

August 2003

DB2. Information Management Software



Oracle Experts say “You Probably Don’t Need RAC”

*IBM Software Group
Toronto Laboratory*

Contents

1. Introduction
2. Why You Probably Don't Need RAC
3. About the Author, Mogens Norgaard
4. Conclusion
5. References

Introduction

In the latest edition of the International Oracle User Group's Select Journal (Volume 10 Number 3, 3Q2003) there is an article by a leading Oracle Expert entitled "You Probably Don't Need RAC"²¹. The article is written by someone not trying to sell you something but rather by someone trying to help readers understand the true facts regarding availability, complexity, manageability and the real costs of deploying an Oracle RAC cluster.

Since DB2 Universal Database (DB2 UDB) first delivered a clustered database in 1994 on the UNIX platform, IBM has honestly represented the requirements for managing and scaling an application to exploit a clustered database. Just like any performance tuning exercise, to realize the maximum possible scalability and performance in a clustered environment, it is recommended that you design the database and your application to optimize the cluster architecture. In 2001, IBM presented a session entitled "DB2 EEE in an OLTP Environment" at the IDUG North American conference describing the design consideration to achieve near linear scaling for transactional systems with DB2 clusters. At that same conference, a presentation by Merrill Lynch described their own successes scaling their transactional system to a 4 node DB2 EEE system. Later that same year, Oracle launched 9i Real Application Clusters (RAC) with the claim that you do not have to do anything to make your database and application scale with 9i RAC.¹¹ There is a difference between an application working without changes in a cluster and actually scaling without changes. Several DB2 customers have stated that managing a DB2 cluster is transparent to the application but that it requires an understanding of a "shared nothing" database architecture by both application developers and DBAs in order to fully exploit the power of a clustered solution. This has been the consistent message from IBM yet Oracle's RAC marketing message contradicts this logical position that you design your database and application to take advantage of clustering technologies. In fact, the Oracle 9i R1 Real Application Clusters Deployment and Performance Guide devoted an entire chapter to describe exactly how to partition your database and application to get the best scalability out of a database cluster. It stated:

- "To use Real Application Clusters to improve overall database throughput, conduct a detailed analysis of your database design and your application's workload."¹
- "A primary characteristic of high performance Real Application Clusters systems is that they minimize the computing resources used for Cache Fusion processing."¹
- "To reduce Real Application Clusters overhead, each instance in a cluster should ideally perform most DML operations against a set of database tables that is not frequently modified by other instances."²

Highlights

- "It is impossible to completely eliminate Real Application Clusters overhead."³

It even goes on to state "If your application attracts more users than you expected, then you may need to add more instances [nodes].

Adding a new instance can also require that you repartition your application."⁴

This was however counter to Oracle's marketing claims and in 9iR2 Oracle removed the information on how to build a scalable application on a database cluster from the technical documentation. Oracle has claimed that the information was no longer valid and yet this same information has resurfaced in technical presentations and papers by third parties. This paper will discuss the issues that must be addressed when you consider a clustered database (namely scalability, manageability, availability and cost) citing Oracle experts who contradict Oracle's own marketing claims and state "You Probably Don't Need RAC"⁵ and that RAC "can still be overwhelmed by the wrong design"⁶.

Why You Probably Don't Need RAC

Much to Oracle's chagrin, an article has been published in the International Oracle Users Group Select Journal (3rd Qtr. 2003) entitled "You Probably Don't Need RAC", authored by Mogens Norgaard, a renowned Oracle expert and member of an elite group of 40 Oracle specialists (known as the Oak Table Network). The conclusion of the article states, "**Most likely, you probably don't need RAC. Alternatives will usually be cheaper, easier to manage and quite sufficient.**"⁷

This Oracle expert's statements appears to counter Oracle's marketing messages which claim that RAC "runs **all** your database applications - at a lower cost."⁸

The Select Journal article discusses four key areas to consider when evaluating a clustered database solution over an SMP solution: Price, Availability, Scalability, and Manageability.

Price

Oracle claims that you can lower your costs by purchasing small Intel-based clusters running Linux in comparison to running your database on a larger SMP. They even went so far as to equate the following hardware configurations at OracleWorld 2001.

