Introduction: Application Response Time is Critical in Cloud Environments

As data centers transition to next generation virtualized & elastic cloud architectures, high performance and resilient cloud networking has become a requirement for delivering always-on, always-available applications. Over the last decade, response time requirements have been rigorously tightened (see Figure 1), in some cases by as much as two to three orders of magnitude. In addition to the financial, education/research institutions and Web 2.0 firms, there is a growing list of customer verticals where fast response time is a top-of-mind business requirement, including rich-media companies (e.g. those involved in video post-production, digital animation and broadcasting), oil & gas producers, and health care.

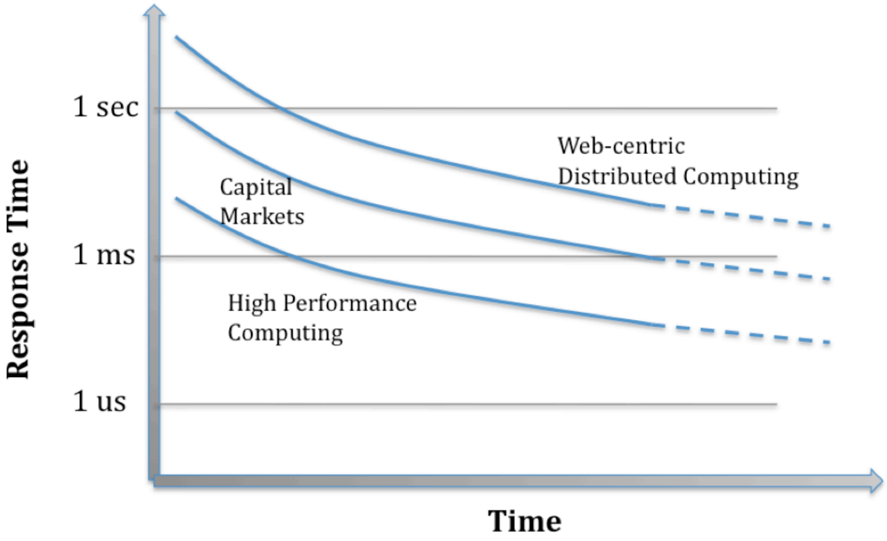


Figure 1: Application Response Time Trend

As response times gravitate towards fractions of a second, both network and compute latencies need to be reduced in parallel. In effect an architectural approach is needed so that response times are consistently and measurably enhanced across a wide set of cloud applications. Unlike prior efforts that often deploy niche and sometimes even proprietary technologies, today it is critical that low latency infrastructure fabric be built using standards-based and widely used technologies for broad-based cloud deployments. Also, because latency optimization is implemented at the infrastructure level, it needs to be future proof to meet latency demands 3 to 5 years out.

Architecting Low-latency Cloud Networks

A key attribute of latency-sensitive workloads is that they are built using distributed compute architecture. Each workload transaction spawns a large number of interactions between compute nodes, in some cases across thousands of machines and data stores, as depicted in Figure 2. For example:

- In algorithmic trading, machines receive market data feeds from other machines, calculate trading decisions and place trades with another set of machines.
- In HPC, thousands of commodity servers are clustered to behave as a “super computer”, connected to a clustered file system (e.g. Lustre) to crunch complex mathematical formulae and simulations.

- In web based cloud applications, workloads operate on thousands of dynamically clustered servers and implement multi-stage processing mechanisms to sift through petabytes of data. Example applications include those implemented by Web 2.0 and Internet Search companies, as well as emerging open-source initiatives such as Hadoop which provides scalable and reliable framework for distributed computing (MapReduce), distributed file system and distributed database.

