White Paper

# Data Center Testing: A Holistic Approach

**May 2009**

## TABLE OF CONTENTS

## *Executive summary*

Data centers today are larger, faster and more complex than ever. New technologies such as virtualization, Fibre Channel over Ethernet and 40/100 gigabit Ethernet aim to help organizations move multiple traffic types – data, storage, video, and voice – onto a single, converged core.

Bearing in mind the old adage that “you can’t manage what you can’t measure,” validation and performance assessment of these new technologies are vital first steps in implementing these new technologies. In a data center context, that means testing each of these technologies not just by itself but also in concert with many other data center components old and new. In short, the key question is: *As I grow my data center, how can I validate that all components will work together as a coherent whole?*

This white paper aims to help network professionals understand the issues involved in data center validation and performance benchmarking. After a review of basic testing concepts and industry-standard testing methodologies, this document discusses each of the new data center technologies in turn, with detailed coverage of the major testing challenges for each.

This paper concludes with an introduction to *holistic testing* of the data center, an end-to-end approach that will help organizations develop confidence in the new technologies that will allow data centers to grow larger and more cost-effective over time.

## *Testing basics*

Before diving in to the new technologies driving data center growth, it may be helpful to review some basic testing concepts. These concepts will come into play in the coverage of emerging data center technologies later in this document.

A sound benchmark must meet four basic criteria: It must be *repeatable*, *reproducible, stressful* and *meaningful*.

*Repeatability* means multiple iterations of the same test, conducted on the same test bed, should produce similar results. Obviously a test that produces wildly different results with each iteration is of little use.

*Reproducibility* is similar to repeatability, but refers to situations where the same test is run on *different* test beds. For many organizations, it is common practice for teams at multiple locations to work on the same set of benchmarks. For example, test engineers working in San Jose, Beijing and Bangalore all should be able to produce similar measurements, assuming all use the same test instrument and system under test (SUT) and follow the same procedures.

As an aside, reproducibility can be an elusive goal in benchmarking. When test engineers at different sites obtain different measurements, the first step should be to verify that all sites have the same software and hardware versions, both in the test instrument and SUT, and follow the exact same procedures.

A benchmark is *stressful* only if it finds the limits of system performance. Throughput tests, for example, seek to find the highest rate at which a device forwards traffic. In stress testing, the goal is to have a successful iteration (the device drops no frames) followed by a failed iteration (the device drops frames); this is the limit of system performance.

A commonly heard refrain when reviewing the results of stress tests is that such benchmarks do not represent “real world” conditions. Leaving aside for the moment that every network has its own definition of reality, such complaints miss a key point. The goal of stress testing is to find system limits, not to find some definition of reality.

The final goal in benchmarking, coming up with a *meaningful* set of results, can be the most difficult to achieve. Tests of networking devices produces vast quantities of data, not all of which will be relevant to the task at hand.

Consider an example where latency tests of two routers produce measurements of 10 and 100 microseconds. If the routers being tested will be deployed on opposite ends of a transcontinental link, with the speed of light introducing propagation delays well into the tens of milliseconds or higher, a difference of 90 µsec is not at all meaningful.

On the other hand, if the routers will be deployed 1 meter apart in a data center, and will carry financial information whose timeliness is worth a million dollars for each microsecond, a tenfold increase in latency is *very* meaningful. The key point is that measurements themselves have no intrinsic value. What makes measurements meaningful is how they apply to the use case being tested.

Above and beyond these principles, a useful rule in network device benchmarking is to *never assume correct behavior on the part of the SUT.* Especially when assessing new network hardware and software, faulty implementations often lead to suboptimal test results. The SUT also may report incorrectly on its status, hence the need for test instruments that provide externally observable validation (or otherwise) of system performance.

Over the years, test engineers have codified these rules and other useful test techniques in a series of industry-standard testing documents issued by the Internet Engineering Task Force (IETF) as requests for comments (RFCs). The most relevant RFCs for data center device testing are as follows:

- RFCs 1242 and 2544 describe terminology and methodology, respectively, for router testing. These are foundation documents for network device measurement, and many other testing RFCs refer to concepts introduced in these RFCs.

- RFCs 2285 and 2889 present terminology and methodology, respectively, for Ethernet switch testing. These RFCs also introduce basic traffic patterns that are essential in data center testing, as discussed later in this section.

