HITACHI
Inspire the Next

High-Performance Nested Virtualization With Hitachi Logical Partitioning Feature

Solutions Enabled by New Intel Virtualization Technology Extension in the Intel Xeon Processor E5 v3 Family

By Hitachi Data Systems

September 2014

Contents

Executive Summary and Introduction

Intel Xeon based Hitachi Compute Blade servers (CB 2000 and CB 500) feature an embedded logical partitioning (LPAR) capability implemented in the blade server firmware. This logical partitioning architecture is based on technology originally developed by Hitachi for mainframe and UNIX servers to provide ultimate flexibility and reliability for enterprise mission-critical workloads. Hitachi is the only vendor to offer LPAR technology on general-purpose blade servers using the Intel x86 architecture. As part of our long-term partnership with Intel, Hitachi was an early collaborator on the Intel Virtual Machine Control Structure (VMCS) Shadowing extension in the Intel Xeon processor E5 v3 family.

This white paper describes the performance improvement realized in an configuration where a Virtual Machine Manager (VMM) is run within LPAR through the VMCS Shadowing extension (also known as Haswell-EP). This improvement is demonstrated by measurement of the performance of a guest operating system (OS), running on a well-known and commonly used software Virtual Machine Manager (VMM), hosted in a logical partition in Hitachi LPAR (VMM on LPAR). This hierarchical structure of OS, VMM and LPAR is often referred to as “nested virtualization.”

We will also review a practical example of the technology in use today in the live production environment of a Hitachi customer.

Hitachi is committed to the continuous enhancement of VMM on LPAR technology, and will continue to work with Intel to expand its application and adoption.

Purpose and Value of VMM on LPAR

Fusion of Virtual Machine (VM) and LPAR Technology

Server virtualization technology can be divided into 2 categories:

- **VM** hides the physical hardware and emulates a virtual machine for each OS running on top of it. This method offers excellent flexibility and ease of management.
- **LPAR** physically partitions the hardware resources, and exposes a subset of hardware resources to each OS in the system. Since each OS directly controls its own hardware resources, and these are physically separated from those running another OS, this method offers superior robustness, integrity, reliability and performance.

In the mainframe market, where virtualization technology was initially developed, the VM category was the first to be productized. Later on, as hardware technology evolved, highly robust and efficient LPAR technology was introduced, to provide the highest flexibility and manageability to users, LPAR further evolved the capability to run VMs within the LPAR partitions. This was the first instantiation of “VMM on LPAR” or “nested virtualization” technology. Over time, LPAR technology has evolved to be essentially equivalent to bare-metal hardware, and VMMs and OSs are able run identically on bare metal or in an LPAR partition without requiring any adaption or code difference to accommodate the specific environment.

In the x86 market today, the combination of the flexibility and manageability of VMs with the robustness, integrity, reliability and high performance of LPAR is the “holy grail” for organizations. In order to deliver this ultimate environment, Hitachi has been developing VMM on LPAR solutions for many years.

Figure 1 shows the concept of VMM on LPAR. This technology enables multiple VMMs to run simultaneously on a single x86 server without each VMM interacting with others, in the same way that multiple guest OSs may run on a single VMM. Considering the LPAR feature from Hitachi as the primary virtualization layer and other VMMs as a secondary virtualization layer, the combination of LPAR and other VMMs can be seen as “nested,” Thus, the technology is often referred to as “nested virtualization.” In this context, the VMM inside the LPAR can be called “Level-1 Guest,” and the OS in the VMM can be called “Level-2 Guest,” as is depicted in Figure 1.

