

# AutoTradeCenter scalability study

Oracle9i Real Application Clusters (RAC) on HP platforms running Linux®



|   |    |
|---|----|
| Executive overview and business motive for this study .....                       | 3  |
| Synopsis .....  | 3  |
| Acknowledgment .....  | 4  |
| Scope .....   | 4  |
| The HP benchmark environment.....   | 4  |
| System and management tools .....   | 5  |
| System and Oracle tools .....   | 5  |
| HP management tools .....   | 5  |
| HP Systems Insight Manager.....   | 5  |
| HP OpenView Performance Manager and Performance Agents .....                      | 7  |
| Load testing methodology .....  | 7  |
| ATC system overview.....  | 7  |
| Test cases.....   | 8  |
| Load generation tool .....  | 8  |
| Transaction response time .....   | 9  |
| Hardware and software environment .....   | 9  |
| Hardware environment .....  | 9  |
| Software stack .....  | 10 |
| Oracle Database configuration.....  | 11 |
| Load testing protocol and performance metrics collection .....                    | 12 |
| Analysis of load testing results .....  | 12 |
| Scalability as measured by transaction response time .....                        | 12 |
| Performance of the DL580 and BL20p as measured by transaction response time ..... | 15 |
| OS and database performance metric analysis.....                                  | 15 |
| Disk I/O analysis .....   | 15 |
| RAC interconnect network usage analysis.....                                      | 16 |
| CPU usage analysis .....  | 18 |
| Conclusions .....   | 26 |

|  |    |
|--|----|
| Appendix 1: hardware and network schematics .....      | 28 |
| Appendix 2. SilkPerformer test cases.....              | 29 |
| Honda Buyer Search .....                               | 29 |
| Subaru Admin .....                                     | 30 |
| Appendix 3. HP Systems Insight Management Agents ..... | 32 |
| For more information.....                              | 35 |

# Executive overview and business motive for this study

AutoTradeCenter (ATC) had completed a total overhaul of its trading systems by summer of 2003. In order to save costs and to ensure scalability, ATC decided to deploy everything on Linux. This included ATC databases, which were migrated from Oracle®/Solaris to Oracle Real Application Clusters (RAC) on Red Hat and HP—with valuable help from the HP Parallel Database Cluster program.

The promise behind the Oracle RAC/Linux/HP back end was one of availability and scalability (to many multiples of our needs at the time)—while at the same time remaining cost competitive. Since ATC makes its money on the actual transactions occurring on our site, it was essential for ATC to be able to respond quickly to business growth, while preferably purchasing the hardware at the last possible instance.

In December of 2003, ATC received a Request for Proposal from one of the major car manufacturers. After the initial feeling of joy evaporated, a scary thought set in: If ATC got this account, it could double and perhaps even triple the business within a short period of time. Although ATC was glad that it was on an adaptive infrastructure that promised growth, it was important to determine at that point how scalable its back-end systems were—and at what cost.

ATC turned to HP to craft an experiment built on the HP Parallel Database Cluster architecture to validate ATC's growth potential, quantify the cost, determine the most cost-effective growth strategy, and also determine the best strategy for maintaining adequate ceilings to allow for performance headroom.

The key drivers for the back-end system are availability, scalability, and cost. Without a strong combination of the three, ATC would not be nearly as successful as it is today. Just minutes of unscheduled downtime can cost ATC a client. In addition, given the rapid pace ATC business is evolving at its back-end systems must be scalable with little effort. And, as a young company just coming out of the gates, keeping costs in line is an important priority.

As you will read in the coming pages, the results ATC obtained in its scalability study were impressive: The ATC, Oracle, and HP teams proved nearly linear scalability with the Oracle RAC/Red Hat/HP back-end systems. ATC will not outgrow this platform anytime soon. As an added bonus, ATC even found a strategy to reduce costs while increasing performance and capacity. From ATC's experience with the RAC clusters and HP servers, back-end system availability goals can continually be achieved.

Oracle RAC and HP have come through on their promise to ATC: scalability and high availability in a grow-as-you-go format. In addition, the support of the people from Oracle and HP who have helped ATC has been invaluable in our quest to have the leanest, best system possible.

Jorge Borbolla, Chief Information Officer, AutoTradeCenter, Inc.

## Synopsis

The purpose of this study is to determine how well Oracle9i Real Application Clusters (RAC) scales on various Intel®-based database server platforms from HP. Instead of using a generic benchmark suite for this study, we selected some typical usage cases from AutoTradeCenter's (ATC's) Web-based online auction application system to generate workload on various hardware configurations running an Oracle9i RAC database. The results reflect real-life performance differences between these system configurations running ATC's Web-based online transaction processing (OLTP) application. The performance results, along with cost differences, will be used to drive ATC's hardware purchase decisions.

# Acknowledgment

The following individuals provided technical and/or organizational support to this project:

- HP: Carolyn Campbell, Parag Joshi, Darrin St. Amant, Terry Wood
- ATC: Jorge Borbolla, Pei Ku, Vladimir Kudlov, Bruce Ludlum, Ulf Zimmermann
- Oracle: Erik Peterson

## Scope

HP's ProLiant DL580 G2 and ProLiant BL20p G2 serve as the Oracle9i RAC servers in this study. The shared storage is the HP Modular SAN Array 1000 (MSA1000). These hardware platforms were selected because their price/performance characteristics match what ATC is interested in for its Oracle9i RAC database servers. The primary goals of this scalability study are as follows:

1. Determine how well Oracle9i RAC database performance scales on a DL580-based cluster as the number of nodes in the cluster scales from 2 to 3, 4, and 6
2. Determine how well Oracle9i RAC database performance scales on a BL20p-based cluster as the number of nodes in the cluster scales from 4 to 6
3. Compare performance of DL580-based Oracle9i RAC database configurations to BL20p-based configurations
4. Evaluate HP management tools: verify and validate the consistency and usefulness of HP management tools as compared to open source Linux tools

In an ideal world, when you double the number of nodes in a cluster, you would expect the workload the system could handle to double while maintaining the same performance level. In reality, however, you rarely get perfect linear scalability. Therefore, the objective of the first two goals listed above is to discover how close to ideal linear scalability the performance increase is as more nodes are added to the cluster.

The objective of the third goal is to compare performance differences between *comparable* DL580 clusters and BL20p clusters. Given the same total number of CPUs, the total cost of a DL580 cluster is roughly the same as a BL20p cluster. For example, total cost of two 4-way DL580 systems is roughly the same as that of four 2-way BL20p systems (for a total of eight CPUs in each example). In addition, Oracle9i RAC software licensing is based on number of CPUs, not hardware types. Therefore, because the total hardware and software price is roughly the same for a given total number of CPUs and Oracle software licenses, it is in the buyer's best interest to find out whether there is a performance difference between DL580 and BL20p clusters (obviously, users want more performance for the same dollar amount spent).

The objective of the fourth goal is to evaluate the usefulness and consistency of HP management tools in an Oracle9i RAC environment. For this benchmark, we used HP Systems Insight Manager and HP OpenView Performance Manager and Performance Agents.

## The HP benchmark environment

The systems used for the benchmark were located in the HP Solution Center in Houston, TX. Installation and configuration of Oracle9i was completed using the HP Parallel Database Cluster Kit (PDC Kit) scripts by the HP Solutions Center team in Houston. The AutoTradeCenter team was located in Menlo Park, CA. The HP team handling system management software was located in Cupertino, CA. A virtual private network (VPN) tunnel was set up to allow remote access to the Houston lab. Installation of ATC application software and tools was completed remotely by the ATC team, and

installation and configuration of HP management tools was done remotely by the HP team in Cupertino. All benchmarks were run remotely by ATC and all data-gathering from performance tools was also done remotely using Remote Desktop for the Windows® servers and SSH for the Linux servers. The HP management tools allow remote management via a browser-based user interface.

## System and management tools

### System and Oracle tools

The following tools were used during the benchmark:

- top—To capture load average and memory information
- iostat—Disk I/O information
- vmstat—CPU information
- iptraf—Network traffic information
- Oracle Statspack—Oracle software information
- Oracle Enterprise Manager—Oracle management tool

### HP management tools

The following HP management tools were used during the benchmark:

- HP Systems Insight Manager and HP Server Management Agents—Management platform used to monitor all servers and storage for hardware and software problems
- HP OpenView Performance Manager and Performance Agents—Collecting performance statistics from database servers

#### **HP Systems Insight Manager**

HP Systems Insight Manager (SIM) is a distributed, client/server software solution. It provides service-driven event and performance management of business-critical enterprise systems, applications, and services through its ability to monitor, control, and report on the health of a system.

HP SIM was used to monitor the benchmark setup. HP ProLiant server management, storage, and network interface card (NIC) agents were installed on all servers in the configuration to monitor for actual or impending component failures. Servers and storage were automatically discovered and identified. HP SIM was used to view the status of all servers and to collect events that were generated as faults were detected. During the benchmark, we experienced NIC failures and configuration problems that were detected by HP SIM.

Figure 1. HP Systems Insight Manager discovered systems

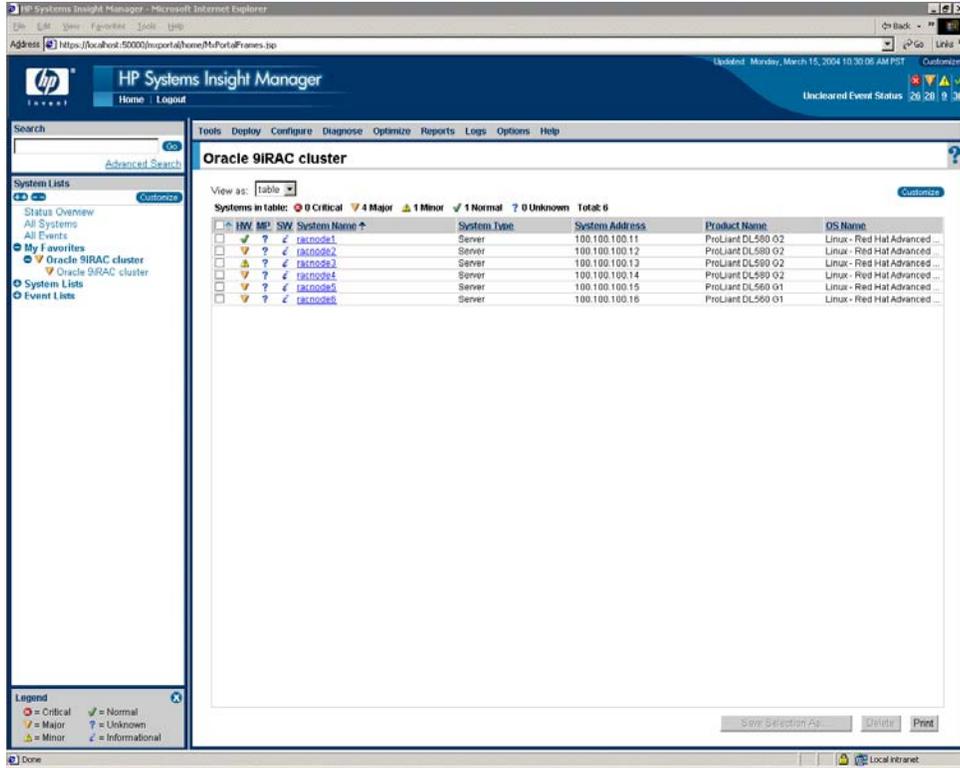


Figure 1 shows the 6-node ProLiant DL580 cluster as discovered by HP SIM. HP SIM automatically identifies the servers and provides a link so users can drill down on each server to view detailed status information.

