LPAR Heterogeneous Workloads on the IBM® eServer pSeries™ 690 System

System performance comparisons and observations of heterogeneous workloads in logical partitions on an IBM eServer pSeries 690 system

Kathleen DeLira, Antonio Garcia, Ric Hendrickson, Lucila Macedo, Raj Patel

IBM AIX Performance and Solutions, Austin TX

February 13, 2002

Abstract

This paper describes a server consolidation experience of combining a variety of significant workloads onto a single large IBM® eServer pSeries™ 690 (p690) system in multiple independent logical partitions (LPARs) running differing levels of AIX® 5.1 and Linux for Power. The LPAR capability of the p690 system provides tremendous flexibility in tailoring each LPAR's configuration to meet the needs of the workload running in the separate LPARs. The paper describes some of the considerations and experiences setting up and managing the LPARs, and the subsequent summary comparisons of each workload running by itself on the system against those workloads executing on a fully loaded system. The experience of the server consolidation testing demonstrated that the LPARs had little affect on each other, allowing production systems to run side by side with LPARs running test scenarios. WebSphere® Application Server, DB2®, HTTP web servers, and an Oracle database were run across the fully configured p690 system.

Server Consolidation

Our testing of the IBM eServer pSeries 690 (referred to here as the p690 system) High Performance Computing (HPC) model was focused on the server consolidation scenarios made available with logical partitioning (LPAR) technology provided in AIX 5L. Server consolidation enables customers to run different AIX levels, application code levels, and workloads (for example, OLTP, Web Serving, and DB2) while using varied resource configurations (memory, CPU, I/O, etc.). With logical partitioning we are able to create multiple partitions within a single server which operate as individual servers. The system under test was a 16-way SMP system with the 1.0 GHz POWER4 processors, with four I/O drawers and 64 GB memory. This system was an early test system which will be updated to "general availability" level with the 1.3 GHz processors in the near future.

For this project, we relied on several key documents available for the p690 system:

- *IBM eServer pSeries 690 System Handbook* (IBM Redbook - SG24-7070)
- *IBM Hardware Management Console for pSeries Operations Guide*
- *IBM eServer POWER4 System Architecture*
- *IBM Partitioning for the IBM eServer pSeries 690 System*
- *IBM eServer pSeries 690 Configuring for Performance*

Included with the p690 system is the IBM Hardware Management Console (HMC) for pSeries. The HMC is the central point for your system management of the p690 system and LPAR. It offers the flexibility of utilizing the p690 as a single server (full partition mode) or as a multiple partitioned server (partition standby mode). It performs several functions including creating and maintaining partitions, power control over partitions as well as the managed system, and providing virtual terminals for working within partitions. It also monitors and stores hardware changes while serving as a focal point for service representatives.

Project Summary

The project produced some specific comparisons of running workloads individually in a partition and then again with the overall system running loaded. Workloads were selected to provide a mix of web serving network driven I/O, together with a heavy Java® workload driving DB2 database transactions with multiple servers connected to the database, and finally an Oracle-

based OLTP workload set intentionally with high I/O wait times simulating a partition with too many CPUs driving an inadequate number of disk drives and adapters.

WebSphere Application Server (WAS) testing was done using the internally developed Trade 2 benchmark workload driving a single DB2 back-end database. This test demonstrated that the p690 system supported running the servers in AIX and Linux partitions with small differences in the overall throughput.

The web serving workload demonstrated that a standard web serving test run with numerous external clients driving web traffic workload on a server in an LPAR demonstrated little difference in measurements when run by itself and again with significant system workload in the other LPARs.

The Oracle OLTP workload demonstrated that running other heterogeneous LPAR workloads have negligible impact on I/O intensive workloads.

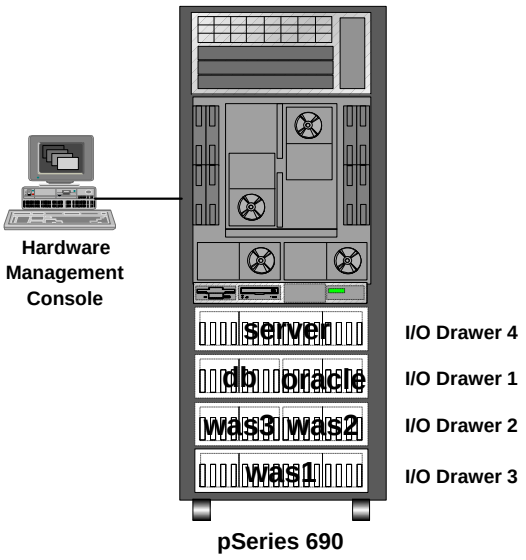
Overall, the project demonstrated that a mix of workloads, including CPU intensive, network intensive, and I/O intensive workloads could be run effectively and flexibly on the new pSeries 690 system from IBM in separate LPARs. We anticipate there may be special cases of applications being run in separate LPARs where the performance characteristics may be affected, but this project with numerous and varied workloads demonstrated no significant impacts.

Planning Considerations

When utilizing the p690 system as a multiple partition server, there are several considerations to be made. We decided to test several varied workload scenarios; OLTP with Oracle 8.1.7, WebSphere Application Server Advanced Edition V4.0.2 with DB2 Enterprise Edition 7.0.1, and a standard Web Server workload. Our hardware resource determinations were based on the application requirements as well as the workload. It is recommended that you refer to The IBM Hardware Management Console for pSeries Operations Guide, which offers several guidelines to assist you in the planning stages.