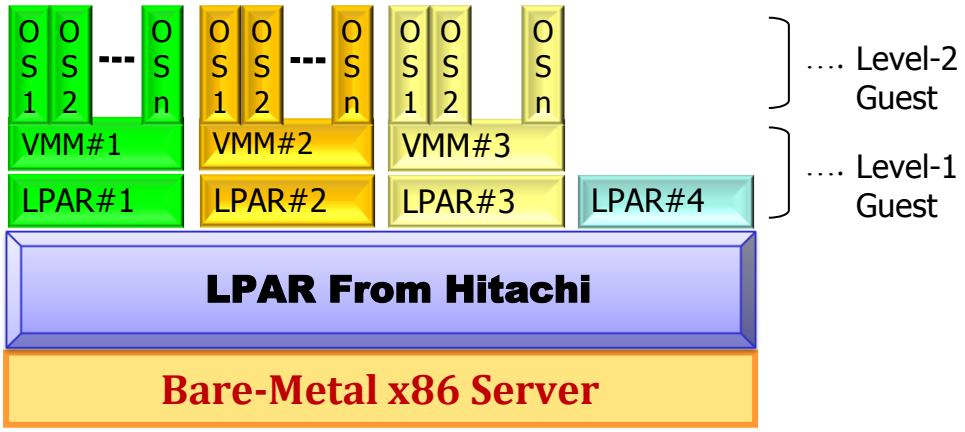

Figure 1. VMM on LPAR Concept

Typical Use Cases

Following are typical use cases of VMM on LPAR:

- **High reliability cloud with fully isolated multitenant platform**. Even in the enterprise cloud, it is becoming common for multiple VMs, each owned by different tenants, to be hosted in a single server. In these cases, the isolation (performance, management, failure impact, and so on) between the VMs belonging to each tenant, is crucial.

  Recent developments in VM management software have tried to address this issue by adding a role-based management capability, allowing different tenants running in a same server to independently manage the operation of their respective VMMs. However, this does not address the impact of VM workload fluctuation (also known as. the "noisy neighbor" problem) or the isolation of hardware or software failures. Because of this situation, it is typical for many enterprise cloud service providers (CSPs) to assign a separate server and VMM to each tenant to achieve the required SLA. The result is higher cost for both CSP and end users. In contrast to this result, VMM on LPAR technology permits multitenant consolidation to be achieved while maintaining the required tenant isolation, dramatically reducing the cost of meeting the required SLA. (See Figure 2.)

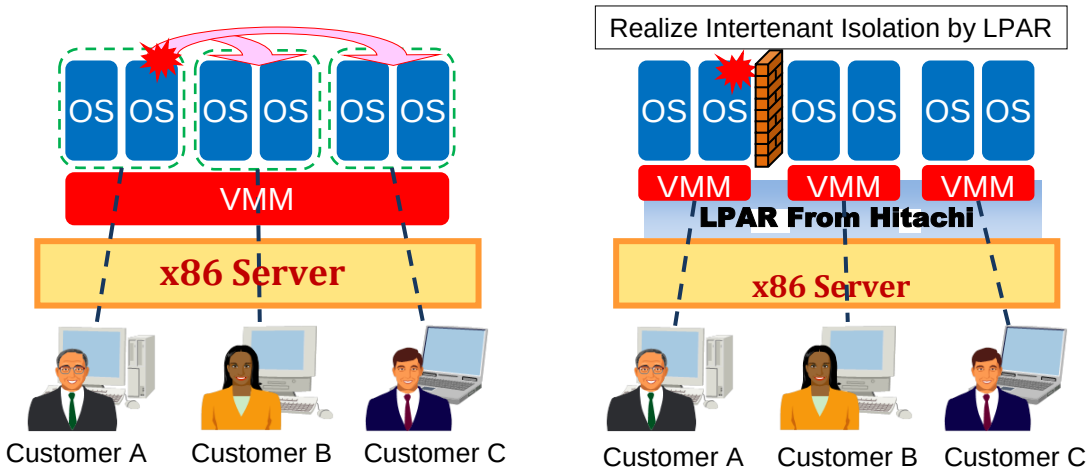

Figure 2. Example of Multitenant Cloud Platform Benefit of VMM on LPAR

- **VMM as a service**. Using the capabilities of VMM on LPAR, a service provider is able provide entire VMMs to its customers, without buying additional physical servers. Figure 3 offers an example of multiple VMMs hosted on a single physical server. This approach enables a hosting service provider to offer end users a more flexible and granular product; for example, customers can specify the number of the cores, amount of memory and disk capacity to be used by the VMM. Then, they can directly manage the VMM to run any number of VMs based on the amount of the dedicated resource for the VMM and the required performance for the VMs. The benefit of this service is not only the reduced cost end users experience by sharing the platform with multiple other users. The end user can also own the entire control of the VMM; they can independently implement any settings they require, flexibly and quickly, without any interaction with the settings of other users running their VMs on the same hardware.

  There are benefits for the hosting service provider, too. The starting price for VMM hosting is much cheaper, which lowers the barriers to entry for new businesses and opening in new markets. The increased flexibility also helps a hosting service provider address the challenges of a major new customer acquisition, or a sudden increase in customer resource usage. The provider can immediately and flexibly deploy resources without the need to purchase a new server.

