

Table of Contents

Challenges 2

Business considerations 2

 Low latency..... 2

 High Throughput..... 2

 Scalability..... 2

 Power, Cooling and Space 3

Design considerations 3

 Low Oversubscription Ratios 3

 Multicast Traffic 3

 Dynamic, Multi-path Routing 3

 Loop-free topology..... 4

Future directions 4

 Flattened Layer 2 architectures 4

 Clos Networks 4

Solutions for Today and Tomorrow: High Bandwidth, Low Latency Gigabit and 10 Gigabit Ethernet Switches 5

 BLADE’s RackSwitch G8000 48-Port GbE Aggregation Switch 5

 BLADE’s RackSwitch G8100: 24 Port 10GbE CX4 Low Latency Switch 5

 BLADE’s RackSwitch G8124: 24 Port 10GbE SFP+ Low Latency Switch 6

Use Cases 6

 Use Case 1: 1GbE Server Aggregation using RackSwitch G8000 7

 Highlights: 8

 Advantages: 8

 Sample Bill of Materials 8

 Use Case 2: 10GbE Server Aggregation 9

 Highlights: 9

 Advantages 10

 Sample Bill of Materials 10

Solution Results 11

 About BLADE Network Technologies..... 12

 For more information 12

Challenges

In the financial services data center, low latency is a key criterion for maintaining a competitive advantage. Stock and options exchanges that provide market data feeds to service distribution networks must ensure their data arrives with minimal and deterministic (fair) latency, within a range of tens of microseconds from end to end. Service distribution networks, financial service providers, and system integrators that aggregate, normalize and provide analytics on market data rely on high speed networks to carry these computational transactions.

The much ballyhooed practice of algorithmic, high frequency trading—using high performance computers to enact millions of trades per second to take advantage of short-term fluctuations in share prices—is at the center of this trend, fueling the 164 percent increase since 2005 in the average daily trading volume in the New York Stock Exchange.¹

In this paper, we propose that some of the fundamental requirements for speeding up and scaling out the next generation high-speed financial services network include use of 1GbE and 10GbE switches to reduce latency and complexity. We provide some example use cases and recommendations for how to use Gigabit Ethernet and 10 Gigabit Ethernet switches from BLADE Network Technologies to achieve a high performance network that is simple, scalable and cost-effective.

Business considerations

Low latency

Low market data latency can be achieved with a combination of the right software and hardware, along with a network architecture that minimizes the number of tiers, or hops for data to traverse. A low latency network should also support technologies that minimize server CPU overhead. Since one of the main sources of latency is the protocol stack and NIC hardware implementation on the server, technologies such as Remote Direct Memory Access and TCP offload can help reduce server processing overhead. Cut-through switching and a latency-optimized data path can ensure that the switches' latency contribution to overall latency is negligible. 10 Gigabit Ethernet can significantly lower latency compared to Gigabit Ethernet. A typical 10GbE NIC will have a latency of under 10 microseconds, where a 1000BASE-T NIC will usually be over 50 microseconds.

High Throughput

High throughput, or the ability to send large files or many small files within a certain amount of time, is critical for market data delivery. Switching and routing equipment should demonstrate line-rate throughput (i.e., zero packet loss) under various configurations and packet sizes of the data transferred.

With prices dropping below \$500 per port, 10 Gigabit Ethernet (10 GbE) provides a cost-effective way to increase throughput in the network, as long as these switches can perform at line rate at all traffic loads and packet sizes. In the past, a 10GbE pipe running at only 50% utilization was still considered more compelling than aggregating two to four 1GbE links together, but today that is no longer the case.

Scalability

The next-generation financial services data center must also be able to support and scale to thousands, or tens of thousands of servers and switch ports, typically at 1GbE and 10 Gigabit Ethernet speeds. Data center switches must support high performance low latency and high bandwidth in a port-dense configuration.

Scalability means that the network can be expanded to meet increasing demands without adding a lot of cost. Scaling out with components that are cost effective, energy efficient and easy to manage generates greater returns than “scaling up” by adding more power and complexity to a smaller number of expensive components.

