White Paper

T-series Routing Platforms: System and Packet Forwarding Architecture

Chuck Semeria
Marketing Engineer

Juniper Networks, Inc.
1194 North Mathilda Avenue
Sunnyvale, CA 94089 USA
408 745 2000 or 888 JUNIPER
www.juniper.net

Part Number 200027-001 04/02
Contents

Executive Summary .................................................. 4
T640 Internet Routing Node Architecture .......................... 4
  Design Objectives .................................................. 4
  Packet Forwarding Performance .................................. 4
  Bandwidth Density .................................................. 5
  IP Service Delivery .................................................. 5
  Multichassis Capability ............................................. 5
  High Availability ................................................... 5
  Single Software Image .............................................. 6
  Security ............................................................... 6
Architectural Framework .............................................. 6
T-series Router Architectural Components ......................... 7
  Physical Interface Card ............................................. 7
  Packet Forwarding Engine ......................................... 8
  Switch Fabric ........................................................ 8
  Routing Engine ...................................................... 8
The T-series Router Chipset .......................................... 8
T640 Internet Routing Node Forwarding Path Architecture ....... 9
  Ingress PFE Forwarding Path Architecture ....................... 10
    Ingress Physical Interface Card ................................ 10
    Ingress L2/L3 Packet Processing ASIC .......................... 10
    Ingress Switch Interface ASIC (Network Mode) ................ 11
    Ingress T-series Internet Processor ASIC ....................... 11
    Ingress Queuing and Memory Interface ASIC (Queue Mode) ... 12
    Ingress Switch Interface ASIC (Fabric Mode) .................. 12
  Egress PFE Forwarding Path Architecture ......................... 12
    Egress Switch Interface ASIC (Fabric Mode) ................. 13
    Egress T-series Internet Processor ASIC ....................... 13
    Egress Queuing and Memory Interface ASIC (Queue Mode) ... 14
    Egress Switch Interface ASIC (Network Mode) ................ 14
    Egress L2/L3 Packet Processing ASIC .......................... 14
    Egress Physical Interface Card ................................ 14
T640 Internet Routing Node Single-Chassis Switch Fabric ....... 14
  Switch Fabric Properties ........................................... 15
    Nonblocking .................................................... 15
    Fair Bandwidth Allocation ..................................... 16
    Maintains Packet Order ......................................... 17
    Low Latency for High-Priority Traffic ......................... 17
    Distributed Control ............................................. 17
    Redundancy and Graceful Degradation .......................... 17
  Switch Fabric Operation .......................................... 18
T-series Multichassis Switch Fabric ................................ 18
  Three-Stage Clos Switch Fabric .................................. 19
    Clos Fabric Topology ........................................... 19
  Juniper Networks Implementation ................................ 20
  T-series Multichassis System ..................................... 20
Summary ........................................................................ 22

Copyright © 2002, Juniper Networks, Inc.
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M-Series Router Architecture</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>T-series Router Architecture</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>T-series Router Entities</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>PFE and Switch Fabric Constructed Using the T-series Chipset</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>Ingress PFE Forwarding Path Architecture</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>Egress PFE Forwarding Path Architecture</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>Single Chassis Switch Fabric Planes</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>A 4x4 Nonblocking Crossbar Switch</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>An Overcommitted Egress Switch Port</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>Three-Stage Clos Topology</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>Multichassis Switch Fabric Planes</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>T-series Multichassis System</td>
<td>21</td>
</tr>
</tbody>
</table>
Executive Summary

In August 1998, Juniper Networks began shipping the M40 Internet router. This was the first Internet-scale router built to offer uncompromising packet forwarding performance, rich packet processing, and reliable IP service delivery. Since then, Juniper Networks has expanded the M-series Internet router family to include the M5, M10, M20, M40e, and M160 routers. All M-series Internet routers are designed to meet the demands of the new public network by delivering high reliability, scalability, consistent JUNOS Internet software features, advanced application specific integrated circuit (ASIC) technology, and the same predictable packet forwarding performance across all platforms.

In April 2002, Juniper Networks began shipping the T640 Internet routing node. The T640 Internet routing node is the first in a new family of Internet routers from Juniper Networks. It delivers breakthrough OC-48/STM-16 and OC-192/STM-64 interface density in a multichassis-capable system. T-series routers use JUNOS software and our next-generation ASICs to provide the ease of use, performance, reliability, and feature richness you have come to expect from all Juniper Networks products.

This paper provides a technical introduction to the architecture of the T640 Internet routing node—both single-chassis and multichassis systems. It describes our design objectives, system architecture, packet forwarding architecture, single-chassis switch fabric architecture, and multichassis switch fabric architecture.

T640 Internet Routing Node Architecture

We begin with an introduction to the architecture of the T640 Internet routing node. This overview is divided into four parts: design objectives, architectural framework, architectural components, and the T-series router chipset.

