

Competitive Analysis of UNIX Cluster Functionality – Part One of a Two-Part HA Study

EXECUTIVE SUMMARY

INTRODUCTION

This report from D.H. Brown Associates, Inc. (DHBA) compares six vendors on their UNIX-based cluster features and functions for high availability (HA). It assesses vendors' offerings for the presence of hundreds of individual features and functions, numerically rates them within six cluster-related functional categories, and then ranks them overall. A subsequent report will evaluate vendor offerings across five in-system-related categories; DHBA's research in this area evaluates features that reduce planned and unplanned downtime within individual servers or cluster nodes. Generally speaking, in-system-related features enhance availability without resorting to failover. Later this year, DHBA plans to combine the results of its in-system research with the cluster results reported here to produce a composite of overall strength.¹

SUMMARY OF OVERALL FINDINGS

Figure 1 presents an overview of DHBA's cluster findings, arranged by vendor. The horizontal line in Figure 1 indicates the average vendor rating based on all features and functions considered in this study. Each column corresponds to a particular vendor and shows that vendor's result vis-à-vis the average. Vendors whose columns extend above the horizontal line beat the average, vendors whose columns fall below it did not. Note that these results must be considered in the context of an industry-wide overview (see "Method and Interpretation," below).

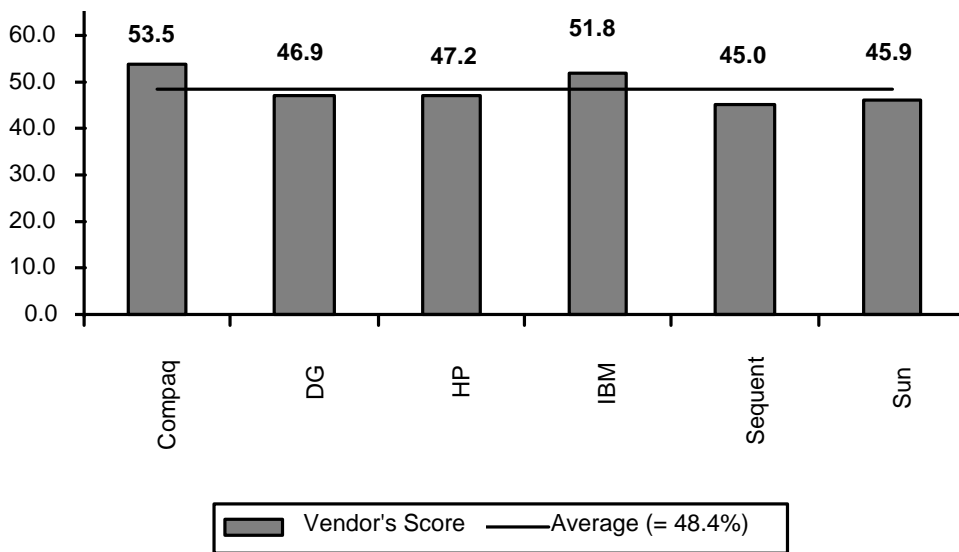
¹ In the composite results, cluster categories (evaluated within this report) will contribute 65% of the total score, and in-system categories will contribute 35%. Thus, the highest achievable result in this evaluation is 65%.

D.H. Brown Associates, Inc.

<http://www.dhbrown.com>

A summary of this report is available to all of our subscribers free of charge. Sponsors of our collaborative program in High Availability Software & Hardware/Cluster (HASH) receive the full report as part of our comprehensive services. Those interested in the program should contact Bill Moran, Research Director – Open Systems at moran@dhbrown.com or 914-937-4302, ext. 230.

**FIGURE 1:
 Overall Results**



In assessing a vendor's result, one must examine not only a vendor's overall ranking, but also how the vendor performed in each functional category. Table 1 summarizes how each vendor placed in the six categories that apply to clustering.

**TABLE 1:
 Accumulated
 Rankings by Vendor**

| Vendor | Firsts | Seconds | Thirds | Fourths | Fifths | Sixths |
|----------------|--------|---------|--------|---------|--------|--------|
| Compaq | 3 | 1 | 1 | 0 | 0 | 1 |
| DG | 0 | 1 | 1 | 1 | 3 | 0 |
| HP | 1 | 1 | 1 | 1 | 2 | 0 |
| IBM | 2 | 2 | 0 | 1 | 0 | 1 |
| Sequent | 0 | 1 | 1 | 2 | 1 | 1 |
| Sun | 0 | 1 | 2 | 1 | 0 | 2 |

(vendors listed in alphabetical order)

METHOD AND INTERPRETATION

This analysis examines clustering technologies that enable multiple nodes to be deployed as a single computing resource. DHBA compared leading vendors' detailed functional capabilities for clustering based on a comprehensive review of each vendor's offerings.²

FIGURE 2:
Functional Categories

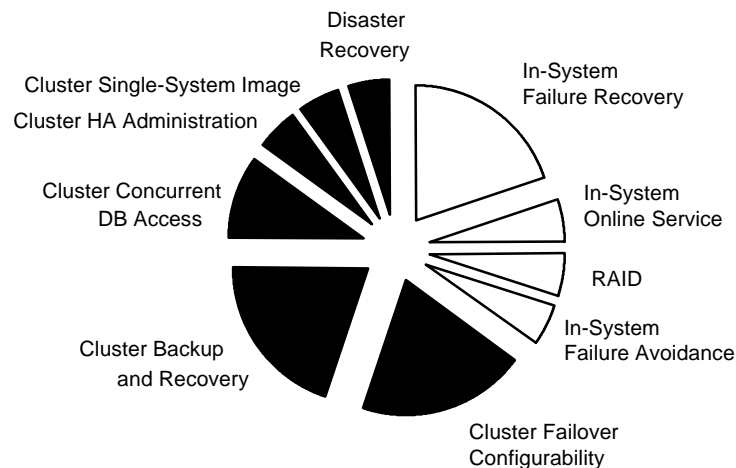


Figure 2 lists the 10 functional categories DHBA identified as the most important strategic tradeoffs for UNIX-based high-availability offerings. Of these 10, six categories (shaded) apply to clustering capabilities and form the basis of this report. The remaining four categories (not shaded), evaluate vendors' in-system (i.e., server) RAS features and RAID; they will be treated in subsequent research.

Each functional category and each feature within a category received a weight to indicate its relative importance within the total score. DHBA assigned these weights based on its understanding of the generalized requirements for HA and clustering. Vendors' results are compared to the vendor average and to leading scores among the products evaluated in this study. These averages should serve as rough guides and may not prove relevant to the particular solutions implemented at any one site. A site examining HA and clusters, for example, must consider its own requirements and should specify its own weights for features and functional categories.

² This study does not attempt to evaluate the fitness of the functional capabilities that each vendor's offerings deliver. Such an analysis would obviously require an enormous investment of time and money. Moreover, it would likely yield results that do not reflect the general quality of vendors' products, but reflect rather the quality of specific implementations.

This report does not evaluate other important considerations such as:

- service and support,
- equipment price and/or total cost of ownership,
- vendor ability to execute and deliver, and
- market penetration.

Any buying decision must take these factors (and others) into account, as well as the information in this report. DHBA plans to do a study of vendors' service and support for mission-critical environments later this year.

While DHBA's methodology requires numerical score calculation and vendor rankings in the examined categories, some of the resulting scores are very close. As a result, the de facto difference between some offerings (vis-à-vis the examined feature or function) may not prove significant in a production environment.

OFFERINGS EVALUATED

Each vendor submitted a cluster offering for DHBA to evaluate. Table 2 lists each vendor's representative operating system and clustering software.

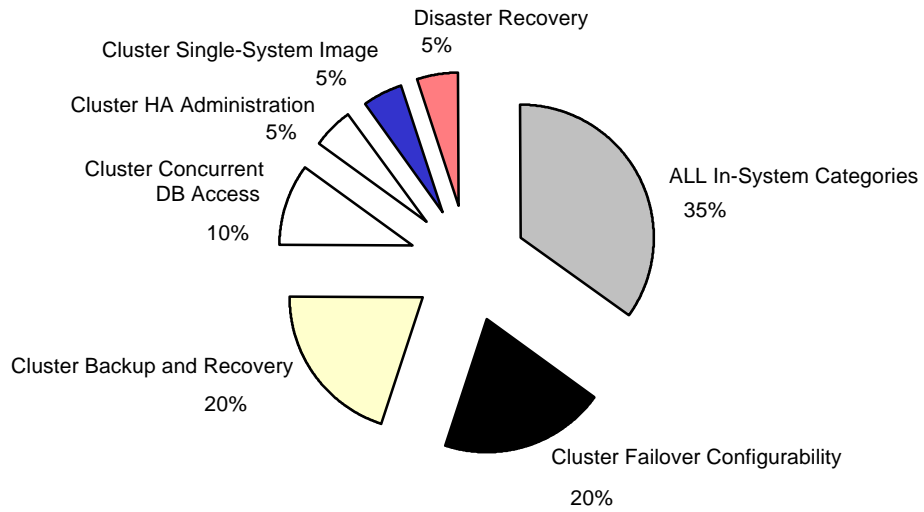
| Vendor | Operating System Software | Cluster Software |
|------------------------|----------------------------------|-------------------------|
| Compaq | Tru64 UNIX | TruCluster Server |
| Data General | DG/UX | DG/UX Clusters |
| Hewlett-Packard | HP/UX | MC/ServiceGuard |
| IBM | AIX | HACMP |
| Sequent | Dynix/ptx | ptx/CLUSTERS |
| Sun | Solaris | Sun Clusters |

TABLE 2:
Operating System and Cluster Software Evaluated

FUNCTIONAL CATEGORIES

As noted above, this study rates vendor offerings according to functional categories; these categories consist of features that DHBA assessed and arranged into groups. Figure 3 (next page) shows on a percentage basis how much each of the six clustering categories contributed to the overall rating.