A modular approach of scaling the data center by building out replicated racks and rows of servers, storage and networking can shrink data center footprint, enable faster communication across fewer hops, and minimize overall application latency.

Power, Cooling and Space

The cost of providing power, cooling and data center space are also considerations when planning for a next generation financial services data center. Seemingly small differences in equipment footprint, power consumption, and cooling efficiency quickly become major differences when multiplied by several hundred or several thousand more switches, routers and servers added to the data center.

For example, a difference of 65 Watts between two different brands of switches can mean more than \$50.00 difference in power costs per year. In a data center deploying hundreds of switches, the energy cost difference can run into the tens of thousands of dollars per year.

As data centers grow in capacity, the need to put ever greater computing capability into a finite space (such as co-location facilities) has led to development of smaller and more space-efficient form factors. Data center equipment like blade servers typically pack 14 to 16 servers into a 9U to 10U high form factor, and new rack designs are available that can hold double the amount of servers found in a typical 42U high, 19 inch wide rack.

Airflow efficiencies such as the hot-aisle, cold-aisle model of data center design can lower energy costs for the data center only as long as all the equipment can be positioned or configured to support these directional airflow patterns. The networking equipment should support both front to rear or rear to front air cooling to match the predominant airflow of the servers; this is an important characteristic to consider for densely packed data centers that rely on air cooling.

Design considerations

Low Oversubscription Ratios

A low oversubscription ratio, or blocking ratio, is also a key requirement for maintaining application performance in the scaled-out data center. Simply put, oversubscription ratio is the amount of uplink bandwidth divided by the amount of server bandwidth. Acceptable oversubscription ratios vary depending on the application and the number of hops in the network. To prevent application bottlenecks, a low oversubscription ratio can increase application performance and decrease latency for market data applications. Higher oversubscription ratios lead to network congestion and drives up latency. In the financial services data center, most traffic is concentrated in east-to-west computational direction—therefore requiring a much lower oversubscription ratio than north-south traffic volume (to and from the Internet). A non-blocking, zero oversubscription ratio is the ideal, and can be designed into the network. However the networking gear should also demonstrate low latency and line-rate throughput in order for a non-blocking architecture to make a difference in application performance.

Multicast Traffic

Support for IP multicast traffic is required in the financial services data center so that traffic streams from multiple exchanges can be efficiently delivered to many end users without overloading the network. In multicast, data intended for many recipients is transmitted from a single stream. Exchanges send out market data broadcasts and other types of traffic in a few-to-many configuration which could quickly overload data center networks if sent out in unicast.

By default, all devices on a Layer 2 network receive multicast traffic floods. To control multicast flows in the Layer 2 network, Internet Group Management Protocol (IGMP) snooping must be supported in all switches in the Layer 2 network. With IGMP Snooping, each recipient that joins an IGMP group receives the multicast stream. While many switches employ up to IGMP v2 snooping, switches that support IGMP v3 snooping enable recipients to specify from which hosts they want to receive multicast traffic. IGMP v3 Snooping can also block traffic from other hosts inside the network; thus freeing up significant network bandwidth.

Dynamic, Multi-path Routing

Use of multipath routing topologies in the network, such as Open Shortest Path First (OSPF) with Equal Cost Multipath Routing (ECMP), can lower overall network latency because it enables optimal load sharing of traffic across multiple routes, while utilizing very low CPU overhead in Layer 3 switching.

Loop-free topology

One limitation in expanding the Layer 2 domain is the use of stacking beyond the access layer. In a Layer 2 domain, Access layer traffic can traverse uplink ports in a switch stack, allowing all switch uplink ports to be active. In an extended Layer 2 domain which includes the Distribution layer, stacked switches can cause a traffic bottleneck due to the large amount of traffic travelling among the stacked access switches. As trunk hashing forces traffic onto a stacking link, the overall cluster bandwidth becomes limited to the stacking bandwidth. A method of configuring switches to support a loop-free topology in an active-active configuration that allows the extension of stacking at the Distribution layer would eliminate the traffic bottleneck. This would free up bandwidth across the entire cluster, and enable a Layer 2 network to extend beyond a single data center.

