Flexible Database Clusters with IBM eServer BladeCenter and Oracle9*i* Real Application Clusters (RAC)

Phil Horwitz and Martha Centeno
IBM eServer xSeries Performance Development & Analysis
Research Triangle Park, NC USA

Kevin Closson
PolyServe, Inc.
Beaverton, OR USA

Abstract

Modern clusters may rival the established UNIX® based server, but are they manageable? What are the ramifications of performance and availability at the application level? What is the impact on total cost of ownership (TCO)? Answering these questions is the focus of this white paper.

Oracle9*i* RAC is designed to help build flexible, high-performance, highly available, clustered database solutions on Linux. Connecting such clusters to a fault-resilient Fibre Channel storage area network (SAN) lays the foundation for the computing infrastructure known as Flexible Database Clusters (FDC).

The audience for this paper is customers who are interested in implementing an FDC infrastructure with the IBM® eServer® BladeCenter® and Oracle9*i* Real Application Cluster (RAC), PolyServe Matrix Server and the Linux operating system.

Introduction

IBM and PolyServe joined forces to build a 14-node (blades) BladeCenter running SUSE Linux Enterprise Server 8 (SLES 8) and attached it to a formidable SAN configured with more than 100 physical disk drives (see Figure 1). Oracle9*i* RAC was installed with PolyServe Matrix Server to simply the deployment and management of RAC and to enable its key features. The entire system was configured and deployed in one standard 42U rack.

The FDC cluster was the target of a series of tests that took an in-depth look at running and managing not just a single application, but three separate applications. The results of the testing confirmed that Flexible Database Clusters provide:

- A means to consolidate and deploy multiple databases and associated applications in a single, easily managed cluster environment
- Simplified management of large database clusters, made possible by the PolyServe Matrix Server clustered file system
- Dynamic “scalability on demand” architecture, enabling near-linear speedup to running applications—with little or no interruption
- Dynamic repurposing of server resources on demand to quickly and easily move processing capacity to where it is most needed.

- An autonomic, always-on operating environment with fast or even immediate self-healing and little or no performance degradation (and therefore increased utilization rates)
- Dramatic incremental TCO benefits from improved manageability, scalability, expandability, availability and asset utilization.

Flexible Database Cluster Concepts

Economics of consolidation

The cost-efficiency factors associated with building Intel® processor-based clusters running Linux in support of Oracle9*i* RAC are both proven and substantial. The common method for deploying Oracle on Linux is to create one small cluster for each database application. However, just as consolidating applications on a large SMP system is a proven cost-saving action, consolidating applications on a large cluster can provide great benefits as well, particularly for scale-out applications and middleware such as RAC. Using RAC to create a Flexible Database Cluster can yield the benefits of consolidation, which include reduced administrative overhead and increased flexibility.

Flexible Database Cluster Components

The Flexible Database Cluster requires flexible architectures such as those of the IBM eServer BladeCenter, PolyServe Matrix Server software, and Oracle9*i* RAC. These components complement each other to create a powerful platform for supporting multiple applications.

IBM eServer BladeCenter architecture

To support the basic computing infrastructure needed by the Flexible Database Cluster proof-of-concept, it was important to choose a hardware platform that would showcase flexibility and manageability. The IBM eServer BladeCenter provides both of these attributes.

The BladeCenter chassis accommodates up to 14 hot-swap 2-way Intel Xeon™ DP processor-based blade servers in its innovative 7U form factor. Key infrastructure components such as Layer 2-7 Gigabit Ethernet switching, SAN switching, and centralized management tools are also integrated in the chassis.

Also, the chassis is designed to provide resources, such as power, switch, management and blower modules, which are shared among all the blades. The chassis provides high-speed I/O capabilities for all of the modules, enabling aggregated I/O throughput and reducing the amount of cabling required in the data center. The BladeCenter Management Module facilitates remote access to control the components in the enclosure.

