VMware® vSphere™: The CPU Scheduler in VMware ESX® 4.1

VMware vSphere™ 4.1

WHITE PAPER

vmware®

Table of Contents

1. Introduction . . . 3
2. CPU Scheduler Overview . . . 3
2.1. Proportional Share–Based Algorithm . . . 3
2.2. Relaxed Co-Scheduling . . . 4
2.3. Distributed Locking with Scheduler Cell . . . 5
2.4. CPU Topology–Aware Load Balancing . . . 6
2.4.1. Load Balancing on NUMA Systems . . . 6
2.4.2. Load Balancing on Hyperthreading Architecture . . . 6
3. CPU Scheduler Changes in ESX 4 . . . 7
3.1. Further Relaxed Co-Scheduling . . . 7
3.2. Elimination of CPU Scheduler Cell . . . 9
3.3. Multicore-Aware Load Balancing . . . 10
3.3.1. CPU Load-Based Migration Throttling . . . 10
3.3.2. Impact on Processor Caching . . . 10
3.3.2.1. vSMP Consolidation . . . 11
3.3.2.2. Inter-VM Cache Affinity . . . 11
3.3.3. Aggressive Hyperthreading Support . . . 11
3.3.4. Extended Fairness Support . . . 12
3.4. Wide-VM NUMA Support . . . 12
3.4.1. Performance Impact . . . 12
4. Performance Evaluation . . . 13
4.1. Experimental Setup . . . 13
4.2. Verifying the Proportional-Share Algorithm . . . 14
4.2.1. Shares . . . 14
4.2.2. Reservation . . . 14
4.2.3. Limit . . . 15
4.3. Impact of Relaxed Co-Scheduling . . . 15
4.4. Impact of Multicore-Aware Load Balancing . . . 17
4.5. Impact of Extended Fairness Support . . . 17
4.6. ESX 4 in Comparison to ESX 3.5 . . . 18
4.7. Impact of Wide-VM NUMA Support . . . 20
5. Summary . . . 22
6. References . . . 22

1. Introduction

The CPU scheduler in VMware vSphere™ 4.x (ESX™ 4.x) is crucial to providing good performance in a consolidated environment. Because most modern processors are equipped with multiple cores per processor, or chip multiprocessor (CMP) architecture, it is easy to build a system with tens of cores running hundreds of virtual machines. In such a large system, allocating CPU resources efficiently and fairly is critical.

In ESX 4, there are some significant changes to the CPU scheduler—for better performance and scalability. This paper describes these changes and their impacts on performance. It also provides certain details of the CPU scheduling algorithms.

This paper is updated to reflect the changes in the ESX 4.1 CPU scheduler. Sections 3.4 and 4.4 discuss wide-VM NUMA support and its performance impact.

It is assumed that readers are familiar with virtualization using VMware® ESX™ and already know the common concepts of the CPU scheduler. However, it is also strongly recommended to read the *vSphere Resource Management Guide* [1], because this paper frequently refers to it.

In this paper, a pCPU denotes a physical CPU; a vCPU denotes a virtual CPU. The former refers to a physical core on a system, and the latter refers to a virtual machine's virtual processor. A host refers to a vSphere server that hosts virtual machines; a guest refers to a virtual machine on a host. In describing CMP architecture, a socket or a processor package can be used to denote a chip that has multiple cores.

Like a process, a vCPU can be in one of the following states: In running state, a vCPU executes on a pCPU. In ready state, it is runnable but waiting in a queue. In wait state, it is blocking on a resource. An idle vCPU can enter wait_idle, a special wait state, which does not depend on a resource. The idle vCPU changes its state from wait_idle to ready when interrupted.

The rest of this paper is organized as follows: Section 2 describes the key features of the CPU scheduler, which comprise ESX 4.0/4.1 and ESX 3.5. The section applies, in most part, to older versions of ESX as well. Section 3 describes the major changes introduced by ESX 4.0/4.1 and discusses their performance impact qualitatively. Section 4 presents experimental results, which show the impact of the changes described in this paper. Finally, Section 5 summarizes this paper.

2. CPU Scheduler Overview

The role of the CPU scheduler is to assign execution contexts to processors in a way that meets system objectives such as responsiveness, throughput and utilization [2]. On conventional operating systems, the execution context corresponds to a process or a thread; on ESX, it corresponds to a world.[1]

Fundamentally, the ESX CPU scheduler shares the same objectives as other operating systems, but it faces unique challenges. First, the allocation of CPU resources among virtual machines is required to be faithful to user specification. Also, providing the illusion that a virtual machine completely owns CPU resources becomes more critical. This section describes the key features of the ESX CPU scheduler and how these features address such challenges.

2.1. Proportional Share–Based Algorithm

Generally, one of the main tasks of the CPU scheduler is to choose which world is to be scheduled to a processor. Also, if the target processor is already occupied, it needs to be decided whether or not to preempt the currently running world on behalf of the chosen one.

The CPU scheduler in UNIX uses a priority-based scheme. It associates each process with a priority and makes a scheduling choice or preemption decision based on the priorities. For example, a process with the highest priority among ready processes would be chosen; then, if it is higher in priority, the process might preempt the currently running process.

[1] In this paper, vCPU and world are used interchangeably to denote the execution context in ESX.

Unlike UNIX, ESX implements the proportional share–based algorithm. It associates each world with a share of CPU resources. This is called entitlement and is calculated from user-provided resource specifications, which are shares, reservation, and limit. See the *vSphere Resource Management Guide* [1] for details.

This entitled resource might not be fully consumed. When making scheduling decisions, the ratio of the consumed CPU resources to the entitlement is used as the priority of the world. If there is a world that has consumed less than its entitlement, the world is considered high priority and will likely be chosen to run next. In ESX, as in UNIX, a numerically lower priority value is considered high; for example, a priority 0 is considered higher than 100. It is crucial to accurately account for how much CPU time each world has used. Accounting for CPU time is also called charging.

The key difference between UNIX and ESX CPU schedulers can be viewed as how a priority is determined. In UNIX, a priority is arbitrarily chosen by the user. If one process is considered more important than others, it is given higher priority. Between two priorities, it is the relative order that matters, not the degree of the difference.

In ESX, a priority is dynamically re-evaluated based on the consumption and the entitlement. The user controls the entitlement, but the consumption depends on many factors including scheduling, workload behavior and system load. Also, the degree of the difference between two entitlements dictates how much CPU time should be allocated.

The proportional share-based scheduling algorithm has a few benefits over the priority-based scheme:

First, one can accurately control the CPU allocation of virtual machines by giving a different number of shares. For example, if a virtual machine, vm0, has twice as many shares as vm1, vm0 would get twice as much CPU time compared to vm1, assuming both virtual machines highly demand CPU resources. It is difficult to achieve this with the UNIX scheduler because the priority does not reflect the actual CPU consumption.

Second, it is possible to allocate different shares of CPU resources among groups of virtual machines. The virtual machines in the group might have different shares. Also, a group of virtual machines might belong to a parent group, forming a tree of groups and virtual machines. With the proportional-share scheduler, CPU resource control is encapsulated and hierarchical. Resource pool [1] is designed for such use.

The capability of allocating compute resources proportionally and hierarchically in an encapsulated way is quite useful. For example, consider a case where an administrator in a company datacenter wants to divide compute resources among various departments and to let each department distribute the resources according to its own preferences. This is not easily achievable with a fixed priority–based scheme.

2.2. Relaxed Co-Scheduling

Co-scheduling, alternatively known as gang scheduling [3], executes a set of threads or processes at the same time to achieve high performance. Because multiple cooperating threads or processes frequently synchronize with each other, not executing them concurrently would only increase the latency of synchronization. For example, a thread waiting to be signaled by another thread in a spin loop might reduce its waiting time by being executed concurrently with the signaling thread concurrently.

An operating system requires synchronous progress on all its CPUs, and it might malfunction when it detects this requirement is not being met. For example, a watchdog timer might expect a response from its sibling vCPU within the specified time and would crash otherwise. When running these operating systems as a guest, ESX must therefore maintain synchronous progress on the virtual CPUs. The ESX CPU scheduler meets this challenge by implementing relaxed co-scheduling of the multiple vCPUs of a multiprocessor virtual machine. This implementation allows for some flexibility while maintaining the illusion of synchronous progress. It meets the needs for high performance and correct execution of guests.

An article, "Co-scheduling SMP VMs in VMware ESX Server," [4] well describes the co-scheduling algorithm in ESX. Refer to the article for more details. The remainder of this section describes the major differences between the strict and the relaxed co-scheduling algorithms.

Strict co-scheduling is implemented in ESX 2.x. The ESX CPU scheduler maintains a cumulative skew per each vCPU of a multiprocessor virtual machine. The skew grows when the associated vCPU does not make progress while any of its siblings makes progress. A vCPU is considered to make progress if it uses CPU or halts.

It is worth noting that there is no co-scheduling overhead for an idle vCPU, because the skew does not grow when a vCPU halts. For example, when a single-threaded application runs in a 4-vCPU virtual machine, resulting in three idle vCPUs, there is no co-scheduling overhead and it does not require four pCPUs to be available.

If the skew becomes greater than a threshold, typically a few milliseconds, the entire virtual machine would be stopped (co-stop) and will only be scheduled again (co-start) when there are enough pCPUs available to schedule all vCPUs simultaneously. This ensures that the skew does not grow any further and only shrinks.

