HACMP/6000 Version 2.1 Overview

By Daniel P. Cox and Frank Lawlor

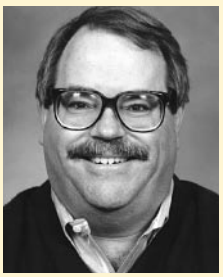
This article provides an overview of Version 2.1 of IBM's High Availability Clustered Multiprocessing/6000 (HACMP/6000) software and describes some new features planned for 1994.

IBM's HACMP/6000 software masks hardware and software failures in clustered RISC System/6000 environments by quickly switching over to backup machines.

HACMP/6000 has two major components: High Availability (HA) and Clustered Multiprocessing (CMP). The *high availability* part of HACMP/6000 is a computing configuration that will survive multiple points of failure. The product detects and recovers from failures of disks, disk adapters, networks, network adapters, and processors. A cluster of loosely coupled machines provides redundancy by transferring control from a failed processor to a backup. A fault-tolerant transfer usually implies the loss of, or interruption to, some in-progress work while the transition takes place. High availability, also referred to as fault-resilient computing, differs in that it is a step toward fault tolerance.

HACMP/6000, which operates on the RISC System/6000, complements and extends the facilities already built into AIX for improved availability, such as the robust Journaled File System and logical volume mirroring.

Generally, HACMP/6000 clusters execute applications without change. Some changes may be necessary for scripts that specify the actions to be taken when a failure occurs and also when components are restored to operation. These scripts are tailored by the system administrator. HACMP/6000 relies on the application, however, to provide any failure recovery transparency or fallover transparency to external users and client machines (restarting the work in-process when a fallover occurs).



Daniel P. Cox

If a system fails, nominal recovery time is approximately 30 to 300 seconds. Actual recovery time depends on the system configuration, application configuration, size of the user databases, and the user's recovery script (if any).

The *cluster multiprocessing* part of HACMP/6000 provides concurrent disk access support to multiple processors in the cluster. This concurrent access provided at the raw disk level requires application support, such as locking, to control access to the shared data. HACMP/6000 provides distributed locking facilities to support this.

By using standard RISC System/6000s, the HACMP architecture provides a lower cost alternative to the expensive and specialized fault-tolerant systems. It is useful when a short application disruption is acceptable.

High-Availability Applications

HACMP/6000 is designed for database and transaction processing applications. The architecture provides reliable, recoverable shared disk resources for database or Online Transaction Processing (OLTP) servers and ultimately to client applications.

Figure 1 shows several examples of applications that use HACMP/6000.

Businesses that support their customers with continuously running computer operations have a vital need for this type of high availability.

Scalable Systems Growth Through Clustered Multiprocessing

As data processing requirements grow beyond the capacity of a single processor, HACMP/6000 can extend the value of an installed RISC System/6000. The clustered multiprocessing feature of HACMP/6000, combined with storage systems such as IBM's 9333 High-Performance Disk Sub-

