Internet Service Provider Networks: Simplifying POP Architectures

Introduction: The Growth of Internet Traffic
While some previous estimates of the Internet’s growth rate were exaggerated, market research firms now generally agree that Internet bandwidth consumption is growing at about 100% per year. Based on its research, IDC expects this growth rate to persist until 2007, with aggregate traffic growing from an estimated 180 petabits/day in 2002 to 5,175 petabits/day by the end of 2007.

Sources of traffic growth will include new subscribers and increased Internet usage by existing enterprise subscribers. However, most of the traffic growth is expected to come from wider adoption of broadband access services by existing subscribers. Broadband access encourages increased usage of the Internet and enables new user services such as IP telephony, digital media distribution, etc. Along with the need for more bandwidth to support these new services, residential and business customers alike are expecting higher levels of network reliability and predictability at lower and lower price per bit levels. The ability to predict and guarantee service levels and maintain QoS will not only add to customer or user satisfaction but also open up new competitive advantages by leveraging the network as a strategic asset.

Favorable government policies coupled with competition among alternative technologies and service providers will eventually result in nearly universal adoption of broadband. For example, in Korea, where broadband has been a national priority, over 94% of Internet subscribers already have broadband connections. According to the International Telecommunications Union, this level of broadband adoption is approximately three years ahead of the global average.

As this rapid growth in access bandwidth occurs, ISPs will need to continue to expand the capacity of their networks while at the same time responding to a number of competitive pressures, including:

- controlling costs
- delivering higher reliability
- offering an attractive mix of broadband access technologies and speeds
- dynamically distinguishing between types of traffic
- offering new value-added service options
- enhancing customer service and customer satisfaction

Priorities for the ISP POP Network
In order to meet these challenges, ISPs are focusing on developing network architectures for the POP that are optimized for scalability, robustness as well as simplicity of operations, and manageability. In order to optimize along these dimensions, ISPs will be designing their networks based on switch/routers with the following general attributes:

- Ability to scale bandwidth gracefully from 10 Mbps, 100 Mbps, 1 Gbps, or 10 Gbps – all with granular 1 Mbps selectivity and with minimal disruption of existing infrastructure
- Support dynamic CIR-like services by rate limiting or shaping bandwidth based on traffic loading demands and then bill according to usage
- Flexible QoS and traffic policing to support simplified provisioning of tiered services
- The network should be able to support multiple upgrades through simple replacement of line cards and modules or deployment of additional devices belonging to the product families already in use
ISP Point of Presence

- Very high switching performance and port density with small footprint to conserve space in the POP and to minimize the number of devices that need to be managed
- Carrier grade hardware and software reliability, plus support for network layer redundancy features to minimize operational costs and to support SLAs guaranteeing performance and availability
- Robustness in standards-based features designed to reduce the number of incidents that require operator attention or have an impact on subscriber services
- Security and traffic control features to protect the network infrastructure and customer networks from disruption by accidental and malicious traffic
- Extensive traffic statistics to manage bandwidth consumption, plan for capacity expansion, and support SLAs

Typical Large POP

Over the last few years, the network architecture for the typical large POP has evolved as a three-tier network design, similar to the one shown conceptually in Figure 1. In the access tier, edge routers consolidate subscriber connections, possibly over a range of diverse access technologies (ATM, Frame Relay, Ethernet, DSL, etc.). Traffic from the edge routers is aggregated by Layer 2 switches in the middle tier and forwarded to core routers that provide connections to Internet Exchanges, the ISP backbone, NAPs, other ISPs, and hosting facilities. Smaller POPs may use a modification of this template in which the aggregation tier is eliminated by direct connections between access and core routers.

An important advantage of this tiered approach is that it facilitates expansion of capacity within any tier with minimal disruption of the other tiers. In addition, the three-tier architecture leverages the specialized platforms that have evolved to focus on specific functions:

- Edge Routing with moderate performance and a diversity of interfaces to support a range of access technologies
- Aggregation Switching with high performance, cost-effective Layer 2 Ethernet switching
- Core Routing with state-of-art performance, Internet scale routing, and optimization for connection to long haul optical networks

However, these advantages have come at the expense of increasing the complexity of the network by increasing the total number of devices that must be individually configured and managed. In addition, the possibility of congestion on each of the inter-tier links may require the configuration of traffic control policies at all three types of devices.
Force10 Solution: High Availability 3-Tier ISP POP

Figure 2 shows how an existing 3-tier POP could be upgraded for enhanced performance and robustness by installing Force10 E-Series switch/routers in the aggregation tier, replacing numerous older aggregation devices, such as ATM switches or earlier generation Layer 2 Ethernet switches. Upgrading the aggregation tier often removes many of the existing performance bottlenecks while preserving the basic architecture of the POP and leaving the existing core and access routers in place.

