Cisco Ultra Traffic Optimization with VPP

Feature Summary and Revision History

Summary Data

Table 1: Summary Data

Applicable Product(s) or Functional Area	5G-UPF
Applicable Platform(s)	VPC-SI
Feature Default Setting	Disabled - Configuration Required
Related Changes in this Release	Not Applicable
Related Documentation	Not Applicable

Revision History

Table 2: Revision History

Revision Details	Release
First introduced.	2022.01.1

Feature Description

The UPF supports Cisco Ultra Traffic Optimization (CUTO) on Vector Packet Processing (VPP).

The Cisco Ultra Traffic Optimization is a RAN optimization technology that increases the subscriber connection speeds in congested cells and, as a result, increases the cell capacity significantly. The result is an optimized RAN, where Mobile Network Operators (MNOs) can deploy fewer cells, on an ongoing basis, and absorb more traffic growth while meeting network quality targets.

Large traffic flows, such as Adaptive Bit Rate (ABR) video, saturate radio resources and swamp the eNodeB scheduler. The Cisco Ultra Traffic Optimization employs machine learning algorithms to detect large traffic flows (such as video) in the network. It also optimizes the Delivery of those flows to mitigate the network congestion without changing the user quality (that is, video works the same for you). In other words, by employing software intelligence at the network core, Cisco Ultra Traffic Optimization mitigates the overwhelming impact the video has on the RAN.

The resulting benefits are seen in congested network sites. The Cisco Ultra Traffic Optimization:

- Increases average user throughput.
- Increases congested cell site capacity.
- Reduces scheduler latency.
- Maintains user quality of experience even when more users and more traffic share a cell.
- Is measured directly by eNodeB performance counters (for example, average UE throughput, scheduler latency). These are the key performance indicators that are used for network capacity planning.
- Provides permanent savings in RAN investment requirements.
- Is integrated in the Cisco StarOS P-GW.
- Requires no new hardware or cabling complexity - it can be turned on for a market in an hour.
- Supports HTTP or HTTPS, and QUIC traffic.

Licensing

The Cisco Ultra Traffic Optimization with VPP is a licensed Cisco solution. Contact your Cisco account representative for detailed information on specific licensing requirements. For information on installing and verifying licenses, refer to the *Managing License Keys* section of the *Software Management Operations* chapter in the *VPC-SI System Administration Guide*.

RCM Support

This feature enables the Redundancy and Configuration Management (RCM) support for the Cisco Ultra Traffic Optimization (CUTO). All relevant configuration to enable CUTO using service scheme and application of the CUTO profile or policy on UPF is supported using RCM.

Sending the GBR or MBR Values to Cisco Ultra Traffic Optimization

If the flow level MBR is greater than the APN-AMBR for a non GBR bearer, traffic is throttled at APN-AMBR. In such a case APN-AMBR is sent as the upper limit to the CUTO library. If there is no valid flow level MBR specific to the flow, APN-AMBR is sent as the upper limit to the CUTO library.

For a GBR bearer, the flow level GBR is sent as the lower limit and flow level MBR is sent as the upper limit to the CUTO library.