Our LPARs were configured with the following:

LPAR	Use	I/O Drawers	CPUs	Memory
server	Web Server	U1.13-P1 (Drawer4)	2	32 GB
was1	WebSphere Application Server	U1.1-P1 (Drawer 3)	4	4 GB
was2	WebSphere Application Server	U1.5-P1 (Drawer 2)	2	4 GB
was3	WebSphere Application Server	U1.5-P2 (Drawer 2)	2	4 GB
db	DB2 backend	U1.9-P2 (Drawer 1)	2	4 GB
oracle	Oracle backend	U1.9-P1 (Drawer1)	4	8 GB



Once you've powered on the managed system in Partition Standby mode, the process of creating an LPAR is very straight forward. First select your managed system from the Contents area, then select Create from the submenu. You will then select Logical Partition. From this point the "Create Logical Partition and Profile" wizard will guide you through several configuration screens. Please refer to the Hardware Management Console for pSeries Operations Guide for explicit details on this process. Keep in mind all LPAR's must have a minimum configuration of one processor, 1 GB memory and one boot device with associated adapter. The final step is activating the partition and installing an operating system.

Note also that there is memory overhead associated with partitioning a system, so not all of the 64 GB of available memory can be directly configured and allocated to the partitions. The p690 planning guides for partitioning describe the overhead associated with the hypervisor and partition support.

Managing Logical Partition Considerations

Keeping track of your system's resources is vital on a multiple partitioned server. There will be instances when an LPAR may require additional resources, if you exceed the amount of available resources in the system you will not be able to activate the partition. This is true also when selecting "desired" resources within your LPARs. If you activate a partition containing "desired" resources that are "required" resources on another LPAR, the first LPAR to be activated will boot with those resources. This could result in one LPAR not having the resources required to be activated. Resources cannot be shared between LPARs. To release a resource requires the LPAR to be shutdown and deactivated, only then can you move the resource. This can be a sticky situation when you are unable to take down an LPAR when “something key” is running there.

However, the "desired" and "required" resource settings do offer flexibility within the LPAR. For example, you have "required" a single processor but "desire" four processors. Your LPAR will activate with however many processors are available between the 1 required and 4 desired. But again, this may affect other LPARs if the amount of available resources has been exceeded.

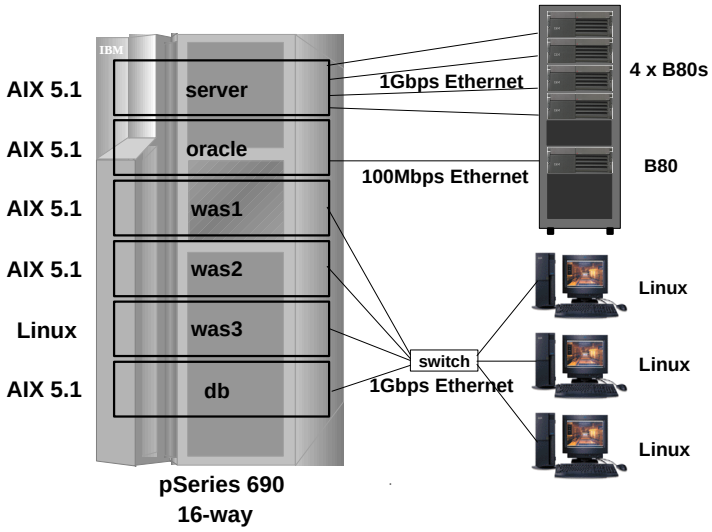
We found it helpful during the initial installation to select the CDROM resource as "desired" in all the LPARs. Since we installed one LPAR at a time the CDROM resource was always available. When the installations were complete we removed the CDROM resource from all the LPARs. This allowed us the ability to move the CDROM resource between LPARs.

Due to the memory overhead associated with partitioning your system, there are considerations to be made prior to activating your partitions. Partitions with greater than 16 GB should be activated first. If you start the partition with greater than 16 GB as the last partition, it may not have enough resources remaining in the system to start.

When activating your partition you can select to open a terminal window. When booting AIX, the terminal display will appear the same as any other AIX system. You will see the AIX banner and the boot method you selected when creating the LPAR. On the HMC you will also see the boot LED's scroll across the status field associated with the booting LPAR . Once the LPAR is up you will see a login prompt on the terminal window and the LPAR status field will read "Running". The terminal windows are generally intended to be used for initial setup/install, firmware menu access, diagnostics, debug, and other limited-use scenarios. For normal operations, the use of network-based mechanisms (telnet, etc.) to access the partitioned operating system is recommended.

When creating your partitions it is best to have all of your adapters in place prior to installing the AIX operating system. In this manner all devices will be configured and the drivers installed accordingly, which has been the normal approach for AIX. However, you may determine later an additional adapter is necessary to your partition. Managing adapters within your LPAR can be performed in AIX through the PCI Hot Plug manager available in SMIT. From here you can remove, replace, add and locate adapters within your LPAR just as you would on a standalone system. Please consult the IBM eServer pSeries 690 System Handbook for additional information.