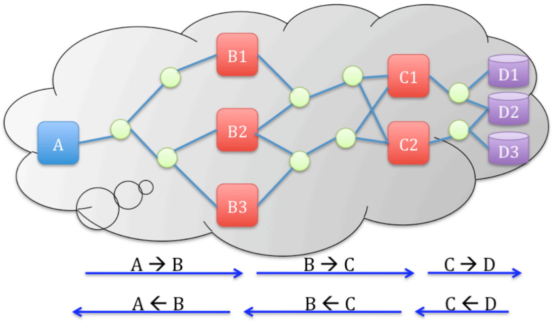


Figure 2: Network based machine-to-machine interactions of a single workload: A transaction

To boost system-wide response time, both compute and network latencies must be minimized. Typically, commodity servers are used for computing where degrees of freedom for latency optimization are limited. One essentially selects a server with denser and faster multi-core CPUs, larger and faster memories/caches, and speedier peripheral interconnects. The other side of latency coin is the cloud network where fabric-wide optimization is called for to minimize transport latencies. In particular, network latencies need to be optimized across four key axes:

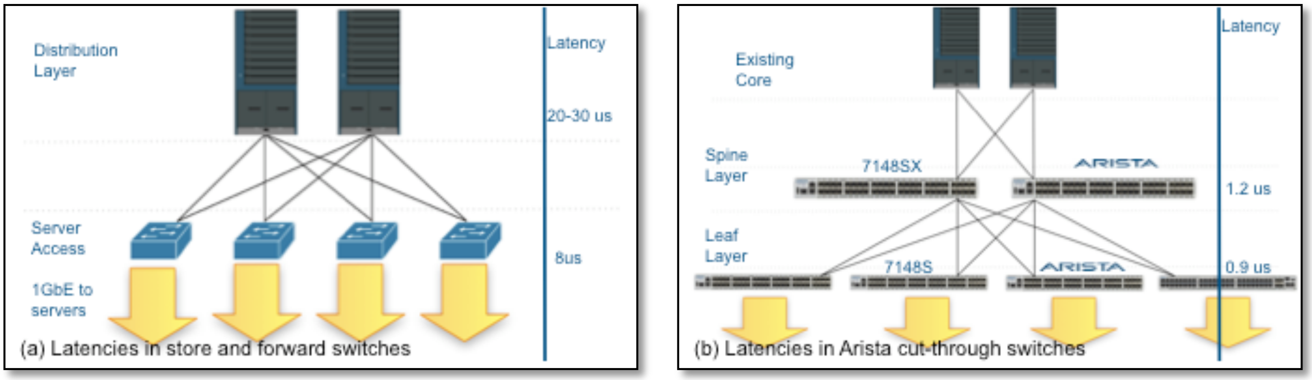
- Reduce latency of each network node
- Reduce number of network nodes needed to traverse from one stage to another
- Eliminate network congestion
- Reduce transport protocol latency

The Arista Advantage:

Arista Networks has adopted an architectural approach to building low latency cloud networking. Its wire-rate, non-blocking and ultra low latency 1Gb/10GbE switches, along with extensible operating system (EOS), enable resilient cloud networks for transporting data, multi-media, storage and compute traffic. Architecting cloud networks based on Arista's 7xxx series switch fabric has the following advantages:

- Reduce network latencies across all four axes (see details below);
- Future-proof the network infrastructure for latency demands of today and tomorrow;
- Implement standard Ethernet, which is widely deployed and operationally well understood.

Network Node Latency: For ultra low latency workloads, such as those in financials and HPC, shaving off every microsecond is important. Deploying the optimal 1/10Gb Ethernet switch can improve latency by 10+ microseconds for 1500B Ethernet frames (even more if the 9KB jumbo frames are used). Also, the cut-through switch architecture can reduce additional tens of microseconds of over older store-and-forward designs. Arista provides dense 10Gb switches with cut-through architecture that yields switching latencies in the range of 0.6 - 1.2 microsecond (see Figure 3).



(a) Legacy Store-and-forward Design	(b) Arista Cut-through Design
Intra-Rack Latency: 8us	Intra-Rack Latency: 0.6us
Inter-Rack Latency: 36us	Inter-Rack Latency: 2.4us

Figure 3: Latency benefits of Arista's cut-through switches over traditional store-and-forward switches

Number of Network Hops: Typical data center architecture consists of a three-tier architecture, with server access, distribution and core switch layers. Often packets need to traverse all the way to the core layer when communicating across server racks. In Arista's two-tier topology with leaf and spine switches (see Figure 4), latency is significantly reduced as packets traverse fewer hops.