Future directions

Networks tend to grow larger as capacity needs change, and often by adding more equipment to augment the existing low port-count switches and routers; this can become unmanageable over time due to cabling complexity and performance differences between the old and the new. While some proposals suggest a rip-and-replace strategy using new and proprietary technology, this approach is not always feasible in any but green field implementations.

Flattened Layer 2 architectures

Stock exchanges are looking to reduce latency by reducing the number of tiers in the network. While traditional networks are based on three tiers; at the access, aggregation and core layers, some new deployments are removing the aggregation layer and connecting access layer switches directly to a 10GbE switch core. Other approaches minimize the use of core routers and instead place more emphasis on creating a high-speed 10GbE aggregation layer to handle compute-intensive traffic within the data center.

One limitation to expanding Layer 2 networks is the need for the networking devices to be able to record and forward traffic to and from large numbers of MAC addresses. There are several proposalsⁱⁱ for how to achieve a large flat Layer 2 address space given the size constraints of MAC address forwarding tables in most switches on the market today. The Transparent Interconnection of Lots of Links or “TRILL” charterⁱⁱⁱ proposes a hybrid Layer 2 routing protocol that lets the switches learn where they are in the network without using MAC forwarding tables. By replacing Spanning Tree Protocol to detect loops in Layer 2 networks, TRILL would enable all links in a Layer 2 domain to be used; effectively increasing the available bandwidth in the Layer 2 domain.

Other architectures propose using a combination of MAC address in MAC address encapsulation, distributed load balancing, and ECMP to achieve network reachability, while employing low-cost, commoditized 10GbE switches as the interconnect for a cluster of hundreds of thousands of servers.^{iv}

Some of these new architectures are proposing to solve this problem by limiting the use of broadcasts and ARPs by utilizing a central directory service for address resolution. Many of these proposals are yet to be fully implemented; yet all point to use of low-cost, high performance 10GbE switching fabric as the key to the next generation high speed networks.

Clos Networks

A decentralized, non-blocking architecture, based on the Clos architecture initially proposed in 1953 has recently been extended into a 10GbE fabric by [Fulcrum's FocalPoint 10GbE switch chips](#). Using a Clos architecture, a port-dense, fully connected fabric can be created that connects high-bandwidth multi-tier switches in a non-blocking fashion. The Clos, or “Fat Tree” architecture, uses standard Ethernet switching with simple extensions to scale the data center.

As of this writing Clos is not being used extensively within Ethernet networks in the financial services data centers. However we do expect to see some Clos architectures with 10G Ethernet come into play in the next couple of years and have proposed several designs. We also expect to see 40G Ethernet in 2010 and 100G Ethernet soon after.

Solutions for Today and Tomorrow: High Bandwidth, Low Latency Gigabit and 10 Gigabit Ethernet Switches

While many vendors claim low latency ideal for the market data application environment, switches from BLADE Network Technologies provide the lowest latency along with significant other logical and physical characteristics to help financial services data center meet today’s challenges for low latency, low cost, low power and easy scalability.

BLADE’s RackSwitch family of extremely low latency, non-blocking and line-rate switches, with industry-proven BLADEOS operating system enable reliable transport for data, storage and computational traffic over a unified Ethernet fabric.

BLADE’s RackSwitch G8000 48-Port GbE Aggregation Switch

The RackSwitch G8000 is a 1-RU, fixed-configuration switch, with 48-ports of 10/100/1000 Mbps Ethernet, plus four optional 10 GbE uplink ports. Four of the 48 1GbE ports can be figured with SFP optics to support fiber connectivity.

