



Cisco Express Forwarding Overview

Cisco Express Forwarding (CEF) is advanced, Layer 3 IP switching technology. CEF optimizes network performance and scalability for networks with large and dynamic traffic patterns, such as the Internet, on networks characterized by intensive Web-based applications, or interactive sessions.

Procedures for configuring CEF or distributed CEF (dCEF) are provided in the [“Configuring Cisco Express Forwarding”](#) chapter later in this publication.

This chapter describes CEF. It contains the following sections:

- [Benefits](#)
- [Restrictions](#)
- [CEF Components](#)
- [Supported Media](#)
- [CEF Operation Modes](#)
- [TMS and CEF Nonrecursive Accounting](#)
- [Network Services Engine](#)
- [Virtual Profile CEF](#)

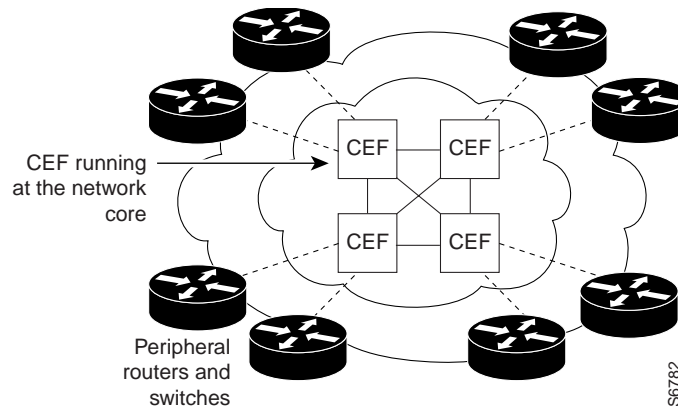
Benefits

CEF offers the following benefits:

- **Improved performance**—CEF is less CPU-intensive than fast switching route caching. More CPU processing power can be dedicated to Layer 3 services such as quality of service (QoS) and encryption.
- **Scalability**—CEF offers full switching capacity at each line card when dCEF mode is active.
- **Resilience**—CEF offers an unprecedented level of switching consistency and stability in large dynamic networks. In dynamic networks, fast-switched cache entries are frequently invalidated due to routing changes. These changes can cause traffic to be process switched using the routing table, rather than fast switched using the route cache. Because the Forwarding Information Base (FIB) lookup table contains all known routes that exist in the routing table, it eliminates route cache maintenance and the fast-switch or process-switch forwarding scenario. CEF can switch traffic more efficiently than typical demand caching schemes.

Although you can use CEF in any part of a network, it is designed for high-performance, highly resilient Layer 3 IP backbone switching. For example, [Figure 8](#) shows CEF being run on Cisco 12000 series Gigabit Switch Routers (GSRs) at aggregation points at the core of a network where traffic levels are dense and performance is critical.

Figure 8 Cisco Express Forwarding



In a typical high-capacity Internet service provider (ISP) environment, Cisco 12012 GSRs as aggregation devices at the core of the network support links to Cisco 7500 series routers or other feeder devices. CEF in these platforms at the network core provides the performance and scalability needed to respond to continued growth and steadily increasing network traffic. CEF is a distributed switching mechanism that scales linearly with the number of interface cards and the bandwidth installed in the router.

Restrictions

- The Cisco 12000 series Gigabit Switch Routers operate only in distributed CEF mode.
- Distributed CEF switching cannot be configured on the same VIP card as distributed fast switching.
- Distributed CEF is not supported on Cisco 7200 series routers.
- If you enable CEF and then create an access list that uses the **log** keyword, the packets that match the access list are not CEF switched. They are fast switched. Logging disables CEF.

CEF Components

Information conventionally stored in a route cache is stored in several data structures for CEF switching. The data structures provide optimized lookup for efficient packet forwarding. The two main components of CEF operation are described in the following sections:

- [Forwarding Information Base](#)
- [Adjacency Tables](#)

Forwarding Information Base

CEF uses a FIB to make IP destination prefix-based switching decisions. The FIB is conceptually similar to a routing table or information base. It maintains a mirror image of the forwarding information contained in the IP routing table. When routing or topology changes occur in the network, the IP routing table is updated, and those changes are reflected in the FIB. The FIB maintains next hop address information based on the information in the IP routing table.

Because there is a one-to-one correlation between FIB entries and routing table entries, the FIB contains all known routes and eliminates the need for route cache maintenance that is associated with switching paths such as fast switching and optimum switching.

Adjacency Tables

Nodes in the network are said to be adjacent if they can reach each other with a single hop across a link layer. In addition to the FIB, CEF uses adjacency tables to prepend Layer 2 addressing information. The adjacency table maintains Layer 2 next-hop addresses for all FIB entries.

