



**IBM TotalStorage Enterprise  
Storage Server™ Model 800 -  
RAID 5 and RAID 10  
Configurations Running  
Oracle® Database Performance  
Comparisons**

May 2003

IBM Systems Group  
Open Storage Systems Laboratory, San Jose, CA  
[www.ibm.com](http://www.ibm.com)

## Contents

### Abstract

#### 1.0 Introduction

#### 2.0 Performance factors

##### 2.1 Number of disks

##### 2.2 Workload attributes

#### 3.0 Synthetic workload measurements

#### 4.0 Oracle 9i workload measurements

##### 4.1 Random update queries analysis

#### 5.0 Conclusions and recommendations

#### 6.0 References

#### 7.0 Acknowledgements

#### 8.0 Appendix

## Abstract

This paper discusses performance considerations of selecting RAID 5 or RAID 10 storage on an IBM TotalStorage Enterprise Storage Server™ Model 800 when using Oracle 9i Database. The paper includes results of laboratory measurements of performance using both RAID levels for a variety of SQL queries.

## 1.0 Introduction

There are many factors that may influence overall performance, such as workload characteristics (write content, cache friendliness, sequentiality, etc.), physical configuration (disk speeds and capacities), and data layout (use of logical volume manager striping).

Before reading this paper we recommend that you read a companion document titled *Configuring the IBM TotalStorage Enterprise Storage Server for Oracle OLTP Applications*, see Reference [1]. That document discusses some of the above-listed factors, and it outlines an approach to data layout called *stripe and spread*, which can be used with either RAID or RAID 10. *Stripe and spread* is a technique that uses AIX logical volume manager to spread data over several striped RAID arrays. This technique may be viewed as a variant of the S.A.M.E (Stripe and Mirror Everything) philosophy outlined in Reference [2].

## 2.0 Performance factors

### 2.1 Number of disks

The number of disks in a configuration is a substantial factor in overall cost and performance. As a result, we should consider this factor carefully when comparing performance of RAID 5 and RAID 10.

If you only had one choice in disk drive capacity, such as 36GB disks, and if you were configuring a disk system solely based on required subsystem storage capacity, then a RAID 10 configuration would require roughly twice the number of physical disks compared to RAID 5<sup>1</sup> just to meet the storage capacity requirement. Because there would be more disks over which to distribute I/O activity, that RAID 10 configuration would likely have better performance than a RAID 5 configuration, but also cost substantially more. This would follow the adage of *you get what you pay for*. It also complicates the comparison, as two variables are changing, both cost and performance.

---

<sup>1</sup> Exact ratio is  $2N/(N+1)$ , where N is the number of data disks in a RAID 5 array. For two ESS 6+P RAID 5 arrays replaced with a 3+3 and a 4+4 RAID 10 array, that ratio would be 1.71.

As a result, the comparisons in this paper focus on a simpler approach by comparing configurations with equal numbers of disks. This not only simplifies the comparison, but also makes sense in real life.

- Although a RAID 10 configuration might have required more disks, a RAID 5 configuration could have been constructed with equal number of disks, and equal cost, although the RAID 5 configuration would have offered unused capacity.
- ESS also offers choices of disk drive capacity, such as 36.4, 72.8 and 145.6 GB. This means that you might consider using a larger capacity disk drive with RAID 10, such as 72.8 GB than you might have used with RAID 5. In that case, the RAID 5 and RAID 10 configuration would have roughly equal capacity and cost.

## 2.2 Workload attributes

There are a number of workload attributes that influence the relative performance of RAID 5 and RAID 10, including the overall demand, use of cache, write content, and sequential content.

- **Application demand.**  
The differences between RAID 10 and RAID 5 do not affect the basic service time in accessing the disks, but do affect the queuing delays introduced by very busy systems. Hence, lightly loaded systems would see little difference.
- **Cache friendly workloads.**  
Workloads that make good use of subsystem cache would generally see very little difference between RAID configurations. In this case, choice of disk probably matters little.
- **Random and sequential reads.**  
For most read-oriented workloads, performance of RAID 5 and RAID 10 should be roughly equal. Data is striped across all of the disks for RAID 5, resulting in uniform disk access. In RAID 10, data is striped across the disks, and load is automatically balanced across the mirrored pairs, also resulting in uniform access.
- **Random writes.**  
Application data is written into subsystem cache. As long as there is room in the cache, the response time seen by the application is based on time to get data into the cache, and not on the characteristics of the disks. When the data is destaged from cache, RAID 10 requires 2 I/Os per write destage, and RAID 5 requires 4 I/Os per write destage. As a result, RAID 10 disks would be half as busy, resulting in twice the throughput of RAID 5 if destages from cache are limiting subsystem throughput.