Figure 2. HP Systems Insight Manager event log

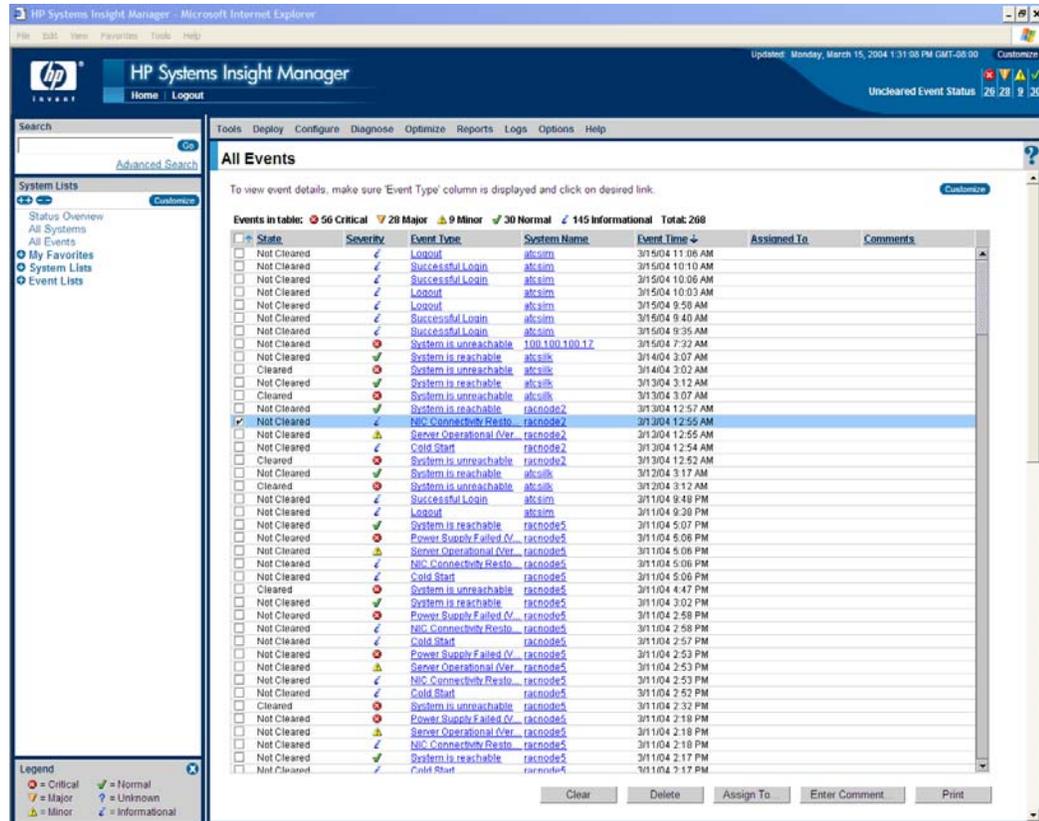


Figure 2 shows the HP SIM event log. HP SIM provides notification of component failure. Automatic event handling allows users to configure policies to notify appropriate users of failures via e-mail, pager, or short message service (SMS) gateway. HP SIM also enables automatic execution of scripts or event forwarding to enterprise platforms.

### HP OpenView Performance Manager and Performance Agents

HP OpenView Performance Manager utilizes data collected from the HP OpenView Performance Agents to isolate performance bottlenecks and help maximize resource uptime. Performance Agents log and collect data, then send alarms about that data when necessary. Performance Agents were installed on all database servers in the benchmark setup, and they collected operating system, network traffic, and shared storage access metrics. Performance Manager was used for performance analysis, generating reports and graphs after the test runs completed.

## Load testing methodology

### ATC system overview

The ATC Web-based online auction system consists of the following major components:

- The Application Server Layer contains a set of Web/application servers that run Apache/Tomcat servers that run Servlets written in Java™.
- The Database Server Layer is where the Oracle9i RAC cluster resides.

ATC end users access the ATC system with Web browsers such as Microsoft® Internet Explorer. The traffic between the ATC end user’s browser and the ATC system is limited to HTML and small graphic files. There are no applets used in the ATC system. In other words, ATC end users interact with the Application Server Layer only, and the Application Server Layer only accesses the Database Server Layer.

## Test cases

We scripted two test cases for the purpose of this study. Each test case represents a series of URLs that a typical ATC end user visits using a Web browser in order to accomplish some typical task in the ATC online auction system. For example, one test case simulates a user looking up Honda vehicles in an auction on the ATC’s Honda “private label” site. After each URL is visited, a fixed “think time” is inserted before the next URL is visited. The test cases are selected based on the following criteria:

- They represent some typical workflows that ATC end users perform
- They perform database-intensive activities

The last point is important: We intentionally select database-intensive workflows as test cases in order to be sure that the database layer experiences contention before the application server and other layers experience contention. By making sure that the bottleneck resides in the database cluster layer for all configurations and all workload levels tested, variation in performance can be attributed to differences in database layer performance characteristics alone.

It is also important to point out that the test cases selected are heavily biased toward database-read activities, as opposed to database-write activities. This bias reflects the real-life usage patterns in the ATC application system: We expect, the majority of the time, that users are performing searches in the ATC system; data modification activities are a small percentage of the typical workflow patterns.

## Load generation tool

We used Segue’s SilkPerformer to simulate ATC end users executing the usage test cases described previously. These virtual users (VUs) are, in essence, software run by SilkPerformer to simulate real users using Web browsers to access the ATC application. Using SilkPerformer, we can vary the workload imposed on the overall system by varying the number of virtual users.

**Table 1.** Load levels

|                                   | Number of VUs for test case<br>“Subaru Admin” | Number of VUs for test case<br>“Honda Buyer Search” |
|-----------------------------------|---|---|
| VU load level #1<br>(factor = 1X) | 2   | 10  |
| VU load level #2<br>(factor = 2X) | 4   | 20  |
| VU load level #3<br>(factor = 3X) | 6   | 30  |

When varying the workload, the ratio of VUs between the two test cases is held constant. As shown in Table 1, by holding this ratio constant, we can directly compare the levels of the load imposed by different workloads (for example, workload C is three times as heavy as workload A).

## Transaction response time

The metric used to compare the performance of different configurations is based on transaction response time. In the context of this benchmark, a transaction is the complete execution of a particular test case by a VU. For each load test run, we collected the following information:

- Oracle9i RAC cluster profile (number of instances, hardware profile, etc.)
- Overall average transaction response time, which is the average of the elapsed time of all the transactions executed by all VUs during the test run (sleep time inserted between the steps in each transaction is not included in calculating the average transaction response time)
- Average transaction response time of the executions of a particularly database-intensive transaction; this transaction (test case) contains mostly database-intensive actions and, therefore, the resulting metric is minimally skewed by no- or low-database-activity steps in the transaction.

For each RAC configuration (such as the cluster of 6 DL580 systems), we run load tests at various workload levels. At the lowest workload level, where there is no resource (CPU, RAM, I/O) contention, the response time represents the “best-case” metric. As the workload increases (for example, the number of VUs increases), at some point resource contention will surface. Response time will begin to increase at that point.

To compare the relative “horsepower” of two configurations, we select an arbitrary response time threshold that is some multiple of the no-contention response time level. For example, for configuration A, assume that it takes X number of VUs to reach that response time threshold; for configuration B, assume that it takes Y number of VUs to reach the same threshold. The ratio between X and Y represents the relative performance ratio between the two configurations.

The reason we collect both overall average transaction response time as well as the average response time of a specific database-query-intensive transaction is that in the former case, the response times of low- and no-database-access steps in the transaction are included in the calculation. We want to see a separate metric that is not significantly skewed by the response time of those steps.

## Hardware and software environment

### Hardware environment

Table 2 shows the configurations of the systems used for running SilkPerformer, the application/Web servers, and the system that runs HP SIM and the Performance Manager console. These system configurations remain the same for all load tests.

**Table 2.** System configurations

| Usage             | Number of servers | Hardware description                   |
|-------------------|-------------------|--|
| App/Web servers   | 2                 | BL20p, 2 x 3.06 GHz, 2.5 GB RAM        |
| Image server      | 1                 | DL580, 4 x 1.6 GHz, 4 GB RAM           |
| Management server | 1                 | DL580, 4 x 1.6 GHz, 4 GB (or more) RAM |
| SilkPerformer     | 1                 | DL580, 4 x 1.6 GHz, 4 GB (or more) RAM |

Table 3 shows the configurations of the DL580 systems and the BL20p systems used in the database layer. Each load test run is executed against an Oracle9i RAC cluster composed of some number of these systems.

**Table 3.** Details of the DL580 and BL20p systems

|                        | <b>DL580</b>         | <b>BL20p</b>        |
|------------------------|----------------------|---------------------|
| <b>Number of CPUs</b>  | 4                    | 2                   |
| <b>CPU</b>             | 2.80 GHz Intel Xeon™ | 3.06 GHz Intel Xeon |
| <b>Processor cache</b> | 2 MB L3 cache        | 512 KB L2 cache     |
| <b>RAM</b>             | 4.0 GB @ 200 MHz     | 2.5 GB @ 266 MHz    |

**Table 4.** Tested Oracle9i RAC cluster hardware configurations

| <b>Machine type</b> | <b>Number of machines in cluster</b> |
|---------------------|--------------------------------------|
| DL580               | 2                                    |
| DL580               | 3                                    |
| DL580               | 4                                    |
| DL580               | 6                                    |
| BL20p               | 4 <sup>1</sup>                       |
| BL20p               | 6 <sup>2</sup>                       |

See Appendix 1 for the schematic of the hardware and network layout.

