July 2006

Scheduler Improvements in VMware ESX Server 2.5.3 and ESX Server 3.0

Performance Tip for Intel Processor-Based IBM NUMA-Class Servers with Hyperthreading Enabled

By Chris Floyd, Joe Jakubowski, and Lily Shi
IBM System x Performance Laboratory
IBM Systems and Technology Group

Introduction

This paper discusses a performance optimization in the VMware ESX Server 2.5.3 and ESX Server 3.0 scheduler that can benefit Intel® processor-based IBM® NUMA-class servers with hyperthreading enabled. This paper describes the test environment and presents the results obtained with and without the performance optimization. The information presented is based on our experience with setting up and running VMware ESX Server configurations in the IBM System x™ Performance Laboratory.

ESX Server Scheduler Optimization

Improvements in VMware's ESX Server scheduler can lead to more optimal processor affinity when hyperthreading is enabled on Intel processor-based IBM System x servers that use the NUMA design. Recent improvements recognize the cost of migrating work within a single core from one hyper-twin to the other. Better estimation of this cost in VMware ESX Server 2.5.3 and ESX Server 3.0 can lead to fewer migrations and improved performance.

Test Configuration and Results

Test Configuration

Tests were conducted on an IBM System x3950 multi-node server and an IBM System x3850 server. A single virtual machine was created in ESX Server. The workload exercised in the virtual machine was SPECjbb2005.[1] SPECjbb2005 is a widely used benchmark for evaluating the performance of servers running typical Java® business applications. This workload is CPU-intensive and memory bandwidth-intensive. The SPECjbb2005 workload metric was measured in steady state over a 20-minute interval. The attributes of the virtual machine and the workload are listed in Table 1.

Guest Operating System	Microsoft Windows 2003 Enterprise Edition SP1
Number of Virtual CPUs	2
Virtual Memory	3600 MB
SPECjbb2005 JVM	Sun JRE 1.5.0_06
Sun JRE Memory Allocation	1600 MB
SPECjbb2005 Number of Warehouses	8 Warehouses

Table 1. Attributes of the ESX Server Virtual Machine and SPECjbb2005 Workload

Measurements were conducted on three different server configurations:

- IBM System x3850 (4 sockets)
- IBM System x3950 (2 nodes, 8 sockets)
- IBM System x3950 (2 nodes, 4 sockets)

In the 2-node, 4-socket configuration, 2 processors were installed in each node. This configuration is referred to as a 2x2, which means two processor sockets were populated in each of two nodes. The 2-node servers are NUMA servers whereas the x3850 is an SMP non-NUMA server. All three servers had

[1] To remain in compliance with SPEC Fair Use Rules and Guidelines, the results data presented in this paper were obtained using the SPECjbb2005 workload in a research setting. The results presented are not compliant with this SPEC benchmark's run and reporting rules. The SPECjbb2005 workload was used only as a driver mechanism to apply load to the server. The following deviation from the run rules is noted: The number of SPECjbb2005 warehouses was fixed at 8.

the Dual Core Intel Xeon® Processor 7040 (3.0 GHz/2x2MB L2 cache) installed. Hyperthreading was enabled in the server BIOS and in ESX Server for all configurations.

Measurements with ESX Server 2.5.2

The data in Table 2 compares the SPECjbb2005 performance throughput metric and CPU affinity between the three servers when running ESX Server 2.5.2 and when hyperthreading is enabled. The CPU affinity data points of interest are highlighted in bold blue text. The throughput for the x3850 and x3950 4x2 servers is nearly identical at approximately 10,000 operations/second. The CPU affinity for the x3950 4x2 server is tightly clustered around two CPU cores on the same socket (CPUs 4 and 5) or two CPU cores on different sockets (CPUs 4 and 6). The ESX Server scheduler optimally affinitizes the threads to maximize performance. However, the x3950 2x2 server has 15-20 percent lower throughput. Note that the CPU affinity is more distributed around CPU cores and their hyper-twins and also across different CPU sockets (CPUs 8-11 and CPUs 12-15 are on different physical CPU sockets).

Server	x3850 Baseline	x3950 (4x2) Run 1	x3950 (4x2) Run 2	x3950 (4x2) Run 3	x3950 (2x2) Run 1	x3950 (2x2) Run 2	x3950 (2x2) Run 3
Throughput (ops/sec)	**10,065**	**10,313**	**10,119**	**10,090**	**8,527**	**8,411**	**8,298**
CPU 0	4.50	3.00	3.00	3.10	2.40	2.50	2.40
CPU 1	0.04	0.11	0.11	1.00	0.07	0.07	0.07
CPU 2	**76.80**	0.10	0.10	0.10	0.07	0.07	0.07
CPU 3	7.80	0.12	0.12	0.11	0.07	0.07	0.07
CPU 4	0.12	**89.10**	**92.40**	**84.50**	0.06	0.06	0.06
CPU 5	**86.80**	**91.00**	**87.30**	7.40	0.07	0.07	0.07
CPU 6	0.06	0.11	0.11	**87.00**	0.06	0.06	0.06
CPU 7	6.80	0.09	0.09	0.09	0.07	0.07	0.07
CPU 8	3.90	0.14	0.14	0.14	**46.10**	**82.10**	0.12
CPU 9	2.90	0.16	0.16	0.15	**17.50**	0.12	0.12
CPU 10	3.40	0.22	0.22	0.16	**49.80**	**88.80**	**90.40**
CPU 11	3.10	0.15	0.15	0.16	**21.50**	1.00	0.10
CPU 12	3.80	0.16	0.16	0.16	1.20	1.30	0.10
CPU 13	5.30	0.16	0.16	0.16	1.10	0.10	0.10
CPU 14	0.90	0.16	0.16	0.16	**28.60**	0.20	0.09
CPU 15	7.40	0.13	0.13	0.13	**10.50**	0.07	**87.90**