The POP configuration shown in Figure 2 eliminates single points of failure by using redundant inter-tier links between the devices. With the E-Series support for IEEE Rapid Spanning Tree Protocol (RSTP), traffic rapidly fails-over from primary to secondary paths in the event of link or device failure. With RSTP, failover periods can be as short as milliseconds to hundreds of milliseconds vs. as much as 30 seconds for the original IEEE 802.1d Spanning Tree Protocol (STP).

The primary benefits of implementing the aggregation tier of the POP with E-Series switch/routers lie in the flawless performance, unmatched scalability, and robustness of the platform. With its 1.68 Tbps switching fabric and industry-leading density of Gigabit Ethernet (GbE) and 10 Gigabit Ethernet (10 GbE) ports, the E-Series can deliver the required switching capacity with a smaller number of devices while offering ample port capacity for scalable inter-device links based on IEEE 802.3ad trunks using multiple GbE or even multiple 10 GbE links.

The next few sections of this document provide an overview of the E-Series product line and the features that make it an ideal choice for deployment as the aggregation switch/router in the POP.

During this discussion, it will become evident that the E-Series switch/router is really an Ethernet-optimized Internet router that also has a complete set of Layer 2 features. The final sections of the document show how the comprehensive Internet routing functionality of the E-Series may be exploited in both the core and access tiers of POPs to help increase their leverage of Ethernet technology. At the end of the document is a discussion how the E-Series could be used in a next generation POP based on a 2-tier architecture that minimizes complexity and cost by taking maximal advantage of Ethernet technology.
Layer 2 Network/Link Redundancy and Resiliency Features

Service assurance and bandwidth continuity require both the network as an ‘organism’ and the devices themselves to be reliable and scalable. In a typical Layer 2 network topology, reliability is achieved through a variety of standards-based protocols.

Link Aggregation (IEEE 802.3ad)
A Link Aggregation Group (LAG) based on the IEEE 802.3ad specification bundles multiple physical Ethernet links of the same speed into a higher bandwidth logical link. The E-Series supports LAGs consisting of up to 16 individual links per group or up to 256 LAGs per chassis. A Force10 hashing algorithm based on the 5-tuple (IP source address, IP destination address, protocol type, TCP or UDP source, and destination port numbers) ensures efficient load balancing of diverse Layer 2 or Layer 3 traffic without packet reordering.

A major benefit of the LAG is that it provides economical scaling of bandwidth for interswitch links within the aggregation tier and between the aggregation and other tiers of the network. For example, a LAG comprised of multiple GbE links offers a granular bandwidth expansion path between 1 GbE and 10 GbE. As 10 GbE links eventually become saturated, the 10 GbE LAG will allow bandwidth scaling until the next generation of Ethernet (100 GbE) is standardized and ready for deployment. LAGs also offer the benefit of greater resiliency, especially when the links in the group are distributed across multiple line cards. Links can be added or deleted from the LAG without disrupting traffic or rebooting the system.

Spanning Tree Protocol (STP) Enhancements
IEEE 802.1d STP is widely used in Layer 2 networks for failure recovery and loop avoidance. The value of standards-based technology enables interoperability among different equipment manufacturers and ensures seamless migration as new enhancements are made. However, STP was conservatively designed for large diameter networks of arbitrary topology and reacts slowly to failures even in relatively simple networks, taking tens of seconds to provide recovery from link and node failures. The E-Series supports IEEE 802.1w Rapid STP (RSTP) that provides fast convergence in case of link or root failure. RSTP allows E-Series platforms to maintain knowledge of multiple paths to the root. When the primary point-to-point link fails, the system fails-over to the secondary link in a matter of milliseconds, placing the secondary link in the forwarding state immediately without previously going through listening and learning states. In the event of root bridge failure, RSTP accelerates the aging of protocol information, allowing rapid failure detection. With RSTP, the E-Series can achieve very fast convergence in simple 2-tier configurations, in the range of tens or hundreds of milliseconds.

