WHITE PAPER

Scaling-out Ethernet for the Data Center

Applying the Scalability, Efficiency, and Fabric Virtualization Capabilities of InfiniBand to Converged Enhanced Ethernet (CEE)

Data center architecture is constantly evolving, with several key trends now emerging:

Data center consolidation: Building larger shared (private or public) data centers instead of many smaller ones

Focus on application and business services: Moving away from manual IT processes

Virtualization anywhere: Servers, I/O, storage, networks, and applications are virtualized and decoupled from physical hardware

Fabric convergence: Networking, storage, and inter-process communication (IPC) from multiple applications traveling over the same physical wire

These trends have significant impact on the fabric architecture of the data center. Fabrics must now support larger-scale Layer 2 (L2) networks since server virtualization and mobility, new storage protocols, and low latency messaging must reside on the same L2 domain. The new data center infrastructure must also be set to overcome the management complexity and address virtualization from the ground up.

The Data Center Bridging Group of the IEEE worked to enhance Ethernet to support the architectural trends outlined above, and in doing so adopted many capabilities from InfiniBand. These include class isolation, low latency, I/O and switch virtualization, lossless traffic flows, congestion control, multi-path L2 routing, and L2 discovery and capability exchange. These new technologies are referred to as Converged Enhanced Ethernet (CEE) or Data Center Ethernet (DCE).

Mellanox's InfiniBand products and fabric management solutions have been addressing such challenges and capabilities for many years. Mellanox delivers the fabric for the world's largest supercomputers and fastest financial trading platforms. Mellanox's CEE switches and software are built on this expertise, allowing end users with less demanding performance requirements to benefit from a far more scalable Ethernet fabric that also lowers overall fabric costs, lowers power consumption, has greater efficiencies, and simplifies management.

This document describes the challenges inherent in traditional Ethernet solutions and how Mellanox's scale-out Ethernet architecture effectively addresses those challenges.

The Need for Scalable Data Center Networks

Computation infiltrates all aspects of our lives. More content and services are now digitized, requiring more storage and associated computation and network resources. As a result, data center capacities are constantly growing at a fast pace while budgets and available power remain at the same levels.

To further increase data center efficiencies, organizations are trying to leverage economies of scale. Rather than hosting multiple smaller data centers, organizations are choosing to consolidate to fewer locations each at a much larger scale. In some cases, organizations are looking to public cloud providers that can host multiple virtual data centers in the same physical location. As a result, public and private clouds are growing at unprecedented rates.

One way to enable large-scale data centers is to increase server densities. New designs allow nearly 100 nodes per rack, and more than 1000 CPU cores per rack. However, these designs require a greater emphasis on power and cooling and create new challenges in switch cabling and switch densities. As an example, 1-2 server racks may have more connections than some of the largest 10GbE switches in the market.

The key technologies enabling increased data center efficiency are virtualization and automation solutions that can squeeze more virtual servers into the same physical resources. These solutions also automate many day-to-day tasks such as delivering a new computation service, conducting maintenance and migrating loads among different hardware platforms.

As data centers grow and become denser, more virtualized and automated, the load over the underlying fabric increases significantly. Today we see the following trends emerging:

- Extensive use of shared external file (NAS) or block (SAN) storage drives significant amounts of traffic throughout the data center, with increasing demands for high reliability, high service quality, and high peak performance.
- Server virtualization forces much higher capacities over the same physical node, requiring equivalent capacity increases in the attached NICs and networks.
- Server and application mobility drives more communication between different physical segments/racks, and server migration—which requires moving the entire virtual server memory footprint from one node to the other—adds to the load.

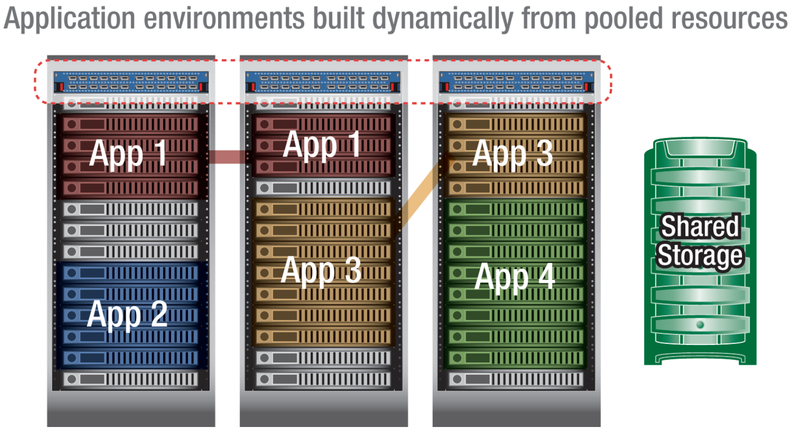
Application environments built dynamically from pooled resources
App 1
App 1
App 3
App 3
App 2
App 4
Shared Storage

Figure 1. *Next generation data center architecture*

With the need for greater efficiency, clustering or scale-out technologies are becoming more widely used. Examples of multiple servers that are interconnected and deliver the same logical function can be found in web clusters, database clusters, clustered file systems or Map-Reduce processing in cloud computing. Scale-out and application clustering technologies in these environments make extensive use of messaging and data movement/replication. This not only increases the load on the fabric, but also requires lower latencies, lossless and predictable behavior, and high burst performance.

It is important to note that much of the aforementioned traffic is limited to the same L2 (bridge) network domain—a virtual machine migrates with its IP address and cannot be mobilized to another IP network segment. In addition, some storage or messaging protocols (such as FCoE, RDMAoE, and PXE) are limited to L2. It is clear that new data center networks need to support much larger L2 switching domains, and cannot use L3 routing for out-of-the-rack communication as they could before.