The capabilities of VMM on LPAR also enable a multitiered hosting model: Tier 1 providers focuses on the large-scale customers and effective platform operation, and hosts Tier 2 providers who can service smaller customers. Tier 2 providers can focusing on customization, fine-tuning and management, based on the lower cost and more flexible VMM hosting platform by the Tier 1 provider.

Technology evolution in x86 processor design has significantly increased the number of the cores built on a processor. As shown in the traditional hosting model in Figure 3, this can result in excess resource in a server, which inevitably results in lower power efficiency. By consolidating multiple small VMMs into a single server and turning off the unused servers, the hosting service provider can significantly improve the power efficiency.

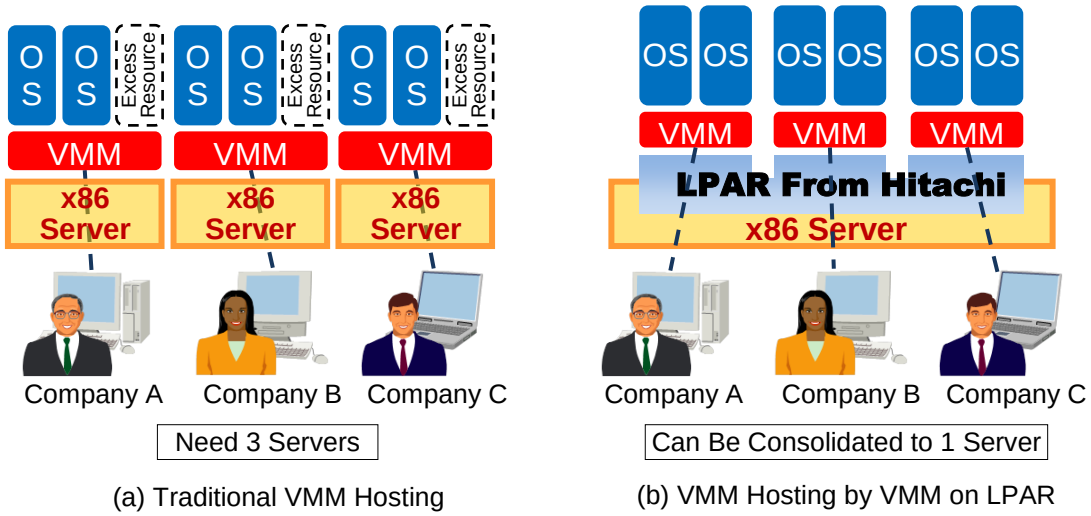

Figure 3. VMM Hosting With VMM on LPAR

In addition, LPAR technology allows a single server to both consolidate applications using VMMs and also run mission-critical systems on bare metal. It can also host multiple versions of VMMs (for example, when testing a new version of a VMM), and provide multiple different VMM environments for VMM transition, all without requiring the purchase of additional servers.

Challenges for VMM on LPAR, and the Effects of VMCS Shadowing in Xeon Processor E5 v3 Family

This section describes the Intel Virtualization Technology (VT) necessary for VMM execution. It examines how the VMCS Shadowing extension in Xeon processor E5 v3 family resolves the challenges of VMM on LPAR experienced by the previous Xeon processor E5 v2 family.

Outline of the Intel VT Hardware Assist Feature for Server Virtualization

In the traditional single-layered VMM case, server virtualization is realized by a single VMM, running on a single physical server, which allows multiple OSs to be executed simultaneously (see Figure 4). A physical server consists of CPU, memory, storage, and so on, that is accessed by a VMM, which provides a layer of abstraction between the hardware and the guest OS. To simultaneously execute multiple guest OSs safely, Intel VT has features to manage certain VMM functions. Specifically, Intel VT saves the status of the guest OS in a VMCS, and provides the VMM with an access method to use the VMCS.

For example, when a guest OS requests access to storage, Intel VT invokes the VMM. The invoked VMM then reads the VMCS to understand the guest OS status, and executes the storage access request on behalf of it, while considering any impacts on any other guest OS that may be running on the VMM at the same time. The result of storage access request is reflected in the VMCS by the VMM, which then resumes the guest OS (see Figure 5).

