Next Generation Terabit Switch/Routers: Transforming Network Architectures

Introduction

Ethernet networks are being called upon to deliver unprecedented volumes of increasingly diverse traffic. Real-time application traffic (including VoIP, network storage, and cluster/Grid interconnect) is highly sensitive to latency and jitter. Enterprise data applications, such as ERP and CRM, often require high priority to protect them from packet loss. Other applications, such as network video and data backups, are not very sensitive to latency or packet loss, but are bandwidth intensive.

As enterprises and service providers upgrade the bandwidth of their networks for growing volumes of aggregated traffic, they also need the ability to support a number of application service classes with predictable and/or guaranteed levels of service. Within each class, the level of service needs to be defined in terms of availability, latency, jitter, packet loss, and bandwidth. In order to deliver predictable service levels and meet required Service Level Agreements (SLAs), network designers need to employ well-balanced network designs characterized by:

- **Ample Bandwidth**: Links carrying highly aggregated traffic should be provisioned to deal with traffic spikes and escalating volumes, using 10 Gigabit Ethernet (10 GbE) as appropriate.
- **QoS and Traffic Control Mechanisms**: Bandwidth by itself will not support high expectations for service quality. On the other hand, combining QoS with traffic control/admittance does support guaranteed service levels.
- **High Availability**: Service interruptions can have a serious impact on customer satisfaction and revenues. Availability of services will increasingly depend on highly resilient networks and devices that can continue operation despite component failures, software errors or upgrades, or malicious attempts to disrupt service.

- **Simplicity and Low TCO**: To offset the added complexity implicit in QoS, traffic control and high availability features, networks should be as simple, robust, and easy-to-manage as possible to control/minimize total cost of ownership.
- **Upgraded Switches and Routers**: Legacy devices need to be upgraded to meet functionality requirements and to eliminate any internal performance bottlenecks. Some device bottlenecks can prevent full utilization of high bandwidth links while others are incompatible with desired service levels.

As a result of these priorities, there is a market need for next generation Ethernet switch/routers with multi-Terabit per second capacity that can allow scalable, full line-rate traffic aggregation while also providing guaranteed low latency, assured data delivery, and the device/network resiliency needed for catastrophic failure prevention.

Next generation devices that can both switch and route at full line rate, irrespective of the L2/L3 services configured, can considerably reduce network complexity and cost by allowing the consolidation of the legacy access and distribution layers into a single traffic aggregation layer. Next generation switch/routers can help to transform network architectures in two ways:

1. By supporting seamless migration from legacy 3-tier to simpler 2-tier network design models
2. By providing the capability to fully exploit 10 Gigabit Ethernet, while also ensuring a graceful migration path to the next generation of Ethernet at 100 Gbps

This paper describes the features that next generation switch/routers must possess if they are to successfully support the evolution of Ethernet as the networking technology of choice for highly scalable, highly reliable, service-aware LANs and MANs. Included in the discussion is a description of some of the key features that make the Force10 E-Series the yardstick against which all other next generation Ethernet switch/routers should be measured.
A Brief History of Ethernet Switch/Routers

Each new generation of Ethernet technology has had significant implications for the internal architecture of LAN forwarding devices, as well as for the architectures of design models used to build Ethernet networks. The following historical overview traces the evolution of the Ethernet switch/router from previous generations of LAN forwarding devices.

Phase I: In the late 1980s and early 1990s most LANs were based on shared Ethernet hubs and general purpose, software-based multiprotocol routers fitted with Ethernet interfaces. Software-based 2-port bridges were used to partition large repeated segments and extend the reach of Ethernet. The era of switched internetworking began in earnest with the emergence of Fast Ethernet and then Full Duplex Ethernet in the mid 1990s. Toward the middle of the decade, advances in integrated circuit technology in the form of denser and faster ASICs and FPGAs helped set the stage for Fast Ethernet while also greatly improving the cost effectiveness of a niche product category, the Ethernet switch introduced by Kalpana in 1990. By 1995, 10/100 Mbps Ethernet switches were available from a large number of vendors and began to gradually replace hubs at the edge of the LAN. As wire-speed Layer 2 switches with Fast Ethernet uplinks became widely deployed, they overwhelmed the forwarding capacity of the core routers of the day. Routers had become the bottleneck in the network because they were essentially special-purpose PCs that relied on a central CPU for forwarding table lookups, maintaining routing tables, and performing management functions. Although some routers were enhanced with ASIC hardware assistance, router performance remained bounded by the basic central CPU and contention bus architecture.

