Spirent™

WHITE PAPER

Key Considerations for Virtual Infrastructure Benchmarking

WHITE PAPER

Key Considerations for Virtual Infrastructure Benchmarking

The world of computing is changing at a rapid pace. Cloud Computing, Software—Designed Data Centers (SDDC), and Network Functions Virtualization (NFV) all rely heavily on Virtualization technologies to deliver their benefits. Virtualization not only enables new functionalities like NFV, VM migration and elastic computing, but also brings with it a tremendous opportunity for capital and operational cost savings. Most of the savings come from utilization optimization. By redeploying more functionality into software over hypervisors using COTS hardware, and also consolidating the number of physical servers deployed, unseen levels of resource optimization and cost savings can be achieved. This hyper-optimized configuration comes at costs however; platform centralization and greater software resource scheduling complexity.

Diversity of Virtualization Platforms

Cloud Services whether SaaS, IaaS, CaaS, VDI, or PaaS are adding incredible diversity in both the number of providers and services offered. However most of this ecosystem is running on a handful of virtual platforms—or Hypervisors—from vendors such as VMware, Amazon, and Microsoft, or from an open source platform like KVM. Compared to traditional server-client applications running over traditional IP networking products, virtualization platforms are comparatively new. They are also considerably less tested in terms of network application performance as they run over standard x86 based server hardware, opposed to purpose built proprietary hardware from networking equipment vendors. In fact the longstanding demarcation between boxes hosting server applications and those performing packet forwarding operations is breaking down. Now a cluster of servers or even a single host can serve as an application server and a major network gateway simultaneously. The challenges this presents comes from the significant differences that exist between application workload processing, and network packet processing and how best to share CPU time and allocate other critical compute resources. This paper focuses on the above mentioned challenges and why testing hypervisor performance for Cloud and NFV use cases is critical, especially in a live cloud and virtual environment.

Hypervisor and Benchmarking its Performance

Virtualization introduces a number of issues in the areas of shared resource allocation, some of which have been more specifically addressed like the use of shadow tables, while others such as the scheduling algorithm implementation are not as clear cut as there is no best fit scheduler for all application job types. While the variables affecting hypervisor performance can be many, we will take a closer look at few factors in the following sections.

Virtual CPUs

Most hypervisor schedulers are modeled with fairness between VMs in mind like round robin servicing, or even a resource allocation budget that prioritizes access time for VMs that have been least active. Design differences between the guest scheduler and the hypervisor scheduler can cause significant application degradation especially for time-sensitive jobs like those associated with network applications.

Hierarchical scheduling problem, also referred to as a semantic gap, between the guest OS scheduler and the hypervisor scheduler can result in Application processing delay beyond that of normal job execution time as potentially two completely different schedulers must play a role in job scheduling. There is also the potential for conflicting scheduling designs. At the heart of the issue is incognizance of the hypervisor scheduler to that of the guest scheduler. For example, guests running Windows and MAC OS X have a multilevel feedback queue implemented. This design uses a number of FIFO queues each assigned a time quantum from lowest to highest value. This enforces shorter duration processes to be executed first in lower order queues, while longer jobs drop into queues with a higher quantum one queue at a time until execution finishes, getting less precedence with each queue change.

