CLUSTERS
FOR HIGH AVAILABILITY

SECOND EDITION

A Primer of HP Solutions

- The start-to-finish guide to high availability clustering
- Includes ways to maximize enterprise application availability—and minimize cost
- Completely updated for the latest tools, technologies, and applications
- Describes high availability solutions in HP-UX®, Linux®, and Windows® environments

Peter S. Weygant

Hewlett-Packard® Professional Books

Library of Congress Cataloging-in-Publication Data

Weygant, Peter.
Clusters for high availability : a primer of HP solutions / Peter S. Weygant.--2nd ed.
p. cm.
ISBN 0-13-089355-2
1. Hewlett-Packard computers. 2. Systems availability. I. Title.
QA76.8.H48 W49 2001
004'.36--dc21 2001021151

Editorial/Production Supervision: *Donna Cullen-Dolce*
Acquisitions Editor: *Jill Harry*
Editorial Assistant: *Justin Somma*
Marketing Manager: *Dan DePasquale*
Manufacturing Buyer: *Maura Zaldivar*
Cover Design: *Talar Agasyan*
Cover Design Director: *Jerry Votta*
Manager, Hewlett-Packard Retail Book Publishing: *Patricia Pekary*
Editor, Hewlett-Packard Professional Books: *Susan Wright*

Prentice-Hall, Inc.
Upper Saddle River, NJ 07458

Printed in the United States of America

ISBN 0-13-089355-2

HP Part Number B3936-90047

Prentice-Hall International (UK) Limited, London
Prentice-Hall of Australia Pty. Limited, Sydney
Prentice-Hall Canada Inc., Toronto
Prentice-Hall Hispanoamericana, S.A., Mexico
Prentice-Hall of India Private Limited, New Delhi
Prentice-Hall of Japan, Inc., Tokyo
Pearson Education Asia Pte. Ltd.
Editora Prentice-Hall do Brasil, Ltda., Rio de Janeiro
Prentice-Hall, Inc., Upper Saddle River, New Jersey

Foreword

The author and I have had the good fortune to work in HP's high availability lab for over six years now. In this time, we have spoken with many, many customers who have expressed a need to understand how to enable their businesses to operate in a world that demands ever-increasing levels of uptime. The explosive development of the Internet and the resulting emphasis on e-business, e-commerce, and e-services has made high availability a requirement for an ever-growing range of systems.

At the same time, Hewlett-Packard's cluster software products—MC/ServiceGuard and ServiceGuard OPS Edition (formerly known as MC/LockManager)—have grown in functionality, and the range of supported hardware devices—disks, disk arrays, networks, and system processing units—has also expanded. In addition, more HP-UX system administrators are now confronting the issues of configuring and monitoring high availability systems.

Today, it is still HP's goal to provide a complete range of solutions that yield an **always-on infrastructure** for commercial computing of the 21st century. Peter Weygant's book is intended to assist readers in developing a greater understanding of the concepts, choices, and recommendations for achieving optimal availability in their environments.

Wesley Sawyer
Hewlett-Packard Computing Systems Group
Availability Clusters Solutions Laboratory Manager

Preface

Since the initial publication of this book in 1996, the installed base of Hewlett-Packard's high availability (HA) clusters has grown to more than 45,000 licenses sold worldwide. The technology itself has matured, incorporating such diverse features as event monitoring, wide area networks (WANs), storage area networks (SANs), clusters of up to 16 nodes, and on-line configuration of most cluster components. Hundreds of thousands of users are employing ServiceGuard clusters for tasks as diverse as Internet access, telecommunications billing, manufacturing process control, and banking, to mention only a few.

Hewlett-Packard HA clusters are now or soon will be available on multiple platforms, including HP-UX, Linux, and Windows. Today, HP's *5nines:5minutes* HA continuum provides products, integrated solutions, and reference architectures that allow users to achieve the highest levels of availability—with as little as five minutes of downtime per year. These efforts show a continuing, ongoing commitment to developing tools and products that approach the vision of 99.999% HA.

The three pillars of Hewlett-Packard's HA computing are robust technology, sound computing processes, and proactive customer support. This guide describes these three aspects of HA solutions in the world of enterprise clusters. It presents basic concepts and terms, then describes the use of cluster technology to provide highly available open systems solutions for the commercial enterprise. Here is an overview of each chapter's topics:

- Chapter 1, “Basic High Availability Concepts,” presents the language used to describe highly available systems and components, and introduces ways of measuring availability. It also highlights the processes needed to keep systems available.
- Chapter 2, “Clustering to Eliminate Single Points of Failure,” shows how to identify and eliminate single points of failure by implementing a cluster architecture.
- Chapter 3, “High Availability Cluster Components,” is an overview of HP’s current roster of HA software and hardware offerings.
- Chapter 4, “Cluster Monitoring and Management,” describes a selection of tools for monitoring, mapping, and managing HA clusters.
- Chapter 5, “Disaster-Tolerant High Availability Systems,” is an introduction to disaster-tolerant cluster architectures, which extend the geographic range of the HA cluster.
- Chapter 6, “Enterprise-Wide High Availability Solutions,” describes hardware and software components and processes that provide the highest levels of availability for today’s environment of e-business, e-commerce, and e-services, including 5nines:5minutes.
- Chapter 7, “Sample High Availability Solutions,” discusses a few concrete examples of highly available cluster solutions.
- Chapter 8, “Glossary of High Availability Terminology,” gives definitions of important words and phrases used to describe HA and HP’s HA products and solutions.

Additional information is available in the HP publications *Managing MC/ServiceGuard* and *Configuring OPS Clusters with MC/LockManager.* The *HP 9000 Servers Configuration Guide* and the *HP Net Servers Configuration Guide* contain detailed information about supported HA configurations. These and other more specialized documents on enterprise clusters are available from your HP representative.

CHAPTER 3
High Availability Cluster Components

The previous chapters described the general requirements for HA systems and clusters. Now we provide more detail about a group of specific cluster solutions provided by Hewlett-Packard. Topics include:

- Nodes and Cluster Membership
- HA Architectures and Cluster Components
- Other HA Subsystems
- Mission-Critical Consulting and Support Services

The solutions described here have been implemented under HP-UX, and many of them have already been ported to Linux, with ports to the Microsoft Windows environment expected shortly.