## Software stack

Following are the software stack details.

Application/Web servers:

- Red Hat Linux 8
- Apache 2.0.45
- Tomcat 4.1.24
- ATC Application, "Daiquiri" release
- HP ProLiant Management Agents
  - HP Server Management Agents for Red Hat Linux 8.0 7.0.0-21
  - HP NIC Agents for Servers 7.0.0-4
  - HP Storage Agents for Linux 7.0.0-16

Image servers:

- Red Hat Linux 8
- Apache 2.0.45

Management server:

- Microsoft Windows 2000 SP4
- HP Systems Insight Manager C.04.00.00
- HP OpenView Performance Manager A.04.04.06

<sup>1</sup> 4-node BL20p cluster has the same number of CPUs (8) and a similar price point as the 2-node DL580 cluster.

<sup>2</sup> 6-node BL20p cluster has the same number of CPUs (12) and a similar price point as the 3-node DL580 cluster.

- HP Insight Management Agents for Windows 2000 7.0.0.0
- Oracle Enterprise Manager

SilkPerformer server:

- Windows 2000 SP4
- HP Insight Management Agents for Windows 2000 7.0.0.0
- Segue SilkPerformer 6.0.1 (console and agent)

Database servers:

- Red Hat Linux AS 2.1, kernel version 2.4.9-e27
- Oracle9i RAC version 9.2.0.4
- Oracle Cluster File System 1.0.9-9
- HP ProLiant Management Agents
  - HP Server Management Agents for Red Hat Enterprise Linux 2.17.0.0-21
  - HP NIC Agents for Servers 7.0.0-4
  - HP Storage Agents for Linux 7.0.0-16
- HP OpenView Performance Agents for Linux C.04.00

## Oracle Database configuration

Following are the relevant non-default init.ora parameter values:

|                               |            |       |
|-------------------------------|------------|-------|
| db_block_size                 | 8192       | bytes |
| db_cache_size                 | 1258291200 | bytes |
| db_file_multiblock_read_count | 16         |       |
| fast_start_mttr_target        | 300        |       |
| log_buffer                    | 2000384    | bytes |
| max_commit_propagation_delay  | 0          |       |
| open_cursors                  | 300        |       |
| pga_aggregate_target          | 104857600  | bytes |
| shared_pool_size              | 218103808  | bytes |

Note: The sum of shared\_buffer\_pool, db\_block\_size, db\_cache\_size, log\_buffer, and pga\_aggregate\_target is about 1.5 MB—less than the total available physical memory in both the DL580 and BL20p database server systems.

# Load testing protocol and performance metrics collection

For each of the 6 Oracle9i RAC cluster configurations (listed in the previous section), up to 5 different VU load levels are executed, as shown in Table 5.

**Table 5.** VU load levels

| VU load level | Number of VUs running "Honda Buyer Search" test case | Number of VUs running "Subaru Admin" test case | Total number of VUs |
|---------------|--|--|---------------------|
| 1             | 20   | 4  | 24                  |
| 2             | 40   | 8  | 48                  |
| 3             | 60   | 12   | 72                  |
| 4             | 80   | 16   | 96                  |
| 5             | 100  | 20   | 120                 |

The following performance metrics are collected for each run. Each run is identified by the number of VUs and the particular Oracle9i RAC hardware configuration used in the load test.

Immediately before the start of a run, the following steps are executed:

- For each of the Oracle9i RAC instances used, take a "begin" Statspack snapshot.
- Start "sar -u", "sar -q", and "iostat" with an interval length of 60 seconds (1 minute) and a sample count of 15 (15 minutes is the duration of each run).

Immediately after the completion of a run, the following steps are executed:

- For each of the Oracle9i RAC instances used, take an "end" Statspack snapshot. The "delta" between the "begin" and "end" Statspack snapshot reflect the Oracle instance performance metrics during the run period.
- Transaction response times (including the "overall average transaction response time" and "average response time of Honda Buyer Search test case") captured by SilkPerformer are recorded.

In addition, HP OpenView Performance Agents continuously collect OS-level, hardware-level, and network-level metrics on all of the database server nodes used during testing.

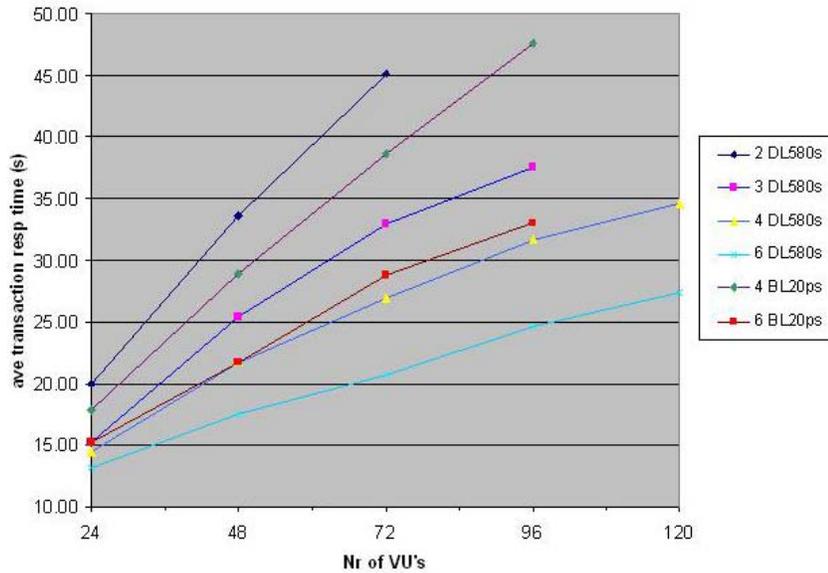
## Analysis of load testing results

This section provides the analysis of the performance metrics generated from the load testing runs.

### Scalability as measured by transaction response time

Figure 3 and Table 6 show the overall average transaction response time of the ATC system tested under various combinations of VU load levels and Oracle9i RAC hardware configurations.

**Figure 3.** Overall average transaction response time of the ATC system



**Table 6.** Number of VUs supported at 25-second overall average response time

|                                  | Number of RAC instances: DL580 systems |      |      |      | Number of RAC instances: BL20p systems |      |
|----------------------------------|--|------|------|------|--|------|
|                                  | 2                                      | 3    | 4    | 6    | 4                                      | 6    |
| RAC instances                    | 2                                      | 3    | 4    | 6    | 4                                      | 6    |
| VUs                              | 33                                     | 47   | 63   | 100  | 39                                     | 59   |
| Perfect scalability              | 1.00                                   | 1.50 | 2.00 | 3.00 |  |      |
| Normalized                       | 1.00                                   | 1.43 | 1.92 | 3.03 | 1.20                                   | 1.80 |
| Perfect scalability (BL20p only) |  |      |      |      | 1.00                                   | 1.50 |
| Normalized (BL20p only)          |  |      |      |      | 1.00                                   | 1.50 |

As shown in Table 6, using a 25-second response time as the common threshold, the number of VUs supported at that threshold on the DL580 configurations tested scales at a rate close to the perfect scalability (“normalized” 1.00/1.43/1.92/3.03 vs. “perfect scalability” of 1.00/1.50/2.00/3.00). Therefore, in the context of this benchmark, we have very good scalability of the ATC application from 2 to 6 DL580 nodes of the Oracle9i RAC database. The BL20p systems reach the perfect scalability: 1.00/1.50 vs. 1.00/1.50.

Please note the following:

- Because of the way the Application Server Layer and the Oracle Java Database Connectivity (JDBC) thin-client load-balancing TNS descriptor mechanism work, it is rare that we achieve perfect load balancing across the available Oracle9i RAC instances (more will be said about this in a later section). Therefore, there might be more contention on some nodes than others, so the average response time for a given run might be somewhat worse than the ideal scenario where perfect load balance had been achieved.

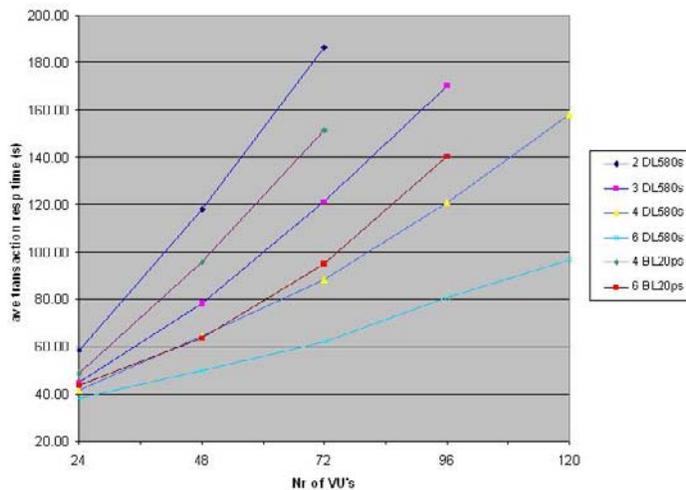
- “Overall” average response time includes response time of low- and no-database-activity steps in the transactions. Since the response times of these steps remain fairly constant (they are determined more by Application Server Layer performance) as more VUs are added (assuming the application servers have not reached the contention threshold), they tend to hold down the rate of increase of overall transaction response time. The effect can be seen in the slight downward curving of the graph lines shown in Figure 3.

The average response time of the Honda Buyer Search test case is shown in Figure 4. This test case consists primarily of database-intensive steps with very few low-database-activity steps.

Figure 4 shows graph lines that are fairly straight with a slight upward curving. This means that, as the number of VUs increases linearly, the rate at which response time increases is slightly faster than the linear rate. This is to be expected; as the level of contention rises at the database level, additional overhead will be consumed to manage the contention.

Table 7 provides good news: Even for this database-read-intensive test case, the performance scales very well (i.e., close to perfect scalability).

**Figure 4.** Average response time of Honda Buyer Search test case



**Table 7.** Number of VUs supported at 80-second response time for Honda Buyer Search

|                                  | Number of RAC instances: DL580 systems |             |      |      | Number of RAC instances: BL20p systems |             |
|----------------------------------|--|-------------|------|------|--|-------------|
|                                  | 2                                      | 3           | 4    | 6    | 4                                      | 6           |
| RAC instances                    | 2                                      | 3           | 4    | 6    | 4                                      | 6           |
| VUs                              | 33                                     | 49          | 64   | 95   | 40                                     | 61          |
| Perfect scalability              | 1.00                                   | 1.50        | 2.00 | 3.00 |  |             |
| Normalized                       | 1.00                                   | 1.49        | 1.94 | 2.90 | 1.22                                   | 1.86        |
| Perfect scalability (BL20p only) |  |             |      |      | 1.00                                   | 1.50        |
| Normalized (BL20p only)          |  |             |      |      | 1.00                                   | 1.52        |
| Normalized to 2 DL580 systems    | <b>1.00</b>                            |             |      |      | <b>1.22</b>                            |             |
| Normalized to 3 DL580 systems    |  | <b>1.00</b> |      |      |  | <b>1.24</b> |

## Performance of the DL580 and BL20p as measured by transaction response time

Based on Table 7, a 4-node BL20p cluster has a 1.22:1 performance advantage over a 2-node DL580 cluster (each of these configurations has a total of 8 CPUs). Similarly, a 6-node BL20p cluster has a 1.24:1 performance advantage over a 3-node DL580 cluster (each of these configurations has a total of 12 CPUs).