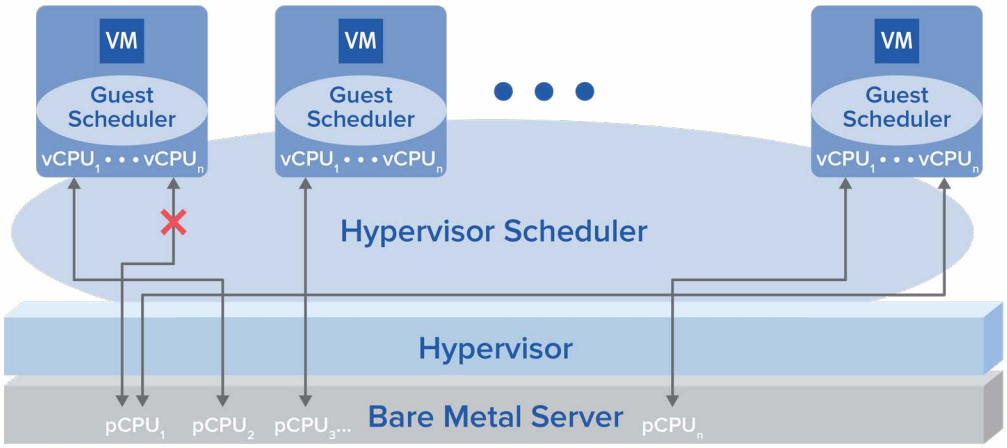

Figure 1: Blocking between two VMs on a monolithic type 1 host configured with 1-n vCPUs.

The potential issues arising from double scheduling that impact application performance are:

- Conflicting precedence between guest and hypervisor scheduling algorithms
- Preemption of critical sections on guest vCPU by hypervisor scheduler
- vCPU stacking where semaphore holder schedules after semaphore waiter
- Task priority inversion and CPU fragmentation due to co-scheduling implementations with multi-processor VMs

Conflicting precedence of processes between the guest and hypervisor operating systems can cause critical tasks to be delayed in favor of lower priority tasks. Preemption of guest processes by the hypervisor can add context switch overhead to processes that should otherwise execute uninterrupted. Virtual processor stacking can cause outright deadlocks in guest OS, forcing a VM restart to recover. Similarly, task priority inversions can cause anything from significant execution delay to exceptions due OS watchdog timer expirations. CPU fragmentation causes delay in multi-CPU VMs as preemption delays to one vCPU can still cause delays to a machine's other vCPU(s), even if underlying pCPUs are available, because they scheduled together as a group.

Virtual platform vendors have implemented or are considering various types of hybrid scheduler designs or adding heuristics to existing ones to mitigate these problems such as relaxed co-scheduling (gang scheduling), borrowed virtual time scheduling (BVT), real-time deferrable server (RTDS), or even combination approaches like using round robin or weighted round robin (WRR) algorithms depending on the process requiring CPU time for their virtual machine monitors (VMMs). The designs continue to evolve, but ultimately hypervisor scheduling comes down to basing on fairness, prioritization, or execution time, or some combination of the three. Regardless of the algorithm(s) chosen, only testing will expose the true performance of a scheduler under real workloads in large-scale environments.

Virtual Memory

One of the performance implication of virtualization is duplication of Virtual Address Translation lookups in the main secondary storage page table, and shadow table instances on VMs. Similar to double scheduling, virtual to physical addresses translations by default are performed unless new technologies are enabled such as Second Level Address Translation (SLAT) or "nested paging". Because memory does not store values contiguously, Virtual addresses are used by applications. The OS is then responsible for mapping these virtual addresses and the data stored there to the corresponding physical regions. This is known as "paging" or using a "swap file". Just as main memory uses cache to speed up the search, paged memory has Translation Lookaside Buffers (TLB) to speed up virtual to physical mapping searches. A miss in cache prompts a much slower search in system memory. A miss in the TLB, produces the slower page file search. Initially hypervisors simply replicated the page table per guest thus duplicating the mapping effort.