Nodes and Cluster Membership

HP's clustering solutions are all designed to protect against data corruption while providing redundancy for components and software. To achieve these goals, the individual nodes in a cluster must communicate with each other to establish and maintain a running cluster group. This is accomplished partly through two important cluster properties:

- Heartbeats
- Cluster quorum

Heartbeats

Heartbeats are the means of communicating among cluster nodes about the health of the cluster. One node that is designated as the **cluster coordinator** sends and receives messages known as heartbeats from the other nodes in the cluster. If heartbeat messages are not sent and received by a certain node within a specific (user-definable) amount of time, the cluster will re-form itself without the node.

Heartbeats can be carried on more than one LAN. Then, if there is a problem with one LAN, the heartbeat can be carried on the other LAN. Figure 3.1 shows heartbeat configured for a simple two-node cluster.

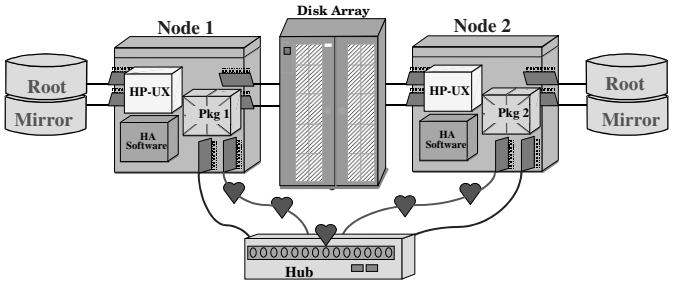

Figure 3.1 *Cluster Heartbeat*

When the heartbeat is lost between cluster nodes, the result is a cluster re-formation, and if one node is unable to communicate with a majority of other nodes, it removes itself from the cluster by causing itself to fail, as shown in Figure 3.2. This failure, known as a Transfer of Control (TOC), is initiated by the cluster software to ensure that only one application is modifying the same data at any one time.

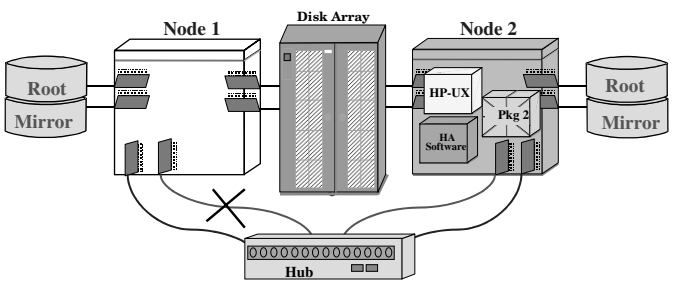

Figure 3.2 *Loss of Cluster Heartbeat*

Cluster Quorum

Under normal conditions, the cluster software monitors the health of individual nodes while applications are running on them. If there is a node failure, the cluster re-forms in a new configuration without the failed node. If there is a communication failure between two sets of nodes, then the set with the greater number of nodes (more than 50%) will be allowed to form a new cluster. This greater number is known as the **cluster quorum**. If the two sets of nodes are of equal size, then both will try to re-form the cluster, but only one can be allowed to succeed. In this case, a **cluster lock** is used.

Cluster Lock

The cluster lock provides a tie-breaking capability in case a communication failure leads to a situation where two equal-sized groups of nodes are both trying to re-form the cluster at the same time. If they were both to succeed in forming new clusters, there would be a "split brain" situation in which the two clusters would access a single set of disks. If this were to happen, data corruption could occur. "Split brain" is prevented by requiring a group of nodes to acquire the cluster lock. The successful group re-forms the cluster; the unsuccessful nodes halt immediately with a TOC. The cluster lock is sometimes implemented as a lock disk, which is a part of an LVM volume group that is used to indicate ownership of the cluster. Figure 3.3 shows a lock disk that has been acquired by one node after a cluster partition.

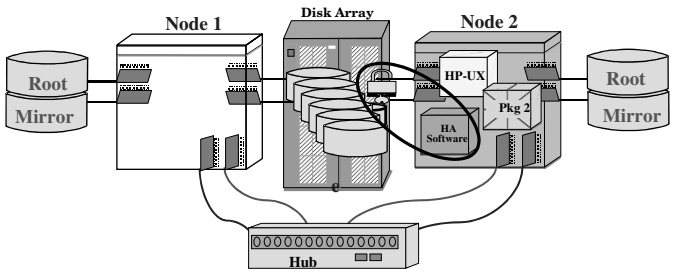

Figure 3.3 *Cluster Lock Acquired by Node 2*

Quorum Server

An alternative to a cluster lock disk is a **quorum server**, which is a software process running on an independent system that provides the tie-breaker when two equal-sized sets of nodes are competing to form a cluster. The quorum server is required in the Linux implementation, but it is available on HP-UX as well. Figure 3.4 shows an example of a quorum server. The quorum server cannot run in the same cluster for which it is providing the tie-breaker, but it can run as a highly available package in another cluster. (Packages are described in detail later in this chapter.)

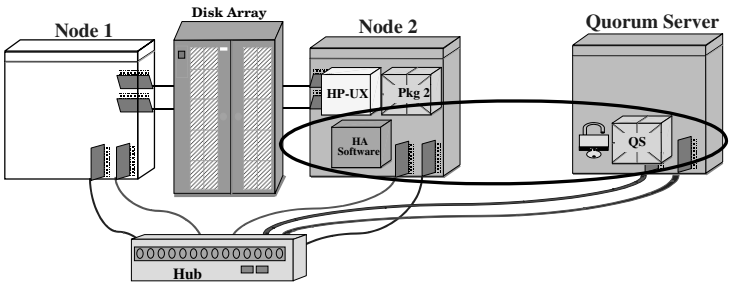

Figure 3.4 *Quorum Server as Tie-Breaker*

The main point of the cluster quorum rule—whether implemented as a lock disk or as a quorum server—is to prevent data corruption that might occur if more than one node tried to run the same application and modfy the same logical volumes. The cluster architectures described throughout the rest of this book have been carefully designed so as to eliminate SPOFs while also eliminating the possibility of data corruption.

HA Architectures and Cluster Components

The cluster shown so far in this book is a generic, loosely coupled grouping of host systems. In fact, each SPU can be connected to another SPU in a variety of highly available cluster configurations. Three basic types are:

- **Active/standby configuration**—one in which a standby SPU is configured to take over after the failure of another SPU that is running a mission-critical application. In an active/standby configuration, two or more SPUs are connected to the same data disks; if one SPU fails, the application starts on the standby. The failed system can then be serviced while the application continues on the standby system. In the active/standby configuration, the backup node may be idle or it may be running another less important application. HP's ServiceGuard product provides active/standby capability.

