IBM System p5 570 Server Consolidation Using POWER5 Virtualization White Paper

November 2006

Hsian-Fen Tsao
IBM Corporation
htsao@us.ibm.com

Bret Olszewski
IBM Corporation
breto@us.ibm.com

TABLE OF CONTENTS

Introduction

This white paper describes how Micro-Partitioning™ can be employed on IBM System p5™ 570 servers for server consolidation. We include an example consolidation scenario and explore the performance robustness of Micro-Partitioning in a demanding transactional environment.

Motivation

Increasing emphasis on the cost of delivering IT services has unleashed an unprecedented interest in server consolidation. Server consolidation refers to the simplification and optimization of IT environments by reducing the number of discrete components of infrastructure. These components start with application environments, such as application servers and databases, which are in turn implemented upon physical hardware, such as servers, routers, and storage. While numerous benefits can be cited for server consolidation, this white paper will focus solely on one practical aspect of server consolidation, namely the performance characteristics of partitioned servers.

Server consolidation has become especially attractive as current generation hardware and logical partitioning allow a number of legacy systems to be hosted within a single frame. With the announcement of the IBM System p5 family of servers, mainframe-inspired IBM Virtualization Engine™ systems technologies have arrived for the UNIX® world. The term virtualization has achieved near universal recognition. It refers to the ability to abstract the physical properties of hardware in a way that allows a more flexible usage model. Virtualization can apply to microprocessors, memory, I/O devices, or storage. Fine grain virtualization permits near instantaneous matching of workload to resources allocated, eschewing the wasted resources common to the one-server/one-application model of computing.

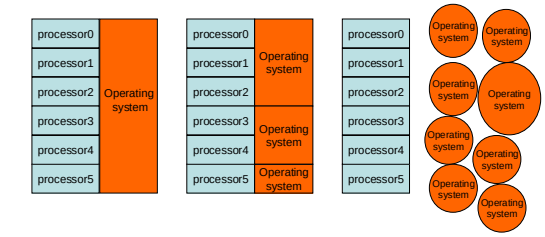
Advanced POWER Virtualization

IBM has long been a leader in virtualization. With the arrival of System p5 servers, new virtualization capabilities extend IBM's leadership in this field. These capabilities include Micro-Partitioning and virtual I/O (disk and communication adapters) are available in the Advanced POWER™ Virtualization option for the System p5 servers. This paper explores the capabilities of Micro-Partitioning.

Micro-Partitioning is the ability to run more partitions on a server than there are physical microprocessors. The concept is not novel. IBM eServer™ zSeries® systems have supported it for years. What is unique is that IBM has implemented Micro-Partitioning as an integrated option in the System p5 servers, bringing this function to a broader class of UNIX clients and applications. The AIX 5L™ Version 5.3 operating system has been adapted and optimized for virtualization.

Micro-Partitioning is the latest in a set of evolutionary steps in server virtualization for System p5 servers. Figure 1 shows the steps of partitioning evolution, beginning with the historical view of a server with one operating system managing the resources of the system. The next step was logical partitioning (LPAR), which was first offered on the IBM eServer pSeries® 690 servers and AIX 5L in 2001. With logical partitioning, it was possible to run more than one partition on a server, with each partition hosting a unique operating system. The CPU granularity of logical partitions was done at the physical processor level. Thus, there could not be more partitions than physical processors. Logical partitioning was extended with AIX 5L Version 5.2 in 2003 to permit resources to be moved between partitions dynamically, though the granularity of partitions was still by physical processor. Micro-Partitioning relaxes the constraint of partition granularity to physical processors; more partitions can operate on a system than there are physical processors. This capability has long been offered on zSeries systems and is referred to as zSeries shared processor partitioning.

Figure 1 – Partitioning Evolution



A Consolidation Proposal

Consider this example of a consolidation opportunity. This opportunity and the consolidation proposal associated with it were measured and validated on actual systems, using IBM derived benchmarks under laboratory testing conditions. The intent of these measurements was to prove the ability of the micro-partitioned server to maintain throughput and quality of service under heavy load.

