Cisco Ultra Traffic Optimization with VPP

Revision History

Note Revision history details are not provided for features introduced before release 21.24.

Revision Details	Release
First introduced	Pre 21.24

Feature Description

Cisco Ultra Traffic Optimization is supported on VPP in the CUPS architecture.

The Cisco Ultra Traffic Optimization is a RAN optimization technology that increases 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 and optimize the delivery of those flows to mitigate the network congestion without changing user quality (that is, video works the same for the end user). In other words, by employing software intelligence at the network core, Cisco Ultra Traffic Optimization mitigates the overwhelming impact 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), which 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(s) 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 *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 the Cisco Ultra Traffic Optimization (CUTO) using service scheme and application of the Cisco Ultra Traffic Optimization (CUTO) profile or policy on User Plane is supported using RCM.

Sending the GBR or MBR Values to Cisco Ultra Traffic Optimization

During the stream create/update, a bearer with valid QER and is GBR bearer, the respective bearer level downlink GBR/MBR values are sent to Cisco Ultra Traffic Optimization (CUTO) library as lower or upper limit values otherwise lower limit or upper limit values are zero. The values of lower limit and upper limit are in Bits Per Second (BPS). Post RCM support, the P-GW sends the downlink flow level GBR and MBR values instead of bearer level GBR and MBR to the optimization library. For GBR bearer, flow level GBR is sent as lower limit and flow level MBR is sent as the upper limit to the Cisco Ultra Traffic Optimization (CUTO) library. For non-GBR bearer 0 is sent as lower limit and flow level MBR is sent as upper limit to the Cisco Ultra Traffic Optimization (CUTO) library. 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 Cisco Ultra Traffic Optimization (CUTO) library. If there is no valid flow level MBR specific to the flow, APN-AMBR is sent as the upper limit to the Cisco Ultra Traffic Optimization (CUTO) library. Optimization library maintains logical flow based on 3-tuple (That is source IP, destination IP and protocol), whereas the non-CUPS architecture considers a flow as 5-tuple (That is source IP, destination IP, source port, destination port and protocol). Hence multiple non-CUPS architecture 5-tuple entries can belong to same

3-tuple entry in optimization library. The PG-W provides GBR and MBR values based on 5-tupple to the optimization library. As part of this feature:

- Optimization library uses the minimum of all MBR values that belong to same 3-tuple entry as upper limit.
- Optimization library uses maximum of all GBR values that belong to same 3-tuple entry as lower limit.

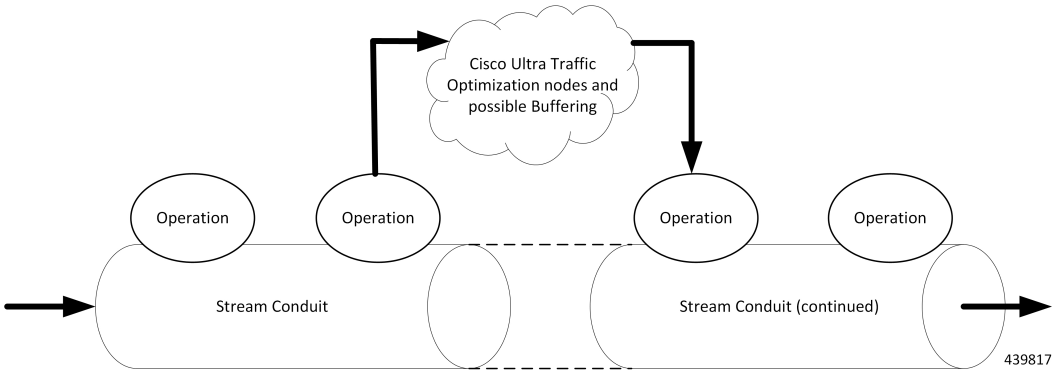
Cisco Ultra Traffic Optimization Library Deinitialization

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 in CUPS.



Cisco Ultra Traffic Optimization is split across Control Plane and User Plane.

