Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

WHITE PAPER

# HP's Strategy for Delivering Cluster Technology to Technical Computing Environments

Sponsored by: HP

Christopher G. Willard, Ph.D. Earl Joseph, Ph.D.

November 2005

## Introduction

Cluster technology has recently grown to take a significant share of the technical computing market. Rather than build or buy powerful special-purpose computers, members of the high-performance computing (HPC) community aim to exploit the opportunities provided by combining multiple general-purpose computers. IDC estimates that scientists and engineers increased their purchases of clustered systems from about $285 million in 1999 to about $2.6 billion in 2004. In addition, clusters currently account for more than half of all technical computer revenue. We believe the rapid adoption of cluster technology is driven by several interrelated factors:

- **Absolute node-level performance.** Individual nodes have become powerful enough to run many technical workloads. Clusters provide a way to leverage Moore's law improvements in industry-standard technologies, thus gaining exponential increases in component capability essentially for free.

- **Price/performance as throughput engines.** Clusters are highly cost-effective in capacity computing (or throughput) environments. In this case, multiple independent but relatively small jobs are run on separate nodes. Capacity computing environments require that jobs be small enough to fit on one or a few nodes.

- **Increasing availability for moderately parallel applications.** Advances in clustering technology combined with high-performance cluster interconnects position clusters to address moderately parallel applications effectively (i.e., applications that use roughly from 16 to 64 processors).

- **Cost-effective solution for highly parallel applications.** Clusters are among the most cost-effective solutions to problems with highly parallel applications. These applications, also referred to as *embarrassingly parallel* or *trivially parallel*, are typified by near-perfect data decomposition, little to no process-to-process communication or synchronization, and high compute-to-I/O ratios.

- **Support for fine-tuned capacity management.** The ability to quickly and easily add nodes to clusters provides a simple method to increase capacity in small increments. Thus, clusters can help system managers avoid the risks associated with overprovisioning, which will result in underutilization, or underprovisioning, which can result in project delays and missed deadlines.

### *IDC Definition of Cluster*

IDC defines a cluster as a set of independent computers combined into a unified system through systems software and networking technologies. In assigning systems to a cluster category, we use the following general rules:

- **Component independence.** Clusters are composed of systems that could operate on their own outside the cluster with minimal additional hardware and software (e.g., monitor, keyboard). This implies that the cluster contains both a physically distributed memory and multiple instances of operating system components.

- **Standard interconnect technology.** Cluster components are generally connected with an industry-standard technology, such as networking or I/O interfaces.

Clusters can be built by either customers or vendors. Customers can purchase cluster components separately and integrate them into systems at their sites. Design and integration responsibility falls to system buyers and not to the hardware, networking, and software component vendors.

Computer systems vendors also field preconfigured clusters or clustered products. These systems are architecturally similar to user-configured clusters. However, in this case, vendors are responsible for system design, configuration, integration, and certification.

### *Challenges When Deploying Clusters*

For end users, assembling cluster systems requires a mastery of multiple technologies. Servers, storage systems, network interconnects, and software must be acquired from multiple vendors and integrated into a working system. In addition, users may need to modify custom applications to run in parallel on their clusters. Packaged applications are often not certified for the cluster platform. Table 1 presents a detailed analysis of cluster design choices.

**TABLE 1**

Design Choices When Building High-Performance Cluster Systems

| Design Issues | Design Choices |
|---|---|
| Mode of use | • **Capacity clusters.** Will the cluster be throughput-intensive, with a large number of smaller jobs running on multiple partitions of a single cluster node or on multiple cluster nodes?<br>• **Capability clusters.** Will the cluster address the most demanding problems, that is, problems that require all or a large proportion of a single system's resources in order to be solved in an acceptable time frame?<br>• **Single application clusters.** Will the cluster be optimized to run one high-value application? Cluster capacity and capability can be matched to the application, often to offload general-purpose servers.<br>• **Multiple workload clusters.** Will the cluster support different types of workloads? If so, the cluster must be configurable for capability or capacity workloads. |
| Processor nodes | • **Processor types.** Which processor will the cluster use — RISC, x86 32-bit, x86 64-bit, or Itanium? Trade-offs are price, performance, and configurability. 64-bit processors allow for larger memory configurations to address more complex problems than 32-bit systems.<br>• **Size of nodes.** How many processors per node will the cluster have? Trade-offs are processor-to-memory performance, the application's absolute memory requirements, and overall price/performance.<br>• **I/O capacity per node.** Are nodes configured with a local disk? Do all nodes have equal access to attached or networked storage devices? Do all nodes support the file system software? Interconnect I/O bandwidth and latency also affect performance, as noted below.<br>• **Operating system.** Will the cluster nodes run Unix, Linux, or Windows? Node operating systems may collide with or match other legacy operating systems. |
| Interconnect | • **Balancing node and interconnect I/O performance.** Are processors and interconnects well matched? Balance is based on computation to communications, both bandwidth and latency.<br>• **Special-purpose direct interconnects.** Are higher-cost point-to-point networks needed for their gigabit speeds and high scalability?<br>• **General-purpose interconnects using PCI, PCI-X, or PCI-Express slot; InfiniBand; and/or Gigabit Ethernet.** Will lower-cost networks provide sufficient speeds and scalability? |
| Cluster software | • **Cluster system management.** Which software will manage the cluster as if it were a single computing resource?<br>• **Load balancing and job management.** Which software will keep the cluster nodes occupied? This functionality is particularly important for capacity workloads, but it is also important for highly parallel workloads.<br>• **Cluster file system.** What software and hardware will provide a cluster of servers with high-capacity and high-performance extended memory? Getting data to nodes quickly enough to keep the nodes busy is the challenge; staging data to the system is the solution.<br>• **Programming environment.** Which compilers, libraries, and other tools will be needed to support parallel programming both within nodes and between nodes?<br>• **Grid computing.** Will cluster management software extend to local grid management and/or easily interface with existing grid structures? |
| Cluster reliability | • **Cluster availability.** How will the cluster be made resilient and available? By their nature, clusters provide multiple points of failure at all levels of the system: interconnect, nodes, and system software.<br>• **Cost of downtime.** How many people depend on the cluster? Downtime cost may be related to any downtime or to downtime during the brief time boxes commonly assigned to engineering projects. |
| Physical infrastructure | • **Ongoing maintenance for components.** How many maintenance procedures are needed? Note that clusters comprise a large collection of smaller systems.<br>• **Environmental concerns.** What additional equipment and costs will accrue from floor space, power, and cooling requirements? |