To summarize, next generation data center networks require:

- A very large number of nodes at higher densities
- Lower power consumption
- Higher bandwidth per port
- Less bandwidth aggregation between tiers
- Lower latencies and predictable behavior
- Multiple, large, L2 (bridge) domains, with fewer switching tiers
- Segment isolation (partitioning) and traffic class isolation (CoS)
- Virtualization and virtualization fabric awareness

Apart from the technical requirements, all of the above capabilities need to carry a reasonable price tag for both operational and capital expenditures. These requirements cannot be met by traditional Ethernet products, so a new category of scalable and data center-optimized switches has been developed to address such challenges.

Mellanox's 20-40Gb/s InfiniBand switching fabric addresses the bandwidth, latency, scale, and density requirements outlined above, and is currently deployed in production environments throughout the world's largest and most traffic-intensive server farms. Mellanox's CEE switches and software, described in detail later in this paper, borrow many of the features of Mellanox's scale-out InfiniBand solutions and also address these requirements.

Traditional Data Center Networking Architecture

In legacy data center designs, the data center was divided into physical silos, with each silo containing a set of servers or a rack that ran a specific application (in one or more application tiers). The application had little communication with the external world since most of its intensive transactional, messaging, and data/storage traffic ran within the rack and only a fraction of that traffic was delivered to consumers outside of each rack (see Figure 2).

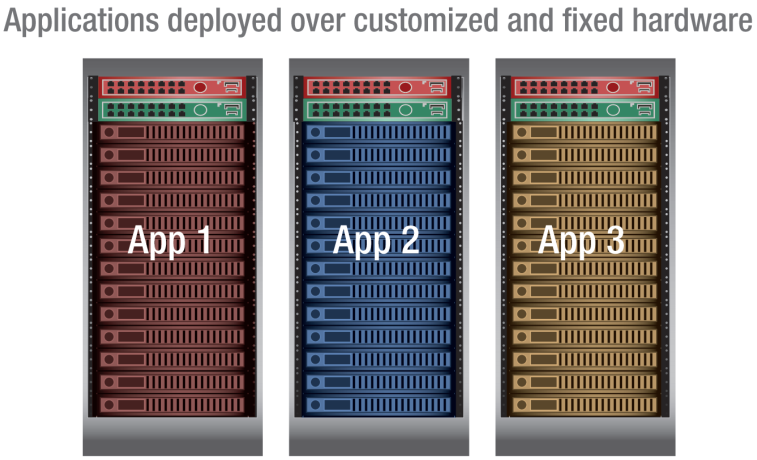

Figure 2. *Legacy data center architecture*

The legacy architecture shown in Figure 3 is enabled by top of rack (TOR) access layer switches that handle internal L2 communication (bridging), with a small set of uplink ports connected to core or distribution aggregation switches, resulting in high oversubscription. In many cases the silos have a unique IP subnet, and the aggregation switches implement L3/L4 routing between those racks/silos and external consumers or other silos.

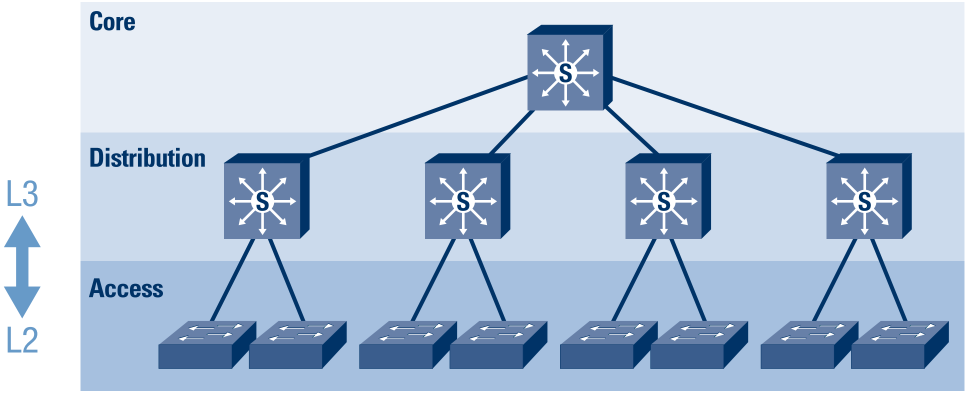

Figure 3. *Traditional three-tiered network design*

In legacy environments, core and distribution switches were always designed to support many complex network services for LAN, WAN, and enterprise communication with deep packet inspection and manipulation capabilities, large IP (L3) routing tables, large content addressable (CAM) tables, and some computation-intensive tasks like web/XML processing, encryption, and long distance optics. Because of their design and evolution, aggregation switches have a very high price per port, and much higher power consumption than access or blade switches. The legacy core switches were used primarily to deliver north-south traffic flowing from the clients to external networks or the Internet, placing less emphasis on application messaging and storage delivery requirements such as low latency, congestion avoidance, and reliability.

In Ethernet environments, there is traditionally a very high oversubscription rate (5:1 to 10:1) between server-facing ports (connected to the low-cost access switch) and the aggregation ports (facing the more expensive and power hungry core/aggregation switches), reducing the weight of the aggregation switches in the overall solution. However, with the growing requirements for CPU capacity and the introduction of server virtualization and mobility, there is much more east-west traffic between server racks and between servers and storage. As a result, such aggregation is no longer acceptable.