The strict co-scheduling might cause CPU fragmentation. For example, a 2-vCPU multiprocessor virtual machine might not be scheduled if there is only one idle pCPU. This results in a scheduling delay and lower CPU utilization.

Relaxed co-scheduling introduced in ESX 3.x significantly mitigated the CPU fragmentation problem. While the basic mechanism of detecting the skew remains unchanged, the way to limit the skew has been relaxed.

The CPU scheduler used to be required to schedule all sibling vCPUs simultaneously when a virtual machine was co-stopped. This was too restrictive, considering that not all vCPUs lagged behind. Now, instead, only the vCPUs that accrue enough skew will be required to run simultaneously. This lowers the number of required pCPUs for the virtual machine to co-start and increases CPU utilization. The CPU scheduler still attempts to schedule all sibling vCPUs for better performance if there are enough pCPUs.

2.3. Distributed Locking with Scheduler Cell

A CPU scheduler cell is a group of physical processors that serves as a local scheduling domain. In other words, the CPU scheduling decision mostly involves a single cell and does not impact other cells. The motivation of the cell structure is to design a highly scalable scheduler on a system with many processors.

The scalability of the CPU scheduler has been an important goal in other operating systems as well. With scalable scheduler, the overhead of the CPU scheduler should not increase too much as the number of processors or the number of processes—or worlds—increases. Considering that the number of virtual machines per processor on ESX is relatively small, the number of worlds per processor should also be small compared to the number of processes on general operating systems. So it is more important to scale well on a system that has many processors.

As the CPU scheduler code can be concurrently executed on multiple pCPUs, it is likely to have concurrent accesses to the same data structure that contains scheduler states. To ensure the integrity of the states, such accesses are serialized by a lock. A simple approach would have a global lock protecting entire scheduler states. While the approach is simple, it serializes all concurrent scheduler invocations and can significantly degrade performance. For better performance, a finer-grained locking is required.

The scheduler cell enables fine-grained locking for the CPU scheduler states. Instead of using a global lock to serialize scheduler invocation from all processors, ESX partitions physical processors on a host into multiple cells where each cell is protected by a separate cell lock. Scheduler invocations from the processors in the same cell would mostly contend for the cell lock.

The size of a cell must be large enough to fit multiprocessor virtual machines because the virtual machine performs best when sibling vCPUs are co-scheduled on distinct processors. The sibling vCPUs might also be required to be co-scheduled to ensure correctness. Instead of acquiring multiple cell locks where those sibling vCPUs would be queued, it would be more efficient to restrict the sibling vCPUs to be scheduled only within a cell at any moment. However, this can be a limiting factor when placing virtual machines.

NOTE: This limitation is eliminated in ESX 4. Section 3.2 describes this change in detail. Also, refer to VMware Knowledge Base article 1007361 [5] for the impact of cell size and examples of how a cell is constructed.

2.4. CPU Topology–Aware Load Balancing

ESX is typically deployed on multiprocessor systems. On multiprocessor systems, balancing CPU load across processors, or load balancing, is critical to the performance. Load balancing is achieved by having a world migrate from a busy processor to an idle processor. Generally, the world migration improves the responsiveness of a system and its overall CPU utilization.

Consider a system that has only two processors, where a number of worlds are ready to run on one processor while none on the other. Without load balancing, such imbalance would persist. As a result, the ready worlds accrue unnecessary scheduling latency and the CPU utilization becomes only half of what could be attainable.

On ESX, the world migration can be initiated by either a pCPU, which becomes idle, or a world, which becomes ready to be scheduled. The former is also referred to as pull migration while the latter is referred to as push migration. With these migration policies, ESX achieves high utilization and low scheduling latency.

However, the migration incurs cost. When a world migrates away from the source pCPU, where it has run awhile and brought instructions and data (the working set[2]) into the on-chip cache, the world has to bring the working set back into the cache[3] of the destination pCPU, or warm up the cache. For a workload that benefits from the cache performance, frequent migrations can be detrimental. To prevent costly migration, the CPU scheduler ensures that the migration only happens for the worlds that have not consumed enough CPU resources in the past, so that the benefit of the migration outweighs the cost.

The CPU scheduler cell on ESX imposes a constraint to the migration policy. As intercell migration is deemed more expensive than intracell migration, the former only happens in much coarser granularity. Whether this constraint has positive or negative performance impact heavily depends on workload behavior. Section 3.3 discusses this issue in detail.

2.4.1. Load Balancing on NUMA Systems

In a NUMA (Non-Uniform Memory Access) system, there are multiple NUMA nodes that consist of a set of processors and the memory. The access to memory in the same node is local; the access to the other node is remote. The remote access takes longer cycles because it involves a multihop operation. Due to this asymmetric access latency, keeping the memory access local or maximizing the memory locality improves performance. On the other hand, CPU load balancing across NUMA nodes is also crucial to performance.

The NUMA load balancer in ESX assigns a home node to a virtual machine. For the virtual machine, the memory is allocated from the home node. Because the virtual machine rarely migrates away from the home node, the memory access from the virtual machine is mostly local.

NOTE: All vCPUs of the virtual machine are scheduled within the home node.

If a virtual machine's home node is more heavily loaded than others, migrating to a less loaded node generally improves performance, although it suffers from remote memory accesses. The memory migration can also happen to increase the memory locality.

NOTE: The memory is moved gradually because copying memory has high overhead. See the vSphere Resource Management Guide [1] for more details.

2.4.2. Load Balancing on Hyperthreading Architecture

Hyperthreading enables concurrently executing instructions from two hardware contexts in one processor. Although it might achieve higher performance from thread-level parallelism, the improvement is limited because the total computational resource is still capped by a single physical processor. Also, the benefit is heavily workload dependent.

It is clear that a whole idle processor, which has both hardware threads idle, provides more CPU resources than only one idle hardware thread with a busy sibling thread. Therefore, the ESX CPU scheduler makes sure the former is preferred as the destination of a migration. ESX provides an option that controls how hardware threads are to be utilized. See the *vSphere Resource Management Guide* [1] for more details.

[2] The working set is usually defined as the amount of memory that is actively accessed for a period of time. In this paper, the working set conveys the same concept, but for the on-chip cache that contains data and instructions.

[3] The on-chip cache is part of the processor. From now on, the cache refers to on-chip cache or processor cache.

3. CPU Scheduler Changes in ESX 4

In ESX 4, the CPU scheduler has undergone several improvements for better performance and scalability. While the proportional-share scheduling algorithm has remained the same, the changes introduced in ESX 4 might impact the performance noticeably. This section describes the major changes in the CPU scheduler and discusses their performance impact.

3.1. Further Relaxed Co-Scheduling

In ESX 4, the relaxed co-scheduling algorithm has been refined such that the scheduling constraint due to the co-scheduling requirement is even further reduced. See Section 2.2 for the background of this discussion.

First of all, the notion of the progress of a virtual machine has been refined. Previously, a virtual machine was considered to make progress if it consumed CPU or halted. This includes the time it spent in the hypervisor. Enforcing synchronous progress including the hypervisor layer is too restrictive because the correctness aspect of the co-scheduling only matters in terms of guest-level progress. Also, the time spent in the hypervisor might not be uniform across vCPUs, which unnecessarily increases the skew.

In ESX 4, a virtual machine is considered to make progress if it consumes CPU in the guest level or halts. The time spent in the hypervisor is excluded from the progress. This means that the hypervisor execution might not always be co-scheduled. This is acceptable because not all operations in the hypervisor will benefit from being co-scheduled. When it is beneficial, the hypervisor makes explicit co-scheduling requests to achieve good performance.

Secondly, the methodology of measuring the accumulated skew has been refined. Previously, the accumulated skew for a vCPU grows if it does not make progress while any of its sibling vCPUs makes progress. This can overestimate the skew and result in unnecessary co-scheduling overhead. For example, consider a case where two sibling vCPUs, v0 and v1, make equal progress but at different moments. While there is essentially no skew between the two vCPUs, the skew still grows.

In ESX 4, the progress of each vCPU in a virtual machine is tracked individually and the skew is measured as the difference in progress between the slowest vCPU and each of the other vCPUs. In the previous example, the skew does not grow in ESX 4 as long as the two vCPUs make equal progress within a period, during which the co-scheduling is enforced. This accurate measurement of the skew eliminates unnecessary co-scheduling overhead.

Finally, the co-scheduling enforcement becomes a per-vCPU operation. Previously, the entire virtual machine was stopped (co-stop) when the accumulated skew exceeded the threshold. The virtual machine was restarted (co-start) only when enough pCPUs were available to accommodate the vCPUs that lagged behind.

In ESX 4, instead of stopping or starting a set of vCPUs, only the vCPUs that advanced too much are individually stopped. Once the lagging vCPUs catch up, the stopped vCPUs can start individually. Co-scheduling all vCPUs is still attempted to maximize the performance benefit of co-scheduling.