Source: Derived from *Technical Cluster and Grid Taxonomy, Part 1: Definitions and Clusters*, IDC #27704

IDC's interviews and discussions with cluster administrators and users have identified several factors that can reduce cluster effectiveness or inhibit use:

- **Strategic fit.** Can the cluster configuration effectively support the application set and meet user requirements? Configuration of the processor nodes (e.g., number of CPUs, amount of memory), performance of the cluster interconnects, and access to mass storage are determining factors in overall application efficiency.
- **Management complexity.** Customers cite cluster management as a major challenge to cluster use and overall cost of ownership. Users are looking for more complete and effective system management toolsets. If clusters can be managed more simply, then the cost of system operations, staffing, and training will be reduced. Similarly, with better management tools, clusters can deliver higher system efficiency, that is, greater performance as a fraction of peak performance.
- **Complexity of parallel programming.** Although it is possible to build a computer of arbitrary complexity, the ability of programmers to work with complexity remains a constant. Parallel applications that do not fall into the *embarrassingly parallel* category can be extremely difficult to develop and verify. In addition, applications may need to be retuned for new cluster configurations or changes in component performance. The software complexity issue extends to the availability of parallel applications provided by independent software vendors (ISVs). Although ISVs are providing parallel applications in some areas and are working to parallelize their applications in other areas, users still note a lack of general availability for these tools. In many cases, the license fee structures for ISV software inhibit scaling.

In summary, cluster systems represent a broad set of computing configurations, with different configurations optimized for different workloads and application types. Successful implementation of cluster strategies requires consideration of system management capabilities and costs and system use modes. Simple savings in hardware can easily be counterbalanced by increased spend on system management, maintenance, physical support, and other operational inefficiencies.

## HP's Strategy for High-Performance Cluster Products

HP's overall strategy for high-performance computing is to leverage the price/performance advantages of volume products with the high-performance advantages of new technology (see *HP's Strategy for High-Performance Computing*, IDC White Paper #3822) The high-performance technical computing community provides a technology test bed that encourages HP to build and field more powerful systems. Mainstream users create a volume market that allows HP to gain economies of scale for foundation technologies (i.e., core technologies for servers, storage systems, networks, and operating environments).

### *HP Unified Cluster Portfolio*

The HP Unified Cluster Portfolio is a family of integrated computation, data management, and visualization technologies that can be combined to create a range of cluster-based solutions. The products comprise multiple platforms, operating systems, software systems, and network interconnects. Products in the portfolio are built with common components and shared architectural principles.