In addition, RAID 10 random writes would cause less contention with other, such as random read workloads.

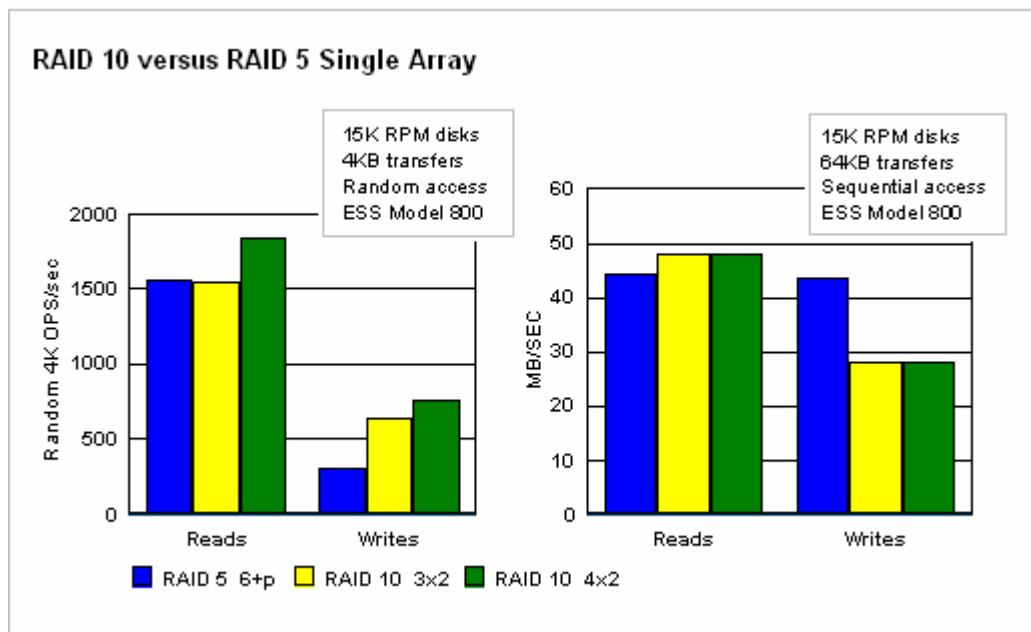
- **Sequential writes.**

Application data is written into subsystem cache. As long as there is room in the cache, the response time seen by the application is based on time to get data into the cache, and not on the characteristics of the disks. When the data is destaged from cache, RAID 10 requires 2 copies of data written to disk. In the case of RAID 5, because the cache destaging design converts sequential writes to “full stripe” writes, parity is generated once for each data stripe (e.g., once for every 6 blocks of data in a 6+P array). As a result, RAID 5 should have better throughput potential for destaging from cache, and also cause less contention with other workloads.

