WHITE PAPER

Achieving Data Center Networking Efficiency

Breaking the Old Rules & Dispelling the Myths

Data centers have been designed for years with the same hierarchical and expensive network design. But as the modern data center evolves to a scale-out, dynamic and virtualized shared services platform, it's time to re-evaluate the old architecture and see if there is a more efficient way.

For years, data center networks have been designed according to the same basic concept:

- Single spanning-tree with three layers (top-of-rack/blade, access, and core switches) and significant oversubscription at each layer
- Applications deployed in self-contained racks, with internal Layer 2 switching connected to Layer 3 aggregation switches/routers

In this design, traffic management is done at the core rather than the edge. Each application silo scarcely crosses rack/edge boundaries, so the only traffic to manage is external. The core switch designs are high-powered and feature-rich, and use a large amount of packet buffering to absorb network spikes and implement traffic shaping. When congestion occurs, packets are dropped to slow down the transmitters. The delay and jitter created by the buffering does not significantly impact application performance since the main purpose of the core network is external (slow, long haul traffic), and the inter-application traffic (IPC, storage) is confined to the rack.

Now, economical challenges are driving data centers to reach new levels of efficiency. IT departments are now measured against business goals, not just up-time. And so they have adopted a service-oriented approach, consolidating and commoditizing server, storage, and network resources overlaid with virtualization and dynamic workload management. This approach lets IT deliver services faster with less effort and fewer physical resources.

Many of the assumptions that drove the old legacy network design have changed dramatically due to recent technological, organizational, and economical changes:

- Virtualization, cloud, and scale-out applications now drive the need for larger Layer 2 domains that span across multiple switches.
- Delay and jitter-sensitive traffic now goes through the core, requiring lower latency and less oversubscription on core switches.
- Storage and IPC protocols (FCoE, RDMA) are bursty and sensitive to packet drops, requiring lossless fabric. When they share the same wire with standard protocols, L2 traffic-isolation mechanisms are needed.

- Multiple applications and tenants can share the same edge ports and wires. Workloads are virtualized and mobile, requiring end-to-end traffic SLA management across virtual, edge, and core switches—and not just in the L3 core.
- Merchant silicon has become denser, cheaper, and loaded with features allowing new vendors to create more innovative and economical solutions.

Heavy network oversubscription is no longer desired since there is now significantly more east-west (rack- to- rack) traffic due to scale-out applications. MapReduce, application and storage clustering, server migration, and fault-tolerance all contribute to this traffic. Adding to the challenge is the inherent oversubscription in the server, with each running multiple virtual machines and acting as a switch with over-subscribed uplinks.

Impact of Traditional Oversubscribed, Single Path Network Design

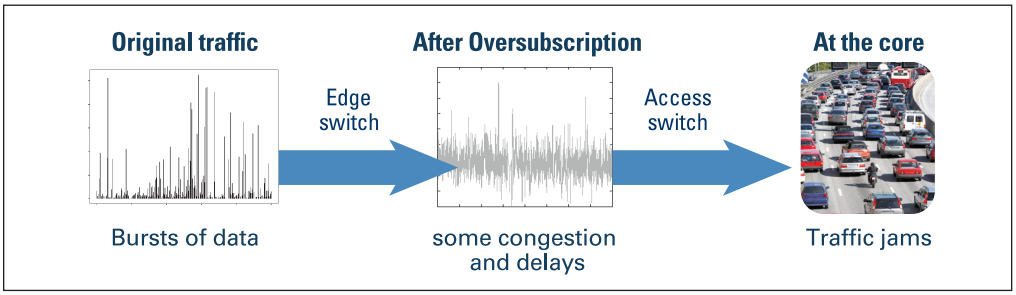

Figure 1. *Original traffic is bursty, and congestion increases after each switching layer—leading to saturated queues and long delays at the core switch. Core switches implement deep buffers and traffic management to prioritize traffic when it's already too late (queues are saturated). When buffers are exceeded, packets must be dropped, which reduces efficiency and increases jitter.*

The New Alternative: Scale-Out Fabrics for Scale-Out Data Centers

A fresh approach is needed to tackle these new data center challenges. Instead of building a hierarchical and oversubscribed network, the data center should be designed as a flat fabric that is dynamically partitioned to address application services. In the fabric, traffic is managed and buffered close to the source/edge, and the fabric core serves as a fast highway with multiple paths/lanes as well as low end-to-end latency. In a fabric, application workloads can span different locations and even migrate dynamically between physical servers while maintaining security and service levels.