- **Active/active configuration**—one in which several nodes may be running mission-critical applications, and some can serve as backups for others while still running their own primary applications. HP's ServiceGuard product also provides active/active capability.
- **Parallel database configuration**—a cluster in which the different nodes each run separate instances of the same database application and all access the same database concurrently. In this configuration, the loss of a single node is not critical since users can connect to the same application running on another node. HP's ServiceGuard OPS Edition product provides the parallel database implementation for use with Oracle Parallel Server.

The following sections describe HP's implementations of each of these cluster architectures.

Active/Standby Configurations Using ServiceGuard

A flexible active/standby configuration which allows the application to start on the standby node quickly, without the need for a reboot, is provided by ServiceGuard. Figure 3.5 shows a two-node active/standby configuration using ServiceGuard. Applications are running on Node 1, and clients connect to Node 1 through the LAN. It is also possible to configure a cluster in which one node can act as a standby for several other nodes.

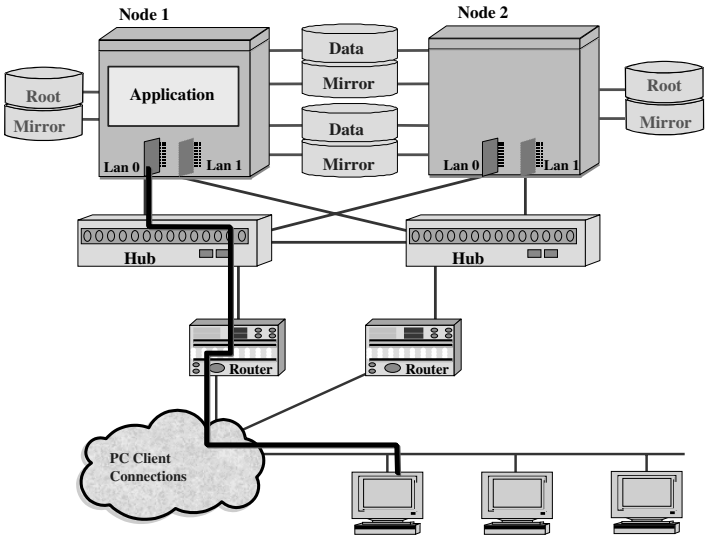

Figure 3.5 *Active/Standby Cluster Before Failover*

In this configuration, the first node is running the application, having obtained exclusive access to the data disks. The second node is essentially idle, though the operating system and the HA software are both running.

The state of the system following failover is shown in Figure 3.6. After failover, the applications start up on Node 2 after obtaining access to the data disks. Clients can then reconnect to Node 2.

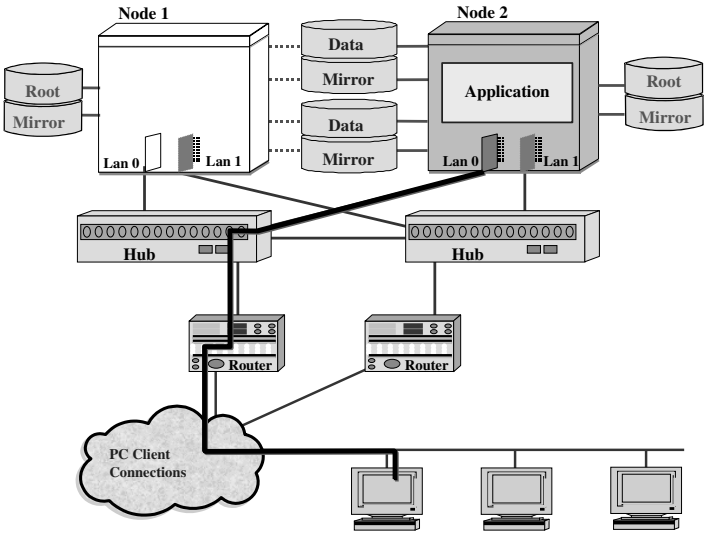

Figure 3.6 *Active/Standby Cluster After Failover*

Note that a failure is not necessary for a package to move within the cluster. With ServiceGuard, the system administrator can move a package from one node to another at any time for convenience of administration. Both nodes remain up and running following such a voluntary switch.

The primary advantage of an active/standby configuration is that the performance of the application is not impaired after a switch to the standby node; all the resources of the standby node are available to the application.

Active/Active Configurations Using ServiceGuard

In an active/active configuration, two or more SPUs are physically connected to the same data disks, and if there is a failure of one SPU, the applications running on the failed system start up again on an alternate system. In this configuration, application packages may run on all nodes at the same time. Figure 3.7 shows a two-node active/active configuration before the failure of one host. Different applications are running on both nodes.

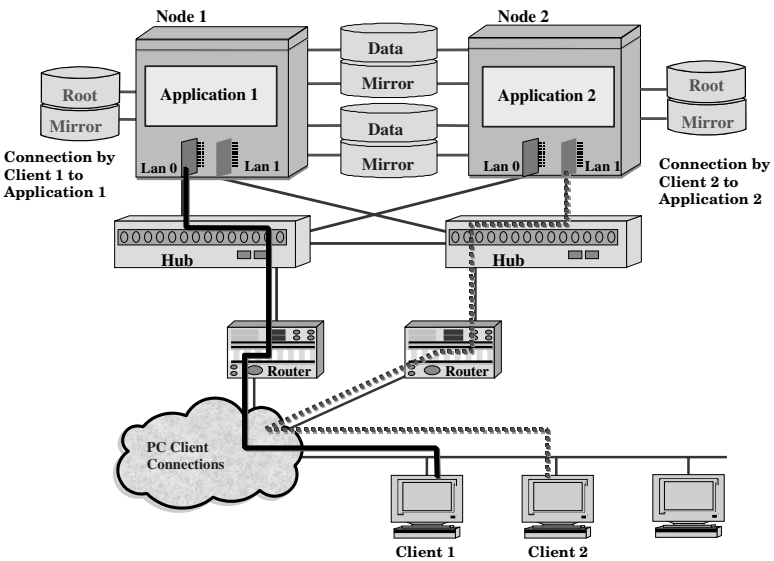

Figure 3.7 *Active/Active Cluster Before Failover*

Figure 3.8 shows an active/active configuration following the failure of one host. The second node still carries on with the applications that were previously running, but it now also carries the application that was running on Node 1 before the failure.