"Most likely, you probably don't need RAC. Alternatives will usually be cheaper, easier to manage and quite sufficient."⁷

Highlights

Clusters Change IT Economics

Cluster Configuration	Cost
2 x Sun E10000 (64 CPU x 450MHz)	\$5,000,000
32 x Sun E420R (4 CPU x 450MHz)	\$1,000,000
32 x Compaq DL580 (4 CPUs @700MHz, 4Gb)	\$750,000

Source : Sun, Compaq Web Site, All Prices Approximate

OracleWorld 2001 – Larry Ellison Keynote

ORACLE

DB2 UDB is certified to run on 1000 servers for a single database.

Of course, there are several flaws in this comparison:

1. Oracle RAC has yet to be certified on 32 Compaq DL580s in a single RAC cluster, thus the comparison is invalid. Note that DB2 UDB is certified to run on 1000 servers for a single database since 1995.
2. The performance characteristics of a cluster of 4way Intel servers does not match the performance of a large SMPs of equal numbers of processors¹⁸
3. The above comparison does not include the price of the Oracle software

In the Select Journal article, Mr. Norgaard makes the following assertions:

“Consider Larry’s vision of cheap Intel-based Linux clusters. For instance, let’s buy those two cheap, 4-cpu Intel boxes and put them together in a cluster with Oracle9i and RAC on top:

- *Price for the hardware: About US \$15,000*
- *Price for the OS (Linux): About US \$50*
- *Price for Oracle w/ RAC: US \$480,000,*

To sum up, that adds up to \$500,000. Moreover, it’s one dollar given to the box movers for every \$32 Oracle receives.”⁹

Mr. Norgaard then goes on to add

“There are other indirect costs associated with implementing RAC”

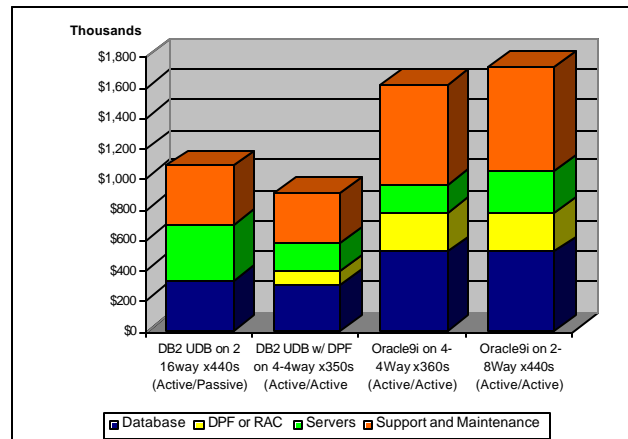
“There are other indirect costs associated with implementing RAC. First, your personnel must be more skilled, with respect to RAC and clusters. Second, you’ll have to consider the availability of a development environment (and possibly a test environment) that consists of both a cluster and RAC”⁹

Clearly looking solely at the hardware costs is not sufficient when deciding to implement a clustered database or an active/passive SMP configuration. Many customers are reluctant to accept an active/passive database configuration as there is a “stigma” around purchasing a machine that will sit idle until a failover occurs. If price

Highlights

when you include software costs, DB2 has a clear price advantage.

is your main concern, can you really justify an active/active clustered RAC solution if the price for doing so adds an additional 60% to the price of the active/passive cluster or is 90% more than a DB2 active/active cluster? Here is a price comparison of DB2 UDB ESE running in an active/passive cluster (2 16ways) or DB2 UDB ESE with DPF running in an active/active cluster (4 4ways) vs. Oracle RAC in an active/active cluster (4 4ways and 2 8ways). All systems have 16 “active” processors. This demonstrates that, when you include software costs, DB2 has a clear price advantage.²⁰



Availability

Mr. Norgaard rightly points out that if you have a standalone UNIX box, you can expect to have around 99.9% availability¹⁰. However, if you have a two node active/active cluster your availability drops to 98% as the likelihood of either of the two nodes failing has now doubled. As well, the article points out that in fact 70% of downtime is caused by human errors and insufficient knowledge.

“...the increased complexity (extra layers of code, extra hardware, etc.) introduced with clusters and RAC are the two main causes of the additional downtime.”