Difficulties Scaling Data Center Networks

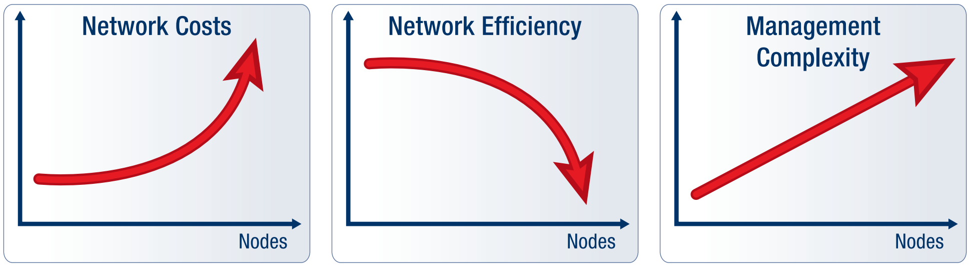

Figure 4. *Results of Ethernet's poor scalability*

Existing network switching and software solutions don't scale well. As networks grow, the cost grows exponentially, efficiency drops, and management complexity increases. As shown in Figure 4, customers building large data centers pay much more but get much less per each additional capacity growth unit.

Cost implications

As described in the previous section, traditional networks are designed with lower cost blade or TOR (top of rack) switches interconnected by much more expensive (and power hungry) aggregation switches. As the size of networks increase, additional aggregation tiers are needed to connect smaller network segments together. As a result, for every usable server node, multiple network ports are used to connect switches. The ratio of server ports to network ports increases with scale, and as we scale, we use more of the expensive (aggregation) switch ports.

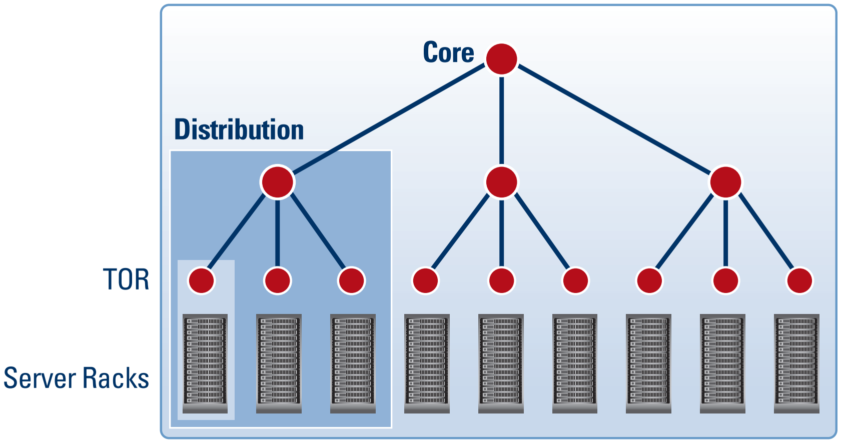

Figure 5. *Ethernet hierarchical (tree) model*

Until recently, users mitigated the cost and scalability problem by using a very high oversubscription ratio between aggregation tiers. As an example, a 48 port switch was connected to 32-36 servers and had only four uplinks to the core switch. In this case a 256 port 1 GbE core switch could support 2,000 connected servers. Anything beyond this number of servers would require another switch tier (Figure 5).

However as described previously, changes in the data center are driving lower oversubscription rates as well as a transition from 1 GbE to 10GbE. This results in much higher network costs in large-scale environments.

Performance degradation

Traditional network designs are not only expensive in large scale, but are also less efficient.

The oversubscribed, hierarchical nature of the network creates bottlenecks as we leave the rack communication path between nodes on different racks traversing multiple aggregation switches. This adds significant latency (especially since aggregation switches are slow) and exposes the communication to network congestion (which can easily occur due to the high oversubscription ratios). When congestion occurs in the network, switches drop packets to notify the source about that congestion, which only causes greater delays that impact application performance.

The traditional Ethernet bridging protocols (spanning-tree) are designed to avoid loops, even at the expense of performance degradation. If the network contains multiple possible paths between end-points, the protocol disables all of the ports that may lead to the same destination. This behavior significantly affects network scalability since the overall network bandwidth (bisectional bandwidth) is limited to the bandwidth supported by the root aggregation switch.

In fabrics such as InfiniBand and Fibre Channel, multiple paths are allowed and managed by a fabric manager, which maximizes the utilization of all ports in the fabric. While guaranteeing a loop-free fabric, multiple root switches can run in parallel (linearly scaling the bandwidth), or different mesh topologies can be applied. The IEEE and IETF organizations are currently defining extensions to Ethernet that will allow similar behavior for Ethernet (such as TRILL). However, it will most likely take several years to develop a multi-vendor standard and interoperable solution.

Device-oriented system management

Most network management applications are built using a device-centric approach. Each device exposes some standard or proprietary APIs/MIBs, and the management station identifies the API/MIB as a managed object and conducts its operations at the device level.

Typical management systems show network topologies (interconnected devices), aggregate events and alarms from multiple devices, and allow some device-level configuration. In this case, routers or firewalls at junction points enforce the networking policies.

However, device-oriented management systems are not suitable for next generation data centers, which require automation and virtualization to deliver the infrastructure as a service. Moreover, there is a disconnect when it comes to translating application-level requirements into fabric policy and configuration and a very low degree of automation. Another challenge is that network provisioning tools are implemented separately from monitoring tools. As a result, it is quite difficult to track the impact of manual or automatic policy changes on the network and application behavior.

As networks grow, the complexity of managing them grows proportionally. Management solutions must focus on the services delivered by the network (such as connectivity, security/isolation, QoS/CoS, availability, and statistics) rather than on the devices forming the network. Such an approach can deliver much greater scalability and can more effectively address data center automation and virtualization requirements.