**FIGURE 3:
Individual Cluster
Categories Percent
Contributions**



1. **Cluster Backup and Recovery** – assesses options for detecting and responding to conditions and failures; also examines failover policies.
2. **Cluster Failover Configurability** – rates a cluster’s ability to operate with variations in hardware configuration and system software.
3. **Cluster Concurrent Database Access** – evaluates support for concurrent access (parallel) databases (e.g., Oracle Parallel Server).
4. **Cluster HA Administration** – examines tools and interfaces provided to ease implementation and management of the HA environment; also examines online cluster HA administration.
5. **Cluster Single-System-Image** – rates features that simplify administration of clusters by presenting the set of distributed cluster resources as if they were resident in a single system.
6. **Disaster Recovery** – measures vendors’ responses to demands for remote data duplication, geographic mirroring, and remote failover, all of which enable recovery from site disasters.

RESULTS OVERVIEW – BY VENDOR

FIRST PLACE – COMPAQ

| Category | Rating | Ranking |
|--|----------------------|--------------|
| Cluster Backup and Recovery | Above Average | Second |
| Cluster Failover Configurability | Above Average | Third |
| Cluster Concurrent Database Access | Above Average | First |
| Cluster High-Availability Administration | Above Average | First |
| Cluster Single-System-Image | Above Average | First |
| Disaster Recovery | Below Average | Sixth |
| Overall | Above Average | First |

SECOND PLACE – IBM

| Category | Rating | Ranking |
|--|----------------------|---------------------------|
| Cluster Backup and Recovery | Below Average | Sixth |
| Cluster Failover Configurability | Above Average | First |
| Cluster Concurrent Database Access | Above Average | Second |
| Cluster High-Availability Administration | Above Average | Second, Tied with Sequent |
| Cluster Single-System-Image | Below Average | Fourth |
| Disaster Recovery | Above Average | First |
| Overall | Above Average | Second |

THIRD PLACE – HP

| Category | Rating | Ranking |
|--|----------------------|----------------------|
| Cluster Backup and Recovery | Above Average | First |
| Cluster Failover Configurability | Above Average | Second |
| Cluster Concurrent Database Access | Below Average | Fifth, Tied with DG |
| Cluster High-Availability Administration | Below Average | Fourth, Tied with DG |
| Cluster Single-System-Image | Below Average | Fifth |
| Disaster Recovery | Above Average | Third |
| Overall | Below Average | Third |

FOURTH PLACE – DG

| Category | Rating | Ranking |
|--|------------------------|----------------------|
| Cluster Backup and Recovery | Slightly Above Average | Third, Tied with Sun |
| Cluster Failover Configurability | Below Average | Fifth |
| Cluster Concurrent Database Access | Below Average | Fifth, Tied with HP |
| Cluster High-Availability Administration | Below Average | Fourth, Tied with HP |
| Cluster Single-System-Image | Above Average | Second |
| Disaster Recovery | Below Average | Fifth |
| Overall | Below Average | Fourth |

FIFTH PLACE – SUN

| Category | Rating | Ranking |
|--|------------------------|---------------------|
| Cluster Backup and Recovery | Slightly Above Average | Third, Tied with DG |
| Cluster Failover Configurability | Below Average | Fourth |
| Cluster Concurrent Database Access | Below Average | Third |
| Cluster High-Availability Administration | Below Average | Sixth |
| Cluster Single-System-Image | Below Average | Sixth |
| Disaster Recovery | Above Average | Second |
| Overall | Below Average | Fifth |

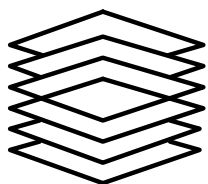
SIXTH PLACE – SEQUENT

| Category | Rating | Ranking |
|--|----------------------|-----------------------|
| Cluster Backup and Recovery | Below Average | Fifth |
| Cluster Failover Configurability | Below Average | Sixth |
| Cluster Concurrent Database Access | Below Average | Fourth |
| Cluster High-Availability Administration | Above Average | Second, Tied with IBM |
| Cluster Single-System-Image | Average | Third |
| Disaster Recovery | Below Average | Fourth |
| Overall | Below Average | Sixth |

TABLE OF CONTENTS

| | |
|---|-----------|
| COMPETITIVE ANALYSIS OF UNIX CLUSTER FUNCTIONALITY – PART ONE OF A TWO-PART HA STUDY | 1 |
| EXECUTIVE SUMMARY | 1 |
| INTRODUCTION | 1 |
| SUMMARY OF OVERALL FINDINGS | 1 |
| METHOD AND INTERPRETATION..... | 3 |
| OFFERINGS EVALUATED | 4 |
| FUNCTIONAL CATEGORIES | 4 |
| RESULTS OVERVIEW – BY VENDOR | 6 |
| FIRST PLACE – COMPAQ | 6 |
| SECOND PLACE – IBM..... | 6 |
| THIRD PLACE – HP..... | 6 |
| FOURTH PLACE – DG | 7 |
| FIFTH PLACE – SUN..... | 7 |
| SIXTH PLACE – SEQUENT..... | 7 |
| SECTION ONE: CLUSTER BACKUP AND RECOVERY | 9 |
| INTRODUCTION..... | 10 |
| FAILOVER MODES | 11 |
| DETECTION METHODS | 11 |
| RECOVERY OPTIONS..... | 12 |
| OTHER RELEVANT FEATURES | 12 |
| GENERAL DISCUSSION..... | 13 |
| VENDOR DETAILS..... | 13 |
| HP – FIRST PLACE | 13 |
| COMPAQ – SECOND PLACE, ABOVE AVERAGE..... | 14 |
| DG, SUN – TIED FOR THIRD PLACE, SLIGHTLY ABOVE AVERAGE | 14 |
| SEQUENT – FIFTH PLACE, BELOW AVERAGE | 15 |
| IBM – SIXTH PLACE, BELOW AVERAGE..... | 15 |
| SECTION TWO: CLUSTER FAILOVER CONFIGURABILITY | 16 |
| INTRODUCTION..... | 16 |
| NODES/HEARTBEAT MEDIUMS..... | 17 |
| RAID | 17 |
| MIDDLEWARE/NFS | 17 |
| LAN/WAN COMMUNICATIONS SWITCHOVER | 18 |
| DISK AND TAPE SWITCHOVER | 18 |
| OTHER RELEVANT FEATURES | 18 |
| GENERAL DISCUSSION..... | 19 |
| VENDOR DETAILS..... | 19 |
| IBM – FIRST PLACE | 19 |
| HP – SECOND PLACE, ABOVE AVERAGE..... | 20 |
| COMPAQ – THIRD PLACE, ABOVE AVERAGE | 20 |
| DG – FOURTH PLACE, BELOW AVERAGE | 20 |
| SUN – FIFTH PLACE, BELOW AVERAGE | 21 |
| SEQUENT – SIXTH PLACE, BELOW AVERAGE..... | 21 |
| SECTION THREE: CLUSTER CONCURRENT DATABASE ACCESS..... | 22 |
| INTRODUCTION..... | 22 |
| CLUSTER CONFIGURATIONS..... | 23 |
| VIRTUAL RAW DISK ACCESS | 23 |
| LOCK MANAGEMENT..... | 23 |

| | |
|---|-----------|
| RELEVANT FEATURES | 24 |
| GENERAL DISCUSSION | 24 |
| VENDOR DETAILS | 24 |
| <i>COMPAQ – FIRST PLACE</i> | 24 |
| <i>IBM – SECOND PLACE, ABOVE AVERAGE</i> | 25 |
| <i>SUN – THIRD PLACE, BELOW AVERAGE</i> | 25 |
| <i>SEQUENT – FOURTH PLACE, BELOW AVERAGE</i> | 25 |
| <i>DG AND HP – TIED FOR FIFTH, BELOW AVERAGE</i> | 26 |
| SECTION FOUR: CLUSTER HIGH-AVAILABILITY ADMINISTRATION..... | 27 |
| INTRODUCTION | 27 |
| ADMINISTRATIVE INTERFACES | 28 |
| RECOVERY SCRIPTS | 28 |
| LOAD-BALANCING | 28 |
| ONLINE CLUSTER MODIFICATIONS | 28 |
| OTHER RELEVANT FEATURES | 29 |
| GENERAL DISCUSSION | 29 |
| VENDOR DETAILS | 30 |
| <i>COMPAQ – FIRST PLACE</i> | 30 |
| <i>IBM AND SEQUENT – TIED FOR SECOND PLACE, ABOVE AVERAGE</i> | 30 |
| <i>DG AND HP – TIED FOR FOURTH, BELOW AVERAGE</i> | 31 |
| <i>SUN – SIXTH PLACE, BELOW AVERAGE</i> | 32 |
| SECTION FIVE: CLUSTER SINGLE-SYSTEM IMAGE | 33 |
| INTRODUCTION | 33 |
| CLUSTER FILE SYSTEMS | 34 |
| OTHER CLUSTER-WIDE RESOURCES | 34 |
| RELEVANT FEATURES | 34 |
| GENERAL DISCUSSION | 35 |
| VENDOR DETAILS | 35 |
| <i>COMPAQ – FIRST PLACE</i> | 35 |
| <i>DG – SECOND PLACE, ABOVE AVERAGE</i> | 36 |
| <i>SEQUENT – THIRD PLACE, SLIGHTLY BELOW AVERAGE</i> | 36 |
| <i>IBM – FOURTH PLACE, BELOW AVERAGE</i> | 36 |
| <i>HP – FIFTH PLACE, BELOW AVERAGE</i> | 37 |
| <i>SUN – SIXTH PLACE, BELOW AVERAGE</i> | 37 |
| SECTION SIX: DISASTER RECOVERY | 38 |
| INTRODUCTION | 38 |
| WIDE-AREA FAILOVER AND CONCURRENT ACCESS | 39 |
| REMOTE DATA REPLICATION/MIRRORING | 39 |
| DISTANCE RESTRICTIONS | 40 |
| RELEVANT FEATURES | 40 |
| GENERAL DISCUSSION | 40 |
| VENDOR DETAILS | 41 |
| <i>IBM – FIRST PLACE</i> | 41 |
| <i>SUN – SECOND PLACE, ABOVE AVERAGE</i> | 41 |
| <i>HP – THIRD PLACE, ABOVE AVERAGE</i> | 41 |
| <i>SEQUENT – FOURTH PLACE, BELOW AVERAGE</i> | 42 |
| <i>DG – FIFTH PLACE, BELOW AVERAGE</i> | 42 |
| <i>COMPAQ – SIXTH PLACE, BELOW AVERAGE</i> | 42 |