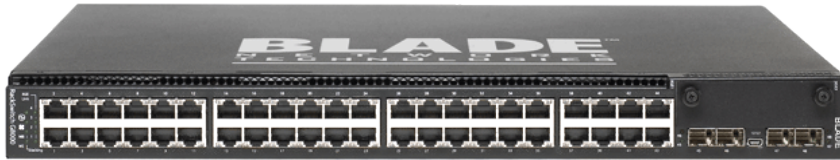


Figure 1 BLADE 48-port 1-GbE RackSwitch G8000—for Top-of-Rack Applications

Table 1: RackSwitch G8000

Feature	Description
Ports	44 x 1GbE RJ-45 Four 1G SFP ports Up to four optional 10GbE SFP+ or CX4 ports
Fabric Capacity	176 Gbps
Average Latency:	5.1 microseconds (port to port)
Expansion Modules:	2-port 10 GbE module
Power Supplies	Dual redundant (standard feature)
Power Consumption	124 Watts
Fans	Dual Redundant (standard feature)
Cooling	Front to Back or Back to Front models available to match predominant data center airflow

BLADE’s RackSwitch G8100: 24 Port 10GbE CX4 Low Latency Switch

The RackSwitch G8100 is a 1-RU fixed-configuration switch, with 20 10 GbE CX4 ports and four 10GbE SFP+ ports. A 480 Gbps switch fabric delivers non-blocking, line-rate performance on all 24 ports. Cut-through switching provides an average port-to-port latency of less than 300 nanoseconds.

The G8100 is ideal for datacenters requiring the lowest latency and the flexibility of supporting either 1GbE or 10GbE through its four SFP+ uplink ports.

BLADE's RackSwitch G8124: 24 Port 10GbE SFP+ Low Latency Switch

The RackSwitch G8124 is a 1-RU fixed-configuration switch, with 24-port 10 GbE SFP+ ports. A 480 Gbps switch fabric delivers non-blocking, line-rate performance on all 24 ports; Cut-through switching provides an average port-to-port latency of less than 680 nanoseconds.



Figure 2 BLADE RackSwitch G8124—for High-Performance, Low-Latency Cluster computing with cost-saving SFP+

Table 2 RackSwitch G8100 and G8124

Feature	G8100	G8124
Ports	24 x CX4+	24 x SFP+
Fabric Capacity	480 Gbps	480 Gbps
Average Latency:	360 nanoseconds (port to port)	680 nanoseconds (port to port)
Power Supplies	Dual redundant	Dual redundant
Power consumption	120 Watts	115 to 168 Watts ¹

The G8124 is ideal for datacenters requiring the flexibility of supporting either 1GbE or 10GbE through its SFP+ interface. The G8124 is low power, consuming between 115W to 168W depending on the speed of the port (1G/10G), type of transceivers (SR or DAC) and number of active ports.

Use Cases

In the financial market data environment, data centers require the highest bandwidth combined with the lowest latency in order to achieve competitive advantage in time to market. Server clusters must combine the processing power of multiple CPU's that appear as a unified or single computational systems, and must connect to the network with low latency, high bandwidth interconnects.

Using a rack-and-roll deployment model such as BLADE's Rackonomics, clusters can scale out a rack at a time by adding pre-installed and configured racks of servers and/or storage with top-of-rack switches which are assembled offsite.

The following two use cases are proposed here for scaling servers and networking a rack at a time. ^{vi}

¹ Power consumption varies depending on the speed of the port (1G/10G), type of transceivers (SR or DAC) and number of active ports