Phase II: By the late 1990s, continued advances in circuit technology and ASICs ushered in two new developments: Gigabit Ethernet and the Layer 3 switch. The combination of these technologies spawned a large number of startups, including Rapid City, Yago Systems, Extreme Networks and Foundry Networks. The Layer 3 switch featured the ability to forward both Layer 2 and Layer 3 traffic at wire speed over Fast Ethernet and Gigabit Ethernet ports. This first generation of switch/router typically relied on a flow-cache architecture to distribute the majority of forwarding decisions to ASICs on intelligent line cards. For Layer 3 forwarding, each line card maintains a cache of recently-accessed routes. In the event of a cache miss, the line card forwards the packet over the "slow path" by first learning the route from the central CPU which maintains the full Forwarding Information Base (FIB). The flow cache switch/router architecture works well in smaller and relatively static networks. However, in large networks, where there are a large number of flows and quite unpredictable flow patterns, cache misses can overwhelm the capacity of the central CPU, causing excessive packet latency over the slow path.

Phase III: As VoIP and other real-time traffic have become more prevalent, traffic flows have become continually less predictable and the tolerance for latency/jitter or catastrophic failure of the network has been greatly reduced. At the same time, the emergence of 10 Gigabit Ethernet meant that switch/routers would be expected to handle approximately ten times as many flows as a Gigabit Ethernet switch/router. Accordingly, the early 2000s saw the emergence of the next generation of Ethernet switch/router that uses a fully distributed architecture to completely eliminate the slow path bottleneck. In the fully distributed switch/router architecture, the central route processor CPU is dedicated to maintaining the FIB and distributing a copy of the full FIB table to each of the line cards. Therefore, in this new generation of Ethernet switch/router, all Layer 2 and Layer 3 packet processing and forwarding decisions are performed by the line card ASIC hardware, allowing all flows to be forwarded with low, consistent latency.

As 10 Gigabit Ethernet becomes widely adopted in the core of enterprise and service provider networks, highly reliable Ethernet switch/routers with fully distributed architectures will continue to evolve to accommodate changes in networked applications and to pave the way for the next generation of Ethernet at 100 Gbps.
Requirements for Next Generation Ethernet Switch/Routers

As a pioneer in the development of line-rate 10 GbE switch/routers that adhere to a fully distributed architecture (the E-Series), Force10 Networks has over six years of experience dedicated entirely to the design and deployment of these next generation switch/routers in enterprise and service provider networks. Based on this experience, Force10 Networks will continue to focus its development efforts on E-Series switch/routers that meet the following general market requirements:

Scalable, Robust Control Plane: The control plane of the switch/router is comprised of the software processes used to build and maintain FIBs, handle L2/L3 control functions, and perform the necessary network management functions. A scalable control plane can support large FIBs and has the performance to prevent contention among control functions from having a negative impact on the system’s response time to topology changes or other changing conditions in the network. The robustness of the control plane design is critical because research has shown that a high percentage of system failures and service interruptions are directly traceable to control plane deficiencies. The overall goal in control plane design is to meet or exceed the scalability/stability of core Internet routers at price points that are comparable to legacy Layer 3 switches.

Highly Scalable Forwarding Performance and Port Density: The scalability of a switch/router is determined by its ability to support application/service requirements in a rapidly expanding network environment. All vendors are driving towards higher density system levels because density and scalability help to drive down the cost per port and maximize the deployable lifetime of the system. In particular, a highly scalable switch/router is designed not only to fully exploit current technology but also to accommodate future steps in technology development. The latter aspect of scalability allows enhancement of the switch/router by a staged program of subsystem upgrades (e.g., line cards, switch fabric modules, and route processor modules) while protecting the prior investment in the chassis, power, cooling, and backplane.