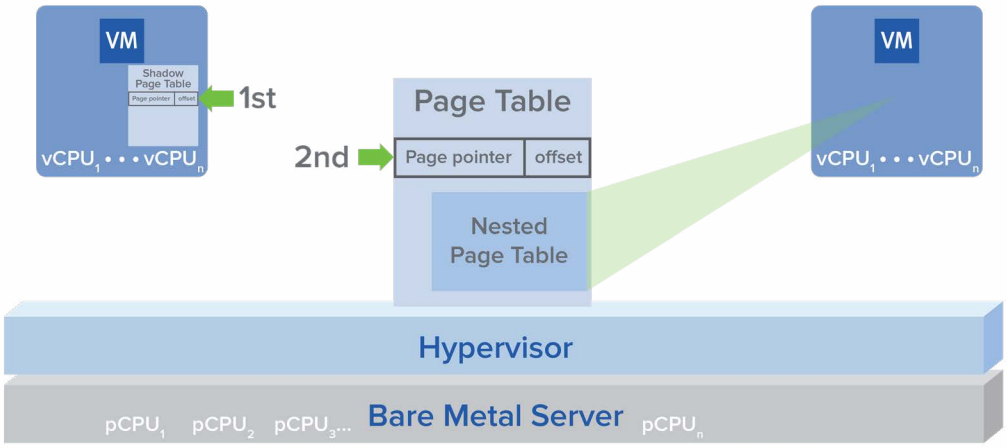

Figure 2: Duplicate address translation lookups via shadow tables vs. nested page tables.

Current hardware-assisted virtualization technology like SLAT is enabled to eliminate double page walks by preventing the overhead associated with virtual machine page table to host machine page table mappings that required frequent updating. However, there are specific application cases where the page size should be set higher to avoid performance loss with SLAT optimization active by increasing the ratio of TLB hits1. Increasing page size has the drawback however of increasing the likelihood of internal fragmentation since the number of pages needed by a process is highly variable and a process requiring only slightly more than one page wastes the majority of the second page space. This wasted page memory can be exacerbated on hosts running many VMs, and especially with different types of applications running many concurrent processes increasing the chance of page allocation failures.

[1] VMWare: Performance Evaluation of Intel EPT Hardware Assist

Storage I/O

Storage I/O performance is another aspect of virtualization that requires significant benchmarking to ensure application performance is acceptable in cloud and NFV deployments. There are many possible storage implementation types—direct attached (DAS), storage area networks (SAN), and network attached (NAS), magnetic, flash, and DRAM all of which have benefits and drawbacks in cost, performance, configuration and durability considerations. For SaaS applications, spinning disks offer acceptable performance at the most attractive price point. For NFV however, typical hardware based appliances rely heavily on NVRAM/SDRAM to perform high-speed lookup operations, and DRAM to store, retrieve and frequently modify large scale control plane tables. Therefore VNFs will require much larger footprints consuming far more DRAM than a typical server application. VNFs typically require setting processor affinity, or CPU pinning as well, to ensure proper performance that further hoards what are intended to be shared compute resources with other adjacent VMs and complicates their placement during provisioning to avoid significant performance degradation from resource contention.

Determining storage performance comes down to three key metrics: throughput, IOPS, and latency. Note each type of storage media supports a maximum data transfer rate that is part of the three metrics discussed here.

- **Throughput** is simply the maximum amount of data that storage will process with respect to time. Bandwidth is often used interchangeably, but more accurately should be described as a storage system's ability to sustain a specific throughput level over time whether that is the maximum or some level below maximum. There will be different measurements for direct attached vs. network-attached storage due to the throughput a storage network will support at a given time. This hits back to the importance of live benchmarking virtual infrastructure.
- **Input/Output Operations per Second (IOPS)** measures read and/or write operations per second and should be considered in the context of data block size (BS). IOPS will typically be higher for smaller block sizes, and lower for larger blocks sizes.
- **Latency** measures the response time of completing a storage operation, and most importantly, factors all sub components that are involved in a data transaction over virtual infrastructure: storage appliances, SAN components, internal buses, controllers, etc.