This fabric architecture is built on a large number of simple, fast, low-power switches that are interconnected and orchestrated to deliver overall network services. This results in faster, more scalable, and cost-effective networks.

Scale-Out, Low-Latency Aggregation Switches

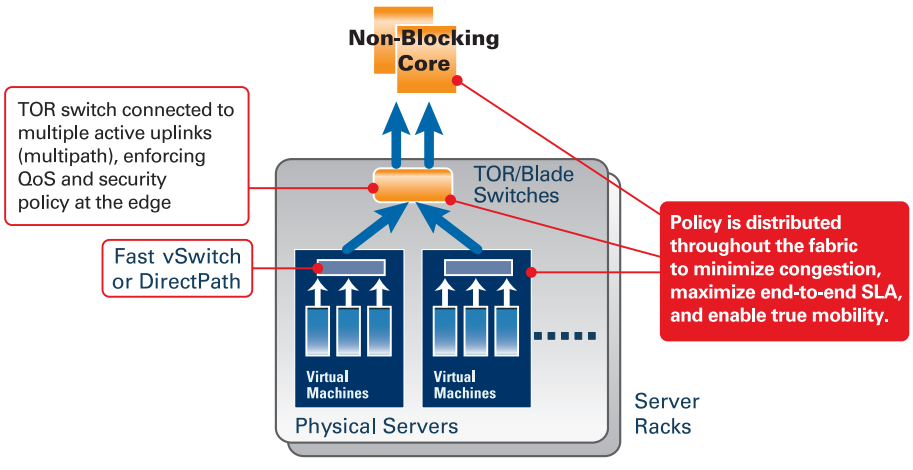

Figure 2. *The Scale-out Data Center Fabric approach shown here uses fast, non-blocking, converged Ethernet core switches with multiple paths between the edge and core switches, as well as distributed policy and network resource management starting at the edge of the fabric (at the virtual and/or TOR switches).*

Legacy Core Switch Evolution, Benefits, and Limitations

Ethernet standards have been enhanced recently to support fabric capabilities taken from InfiniBand and Fibre Channel, such as multi-pathing, class isolation, lossless behavior, dynamic congestion control, virtual end-points, and capability discovery. These capabilities allow new network designs that are more efficient and facilitate fabric convergence. They also fit better within virtualized data center environments. These new standards are gradually adopted with pre-standard solutions already delivered to the market.

Enterprise core switches usually function as a junction between multiple communication sources such as desktops, servers, and the Internet, as well as multiple users with different access priorities and rights. They are designed to address the many flows that cross this individual junction and focus on feature richness and throughput (not response time) while providing security, QoS, and monitoring.

Most vendors have created solutions using a similar architecture that consists of line interfaces that accept and classify the traffic. The traffic is then placed in queues for each destination and/or service (Virtual Output Queues), and traffic is piped to the destination port at the desired speed and priority. To improve utilization, packets may be broken into small cells that are load-balanced across the internal fabric. Later on, the packets are reassembled and sent to the destination.

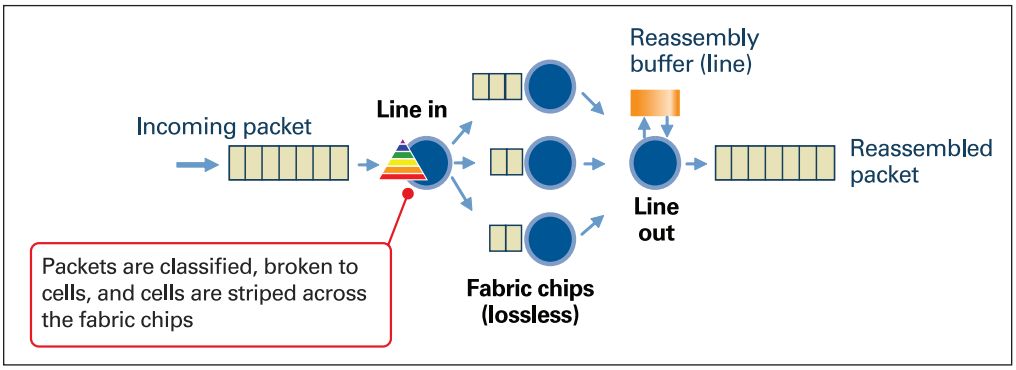

Figure 3. *Legacy core switch internal architecture*

This architecture has a number of key advantages. It is able to manage many flows going through the same junction and improve the switch utilization in oversubscribed environments, and it provides sufficient buffering to overcome unpredictable WAN performance.

