Verizon LTE Mobile Private Network
Cisco Integrated Services Services Router

Router Configuration Guide for
Private Network Traffic Management (LTE QoS)
on Verizon Wireless MPN

Revision 1.1

March 2018
Introduction

Verizon offers Private Network Traffic Management (PNTM), LTE Quality of Service for the Verizon 4G LTE Mobile Private Network service (MPN). MPN provides private last-mile access to an organization's internal network via 4G LTE or 3G cellular wireless. With MPN, enterprise traffic over Verizon LTE is not exposed to the public Internet. The router's LTE IP address and LAN addresses are part of the organization's private address space, allowing for native routing and static IP addressing for manageability. Verizon MPN can be used by organizations in two ways:

- Verizon Wireless Dynamic Mobile Network Routing (DMNR) is a network-based, mobile technology providing dynamic routing to ISRs in primary wireless access or automatic wireless backup configurations. DMNR integrates wireless and wireline networks by leveraging Mobile IP Network Mobility (NeMo) protocol. DMNR provides end-to-end routing without the need for prem-based overlay tunnels/VPNs.

- MPN also supports the use of prem-based end-to-end VPNs, such as GRE, IPsec and DMVPN.

QoS configuration for LTE is different than for wireline last mile access. This is due to the differences between wireline and today's cellular wireless technology. Wireline last mile can offer a consistent bandwidth/bit rate and a committed information rate. LTE as last mile is a shared medium; the attained bandwidth varies based on congestion (devices in the cell radio sector) and radio signal quality and power. As LTE does not offer guaranteed access for user data, QoS is based on providing enhanced access to radio resources.

LTE’s advantages in provisioning time, flexible deployment, on-the-move ability, and cost/performance characteristics make it a fast growing option for last mile backup and primary access. Organizations today use best effort LTE service and deploy ISRs with QoS configurations to prioritize important traffic over less important traffic. With LTE QoS, performance and response time for important applications can be improved.

Below is a figure depicting Verizon MPN with DMNR and Private Network Traffic Management (PNTM).
4G LTE technology was originally designed to support single-user devices such as cell phones and tablets. LTE QoS is based on a traffic flow template pushed to the LTE modem from the network. Historically, the flow template is a 5-tuple (source and destination IP addresses, protocol, source and destination port numbers). A use case was streaming of pay-per-view video content. This method of classification by the modem is not scalable for network-based multi-user deployment (Routers with LTE as WAN).

Verizon and Cisco have partnered to leverage LTE technology for use as a last mile Wide Area Network access with Private Network Traffic Management. The extensions are deployment options, network enhancements, and ISR enhancements, which meet today’s LTE open 3GPP standards. Below are the key characteristics:

- Verizon offers two bearer channels across an LTE Private Network connection. Both bearers are part of a single logical connection via a single ISR interface and IP address.
- Traffic is classified and placed onto the appropriate LTE bearer based on the ToS byte of an IP packet (e.g. DSCP value). This allows integration with an organization’s existing QoS model per generally accepted practices (mark DSCP close to the source and treat traffic across the network based on DSCP in each packet). Multiple classes can be mapped to the bearer offering enhanced QoS.
- The PNTM bearers include the dedicated bearer (supports all packets marked DSCP CS3, AF3x, CS4, AF4x, CS5, EF), and the default bearer (supports all other packets). The default bearer is provided best effort service up to line rate, while the dedicated bearer is provided prefential treatment (Mission Critical CoS) up to the service rate (available at 512Kbps or 2Mbps). Note that traffic above the service rate in the dedicated bearer is dropped by the network.