Mr. Norgaard states “the increased complexity (extra layers of code, extra hardware, etc.) introduced with clusters and RAC are the two main causes of the additional downtime. This coupled with the fact that it just takes longer to boot a cluster, start-up RAC, and perform some other management tasks impacts availability.”¹⁰

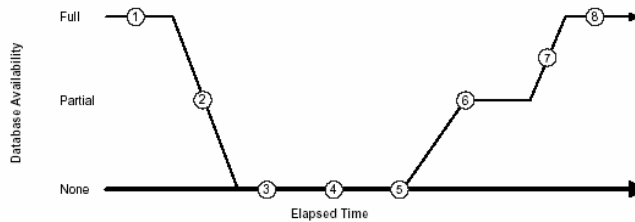
The article goes on to describe a client system that Mr. Norgaard recently worked with that was down for nearly **five hours** because they needed to upgrade their RAC system from 9.2.0.1 to 9.2.0.3 (a simple service pack installation). This is in stark contrast to the installation steps involved in applying a DB2 fix pack. With DB2 the fix pack is installed online. It is then activated by running db2instupdt with the database offline, which takes approximately a minute. This real world customer example from an Oracle expert illustrates that availability encompasses more than just marketing messages.

The reality is that a single node failure in a RAC cluster will freeze the entire database while the global cache service resources are remastered.

It should also be noted that the failover process in a RAC cluster is not instantaneous. The reality is that a single node failure in a RAC cluster will freeze the entire database while the global cache service resources are remastered. Here is how the Oracle 9i RAC Administration Guide describes the database availability during an instance failure.¹² (Note that between points 3 and 5 there is no database availability for *any* user request on *any* node.)

Highlights

Figure 7-1 Steps of Oracle Instance Recovery



The steps in recovery are:

1. Real Application Clusters is running on multiple nodes.
2. Node failure is detected.
3. The **Global Cache Service (GCS)** is reconfigured; resource management is redistributed onto the surviving nodes.
4. SMON reads the redo log of the failed instance to identify the database blocks that it needs to recover.
5. SMON issues requests to obtain all of the database blocks it needs for recovery. After the requests are complete, all other blocks are accessible.

Note: The **Global Cache Service Processes (LMSn)** only re-master resources that lose their masters.

6. Oracle performs roll forward recovery; redo logs of the failed threads are applied to the database.
7. Oracle performs roll back recovery; undo blocks are applied to the database for all uncommitted transactions.
8. Instance recovery is complete and all data is accessible.

Add this to the time it takes for Oracle RAC to detect a failure (time between points 1 and 2 above), which by default is set to 20 seconds (minimum detection time is 7 seconds)¹³ and it is apparent that a node failure would impact the system for tens of seconds. Oracle's latest collateral confirms this and states "RAC also provides ... hot failover in less than 30 seconds."¹⁴ This is likely acceptable to almost all customers but the question remains: Why should I pay the extra costs to purchase and manage a clustered database when I can achieve this same level of failover using a simpler active/passive cluster which costs less? DB2 UDB V8.1 has demonstrated 9-second failover with a Telco customer that was running 3500+ transactions per second (see <http://www.ibm.com/software/data/highlights/availability.html> for more details). If you can fail over the entire database and workload in 9 seconds, why would you implement a RAC solution that requires a minimum of 7 seconds before it even detects the failure and begins the failover process? Note that different applications will have differing failover characteristics.

DB2 UDB V8.1 has demonstrated 9-second failover

[an active/passive cluster] allows for 100% capacity even after a failover has occurred.

It is important to note that the 2 node active/passive failover cluster is the traditional model for HA systems, and is by far more proven over time with more implementations than any other approach. This is the HA model where the vast majority of features and practices, involving both hardware and software, have been focused in the distributed world. It also allows for 100% capacity even after a failover has occurred.

The active/passive terminology is used here to differentiate from an active/active cluster (such as DB2 with DPF or RAC) where both nodes are working on the *same* application. But active/passive is a bit of a misnomer. While the secondary (or failover) node is not sharing in the primary workload, it need not be completely idle. With

Highlights

proper planning and procedures, the failover node can be utilized to run a different application. Typically this is a non-critical development workload that can be shut down in the event of a failover.

