. This means the user must configure all the naming service, user and host authentication configuration, I/O redirection to the Master node, network and file server configuration, etc. As you can imagine, logging into each and every compute node to manually set up the various aspects of a cluster is an excruciatingly slow process. The innovative Scyld ClusterWare architecture simplifies this process dramatically.

"Stateless" Provisioning: Bringing Up Bare Metal

The maturity of high speed networking and network booting mechanisms combined with the growing complexity of managing increasingly large pools of servers has caused the number of IT architects to recognize the benefits of "stateless provisioning" (i.e., direct to memory via a network boot) of the server operating environment to grow rapidly. It is dramatically easier and faster to provision and manage large server pools when you simply decide never to install a full operating environment to the hard disks.

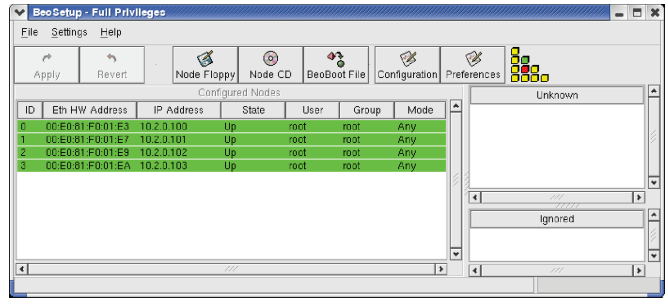
A full OS installation to the disk drive is relatively slow, generally taking 15 to 30 minutes to complete, depending on a variety of factors. This can be scripted and even parallelized but it will still take hours to provision the entire cluster once all of the preparation is completed. Also, many automated provisioning tools assume an initial network boot method, which then must be converted back to a "boot from disk" method after the cluster is installed- an extremely time-consuming process even if a good utility exists.

On the other hand, Scyld ClusterWare bundles, by default and for convenience, a complete standard Linux distribution from CentOS with optional support for installing on Red Hat ES 4.0. The installation is completely integrated and needs to be done only once and only on the designated Master node, regardless of the size of the cluster. The compute nodes are then auto-provisioned with a cluster-aware preconfigured operating system environment directly to memory and no further

configuration is generally required since Scyld ClusterWare sets up the entire necessary configuration. With Scyld, this process takes approximately 20 seconds for each node and they are ready to run.

Figure 2 depicts the graphical control screen supplied with Scyld ClusterWare for managing the cluster setup, showing compute node status and offering direct control if needed, as nodes rapidly join the cluster. It also provides access to diagnostic information should a node fail to boot.

Figure 2.



The entire process of initial installation and configuration is effortless with Scyld, especially since you install once, on the Master, and run everywhere automatically. The ROI of Scyld ClusterWare begins paying off on the very first day with the significantly reduced staff time needed to set up the cluster.

Adding, Updating or Re-provisioning Compute Nodes Effortlessly

The value of rapid provisioning is even more apparent during ongoing cluster operation when compute nodes fail, are updated or re-provisioned or new nodes are added. The diskless approach is effortless and nimble in comparison to a traditional local full-install.

First of all, with Scyld, a local compute node disk is only used if the application requires it. Since disk drives are among the least reliable and most power hungry components in the cluster, requiring them outright means more compute node failures will occur and higher capital and operational expenses will be incurred.

Furthermore, a full disk-based install on a replacement or new server is going to take another 15-30 minutes and will generally have to be scheduled by IT staff, when they can get to it. This increases the time required to restore the cluster to full operation. With Scyld ClusterWare, simply plug the node in and power it up. Once it contacts the Scyld Master node for provisioning, it is up and running in 20 seconds.

When software updates are required on Scyld ClusteWare, you apply them only on the Master, which will then re-provision the compute nodes automatically and quickly. The method used in traditional architectures requires elaborate scripting at best and increases the chance that something can go wrong part way through. This will result in another problem, version skew, the potential for which is eliminated with Scyld ClusterWare.

