Issue Date: IBM edition for Unix
February | March 2005, Posted On: 2/16/2005

Virtual Explanation

Jaqui Lynch

Today’s big buzzword when discussing on demand solutions is virtualization. Although LPAR and dynamic LPAR (DLPAR) provide a form of virtualization, it’s the introduction of IBM* POWER5* technology that adds true virtualization and provides flexibility in resource allocation while minimizing constraints. Virtualization encompasses two major features: the capability to share physical resources and the capability to move those resources dynamically between workloads or partitions.

POWER5 virtualization is built on the basics of LPAR technology. Unlike POWER4* systems, POWER5 systems are always running under a Hypervisor*, even if all of the resources are assigned to just one partition. The Hypervisor is a piece of firmware that handles time slicing and dispatching for the partition workloads between processors. There’s a slight overhead associated with using the Hypervisor because it needs both memory and processor resources to run. Actual overhead depends on the number of workloads and LPARs and the amount of page mapping that occurs. This can be monitored using the lparstat command. By using lparstat -h, it’s possible to see the percentage of processor time spent in the Hypervisor, as well as the number of Hypervisor calls.

POWER5 and POWER4 systems running AIX* v5.2 require the Hardware Management Console (HMC) for partitioning. The HMC allows partitions to be created with a minimum granularity of one processor and with I/O devices being assigned at the slot level. However, once the system is running AIX v5.3 or an enabled version of Linux* along with the Advanced POWER* Virtualization (APV) feature, the options change dramatically.

POWER5 technology takes this to the next level with the APV feature. Virtualization in the POWER5 world has several prerequisites. First is the combination of POWER5 hardware and AIX v5.3 or an enabled version of Linux. For the purposes of this article, I’ll use AIX v5.3 as a reference. Additionally, virtualization requires the aforementioned APV feature. This comes standard with the pSeries* 590 and 595 servers, but is a fee-based feature on the pSeries 570, 550 and 520 servers. Lastly, the HMC is required to implement virtualization and the new Capacity on Demand (CoD) options.

The key components of virtualization in the POWER5 world include: the POWER5 servers, AIX v5.3 (or an enabled version of Linux), Micro-Partitioning*, simultaneous multi-threading (SMT), Virtual I/O Server, virtual Ethernet, shared Ethernet adapter (SEA) and virtual SCSI. The APV is needed to provide for Micro-Partitioning, SEA, virtual SCSI server and Partition Load Manager (PLM).

Micro-Partitioning

Micro-Partitioning changes the whole planning structure for POWER5 servers. Bottlenecks are still identified in the same way and tend to occur in the same places, however, the solutions for these bottlenecks are often different from those in POWER4. The on demand world, implemented in POWER5 with APV, allows greater flexibility and granularity in allocating resources. The ability to add fractions of processors instead of whole processors also helps customers make better use of the servers. The use of Workload Manager (WLM) and PLM helps better optimize those resources.

Partitions are now either dedicated processor partitions (processors are assigned in increments of a full processor), or they’re shared processor partitions (these use Micro-Partitioning); they can’t be a combination of the two. In the case of Micro-Partitioning, a set of processors can be assigned to the shared processor pool (SPP) and then made available to LPARs that access the processors based on their entitled capacity and priority within the pool. At this point, there’s only one SPP available on the server, but multiple LPARs can share those resources. So, on a 16-way we could have seven processors dedicated to LPARs as we did previously, and the other nine could be in the SPP or CoD pool. For now, let’s assume they’re in the SPP.

Like regular processors, processors in the SPP can be assigned to LPAR. However, the granularity is different. The initial assignment is a minimum of 1/10th of a processor with increments of 1/100th of a processor. Keep in mind that these partitions still need dedicated memory and I/O resources. The I/O and networking resources can be dedicated as they were previously, virtualized through the Virtual I/O Server or you can combine these two options.

When defining a partition, you must choose between a shared processor and a dedicated processor partition. For a shared partition it’s also necessary to determine whether the partition is capped or uncapped. This is an important decision because it affects both performance and license charges. An uncapped partition can use all of the processors in the pool if it needs to and if it’s more important than the other workloads running. A capped partition can use processors up to the number of processors specified in the cap. License charging is typically based on the maximum processors that this workload could use—in the case of uncapped partitions, this would be all of the processors in the pool, whereas with a capped partition, this would be the number of processors specified in the cap. Additionally, the number of processors allocated to a partition is referred to as the entitled capacity.

When defining an uncapped partition you also must define its variable capacity weight. This is a number between zero and 255 that’s used to determine the share of extra capacity the partition can receive when it’s competing for available resources with other uncapped partitions.

All of the above significantly adds to the planning cycle but also provides far greater flexibility in assigning resources.

Simultaneous Multi-Threading
SMT is a POWER5 enhancement that allows two threads to execute concurrently on a single physical processor. For many workloads, SMT can lead to a 30-percent improvement in throughput and/or response time with no user cost. It requires POWER5 hardware and AIX v5.3 or an enabled version of Linux.