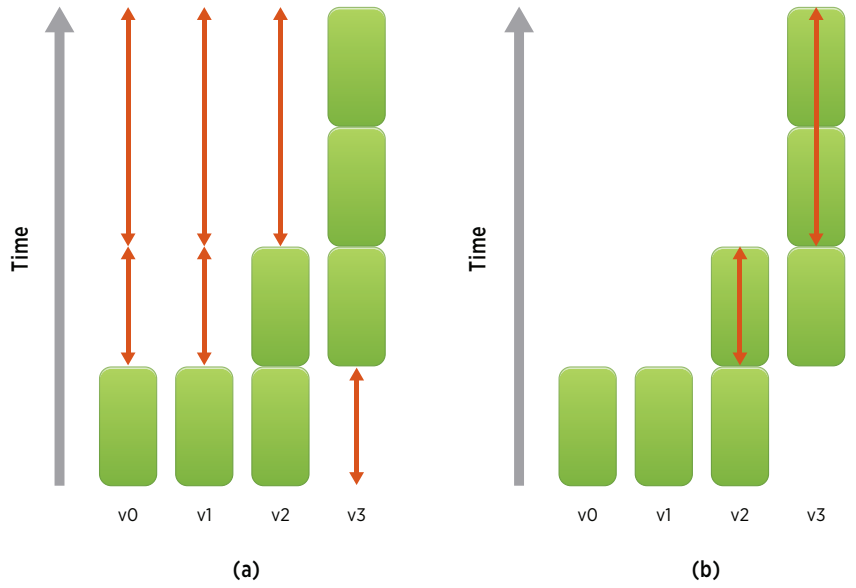

Figure 1. Illustration of Measuring the Accumulated Skew in (a) ESX 3.x and (b) ESX 4

Figure 1 illustrates how the skew is measured differently in ESX 3.x and ESX 4, using the example of a 4-vCPU virtual machine. CPUs v0 through v3 denote four vCPUs of the virtual machine, and the time goes upward. The green bars for each vCPU represent the duration while the vCPU makes progress. The vCPUs make progress at different times. This might happen when the system is overcommitted or interrupted. The size of the bars is a multiple of unit time, T. The red arrows represent the amount of skew accumulated for the corresponding vCPU.

Figure 1 (a) shows that the accumulated skew for v0 and v1 is 3T; the actual difference in the guest progress is only 2T. This is a 50-percent overestimation of the skew and might result in unnecessary co-scheduling overhead. Figure 1 (b) shows the new implementation that tracks the difference in progress from the slowest one.

NOTE: The skew for v3 is 2T, which is an accurate estimation.

In this example, assume that the skew between v0/v1 and v3 is greater than the threshold. The old implementation would co-stop the entire virtual machine and require v0 and v1 to be scheduled together until the skew decreased sufficiently. In the new implementation, only the vCPU that advanced too much would be stopped. Scheduling multiple vCPUs together is no longer required. Table 1 compares available scheduling choices.

NOTE: ESX 4 has more scheduling choices. This leads to higher CPU utilization.

	ESX 3.X	ESX 4.X
SCHEDULING CHOICES	(v0, v1, v2, v3)	(v0, v1, v2, v3)
	(v0, v1, v2)	(v0, v1, v2)
	(v0, v1, v3)	(v0, v1, v3)
	(v0, v1)	(v0, v1)
		(v0, v2)
		(v1, v2)
		(v0)
		(v1)
		(v2)

Table 1. Scheduling Choices When a Multiprocessor Virtual Machine Is Skewed Where v3 Advanced Too Much Compared to Both v0 and v1

3.2. Elimination of CPU Scheduler Cell

As discussed in Section 2.3, previous versions of ESX have been using the CPU scheduler cell to achieve good performance and scalability. Although it has been working quite well so far, this approach might limit the scalability in the future.

First, the cell size must be increased to accommodate wider multiprocessor virtual machines. The width of a virtual machine is defined as the number of vCPUs on the virtual machine. In ESX 4, a virtual machine can have up to eight vCPUs, which means the size of a cell would also be increased to eight. On a host with eight pCPUs, there would be only a single cell, with the cell lock effectively becoming the global lock. This would serialize all scheduler invocations and seriously limit the scalability of ESX.

Also, the cell approach might unnecessarily limit the amount of the cache and the memory bandwidth on state-of-the-art multicore processors with shared cache. For example, consider a case where there are two sockets each on a quad-core processor with shared cache, so that a cell[4] corresponds to a socket. If a 4-vCPU virtual machine is scheduled, it might perform better by utilizing cache or memory bandwidth from two sockets rather than one socket. Under the existing cell approach, the virtual machine can only be scheduled within a cell or, in this example, a socket. As a result, the memory subsystem on the host is not fully utilized. Depending on workloads, larger cache or higher memory bandwidth might significantly improve performance. Figure 2 illustrates this example.

NOTE: In ESX 4, a virtual machine can have higher memory bandwidth.

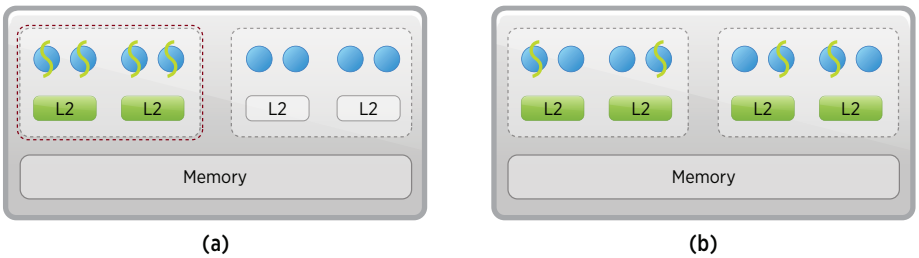

Figure 2. The Scheduler Cell Limits the Amount of the Cache and the Memory Bandwidth to a 4-vCPU Virtual Machine in (a). By Eliminating the Cell, the Virtual Machine Has Bigger Aggregated Cache and Memory Bandwidth in (b)

To overcome these limitations of the cell approach, the CPU scheduler cell is eliminated and is replaced with finer-grained locks in ESX 4. A per-pCPU lock protects the CPU scheduler states associated with a pCPU, and a separate lock per virtual machine protects the states associated with a virtual machine. With these finer-grained locks, the lock contention is significantly reduced.

3.3. Multicore-Aware Load Balancing

The CMP architecture poses an interesting challenge to the load-balancing algorithm, due to the various implementations of on-chip cache hierarchy. Previous generations of multicore processors are usually equipped with private L2 cache, while newer generations are equipped with shared L2 or L3 cache. Also, the number of cores sharing the cache varies from two to four or even higher. The last-level cache (LLC) denotes the cache beyond which the access has to go to the memory. Because the access latency to the LLC is at least an order of magnitude smaller than that of the memory, maintaining a high cache-hit[5] ratio in LLC is critical to good performance.

As previously discussed, load balancing is achieved by the migration of vCPUs. Because the migrated vCPU must warm up the cache, the cost of the migration can greatly vary depending on whether or not the warm-up requires the memory accesses that LLC cannot satisfy. In other words, the intra-LLC migration tends to be significantly cheaper than the inter-LLC migration.

Previously, this on-chip cache topology was not recognized by the CPU scheduler, and the load-balancing algorithm was considered only a fixed migration cost. This has been working fine with the scheduler cell because throttling intercell migrations results in throttling inter-LLC migrations. A cell usually matches with one or, at most, two LLCs.

Because the cell is eliminated in ESX 4, the inter-LLC migration is no longer throttled. Therefore, the load-balancing algorithm is improved such that the intra-LLC migration is always preferred to the inter-LLC migration. When the CPU scheduler looks for a pCPU as the destination of the migration, a local processor that shares LLC with the origin is preferred to a remote processor that does not share LLC. It is still possible to select the remote processor when choosing a local pCPU cannot resolve the load imbalance.

3.3.1. CPU Load-Based Migration Throttling

In ESX 4, a vCPU might not migrate if it contributes significantly high CPU load to current pCPU. Instead, a vCPU with a lower contribution tends to migrate. This is to prevent too frequent migrations due to fake imbalance of CPU load. In a fairly undercommitted environment, it is possible to have only a few processors busy. This is because there are not enough vCPUs to fully utilize the system. Blindly attempting to balance the CPU load in such a situation might result in unnecessary migrations and degrade performance.

For example, consider a case where a vCPU contributes 95 percent load of the current pCPU (P_A), and other pCPUs are idle. If the vCPU is pulled by the other idle pCPU (P_B), the original pCPU (P_A) likely becomes idle soon and wants to pull a vCPU from others as well. It might pull the same vCPU from P_B when the vCPU is temporarily descheduled. This can significantly increase the number of transient migrations. By throttling migration of the vCPU, this ping-pong situation will not happen.

The high contribution of CPU load can be translated into a large cache working set, because it has enough time to bring it into the cache. It is not ideal, but it is considered to be a reasonable approximation. Throttling migrations based on CPU load can be viewed as a necessary condition to the throttling based on the size of the cache working set. Accurately estimating the cache working set and utilizing it as a scheduling hint constitute future work.

A vCPU migration is throttled only when the vCPU contributes a significantly high portion of current pCPU load. The threshold is set high enough to allow migrations that actually improve the load balancing. As the CPU overcommitment level increases, the vCPU migrations are less likely throttled because the contribution by each vCPU decreases. This makes sense because the fake imbalance problem is only likely in the undercommitted environment.

3.3.2. Impact on Processor Caching

In ESX 4, the vCPUs of a virtual machine can be scheduled on any pCPUs because the scheduler cell has been removed. Under the load-balancing algorithm that tries to balance CPU load per LLC, the vCPUs tend to span across multiple LLCs, especially when the host is undercommitted.