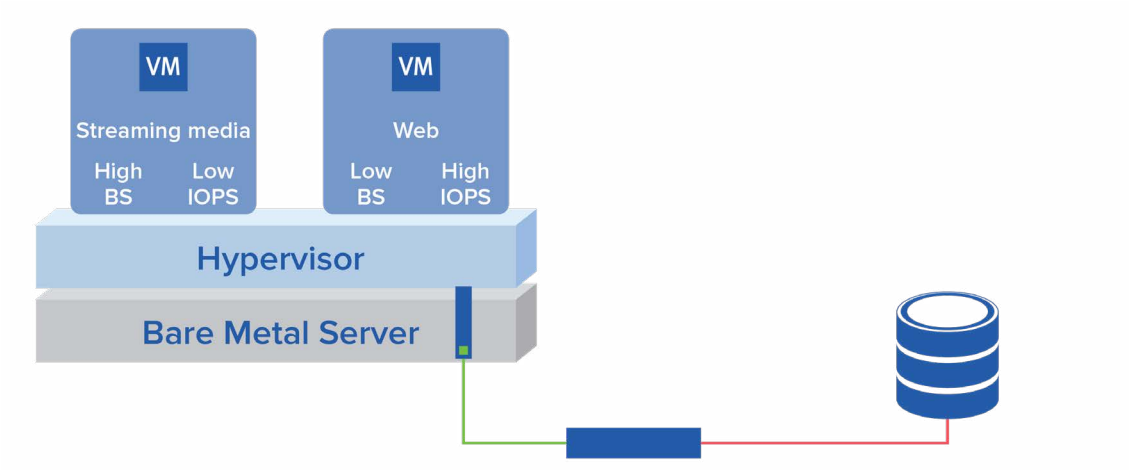

Figure 3: Impact of varying IOPS and BS profiles on volume Storage hitting same SAN target.

Cost considerations aside, one of the tougher choices to make is choosing between flash and conventional magnetic solutions for a given software application. While magnetic disks offer parity between read and write latency, flash offers superior access times under specific conditions. Some of the key metrics needed when evaluating whether to use flash based storage in virtual infrastructure are:

- Storage access profile—flash excels at reads, has a write penalty
- Block sizes of data transfers—larger, smaller, varied
- Queue depths—flash performs better at larger queue depths
- I/O pattern—read, write, sequential, burst, random, etc.
- Utilization of storage—flash performance degrades as storage approaches capacity

Regardless of storage type used in a cloud or NFVI, performance is largely dependent on the variation of block sizes and IOPS read and written by applications, the total latency between the source VM and target storage media, and especially contention from other VMs accessing the same storage resources.

Network I/O

The final piece of the puzzle we will examine is network I/O implementations in virtualization. Outside of special applications like video processing, network I/O is one of the most important resources where virtualized application workloads require near machine level performance for a good user experience. Like schedulers and storage, there are a number of options for virtualizing I/O devices. Two of the earliest methods, device emulation and para-virtualized drivers, were strong in sharing but weaker in performance. The performance degradation comes from excessive overhead of packet processing on the host CPU. The performance cost from device emulation of an Ethernet card could be higher than 50% compared to running on bare metal. This was quickly answered with hardware "pass-through" options like Intel® VT-x and VT-d which provide the VM's network driver direct access to the network card's resources installed on the host. However configuration restrictions, namely one to one virtual port to physical port mappings reduced the benefits of device sharing provided by emulation and para-virtualization. This is where Single Root I/O Virtualization (SR-IOV) and Sharing was introduced by PCI-SIG as a standard to bring the performance gains of pass-through with the flexibility of sharing seen with emulation and para-virtualization.