Use Case 1: 1GbE Server Aggregation using RackSwitch G8000

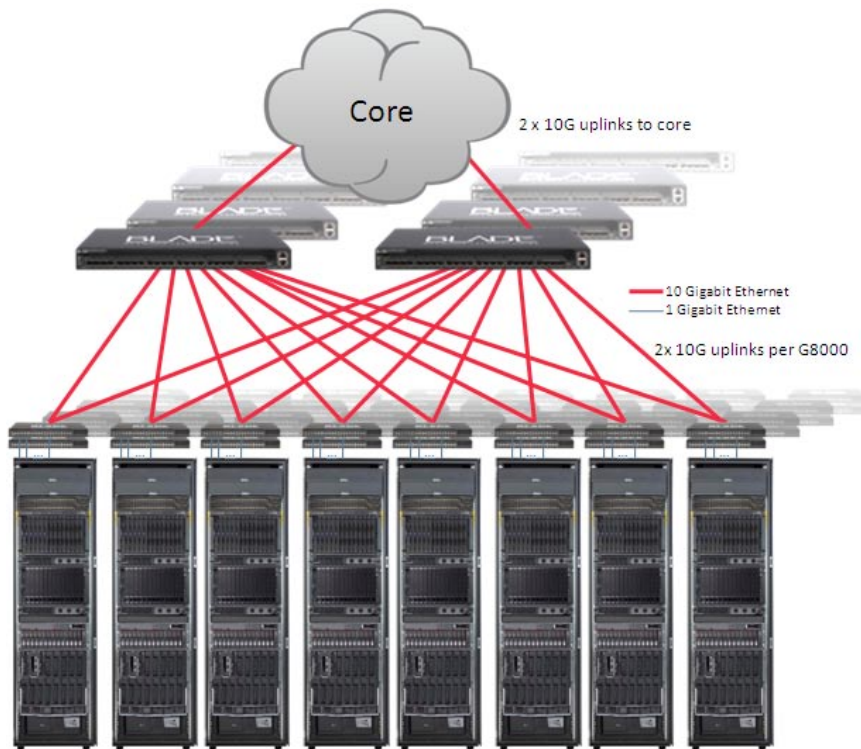


Figure 3 1GbE Server Aggregation

As shown in Figure 1, racks of servers with 1 Gigabit Ethernet interfaces are connected to Top of Rack BLADE RackSwitch G8000 1/10Gb Switches as the access layer interconnects, and to 10GbE G8124's at the aggregation layer.

This configuration uses two of the four available RackSwitch G8000's 10GbE uplink ports to provide uplink bandwidth of 20 Gb per switch.^{vii} A low blocking ratio² between access and aggregation layer is used to minimize the latency in east-west application traffic within the data center. A much higher oversubscription ratio is acceptable between aggregation layer switching to the core routers; the amount of bandwidth needed to send traffic destined to the Internet in most financial calculations is much smaller than the bandwidth needed to carry the east-west computational traffic from within the data center.

In this example, maximum flexibility can be achieved initially and at scale while maintaining the bandwidth, latency, and oversubscription ratios initially designed.

² Alternately this configuration can be modified to use only two of the 10GbE uplink ports and re-allocate some of the uplinks to servers to accommodate up to 48 servers and a higher blocking ratio of 2.4:1

Highlights:

- 40 servers per rack (node)
- Each server has dual redundant 1 x 1GbE connection to a RackSwitch G8000
- Cluster is scalable up to 32 nodes to support up to 1,280 servers.
- Uses 2x10GbE uplinks per G8000 switch to aggregation layer RackSwitch G8124s
- Line-rate from servers to access layer
- 2:1 oversubscription ratio from access to aggregation layer
- 8:1 oversubscription from aggregation G8124 switches to core network
- Server bandwidth: 20G/40=500 Mbps per server
- BLADE Active Multipath used for loop free topology without Spanning Tree Protocol

Advantages:

- Design allows flexibility in scaling out by allowing administrator to start with a minimum number of switches and servers, to meet bandwidth, latency and oversubscription requirements and grow flexibly to maximum scale while maintaining these same values
- Low 2:1 blocking between access and aggregation layers for best performance in east-west computational traffic
- Works in many data centers still employing 1GbE rack servers
- Active Multipath (AMP)^{viii} eliminates blocking of looped paths, and effectively doubles uplink bandwidth

Sample Bill of Materials

Item	Description	Quantity
Servers	Rackable servers with dual 1GbE Interfaces	40 per rack Up to 1280 servers in a 32 node cluster
10GbE Aggregation Switch	BLADE RackSwitch G8100 or G8124	2 to 8
Top of Rack 1/10G Switch	BLADE RackSwitch G8000	8 to 32
Cabling	RJ-45 dual homed server to G8000s	80 per rack
	SFP+ DAC G8000 10GbE Uplinks	2 per switch=16 to 64
	SFP+ DAC G8124 10GbE Uplinks Or G8100 CX4 10GbE Uplinks	2 per switch= 4 to 32