The example client has five legacy servers, each an IBM RS/6000® 44P-270, running at 375 MHz with 4MB L2 caches and 7GB of memory. Each server was attached to a single drawer of IBM Serial Disk System D40 storage. The servers were running identical software stacks, each with AIX 5L V5.2, WebSphere® WAS5.0, and DB2® Universal Database™ V8.1 FP4. The bulk of the application execution was J2EE™ code in the application server. There were multiple transaction types, some fairly lightweight and some heavyweight.

The process of defining the consolidation proposal began with identifying the processor, memory, and I/O requirements of the legacy servers. We measured the environment to determine the peak CPU utilization of the server. Most servers experience highly variable CPU utilization when viewed over days. But, traditionally, peaks occur during certain periods of the day. We sized our consolidation to peak usage. Sizing systems to peak usage is usually very conservative, as it would probably be rare that all five partitions would hit peak at the same time. But, sizing to peak ensures that quality of service arrangements can be maintained even under the most difficult circumstances. For this case the system was sized for a possible 33% peak over current measurements.

In order to exploit Micro-Partitioning, an environment must be migrated to AIX 5L V5.3. While the levels of WebSphere and DB2 had not been officially qualified on AIX 5L V5.3 or with Micro-Partitioning at the time of this writing, it was our experience that they operated and scaled in the virtual environment unaltered. This was the design point of Micro-Partitioning -- transparency to application software.

We planned to migrate the SSA disk subsystem from the 44P-270 to the consolidation platform, as our analysis showed that they supplied sufficient performance to satisfy the new environment. This simplified the client migration and reduced one variable when comparing performance between the legacy servers and the partitioned server.

The consolidation proposal was loosely based on IBM's rPerf performance ratings www.ibm.com/servers/eserver/pseries/hardware/system_perf.html for the legacy platform as well as for the System p5 offerings. We began with the rPerf rating of 3.59 for the 44P-270 system. Since our legacy servers peaked at approximately 60% CPU utilization, our consolidation server required slightly more than (3.59×0.60% utilization×5 servers×1.33 peak growth)=14.3 rPerf performance. The most obvious candidate for this level of performance was a four-processor 1.65 GHz System p5 570 with an rPerf rating of 19.66. While it would appear that the p5-570 might have more performance than required, we built some contingency into our sizing. This contingency compensates for two levels of uncertainty. The first is that rPerf is not a perfect indicator of the relative performance between two systems running arbitrary workloads. The second incorporates the fact that Micro-Partitioning is not without some overhead. The concurrent execution of many partitions at high system load impacts the efficiency of the hardware to some degree. The exact overhead is very dependent on the system utilization and application characteristics.

Though the legacy system contained 7GB, it was over provisioned. This is a common finding in consolidations; servers that have more resources than are actually used. The memory requirement for the five partitions could be handled with 16GB on the p5-570. This allowed 3GB of memory for each partition which was sufficient for the environment.

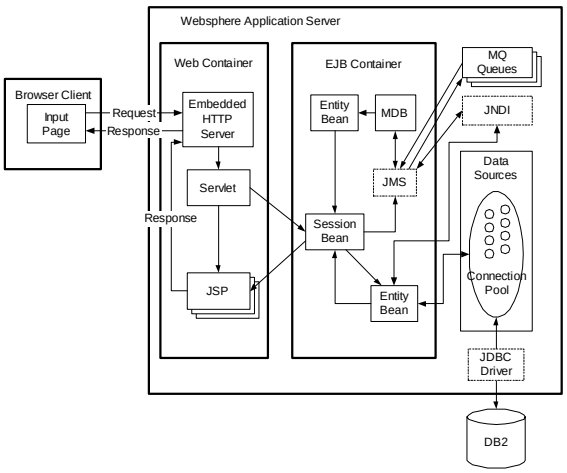
Test Workload Description

A performance test application was used to measure the performance of the p5-570 and a legacy system in a 2-tier test environment. Tier-1 consisted of the machine used to drive the System Under Test (SUT). Tier-2 consisted of WebSphere Application Server and DB2 database server running concurrently under AIX 5L with 32-bit kernel. Since the application server and the database server were executing within the same operating system instance, no external communication was required between them.