Blade technology, such as IBM eServer BladeCenter, enables sites to reduce “server sprawl” and greatly reduce the complexity of their distributed IT infrastructure. Blades also deliver better management software and provide more expansion possibilities with smaller footprint requirements.

The Flexible Database Cluster proof-of-concept required that large amounts of highly available disk storage be connected to the BladeCenter. The storage subsystem was designed around the IBM TotalStorage® FAStT700 Storage Server. The FAStT700 is a RAID storage subsystem that contains Fibre Channel interfaces to connect both the host systems and the disk drive enclosures. With its 2Gbps controllers and high-availability design, the FAStT700 delivers the necessary throughput to support the FDC proof-of-concept.

Single points of failure were greatly reduced in the FDC test system through the combination of a fully redundant switched Fibre Channel SAN and the multi-path I/O feature provided by PolyServe Matrix Server.

For more information about the IBM eServer BladeCenter, visit IBM's Web site:

http://www-1.ibm.com/servers/eserver/bladecenter/

PolyServe Matrix Server

PolyServe Matrix Server enables multiple low-cost Linux- or Windows-based servers to function as a single, easy-to-use, highly available system. Matrix Server includes a fully symmetric cluster file system that enables scalable data sharing, high-availability services that increase system uptime and utilization, and cluster and storage management capabilities for managing servers and storage as one. With Matrix Server, customers gain an unparalleled level of scalability, availability, and manageability for scale-out application and middleware deployments such as RAC.

The Matrix Server cluster file system component is both general-purpose and optimized for Oracle. It provided the following advantages in the FDC analysis:

- Improved management of applications and a shared Oracle Home.
- Simple, contained database movement between applications (for example, transportable tablespace from OLTP to DSS without accessing the network).
- Large database loads with External Tables and Parallel Query.
- Dynamic addition or repurposing of servers to improve throughput and response time for specific workloads.

The "shared Oracle Home" functionality is one of the keys to the Flexible Database Cluster architecture. Matrix Server's cluster file system component supports setting up a single directory for Oracle Home. Oracle needs to be installed only once—on the Matrix Server cluster file system as a single-node install. The single directory Oracle Home is then "converted" to a shared Oracle Home through methodology documented on Oracle's MetaLink Web site, allowing all executables to be stored in one place and all nodes in the cluster to use the same executables. Also, configuration files are located in the Matrix Server cluster file system and can be edited from any node in the cluster.

In addition, of course, Matrix Server can provide "shared home" for applications and middleware, other than Oracle9*i* RAC, running alongside Oracle (or in separate clusters), and can also provide high availability for all of those applications and middleware services.

With shared installation of Oracle9*i* RAC and applications, adding replacement nodes is greatly simplified. The only software that needs to be installed on the added or replaced node is the operating system and the easily installed Matrix Server RPM. Once a node has access to the SAN, it can join the cluster within five minutes. If it were necessary to install Oracle on the private drives of the replacement or added node, the time to join the Oracle9*i* RAC cluster would be in the range of 45 to 60 minutes.

The shared Oracle Home and Rapid Patch Methodology capabilities provided by Matrix Server greatly simplify deployment and management of RAC clusters for single or multiple Oracle workloads, and reduce storage requirements as well.

Matrix Server also provides these additional benefits in an Oracle9i RAC environment:

- With Matrix Server, all Oracle files can be stored in the file system. This includes, but is not limited to, the Oracle Cluster Management quorum disk, srvconfig file, control files, online and archived redo logs, datafiles, imp/exp files, SQL*Loader source files and External Tables.