Use Case 2: 10GbE Server Aggregation

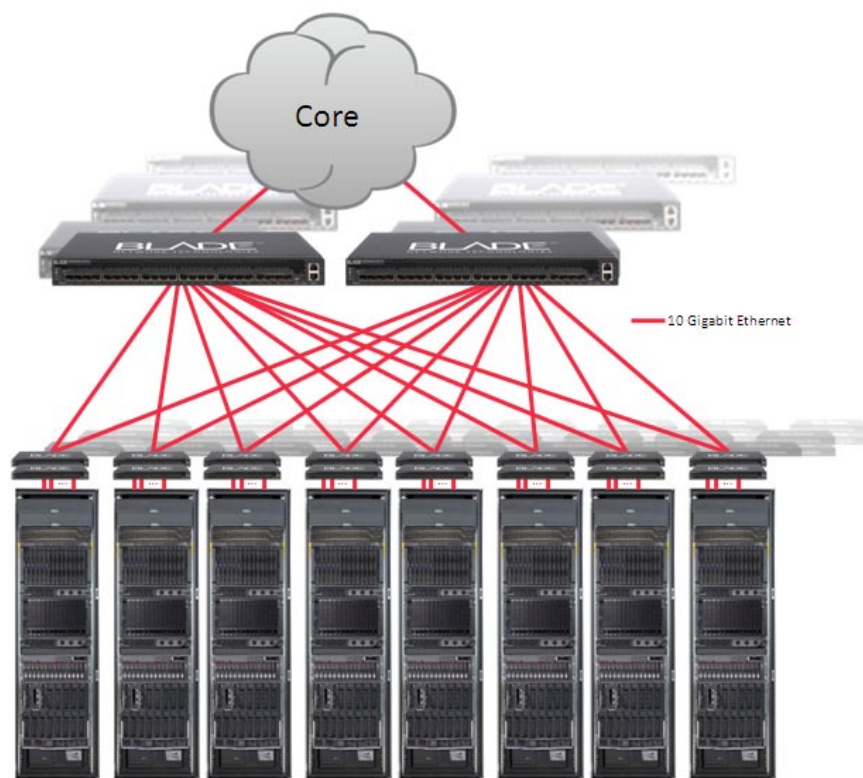


Figure 4: 10GbE Server Aggregation

Similar to the first example, this is a rack-based solution but relies on the RackSwitch G8100 or G8124 10GbE Low Latency switch to act as a top of rack access switches for 10GbE rack servers. It is assumed that all 1RU servers come with dual-homed 10GbE NICs. RackSwitch G8100 or G8124 switches are also used at the aggregation layer to provide a low-latency, high throughput switching fabric for high-speed computational traffic within the data center.

Highlights:

- 20 servers per rack (node)
- Each server has dual redundant 10GbE connections to a RackSwitch G8124
- Cluster is scalable up to 32 nodes to support up to 1280 servers.
- Uses 4x10GbE uplinks per G8124 top of rack access switch to connect to aggregation layer RackSwitch G8124s
- Line rate from servers to access layer
- 5:1 oversubscription from access to aggregation
- 8:1 oversubscription from aggregation G8124 switches to core network
- Server bandwidth: 20G/40=500 Mbps per server
- High Availability with Uplink Failure detection
- BLADE Active Multipath used for loop free topology without Spanning Tree Protocol

Advantages

- Design allows flexibility in scaling out by allowing administrator to start with a minimum number of switches and servers, to meet bandwidth, latency and oversubscription requirements and grow flexibly to maximum scale while maintaining these same values.
- Non-blocking between access and aggregation layers for best performance in east-west computational traffic
- Keeps east-west computational traffic away from core network
- Works in many data centers employing 10GbE rack servers
- Active Multipath (AMP) effectively doubles uplink bandwidth as no paths are blocked

Sample Bill of Materials

Item	Description	Quantity
Servers	Rackable servers with 10GbE Interface	20 per rack Up to 640 servers in a 32-node cluster
10GbE Top of Rack Access Switch	BLADE RackSwitch G8124	Two per Rack Up to 64 in a 32-node cluster
10GbE Aggregation Switch	BLADE RackSwitch G8100 or G8124	2 to 8
Cabling	SFP+ DAC (G8124)	48 uplink 40 server to switch

Solution Results

Today's financial services data center needs an environment with minimal application latency that can still scale out easily and effectively. As networks move away from hierarchical core-centric models towards a flatter architecture, scale out will be achieved through the lowest cost, highest performance interconnects available—and today that is the 1GbE and 10GbE switches offered by BLADE Network Technologies.