However, there are also some drawbacks:

- The improved utilization comes at the expense of high delays/latency due to extensive queuing, packet fragmentation and reassembly, and store-and-forward architecture. With many flows traversing the switch, and as the switch becomes more loaded, the delays increase.
- Traffic management is handled at the congestion point (after the network is already jammed) not at the source.
- More components and a complex architecture lead to expensive, high-powered, and lower MTBF switch designs.
- This solution assumes there is one central junction that all the data flows through. This is not the case in large, scale-out data centers that contain east-west traffic and/or when using multi-path.
- Designing with a proprietary internal fabric limits the product and component options and locks in customers.

In legacy networks, there is only one core/root switch, which can easily become a bottleneck. That's also why there is a lot of emphasis on how to prioritize and schedule traffic in the core since important traffic should not be jammed behind less critical data. However, when building larger configurations that have a first layer of edge switches, the edge switches are typically not synchronized with the core policy—or they implement their own load-balancing/trunking—which undermines the efficiency of the core.

Achieving New Levels of Efficiency with Scale-Out Fabrics

In new scale-out fabric solutions, there are often multiple cores handling the load. When there are two or more such active-active core switches (rather than an active and a standby core switch as is common in legacy systems), the traffic divides between the cores, which lowers the traffic to less than 50% load per core. This leads to far fewer delays and better throughput than a 100% utilized single core, where any sudden burst causes a delay. The solution cost is often the same or even less due to the simpler design of the fabric core switches. Even if the traffic is not uniformly distributed between all cores due to hash inefficiency, it is still far better than having a saturated core switch.

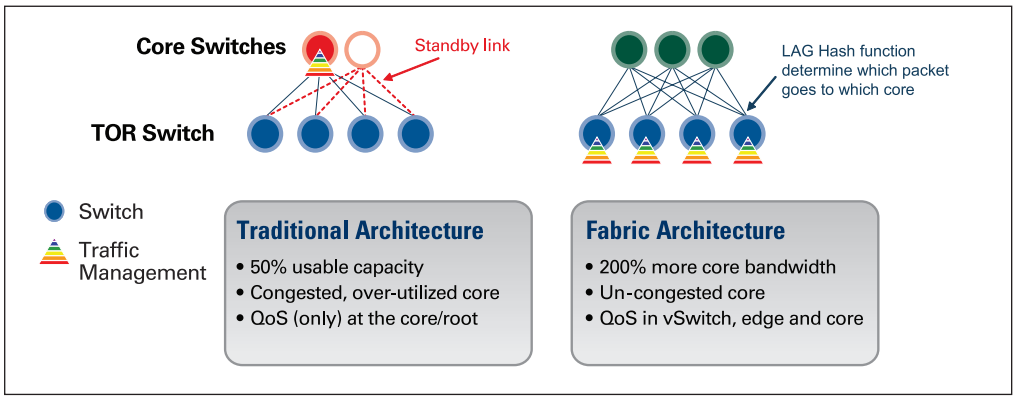

Figure 4. *Comparison of traditional data center architectures vs. a scale-out fabric architecture*

This fabric design imitates the design of a legacy core switch by scaling it out over multiple boxes. In this scheme, the edge or virtual switches function as line cards that classify and manage the traffic. The core switches function as the fabric cards that focus on low latency and lossless-traffic exchange between line cards. To orchestrate all the fabric switches, there is a fabric manager or operating system, which acts as the scale-out version of a supervisor module in legacy core switches. Fabrics eliminate the need for deep buffers at the core as data traverses the core without stopping and without being dropped. If data needs to be throttled down to eliminate fabric congestion, fabrics use dynamic and fabric-wide congestion-control mechanisms that keep traffic out of the center while buffering the traffic at its injection point (such as server-host memory). Just as in legacy core switches, traffic is buffered in the line card and not in the fabric card.

As an analogy, think of a busy road with many traffic lights. The road is fully utilized and prone to traffic jams. Wouldn't it be better to have a highway with two or more such roads or lanes that the traffic can be directed to? Such a highway, by definition, should not be over-utilized since the effectiveness of a highway is measured by how fast cars can travel the length of the highway (delay), and not how many cars can be squeezed onto it (utilization). If the cars on a highway are only few feet apart (i.e. the highway is 80% utilized), it is likely that drivers will experience delays.