Cisco Ultra Traffic Optimization Library Deinitialization

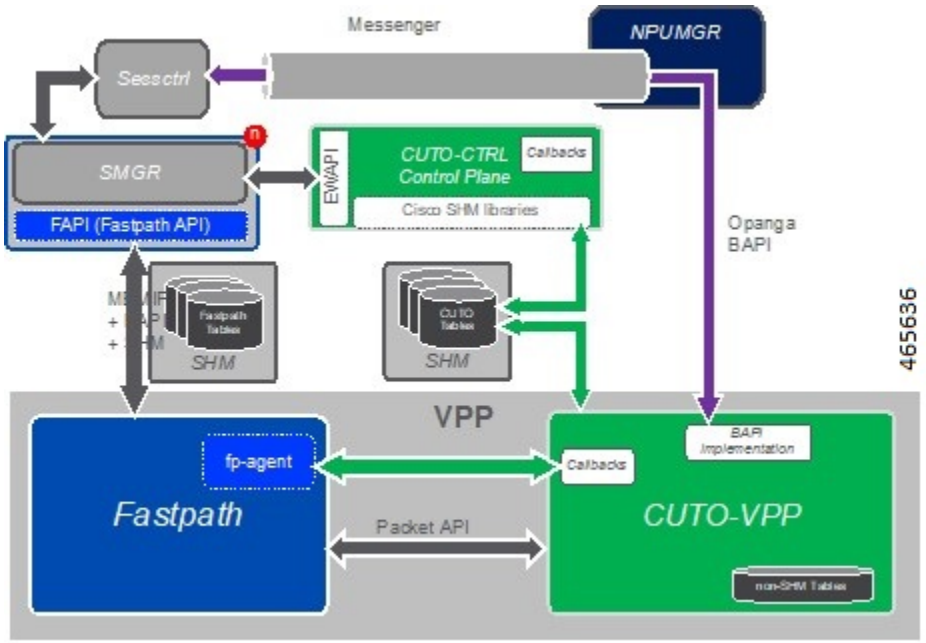
This feature currently doesn't support the Deinitialization. Deinitialization happens when the Cisco Ultra Traffic Optimization (CUTO) license is removed from the system.

How it Works

Architecture

The following figure illustrates the architecture of Cisco Ultra Traffic Optimization on VPP.

Figure 1: Architecture



CUTO-CTRL

- CUTO-CTRL receives guidance and requests from SMGR through the East-West API (EWAPI), through which clients (SMGR instances) are registered and deregistered, and new streams or flows are created and terminated.
- CUTO-CTRL manages a set of shared memory (SHM) tables using a North-South API (NSAPI) consisting of Cisco-provided SHM infrastructure.
- It is through this SHM environment that CUTO-VPP can read and write content that is visible to both CUTO-VPP and CUTO-CTRL.
- The SHM is used for all high volume, scalable/mutable content necessary for the high-performance configuration and administration of the CUTO solution in VPP.

NPUMGR

NPUMGR is the management layer that is responsible for the overall VPP operation. It sends Binary API (BAPI) requests to CUTO-VPP for initialization, global runtime configuration, and policy configuration.

SMGR

SMGR is the main subscriber control plane. There are N SMGR instances, and all instances are managed by the SessCtrl. In the context of VPPMOB/Fastpath, the SMGR instances are also known as “Clients”, and each client has a unique ID.

SMGR issues policy directives to SessCtrl through the Messenger tunnel, and sends updated Policy guidance to CUTO-VPP through the Binary API.

SMGR communicates with Fastpath for pre-existing functionality with a set of MEMIF, Binary API, and shared memory (SHM) infrastructures.

Session Control (SessCtrl)

Session Control is the management layer responsible for overseeing the set of SMGR instances.

The BAPI requests are tunnelled from SessCtrl to NPUMGR through Messenger.

Fastpath

VPP is responsible for packet processing. Fastpath performs subscriber-related packet processing within the VPP environment. Subscriber flows are divided into unidirectional Streams, and a Stream conduit is the pipeline of functions through which a packet is transformed and egressed from subscriber processing.

A packet API between the Fastpath and CUTO-VPP facilitates the exchange of packets traversing the Fastpath conduit.

CUTO-VPP

- CUTO-VPP is the packet processing engine in the UPF.
- In fastpath, Cisco Ultra Traffic Optimization is applied to packets on a stream configured with its operation.
- Packets are sent from the Stream conduit to a particular CUTO-VPP operation, and after some potential delay (0-N milliseconds), traffic is returned to the same Conduit.
- Packets are never dropped by the Cisco Ultra Traffic optimization library.

CUTO-TODR

Traffic Optimization Data Records (TODR) can only be generated as events, and are enabled only when the configuration is available.