In deployments that can benefit from multiple tagged 802.1Q VLANs, the E-Series offers support for Multiple Spanning Tree Protocol (MSTP) and stacked 802.1Q VLANs. The combination of 802.1Q VLANs and MSTP allows active-active redundancy and load sharing over parallel paths through the Layer 2 switched network.
Switch/Router Device Reliability
To increase network reliability to ever more stringent ‘less than 100 millisecond’ outage demands with no packet loss requirements – all at ever increasing data rates – next generation terabit scale devices need to be considered. Beyond zero packet loss hitless failover of redundant components, protected memory, and modular software, thought must be given to preventing catastrophic lock-up of a device causing the need to reboot.

Robust Control Plane: Maximizing the robustness of the control plane was one of the primary design goals for the E-Series product line. Each Route Processor Module (RPM) dedicates a microprocessor with its own pool of ECC/parity-protected memory to each control function: Layer 2 switching, Layer 3 routing, and system management. The use of three independent microprocessors increases the aggregate capacity of the control plane, prevents one function (e.g., Spanning Tree) from depriving processing cycles from other functions (e.g., routing updates), and isolates problems that could otherwise lead to catastrophic failures. ECC memory also greatly reduces the possibility of parity-related crashes, a fairly common problem in the Internet infrastructure. Control traffic to each microprocessor may also be classified and prioritized, with the lower priority traffic rate-limited as needed to protect critical control tasks.

Redundant Switch Fabric: The E-Series switch fabric design uses 8:1 redundant Switch Fabric Modules (SFMs) that together provide over 56 Gbps of non-blocking bandwidth to each line card slot over a passive copper backplane. Unlike optical backplanes or active copper backplanes, the E-Series backplane has no single points of failure. In fact, when a single SFM fails, backplane bandwidth is unaffected. With other backplane architectures that use active dual redundant SFMs, the failure of a single SFM can reduce backplane capacity as much as 50%. The switch fabric also eliminates costly electrical-optical-electrical (EOE) conversions. As a result, the E-Series backplane provides simple, bullet-proof reliability at reasonable cost.

Hot Swappable Redundant System Components: All key systems in the E-Series are redundant and hot swappable including the RPMs, SFMs, cooling, and power. Line cards are hot swappable with redundancy achievable through configuration of LAGs distributed across multiple line cards.

Software Resiliency: The Force10 Operating System (FTOS) is a modularized switch operating system optimized for a multiprocessing control plane with passive dual-redundancy. FTOS provides automatic synchronization of configuration information between redundant RPMs in order to minimize recovery time in the case of an RPM or SFM failure. With full synchronization, the non-stop, "hitless" forwarding feature ensures that the system continues to forward traffic without packet loss in the event of an RPM failover. Furthermore, the hitless forwarding feature enables users to perform hitless software upgrades by loading a new version of the FTOS software on the standby RPM, and then initiating an RPM failover to begin operation of the new version.

To reduce system downtime during replacement of line cards, FTOS supports persistent configuration and pre-configuration of line card slots. When a line card is removed, FTOS stores the line card type, MAC address assignments, and configuration information. When the replacement card is inserted in the slot, FTOS senses the insertion and automatically gives the new card the stored configuration. With pre-configuration, the system administrator can configure an empty slot as if a line card was present. When a card is
inserted in the pre-configured slot, the stored configuration is automatically loaded by FTOS. These features dramatically reduce the expertise level and time required for line card swap-outs or for provisioning of new line card capacity.

QoS, Security, and Traffic Statistics

QoS: The E-Series provides extensive QoS and traffic management capabilities designed to comply with QoS standards, including 802.1p and IP DiffServ specifications for traffic marking. With the service-aware QoS capabilities of the E-Series architecture, ISPs can honor customer-defined traffic priorities, or assign their own class of service policies to enforce simple and manageable Service Level Agreements (SLAs). For the ISP, the simplest approach would be to assign and enforce priorities based on the subscriber’s service tier.