- RFCs 2432 and 3918 offer terminology and methodology, respectively, for IP multicast testing. Multicast testing belongs in any assessment of data center devices, especially considering it is a common transport for video and triple-play services. These documents also describe tests that blend unicast and multicast traffic.

- RFCs 2647 and 3511 document terminology and methodology, respectively, for firewall performance measurement. These documents introduce the concept of *goodput*, a far more meaningful metric than throughput when measuring the performance of loss-tolerant TCP traffic through a firewall (or performance of any layer-4/layer-7 device, for that matter).

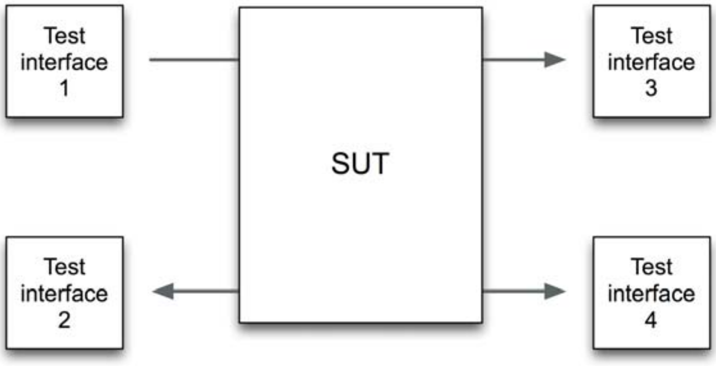
- Several other RFCs also may be useful in assessing performance of data center devices. RFC 5180 extends RFC 2544 with IPv6-specific tests. RFC 4814 recommends pseudorandom traffic patterns for use in testing, unlike the artificial and static patterns generated by many test instruments, and presents formulas for calculating the overhead introduced by bit- and byte-stuffing on SONET/SDH links. And RFC 4689 describes terminology used in testing network-layer quality-of-service (QoS) mechanisms. The benchmarking working group is reviewing several other testing documents, all currently in draft status.

As noted, RFCs 2285 and 2889 discuss traffic patterns relevant for testing data center devices. RFC 2285 defines two “traffic orientations” and three “traffic distributions.”

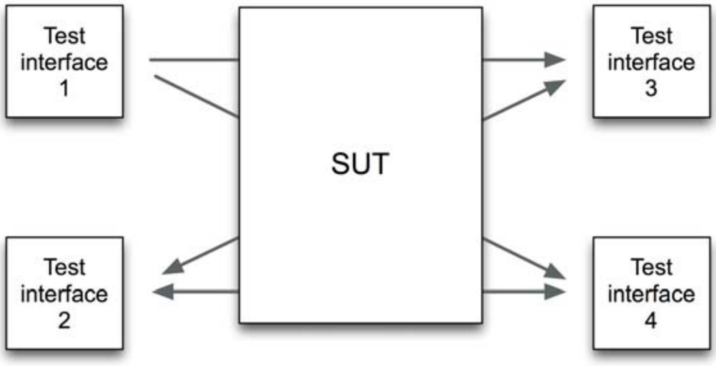
The traffic orientations simply refer to unidirectional and bidirectional traffic. In a unidirectional test pattern, one test interface offers traffic destined for another interface. In a bidirectional pattern, every receiving interface is also a transmitting interface.

The RFCs’ three traffic distributions are “non-meshed,” “partially meshed” and “fully meshed.”

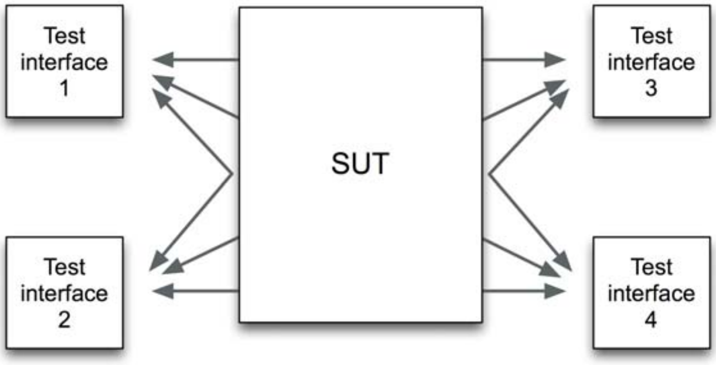
In a non-meshed pattern, as shown below, every pair of transmitting and receiving ports on the test instrument is mutually exclusive. This traffic may be unidirectional (as with interfaces 1 and 3) or bidirectional (as with interfaces 2 and 4). Some test instruments refer to this as a “port pair” pattern:



In a partially meshed distribution, sometimes referred to as a “backbone” pattern, one set of test interfaces offers traffic to a different set of interfaces, but not to any interfaces within its own set. Again, traffic may be unidirectional (interface 1) or bidirectional (interfaces 2, 3 and 4):



In a fully meshed pattern, all test interfaces offer traffic destined to all other test interfaces. With fully meshed patterns, all flows are unidirectional; they only appear to be bidirectional in the figure below because each test interface sends and receive traffic from all other interfaces. A fully meshed pattern puts the most stress on a switch or router fabric and thus is the preferred traffic distribution when testing these devices.



The illustrations given here use only the simplest possible traffic distributions. Within each test interface, it is possible to generate thousands of unique traffic streams, perhaps with varying traffic distributions. These complex patterns can be useful in assessing quality of service (QoS) mechanisms including prioritization of Fibre Channel over Ethernet (FCoE) traffic, as discussed later in this document.

RFC 2285 introduces one more concept that will be useful in data center benchmarking: It distinguishes between *intended load* and *offered load.* As the name implies, intended load is simply the rate at which the tester would *like* to offer traffic.

However, because of congestion control mechanisms such as 802.1Qbb priority flow control, there may be a difference between the transmission rate the tester *expects* to use, and what the test instrument *actually* uses. The latter is the offered load. In some cases offered load may be lower than the intended load because the transmitting test interface was throttled during the test by a congestion control mechanism.

Each of these basic testing concepts will play a part in data center device benchmarking. The remainder of this document will cover how these concepts apply to three new technologies at work in the data center: Virtualization, Fibre Channel over Ethernet and 40- and 100-Gbit/s Ethernet.

## *Virtualization*

Virtualization, which for years has brought benefits to servers, is now coming to networked devices. Three factors are driving the growth of virtualization in data centers. First is *consolidation:* As organizations move resources into a few large data centers, server virtualization is a logical choice.

Second is *convergence:* Data center backbones now carry not only Ethernet data but also storage traffic, encapsulated using Fibre Channel over Ethernet (FCoE). In addition, data centers handling streaming video and triple-play applications also are moving from multiple distinct networks onto a single converged core.

Third is *virtual networking devices:* These new virtual switches and appliances lack physical points of attachment, introducing a new requirement for virtual testing capabilities.

The explosive growth in virtual server deployments also heightens the need for scalability testing. Where network architects previously specified a maximum four to eight virtual machine (VM) instances per physical server, new virtualization products will push that number to 64 VMs or beyond. Considering that a standard rack holds up to 42 physical servers, and that there may be hundreds or thousands of racks in some data centers, the implications are significant for network traffic.

Moreover, a VM instance often uses more bandwidth than a physical server. There is the extra traffic involved in managing a virtual machine from a central location. Also, many data centers use products such as VMware Vmotion to move virtual machines within the data center. This greatly enhances uptime and reliability, but it also generates considerable network load.

Testing virtual network devices and servers poses several interesting new questions:

**How can a test instrument attach to a virtual network device?** For many types of tests, a connection to the physical server hosting a virtual switch will not be sufficient. One physical interface may handle traffic for dozens of VM instances, making it difficult to isolate and measure the performance of each VM instance.

What's needed is to virtualize the capabilities of the test instrument. A virtual test instrument resides in software, and thus runs *inside* the physical machine hosting virtual network and server instances. From the standpoint of the virtual network device, a test port looks exactly the same as it would in the physical world.

Of course, a virtual test instrument should offer the same capabilities as its physical counterpart. As discussed later in this document, the test instrument should be able to offer traffic between any number and any combination of virtual and physical interfaces, and measure the whole system as a single entity. It also should be able to offer any layer-2 through layer-7 traffic pattern the tester desires, both on virtual and physical interfaces.

**Can test instruments on virtual machines be trusted?** The concept of test instrumentation running in software is certainly nothing new; indeed, software-based networking test tools predate hardware-based instruments by decades. Unlike hardware-based instruments, however, software-based test tools often produce measurements that say as much or more about underlying components – the networking stack, host operating system, drivers and network interface card – as the system they purportedly measure.

Software-based tools also can produce results that are either nonrepeatable or nonreproducible.