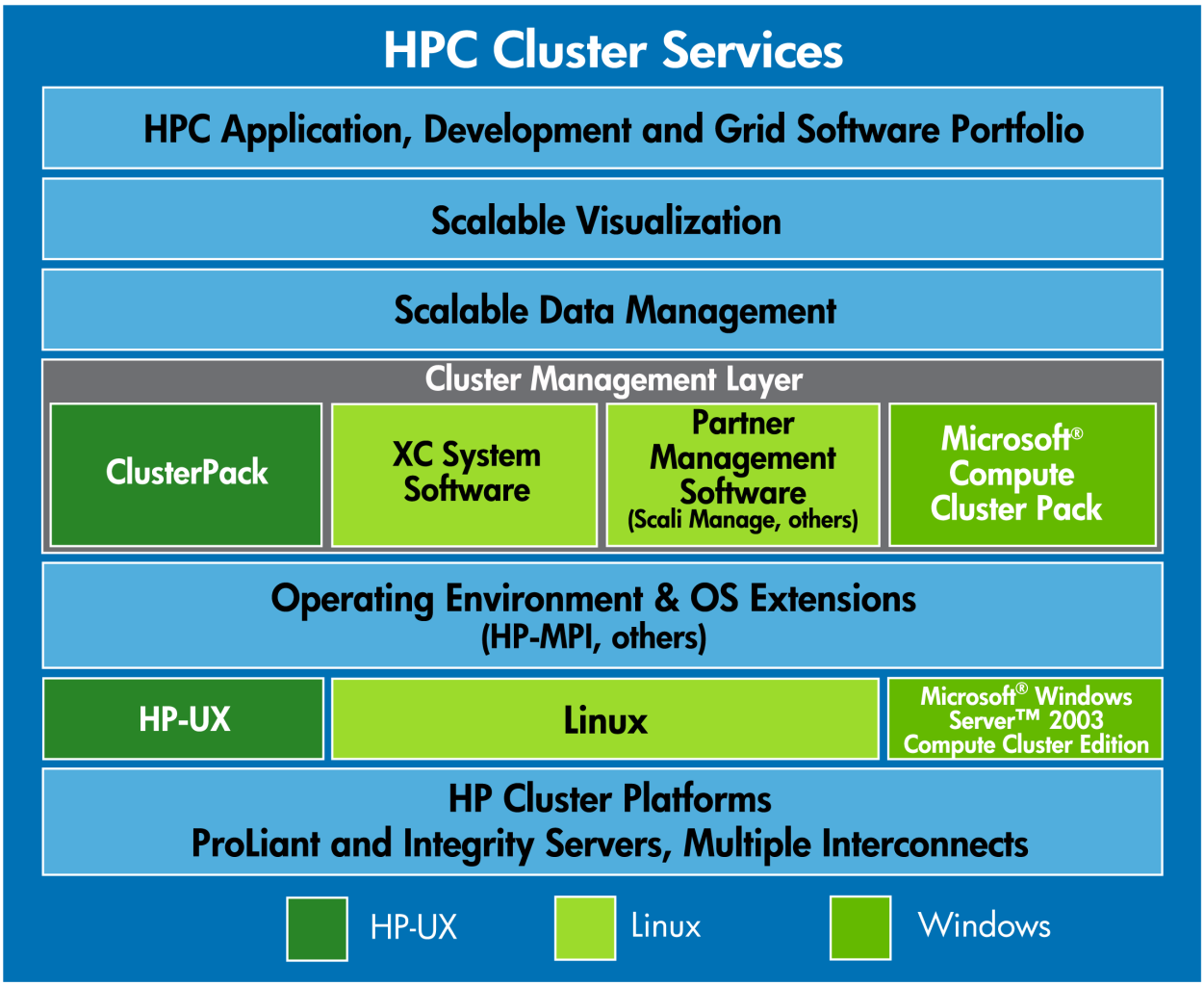
 

A key objective for the HP Unified Cluster Portfolio is to reduce the risks and challenges associated with designing and deploying a cluster. By preassembling clusters of several kinds, HP can offer customers nearly turnkey hardware and software solutions. HP expects that with added ease of deployment and operations, cluster technology will serve the needs of the most demanding high-performance computing customers and of a broader audience of users in the middle or mainstream market.

Figure 1 provides an overview of the layered architecture and technology choices offered in the HP Unified Cluster Portfolio.

**FIGURE 1**

The HP Unified Cluster Portfolio




Source: HP, 2005

The foundation layer, HP Cluster Platforms, comprises high-performance technical computing servers drawn from HP ProLiant and Integrity server product lines, including HP BladeSystems, and configured into build-to-order cluster hardware solutions, with multiple operating system and interconnect choices. Taken together, HP ProLiant and Integrity servers and HP BladeSystems support three major operating systems: Microsoft Windows, Linux, and HP-UX.

HP provides a common operating environment (COE) consisting primarily of HP-MPI message passing middleware to enable developers and ISVs to support an application across all of HP's processor, server, operating system, and interconnect choices, including ProLiant and Integrity servers and BladeSystems running Linux and HP-UX. HP will also provide support for HP-MPI on Microsoft Windows Compute Cluster Server 2003 at or shortly after its release.

HP defines a series of cluster software reference stacks that span all of the layers of a cluster solution, with HP qualified combinations of hardware and software components to ensure that they work together. By predefining these reference stacks, and basing them on an HP Cluster Platform, the company is able to lower the risk of cluster deployment. These stacks include:

- **HP-UX Cluster Reference Stack.** An integrated cluster management software stack for the HP-UX 11i v2 operating system and HP Integrity servers, this reference stack supports the Itanium 2–based Cluster Platform 6000, HP's Technical Computing Operating Environment (TCOE), and HP Cluster Pack — a collection of management utilities for clusters based on Integrity servers running HP-UX.

- **HP XC Cluster Software.** This integrated cluster management software stack for HP Cluster Platforms is based on ProLiant and Integrity servers running Linux. This product is based primarily on open source software and includes a Linux distribution designed to be compatible with Red Hat Enterprise Linux v4 (RHEL), HP-MPI, and fully integrated resource management based on Platform Computing's LSF. It forms a Linux cluster solution that is integrated, qualified, and supported by HP.

- **Novell Validation Suite.** This Linux cluster solution, developed as a collaborative effort between Novell and HP, includes SUSE SLES 9 and ISV and HP qualified middleware and industry applications. The suite includes Scali Manage from Scali Inc. as the cluster management solution, a choice of Scali Connect or HP-MPI as the message passing library, Altair Engineering PBS Professional for resource management, PolyServe Matrix Server and Cluster Volume Manager for data management, and a variety of development tools and utilities.

- **Microsoft Windows Compute Cluster Server 2003.** This new product will be introduced by Microsoft in the first half of 2006 and will be available for Cluster Platform 3000 and 4000 systems. It is made up of Windows Server 2003 Compute Cluster Edition (CCE), a 64-bit operating system derived from Windows Server 2003, Standard x64 Edition, and Microsoft Compute Cluster Pack (CCP). CCP provides tools to deploy and manage a cluster (e.g., Job Scheduler, MPI) in an environment that is highly integrated with existing Windows infrastructure (e.g., Active Directory, Microsoft Management Console).

- **Open source cluster solutions.** Packages such as NPACI Rocks or OSCAR can be deployed, providing a low-cost alternative for users with resources and expertise to maintain custom environments. In addition, Platform Computing is announcing Platform Rocks support for HP Cluster Platform 3000 systems at SC2005, providing an open source cluster management solution with available vendor support from Platform.

Other elements within the portfolio include:

- **Data management solutions.** Many cluster applications have requirements for large volumes of data to be available across the cluster, for which a number of cluster file systems have been developed. HP provides its StorageWorks Scalable File Share (HP SFS) and HP Cluster Gateway, as well as products from partners, including ADIC, IBRIX, PolyServe, Terrascale, VERITAS, and others. The HP SFS solution is an integrated hardware and software solution for high-bandwidth and high-capacity cluster data management. It is based on open source Lustre technology combined with HP StorageWorks systems and HP servers. SFS is supported for all HP Cluster Platforms running Linux.