The Force10 ASICs have the ability to read, set, and re-map the priorities for the Ethernet and IP frames. Traffic conditioning is based upon two-rate, three-color, token-bucket metering and marking. Eight queues per destination port map directly to class-based DiffServ and IEEE 802.1p queuing models. Congestion avoidance is enabled by configurable drop-precedence probability curves of Weighted Random Early Discard (WRED). The combination of these features enables Committed Access Rate (CAR)-based service offerings with rate policing and limiting. For example, the ISP could offer multiple CAR bandwidths on Fast Ethernet of Gigabit Ethernet access links (possibly 25 or 50 Mbps over FE and 250 or 500 Mbps over GbE).

QoS is also built into the switch fabric. Both ingress and egress buffering are provided, including back-pressure mechanisms that ward off the possibility of head-of-line blocking. Separate unicast and multicast queues with up to 200 milliseconds of buffering enable minimal packet loss even in oversubscribed network conditions. The E-Series switch fabric uses Interleaved Weighted Fair Queuing (IWQ) to schedule traffic out of the ingress and egress queues, and programmable queue sizes allow seamless handling of both real-time and bursty traffic patterns.

Security: The E-Series supports up to 1.1 million Layer 2 and Layer 3 Access Control Lists (ACLs). Because the ACL filtering is performed by the line card ASICs in parallel with packet forwarding, any number of ACLs may be configured without affecting the throughput or the latency of the E-Series device.

As mentioned earlier, another E-Series feature that can be used to enhance security is its ability to prioritize and rate limit traffic sent to the control processors on the RPM. For example, in the unlikely event that an attacker learns the IP address of the E-Series Layer 2 switch and launches a DOS attack based on flooding it with ICMP messages, the attack can be mitigated by rate limiting low priority ICMP traffic, allowing the management processor to continue to receive high priority traffic and handle critical tasks, such as gathering statistics and sending SNMP traps. The E-Series rate limiting could also be applied to customer traffic to throttle DOS attacks based on ICMP or SYN flooding.

Traffic Statistics: Based on IETF RFC 3176, sFlow is a standards-based sampling technology embedded in the forwarding ASICs of E-Series switch/routers. sFlow provides the ability to continuously monitor Layer 2-Layer 7 traffic flows at wire speed simultaneously on all ports. The sFlow Agent is a software process that runs on the network management processor that aggregates interface counter values, forwarding table information, and traffic samples into sFlow datagrams that are forwarded across the network to an sFlow Collector, where statistics are stored and used for analysis and report generation. The
sFlow traffic statistics can be used in a variety of ways, including:

- Real-time congestion management
- Understanding bandwidth consumption by application type (e.g., P2P, Web, FTP, email, etc.)
- Usage accounting for billing and charge-back
- Audit trail analysis to identify unauthorized network activity and trace the sources of Denial-of-Service attacks
- Route profiling and peering optimization
- Trending and capacity planning

**E-Series Deployments Throughput the POP**

Over the next few years, the cost-effectiveness of Ethernet technology will result in its becoming the dominant transmission technology both in both the Metro and the WAN, in addition to the LAN. This technology transition will make it possible to greatly simplify network infrastructures, including ISP networks and their POPs. For example, when 10 GbE is deployed both within the POP and for WAN/Metro links, it will be possible to leverage the simplicity and cost-effectiveness of the E-Series in both the core and access tiers of the POP. Figure 3 shows a POP in which the E-Series is used wherever the switch/router is dedicated to 10 GbE (LAN or WAN) for uplinks or inter-device links. The links shown in blue and black could be either single links or 802.3ad trunks.

As has been noted several times in the earlier sections of this document, the E-Series combines the functionality of a full-featured Layer 2 switch with that of an Internet core router. In a comparison of the E1200 with high-end Internet routers from the two leading vendors of this class of device, one finds that they are comparable. In fact, for many of these basic specifications, plus the overall robustness of the device and its control plane, the E1200 enjoys an industry-leading position.

In addition to having switching/forwarding capacity, port density, and redundancy features that compare favorably with other high-end Internet switch/routers currently available on the market, the E-Series offers very robust implementations of the comprehensive routing functionality that is required for Internet core routing.

