White Paper

Scaling to Huge:

Load Testing a Million Virtual Machines in the Cloud

Consider a powerful data center set up to host a cloud. On the hardware side, there are dozens of racks filled with servers and switches. Maybe hundreds of racks, each packed with 1U pizza boxes and blade chasses. Millions of dollars of CPU, memory, storage, cabling, load balancing, redundant power, and Ethernet connectivity. For software, a bare-metal virtual machine hypervisor such as VMware's vSphere ESXi or Mirantis OpenStack, with a management console such as VMware's vCenter Server or HP's Horizon Management. For network architecture, possibly a Software Defined Network (SDN) managed by a controller such as VMware NSX or Cisco's ACI.

If you don't measure the end-to-end data center stack, you don't know how it will perform in the real world. Here's how you measure a cloud-scale platform.

Huge is the Ultimate Road Test

When someone has a high-performance car, it's natural to want to make it even more powerful. Straighter headers, high-flow exhaust, forced-air induction turbochargers, bigger air intakes, upgraded valves and pistons, expensive fuel injectors, spark plugs made of exotic metals, and even improved engine management software. Each of those upgrades is rated for a certain number of horsepower improvements. At the end of the day, however, automotive tuners place the car onto a testing dynamometer and measure the actual horsepower being delivered to the drive wheels under variable conditions. Based on those readings, the mechanics tweak and refine to get the best power, efficiency. Dyno testing is the only way of knowing… until, of course, the car hits the race track.

In the IT industry, as in the auto industry, "You don't know until you test" has been the mantra of load testing for decades. Whether the system under test is a huge cloud data center, a mobile app, a website, a network switch, an applications programming interface (API) or an embedded system, load testing is fundamental to both software development and data center operations.

Indeed, load testing reveals much more than the raw capacity of a system. It shows how the system actually performs under a variety of workloads. What's the response time? Does data get dropped? Do load balancers perform as expected? Those are only three questions that load balancers can answer.

It's All About Scale

The lights are blinking, the fans are humming, and there's a lovely new-server smell. This shiny new cloud is ready to host internal enterprise applications, or perhaps it will be used by a cloud service provider to begin offering multitenant services backed by service level agreements (SLAs).

Does it scale? Does the data center operator have a realistic sense of how much load this rack can handle when fully populated, with tens or hundreds of thousands of virtual machines? Does the owner have confidence enough in the capacity in order to effectively price its services, and back them with money-back SLAs?

For most data center operators, the answer is no: They cannot accurately predict the performance of a stressed cloud, especially one that spans hundreds or thousands of servers, as well as the requisite storage and networking infrastructure. Add in the additional layers of hypervisors and SDN, and capacity is an educated guess at best, calculated by studying product spec sheets.

Because cloud operators don't know what their cloud can actually do, they miscalculate the capacity of their systems. If they conservatively underestimate, the system will be underutilized, which not only adds to expenses, but reduces the number of services that can be sold and provisioned into that cloud – thereby wasting money in a business with tight margins. If they aggressively overestimate, they run the risk of service degradation or failure, which not only hits the bottom line with SLAs, but can be expensive and time-consuming to troubleshoot and remediate.

The solution: Move beyond back-of-envelope projections and load up and measure the cloud. Stress and test the actual compute resources, storage, network, infrastructure, and virtual machine management. Measure performance and capacity with realistic loads and full instrumentation. And analyze the results to understand exactly what the cloud can do – and when (and how) it will fail.

When the cloud system has 100,000 or a million virtual machines, there's only one way to test it: Spirent HyperScale. It scales to huge.

In a complex system – and it doesn't get more complex than a huge cloud data center – the load tester instruments many points within the stack, and indicate not only the full end-to-end performance of the data center, but also the weak points and vulnerabilities within it. Those vulnerabilities can be everywhere, encompassing not only hardware and software components, but also how they are configured, as in the example above. Architecture matters.

Business efficiency also matters, as discussed above. For enterprises, data centers are a tremendous investment, and its owners must understand the capacity of that data center to run applications today, as well as how much headroom is available for future growth. Understanding how services and performance will degrade as load increases is vital for accurately predicting future investments in hardware, software and infrastructure.

For cloud operators and service providers, the data center is both an investment and direct revenue generator. A commercial real estate business wouldn't construct a multitenant apartment building without knowing how many units can be rented, and how much to charge for that rent. Similarly, one shouldn't deploy a commercial VM-based multitenant data center without knowing, with confidence, how many services can be offered in that cloud.

Part of testing goes beyond today's anticipated workload into future planning. When capacity begins to be reached, operators need to know which portions of the systems will need investment prior to service degradation. Will it be the network switches? The CPUs? Memory? Storage I/O? SDN configuration? OpenStack settings? Bottlenecks can be literally anywhere – and can vary depending on the type and intensity of the workload. A VM cluster that's CPU-bound running computations will impact the cloud differently than a VM that's I/O bound running database queries.