Design Objectives

The T640 Internet routing node was developed to support seven key design objectives:

- Packet forwarding performance
- Bandwidth density
- IP service delivery
- Multichassis capability
- High availability
- Single software image
- Security

Packet Forwarding Performance

Raw packet forwarding performance determines the ability of your network to support the services provided at the edge of your network, grow these services, and deliver new revenue-generating services. If core routers cannot provide the necessary forwarding performance, your network will be unable to deliver multiple services over a common infrastructure (a multiservice network), support mission-critical applications, or meet subscriber service level agreements (SLAs).
A single-chassis T640 Internet routing node can forward an aggregate 640 million packets per second (Mpps). When compared to the M160 router, which can forward an aggregate 160 Mpps, the T640 provides a fourfold increase in aggregate packet forwarding performance in just two years.

**Bandwidth Density**

Bandwidth density determines your ability to find sufficient rack space and adequate power to deploy equipment at a rate commensurate with circuit availability and subscriber demand. Bandwidth density also simplifies the management of your network because fewer routers are required to deliver a given amount of aggregate bandwidth.

A single-chassis T640 Internet routing node supports 32xOC-192/STM-64 or 128xOC-48/STM-16 interfaces in half a rack form factor. When compared to the M160 router, which supports 8xOC-192/STM-64 or 32xOC-48/STM-16 interfaces in half a rack, the T640 provides a fourfold increase in bandwidth density in just two years.

**IP Service Delivery**

For high-performance core routers, IP service delivery is concerned with the ability to support aggregate DiffServ traffic classes, filtering, policing, rate limiting, and monitoring traffic on high-speed trunks. All T-series and M-series routers support these control functions in hardware without significantly impacting packet forwarding performance. These hardware-based tools allow you to create and maintain an environment that can support the services offered at the edges of your network.

**Multichassis Capability**

The typical lifetime of a core router is 18 to 24 months. Multichassis systems allow providers to extend the deployable lifetime of their core routers to five or more years by allowing them to increase the size of their systems in whatever increments their budget and traffic load require. Multichassis systems simplify point of presence (POP) design by merging multiple layers of core routers into a single cluster while minimizing deployment costs because the cluster components remain in place for a longer period of time. The T-series family of routers has been designed to provide a seamless transition to a multichassis system at any time.

**High Availability**

High availability ensures the dependability of subscriber services running across your network. High availability results from the deployment of both redundant hardware and stable software—the deployment of one without the other is meaningless. Reliability goals are achieved by isolating the effect of hardware and software component failures to the smallest possible portion of the system. This allows you to add, remove, and install new components without negatively impacting the system.

Single-chassis and multichassis T-series routers provide high availability because there is no single point of system failure. Hardware serviceability is supported by hot-swappable Physical Interface Cards (PICs) and Flexible PIC Concentrators (FPCs). Hardware reliability is supported by the use of redundant routing engines, switch fabric planes, power supplies, and so on. All single-chassis and multichassis T-series routers inherit the fundamental features and reliability provided by JUNOS software.
Single Software Image

Software determines the features that are available to control the operation of your network. The use of a single software image means that all features are supported in hardware on all platforms and all interfaces. You don’t have to worry about certain features appearing on one platform before they are available on other platforms. A single software image eliminates the need to understand complex feature, software, line card, and chassis matrices and provides service continuity with predictable functionality. Finally, a single software image using a single user interface makes it much easier to transfer your operational expertise with M-series routers to the T640 Internet routing node.

Security

In today’s environment, security is a critical aspect of network operations. If you cannot provide a secure environment, then you are unable to satisfy one of the most fundamental requirements of your subscribers. All T-series and M-series routers provide a consistent and comprehensive set of security tools (filtering, rate limiting, tracing, logging, source address verification, and so on) that allow you to manage security from any location (core or edge) or any interface (DS-1 to OC-192/STM-64) in your network.

Architectural Framework

The M-series router architecture successfully provides the performance and control that carriers need to support their transition from OC-12/STM-4 to OC-48/STM-16 and from OC-48/STM-16 to OC-192/STM-64. In fact, other router vendors are still struggling to achieve the performance, reliability, and feature richness delivered by the M-series routers. However, the challenges facing carriers in 2002 are different from the challenges they faced in 1998. Today, carriers demand extremely scalable, fault-tolerant, multichassis systems that have a deployable lifetime of five or more years.

Figure 1 illustrates the architecture of an M-series router. In an M-series router, each input interface is logically connected to a switch which distributes the cells of a packet into a centralized shared memory system. The cells are written into memory once, read from memory once, and sent through a switch to the proper output interface for transmission to the network.