SECTION ONE: *Cluster Backup and Recovery*

INTRODUCTION

Since detection and recovery mechanisms fulfill a primary objective for availability, DHBA weighted the Cluster Backup and Recovery functional category as 20% of the total score. Significant features evaluated within this category concern failover modes, detection methods, and recovery options. This study assessed backup and recovery options, including assignment of backups for services, flexibility in selecting local versus failover recovery, parallel versus serial recovery, and a range of such cluster-detectable conditions as failures and resource shortages.

Figure 4 shows the results of the Cluster Backup and Recovery functional category, displayed by vendor. The horizontal line indicates the vendor average.

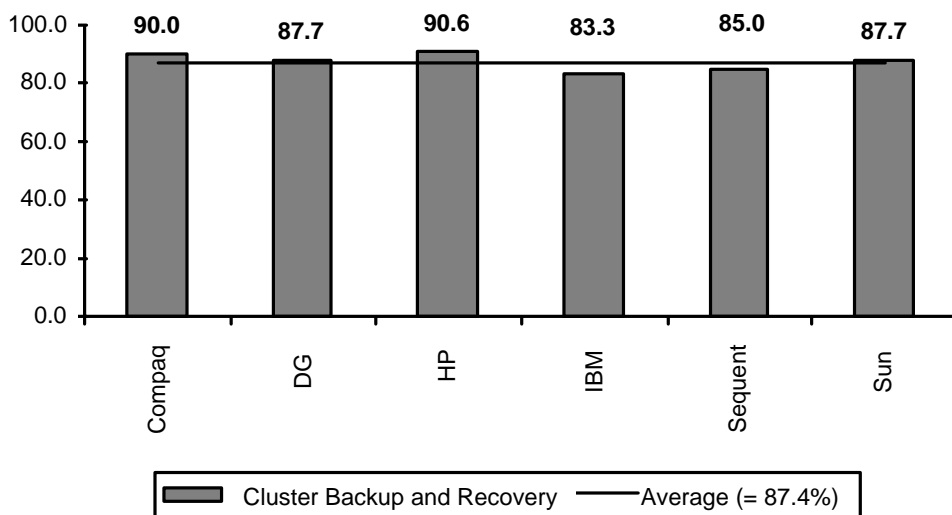


FIGURE 4:
Cluster Backup and Recovery

FAILOVER MODES

This study scored vendor products on the following failover-mode capabilities:

- Multiple-node backup – should enable one node to backup multiple nodes, providing flexibility to assign backups to resources.
- Multiple-node failover – should enable one node to failover to multiple nodes, allowing flexible load balancing after a node fails.
- Cascading failover (multiple backups) – should enable recovery from multiple subsequent node failures by defining a prioritized list of backup nodes to each resource. This capability ensures that recovery will occur if a defined backup system is down at the time of failure.
- Selective failover (partial failover) – should enable failover of a subset of resources (e.g., applications, services), generally those affected by a failure, rather than failing over all resources. This function allows other resources deployed on the node to remain unaffected.
- Dynamic backup selection – should support backup selection, simplifying load balancing after a resource fails by allowing the cluster to select a backup node at the time of failure, as opposed to predefining the backup.

DETECTION METHODS

This study scored vendor products on the following detection methods:

- Node-failure detection – should use multiple paths and multiple mediums (i.e., protocols) to ensure reliable detection of failed nodes via a heartbeat mechanism. If TCP/IP alone is used for heartbeat messages, a TCP/IP failure may be mistaken for a node outage.
- Failed/hung application detection – should allow application failures to be detected and should support subsequent recovery. This study also evaluated products' ability to detect application failures without modifying applications.
- Predictive failover – should enable cluster software to initiate recovery actions in advance of a predicted pending condition, not just in response to a failure. The user may need to create custom agents for examining system conditions. Advance warning of pending failure conditions may result in faster recovery and enables advance notification to users of imminent failovers.
- Supplied software (e.g., agent) – should detect when CPU utilization, file system free space, or system swap space exceeds a defined threshold, so cluster software can invoke a customer-defined action, such as generating an alert or an automated recovery procedure (e.g., failover).
- Agents – All agents should be monitored and automatically restarted by the cluster software if one fails.
- Split brain – Systems should avoid a split brain, which can result when a network failure isolates one or more nodes from the cluster. Each cluster fragment may think it is the cluster, which can cause operational problems.

RECOVERY OPTIONS

This study scored vendor products on the following recovery options:

- Choice of local or failover recovery – should allow an administrator to specify local or failover recovery in response to various hardware and application failures.
- Manual failover confirmation option – should suspend recovery until an operator manually confirms, allowing an administrator to determine whether the pending recovery action is desirable at the time of failure. A failover could result in an overloaded cluster node, for example, and this feature could allow an administrator to avoid such a situation.
- Configurable delay before takeover – should allow a failed system to reboot and restart its assigned applications, rather than failing them over to a backup node.
- Parallel recovery of resources – should support parallel recovery, both within each node and across nodes, which may enable faster recovery compared with serial recovery.
- Restored system becomes backup – should support this capability upon restoration, ensuring that resources remain backed up. The user should be able to define whether this occurs automatically or if it requires operator intervention.
- Resources return to original node – should support this capability upon restoration, ensuring that available capacity is made available to applications as soon as possible. The user should be able to define if this occurs automatically or if it requires operator intervention.

OTHER RELEVANT FEATURES

This study scored vendor products on these other relevant features in the Cluster Backup and Recovery functional category:

- Disk ownership changes without remount – should support this capability to enable transparent recovery for disk data upon failover.
- File System Check (FSCK) not required³ – should not require a time-consuming FSCK in order to repair a damaged file system, yielding faster recovery after a system crash.
- Configuration status monitoring API – should support this API, allowing clients to become cluster aware and to participate in recovery actions.
- Disk corruption – should be avoided using a mechanism that prevents a failed or failing system from writing to a disk which has had its ownership transferred to another node.

³ When using a journaled file system.

GENERAL DISCUSSION

The relatively high overall average in this important category and the small divergence among vendors' scores illustrate the maturity of these cluster solutions. Features common to all vendors include all failover modes evaluated, with the exception of dynamic backup selection. All vendors also support journaled file systems and the ability to detect failed disk adapters, LAN adapters, and applications, and to recover either locally or via failover. Further, all vendors support dual heartbeat paths and parallel recovery within each node and across nodes. Moreover, all support a configuration monitoring API, predictive failover, and mechanisms to avoid disk corruption (resulting from failover) and split-brain conditions.⁴ Finally, all vendors support recovery from hung data services and applications. In this category, no vendors deliver application failover with preserved application context. Otherwise, one or more vendors offer all features in the category.

Areas of difference include the following:

- the range of detectable conditions, including those that enable prediction of pending failures;
- the ability to integrate customer-defined agents for monitoring specific system conditions, as well as the number of agents provided by the vendor;
- the ability to recover from WAN adapter failures;
- some vendors offer dynamic backup selection, while others do not;
- some vendors support disk-volume ownership changes without a remount required by the backup node, while others do not;
- some vendors support configurable manual intervention before failover and configurable threshold delay before takeover, while others do not; and
- vendors differ in the ability to failover disk volumes without requiring a remount.

VENDOR DETAILS

The following sections describe individual vendors' capabilities in the Cluster Backup and Recovery functional category:

HP – FIRST PLACE

HP scores highest in this category. In addition to features common to all vendors, HP supports dynamic backup selection. This is achieved by defining a MC/ServiceGuard "package" failover to the operational node with the fewest number of "packages" currently deployed. Administrators define a "package" as the group of resources required for some application to operate. Package

⁴ See the definition of split brain on page 11.

contents may include application processes, disk drives, logical volumes, IP addresses, and databases. In the event of a failover, all resources within a package failover together.

HP also supports recovery from WAN adapter failures, both local and via failover. Further, HP supplies agents that monitor CPU queue length and file system free space. Finally, MC/ServiceGuard support manual intervention to confirm failover and configurable threshold before takeover.

HP could also improve by delivering an agent for monitoring low swap space conditions. Further, MC/ServiceGuard could be enhanced to monitor agents for failure and automatically restart failed agents. Moreover, as MC/ServiceGuard does not support a clustered file system, HP has yet to address changes in disk-volume ownership without requiring a remount.