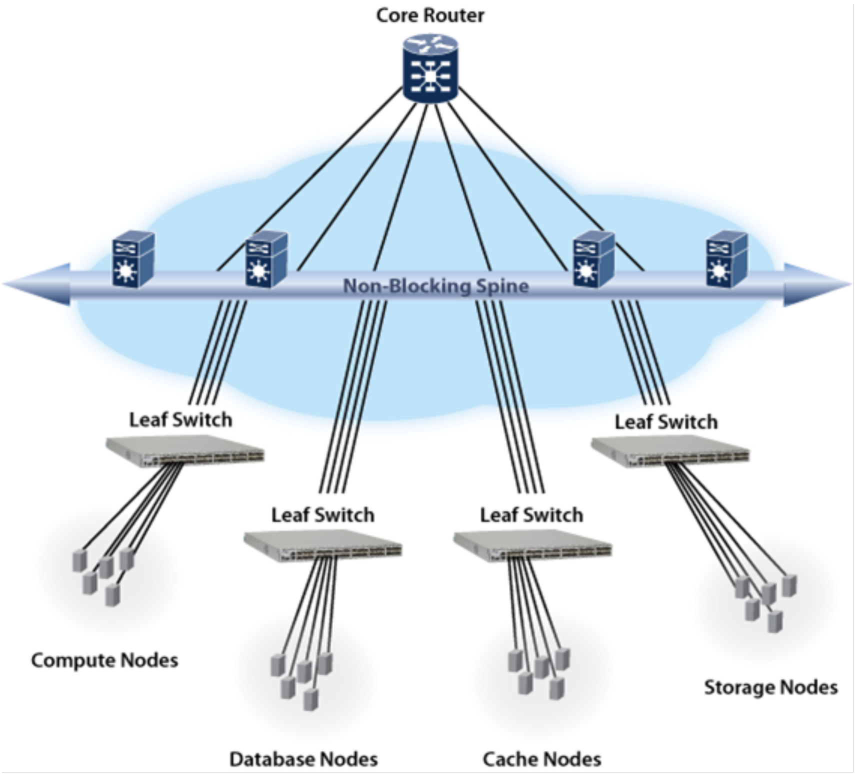


Figure 4: Arista's 2-tier (leaf and spine) topology reduces latency over traditional 3-tier (access, distribution, core) topology

Network Congestion: In most applications, IP/Ethernet packets are transported using the TCP protocol. To guarantee reliability and maintain throughput, TCP state machine uses windowing to adjust to dynamic network congestion. During congestion periods, TCP uses smaller windows, thus decreasing throughput and increasing overall transmission latency. Traditional data center architecture is oversubscribed, thus leading to congestions & dropped packets and hence large TCP latencies. Arista's 7100 switches are wire-rate switches (line rate for every 10Gb Ethernet port), thus enabling non-oversubscribed cloud network architectures with 1:1 bisectional bandwidth (as shown in Figure 4). Large, congestion-free, fabric can be architected that allows TCP to operate using large window sizes and hence at high throughput and low latency. Hundreds or thousands of machine-to-machine TCP interactions of application workloads, be they web, file/storage, multi-media or database, benefits immensely in terms of TCP throughput and latency.

Transport Protocol Latency: Certain low latency environments (e.g. HPC, applications leveraging Lustre file system algorithmic trading), it is desired to bypass TCP altogether and instead leverage direct machine-to-machine communication using message-passing (MPI) or remote direct memory access (RDMA). Infiniband and other proprietary interconnects have been used because of their ultra low latency, high throughput and lossless characteristics. Customers have longed for standards-based Ethernet to provide such capabilities as well as price-performance advantage so that they can benefit from the cost-efficiencies of managing a single Ethernet fabric across the public or private cloud. Arista's 7100 series of 10Gb Ethernet switches, with line rate and cut-through low latency performance, enables a single low latency cloud network. Arista's architecture also supports current and emerging Ethernet standards in IEEE and IETF on Layer-2/3 multi-pathing and congestion control to span large congestion free domains. Large Layer-2 domains, along with Arista's low latency, wire-rate and congestion-free cloud network, enable rapid migration of application workloads anywhere in the cloud.