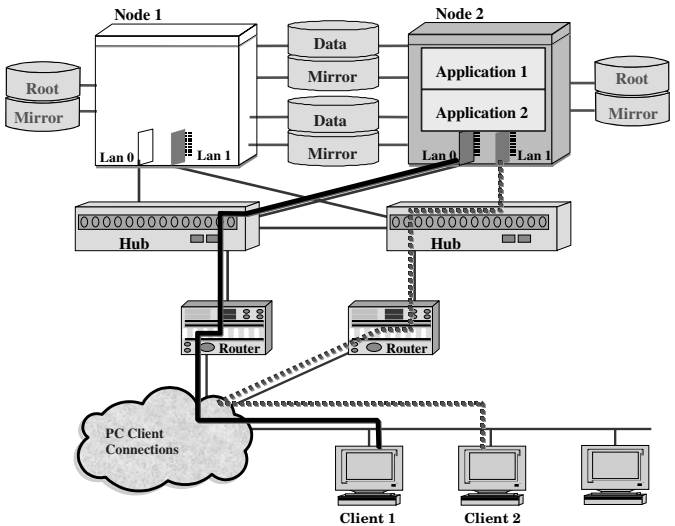

Figure 3.8 *Active/Active Cluster After Failover*

In an active/active configuration, ServiceGuard does not use a dedicated standby system. Instead, the applications that were running on the failed node start up on alternate nodes while other packages on those alternate nodes continue running.

Parallel Database Configuration Using ServiceGuard OPS Edition

In a parallel database configuration, two or more SPUs are running applications that read from and write to the same database disks concurrently. Special software (Oracle Parallel Server, or OPS) is needed to regulate this concurrent access. In the event one cluster node fails, another is still available to process transactions while the first is serviced. Figure 3.9 shows a parallel database configuration before the failure of one node.

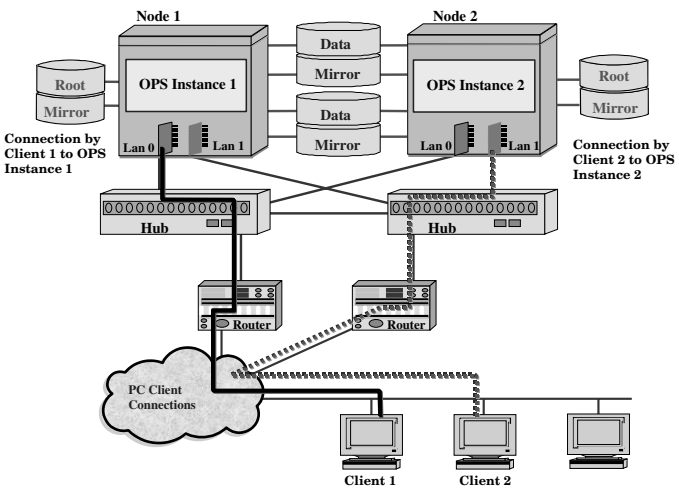

Figure 3.9 *Parallel Database Cluster Before Failover*

Figure 3.10 shows the parallel database cluster after the failure of one node. The second node remains up, and users now may access the database through the second node.

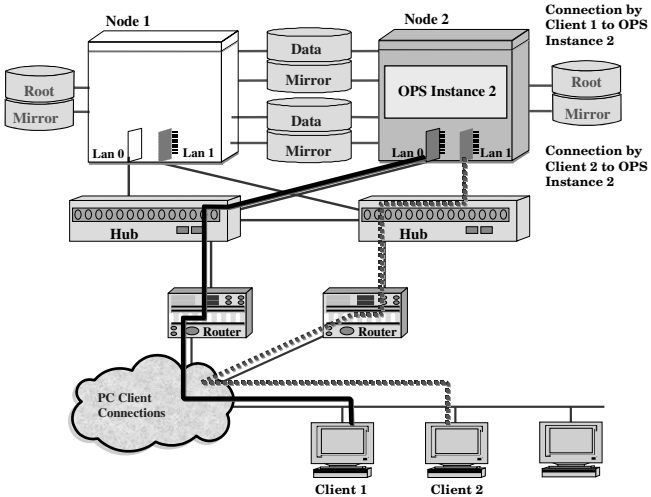

Figure 3.10 *Parallel Database Cluster After Failover*

How ServiceGuard Works

Applications, together with disk and network resources used by applications, are configured in **packages,** which can run on different systems at different times. Each package has one or more application **services** which are monitored by ServiceGuard; in the event of an error in a service, a restart or a failover to another node may take place. A particular benefit of ServiceGuard is that you can configure failover to take place following the failure of a package, or following the failure of individual services within a package. You can also determine whether to try restarting a service a certain number of times before failover to a different node.

With ServiceGuard, there need not be any idle systems; all of the nodes can run mission-critical applications. If one node fails, the applications it supports are moved and join applications that are in progress on other nodes.

Under normal conditions, a fully operating ServiceGuard cluster simply monitors the health of the cluster's components while the packages are running on individual nodes. Any node running in the ServiceGuard cluster is called an **active node**. When you create a package, you specify a **primary node** and one or more **adoptive nodes**. When a node or its network communications fails, ServiceGuard can transfer control of the package to the next available adoptive node.

The primary advantage of the active/active configuration is efficient use of all computing resources during normal operation. But during a failover, performance of applications on the failover node may be somewhat impacted. To minimize the impact of failover on performance, ensure that each node has the appropriate capacity to handle all applications that might start up during a failover situation.

Use of Relocatable IP Addresses

Clients connect via the LAN to the server application they need. This is done by means of IP addresses. The client application issues a **`connect()`** call, specifying the correct address. Ordinarily, an IP address is mapped to an individual hostname—that is, a single HP-UX system. In ServiceGuard, the IP address is assigned to a package and is temporarily associated with what-

ever host system the package happens to be running on. Thus, the client's **`connect()`** will result in connection to the application, regardless of which node in the cluster it is running on.

Figure 3.11 shows a cluster with separate packages running on each of two nodes. Client 1 connects to a package by its IP address. The package is shown running on Node 1, but the client need not be aware of this fact.