Overall system scalability has a number of dimensions including the number of peers/neighbors supported, the number of routing and forwarding table entries, the number of MAC addresses and VLANs supported, the depth of VLAN stacking supported, the number of standard and extended ACLs that can be configured, the amount of packet buffering allocated to each output interface, the aggregate bandwidth supported by the switch fabric, and the packet forwarding performance/capacity provided by each line card.

Scalable forwarding performance implies a distributed non-blocking switch/router architecture with line card implementations that maintain full line-rate performance and predictable latency irrespective of the traffic load, QoS features, ACLs, traffic monitoring, or other services that have been configured for traffic control or security. Scalable port density is achieved through the combination of line cards with scalable forwarding performance and a scalable switch fabric which has ample capacity to deliver non-blocking full line-rate performance for large numbers of high density line cards.

With integrated hardware-based packet processing and forwarding on all ports, full line-rate performance can be maintained even as the market evolves to require additional protocol formats and encapsulations (e.g., tunneling). With fully integrated bridging and routing functionality selectable on every port, switch/routers have the flexibility to be deployed in the core and aggregation layers of any multi-tiered network architecture and can readily support transitions from one design model to another by simple device re-configuration.

Device Resiliency: As switch/routers continue to scale in terms of link speeds, forwarding performance and port density, device resiliency is becoming an indispensable system attribute. For example, high degrees of traffic aggregation mean that even short periods of interrupted operation can affect a large number of traffic flows and user/subscribers, potentially violating numerous SLAs.

A high availability network design always involves a degree of redundancy at the network level, but a network designed for automatic failover among fully
redundant links and nodes is generally overly complex and expensive. Furthermore, even with careful network design and configuration, failover times of several seconds or more are quite common. Therefore, the best approach to designing networks for very high levels of availability is to start with highly resilient network devices.

Resiliency for a switch/router device implies an ability to continue network operations in the face of the full gamut of fault conditions, including the failure of hardware components, software faults and restarts, link failures, protocol restarts, and attempts by intruders to disrupt normal traffic flow.

The remainder of this paper focuses on the Force10 E-Series architecture and the specific features next generation switch/routers should have to meet market requirements for scalable/robust control planes, scalable forwarding performance/port density, and resiliency. For each of these feature categories, the paper will include a description of how the Force10 E-Series has been architected and designed from the inception to meet these requirements.

**E-Series Control Plane**

The Force10 E-Series family of switch/routers shares a common fully distributed architecture (shown in Figure 1) based on a virtual output queue (VOQ) cross bar matrix switch fabric coupled with line cards that perform integrated switching and routing (for L2, IPv4, IPv6) supported entirely in wire-speed ASIC hardware. Control plane functions are performed on the Route Processor Modules (RPMs) and on separate control processors on the line cards.

For a complete discussion of the E-Series architecture and features please refer to:

As noted earlier, a scalable control plane can support large FIBs and has the performance to prevent contention among control functions from having a negative impact on the system’s response time to topology changes or other changing conditions in the network.

The E-Series features a high-performance multiprocessor control plane that supports line-rate L2 switching and line-rate L3 routing with full L2/L3 protocol and features enabled. Designed to meet the needs of Internet-scale networks, the E-Series control plane supports millions of routing table entries, with up to 512K forwarding table entries, and tens of thousands of ACLs on every line card.

The modularity of the Force10 Operating System (FTOS) complements the distributed multiprocessor architecture comprised of three control processors on the Route Processor Modules and an additional control processor on each line card. One RPM CPU is dedicated to Layer 3 processes, with the second CPU dedicated to Layer 2 processes, and the third dedicated to local control and management functions. Each of these CPUs runs an independent software image, maximizing performance and minimizing the possibility that a fault in any one process could lead to catastrophic system failure. For example, a fault in a management process would not interfere with L2/L3 forwarding table updates.