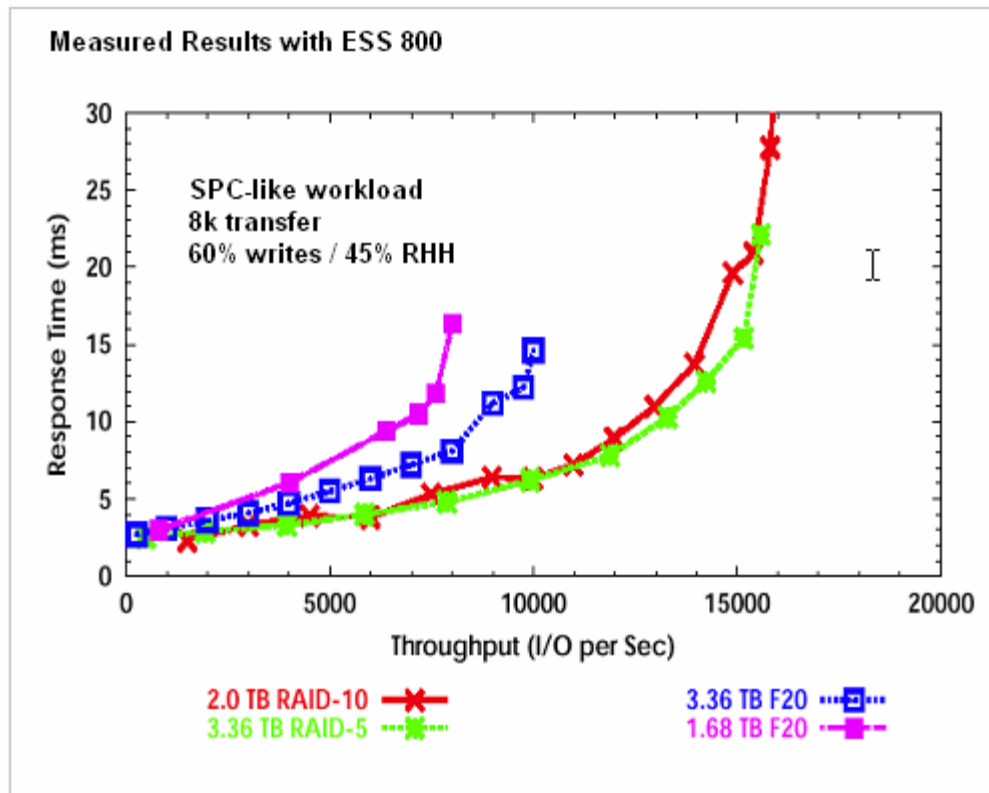
### 3.0 Synthetic workload measurements

Actual measurements on an ESS Model 800 comparing RAID 5 and RAID 10 in a synthetic workload support these general expectations, as shown in the next charts.

The following chart shows the results of measurements of maximum throughputs achieved using a synthetic workload consisting entirely reads or writes against a single RAID 5 array with 7 disks, compared to RAID 10 arrays with either 6 or 8 disks.



Another example of comparison of RAID 5 and RAID 10 configurations can be found in Reference[3]. The following chart has been extracted from that report, showing comparisons of RAID 5 and RAID 10 configurations using a synthetic workload with mixtures of random and sequential accesses. These results show very similar performance between RAID 5 and RAID 10.



As you can see from the results of these synthetic measurements:

- RAID 5 and RAID 10 perform roughly equal for workloads dominated by reads
- RAID 10 outperforms RAID 5 for workloads dominated by random writes
- RAID 5 performs better than RAID 10 for sequential writes
- RAID 5 and RAID 10 might perform about the same for mixed workloads

So what will happen to more complex workloads such as those accessing Oracle databases?

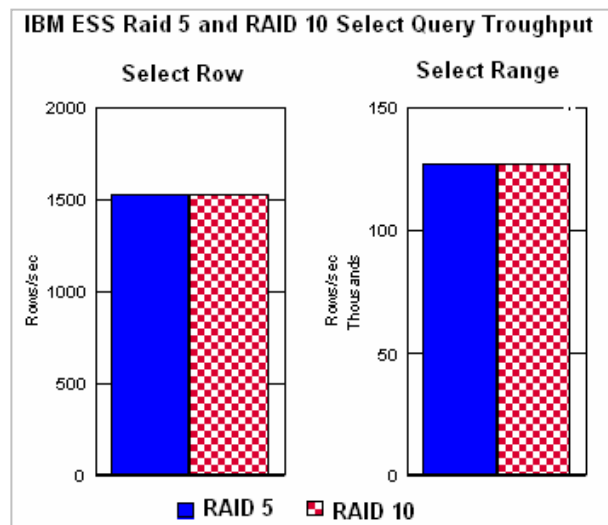
#### 4.0 Oracle 9i workload measurements