Test Environment



The systems were configured with the following environments for web serving, and Oracle database, and three WebSphere servers, and a DB2 database.

The six partitions each had network connections defined to attach to external systems which drove the workloads.

The web server partition was assigned four Gigabit network adapters which were each connected directly to peer pSeries B80 systems.

The Oracle partition was connected directly to another B80 system via a 100 Mbps Ethernet connection.

The WebSphere partitions and DB2 database and Linux client systems were connected via a Gigabit switch.

The specific configurations of each partition are detailed next.

WebSphere Application Server / Trade 2 Workload

Level	Model	Processors	Memory	OS
Tier 1	2 x IBM Netfinity 6000R - Drivers 1 x IBM Netfinity 4000R - Driver	4 x 700 MHz Intel 2 x 650 MHz Intel	8 GB 4 GB	Linux RedHat 6.2
Tier 2	IBM eServer pSeries 690 - WAS 1 LPAR IBM eServer pSeries 690 - WAS 2 LPAR IBM eServer pSeries 690 - WAS 3 LPAR	4 x 1 GHz Power4 2 x 1 GHz Power4 2 x 1 GHz Power4	4 GB 4 GB 4 GB	AIX 5.1 AIX 5.1 SuSe Linux 7.1
Tier 3	IBM eServer pSeries 690 - DB2 LPAR	2 x 1 GHz Power4	4 GB	AIX 5.1

Network communication was achieved using Gigabit Ethernet routed through an Alteon gigabit switch. Each of the three WebSphere partitions was allocated a Gigabit Ethernet adapter connected separately to the switch.

Standard Web Server Workload

Level	Model	Processors	Memory	OS
Tier 1	4 x IBM eServer pSeries 640 Model B80 - Clients	4 x 375 MHz	4 GB	AIX 4.3.3
Tier 2	IBM eServer pSeries 690 - Server LPAR	2 x 1 GHz Power4	32 GB	AIX 5.1

Network communication was achieved using Gigabit Ethernet peer to peer connection between the partition on the p690 and the standalone B80 client system.

OLTP Workload

Level	Model	Processors	Memory	OS
Tier 1	IBM eServer pSeries 640 Model B80 - Client	4 x 375 MHz	4 GB	AIX 5.1
Tier 2	IBM eServer pSeries 690 - Oracle LPAR	4 x 1 GHz Power4	8 GB	AIX 5.1

Network communication was achieved using a 10/100 BaseT Ethernet peer to peer connection between the p690 LPAR and the standalone B80 client system. Network communications were minimal considerations since the B80 client attached via remote shell (rsh).

Setting Up the Web Server and Clients

The Web Server LPAR

The standard Web Server workload was installed in a logical partition on the p690. This LPAR was configured with two processors and 32 GB memory. The partition was installed with the latest AIX 5.1 level. Four Gigabit Ethernet adapters were configured, one for each of the web

servers which would be installed in this partition. Four external B80 model systems were used as clients to drive the respective web servers.

Our workload was configured using methods consistent with Commercial Web servers. It is generated with a predetermined number of connections performing dynamic gets, custom ad rotations, and posts, while maintaining specific throughput and error rate requirements. These connections must also sustain a specified maximum bit rate and segment size. Failure to conform to the connection specifications would result in a failed run.

The following attributes were set on the Gigabit Ethernet adapters:

copy_bytes=256 (copy packet into contiguous buffer on transmit if greater or less than)
large_send=yes (turns on the large send capability of TCPIP)

The following network tuning options were set:

no -o tcp_timewait=5	(ensures TIME_WAIT is at least 60 seconds)
no -o send_file_duration=100000	(cached send_file duration time)
no -o nbc_pseg=200000	(max file entries in private segment)
no -o nbc_max_cache=49060	(max file size for network buffer cache)

The Web Server Clients

Our clients consisted of four IBM eServer pSeries 640 Model B80s. Each of the client systems had four processors and 4 GB memory. Each client was installed with AIX 4.3.3 and the client workload to drive the associated web servers in the p690 LPAR. Each B80 system contained a Gigabit Ethernet adapter and was configured with peer-to-peer connections to the p690 system. The following attribute was set on the Gigabit Ethernet adapters:

copy_bytes=256 (copy packet into contiguous buffer on transmit if greater or less than)

The following network tuning options were set:

no -o tcp_timewait=5	(ensures TIME_WAIT is at least 60 seconds)
no -o delayack=3	(delay ack for connection setup and shutdown)
no -o delayackports={80}	(delay ack ports)