- **HP Scalable Visualization Array (SVA).** Technical applications tend to generate large volumes of data, and users look to visualization solutions to display, manipulate, and interpret that data. HP is developing visualization technology utilizing industry-standard visualization nodes and graphics cards with HP-developed SVA technology. SVA is designed to integrate closely coupled workstations within a cluster and to provide parallel rendering and compositing that enable visualization of very large data sets, very high resolution images, and dynamic data including computational steering. The HP Scalable Visualization solutions are closely coupled with the HP Unified Cluster Portfolio.

- **Development environment.** Through product development, partners, and open source collaborations, HP provides a development environment for building high-performance applications for shared memory and distributed memory computers. HPC unique components of HP's product set include HP-MPI message passing library, Unified Parallel C (UPC) compiler, HP MLIB mathematics library, and others.

- **Grid and resource management.** HP has historically contributed to grid development through work at its HP Labs. HP participates in the grid consortia, grid-enables its products, partners with grid and resource management ISVs, and delivers grid implementation services through its consulting and integration services organization.

- **Application portfolio.** HP and its partners provide a broad portfolio of applications in target application segments, including government and defense, scientific research, computer-aided engineering, life and materials sciences, and geosciences and energy.

- **Services.** For all layers and components in the HP Unified Cluster Portfolio, members of HP Services provide a spectrum of design, installation, training, and support services.

The following sections describe each offering in the HP Unified Cluster Portfolio in greater detail.

### *HP Cluster Platforms*

An HP Cluster Platform is a factory preassembled product comprising racks of server nodes, interconnects, network interfaces, and (optionally) software. Thus, HP delivers fully integrated, tested, and supported turnkey systems rather than components to be assembled by HP's customers.

The catalog of offerings includes platforms ranging from 4 to 512 nodes, with larger systems available by special request. Three Cluster Platform solutions released in 2004 are being updated for 2005. The revisions include new models of systems, a new blades-based cluster platform, and a new visualization cluster platform:

- **HP Cluster Platform 3000.** This cluster is based on the ProLiant DL140 G2 and DL360 G4 servers and xw8200 workstations with dual Intel Xeon processors with EM64T technology. Node options include:
  - **ProLiant DL140 G2.** This 1U server is configured to address HPC requirements for high price/performance by eliminating features not generally required by technical users (e.g., hot-swap disk drives). System management is based on the Intelligent Platform Management Interface (IPMI) v2 system monitoring protocol.
  - **ProLiant DL360 G4.** This 1U server is configured to address HPC requirements for performance and functionality and includes such features as HP's Integrated Lights Out (iLO) management hot-swap disk drives and redundant power supplies.
  - **HP xw8200 workstations.** These nodes support remote visualization, scalable resolution for multipanel displays, and distributed rendering. When not being used for visualization, these nodes are part of the general-purpose compute resources in the cluster.
- **HP Cluster Platform 4000.** This cluster is based on the ProLiant DL145 G2 and DL585 servers and xw9300 workstations with two or four AMD Opteron processors, including dual-core processors. This series also includes a blade-based cluster using BL35p nodes. Node options include:
  - **ProLiant DL145 G2.** This 1U server has dual AMD Opteron 200 series processors, including dual-core CPUs, configured to address HPC requirements for high price/performance by eliminating features not generally required by technical users (e.g., hot-swap disk drives). System management is based on the Intelligent Platform Management Interface (IPMI) v2 system monitoring protocol.
  - **ProLiant DL585.** This 4U server features quad AMD Opteron 800 series processors, including dual-core CPUs, configured to address HPC requirements for four-processor scalability, performance, and functionality, and including such features as HP's Integrated Lights Out (iLO) management, hot-swap disk drives, and redundant power supplies.

- **ProLiant BL35p.** This p-Class, double-density blade server is based on dual AMD Opteron 200 processors, including dual-core CPUs. The p-Class BladeSystems use a 6U blade chassis that can accommodate up to 16 BL35p blades.

- **HP xw9300 workstations.** In early 2006, HP will bring HP Cluster Platform 4000 to market based on the xw9300. These systems will support remote visualization, scalable resolution for multipanel displays, and distributed rendering. When not being used for visualization, these nodes are part of the general-purpose compute resources in the cluster.

**HP Cluster Platform 6000.** This cluster is based on the HP Integrity rx1620 and rx2620 servers with dual Intel Itanium 2 processors, providing 64-bit floating point performance. Node options include:

- **Integrity rx2620.** This 2U server provides dual Itanium 2 processors based on HP's zx1 chipset for high-performance, iLO management, and RAS.

- **Integrity rx1620.** This 1U server is configured for high processor density with dual Itanium 2 processors based on HP's zx1 chipset for high-performance, dense configurations, iLO management, and RAS.

The HP Cluster Platforms support HP-UX (11i v2) on the CP6000 platform and Linux (Red Hat EL3 and EL4, SUSE SLES 9) (on the CP3000, CP4000, CP4000BL, and CP6000 platforms). Support will be provided for Microsoft Windows Compute Cluster Server 2003 when it ships. (This operating system will be available on the CP3000, CP4000, and CP4000BL platforms.)

Interconnect choices for HP Cluster Platforms are Gigabit Ethernet (on the CP3000, CP4000, CP4000BL, and CP6000 platforms), Myricom Myrinet (on the CP3000 and CP4000 platforms), Quadrics QsNet (on the CP4000 and CP6000 platforms), and InfiniBand (Voltaire) (on the CP3000, CP4000 and CP6000 platforms). These platforms provide the foundation for all HP cluster products. HP will integrate emerging nodes and interconnect technologies into its Cluster Portfolio.

