Enterprise networks for low latency, high frequency financial trading

Casimer DeCusatis
Distinguished Engineer, IBM Corporation 2455 South Road, Poughkeepsie, NY

Abstract

Emerging applications such as high frequency trading are driving new requirements for ultra-low latency data center network designs. In this paper, we review low latency messaging applications for enterprise server platforms and present the results of a recent audited industry standard benchmark test designed to simulate the effects of real time financial trading systems. Design best practices for networks within the data center and between redundant data centers are also presented.

Introduction

In recent years we have seen many fundamental and profound changes in the architecture of modern data centers, which host the computational power, storage, networking, and applications that form the basis of any modern business. Proper planning for the data center infrastructure is critical, including consideration of such factors as latency and performance, cost effective scaling, resilience or high availability, rapid deployment of new resources, virtualization, and unified management. The broad interest that information technology (IT) organizations have in redesigning their data center networks is driven by the desire to reduce cost while simultaneously implementing the ability to support an increasingly dynamic data center environment.

For example, modern data center designs should enable users to treat computing, storage, services, and network resources as fully fungible pools of assets that can be dynamically and rapidly partitioned without the infrastructure or the applications knowing details about each other. This concept of a federated data center is most typically associated with public or private cloud computing environments. Consider the case of a multi-tenant data center, in which the tenants represent clients sharing a common application space. This may include a private cloud which enables sharing data center resources among different divisions of a company (accounting, marketing, research), stock brokerages sharing a real time trading engine, government researchers sharing virtual machines on a supercomputer, or clients sharing video streaming from a content provider. In any sort of multi-tenant data center, it is impractical to assume that the infrastructure knows about the details of any given application or that applications know about details of the infrastructure. This basic design concept leads to greater simplicity and efficiency in the data center.

As another example, modern data centers should provide connectivity between all available data center resources with no apparent limitations due to the network. An ideal network would offer infinite bandwidth and zero latency, be available at all times, and be free to all users. Of course, in practice there will be certain unavoidable practical

considerations; each port on the network has a fixed upper limit on bandwidth or line rate and a minimal, nonzero transit latency, and there are a limited number of ports in the network for a given level of subscription. Still, a well designed network will minimize these impacts or make appropriate tradeoffs between them in order to make the network a seamless, transparent part of the data processing environment.

One application which has received considerable interest lately is the design of data centers to accommodate real time electronic financial transactions. Sometimes known as high frequency trading (HFT), this approach is currently responsible for over 1/3 of all stock transactions and is expected to grow significantly in coming years. The overriding design consideration for HFT applications is lowering latency, which refers to the total end to end time delay within the data center network due to a combination of time of flight and processing delays within the network equipment. Financial applications are especially sensitive to latency; a difference of microseconds or less can mean millions of dollars in lost revenue [1]. High latency translates directly to lower performance because applications stall or idle when they are waiting for a response over the network. Further, new types of network traffic are particularly sensitive to latency, including virtual machine migration and storage traffic.In the case of HFT, both the magnitude and consistency of the latency (jitter, or variation in packet arrival times) are important. Low latency is critical to high performance, especially for modern applications where the ratio of communication to computation is relatively high compared to legacy applications. Today there is a tradeoff between virtualization and latency, so that applications with very low latency requirements do not virtualize their applications. In the long term, this may change as increased speeds of multi-core processors and better software reduce the latency overhead associated with virtualization.

In this paper, we will describe several design principles for ultra-low latency networks, including connectivity between data centers over long distances as well as within a single data center. Recent industry standard benchmarking data for several different types of networking switches will be presented.

WebSphere LLM

WebSphere Low Latency Messaging (LLM) is a peer-to-peer messaging system based on patented technology from IBM [2]. It includes batching algorithms specifically designed to handle bursty traffic, as well as congestion management features to insure stability. For example, it can automatically detect slow consumers of data and suspend their traffic if required, as well as automatically monitor traffic queue depths in real time and identify anomalous behaviors such as high re-transmission rates. System or application bottlenecks can be identified using the built-in performance and latency statistical monitoring features. The configuration and monitoring are centralized, with remote site support for business continuity applications. Millions of logical flows can be supported on a single topic. To control latency, a messaging selection process attaches meta-data where required in the traffic stream. High availability features include sub-second failover with guaranteed zero lost messages.

