Hyper-scale WAN Optimization

Jim Metzler

Driven by factors such as storage replication and virtual machine migration, the amount of traffic traversing inter-data center WANs has been growing at a dramatic rate and this rate of growth will likely increase over the near term. Supporting this traffic presents both economic and performance challenges. The traditional WAN optimization controller (WOC) was not designed to respond to these challenges, but rather to the challenges of branch office to data center communications. Because inter-data center traffic has a set of characteristics that are not addressed by traditional WOCs, what is needed is a new class of device. This new class of device, the Hyper-scale WOC, must be purpose built to support inter-data center traffic. An example of such a product is Infineta's Data Mobility Switch (DMS).

June 2011

Replicating data between storage arrays and NAS filers is a critical element of disaster recovery strategies because businesses need their critical data to be preserved for business continuity purposes. Disk array and filer volumes are expanding rapidly, driven by the combination of business requirements and advances in disk technology. Volume sizes of up to 500 TB –1 PB are becoming increasingly common. Like most other inter-data center traffic, storage replication traffic is characterized by a relatively small number of flows or connections, but very high traffic throughput per flow.

System Backups

Backups are an important component of a disaster recovery strategy because they focus on the continuity of physical and virtual application servers as well as database servers. The increased complexity of application software and the underlying operating systems is causing server image sizes to grow dramatically. Backup solutions that can minimize the backup window, thus allowing for more frequent backups, can help make a backup strategy more effective.

Virtual Machine Migration

Enabled by the wide-spread adoption of virtualization and cloud computing, the migration of VMs between data centers is becoming increasingly common. Live migration of production virtual machines (VMs) between physical servers provides tremendous value. It allows automated optimization of workloads across resource pools and makes it possible to transfer VMs away from physical servers that are undergoing maintenance procedures or experiencing faults or performance issues. During VM migration the machine image, which typically runs 10 Gigabytes or more in size; the active memory; and the execution state of a virtual machine is transmitted over a high-speed network from one physical server to another. For VM migrations between data centers, the virtual machine disk space may either be asynchronously replicated to the new data center or accessed from the original data center over the WAN. Yet another approach is to use synchronous replication between the data centers, which allows the data to reside at both locations and to be actively accessed by VMs at both sites; a.k.a., active-active storage. In the case of VMotion, VMware recommends that the network connecting the physical servers involved in a VMotion transfer have at least 622 Mbps of bandwidth and no more than 5 ms of end-to-end latency[2].

High Performance Computing (HPC)

A significant portion of supercomputing is performed on very large parallel computing clusters resident at R&D labs, universities, and Cloud Service Providers that offer HPC as a service. Transferring HPC jobs to these cluster data centers for execution often involves the transfer of huge datasets over the WAN. Particularly for applications that tend to run over clusters that are relatively loosely coupled together (e.g. those based on Hadoop or MapReduce), inter-node communications may involve large data transfers of interim results.

One way to support the huge inter-data center traffic flows described above is to connect the data centers with WAN links running at speeds of 10 Gbps or higher. Unfortunately,

[2] http://www.vmware.com/pdf/vsphere4/r41/vsp_41_dc_admin_guide.pdf

a critical limitation of this approach is that these WAN links are not always available when and where they are needed. Another more fundamental limitation is that these WAN links are inordinately expensive.

A far more practical way to support these large inter-data center traffic flows is to implement techniques that reduce the amount of data that gets transferred over the WAN and that guarantee sustained performance for critical traffic. Over the last few years, many IT organizations have deployed WAN optimization controllers (WOCs) to reduce the amount of data that gets transferred over the WAN and to guarantee performance for critical traffic. Unfortunately, the traditional WOC is unable to effectively optimize these large inter-data center traffic flows, because while the functionality they provide appears to be what is needed, WOCs were originally designed to support traffic between branch offices and a data center. Branch traffic is comprised of tens, if not hundreds, of slow-speed connections. As previously mentioned, inter-data center traffic is comprised of a small number of very high-speed connections.

What is needed is a new class of device - one that is distinct from the traditional high-end data center WOCs because this new class of systems must be purpose-built to support inter-data center traffic, and not branch to data center traffic. This new class of device will be referred to as Hyper-scale WAN Optimization Controllers ("Hyper-scale WOCs").

The remainder of this white paper provides a discussion of the requirements for a Hyper-scale WOC and then describes an example of this new class of device - the Infineta Data Mobility Switch (DMS).

Requirements for a Hyper-scale WOC

Hyper-scale WOCs must have the following characteristics and functionality:

High Throughput

Hyper-scale WOCs must be capable of saturating a multi-gigabit WAN link, even if the number of concurrent flows between the data centers is quite small. For example, if storage replication is the only active flow on the WAN, the device should have the processing power and the TCP protocol optimization functionality that is needed to fill a multi-gigabit/sec WAN link with traffic. This eliminates any significant amount of expensive and unused WAN bandwidth and improves the efficiency of operations such as storage replication, backups and VM migration.

Although it might be technically feasible to achieve high throughput by load balancing a very large number of optimization solutions, this approach is not economically or operationally feasible. A Hyper-scale WOC must therefore be able to achieve high throughput without resorting to this approach.

Transport Optimization

The Hyper-scale WOC's congestion control mechanism for TCP needs to be very aggressive in its control of window sizes in order to achieve high throughput and to consume all of the bandwidth allocated to each type of traffic flow. The Hyper-scale WOC must have very large buffers in order to shield the end systems at each data center from the effects of WAN propagation latency and any WAN packet loss. Additionally, it must be able to monitor and ensure optimal distribution of WAN resources across active flows, so that the WAN is full regardless of the number of flows active on the WAN at any given time.