Alternatively, the second server can host its own critical production application and failover to the first box. This mutual takeover configuration achieves the same resource utilization as active/active clusters, while trading-off capacity of both applications when sharing a single server in failover mode. That is, an active/active cluster or a mutual takeover cluster will not achieve 100% application performance as there is now more work running on the surviving node(s). Of course the servers could be sized for spare capacity to better accommodate these limited periods.

As an example, Deutsche Post relies upon two xSeries 330 servers in a Linux cluster with an active/passive cluster using DB2 Universal Database Enterprise Edition. If one server fails, the high availability cluster switches over to the standby server within ten seconds--making the interruption of service almost imperceptible for end users.¹⁹

Failover in a DB2 UDB ESE database using the Database Partitioning Feature (DPF) behaves the same as an active/passive cluster with the exception that the takeover node may or may not be an active participant in the cluster. With DPF, you can configure the system such that one node can be the standby for all of the other nodes in the cluster or each active node in the cluster can assume the work of a failed node.

There are many options, all involving their own unique set of trade-offs, to achieve an optimal balance of scalability, availability, and resource utilization within a clustered configuration. It is overly simplistic to suggest that any single architecture or configuration solves all of these competing needs ideally for all applications.

Scalability

The Oracle 9i R1 Real Application Clusters Deployment and Performance Guide states the following:¹⁵

“To reduce Real Application Clusters overhead, each instance in a cluster should ideally perform most DML operations against a set of database tables that is not frequently modified by other instances.”²

“In general, however, you can apply several strategies to partition application workloads. These strategies are not necessarily mutually exclusive and are discussed in the following sections:

- Functional Partitioning
- Separating E-Commerce and Data Warehousing Processing
- Departmental and User Partitioning
- Physical Table Partitioning
- Transaction Partitioning”²

“If you have properly partitioned your application for Real Application Clusters, then as the size of your database increases you can maintain the same partitioning strategy and simultaneously achieve optimal performance. The partitioning method to use when adding new functionality depends on the types of data the new functions

“To reduce Real Application Clusters overhead, each instance in a cluster should ideally perform most DML operations against a set of database tables that is not frequently modified by other instances”

Highlights

Adding a new instance can also require that you repartition your application

access. If the functions access disjoint data, then your existing partitioning scheme should be adequate. If the new functions access the same data as the existing functions, then you may need to change your partitioning strategy. If your application attracts more users than you expected, then you may need to add more instances. Adding a new instance can also require that you repartition your application.”⁴ This is counter to Oracle claims that “Oracle9i Real Application Clusters (RAC) is the only clustered database available that can transparently scale and protect packaged applications with literally no changes necessary for the application or the organization of its associated data.”¹¹

Clearly, *transparent* is a relative term. Any system architecture benefits when the hosted application has been designed or tuned to optimize particular characteristics of the platform. This is something DBAs intuitively understand, and it applies equally to RAC as to every other RDBMS architectures. Optimizing the application/database design to the platform invariably helps performance and scalability while trading-off some of the benefits of total application transparency. The laws of physics don’t change (locality of reference will always boost database performance) but sometimes creative marketing can make it seem that way.

Oracle removed this information from their technical manuals

The Smoking Gun?

In 9iR2, Oracle removed this information from their technical manuals. Since RAC did not change significantly between 9iR1 and R2, one could assert that the R1 manuals were simply inaccurate and that Oracle forgot to take the information out of the old OPS manuals when they moved to 9i. However, looking at the 9iR1 Real Application Clusters Deployment and Performance Guide one can see that it differs significantly from the Oracle 8i Parallel Server Administration, Deployment and Performance manual. Here is the chapter overview from the 9iR1 manual¹⁵

3

Scaling Applications for Real Application Clusters

This chapter describes methods for scaling applications for deployment in [Oracle Real Application Clusters](#) environments. This chapter provides a methodical approach to application design as well as procedures for resolving application performance issues. Topics in this chapter include:

- [Overview of Development Techniques in Real Application Clusters](#)
- [SQL Statement Execution in Real Application Clusters](#)
- [Workload Distribution Concepts in Real Application Clusters](#)
- [Workload Characterization in Real Application Clusters](#)
- [Scaling-Up and Partitioning in Real Application Clusters](#)