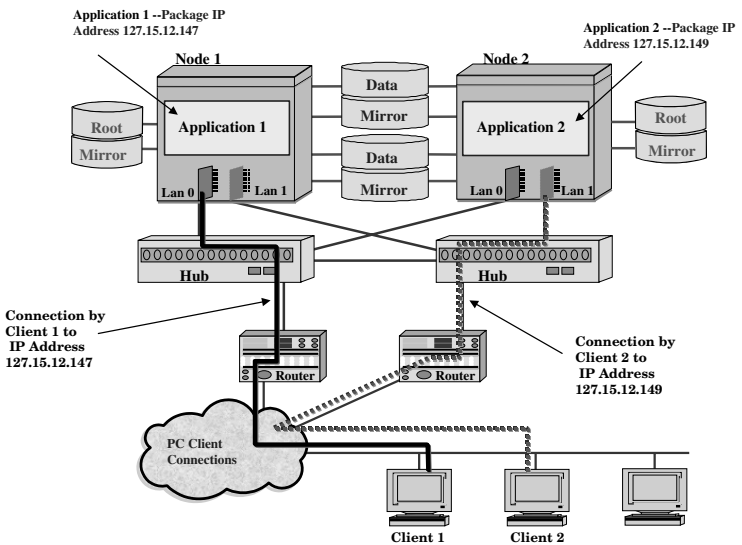

Figure 3.11 *IP Addresses Before Package Switching*

After a failure on Node 1, the package moves over to Node 2. The resulting arrangement of packages is shown in Figure 3.12. Note that the IP address of the package is the same.

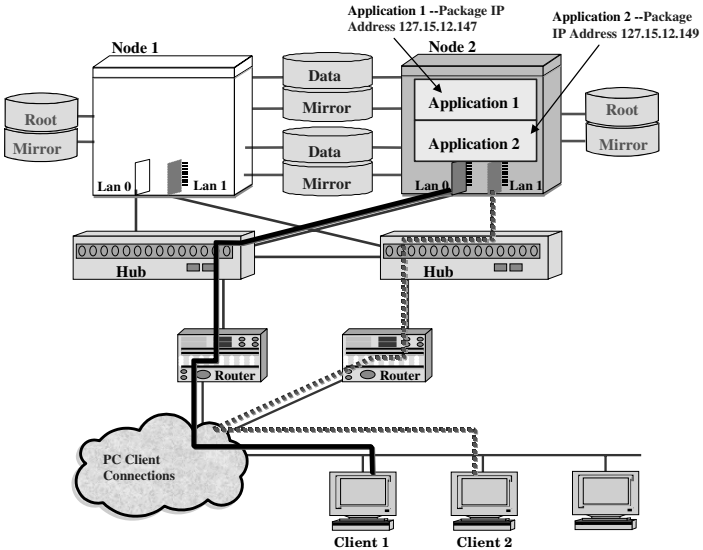

Figure 3.12 *IP Addresses After Package Switching*

The key benefit of using relocatable IP addresses with packages is transparency. The client is unconcerned with which physical server is running a given application. In most cases, no client or server code changes are needed to take advantage of relocatable IP addresses.

Application Monitoring

Central to the functioning of ServiceGuard is the monitoring of user applications. When a package starts, its applications are started with a special cluster command that continues to monitor the application as long as it is running. The monitor immedi-

ately detects an error exit from the application, and alerts ServiceGuard. Depending on the kind of error condition, ServiceGuard can restart the application, halt the application, or fail it over to a different node.

Fast Recovery from LAN Failures

ServiceGuard monitors the status of the LANs used within each node of the enterprise cluster. If any problem affects the LAN, ServiceGuard will quickly detect the problem and activate a standby LAN within the same node. This detection and fast switch to an alternate LAN is completely transparent to the database and attached clients. This feature eliminates the downtime associated with LAN failures and further strengthens the enterprise cluster environment for supporting mission-critical applications.

Workload Balancing

The use of application packages provides an especially flexible mechanism for balancing the workload within the cluster after a node failure. Individual application packages within a single node can be moved to different alternate nodes, distributing the workload of one node across the surviving nodes of the cluster. For example, a cluster with four nodes is configured and each node is running three packages. If a node fails, each of the three packages running on that node can be moved to different nodes. This distributes the workload of the failed node among all of the remaining nodes of the cluster and minimizes the performance impact on the other applications within the cluster.

This same package capability also allows the workload of a cluster to be balanced according to the processing demands of different applications. If the demand of one application package becomes too high, the system administrator can move other application packages on the same node to different nodes in the cluster by using simple commands, thus freeing processing power on that node for meeting the increased demand.

Workload tuning within individual nodes of an enterprise cluster can be further refined by using HP's Process Resource Manager (HP PRM), described in a later section.

Failover Policy and Failback Policy

Added flexibility in the cluster workload configuration is provided by the ability to choose a failover policy for packages. You can choose to fail a package over to the node with the fewest packages currently running on it, or you can have it fail over to a specific node whatever its load.

You can also control the way packages behave when the cluster composition changes by defining a failback policy. If desired, a package can be set up in such a way that it fails over to an adoptive node and then fails back to the original node as soon as the original node becomes available again.

Figure 3.13 shows a cluster with a package failing over to an adoptive node, then failing back to its original node.

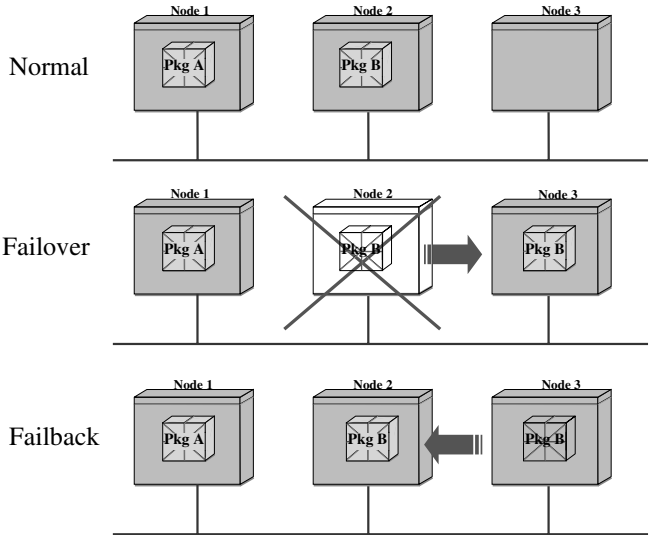

Figure 3.13 *Failover and Failback Policies*

Rolling Upgrades

Another useful feature of ServiceGuard is the ability to upgrade the software on a given node—including the OS and the HA software—without bringing down the cluster. You carry out the following steps for every node in the cluster:

1. Move applications from the node that is to be upgraded to some other node in the cluster.
2. Remove the node from the cluster.
3. Perform the upgrades.
4. Allow the node to rejoin the cluster.
5. Move applications back to the upgraded node.