NOTE: The virtual machine is still scheduled within a NUMA node if it is managed by the NUMA scheduler. See the Resource Management Guide [1] for more details on the NUMA scheduler.

The more aggregated cache and memory-bus bandwidth significantly improves the performance of most workloads. However, certain parallel workloads that have intensive communications between threads can suffer from performance loss when the threads are scheduled across distinct LLCs.

For example, consider a parallel application that has a small cache working set but very frequent producer-consumer type of communications between threads. Also, assume that the threads run on distinct vCPUs. When the host is undercommitted, the vCPUs likely span across multiple LLCs. Consequently, the communication between threads might suffer more LLC misses and degrade performance. If the cache working set were bigger than a single LLC, the default policy would likely provide better performance.

The relative benefit of providing larger cache capacity and enabling more cache sharing is very workload dependent. Also, detecting such performance dynamically and transparently is a challenging task and constitutes future work. Meanwhile, users might statically prefer cache sharing by using vSMP consolidation, which is discussed in the next section.

3.3.2.1. vSMP Consolidation

If it is certain that a workload in a virtual machine will benefit from cache sharing and does not benefit from larger cache capacity, such preference can be specified by enabling vSMP consolidation, which causes sibling vCPUs from a multiprocessor virtual machine to be scheduled within an LLC. Such preference might not always be honored, depending on the availability of pCPUs.

To enable vSMP consolidation for a virtual machine, take the following steps in vSphere Client:

1. Right-click the virtual machine and select **Edit Settings**.

2. Select the Options tab.

3. Under **Advanced**, click **General**, and on the right, click the **Configuration Parameters** button.

4. Click **Add Row**.

5. Add `sched.cpu.vsmpConsolidate` set to `true`.

3.3.2.2. Inter-VM Cache Affinity

When there are two virtual machines on the same host that communicate frequently, those virtual machines might benefit from sharing the cache. This situation applies to intermachine sharing; vSMP consolidation applies to intramachine sharing. Also, the latter applies only to multiprocessor virtual machines; the former applies to uniprocessor virtual machines as well.

The CPU scheduler can transparently detect such communicating virtual machines in the same host and attempts to schedule them in the same LLC. The attempt might fail, depending on the system load.

3.3.3. Aggressive Hyperthreading Support

Section 2.4.2 briefly describes the load balancing on a hyperthreading architecture. The migration policy is to prefer whole idle core, where both hardware threads are idle, to partial idle core, where one thread is idle while the other thread is busy. Since two hardware threads compete for a single processor, utilizing a partial core results in worse performance than utilizing a whole core. This causes asymmetry in terms of the computational capability among available pCPUs, depending on whether or not its sibling is busy.

This asymmetry degrades fairness. For example, consider a vCPU that has been running on a partial core and another vCPU running on a whole core. If two vCPUs have the same resource specification and demand, it is unfair to allow such a situation to persist.

To reflect the asymmetry, the CPU scheduler charges CPU time partially if a vCPU is scheduled on a partial core. When the vCPU has been scheduled on a partial core for a long time, it might have to be scheduled on a whole core to be compensated for the lost time. Otherwise, it might be persistently behind compared to vCPUs that use the whole core. The compensation keeps the sibling thread idle intentionally and might not reach full utilization of all hardware threads. However, the impact should be low because the added benefit of the extra hardware thread is limited.

Internal testing shows that the recent generation of hyperthreading processors displays higher performance gain and lower interference from the neighboring hardware thread. This observation encourages a more aggressive use of partial core in load balancing. The whole core is still preferred to the partial core as a migration destination; but if there is no whole core, a partial core is more likely to be chosen. Previously, a partial core might not have been chosen, to make sure that fairness would not be affected. Experiments show that the aggressive use of partial core improves the CPU utilization and application performance without affecting the fairness.

It is worth noting that the accounting of CPU usage on a hardware thread is discounted if its sibling thread is busy. This causes the utilization of hardware threads to appear lower than actual. For example, consider that there are two hardware threads, with both threads busy all the time. In terms of CPU utilization, the system is fully utilized, but the accounted time is less than full utilization because of the discount. A new counter, “PCPU Util”, has been introduced in esxtop to show the system utilization; “PCPU Used” still

shows the current accounted time. Check the esxtop man page for more details.

3.3.4. Extended Fairness Support

As mentioned in Section 2.1, the basic scheduling algorithm implemented in the ESX CPU scheduler ensures fair allocation of CPU resources among virtual machines to their resource specification. In some cases, this fairness in CPU time allocation might not directly translate into the application-metric fairness because many aspects other than CPU resources, including the on-chip cache and the memory bandwidth, affect the performance.

To extend fairness support, the long-term fairness migration might happen to fairly allocate on-chip cache or memory bandwidth. Although CPU resources are fairly allocated, there still can be migrations to improve fair allocation of on-chip cache or memory bandwidth. Although there is a cost incurred from this type of migration, it happens infrequently enough, typically at a few seconds, so that the performance impact is minimized.

NOTE: The extended fairness support is also implemented in the NUMA load balancing algorithm [1].

3.4. Wide-VM NUMA Support

Wide-VM is defined as a virtual machine that has more vCPUs than the available cores on a NUMA node. For example, a 4-vCPU SMP virtual machine is considered "wide" on an AMD Opteron 82xx system (dual core) but not as such on an AMD Opteron 83xx (quad core) system. Only the cores count, and hyperthreading threads don't; that is, an 8-vCPU SMP virtual machine is considered wide on an Intel Xeon 55xx system because the processor has only four cores per NUMA node.

ESX 4.1 allows wide-VMs to take advantage of NUMA management. NUMA management means that a virtual machine is assigned a home node where memory is allocated and vCPUs are scheduled. By scheduling vCPUs on a NUMA node where memory is allocated, the memory accesses become local, which is faster than remote accesses.

Wide-VM NUMA support is accomplished by splitting a wide-VM into smaller NUMA clients whose vCPU count does not exceed the number of cores per NUMA node, and assigning a home node to each client. For example, a 4-vCPU SMP virtual machine on an AMD Opteron 82xx system has two 2-vCPU NUMA clients, and an 8-vCPU SMP virtual machine on an AMD Opteron 83xx system has two 4-vCPU NUMA clients. The wide-VM has multiple home nodes because it consists of multiple clients, each with its own home node.

The vCPU scheduling of each smaller NUMA client is performed the same as for a nonwide VM; that is, the vCPUs of the client are scheduled within the home node. However, the memory is interleaved across the home nodes of all NUMA clients of the VM. This is because the memory access pattern from a NUMA client to memory nodes can change so frequently by guest OS that any heuristics based on past memory accesses do not work well consistently. By interleaving memory allocation, the average memory access latency is expected to be better than other allocation policies based on simple heuristics.

3.4.1. Performance Impact

Typically, most of the virtual machines fit into a NUMA node and perform optimally. Since most modern processors have at least four cores per NUMA node, 4-vCPU VMs or smaller VMs would not be affected by this change. Also, both CPU and memory scheduling do not differ much on two-node NUMA systems, with or without this feature. This feature aims to improve the performance of wide SMP VMs on large systems.

Consider an 8-vCPU VM on a two-socket AMD Opteron 83xx (quad core) system. Assuming uniform memory access pattern, about 50% of memory accesses will be local, because there are two NUMA nodes without wide-VM NUMA support. In this case, wide-VM NUMA support won't make much performance difference. On a four-socket system, only about 25 percent of memory accesses will be local; wide-VM NUMA support improves it to 50 percent local accesses. The performance benefit will be even greater on bigger systems.

In an undercommitted case, wide-VM NUMA support might result in underutilizing on-chip cache resource or the memory bandwidth. Consider a case where there is only a single 8-vCPU VM running on a four-socket quad-core NUMA system. In that case, the default wide-VM NUMA policy assigns two NUMA nodes to the tCPU VM, leaving the other two nodes not utilized.

This case should be rare in typically consolidated servers. Also, in moderately loaded systems, improved memory locality tends to outweigh the lower utilization of the on-chip cache or the memory bandwidth. For this reason, the default policy is to split the wide VMs into as few NUMA nodes as possible. However, this policy can be overridden by setting the advanced option numa.vcpu.maxPerMachineNode, which limits the number of vCPUs per NUMA client. If the parameter is set to 2 on a quad-core NUMA system, an 8-vCPU VM will be split so as to use four NUMA nodes and to fully use on-chip cache and the memory bandwidth in an undercommitted case. The performance trade-off should be carefully evaluated per workload. To set the maxPerMachineNode parameter:

1. Right click the virtual machine and select **Edit Settings**.
2. Select the **Options** tab.
3. Under **Advanced**, click **General**, on the right, click the **Configuration Parameters** button.
4. Click **Add Row**.
5. Add `numa.vcpu.maxPerMachineNode` set to new value.

4. Performance Evaluation

This section presents data that verifies the effectiveness of CPU resource controls, including shares, reservation and limit. Then it compares different coscheduling algorithms. Lastly, this section evaluates the performance impact of CPU scheduler changes in ESX 4.

4.1. Experimental Setup

The following table summarizes the experimental setup.