Here is the similar chapter overview from the Oracle 8i OPS manual which shows a significant difference between the two documents.¹⁶

Application Analysis and Partitioning

This chapter explains application design optimization techniques for Oracle Parallel Server. It includes the following sections:

- [Overview of Development Techniques](#)
- [Application Transactions and Table Access Patterns](#)
- [Selecting A Partitioning Method](#)
- [Application Partitioning Techniques](#)
- [Departmental and User Partitioning](#)
- [Departmental and User Partitioning](#)
- [Physical Table Partitioning](#)
- [Transaction Partitioning](#)
- [Scaling Up and Partitioning](#)
- [Adding Instances](#)
- [Design-Related Batch Processing Issues](#)

“...you can still get into situations where traditional OPS workarounds are needed”

“It was allegedly, deliberately removed from Oracle documentation because of the message it sent to customers.”

“RAC still requires more skills and more time to manage. As discussed earlier, added complexity means additional skills and additional time”

The original 9iR1 documentation is more inline with Mr. Norgaards article which states “you can still get into situations where traditional OPS workarounds are needed (data partitioning, etc.) in order to achieve maximum performance. Even the wonderfully complex and mythic GC_FILES_TO_LOCKS parameter can be useful at times. **It was allegedly, deliberately removed from Oracle documentation because of the message it sent to customers.**”¹⁰

Similarly, a presentation by Jonathan Lewis entitled “Wrecking RAC Will “anything” run on it?” concluded “RAC is much better than OPS - but it can still be overwhelmed by the wrong design.”⁶

Manageability

Managing a single SMP or an active/passive cluster has become a skill that virtually all DBAs today have acquired. However, managing, tuning and troubleshooting an active/active cluster is far more complex with Oracle RAC and it is a skill that is much rarer according to Mr. Norgaard as he states “RAC still requires more skills and more time to manage. As discussed earlier, added complexity means additional skills and additional time.”⁷ Although Oracle has claimed to have made RAC easier to manage, Norgaard believes that “when RAC is pushed to the limit, you could still need to do the same things that were required with OPS”.⁷

And in terms of troubleshooting a problem on a cluster:

“[when a cluster freezes] the resulting trace/log file from the cluster will normally be the size of Texas and only one or two people in the entire vendor organization can truly understand them, or at least this is what you’ll likely be told. Next the files (often with sizes measured in GB) are shipped to the vendor and some months later they will report back that it wasn’t possible to pinpoint the exact reason for the

Highlights

complete cluster freeze or crash, but that this parameter was probably a bit low and this parameter was probably a bit high. That's what always happens. I have never met a vendor who could correctly diagnose and explain a hanging cluster or a cluster that kept crashing".⁷

About the Author, Mogens Norgaard

Mogens Norgaard is the co-founder and technical director of Miracle A/S, providing consulting, support and training on Oracle and SQL Server. He is also the originator of the Oak Table Network, which is an invitation only group of 40 highly respected Oracle practitioners. Mogens also worked for Oracle in their support organization for 10 years and is considered an expert on the Oracle RDBMS by the Oracle user community.¹⁷

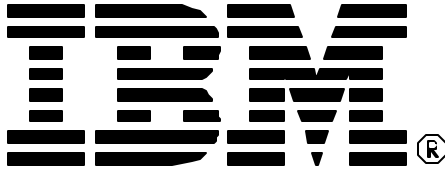
Conclusions

IBM is focused on understanding your business needs and delivering technology to enable your business to grow. We recognize and advise customers on the trade-offs involved in different approaches to database architecture, and the optimizations that are available to tailor any application to maximize scalability and performance on a given platform. We don't promise one size fits all. IBM DB2 UDB is able to take advantage of the availability characteristics of both active/passive as well as active/active clusters. In addition, DB2 can scale to a 1000 nodes cluster (125X larger than Oracle's largest certified cluster). Oracle experts caution you to review your options before considering clustered solutions. They ask you to account for the price, availability, scalability and manageability requirements of the full solution. "Mostly likely, you probably don't need RAC. Alternatives will usually be cheaper, easier to manage and quite sufficient".⁷ A useful reminder that the laws of physics haven't changed. Call IBM to see what alternatives exist for you, both in the active/passive and active/active clustering environments and then make an informed decision.