WebSphere LLM features native Open Fabrics Enterprise Distribution (OFED) version 1.2 support for both InfiniBand transport and 10 Gigabit Ethernet transports using remote direct memory access (RDMA) technology. Both of the major industry standards, iWarp or RDMA over Converged Ethernet (RoCE), are supported. InfiniBand is the highest performance, ultra low latency protocol (1-9 microsecond range) with throughput up to 120 Gbps. The InfiniBand interface provides support for TCP-like and UDP-like communications with multicast. RDMA over Ethernet is slightly higher latency (10-15 microseconds or less), and throughput is only about 25% of the levels achievable with InfiniBand. However, it offers a very low cost, industry standard networking infrastructure which is simple to deploy and manage. RDMA over Ethernet also supports UDP-like communications and multicast (using RoCE).

Currently included as part of WebSphere Front Office for Financial Markets, this technology is being used in production environments such as market data distribution systems, internal trade execution systems, and exchange execution systems. Potential applications extend across all facets of the financial enterprise data center, including front office (sales and trading operations), middle office (market management, credit risk, regulatory compliance reporting, and profit/loss calculations), and back office (settlements, clearances, record maintenance, and accounting).

A network for low latency messaging solutions is based on the IBM-BNT 64 port, 1/10 Gbps Ethernet switch model G8264 [3]. This top of rack switch is the first single-chip switch to exceed 1 Terabit per second throughput (1.28 Tbps full duplex). A single-chip switch offers not only lower latency than a multi-chip switch, but also provides more consistent, deterministic latency to every switch port. Single-chip solutions also offer higher reliability and lower power dissipation (5.8 W/port or 375 W maximum). This is also the first production ready switch to offer 40 Gbps Ethernet for the data center using four QSFP ports (which can be reconfigured as 10 Gbps Ethernet using a fanout cable). The switch support converged enhanced Ethernet (CEE), and is compatible with iWarp, RoCE, and FCoE. High availability is insured through the use of concurrently maintainable dual redundant power and cooling.

The combination of WebSphere LLM and the IBM-BNT switch has demonstrated the lowest, most consistent latency on a recent audited industry-standard benchmark [4]. The Securities Technology Analysis Center (STAC™) is a vendor neutral benchmarking organization comprised of leading financial market firms, who write and maintain a library of test suites which represent customer-defined, simulated market trading environments. Testing with this benchmark is observed and audited by STAC™ and made available to their members and subscribing companies. This testing used WebSphere MQ LLM (version 2.4.0.2) running on IBM 3550 servers (model 7946-E2U) with dual Intel Xeon quad core processors (X5570, 2.93 GHz), using Red Hat Linux (version 5.5, 64 bit). Network attachment of the servers was provided by either a Mellanox ConnectX2 adapter or a Solarflare NIC with OpenOnLoad, connected through an IBM/BNT 8264 Ethernet switch in both cases.

Results of this test are shown in tables 1 as a function of message rate, with several other recently published STAC™ benchmarks included for comparison. Latency was measured in a "supply to receive" test in which one data source is transferred to a group of five consumers. In this case, mean latency was measured as 9 microseconds, with near zero standard deviation, and a peak supply rate of 1,500,000 messages per second. End-to-end latency is measured using a reflector test [4].

STAC-M2 Benchmark™ BASELINE Test Comparison (1 Producer, 5 Consumers)	Mean (us)	99P (us)	STDV (us)	Highest Supply Rate (msg / sec)
IBM LLM / IBM-BNT 8264 / Mellanox ConnectX-2 RoCE	**6**	**9**	**1**	**1,400,000**
IBM LLM / IBM-BNT 8264 / Solarflare w/ OpenOnload	**9**	**11**	**0**	**1,500,000**
IBM LLM / Juniper QFX 3500 / Solarflare w/ OpenOnload (Required additional EX4500 as IGMP Querier)	**9**	**11**	**0**	**1,500,000**
IBM LLM / Voltaire IB / Mellanox ConnectX IB	**8**	**11**	**1**	**1,000,000**
29W LBM / Cisco N5010 / Solarflare w/ OpenOnload	**14**	**17**	**1**	**1,300,000**
29W LBM / Cisco 4900M / Solarflare w/ OpenOnload	**15**	**18**	**1**	**1,300,000**

Table 1 – STAC-M2 benchmark™ testing results for IBM LLM solution; other recent benchmark tests shown for comparison.

