

IBM @server Cluster 1600 for High Performance Computing



IBM@server Cluster 1600

Highlights

- **Offers scalable, high-performance parallel computing systems for numerically and I/O intensive workloads**
- **Builds on proven IBM RS/6000® SP™ technology to deliver high performance computing solutions**
- **Reduces the complexity and total cost of ownership for managing multiple systems**

Moving in parallel

One of the best approaches to solving scientific and technical computing challenges is parallel processing. By harnessing the power of many microprocessors, it is possible to solve computationally or I/O intensive problems that would simply take too long to complete on a single symmetric multiprocessor (SMP) system, no matter how powerful.

In particular, organizations have found that clustering (i.e. inter-connecting two or more computers into a single, unified and managed

computing resource) provides an approach to parallel processing that yields not just superior performance, but excellent price/performance as well.

However, clusters present their own set of challenges. Simply grouping together hundreds of processors isn't enough to handle complex workloads such as computational fluid dynamics, oceanographic modeling or petroleum reservoir simulation. These tasks require a balanced system—with high throughput, outstanding reliability, robust memory, sustainable input/output and leading-edge productivity tools—that is easy to manage and maintain.

The IBM @server Cluster 1600 delivers the power, throughput and manageability required in the high performance computing (HPC) environment.

Building on the best

The Cluster 1600 leverages and extends the capabilities of the very successful supercomputer, from IBM, the IBM RS/6000 SP server.¹ The U.S. Department of Energy's Lawrence Livermore National Laboratory is currently using the RS/6000 SP "ASCI White"—one of

the world's fastest computers at 12 teraflops—to conduct complex simulations of the operation of a nuclear weapon.² SP systems are also installed at the supercomputer centers at the Universities of Minnesota and Karlsruhe, North Carolina Supercomputer Center, Atomic Weapons Establishment in the United Kingdom, National Center for Environmental Prediction, San Diego Supercomputer Center and the Department of Energy's Oak Ridge and Lawrence Berkeley National Laboratories.

With the introduction of the Cluster 1600, IBM has enhanced SP technology and extended its benefits to more hardware building blocks, including SMP systems with the award-winning³ POWER4™ micro-processor from IBM.

A Cluster 1600 system can be built with up to 32 IBM **@server** pSeries™ 690 or 670 servers, the most powerful UNIX® servers offered by IBM. In addition, the Cluster 1600 can be built using existing SP nodes and other cluster-enabled pSeries servers such as pSeries 660 and 680 systems. In the fourth quarter of 2002, the entry rack-optimized pSeries 630 is planned to be supported as a Cluster 1600 node.

A choice of superior hardware

The Cluster 1600 design uses tightly interconnected, robust, shared-memory servers and nodes in parallel to provide scalable performance for HPC workloads. Configurations can include up to 32 p690 or p670 servers, 128 SP nodes or a combination of the two, not to exceed 128 total servers—all managed from a control workstation. Scalability limits can be increased through special order (for example, 512 SP nodes or more than 32 p690 servers).

pSeries 690 or 670 servers and SP nodes can be mixed and matched in a Cluster 1600 configuration to support a broad range of technical and scientific applications, as well as infrastructure workloads including Web serving, e-mail and file/print sharing. The entire cluster is assigned a single ID number, which helps an organization manage assets more effectively.

The pSeries 690 and 670 incorporate the latest advances in chip technology from IBM, including POWER4 micro-processors. POWER4 represents the first "SMP-on-a-chip" design for high-end UNIX servers. The p690 features

two 1.1 or 1.3 GHz processors with Level 2 (L2) cache incorporated on each chip. In this configuration, a single POWER4 chip is capable of delivering 124.8GB of data per second from L2 cache to the processors. An optional HPC configuration features high memory bandwidth and dedicated cache that can further improve throughput. A single system has peak performance of 166 gigaflops, resulting in a 2.6 teraflop peak performance of a 16-node cluster.