When using this feature of ServiceGuard, you must carefully plan the capacity of the nodes in the cluster so that moving an application from one node to another during upgrades will not degrade performance unacceptably.

How ServiceGuard Works with OPS

ServiceGuard OPS Edition is a special-purpose HA software product that allows host servers to be configured with Oracle Parallel Server (OPS), which lets you maintain a single database image that is accessed by the HP 9000 servers in parallel, thereby gaining added processing power without the need to administer separate databases. Oracle Parallel Server is a special relational database design that enables multiple instances of the Oracle database to function transparently as one logical database. Different nodes that are running OPS can concurrently read from and write to the same physical set of disk drives containing the database.

Oracle Parallel Server handles issues of concurrent access to the same disk resources by different servers and ensures integrity of Oracle data.

The OPS Edition uses the same underlying cluster mechanism as basic ServiceGuard. This means that you can create and manipulate packages as well as OPS instances on the cluster. Note the following difference, however: In basic ServiceGuard, packages run on only one node at a time, whereas in the OPS Edition, OPS applications may run concurrently on all nodes in the OPS cluster.

Fast Recovery from LAN Failures

Like ServiceGuard, ServiceGuard OPS Edition monitors the status of the LANs used within each node of the OPS cluster. Problem detection and fast switching to an alternate LAN is completely transparent to the database and attached clients.

Protecting Data Integrity

When a node fails, ServiceGuard OPS Edition instantly prevents the failed node from accessing the database. This capability prevents a hung node or a node that has rebooted itself after a failure from inadvertently (and incorrectly) attempting to write data without coordinating its actions with the other node (this situation, which was described previously, is called split-brain syndrome).

Reduced Database Administration Costs

OPS clusters can also help reduce administrative costs through the consolidation of databases. In networks that employ multiple independent databases or partitions of the database on different nodes, an OPS cluster can substantially reduce database administration costs by allowing the multiple databases to be consolidated into one logical database. Even though two nodes are accessing the database from within the cluster, the database is managed as a single unit.

Oracle Parallel FailSafe

A new cluster design from Oracle is known as Oracle Parallel FailSafe. Parallel FailSafe is built upon OPS running in a primary/secondary configuration. In this configuration, all connections to the database are through the primary node. The secondary node serves as a backup, ready to provide services should an outage at the primary occur. Unlike OPS, Parallel FailSafe is tightly integrated with HP clustering technology to provide enhanced monitoring and failover for all types of outages.

Parallel FailSafe can support one or more databases on a two-node cluster. All databases are accessed in a primary/secondary mode, where all clients connect through a single instance on one of the nodes. In a primary/secondary configuration, only one server, the primary, is active at any time. The secondary is running an Oracle instance, but that instance is not doing any work. During normal operations, clients are connected to the primary instance using a primary IP address. In addition, they can be preconnected to the secondary instance using a secondary IP address.

The result of integrating ServiceGuard OPS Edition with Oracle Parallel Failsafe is the simplicity of traditional cluster failover in which the database and applications run on only one node in the cluster. This is important because it means the solution works with no code changes for all types of third-party applications that are designed to work with a single instance of Oracle, but are not parallel server-aware. In addition, the Parallel FailSafe solution provides much faster detection and bounded failover in

the event of a failure, with improved performance after failover thanks to pre-connected secondary connections and a pre-warmed cache on the secondary node.

An example showing a single primary instance is depicted in Figure 3.14.

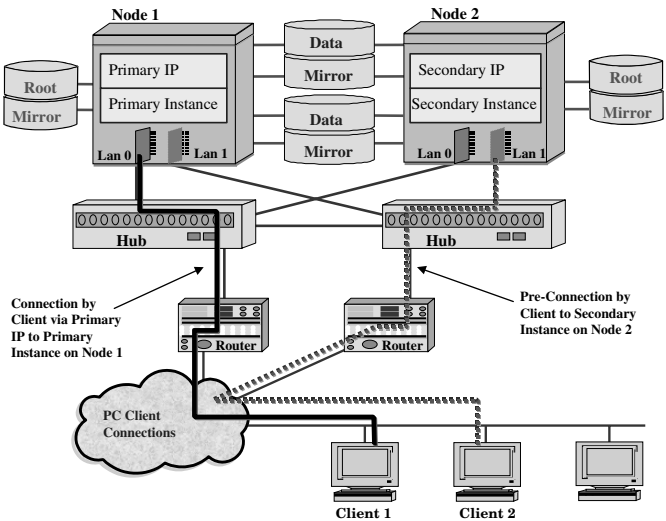

Figure 3.14 *Oracle Parallel FailSafe Before Failover*

After failover, the failover node starts up the primary instance, as shown in Figure 3.15.

Failover time depends on a number of factors in Service-Guard clusters, including the time it takes to obtain a cluster lock, the time required for running startup scripts on the failover node, and the database recovery time required before transactions can

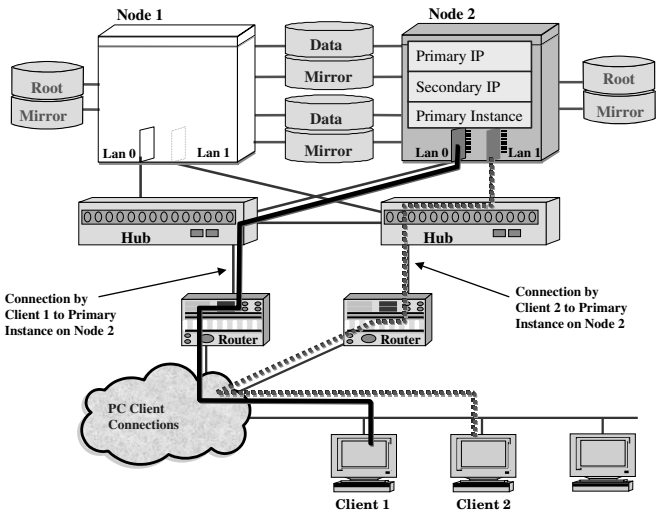

Figure 3.15 *Oracle Parallel FailSafe After Failover*

be processed on the failover node. Oracle Parallel FailSafe can reduce this recovery time significantly. In addition to the standby Oracle instance, other applications can run on the standby node. Note also that on a cluster with more than two nodes, each node can run an instance of Oracle 8i that may fail over to the standby node.