### Computing Operating Environment for HP-UX

HP offers a Technical Computing Operating Environment (TCOE) that is based on the current version of its HP-UX 11i operating system and includes features targeted for compute-intense applications, such as HP MLIB scientific and math libraries, support for parallel programming including HP-MPI, and support of drivers for the InfiniBand interconnect. The TCOE also includes such basic Internet tools as Apache Web server, Netscape Communicator, Java Runtime Environment, and Java Developer's Kit, as well as support for the Common Internet File System (CIFS) — a file-sharing protocol that client workstations use to request file access from servers.

HP-UX systems also use HP MC/ServiceGuard to address issues of system resilience and scalability. For technically oriented large-batch application loads, HP offers solutions through partners such as Platform Computing (LSF) and Altair (PBS/Pro).

**Computing Operating Environment for Linux-Based Clusters**

For Linux-based cluster products, HP augments the Linux distributions from Red Hat and SUSE. HP supports its MPICH-compatible message passing interface (HP-MPI) and mathematics library (HP MLIB) for the Linux environment. By providing HP-MPI and HP MLIB across all of the HP Cluster Platforms, HP enables software developers and ISVs to support its multiple system architectures, interconnects, and operating systems (Linux and HP-UX) with minimal effort. This is a major factor in HP's commitment to provide an extensive ISV portfolio for the HP Unified Cluster Portfolio.

HP's clusters also leverage the HP Systems Insight Manager (SIM) and the integrated server management processors embedded in the ProLiant and Integrity servers. The core Systems Insight Manager software delivers the essential capabilities required to manage all HP server platforms, including device management with plug-ins for rapid deployment, performance management, partition management, and workload management.

### *HP XC Clusters*

HP XC Clusters are a family of integrated, scalable Linux clusters, combining XC System Software with predefined implementations of HP Cluster Platforms 3000, 4000, and 6000. XC Clusters are industrial-grade, production-ready systems specifically configured for high-performance technical computing. These systems provide a comprehensive, supported solution for serial and parallel applications.

The XC System Software is based on a standard release of Linux. This release is enhanced with clustering technology from the open source community. HP adds capabilities through cluster tools and utilities developed by HP and its partners.

The XC Cluster environment works to combine a collection of servers into a single, scalable production system through such features as unified scheduling and a single point of administration and control. The XC System Software includes resource management, scheduling, and job launch technology from such partners as Platform Computing and the open source–based Simple Linux Utility for Resource Management (SLURM). Access to existing applications is preserved through use of underlying Linux API/ABI and integration of HP-MPI standards. HP provides an interconnect-neutral (but MPICH-compatible) application MPI library.

With the open source Lustre file system, XC Clusters provide high-performance and highly available, global I/O capability. HP's implementation of Lustre, which is called HP StorageWorks Scalable File Share, is fully integrated with HP XC Clusters. XC System Software also supports NFS, including both server and client functionality.

XC Clusters are HP's recommended offering for large-scale, compute clusters. They are being deployed at several top 500 sites. In July 2005, HP announced the installation of an XC Cluster based on a 1,024-node CP4000 and InfiniBand at the Aeronautical Systems Center for the U.S. Department of Defense. In addition, in June 2005, HP announced plans to deploy this technology in the expansion at Canada's SHARCNET, with over 2,000 new nodes going into this grid-based consortium.

### *Partner Options for Linux Clusters*

HP looks to open source and partner cluster management software to provide customers with additional cluster management choices and expand the market for HP Cluster Platforms into multiple sales channels. HP has qualified the open source NPACI Rocks across the CP3000 and CP4000 platforms. Open source offerings, including OSCAR and NPACI Rocks, have been installed by customers and by HP's consulting and integration services organization.

HP established a strategic partnership in 2002 with Scali to deliver Scali Manage and Scali MPI Connect on Linux-based HP servers and Cluster Platforms. HP and Scali collaborate to test and qualify Scali Cluster solutions on HP Cluster Platforms. In addition, as part of the Novell Validation Suite, Scali provides heterogeneous cluster capability where customers have mixed servers and operating systems environments. In these cases, Scali provides a common computing utility tools for managing and monitoring the multiple systems. Scali Manage provides tools for system installation, configuration, management, and monitoring in addition to support for the cluster interconnects and platforms. Scali Manage is fully integrated with the iLO and IPMI management infrastructure provided on HP Cluster Platforms.

Additional partner products available for HP Cluster Platforms include:

- Grid and resource management products from Altair, Axceleon, DataSynapse, Platform Computing, TurboWorx, and United Devices
- Data management solutions from ADIC, IBRIX, PolyServe, Terrascale, VERITAS, and others
- Development tools from Intel, Portland Group, PathScale, Etnus, Allinea, and others

### *HP-UX Integrity Clusters*

The most mature HP cluster offerings are products based on HP-UX and ClusterPack. The software stack rests on the HP Cluster Platform 6000 built with servers from the Integrity line. ClusterPack provides cluster management functionality, including automated installation and configuration as well as single point of system administration, and it also manages the cluster's distributed resources. ClusterPack offers a set of tools that includes system image creation and installation services, cluster event monitoring, configuration and management, clusterwide command line interface (CLI), and comprehensive summarization support. HP-UX Integrity Clusters can support an HP Cluster Platform 6000 with up to 128 rx1620 or rx2620 server nodes with a total of 256 Itanium 2 processors. ClusterPack also supports the full range of Integrity servers up to and including the Integrity Superdome server.

### *HP Windows-Based Clusters*

HP will support clusters using HP Cluster Platforms 3000 and 4000 with the Windows operating system. Microsoft has released the beta version of Windows Compute Cluster Server (CCS) 2003 for HPC. Windows CCS is a two-component product: The first component is Windows Server 2003 Compute Cluster Edition (CCE), a 64-bit operating system derived from Windows Server 2003, Standard x64 Edition. The second component is the Microsoft Compute Cluster Pack (CCP), which provides tools to deploy and manage a cluster, including a Job Scheduler, MPI, and consoles and interfaces to manage cluster nodes and jobs. These tools are integrated with existing Windows infrastructure. For example, users and security are managed using Active Directory, and the cluster management consoles utilize Microsoft Management Console.