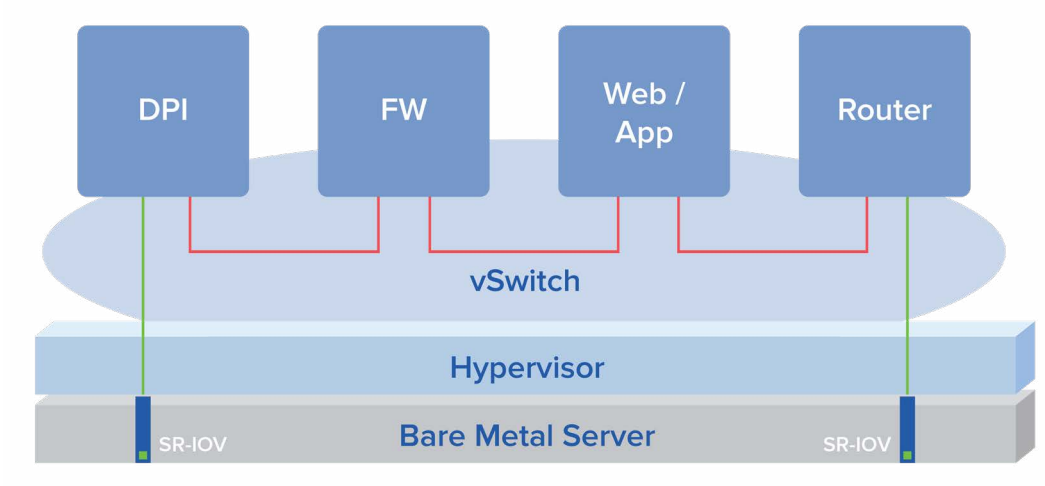

Figure 4: NFV service chain throughput bottleneck between SR-IOV and vSwitch connectivity.

About Spirent

Spirent Communications (LSE: SPT) is a global leader with deep expertise and decades of experience in testing, assurance, analytics and security, serving developers, service providers, and enterprise networks. We help bring clarity to increasingly complex technological and business challenges. Spirent's customers have made a promise to their customers to deliver superior performance. Spirent assures that those promises are fulfilled.

For more information visit: www.spirent.com

COMPUTER CONTROLS

Computer Controls AG distributes electronic components, IoT applications, software, test and measurement solutions. We support our customers in selection, integration and maintenance, translating requirements into complete electronic solutions and customizing systems according to individual specifications. We are a qualified partner for solution-orientated cutting-edge technology.

Computer Controls AG
Industriestrasse 53
8112 Otelfingen
Switzerland
Phone: +41 (0) 44 308 66 66
E-mail: info@ccontrols.ch
Web: https://www.ccontrols.ch

Throughput performance difference between hardware assisting solutions like SR-IOV and internal software based virtual switches can be extreme, especially for open source virtual switches. This can especially impact virtual service chains where VNFs process packets sequentially within the service chain. If the vSwitch introduced latency is excessive, the buffering at the physical network cards can become congested even leading to drops. To address this performance discrepancy, VNF vendors are enhancing performance with data plane optimizations via toolkits. This improves buffering and handling of packets, particularly for smaller packet sizes by reserving specific areas of main memory and utilizing huge page support. Many of these toolkit enhancements are still underway by most VNF vendors so validating their packet forwarding performance is crucial to determine if the VNF meets their SLAs and QoE expectations. It's imperative such testing can occur in live environments as well.

Summary

In conclusion, the demands and complex interactions placed on virtual infrastructure necessitate considerable amounts of testing and benchmarking. It helps determine and validate realistic expectations of cloud services or virtual service chains delivered by NFV. Hypervisor is by far the most critical component of the infrastructure. From the chosen scheduling algorithm to the memory management, to storage and network configurations, the potential for conflicts and resulting discord can wreck havoc on application performance. The hypervisor's ability to handle software workloads across massive scale deployments over a large number of compute nodes, is key to ensuring that the performance of a particular service offering or service chain meets the customer's needs. The demands placed on any given hypervisor are so excessive that only live testing of NFV service chains and on cloud platforms themselves can provide accurate assessments of their true performance. Cloud and Virtual testing needs are greatly simplified with Spirent Cloud solutions by making performance and benchmarking optimization of virtualized networks and cloud infrastructure—more transparent and effective—in turn maximizing your cloud investments and delivering money-back SLAs to your customers. For more information about Spirent Cloud solutions, please visit https://www.spirent.com/Solutions/Cloud-Data-Center.

Americas 1-800-SPIRENT
+1-800-774-7368 | sales@spirent.com

Europe and the Middle East
+44 (0) 1293 767979 | emeainfo@spirent.com

Asia and the Pacific
+86-10-8518-2539 | salesasia@spirent.com

Rev D | 02/24 | www.spirent.com

spirent