One strategy to ensure measurements can be trusted is to implement the *entire* test instrument – including software-based emulation of hardware components – in software. This approach requires a far more rigorous approach to system design than does software-only tool design. But the benefit is clear: By emulating the entire test instrument in software, the instrument's measurements are far less dependent on extraneous factors. Such an instrument will produce more meaningful measurements than software-only tools.

**Do virtual and physical switches offer comparable performance?** Line-rate throughput and low latency and jitter have long been the hallmarks of physical Ethernet switches, but their virtual counterparts may not compare. Tests of early virtual switches show frame loss with offered loads as low as 50 Mbit/s. Moreover, these tests involved just a single pair of interfaces on a single virtual switch; in contrast, the standard practice for switch and router testing is to attach test interfaces to *all* switch ports and generate traffic in a fully meshed pattern – a far more stressful test pattern than using just a single port pair.

Even if virtual switches will never handle loads as heavy as physical switches (a dubious assumption in these early days of virtual networking), it is still important to conduct stress tests to describe the limits of system performance. As discussed in the “Testing basics” section of this document, test engineers for years have relied on industry-standard methodologies for assessing unicast and multicast performance in switches and routers. These same methodologies still apply in the virtual world.

**Does a virtual switch support the same protocols and functions as a physical switch?** When assessing Ethernet switches, network managers often put at least as much emphasis on reliability and features as on performance. Functional testing is just as important for virtual switches as performance and scalability testing, and should be a part of any data center test methodology.

A network manager can reasonably expect any modern Ethernet switch to support features such as virtual LANs (VLANs), access control lists (ACLs), and Internet group management protocol (IGMP) for forwarding multicast traffic. These protocols (and often many others) are often included as part of physical switch performance testing; they should also be included as well when testing virtual switches.

## *Fibre Channel over Ethernet (FCoE)*

Fibre Channel (FC), by far the most widely used transport in storage-area networks (SANs), presents special challenges when it is converged into Ethernet-based data centers. FC employs management frames to identify endpoints and switch fabrics and to provide flow-control features not found in Ethernet. Also unlike Ethernet, FC is intended to operate in a loss-free manner; in contrast, Ethernet networks can tolerate loss. And FC traffic is highly sensitive to increases or changes in latency and jitter.

Fibre Channel over Ethernet (FCoE) encapsulates FC traffic into Ethernet frames, thus substantially reducing interface count and cabling in the data center. However, it does not by itself protect the FC-specific features mentioned above. The IEEE has developed several new specifications to ensure reliable delivery of FC traffic – and each of these require testing, especially in the mixed Ethernet/FCoE deployments that will be increasingly common in many data centers.

The new IEEE specifications include the following:

- **802.1Qbb priority flow control (PFC):** This congestion control mechanism that allows multiple traffic types, such as FCoE and non-FCoE, to share an Ethernet link. PFC uses Ethernet pause frames to delay non-preferred traffic when a transmitter has a preferred-class frame ready to send.

  Like the earlier version of Ethernet flow control defined in IEEE 802.3x, PFC works by sending XOFF and XON messages to signal that an interface should stop and resume transmitting, respectively. Unlike 802.3x flow control, PFC works on a per-priority basis, allowing different traffic classes to use different XOFF/XON intervals.

- **802.1Qaz priority groups:** This scheduling mechanism aims to ensure consistent quality of service levels for multiple traffic classes.

- **Data center bridge exchange (DCBX):** This set of extensions to the IEEE's link-layer discovery protocol (LLDP) allows data-center devices to exchange capabilities information upon link establishment. DCBX uses LLDP to carry messages specific to data-center networking, such as the use of PFC or 802.1Qaz priority groups.

Basic FCoE testing covers functional validation of the new protocol. An FCoE-capable test instrument should help answer questions such as whether FCoE interfaces correctly use the FC initialization protocol (FIP) to discover and then log in and out of switch fabrics; whether FC endpoint IDs (FCIDs) are correctly mapped into Ethernet MAC addresses; and whether FCoE devices will accept static assignment of Fibre Channel's World-Wide Names (WWNs).