HP is targeting its Windows cluster solutions at organizations with primary computing environments based on Windows and at organizations that are using applications optimized by ISVs and Microsoft for Windows 2003.

### *HP StorageWorks Scalable File Share*

HP StorageWorks Scalable File Share (HP SFS) is a file server for Linux clusters that delivers scalable bandwidth to a shared file system. The ability to increase bandwidth helps to avoid or eliminate I/O bottlenecks. The file server runs a combination of open source Lustre technology and HP software. It is built from a combination of HP ProLiant servers and HP StorageWorks SFS20 disk arrays. HP SFS is designed to scale in both storage capacity and bandwidth based on building blocks that HP calls data smart cells and metadata smart cells:

- Data smart cells are the main data storage for HP SFS. The smart cells are servers supporting SCSI-connected SATA disk arrays that are coupled to a cluster interconnect fabric. The smart cells are deployed in pairs with cross-connected storage for redundancy and resiliency. Smaller files are typically stored on individual cells, and larger files are striped across multiple cells. Striping is a method for scaling I/O bandwidth at the file level, as multiple cells with multiple data servers and network interconnections can work simultaneously.

- Metadata smart cells store information about the files, such as file names, directory structures, file status, and security information. Pairs of metadata smart cells are cross-connected to disk arrays and to the cluster interconnect fabric. Metadata cell servers are also designed to support failover for improved system resiliency.

Users scale HP SFS systems by adding data cells and metadata cells. As a result, bandwidth is added along with capacity. When new data cells are added, existing files remain striped across existing cells. New cells join with existing cells as candidates for file striping, thus providing performance scalability as well.

Storage resiliency options include the use of RAID5 disk arrays cross-connected to two smart cells to maintain redundant paths to the data. For even higher resiliency, HP SFS offers this RAID5 storage mirrored (i.e., RAID5 + RAID1).

Client software runs on the cluster served by HP SFS and presents a POSIX-compliant file system interface to applications. Lustre software maintains consistency among buffers in the many distributed clients so that file storage behaves logically as if it were a single shared memory buffer on an SMP server.

### *HP Scalable Visualization*

HP Scalable Visualization Array (HP SVA) is a hardware and software system that is based on open source technology and designed and implemented to transform traditional serial graphics and render applications to run on parallel computer architectures. The SVA is closely coupled to the cluster's computational and data management resources. HP SVA is based on workstation-based HP Cluster Platforms and is fully integrated with the XC System Software, HP's cluster management software. All HP SVA nodes are equipped to be Scalable File Share clients. HP SVA technology supports multiple displays, immersion applications, 3D perspectives, and stereo displays. HP clusters with SVA nodes work as both visualization and compute clusters with the SVA nodes being available for computational operations when not addressing visualization tasks.

HP SVA closely couples visualization to the cluster's scalable computation and data management. The parallel approach used by HP SVA supports:

- Visualization of very large data sets (Data can be rendered on parallel workstations within the cluster, and more workstations can be added as data set size increases.)
- Visualization of very high resolution images, including the use of immersive technologies such as multidisplay walls and caves
- Visualization of dynamic data, including user interaction with the simulations producing the data, sometimes called computational steering

The HP Remote Graphics Software (HP RGS) package extends HP SVA functionality to remote desktop or laptop systems, allowing users at remote sites to access the same computational power that local users access. This virtualization of the clustered workstations offers a remote working environment. In principle, the same hardware can be used in 8-hour shifts 24 hours a day, with the users in various areas of the world. This enables collaboration and simplifies remote access to cluster resources, applications, and data.

SVA can operate in multiple modes:

- Individual workstations can be virtualized with up to four virtual workstations that are accessed remotely via HP RGS. This allows the cluster's visualization resources to be efficiently shared around an organization or grid.
- Multiple workstations can be used together to support parallel rendering. This enables the visualization of very large or high-resolution data sets beyond the capabilities of a single workstation.
- Multiple workstations, possibly with the addition of a second interconnect acting as a pixel network, can be used together to support parallel rendering and compositing. This enables visualization of very large or high-resolution data sets, combined with the ability to support dynamic graphics, such as rotation of very complex 3D images or dynamic data with higher frame rates.

### *Development Environments*

High-performance technical computing is accomplished with programming models that are different from business data processing models. Technical organizations tend to have a higher level of custom program development. System performance and application functionality must be continually upgraded to support ever-changing scientific and engineering requirements.

Developing technical applications requires optimizing compilers, highly tuned scientific and mathematics libraries, support for parallel applications development, debuggers to hunt down problems, and analysis tools to optimize performance. HP provides a suite of development tools and environments in support of technical applications development:

- **Compilers and libraries.** HP provides HP-UX compilers and development environments for Fortran, C/C++, HPC, and Java. In addition, HP supports MPI standards with its HP-MPI. HP's Math Library, MLIB, is available with HP-UX. Open source GNU compilers and compilers from Intel, Portland Group Compiler Technology, and PathScale provide similar functionality under Linux. HP also supports higher-level languages, such as MATLAB and Mathematica available from partners The MathWorks and Wolfram Research, respectively.
- **Debuggers.** HP's WDB debugger analyzes problems with multithreaded programs. Debuggers are also available from HP partners, such as Etnus' TotalView, which can trace parallelism embedded in message passing as well as debug multithreaded programs. DDT from Allinea Software (formerly Streamline Computing) provides graphical views of parallel code.
- **Tuning and performance analysis tools.** HP Caliper provides performance analysis support for code that runs on Itanium processors. MPI Trace Analyzer and Trace Collector and other tools are available from Intel. A number of open source products, such as TAU from the University of Oregon, are also available.