Disaster-Tolerant Architectures

Many of the cluster designs presented so far in this chapter can be extended and adapted for use in long-distance and wide area environments as well. Such clusters are known as disaster-

tolerant clusters because they protect against the kinds of events that affect entire data centers, sometimes even entire cities or regions. Examples are natural disasters like hurricanes.

Three types of modified architecture have been developed based on the standard ServiceGuard cluster:

- Campus cluster
- Metropolitan cluster
- Continental cluster

These types are described fully in Chapter 6.

Other HA Subsystems

To help you further enhance the overall availability, flexibility, and ease of management of your mission critical environment, the following products and services are suggested:

- Software mirroring
- Automatic port aggregation (APA)
- HA disk enclosure
- HP disk arrays
- HP SureStore Disk Array XP Series
- EMC disk arrays
- Enterprise Cluster Master Toolkit
- NFS Toolkit

- ServiceGuard Extension for SAP
- Journaled file system (JFS)
- OnLineJFS
- Veritas Volume Manager
- Transaction processing (TP) monitors
- PowerTrust uninterruptible power supplies (UPS)
- System management tools

Each of these is described further below.

Software Mirroring

Basic mirroring of individual disks is provided with MirrorDisk/UX. Operating through the HP-UX Logical Volume Manager (LVM), MirrorDisk/UX transparently writes data to one primary volume and up to two mirror volumes. By mirroring across different disk adapters (channels), MirrorDisk/UX provides protection against the failures of all major components associated with data.

An added benefit of MirrorDisk/UX is the capability for on-line backup. The mirror is “split” from the primary, resulting in a static copy of the data that is used for a backup. After the backup is performed, MirrorDisk/UX will transparently handle the resynchronization of the mirror and primary data. In cases where you need the highest levels of protection, three-way mirroring allows the on-line backup function to be performed while the primary disk is still being mirrored.

Software from Veritas Corporation also provides extensive mirroring capability for logical volumes. Veritas VxVM is an alternative volume manager provided on HP-UX 11i and later systems; and CVM (described below) provides mirroring support for HP-UX clusters.

Automatic Port Aggregation

Autoport aggregation (APA), available as a separate product on HP-UX 11.0 and later systems, allows you to group multiple physical fast Ethernet or Gigabit Ethernet ports into a logical link aggregate. Once enabled, each link aggregate can operate as a single logical link with only one IP and MAC address. This technology provides automatic fault detection and recovery for HA as well as increased bandwidth with load balancing among the physical links.

High Availability Disk Storage Enclosure

Conventional disk mechanisms may be mounted in a special HA storage enclosure that permits hot plugging, that is, removal and replacement of one disk in a mirrored pair without loss of service while the OS is running and the device is powered on. HP-UX system administration is required during the replacement to allow the old disk to be removed from the mirrored group and to allow the new disk to be added back to the mirrored group.

High Availability Disk Arrays

HP's HA disk array products provide an alternate method of protecting your vital application data. These are hardware RAID units that can be configured so that all major components are redundant, including not just the disk drives, but also the power supplies, fans, caches, and controllers. The dual controllers can both be used simultaneously from the server to read and write volumes on the array, improving performance as well as enhancing availability. If any major component fails, the redundant component picks up the workload without any loss of service. Furthermore, the repair of the failed component does not require any scheduled downtime for maintenance. All the major components are hot-swappable, that is, they can be replaced on-line.

These units also support a global spare for the disk volumes. This means that one spare drive unit can be used as a backup of all the RAID volumes that have been configured in the array. If one drive fails in any of the defined volumes, the global spare is used to quickly re-establish full redundancy.

HA disk arrays support RAID 0 (striping), RAID 1 (mirroring), RAID 0/1 (striping and mirroring), and RAID 5 (rotating parity). They support up to 64MB of read/write cache per controller. Drives can be added on-line, up to the maximum system capacity.

HP SureStore Disk Array XP Series

The HP SureStore Disk Array XP Series offers large data capacity together with many enhanced HA features. This array provides full redundancy of components, including power supplies, fans, disk mechanisms, processors, and caches.

The XP Series lets you define storage units (LUNs) for use on individual systems and cluster nodes. Raid Manager software provides additional capabilities: Business Copy allows you to split off duplicate volumes for off-line backup or other purposes; and Continuous Access lets you establish and maintain links to other XP arrays for physical data replication. These capabilities underlie the physical data replication used in specialized metropolitan and continental clusters (described further in Chapter 5).

EMC Disk Arrays

High-capacity disk arrays are also available from other vendors. A notable example is the Symmetrix product family of cached disk arrays from EMC Corporation. In addition to providing very high capacity, Symmetrix arrays allow connections from the same data disks to multiple cluster nodes across different FibreChannel or SCSI buses.

The Symmetrix EMC SRDF facility, like Continuous Access on the XP Series, allows direct hardware connection to other disk arrays for physical data replication.

Enterprise Cluster Master Toolkit

The Enterprise Cluster Master Toolkit contains a variety of tools for developing HA solutions using ServiceGuard clusters. Individual toolkits included in the Master Toolkit product include:

- Netscape Internet Server toolkit
- Netscape Messaging Server toolkit
- Netscape Calendar Server toolkit
- Oracle toolkit
- Oracle Standby Database toolkit
- Informix toolkit
- Progress toolkit
- Sybase toolkit
- Foundation Monitor toolkit

Each toolkit consists of a set of sample scripts or other tools plus documentation on using it to build a specific HA solution.

A separate NFS toolkit product is also available.

ServiceGuard Extension for SAP R/3

The specialized environment of SAP R/3 requires special customization when implemented in a ServiceGuard cluster. One component, the Enqueue Server, can be made highly available in different ways—by using a standard package arrangement or by building a special component known as Somersault into the appli-

cation. Somersault maintains parallel state tables on multiple nodes for a crucial SAP service; in the event of failover, this state table remains immediately available on the failover node.

An SAP R/3 configuration is especially complex, and should be implemented by specialists in SAP and ServiceGuard configuration. Consulting services are available specifically for this configuration.

Journaled File System

The journaled file system (JFS), a standard feature of HP-UX, is an alternative to the UNIX high-performance file system (HFS). JFS uses a special log to hold information about changes to file system metadata. This log allows JFS to improve availability by reducing the time needed to recover a file system after a system crash. With JFS, the file system can be restarted after a crash in a matter of seconds, which is much faster than with HFS.