The fabric design takes the highway approach:

- Use many simpler switching elements to reduce costs, power, and space
- Eliminate congestion with fast-forwarding and multiple paths/lanes
- Manage and buffer the traffic at the entrance (eliminate "traffic lights" inside)
- Minimize packet drops

Using this method, the fabric cores are not saturated, and delays are not accumulated throughout the fabric. The fabric's bi-sectional bandwidth (overall bandwidth) scales linearly as cores are added, allowing greater scalability and performance at lower costs.

Managing Congestion Efficiently In Real-time

The traditional way to manage congestion is for an oversubscribed network junction to drop packets. Once the source detects the drop, it retransmits the data and automatically slows down. To reduce drops due to traffic spikes, core switches use deeper buffers. Congestion accumulates, and if the congestion does not subside, packets are dropped after long delays. The traditional TCP protocol implements a slow-start mechanism that gradually increases performance to minimize drops. However, when an application

wants to send a large burst of data, it takes a much longer time compared to fabrics with hardware flow/congestion control (like FC, InfiniBand, or Converged Ethernet).

The benchmark shown in Figure 5 demonstrates how more buffers increase the network's ability to absorb oversubscription. At the same time, the legacy lossless Ethernet mechanism behaves just like a switch with deep buffers. Packets are not dropped, which eliminates the traffic collapse problem.

With Enhanced Ethernet standards, several new hardware mechanisms were introduced to provide better hardware flow-control and congestion-management to reduce congestion, eliminate collapse, and allow fast bursts without slow starts.

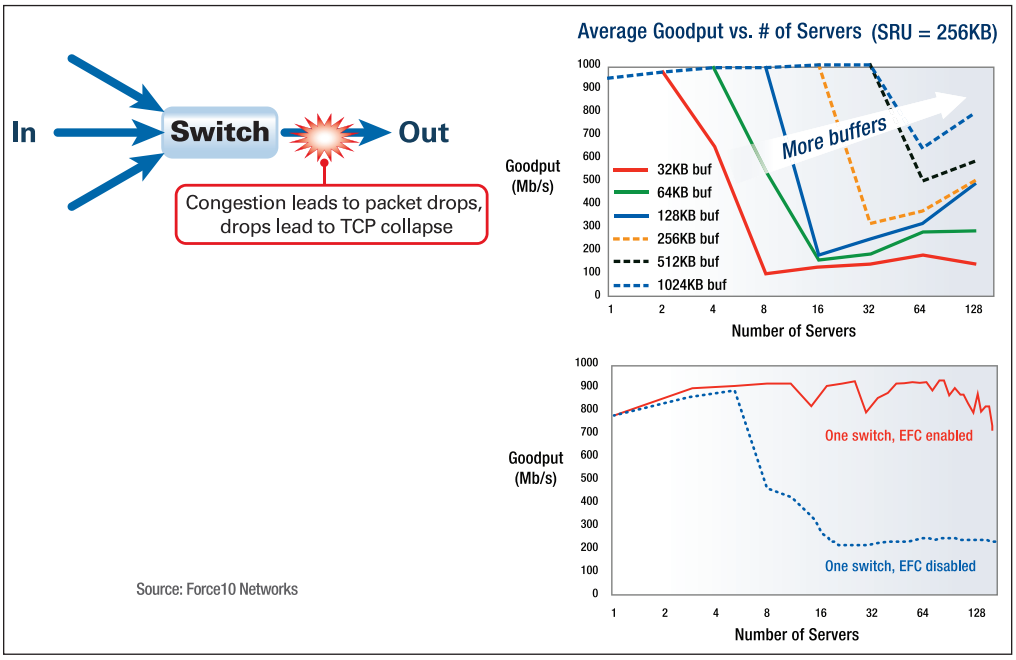

Figure 5. *Comparing switch behavior with different buffer depths and with lossless flow-control*

The following diagram shows a common head-of-line blocking scenario.