Adjacency Discovery

The adjacency table is populated as adjacencies are discovered. Each time an adjacency entry is created (such as through ARP), a link-layer header for that adjacent node is precomputed and stored in the adjacency table. Once a route is determined, it points to a next hop and corresponding adjacency entry. It is subsequently used for encapsulation during CEF switching of packets.

Adjacency Resolution

A route might have several paths to a destination prefix, such as when a router is configured for simultaneous load balancing and redundancy. For each resolved path, a pointer is added for the adjacency corresponding to the next hop interface for that path. This mechanism is used for load balancing across several paths.

Adjacency Types That Require Special Handling

In addition to adjacencies associated with next hop interfaces (host-route adjacencies), other types of adjacencies are used to expedite switching when certain exception conditions exist. When the prefix is defined, prefixes requiring exception processing are cached with one of the special adjacencies listed in [Table 4](#).

Table 4 *Adjacency Types for Exception Processing*

This adjacency type...	Receives this processing...
Null adjacency	Packets destined for a Null0 interface are dropped. This can be used as an effective form of access filtering.
Glean adjacency	When a router is connected directly to several hosts, the FIB table on the router maintains a prefix for the subnet rather than for the individual host prefixes. The subnet prefix points to a glean adjacency. When packets need to be forwarded to a specific host, the adjacency database is gleaned for the specific prefix.

Table 4 *Adjacency Types for Exception Processing (continued)*

This adjacency type...	Receives this processing...
Punt adjacency	Features that require special handling or features that are not yet supported in conjunction with CEF switching paths are forwarded to the next switching layer for handling. Features that are not supported are forwarded to the next higher switching level.
Discard adjacency	Packets are discarded.
Drop adjacency	Packets are dropped, but the prefix is checked.

Unresolved Adjacency

When a link-layer header is prepended to packets, the FIB requires the prepend to point to an adjacency corresponding to the next hop. If an adjacency was created by the FIB and not discovered through a mechanism, such as ARP, the Layer 2 addressing information is not known and the adjacency is considered incomplete. Once the Layer 2 information is known, the packet is forwarded to the Route Processor (RP), and the adjacency is determined through ARP.

Supported Media

CEF currently supports ATM/AAL5snap, ATM/AAL5mux, ATM/AAL5nlpid, Frame Relay, Ethernet, FDDI, PPP, HDLC, and tunnels.

CEF Operation Modes

CEF can be enabled in one of two modes described in the following sections:

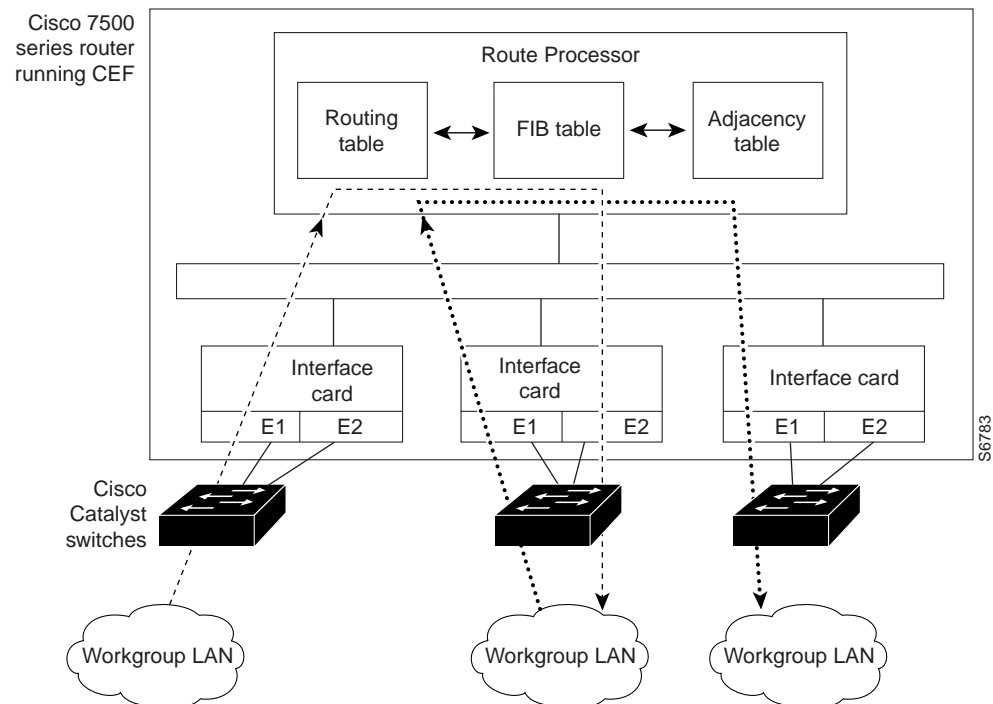
- [Central CEF Mode](#)
- [Distributed CEF Mode](#)

Central CEF Mode

When CEF mode is enabled, the CEF FIB and adjacency tables reside on the RP, and the RP performs the express forwarding. You can use CEF mode when line cards are not available for CEF switching or when you need to use features not compatible with dCEF switching.