Application Operating Environment Consistency: The Dreaded Version Skews

The correctness of applications that run on multiple compute nodes is often dependent on everything being precisely the same on each processing element. The tiniest difference in a driver or library can render the results useless, after days of calculation — if the application runs at all.

Depending on local full-install operating environments is one sure way to create problems of version skew. You can spend hours or days trying to figure out which node has the wrong version of something. In the end, you may end up spending hours reinstalling everything fresh and getting back to a known state. Local copies get stale and inconsistent, causing wasted time and costly rework.

Generally speaking, Scyld ClusterWare provisions to the nodes exactly what is running on the master, so there is nothing to get out of sync. Even if someone pushed inconsistent libraries to a compute node, rebooting the node from the Scyld Clusterware GUI corrects the problem in seconds and returns the node to its pristine state.

One begins to see that avoiding the problems in the first place with the correct fundamental architecture makes set up much simpler. Because of this architecture, further optimizations on this concept of stateless provisioning are possible. The dedicated nature of the cluster's compute resources also means that it is unnecessary for a full Linux distribution to be provisioned to the compute nodes, on of the many benefits of right-sizing the compute node operating environment.

Trimming Out The Fat:
The Lightweight Linux Compute Server

Dedicated Linux clusters employ an array of compute nodes that support a smaller set of specific applications than do general-purpose machines. If architected properly, as in Scyld ClusterWare, there is no real need for them to be set up as standalone general-purpose machines. Therefore the Scyld approach is to provision the compute nodes with a minimal operating system, omitting extraneous packages from a standard enterprise Linux distribution, as well as the standard user shell and most of the standard Linux services (i.e., daemons such as Telnet, FTP, SNMP, NTP, etc.) to realize benefits in performance, scalability and security.

Dynamic "Just-in-Time" Provisioning

The default operating environment on the compute nodes may not always have all of the software support required by certain applications. Scyld ClusterWare addresses these cases by incorporating a powerful infrastructure into the mechanism for pushing application jobs out to multiple nodes whereby the dependencies on, for example, shared libraries that the program needs, are send out along with the job. The compute nodes then automatically "cache" the necessary shared libraries from the full Master node installation, on demand. In this manner, the compute nodes are always provisioned with precisely the environment they need and only what they need.

Greater Performance

One of the immediate benefits of the lightweight compute nodes is performance. Part of the reason Scyld compute nodes are provisioned in less than 20 seconds is that the operating system (OS) is significantly smaller. Real world applications, run in an apples-to-apples comparison, have shown short application iterations exhibiting 5-10% performance improvements with the Scyld ClusterWare compute node OS configuration versus the same Linux OS in a full OS configuration.

Related to performance is improved memory usage, especially important if insufficient memory forces an application to swap space. On a typical cluster compute node with 1 GB of RAM memory, Scyld ClusterWare consumes less than 1% (8MB) of memory compared to 40% (800MB) for a full Linux installation. These improvements may seem small but in very large clusters with long running jobs even small improvements can have a substantial positive impact on an organization's ROI for their cluster.

A more significant issue is the scheduling latencies that many of the standard Linux services can introduce over long running applications. It has been shown that these scheduling latencies can cut cluster performance of real world applications by 50% in large cluster configurations and they can be impractical to isolate, as they are very application dependent. Performance also improves in regard to scale-out capabilities resulting from the Scyld ClusterWare architecture.

Enhanced Scalability

Scale-out performance is greatly improved due to the unique design of the Scyld ClusterWare architecture. Scyld's design employs a single primary daemon on compute slaves and leverages this daemon in order to run jobs, get standard I/O, logs and statistics, etc. out on those slaves, all from the Master node. Enhanced scalability occurs because compute nodes can be added on demand and common tools can be optimized to leverage the support already built into the Scyld architecture.