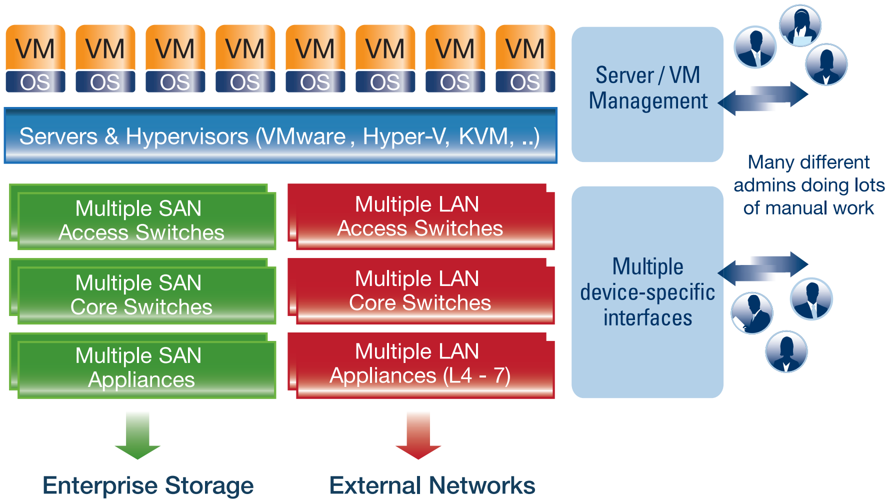
VM
OS
VM
OS
VM
OS
VM
OS
VM
OS
VM
OS
VM
OS
VM
OS
Servers & Hypervisors (VMware, Hyper-V, KVM, ..)
Server / VM Management
Many different admins doing lots of manual work
Multiple SAN Access Switches
Multiple LAN Access Switches
Multiple SAN Core Switches
Multiple LAN Core Switches
Multiple SAN Appliances
Multiple LAN Appliances (L4 - 7)
Multiple device-specific interfaces
Enterprise Storage
External Networks

Figure 6. *Traditional networks are managed manually using individual device interfaces*

Scale-Up vs. Scale-Out

When more capacity is needed, there are two possible approaches to take:

1. Scale-up: Choose a bigger monolithic machine
2. Scale-out: Scale horizontally by clustering multiple smaller machines

Formerly, when larger computation tasks were required, users bought large mainframes or proprietary supercomputers. This scale-up approach allowed greater simplicity with a single machine to manage and control, and typically had greater reliability.

However, the scale-up approach does have a few key limitations:

- **Expense:** Larger machines cost much more than multiple, smaller machines with the same aggregate capacity.
- **Limited Scalability:** Even the largest machine is still not large enough in many cases.
- **Lack of modularity:** It is expensive and complex (or even sometimes impossible) to resize the machine after it is deployed.
- **Availability:** Once a system fails the service is disrupted, and chassis-level redundancy cannot guarantee complete fault-tolerance.

As the demand for new IT services and storage grows beyond available machine capacity, it is evident that scale-out designs are required. Current scale-out solutions address greater application capacities and/or larger storage capacities at a reduced cost, while also addressing some of the manageability challenges.

In a scale-out design, multiple industry-standard elements are stacked and coupled using clustering software that forms a unified system view. Interestingly enough, since industry-standard components are produced much more quickly than high-end systems, they usually feature the latest technology. So, these single element capacities are typically not far from the capacity offered by higher-end systems.

There are many examples of scale-out solutions in the data center—such as clusters of web servers, clustered applications and databases, and clustered file or block storage—that demonstrate the clear advantage of such an architecture. All of these examples show improved performance and capacity with fewer resources. The success of any cluster solution is dependant on its software architecture. How much does the cluster behave like a single unified system? Does it scale linearly? And how seamlessly does it address failures?

Many data center network configurations are still designed with a scale-up approach. In this type of configuration, a large, heavyweight, expensive core switch sits at the center of the network with all traffic feeding into it. Because these large core switches are overloaded with functionality, they are too slow, too expensive, and require too much power—and become a bottleneck.

A scale-out data center design coupled with a unified fabric management application serves the next generation data center better than the scale-up architecture and offers significant cost performance benefits.

Scale-out Data Center Fabrics and CEE

InfiniBand (and to some degree Fibre Channel) fabrics were designed from day one for scale-out data centers with a very lightweight switching infrastructure, the ability to run mesh topologies and multiple paths, fabric and I/O partitioning capabilities, and central discovery and policy management. The Converged Enhanced Ethernet (CEE) standard incorporates InfiniBand-like capabilities and enables Ethernet to be used as a unified and scale-out data center fabric, making technology developed by InfiniBand vendors—such as Mellanox—very relevant to Ethernet.

Mellanox InfiniBand and CEE solutions are designed for scale, allowing horizontal scalability without performance degradation while maintaining a linear cost model. Like other scale-out solutions, Mellanox's fabric (and any third party physical or virtual switches and I/O adapters that may be attached) is managed as a whole with central monitoring, central fabric and I/O partitioning, central QoS/CoS, central security and HA enforcement, and central administration—allowing for greater efficiencies, simplicity, and data center automation.

Mellanox Scale-out Data Center Fabric Architecture

Mellanox's data center fabric architecture is designed to address the new challenges IT organizations face, such as:

- Data center consolidation
- Extensive use of server virtualization
- High density racks and the use of powerful CPUs
- Automation and cloud computing service orientation

The following sections will describe Mellanox's scale-out fabric architecture based on Converged Enhanced Ethernet (CEE), which delivers a scalable and efficient fabric for servers and storage in the data center.

Efficiency: Cost, Power, Latency & Density

Mellanox's CEE fabric solution comprises very high-density top-of-rack (TOR) and core switches that can switch many 10GbE ports at wire speeds and without oversubscription. In addition, the switch architecture provides several unique traffic management capabilities that reduce the latency to approximately one microsecond, while at the same time guaranteeing application performance and lossless or lossy behavior for the various traffic classes.