Figure 9 shows the relationship between the routing table, FIB, and adjacency table during CEF mode. The Catalyst switches forward traffic from workgroup LANs to a Cisco 7500 series router on the enterprise backbone running CEF. The RP performs the express forwarding.

Figure 9 CEF Mode



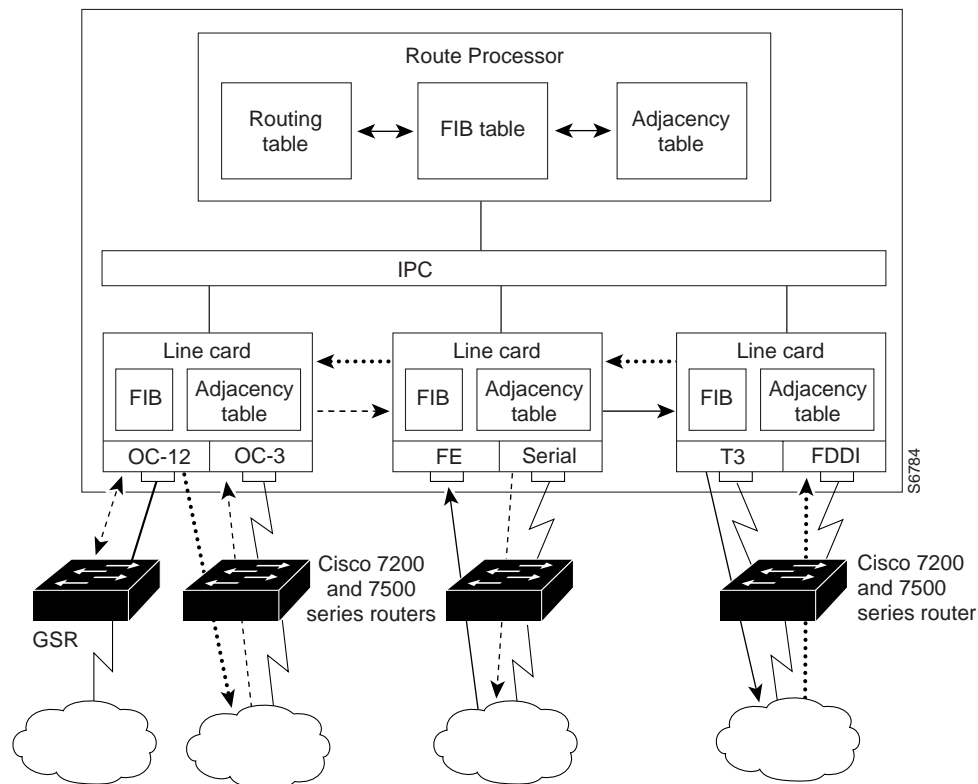
Distributed CEF Mode

When dCEF is enabled, line cards, such as VIP line cards or GSR line cards, maintain an identical copy of the FIB and adjacency tables. The line cards perform the express forwarding between port adapters, relieving the RSP of involvement in the switching operation.

dCEF uses an Inter Process Communication (IPC) mechanism to ensure synchronization of FIB tables and adjacency tables on the RP and line cards.

Figure 10 shows the relationship between the RP and line cards when dCEF mode is active.

Figure 10 dCEF Mode



In this Cisco 12000 series router, the line cards perform the switching. In other routers where you can mix various types of cards in the same router, all of the cards you are using may not support CEF. When a line card that does not support CEF receives a packet, the line card forwards the packet to the next higher switching layer (the RP) or forwards the packet to the next hop for processing. This structure allows legacy interface processors to exist in the router with newer interface processors.



Note

The Cisco 12000 series GSR operate only in dCEF mode; dCEF switching cannot be configured on the same VIP card as distributed fast switching, and dCEF is not supported on Cisco 7200 series routers.

CEF and dCEF Additional Capabilities

In addition to configuring CEF and dCEF, you can also configure the following features:

- Distributed CEF switching using access lists
- Distributed CEF switching of Frame Relay packets
- Distributed CEF switching during packet fragmentation
- Load balancing on a per-destination-source host pair or per-packet basis
- Distributed CEF switching across IP tunnels

For information on enabling these features, see the chapter “Configuring Cisco Express Forwarding.”

TMS and CEF Nonrecursive Accounting

Traffic matrix statistics (TMS) data is counted during packet forwarding by CEF nonrecursive accounting. TMS enables an administrator to capture and analyze traffic data entering a backbone that is running the Border Gateway Protocol (BGP). This feature also allows an administrator to determine the neighbor autonomous systems of a BGP destination.

The following paragraphs explain how CEF nonrecursive accounting aggregates packet statistics for IGP routes and their dependent BGP routes.