COMPAQ – SECOND PLACE, ABOVE AVERAGE

Compaq scores above average. Compaq supports recovery from WAN adapter failures – both local and via failover. Further, Compaq's cluster file system enables disk volume ownership changes without requiring a remount. Moreover, TruCluster software features dynamic backup selection. Like HP, TruCluster software uses the number of resource groups deployed on each node as the metric to determine which node takes over a resource group.

Compaq could improve by offering agents that monitor CPU utilization, swap space, and file system free space. Further, Compaq has yet to address automatic restart of failed agents. Moreover, TruCluster Server software could be enhanced with optional manual intervention to confirm failover and configurable threshold before takeover.

DG, SUN – TIED FOR THIRD PLACE, SLIGHTLY ABOVE AVERAGE

DG scores slightly above average. DG delivers an agent for monitoring file system free space. Further, DG/UX Clusters monitor all agents for failure, enabling restart of a failed agent, a feature only offered by one other vendor, Sun.

DG/UX clusters support manual intervention to confirm failover and configurable threshold before takeover. Finally, DG's cluster file system allows disk volume ownership changes (i.e., path changes) without requiring a remount.

DG has yet to offer agents that monitor CPU utilization and swap space. DG could also improve by supporting recovery from WAN adapter failures.

Sun scores slightly above average. Sun enables recovery from WAN adapter failures via both local recovery and failover. Further, Sun Clusters monitor and restart failed agents ("fault monitors").

Sun could improve by offering agents to monitor CPU utilization, swap space, and file system free space. Sun has also yet to address optional manual intervention to confirm failover and a configurable threshold before takeover. Finally, Sun has yet to offer a cluster file system allowing changes of disk-volume ownership without requiring a remount.

SEQUENT – FIFTH PLACE, BELOW AVERAGE

Sequent scores below the average score for vendors in this study. The company's ptx/CLUSTERS support dynamic backup using ptx/CTC to manage the location of resources when a failover occurs.

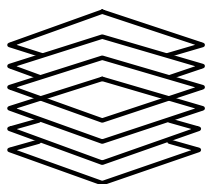
Sequent also supports configurable threshold before takeover. Further, Sequent's cluster file system enables changes to disk-volume ownership without requiring a remount.

Sequent could improve by supporting recovery of WAN adapter failures. Further, Sequent's ptx/CLUSTERS do not support multiple protocols for heartbeat messages; it supports only TCP/IP. Sequent could also improve by offering integrated agents for monitoring CPU utilization, swap space, and file system free space, as well as monitoring agents for failure. Finally, Sequent has yet to offer optional manual intervention to confirm failover.

IBM – SIXTH PLACE, BELOW AVERAGE

IBM follows Sequent, scoring below the average score for vendors in this study. IBM offers agents that monitor CPU utilization, swap space, and file system free space. Further, HACMP features configurable threshold before takeover.

IBM could improve by adding the capability to look for and restart failed agents. Further, IBM has yet to address supporting recovery from WAN adapter failures; optional manual intervention to confirm failover, and disk-volume ownership changes without requiring a remount.



SECTION TWO: *Cluster Failover Configurability*

INTRODUCTION

The Cluster Failover Configurability functional category covers options for hardware and system software in clusters configured for failover.⁵ Hardware should include different classes of servers, disks, and networks for heartbeat and client connectivity. Software should include mixed operating-system versions (enabling rolling upgrades), middleware alternatives, and RAID software. Capability for flexible configuration protects users' existing investments in peripherals, LANs, servers, and applications.

Since Cluster Failover Configurability functions enable important flexibility for users in configuring clusters, DHBA weighted this functional category as 20% of the total score. Significant features evaluated within this category concern nodes/heartbeat mediums, RAID, middleware/NFS, TCP/IP address takeover, and disk interconnects.

Figure 5 shows the results of the Cluster Failover Configurability assessment on a per-vendor basis. The horizontal line indicates the vendor average.

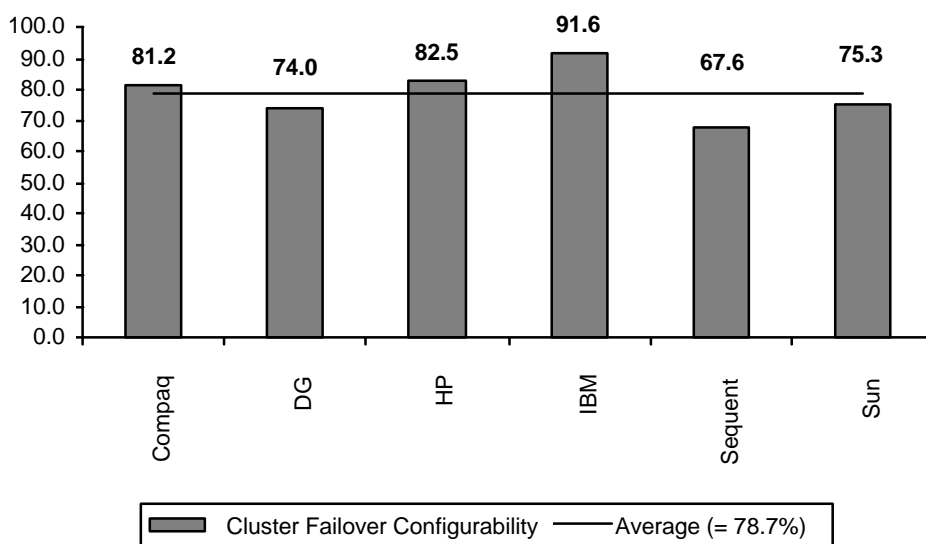


FIGURE 5:
Cluster Failover Configurability

⁵ As opposed to concurrent access, which is evaluated in the “Cluster Concurrent Database Access” chapter.

NODES/HEARTBEAT MEDIUMS

This study scored vendor products on the following criteria for nodes/heartbeat mediums:

- Number of nodes (and CPUs) supported in HA configuration⁶ – should support as many nodes as possible, which (up to a limit) provide flexibility and enable deployment of high-volume applications on a single cluster.
- The points credited to a vendor increase as the maximum number of supported nodes increases above two, above four, above eight, and above 16.
- Number of nodes connected to disks – should optionally allow (but not require) as many nodes as possible to physically connect to disks, increasing flexibility in assigning primary and backup owners of the disks. This number supported is also important because cascading failover of data services is limited to the number of nodes connected.
- Range of node interconnects supported – support for heartbeat messages should include Ethernet, Fast Ethernet, Gigabit Ethernet, FDDI, FCS, ATM, token ring, memory-based interconnects, high-speed switches, target mode SCSI (or other disk interconnect based), and serial line. The last two mediums offer the benefit of not requiring an additional I/O slot (as would be required for a LAN medium).

RAID

This study scored vendor products on the following criterion for RAID:

- RAID 1 and RAID 5 – clusters should support both hardware and software-based implementations.

MIDDLEWARE/NFS

This study scored vendor products on the following criteria for middleware/NFS:

- NFS locked failover – should support preserving NFS lock data upon failover.
- TP monitors – should be certified to operate within clusters.

⁶ i.e., failover versus concurrent access.

LAN/WAN COMMUNICATIONS SWITCHOVER

This study scored vendor products on the following criteria for LAN/WAN communications switchover:

- TCP/IP address takeover – should be supported on Ethernet, Fast Ethernet, Gigabit Ethernet, FDDI, token ring, FCS, and ATM LANs.
- SNA and X.25 WAN lines – should be switchable between nodes.

DISK AND TAPE SWITCHOVER

This study scored vendor products on the following criteria for disk and tape switchover:

- Support for a range of disk and tapes – clusters should support such disk types as SCSI, UltraSCSI, FCS, FC-AL, and FC switched multipath.
- Tape drive failover – should support tape failover.

OTHER RELEVANT FEATURES

This study scored vendor products on these other relevant features in the Cluster Failover Configurability functional category:

- Single boot disk for all cluster nodes – should support this capability to minimize the number of boot disks, reducing overall cost.
- Single boot server for all cluster nodes – should support single boot server, as this enables use of smaller servers with fewer I/O slots.
- Diskless nodes – should be supported, allowing customers to deploy lower cost nodes.
- Optimized protocol – for heartbeat traffic should be used to lower network overhead.
- Heartbeat messages – should be supported on Ethernet, Fast Ethernet, Gigabit Ethernet, FDDI, token ring, ATM LANs, serial lines, and through disk subsystems (i.e., target mode SCSI)

GENERAL DISCUSSION

The divergence between high and low scores indicates that this category constitutes an area of differentiation among vendors. All vendors support four or more nodes⁷; hardware-based RAID 1 and RAID 5; and TCP/IP address takeover on Fast Ethernet, FDDI, token ring, and ATM networks. This means that all vendors evaluated support configurations large and flexible enough for a broad range of application requirements. Further, all support SCSI disks and at least one TP monitor. None of the vendors evaluated support a single boot server or diskless cluster nodes.

Solutions differ in support for larger numbers of nodes, mediums supported for heartbeat messages, optimized protocol for heartbeat, TCP/IP address failover, disk interconnects supported, switchover of WAN lines and tape drives, software-based RAID 5, and support for mixed operating-system versions.

VENDOR DETAILS

The following sections provide information on the status of individual vendors in the Cluster Failover Configurability functional category:

IBM – FIRST PLACE

IBM scores highest in this category. In addition to features common to all vendors, IBM gets credit for supporting greater than 16 nodes in failover configurations. Of the products evaluated, only HACMP supports up to 32 nodes. Using the 24-way SMP RS/6000 Model S80, a HACMP cluster may be configured with up to 768 CPUs.