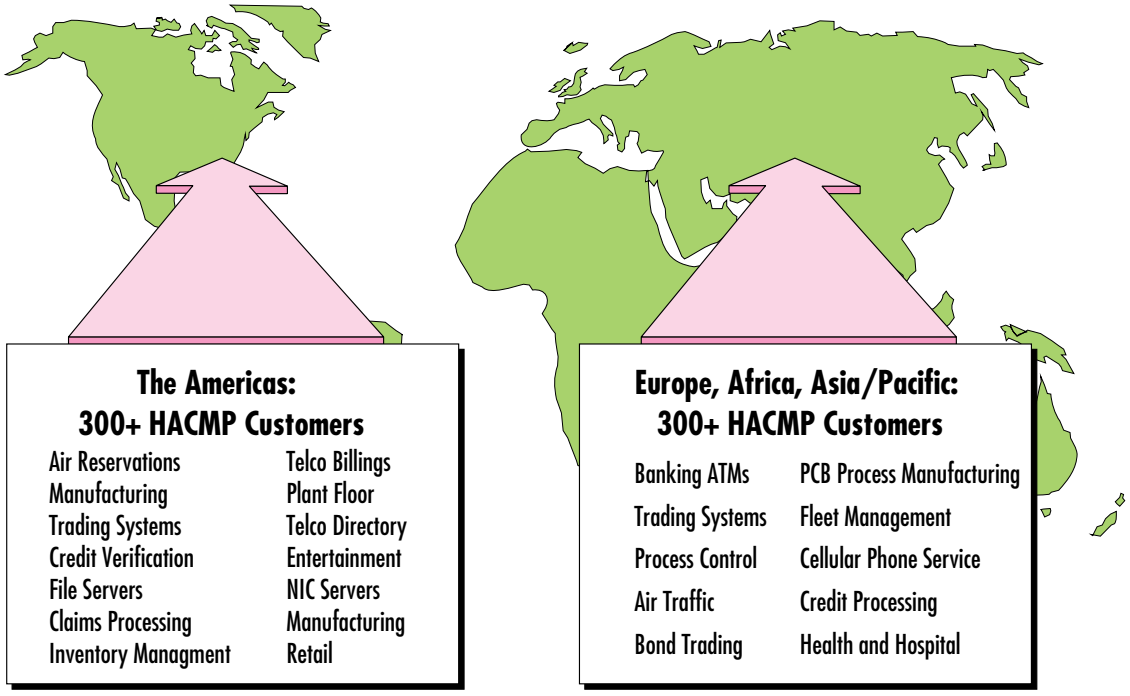


Figure 1. HACMP/6000—a worldwide product

system, allows concurrent disk sharing by multiple processors in the cluster.

HACMP/6000 currently supports up to four processors in a cluster. Later this year, it is expected to support eight-way clusters.

The Cluster Lock Manager (CLM) is especially useful for controlling access to concurrent disks. It provides programming interfaces that applications can use to create a single lock image that can be shared among all the nodes in the cluster. Transaction-oriented applications such as databases can benefit from the CLM.

HACMP/6000 has two locking models. The UNIX lock model supports standard UNIX System V region locking. The CLM lock model provides the following functions:

- Six locking modes that increasingly restrict access to a resource
- Asynchronous lock completion
- Global data through lock value blocks

System Management

New HACMP/6000 installation and configuration facilities allow administrators to install and configure a cluster of RISC System/6000 processors from a single processor. This is easier than individually installing each of the systems in the cluster. HACMP/6000 also includes new installation verification services.

Problem determination capabilities help to trace cluster problems more effectively. More comprehensive, data-driven scripts now minimize the need to modify HACMP/6000 fallover scripts.

HACMP/6000 Runtime Environment

The cluster manager and related services run on each server and perform the following functions:

- Constantly send a series of heartbeat messages through the networks to determine the status of other servers in the cluster (All nodes share these messages to provide information about server failures. These messages can transmit over a LAN such as Ethernet™, Token Ring, or Fiber-optic Distributed Data Interchange, or a serial link.)
- Maintain "state machine" or topology information showing the status (up or down) of processors in a cluster
- Recognize changes in clustered processor states (up or down)

- Execute fallover scripts specific to the resource that changes state (such as a network interface failure or a processor being restored to service)
- Coordinate system booting and failure-recovery operations across the processors in a cluster
- Provide cluster status to the administrator and interfaces to other AIX and RISC System/6000 applications, such as NetView/6000 (Cluster status is also available using Simple Network Management Protocol (SNMP.)

HACMP/6000 currently supports up to four processors in a cluster. Later this year, it is expected to support eight-way clusters.

Cluster Management Operations

The cluster manager is automatically started at boot time. During startup, each machine in the cluster contacts the others over a network, and together they elect a primary or master cluster manager. From this point, all cluster managers send periodic heartbeat messages back and forth among themselves to ensure that all the cluster members are "alive."

If one cluster manager stops sending these heartbeat messages, the other cluster managers try to contact the silent machine through alternate network interfaces or alternate networks. (All cluster nodes should be equipped with alternate adapters and attached to multiple networks.) If these attempts are unsuccessful, the other processors in the cluster will assume that the machine has failed and drop it from the cluster.