The slightly faster CPU speed and 33% faster RAM speed may be the primary contributing factors for the superior performance of the BL20p. A larger processor cache in the DL580 is overshadowed by these other factors (refer to Table 4 for configuration information).

**Note:** Be careful not to generalize the apparent advantage of the BL20p over the DL580. Remember that the ATC application test cases selected can be characterized by the following:

1. The test cases had high database-read activities and minimal database-write activities. Furthermore, since the system global area (SGA) has been sized and the data access patterns of the test cases are such that once a test case run “ramps up,” most database-read accesses are buffer cache hits in the SGA (i.e., minimal disk read I/Os). In other words, the benchmark is, by and large, comparing the processor and memory subsystems between DL580 and BL20p clusters.
2. Statistics collected confirm that there is minimal Oracle9i RAC interconnect traffic. Again, this is due to the minimal database-write characteristics of the benchmark. It is conceivable that, with a different workload in which high database-write activities dominate, the interconnect traffic between four BL20p systems will create significant overhead compared to two DL580 systems. However, validation of this theory is beyond the scope of this benchmark study.

## OS and database performance metric analysis

### Disk I/O analysis

For this study, Oracle9i RAC database init.ora parameters are unchanged for all load test runs conducted. The SGA and other parameters are sized to minimize disk I/O. In essence, the system configurations are designed with the intention of comparing primarily CPU/memory subsystems.

Based on the iostats and Statspack data collected, the load tests do incur minimal I/O activities. The data from iostats shows less than 1% I/O utilization on the device Oracle Database resides on (an Oracle cluster file system [OCFS]), with a maximum I/O utilization of 3%.

Figure 5. OpenView Performance Manager shared storage access statistics

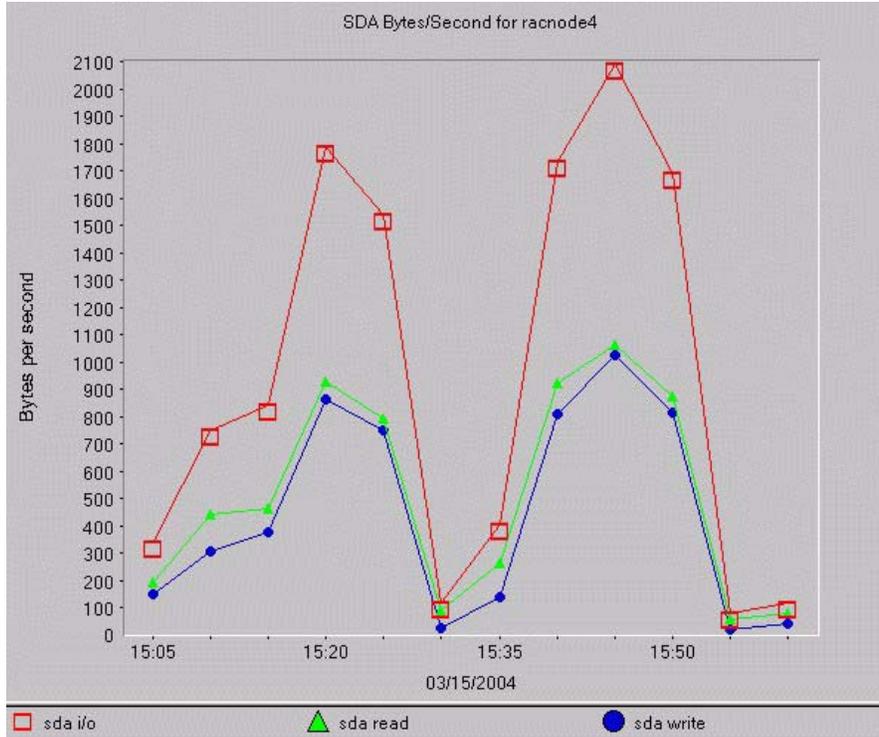


Figure 5 shows access to shared storage (MSA1000) by the fourth node in a 6-node DL580 cluster with 96 and 120 virtual users. I/O was only slightly more than 2 KB per second during the heaviest load on node 4. I/O traffic was not equal on all servers, but the highest rate was less than 3 KB per second on a 6-node cluster with 120 virtual users.

#### RAC interconnect network usage analysis

HP OpenView Performance Manager statistics on the Gigabit Ethernet interfaces used for passing Oracle9i RAC interconnect ("cache fusion") traffic show low packet counts (on average, less than 25 packets per second with a maximum of 50 packets per second). Oracle Statspack also shows that the instances are spending a negligible amount of time waiting for cache fusion events.

Figure 6. OpenView Performance Manager network traffic report

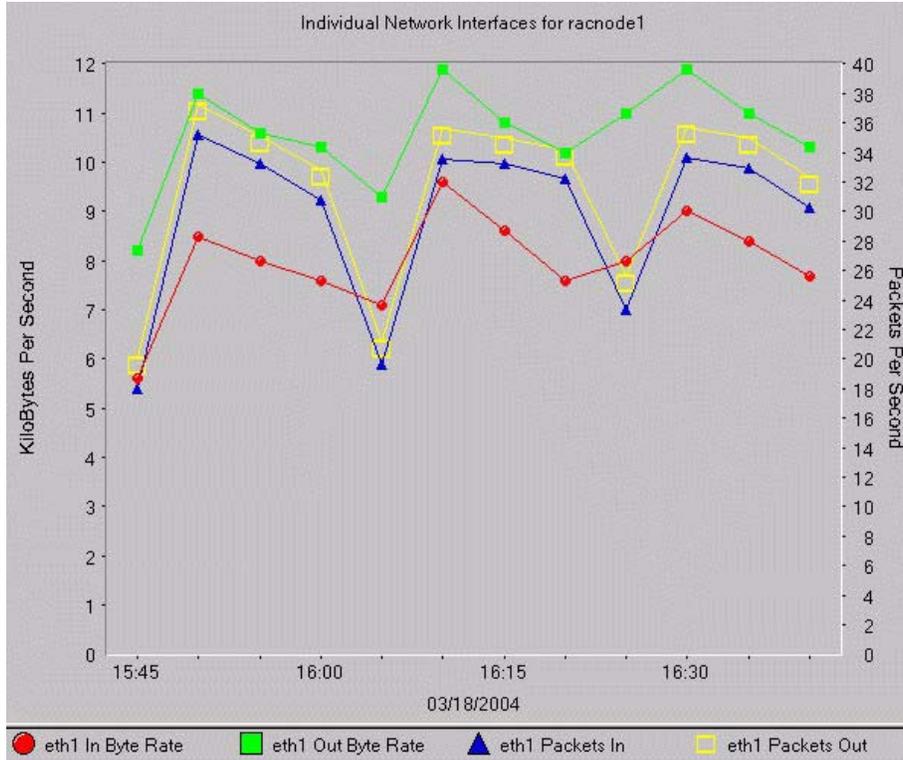


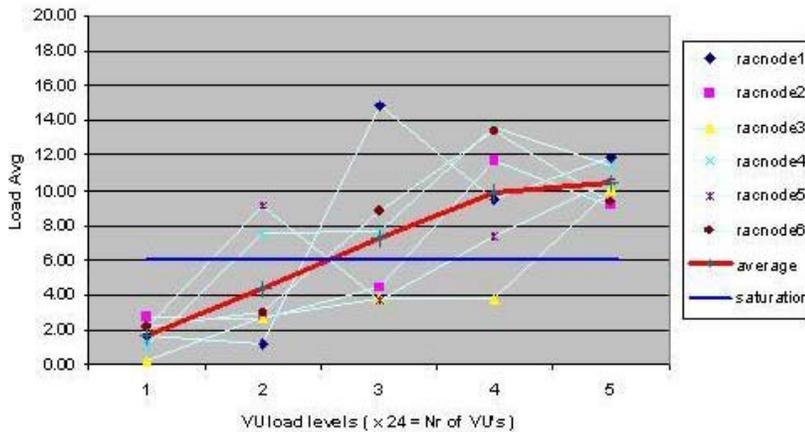
Figure 6 shows the network traffic for the first node in the 6-node cluster configuration. The statistics cover a window of 3 benchmark tests, from 48 to 72 to 96 virtual users. The network traffic peaked (total of In and Out bytes) at less than 22 KB per second on a Gigabit Ethernet link.

One can conclude that, for the test cases used in this study, Oracle9i RAC's cache fusion mechanism poses negligible overhead.

## CPU usage analysis

Since disk I/O and interconnect network traffic are not major factors in this study, the CPU and memory subsystems are the dominant factors in performance variation as the load varies. This section provides analysis of CPU usage.

**Figure 7.** 1-minute load average on 6-node DL580 cluster



**Figure 8.** CPU idle percentage vs. number of VUs on 6-node DL580 cluster

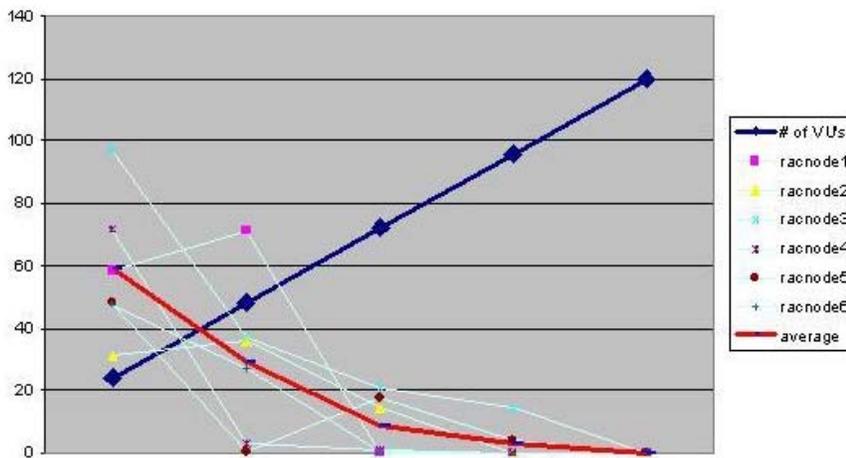


Figure 7 depicts the load average of the 6 DL580 systems at various VU load levels. Figure 8 depicts the CPU idle percentage statistics on the same set of systems.