For example, a BGP network deployed by a service provider has the following components:

- IGP routes that describe the next hop to which traffic should be sent.
- BGP routes that specify an intermediate address to which traffic should be sent.

In this example, the intermediate address might be several hops away. The next hop for the BGP route is the next hop for the intermediate address of the BGP route. The BGP route is called recursive, because it points (through its intermediate address) to an IGP route that provides the next hop for forwarding.

CEF represents IGP routes as nonrecursive entries and BGP routes as recursive entries that resolve to nonrecursive entries.

CEF nonrecursive accounting counts the packets for all the CEF recursive entries that resolve to a CEF nonrecursive entry and the packets for the nonrecursive entry. The number of packets is collected and totalled in one location.

The following example networks show how CEF nonrecursive accounting counts packets when BGP routes resolve to one IGP route and when they do not. A multiaccess network access point (NAP) has BGP routes referring to hosts on that network.

- If the network is advertised as a single IGP route, all the BGP routes to the various hosts at that NAP resolve to a single IGP route. CEF nonrecursive accounting summarizes the packets to all of the BGP destinations.
- If a network administrator instead advertises individual host routes from the NAP network to the IGP, CEF nonrecursive accounting will count packets to those hosts separately.

The count of packets forwarded based on a nonrecursive CEF entry can be split into two bins based on whether the input interface of the backbone router is configured as internal or external. Thus, all packets that arrive on external interfaces (external to the region of interest) and are forwarded based on a given IGP route (either directly or through a recursive BGP route) are counted together.

TMS Data

The TMS feature allows an administrator to gather the following data:

- The number of packets and bytes that travel across the backbone from internal and external sources. The packets and bytes are called traffic matrix statistics and are useful for determining how much traffic a backbone handles. You can analyze the traffic matrix statistics using the following methods:
 - Collecting and viewing the TMS data through the application of the Network Data Analyzer (NDA).
 - Reading the TMS data that resides on the backbone router.

The following sections explain how to collect and view the traffic matrix statistics using the command-line interface (CLI) and the NDA. For detailed instructions on using the NDA, see the *Network Data Analyzer Installation and User Guide*.

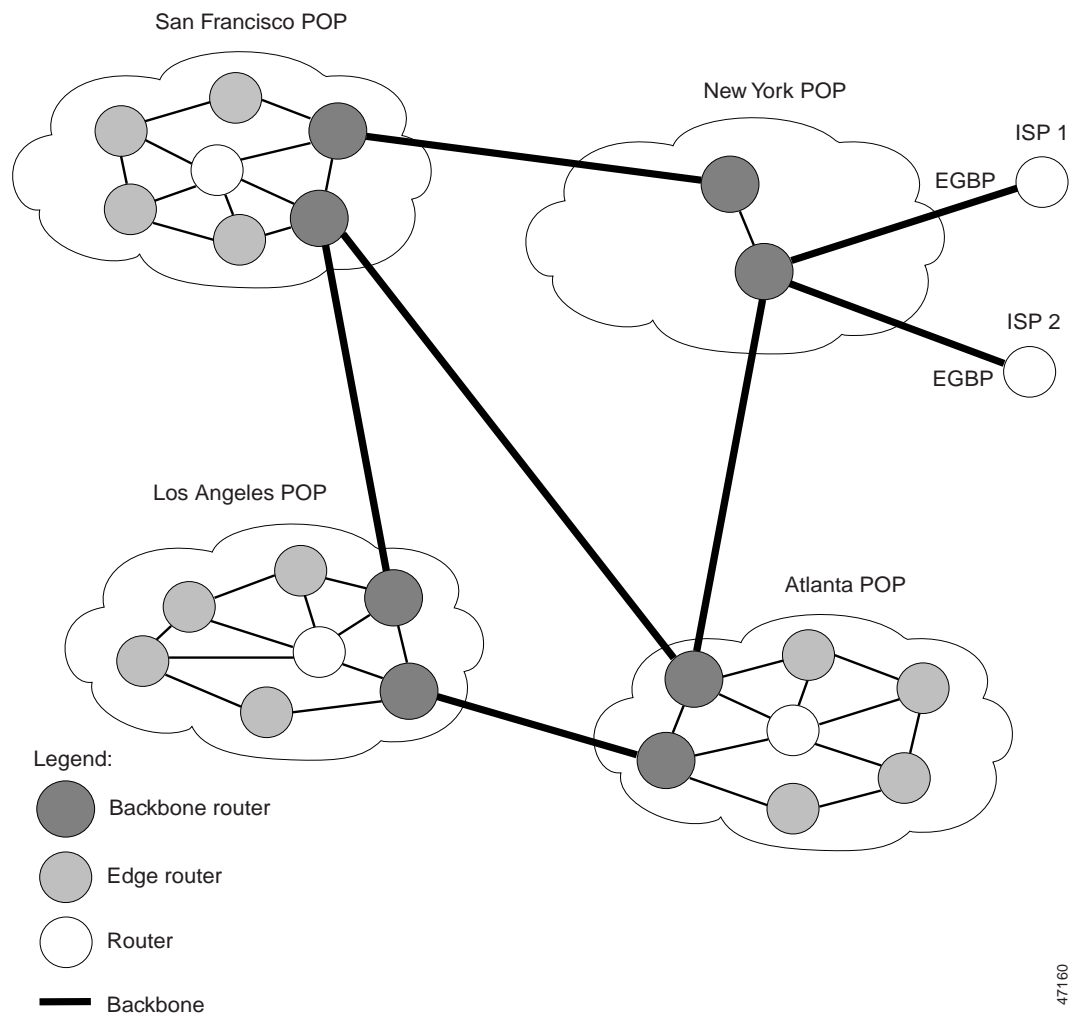
- The neighbor autonomous systems of a BGP destination. You can view the neighbor autonomous systems of a BGP destination by reading the `tmashinfo_ascii` file that resides on the backbone router.