Ultra Low Latency Connectivity Between Data Centers

Most of the latency associated with data center networks is incurred by the upper layer protocols (TCP windowing, flow control, packet retransmission and routing, store and forward, etc.). However, a significant amount of latency is also incurred from wide are network transport. There are three major sources of latency in the wide area network (WAN); fiber latency, WAN equipment latency, and the contributions of equipment in the fiber path (signal regenerators, amplifiers, and dispersion compensators). The fiber latency is fixed at 5 microseconds per km, and will be dominated by the WAN distance rather than distances within the data center. This is particularly difficult to adjust, since fiber paths are oftern indirect and much longer than the geographic distance between two locations. For connections between major cities, existing fiber routes are not particularly direct, and new, more direct fiber builds are often not economically justified since it is much easier to reinforce existing fiber routes. One notable exception is a 1,325 km direct fiber route constructed between New York and Chicago on the straightest practical path, which significantly reduces latency [5].

There are also potentially significant sources of latency in the long distance optical transport equipment. For example, optical transponders are used to convert an incoming data signal to a specific modulated optical wavelength for multiplexing purposes, or to aggregate lower data rates using time division multiplexing. The electronic time multiplexing, performance monitoring, protocol conversion, clock recovery, and forward error correction (FEC) algorithms used in this application are all sources of added latency. While this is usually negligible for typical applications, it can be significant for latency sensitive applications. Higher data rates (over 10 Gbit/second) require FEC in order to detect and correct bit errors, but this can add tens to hundreds of microseconds latency. Similarly, the convergence of optical and electrical signals in a sub-rate multiplexing architecture can be achieved using the industry standard IETF G.709, known as Optical Transport Network (OTN). This approach encapsulates user data in a digital wrapper to decouple the server links from the long haul links, and is commonly used to encapsulate lower data rate traffic into a 40-100 Gbit/second backbone. However, OTN encapsulation also introduces tens of microseconds additional latency, and should be disabled for ultra low latency networks. We also note that many vendor proprietary inter-switch links (ISLs) on Fibre Channel switches are not fully compatible with OTN, and thus OTN should be disabled if these interconnects are used for long distance transmission.

For distances exceeding 80-100 km, optical amplification and dispersion compensation are required. Optical fiber amplifiers consist of specially doped sections of fiber which may be tens to hundreds of meters or more in length. Optical signals passing through this fiber are amplified without requiring electronic to optical signal conversion, so the overall latency from an optical amplifier is lower than a corresponding electronic amplifier; there is a tradeoff in signal integrity since the optical amplifier cannot retime a signal like the electronic amplifier. Although the latency introduced by a single optical amplifier is typically very low (less than a few microseconds), for fiber links with poor noise figures, many amplifiers placed close together may be required, thus increasing the aggregate latency. The type of optical amplifier will also make a difference. Erbium doped fiber amplifiers (EDFAs) require longer fiber lengths within the amplifier, and thus add more latency compared with Raman amplifiers.

Extended distance links also require dispersion compensation, to overcome the fixed levels of chromatic dispersion associated with long distances of installed fiber. The type of dispersion compensator can make a significant difference in latency. One approach involves inserting spools of specially treated dispersion compensating fiber into the link, which have a negative dispersion shift and cancel out the positive dispersion associated with the rest of the fiber. A typical 100 km link can be compensated with about 14 km of dispersion shifted fiber, which adds about 70 microseconds to the link latency [6]. If the dispersion compensating fiber is not optimally placed, additional optical amplifier stages may be required, which further increases the link latency. Another approach is the use of dispersion compensation gratings, which are short lengths of optical fiber fabricated with a chirped fiber Bragg grating in their core. This diffraction grating is able to induce high levels of negative dispersion proportional to the optical wavelength; several possible designs have been proposed [7]. A 100 km length of fiber can be compensated using

only about 20 meters of fiber Bragg grating, with an additional latency of less than 0.15 microseconds. Although dispersion compensating grating are currently more expensive, the cost difference may be justified in cases where ultra-low latency is required. Additional latency tuning can also be achieved through tuning of the application environment, operating system, and hardware environment of the servers attached to the network [8], although these details are beyond the scope of this paper.