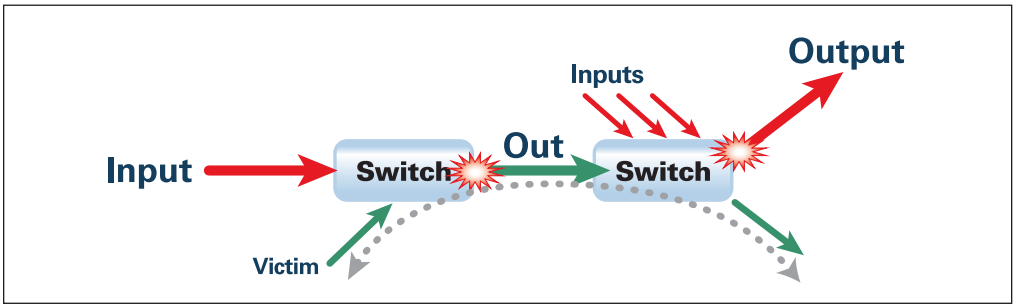

Figure 6.

In this scenario, output is oversubscribed with traffic from the four red sources. Green traffic is delayed since it is waiting behind packets leading to the congested output.

InfiniBand and Enhanced Ethernet fabrics incorporate congestion-control mechanisms (802.1 Qau/QCN) to address the congestion by throttling the sources of congestion and allowing victim traffic to pass. The following diagrams demonstrate the network behavior using the Mellanox Vantage™ 8500, an Enhanced Ethernet switch:

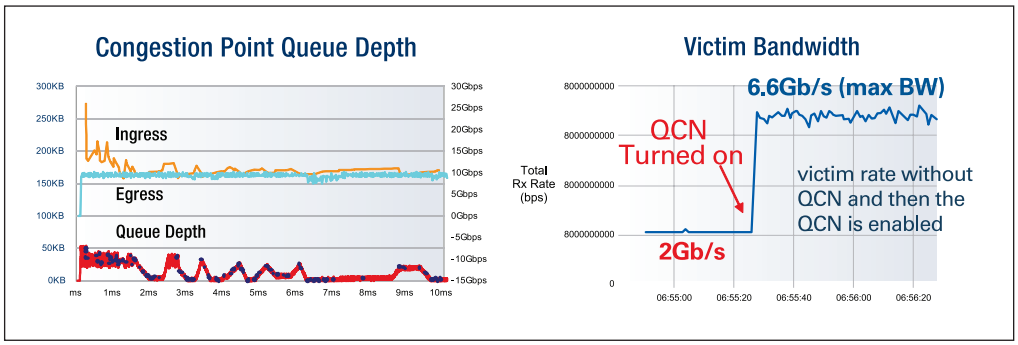

Figure 7. *Performance improvements with dynamic congestion control*

As the left diagram shows, the QCN mechanism reduces the queue depth at the congestion point. This ensures victim traffic is not stuck behind traffic going to an oversubscribed destination. The right diagram demonstrates how enabling the QCN causes victim traffic to jump to its maximum: 6.6Gb/s, with ~3Gb/s still flowing to the oversubscribed output.

Maximum performance is achieved when the queue depth is short. This is counter-intuitive as people have been trained to think that deeper buffers lead to better performance. In reality, the best way to improve application performance is to make sure the core is not oversubscribed too much using multi-path architectures. It's also important to control the traffic closer to the edge of the fabric while using host memory as a buffer along with dynamic, hardware-based, congestion-control mechanisms.

Summary

Legacy network architectures trained people to think that the right approach was to create longer lines (deep buffers). However, if we look at lessons learned in other fields, it is better to proceed taking the following approach:

- Shorten the process time to reduce switch-hop latency
- Add more lanes using multipath and reduced oversubscription
- Tell traffic when to approach the line (and when not to) with dynamic congestion notifications and management, as well as QoS at the edge
- Address the bottlenecks at all stations by managing traffic fabric-wide
- Do not focus on a single (core) device

Networks are built to serve applications and services, providing a mechanism to move data among applications, users, and storage in a fast, secure and reliable manner. When building a data center network, it's important to evaluate which solution is the most efficient and cost effective, and which solution meets the desired application performance, security, and reliability objectives. Solutions should not be evaluated by how well they fit the OLD design, but rather how they fit into the new model of a shared, scale-out data center.

This new fabric architecture built using Enhanced Ethernet standards is well-suited for scale-out and virtualized data center environments. It allows efficient horizontal scaling and faster performance, as well as guaranteed end-to-end network service levels and isolation. It also has inherent high-availability and resiliency, with multiple possible paths to address high workloads or network failures, and it is simpler to manage through a centralized and service-oriented fabric OS.

Mellanox TECHNOLOGIES

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