HP's strategy for development environments is to integrate and augment its own product line with offerings from both the open source community and HP partners in order to provide customers with a wide range of choices.

### *Summary of the HP Unified Cluster Portfolio*

The HP Unified Cluster Portfolio offers customers integrated cluster systems that provide computation, data management, and visualization capabilities and that are packaged for immediate deployment. Servers using Opteron, Xeon, and Itanium processors and interconnect technology from HP and others provide platforms for alternative software stacks that use HP-UX, Linux, and Windows operating systems. Customers can choose among cluster management software products from HP and others. For Linux-based clusters, HP offers StorageWorks Scalable File Share (HP SFS) and the new HP Scalable Visualization Array.

### *Future Directions for the HP Unified Cluster Portfolio*

HP's High Performance Computing (HPC) division has a formal advanced development program that fosters collaboration with strategic customers and partners to extend HP's clustering environment and to add new technologies. Through its advanced development programs, the HPC division has introduced compilers, languages, tools, operating system improvements, and cooling management techniques with the assistance of research sponsorship. Major past programs include work on high-performance cluster interconnect networks, scalable data visualization hardware and software, and high-performance parallel file systems research.

Other HP HPC-related collaborative efforts include:

- **The HP Collaboration and Competency Network (HP CCN).** HP CCN is a forum that facilitates collaboration, innovation, discovery, and competency sharing between HP HPC customers and partners. The goal of HP CCN is to enable members to share experience and expertise in order to more rapidly advance research, development, and use of HPC technologies in areas such as computational and data grids, Linux, global file systems for Linux, and scientific visualization.

- **HP University Relations.** HP University Relations has more than 70 active collaborations with institutions in the United States, Europe, and Asia. It sponsors and collaborates with outside research institutions for the advancement of public knowledge. Such sponsorships may include grants for academic study or equipment, or they may join together HP Labs, HP product divisions, and nonacademic sponsors. HP University Relations contributes to HPC with collaborations on architectures, methods, compilers and languages, and application methods for solving large-scale problems.

- **HP Labs.** The function of HP Labs is to deliver new technologies and create business opportunities that go beyond HP's current strategies. The HPC division and HP Labs are collaborating in areas such as Linux, grid computing, utility computing, system and storage architectures, parallel programming tools and methodologies, security, and biosciences.

  The Smart Cooling research program is an example of the collaboration with HP Labs. Compute clusters, with large numbers of processors in high-density configurations, create significant thermal challenges for systems and datacenter design. HP Labs has developed a set of services for datacenter thermal management using computational fluid dynamics (CFD) and 3D techniques to model thermal characteristics of datacenters, including the computing and cooling systems within them. This model can be used to improve the design of datacenters as well as to provide dynamic modeling that enables datacenter managers to adapt quickly to changes in system usage or cooling availability. The goal of HP's cooling services is to reduce high cooling costs caused by delivering too much cooling and to maximize system reliability by delivering adequate cooling overall or to system hotspots.

## IDC Analysis

### *Technology Versus Market Structure*

Cluster technology is unique in technical computing markets in that designing and assembling a cluster can be done at any level in the market, ranging from end users working on nights and weekends up through the largest vertically integrated computer solutions company. Prior to productized clusters, the market for cluster technology was limited to those organizations with large R&D budgets, strong technical skills, and manufacturing capabilities.

One indication of the importance of this technology is its growing market share. In 2000, IDC estimated that clustered systems sales accounted for 5% of the total HPC market. By 2004, cluster systems accounted for 36% of the market. By mid-2005, clusters represented just over 50% of the market, and they are expected to continue to capture market share through the rest of the decade.

### *Cosmological Market View of Clusters*

Table 2 presents IDC's *cosmological* model of the cluster market. This model was designed to help estimate the overall size of the cluster market; however, it also provides a view of actual market structures. Our framework is drawn from the classification of matter in the universe. Types of matter are as follows:

- **Bright matter.** This category refers to objects that glow and can thus be easily seen. By analogy, bright clusters are systems sold by vendors as clusters and can be easily counted.

- **Dim matter.** This category refers to objects that do not generate their own light but can be seen by light reflected from or through them. HPC vendors sell dim cluster components (e.g., individual nodes), and end users or integrators assemble the cluster.

- **Dark matter.** This category refers to objects or particles that cannot be seen but are largely deduced through theory. Dark clusters are either sold by smaller second-tier computer vendors or integrators, or they are assembled by end users from components purchased through standard commodity channels.

- **Invisible matter.** We do not know of any invisible matter *per se*; however, clusters can also be developed by recycling components that have outlived their usefulness in their original tasks and add no sales revenue to the market. These clusters can be viewed as acting as a damper on the market. Thus, they could be designated as black holes.

IDC's long-term objective is to throw as much light on the dim and dark clusters as possible and move them into the bright category.

These options for acquiring clusters provide end users with a very broad range of purchasing choices. Historically, HPC customers largely created this market — with virtually all early clusters being assembled by end users and thus falling into the dark, dim, and invisible categories described in Table 2. As the technology gained momentum, systems integrators and second-tier vendors entered the market. These companies provide combinations of integration skills, specialized system management, application knowledge, and, in some cases, highly evolved products. In the most recent phase, traditional technical systems vendors (such as HP) have offered the market a broad array of hardware, software, and services options as well as the ability to extend cluster technology through R&D investments.