HOST1	Dell PE 2950, 2-socket quad-core Intel Xeon 5355, 32GB 4MB L2 shared by two cores		
HOST2	Dell PE R905, 2-socket quad-core AMD Opteron 8384, 64GB 4 pCPUs, 32GB per NUMA node		
HOST3	Dell PE 6950, 4-socket dual-core AMD Opteron 8222, 64GB 10MB private L2 cache, 2 pCPUs, 16GB per NUMA node		
HOST4	Dell PE R905, 4-socket quad-core AMD Opteron 8378, 72GB 6MB L3 shared by four cores , 4 pCPUs, 18GB per NUMA node		
GUEST	Red Hat Enterprise Linux 5.1, x64 1, 2, 4 vCPUs; 1, 2, 4GB		
WORKLOAD	SPECjbb 2005	JVM Java_heap Warehouses Runtime	jrockit-R27.5.0-jre1.5.0_14 50% of guest memory size Equal to the number of vCPUs 10 minutes, repeat three times
	kernel-compile	Command #threads	make –j #threads bzImage Equal to twice the number of vCPUs

Table 2. Hosts, Guest, and Workload Setups

NOTE: Most experiments are conducted on Host1; Host2 is used for NUMA-related experiments only.

In this paper, mostly CPU-intensive workloads are used. For larger-scale workloads, refer to other white papers [6] [7]. For wide-VM NUMA support, Host3 and Host4 are used. Both systems have four NUMA nodes.

4.2. Verifying the Proportional-Share Algorithm

The goal of this experiment is to verify whether the proportional-share scheduling algorithm works as designed in ESX 4. Although there is no change in the basic algorithm, it should be useful to confirm that ESX CPU scheduler accurately reflects the user-specified resource allocation.

4.2.1. Shares

To verify whether the CPU time is allocated proportionally according to the shares, different shares are assigned to two 4-vCPU virtual machines, vm0 and vm1, on a 4-pCPU host. A multithreaded CPU-intensive micro benchmark runs in both virtual machines, making all four vCPUs busy.

Shares to vm0 and vm1 are given in a way such that the ratio of the shares between vm0 and vm1 is 1:7, 2:6, 3:5, 4:4, 5:3, 6:2, and 7:1. Both virtual machines are busy; CPU time used by each virtual machine is measured and plotted in Figure 3.

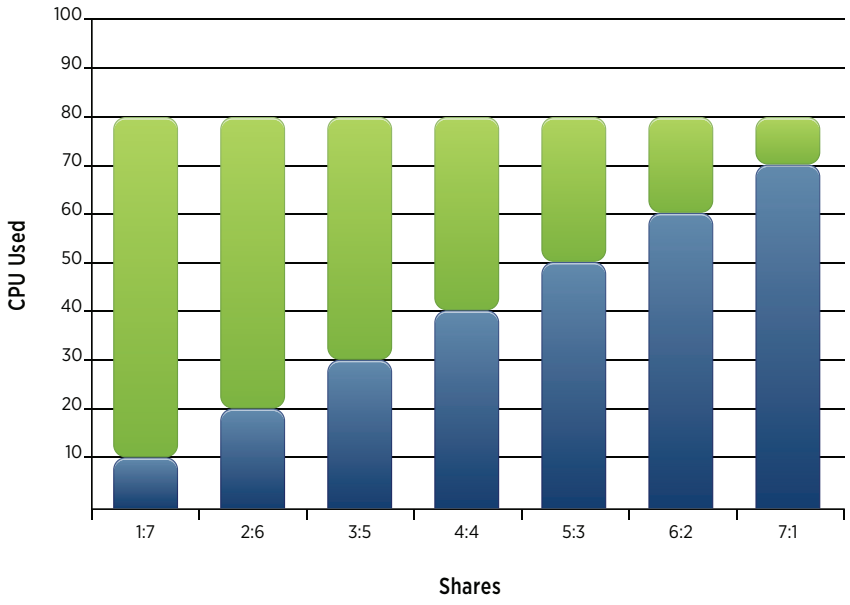

Figure 3. CPU Time Between vm0 and vm1 Having Different Shares with the Ratios of 1:7, 2:6, 3:5, 4:4, 5:3, 6:2, and 7:1.

As you can see from Figure 3, CPU time is distributed between two virtual machines, proportional to the shares. For example, when the ratio of shares between vm0 and vm1 is 1:7, 1/8 CPU time is given to vm0 and 7/8 CPU time is given to vm1. This is consistent across all ratios.

4.2.2. Reservation

A reservation specifies the guaranteed minimum allocation for a virtual machine. To verify whether the reservation properly works, a 1-vCPU virtual machine, vm0, reserves the full amount of CPU. On a 4-pCPU host, vm0 runs with a varying number of virtual machines that are identical to vm0 but do not have any reservation. The number of such virtual machines varies from one to seven.

Figure 4 shows the CPU used time of vm0 and the average CPU used time of all other virtual machines. The CPU used time is normalized to the case where there is only vm0. With up to three other virtual machines, there would be little contention for CPU resources. CPU time is therefore fully given to all virtual machines. All virtual machines have the same shares. However, as more virtual machines are added, only vm0, with the full reservation, gets the consistent amount of CPU, while others get a reduced amount of CPU.

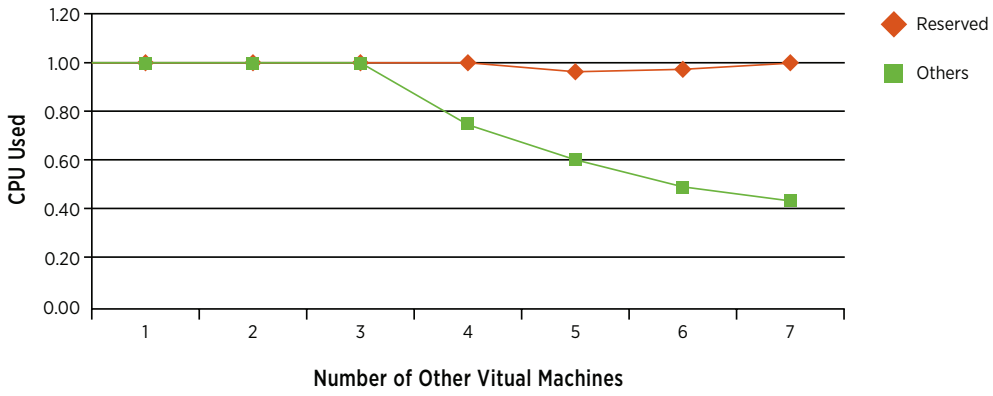

Figure 4. CPU Time of a Fully Reserved Virtual Machine and the Average CPU Times of Others Without Reservation

4.2.3. Limit

A limit specifies an upper boundary for CPU resources that can be allocated to a virtual machine. To verify that the limit works correctly, two identical 4-vCPU virtual machines, vm0 and vm1, run on a 4-pCPU host, where vm0 is limited to varying amounts of CPUs from 50 to 200 percent. The host has 400 percent available because there are four pCPUs.

Figure 5 shows the relative CPU time used by vm0 and vm1; vm0 does not consume more than the limit specified. The total used time is scaled to 400 percent.

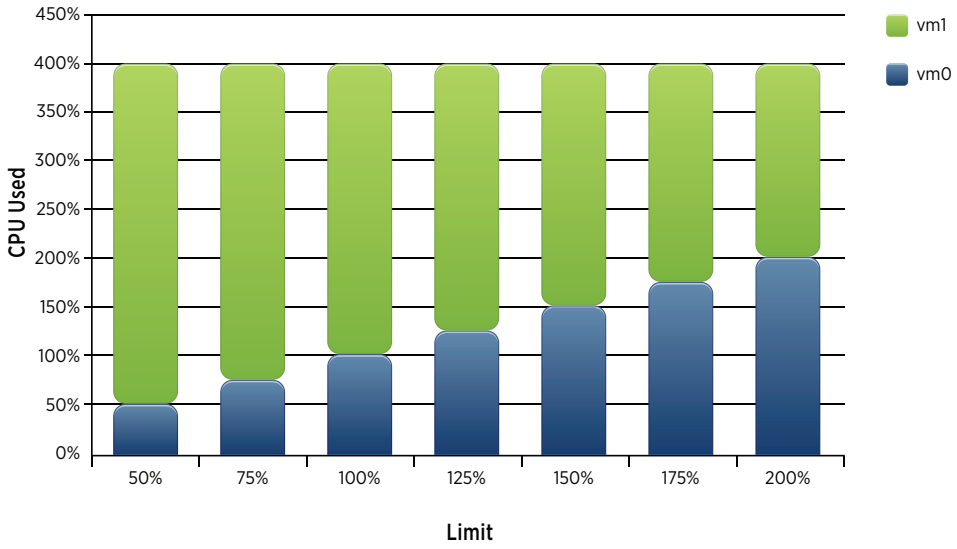

Figure 5. Relative CPU Time Used by vm0 and vm1 When vm0 Is Limited by 50% to about 200%

4.3. Impact of Relaxed Co-Scheduling

Isolating the impact of the co-scheduling algorithm is difficult because it heavily depends on the workload and the load of the system, which vary dynamically. In this experiment, a 2-vCPU and a one-vCPU virtual machine are affined to two pCPUs so as to stress the co-scheduling algorithm. When the one-vCPU virtual machine utilizes a pCPU, the other pCPU might or might not be utilized, depending on the co-scheduling algorithms.

To evaluate whether the relaxed co-scheduling introduced in ESX 3.x really improves utilization, strict co-scheduling is emulated in ESX 3.5 and is used as a baseline. The emulation is not representative of releases prior to ESX 3.x.