IBM supports heartbeat and TCP/IP address takeover on all networks evaluated, recovery of X.25 and SNA sessions (via re-dialing), and FC and FC-AL disk interconnects. Moreover, HACMP allows mixed-heartbeat mediums, uses an optimized protocol for heartbeat, allows mixed operating system versions to be deployed, and supports SCSI and UltraSCSI disk interconnects.

IBM could improve by also supporting FC switched multipath. IBM has also yet to address failover of dedicated (versus dial-up) X.25 and SNA WAN lines.

⁷ Note that this study assigns points based on number of nodes configurable. The total number of CPUs configurable, a function of both maximum number of nodes and maximum CPUs per server, is presented for information purposes.

HP – SECOND PLACE, ABOVE AVERAGE

HP follows IBM, scoring above average. MC/ServiceGuard supports up to 16 nodes. Using 32-way SMP V-class servers, up to 512 CPUs may be deployed.

MC/ServiceGuard heartbeat messages may travel over Ethernet, Fast Ethernet, FDDI, token ring, and serial lines. Moreover, mixed-heartbeat mediums are supported, as are mixed operating system versions. HP also supports failover of dedicated X.25 and SNA WAN lines between cluster nodes. Finally, HP supports all disk interconnects evaluated, excluding FC switched multipath.

HP could improve by supporting greater than 16 nodes. Further, HP has yet to address heartbeat over Gigabit Ethernet and ATM, disk-based heartbeat, and optimized protocol for heartbeat. Finally, HP could also improve by supporting FC switched multipath disk interconnects.

COMPAQ – THIRD PLACE, ABOVE AVERAGE

Compaq follows closely behind HP, also above average. TruCluster Server software supports up to eight nodes. Up to 256 CPUs may be deployed using 32-way SMP servers.

TruCluster Server heartbeat messages may travel over Ethernet, Fast Ethernet, FDDI, ATM LANs, serial lines, and via disk. Further, Compaq's Memory Channel interconnect features an optimized protocol for heartbeat. Mixed-heartbeat mediums are also supported. Compaq also enables failover of dedicated X.25 and SNA WAN lines using LAN-based routers.

TruCluster Server solutions may be implemented using a shared boot disk. Further, configurations may be deployed using SCSI or UltraSCSI disk interconnects. Finally, shared tape drive failover is also supported.

Compaq could improve by supported more than eight nodes. Further, Compaq has yet to address heartbeat over Gigabit Ethernet and token ring. Finally, Compaq could improve by supporting FC, FC-AL, and FC switched multipath disk interconnects.

DG – FOURTH PLACE, BELOW AVERAGE

DG finishes fourth, below the average score for vendors in this study. DG/UX Clusters may be configured with up to four nodes. Using 32-way SMP AViiON servers, up to 128 CPUs may be configured.

Ethernet, Fast Ethernet, and disk-based heartbeat mediums are supported. Further, mixed-heartbeat mediums and mixed operating system versions are allowed. DG/UX Clusters support SCSI, FC, and FC-AL disk interconnects.

Finally, DG supports a single boot disk and is one of only two vendors (DG and Sequent) supporting a FC switched multipath disk interconnect.

DG could improve by increasing the number of nodes configurable above four. Further, DG has yet to address heartbeat over Gigabit Ethernet, FDDI, ATM LANs, token ring, and serial lines. Moreover, DG/UX Clusters could improve by enabling failover of X.25 and SNA WAN lines between nodes. Finally, DG could also improve by supporting UltraSCSI disk interconnects.

SUN – FIFTH PLACE, BELOW AVERAGE

Sun places fifth, closely behind DG and below the average score for vendors in this study. Sun Clusters support a maximum of four nodes. Using 64-way SMP E10000 servers, up to 256 CPUs may be deployed.

Ethernet, Fast Ethernet, or Gigabit Ethernet may be deployed as the heartbeat message medium. Further, Sun's SCI (node interconnect) features an optimized protocol for heartbeat messages. Sun Clusters support all disk interconnects evaluated, excluding FC switched multipath. Of all vendors evaluated, only Sun offers software-based RAID5.

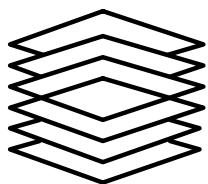
Sun Clusters could improve by supporting greater than four nodes. Sun could also improve by offering FDDI, ATM, token ring, disk-based, and serial-line heartbeat. Sun has also yet to address mixed-heartbeat mediums and mixed operating systems. Finally, Sun could enhance its offering with shared tape failover (the physical connection of a tape drive to two or more nodes), a feature offered by all other vendors evaluated.

SEQUENT – SIXTH PLACE, BELOW AVERAGE

Sequent finishes sixth, below the average score for vendors in this study. Ptx/CLUSTERS may be configured with up to four nodes. Using 64-way SMP NUMA-Q 2000 servers, up to 256 CPUs are supported.

All heartbeat mediums evaluated are supported, excluding Gigabit Ethernet and disk-based, mixed-heartbeat mediums; mixed operating system versions are also allowed. Finally, all disk interconnects evaluated may be deployed, excluding UltraSCSI.

Sequent could improve by supporting more than four nodes. Further, of all vendors evaluated, only Sequent has yet to address NFS locked failover. Finally, Sequent could improve by addressing failover of X.25 and SNA WAN lines.



SECTION THREE: *Cluster Concurrent Database Access*

INTRODUCTION

Concurrent access allows multiple cluster nodes to share a single physical database. It requires a Distributed Lock Manager (DLM), which coordinates simultaneous requests to update shared data so multiple users do not overwrite each other's updates. Concurrent access is a scalability enabler, but also promotes faster recovery, thus providing availability benefits.

Features evaluated in this functional category concern cluster configurations, virtual raw disk access, and lock management. Since concurrent access is not required within all clusters, DHBA weighted this functional category as 10% of the total score.

Figure 6 shows the results of the Concurrent Access functional category analysis by vendor. The horizontal line indicates the vendor average.

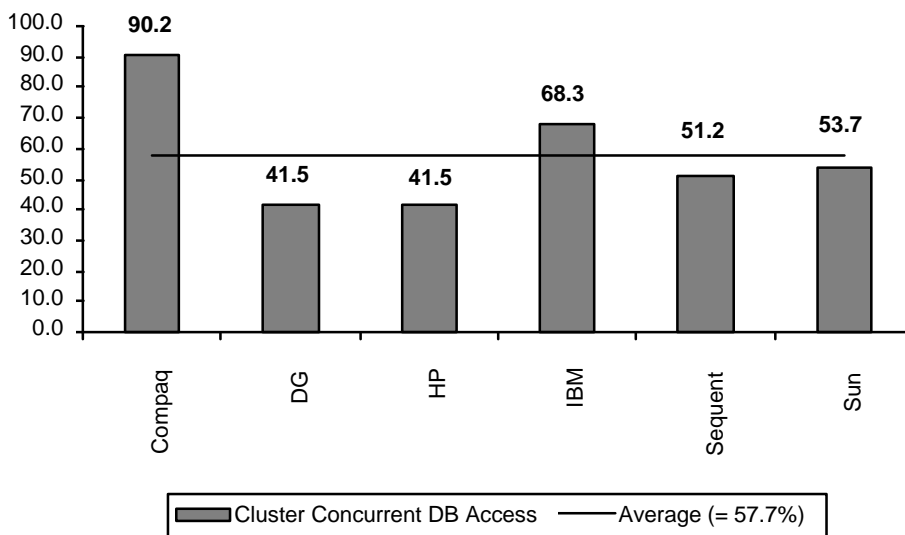


FIGURE 6:
Concurrent Access

CLUSTER CONFIGURATIONS

This study scored vendor products on the following criteria for cluster configuration:

- Concurrent access configurations – should support at least four nodes.
- RAID 5 – should support RAID 5 in both hardware- and software-based implementations.
- Shared-nothing parallel databases – should support shared-nothing implementations such as Informix XPS.
- Node interconnects – should support interconnects that provide high performance.
- Lock manager messages – should use optimized protocols, rather than the full TCP/IP stack.

VIRTUAL RAW DISK ACCESS

This study scored vendor products on the following criterion for virtual raw disk access:

- Virtual raw disk access – should be supported so nodes do not require physical connections to disks. Virtual raw disk makes transparent the fact that a disk is not physically connected to a node. With virtual raw disk, nodes require fewer I/O slots for disk. Moreover, the number of nodes in concurrent access configurations is not subject to disk multiporting limitations. Virtual raw disk should be implemented using Direct Memory Access (DMA) for enhanced performance. Further, nodes with physical connection to disks should be able to use direct (true concurrent) disk access, while other nodes use indirect access.

LOCK MANAGEMENT

This study scored vendor products on the following criteria for lock management:

- Multithreaded and kernel-based DLM – should support this capability to enable higher DLM performance.
- Redundant DLM paths – should be used.
- The DLM API – should be available to user applications.

RELEVANT FEATURES

This study scored vendor products on these other relevant features in the Cluster Concurrent Database Access functional category:

- Wide-area concurrent access – should be supported so remote nodes can concurrently access a database. This requires virtual raw disk capabilities.
- Mixed modes – should support both concurrent access and failover on the same cluster.

GENERAL DISCUSSION

The overall low scores suggest this functional category as an area where vendors can improve. The divergence in scores shows that concurrent access is a vendor differentiator.

All vendors support four or more nodes, offer both shared disk and shared nothing databases, redundant DLM message paths, multithreaded DLMs, mixed failover (HA) and concurrent access modes, and hardware-based RAID 5.

Vendors differ in number of nodes supported, virtual raw disk, DMA, memory-based interconnects for DLM messages, optimized protocol for DLM messages, and software-based RAID5.