Figure 1: M-Series Router Architecture

Figure 2 illustrates the architecture of a T-series router. A T-series router uses a distributed architecture with packet buffering at the ingress Packet Forwarding Engine (PFE) before the switch fabric as well as packet buffering at the egress PFE before the output port. As a packet enters a T-series router from the network, the ingress PFE segments the packet into cells, the
cells are written to ingress memory, a route lookup is performed, the cells representing the packet are read from ingress memory and are sent across the switch fabric to the egress PFE. When the cells arrive at the egress PFE, they are written to the second memory, a second route lookup is performed, the cells representing the packet are read out of the second memory, reassembled into a packet, and transmitted on the output interface to the network.

**Figure 2: T-series Router Architecture**

---

**T-series Router Architectural Components**

A single-chassis or multichassis T-series system consists of four major components: PICs, PFEs, the switch fabric, and one or more Routing Engines (Figure 3).

**Figure 3: T-series Router Entities**

---

**Physical Interface Card**

The PICs connect a T-series router to the network and perform both physical and link-layer packet processing. They perform all of the functions that are required for the router to receive packets from the network and transmit packets to the network.
Packet Forwarding Engine

Each PFE extracts the appropriate packet headers, performs routing table lookups and packet filtering, forwards packets to the correct output interface, and manages output queues for packet buffering. Logically, you can think of each PFE as a highly integrated packet processing engine (using Juniper Networks-developed custom ASICs) that can perform 80 million route lookups per second. Each PFE provides the equivalent of 2xOC-192/STM-64 interfaces to the network and more than two OC-192/STM-64 interfaces to the switch fabric to ensure high performance and eliminate head-of-line blocking.

When a packet arrives from the network, the ingress PFE performs a routing table lookup to determine the egress PFE connected to the egress PIC. The ingress PFE forwards the packet across the switch fabric to the egress PFE. The egress PFE performs a second routing table lookup to determine the output PIC and the packet is forwarded to the network.

Switch Fabric

For a single-chassis or a multichassis T-series system, the switch fabric provides connectivity between the PFEs. In a single-chassis system, the switch fabric provides connectivity among all of the PFEs residing in the same chassis. In a multichassis system, the switch fabric provides connectivity among all of the PFEs in the different chassis of the router cluster. In a single-chassis or a multichassis system, each PFE is considered contiguous to every other PFE connected to the switch fabric.

Routing Engine

The Routing Engine executes the JUNOS software and creates the routing tables which are downloaded to each PFE. The Routing Engine is connected to the other subsystems of a T-series router by an internal Ethernet. Each subsystem includes one or more embedded microprocessors for controlling and monitoring the custom ASICs, and these microprocessors are also connected to the internal Ethernet.

The T-series Router Chipset

The use of highly integrated ASICs is critical to the delivery of industry-leading forwarding performance and packet processing. In 1998, Juniper Networks shipped the M40 Internet backbone router—our first system developed using the M-series chipset. This chipset included a number of media-specific ASICs, the I/O Manager ASIC, the Distributed Buffer Manager ASIC, and the Internet Processor ASIC. In 2000, we extended the capabilities of the Internet Processor ASIC by releasing the Internet Processor II ASIC which provided hardware-based packet filtering, counting, logging, sampling, and deterministic load balancing. The use of the M-series chipset allowed us to rapidly develop and ship the M5, M10, M20, M40e, and M160 routers to help service providers build the New Public Network. All of the ASICs in the M-series chipset were developed in-house by the Juniper Networks design team.

The T-series chipset is our second generation of ASICs specifically designed and developed for a new generation of carrier-class routers. These ASICs leverage the conceptual framework, many of the building blocks, and our extensive operational experience with the M-series chipset. Figure 4 illustrates how the T-series chipset is arranged to implement a single instance of a PFE.
Figure 4: PFE and Switch Fabric Constructed Using the T-series Chipset

The T-series chipset includes the following ASICS:
- Media-specific ASICs (M-S)
- L2/L3 Packet Processing ASIC (L2/3)
- Switch Interface ASIC (SI)
- T-series Internet Processor ASIC (T-series IP)
- Queuing and Memory Interface ASIC (Q&M)
- Fabric ASIC (Fabric)

The T-series chipset provides hardware-based forwarding performance and service delivery for Internet Protocol version 4 (unicast and multicast), IPv6 (unicast and multicast), and Multiprotocol Label Switching (MPLS) while also having the flexibility to support other protocols in the future. The chipset was designed to provide the functionality needed for the development of single-chassis or multichassis systems.

Some of the ASICs operate in multiple modes. For example, the Switch Interface ASIC can operate in network mode or fabric mode, while the Queuing and Memory Interface ASIC can operate in queue mode or memory mode. The Fabric ASIC can operate in one of four modes depending on its function in the system: crossbar, Clos stage 1, Clos stage 2, or Clos stage 3.