Here is how BLADE's RackSwitch products meet the requirements for today's high-speed financial services data center:

Requirement	BLADE's Solution
Low latency	Port to port latency: <ul style="list-style-type: none"> • G8100: 360 nanoseconds • G8124: 680 nanoseconds
High Throughput	Line-rate throughput at 100% traffic load
Scalability	Best in class price and performance
Power, Cooling and Space	Power: Best in class power consumption: 120W-165W depending on model Cooling: Server friendly airflow: All BLADE RackSwitch models come with front-to-rear or rear-to-front airflow to match the airflow direction of your servers Space: shorter depth for horizontal or vertical placement in standard or high density racks.
Low Oversubscription	High performance cut-through 10GbE switching fabric Ideal in low or non-blocking configurations
Multicast Traffic	Supports IGMP v1, v2 and v3 Snooping for efficient use of Multicast bandwidth
Dynamic, Multi-path Routing	Support for OSPF with ECMP provides optimal load balancing
Loop-free topology	Support for Active Multipath effectively doubles usable bandwidth as loops blocked by Spanning Tree are available in active-active mode. Built-in fault tolerance and 100% active uplinks is ideal for east-west computational traffic
Robust hardware	Redundant power supplies and redundant fans, which usually are additional cost add-ons with other vendors, come standard in all BLADE RackSwitch models
Simple and Compatible	Scriptable CLI for easy customization and deployment Easy to use Web GUI Compatible – tested to work with Cisco core infrastructure. BLADEOS ISCLI similar to Cisco Configurable with BLADEHarmony Manager network management software for one-click updates to entire network of BLADE switches. BLADEHarmony Manager also integrates into existing network management tools

About BLADE Network Technologies

BLADE Network Technologies is the industry leading provider of Gigabit & 10Gb Ethernet switches for cloud networks and data center virtualization. BLADE's RackSwitch family demonstrates "Rackonomics," a low cost, low power and efficient way for scaling out data center networks. BLADE's customers include over 300 of the Fortune 500, representing 260,000+ network switches and 6M+ switch ports connecting over 1.3M servers and storage systems.

For more information

For benchmark results and more information, visit BLADE on the web at

<http://www.bladenetwork.net/financial-services.html>

©2009 BLADE Network Technologies, Inc. All rights reserved. Information in this document is subject to change without notice. BLADE Network Technologies assumes no responsibility for any errors that may appear in this document. All statements regarding BLADE's future direction and intent are subject to change or withdrawal without notice, at BLADE's sole discretion.
<http://www.bladenetwork.net>.

MKT090917

i <http://dealbook.blogs.nytimes.com/2009/07/24/traders-profit-with-computers-set-at-high-speed/?scp=2&sq=high%20speed%20trading&st=cse>

ii: <http://conferences.sigcomm.org/sigcomm/2008/workshops/presto/papers/p57.pdf>

<http://nfarring.wnmh.net/pdf/portland-sigcomm09.pdf>

iii <http://www.ietf.org/dyn/wg/charter/trill-charter.html>

iv Albert Greenberg, Prantap Lahiri, David A. Maltz, Parveen Patel, Sudipta Sengupta, Microsoft Research, Redmond, WA USA: "Towards a Next Generation Data Center Architecture: Scalability and Commoditization" <http://conferences.sigcomm.org/sigcomm/2008/workshops/presto/papers/p57.pdf>

vi . Configurations for implementing firewalls, intrusion detection, storage and other key elements of the network are beyond the scope of this paper.

vii Alternately this example may be modified to employ all four available 10GbE uplinks for a non-blocking configuration from access to aggregation, with appropriate adjustments being made to cluster size given the number of available ports.

viii Optional future feature