The line card CPUs currently perform local control functions including sFlow sample aggregation and reporting. In the future, the modularity of the FTOS operating system will allow some central control processes to be distributed across the line card CPUs. This distributed multiprocessing will add even greater scalability to the processing capacity of the control plane.

**E-Series Scalable Forwarding Performance and Port Density**

This section of the document provides an overview of the E-Series focused on the features that help to ensure scalable forwarding performance and port density.

**Control Plane Implementation**

The E-Series control plane, described in the previous section, has the scalable performance to ensure that the FIBs provided to the line cards accurately reflect the current topology of the network. This ensures that the line card forwarding performance is not wasted due to stale copies of the FIB.

**Backplane**

The backplane of the switch/router provides the physical interconnect between the line cards, the route processor cards, and the switch fabric as shown in Figure 1. A scalable backplane has ample reserve transmission capacity to accommodate higher performance switch fabric implementations as well as the future introduction line cards with higher port density.

The Force10 E-Series features a multi-channel passive copper backplane with 5 Tbps of transmission capacity. Five Tbps is enough capacity to support a chassis switch with 14 line card slots with over 330 Gbps per slot. This means that the current E1200 chassis has more than
enough backplane capacity to be able to support 14 ports of 100 Gbps Ethernet (which will require 200 Gbps per line card slot, assuming 1 port per line card). It should be noted that in addition to 100 Gigabit Ethernet line cards, this future system upgrade would also require replacement of the switch fabric cards.

Compared to optical backplanes or active copper backplanes, the patented Force10 TeraScale passive backplane eliminates costly electrical-optical-electrical conversions and has no single points of failure. The result is a simple system design with unparalleled reliability and minimum cost.

**Switch Fabric**

The switch fabric of the switch/router utilizes the underlying backplane to switch packet traffic among the line cards and other modules of the system. A scalable switch fabric has ample packet switching capacity to accommodate current and future generations of higher density line cards.

The Force10 E-Series E1200 features a Modular Cross-Bar Switch Fabric with 1.68 Tbps of non-blocking packet switching capacity. The fabric is implemented as nine load-sharing Switch Fabric Modules (SFMs). This multi-module architecture was implemented in order to reduce sparing/manufacturing costs and to maximize the resiliency of the system. In an E1200 chassis fully loaded with 48 port GbE TeraScale line cards, the effect of an SFM failure is an 12% loss in switching capacity (8:1 redundancy). In a partially loaded chassis (e.g., with 40 or fewer GbE ports per line card or four 10 GbE ports per line card), the loss of a single SFM does not result in any reduction in switching capacity (8+1 redundancy).

**Data Plane Hardware Implementation**

The data plane implementation determines how much of the potential capacity of the switch fabric can be achieved while forwarding user traffic through the switch/router. A scalable data plane supports full line-rate forwarding and low, consistent latency for both Layer 2 or Layer 3 traffic irrespective of the L2/L3 services that are configured.

The E-Series architecture completely separates the data plane from the control plane. In the data plane, ASICs are used exclusively to implement all packet processing functions in hardware. Ternary content addressable memories (TCAMs) on the line cards perform packet classification while other ASICs perform the buffer and traffic management functions including packet-rate policing, two-rate three color queuing with dual token buckets, queue scheduling, rate limiting, ACLs, statistical sampling of traffic with sFlow, and congestion avoidance. Even with the complete range of QoS, traffic management and monitoring services enabled, full line-rate performance is maintained.

According to Tolly Group tests, the Force10 E1200 achieves 100 percent line-rate, zero-loss Layer 2 and Layer 3 throughput across 672 Gigabit Ethernet and 56 10 Gigabit Ethernet ports for all packet sizes. These tests verify the E1200 forwarding performance of 1.34 Tbps or 1 billion pps. The Tolly Group also verified that line-rate performance is maintained even when over 1 million ACLs are configured on the E1200.

Since the Tolly Group testing concluded, Force10 has transformed the E-Series into the industry’s first Terabit switch/router, delivering 1260 GbE or 224 10 GbE ports per chassis. With the Force10 TeraScale technology, the E-Series can now deliver up to 90 GbE ports or sixteen 10 GbE ports per line card slot and process 1 billion packets per second.