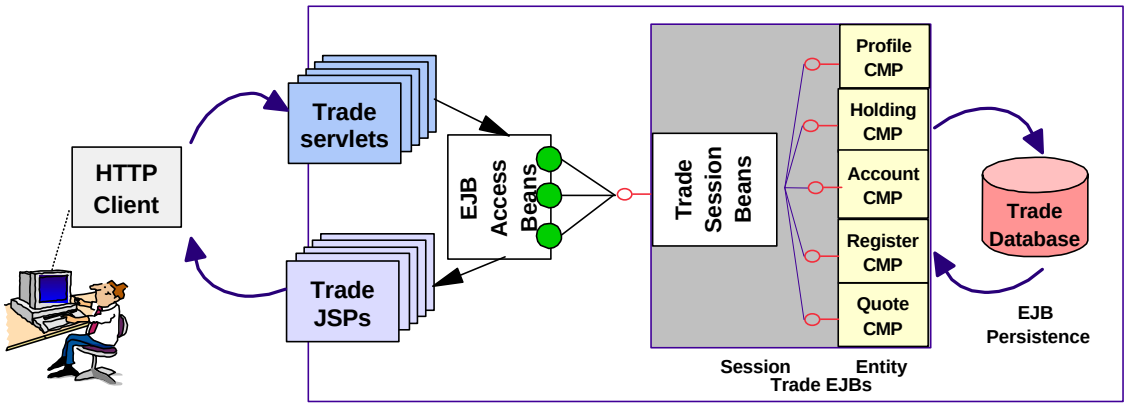
Setting Up the Trade 2 WebSphere Benchmark

The Trade 2 benchmark, also called the WebSphere Performance Benchmark Sample, is a sample benchmark that measures the performance of servers running the IBM WebSphere Application Server software. Trade 2 was built to emulate an online brokerage firm. This workload exercises the entire solution stack that consists of the WebSphere Application Server, the Java Virtual Machine (JVM) and the Just-In-Time (JIT) compiler, the HTTP server, the DB2 Database Server and the DB2 client, the AIX operating system, and the system hardware.

The WebSphere Performance Benchmark Sample is available at:
http://www.ibm.com/software/webservers/appserv/wpbs_download.html

The Trade 2 application is a collection of Java classes, Java servlets, Java Server Pages and Enterprise Java Beans (EJBs) that service requests made by registered users. This application runs as a single Java process which is managed by the WebSphere Application Server.

The following chart demonstrates the flow of information that occurs once a user makes a request of the WebServer.



Trade 2 performance is measured by how many requests the server can handle per second (throughput). This is accomplished by using a HTTP request generator to simulate multiple users making hundreds of requests to the server.

For the purposes of this test, the internally developed "AKstress" tool was used to generate the requests using a Trade 2 supplied script for the user simulation. The Trade 2 servlet on the WebSphere server has a function where it generates an access to the server. AKstress generates hits to this servlet via web requests and then generates a report based on the results it received from the server.

System Configuration - AIX LPAR

The test systems were setup as a normal three-tier environment. Tier 2 consisted of three WebSphere Application Server LPARs. Tier 3 contains the DB2 database server running in a fourth LPAR. Tier 1 (the clients) consists of the external systems used to drive the p690 system.

For the purposes of demonstrating a mixed operating system environment, the LPARs were configured differently. Two of the WebSphere Application Server LPARs were installed with the latest AIX 5.1. One of these LPARs was configured as a 4-way SMP configuration, the other as a 2-way SMP configuration. The third LPAR was installed with Linux. All three were configured with 4 gigabytes of memory for each LPAR. Each LPAR had one SCSI disk and four SSA disks connected. The SCSI disk was used only for the operating system and software. Certain standard AIX environment variables are also set in order to improve performance of the system.

```
AIXTHREAD_COND_DEBUG=OFF
AIXTHREAD_MUTEX_DEBUG=OFF
AIXTHREAD_RWLOCK_DEBUG=OFF
AIXTHREAD_SCOPE=S
```

Several drivers were used to drive the WebSphere Application servers. The driver used on two of the systems were IBM Netfinity® 6000R 4-way SMP Intel-based machines running the Linux Red Hat 6.2 operating system. Each AKstress system running on the Linux drivers to simulate 100 users (100 threads) was used for testing the EJB (Enterprise Java Beans) method of accessing the DB2 database.

Each WebSphere Application Server uses its own configuration database stored on the DB2 server and accesses the same Trade 2 database. The DB2 database LPAR is running the AIX 5.1 operating system and uses DB2 v 7.01 Fixpack 5.

System Configuration - Linux for PowerPC LPAR

The setup for the Linux LPAR was similar to the AIX LPARs. At the time of writing this article, the p690 support for Linux in a partition is available only as a technology preview. Even so, we were particularly interested in using this in the LPAR environment as part of the varied workload and experienced no problems with the approach.

The WebSphere Application Server LPAR running Linux runs as a 2-way SMP system configured with 4 gigabytes of memory. The Linux LPAR was running SuSE 7.1 with a 2.4.13 64-bit kernel. At this time, SuSE does not provide support for their Linux distributions running in a partition on the p690 system.

To test Linux running in an LPAR, the team took an already running Linux disk from a pSeries 640 model B80 system and placed it in the p690 partition. Another choice for a system administrator would be to simply install the SuSE SLES 7 distribution of Linux directly onto the disk on the p690 system when that support is available.

Before moving the Linux disk from the B80 to the p690, the hypervisor virtual console was enabled. This is accomplished by updating the Linux configuration files on the B80 before removing it and putting it in the p690. This was done by adding *hvc0* to */etc/securetty*, spawning a getty on it from */etc/inittab*, and making the node for the *hvc0* device, which would be "seen" by the HMC when that Linux was booted in a partition. While SuSE does not formally support the Linux distribution running in a partition at this time, the SLES 7 distribution should now recognize the partition at boot time and handle defining the *hvc0* device automatically.