CUTO-CTRL

- CUTO-CTRL receives guidance and requests from SMGR through the East-West API (EWAPI), through which clients (SMGR instances) are registered and de-registered, and new streams/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.

CUTO-VPP

- CUTO-VPP is the packet processing engine in the user plane.
- 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 application.

Limitations

The Cisco Ultra Traffic Optimization feature in CUPS 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.
- Dynamic memory allocation between SMGR and CUTO-VPP.
- Bearer-related triggers for enabling Cisco Ultra Traffic Optimization are not supported.
- Rule match change trigger must be configured for CUTO in CUPS.
- Disabling of Traffic optimization is not supported on 'loc-update' trigger.
- Enabling Cisco Ultra Traffic Optimization via Gx is not supported.
- Removal of CUTO license will not 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 in CUPS.

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

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
- Link-Profile:
  - Initial-Rate
  - Max-Rate
  - Peak-Lock
- Session-Params:
  - Tcp-Ramp-Up
  - Udp-Ramp-Up

Bulkstats

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

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
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

Bulk Statistics	Description
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.
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.

Bulk Statistics	Description
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.

The following statistics for Cisco Ultra Traffic Optimization, that are part of the legacy (StarOS) implementation, are not applicable to the CUPS implementation.

- tcp-uplink-drop
- tcp-uplink-hold
- tcp-uplink-forward
- tcp-uplink-forward-and-hold
- tcp-uplink-hold-failed
- tcp-uplink-bw-limit-flow-sent
- tcp-dnlink-drop
- tcp-dnlink-hold

- tcp-dnlink-forward
- tcp-dnlink-forward-and-hold
- tcp-dnlink-hold-failed
- tcp-dnlink-bw-limit-flow-sent
- tcp-dnlink-async-drop
- tcp-dnlink-async-hold
- tcp-dnlink-async-forward
- tcp-dnlink-async-forward-and-hold
- tcp-dnlink-async-hold-failed
- tcp-process-packet-drop
- tcp-process-packet-hold
- tcp-process-packet-forward
- tcp-process-packet-forward-failed
- tcp-process-packet-forward-and-hold
- tcp-process-packet-forward-and-hold-failed
- tcp-pkt-copy
- tcp-pkt-Copy-failed
- tcp-process-pkt-copy
- tcp-process-pkt-copy-failed
- tcp-process-pkt-no-packet-found-action-forward
- tcp-process-pkt-no-packet-found-forward-and-hold
- tcp-process-pkt-no-packet-found-action-drop
- tcp-todrs-generated
- udp-uplink-drop
- udp-uplink-hold
- udp-uplink-forward
- udp-uplink-forward-and-hold
- udp-uplink-hold-failed
- udp-uplink-bw-limit-flow-sent
- udp-dnlink-drop
- udp-dnlink-hold
- udp-dnlink-forward

- udp-dnlink-forward-and-hold
- udp-dnlink-hold-failed
- udp-dnlink-bw-limit-flow-sent
- udp-dnlink-async-drop
- udp-dnlink-async-hold
- udp-dnlink-async-forward
- udp-dnlink-async-forward-and-hold
- udp-dnlink-async-hold-failed
- udp-process-packet-drop
- udp-process-packet-hold
- udp-process-packet-forward
- udp-process-packet-forward-failed
- udp-process-packet-forward-and-hold
- udp-process-packet-forward-and-hold-failed
- udp-pkt-copy
- udp-pkt-Copy-failed
- udp-process-pkt-copy
- udp-process-pkt-copy-failed
- udp-process-pkt-no-packet-found-action-forward
- udp-process-pkt-no-packet-found-forward-and-hold
- udp-process-pkt-no-packet-found-action-drop
- udp-todrs-generated

Sample Configuration

Sample configuration to enable CUPS 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 300 min-flow-control-rate 150
      heavy-session threshold 200000
      link-profile max-rate 20000
    #exit
    traffic-optimization-policy default
    #exit
end
```