As JFS receives standard read/write requests, it maintains an intent log in a circular file that contains file system data structure updates. If a file system restart is performed, `fsck` only needs to read the intent log and finish the outstanding updates to the data structures. Note that this does not normally include user data, only the file system data structures. This mechanism assures that the internal structure of the file system is consistent. The consistency of user data is achieved by a transaction logging mechanism.

OnLineJFS

OnLineJFS is an optional product that adds extensions to JFS. This product eliminates the planned downtime that is associated with typical file system maintenance activities. With OnLineJFS, activities such as defragmentation, reorganization, and file system expansion can all be performed while applications are accessing the data. (The conventional HFS requires that applications be halted before performing these kinds of maintenance activities. HFS does not support or require defragmentation.)

The on-line backup feature is provided by designating a *snapshot* partition. As writes are made to the data, a copy of the old data is copied to the snapshot. This allows applications to access the latest data while the backup process accesses a static copy of the data. The size of the partition needed to hold the snapshot will vary depending on the number of writes performed to the data during the time that the snapshot is maintained; typically, a snapshot will require 2-15% of the disk space of the original data.

Veritas Cluster Volume Manager

With the HP-UX 11i release, the Veritas Volume Manager (VxVM) became a standard software component shipped with all systems. For clusters, a special volume manager known as the Cluster Volume Manager (CVM) offers several important advantages over the earlier LVM provided with HP-UX. LVM requires you to import a volume group separately on each cluster node;

CVM lets you set up a disk group only once and have it immediately visible to all nodes in the cluster. Tools are provided for converting from LVM to CVM disk storage.

Transaction Processing Monitors

Transaction processing (TP) monitors ensure availability in a matter of seconds when used in conjunction with HA clusters by resubmitting transactions to another node when the first node fails. Transaction Processing monitors enable quick restart after any failures and guarantee that incomplete transactions are rolled back. Furthermore, in a mission-critical environment, the TP monitor combines operations of subsystems into one transaction, and integrates the various resources residing in different locales into global transaction services. This ability to globally manage heterogeneous subsystems cannot be achieved by databases alone.

TP monitors available on HP systems include CICS/9000, Encina/9000, TUXEDO, Top End MTS (MicroFocus Transaction System), and UniKix.

Uninterruptible Power Supplies

Installing an HP PowerTrust uninterruptible power supply (UPS) in an HP-UX computer system ensures that power is maintained to your computer, preventing problems such as networking time-outs and tape rewinds. PowerTrust provides at least 15 minutes of continuous backup power, ensuring that data is not lost in the event of a power failure.

A PowerTrust UPS can be configured to bring a system down gracefully before its batteries deplete, thus maintaining data integrity and ensuring a clean reboot and reasonably fast file system recovery. For larger installations, you may wish to use passthrough UPS power protection.

System and Network Management Tools

HP offers a comprehensive set of software tools that allow for centralized, automated management of a wide-ranging network of servers and workstations from many different vendors. See Chapter 5 for complete details.

Mission-Critical Consulting and Support Services

The use of consulting and support services is highly recommended when you are developing an HA system. HP's HA product family encompasses the following consulting and support services:

- Availability management service
- Business continuity support
- Business recovery services

HP has vast experience in creating HA solutions for the UNIX environment, and all HA customers are encouraged to take advantage of this specialized know-how.

Availability Management Service

One main focus of this consulting service is to perform a comprehensive operational assessment of a mission-critical processing environment. This includes analyzing all aspects of the environment such as the hardware being used, software versions and tools, business processes related to the computing environment, as well as the skill set of data processing personnel. This service will identify weaknesses in the processing environment that might cause service outages and will create a plan to eliminate the identified weaknesses.

A second area of consulting is the design and implementation of an availability management plan. Consultants can assist in the following areas:

- Project management
- System and network software installation
- System performance testing
- Full site design, cabling, and testing
- Tuning and enhancement of operations
- Customization, integration, and testing of availability management tools and processes

Business Continuity Support

Business continuity support (BCS) is HP's most comprehensive support offering. It is designed for use in mission critical environments where unacceptable financial or business damage results from even short outages. Business Continuity Support has

been crafted to ensure maximum application availability by targeting potential planned and unplanned outages at their source and taking direct action to prevent them or minimize their duration and impact on your business.

The first step in the delivery of BCS is an operational assessment. This is a consulting engagement in which an HP availability expert reviews your system and operations environment, analyzes its strengths, identifies gaps that could lead to outages, and makes recommendations to help you reach your availability goals. Next, a service level agreement (SLA) is developed with you, and your account team of HP experts is identified and made thoroughly familiar with your environment, business, and support needs.

The BCS account team then provides several *proactive* services:

- Change management planning services to carefully plan, script, and assist in changes of any kind to the computing environment.
- Daily review, communication, and planning for applying patches that are appropriate for your environment, or otherwise heading off potential problems.
- Regular technical reviews to advise and convey information on a variety of HA topics.
- Continuous monitoring of your HP system for anomalies that could escalate into outages if action is not taken.

The BCS team also provides *reactive* services in the event of an outage. HP provides a commitment to restoring your business operation in four hours or less, usually a lot less. This commitment is possible because of HP's investment in a large global response infrastructure of tools, resources, and processes, and HP's staff of experienced recovery experts.

Business Continuity Support is delivered with a commitment to an intimate understanding of your business and IT environment, and total care for that environment. BCS complements the other investments you make in HA technology and operations processes, and offers peace of mind for your users.

Network Availability Services

To meet the needs of IT managers and network administrators, HP offers HP network availability services and HP assessment services for networks to extend mission-critical support to the network. These network services can be purchased standalone or as a complement to HP critical systems support and BCS to help you realize maximum system and network uptime. Network availability services provide a suite of scalable services that utilize assigned network specialists to provide reactive and proactive support for your network.

Assessment services for networks provide three levels of services to optimize network availability. The assessment services include proactive identification of potential critical points of failure, and analysis and recommendations to optimize network performance and improve operational and IT processes.

Business Recovery Services

Another group of services for HA systems is business recovery services. These services are used to provide protection against disasters and large-scale failures. Business recovery services are discussed in Chapter 6, “Disaster-Tolerant Solutions.”

Index

Numerics

A

E

F

G

H

I

J

L

M

N

O

P

Q

R

S

T

U

V

W

X