- Matrix Server provides cluster-wide uniform device naming, which reduces "device slippage" and related problems. Device slippage complicates cluster administration and, if not handled carefully, can threaten to corrupt data.
- Matrix Server enables the Oracle Managed Files feature in an Oracle9*i* RAC environment. With Oracle Managed Files, the database itself creates and extends tablespaces dynamically, as needed, simplifying database administration.
- Matrix Server enables database tablespaces to be stored in standard file system files, and supports access by standard backup tools and utilities. This permits standard third-party backup tools to be used for Oracle database tables.
- Matrix Server enables Oracle's External Table feature to allow all cluster nodes to access data stored in flat files.
- Matrix Server permits Extract/Transform/Load (ETL) processes to run in parallel across all nodes in the cluster.
- Matrix Server extends Oracle's availability capabilities by providing system-wide wellness and failover for applications, middleware, servers, networking and file systems.
- Matrix Server also improves availability by supporting integrated multi-path I/O for multiple Fibre Channel connections from servers to the SAN and multiple switches within the SAN. In such a configuration, all cluster nodes can continue to operate even in the presence of multiple cable, Host Bus Adapter (HBA), or switch failures.
- Matrix Server integrates with fabric access control mechanisms to ensure that only correctly functioning cluster members can access shared data.

For more information on the PolyServe Matrix Server value proposition for Oracle9*i* RAC visit:
http://www.polyserve.com/ibm/ibm_oracle.html or
http://www.polyserve.com/products_literature.html

Matrix Server Oracle Disk Manager

PolyServe Matrix Server also provides an implementation of the Oracle Disk Manager (ODM) interface. The Matrix Server ODM implementation offers improved datafile integrity through cluster-wide file keys for access and enables Oracle9*i* with asynchronous I/O on the direct I/O mounted file systems where it stores datafiles and other database files such as redo logs. The monitoring capability of Matrix Server ODM is a major benefit in an FDC architecture deployment.

The I/O statistics package of Matrix Server ODM provides I/O performance information at a cluster-wide level (all databases in aggregate), database global level, instance, or node level. Because Matrix Server ODM understands Oracle file, process, and I/O types, it offers specialized reporting that focuses on key Oracle "subsystems" such as the Parallel Query Option (PQO), Log Writer, and Database Writer.

Oracle9*i* RAC

Oracle9*i* RAC can be used in a large, flexible cluster, or in the consolidation of multiple Oracle workloads (or clusters) to a single cluster. Oracle9I RAC capabilities include:

- Availability— Oracle9*i* RAC is fault-resilient and allows nodes to join an application in the event of a down server.
- Scalability—Applications scale well due in part to Oracle's Cache Fusion technology.
- Flexibility—Multiple Oracle database applications can share a SAN from within a single cluster, reducing administrative overhead, and nodes can be reprovisioned from one application to another.

Proof of Concept

Figure 1 shows the components used for the Flexible Database Cluster test system. The entire environment is configured and deployed in one standard 42U rack.