How Backbone Routers Collect TMS Data

By enabling a backbone router to gather traffic matrix statistics, you can determine the amount of traffic that enters the backbone from sites outside of the backbone. You can also determine the amount of traffic that is generated within the backbone. The traffic matrix statistics help you optimize and manage traffic across the backbone.

Figure 11 shows a sample backbone, represented by darkly shaded routers and bold links. The lighter shaded and unshaded routers are outside the backbone. The traffic that travels through the backbone is the area of interest for TMS collection.

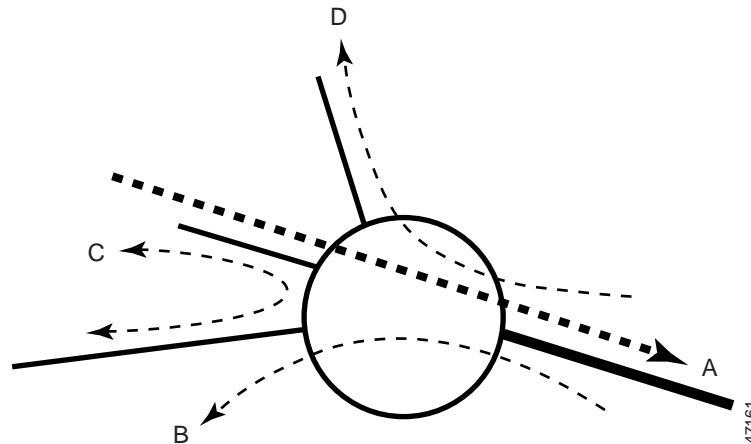
Figure 11 Network Backbone and Routers



47160

Figure 12 shows an exploded view of the backbone router that links the Los Angeles point of presence (POP) in Figure 11 to the Atlanta POP. The bold line represents the backbone link going to the Atlanta POP.

Figure 12 Traffic Traveling Through a Backbone Router



The following types of traffic travel through the backbone router shown in Figure 12:

- The dotted line marked A represents traffic entering the backbone from a router that is not part of the backbone. This is called external traffic.
- The dotted lines marked B and D represent traffic that is exiting the backbone. The router interprets traffic from paths B and D as being generated from within the backbone. This is called internal traffic.
- The dotted line marked C represents traffic that is not using the backbone and is not of interest to TMS.

You can determine the amount of traffic the backbone handles by enabling a backbone router to track the number of packets and bytes that travel through it. You can separate the traffic into the categories “internal” and “external.” You separate the traffic by designating incoming interfaces on the backbone router as internal or external.

Once you enable a backbone router to collect traffic matrix statistics, it starts free running counters, which dynamically update when network traffic passes through the backbone router. You can retrieve a snapshot of the traffic matrix statistics, either through a command to the backbone router or through the NDA.

External traffic (path A) is the most important for determining the amount of traffic. Internal traffic (paths B and D) is useful for ensuring that you are capturing all the TMS data. When you receive a snapshot of the traffic matrix statistics, the packets and bytes are displayed in internal and external categories.

Viewing the TMS Data

Once TMS data is collected, you have the following options for viewing the data:

- Viewing the data in a graphical format, using the NDA Display module. The Display module is useful for graphing the traffic matrix data and comparing statistics. See the section “Viewing the TMS Data Through the NDA” for more information.
- Entering the **more system:vfiles/tmstats_ascii** EXEC command on the backbone router. This command displays a table of traffic matrix statistics. See the section “Viewing the TMS Data by Reading the Virtual Files that Reside on the Backbone Router” for more information.
- Entering the **show ip cef** EXEC command on the backbone router. This command displays nonrecursive accounting data for the backbone router. Included in the output is the number of packets and bytes of internal and external traffic that have been collected. See the section “Viewing TMS Data Through the show ip cef Command” for more information.