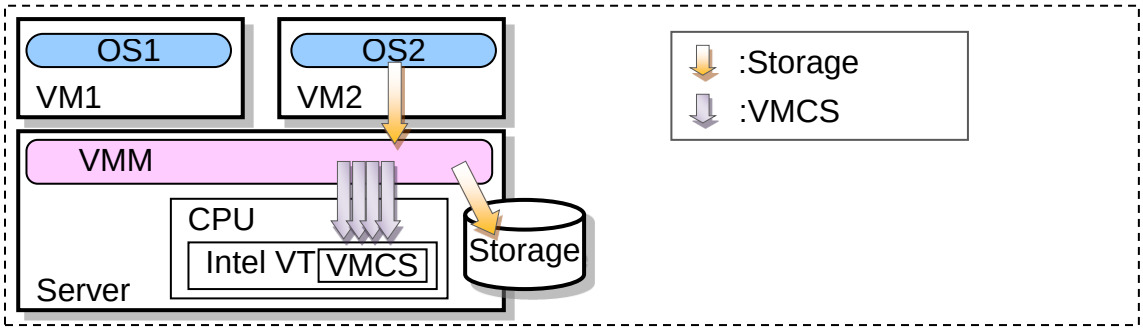

Figure 4. Example of Typical Server Virtualization

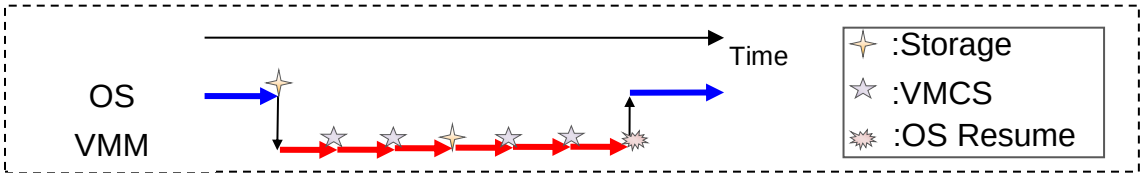

Figure 5. Typical Time Chart of General Server Virtualization

VMM on LPAR Architecture and Its Challenges Without VMCS Shadowing Extension

When VMMs were used within an LPAR on previous generations of Intel Xeon processors, VMMs were executed in an LPAR using VMCS, as previously described. However, in addition, some of the internal processes of the VMM also require the use of a VMCS. Therefore, the LPAR feature from Hitachi created another VMCS exclusively for a VMM, called a Shadow VMCS. When a VMM accesses the VMCS, LPAR accesses the Shadow VMCS on behalf of the VMM (see Figure 6).

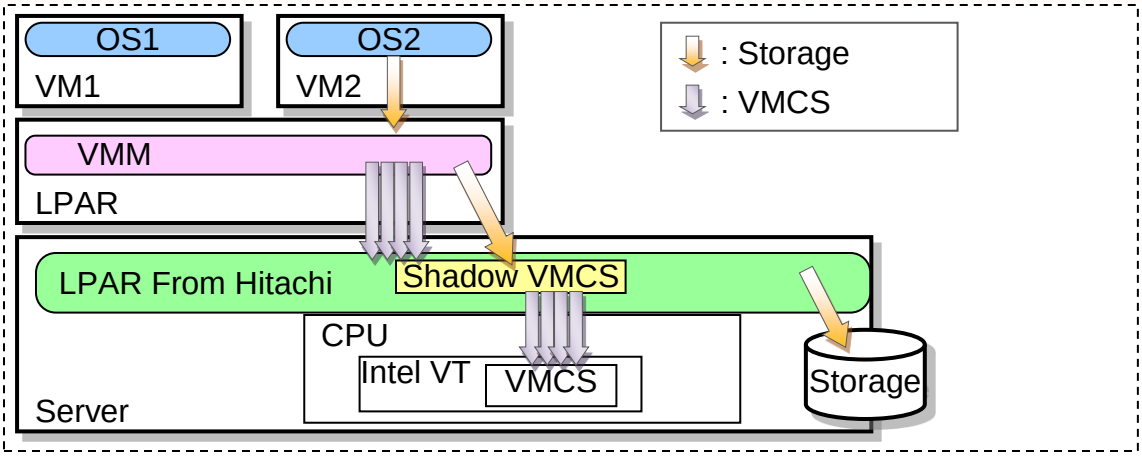

Figure 6. VMM on LPAR Operation Without VMCS Shadowing

As an example, a guest OS storage access is processed as follows. When a storage access request occurs in the guest OS, Intel VT recognizes it and invokes the LPAR feature from Hitachi. LPAR copies the OS status information in VMCS to Shadow VMCS, and invokes the VMM. Thereafter, when the invoked VMM accesses the VMCS, LPAR actually accesses the Shadow VMCS on behalf of the VMM. After the storage access is completed and before the VMM resumes the guest OS execution, LPAR first writes back the Shadow VMCS information into the VMCS, and then the OS resumes. (see Figure 7).