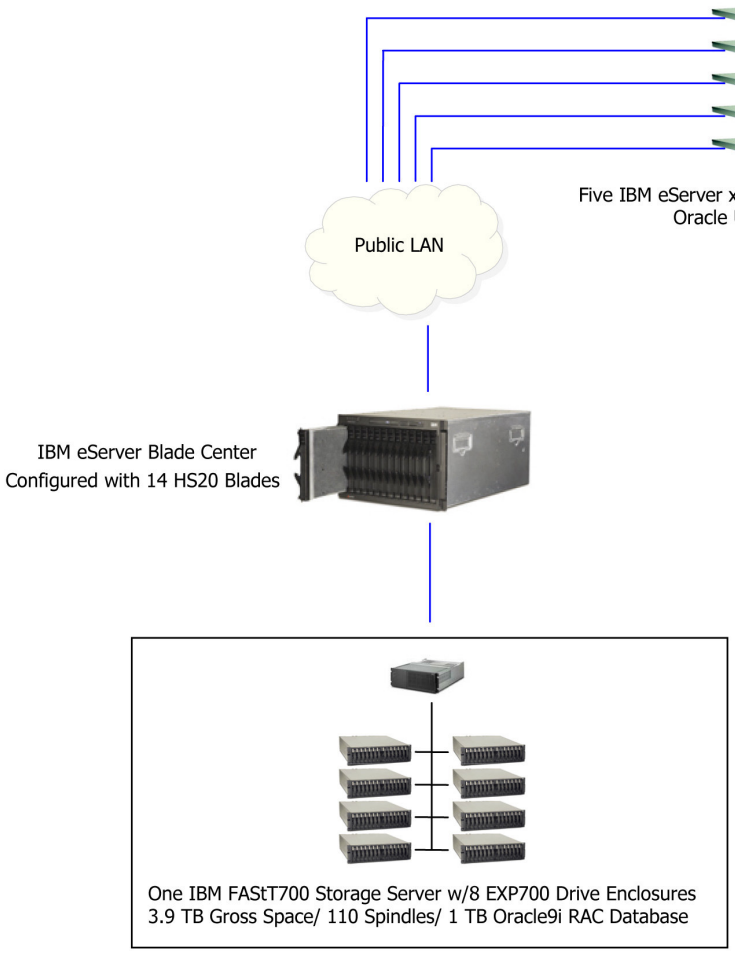

Figure 1. Components of the Flexible Database Cluster System

System Overview

The IBM eServer BladeCenter chassis was configured as follows:

- **Server nodes:** 14 IBM eServer BladeCenter HS20 blade servers. These servers are two-way SMP-capable Xeon™ processor-based and are highly scalable. The architecture of the servers contributed to the high availability of the FDC environment, and, when combined with the multi-path I/O feature provided by Matrix Server, made it possible to build a SAN subsystem with few (if any) single points of failure.
- **BladeCenter I/O modules:** The 4-port Gigabit Ethernet Switch Module and 2-port Fibre Channel Switch Module provided standard Gigabit Ethernet connectivity and connection to a fault-resilient Fibre Channel SAN.

- **Storage:** IBM TotalStorage FAStT700 Storage Server with eight EXP700 expansion enclosures held a total of one hundred ten (110) 36.4GB HDDs. The IBM FAStT Storage Manager software was used to configure the arrays and drives.

The BladeCenter includes a built-in Web-based GUI that can be used for configuration and management tasks and for viewing system status. It also allows remote access to the BladeCenter to remotely power on and power off blades and to manage I/O modules.

Database Overview

To test the Flexible Database Cluster architecture, three databases (i.e., OLTP, DSS, and DEV) were created in the Matrix Server cluster file system using Oracle9*i* Release 2 version 9.2.0.4. The particular workloads chosen for the FDC measurements were not as important as the fact that there were three of them. The goal was to have a realistic mix of processing running on the system while assessing the manageability of the FDC architecture.

OLTP Database (PROD): On-Line Transaction Processing

The OLTP database schema was based on an order entry system and contained Customers, Orders, Line Items, Product, Warehouse, and Credit Card application tables. The total database size was approximately 1.1 TB.

The application workload accessing the PROD database connected 200 users per node. The nodes under test were evenly loaded. Each user cycled through a set of transactions. At the end of each transaction, the client process slept for a short period of time randomly determined to simulate human interaction.

DSS Database: Decision Support

The DSS database performed analytical queries about customer credit. The fact table used for this decision support was the Credit Card activity table from the OLTP schema. The Card tablespace was set up as a transportable tablespace and accessed directly—without copy—by the DSS database.

This configuration shows the power of a large cluster on a SAN. It becomes very efficient to take data from one database to another without copying across a network.

DEV Database: Development

The DEV database was a simple insert engine designed to test scalability while inserting records 2 KB in size. The database was approximately 10 GB. It had only two threads defined; therefore, only two instances could access this database at one time.

The DEV workload was a zero think time program that inserted 2K rows via pipe to SQL*Loader. The streams of loader processes executed on up to two nodes when DEV was being tested along with the other workloads.