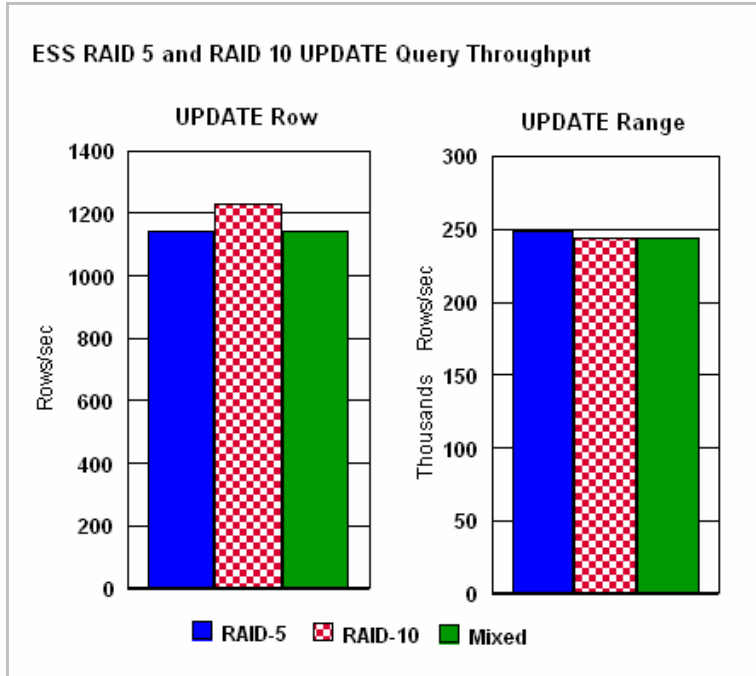
Workloads accessing Oracle databases are more complex than depicted by these synthetic workloads. Many factors can contribute to system level performance. To further understand

these factors, a set of experiments were conducted using a workload driving Oracle queries using ESS RAID 5 and RAID 10 options. The following summarizes these tests.

- A workload was generated that tested different variations of SELECT and UPDATE queries, in which individual rows or ranges of rows were selected or updated. This was done to understand the sensitivity to the amount of write activity.
- The workload generated concurrent activity through multiple users and processes accessing and updating a shared database. Sufficient processes were run to generate a reasonable level of stress on the disk subsystem. 32 processes were run for the “range” queries, and 80 processes were run for the individual row queries.
- The workload used a moderate-size configuration consisting of 7 ESS RAID arrays. The databases were spread across the six of the arrays using the AIX Logical Volume Manager, consistent with the methodology of “stripe and spread” discussed earlier in this document. In all of the cases, the control files, redo logs, and rollback segments were isolated onto one of the seven RAID arrays. Oracle resources, such as buffer size, were constrained at artificially small values to force more contention onto the disk subsystem.
- Three scenarios were tested: all RAID 5 arrays, all RAID 10 arrays, and a mixed configuration consisting of data files on RAID 5 with logs on RAID 10. Measurements were taken of transaction throughput, as well as disk I/O characteristics from IOSTAT and ESS Expert.

The following charts summarize the results of comparisons of transaction throughput, as well as information gathered from IOSTAT and from the benchmark





The following table illustrates the IOSTAT comparisons:

Query	Configuration	Rows per sec	MB per sec	IO per sec	MB read per sec	MB write per sec	Log MB per sec	Log IO per sec
UPDATE row	RAID 5	1,147	37.7	4,735	26.0	11.7	0.6	120.4
	Mixed	1,146	36.9	4,643	25.3	11.7	0.6	117.9
	RAID 10	1,231	38.6	4,855	26.5	12.2	0.6	124.7
UPDATE range	RAID 5	249,150	41.4	2,952	14.5	26.9	9.3	154.1
	Mixed	243,700	40.6	2,882	14.0	26.6	9.1	145.7
	RAID 10	244,100	40.3	2,818	13.6	26.7	9.3	148.2
SELECT row	RAID 5	1,524	35.8	4,463	33.7	2.1	0.0	7.1
	RAID 10	1,524	35.9	4,473	32.4	3.5	0.0	4.5
SELECT range	RAID 5	127,250	8.8	952	6.8	2.0	0.0	3.2
	RAID 10	127,000	9.8	1,000	6.6	3.2	0.0	1.4

The following are some general observations of the results:

- The results for the read-oriented queries (SELECT Row and SELECT Range) are nearly identical.
- UPDATE Range shows a very slight difference in throughput (2%), with a very slight advantage for the RAID 5 configuration. This is the only measurement with any significant logging rate, approximately 9MB/sec.

Although this does not approach saturation of the sequential write rate for an ESS RAID array (approx. 30-50MB/sec), it does demonstrate the viability of RAID 5 for logging. (Note: the MIXED configuration used RAID 5 for data files, and RAID 10 for logging.)

- UPDATE Rows show the greatest difference in application throughput, approximately 7% more throughput for the RAID 10 configuration). This case is further analyzed below.