One striking example of tool optimization is the Scyld implementation of the popular Ganglia web-based monitoring tool available in open source and used by many HPC administrators. Normally Ganglia employs its own daemon on every compute node and generates large amounts of XML traffic on the cluster network each time it updates. Clusters over 100 nodes must trim the refresh rate down to nearly unusable levels and it will not work at all much over 200 nodes.

However, Scyld ClusterWare already caches all of the cluster statistics needed by Ganglia on the Master node and uses its own built-in mechanism, which generates orders of magnitude less traffic. The version of Ganglia shipped with Scyld ClusterWare simply picks up the information out of memory on the Master node. The Scyld ClusterWare architecture makes it possible to run Ganglia on virtually any size cluster.

Enhanced Security in a Master/Slave Architecture

As mentioned earlier, the compute slaves of a Scyld ClusterWare cluster do not have most of the standard Linux daemons and do not have their own shell. The Master node has a special shell mechanism which sends commands out to the slaves, making the Scyld architecture inherently more secure since the compute nodes cannot be logged into or attacked in any of the standard ways.

The Power of One:
Virtualizing the Linux Cluster into a Single, Powerful Machine

A great deal of emphasis and discussion is underway in the industry regarding virtualization. Virtualization can be applied at many different levels to gain benefits in various areas. Virtualization is possible on the individual server for server consolidation, on the network and storage infrastructure for flexibility and availability of those resources, etc.

The purpose of the Scyld ClusterWare architecture is to make large pools of servers act and feel as if they were a single, consistent, virtual system. Scyld ClusterWare leverages virtualization techniques in several important areas to reduce complexity in exploiting large pools of servers, in support of the master-slave architecture of lightweight compute servers and in optimizing the resource utilization of the cluster according to an enterprise's business priorities.

In this way, Scyld ClusterWare makes it much more easy and intuitive to install, update, manage and monitor a Linux cluster with this software. We have seen several very real examples of just how effortless it is to provision and update with Scyld ClusterWare in previous sections. We have also alluded to the fact that the compute nodes in this architecture are truly logical extensions of the Scyld Master node. Now we see why.

The Elegantly Simple Single Virtual System

Scyld ClusterWare employs a powerful technique built upon the standard, out-of-the-box enterprise Linux distributions based on the 2.6.x Linux kernel, to create "single system image" behavior with the Linux you already know. It does this by extending the Linux configuration on the Scyld Master node to have a single unified process space. From both the administrator and the user point of view, a 100 node cluster with 400 processors appears very much like a 400 processor SMP machine at the cost of commodity Linux x86 cluster computing.

The compute servers are fully transparent and directly accessible if need be. However, if you are interested in the compute capacity presented at the single Master node, you need never look further than this one machine. The simplest way to convey how this might manifest itself to the typical user is to give the example of the everyday task of issuing the ubiquitous "ps" process list command. After issuing the command, what you get back is a listing of all processes running on all machines as if it were just one machine. You can still tell which processes are running where in the cluster with a simple addition to the command, but only if you really care to know it. Other standard Linux commands work in the same, intuitive way as on a single machine.

When you wish to add a user and set up passwords, you need only do it on the Master. When you want to run a job, you run it on the Master and simply tell it how many processors you need (even non-MPI jobs). If you need to terminate a job, it is done on the Master node and automatically removed cleanly on the compute node(s) it actually resides on. Of course, you can run jobs or general commands on specific nodes if you need to. If you need to see the vital statistics of load, memory usage, disk usage, etc. on any or all nodes, add one command line or GUI invocation on the Master node and you get it.

With Scyld ClusterWare you focus on the work throughput you need to achieve, not a cluster that requires massive scripting to iterate commands over each and every node.

Advanced Workload Virtualization: Scyld TaskMaster

Scyld ClusterWare provides the optimal architectural foundation for clustered computing. Built into Scyld ClusterWare is an intelligent resource management and mapping facility that has a default policy to direct jobs to the least used processors in the cluster- unless directed to run them on specific nodes. Building upon this base, however, is the optional complimentary product Scyld TaskMaster Suite that, when combined with Scyld ClusterWare, delivers a fully integrated and virtualized HPC cluster environment that delivers the right resources at the right time — every time.