CPU idle percentage describes the percentage of time that particular CPU is not in use. Load average value (on Linux) indicates the average number of running and runnable-but-waiting processes for a given time period.

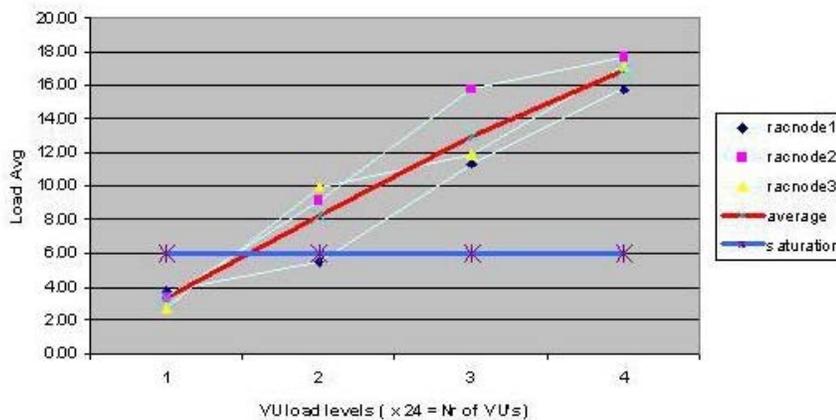
For a CPU-intensive workload (as opposed to a disk-intensive workload), the higher the load average, the lower CPU idle percentage. The lowest level of load average at which CPU idle percentage becomes 0 is the "CPU saturation" point.

Due to the less-than-perfect database connection load balancing via the JDBC Oracle thin-client TNS descriptor mechanism and the way in which Tomcat utilizes the established database connections, load average differs quite significantly between the available nodes for some of the load test runs. This fairly wide range of load average values (at various VU load levels) explains an equally wide standard deviation in average transaction response time (not shown in this paper). While not ideal for benchmark purposes, this is the not-so-desirable “real-life” behavior of the ATC system.

Nevertheless, when the load average and CPU idle percentage data points are averaged at each VU load level, the resulting “curves” fit expected performance behavior. As the VU load increases, CPU idle percentage decreases and eventually approaches zero, and load average increases at a steady rate, eventually crossing the “CPU saturation” level.

Figures 9 and 10 show similar metrics for load runs against other cluster configurations.

**Figure 9.** 1-minute load average on 3-node DL580 cluster



**Figure 10.** CPU idle percentage vs. number of VUs on 3-node DL580 cluster

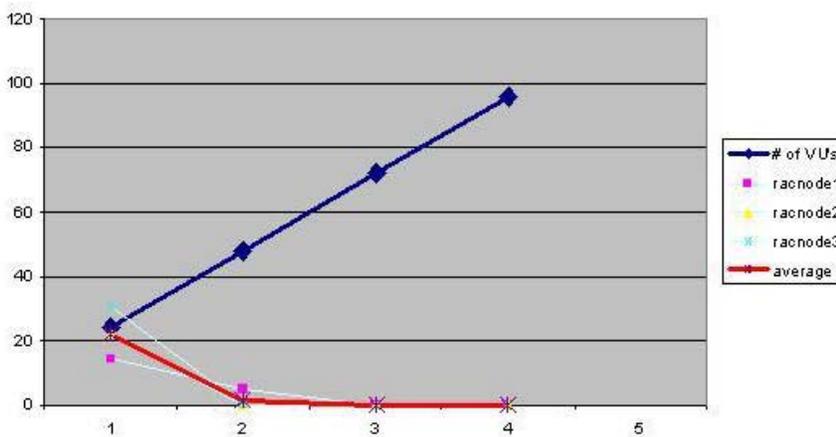


Figure 11. 1-minute load average on 6-node BL20p cluster

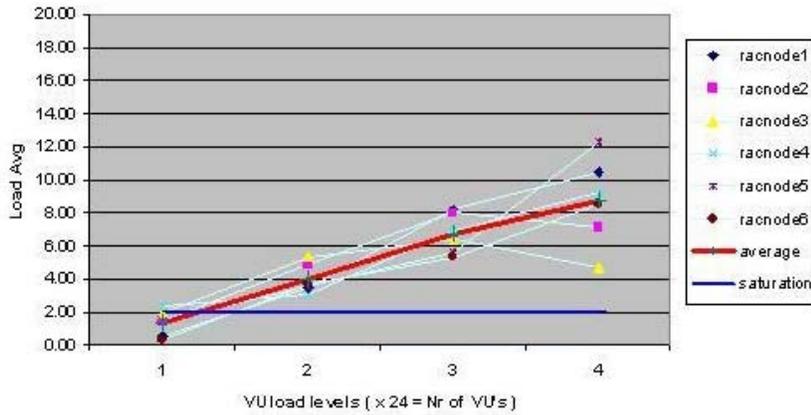
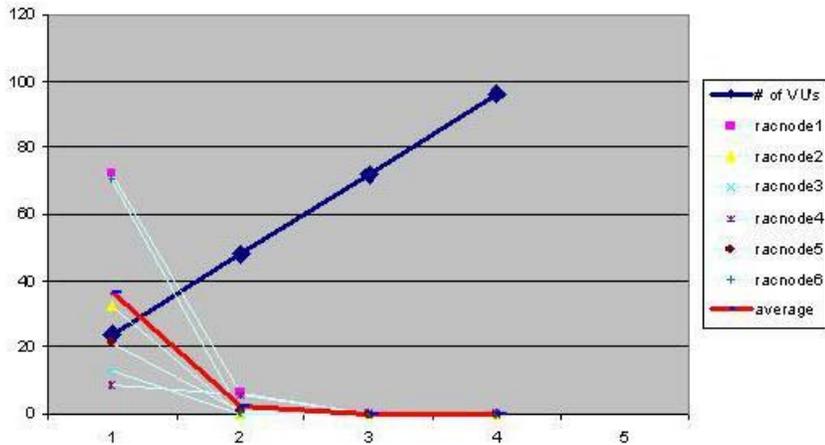


Figure 12. CPU idle percentage vs. number of VUs on 6-node BL20p cluster



Figures 11 and 12 illustrate the performance difference between the 6-node BL20p cluster and the 3-node DL580 cluster. As the VU load level increases, load average increases at a faster rate on the DL580 cluster than on the BL20p cluster. This means that higher CPU contention exists on the DL580 cluster, contributing to the slower overall response time.

**Figure 13.** System and user CPU utilization

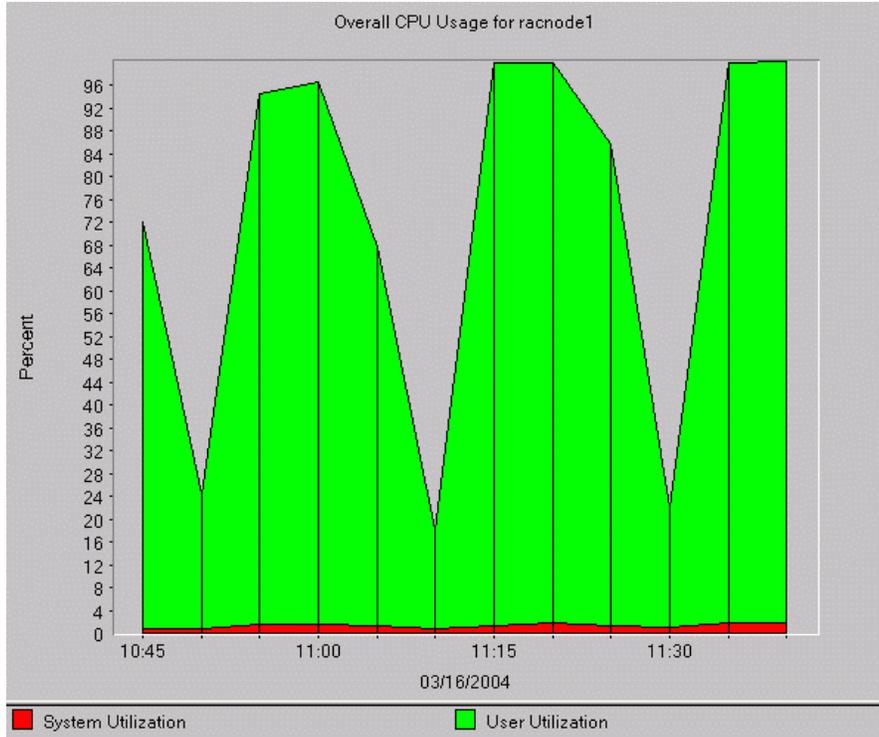


Figure 13 shows CPU utilization, separated into system and user-level software, during the benchmark tests on a DL580. This chart shows that virtually all of the CPU load was generated by the user-level code (Oracle), and operating system overhead was negligible. Similar ratios of system utilization and user utilization were seen for all tests on both platforms.

**Figure 14.** HP OpenView Performance Manager CPU utilization for a 3-node ProLiant DL580 cluster with 96 virtual users

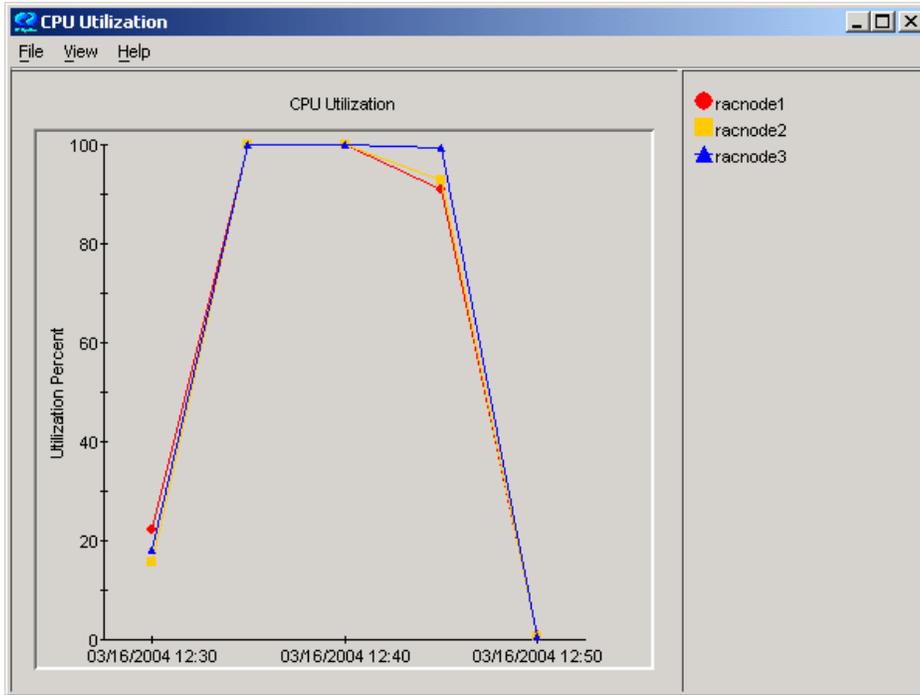


Figure 14, generated by OpenView Performance Manager, shows the CPU utilization on a 3-node DL580 cluster for a 96-virtual-user benchmark run. CPU utilization on all CPUs for all DL580 benchmark tests of greater than 48 virtual users was over 99.9%.