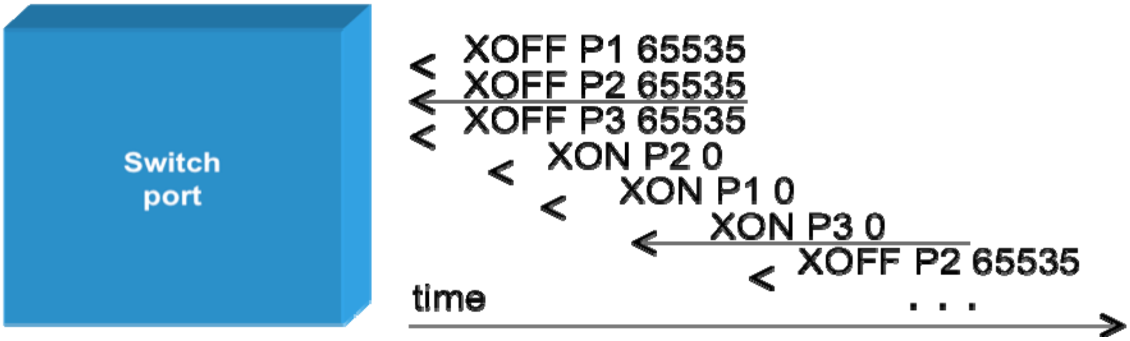
In more advanced FCoE benchmarking, the test instrument generates and analyzes a mix of multiple traffic FCoE and non-FCoE classes. Of key importance here is how well FCoE-compliant devices observe priority flow control messages during periods of congestion. Timing is a critical factor in assessing PFC efficiency.

Like conventional Ethernet pause frames, PFC messages contain a *pause quanta* indicating how long a device should refrain from transmitting. One pause quanta equals 512 bit times, which is equivalent to 51.2 nanoseconds at 10-Gbit/s rates. Note that a pause quanta indicates the maximum amount of time an interface should defer transmission; the actual time may be shorter if the device sends an XON message indicating congestion has cleared.

After generating PFC frames, the test instrument can measure both *pause duration* and also *pause response time*, the interval between receipt of the PFC message and the actual pause. Also, pause response times may differ for PFC XOFF and XON messages.

By generating multiple traffic classes with different pause quanta (and also using multiple frame sizes), the test instrument can create complex loads to simulate the stresses an FCoE switch may experience in production networks.

Consider the example of a test instrument generating three traffic classes labeled P1, P2 and P3, initially all at the same time:



The test instrument uses different XOFF/XON intervals for each class, and repeats each at different intervals:

| Traffic class | XOFF/XON interval (µsec) | Inter-PFC burst interval (µsec) |
|---|---|---|
| P1 | 200 | 500 |
| P2 | 150 | 450 |
| P3 | 300 | 700 |

In this example, all three priorities initially send PFC XOFF messages at the same time, each with pause quanta of 65535. Some 150 µsec later, the P2 class sends an XON message, followed 50 µsec after that by an XON message for P1 traffic. At 300 µsec, the test instrument then sends an XON message for P3 traffic. The entire cycle then repeats, beginning with an XOFF message for P2 traffic 450 µsec after the first message.

By using a mix of frame sizes and inter-PFC burst intervals, the different traffic classes quickly will become desynchronized, placing a heavy burden on FCoE devices.

Moreover, the test grows significantly more stressful as port count increases. When testing data center switches with hundreds or thousands of ports, each handling a mix of multiple FCoE and non-FCoE frames, the switch's flow control logic will need to keep up with a constant barrage of PFC messages at 10-Gbit/s Ethernet line rate (or beyond, as discussed in the next section). As with any emerging technology, it is prudent to stress-test PFC functionality under load to determine the limits of system performance.

### *40- and 100-Gbit/s Ethernet*

Another major technology driver for data centers is the impending introduction of 40-Gbit/s and 100-Gbit/s versions of Ethernet. In one sense, these new transports are "just Ethernet," only faster. In another, they will pose fundamental challenges for test equipment, even including the ability to count packets.

A major driver for higher-speed Ethernet in the data center is that edge ports are getting faster. Just as gigabit Ethernet interfaces in servers has driven the deployment of 10-Gbit/s Ethernet uplinks in switches and routers, so too will the upcoming introduction of 10-Gbit/s connections in servers and access switches drive a need for 40- and 100-Gbit/s Ethernet in data center backbones. Even when servers and downstream switches use interim solutions such as IEEE 802.3ad link aggregation instead of 10-Gbit/s connectivity, the extra traffic still places an additional burden on network backbones.

Another major driver for faster Ethernet versions is the growing popularity of high-bandwidth video traffic. With the growing use of high-definition television and on-demand video services, data centers tasked with providing video-only or triple-play content may find gigabit and 10-gigabit links saturated. Higher-capacity backbones provide a natural solution for data centers serving up video traffic.