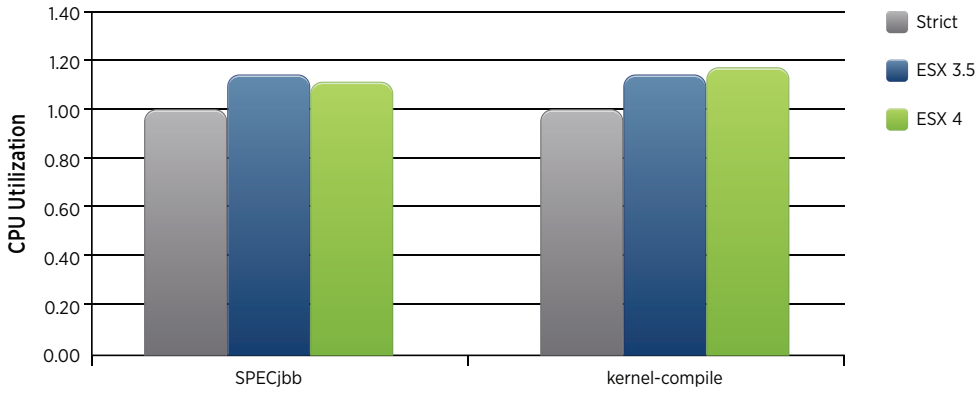

Figure 6. CPU Utilization Normalized to the Strict Co-Scheduling Emulation

Figure 6 compares CPU utilization between ESX 3.5 and ESX 4. Both are normalized to the emulation of the strict co-scheduling algorithm. The results clearly show that the relaxed algorithm achieves higher utilization. This higher utilization improves performance. Figure 7 compares the relative performance of ESX 3.5 and ESX 4, which is normalized to the strict algorithm.

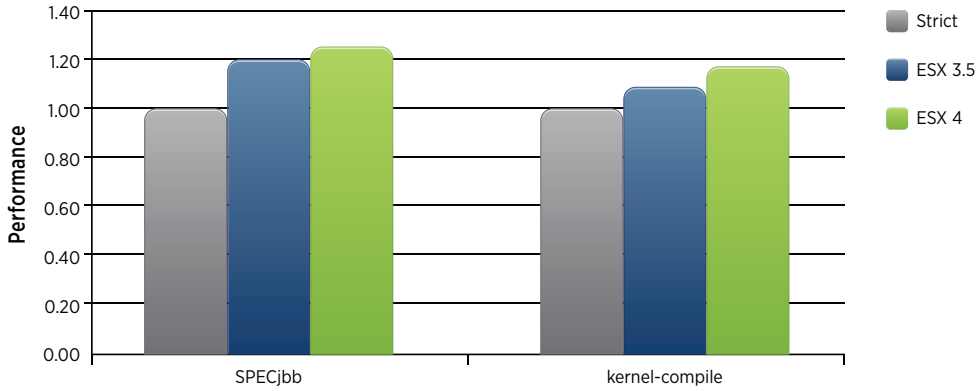

Figure 7. Performance Normalized to the Strict Co-Scheduling Emulation

Although the CPU utilization is about the same with ESX 3.5 and ESX 4, the performance is higher in ESX 4. This is mostly because ESX 4 achieves the same utilization with less co-scheduling overhead. Figure 8 compares the number of co-stops between ESX 3.5 and ESX 4. It is clear from the figure that ESX 4 has far fewer co-stops than ESX 3.5. Because the co-stop means that vCPUs are not schedulable due to the co-scheduling restriction, fewer co-stops indicate less co-scheduling overhead.

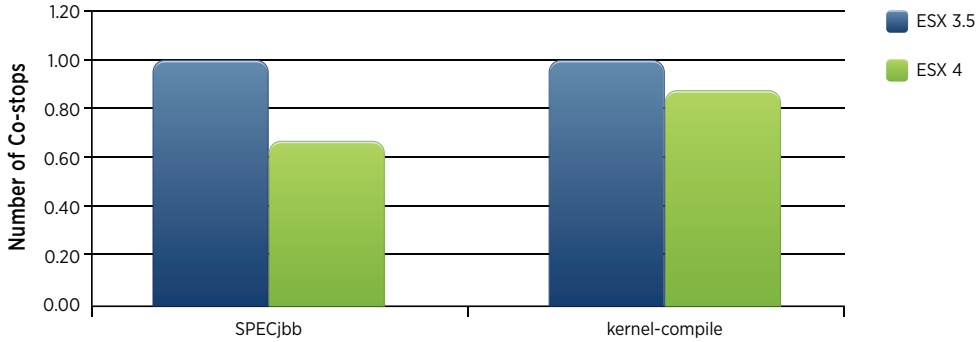

Figure 8. Comparison of the Number of Co-Stops Between ESX 3.5 and ESX 4

The results indicate that ESX 4 achieves high utilization with less co-scheduling overhead compared to ESX 3.5. Depending on the workload and the system load, the resulting performance improvement can be significant.

4.4. Impact of Multicore-Aware Load Balancing

The benefit of multicore-aware load balancing is most pronounced when the host is lightly utilized; that is, some pCPUs are idle. If all pCPUs are fully utilized, it is highly likely that the on-chip cache and the memory bandwidth are already saturated. Secondly, a multiprocessor virtual machine would benefit more because it used to be confined in a scheduler cell.

Figure 9 shows the aggregated throughput of SPECjbb and kernel-compile workloads normalized to ESX 3.5. A mix of one-, two-, 4-vCPU virtual machines run on a two-socket quad-core Intel Xeon system. The cache architecture of the system is similar to Figure 2.

The result clearly shows that utilizing more aggregated cache and memory bandwidth improves the performance of tested workloads. Especially, SPECjbb throughput has significantly improved, up to 50 percent. The improvement of kernel compile is modest, up to 8 percent. The benefit heavily depends on workloads. As discussed in Section 3.3.2, some workloads might benefit from being scheduled within LLC. Default scheduling policy might need to be reviewed carefully.

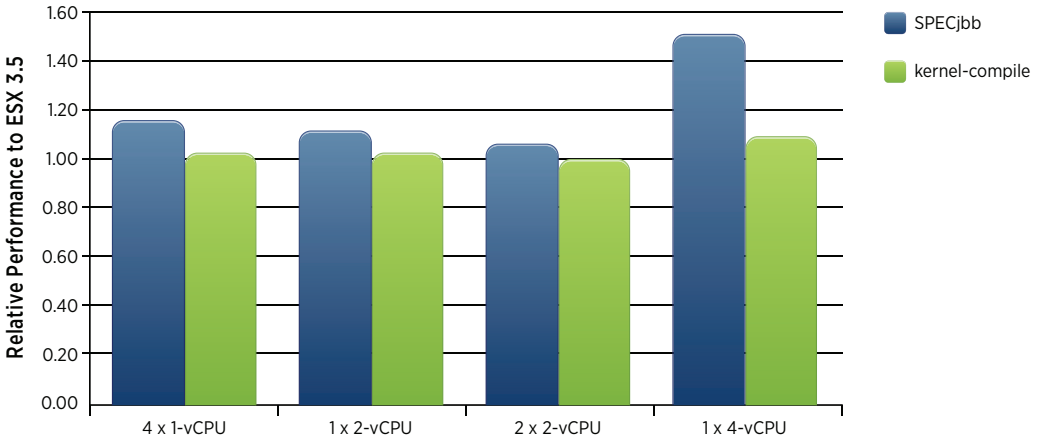

Figure 9. Relative Performance Improvement to ESX 3.5 When Virtual Machines Are Provided with Larger Aggregated On-Chip Cache and Memory Bandwidth

4.5. Impact of Extended Fairness Support

The proportional share–based algorithm in ESX allocates CPU time to virtual machines fairly, according to their resource specification. Experimental results in Section 4.2 corroborate that the algorithm works as designed. However, the performance of the virtual machines might depend on other resources, such as on-chip cache or the bandwidth of memory, and might perform disproportionately. Section 3.3.4 discusses the extended fairness support introduced in ESX 4.

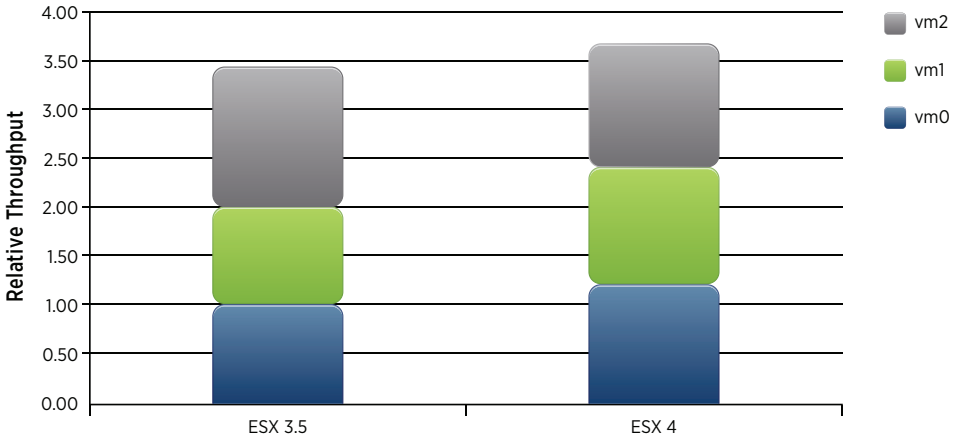

Figure 10. Relative Throughput of SPECjbb2005 Benchmark of Three 2-vCPU Virtual Machines on a Two-Socket Quad-Core System—Throughput Is Normalized to vm0 of ESX 3.5

Figure 10 shows the impact of the extended fairness support in ESX 4. On a two-socket quad-core Intel Xeon system, three 2-vCPU virtual machines run the SPECjbb2005 workload. Because the three virtual machines run identical workloads and have equal shares, they are expected to generate the same throughput. However, vm2 in ESX 3.5 generated 40 percent higher throughput. This is because vm2 is mostly scheduled in one socket, and vm0 and vm1 are scheduled in the other socket. Although not presented, CPU time is still fairly allocated among all three virtual machines. In ESX 4, all three virtual machines generate the same throughput because the ESX CPU scheduler also considers the fair allocation of the cache and the memory bandwidth.