#### 4.1 Analysis of Random Update Queries

The RANDOM UPDATE workload shows the greatest difference in performance between RAID 5 and RAID 10, and is worth examining a little closer. For random updates, RAID 10 shows a modest increase in transaction throughput (7%) relative to RAID 5. Is this a reasonable expectation? Why would we not see throughput improvements more like those seen when comparing random write I/O streams? (Keep in mind that this workload consists of nothing but random updates, and is therefore a relatively extreme case to consider.)

To understand this better, it is helpful to examine how Oracle is exercising the disk subsystem. The following table shows information gathered from the ESS Expert for representative 5-minute samples of the RANDOM UPDATE experiment.

	Host I/Os	I/O per second Per GB	Host reads	Host writes	Cache misses	NVS Delay Percentage	RAID rank read	Rank read delay (ms)	RAID rank destage	Rank destage delay (ms)	Estimated I/O per disk
<b>RAID 5</b>	4770	1.4	3086 (65%)	1684 (35%)	1525 (32%)	0	1993 (42%)	31	1456	72	160
<b>RAID 10</b>	4867	2.7	3072 (63%)	1795 (37%)	1611 (33%)	0	2080	25	1491	55	98

The following are some observations on this data:

- The I/O workload generated by the random updates is certainly not trivial. The I/O rates are approximately 4800 I/O per second, spread over 7 ESS arrays. Since this test used 72GB disks, this workload represents 1.4-2.7 I/O per second per gigabyte of available capacity. Some internal studies have shown the average in open systems workloads below 1 I/O per second per GB.
- Because this experiment employed a “stripe and spread” storage allocation strategy using AIX LVM, examination of the ESS Expert disk statistics showed that workload has been spread very evenly across the available RAID

arrays without requiring any detailed analysis or tuning. This approach tends to better distribute workload, and reduce the chances of a single array creating a workload bottleneck. In other words, no single array is totally saturated. This also contributes to the fact that there are virtually no “NVS Delays”, meaning that writes are not delayed for the application due to inability of the disk subsystem to destage data from cache.

- Although this workload consists of entirely random updates to Oracle databases, it does not generate entirely random writes at the disk level. Updating rows requires accesses to both the data blocks and index blocks. The disk workload consists of reads and writes, cache hits and misses, and both random and sequential activity.
- From this data we have also estimated the average number of physical disk accesses. The physical accesses for RAID 10 (98 I/O per second per disk) are certainly lower than for RAID 5 (160 I/O per second per disk), and should be able to absorb more workload. (The disks used in these experiments should be able to achieve around 200 or more I/O per second per disk.) Note that in this set of experiments, we ran with a fixed number of users and processes. In absence of other system bottlenecks, that RAID 10 configuration would likely have been able to support more users and processes.
- The “Rank Read Delay” is lower for RAID 10 (25ms vs. 31ms). This delay would likely be seen in the response time for read misses, and is likely the dominant factor in the improved throughput. Although the “Rank Destage Delay” is higher for RAID 5, this occurs asynchronously for writes, and is likely not a factor in the application throughput.
- The above observations tend to support a conclusion that this RAID 10 configuration should see some modest improvement in performance.

## 5.0 Conclusions

The following are some conclusions and recommendations regarding selection of RAID 5 and RAID 10 with IBM ESS.

- There are many factors that influence selection of RAID 5 or RAID 10, including overall cost and customer preference. The performance results suggest that both are viable options for most workloads.
- Following a data layout strategy of stripe-and-spread by making use of AIX Logical Volume Manager on top of RAID 5 or RAID 10 allows workload to be well-balanced without requiring detailed tuning and analysis.

- The performance of RAID 10 may be better for very high random write workloads. The amount of improvement will vary.

## 6.0 References

1. *Configuring the IBM Enterprise Storage Server for Oracle OLTP Applications*, (Martin), available from IBM representative
2. *Optimal Storage Configuration Made Easy* (Loiza), available from Oracle at [http://technet.oracle.com/deploy/availability/pdf/oow2000\\_same.pdf](http://technet.oracle.com/deploy/availability/pdf/oow2000_same.pdf)
3. *IBM TotalStorage Enterprise Storage Server™ Model 800 Performance* (McNutt), July 2002, available from IBM representative

## 7.0 Acknowledgements

- John Aschoff, author, Database and Performance Lead Engineer
- Neena Cherian
- Kurtis Dorsey
- Hemanth Kalluri
- Ravisankar Shanmugam
- Brian Smith.