The scale-out design of the switching fabric couples less complex and less expensive hardware elements with a powerful scale-out software stack, enabling lower costs, lower power consumption, and higher switching density than other aggregation switches available today. Furthermore, multiple TOR and core switches can be meshed together to form enormous topologies without losing these advantages.

Mellanox's line of Vantage™ 10GbE switches enable new levels of efficiency, scalability and real-time application performance, while at the same time consolidating multiple/redundant network tiers and significantly reducing infrastructure expenses. The Mellanox Vantage switch family includes the following:

Mellanox Vantage 8500:

The Vantage 8500 switch is a completely modular, high-performance, high density Layer-2 core switch, optimized for enterprise data center and cloud computing environments. The Vantage 8500 delivers up to 288 ports of 10GbE connectivity and features an impressive 11.52 Tb/s non-blocking backplane, allowing applications to perform at maximum bandwidth and efficiency. As a focused, L2 core switch, it seamlessly integrates with existing switch infrastructures, and enables application performance at the lowest bandwidth to power ratio—with only 600-1200 nanoseconds of port-to-port latency and power consumption as low as 10 watts per port.

Mellanox Vantage 6048:

The Vantage 6048 switch is a high performance Layer 2/3 10GbE top-of-rack switch optimized for enterprise data center and cloud computing environments. With 48 ports of 10GbE line-rate connectivity and 960 Gb/s non-blocking switching throughput, the Vantage 6048 features the most comprehensive Layer 2 protocol stack, enabling the industry's most robust and advanced 10GbE fabric. It features the industry's highest density and converged Ethernet capabilities, and is equipped with the most advanced congestion management and flow control protocols.

Mellanox Vantage 6024:

The Vantage 6024 switch is a low latency, high performance Layer 2/3 top-of-rack switch that features the industry's most power-efficient and lowest latency capabilities on 10 Gigabit Ethernet. With 24 ports of 10GbE line-rate connectivity and power consumption as low as 115 Watts (4.8 Watts/port), the Mellanox Vantage 6024 switch features an impressive Layer 2 and 3 protocol stack. With its low latency silicon core, the Mellanox Vantage 6024 enables the fastest Ethernet fabric, featuring less than 700 nanoseconds of port-to-port latency under a full load and with a wide range of transceivers.

Linear Scalability

From their inception, Mellanox's solutions were designed with a scale-out architecture that allows multiple switches to appear as one very large Ethernet switch to any external device. This transparent architecture allows simple integration in heterogeneous switching environments while delivering on the core benefits of scale-out architectures.

InfiniBand-like enhancements to Ethernet (such as the TRILL protocol) would allow mesh configurations based on simpler building blocks. However, these enhancements are still incomplete, and will not work with existing switches or multi-vendor environments due to the new packet formats used.

This model is unprecedented in a market where the typical solution has an exponential cost model and demonstrates server performance degradation when trying to scale-out an environment.

Figure 7 compares a Mellanox-based fabric with the alternative, which uses traditional Ethernet aggregation switches in a hierarchical design. The results clearly demonstrate the advantages in cost and efficiency in larger scale configurations.

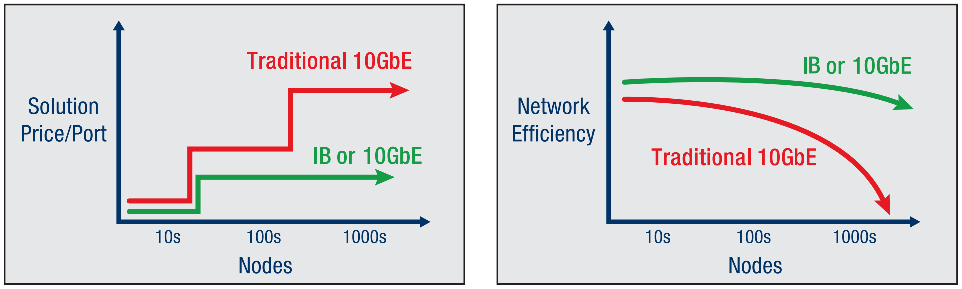

Figure 7. *Mellanox's scalability (green) vs. other networking solutions (red)*

Thus, the Mellanox solution allows users to build larger consolidated server and storage farms at much lower costs while maintaining the highest performance levels.

Virtualized from the Ground Up

Mellanox's Unified Fabric Manager™ (UFM™) software is a management software platform that delivers what is required in today's virtualized networks. UFM discovers all of the physical switches (including supported third party switches) and virtual switches on the network, and dynamically configures them to deliver on the desired service or application requirements. UFM constantly monitors the switches and the traffic flowing through them to ensure a properly managed network and adherence to the user policies.

UFM accepts application level requirements or definitions of virtual entities and constantly adjusts the end-to-end fabric policy across all switches to accommodate the requirements, eliminating the need to manually configure the individual switches, track the physical to virtual relations, or understand how the switches are interconnected.

One of the key services delivered by a switching infrastructure is to connect multiple end-points and allow traffic to flow between them, as well as between the end-points and external connections/ports. UFM follows this concept by defining and managing virtual or physical end-points, and enforcing the traffic policy between them, regardless of their physical locations. UFM can also segregate a single physical end-point to multiple virtual end-points (in case multiple virtual machines or VNICs reside on the same node).

Examples of managed end-points include:

- A NIC, HCA or HBA on a virtual or physical machine (Virtual I/O)
- Storage elements (target, LUN/Volume, or file server)
- Router/gateway port or uplink port

Multiple end-points can be grouped as well, such as when a set of end-points in a virtual network needs to communicate on the same L2 domain and when an application cluster contains a set of network interfaces—one per every node in that cluster.