Table 2. SPECjbb2005 Measurements with ESX Server 2.5.2 and Hyperthreading Enabled

As shown in Table 2, the ESX Server 2.5.2 scheduler behaves differently on the x3950 4x2 and the x3950 2x2 servers even though both are NUMA servers. The scheduler differences are a result of the 16-processor limit in ESX Server 2.5. The x3950 4x2 presents 32 logical CPUs to ESX Server when hyperthreading is enabled. Since ESX Server has a limit of 16 CPUs, on detecting 16 physical CPUs, the additional logical CPUs due to hyperthreading are disabled by the ESX Server. The x3950 2x2 presents 16 logical CPUs to ESX Server when hyperthreading is enabled, so threads can be scheduled on all of the logical CPUs.

Table 3 shows the ESX Server 2.5.2 scheduler's behavior on the x3950 2x2 when hyperthreading is disabled in the system BIOS and in ESX Server. With hyperthreading disabled, throughput approaches the levels measured on the x3850 and x3950 4x2 server configurations. The data in Tables 2 and 3 shows that the ESX Server 2.5.2 scheduler is not optimally affinitizing the threads for the x3950 2x2 server when hyperthreading is enabled.

Server	x3950 (2x2) Run 1	x3950 (2x2) Run 2
Hyperthreading	**Disabled**	**Disabled**
Throughput (ops/sec)	10,068	10,090
CPU 0	5.80	5.90
CPU 1	**88.30**	**88.10**
CPU 2	**90.50**	**90.45**
CPU 3	0.12	0.12
CPU 4	0.13	0.14
CPU 5	0.13	0.13
CPU 6	0.14	0.14
CPU 7	0.13	0.13

Table 3. SPECjbb2005 Measurements with ESX Server 2.5.2 and Hyperthreading Disabled

Measurements with ESX Server 2.5.3

Table 4 data illustrates the effect of scheduler improvements with ESX Server 2.5.3 compared to ESX Server 2.5.2 on the x3950 2x2 server. The ESX Server 2.5.3 scheduler consistently affinitizes threads to logical CPUs on the separate cores (CPUs 5 and 7) and there is no random migration to their hyper-twins. The scheduler improvements result in more consistent throughput and higher performance with the SPECjbb2005 workload running on the x3950 2x2.

For completeness, NUMA locality was examined on the x3950 2x2 using IBM proprietary tools. Data shows that 99.94 percent of the memory accesses were local to the node where the virtual machine's threads were affinitized. This indicates that the ESX Server NUMA code was functioning optimally.

Server	x3950 (2x2) Run 1	x3950 (2x2) Run 2	x3950 (2x2) Run 3	x3950 (2x2) Run 1	x3950 (2x2) Run 2	x3950 (2x2) Run 3
ESX Server Build	2.5.2	2.5.2	2.5.2	2.5.3	2.5.3	2.5.3
Throughput (ops/sec)	8,527	8,411	8,298	10,042	10,072	10,120
CPU 0	2.40	2.50	2.40	6.30	6.50	2.90
CPU 1	0.07	0.07	0.07	0.15	0.15	0.15
CPU 2	0.07	0.07	0.07	0.08	0.08	0.08
CPU 3	0.07	0.07	0.07	0.18	0.18	0.20
CPU 4	0.06	0.06	0.06	0.10	0.10	0.10
CPU 5	0.07	0.07	0.07	**91.80**	**88.80**	**91.00**
CPU 6	0.06	0.06	0.06	0.09	0.09	0.09
CPU 7	0.07	0.07	0.07	**85.70**	**87.70**	**87.20**
CPU 8	**46.10**	**82.10**	0.12	0.11	0.11	0.11
CPU 9	**17.50**	0.12	0.12	0.12	0.12	0.12
CPU 10	**49.80**	**88.80**	**90.40**	0.08	0.08	0.08
CPU 11	**21.50**	1.00	0.10	0.10	0.10	0.10
CPU 12	1.20	1.30	0.10	0.08	0.09	0.08
CPU 13	1.10	0.10	0.10	0.09	0.09	0.09
CPU 14	**28.60**	0.20	0.09	0.08	0.08	0.09
CPU 15	**10.50**	0.07	**87.90**	0.10	0.10	0.10

Table 4. SPECjbb2005 Measurements with ESX Server 2.5.3 and Hyperthreading Enabled on the IBM System x3950 2x2

Conclusion

Scheduler improvements in VMware ESX Server 2.5.3 have been shown to better optimize thread affinity on Intel processor-based IBM NUMA-class servers when hyperthreading is enabled. The optimizations are effective when the logical CPU count does not exceed the maximum number of logical CPUs supported by ESX Server. IBM systems meeting this criteria include the x3950 2x2 with dual core CPUs and the x460 8-way with single core CPUs. Using the SPECjbb2005 workload, throughput increased by as much as 20 percent when running in an ESX Server 2.5.3 virtual machine compared to the same workload running in an ESX Server 2.5.2 virtual machine.

ESX Server 3.0 also contains the scheduler improvements for Intel processor-based IBM NUMA-class server environments with hyperthreading enabled. While this paper does not present measured data with ESX Server 3.0, similar results are expected.

Customers using VMware ESX Server 2.5.2 or earlier should be advised to upgrade to ESX Server 2.5.3 or later if the underlying physical server is an Intel processor-based IBM NUMA-class server with hyperthreading enabled such as the x3950 2x2 with dual core CPUs or the x460 8-way with single core CPUs.

IBM Systems and Technology Group
3039 Cornwallis Road
Research Triangle Park, NC 27709

Produced in the USA
07-06