T640 Internet Routing Node Forwarding Path Architecture

The forwarding path architecture discussed in this section details the functions performed by each ASIC in the T-series chipset. It describes the packet processing that occurs as an IPv4 unicast packet is received from the network, processed by the ingress PFE, forwarded across the switch fabric, processed by the egress PFE, and forwarded to the network. Note that the T-series chipset and all T-series routers are equally proficient at forwarding IPv4 multicast, IPv6 unicast, IPv6 multicast, and MPLS traffic.
Ingress PFE Forwarding Path Architecture

Figure 5 illustrates the forwarding path as an IPv4 packet is received from the network by an OC-192/STM-64 PIC (one of a number of PICs supported by T-series routers), processed by the ingress PFE, and forwarded to the switch fabric.

Figure 5: Ingress PFE Forwarding Path Architecture

Ingress Physical Interface Card

An IPv4 packet arrives from the network, enters the ingress PIC, and passes through an opto-electronic converter. Assuming that the packet is received by an OC-192/STM-64 SONET PIC, a Juniper Networks-designed media-specific ASIC detects frame boundaries, interprets SONET control information, and extracts the data payload. As the media-specific ASIC extracts the data payload, it immediately forwards the data across the PIC-PFE interface to the L2/L3 Packet Processing ASIC.

Ingress L2/L3 Packet Processing ASIC

The L2/L3 Packet Processing ASIC:

- Removes the L2 header from the packet
- Locates and checks the L3 packet header
- Extracts control information (protocol type, DiffServ, and so on) from the L2 and L3 packet headers
- Performs input accounting
- Segments each packet into data cells
- Forwards the cells to the network mode of the Switch Interface ASIC
The ingress PFE, switch fabric, and egress PFE operate exclusively on fixed-length cells. The original packet is not reassembled until it arrives at the L2/L3 Packet Processing ASIC of the egress PFE.

**Ingress Switch Interface ASIC (Network Mode)**

The cells from the packet are received by the network mode of the Switch Interface ASIC. The Switch Interface ASIC operates in two modes: network mode and fabric mode. When operating in network mode, the Switch Interface ASIC of the ingress PFE:

- Extracts the route lookup key for the packet. While a typical lookup key contains the source and destination address, the network mode of the Switch Interface ASIC can be programmed to extract any information from the cells of the packet.
- Distributes the cells of the packet across the local memory subsystem. The local memory subsystem contains four Queuing and Memory Interface ASICs operating in memory mode.
- Collects the address of each cell after it is written to local memory by the memory mode of the Queuing and Memory Interface ASICs. The network mode of the Switch Interface ASIC stores the addresses of the cells as they are written to memory.
- Combines the addresses with the route lookup key to create a packet “notification.” T-series routers use a data structure called a packet notification to keep track of the memory locations of the cells that make up a single packet. The notification also contains control information such as the packet’s arrival interface, priority, and protocol type.
- Forwards the packet notification to the T-series Internet Processor ASIC for route lookup when the end of the packet arrives and all data cells have been written to memory.

The Queuing and Memory Interface ASICs operate in two modes: memory mode and queue mode. When operating in memory mode, the Queuing and Memory Interface ASICs accept write requests from the network mode of the Switch Interface ASIC and write cells to memory. After a cell is written to memory, the memory mode of the Queuing and Memory Interface ASIC returns the write address of the cell to the network mode of the Switch Interface ASIC.

Each Queuing and Memory Interface ASIC operates as a memory controller for storing data cells and is connected to 4 banks of DRAM, providing a total of 16 banks of DRAM for each PFE.

**Ingress T-series Internet Processor ASIC**

The notification representing the packet is received by the T-series Internet Processor ASIC which performs the route lookup, packet filtering, traffic policing, and packet classification. The T-series Internet Processor ASIC is based on a new design that incorporates our experience in the Internet classroom with the Internet Processor II ASIC. The key enhancements supported by the T-series Internet Processor ASIC include:

- Delivering route lookup and packet processing performance of 80 Mpps to support the design goals of the T-series architecture.
- TCP-friendly support for rate limiting using hardware-based packet filtering.

After completing the route lookup and packet processing, the T-series Internet Processor ASIC writes additional information into the notification (such as the next hop, egress PFE, output queue number, and other control information) and forwards the notification to the queue mode of the Queuing and Memory Interface ASIC.
Ingress Queuing and Memory Interface ASIC (Queue Mode)

The notification is received by the queue mode of the Queuing and Memory Interface ASIC. Recall that the notification from the T-series Internet Processor ASIC contains the egress PFE and a queue number. This allows the T-series Internet Processor ASIC to select not only the egress PFE but also the priority level of each packet.