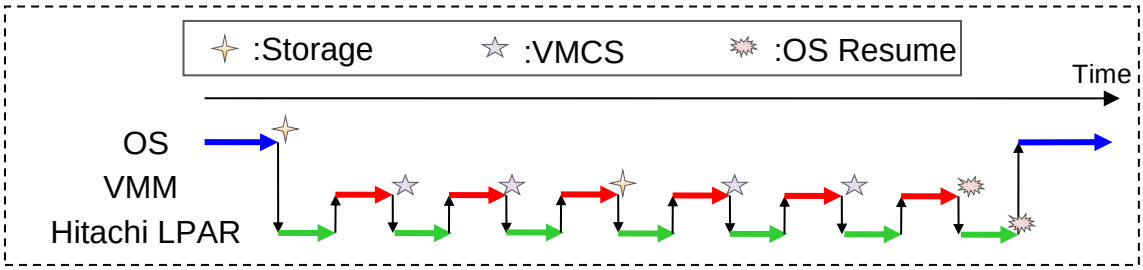

Figure 7. Execution Time Chart of VMM on LPAR Without VMCS Shadowing

During this process, the VMM refers to, and updates, the VMCS multiple times. Because of this, the number of storage accesses and other I/O operations can significantly impact system performance when nested virtualization, such as VMM on LPAR is used. In use cases requiring a large number of I/O accesses, the impact on system performance will be greater because the proportion of processing time required by the level 1 VMM (LPAR in this case) increases, reducing the processing time available to service the guest OS.

Solving the Problem Using VMCS Shadowing Extension in Intel VT

To resolve the problem of performance degradation caused by frequent I/O in nested VMMs, the Intel Xeon processor E5 v3 family has introduced a new hardware extension: VMCS Shadowing allows the operation of the Shadow VMCS to be performed in the processor hardware. Because of its long experience with virtualization and experience with nested virtualization on LPAR, Hitachi was an early collaborator with Intel during architecture definition and development of this extension as part of our continuing long-term partnership.

When the VMCS Shadowing extension is enabled, the Intel Xeon processor E5 v3 family retains not only the VMCS itself but also a Shadow VMCS (see Figure 8). Now, when a VMM on LPAR accesses VMCS, LPAR is no longer invoked. Instead, the processor directly reads or writes the Shadow VMCS. By using the VMCS Shadowing extension, the total execution time of LPAR is not affected by the number of VMCS accesses, which improves the performance of VMM on LPAR significantly (see Figure 9).