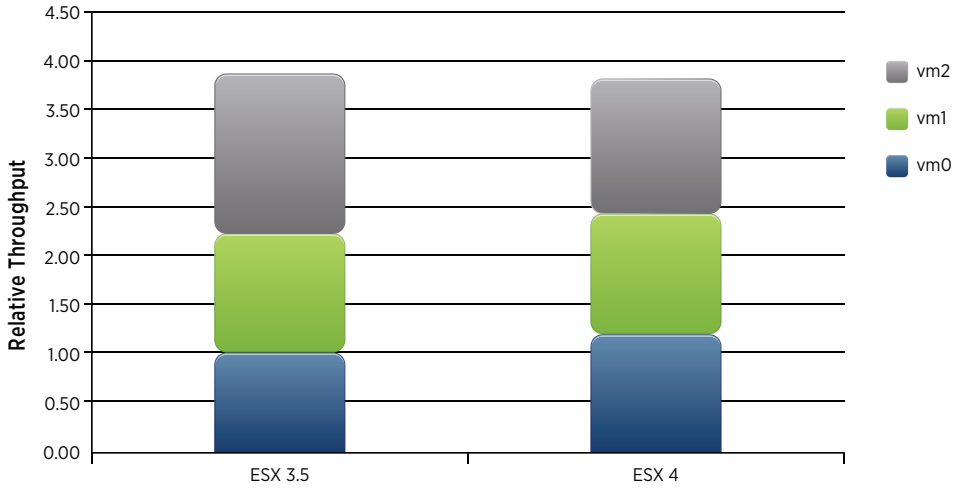

Figure 11. Relative Throughput of SPECjbb2005 Benchmark of Three 4-vCPU Virtual Machines on a Two-Node NUMA System—Each Node Has Four Cores—Throughput Is Normalized to vm0 of ESX 3.5

Figure 11 shows the extended fairness support in the NUMA scheduler. On a two-socket quad-core AMD Opteron system, three 4-vCPU virtual machines run SPECjbb2005 benchmarks. Each socket forms a NUMA node. Likewise, three virtual machines are expected to generate equal throughput. On ESX 3.5, vm1 and vm2 generate higher throughput compared to vm0. Because there are three virtual machines active on two NUMA nodes, one node is occupied by two virtual machines at any moment, which results in less cache and memory bandwidth. On ESX 4, the three virtual machines generate almost equal throughput.

4.6. ESX 4 in Comparison to ESX 3.5

This section compares the performance of ESX 4 and ESX 3.5 at various levels of system load. The system load is defined as the number of active vCPUs on a host. To vary the system load, different numbers of virtual machines were tested. Also, different mixes of 1-, 2-, and 4-vCPU virtual machines were used to achieve the same load. The following table explains the mix of test configurations.

LABEL	DESCRIPTION	TOTAL NUMBER OF VCPUS
{m} x {n} vCPU	{m} instances of {n}-vCPU virtual machines	{m} x {n}
{m} x4vCpuT	{m} instances of {2x1vCPU + 1x2vCPU}	{m} x 4
{m} x8vCpuT	{m} instances of {2x1vCPU + 1x2vCPU + 1x4vCPU}	{m} x 8
{m} x16vCpuT	{m} instances of {2x1vCPU + 1x2vCPU + 1x4vCPU + 1x8vCPU}	{m} x 16

Table 3. Test Configuration Mix

Figure 12 shows the throughput of SPECjbb2005 workload in ESX 4, normalized to that of ESX 3.5. The X-axis represents various configurations sorted from light to heavy loads. The Y-axis represents the normalized aggregated throughput. A bar that is greater than 1 means improvements over ESX 3.5.

It is clear from the result that ESX 4 scheduler achieves significant performance improvement in both lightly loaded and heavily loaded cases. When the system is lightly loaded, the intelligent load-balancing algorithm achieves higher performance by maximizing on-chip cache resources and memory bandwidth. When the system is heavily loaded, improved co-scheduling and lower locking overhead improve performance.

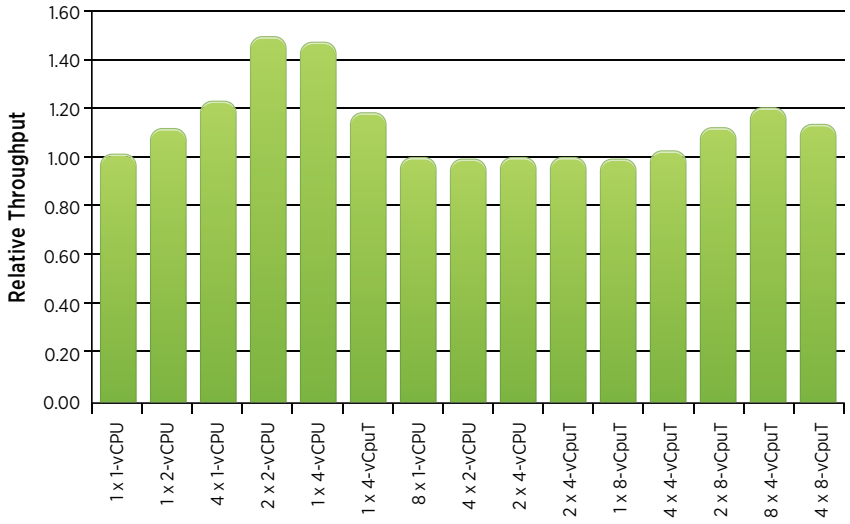

Figure 12. ESX 4 Throughput of SPECjbb2005 Normalized to ESX 3.5—Greater than 1 Means Improvement

Figure 13 shows the throughput of the kernel-compile workload normalized to ESX 3.5. The same configurations were used as SPECjbb2005. Similarly, performance is improved in both lightly loaded and heavily loaded configurations. The improvement is not as high as that of the SPECjbb2005 workload. This is probably because kernel-compile is less sensitive to its cache performance. Therefore, the impact of having higher memory bandwidth is less beneficial to this workload. Still, approximately 7–11 percent improvement is observed in many cases.

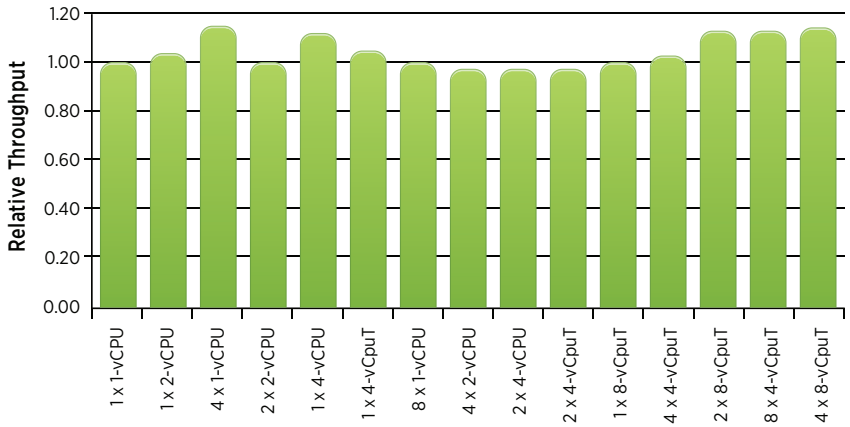

Figure 13. ESX 4 Throughput of Kernel-Compile Workload Normalized to ESX 3.5—Greater than 1 Means Improvement

Although not supported, ESX 3.5 can run 8-vCPU virtual machines by increasing the scheduler cell size. Figure 14 compares the performance of SPECjbb2005 (a) and kernel-compile (b) between ESX 4 and ESX 3.5 using configurations that include 8-vCPU virtual machines.