The test environment used a typical request and response model with clients making requests via a Web browser and the server processing the requests and sending the responses back to clients. The application was built on Java™ 2 platform using

Enterprise Edition (J2EE) technology. Figure 2 shows a diagram of transaction flow for the workload. The load driver accessed the application through the Web layer via Enterprise Java Beans (EJBs) exercising the IBM WebSphere Application Server (WAS) and DB2 database servers (See Figure 2). This environment also made use of the Java Messaging Service (JMS) and the Message Queue (MQ) infrastructure to increase its complexity and approximate real-life scenarios. The workload ran a mix of transactions, running a gamut from fairly lightweight to very heavyweight CPU operations.

Figure 2 – High Level Conceptual View of the Test Environment Data Flow



For this white paper, the performance of each server or partition was measured by using "Operations Per Second" (OPS) to represent the rate of requests the server processed over the measurement interval. The OPS metric included the sum of all of the transactions in the workload over time, including the full mix of light-weight and heavy-weight transactions.

Test Environment Description

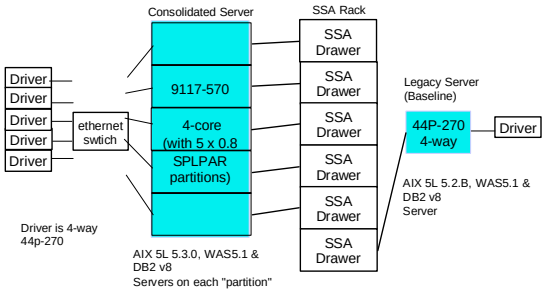
The test environment consisted of a baseline platform and a target platform. Each platform included a server (SUT) and a driver. The legacy platform consisted of one server and one driver. Both were 44P-270's, running AIX 5L V5.2 with 32-bit kernel and configured as 375 MHz 4-way servers each with one integrated 10/100 Mbps Ethernet adapter. The legacy server had a single SSA adapter to access external storage. The consolidation platform consisted of one IBM System p5 570 server, running at 1.65 GHz,that had been virtualized into five micro-partitions. As is the case for the legacy platform, each partition was driven by an identical 44P-270 server for measurement purposes. Each partition used AIX 5L V5.3 with 32-bit kernel, and had a partition configuration of 8/10th of an IBM POWER5™ CPU, two virtual CPU's, and runs capacity capped. Each partition contained one SCSI adapter, one SSA adapter, one 10/100 Mbps Ethernet adapter, and 3GB of physical memory. A single Cisco 100 Mbps full-duplex Ethernet switch was used to provide a private network for the test environment. The servers and clients communicated via the Cisco switch.

Each of the five partitions and the legacy system used a single 9GB SCSI disk for their AIX 5L image and 16x9GB SSA disks configured into two loops for the database, DB2 log, Web-based Java application, and data analysis repository. The write intensive database log is placed at the second loop to ensure optimal performance.

The five partitions and legacy system contained identical application stacks, with JDK1.4, WAS 5.1GM, MQ, DB2 V8+FixPack 4, and the DB2 JDBC driver installed in order to run the performance workload.

The virtualization server consolidation topology is depicted in Figure 3 below:

Figure 3 - Virtualization Server Consolidation Topology

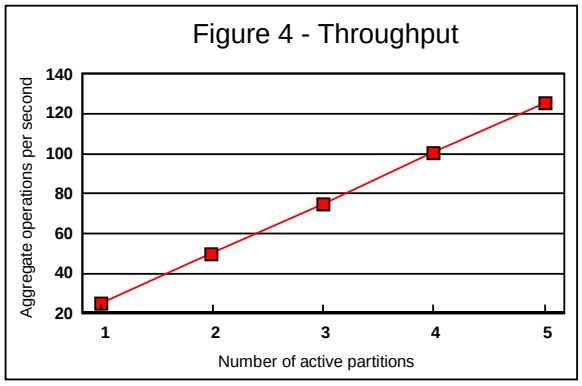


Evaluating Consolidation – Part 1

This section details measurements that show how the partitioned server performance compared to the legacy server under increasing client load. These measurements involved driving an increasing number of clients at the peak load measured on the legacy server. The load was ramped by increasing the number of clients that are driving the workload. The initial measure is of a single server partition driven to 25 operations per second with the other four

partitions running AIX 5L, but not running active workload. The overhead of idle partitioning running AIX 5L V5.3 was negligible, a result of optimizations for the Micro-Partitioning environment. Subsequently, we drove two partitions, each to 25 operations per second, with three idle partitions. The number of active partitions was increased incrementally under all five are under load.