Limitations

The Cisco Ultra Traffic Optimization feature has the following limitations:

- CUTO configuration changes done in Service Schema do not take effect immediately for existing flows.
- Cisco Ultra Traffic Optimization VPP global deinitialization is not supported.
- Bearer-related triggers for enabling Cisco Ultra Traffic Optimization are not supported.
- Rule match change trigger must be configured for CUTO in UPF.
- Enabling/Disabling of Traffic optimization is not supported on "loc-update" trigger.
- Removal of CUTO license doesn't trigger global deinitialization. CUTO configurations must be removed to disengage CUTO functionality for new flows.

Show Commands and Outputs

This section provides information regarding show commands and their outputs in support of Cisco Ultra Traffic Optimization.

For information on other supporting show commands, refer to *Monitoring and Troubleshooting* section under the *Cisco Ultra Traffic Optimization* chapter in the *P-GW Administration Guide*.

Show Commands and Outputs

show user-plane-service traffic-optimization counters sessmgr all

The output of this command includes the following fields:

TCP Traffic Optimization Flows:

- Active Normal Flow Count
- Active Large Flow Count
- Active Managed Large Flow Count
- Active Unmanaged Large Flow Count
- Total Normal Flow Count
- Total Large Flow Count
- Total Managed Large Flow Count
- Total Unmanaged Large Flow Count
- Total IO Bytes

- Total Large Flow Bytes
- Total Recovered Capacity Bytes
- Total Recovered Capacity ms

UDP Traffic Optimization Flows:

- Active Normal Flow Count
- Active Large Flow Count
- Active Managed Large Flow Count
- Active Unmanaged Large Flow Count
- Total Normal Flow Count
- Total Large Flow Count
- Total Managed Large Flow Count
- Total Unmanaged Large Flow Count
- Total IO Bytes
- Total Large Flow Bytes
- Total Recovered Capacity Bytes
- Total Recovered Capacity ms

show user-plane-service traffic-optimization info

The output of this command includes the following fields:

- CUTO Ctrl Library Version
- CUTO VPP Library Version
- Mode
- Configuration
  - Data Records (TODR)
  - Statistics Options
  - EFD Flow Cleanup Interval
  - Statistics Interval

show user-plane-service traffic-optimization policy all

The output of this command includes the following fields:

- Policy Name
- Policy-Id

- Bandwidth-Mgmt
  - Backoff-Profile
  - Min-Effective-Rate
  - Min-Flow-Control-Rate
- Curbing-Control:
  - Time
  - Rate
  - Max-Phases
  - Threshold-Rate
- Heavy-Session:
  - Threshold
  - Standard-Flow-Timeout
  - Seed-Time
- Detection-Mode
- Link-Profile:
  - Initial-Rate
  - Max-Rate
  - Peak-Lock
- Session-Params:
  - Tcp-Ramp-Up
  - Udp-Ramp-Up
- Total traffic-optimization-policies found

Bulkstats

The following existing bulk statistics are supported by Cisco Ultra Traffic Optimization in UPF:

Bulk Statistics	Description
cuto-uplink-drop	Indicates the total number of uplink packets dropped by CUTO library
cuto-uplink-hold	Indicates the total number of uplink packets held by CUTO library
cuto-uplink-forward	Indicates the total number of uplink packets forwarded by CUTO library