Below is a diagram depicting the Verizon PNTM service
Today’s LTE cellular modems do not provide the granularity in congestion feedback that wireline interfaces offer. Enterprise QoS traffic congestion control and prioritization mechanisms are engaged when there is congestion (too many packets in an egress queue). The current lack of granular notification from the LTE modem to the device (e.g. a router interface) is handled in two ways with Cisco ISRs

- The ISR 4200, 4300 and 4400 series uses an advanced multi-gigabit internal fabric for connection between its separate data plane cores, Network Interface Modules (NIM) and Service Modules (SM-X). The architecture allows for indirect detection of LTE modem congestion and congestion notification to IOS-XE for a cellular interface. This allows for a dynamic QoS policy, providing the maximum bandwidth from an LTE connection with QoS. Configuration is detailed in the “ISR 4200, 4300, 4400” section further below.
  - The ISR 1100 series uses a similar architecture to the ISR 4000 series and also allows for dynamic QoS policy, providing the maximum bandwidth from an LTE connection with QoS. The configuration is the same as the ISR 4000 series (with cellular interface number of 0/2/x).

- For 1900, 2900, 3900, 8x9 series, deploying a method to generate the congestion notification allows QoS mechanisms to be engaged. A configuration method to enable this has been developed by Cisco and Verizon, and is explained in the “ISR G2” section further below. Note that 1900, 2900 and 3900 models are end-of-sale, as of March 2018.
**LTE General QoS Planning Guidelines**

- The ISR prerequisite for PNTM on Verizon 4G LTE MPN is an ISR with embedded Verizon LTE interface running recommended IOS and modem firmware releases. All other Verizon and Cisco recommendations for MPN and DMNR are also applicable. The following guide provides the specifics.


- The physical deployment prerequisite is a site survey, which can be done before or after installation. Antenna placement and type can make a significant difference in LTE throughput. ISRs provide details on LTE signal strength and quality. The details change within a minute of moving the antennas. The values can be seen from the ISR console, or by text message from a cell phone using this no charge application [https://supportforums.cisco.com/t5/network-infrastructure-documents/command-over-sms/ta-p/3340391](https://supportforums.cisco.com/t5/network-infrastructure-documents/command-over-sms/ta-p/3340391) documented here


- Mark packets as close to the source as can be trusted to differentiate traffic into sub-classes to be used in the third level policy. Methods for marking are provided in Cisco best practices QoS guides such as

  [link](https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Campus/Borderless_Campus_Network_1-0/Borderless_Campus_1-0_Design_Guide/BN_Campus_QoS.html#wp1232603)


- Use a CS5/EF class for voice traffic and/or the most critical traffic (second level policy). It is recommended to set this traffic class to less than or equal to 25% of the service rate (either 0.5Mbps or 2Mbps, i.e. no greater than 125Kbps or 512Kbps). For voice traffic, best practices for call admission control should be deployed to limit the number of simultaneous calls and ensure sufficient call quality.

- Use CS3/AF3x/CS4/AF4x classes for important data or video applications (second level policy).

- The shaped rate for the priority class should be set to the service rate (0.5Mbps or 2Mbps).

- Multiple classes can be defined in the 2nd level policy, providing granular minimum bandwidth to multiple classes of important traffic (e.g. AF31, AF32, AF33, AF41, AF42, AF43, CS3, CS4).

- There may be a desire to not limit important traffic to 2Mbps. As that traffic class/bearer will be policed in the network, there is an option for the traffic class to police instead of shape. The policer on the ISR can be configured to re-mark DSCP in packets exceeding the service rate, instead of dropping them. The packets over the service rate would then be carried in the default bearer. The important but over-contract traffic can still gain better treatment than traffic with DSCP 0, via the “random-detect dscp-based” treatment defined under “class-default” in the recommended configurations below. This can be done by remarking the over-contract important traffic with a DSCP value of CS2 for example. An example of this method: edit the 2nd level as follows:

  - Replace “shape average 2000000”
    with “police 2000000 conform-action transmit exceed-action set-dscp-transmit CS2”

- QoS configuration for 19xx/29xx/39xx, 8x9, 42xx/43xx/44xx and 11xx series are the same except for interface names/numbers.
Voice Considerations: To support voice, the bandwidth needed to support the maximum number of simultaneous calls must be calculated. This bandwidth need is the value for the strict priority bit rate defined in the 2nd level policy's voice traffic class. As voice packets are small in size, packet overhead has a greater effect on the bandwidth required for a voice call. To determine the total bandwidth needed in the strict priority queue, the amount of bandwidth per call is multiplied by the maximum number of simultaneous calls.