A "deadman" switch in each cluster manager ensures that if contact is lost with the other cluster managers, it will terminate system activity. This prevents a rogue system from corrupting shared data.

When a machine is dropped from the cluster, a recovery shell script is executed in the remaining processors. This allows the cluster to reconfigure itself to deal with the failed processor. Backup machines take control of the failed processor's disks, and one may masquerade as the failed machine on the network (using IP address takeover).

The HACMP/6000 cluster manager makes no assumptions about the actions that should occur when the cluster configuration is modified due to a fallover. The cluster behavior is left to the person implementing the cluster. A system administrator or programmer tailors the fallover scripts that define the system fallover and recovery procedures. These scripts—modifiable shell procedures—can include any appropriate commands.

For example, suppose two HACMP/6000 servers are functioning as database servers to several clients. If one machine fails, the cluster manager fallover script would allow the surviving processor to act for the failed machine. It gains access to all shared disks and starts a second database server identical to the one that was running on the failed machine. At that point, client applications would check for cluster status and reconnect to the new database server.

Clients of HACMP/6000 clusters can be connected by serial links or attached to the network. However, serial devices must be switched to the fallover host manually unless they have been connected to a network-attached terminal server or automated switching device.

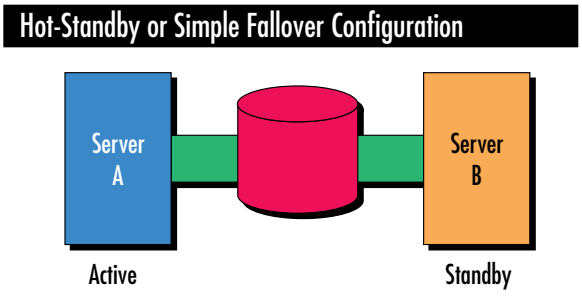


Figure 2. A hot-standby or simple fallover configuration

High-Availability Configurations

There are many ways to configure a highly available RISC System/6000 cluster.

Hot standby or simple fallover: In a hot-standby or simple fallover configuration (shown in Figure 2), the active processor executes the application and the standby processor waits for a failure. The standby machine is not necessarily idle; the work being done on this machine may be stopped if the total capacity is needed to take over for the primary machine. When the primary processor is returned to service, it reclaims the application and resources.

Rotating standby: Rotating standby is like hot standby except that the roles of primary and standby are not fixed. When the previous primary is returned to service, it does not reclaim the application.

Mutual takeover or partitioned workload: This configuration allows each processor to back up the applications running on each of the other processors. In Figure 3, Server A runs applications 1 and 2. Server B runs applications 3 and 4. Server C runs applications 5 and 6. If Server B fails, Server A can restart application 3, while applications 1 and 2 continue to execute. Server

C can restart application 4 while continuing to run 5 and 6.

HACMP Benefits

The following are some of the unique benefits of HACMP/6000 for businesses with mission-critical applications.

- HACMP/6000 improves the utility of installed RISC System/6000s. It is a cost-effective addition to existing hardware and software. The incremental cost to implement HACMP/6000 is low compared to buying additional hardware.
- HACMP/6000 can expand the capacity of a current installation. HACMP/6000 provides scalability without replacing installed hardware and software. This benefit comes from using HACMP/6000 2.1 in the mutual takeover configuration—effectively doubling system capacity by splitting the workload between two or more systems. The system also provides increased availability.