When operating in queue mode, the Queuing and Memory Interface ASIC functions as a memory controller with four banks of DRAM, however this memory is used to maintain queues of packet notifications rather than data cells. The queue mode of the Queuing and Memory Interface ASIC:

- Performs priority queuing for cells departing the ingress PFE for the switch fabric
- Performs random early detection (RED) drops if the switch fabric experiences congestion
- Polices and rate-shapes traffic
- Forwards a notification to the fabric mode of the Switch Interface ASIC when each notification arrives at the head of its queue

Ingress Switch Interface ASIC (Fabric Mode)

When operating in fabric mode, the Switch Interface ASIC provides the interface between the ingress PFE and the switch fabric. The fabric mode of the Switch Interface ASIC:

- Uses the cell addresses carried in the packet notification to fetch the corresponding data cells from the memory mode of the Queuing and Memory Interface ASICs
- Prepends a header and cyclic redundancy check (CRC) trailer to each cell. The header contains the source PFE, the destination PFE, control codes, and a cell sequence number that is used by the egress PFE to ensure ordering
- Forwards the cells that make up the packet across the switch fabric to the egress PFE
- Forwards the notification representing the packet across the switch fabric to the egress PFE
- Participates in switch fabric flow control and fault tolerance algorithms

For the moment, assume that the switch fabric is a black box and that all of the cells in the packet are transported across the fabric from the ingress PFE to the egress PFE. A more detailed description of the operation of the switch fabric is provided later in this document.

Egress PFE Forwarding Path Architecture

Figure 6 illustrates the forwarding path as a packet is received from the switch fabric, processed by the egress PFE, and forwarded to the network.
**Egress Switch Interface ASIC (Fabric Mode)**

The cells of the packet arrive at the egress PFE from the switch fabric and are received by the fabric mode of the Switch Interface ASIC. As the cells arrive, the fabric mode of the Switch Interface ASIC:

- Maintains a reorder buffer to ensure that packets (and the cells in a given packet) are reassembled in the correct order.
- Extracts the route lookup key for the packet.
- Writes all of the cells that make up the packet into the local memory system (implemented by the four Queuing and Memory Interface ASICs operating in memory mode) of the egress PFE.
- Collects the new addresses returned from the memory mode of the Queuing and Memory Interface ASICs and combines them with the route lookup key to update the packet notification. The cell addresses must be updated because the cells are stored at different memory locations in the egress PFE than they were in the ingress PFE.
- Forwards the updated packet notification to the T-series Internet Processor ASIC for a second route lookup.

**Egress T-series Internet Processor ASIC**

The packet notification is received by the T-series Internet Processor ASIC of the egress PFE. The T-series Internet Processor performs a second route lookup based on the lookup key carried in the packet notification. The second route lookup provides the opportunity to perform additional packet filtering, packet policing, and packet classification. After completing its tasks, the T-series Internet Processor ASIC updates the information in the notification and forwards the notification to the queue mode of the Queuing and Memory Interface ASIC.
Egress Queuing and Memory Interface ASIC (Queue Mode)

The queue mode of the Queuing and Memory Interface ASIC of the egress PFE:

- Implements a weighted deficit round-robin algorithm for packets exiting the egress PFE for the network
- Performs RED drops per queue if the queue experiences congestion
- Forwards the notification to the network mode of the Switch Interface ASIC when each notification arrives at the head of its queue

Egress Switch Interface ASIC (Network Mode)

When operating in network mode, the Switch Interface ASIC of the egress PFE:

- Uses the recomputed cell addresses carried in the packet notification to fetch the corresponding data cells from the memory mode of the Queuing and Memory Interface ASICs
- Forwards the cells that make up the packet to the L2/L3 Packet Processing ASIC

Note that the processing performed by the network mode of the Switch Interface ASIC on the egress PFE is similar to the functions performed by the fabric mode of the Switch Interface ASIC on the ingress PFE. The primary difference is that the network mode of the Switch Interface ASIC on the egress PFE interfaces with the L2/L3 Packet Processing ASIC rather than the switch fabric.

Egress L2/L3 Packet Processing ASIC

The L2/L3 Packet Processing ASIC of the egress PFE:

- Updates output accounting statistics
- Updates the L3 header
- Adds the L2 header
- Reassembles the data cells into an IPv4 packet (stream of bytes) as they are forwarded to the egress PIC

Egress Physical Interface Card

The stream of bytes is received by the PIC connected to the egress PFE. The OC-19/STM-64 media-specific ASIC on the egress PIC performs SONET encapsulation and transmits the IPv4 packet on the appropriate output interface to the network.

T640 Internet Routing Node Single-Chassis Switch Fabric

Initially, T-series routers are available as a single-chassis system. In a single-chassis system (see Figure 7), the switch fabric is implemented using four operationally independent but identical switch planes (labeled A through D) that are simultaneously active and an identical fifth plane (labeled E) that acts as a hot spare to provide redundancy. Each plane contains a single, Juniper Networks-designed 16-port Fabric ASIC that operates in crossbar mode.
The 16-port Fabric ASIC provides nonblocking connectivity among the 16 PFEs that can populate a single-chassis system. Each chassis contains a maximum of 8 FPCs, with each FPC supporting a maximum of 2 PFEs, for a total of 16 PFEs that communicate across the switch fabric.