Analysis

The goal of the FDC proof-of-concept was to validate the FDC architecture and to ascertain value-add in key areas such as high availability and “on-demand” scalability. These are the key points learned from the testing:

- The IBM BladeCenter architecture and technology provide a high-availability platform for the Flexible Database Cluster.
- Capacity can be increased dynamically and transparently, without user interruption, to reduce workload completion time.

- Scalability is directly related to I/O throughput. With the FAStT SAN architecture, adding disk drives to the array may resolve performance bottlenecks.

High Availability and Manageability

Using a Flexible Database Cluster with the IBM eServer BladeCenter and PolyServe Matrix Server adds both architectural and operational value to the capability of Oracle9i RAC for high availability. This test was designed to measure the added availability provided by the FDC architecture.

An OLTP workload was running on the first 10 nodes of the BladeCenter, while the remaining four nodes were running a light DSS workload. A node was powered off, and all users connected to the nine remaining nodes maintained their connection to Oracle. In only 71 seconds after the node was powered off, a replacement node was redeployed, running with an instance of Oracle and accepting connections from users.

Figure 2 is a timeline of the events that occurred during the test.

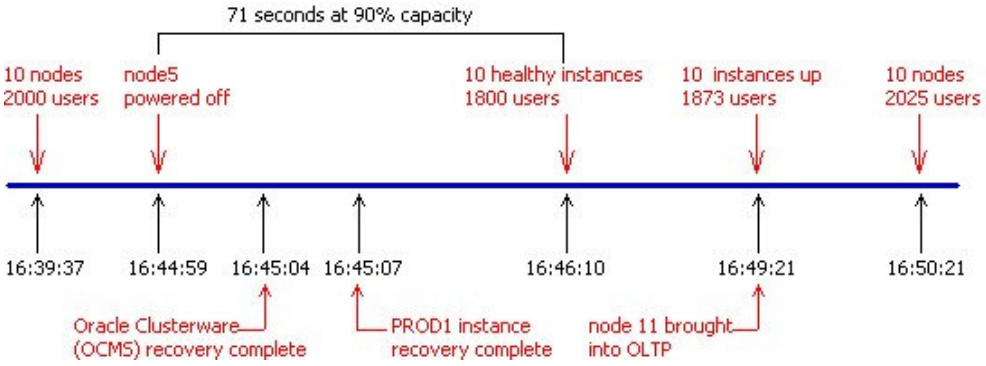

Figure 2. Timeline of events during test

The test proved the ability of Oracle9i RAC to handle a crash of one of the 10 nodes, as well as its ability to reprovision a DSS node to take the place of the crashed node. These are highlights of the test:

- Application reconfiguration was unnecessary.
- The crash and replacement of a node were completed transparent to users.
- All operations were completely dynamic.

Scalability in an OLTP Environment

This test confirmed that the FDC architecture provides a flexible pool of server resources that can be added to a workload without interruption. In the test, 100 pseudo-users executed the Pro*C benchmark code while connected via SQL*Net to dedicated servers. When the workload completed on one node, an instance of the PROD database started on the next node, and the workload was run there. The measurements were replicated across all 14 nodes.

When the results were examined for this measurement, scalability was determined to be directly related to I/O throughput, as Figure 3 shows.