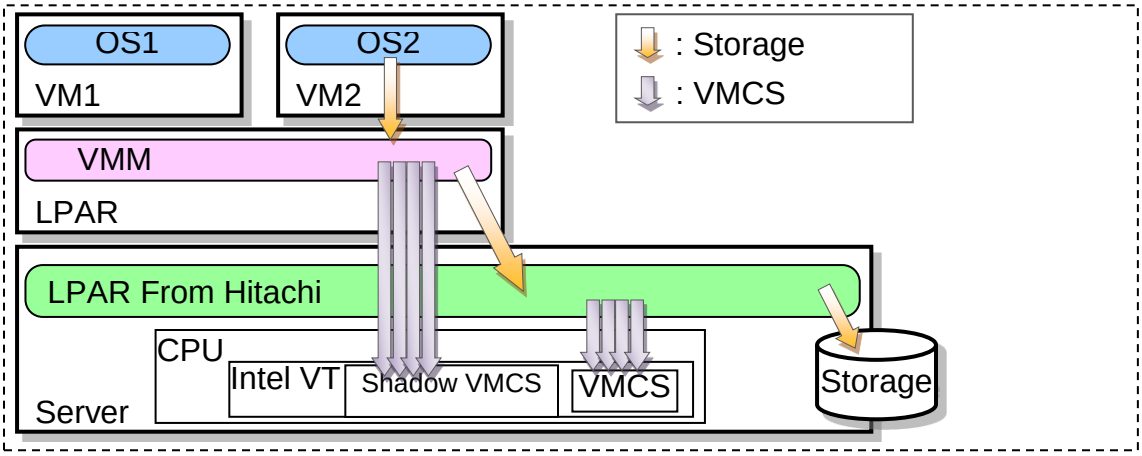

Figure 8. VMM on LPAR Operation With VMCS Shadowing

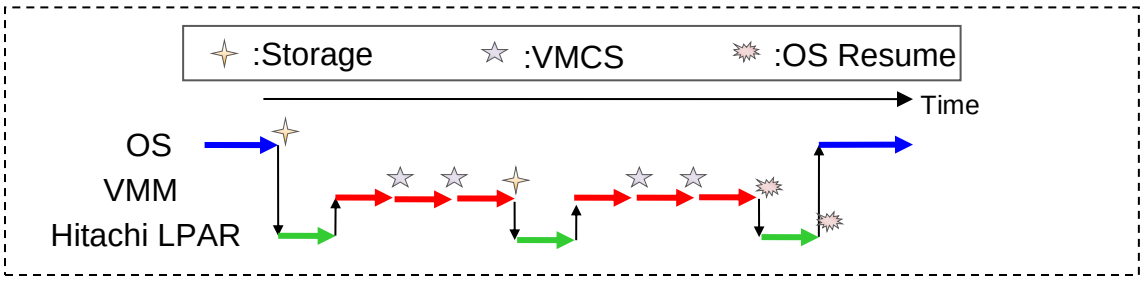

Figure 9. Execution Time Chart of VMM on LPAR With VMCS Shadowing

Performance Measurement Environment and Results

System Configuration for Performance Measurement

To measure the performance improvement resulting from the VMCS Shadowing extension, Hitachi configured the latest Compute Blade 500 system with Intel Xeon processor E5-2667 v3 and the VMCS Shadowing extension prototype version of the LPAR feature (see Table 1).

On LPAR, the VMM on LPAR configuration was set up as described in Table 2. The VMM used was Red Hat KVM, within which the VMs described in Table 2 were configured, and Red Hat Enterprise Linux 6.4 (64bit) was installed and executed. To clearly demonstrate the performance improvement in VMM on LPAR, 2 highly storage dependent

benchmarks were chosen: IOmeter, which issues very high frequent storage accesses, and sysbench, which simulates typical real-life database use (see Table 3).

Table 1. Hardware Specification

Chassis		CB 500
Blade		CB520H B3
	Hitachi LPAR	VMCS Shadowing extension prototype version
	Processor	Intel Xeon CPU E5-2667 v3 @ 3.20GHz (8 core) x2 (Hyper-Threading Disabled)
	Memory	64GB
Storage		Hitachi Adaptable Modular Storage 2300

Table 2. LPAR,VMM and VM Configuration

Operating System		RHEL 6.4 x64 (Kernel 2.6.32-358.el6.x86_64)
Virtual Machine	CPU	4 core (cpu64-rhel6)
	Memory	8GB
Virtual Machine Manager		KVM (bundled in RHEL 6.4)
LPAR	CPU	6 core
	Memory	10GB

Table 3. Benchmark Setting

Benchmark	Configuration
iometer-2006_07_27	1 thread 4K; 100% read; 0% random
sysbench 0.4.12	--test=oltp --oltp-table-size=1000000 --num-threads=4 --max-requests=0 --max-time=180 --oltp-test-mode=complex (database server is MySQL 5.1.66)

Performance Measurement Result

Figure 10 shows the performance of VMM on LPAR both without VMCS Shadowing extension (equivalent to the Intel Xeon processor E5 v2 family) and with VMCS Shadowing extension (equivalent to the Intel Xeon processor E5 v3 family). When the VMCS Shadowing extension is enabled, *IOmeter* (highly frequent storage access) records 1.8 times higher performance. *Sysbench* (database access simulation) records 1.4 times higher performance.