The required input and output aggregate bandwidth of the switch fabric exceeds the I/O capabilities of a single Fabric ASIC. Each PFE may transmit two OC-192/STM-64s worth of data cells across the fabric and speedup is required to compensate for cellification overhead, scheduling overhead, and to reduce congestion in the fabric. To overcome this limitation, the fabric is implemented as multiple switch planes, each PFE is connected to four active switch planes, and each switch plane carries a portion of the required bandwidth. To guarantee that cells are evenly load balanced across the active switch planes, each PFE distributes cells equally across the four switch planes on a cell-by-cell basis rather than a packet-by-packet basis.

**Switch Fabric Properties**

The T-series router switch fabric for both single-chassis and multichassis systems is specifically designed to provide the following attributes:

- Nonblocking
- Fair bandwidth allocation
- Maintains packet order
- Low latency for high-priority traffic
- Distributed control
- Redundancy and graceful degradation

**Nonblocking**

A switch fabric is considered nonblocking if two traffic flows directed to two different output ports never conflict. In other words, the internal connections within the switch allow any ingress PFE to send its fair share of bandwidth to any egress PFE simultaneously.
Figure 8: A 4x4 Nonblocking Crossbar Switch

Figure 8 illustrates the internal topology for a nonblocking, single stage, four-port crossbar switch. The challenge when building a crossbar is that it requires $n^2$ communication paths internal to the switch. In this example, the four-port crossbar requires a communication path connecting each input port to each output port, for a total of 16 communication paths. As the number of ports supported by the crossbar increases, the $n^2$ communication path requirement becomes an implementation challenge.

Fair Bandwidth Allocation

In a production network, it is impossible to control the pattern of ingress traffic so that an egress port of the crossbar switch is never overcommitted. An egress port becomes overcommitted when there is more input traffic destined for the egress port than the egress port can forward. In Figure 9, the aggregate amount of traffic that ingress ports 1, 2, and 3 forward to egress port 4 is greater than the capacity of egress port 4. Fair bandwidth allocation is concerned with the techniques that a switch uses to share the bandwidth among competing ingress flows to an overcommitted egress port.

Figure 9: An Overcommitted Egress Switch Port

The T-series router switch fabric provides fairness by ensuring that all ingress PFEs receive an equal amount of bandwidth across the switch fabric when transmitting cells to an oversubscribed egress PFE. Providing this type of fairness across all streams is difficult to support because it is difficult to keep track of all users of the bandwidth to the egress PFE. The challenge is that if you have $n$ ports on your switch then there are $n^2$ streams of traffic through
the switch. Because most switch architectures are not capable of keeping track of \( n^2 \) individual streams, they are forced to aggregate traffic streams making it impossible to be completely fair to each individual stream. The T-series switch fabric can monitor the \( n^2 \) streams so that each stream receives its fair share of the available fabric bandwidth to an oversubscribed egress PFE.

**Maintains Packet Order**

The potential for misordering cells as they are transmitted across parallel switch planes to the egress PFE is eliminated by the use of sequence numbers and a reorder buffer. In this design, the fabric mode of the Switch Interface ASIC on the ingress PFE places a sequence number into the cell header of each cell that it forwards into the fabric. On the egress PFE, the fabric mode of the Switch Interface ASIC buffers all cells that have sequence numbers greater than the next sequence number it expects to receive. If a cell arrives out of order, the fabric mode of the Switch Interface ASIC buffers the cells until the correct in-order cell arrives and the reorder buffer is flushed. The reorder buffer and sequence number space are large enough to ensure that packets (and the cells in a given packet) are not reordered as they traverse the switch fabric.

**Low Latency for High-Priority Traffic**

Some types of traffic, such as voice or video, have latency requirements as well as bandwidth requirements. The T-series router switch fabric is designed so that blocking in the fabric is extremely rare because the size of the ports into the fabric is considerably larger than the network ports into an ingress PFE. In the rare case that congestion does occur, each ingress PFE is allocated priority queues into the switch fabric.

As discussed above, the switch fabric fairly allocates bandwidth among all of the ingress PFEs that are competing to transmit cells to an overcommitted egress PFE. As long as each ingress PFE does not forward an excessive number of high-priority cells into the fabric and exceed its fair bandwidth allocation, high-priority traffic is guaranteed a path through the switch fabric.