Once the Linux operating system disk migration was complete, internal test copies of the DB2 Administration Client Ver 7.01 Fixpack 5, the WebSphere Application Server Version 4.0.2, IBM HTTP Server 1.3.19 and Trade 2 were installed on the Linux disk. These IBM products are currently not generally available for Linux on pSeries, but we were able to run these in our test setup.

For information on the SuSE Linux Enterprise Server (SLES) 7 see:
http://www.suse.com/us/products/suse_business/sles/sles_iSeries_pSeries/index.html

DB2 LPAR Configuration

DB2 resides in its own LPAR. This LPAR was configured with eight SSA drives. DB2 was configured to use four SSA drives spread across two SSA loops. All drives are part of the same volume group (*db2vg*). The log for the *db2vg* volume group and the filesystem, */db2* for storing the

databases, are located on the first drive. The other three drives are used as a raw logical volume for the DB2 logs and are on a different SSA loop from the first drive. The data on these two drives is striped across the drives for better performance.

The normal rule of thumb for setting up the Trade 2 database is 500 defined users for 100 active users. Following this rule, the Trade 2 database size was increased to 1500 users for use by the three WebSphere servers, each handling 100 active users.

DB2 has a number of configuration settings for each database, as well as the database manager instance. The following settings were applied to the trade database:

applheapsz 256
buffpage 40000
maxappls 256
avg_appls 256
maxlocks 75
logbufsz 512

WebSphere - AIX and Linux LPARs

In order to achieve the best results, the Prepared statement cache for the Trade 2 database must be set. For the purposes of our test, we have set the statement cache to 1000 on each LPAR WebSphere server. The number of WebSphere container threads is set to 50 and the number of DB2 threads is set to 10. The Pass-by-Reference option was also set.

IBM HTTP Server - AIX and Linux LPARs

The IBM HTTP Server (IHS) Version 1.3.19 was used as the web server. This server was used in both the AIX LPARs and the Linux LPAR. The installation procedure for the WebSphere Application Server provides an option to select the plug-in module for this web server. The only changes made to the *httpd.conf* file for these tests are:

ServerName <hostname>
MinSpareServers 5
MaxSpareServers 107
StartServers 107

Setting the number of server processes is not recommended for customer environments. The only reason for setting them here is to maintain a constant number of *httpd* processes between runs as we know that we will be hitting the server with 100 connections (one *httpd* process for each connection).

Trade 2 Setup Considerations

For the tests, Trade 2 servlet was run with Java min and max heap size of 512 M. The stack size is also set to 1024 K allowing for a greater number of instructions to be placed on the stack for execution.

For the purposes of this test, the option "No Register" is set when configuring the Trade 2 Database. This tells Trade 2 to perform all database operations, but not actually updating the database. This allows us to control the size of the database during a long run.

Trade 2 Test Procedure

The Trade 2 benchmark functions in several different modes of operations. For the purposes of this test, only the EJB mode was used. For the EJB mode, Trade 2 uses the VisualAge for Java EJB access beans to access the trade database.

The EJB mode accepts a request from the user and then handle that request. Several actions must then occur. The request is made by a end-user, using a web browser to makes the selections and submits the request. The request is handled by the HTTP web server which hands it over to the WebSphere Application Server. The WebSphere Application Server then performs the action, usually accessing and updating the Database. Then returns its response to the HTTP web server and finally returns to the end-user.

The goal of these test is to measure the performance of several LPAR systems under stressful situations. For a baseline, the systems were run with only the three WebSphere server LPARs and the DB2 LPAR. Each were run for 2½ hours. Once everything was set, all LPARs were run in tandem and the performance difference between the two was only about 2%.

Setting Up the Oracle 8.1.7 OLTP Test

The Oracle 8.1.7 OLTP test is representative of an enterprise back-end office environment supporting an order entry system.

Papers under development for the future includes one which describes the process of migrating an Oracle 8.1.7 database from an AIX 5.1 system to a logical partition on the p690 system. For examples of other LPAR and p690 related white papers see the web site at:
http://www.ibm.com/servers/eserver/pseries/hardware/datactr

The Oracle database OLTP instance was setup using raw logical devices which offers performance improvements over JFS file systems. Each logical volume containing datafiles were placed on separate disks. The more I/O intensive datafiles were placed on the outer edge portion of the disk across multiple volumes. No AIX striping was used. Each Oracle redo log file was striped across four disks using SSA raid adapters (raid-0). Six SSA adapters were used. Each adapter port (A/B loop) was connected to a single disk drawer (containing 16 drives each). Each logical volume that was spread across multiple drives was contained within a single SSA loop. An SSA adapter was dedicated to the redo logs. Fast-write cache was turned on for the redo logs and selected disks with heavy I/O write activity.