No vendor supports more than eight nodes in concurrent access mode. Further, none support wide-area concurrent database access.

VENDOR DETAILS

The following sections provide information on the status of individual vendors in the Concurrent Access functional category:

COMPAQ – FIRST PLACE

Compaq significantly outscores competitors in this category. TruCluster Server software supports up to eight nodes in concurrent database access configurations. Concurrent database access uses virtual raw disk access, a feature that requires a fast node interconnect such as Compaq's Memory Channel.

Compaq's Memory Channel interconnect uses an optimized protocol for DLM messages. Further, both Compaq and IBM use a mixed virtual raw disk and true concurrent access, so any node with a physical connection to disks can use that direct connection, rather than accessing data indirectly over a node interconnect.

Of the vendors evaluated, only Compaq's implementation supports virtual raw disk access using DMA (using Compaq's Memory Channel node interconnect) and dynamic load balancing of requests between hosts with physical disk connections. Only one other vendor, Sun, offers a memory-based node interconnect. Finally, Compaq's DLM is kernel-based and the DLM API is available to user applications

Compaq could improve by offering wide-area concurrent database access and supporting software-based RAID5.

IBM – SECOND PLACE, ABOVE AVERAGE

IBM follows in second place. HACMP allows configurations of up to eight nodes in concurrent database access mode. Further, IBM features optimized protocol for DLM messages, kernel-based DLM, and DLM API for user applications.

IBM could improve by offering a memory-based node interconnect, DMA raw virtual disk, and virtual raw disk access dynamic load balancing.

SUN – THIRD PLACE, BELOW AVERAGE

Sun finishes in third place, below the average score for vendors in this study. Sun Clusters support four nodes in concurrent database access mode. Further, Sun Clusters use the SCI memory based node interconnect and use an optimized protocol for DLM messages. Of the vendors evaluated, only Sun offers software-based RAID (supported with shared-nothing databases – i.e., INFORMIX XPS, not OPS).

Sun could improve by supporting greater than four nodes, virtual raw disk access, DLM API, and kernel-based DLM.

SEQUENT – FOURTH PLACE, BELOW AVERAGE

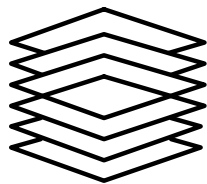
Sequent places fifth, below the average score for vendors in this study. Ptx/CLUSTERS may be configured with up to four nodes. Ptx/CLUSTERS use an optimized protocol for DLM Messages, UDP versus TCP/IP, and use a kernel-based DLM.

Sequent could improve by supporting greater than four nodes, virtual raw disk access, DLM API for applications, and a memory-based node interconnect.

DG AND HP – TIED FOR FIFTH, BELOW AVERAGE

DG and HP tie in fourth place, below the average score for vendors in this study. HP's MC/ServiceGuard OPS edition supports eight nodes, while DG/UX Clusters support four. DG offers a kernel-based DLM.

DG could improve with support for greater than four nodes. HP could improve with a kernel-based DLM. Both DG and HP could improve with virtual raw disk access, memory based interconnect, and a DLM API for applications.



SECTION FOUR: Cluster High-Availability Administration

INTRODUCTION

This functional category evaluates the tools provided to ease overall administration and allow online management of the HA environment. Features evaluated in this functional category concern administrative interfaces, recovery scripts, load balancing, and online cluster modifications. DHBA weighted it as 5% of the total.

Figure 7 shows the results of the Cluster High-Availability Administration functional category analysis by vendor. The horizontal line indicates the vendor average.

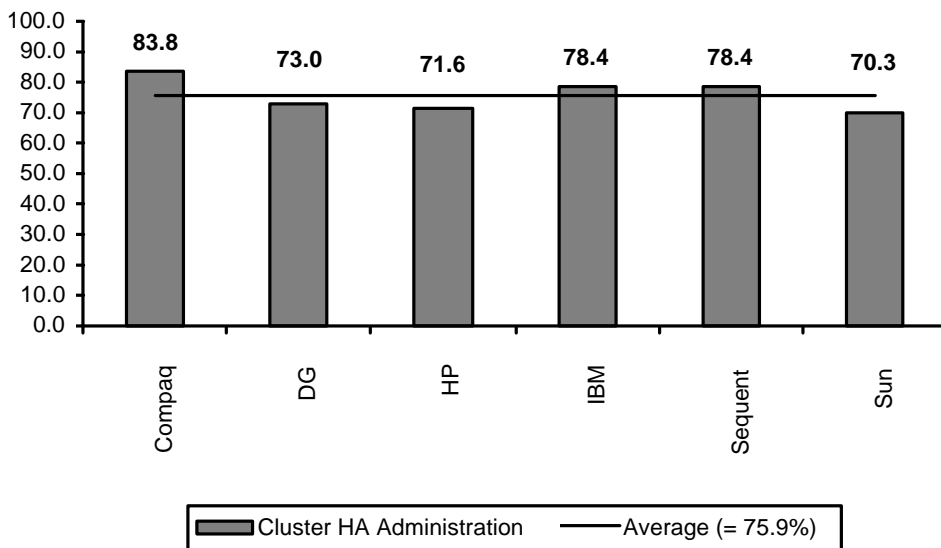


FIGURE 7:
**Cluster High-Availability
Administration**

ADMINISTRATIVE INTERFACES

This study scored vendor products on the following criteria for administrative interface:

- Single point of operations – should provide this capability for such tasks as boot, cluster definition, cluster monitoring, and cluster management.
- GUI-based controls – should support GUI and/or browser-based monitors and management interfaces.
- Command-line controls – should also support command-line controls enabling scripting and optimized remote communications.
- Multiple clusters – should be managed from a single point.

RECOVERY SCRIPTS

This study scored vendor products on the following criterion:

- Vendor-supplied scripts – should offer scripts for resource recovery.

LOAD-BALANCING

This study scored vendor products on the following criteria for load balancing:

- Manual load balancing – should support manual load balancing via failover of selected applications.
- Dynamic assignment of processes – should enable dynamic assignment to nodes at process startup time.
- Active process migration – should support migration between nodes, enabling non-disruptive load balancing.
- Dynamic/automatic load balancing – should make this type of load-balancing possible without operator intervention.

ONLINE CLUSTER MODIFICATIONS

This study scored vendor products on the following criteria for online cluster modification:

- Addition of resources – should enable additions to the cluster configuration, including nodes, devices, and applications, without taking down the cluster software.
- Recovery scheme modifications – should make possible modifications such as the assignment of backups to resources without taking down the cluster software.

OTHER RELEVANT FEATURES

This study scored vendor products on this other relevant feature in the Cluster HA Administration functional category:

- Process resource allocation – should support assignment of resources to groups of processes.
- SNMP traps – should generate traps by the cluster software, allowing SNMP-compliant network monitoring solutions to track cluster events.
- Logging – the system should log configuration changes.
- Availability tools – should provide tools to measure uptime and to track causes of downtime by application.
- Automatic failback – should allow configuration of automatic failback based on time of day.
- Offline support – should allow systems and cluster management operations to be performed on offline nodes so the operation need not be repeated when a node comes online.
- Rolling upgrades – should support rolling upgrades of both operating system and cluster software to reduce planned downtime.

GENERAL DISCUSSION

The relatively small divergence in vendor scores within DHBA's overview analysis shows only some vendor differentiation in the area of Cluster HA Administration.

All vendors support GUI-based interfaces for command and monitoring, configuration capturing to a file, management of multiple clusters from a single point, and grouping of resources in to failover groups. Further, all offer recovery scripts for recovery of databases and LANs and support manual load-balancing and online cluster administration. Moreover, all offer process-resource-allocation management, online addition/removal of services and nodes to/from the cluster configuration, and rolling operating-system upgrades.

Vendors differ in support for cloning configurations, browser-based interfaces, command-line interfaces, SNMP-compliant traps, application recovery scripts, load balancing, configuration change auditing, rolling cluster software upgrade, automatic failback based on time of day, and the ability to manage offline nodes. Further, vendors differ in delivering tools to monitor and track availability.

No vendor currently offers active process migration and dynamic load balancing using active process migration. Further, none offer tools that track uptime by application.

VENDOR DETAILS

The following sections provide information on the status of individual vendors in the Cluster HA Administration functional category.

COMPAQ – FIRST PLACE

Compaq outscores the competition in this category. Compaq offers the full range of command and monitoring interfaces evaluated. The TruCluster Server design layers web, GUI, and command-line management interfaces on top of a single management application.

Compaq offers TruCluster Server scripts for all resources evaluated. Compaq also features configuration cloning and drag-and-drop manual load balancing. Other Compaq features include cluster MIB, SNMP traps, and rolling upgrades to operating-system and cluster software. Further, systems management may be performed on offline nodes by deploying system and cluster configuration information on a shared cluster file system. Finally, TruCluster Server software provides load balancing upon failover,⁸ under the dynamic backup selection feature.

Compaq could improve by supporting automatic load balancing based on node resource utilization (or user-supplied algorithm), dynamic process-to-node assignments, system and cluster configuration change audit trails, uptime-measuring tools, and automatic failback based on time of day.

IBM AND SEQUENT – TIED FOR SECOND PLACE, ABOVE AVERAGE

IBM and Sequent tie for second in this category. IBM's HACMP allows configurations to be cloned by taking a snapshot of an existing configuration file and deploying it on an identical hardware configuration. A browser-based interface is provided for configuration definition, as well as a GUI-based and a command-line-based command and monitoring interface.