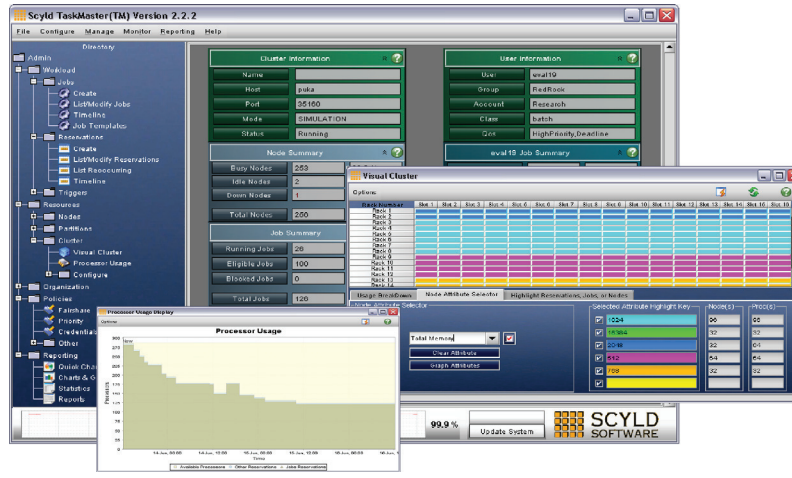
Scyld TaskMaster Suite provides advanced job scheduling and workload management capabilities with an easy-to-use graphical interface and a flexible policy engine to ensure that cluster utilization is always optimized and aligned with individual and/or departmental priorities within your organization.

The Scyld TaskMaster Suite allows organizations to share resources, applying policies that ensure agreed upon service levels to departments, groups and individual users. Its integrated policy-based job scheduler and event engine combine intelligent and flexible scheduling with advanced reservation capabilities for use by any group or individual. Event triggers can also be configured to automate maintenance tasks, adjust policies, or send notification. Scyld TaskMaster even includes a simulation mode where different policies can be tried in "what-if" scenarios to help understand how best to use the engine's capabilities.

The powerful, easy to use interface, accessible locally on the Master node or remotely via the web, allows immediate and intuitive access to jobs, nodes, statistics and policies, as well as instant visibility into what resources are available for

immediate and future use. Users submit and manage jobs with an intelligent, personalized submission screen indicating what resources and special needs are required. Then Scyld TaskMaster schedules jobs according to policies you set to maximize utilization rates in the 90-99% ranges while respecting requested deadlines and other considerations. Figure 4 gives a sample of the cluster visibility and control that Scyld TaskMaster Suite offers.

Figure 4. Scyld TaskMaster: Powerful Policy-Based Workload Management- Comprehensive Cluster Visibility



The powerful monitoring and reporting mechanisms built into Scyld TaskMaster, capabilities that are an additional cost in competing products, enable instant visualization of live and historical utilization, resource allocation, and backlog data and the creation of numerous reports on this information that are customizable for your needs, all with a few mouse clicks instead of the hours or days required with command line alternatives with text reporting only. Scyld TaskMaster vastly increases both user and administrator productivity and delivers an exceptionally high return on investment every day it is utilized.