The setup of the OLTP consisted of the following high-level steps:

1. Create, configure and activate an LPAR profile identifying required system resources
   - 4 processors, 8 GB memory, allocation of 10 PCI ports
2. Connect and configure disk shelves to SSA adapters
   - each SSA loop (A/B) was attached to a single drawer (16 drives per drawer)

3. Install and configure the latest AIX 5.1 operating system and latest AIX file set
   - external interface (ent0) and internal peer to peer interface (ent1)
4. Install Oracle 8.1.7 32-bit database and 8.1.7.2.0 latest patch
5. Create Oracle 8.1.7 OLTP database instance and generate workload data
6. Configure OLTP Client Driver
   - application server to send SQL transactions to Oracle LPAR database server

OLTP Workload Measurements

The OLTP workloads were initially run on an unloaded p690 system, and then again on a fully loaded system running the heterogeneous workloads. The measurements selected to use for the runs were the simple average number of transactions and response time processed in each run. **vmstat** and **iostat** statistics captured the normal system load characteristics during the runs. The focus of the runs was to compare the statistics with and without significant system workload.

Project Results Running Full System Load

Each of the workloads were run individually with no other system load to establish baseline measurements for comparison. Then all of the workloads were run simultaneously putting significant overall system demand on the 16-way SMP system. The same measurements for each workload were taken again and compared against the baseline measurements. For these workloads, we did not see any significant affect between the runs.

None of the workloads in each partition were specifically tuned for optimal use. The focus of the project efforts were on simultaneously running a wide variety of applications, middleware, and databases. The numbers gathered are not intended be "over-analyzed" or representative of a workload tuned for proper use and effective resource usage.

WebSphere Application Server with DB2 Database

The pages per second (throughput) for the three WebSphere Application Server LPARs remained consistent. The difference between running in a single workload base and a multiple workload comparison was less than 2% for each of the LPARs. The system load for the four LPARs during the runs had characteristics of:

LPAR	usr	sys
WebSphere #1	71% - 75%	25% - 29%
WebSphere #2	68% - 71%	29% - 32%
WebSphere #3	75% - 78%	22% - 25%
DB2 database	13% - 17%	18% - 23%

The data gathered for DB2 using **vmstat** and **iostat** did not show any significant variance and showed that the DB2 back-end remained consistent between the single workload base and the multiple workload comparison. The DB2 LPAR had CPU idle 55% - 60% and 4% - 7% iowait.

Standard Web Server Workload

Our web server workload was generated with 1000 connections at once across the four clients (250 connections per client).

The multiple workload comparison run was also initialized with the same parameters, 1000 connections across the same four clients (250 connections per client). This run completed successfully as well, again meeting all specifications and requirements of the workload.

The system load for the web server workload:

LPAR	usr	sys
Web Server	9% - 13%	80% - 85%

The **vmstat** and **iostat** measurements taken show less than 3% variance in CPU utilization and consistent disk usage between these two workload runs.

OLTP Workload

The delta differences between the unloaded OLTP test and the loaded OLTP was almost negligible, between .3% and .5% for the transactions per minute rate and average response time. **vmstat** measurements were also minimal with identical low range values and less than a 1% difference for high range values. SSA adapter throughput decreased less than 1%.

The system load for the Oracle OLTP workload was:

LPAR	usr	sys
Oracle database	41% - 45%	21% - 25%

Idle percentages were consistently less than 4% - 6%. Iowait for this scenario was 30% - 38% . The workload was specifically defined to cause a high iowait condition by overloading the number of users. The difference in iowait between the two runs was negligible.

Summary

This project successfully utilized a good variety of application and server workloads to demonstrate the affects of an overall heterogeneous workload running on an early 16-way p690 HPC system with 64 GB of memory. Each of the workloads ran essentially the same with and without a significant workload on the overall system.

We found the LPAR management support on the p690 easy to use which provided great flexibility in day-to-day management of the overall system. It was easy to shift CPUs, memory, network connections, and disks from one partition to another as we worked with varying configurations and testing. The team quickly determined that other workloads running in other partitions had little affect on “their” partition, which reinforced the notion of having many separate servers running on a single platform.

A variety of configurations were intentionally used to demonstrate differing characteristics of LPAR support. Partitions were defined and redefined with varying CPU processors, network connections, memory, and disk configurations. The AIX partitions ran numerous levels of AIX 5.1, including early test versions, the shipping copy, and subsequent levels with different fixes applied. An early Linux version was run in a partition which ran WebSphere alongside the AIX 5.1 partitions. The project reinforced the ability of the p690 system and AIX 5.1 to flexibly and easily deploy independent levels of operating systems in the separate LPARs.

The overall project had several independent teams running together in parallel over the course of the project. The paper discusses some of the experiences the team had with managing the p690 system and the partitions. Most of the time and energy of the teams was focused on working with their particular workloads, with little additional time required to run and support and manage the overall p690 system.

References

1. IBM eServer pSeries 690 System Handbook (IBM Redbook - SG24-7040)
2. IBM Hardware Management Console for pSeries Operations Guide (IBM Manual - SA38-0590-00). Related white papers and technical reports: http://www-1.ibm.com/servers/eserver/pseries/library/wp_systems.html
3. POWER4 System Microarchitecture (white paper): http://www-1.ibm.com/servers/eserver/pseries/hardware/whitepapers/power4.html
4. IBM Partitioning for the IBM eServer pSeries 690 System (white paper): http://www-1.ibm.com/servers/eserver/pseries/hardware/whitepapers/lpar.html
5. IBM eServer pSeries 690 Configuring for Performance (white paper): http://www-1.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_config.html

Special Notices

The information contained in this document represents the current views of IBM on the issues discussed as of the date of publication. IBM cannot guarantee the accuracy of any information presented after the date of publication.