The bandwidth per call is determined by taking the codec type and IP overhead into account. Two popular codecs are G.711 and G.729. The bandwidth required for various scenarios is shown below.

<table>
<thead>
<tr>
<th>Codec</th>
<th>LAN</th>
<th>LTE, MPN + DMNR</th>
<th>LTE, MPN + DMNR + GETVPN</th>
<th>LTE, MPN + DMVPN</th>
<th>LTE, MPN + IPsec</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.711</td>
<td>87 kbps</td>
<td>96 kbps</td>
<td>120 kbps</td>
<td>131 kbps</td>
<td>125 kbps</td>
</tr>
<tr>
<td>G.729</td>
<td>31 kbps</td>
<td>41 kbps</td>
<td>65 kbps</td>
<td>76 kbps</td>
<td>70 kbps</td>
</tr>
</tbody>
</table>


As an example, if 3 simultaneous G.729 calls over MPN with DMNR are desired, 3 x 41kbps or 123 kbps would be set in the 2nd level policy map for the strict priority class (class “CS5-EF” in the examples that follow). Call admission control would define this site as a “location” and limit the voice traffic to 125kbps, the phones at this site would be in a device pool assigned to a region that supported only G.729.

ISR 1000, 4200, 4300 and 4400 Series LTE QoS Guidelines

These platforms provide for the maximum upstream throughput for LTE. Upon egress/upstream congestion detection on the cellular interface, the ISR 4K will engage the defined QoS policing in order to queue less important traffic and prioritize more important traffic. This benefits users and application for best effort traffic, and may also benefit more important traffic that exceeds the service level (512Kbps or 2Mbps) by dynamically allowing the maximum upstream bandwidth to be used while still “protecting” more important traffic if there’s more to send than the upstream bandwidth allows. As LTE is a shared medium and the cellular network dynamically increases or decreases an LTE connection’s radio resources, the ISR 4K adapts to offer the maximum throughput with QoS.

Traffic shaping is a method of rate limiting traffic without dropping the excess packets. Packets that would exceed a particular bit-per-second rate are momentarily queued. This provides traffic smoothing, with TCP-based flows adapting to a lower throughput rate, so those applications can operate without retransmitting packets and exacerbating traffic congestion. Traffic shaping also induces congestion notification.

The 4000/1000 series configuration in this document is based on class-based traffic shaping and multi-level traffic treatment. The configuration is constructed to support maximum overall LTE upstream throughput while protecting more important traffic. The QoS configuration is described below.

- First level: if LTE upstream congestion is detected, any traffic marked with DSCP CS5 or EF is provided strict priority over all other traffic up to a defined bit rate. Any traffic marked with CS3, CS4, AF3x, AF4x is prioritized above remaining traffic. The combination of the above important traffic is rate limited to not exceed the service rate (either 512Kbps or 2Mbps).
- First level: if the aggregate of CS3/AF3x, CS4/AF4x and CS5/EF traffic exceeds the service rate, it will be shaped (rate limited) so that the service rate (max allowed bit rate on dedicate bearer) is not exceeded. The shaped rate for CS3/AF3x, CS4/AF4x and CS5/EF traffic will be set to the service rate.
- Second level: if important traffic will exceed the service rate, multiple AF3x/CS3 and AF4x/CS4 traffic classes along with a voice class can each be provided a minimum bandwidth. The important data classes can transmit greater than the minimum defined, but together with voice will not exceed the service rate.

The figure below depicts a 2-level policy. “T” represents the total upstream bit rate that is dynamically available. “S” is the service rate, the bit rate that Verizon provides for important traffic (Mission Critical CoS). This is the ontract rate for the dedicated bearer (2Mbps or 0.5Mbps). “V, X, Y” represent traffic class bit rates for traffic classes that are within the service rate: “V” for voice, X for key business apps, Y for business apps.