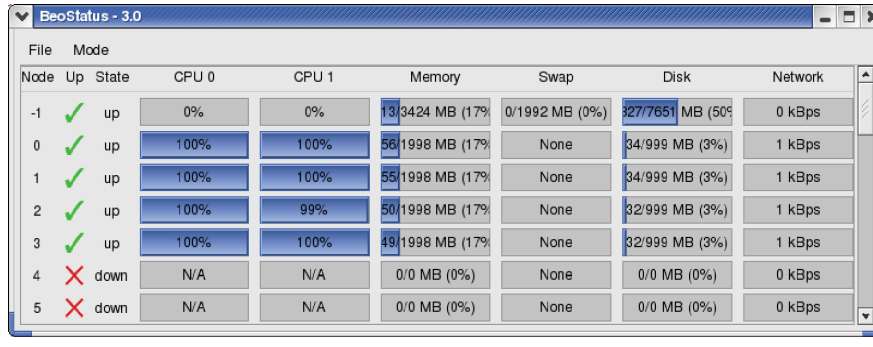
Putting It All Together

Fully Integrated Platform

Scyld ClusterWare is a fully integrated software platform for HPC that includes a standard Linux distribution as well as a complete HPC software tool set, pre-integrated and pre-tested for compatibility, ready to run. Tools include optimized open source MPI libraries and Ganglia web-based monitoring as well as open source cluster file systems and compilers. Scyld ClusterWare optionally supports commercial MPI tools, compilers and Red Hat Linux Enterprise Server.

A simple, yet effective user interface into the entire cluster's interface is provided for comprehensive visibility into the cluster status. Resource utilization is included with the product as well as a set of powerful command line utilities for power users. Figure 3 gives a sample view of a cluster with node by node statistics on CPU usage, node status, memory/ disk usage, etc.

Figure 3. Scyld ClusterWare Status View



This integrated package provides a complete environment for users to run a wide variety of custom and commercial applications. Scyld ClusterWare allows your Linux cluster to be deployed rapidly with users becoming productive immediately. It is a fully supported, working system — out of the box.

Service and Support: Dependable Linux and Clustering Expertise
In order to provide real value to customers with real challenges to solve, it not sufficient to simply offer good software products with helpful documentation. It is the expertise behind the technology that ultimately what makes the difference. Penguin Computing excels at offering not only great software products, but great expertise as well and a willingness to help customers solve their problems. Our Scyld software represents a critical part of the ecosystem in which our customers work, but the experience and expertise we offer everyday helps customers achieve their goals quickly and with far fewer unknown elements. We also offer a wide variety of on-site professional services to create tailored working solutions in our customer's particular environment and applications needs to produce results quickly. This real help from real experts who have the experience from making it happen themselves.

Customer Case Study:
Sandia National Laboratory

Even though Sandia National Laboratory's Combustion Research Facility (CRF) was doing science for the Department of Energy (DOE) with real world pocket-book impact, they were often limited by a lack of supercomputer resources required to conduct numerical simulations and analyze data. A cost-effective solution was required to enable Sandia to complete more research in-house, without adding staff to manage a large departmental cluster.

The Challenge:
The DOE funds Sandia to find high efficiency, low emission solutions to complex combustion problems. As part of this mission, the CRF conducts complex simulations and analysis of turbulent reacting flows to study interactions between fluid dynamics and combustion chemistry that affect the performance and emissions of combustion devices.

Simulations done by CRF are so complex that a typical baseline case takes one to two weeks. More sophisticated jobs need up to eight weeks of compute time. Nothing in-house was powerful enough to do that sort of work so CRF had to rely on off-site supercomputer grants to complete their research. But because the CRF had to compete for resources at the handful of supercomputing centers nationwide against other government facilities, academic researchers, etc., CRF previously only got a small fraction of the total compute hours per year needed to perform the required calculations.

The Solution:
Two members of CRF's technical staff, Senior Member Joe Oefelein and Distinguished Member Jackie Chen, realized high performance technology had evolved to the point where CRF might be able to cost-effectively create a departmental scale Linux cluster with 72 nodes (144 processors) that would allow them to run many of their initial calculations. Supercomputer time could be reserved for larger simulations that require substantial system support.

"We purchased hardware from Penguin Computing [because] their system engineering appeared to be the most robust," Oefelein said. "We chose Scyld [software] because it is easier to use [than other options] and offered us a turn-key solution. The Scyld interface emulates a workstation."