All prices shown are IBM's suggested list prices; dealer prices may vary.

IBM products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Information provided in this document and information contained on IBM's past and present Year 2000 Internet Web site pages regarding products and services offered by IBM and its subsidiaries are "Year 2000 Readiness Disclosures" under the Year 2000 Information and Readiness Disclosure Act of 1998, a U.S statute enacted on October 19, 1998. IBM's Year 2000 Internet Web site pages have been and will continue to be our primary mechanism for communicating year 2000 information. Please see the "legal" icon on IBM's Year 2000 Web site (www.ibm.com/year2000) for further information regarding this statute and its applicability to IBM.

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

Trademarks

The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both:

IBM, AIX, DB2, e(logo), Netfinity, pSeries, WebSphere

IBM Trademarks information can be found at: http://www.ibm.com/legal/copytrade.shtml.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries

SPEC, SPECjbb, SPECint, SPECfp, SPECweb, and SPECsfs are trademarks of the Standard Performance Evaluation Corporation and information can be found at: http:\\www.spec.org.

TPC, TPC-R, TPC-H, TPC-C, and TPC-W are trademarks of the Transaction Processing Performance Council. Information can be found at: http:\\www.tpc.org.

Other trademarks are the property of their respective owners.

Notes on Benchmarks and Values

The benchmarks and values shown here were derived using particular, well configured, development-level computer systems. Unless otherwise indicated for a system, the values were derived using 32-bit applications and external cache, if external cache is supported on the system. All benchmark values are provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Actual system performance may vary and is dependent upon many factors including system hardware configuration and software design and configuration. Buyers should consult other sources of information to evaluate the performance of systems they are considering buying and should consider conducting application oriented testing. For additional information about the benchmarks, values and systems tested, contact your local IBM office or IBM authorized reseller or access the following on the Web:

TPC	http://www.tpc.org
GPC	http://www.spec.org/gpc
SPEC	http://www.spec.org

Pro/E	http://www.proe.com
Linpack	http://www.netlib.no/netlib/benchmark/performance.ps
Notesbench Mail	http://www.notesbench.org
VolanoMark	http://www.volano.com
Fluent	http://www.fluent.com

Unless otherwise indicated for a system, the performance benchmarks were conducted using AIX V4.2.1 or 4.3, IBM C Set++ for AIX/6000 V4.1.0.1, and AIX XL FORTRAN V5.1.0.0 with optimization where the compilers were used in the benchmark tests. The preprocessors used in the benchmark tests include KAP 3.2 for FORTRAN and KAP/C 1.4.2 from Kuck & Associates and VAST-2 v4.01X8 from Pacific-Sierra Research. The preprocessors were purchased separately from these vendors.

The following SPEC and Linpack benchmarks reflect the performance of the microprocessor, memory architecture, and compiler of the tested system:

- SPECint95 - SPEC component-level benchmark that measures integer performance. Result is the geometric mean of eight tests that comprise the CINT95 benchmark suite. All of these are written in the C language. SPECint_base95 is the result of the same tests as CINT95 with a maximum of four compiler flags that must be used in all eight tests.

- SPECint_rate95 - Geometric average of the eight SPEC rates from the SPEC integer tests (CINT95). SPECint_base_rate95 is the result of the same tests as CINT95 with a maximum of four compiler flags that must be used in all eight tests.

- SPECfp95 - SPEC component-level benchmark that measures floating-point performance. Result is the geometric mean of ten tests, all written in FORTRAN, that are included in the CFP95 benchmark suite. SPECfp_base95 is the result of the same tests as CFP95 with a maximum of four compiler flags that must be used in all ten tests.

- SPECfp_rate95 - Geometric average of the ten SPEC rates from SPEC floating-point tests (CFP95). SPECfp_base_rate95 is the result of the same tests as CFP95 with a maximum of four compiler flags that must be used in all ten tests.

- SPECint2000 - New SPEC component-level benchmark that measures integer performance. Result is the geometric mean of twelve tests that comprise the CINT2000 benchmark suite. All of these are written in C language except for one which is in C++. SPECint_base2000 is the result of the same tests as CINT2000 with a maximum of four compiler options that must be used in all twelve tests.

- SPECint_rate2000 - Geometric average of the twelve SPEC rates from the SPEC integer tests (CINT2000). SPECint_base_rate2000 is the result of the same tests as CINT2000 with a maximum of four compiler options that must be used in all twelve tests.

- SPECfp2000 - New SPEC component-level benchmark that measures floating-point performance. Result is the geometric mean of fourteen tests, all written in FORTRAN and C languages, that are included in the CFP2000 benchmark suite. SPECfp_base2000 is the result of the same tests as CFP2000 with a maximum of four compiler options that must be used in all fourteen tests.

- SPECfp_rate2000 - Geometric average of the fourteen SPEC rates from SPEC floating-point tests (CFP2000). SPEC_base_rate2000 is the result of the same tests as CFP2000 with a maximum of four compiler options that must be used in all fourteen tests.

- SPECweb96 - Maximum number of Hypertext Transfer Protocol (HTTP) operations per second achieved on the SPECweb96 benchmark without significant degradation of response time. The Web server software is ZEUS v.1.1 from Zeus Technology Ltd.