"Most likely, you probably don't need RAC. Alternatives will usually be cheaper, easier to manage and quite sufficient"

References

1. Oracle 9i Real Application Clusters Deployment and Performance manual page 3-2.
2. Oracle 9i Real Application Clusters Deployment and Performance manual page 3-7.
3. Oracle 9i Real Application Clusters Deployment and Performance manual page 3-8.
4. Oracle 9i Real Application Clusters Deployment and Performance manual page 3-16.
5. International Oracle Users Group Select Journal. Volume 10, No.3 3rd Qtr. 2003 (page 39)
6. Wrecking RAC - Will "anything" run on it?" <http://www.ukoug.org/calendar/may03/Racday.htm>
7. International Oracle Users Group Select Journal. Volume 10, No.3 3rd Qtr. 2003 (page 42)
8. http://www.oracle.com/ip/index.html?rac_home.html
9. International Oracle Users Group Select Journal. Volume 10, No.3 3rd Qtr. 2003 (page 40)
10. International Oracle Users Group Select Journal. Volume 10, No.3 3rd Qtr. 2003 (page 41)
11. http://www.oracle.com/ip/dep/otn/database/oracle9i/transaction_processing.html
12. Oracle 9i Real Application Clusters Administration Guide (page 7-12)
13. Failure detection with Oracle RAC is determined by two configuration parameters (specified in the cmcfig.ora configuration file). MissCount which is the time that the Cluster Manager waits for a heartbeat from the remote node before declaring that node inactive (default value of 15seconds with a minimum value of 6 seconds). WatchdogSafetyMargin which specifies the time between when the cluster manager detects a remote node failure and when the cluster reconfiguration is started. (default value is 5 seconds with a minimum of 1 second) http://www.mamiyami.com/doc/oracle9i/a97297/appf_ocrm.htm
14. <http://www.oracle.com/ip/dep/otn/database/oracle9i/reliability2.html>
15. <http://www.engin.umich.edu/caen/wls/software/oracle/rac.901/a89870.pdf>
16. <http://servo.oit.gatech.edu/docs/oracle/816/paraserv.816/a76970.pdf>
17. http://searchdatabase.techtarget.com/originalContent/0,289142,sid13_gcj914391,00.html
18. Based on TPC-C results as of August 5, 2003, the clustered results (using the same number of CPUs) do not match the performance results of the single systems.
19. <http://www-3.ibm.com/software/success/cssdb.nsf/CS/NAVO-5CTQX4?OpenDocument&Site=dmmain>
20. Figures are based on published prices as of 06/05/2003 and are subject to change without notice. Discounts listed are for comparison purposes only. DB2 pricing Includes 15% discount provided for DB2 UDB under Passport Advantage. Oracle pricing includes 20% discount - standard discounting provided by the Oracle online store for Oracle 9i. and RAC. Hardware pricing for IBM xSeries 440 and 360 with 15% discount applied. Service is based on three years of software and server support / maintenance which is 15% of total server cost for years 2 & 3 (1 year warranty included) for hardware, 20% of total cost for DB2 UDB and 22% of total cost for Oracle 9i.
21. International Oracle Users Group Select Journal Volume 10 Number 3 © 2003



© Copyright IBM Corporation 2003
IBM Canada
8200 Warden Avenue
Markham, ON
L6G 1C7
Canada

Printed in United States of America
08-2003
All Rights Reserved.

IBM, DB2, DB2 Universal Database, OS/390, z/OS, S/390, and the ebusiness logo are trademarks of the International Business Machines Corporation in the United States, other countries or both.

UNIX and Unix-based trademarks and logos are trademarks or registered trademarks of The Open Group. Intel and Intel-based trademarks and logos are trademarks or registered trademarks of Intel Corp. Other company, product or service names may be the trademarks or service marks of others.

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:
INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

The information in this white paper is provided AS IS without warranty. Such information was obtained from publicly available sources, is current as of 08/26/2003, and is subject to change. Any performance data included in the paper was obtained in the specific operating environment and is provided as an illustration. Performance in other operating environments may vary. More specific information about the capabilities of products described should be obtained from the suppliers of those products.