The use of priority queues from the ingress PFE into the switch fabric provides two important benefits:

- The latency for high-priority traffic is always low because the fabric will never become congested.
- The class of service (CoS) intelligence required to perform admission control into the fabric is implemented in the PFEs. This allows the switch fabric to remain relatively simple because CoS is not implemented inside the fabric, but at the edges of the switch fabric.

**Distributed Control**

A T-series router does not have a centralized controller that is connected to all of the components in the switch fabric. This means that within the fabric any component can fail and the other components around the failed component will continue to operate. Additionally, a centralized control channel does not need to be operational for the switch fabric to function.

**Redundancy and Graceful Degradation**

The fabric mode of each Switch Interface ASIC monitors a request-grant mechanism. If the ingress PFE expects a grant for an outstanding request but the grant does not return after a reasonable amount of time, then the ingress PFE assumes that the destination PFE is unreachable on the plane that was used to send the request. If a switch plane fails, only the
cells that are currently in transit across the switch plane are lost because buffering does not occur within a switch plane and the request-grant mechanism ensures that cells are never transmitted across a failed switch plane.

The request-grant mechanism allows a failed component within a switch plane to be removed from service by diverting traffic around the faulty component, or all traffic using the faulty plane can be switched to a redundant plane. If there is a significant number of faults on a given switch plane or the plane must be swapped out for maintenance, the chassis manager coordinates moving traffic to the redundant switch plane. Each step in the migration of traffic to the redundant plane involves moving only a small fraction of overall traffic. This allows the system to remain operational with no significant loss in fabric performance. If a second switch plane fails resulting in a system with only three active planes, the fabric continues to operate but with a reduction in performance.

Switch Fabric Operation

The process of transmitting a data cell across the switch fabric involves a request and grant protocol:

- The source PFE transmits a request across the switch fabric to the destination PFE. Each request for a given destination is transmitted across a different switch plane in a round-robin order to distribute the load equally.
- When the request is received by the destination PFE, the destination PFE transmits a grant to the source PFE across the same switch plane on which the corresponding request was received.
- When the grant is received by source PFE, the source PFE transmits the data cell to the destination PFE across the same switch plane on which the corresponding grant was received.

This approach provides both a flow control mechanism for transmitting cells into the fabric and a mechanism to detect broken paths across the switch fabric.

T-series Multichassis Switch Fabric

A popular approach to increasing forwarding performance, boosting bandwidth density, and extending the deployable lifetime of core routers is to use a multichassis system. Multichassis systems are designed with an expandable switch fabric that allows the provider to grow their systems in the increments that budget and traffic loads require. A number of switch fabric architectures can be used to support the construction of a multichassis router system:

- Toroidal meshes
- Hypercubes
- Banyan (a.k.a, Butterfly, Omega, and others) topologies
- Clos topologies

Many of these switch architectures were originally developed to support the construction of large multiprocessor computer systems. The challenge with simply selecting one of these technologies and implementing it is that the requirements for computer systems and networking systems are very different. Computer systems have the principle of locality where communication patterns are regular and very predictable. Communication patterns in large IP networks can change quickly and it is important that the switch fabric be nonblocking without making assumptions about traffic patterns on a network-wide or port-to-port basis.
Over the past few years, a number of router vendors have attempted to adapt these traditional switch fabrics to build large multichassis networking systems. An oversight in the selection or implementation of the switch fabric architecture can severely limit the future scalability of a multichassis router that communicates across the switch fabric. A number of these traditional switch fabric architectures have well-known limitations when attempting to build an any-to-any networking system:

- Toroidal meshes are blocking.
- Hypercubes have scaling problems because large systems require an impractical number of links per line card or an impractical amount of fabric speedup.
- Conventional Banyan topologies are blocking.

**Three-Stage Clos Switch Fabric**

When considering the fundamental requirement for a nonblocking fabric that can scale to support the any-to-any connectivity requirements of large IP networks, we chose the Clos topology as the switch fabric for our multichassis system. The Clos topology was originally developed in early 1950s for the design of the then next generation of scalable telephone switching equipment.

**Clos Fabric Topology**

Figure 10 illustrates the topology of a typical three-stage Clos fabric. The blue squares represent individual 16-port single-stage crossbar switches. The topology consists of multiple rows of single-stage crossbar switches arranged into three columns. The benefit of the Clos topology is that it facilitates the construction of large, scalable, and nonblocking switch fabrics using smaller switch fabrics as the fundamental building block.
The topology of the Clos network has each output port of the first stage connected to one of the crossbars in the second stage. Observe that a two-stage switch fabric would provide any-to-any connectivity (any ingress port to the fabric can communicate with any egress port from the fabric) but the path through the switch is blocking. The addition of the third stage creates a nonblocking topology by creating a significant number of redundant links. It is the presence of the redundant paths that provides the nonblocking behavior of a Clos fabric.