These new versions of Ethernet pose new testing challenges, including the following:

- Can my test instrument count?
- Can my test instrument provide accurate latency and jitter measurements?
- Can my test instrument measure 40/100-Gbit/s Ethernet as a single entity?
- Can my test instrument determine sequencing?

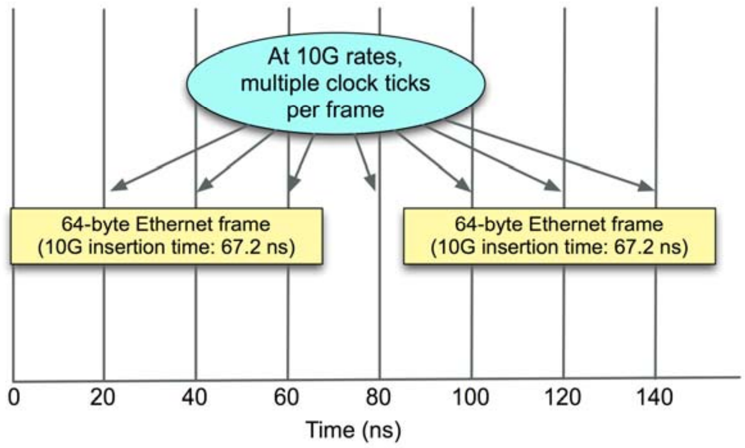
The first of these points, about counting frames, seems almost too obvious to mention. But obtaining accurate packet counts on 40/100-Gbit/s Ethernet will pose a significant technical challenge for test instruments. To understand why, it is useful to explore the concept of *timestamp resolution.*

Timestamp resolution describes the precision with which a test instrument clocks the departure and arrival of frames on each test interface. To measure transmission time and count packets, test instruments embed a “signature field” in every frame. For example, to measure latency a test instrument subtracts the time when each frame is received from the time when it is transmitted. To do this, the instrument compares the timestamp in the received frame (the transmit time) with its system time (the receive time).

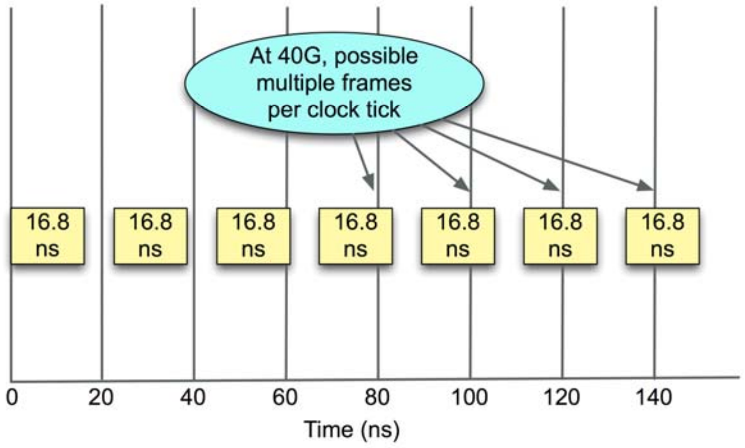
For any time-based measurement, it also is important to know the timestamp resolution. For example, if a latency measurement is 1.000 microsecond and the test instrument has a timestamp resolution of 20 nanoseconds (a commonly used value in the field), then the actual measured latency may be anywhere in the range of 0.980 to 1.020 microseconds.

At 40/100-Gbit/s speeds, a 20-ns resolution is not sufficiently precise for latency and jitter – or, for that matter, to see all traffic in the first place. The reason: The time needed to create a frame is less than 20 nanoseconds, both for 40- and 100-Gbit/s Ethernet.

Some simple illustrations make the problem clear. With 10-Gbit/s Ethernet, the minimum time to insert a 64-byte frame onto the medium is 67.2 ns. In this case a 20-ns resolution is adequate since there are always multiple clock “ticks” per frame:

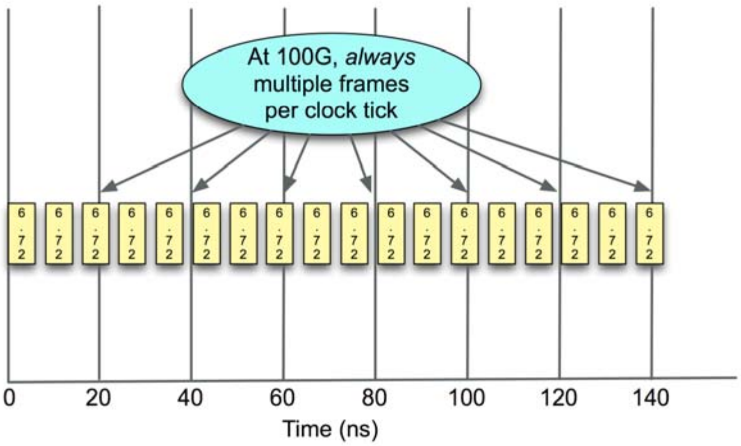

At 10G rates, multiple clock ticks per frame
64-byte Ethernet frame (10G insertion time: 67.2 ns)
64-byte Ethernet frame (10G insertion time: 67.2 ns)
0
20
40
60
80
100
120
140
Time (ns)

With the higher transmission rates of 40- and 100-Gbit/s Ethernet, a 20-µsec timestamp resolution will not be sufficient. At 40-Gbit/s rates, the minimum frame insertion time falls to 16.8 ns. While 20-ns clock ticks are sufficient to measure the first few frames, the test instrument’s clock ticks and the frame rate will quickly become desynchronized, with two frames seen *within* a single tick of the clock:



In this scenario, the test instrument may not be able to count all frames, let alone provide accurate latency and jitter measurements for those frames.

The situation grows even more dire with 100-Gbit/s Ethernet, where the minimum frame insertion time drops to 6.72 ns. At 100-Gbit/s rates, there will *always* be multiple frames present per 20-ns clock tick:



In this situation, the test instrument is virtually blind. Since it cannot provide an accurate count of the number of frames seen per clock tick, there is no chance for accurate measurements of any kind. This applies to all measurements provided by the test instrument – not just time-based metrics such as latency and jitter, but other key metrics such as throughput and frames in sequence.

Clearly, 40- and 100-Gbit/s Ethernet will require finer timestamp resolution. A key question in selecting test equipment will be what level of precision the instrument provides.

Another consideration specific to 40-Gbit/s Ethernet is whether the test instrument can track traffic from each interface *as a single entity*. This requirement may seem counterintuitive; after all, test instruments have provided per-port measurements for many years. But the IEEE specification for 40-Gbit/s Ethernet creates a flow by aggregating four 10-Gbit/s streams into one. Test instruments will need to reassemble these four streams, potentially at line rate, while still providing accurate measurements.

A final issue that applies to both 40- and 100-Gbit/s Ethernet involves sequence counting. Applications such as FCoE-based storage and high-bandwidth video aggregation expect frames transmitted in order to be received in that same order, with even a small amount of reordering leading to degraded performance[1].

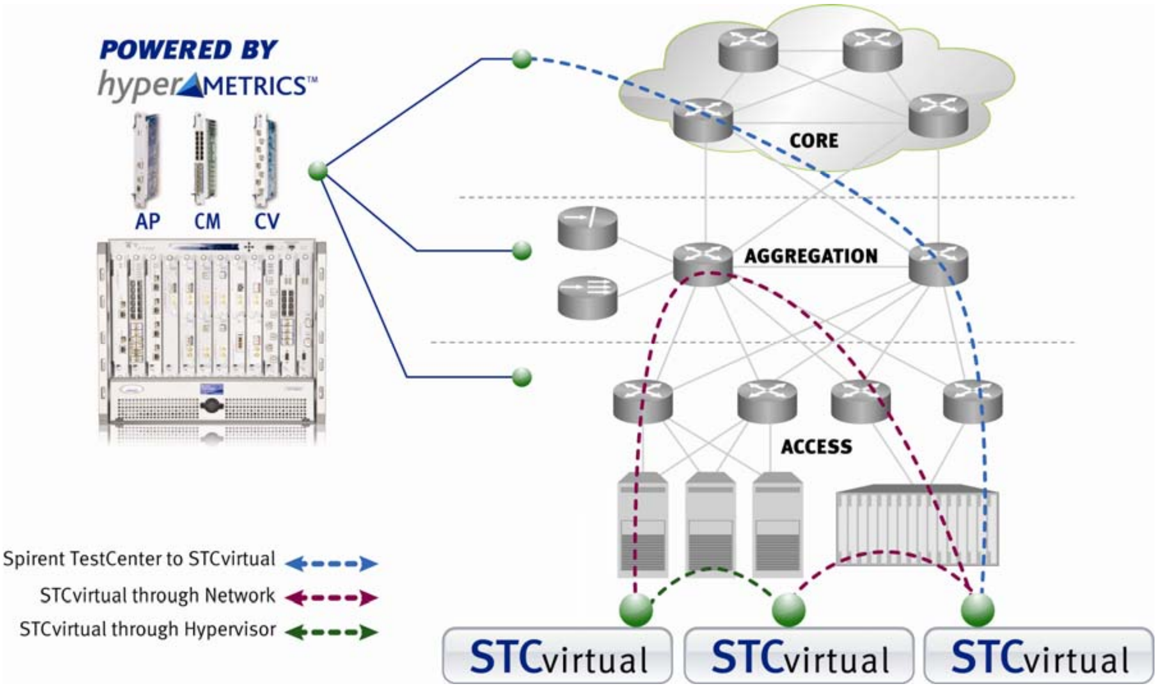
The signature fields that test instruments embed in each Ethernet frame contain sequence numbers the instrument uses to track frame order. Given the concerns discussed earlier about timestamp resolution and reassembly of multiple streams, it is reasonable to ask when assessing 40- and 100-Gbit/s test instruments whether they can provide comprehensive sequence analysis for all frames.