Figure 4 shows the throughput scaling linearly with the number of partitions, up to our maximum measurement of five active partitions.



To evaluate quality of service, we monitored the response times of each of the transaction types. Since the response times of the transactions scaled linearly with increasing active partitions, we chose to concentrate our analysis on the middle-weight transaction in the workload. Figure 5 shows the responsiveness scaled as we increased the number of loaded partitions. The figure shows our average response time did not increase dramatically with more partition load. The 90th percentile response times scaled similarly to what we would observe on a traditional server environment with increasing CPU utilization.

The fact that response time was not measurably impacted for this workload by Micro-Partitioning is important, though expected. The hypervisor dispatching mechanism is granular enough to allow partitions to be responsive, but efficient enough not to dramatically impact partition throughput.

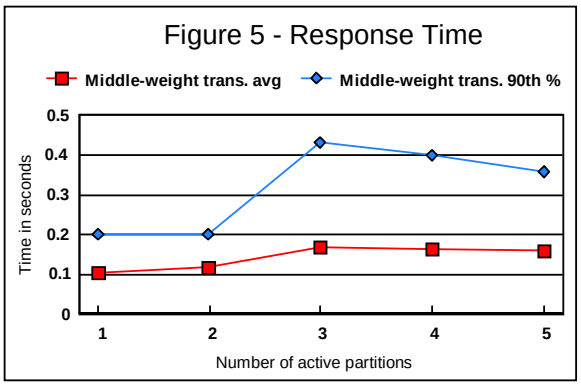


Figure 6 compares the responsiveness of the partitioned server to the legacy server at the same operation rate. Under the same constant workload, the response time and throughput of each active partition were similar or better compared to that of the legacy system. This is not surprising because the p5-570 has much more powerful processors than the legacy server which allows it to process CPU intensive workloads much more efficiently.

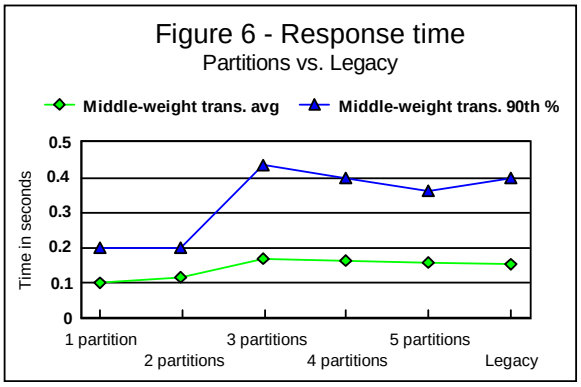
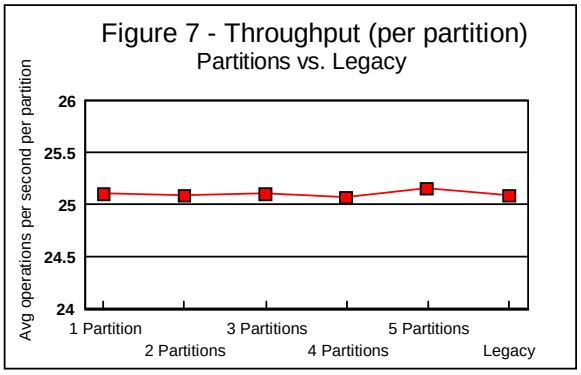


Figure 7 shows that the workload driven in our experiments was fairly consistent across all of the measurements. It shows that the comparisons to the legacy server are very precise in terms of the workload measured.