Viewing the TMS Data Through the NDA

The Network Data Analyzer collects TMS data from the backbone router and displays it using the NDA Display module. The TMS data can look similar to the data shown in [Figure 13](#) and [Figure 14](#). The display format depends on the aggregation scheme you selected. Refer to the *Network Data Analyzer Installation and User Guide* for more information.

(The NDA Display module is wide. You must slide the scroll bar to the right and left to see all of the data. [Figure 13](#) and [Figure 14](#) taken together show all the columns of data.)

Figure 13 Displaying TMS Data Through the NDA (Part 1)

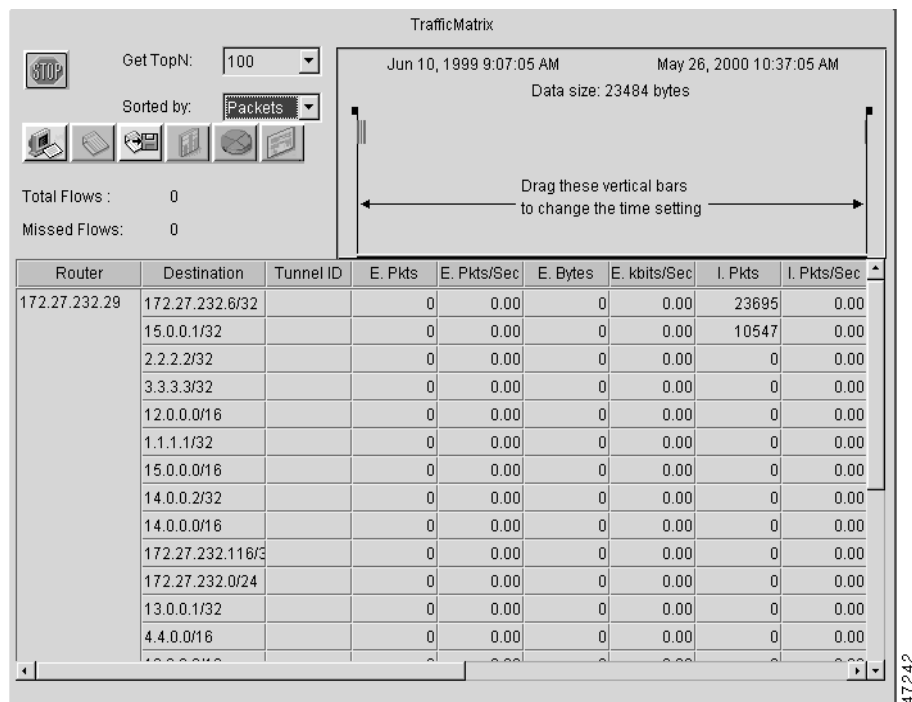
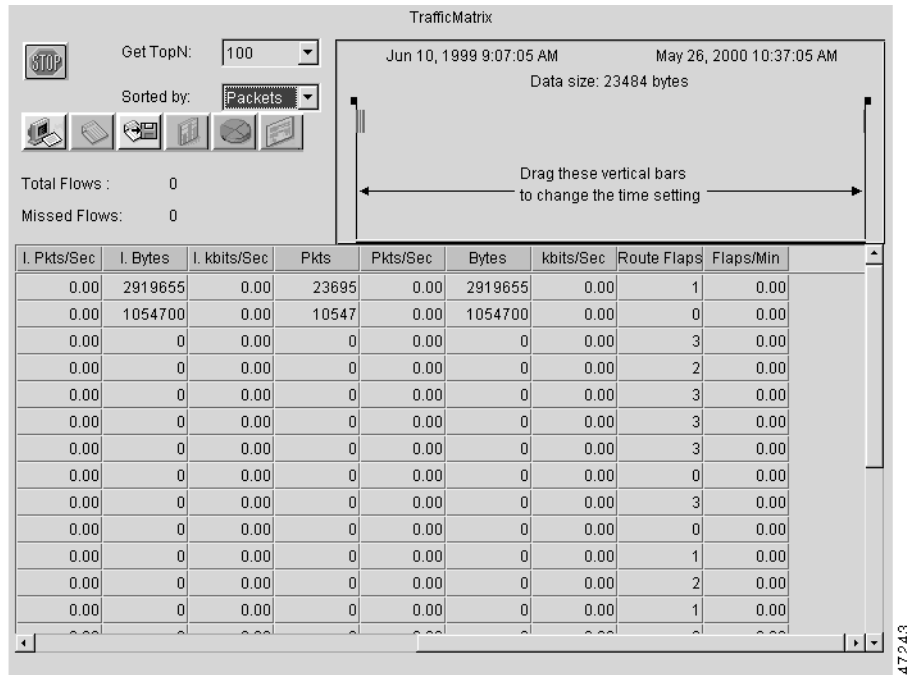


Figure 14 Displaying TMS Data Through the NDA (Part 2)



Viewing the TMS Data by Reading the Virtual Files That Reside on the Backbone Router

You can read the TMS data that resides on the backbone router and is stored in the following virtual files:

- `tmstats_ascii`—TMS data in ASCII (human readable) format.
- `tmstats_binary`—TMS data in binary (space-efficient) format.