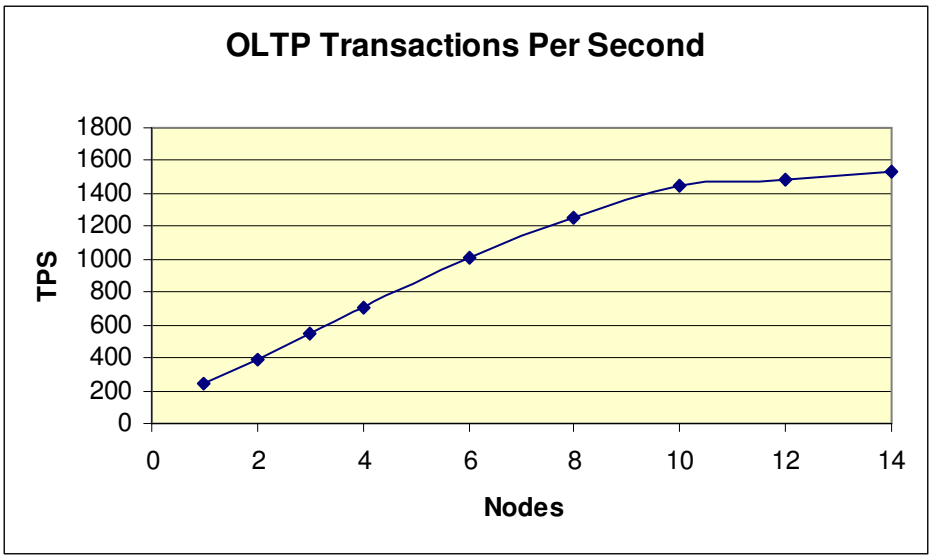

Figure 3. Scalability in an OLTP Environment

Oracle9*i* RAC on PolyServe Matrix Server experienced no scalability problems. With the tremendous bandwidth available in the BladeCenter nodes, the peak cluster-wide CPU utilization never exceeded 82%. With Oracle9*i* RAC on BladeCenter, the challenge for building balanced, scalable systems running this workload may not be a processor- or memory-level concern, but may instead be a storage subsystem issue. Unlike other OLTP system architectures, where the bottleneck was usually at the system level, resolving this performance issue is straightforward. With the FAStT SAN architecture, adding disk drives to this array is simple and would have relieved the bottleneck under this workload.

Scale on Demand, Dynamically and Transparently

This DSS-style testing confirmed that Oracle9*i* RAC scales on demand when incorporated into the FDC architecture. The testing essentially consisted of adding nodes to speed up a DSS query or a large administrative task such as index creation. The DSS database was never shut down. Instead, servers were simply allocated and the test was executed again.

Three main tests were executed at various node-count levels. As an example, in DSS test 1, the query tested the ability of the FDC architecture to scan and sort a large amount of data. The query consisted of a select count(distinct(cardnum)) from the card table. The table had roughly 1.6 billion rows, which required a great deal of sorting and merging to eliminate duplicates. The sorting could not be performed entirely in memory.

Figure 4 shows the query completion times for the test. On the first node, the query completion time was 101 minutes. Without interruption, another instance of Oracle was added and the completion time improved with 100% linear scalability. Additional Oracle instances were booted and the test was rerun in succession for four, six and eight nodes. The completion time for eight nodes was 15.9 minutes—79% of linear scale.