A challenge for all complex systems, including clouds, is that performance under stress is not linear. Performance at 70% of capacity might be at or near maximum projected throughput. Performance at 71% of capacity might be the same. Performance at 72% of capacity might drop by a quarter, and at 73% of capacity might slow to a snail's pace.

Operators can't know what they can't measure – and they can't remediate potential bottlenecks that they can't see. A simple software configuration change might move the performance inflection point significantly… in either direction. You don't know if you don't test each configuration change to understand its impact.

Architecturally, the tests run on a configured cloud, with servers, storage, switches, and hypervisors. Ideally, the cloud will already be divided into functional units, which in the VMware world are called pods, and in the OpenStack world are called tenants. Both terms are functionally equivalent. In addition to the workload VMs, each pod will also have one or more "helper" virtual machines that manage the pod's resources.

Spirent HyperScale installs an API Client into each pod; the API Client communicates with the rest of the test management system in order to populate the pods with test VMs, synchronize their starting and stopping, and communicate with the test management software outside the pod.

At the highest level, the Spirent HyperScale GUI presents a user interface for designing tests, running the tests, and analyzing the results. It also manages and abstracts two additional software resources necessary to manage the tests: the Virtual Deployment Service (VDS) and the Synchronized Test Orchestrator (STO).

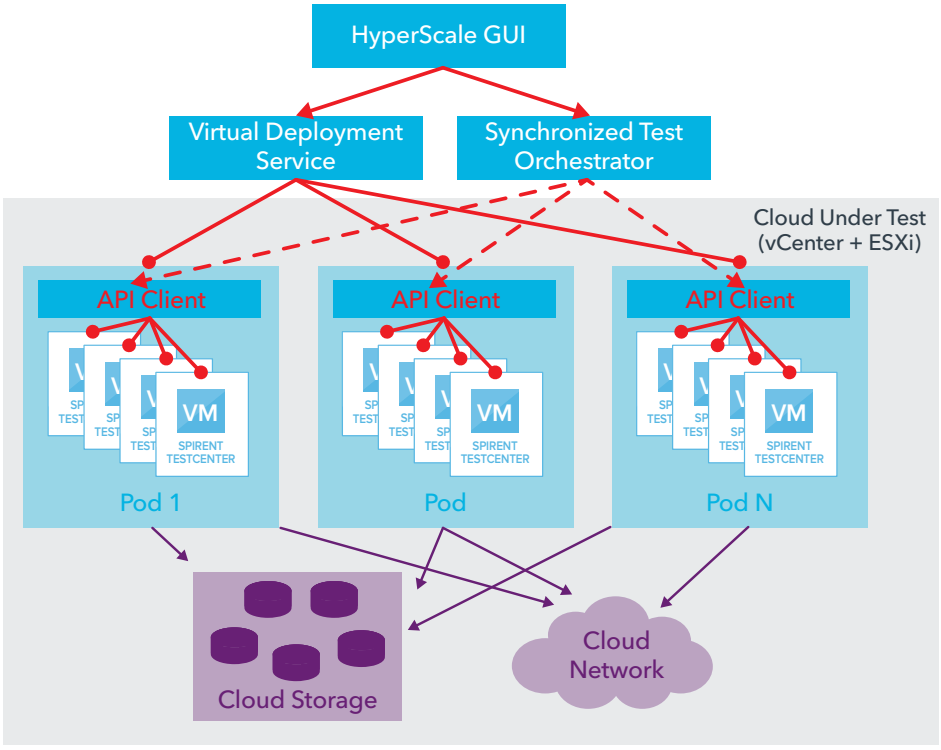
When it is time to set up the test, the VDS communicates with vCenter or OpenStack Nova to install the properly configured virtual machines, each with an instance of Spirent TestCenter Virtual (STCv). STCv generates the workload for each VM, while consuming network resources and providing measurement instrumentation. VDS also connects the virtual network interfaces to the networks. When the entire cloud is ready, VDS sends a green-light signal back to the Spirent HyperScale GUI, indicating that tests can be run.

Testing Huge What-If Scenarios

While Spirent HyperScale is most often used for measuring and tuning a complete or near-complete cloud system, it can also be used much in the process during prototyping and even vendor evaluation and product selection.

Spirent HyperScale can be installed onto a testbed and used to run what-if scenarios without requiring hardware changes. For example, test systems can be examined with different types of VM configurations, SDN network architectures or OpenStack or VMware settings.

Similarly, the testbed can evaluate solutions providers' hardware, such as servers, storage arrays, network interface cards, switches, routers, as well as software stacks. Set up a realistic scenario, run the tests, note the results. Charge hardware, rinse, repeat. Move beyond vendor claims and bench tests of standalone hardware, to see exactly how the products will perform in your data center, in your hardware and software stack, and with your own VMs and applications.