**E-Series Resiliency**

The E-Series switch/routers have been designed from the outset to achieve the highest levels of resiliency by eliminating as many sources of potential catastrophic system failure as possible. After a comprehensive analysis of the causes of system crashes and of all potential sources of catastrophic system failure, the Force10 design engineers adopted the multi-layered resiliency architecture shown in Figure 2.
Across the layers of the model, the Force10 Resiliency Architecture leverages redundancy to eliminate single points of failure and incorporates robustness features that ensure that the system can gracefully continue operations in the face of a wide variety of traffic anomalies, fault conditions, or operational changes to the network.

**Hardware Resiliency**

Resiliency at the hardware level of the E-Series switch/routers is focused on eliminating single points of failure within the hardware platform itself. As shown in Figure 3, the E1200 switch/router supports a high degree of subsystem redundancy, including:

- Cooling (fans and fan modules)
- Power supplies (1+1)
- RPMs (1+1), optional
- Switch Fabric Modules (8:1)

The line cards and all of the redundant subsystems support Online Insertion/Removal (OIR) and are fully monitored for fault conditions or out-of-range environmental parameters. Hardware reliability is further enhanced by the passive nature of the copper backplane, since there are no active components that could become potential single points of failure.

Within the control plane itself, hardware resiliency features include:

- Distributed multiprocessor architecture of the RPMs
- Protected memory among CPUs to prevent process corruption
- Error correcting code (ECC) or parity protected memory to ensure data integrity
- Out-of-band switched Fast Ethernet control plane communications
- Hitless failovers between redundant RPMs. The independence of the data plane allows each line card to continue forwarding using its own copy of the FIB during an RPM failover or software upgrade. As a result, no traffic is lost and all current traffic flows are protected.
- Control plane ACLs in addition to hardware-based prioritization and rate-limiting of control traffic provide network stability and protect the control plane from DDOS attacks

![Figure 3. TeraScale E-Series E1200 hardware redundancy](image)
**Software Resiliency**

As noted earlier, much of the system resiliency at the software layer is due to the modularity of FTOS operating system software. With software modularity all the major functions are separated into separate processes, with each process supported by its own protected segment of memory. This degree of modularity prevents a fault in one process from affecting or corrupting other processes. As FTOS continues to evolve, its modularity will be further enhanced to allow individual processes to be started or stopped independently and to allow individual software components to be upgraded without stopping or disrupting the remaining processes.

Additional planned enhancements of FTOS include support for preemptive process scheduling (to prevent a single process from monopolizing the CPU) and advanced Inter-process Communications (IPC) mechanisms for control plane communications. IPC will provide the basis for leveraging the distributed processing architecture (encompassing the RPM CPUs and the line card CPUs) to enhance control plane resiliency and scalability.

**Link Resiliency**

Link Resiliency features allow traffic on a failing link to be re-directed to a parallel redundant link or path. Failover or recovery time needs to be minimized to prevent or limit the disruption of application traffic.

At the IP layer, the E-Series supports Equal Cost Multi-Path routing (ECMP). With ECMP, Layer 3 traffic can failover to a redundant link or path in less than a second based on detection of link/path errors. At the Ethernet layer, the E-Series supports IEEE 802.3ad link aggregation (LAG). LAG can achieve failovers of less than a second based on physical layer detection/signaling of failures.

Fault detection/signaling at the physical layer for ECMP and LAG are provided by E-Series support for auto-negotiation for Gigabit Ethernet and Link Fault Signaling (LFS) for 10 Gigabit Ethernet. On 10 GbE WAN PHY links, the E-Series supports Alarm Indication Signal and Remote Defect Indication for line and path faults (AIS-L, AIS-P, RDI-L, and RDI-P).