Virtual I/O Server
One key component of the APV feature is the Virtual I/O Server. This facility allows you to virtualize both I/O and network resources. The Virtual I/O Server is a custom AIX v5.3 partition that’s used to provide I/O resource sharing. The server can be used for SEAs and disks (Virtual SCSI) for multiple client LPARs. The I/O Server acts as the host system and actually owns the physical resources. These resources are then virtualized out to the client LPARs. Typically, at least two Virtual I/O Servers are recommended per server to provide redundancy and balance performance. This server is a highly customized AIX v5.3 partition that comes on a preconfigured CD. It can’t be used for anything else, and the only item IBM will support for installation is a backup client such as a Tivoli Storage Manager (TSM) client. It does support IBM’s Shared Device Driver (SDD) and EMC’s Powerpath software to allow for multiple paths to I/O devices.

Virtual Ethernet
Virtual Ethernet isn’t part of the APV—it only requires AIX v5.3 or an enabled version of Linux on a POWER5 system. The HMC is used to define virtual Ethernet devices such that LPARs are connected via memory rather than real Ethernet cards. An LPAR can support up to 256 virtual Ethernets, each running between 1 and 3 Gbps. (Note: The use of virtual Ethernet will incur some system overhead because it uses the system’s processors for the communications that would normally be offloaded to an Ethernet card.)

Shared Ethernet Adapter
SEA, a new service in POWER5 that comes with the APV, allows an adapter that’s physically in a Virtual I/O Server to be shared among several partitions. Effectively, it acts as a network switch that routes traffic between the virtual Ethernet adapters in the clients and the real adapter in the hosting server. Network adapter sharing is done using a SEA that’s configured in the Virtual I/O Server. The Ethernet adapter belongs to the Virtual I/O Server and is exported to the client partitions. IP addresses are mapped accordingly, and the traffic is routed through the one physical adapter in the Virtual I/O Server. From the Virtual I/O Server, traffic either flows over the real network or through the Hypervisor to the actual partitions. (Note: Adapters in the hosting partition can still be trunked into EtherChannels to increase performance and availability. A single or trunked interface can support up to 16 virtual Ethernets.) As with virtual Ethernet, there’s some processor overhead in the client systems because they can’t take advantage of the processors on the Ethernet cards.

Virtual SCSI
In the case of I/O, disks and adapters are shared using the virtual SCSI server. This runs in the Virtual I/O Server and effectively allows you to have a physical disk that has multiple logical volumes (LVs) allocated on it. You can then export each LV to a different client LPAR. The client LPAR would see these LVs as normal SCSI disks, even though they may be Fibre attached to the server. This allows you to take a 146 Gb disk drive and carve it into 3 x 45 Gb LVs that could be used as boot disks by three different LPARs.

It should be noted that virtual SCSI will perform slower than real devices. According to IBM at announce time, it takes roughly double the processor time to perform I/O as compared to using real disks. This load is split evenly between the Virtual I/O Server and the virtual SCSI client.

Updates to Tools

The smctl command shows the status of SMT for the LPAR and controls SMT for that LPAR. lparstat -i can also be used to obtain information, including SMT status, for an LPAR. Additionally mpstat -s reports on statistics for a physical processor across its two logical processors. Both vmstat and iostat have been updated to report on physical processor consumed by the LPAR, percentage of entitlement consumed, asynchronous I/O and per-file-system adapter queues. The sar -P ALL has been updated to report on entitled capacity consumed, and the -L option was added to topas to report on LPAR information.

A Virtual Success

These features are some of the key components of virtualization. It's important to note that virtualization takes additional processor power and memory—it's not free. Part of any new server strategy should include an analysis of where virtualization might work for your site. Definite cost savings can be made by sharing resources using virtualization, but it's important to understand the workloads that are running and whether they need dedicated bandwidth or can work happily in a shared or virtualized environment.

Examples of where virtualization can be powerful and beneficial to an IT environment include:

1. Server consolidation of mixed workloads (i.e., some using I/O and some using CPU)
2. Consolidation of many small non-busy workloads

The key to success in the virtual world is good planning. Examine all of your workloads and determine whether they will play well together and benefit from sharing resources. You'll also need to figure out which workloads will benefit from Micro-Partitioning versus dedicated processors. POWER5 technology provides great flexibility and multiple ways to accomplish the same goal, hopefully while saving resources and money.

About the Author(s):

Jaqui Lynch: Jaqui Lynch, an *eServer Magazine, IBM edition for UNIX* technical editor, is a senior systems engineer focusing on pSeries and Linux at Mainline Information Systems. During her more than 26 years in the IS industry, she's been responsible for a wide variety of projects and OSs across multiple vendor platforms, including mainframes, UNIX systems, midrange systems and personal workstations. Jaqui can be reached at jaqui.lynch@mainline.com.