[1] TCP, which carries 90 percent or more of Internet backbone traffic according to studies by CAIDA and Sprint, can tolerate some reordering. However, excessive reordering of TCP traffic also can degrade performance and even lead to connection loss.

## *Putting it all together*

The emerging model for data center network design converges many types of traffic – data, voice, video, storage – onto a single high-speed Ethernet platform, using a mix of virtual and physical components. Testing this complex design is greatly simplified with a *holistic* approach to protocol validation and performance measurement.

In the new data center model, there are thousands of potential points of attachment for test instruments, some virtual and some physical. Consider the three-layer data center design – access, aggregation and core – in the following figure:


POWERED BY
hyperMETRICS™
AP
CM
CV
CORE
AGGREGATION
ACCESS
Spirent TestCenter to STCvirtual
STCvirtual through Network
STCvirtual through Hypervisor
STCvirtual
STCvirtual
STCvirtual

Working from the access layer up, this data center design includes a mix of virtual servers running on blade chassis along with physical servers and mainframes, all connected to the aggregation layer over a combination of gigabit and 10-Gbit/s Ethernet links. The access layer also includes some physical switches and routers, and also virtual switches running on some of the blade chassis. The aggregation and core layers use a combination of switches and routers linked with 10-Gbit/s Ethernet (and 40- and 100-Gbit/s Ethernet in the future) to interconnect all hosts with the public Internet.

A holistic approach to testing can help make sense of this a complex mix of interconnected devices. In this context, "holistic testing" means the ability to measure performance not only of individual data center components, but also of the entire data center, and to make sense of the results. Holistic testing covers all layers of the networking stack and measures traffic across any arbitrary path through the data center.

Holistic testing delivers many benefits:

- **The test instrument provides a single unified environment for generating and analyzing test traffic.** Test engineers should not need to "post-process" test results from multiple applications or test platforms to characterize the performance of the entire data center.

- **Test instruments offer any-to-any connectivity between virtual and physical endpoints.** This may mean testing between virtual servers; between physical servers; between virtual or physical network devices; or any permutation of multiple virtual and physical devices.

- **Flow counts are highly scalable to assess the quality of service and quality of experience delivered by the data center.** A complex data center design may handle thousands or millions of distinct flows, each with different service-level requirements. A holistic approach to testing allows the generation and analysis of many distinct traffic classes, all on a single platform.

- **The entire test instrument scales so that measurements from the Nth port are just as accurate as those from the first port.** A holistic approach requires a test instrument design that does not affect test measurements as test instrument interfaces and/or chassis are added. For example, a latency measurement between two test ports in the same chassis should yield the same result as a measurement between any two test ports on multiple, linked chassis.

- **The test instrument can accommodate future bandwidth growth.** The forthcoming introduction of 40- and 100-Gbit/s Ethernet not only requires test interfaces running at these speeds; it also demands that the test instrument will be able to provide the same measurements, with the same level of precision, as at lower speeds. This requires a test architecture designed from the ground up to accommodate 100-Gbit/s rates across the data center.

In short, holistic testing means testing the *entire* data center – and that, in turn, requires a test instrument that can measure performance of the data center as a unified whole.

TESTED with SPIRENT