As data centers adopt converged fabrics, it is critical to define the specific class of a virtual I/O end-point. This will determine its default policy and behavior. The three main types of virtual I/O end-points are:

1. Network adapter (NIC), with traditional LAN characteristics
2. Storage adapter (HBA) (requires lossless fabric behavior)
3. Messaging/IPC adapter (HCA) (requires low latency and lossless behavior)

Virtual end-points may change their physical locations dynamically, such as when a virtual machine migrates from one node to another. In this case, the policy that describes the end-point behavior, or the connectivity between any two or more end-points, is maintained, and traffic monitoring is kept in sync.

Focus on Application Performance and SLA

Mellanox's UFM allows users to define applications, application I/O and network requirements, and application flow requirements (connections between application entities). The intelligent resource manager in UFM acts as an optimizer, and automatically produces the traffic policy for all switches to guarantee the application behavior. It also maximizes performance and reports resource conflicts to users or external automation tools.

As an example, an application may have a few tiers connected in a certain topology. It may also have certain requirements for storage traffic and external (uplink) traffic, and may require low latency for its inter-messaging communication. These requirements are easily modeled and stored using UFM. When an application is started and physical server and storage resources are assigned to these applications, UFM will configure the switching infrastructure to provision the desired topology by partitioning the fabric and creating virtual I/O end-points, and by enforcing the traffic between the end-points according to the specified policy.

UFM constantly samples the traffic on the virtual I/O entities and application flows, reporting bottlenecks or statistics back and mapping to the application objects. A user may then decide to change his/her preferences to improve application performance.

For messaging traffic, Mellanox provides a variety of fabric protocols and protocol extensions that link to the application and allow the fastest inter-application messaging and communication, resulting in a significant application boost. Some examples include multicast acceleration, message queue acceleration, MPI enhancements, Oracle RDS, and storage acceleration.

UFM application-driven fabric resource management and monitoring enables increased fabric utilization, increased application performance, isolation of applications, minimized cross interference, and much greater visibility into the application performance.

Service Oriented Management

As illustrated earlier in this paper, Mellanox's complete fabric solutions focus on the services delivered by the fabric, which can consist of hundreds of interconnected switches. This is achieved through application and service modeling, and through application-oriented monitoring.

As illustrated in Figure 8, the fabric is mapped to three layers, including physical objects (such as servers, switches, storage, and ports), virtual objects (such as virtual machines, virtual I/O, and volumes), and application objects (such as application tiers and connections). In addition, there are physical or logical group objects that define a collection of physical or logical individual objects. A rack, for example, is a group of servers in the same location. A user can monitor, read, or modify the objects and their attributes, or even get notified of status changes or thresholds associated with those objects.

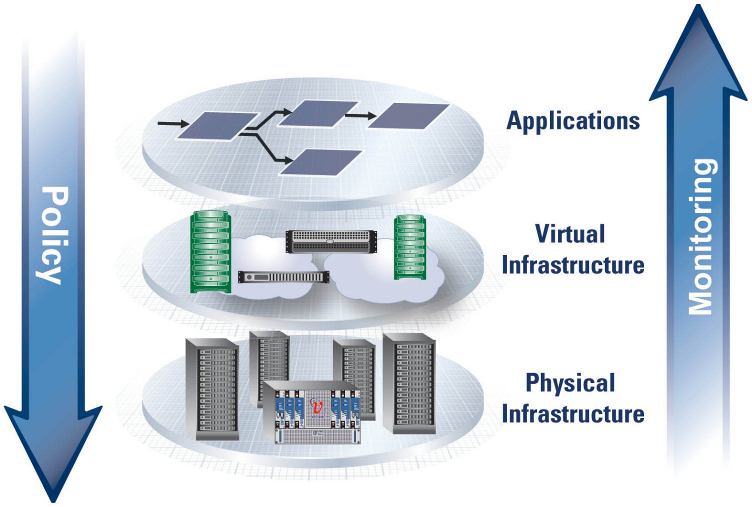

Figure 8. *Mellanox UFM three-layer data center model*

Users or external automation tools do not need to manage individual devices and their attributes. Instead, they can define the desired fabric services and allow the fabric manager to enforce these services and provide high-level feedback. The fabric can be easily managed through an extensible web-services API, through an object-oriented context aware CLI, or through a web-based graphical interface.

Delivering Fabric as a Service (FaaS™)

The Need for FaaS

As more and more IT organizations focus on consolidating data centers, simplifying their operations, and implementing automation concepts, they can pool server and storage resources and assign them to applications on demand.

New automation and virtualization technologies allow treating Infrastructure as a Service (IaaS), and brokering user requests with available resources. This is further enhanced with private or public cloud architectures that allow "renting" of server infrastructure or even complete applications.

Data center consolidation cannot be achieved without controlling the underlying fabric, and since the environments are automated and managed through service concepts, one cannot manually configure the fabric. Thus, a service-oriented fabric management paradigm is needed.

The following are key challenges in data center consolidation and cloud computing that relate to the fabric:

- There is a lack of isolation and security between virtual data centers
- There is a lack of service level monitoring and enforcement for shared fabrics
- Virtualization for servers and storage have a weak correlation to fabric policies
- Application mobility and virtual machine (VM) migration require synchronization with fabric policies
- With CPU consolidation and virtualization the bottleneck transfers to I/O; thus I/O and fabric resource optimization is needed to remove the bottlenecks and ensure application performance
- The fabric layout, current load, and application I/O must be taken into consideration when placing jobs and VMs over the fabric
- Administration and troubleshooting are difficult in scale-out environments
- The impact of fabric congestion or oversubscription on application performance cannot be easily measured

These challenges are addressed by Mellanox's Fabric as a Service (FAAS) technology, which extracts server interconnect networks from their physical elements and controls them as variable logical entities.