IBM offers sample recovery scripts for databases, TP monitors, LAN, and application recovery (for selected applications). Other IBM features include SNMP trap generation, dynamic process to node assignment (for batch), auditing of system and cluster configuration changes, and cluster MIB.

IBM could improve by offering a browser-based cluster monitor, uptime measuring tools, enabling management of HACMP and NT clusters from a single point, allowing systems management to be performed on offline nodes, and supporting automatic failback based on time of day.

⁸ As discussed in the "Cluster Backup and Recovery" chapter.

Sequent's ptx/CLUSTERS feature GUI-based, browser-based, and command-line based interfaces for command and monitoring. Sequent offers recovery scripts for all resources evaluated, excluding applications. The ptx/CTC system allows load balancing decisions upon failure to be made dynamically by the cluster software based on node utilization. Further, ptx/CTC enables automatic load balancing upon addition of new nodes.

Sequent offers SQLINK, allowing tracking of uptime percentages for individual customer systems. Other features include cluster MIB, automatic failback based on time of day using ptx/CTC, and the ability to perform systems management operations on offline nodes by virtue of a shared cluster file system.

Sequent could improve by offering a high-level interface (i.e., GUI) for script creation. Further, Sequent has yet to address configuration cloning, SNMP traps, drag-and-drop manual load balancing, dynamic process-to-node assignments (at process startup) for load balancing, and tracking causes of downtime.

DG AND HP – TIED FOR FOURTH, BELOW AVERAGE

DG features GUI and command-line interfaces. Further, DG offers scripts for all resources evaluated, excluding TP monitors and applications. Moreover, DG allows management functions to be performed on offline nodes, due to its cluster file system.

DG could improve by allowing configuration cloning, offering browser-based interfaces, SNMP traps, automatic load balancing upon failure and node addition, drag-and-drop load balancing, uptime tracking on an application basis, cluster MIB, a common UNIX/NT management interface, and automatic failback based on time of day.

HP offers GUI-based cluster command and monitoring interfaces. Further, HP offers scripts for all resources evaluated. Further, HP features SNMP-compliant traps, automatic load balancing upon failure (based on number of MC/ServiceGuard "packages" deployed on each node), cluster MIB, and rolling upgrades to operating system and cluster software.

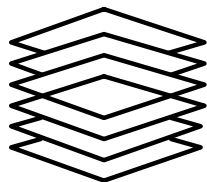
Moreover, HP offers the HP Observatory (HAO), enabling auditing of system and cluster configuration changes. HAO will periodically transfer collected information to HP's service organization in order to assist in any subsequent problem diagnosis.

HP could improve by offering browser- and command-line-based interfaces for monitoring and commands. Further, HP has yet to address configuration cloning. Moreover, HP could enhance its offering with dynamic process-to-node assignments, uptime-measuring tools, automatic failback based on time of day, and the ability to perform systems management on offline nodes.

SUN – SIXTH PLACE, BELOW AVERAGE

Sun supports all the cluster-management interfaces evaluated, excluding a browser-based command interface. Further, Sun offers recovery scripts for all resources evaluated and features SNMP-compliant traps, cluster MIB, and change auditing for system and cluster configuration.

Sun could improve by offering configuration cloning, drag-and-drop manual load balancing, automatic load balancing upon failure, management of offline nodes, uptime-measuring tools, and automatic failback based on time of day. Sun also has yet to address rolling cluster software upgrades and the ability to upgrade a single node to a two-node cluster without re-booting.



SECTION FIVE: *Cluster Single-System Image*

INTRODUCTION

This functional category examines features that ease use and administration by presenting distributed cluster resources as if they were resources within a single server. Features evaluated in this functional category concern cluster file systems and other cluster-wide resources. DHBA weighted it as 5% of the total score.

Figure 8 shows the results of the Cluster Single-System Image category analysis by vendor. The horizontal line indicates the vendor average.

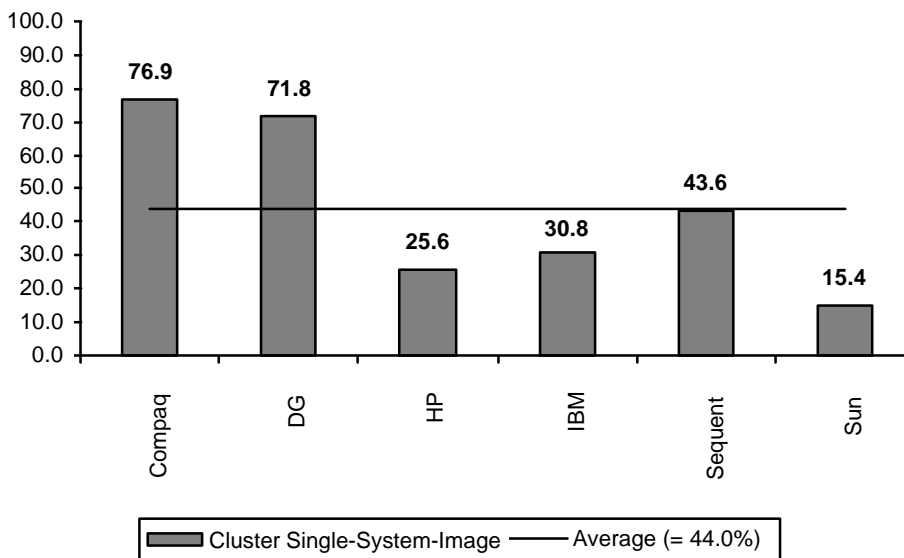


FIGURE 8:
Cluster Single-System Image

CLUSTER FILE SYSTEMS

This study scored vendor products on the following criteria for cluster file system considerations:

- Cluster-wide devices – should support devices that appear locally attached to all cluster nodes. Any application on any node should be able to access any hardware device as if it were a local device.
- Cluster-wide file systems – should support UNIX file systems that are shared across multiple nodes as if they were a single system.
- POSIX semantics – should support POSIX on the cluster file system.
- Root disks – should support root disks on cluster file systems.

OTHER CLUSTER-WIDE RESOURCES

This study scored vendor products on the following criteria for other cluster-wide resources:

- Print and batch queues – should support shared queues by cluster nodes.
- Single password file – should support single password to simplify administration.
- Single IP host table – should support single IP host table for all cluster nodes.
- Tape drives – should be attached to and accessible by multiple nodes, enabling tape-drive failover.
- Inter-process communications (IPC) – should support IPC using shared memory (e.g., streams and pipes) to span nodes.

RELEVANT FEATURES

This study scored vendor products on these other relevant features in the Cluster Single-System Image functional category:

- Single IP address – should support single addresses for an application deployed in multiple instances across multiple nodes. The software should load-balance requests across these multiple instances transparently to clients.
- Distributed process startup – should support distributed process startup to allow processes to be started on other nodes.
- Local swap partition – should not be required on each node, because swapping is performed to a disk attached to another node.
- Inter-process relationships – should apply across nodes (e.g., parent-sibling relationship).

GENERAL DISCUSSION

As illustrated by the low overall scores, vendors face a challenge in this functional category. The divergence in scores reveals single-system image as an area of vendor differentiation.

All vendors offer single boot point (for cluster nodes), single point of cluster definition, shared tape drive (logical sharing), cluster-wide time synchronization, and cluster-wide print queues.

No vendors support cluster-wide devices, cluster-wide shared memory (i.e., streams/pipes), single boot server, or non-local swap partition (i.e., swapping over node interconnect to disk owned by other node). Further, none preserve interprocess relationships across nodes or support a single boot server.

Vendors differ in support for cluster file systems; (root, journaling and POSIX semantics on a cluster file system); indirect and direct physical access to a cluster file system; making all devices appear local; global networking (i.e., a single IP address for multiple instances of a single application); single password and IP host table; distributed process startup; inter-process relationships preserved across nodes; consistent storage path names from any node; cluster-wide process status; and requirements for local boot disk and swap partition.

VENDOR DETAILS

The following sections provide information on the status of individual vendors in the Cluster Single-System Image functional category:

COMPAQ – FIRST PLACE

Compaq scores highest in this category. TruCluster Server software supports a cluster file system (CFS) with POSIX semantics. Further, Compaq's Advanced File System (AdvFS) enables journaling of the CFS, eliminating the need for a time-consuming FSCK (file system check) to be performed after a crash.

Compaq allows deployment of the root file system on its CFS and may be deployed with a single shared logical boot disk. Moreover, only Compaq allows all both direct and indirect access to the CFS using a Device Request Dispatcher (DRD). Nodes with a physical path to the file system may use that path while others access indirectly using the DRD (i.e., requests are served by a node with a physical path).

Compaq also supports a cluster-wide password file, IP host table, cluster alias, single security domain, process status, and volume management. Further, Compaq's Tru64 UNIX operating system provides consistent path names to storage devices from all cluster nodes.

Compaq, as well as all other vendors, could improve by adding any of the features listed above that currently are not offered by any vendor.

DG – SECOND PLACE, ABOVE AVERAGE

DG follows Compaq in this category. Of the products evaluated, DG's DG/UX Clusters was the first to offer a CFS. Further, DG/UX Clusters allow all cluster nodes to share the same root disk file system and support a single boot disk. Further, DG supports a cluster-wide password file, IP host table, cluster alias, single security domain, process status, volume management, and consistent path names.

DG could improve by supporting indirect access to its CFS. Further, DG has yet to support journaling using its CFS.

SEQUENT – THIRD PLACE, SLIGHTLY BELOW AVERAGE

Sequent finishes third, slightly below the average. Sequent's ptx/CLUSTERS support a CFS with POSIX semantics and journaling. Further, Sequent features cluster-wide volume management.