- Flexible to match an organization’s QoS design. Full LTE B/W can be used
- Traffic on dedicated bearer shaped to its service rate, reduced dropped packets
- Class-based shaping if SP polices the dedicated bearer. Default bearer not policed
- If the bandwidth allotted to a traffic class is unused, it can be used by other classes
Typical Deployment: The following scenario may be seen in a typical deployment. After an ISR installation with LTE site survey, there is sufficient upstream bandwidth to exceed the 2Mbps service rate for mission critical traffic. Leveraging Verizon PNTM with the 2Mbps service rate, below is a feasible logical and configuration model. It assumes 200Kbps of voice traffic (e.g. 5 simultaneous 8.729 calls), a minimum of 1.8Mbps of important traffic (CS3/AF3x/CS4/AF4x), with 2 important traffic sub-classes (CS3/AF3x, CS4/AF4x) each of which is offered a minimum of 800Kbps and 1Mbps respectively. More granular policies are feasible.

The ISR QoS egress policies depicted in the diagram below implement the logical design shown above.
ISR Configuration for LTE NIM (ISR 4K) and ISR 1K – LTE QoS

### IOS XE 16.3.3 or later recommended as of March 2018 ###

```
service internal
!
hostname C4321-4G
!
boot-start-marker
boot system bootflash:isr4300-universalk9.16.03.03.SPA.bin
boot-end-marker
!
ip dhcp pool 10dot250dot1
  network 10.250.1.0 255.255.255.0
  default-router 10.250.1.1
  dns-server 10.20.45.20
!
username cisco privilege 15 secret 5 xxxxxxxxxxxxxx
!
controller Cellular 0/2/0
!
class-map match-any PREC-345
  match ip precedence 3 4 5
class-map match-any PREC-5
  match ip precedence 5
class-map match-any PREC-4
  match ip precedence 4
  match access-group name PRO-NEMO
class-map match-any PREC-3
  match ip precedence 3
class-map match-all MARK-CS3
  match access-group name DEV-CS3
class-map match-all MARK-CS4
  match access-group name DEV-CS4
class-map match-all MARK-CS5
  match access-group name DEV-CS5
!
policy-map SUB-CLASS-345
  class PREC-5
    priority level 1 percent 10
  class PREC-4
    bandwidth percent 50
  class PREC-3
    bandwidth percent 40
!
policy-map PNTM-A
  class PREC-345
    shape average 2000000
    bandwidth remaining ratio 3
    service-policy SUB-CLASS-345
  class class-default
    bandwidth remaining ratio 1
    fair-queue
    random-detect dscp-based
```
policy-map MARK-IN
  class MARK-CS5
    set dscp cs5
  class MARK-CS4
    set dscp cs4
  class MARK-CS3
    set dscp cs3

! interface Loopback1234
  description ### NEMO Router Home Address
  ip address 1.2.3.4 255.255.255.255

! interface GigabitEthernet0/0/0
  description ### on ISR 4K, for ISR 1K, this would be VLAN 1
  ip address 10.250.1.1 255.255.255.0
  ip tcp adjust-mss 1390
  load-interval 30
  media-type rj45
  negotiation auto
  service-policy input MARK-IN

! interface GigabitEthernet0/0/1
  ip address 10.0.3.1 255.255.255.0
  ip nat inside
  ip tcp adjust-mss 1390
  load-interval 30
  negotiation auto

! interface Cellular0/2/0
  bandwidth 1500
  ip address negotiated
  ip mobile router-service roam
  ip mobile router-service collocated ccoa-only
  load-interval 30
  dialer in-band
  dialer idle-timeout 0
  dialer enable-timeout 2
  dialer watch-group 1
  pulse-time 1
  service-policy output PNTM-A

! interface GigabitEthernet0
  vrf forwarding Mgmt-intf
  ip address 10.0.0.2 255.255.255.254
  negotiation auto