How Spirent HyperScale Works

Spirent HyperScale is a software-based load testing solution that, in its initial offering, is designed for huge clouds based on either VMware vCenter or on OpenStack. (Future versions may support additional cloud and virtual machine platforms.) Spirent HyperScale was specifically designed to design and implement tests on clouds with thousands, tens of thousands, hundreds of thousands, or even a million virtual machines.

The design of Spirent HyperScale allows the test team to create tremendous numbers of virtual machines without needing to fine-tune or tweak them at a granular level. Instead, testers create profiles of VMs with broad characteristics in terms of CPU and memory utilization, storage consumption, LAN and WAN bandwidth, and so-on. The configuration of those test-load VMs is fast and easy – as easy as editing a comma-delimited text file.

Spirent HyperScale populates those VMs across the cloud data center, measures performance and resource utilization, and aggregates the results into a graphical user interface.

Ready, set, go: When the test operator begins the test, STO kicks into action. It synchronizes the pods and spools up each of the VMs, again working through the API client. When all the pods and STCv instances are ready, STO broadcasts the "start test" message across the entire cloud, reaching every STCv instance nearly simultaneously. This is an important aspect of Spirent HyperScale: The "start test" signal must be a single atomic operation, because otherwise the first few test cases on one pod's VMs might have finished running before the last pod's VMs have even started.

The tests, while running, use live data and full network resources across the entire network, and even utilizing external resources such as storage arrays and Internet connections: All ports, all services, network traffic, and storage I/O. Everything is fully engaged, so if a bottleneck appears, it will be found; if packets drop, they will be detected.

As the tests are running, real-time data is sent from the pods' API Clients back to the Spirent HyperScale GUI, where it is curated and presented in a meaningful way to the test operator. As the purpose of the exercise is to stress the entire cloud, with its hundreds of thousands of VMs, the default aggregate reports hide unnecessary details about individual pods, components and VMs. Because all test data is captured and logged, the test operator and systems analysts always can dive as deep as desired to understand performance issues and resolve bottlenecks, both during and at the conclusion of the test.

Scaling to Huge:

Load Testing a Million Virtual Machines in the Cloud

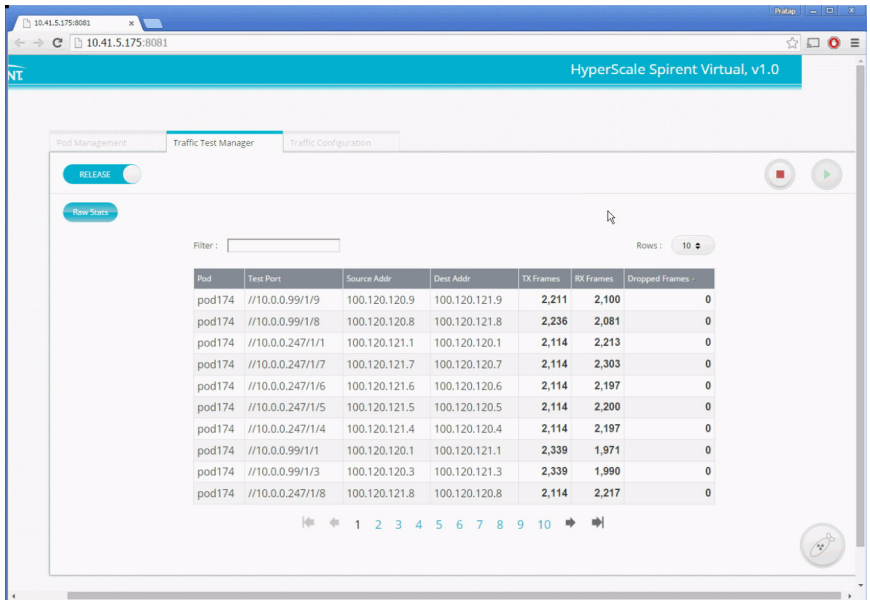
About Spirent Communications

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



Spirent HyperScale showing results with actionable insights

Scaling to Huge Means Testing Huge

A cloud-scale data center costs millions of dollars or more, and its owners have tremendous expectations for that investment. What's more, when that data center is used to offer commercial hosting or multitenant applications, customers have significant expectations as well, backed up by SLAs.

How will that data center perform when running thousands or even millions of virtual machines? No matter the specs of the individual hardware, software and infrastructure components, the only way to know is to load test the actual system, end to end, top to bottom. You don't know what you don't test, not only in terms of raw capacity, but also to determine weaknesses and bottlenecks.

Spirent HyperScale is the only solution that can test a million VMs in a cloud data center, and allow data center operators to run what-if scenarios on different configurations of the solution stack. Forget guessing and back-of-the-envelope calculations: Spirent HyperScale gives data center operators and owners the confidence to know exactly what their brand-new cloud can do.

spirent™ Promise. Assured.

Contact Us

For more information, call your Spirent sales representative or visit us on the web at www.spirent.com/ContactSpirent.

www.spirent.com

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

US Government & Defense
info@spirentfederal.com | spirentfederal.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 A | 10/18