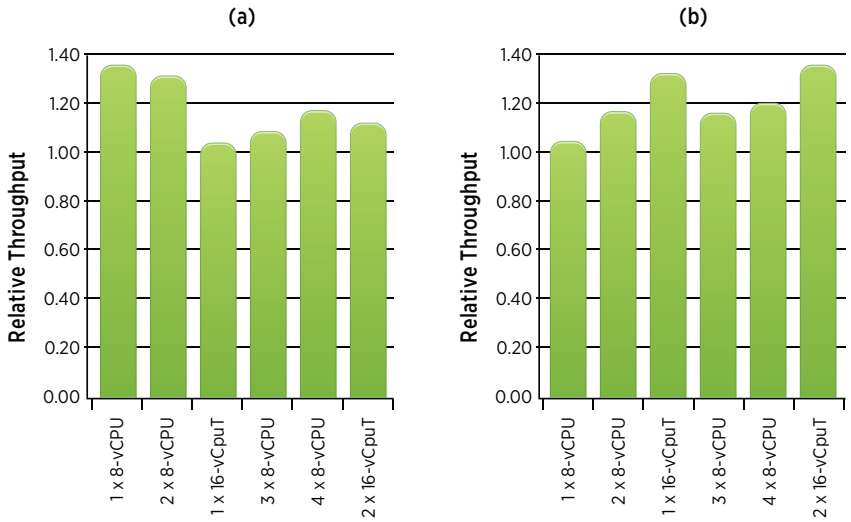

Figure 14. The Aggregated Throughput of SPECjbb2005 (a) and Kernel-Compile (b) in ESX 4 Normalized to ESX 3.5 with Configurations That Include 8-vCPU Virtual Machines—8-vCPU Virtual Machines Are Not Supported in ESX 3.5

It is clear from the results that ESX 4 provides much better performance compared to ESX 3.5. Depending on workloads and configurations, the improvement can be higher than 30 percent.

4.7. Impact of Wide-VM NUMA Support

To show the performance impact of wide-VM NUMA support, systems with four NUMA nodes (Host3 and Host4) were chosen and CPU-intensive applications were tested. Section 3.4.1 discusses why wide-VM NUMA support has a performance impact on NUMA systems with more than two nodes. For Host3, a 4-vCPU VM was tested because the system had two cores per NUMA node; for Host4, an 8-vCPU VM was tested (four cores per NUMA node).

Figure 15 shows the relative performance of SPECjbb2005 and kernel-compile on Host3. The baseline performance was collected on ESX 4. Because the SPECjbb2005 workload heavily accesses memory, it significantly benefits from wide-VM NUMA support. ESX 4.1 gives 11 percent higher throughput when there are two 4-vCPU VMs. Meanwhile, the kernel-compile workload does not benefit much from this feature because the workload is not memory intensive. The improvement becomes negligible in SPECjbb2005, probably due to the underutilization of on-chip cache resources and the memory bandwidth. For a similar reason, kernel-compile slowed down about 2 percent, although it is quite close to workload variation.

Figure 16 shows relative performance from the similar experiments on Host4. The baseline is also ESX 4. The performance trend is quite similar to that of Host3. SPECjbb2005 benefits most from wide-VM NUMA support, about a 17 percent increase in the aggregated throughput. However, the throughput drops by 2 percent in a single-VM case. Kernel-compile is improved by up to 5

percent.

Figure 17 shows the impact of changing the default behavior of splitting wide-VMs. By default, wide-VMs are split into the fewest possible NUMA nodes to maximize memory locality. By setting `numa.vcpu.maxPerMachineNode` to 2 on Host4, an 8-vCPU VM uses four nodes. Because SPECjbb2005 greatly benefits from larger on-chip cache, using 4 nodes increased the throughput by 9 percent. However, adding a second 8-vCPU VM not only removes the benefit but actually yields worse throughput, due to the poor memory locality. Also, kernel-compile does not benefit at all by using more nodes. The default configuration works well in most cases. Although not shown, the benefit of overriding the default policy on Host3 is negligible, probably because the last-level cache is private, so there is no benefit of a larger cache by using four NUMA nodes.

By improving the memory locality, wide-VM NUMA improves the performance of a memory-intensive workload. The benefit is more pronounced when the host becomes moderately utilized. Finally, the benefit varies depending on underlying hardware architectures.

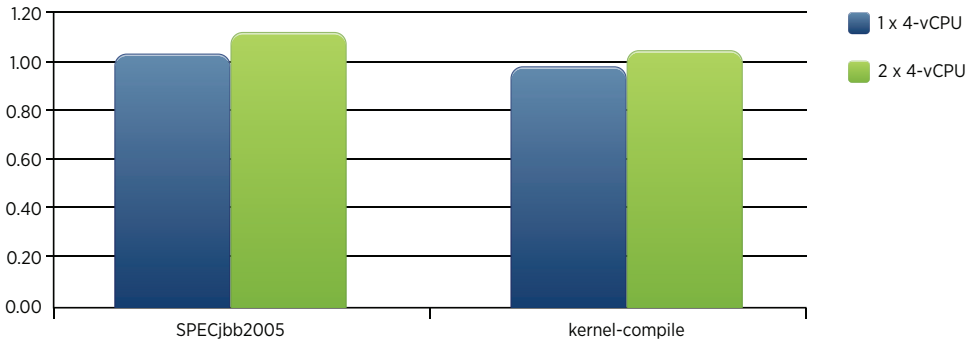

Figure 15. Relative Performance Normalized to ESX 4 on Four-Socket Dual-Core AMD Opteron (Greater than 1 Means Improvement)

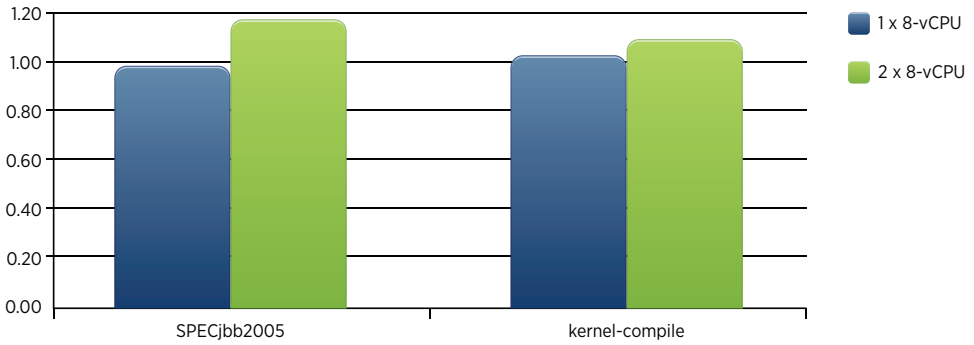

Figure 16. Relative Performance Normalized to ESX 4 on Four-Socket Quad-Core AMD Opteron (Greater than 1 Means Improvement)

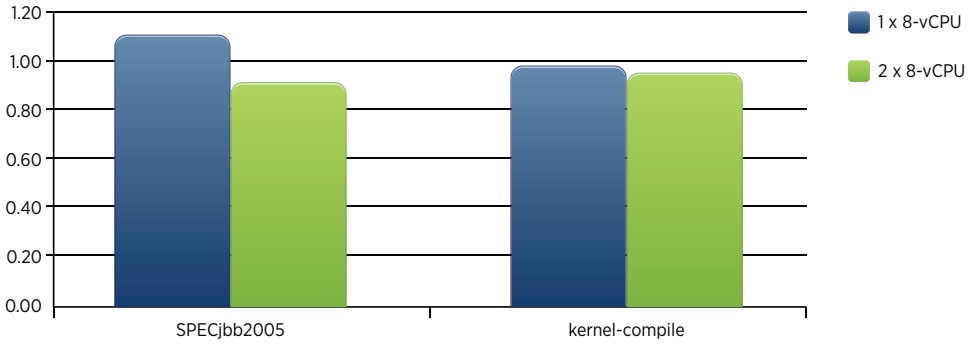

Figure 17. Impact of Setting numa.vcpu.maxPerMachineNode to 2 on Host4, Normalized to the Default (Greater than 1 Means Improvement)

5. Summary

In ESX 4, many improvements have been introduced in the CPU scheduler. This includes further relaxed co-scheduling, lower lock contention, and multicore-aware load balancing. Co-scheduling overhead has been further reduced by the accurate measurement of the co-scheduling skew and by allowing more scheduling choices. Lower lock contention is achieved by replacing scheduler cell lock with finer-grained locks. By eliminating the scheduler cell, a virtual machine can get higher aggregated cache capacity and memory bandwidth. Lastly, multicore-aware load balancing achieves high CPU utilization while minimizing the cost of migrations.

Experimental results show that the ESX 4 CPU scheduler faithfully allocates CPU resources as specified by users. While maintaining the benefit of a proportional-share algorithm, the improvements in co-scheduling and load-balancing algorithms are shown to benefit performance. Compared to ESX 3.5, ESX 4 significantly improves performance in both lightly loaded and heavily loaded systems.

ESX 4.1 introduces wide-VM NUMA support, which improves memory locality for memory-intensive workloads. Based on testing with micro benchmarks, the performance benefit can be up to 11–17 percent.

6. References

[1] VMware, Inc., *vSphere Resource Management Guide*, http://www.pubs.vmware.com

[2] William Stallings, *Operating Systems*, Prentice Hall.

[3] D. Feitelson and L. Rudolph, *Mapping and Scheduling in a Shared Parallel Environment Using Distributed Hierarchical Control*, International Conference on Parallel Processing, Vol. I, pp. 1–8, August 1990.

[4] VMware, Inc., *Co-Scheduling SMP VMs in VMware ESX Server*, http://www.communities.vmware.com/docs/DOC-4960

[5] VMware, Inc., *Scheduler Cell Size on Six-Core Processors*, http://www.kb.vmware.com/kb/1007361

[6] VMware, Inc., *Virtualized SAP Performance with VMware vSphere 4*, http://www.vmware.com/resources/techresources/10026

[7] VMware, Inc., *Microsoft Exchange Server 2007 Performance on VMware vSphere 4*, http://www.vmware.com/resources/techresources/10021

vmware

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com