Within a day of arrival, the new Penguin/Scyld system was in place and operational which is a strong testament to how easy it is to bring up a large cluster with Scyld ClusterWare. Since then, the cluster has been in continuous operation. The Scyld cluster virtualization software dramatically eased the deployment and management of the cluster for CRF. CRF Principal Investigators simply access the Scyld cluster from their workstations just by using an informal queue to manage shared use of the cluster. A single system experience for the entire cluster on the master node also means that it is extremely intuitive and effortless for the CRF team to scale up incrementally without redesigning or administrative effort.

The net result for Sandia was that the cluster could be completely administered part time by one of the Principal Investigators, saving CRF $150,000 a year in specialized administration overhead. Also, since they spend minimal time and effort in managing the cluster, this gave them over five million compute hours per year and a five-fold increase in the amount of research they could complete.

Conclusion:
The Scyld ClusterWare Advantage
We began this paper by noting that the top challenges facing any enterprise managing a guaranteed high performance computing service are getting the right solutions to deploying and managing these resources, scaling on-demand to ever-changing workload requirements, and achieving the highest utilization levels matched to business priorities.

We also noted that emerging solutions to these challenges are based on the fundamental assumption that servers must be provisioned as a static, disk-based, full operating system installation and the belief that the use of extensive scripting and middleware can eliminate most of the resulting complexity.

Throughout this paper we have walked through the unique architecture of Scyld ClusterWare, which by employing several well-recognized concepts in combination, delivers virtualized cluster systems that make large pools of servers appear and act like a single virtual system. Along the way, we have cited real world examples of how these concepts have a tremendous, positive impact on rapid provisioning, manageability, scalability, security and reliability within the cluster. Table 1 below summarizes the key elements of the Scyld ClusterWare solution along with the delivered benefits of each capability.

Scyld ClusterWare Features	Delivered Benefits for Individual Features
Included CentOS Linux w/ option for Red Hat Enterprise Server	• ISV Application Compatibility • Flexibility to fit into your ecosystem
Complete, Architected and Integrated Solution that is commercially supported and works out of the box	• "Day 1" out-of the box productivity • No need to find and cobble together a solution • Ongoing expert support there when you need it to remain productive, competitive
Stateless Provisioning Architecture	• Installations/Updates on Master Node only • Add/Update/Re-Provision Compute nodes in seconds • If Applications do not require disks, save capital, operational costs and enjoy increased service availability
Lightweight Compute Operating Environment	• Enhanced Performance • 50x more memory for the application • Enhanced Scalability • Enhanced Security; no traditional attack points
Single System Process Space	• Cluster acts and feels like a single virtual machine • Single point of Installation • Single point of Management • Single point to Launch Jobs • Instant visibility into entire cluster on Master • Less scripting as both MPI jobs and non-MPI jobs intuitively scheduled to a virtual pool of processors • Supports JIT provisioning: enables compute nodes to provision on demand, only what they need
Scyld TaskMaster Virtualized Workload Management • Intuitive GUI- local/Web portal • Policy-based job scheduler & event engine • User Reservations/Triggers • Graphical reporting- live & historical • Built-in accounting for shared usage	• Instant Visibility across dozens of cluster parameters maximizes user productivity • Increase cluster utilization to 90%-99% • Prioritize workloads by organizational priorities • Helps plan current & future workloads

Table 1. Summary of Scyld ClusterWare Capabilities/Benefits

Built upon industry standard Linux distributions from CentOS and Red Hat for maximum ecosystem compatibility, Scyld ClusterWare extends the operating system platform to deliver an elegantly simple and powerful new paradigm of virtualized clustered computing. This new paradigm eliminates the need for multiple levels of cost and support and delivers everything needed for users and administrators to be productive immediately running HPC applications — out of the box. It also dramatically increases efficiency and reduces operating costs while delivering a dependable HPC service to your organization, thereby maximizing the return on investment for Linux Clustering in your highly competitive business environment.

For more information, visit www.penguincomputing.com or contact a Penguin Cluster Expert at:

1-888-PENGUIN (1-888-736-4846) or sales@penguincomputing.com