- FIB-based routing with the entire FIB replicated on each line card. Support for up to 384K IP routes per line card for Internet scale routing
- Robust, standards-compliant implementations of RIP v1 and v2, OSPF v2, IS-IS, and BGP-4
- OSPF and BGP graceful restart mechanisms to allow the data plane to continue forwarding packets while the router's control plane software is reloaded or restarted. Graceful restart complements the hitless RPM failover functionality described earlier
- Equal Cost Multi-path routing (ECMP) allows active-active redundancy and load sharing within logical groupings of up to 16 equal cost IP links
- Robust, standards-compliant implementations of IGMP, MBGP, PIM-SM v2, and PIM BSR to support IP multicast applications if these are required
- Support for line rate forwarding of both IPv4 and IPv6
- Support for the Virtual Router Redundancy Protocol (VRRP) to eliminate a single point of failure in data center and server farm deployments
- Support for 5-tuple and IETF DiffServ traffic classification to allow traffic prioritization (QoS), traffic shaping, policing, and rate limiting. Congestion management functionality includes 8 queues per port with buffer management based on Weighted Random Early Detection (WRED), Weighted Fair Queuing (WFQ), or strict priority queuing, line rate packet forwarding performance is maintained with these and all other Layer 3 features enabled
- Support over a million of standard and extended ACLs to classify and control Layer 3 traffic
ISP Point of Presence

Next Generation POPs based on the E-Series

After ISPs have implemented POPs similar to the one shown in Figure 3, the next logical step in the migration to an Ethernet-centric network architecture would be to further simplify the POP. For example, by exploiting the 10 GbE optimization and Layer 3 routing functionality of the E-Series switch/router, it would be possible to collapse the core and aggregation tiers of the POP into a single tier as shown in Figure 4. The POP is simplified through reduction of the number of devices required and the elimination of the need to manage and configure the separate Layer 2 functionality of an aggregation tier. Furthermore, if the Ethernet uplinks from the access tier routers are over-provisioned (i.e., are under-subscribed), the possibility of congestion (and the need to configure QoS policies) will be eliminated for this tier. This would allow the ISP to administer its QoS and traffic control policies entirely on the E-Series devices in the core/aggregation tier. In addition to simplicity, this approach has the advantage of applying policies where they will have absolutely no impact on the forwarding capacity of the POP.

Undersubscribed access uplinks also provide an opportunity for the ISP to offer tiered services based on sub line rate bandwidths, using the traffic shaping and policing features of the E-Series to control subscriber traffic and provision bandwidth increases.

Figure 3: E-Series in the core and access tiers

Figure 4: 2-Tier ISP PoP with the E-Series
Conclusion/Summary

The Force10 E-Series switch/router is truly a next generation device, specifically designed to meet the challenges of providing full-featured Layer 2 switching and Internet routing in an Ethernet-optimized platform made to deliver forwarding capacity that scales into the multi-Tbps range. For application in the ISP POP, the E-Series offers a number of unique advantages:

**Highest port densities:** Up to 672 GbE and 56 Ten GbE ports.

**Choice of chassis configuration:** A range of modular platforms (1/6 to 1/2 rack) to suit different applications and different-sized networks.

**Uncompromising predictable performance:** True non-blocking, wire-speed performance regardless of port density or feature utilization, including any combination of QoS processing, ACLs, and rate-limiting/policing.

**Ultra robust hardware and software:** Subsystem redundancy features include the ultimate in fault-tolerant switch fabrics with 8:1 redundancy, a unique multi-processing control plane that supports hitless RPM failovers and software updates, as well graceful protocol restart for BGP and OSPF.

**Carrier-class Internet routing and Layer 2 switching:** Layer 2 switching is fully aware of Layer 3/4 packet information and the E-Series supports the gamut of standards-based Layer 2 and Layer 3 resiliency and redundancy features. The Layer 3 routing architecture features distribution of the complete Internet-scale FIB (up to 384 K route entries) to each line card. With support for Internet routing (BGP-4, OSPF, and IS-IS), the E-Series can provide a single solution for the implementation of highly robust Layer 2/Layer 3 networks.

**Clear migration path to next generation POPs:** The E-Series is the only true Internet router product available that is optimized for Ethernet technology, including 10 GbE today and 100 GbE in the future.

**Support of stringent SLAs:** Through providing both system resiliency at the control and data forwarding level and predictable terabit scale switching and routing, the E-Series can support SLAs that not only set new uptime benchmarks but also enable service differentiation and therefore offer a competitive advantage.