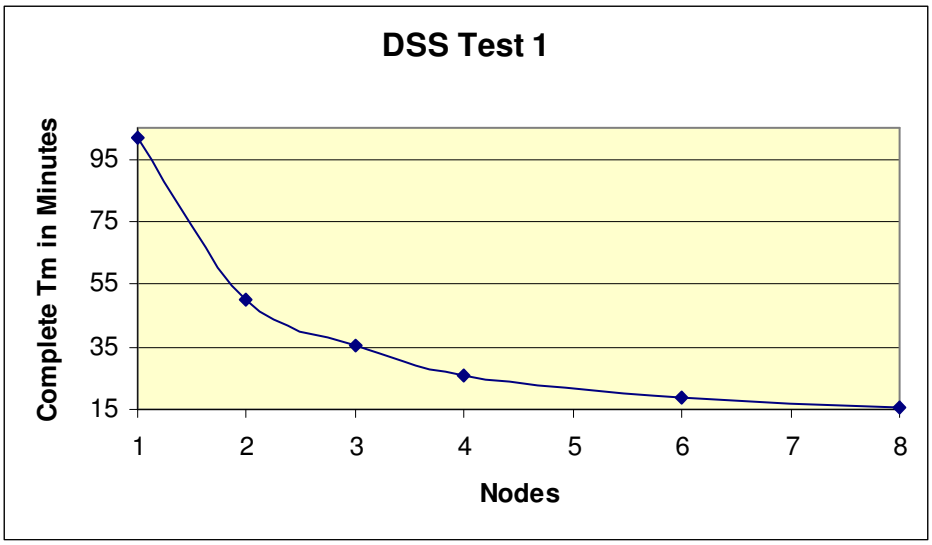

Figure 4. Query Completion Times as Nodes Are Added

Following are highlights from the DSS tests:

- Adding nodes to speed up the workload was a non-intrusive effort. An instance of the DSS database was started on another blade, the task was executed again, and completion times improved.
- Oracle9i RAC and PolyServe Matrix Server take full advantage of all available disk subsystem bandwidth and do not exhibit scalability limits at the software level.
- The scalability attributes were the same although the CPU and memory requirements of each test were dramatically different. If there had been a server bottleneck, substantial performance variance would have been apparent.

The FDC architecture is ideal to accommodate growth. If more server bandwidth is needed, simply add another server. If more disk subsystem bandwidth is needed, simply add it to the SAN.

Summary

The synergy of IBM eServer BladeCenter, PolyServe Matrix Server, and Oracle9*i* RAC makes the Flexible Database Cluster a powerful platform for supporting multiple applications. The FDC analysis presented in this paper validates the FDC architecture and technology and confirms that:

- PolyServe Matrix Server and Oracle9*i* RAC are fully supported on the BladeCenter.
- The BladeCenter architecture and technology provide an unparalleled high-availability platform for implementing Flexible Database Clusters.
- IBM, PolyServe and Oracle are leaders in the development of technology for scale-out computing.
- The architecture and technology of the Flexible Database Cluster enables on-demand computing. Cluster nodes provide a pool of flexible resources for use among applications, and the availability of Oracle9*i* RAC is enhanced, because nodes can be dynamically reprovisioned using Matrix Server to cover the loss of another node.

- The Flexible Database Cluster provides strong management tools such as Matrix Server for performance and availability. A single large cluster is now easier to manage than many small clusters.
- A general-purpose cluster file system such as the one included with Matrix Server provides a single-system feel and greatly enhances manageability. A shared Oracle home used by all nodes also simplifies management. Support is available for all database operations that require a file system.
- Improved manageability, scalability, expandability, availability and asset utilization in an FDC cluster can also dramatically improve TCO.

IBM Systems and Technology Group

Department 23U

Research Triangle Park, NC 27709

Produced in the USA.

5-04

Visit *www.ibm.com/pc/safecomputing* periodically for the latest information on safe and effective computing. Warranty Information: For a copy of applicable product warranties, write to: Warranty Information, P.O. Box 12195, RTP, NC 27709, Attn: Dept. JDJA/B203. IBM makes no representation or warranty regarding third-party products or services including those designated as ServerProven or ClusterProven.

warranties in certain transactions; therefore, this statement may not apply to you.

This publication may contain links to third party sites that are not under the control of or maintained by IBM. Access to any such third party site is at the user's own risk and IBM is not responsible for the accuracy or reliability of any information, data, opinions, advice or statements made on these sites. IBM provides these links merely as a convenience and the inclusion of such links does not imply an endorsement.

This document is for informational purposes only and does not set forth any warranty, expressed or implied, concerning any software, software feature, or service offered or to be offered by PolyServe, Inc. PolyServe, Inc., reserves the right to make changes to this document at any time, without notice, and assumes no responsibility for its use. This informational document describes features that may not be currently available. Contact PolyServe corporate headquarters for information on feature and product availability. The PolyServe Matrix Server product uses software developed by Spread Concepts LLC for use in the Spread toolkit. For more information about Spread, see http://www.spread.org.