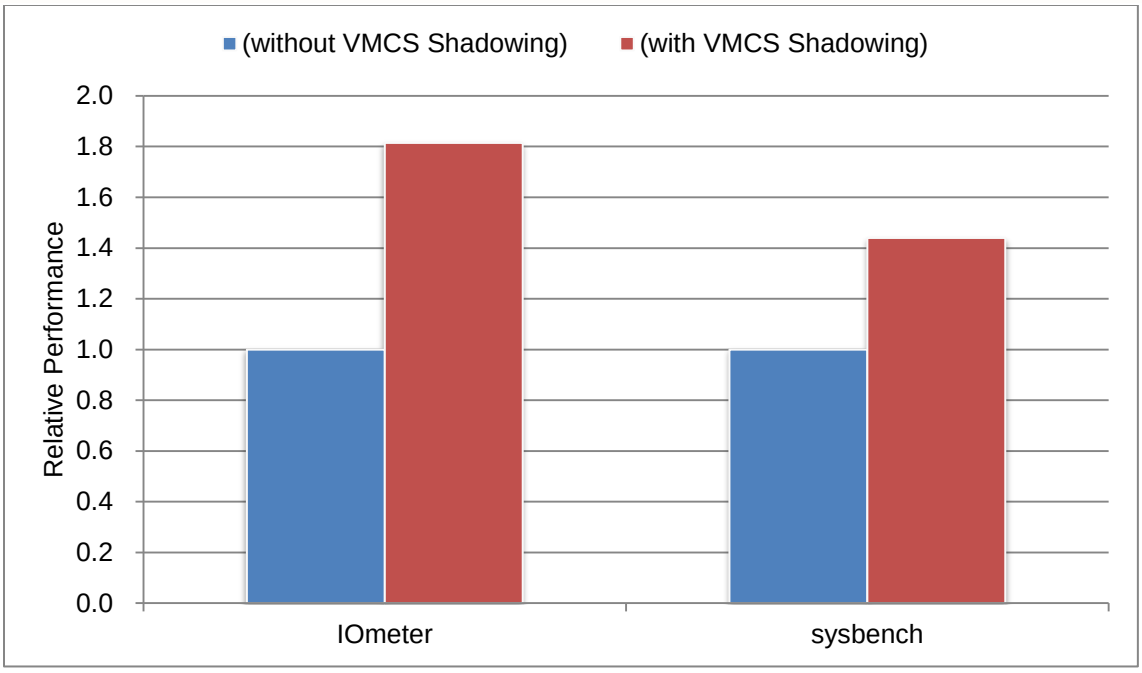

Figure 10. Performance Difference Using VMCS Shadowing Extension

Benchmark Observations

In these benchmark measurements, *IOmeter*, which issues storage accesses more frequently, recorded a higher performance improvement due to the VMCS Shadowing extension. This increase occurs because more storage access results in more frequent usage of VMCS Shadowing extension. This measurement shows that the VMCS Shadowing extension is more effective in reducing the performance overhead for I/O intensive workloads than for simple CPU- or memory-intensive workloads in the VMM on LPAR environment.

Using the previous generations of Intel Xeon processor E5 family, the peak performance design of VMM on LPAR was sometimes difficult to predict, because the virtualization overhead varied greatly depending on the frequency of storage access. Now, with the new VMCS Shadowing extension, not only is the overhead much smaller but also the variability is reduced, resulting in more predictable sizing of server and storage configurations. We believe this change will rapidly accelerate the market adoption of VMM on LPAR and allow this technology to be more widely used.

Customer Use Case

In this section, the system configuration and live experience of a Hitachi customer, TAKAOKA TOKO Co. Ltd., is described. By using VMM on LPAR, the both capital and operating expenditure (capex and opex) were reduced without comprising the functionality and reliability of the IT environment.

TAKAOKA TOKO Co., Ltd., is a leading provider of electric power distribution and transformation equipment. Much of its equipment is sold to large electric power companies and companies in the public sector. TAKAOKA TOKO Co., Ltd., has been using Hitachi Compute Blade 500 (CB 500) and Hitachi Unified Storage 110 (HUS 110), as well as other cutting-edge technologies and products, for its mission-critical systems, such as accounting and production management. The

LPAR feature from Hitachi is used to host the mission-critical workloads, which require the highest reliability and availability, while VMM environments are used to support the legacy environment consolidation. TAKAOKA TOKO Co., Ltd., chooses the appropriate environment based on the required service level. However, because of business expansion including acquisition of other companies, both capex and opex costs in IT have continuously increased.