The SP node choices include 375 MHz or 450 MHz thin and wide and 375 MHz high nodes based on POWER3™ architecture, each of which fits inside SP system frames. Each thin and wide node contains two or four processors. High nodes are available with a choice of 4, 8, 12 or 16 processors.

In an HPC environment, overall performance also depends heavily on robust internode communications. That's why IBM offers the SP Switch2, a high-bandwidth, low-latency interconnect for the Cluster 1600. The SP Switch2 provides one-way bandwidth of up to 500MB per second and bi-directional bandwidth of up to 1GB per second.

With the introduction of the Cluster 1600, the SP Switch2 has been enhanced to double the switch configuration through two-plane support for cluster-enabled pSeries servers and 375 MHz high nodes. This helps improve bandwidth as well as reliability, availability and serviceability characteristics across the switch fabric. In addition, the SP Switch2 is now designed to support all currently available SP nodes and cluster-ready pSeries servers. For greater operational efficiency, multiple SP Switch2 assemblies can be mounted in 24" or 19" frames.

Outstanding system management software

IBM recognizes that software is a key element of clustering, and Cluster 1600 software has distinct advantages in terms of manageability and performance.

AIX[®], the high-performance, open UNIX operating system with Linux[®] affinity from IBM, offers Web-based remote management tools to control the system and monitor key resources such as adapter and network availability, file system status and processor workload. AIX incorporates Workload Manager, which can help to ensure that critical applications

remain responsive even during periods of peak system demand. AIX runs across all pSeries and RS/6000 servers, and all Cluster 1600 nodes, for greater compatibility and investment protection.

The latest release of AIX, AIX 5L[™] Version 5.1, adds new functionality to further enhance security, system availability and workload management. In fact, the system management and Internet/Web-application services of AIX 5L rank as industry leaders.⁴

Cluster manageability is provided by **Parallel System Support Programs** (PSSP) for AIX. These are a collection of functionally rich, cluster software tools designed to provide a foundation on which to scale workloads and cost-effectively manage hundreds of Cluster 1600 nodes and servers. Designed to deliver high-performance and extreme horizontal and vertical scalability, PSSP offers low cost, highly effective clustered system management; easy, continuous upgrades for today's growing workload requirements; and high levels of system, application and data availability.

PSSP 3.4, available with a Cluster 1600, is designed to build on the systems management tools and commands of AIX, providing "cluster-aware" tools for hardware and software configuration and installation, device management, security administration, error logging, problem management, system recovery and resource accounting—all from the control workstation.

PSSP offers several components that help deliver extreme scalability for parallel applications in high-performance computing, including:

- **Communication Subsystems Support**, which contains switch device driver and configuration methods, parallel communication application programming interfaces (APIs), switch initialization and fault-handling software, plus switch adapter diagnostics
- **Virtual Shared Disk (VSD)**, an API that creates logical disk volumes for parallel application access of a real disk device. These can be attached locally or to another node in the cluster. VSD can enhance the performance of applications that provide concurrency control for data integrity

- **IBM Recoverable Virtual Shared Disk**, which provides recovery from failures of virtual shared disk server nodes, and takes advantage of the availability services provided by PSSP to determine which nodes are up and operational
- **Applications programming models** supported by PSSP are multi-threaded, standards-compliant Message Passing Interfaces via IBM Parallel Environment for AIX, as well as single-threaded MPI support. PSSP includes a Low-Level Application Programming Interface

Providing node grouping, PSSP permits cluster nodes to be managed on the cluster level, the individual level or as groups defined by the administrator. This permits a Cluster 1600 to be partitioned so that one group of nodes can run a job in parallel, while other nodes are dedicated to interactive work, serial processing or I/O and networking services. Thus, multiple mixed workloads may be handled independently with no impact on each other.

PSSP includes optional switch connectivity. A cluster using the SP Switch2 can have some nodes that are not connected to the switch. That makes it possible to upgrade to the SP Switch2 and still keep older nodes that are not connected to the switch in the cluster.