**Figure 15.** CPU utilization for a 6-node BL20p cluster with 96 virtual users

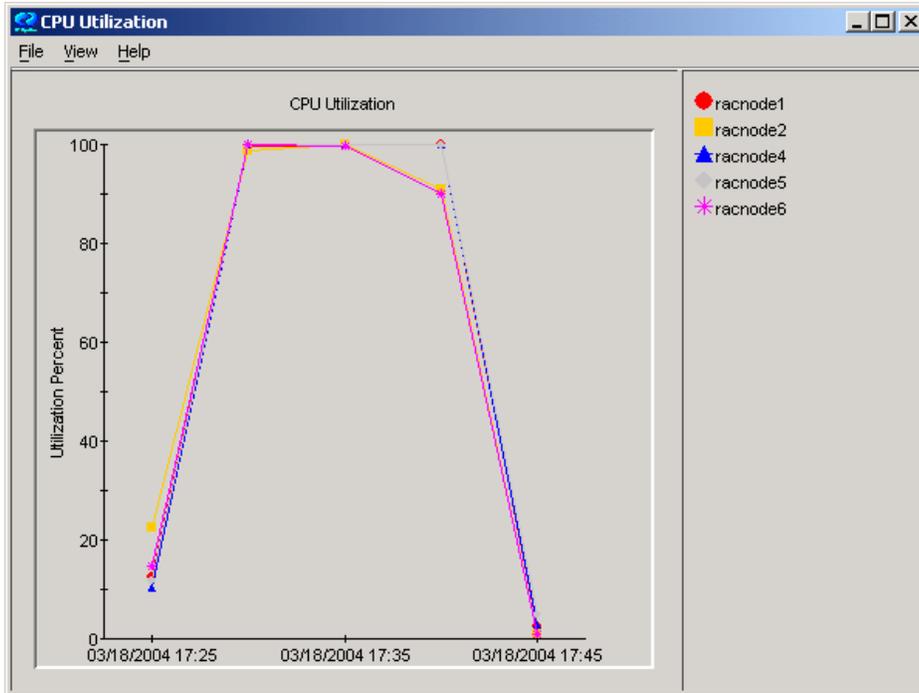


Figure 15 shows the CPU utilization on a 6-node BL20p cluster for a 96-virtual-user benchmark run. CPU utilization on all CPUs for all BL20p cluster benchmark tests of greater than 48 virtual users was over 99.9%.

Figures 16 and 17 show the global run queue for the BL20p and DL580 clusters with a load of 96 virtual users. The global run queue indicates the number of runnable processes during an interval. During this time period, CPU utilization was 99.9%. Together, these two metrics indicate a CPU bottleneck.

Figure 16. Global run queue for a 6-node BL20p cluster with 96 virtual users

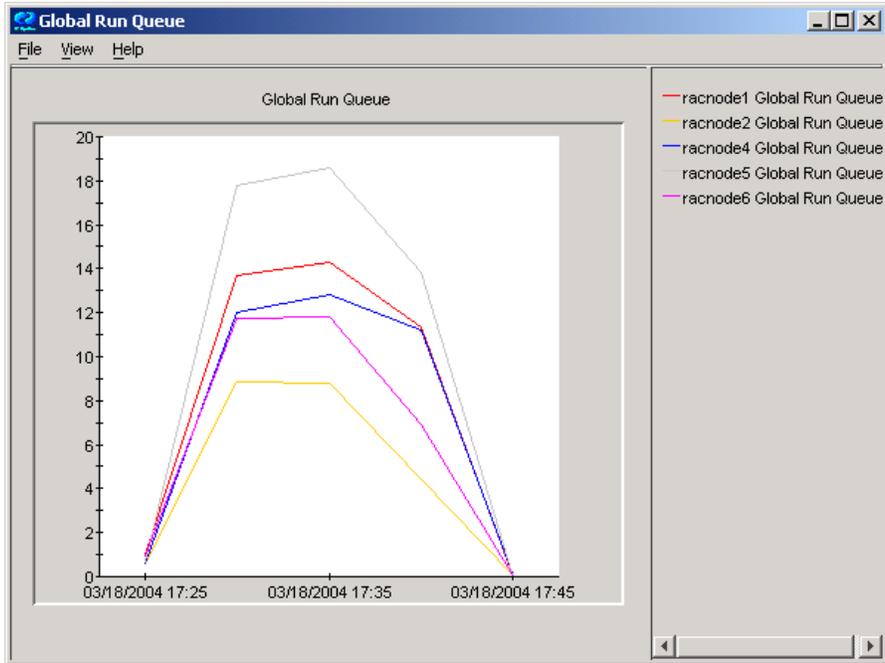
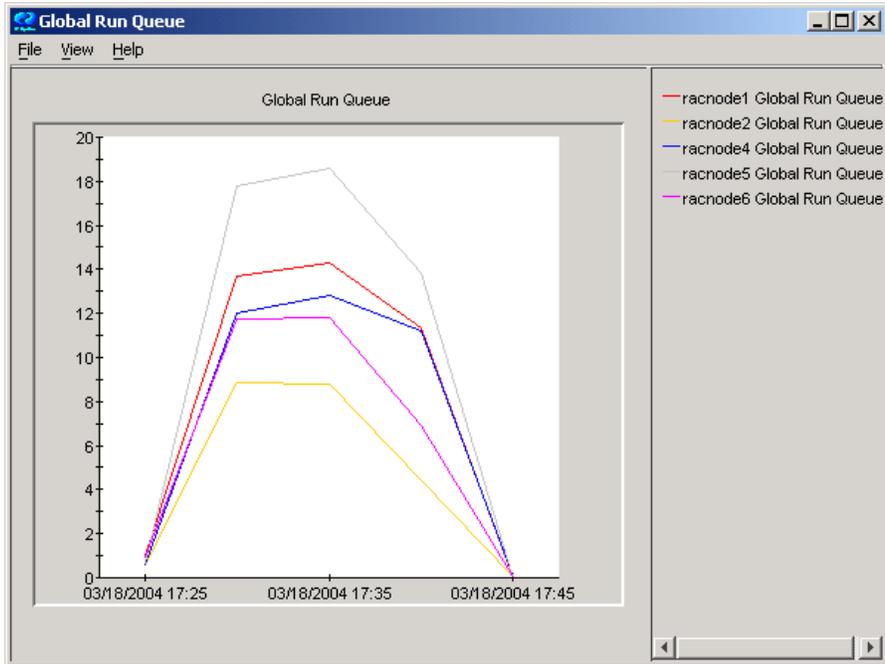


Figure 17. Global run queue for a 3-node DL580 cluster with 96 virtual users



Figures 18 and 19 show memory paging for both clusters during benchmark testing with a load of 96 virtual users. Memory paging is the total number of disk blocks paged into memory and paged out to disk per second during the interval. This includes pages paged in from paging space and from the file system, and pages paged out to paging space and to the file system.

Figure 18 shows the memory page request rate for a 6-node BL20p cluster running a load of 96 transactions per minute.

**Figure 18.** Memory paging for a 6-node BL20p cluster with 96 virtual users

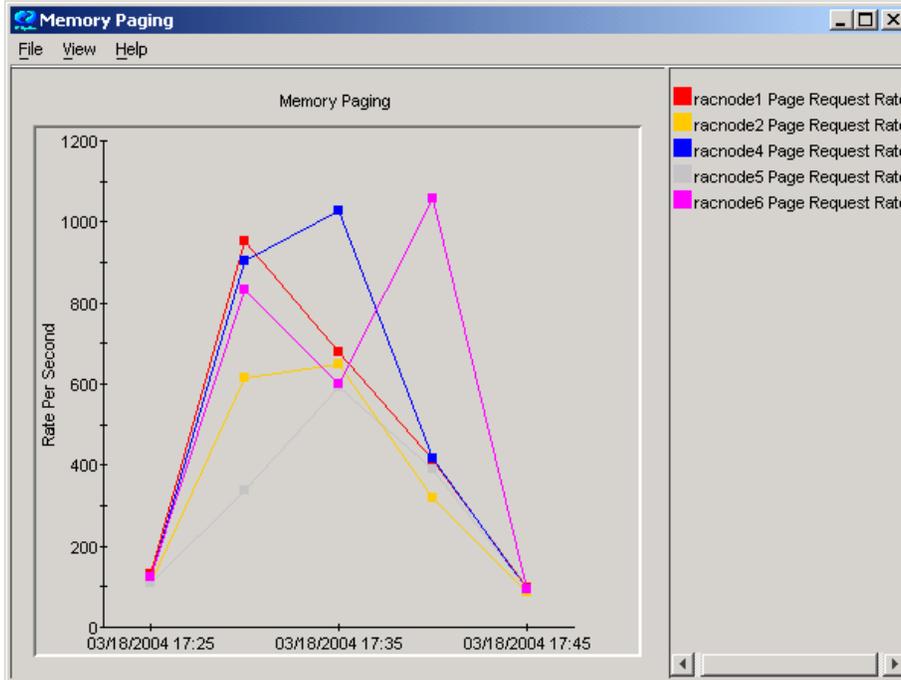


Figure 19. Memory paging for a 3-node DL580 cluster with 96 virtual users

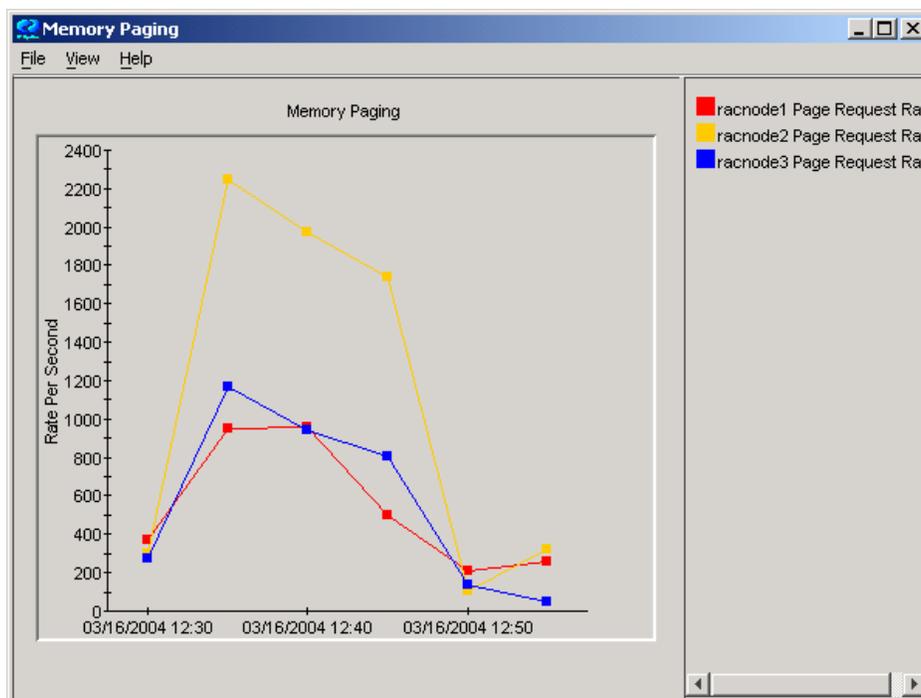


Figure 19 shows the memory page request rate for a 3-node DL580 cluster running a load of 96 transactions per minute.