To address the problem of increasing expenditure, TAKAOKA TOKO Co., Ltd., decided to build a new system shown in Figure 11. This system uses 2 LPARs: The 1st LPAR is used to host mission-critical production systems, such as corporate accounting. The 2nd is used to host a VMM, which supports Microsoft® Windows Server® 2003 as a Level-2 Guest. Under this OS, an older generation legacy accounting system is running. Although a high performance level is not required to support this application, it is essential that it runs without failure. To migrate the application to a new OS and retain the required level of availability and reliability would be very expensive and involve high risk. Using VMM on LPAR, the existing mature application can continue to be used for a longer time, saving rewrite and maintenance costs, and both legacy systems and new systems can be run on the same single server. If there is a new service request, TAKAOKA TOKO Co., Ltd., can add a new LPAR or VM flexibly and quickly, at any time.

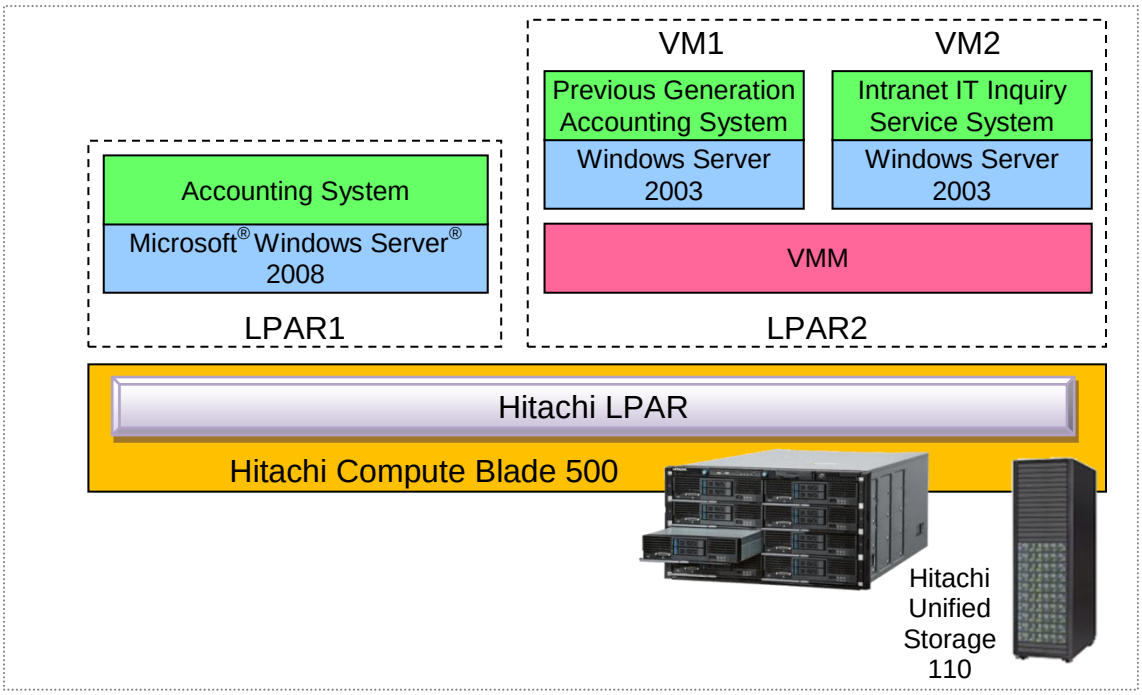

Figure 11. System configuration in TAKAOKA TOKO CO., LTD.

TAKAOKA TOKO Co., Ltd., commented: “We are very satisfied with the VMM on LPAR technology provided by Hitachi. LPAR is an optimal platform for mission-critical system consolidation. VMM is optimal for legacy system consolidation. The fusion of these 2 technologies realized the highest efficiency and server resource utilization for us, and resulted in the reduction of both capex and opex at the same time. Especially, the reduction of data center floor area and power consumption is significant. This technology delivered exactly what we needed for this IT platform renovation project.”

The company added: “We have been hearing that Intel and Hitachi have been collaborating closely for a long time, and we will soon have the processor innovation to expand the applicability of VMM on LPAR technology. We are looking forward to more from this technology and can hardly wait.”

Hitachi Data Systems

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 96050-2639 USA
www.HDS.com community.HDS.com

Regional Contact Information
Americas: +1 408 970 1000 or info@hds.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hds.com
Asia Pacific: +852 3189 7900 or hds.marketing.apac@hds.com

WP-494-A N Winkworth September 2014