PSSP 3.4 has been enhanced to include:

- *Support for two-plane SP Switch2 configurations*
- *Boot-install from Fibre Channel SAN DASD*
- *The Low-level API*
- *Secure remote command processing*
- *VSD support for subsystem device driver*
- *Support for 64-bit applications running AIX 5L Version 5.1*
- *PSSP-related licensed programs*

Software to maximize performance

In a typical cluster, it's common for some processor nodes to be overworked while others are under utilized. Valuable resources can be left unused, especially during off hours. **LoadLeveler**[®] software is designed to optimize cluster resources through dynamic job scheduling and workload balancing, supporting thousands of jobs across a Cluster 1600.

LoadLeveler provides a facility for building, submitting and processing jobs—batch, interactive, serial and parallel—in a dynamic environment. It is designed to match application processing needs with available resources for improved performance and rapid turnaround. Job requirements may include a combination of memory, disk space, processing

type, operating system and application programs. LoadLeveler collects resource information and dispatches the job when it locates suitable nodes.

IBM LoadLeveler for AIX 5L Version 3 is designed to improve throughput across the Cluster 1600 with the following enhancements:

- *Speeds job processing with striped communication between nodes with enhanced switch adapter support. Striping allows the job to use all communication paths to the node for increased data bandwidth and reliability*
- *Allows more windows per SP Switch2 PCI Attachment Adapter, thus enabling more User Space jobs to run simultaneously*
- *Provides consistent resource management across all nodes by integrating AIX 5L Workload Manager*
- *Provides overall system utilization and responsiveness to interactive workloads with Gang Scheduling*
- *Includes checkpoint restart for parallel jobs⁴*
- *Supports 64-bit applications for interactive and batch jobs that run with AIX 5L Version 5.1*

File system performance on the Cluster 1600 is enhanced with **General Parallel File System (GPFS)** for AIX Version 1.5. GPFS is a high-performance, shared-disk file system that can provide fast data access to all nodes in a cluster.

Most UNIX file systems are designed for a single-server environment. Adding additional file servers typically doesn't improve file access performance. GPFS is designed to deliver scalable performance and failure recovery across multiple file system nodes, while complying with UNIX file standards.

In addition to existing AIX administrative file system commands, GPFS has functions that simplify multinode administration. A single GPFS multinode command can perform a file system function across the entire GPFS cluster and can be performed from any node in the cluster. Because GPFS supports the file system standards of X/Open 4.0 with only minor exceptions, most AIX and UNIX applications can use GPFS without modification, and most existing UNIX utilities will run unchanged.

GPFS delivers high-performance by permitting shared access to the disks that make up the file system—whether the disks are physically attached or shared through software simulation provided by VSD and the SP Switch2. Additional performance gains are realized through client-side data caching, large file block support and the ability to perform read-ahead and write-behind file functions. As a result, GPFS can outperform Network File System, Distributed File System and Journalled File System. Unlike these alternatives, GPFS file performance scales as additional file server nodes and disks are added to the Cluster 1600 system.

GPFS version 5.1 supports IBM Enterprise Storage Server™ (ESS) and Hitachi HDS Fibre Channel attached disks. In addition, the Subsystem Device Driver software for the ESS multi-pathing capability is supported. Other enhancements provide improved quality and support for the enablement of Cluster 1600 nodes, including the pSeries 690 and 670 servers.

Parallel Environment for AIX is a high function development and execution environment for parallel applications using Cluster 1600

systems. It is designed to provide a complete solution for organizations that need to develop, debug, analyze, tune and execute parallel programs on AIX. Features include:

- *Parallel Environment Benchmarking Tools—a suite of applications and utilities to analyze the performance of programs*
- *Parallel checkpointing capabilities⁴*
- *Support for 64-bit applications*
- *MPI enhancements, including an additional command for starting MPI jobs*

Finally, IBM offers the **Engineering and Scientific Subroutine Library (ESSL)**, a state-of-the-art collection of mathematical subroutines that provide optimum performance for floating-point engineering and scientific applications. ESSL contains over 400 high-performance mathematical subroutines; **Parallel ESSL**, an additional offering, contains over 100 high-performance mathematical subroutines specifically designed to exploit the full parallel power of Cluster 1600 systems. Parallel ESSL is designed to run on the SP Switch and SP Switch2. To take advantage of this increased performance, programs need only be relinked, not recompiled.