Key benefits of Mellanox's FaaS architecture include:

1. Delivering application-level SLA, performance monitoring and optimization
2. Controlling multi-tenant and virtualized applications in a single large fabric
3. Enabling complete data center automation

Mellanox Unified Fabric Manager (UFM): Delivering FaaS

Mellanox's UFM delivers FaaS by pooling and owning all of the fabric resources—including virtual or physical switching elements and I/O adapters—and by providing central fabric monitoring and service-oriented fabric policy enforcement.

As seen in the screenshots below, UFM manages the complete lifecycle of the fabric.

Dashboards, Physical & Logical Monitoring

Central Alarms and Logs

End-to-End Virtualization and Provisioning

Central Administration and Troubleshooting

Figure 9. *Complete fabric lifecycle managed by UFM*

UFM delivers four main services:

1. Physical, logical, and application level monitoring

- Collect statistics and information from physical and virtual switches
- Generate statistics and traffic analysis per VM, specific I/O, or traffic flow
- Get high-level, aggregated information via dashboards and APIs

2. End–to-end fabric provisioning and virtualization

- Carve the fabric into multiple classes of service for LAN, IPC, and storage
- Apply QoS (priority, limits, guarantee) per I/O or traffic flow
- Secure physical partitioning, VLAN, virtual I/O, and ACL provisioning
- Automatically configure hypervisors, virtual switches and vNICs, and automate VM migration tracking
- Guarantee HA (multi-rail, multi-path configuration, and policy synchronization)
- Ensure congestion isolation, control/throttling, and monitoring
- Suggest optimal placement based on fabric allocation or load

3. Central alarms and logs

- Trigger alarms automatically following specified physical or logical/aggregated events
- Collect statistics and status from multiple and diverse sources
- Auto-correlate raw data with virtual, logical or application objects
- Invoke any external tool based on physical or logical fabric events

4. Central administration and extensibility

- Centrally manage multiple switches through a single console
- Enable launch-in context and remote shell/scripting for any device or application in the fabric
- Store device, appliance, or object specific meta-data using an extensible model and database
- Benefit from an open web-services based framework with a rich set of APIs and SDK

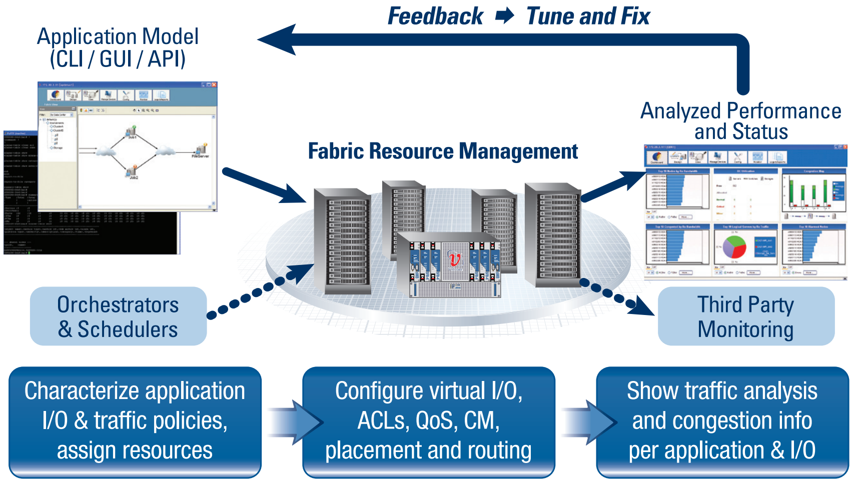
Feedback ➡ Tune and Fix
Application Model (CLI / GUI / API)
Analyzed Performance and Status
Fabric Resource Management
Orchestrators & Schedulers
Third Party Monitoring
Characterize application I/O & traffic policies, assign resources
Configure virtual I/O, ACLs, QoS, CM, placement and routing
Show traffic analysis and congestion info per application & I/O

Figure 10. *Fabric management lifecycle*

Figure 10 illustrates the fabric management flow controlled by UFM:

1. Application requirements are characterized by users or external orchestration or automation tools via the GUI, CLI, or web-services API
2. Physical or virtual resources are assigned to the application templates manually or automatically via an external scheduling or resource management tool
3. Fabric is configured and optimized to deliver on the desired application policy and maximize application performance
4. Statistics and status information is gathered from all switching elements and optional agents, and mapped to the applications and application flows, generating alarms if needed
5. Statistics and fault information can be used to manually or automatically adjust fabric behavior
6. Users or automation tools can apply changes such as migrating virtual machines, increasing or decreasing capacities, and changing connectivity to running objects, resulting in fabric re-adjustments

Fabric Management as Part of Cloud Management

Fabric management is a critical element in automated data center management and cloud computing, since data center elements cannot be provisioned without establishing and configuring the underlying connectivity.

Figure 11 demonstrates how fabric management is integrated as part of the overall cloud management infrastructure.