Low Latency

A number of inter-data center operations are improved if the Hyper-scale WOC has very low internal port to port latency. For example, synchronous storage replication guarantees no loss of data because a write operation is not considered to be complete without an acknowledgement handshake between the local and the remote storage. In most cases, the subsequent write operation will not start until the acknowledgement of the previous write operation is received, which creates a ping-pong effect that magnifies the effect of WAN latency. Any significant device latency reduces the inter-data center distance over which synchronous replication is feasible, or in the worst case, makes the use of the device a practical impossibility.

A Hyper-scale WOC's internal latency can also be a significant factor affecting the inter-data center distances over which VM migration can be reliably performed.

Maximal Data Reduction

Data Reduction based on de-duplication and compression decreases the need for WAN bandwidth and further reduces the time it takes to complete the previously described inter-data center tasks. Storage replication and backup applications, as an example, typically send only those blocks of data that have changed since the previous transfer. In these cases, good de-duplication ratios depend on identifying patterns that are far smaller than the typical data block addressed by a disk system, which is typically 4 KB or larger. For maximal data reduction, the de-duplication implementation should be able to find repetitions all the way down to sub-10 byte packet segments - both within and across individual streams or flows. The efficiency of the de-duplication process should be independent of throughput, up to at least 10 Gbps. This means that the de-duplication engine within the Hyper-scale WOC has to have the processing power to look for short duplicate strings even at extremely high data rates. Data compression to provide further data reduction may occur after de-duplication has been performed, in order to make the data transfer even more efficient.

QoS and Traffic Management

Inter-data center WAN links typically carry a number of different traffic types with varying requirements for latency and bandwidth. Therefore, the Hyper-scale WOC must have a hardware-based QoS and traffic management system that can classify and prioritize traffic at multi-gigabit line rates and allocate bandwidth in accordance with

configured QoS policies. Based on these policies, the product must then apply the appropriate acceleration techniques and priorities to the various traffic classes.

High Availability

Given the business critical nature of accelerating inter-data center traffic, the Hyper-scale WOC must support high availability. For example, in addition to providing a number of internal high availability features, such as redundant power supplies, the device must support high availability network designs based on in-line or out-of-path redundant configurations.

The Infineta Data Mobility Switch

The Infineta Data Mobility Switch (DMS) is an example of a Hyper-scale WOC. Each of the core functions of the DMS (i.e., QoS, TCP Transport Optimization, De-duplication, and Compression) is supported in a separate hardware complex. These hardware complexes are arranged in a pipelined, distributed-processing architecture and are interconnected with a high speed, non-blocking switch fabric that boasts an aggregate capacity of 160 Gbps. Each DMS has four 10 Gbps ports for network connectivity and is capable of accelerating inter-data center traffic at up to 10 Gbps wire speed. Internal (port-to-port) latency on the DMS is measured in the 10s of microseconds, with many per-packet operations averaging approximately 50 microseconds.

The QoS complex employs a hardware-based policy engine to classify and prioritize incoming traffic. Business critical traffic that requires acceleration is directed into a pipeline of hardware complexes for optimization, while other traffic is forwarded downstream at line rate without incurring any extra latency. The QoS complex also can be configured to allocate minimum bandwidth guarantees for accelerated traffic to enable IT to advertise and deliver on service level guarantees to their internal customers.

The Transport Layer Optimization complex, called the Velocity Transport Engine™, uses multi-core network processors to perform transport-level acceleration at multi-gigabit speeds. Infineta's proprietary congestion control and avoidance algorithm, which is implemented within this complex, is unique in the way it allocates bandwidth across a number of TCP connections. By maintaining a dynamically updated view of competing flows and the status of the WAN link, the system can enable multiple high priority flows to aggressively take up large amounts of bandwidth without allowing inter-flow conflict. For example, connections for workflows such as replication and migration can be assured 1Gbps or higher throughput if needed. Lower priority connections can also be protected and assured lower, but sustained transfer rates consistent with configured policies. This complex acts as a transparent TCP proxy between the end systems and has very large buffers in order to shield the end systems from the effects of WAN packet loss and high WAN latencies.

Data Reduction is performed in three pipelined hardware complex stages: Repeating Byte Suppression, which removes repeating byte sequences from incoming data; the Velocity Dedupe Engine™; and standards-based Compression. The Velocity Dedupe Engine

implements a patented parallel processing de-duplication algorithm that leverages massively parallel programmable logic in the Velocity hardware complex. The parallelism allows chunks of received data as small as 8-bytes to be compared to data in the de-duplication dictionary at up to 10 Gbps rates. Overall, the three hardware complexes working together can reduce incoming data volume by as much as 80-90%

Summary

There is no doubt that the amount of traffic traversing inter-data center WANs has been growing at a dramatic rate and that it is highly likely that this rate of growth will increase over the near term. Given the critical nature of this traffic, IT organizations must find a cost-effective way to support it.

The traditional WOC was designed to solve the problems associated with supporting branch office to data center traffic. As such, these solutions are not effective at supporting the very high data rates that are characteristic of traffic traversing inter-data center WANs, while also performing massive data reduction and mitigating the negative effects that latency and packet loss have on standard TCP throughput.

Because inter-data center traffic has a unique set of characteristics, what is needed is a new class of device. This new class of device, the Hyper-scale WOC, should be designed to support inter-data center traffic. An example of such a product is Infineta's Data Mobility Switch (DMS). In addition to improving the effectiveness of inter-data center and cloud computing operations, the DMS can increase the effective capacity of large, expensive WAN links by a factor of five or more.