Sequent could improve by supporting the root file system on the CFS, as well as supporting indirect (i.e., served) access to its CFS. Further, Sequent has yet to address cluster alias, cluster-wide password, IP host tables, security, and process status. Finally, Sequent could improve with consistent storage path names from all cluster nodes.

IBM – FOURTH PLACE, BELOW AVERAGE

IBM places fourth, below the average score for vendors in this study. IBM supports a CFS, cluster-wide security, volume management, and process status.

IBM could improve with POSIX compliance for its CFS, as well as support for the root file system. IBM has also yet to address cluster alias and cluster-wide password, IP host table, and process status. IBM could also improve by supporting consistent storage path names from all nodes.

HP – FIFTH PLACE, BELOW AVERAGE

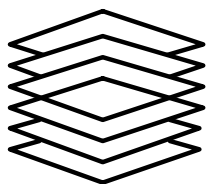
HP finishes fifth, below the average score for vendors in this study. In addition to the features offered by all vendors, HP supports cluster aliases.

HP has yet to address a cluster file system, as well as the remainder of items evaluated in this category.

SUN – SIXTH PLACE, BELOW AVERAGE

Sun scores sixth, below the average score for vendors in this study. Sun offers those features discussed earlier as delivered by all vendors.

Sun could improve by delivering other items evaluated in this category.



SECTION SIX: *Disaster Recovery*

INTRODUCTION

This functional category examines remote data replication and remote (wide-area) failover of applications. Since disaster recovery is a relatively uncommon requirement, DHBA weighted it as 5% of the total score. Features of concern in this functional category include wide-area failover and concurrent access, remote data duplication/mirroring, and distance limitations between disks/nodes.

Figure 9 shows the results of the Disaster Recovery functional category analysis, displayed by vendor. The horizontal line indicates the vendor average.

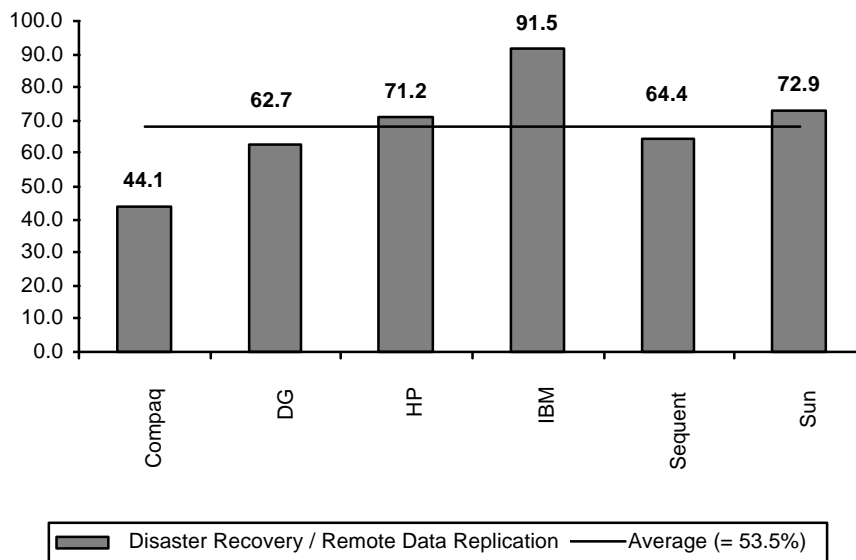


FIGURE 9:
Disaster Recovery

WIDE-AREA FAILOVER AND CONCURRENT ACCESS

This study scored vendor products on the following criteria for wide-area failover and concurrent access:

- Wide-area application failover – should support wide-area application failure, so a node in another geographic location can take over applications in the event of a site disaster.
- Wide-area TCP/IP takeover – should support wide-area TCP/IP takeover in conjunction with wide-area failover, so clients can transparently reconnect to applications after recovery.
- Concurrent access – should enable access to databases and file systems from remote cluster nodes.

REMOTE DATA REPLICATION/MIRRORING

This study scored vendor products on the following criteria for remote data replication/mirroring:

- Remote mirroring – should support remote mirroring, so that data updates to volumes and file systems are also applied to a mirror controlled by a LAN- or WAN-connected node.
- Remote duplication – should support remote duplication of files and databases.
- Asynchronous mode – should support asynchronous mode for files and databases to optimize performance.
- Synchronous mode – should support synchronous data replication to optimize data synchronization.
- Duplication to multiple sites – should support concurrent duplication to multiple sites.
- High-performance resynchronization – should enable high-performance resynchronization after a failure results in unsynchronized data. In other words, the system should not require all data to be recopied.
- Static data duplication – should enable static duplication for periodic batch database updates. This is most useful where identical read-only databases are deployed on multiple nodes for performance reasons.
- Hardware-based remote mirroring – should be supported (within the storage array), offloading this operation from the server CPUs.

DISTANCE RESTRICTIONS

This study scored vendor products on the following criteria for distance restrictions:

- Distance between servers and disks – should be maximized to protect against site disaster.
- Distance between cluster nodes – should be maximized to protect against site disaster.

RELEVANT FEATURES

This study scored vendor products on these other relevant features in the Disaster Recovery functional category:

- Single boot console/interface – should include single boot console/interface for local and remote cluster nodes.
- Split brain – should provide a mechanism to prevent a split-brain syndrome in the event of lost communications between sites.⁹

GENERAL DISCUSSION

The low average in this category suggests vendors have room for improvement. The large divergence in scores identifies disaster recovery as an area of vendor differentiation.

All vendors support database replication products provided by database vendors. Further, all support a single boot console and separation of varying distances between servers and disks, and between cluster nodes. Moreover, all support hardware-based remote mirroring. No vendor supports wide-area concurrent access.

Vendors differ relative to separation distances between servers and disks and between cluster nodes. Further, not all vendors offer a high-level interface for staging distribution of static data to distributed nodes nor the ability to mirror to two or more remote sites in real-time. Lastly, vendors differ in providing remote duplication of logical file systems and logical volumes to remote nodes.

⁹ See the definition of split brain on page 11.

VENDOR DETAILS

The following sections provide information on the status of individual vendors in the Disaster Recovery functional category.

IBM – FIRST PLACE

IBM places first in this category. IBM's HAGEO product enables failover between nodes separated by unlimited distance. This is possible as the GeoMirror function of HAGEO propagates volume, file system, and database updates to a remote system via a wide-area communications link. Therefore, HAGEO does not rely on shared disk cables that limit the distance between nodes in geographic failover configurations.

Using the Enterprise Storage System (ESS), IBM allows a server and its disk drives to be separated by up to 103KM using fiber-optic connections, further than any other vendor's solution allows. A failover solution may also be deployed using the ESS when sites are within the 103KM distance limit. The ESS further enables static data distribution using the "Flash Copy" feature and allows real-time remote mirroring to multiple sites (i.e., more than one other site).

IBM could improve by supporting unlimited distance between servers and disks.

SUN – SECOND PLACE, ABOVE AVERAGE

Sun places second. Sun supports node separation distances of up to 10KM in Sun Cluster Campus Cluster configurations. The Remote Dual Copy (RDC) feature of Sun's StorEdge A7000 Intelligent Storage Server allows individual A7000 arrays to be placed up to 50KM from apart, with one in a remote location. The host-based RDC feature allows duplication of logical volumes and static data distribution.

Sun could improve by allowing servers, disk drives, and cluster nodes to be separated by unlimited distances. Further, Sun has also yet to address file system duplication.

HP – THIRD PLACE, ABOVE AVERAGE

HP finishes third, above average. HP MetroClusters allows cluster nodes to be separated up to 50KM. Further, Continental Clusters allow unlimited separation distances between cluster nodes sharing disks, a feature unmatched by competitors. EMC SRDF allows data to be mirrored to multiple sites (i.e., to other EMC arrays in more than one other location).

HP could improve by offering a tool for static data distribution and by supporting duplication of volumes and file systems to remote nodes.

SEQUENT – FOURTH PLACE, BELOW AVERAGE

Sequent places fourth, below the average score for vendors in this study. Sequent allows cluster nodes sharing disks to be up to 10KM apart. Further, SRDF allows individual EMC arrays to be placed up to 60KM apart, with one in a remote location, and allows data to be replicated to EMC arrays in multiple locations. Moreover, volumes may be duplicated to a remote node/location by placing half a mirror in each location. Sequent's cluster file system allows both nodes to own all disks. Any update to the CFS initiated by a node in one location also gets propagated to the mirror in the remote location.

Sequent could improve by increasing the distance supported between cluster nodes. Sequent has also yet to address remote file system duplication.

DG – FIFTH PLACE, BELOW AVERAGE

DG supports up to 30KM separation distances between DG/UX Cluster nodes using Gigabit Fibre Channel. Volumes may be duplicated to a remote node/location by placing halves of a mirror in each location. DG's cluster file system allows both nodes to own all disks. Any update to the CFS initiated by a node in one location is also propagated to the mirror in the remote location.

DG could improve by offering allowing cluster nodes to be separated by unlimited distances. Further, DG has yet to address a high-level interface for staging static data distribution to remote nodes and real-time remote mirroring to multiple sites.

COMPAQ – SIXTH PLACE, BELOW AVERAGE

Compaq places sixth, below the average score for vendors in this study. Compaq supports up to 15KM separation distances between servers and disk drives using SCSI-fibre extenders.

Compaq could improve by increasing this distance. Further, Compaq has yet to address remote volume and file system duplication. Finally, Compaq could also improve by increasing the distance between servers and disks.

Compaq's TruCluster Server V5.0 will significantly improve its disaster-recovery capability with support of FC disk interconnects later this year.