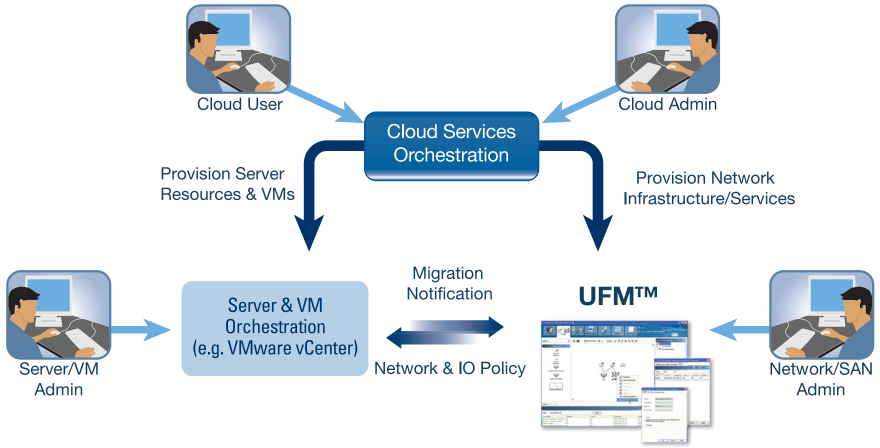
Cloud User
Cloud Admin
Cloud Services Orchestration
Provision Server Resources & VMs
Provision Network Infrastructure/Services
Server/VM Admin
Server & VM Orchestration (e.g. VMware vCenter)
Migration Notification
Network & IO Policy
UFM™
Network/SAN Admin

Figure 11. *Overall cloud management architecture*

In this architecture, users submit requests for new applications/resources to the cloud service orchestration/scheduling layer. This is performed via a cloud user portal, through which some resource definitions, service level requirements, or business goals are specified. Typically, user portals will also address account management, credentials, and billing. Then, a set of provisioning tasks are initiated to optimize the infrastructure for running these applications based on some workflows and service templates.

At the same time, infrastructure and applications are constantly being monitored. Feedback is provided to the service automation layer, which may throttle server or fabric resources up or down in order to meet service objectives, automatically handle service interruption, or use any reported data for billing and accounting.

The storage is typically divided between application/OS images and data storage. The application software images and the application data can be maintained in a shared NAS or SAN storage. All of these storage elements are connected over the same fabric, using proper connection provisioning and isolation.

Two types of infrastructure resource management fulfill the infrastructure requirements:

1. **Server and VM Orchestration:** Managing servers and virtual machines (e.g. VMware vCenter)
2. **Fabric management using UFM software:** Managing the connectivity between VMs, servers, storage, and networks

The combination of server and fabric management provides a set of powerful capabilities that establish a complete solution for large-scale virtualized data centers, and create one enormous and unified system that is easy to manage—providing better performance, utilization, isolation, and superior service levels.

A test-case scenario that combines some of the steps above involves process migration in a VM container over the fabric. In this advanced and common scenario, a VM must move across the fabric onto another compute resource, while having to maintain all its fabric SLA properties. This calls for a "VM-aware fabric" that is capable of adapting itself to process migration, without the manual steps of reconfiguring separate network switches.

Open Architecture and Extensibility

A critical element in the success of data center consolidation and automation is having an open and extensible architecture. With this type of architecture, users can customize the solution according to their specific needs and requirements and expand management functionality to support additional hardware elements, software elements, and use cases.

Mellanox's fabric solution models various physical, virtual, and application resources within the data center. It also preserves the configuration, policies, real-time status, and performance data of those resources, and it allows users and OEMs to extend a model through south- and north-bound interfaces:

- On the south-bound side, the fabric manager can be extended to support more physical or virtual switching, I/O, and server devices (to be monitored and managed under the same consistent fabric policy).
- On the north-bound side, the fabric manager provides a rich and extensible web services-based API or object-based command line interface (CLI) to monitor and provision the entire fabric. Additionally, via its FaaS™, it provides services and can obtain aggregate information on physical or virtual entities comprising the fabric, or be notified of a variety of events in real-time.

The open nature of the solution allows OEMs to create powerful, unified computing solutions, and provides end users and ISV partners with an easy way to extend the fabric management for their particular requirements or existing application environments.

Summary

As data centers grow and become denser, more virtualized, and automated, the load over the underlying fabric increases significantly along with the need for greater efficiency. As a result, clustering or scale-out technologies are now more widely used. InfiniBand fabrics were designed from day one as a scale-out data center fabric with a very lightweight switching infrastructure, the ability to run mesh topologies and multiple paths, and with fabric and I/O partitioning capabilities and central discovery and policy management. However, traditional Ethernet products do not scale well. As networks grow the costs grow exponentially, efficiency drops, and management complexity is increased.

Mellanox's scale-out fabric architecture based on Converged Enhanced Ethernet (CEE) delivers the following advantages:

- **Efficiency:** The scale-out design couples less complex and less expensive hardware elements with a powerful scale-out software stack, enabling lower costs, lower power consumption, and higher switching density than other aggregation switches currently available.
- **Linear scalability of latency, power and performance:** When using Mellanox's solutions, both cost and performance are linear, so whether the topology consists of a few hundred connected nodes or a few thousand connected nodes, it will have exactly the same price per port, exactly the same latency, and, the same total bandwidth per port.
- **Virtualized from the ground up:** Mellanox's Unified Fabric Manager (UFM) discovers all of the physical and virtual switches on the network, and dynamically configures them to deliver on the desired service or application requirements.
- **Focus on application performance and SLA:** The intelligent resource manager in UFM acts as an optimizer, and automatically produces a traffic policy for all switches on the network, guaranteeing the application behavior.
- **Service-oriented management:** Users define the desired fabric services and allow the fabric manager to enforce these services and provide high-level feedback.
- **Open and extensible architecture:** Mellanox's fabric solution models various physical, virtual, and application resources within the data center; preserves the configuration, policies, real-time status, and performance data of those resources; and allows users and OEMs to extend a model through south- and north-bound interfaces.

As Ethernet starts borrowing technologies from InfiniBand, InfiniBand is also beginning to leverage traditional Ethernet technologies such as FCoIB and some software stack elements. While InfiniBand is inherently faster and cheaper per Gb/s, Ethernet is often chosen by end users with less demanding performance requirements. With its CEE fabric architecture, Mellanox allows customers to choose the fabric that is right for them, without sacrificing scalability and performance.

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com