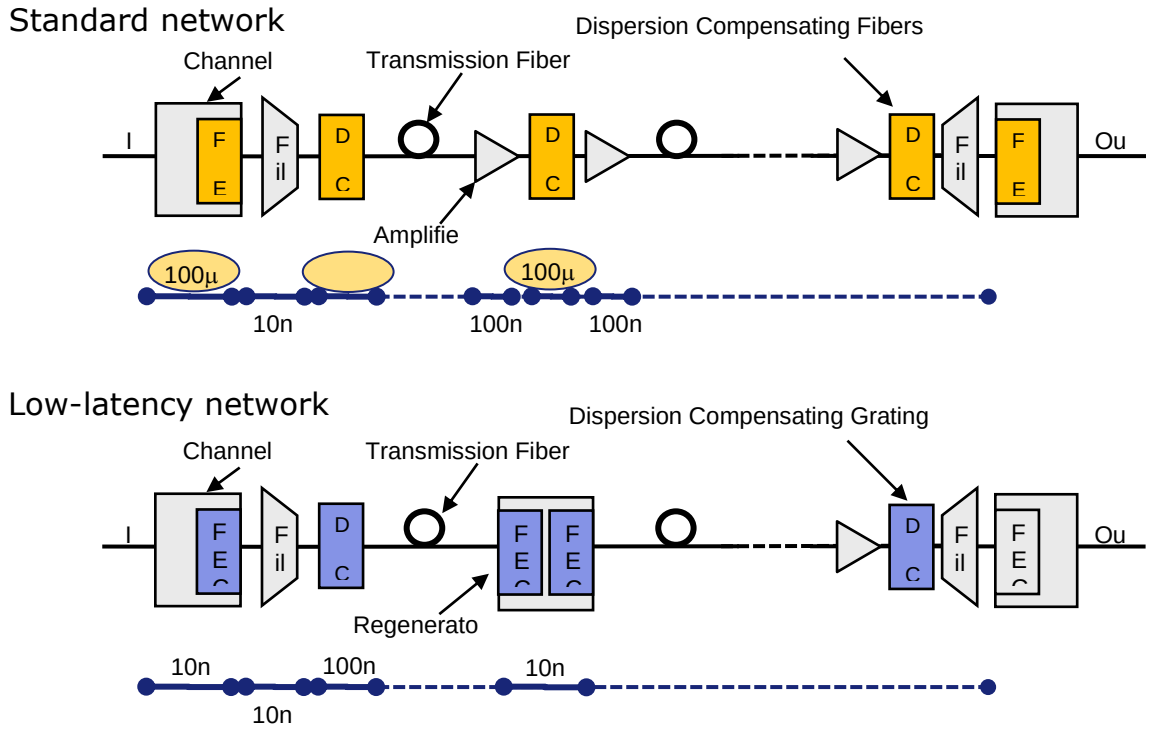

Figure 1 – Typical latency reductions achievable in long distance, ultra low latency networks (after ref. [5]).

Conclusions

For ultra-low latency applications such as high frequency financial trading, the data center network can introduce significant amounts of latency. Within a data center, the entire network stack must be considered, including the server adapter, top of rack switches, and core switches; end to end solutions which perform well on independently audited latency benchmark tests are recommended. For links between data centers, latency can be optimized by selecting the shortest possible physical fiber path, disabling FEC and OTN, using Raman amps instead of EDFAs, and using dispersion compensating Bragg gratings instead of dispersion compensating fiber. In the future, as transaction rates increase, we expect further reductions in latency will be possible through faster processors and network interface controllers, accelerated middleware appliances, and ultra-low latency switches, combined with a certain amount of tuning and design optimization.

References

1) A. Bach, “High speed networking and the race to zero”, Hot Interconnects (HOTI) conference, 11 Madison Ave, New York, NY, August 25-27, 2009; http://www.hoti.org/hoti17/program/

2) WebSphere MQ Fundamentals, IBM Redbook SG24-7128, 440 pages, first published December 2005, available from http://www.redbooks.ibm.com/abstracts/sg247128.html?Open ; see also “WebSphere MQ Low Latency Messaging”, IBM white paper, http://www.ibm.com/software/integration/wmq/llm/

3) “OpenFlow: the next generation in network interoperability”, IBM white paper, InterOp 2011 (May 8-10), Las Vegas, NV; also available from http://www.bladenetwork.net/userfiles/file/OpenFlow-WP.pdf ; see also “Low Latency Solution Stack for HFT”, IBM/BNT white paper, bladenetwork.net/userfiles/file/WP_IBM_SolarFlare_G8264.pdf

4) STAC™ benchmark reports, www.stacresearch.com/reports

5) T. Bundy et.al. , “Virtualized converged data centers in the cloud and their effect on optical networks”, Internet2 spring workshop, Arlington, VA (April 2011)

6) M. Loro and J. Gerrity, “Optical networks for low latency applications”, Ciena webinar and white paper, March 29, 2011

7) C. DeCusatis, “Low differential delay chromatic dispersion compensator”, US patent 7,689,077 (issued March 30, 2010)

8) “Best practices for tuning system latency”, IBM White Paper, March 2011, 17 pp., http://publib.boulder.ibm.com/infocenter/lnxinfo/v3r0m0/topic/performance/rtbestp/rtbestp_pdf.pdf