- SPECweb99 - Number of conforming, simultaneous connections the Web server can support using a predefined workload. The SPECweb99 test harnass emulates clients sending the HTTP requests in the workload over slow Internet connections to the Web server. The Web server softwre is Zeus from Zeus Technology Ltd.

- LINPACK DP (Double Precision) - n=100 is the array size. The results are measured in megaflops (MFLOPS).

- LINPACK SP (Single Precision) - n=100 is the array size. The results are measured in MFLOPS.

- LINPACK TPP (Toward Peak Performance) - n=1,000 is the array size. The results are measured in MFLOPS.

- LINPACK HPC (Highly Parallel Computing) - solve largest system of linear equations possible. The results are measured in GFLOPS.

VolanoMark is a 100% Pure Java server benchmark characterized by long-lasting network connections and high thread counts. In this context, long-lasting means the connections last several minutes or longer, rather than just a few seconds. The VolanoMark benchmark creates client connections in groups of 20 and measures how long it takes for the clients to take turns broadcasting their messages to the group. At the end of the test, it reports a score as the average number of messages transferred by the server per second.

VolanoMark 2.1.2 local performance test measures throughput in messages per second. The final score is the average of the best two out of three results.

The following SPEC benchmark reflects the performance of the microprocessor, memory subsystem, disk subsystem, network subsystem:

- SPECsfs97_R1 - the SPECsfs97_R1 (or SPEC SFS 3.0) benchmark consists of two separate workloads, one for NFS V2 and one for NFS V3, which report two distinct metrics, SPECsfs97_R1.v2 and SPECsfs97_R1.v3, respectively. The metrics consist of a throughput component and an overall response time measure. The throughput (measured in operations per second) is the primary component used when comparing SFS performance between systems. The overall response time (average response time per operation) is a measure of how quickly the server responds to NFS operation requests over the range of tested throughput loads.

The following Transaction Processing Performance Council (TPC) benchmarks reflect the performance of the microprocessor, memory subsystem, disk subsystem, and some portions of the network:

- tpmC - TPC Benchmark C throughput measured as the average number of transactions processed per minute during a valid TPC-C configuration run of at least twenty minutes.

- $/tpmC - TPC Benchmark C price/performance ratio reflects the estimated five year total cost of ownership for system hardware, software, and maintenance and is determined by dividing such estimated total cost by the tpmC for the system.

- QppH is the power metric of TPC-H and is based on a geometric mean of the 17 TPC-H queries, the insert test, and the delete test. It measures the ability of the system to give a single user the best possible response time by harnessing all available resources. QppH is scaled based on database size from 30GB to 1TB.

- QthH is the throughput metric of TPC-H and is a classical throughput measurement characterizing the ability of the system to support a multiuser workload in a balanced way. A number of query users is chosen, each of which must execute the full set of 17 queries in a different order. In the background, there is an update stream running a series of insert/delete operations. QthH is scaled based on the database size from 30GB to 1TB.

- $/QphH is the price/performance metric for the TPC-H benchmark where QphD is the geometric mean of QppH and QthH. The price is the five-year cost of ownership for the tested configuration and includes maintenance and software support.

The following graphics benchmarks reflect the performance of the microprocessor, memory subsystem, and graphics adapter:

- SPECxpc results - Xmark93 is the weighted geometric mean of 447 tests executed in the x11perf suite and is an indicator of 2D graphics performance in an X environment. Larger values indicate better performance.

- SPECplb results (graPHIGS) - PLBwire93 and PLBsurf93 are geometric means of literal and optimized Picture Level Benchmark (PLB) tests for 3D wireframe and 3D surface tests, respectively. The benchmark and tests were developed by the Graphics Performance Characterization (GPC) Committee. The results shown used the graPHIGS API. Larger values indicate better performance.

- SPECopc results - CDRS-03, CDRS-04, DX-03, DX-04, DX-05, DRV-04, DRV-05, DRV-06, Light-01, Light-02, Light-02, AWadvs-01, AWadvs-02, Awadvs-03, and ProCDRS-02 are weighted geometric means of individual viewset metrics. The viewsets were developed by ISVs (independent software vendors) with the assistance of OPC (OpenGL Performance Characterization) member companies. Larger values indicate better performance.

The following graphics benchmarks reflect the performance of the microprocessor, memory subsystem, graphics adapter, and disk subsystem:

Bench95 and Bench97 Pro/E results - Bench95 and Bench97 Pro/E benchmarks have been developed by Texas Instruments to measure UNIX® and Windows NT ® workstations in a comparable real-world environment. Results shown are in minutes. Lower numbers indicate better performance.

The NotesBench Mail workload simulates users reading and sending mail. A simulated user will execute a prescribed set of functions 4 times per hour and will generate mail traffic about every 90 minutes. Performance metrics are:

- NotesMark - transactions/minute (TPM).
- NotesBench users - number of client (user) sessions being simulated by the NotesBench workload.
- $/NotesMark - ratio of total system cost divided by the NotesMark (TPM) achieved on the Mail workload.
- $/User - ratio of total system cost divided by the number of client sessions successfully simulated for the Mail NotesBench workload measured.

Total system cost is the price of the server under test to the customer, including hardware, operating system, and Domino Server licenses.