Low-latency Application Example: Capital Markets

Algorithmic trading and market data feeds are the most celebrated use cases of ultra low latency designs where it is said that "1-millisecond advantage in trading applications can be worth $100 million a year to a major brokerage firm"[1]. Shaving off milliseconds is so critical that brokerage firms co-locate their trading systems right on the exchange floor for direct access to market data feeds, thus eliminating distance-based latency barriers. Modern designs strive for one millisecond response time for trading transactions, with the eye towards deep sub-millisecond designs in few years. Quest for speed has often led to deployment of exotic interconnects, such as Infiniband that are operationally difficult to deploy & manage at cloud scale.

Arista's line-rate Ethernet switch fabric, along with eco system partners such as Solace Systems, provides a complete algorithmic trading solution with 31-microsecond latency, as shown in Figure 5(b).

[1] "Business at Light Speed" by Richard Martin, InformationWeek, April 2007 (http://www.informationweek.com/news/infrastructure/showArticle.jhtml?articleID=199200297)

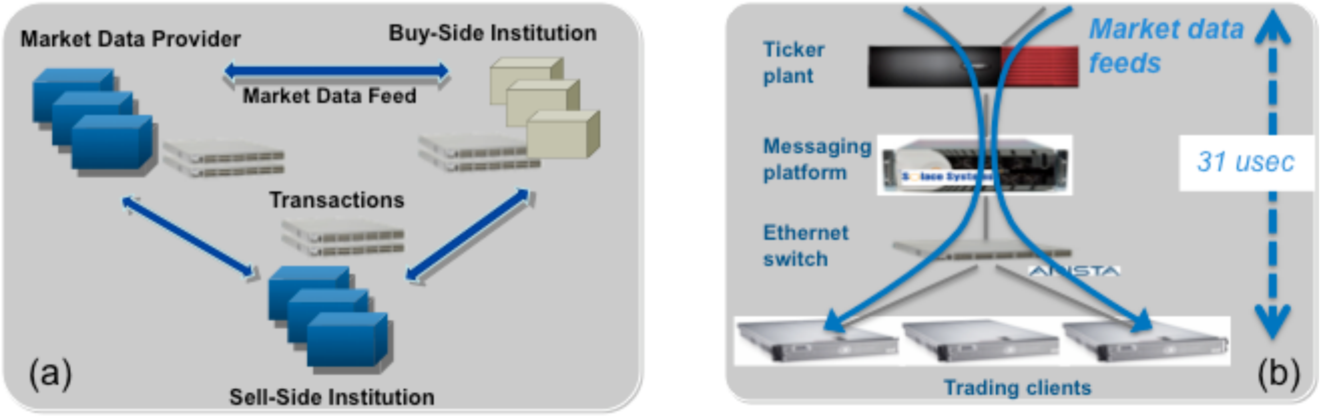


Figure 5(a): Market data feed and algorithmic trading 5(b): Arista + Solace Systems provides ultra low latency algorithmic trading solution

Summary

The drive for scalable and efficient cloud networks has led to 10Gb Ethernet as the interface of choice for simultaneously transporting data, storage and compute information over a common Ethernet cloud. Multiple classes of applications, including those in capital markets, high performance computing and web-centric computing, rely on latency-optimized architectures for ultra fast response times. After all, it is faster response times that ensure users' stickiness to a service, enhance scalability of HPC clusters and increase trading profits by speeding up algorithmic trade execution during price fluctuations. The Arista Networks 7XX series is designed solve these real word cloud-computing challenges for latency and scale.