## 8.0 Appendix

### Benchmark Program

A benchmark program developed by IBM, called the *Database Performance Analysis Workload Suite* or DPAWS for short, was used to drive the Oracle Database with a high volume of transactions. The database setup for DPAWS uses 15 tables, each with the same number of rows, spread across 12 table spaces on 12 RAW logical volumes. Each Table space is 32 GB in size, and each table has row sizes that range from 14 bytes to 805 bytes. A 1 row database takes up 4 K of space. For this environment a 140 GB database was used which meant that each table contained 31,901,312 rows.

The types of queries generated by this workload include:

- Update Row simply randomly selects 1 of the 15 tables, and randomly selects one of 31 million rows and updates that row.
- Update Range randomly selects 1 of the 15 tables, and then randomly selects one of the 31 million rows and updates from that row sequentially for 10,000 rows.
- Select Row randomly selects 1 of the 15 tables, and randomly selects one of 31 million rows and just does a SELECT SQL statement on it.
- Select Range randomly selects 1 of the 15 tables, and then randomly selects one of the 31million rows and does a SELECT SQL statement from that row sequentially for 10,000 rows.

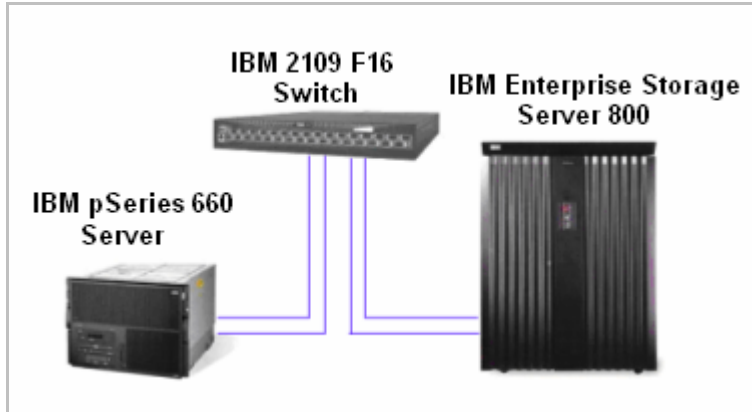
The script for running the tests ran each scenario 4 times to confirm the repeatability in the experiment and each time through it would sequentially go through the different workload types (Update Row -> Update Range -> Select Row -> Select Range -> Update Row -> ... -> Select Range). Since the DPAWS workload reaches a steady state in a matter of minutes, each workload type was run for an hour.

### Hardware Layout

For this experiment we chose 3 different storage layouts on the ESS 800 and drove the ESS with an IBM eServer® pSeries® server. The hardware configuration was as follows:

- pSeries 660 Model 6H1 with 4 GB memory
- ESS 800 with 16 GB cache
- 2 Fibre Channel 2 Gigabit cards
- 1 IBM TotalStorage switch 2109 – F16

A pSeries with 4 GB of memory was attached via 2, 2Gbit Fibre channel connections through a switch attached to the ESS. Using the Subsystem Device Driver (SDD) for load balancing across the fibre, it translated to 4 paths per ESS LUN.



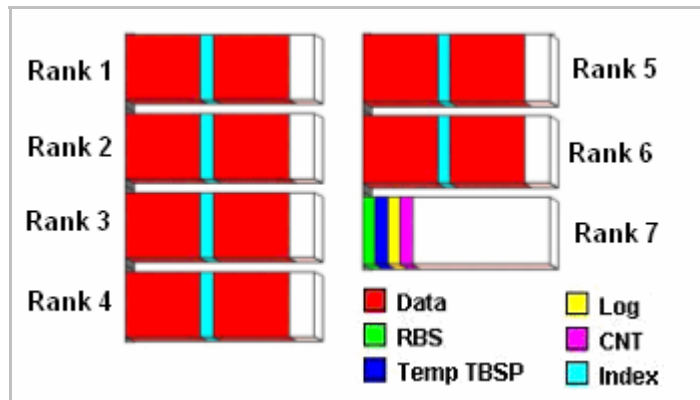
### Software

For this experiment we chose the newest releases for both Oracle and AIX. The software environment was as follows:

- AIX 5L (5.1)
- Oracle 9i (9.2)
- SDD 1.3.2.11
- DPAWS (Database Performance Analysis Workload Suite) (described previously)

### ESS Data Layout

To achieve an optimal storage set up for the database files the data was spread over 6 ranks and each rank was broken up into 8 GB LUNs. A seventh rank was used to hold the Log files, and this was also broken up into 8 GB LUNs. Customers would most likely choose LUN sizes greater than 8GB. However, using this size allowed us to more easily distribute the active data in this benchmark (approx. 140GB) across the entire RAID array, thereby creating a more realistic data distribution.



### AIX Logical Volumes

The AIX volume environment was set up as follows

- 13 volume groups, 12 for data and 1 for logs
- Each of the 12 Data volume groups contain 12 - 8 GB LUNs, 2 from each LSS/array (ranks 1 – 6)

- The log volume group contains 6 – 8 GB LUNs all from the same rank (rank 7)

## Tools

To monitor the performance of the workload and of the storage, the standard UNIX command *iostat* and for the ESS StorWatch Expert™ were used to monitor the disk activity.

Iostat is an AIX tool that monitors all disks that are connected to a pSeries machine. This program measures intervals -- in this case, every 1 minute -- and collects the total read/write bytes during the interval. It also measures the CPU utilization for the interval, as well as the disk transactions per second.

ESS StorWatch Expert is an IBM product that allows gathering of many internal statistics on the performance and utilization of the ESS and the disks inside.

## RAID Variations

To determine whether there is a significant performance difference resulting from the use of different types of RAID with an Oracle database, several different scenarios were put to the test.

- Data and Logs RAID 5
- Data on RAID 5, Logs on RAID 10
- Data and Logs RAID 10

For each of these tests the same workload was run, and a new database was loaded in-between the migration from the data on RAID 5 to RAID 10

- **Data and Logs RAID 5**

For this test all the ranks in the ESS were formatted using RAID 5. The formatted arrays were then broken into 24 – 8 GB LUNs and attached via two fiber connections to a pSeries server. All of the volume groups were created and the DPAWS database was loaded. After the database was loaded the 4 tests were performed and data was taken.

- **Data on RAID 5 Logs on RAID 10**

For this test the logs were transferred to RAID 10. To do this another rank on the same LSS was formatted to RAID 10 and then attached to a pSeries server. A physical volume move was then performed through LVM to transfer the data from the RAID 5 array to the new RAID 10 array. Once this was completed the 4 tests were again performed.

- **Data and Logs on RAID 10**

For this test all for the ranks in the ESS were formatted using RAID 10. The formatted arrays were then broken into 24 – 8 GB LUNs and attached via two fibre connections to the pSeries server. All of the volume

groups were created and the DPAWS database was loaded. The set up in the end looked exactly like the previous RAID 5 setup to the pSeries server and to Oracle. After the database was loaded the 4 tests were performed and data was taken.

© Copyright IBM Corporation 2002  
IBM Storage Systems Group  
5600 Cottle Road  
San Jose, CA 95136  
Produced in the United States

April 2003

All Rights Reserved

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This information could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or programs(s) at any time without notice.

The performance data contained herein was obtained in a controlled, isolated environment. Actual results that may be obtained in other operating environments may vary significantly. While IBM has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead. It is the user's responsibility to evaluate and verify the operation of any non-IBM product, program or service.

The information provided in this document is distributed "AS IS" without any warranty, either express or implied. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

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 in connection with this publication 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 provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing  
IBM Corporation  
North Castle Drive  
Armonk, NY 10504-1785  
U.S.A.

Trademarks

The following terms are trademarks of International Business Machines Corporation in the United States, other countries, or both: AIX, IBM, AS/400, DFSMSdfp, DFSMSdss, DFSMSshm, DFSMSrmm, DFSORT, Enterprise Storage Server, ESCON, FICON, FlashCopy, iSeries, Magstar, MVS/ESA, Netfinity, OS/390, OS/400, pSeries, RS/6000, S/390, SANergy, Tivoli, TotalStorage, VM/ESA, VSE/ESA, xSeries, z/OS, z/VM and zSeries

Other company, product or service names may be trademarks or service marks of others.