**TABLE 2**

Cluster Market Model Topology

| Category | Description |
|---|---|
| Clustered products<br><br>(Bright clusters) | Systems in this category comprise actual models offered by systems vendors or integrators. Also counted in this category are clusters of servers or workstations that a vendor also sells as standalone units but without a specific product designation. In both cases the cluster is assembled and delivered to the client by the vendor and the vendor is able to identify that the system is sold as a cluster and can account for all components in the cluster as a single system sale. |
| Clustered technical systems<br><br>(Dim clusters) | Systems in this category comprise clusters that end users assembled from systems sold as technical servers. The vendor accounts for and reports the component servers as individual technical systems sales, and the cluster is subsequently accounted for as multiple shipments of lower-priced systems rather than a single larger system sale. |
| Off-the-shelf clustered systems<br><br>(Dark clusters) | Systems in this category comprise clusters that end users assembled from components purchased through nontraditional channels. Components can include boards, PCs, workstations, and/or servers. Sources can range from off-the-shelf sales from consumer outlets, to direct sales over the Web, to systems sold through standard vendor channels but simply not accounted for or reported as technical system shipments. |
| Repurposed clusters<br><br>(Invisible clusters) | These clusters are built out of previously purchased and (effectively) fully amortized components; in other words, PCs, workstations, or low-end servers not originally purchased for use in a technical cluster, but which have been reassigned to use in clusters. Reassignment can result from any of a wide variety of reasons ranging from systems being freed up as projects end to older systems being replaced by newer systems. Invisible clusters generate virtually no new revenue in the market. |

Source: IDC, 2005

IDC believes that, over time, the dark and dim cluster markets will decline as clustered products provide advantages in technology sophistication, ease of implementation, and long-term product support. As clusters replace a larger portion of the technical server market, users will expect and purchase more complete solutions. We also expect to see the stronger second-tier vendors become more visible in the market and join the bright cluster ranks. Finally, we expect dark clusters to remain of interest to those customers who require highly specialized system configurations and who have the skill and time to develop these systems.

During this time, the challenge for users will be to choose a cluster implementation strategy that balances such factors as implementation time, internal management and integrations costs, long-term support, system software growth paths, system reliability, initial system price, and total cost of ownership.

### *Challenges for Vendors*

The combination of high rates of technology innovations, fluidity of system architecture designs, ease of cluster constructions, and the availability of open software generates a volatile market. Vendors such as HP will face a number of challenges in the evolving market for high-performance cluster systems:

- **Competition everywhere you look.** Cluster technology provides a relatively easy entry path into technical markets. Thus, new competitors can be expected to continually appear in the market. These new entrants will require established vendors to compete against a variety of cluster marketing strategies, including:

  - **The margin squeeze.** Smaller new entrants may accept lower margins, thus reducing the profits from clusters for larger companies, placing pressure on their business models, and potentially sending them in search of greener pastures.

  - **Loss leaders.** Better-funded competitors may take the low-margin strategy one step further by engaging in market share buying. In this case, a vendor sells systems at or near cost in order to attract customers, gain market share, and eliminate less well-funded companies. Once loss-leader companies have captured share, they plan to reinstate healthy margins in a market with a reduced level of competition. Unfortunately, users generally resist any such price increases, and new market leaders may find that they have helped to establish a permanent low-margin market.

  - **Niche snipers.** Smaller entrants with specialized knowledge or skills may target specific subsegments of the market where they can demonstrate a differential advantage. The effect of such sniping is to ultimately reduce the size of the market and to isolate general-purpose vendors from more profitable subsegments.

- **Keeping pace with technology.** The cluster market will continue to evolve over the next several years as new systems software, interconnect, and processor options become available. We believe that technology will continue to advance, at least over the short term. Such technical fluidity will require vendors to continually incorporate new functionality into their product lines and at the same time continue to support client investment in earlier solutions. Such rapid changes in technology can lead to bloated product portfolios, high internal training costs, and a combinatorial explosion of product offerings.

- **Being different when everything is the same.** At their cores, standards and cluster concepts provide all competitors with the same set of technologies with which to build their systems. By and large all players in the cluster areas will have access to the same processors, the same interconnects, and the same basic system software tools. In such environments, differentiation may seem impossible. Thus, vendors will need to expand their strategies for differentiating their products. We believe that, to a growing extent, future vendors will differentiate their products based on nontechnology factors such as reducing integration risk and costs, certifying solutions stability and reliability, reducing operations costs, and providing long-term comprehensive support.

### *HP Meeting the Challenges*

We believe that HP's clustering product strategy contains the elements necessary to compete in the technical cluster market over the long term. These elements include:

- **Corporate commitment.** HP views technical computing as a laboratory where new information technology concepts are developed, tested, and deployed. Such a view allows the company to continue to support technical markets over the long term. That said, we believe that HP also views technical computing as a real market opportunity that should provide acceptable levels of return on investment.
- **Broad product line.** HP's product strategy is designed to incorporate a relatively large number of technology options at the processor, interconnect, and system software levels. Such a strategy should position the company to advance its products in step with advances in technology and changes in market direction.
- **Partnerships.** IDC end-user research indicates that it is unlikely that any single vendor can provide all elements of a complete cluster product solution and that end users look to primary vendors to put them in touch with organizations that can provide specialized knowledge and products. HP's overall strategy leverages the company's ability to extend its technology and product expertise through partnerships with companies that can add value through application and industry knowledge, local support, system software packages, and so on.

## Conclusion

Technical computing vendors must maintain a close focus on their overall mission, which is to provide the tools, such as clusters, that scientists and engineers need in order to address the next generation of problems. We believe that over the long term successful vendors will continue to advance the state of the art as opposed to slipping into a margin war by simply providing less expensive cycles to address the current generation of problems.

HP has developed a comprehensive strategy to address clustering requirements for the technical computing community. The company's strategy shows that HP takes technical computing seriously and plans to compete aggressively in this market. The breadth of HP's offerings, in terms of platforms, software products, partnerships, and service capabilities, is indicative of HP's commitment to the high-performance technical computing market.

## Copyright Notice