Comparing the number of nodes could explain the memory paging differences. If each node has 3 GB of SGA, the 3-node DL580 cluster would have 9 GB of SGA. The 6-node BL20p cluster would have 18 GB of SGA. The BL20p cluster is more likely to find the data in cache using cache fusion than the DL580 cluster.

## Conclusions

This benchmark based on ATC's OLTP system has led to the following major conclusions, drawn from the results of load testing:

- Oracle9i RAC scales very well (i.e., close to perfect linear scalability).
- For the same number of CPUs, the 2-way BL20p cluster performs better than the 4-way DL580 cluster.
- HP management tools provided information that was critical to remotely managing and monitoring the benchmark configuration.

From a purely price/performance perspective, BL20p-based clusters have an advantage over DL580-based clusters. However, since the ProLiant BL20p G2 cluster provides only dual-processor support, more systems are needed in order to achieve the same number of CPUs as a comparable ProLiant DL580 G2 cluster. Physical footprint is not an issue since each BL20p is just an additional blade in the chassis, but there may be additional management overhead for managing more systems.

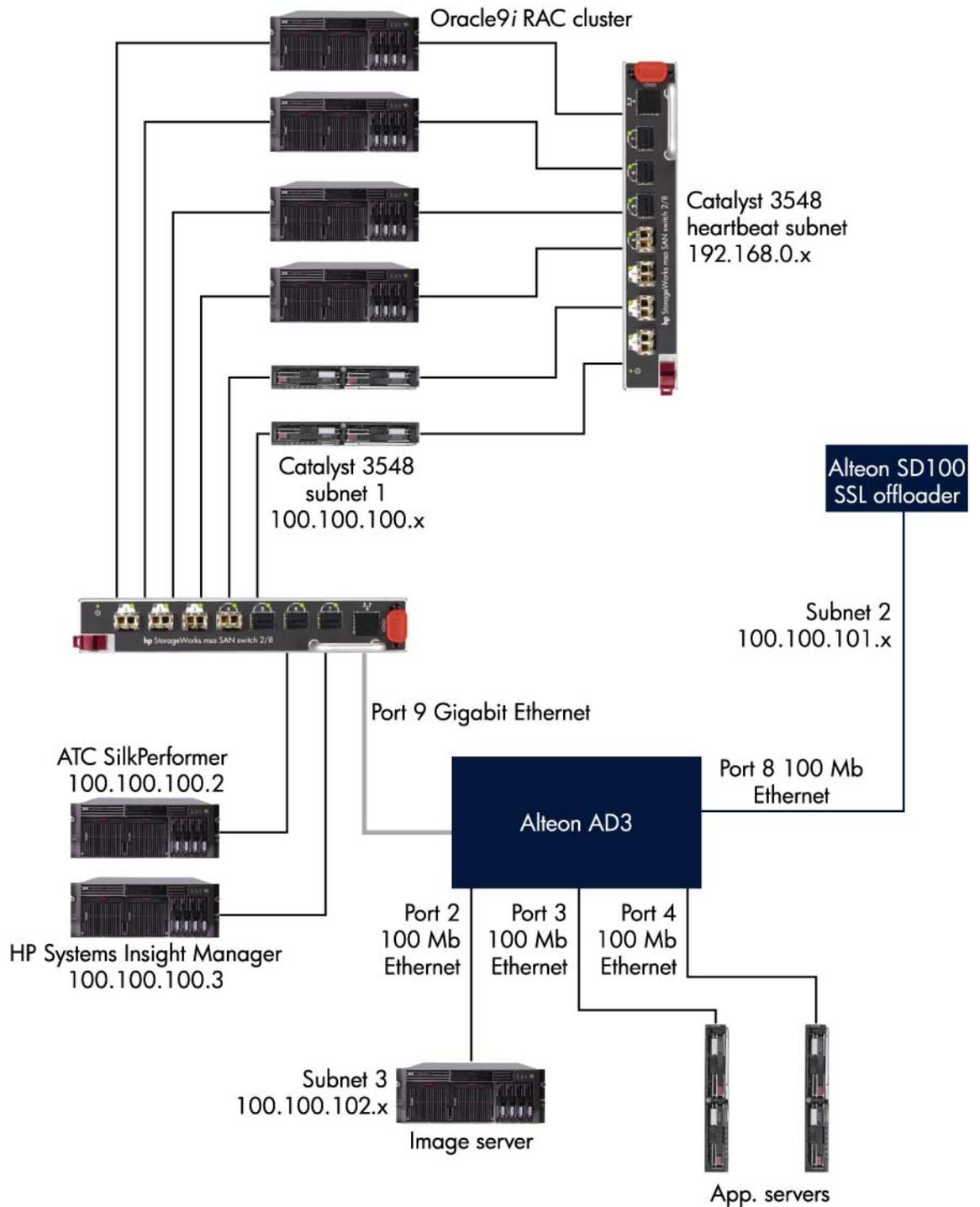
The DL580 does have more internal redundancy (for example, RAM can be mirrored). However, because Oracle9i RAC redundancy can be built up by adding more machines, per-machine advantages in redundancy may not be a major determining factor.

Database-write-intensive and disk I/O-intensive workloads may have different performance characteristics on the configurations tested in this study. However, they are beyond the scope of this study.

Including management tools in the benchmark process helped to isolate problems and evaluate the test results. HP Systems Insight Manager discovered the cluster servers and storage. Events were generated when failures were encountered, enabling quick isolation of hardware problems. Using the information provided from the Insight Management Agents, we were able to verify remotely that our system configurations were consistent and correct.

HP OpenView Performance Manager and Performance Agents provided detailed graphs confirming and explaining the results seen from the Linux command line tools. Using the OpenView performance tools, all operating system data is available and easily correlated and graphed over any given time period. All of the HP management tools used during the benchmark allowed remote access via a browser-based user interface, which was critical for our widely distributed benchmark team.

# Appendix 1: hardware and network schematics



# Appendix 2. SilkPerformer test cases

## Honda Buyer Search

```
//-----  
// Recorded 02/06/2004 by SilkPerformer Recorder v6.0.1.1816  
//-----  
  
benchmark SilkPerformerRecorder  
  
use "WebAPI.bdh"  
  
dcluser  
  user  
    VUser  
  transactions  
    TInit          : begin;  
    TMain          : 1;  
  
var  
fTIME: float;  
dclrand  
  
dcltrans  
  transaction TInit  
  begin  
    WebSetBrowser(WEB_BROWSER_MSIE6);  
    WebModifyHTTPHeader("Accept-Language", "en-us");  
    //WebSetUserBehavior(WEB_USERBEHAVIOR_FIRST_TIME);  
    //WebSetDocumentCache(true, WEB_CACHE_CHECK_SESSION);  
  end TInit;  
  
  transaction TMain  
  var  
  begin  
    fTIME := 5.0;  
    WebPageUrl("http://honda.hp.autc.com/", "Welcome to Honda VIPS");  
  
    // Redirecting https://honda-app.stg.autc.com/login.html  
    // -> https://honda-app.stg.autc.com/honda_home.html  
    ThinkTime(fTIME);  
    //ThinkTime(fTIME);  
    WebPageSubmit("t", FORM001, "Honda VIPS - start"); // Form 1  
  
    ThinkTime(fTIME);  
  
    WebPageLink("Find Vehicles", "Honda VIPS - Find Vehicles"); // Link 3  
  
    ThinkTime(fTIME);  
  
    WebPageSubmit("form1", FORM1002, "Honda VIPS - Search Results"); // Form 1  
  
    ThinkTime(fTIME);  
    WebPageLink("SearchResultsUpArrow_volvo_pl", "Honda VIPS - Search Results (#1)"); //  
Link 14  
  
    ThinkTime(fTIME);  
    WebPageLink("SearchResultsUpArrow_volvo_pl", "Honda VIPS - Search Results (#2)", 2); //  
Link 16  
  
    ThinkTime(fTIME);  
    WebPageLink("Logout", "Honda VIPS - Logout"); // Link 5  
  
    ThinkTime(fTIME);  
    WebPageLink("click here", "Welcome to Honda VIPS (#1)"); // Link 1  
  end TMain;  
  
dclform  
FORM001:  
  "userid"          := "208199_jshort", // changed  
  "password"        := "welcome1", // changed  
  "go.x"            := "0", // added  
  "go.y"            := "0"; // added
```

```

FORM1002:
"cargroup_id" := "153", // changed
"begin_years" := "-1", // changed
"end_years" := "-1", // changed
"make_id" := "-1", // changed
"model_id" := "" <SUPPRESS> , // suppressed, value: ""
"price" := "20000", // changed
"color" := "-1", // changed
"availability" := "-1", // changed
"state" := "5", // changed
"max_delivery_miles" := "-1", // changed
"mileage" := "30000", // changed
"book" := "" <USE_HTML_VAL> , // hidden, unchanged, value: "Fake"
"page" := "" <USE_HTML_VAL> , // hidden, unchanged, value: "1"
"sort" := "" <USE_HTML_VAL> , // hidden, unchanged, value: "5"
"stype" := "" <USE_HTML_VAL> , // hidden, unchanged, value: "r"
"Savey" := "" <USE_HTML_VAL> , // hidden, unchanged, value: "1"
"Savex" := "" <USE_HTML_VAL> , // hidden, unchanged, value: "1"
"Search.x" := "64", // added
"Search.y" := "10"; // added