Reading the ASCII File

To view statistics in the ASCII file, enter the following command on the backbone router:

```
Router# more system:/vfiles/tmstats_ascii
```

Each file displayed consists of header information and records. A line of space follows the header and each record. A bar (|) separates consecutive fields within a header or record. The first field in a record specifies the type of record. The following example shows a sample TMSTATS_ASCII file:

```
VERSION 1|ADDR 172.27.32.24|AGGREGATION TrafficMatrix.ascii|SYSUPTIME 41428|routerUTC
3104467160|NTP unsynchronized|DURATION 1|
p|10.1.0.0/16|242|1|50|2|100
p|172.27.32.0/22|242|0|0|0|0
```

The following sections describe the header and the various types of records you can display.

File Header

The ASCII file header provides the address of the backbone router and information about how much time the router used to collect and export the TMS data. The header occupies one line and uses the following format:

```
VERSION 1|ADDR<address>|AGGREGATIONTrafficMatrix.ascii|SYSUPTIME<seconds>|
routerUTC<routerUTC>|NTP<synchronized|unsynchronized>|DURATION<aggregateTime>|
```

Table 5 describes the fields in the file header of the TMSTATS_ASCII file.

Table 5 TMSTATS_ASCII File Header

Maximum Field Length	Field	Description
10	VERSION	File format version.
21	ADDR	The IP address of the router.
32	AGGREGATION	The type of data being aggregated.
21	SYSUPTIME	The time of export (in seconds) since the router booted.
21	routerUTC	The time of export (in seconds) since 1900-01-01 (Coordinated Universal Time (UTC)), as determined by the router.
19	NTP	Whether Coordinated Universal Time (UTC) of the router has been synchronized by the Network Time Protocol (NTP).
20	DURATION	The time needed to capture the data (in seconds).

Destination Prefix Record

The destination prefix record displays the internal and external packets and bytes for the IGP route and uses the following format:

```
p|<destPrefix/Mask>|<creationSysUpTime>|<internalPackets>|
<internalBytes>|<externalPackets>|<externalBytes>
```

Table 6 describes the fields in the destination prefix record.

Table 6 Destination Prefix Record Fields

Maximum Field Length	Field	Description
2	<recordType>	p means that the record represents dynamic label switching data or traffic engineered (TE) tunnel traffic data.
19	destPrefix/Mask	The IP prefix address/mask (a.b.c.d/len format) for this IGP route.
11	creationSysUpTime	The sysUpTime when the record was first created.
21	internalPackets	Internal packet count.
21	internalBytes	Internal byte count.
21	externalPackets	External packet count.
20	externalBytes	External byte count (no trailing).

Tunnel Midpoint Record

The tunnel midpoint record displays the internal and external packets and bytes for the tunnel head and uses the following format:

```
t|<headAddr><tun_id>|<creationSysUpTime>|
<internalPackets>|<internalBytes>|<externalPackets>|<externalBytes>
```

Table 7 describes the fields in the tunnel midpoint record.

Table 7 Tunnel Midpoint Record Fields

Maximum Field Length	Field	Description
2	<recordType>	t means that the record represents TE tunnel midpoint data.
27	headAddr<space>tun_id	The IP address of the tunnel head and tunnel interface number.
11	creationSysUpTime	The sysUpTime when the record was first created.
21	internalPackets	Internal packet count.
21	internalBytes	Internal byte count.
21	externalPackets	External packet count.
20	externalBytes	External byte count (no trailing).

Reading the Binary File

The binary file `tmstats_binary` contains the same information as the ASCII file, except in a space-efficient format. You can copy this file from the router and read it with any utility that accepts files in binary format.

Viewing TMS Data Through the `show ip cef` Command

You can use the **show ip cef EXEC** command to display nonrecursive accounting information, including the internal and external packets and bytes that have traveled through the IP prefix address/mask (a.b.c.d/len format) for an IGP route.

```
router# show ip cef 192.168.1.8

192.168.1.8/32, version 220, per-destination sharing
0 packets, 0 bytes
tag information set
local tag:17
via 192.168.67.8, FastEthernet6/0, 0 dependencies
next hop 192.168.67.8, FastEthernet6/0
valid adjacency
tag rewrite with Fa6/0, 192.168.67.8, tags imposed {}
1143 packets, 56702 bytes switched through the prefix
30 second output rate 0 Kbits/sec
tmstats:external 0 packets, 0 bytes
internal 1144 packets, 56742 bytes
```

Viewing the BGP Neighbor Autonomous Systems

The TMS feature also displays the BGP neighbor autonomous system (AS) associated with each IGP destination. You can display all the neighbor autonomous systems for any IGP destination.