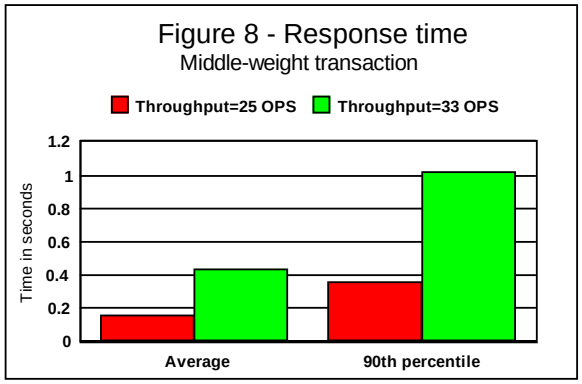


This data shows that the partitioned server scales exceptionally well under increasing load at the peak workload measured on the legacy server. It also shows that the sizing ratio between the p5-570 partitions and our legacy system is close to one-to-one. Quality of service is comparable between the System p5 server with five active partitions and the dedicated legacy server at this workload level.

Evaluating Consolidation – Part 2

The next data shows the partitioned server's ability to run the workload at a higher throughput. This overdriven workload shows that the partitioned server in this configuration actually provides the headroom assumed in the initial sizing with satisfactory quality of service.

For this comparison, we have measured all five partitions under load. In the first case, as in the previous section, each partition is driven to 25 operations per second. In the second case, each partition is driven to 33 operations per second. Figure 8 shows the relative average response times as well as the 90th percentile response times for the middleweight transaction. In the case of the 33 operations per second measurement, the partitioned server is effectively more than 90% utilized. The increase in response times is related to that fact, the more heavily utilized the server becomes, the more response time is impacted. But, the responsiveness of the partitioned server, even at this high utilization, is within the sizing goals.



Conclusion

This exercise results in three key points.

First, virtualization using a single System p5 570 with five partitions can provide the same level of performance as five heavily loaded dedicated 44P-270 servers. Fundamentally, System p5 platforms with IBM Virtualization Engine systems technology provides scalability and granularity that effectively allow four POWER5 processors to do the work of twenty legacy POWER3™ processors in a demanding Web application serving environment.

Second, System p5 virtualization technology allows a partitioned server to have comparable quality of service to dedicated legacy servers, even at fairly high machine utilizations. The Micro-Partitioning solution does not inhibit the responsiveness of typical workloads.

Finally, it is possible to do direct system sizing based on relative ratios of System p™ servers using partitioning. This allows the relatively simple leverage of Micro-Partitioning in server consolidation environments.

Glossary

90th Percentile Response Time: The time elapsed from when the first transaction is sent from the driver to the SUT until the driver from the SUT receives the response from slowest 90% of the transactions.

Capped: A partition is defined as either capped or uncapped. A capped partition is not allowed to exceed its entitlement. An uncapped partition is allowed to exceed its entitlement if CPU resources are available in the shared processor pool.

Entitlement: The entitled processor capacity is an attribute of each partition. The value represents a commitment of capacity reserved for the partition.

Hypervisor: A trusted firmware program that controls access to the key hardware resources. It enforces partitioning boundaries and maintains integrity once the resources are assigned and partitions are up and running.

Micro-Partitioning: A technology provided by IBM System p5 systems allowing more active partitions than physical processors. A layer of firmware known as the hypervisor manages the allocation of physical resources to partitions and handles scheduling of partitions.

Partition: The definition of an instance of an operating system. Partitions may be active (booted) or inactive. The number of active partitions is constrained by the partition resource definitions.

rPerf (Relative Performance): An estimate of commercial processing performance relative to other System p servers
http://www.**ibm.com**/servers/eserver/pseries/hardware/rperf.html

Simultaneous Multithreading: An advanced feature of the POWER5 microprocessor. Each microprocessor has 2 hardware threads, which increases the potential instruction execution parallelism

Virtual CPU: An attribute of a partition. It defines the number of processors exposed to the partition. With simultaneous multi-threading, the number of processors observed by an AIX 5L V5.3 partition is twice the number of physical CPUs.

IBM Corporation
Marketing Communications
Systems and Technology Group
Route 100
Somers, New York 10589

Produced in the United States of America
November 2006

PSW3006-USEN-00