Bulk Statistics	Description
cuto-uplink-rx	Indicates the total number of uplink packets received by CUTO library
cuto-uplink-tx	Indicates the total number of uplink packets sent by CUTO library
cuto-dnlink-drop	Indicates the total number of downlink packets dropped by CUTO library
cuto-dnlink-hold	Indicates the total number of downlink packets held by CUTO library
cuto-dnlink-forward	Indicates the total number of downlink packets forwarded by CUTO library
cuto-dnlink-rx	Indicates the total number of downlink packets received by CUTO library
cuto-dnlink-tx	Indicates the total number of downlink packets sent by CUTO library
cuto-todrs-generated	Indicates the total number of TODRs generated.
tcp-active-normal-flow-count	Indicates the number of TCP active-normal-flow count for Cisco Ultra Traffic Optimization.
tcp-active-large-flow-count	Indicates the number of TCP active-large-flow count for Cisco Ultra Traffic Optimization.
tcp-active-managed-large-flow-count	Indicates the number of TCP active-managed-large-flow count for Cisco Ultra Traffic Optimization.
tcp-active-unmanaged-large-flow-count	Indicates the number of TCP active-unmanaged-large-flow count for Cisco Ultra Traffic Optimization.
tcp-total-normal-flow-count	Indicates the number of TCP total-normal-flow count for Cisco Ultra Traffic Optimization.
tcp-total-large-flow-count	Indicates the number of TCP total-large-flow count for Cisco Ultra Traffic Optimization.
tcp-total-managed-large-flow-count	Indicates the number of TCP total-managed-large-flow count for Cisco Ultra Traffic Optimization.
tcp-total-unmanaged-large-flow-count	Indicates the number of TCP total-unmanaged-large-flow count for Cisco Ultra Traffic Optimization.
tcp-total-io-bytes	Indicates the number of TCP total-IO bytes for Cisco Ultra Traffic Optimization.
tcp-total-large-flow-bytes	Indicates the number of TCP total-large-flow bytes for Cisco Ultra Traffic Optimization.

Bulk Statistics	Description
tcp-total-recovered-capacity-bytes	Indicates the number of TCP total-recovered capacity bytes for Cisco Ultra Traffic Optimization.
tcp-total-recovered-capacity-ms	Indicates the number of TCP total-recovered capacity ms for Cisco Ultra Traffic Optimization.
udp-active-normal-flow-count	Indicates the number of UDP active-normal-flow count for Cisco Ultra Traffic Optimization.
udp-active-large-flow-count	Indicates the number of UDP active-large-flow count for Cisco Ultra Traffic Optimization.
udp-active-managed-large-flow-count	Indicates the number of UDP active-managed-large-flow count for Cisco Ultra Traffic Optimization.
udp-active-unmanaged-large-flow-count	Indicates the number of UDP active-unmanaged-large-flow count for Cisco Ultra Traffic Optimization.
udp-total-normal-flow-count	Indicates the number of UDP total-normal-flow count for Cisco Ultra Traffic Optimization.
udp-total-large-flow-count	Indicates the number of UDP total-large-flow count for Cisco Ultra Traffic Optimization.
udp-total-managed-large-flow-count	Indicates the number of UDP total-managed-large-flow count for Cisco Ultra Traffic Optimization.
udp-total-unmanaged-large-flow-count	Indicates the number of UDP total-unmanaged-large-flow count for Cisco Ultra Traffic Optimization.
udp-total-io-bytes	Indicates the number of UDP total-IO bytes for Cisco Ultra Traffic Optimization.
udp-total-large-flow-bytes	Indicates the number of UDP total-large-flow bytes for Cisco Ultra Traffic Optimization.
udp-total-recovered-capacity-bytes	Indicates the number of UDP total-recovered capacity bytes for Cisco Ultra Traffic Optimization.
udp-total-recovered-capacity-ms	Indicates the number of UDP total-recovered capacity ms for Cisco Ultra Traffic Optimization.

Sample Configuration

Sample configuration to enable the CUTO feature:

```
configure
  active-charging service ACS
    trigger-action TA1
       traffic-optimization policy custom1
    #exit
    trigger-condition TC1
```

```
      rule-name = dynamic-rule2
    #exit
    service-scheme SS1
      trigger rule-match-change
        priority 5 trigger-condition TC1 trigger-action TA1
      #exit
    subs-class SB1
      rulebase = cisco
    #exit
    subscriber-base default
      priority 5 subs-class SB1 bind service-scheme SS1
    #exit
    traffic-optimization-profile
      mode active
      data-record
    #exit
    traffic-optimization-policy custom1
      bandwidth-mgmt min-effective-rate 800 min-flow-control-rate 250
      heavy-session threshold 200000
      link-profile max-rate 20000
    #exit
    traffic-optimization-policy default
    #exit
end
```

CUTO-TODR

Sample configuration to enable the CUTO-TODR:

```
context ISP
edr-module active-charging-service
file name EDR directory TODR_CUTO rotation volume 51200 headers
cdr use-harddisk
```