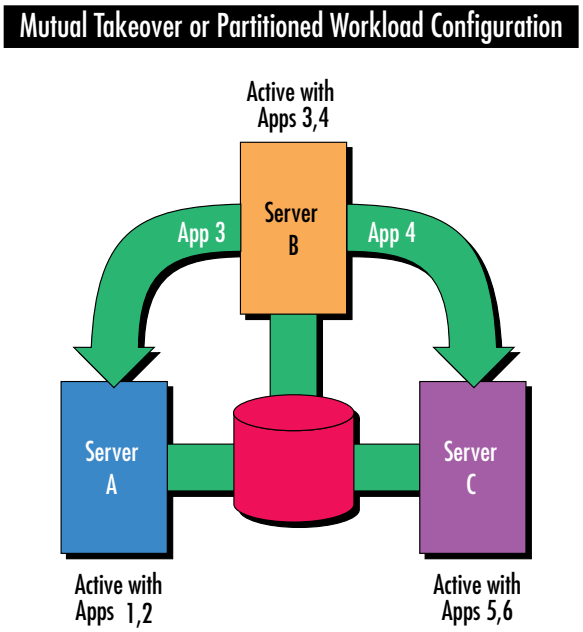


Figure 3. A mutual takeover or partitioned workload configuration

HACMP/6000 provides scalability without replacing installed hardware and software.

- The concurrent disk access configuration also provides availability and horizontal scalability when a single copy of the data is required for executing applications. The data cannot be partitioned as in the mutual takeover configuration.
- HACMP/6000 is an alternative to symmetric multiprocessing configurations when clustering can provide the increased availability that ordinary symmetric multiprocessors cannot.

Performance Considerations

HACMP/6000 cluster performance can be measured and reported in many ways. In a mutual takeover/partitioned workload cluster environment, the applications and data are spread across two to four machines. With minimal interaction between nodes in a cluster, this partitioned environment will have almost 100% efficiency in each node. When a fallover occurs and the backup machines take over for a failed node, performance will be degraded during the time the node is down.

With concurrent disk sharing, distributed locking reduces the efficiency of the scaling. The scaling with Oracle7 Parallel Server was measured at 80% efficiency for a cluster of two RISC System/6000s Model 980. That means the cluster performed at 1.6 times the performance of one machine. IBM estimates four-way cluster efficiency at 2.6 times one machine for transaction processing applications. IBM is conducting four-way tests with results expected in the third quarter of 1994. In long transactions such as the TPC-C benchmarks, we anticipate more efficient scaling because the lock overhead is less. Estimates range from 1.8 times in two-way clusters to 3.2 times in four-way configurations.

Future Directions

IBM expects to provide the following additional capabilities in HACMP/6000 during 1994:

- Eight-way clusters
- Concurrent file system providing a single system image to applications
- Enhanced cluster manager administrative tools
- Certification of HACMP/6000 interfaces to:
  - Distributed Computing Environment (DCE), CICS/6000™, and Encina®
  - Load-leveler queuing and load-balancing extensions
  - DB2/6000™

Summary

HACMP/6000 can provide commercial users with quick recovery from system failures at a reasonable cost. Although HACMP/6000 does not provide complete fault tolerance and continuous operation, it does provide a minimal recovery time after a system failure.

HACMP/6000 provides horizontal scalability by allowing applications to share the disks and CPUs of clustered RISC System/6000 processors.

HACMP/6000 should be considered in situations where users require the following:

- High availability for mission-critical databases or OLTP applications
- Open systems functionality and portability
- Growth in system resources, both horizontally and vertically

Daniel P. Cox, IBM Corporation, 11400 Burnet Road, Austin, TX 78758. Mr. Cox is the brand manager for clustered systems in the RISC System/6000 Division. After joining IBM in 1968 as a programmer in the IBM San Jose Laboratory, he has held positions in development programming, systems engineering, product planning, and management. Before becoming brand manager, Mr. Cox was the product manager for high-availability systems in the RISC System/6000 Division. He has been involved with HACMP/6000 since its beginning—in both planning and implementation. Mr. Cox has BS and MS degrees from San Jose State University.

Frank Lawlor, IBM Corporation, 11400 Burnet Road, Austin, TX 78758. Mr. Lawlor is the AIX architect for high-availability and clustered systems. Since joining IBM in 1970 in Poughkeepsie, New York, he has held several lead architectural positions and also positions in hardware and software development for S/370, AS/400®, and the RISC System/6000. Mr. Lawlor has a BS in Physics from Fordham University and an MS in Electrical Engineering from Columbia University.