```

## Subaru Admin

```

//-----
// Recorded 01/05/2004 by SilkPerformer Recorder v6.0.1.1816
//-----

```

```
benchmark SilkPerformerRecorder
```

```
use "WebAPI.bdh"
```

```
dcluser
```

```

user
  VUser
transactions
  TInit      : begin;
  TMain      : 1;

```

```
var
```

```
fTIME: float;
dclrand
```

```
dcltrans
```

```

transaction TInit
begin
  WebSetBrowser(WEB_BROWSER_MSIE6);
  WebModifyHTTPHeader("Accept-Language", "en-us");
  //WebSetUserBehavior(WEB_USERBEHAVIOR_FIRST_TIME);
  //WebSetDocumentCache(true, WEB_CACHE_CHECK_SESSION);
end TInit;

```

```
transaction TMain
```

```
var
begin
```

```

fTIME := 5.0;
ThinkTime(fTIME);
WebPageUrl("http://subaru.hp.autc.com/", "Welcome to SubaruSOLD");

```

```

// Redirecting https://subaru-app.stg.autc.com/login.html
// -> https://subaru-app.stg.autc.com/subaru_home.html

```

```

//ThinkTime(fTIME);
ThinkTime(fTIME);

```

```

WebPageSubmit("t", FORM001, "SubaruSOLD -"); // Form 1
ThinkTime(fTIME);
WebPageLink("Sell", "SubaruSOLD - Vehicle Work List"); // Link 3

```

```

ThinkTime(fTIME);
WebPageLink("Create Vehicle", "SubaruSOLD - Vehicle Work Page"); // Link 8

```

```

ThinkTime(fTIME);
WebPageLink("Release Vehicles", "SubaruSOLD - Vehicles Awaiting Release"); // Link 11

```

```
ThinkTime(fTIME);
```

```

WebPageLink("Administration", "SubaruSOLD - Find An Organization"); // Link 2

ThinkTime(fTIME);
WebPageLink("Post Sales", "SubaruSOLD - Post Sales Report"); // Link 14

ThinkTime(fTIME);
WebPageLink("Send Transport", "SubaruSOLD - Waiting to Send Transport Order"); // Link
13

ThinkTime(fTIME);
WebPageLink("Transport Dropoff", "SubaruSOLD - Waiting to Receive Drop Off
Confirmation"); // Link 15

ThinkTime(fTIME);
WebPageLink("Receive Payment", "SubaruSOLD - Waiting to Receive Payment from Buyer"); //
Link 19

ThinkTime(fTIME);
WebPageLink("Finished", "SubaruSOLD - Finished"); // Link 24

ThinkTime(fTIME);
WebPageLink("Messages", "SubaruSOLD - Messages"); // Link 8

ThinkTime(fTIME);
WebPageLink("Logout", "SubaruSOLD - Logout"); // Link 5
ThinkTime(fTIME);
WebPageLink("click here", "Welcome to SubaruSOLD (#1)"); // Link 1
end TMain;

dclform
FORM001:
  "userid"           := "zhallowell-a", // changed
  "password"        := "welcome1", // changed
  "go.x"            := "0", // added
  "go.y"            := "0"; // added

```

## Appendix 3. HP Systems Insight Management Agents

HP Systems Insight Manager allows the user to drill down to the devices being managed. The HP Systems Insight Management Agents running on each server provide detailed subsystem information to the user. Figure 20 shows the links provided by each subsystem that allow users to drill down to the detailed information available for a ProLiant DL580 G2 cluster. Each server in the cluster can see the shared storage. Drilling down from the “Storage System ATC (1)” link brings up the storage system information.

Figure 20. System Management Homepage for DL580 cluster

The screenshot displays the HP System Management Homepage for a ProLiant DL580 G2 cluster. The browser window shows the URL <https://racnode1:2381/77103>. The page header includes the HP logo, the title "System Management Homepage for racnode1", and system information: System Model: ProLiant DL580 G2, System Status: OK, and Current User: administrator. Navigation tabs for Home, Settings, Tasks, and Logs are visible. A left sidebar lists various components like Integrated Agents, Other Agents, Management Processor, and Other Software. The main content area features a "Failed & Degraded Items" section which is currently empty. Below this, there are several subsystem status boxes, each with a green checkmark indicating they are OK. These include: 

- NIC**: HP NC7131 Gigabit Server Adapter in Slot 2, HP NC7770 Gigabit Server Adapter in Slot 6, and a link to Virtual interface 1.
- Recovery**: Autorecovery, Environment, Power converter, Power Supply, and Remote Communications.
- Storage**: External Storage Connections, Smart Array 5i Controller on System Board, Standard IDE Controller on System Board, Storage System ATC (1), and Floppy Drives.
- System**: System Board, File System Usage, Security, Software Version, and Fan Boards.

 A "Utilization" section for Processor Utilization is also present. The footer of the page includes "HTTP Server v5.91.0" and "© 2002, 2003 Hewlett-Packard Development Company, L.P.".

Figure 21. ProLiant DL580 G2 cluster storage subsystem information

The screenshot shows the HP System Management Homepage for 'racnode1'. The main content area is titled 'Storage System Information' and contains the following details:

**Storage System ATC (1)**

- Model: StorageWorks Modular SAN Array 1000
- Name: ATC
- Connection: Fibre Attached
- IO Slot 1: Fibre Array
- IO Slot 2: Not Installed
- Serial Number: 9J1CJN71A64P
- IO Module Type: 2 Port Ultra3 SCSI Module

**Internal Storage System Bus 1**

- Vendor: COMPAQ
- Model: PROLIANT 4L71 DB
- Serial Number: 9J1CJN71A64P
- Firmware Revision: 1.86
- Board Revision: 0
- Version: 5
- Placement: Internal
- Temperature Status: OK
- Fan Status: OK
- Power Supply Status: OK
- Speed: Ultra3
- Drive Bays: 7
- Duplex Option: Duplex Bottom

**Internal Storage System Bus 2**

- Vendor: COMPAQ
- Model: PROLIANT 4L71 DT
- Serial Number: 9J1CJN71A64P
- Firmware Revision: 1.86
- Placement: Internal
- Board Revision: 0
- Version: 5
- Speed: Ultra3
- Drive Bays: 7
- Duplex Option: Duplex Top

**External Storage System Bus 3**

- Vendor: COMPAQ
- Model: PROLIANT 4LEE
- Serial Number: 9J36FLW1JSTZ
- Firmware Revision: JB4H
- Board Revision: 32
- Temperature Status: OK
- Fan Status: OK
- Power Supply Status: Not Redundant
- Speed: Ultra3
- Drive Bays: 14

The left sidebar shows a tree view for 'Storage System ATC (1)' with a list of physical drives (Part 1 Drive 1 through Part 3 Drive 1) and their capacities (e.g., 34732 MB).

From the System Management Homepage (Figure 20) the user can drill down on a NIC link to view detailed status of their Ethernet connections. (Figure 22)

Figure 22. ProLiant DL580 G2 cluster NIC subsystem information

The screenshot shows the HP System Management Homepage for a ProLiant DL580 G2 server. The browser address bar shows the URL: https://racnode1.2381/77131. The page title is "System Management Homepage for racnode1". The current user is "administrator" and the system status is "OK".

The main content area displays "HP NC7131 Gigabit Server Adapter in Slot 2". Below this, there are two sections:

#### ETHERNET-CSMACD Controller Information

|                       |                                  |                          |                   |
|-----------------------|----------------------------------|--------------------------|-------------------|
| <b>Model:</b>         | HP NC7131 Gigabit Server Adapter | <b>Role:</b>             | Primary           |
| <b>Status:</b>        | OK                               | <b>Port:</b>             | 1                 |
| <b>Slot:</b>          | 2                                | <b>Base Mem Addr:</b>    | F7FE0000h         |
| <b>Duplex:</b>        | Full                             | <b>DMA Channel:</b>      | N/A               |
| <b>Base I/O Addr:</b> | N/A                              | <b>Physical Address:</b> | 00 02 A5 42 1F AC |
| <b>IRQ:</b>           | 3                                |                          |                   |

| Receive Statistics |         | Transmit Statistics  |         |
|--------------------|---------|----------------------|---------|
| Good Frames:       | 9074251 | Good Frames:         | 9074631 |
| Bad Frames:        | 0       | Bad Frames:          | 8975522 |
| Alignment Errors:  | 0       | Carrier Sense Err:   | 8975522 |
| Frame Too Long:    | 0       | Multi. Coll. Frames: | 0       |
| FCS Errors:        | 0       | Late Collisions:     | 0       |
| MAC Recv Errs:     | 0       | Single Coll. Frames: | 0       |
|                    |         | Excessive Coll.:     | 0       |
|                    |         | Deferred Xmits:      | 0       |
|                    |         | MAC Xmit Errs:       | 0       |

#### Interface Information - eth1

|                          |                 |                            |                                 |
|--------------------------|-----------------|----------------------------|---------------------------------|
| <b>IP Address:</b>       | 192.168.0.1     | <b>Speed:</b>              | 1000 Mbit/Sec                   |
| <b>Type:</b>             | ETHERNET-CSMACD | <b>Physical Addr:</b>      | 00 02 A5 42 1F AC               |
| <b>Interface Status:</b> | OK              | <b>Last Status Change:</b> | 4 Days, 2 Hours, 20 Minutes ago |
| <b>Max Packet Size:</b>  | 1500 bytes      |                            |                                 |

| Receive Statistics   |            | Transmit Statistics  |            |
|----------------------|------------|----------------------|------------|
| Bytes:               | 1166938762 | Bytes:               | 1065927570 |
| Total Packets:       | 9074251    | Total Packets:       | 9074631    |
| Unicast Packets:     | 9074251    | Unicast Packets:     | 9074631    |
| Non-Unicast Packets: | N/A        | Non-Unicast Packets: | N/A        |

## For more information

To learn more about HP ProLiant clusters, contact an HP sales representative or visit:  
[www.hp.com/solutions/highavailability/oracle](http://www.hp.com/solutions/highavailability/oracle)

© 2004 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Intel and Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. Java is a U.S. trademark of Sun Microsystems, Inc. Linux is a U.S. registered trademark of Linus Torvalds. Microsoft and Windows are U.S. registered trademarks of Microsoft Corporation. Oracle is a registered U.S. trademark of Oracle Corporation, Redwood City, California.

5982-7009EN, 06/2004