The `tmasinfo` file is in the ASCII format, which is the only one provided for this data. Enter the following command to read the `tmasinfo` file:

```
Router# more system:/vfiles/tmasinfo
```

Each file consists of header information and a number of records. A line of space follows the header and each record. A bar (|) separates consecutive fields within a header or a record.

Header Format

The file header provides the address of the router and indicates how much time the router used to collect and export the data. The file header uses the following format:

```
VERSION 1|ADDR<address>|AGGREGATION ASList.ascii|SYSUPTIME<seconds>|routerUTC
<routerUTC>|DURATION<aggregateTime>|
```

Table 8 describes the fields in the file header.

Table 8 *TMASINFO File Header*

Max. Length	Field	Description
5	VERSION	File format version.
15	ADDR	The IP address of the router.
20	AGGREGATION	The type of data being aggregated.
10	SYSUPTIME	The time of export (in seconds) since router booted.
10	routerUTC	The time of export (in seconds) since 1900-01-01, as determined by the router.
10	DURATION	The time needed to capture the data (in seconds).

Neighbor Autonomous System Record

The neighbor autonomous system record displays the neighbor autonomous system and the underlying prefix/mask for each BGP route. The record uses the following format:

```
<nonrecursivePrefix/Mask>|<AS>|<destinationPrefix/Mask>
```

Table 9 describes the fields in the neighbor autonomous system record.

Table 9 *Neighbor Autonomous System Record Fields*

Maximum Field Length	Field	Description
18	nonrecursivePrefix/Mask	The IP prefix address/mask (a.b.c.d/len format) for this IGP route.
5	AS	The neighbor autonomous system.
18	destinationPrefix/Mask	The prefix/mask for the FIB entry (typically BGP route).

Network Services Engine

The network services engine (NSE) is a processor engine for Cisco series routers. The NSE delivers wire rate OC-3 throughput while running concurrent high-touch WAN edge services. The NSE takes advantage of a new technology called Parallel eXpress Forwarding (PXF).



Note

Before enabling the PXF processor, you must have IP routing and IP CEF switching turned on.

For information on configuring NSE, see the “[Cisco Express Forwarding Overview](#)” chapter later in this publication.

Network Services Engine benefits and requirements are as follows:

- Accelerated services—The following features are accelerated on the NSE: Network Address Translation (NAT), weighted fair queueing (WFQ), and NetFlow for both enterprise and service provider customers.
- PXF field upgradable—PXF is based on microcode and can be upgraded with new software features in future Cisco IOS releases.

The PXF processor enables IP parallel processing functions that work with the primary processor to provide accelerated IP Layer 3 feature processing. The PXF processor off-loads IP packet processing and switching functions from the RP to provide accelerated and highly consistent switching performance when coupled with one or more of several IP services features such as access Control Lists (ACLs), address translation, quality of service (QoS), flow accounting, and traffic shaping.

PXF offers the advantage of hardware-based switching power, plus the flexibility of a programmable architecture. The PXF architecture provides future-proofing—if additional features are added, an application-specific integrated circuit (ASIC) will not be required. New features for accelerated services can be added by reprogramming the PXF processor.

- System requirements—An NSE-1 can be used on existing Cisco 7200 VXR series routers with Cisco Release IOS 12.1(1)E or a later version of Cisco IOS Release 12.1 E, and with Cisco IOS Release 12.1(5)T or a later version of Cisco IOS Release 12.1 T.
- High performance—Network-layer services such as traffic management, security, and QoS benefit significantly from NSE-1 high-performance. NSE-1 is the first Cisco processing engine to offer integrated hardware acceleration, increasing Cisco 7200 VXR series system performance by 50 to 300 percent for combined “high-touch” WAN edge services.

Virtual Profile CEF

The Virtual Profile CEF feature allows you to enable asynchronous and ISDN interfaces in CEF switching. This feature allows you to create a datagram prefix and cache it in an adjacency table for fast reference and rewrite during the call setup. For information on configuring the Virtual Profile CEF feature, see the “[Configuring Cisco Express Forwarding](#)” chapter later in this publication.

Virtual Profile CEF benefits are as follows:

- FIB—Virtual Profile (VP) CEF switching allows the user to use the FIB to look up a route for a forwarding packet. Because the FIB is populated by routing topology, not by traffic, the FIB is a performance enhancement over cache tables used in fast switching.
- MPLS VPN/BGP integration—VP CEF switching enables VP to be used in other technologies that require CEF switching, such as MPLS Virtual Private Network/Border Gateway Protocol (VPN/BGP).
- ISDN interfaces—VP CEF allows you to enable ISDN interfaces in CEF switching.