ESSL Version 3.3 and Parallel ESSL Version 2.3 provide the following enhancements:

- *Libraries are tuned for POWER4 processors*
- *ESSL supports the AIX 5L Version 5.1 32-bit and 64-bit kernels; Parallel ESSL supports the 32-bit kernel, but not the 64-bit kernel*
- *The ESSL and Parallel ESSL header file now supports the C++ Standard Numerics Library facilities*

Backed by IBM

The IBM @server Cluster 1600 is backed by worldwide service and support from IBM. Hardware installation support is included with the Cluster 1600 system. Software installation and customization services are also available. In addition, scientific and technical teams at IBM excel at creating targeted solutions and supporting organizations through competency centers, consulting services, workshops and other technical support resources.

Summary

The technology behind the Cluster 1600 is the same technology behind some of the world's most powerful supercomputers. These systems provide outstanding performance and extreme scalability that have been demonstrated repeatedly in some of the most demanding massively parallel application

environments, including global climate modeling, engineering design and analysis, petroleum exploration and production, financial modeling and pharmaceutical design.

In addition, the Cluster 1600 is flexible enough to build on existing assets while incorporating the latest technology, thus offering a clear, smooth upgrade path for the future. A well-planned road map for Cluster 1600, including the RS/6000 SP, allows organizations to start small and scale up to even larger, more powerful systems—adding nodes without having to replace existing hardware and ensuring long-term investment protection as operating needs grow.

The innovative architecture of the Cluster 1600 makes it ideally suited for high-performance computing tasks. IBM is uniquely positioned to deliver Cluster 1600 solutions that meet scientific and technical computing needs.

For more information

To learn more about the IBM @server Cluster 1600, contact your IBM representative or certified IBM Business Partner or visit the following Web sites:

- ibm.com/eserver/clusters
- ibm.com/eserver/pseries
- ibm.com/ibmlink



© Copyright IBM Corporation 2002

IBM Corporation
Marketing Communications,
Server Group
Route 100
Somers, NY 10589

Printed in the United States of America
07-02
All Rights Reserved

This publication was developed for products and/or services offered in the United States. IBM may not offer the products, features or services discussed in this publication in other countries. The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM's future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only.

IBM, the IBM logo, the e-business logo, AIX, AIX 5L, Enterprise Storage Server, LoadLeveler, POWER3, POWER4, pSeries, RS/6000 and SP are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries or both.

Linux is a registered trademark of Linus Torvalds.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product and service names may be trademarks or service marks of others.

IBM hardware products are manufactured from new parts, or new and used parts. Regardless, our warranty terms apply.

Photographs show engineering and design models. Changes may be incorporated in production models.

Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM.

This equipment is subject to FCC rules. It will comply with the appropriate FCC rules before final delivery to buyer.

Information concerning non-IBM products was obtained from the suppliers of these products. Questions on the capabilities of the non-IBM products should be addressed with the suppliers.

All performance estimates are provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Buyers should consult other sources of information, including system benchmarks, to evaluate the performance of a system they are considering buying.

¹ Source: *TOP500 Supercomputer List 06/02* available at www.top500.org

² Source: visit ibm.com/servers/eserver/pseries/news/pressreleases/2001/aug/ascii.html

³ Best workstation/server processor award from MicroDesign Resources; www.microdesign.com; January 30, 2002.

⁴ *2001 UNIX Function Review*, D.H. Brown Associates, Inc., March 2001, and *IBM Flexes UNIX Muscle with AIX 5L*, D.H. Brown Associates, Inc., May 2001