**Juniper Networks Implementation**

The Juniper Networks implementation of a Clos fabric for a multichassis T-series router is specifically designed to support the following attributes:

- Nonblocking
- Fair bandwidth allocation
- Maintains packet order
- Low latency for high-priority traffic
- Distributed control
- Redundancy and graceful degradation

The existence of multiple parallel paths through the Clos fabric to any egress port gives the switch fabric its nonblocking behavior. Technically, a Clos topology is referred to as “rearrangably nonblocking.” A proof from the early 1950s states, “If you use a Clos fabric in circuit switched telephone equipment, set up some existing connections across the fabric, and you are required to add a new connection between an unused ingress and egress port, THEN there will always be a path to support the new connection by moving your existing connections to currently unused paths and then adding the new connection.” Thus, a circuit-switched Clos network is rearrangably nonblocking.

However, is a packet- or cell-switched Clos network nonblocking? Dividing packets into cells and then distributing them across the Clos switch fabric achieves the same effect as moving connections in a circuit-switched network because all of the paths are used simultaneously. Thus, a Clos fabric proves to be nonblocking for packet or cell traffic. This allows PFEs to continue to send new traffic into the fabric and, as long as there are no inherent conflicts such as an overcommitted output port, the switch remains nonblocking.

Finally, an important property of a Clos fabric is that each crossbar switch is totally independent of the other crossbar switches in the fabric. This means that a Clos fabric does not require a centralized scheduler to coordinate the actions of the individual crossbars. Each crossbar acts independently of the others which results in a switch fabric that is highly resilient to failures, and scales to support the construction of extremely large fabrics.

**T-series Multichassis System**

The T-series router multichassis fabric is constructed using a Juniper Networks-designed 16x16 crossbar called the Fabric ASIC. The same Fabric ASIC used in a single-chassis T-series router. Depending on its placement in the Clos topology, each Fabric ASIC is a general building block that can perform stage 1, stage 2, or stage 3 functionality within the switch fabric.

The Clos fabric provides the interconnection among the PFEs in a multichassis T-series router (Figure 11). Similar to a single-chassis system, the switch fabric for a multichassis system is implemented using four operationally independent but identical switch planes (labeled A
through D) that are simultaneously active and an identical fifth plane (labeled E) that acts as a hot spare to provide redundancy. Each plane contains a three-stage Clos fabric built using Juniper Networks Fabric ASICs.

**Figure 11: Multichassis Switch Fabric Planes**

The required input and output bandwidth of the switch fabric exceeds the I/O capabilities of a single plane. To overcome this limitation, the fabric is implemented as multiple switch planes, each PFE is connected to four active switch planes, and each switch plane provides a portion of the required bandwidth. To guarantee that cells are evenly load balanced across the active switch planes, each PFE distributes cells equally across the four switch planes on a cell-by-cell basis rather than a packet-by-packet basis.

Figure 12 illustrates the components of a T-series multichassis system.

**Figure 12: T-series Multichassis System**

A T-series multichassis system consists of a switch chassis and a number of line-card chassis. The similarities between a T-series multichassis system and a single-chassis system include:

- The chipset is the same T-series chipset used in a single-chassis system.
- The chassis components (routing engines, SONET clock generators, power supplies, cooling, and so on) are the same as those used in a single-chassis system.
The FPCs are the same as those used in a single-chassis system.

The PICs are the same as those used in a single-chassis system.

The JUNOS software and all of the features are on a single-chassis T-series router or an M-series router.

The differences between a multichassis system and a single-chassis system include:

- While the same T-series chipset is used, the Fabric ASICs are arranged differently and operate in different modes than in a single-chassis system.

- In a multichassis system, the switch interface boards (SIBs) from a single-chassis system are replaced with new SIBs that contain two Fabric ASICs and a set of optical interfaces that provide connectivity to the switch chassis. One of the Fabric ASICs operates in stage 1 (F1) mode and the other Fabric ASIC operates in stage 3 (F3) mode.

- The line-card chassis are connected to the switch chassis with fiber-optic cable. This connectivity uses the latest VCSEL technology to provide extremely high throughput, low power consumption, and low bit-error rates.

- The Fabric ASICs in the switch chassis provide the stage 2 (F2) functionality of the Clos network for the router cluster.

Summary

This paper presented a technical description of the architecture of the T640 Internet routing node for both single-chassis and multichassis systems. It detailed our fundamental design objectives, system architecture, packet forwarding architecture, single-chassis switch fabric, and multichassis switch fabric.

The T640 Internet routing node demonstrates how we have evolved our router architecture to achieve substantial technology breakthroughs in packet forwarding performance, bandwidth density, IP service delivery, and system reliability. Not only do T-series routers deliver industry-leading scalability, they provide the required scalability while still maintaining feature and software continuity across all router platforms. When the time to deploy a multichassis system arrives, you can be assured that the T-series router architecture will satisfy all of your networking requirements.