With either ECMP or LAG, up to 16 parallel paths/links can share the traffic load and provide mutual redundancy. In the E-Series implementations, ECMP and LAG use the same load balancing algorithm to distribute the traffic across redundant links. This Force10 algorithm is based on a unique IP 5-tuple hash (IP source and destination addresses, IP protocol, Layer 4 source and destination ports) to distribute traffic flows across the links. Since this hash algorithm does not use MAC addresses, LAG load balancing can accommodate either Layer 2 or Layer 3 traffic. As an alternative, Force10 also supports a LAG load balancing algorithm that is based on MAC addresses, which may be used for non-IP traffic in Layer 2 segments of the network.

**Protocol Resiliency**

When network links fail or switch/routers fail or require restarts, the result often involves convergence to a new network topology with temporary sub-optimal routing of traffic and possible packet loss. Protocol resiliency includes a number of features to minimize re-convergence times and prevent local failures from affecting the overall network.

The E-Series achieves protocol resiliency with the following mechanisms:

- Rapid STP (RSTP), in conjunction with Multiple STP (MSTP) or Per VLAN STP (PVST+), allow Layer 2 tiers of the switched network to be designed for fast (sub-second) STP convergence and active-active load sharing across redundant Layer 2 links and switches.
- OSPF and BGP graceful restart mechanisms allow the data plane to continue forwarding packets while the router’s control plane software is reloaded or restarted. The main goal of these restart protocols is to allow a router to restart gracefully without causing a routing flap or transient loops across the network.
- Virtual Router Redundancy Protocol (VRRP) eliminates statically defined default gateways as single points of failure in the network. Without VRRP, traffic could be disrupted for a considerable time (seconds or even minutes) while IP configurations are renewed or the default router is restored.

**Manageability and Serviceability**

Resiliency at the manageability/serviceability layer implies that system upgrades, re-configuration, fault correction, and component repair can all be accomplished without taking the switch/router out of service. Other manageability features that help to reduce the time to diagnose and repair faults also have an impact on resiliency.
The E-Series supports the following manageability/serviceability resiliency features:

- Online Insertion/Removal (OIR) of all critical components
- Hitless software upgrades
- Hitless protocol restarts
- Persistent configuration and pre-configuration of line card slots
- Control protocol tracing
- Proactive monitoring via the Force10 Service Agent, SNMP, and integrated hardware monitoring
- sFlow traffic monitoring

Resiliency of E-Series switch/ routers at the management level is further ensured by mechanisms, including RADIUS and TACACS+ that provide authentication, authorization and accounting (AAA) services to control remote access to the device by network managers and administrators. In addition, Secure Shell (SSH) and Secure Copy (SCP) are supported to provide strong encryption, robust authentication, and data integrity for remote system access or file transfers to/from the switch/router.

**Conclusion**

Next generation Ethernet switch/routers will continue to evolve to enable widespread adoption of 10 GbE, followed by the next generation of Ethernet at 100 Gbps. As this evolution takes place, scalable/robust control planes, scalable forwarding performance/density, and device/network resiliency will be the most critical attributes for switch/ routers designed for the core of LANs/MANs or for the data center.

The high degree of synergy among these product attributes will give network designers the freedom to adopt LAN design models that have significantly reduced complexity and lower overall life cycle cost. For example, the L2 access and L2/L3 distribution layers of legacy LAN designs can be collapsed into a single L2/L3 aggregation layer using next generation Ethernet switch/routers exemplified by the E-Series, as shown in Figure 4. Some of the total cost of ownership (TCO) benefits of transforming network architectures by exploiting the density and resiliency of E-Series switch/routers are documented in a white paper authored by the Meta Group.

In conclusion, while there are some other models of Ethernet switch/routers and Internet routers that employ a fully distributed architecture, the E-Series is the only 10 Gigabit Ethernet switch/router that can offer the following combination of leadership characteristics:

- Uniquely scalable and robust multi-processor distributed control plane
- Highest forwarding performance plus highest GbE and 10 GbE port density
- Future-proof migration path to 100 Gbps Ethernet with a 5 Tbps passive backplane
- Highest level of device and network resiliency
- Lowest total cost of ownership per Terabit of switching capacity

![Figure 4. Network architecture transformation enabled by next generation switch/routers](image)