! router mobile

! ip forward-protocol nd

! ip mobile secure home-agent 66.174.251.2 spi decimal 256 key ascii VzWNeMo
  algorithm hmac-md5
  ip mobile router
    address 1.2.3.4 255.255.255.0
    collocated single-tunnel
  home-agent 66.174.251.2
mobile-network GigabitEthernet0/0/1
mobile-network GigabitEthernet0/0/0
non-connected-network 192.168.222.0 255.255.255.0
register extend expire 10 retry 3 interval 5
reverse-tunnel
tunnel mode gre

ip access-list extended DEV-CS3
  permit ip host 10.250.1.12 any
ip access-list extended DEV-CS4
  permit ip host 10.250.1.11 any
ip access-list extended DEV-CS5
  permit ip host 10.250.1.10 any
access-list extended PRO-NEMO
  permit udp any any eq mobile-ip

!"dialer watch-list 1 ip 5.6.7.8 0.0.0.0
  dialer watch-list 1 delay route-check initial 60
  dialer watch-list 1 delay connect 1

  control-plane

!"alias exec spc show policy-map interface cell 0/2/0 out
alias exec sic show interface cell 0/2/0
alias exec sig show interface gi0/0/0
alias exec spg show policy-map interface gi0/0/0 in

!"line con 0
  exec-timeout 0 0
  login local
  stopbits 1
line aux 0
  stopbits 1
line vty 0 4
  login local
!"end

*** Note that the ingress QoS policy on Gigabit Ethernet 0/0/0 (MARK-IN) is configured to re-mark traffic for QoS lab testing. For an organization’s production configuration, there may be no policy (marking may be done by trusted devices or LAN switches) or a policy that integrates with an organization’s existing QoS design.
Traffic shaping is a method of rate limiting traffic without dropping the excess packets. Packets that would exceed a particular bit-per-second rate are momentarily queued. This provides traffic smoothing, with TCP-based flows adapting to a lower throughput rate, so those applications can operate without retransmitting packets and exacerbating traffic congestion. Traffic shaping also induces congestion notification.

The 1900, 2900, 3900, 8x9 series configuration guidelines in this document are based on traffic shaping and multi-level traffic treatment. The configurations are constructed in a way that allow for simple modification when current limitations of LTE modem congestion notification are mitigated (by removing the top level shaper). The QoS configuration includes 3 levels of traffic treatment, and are described below.

- First level: all traffic sent over an LTE connection is shaped. Recommendations regarding determining the shaped rate are included in the subsequent deployment section. If the traffic sent is below the shaped rate, the overall traffic is not conditioned. If the traffic to be sent will exceed the shaped rate, the traffic is temporarily queued to smooth bursts. The traffic is also then prioritized based on the second level.

- Second level: if congestion for overall traffic has occurred, any traffic marked with DSCP CS5 or EF is provided strict priority over all other traffic up to a defined bit rate. Any traffic marked with CS3, CS4, AF3x, AF4x is prioritized above remaining traffic. The combination of the above important traffic is rate limited to not exceed the service rate (either 512Kbps or 2Mbps).

- Second level: if the important traffic rate (CS3/AF3x, CS4/AF4x, CS5/EF) will exceed the service rate, it will be shaped so that the service rate (max allowed bit rate on dedicate bearer) is not exceeded. The shaped rate for CS3/AF3x /CS4/AF4x will be set to the service rate less the bit rate defined for CS5/EF.

- Third level: if important traffic will exceed the service rate, multiple AF3x/CS3/AF4x/CS4 traffic classes can each be provided a minimum amount of bandwidth. These classes can transmit greater than the minimum defined, but together these will not exceed the service rate less the bit rate defined for CS5/EF.

The figure below depicts a 3 level policy. The value T represents the Total upstream bit rate that is always available (the achievable upstream bandwidth is never below this value). The value S is the service rate, the bit rate that Verizon provides for important traffic (Mission Critical CoS). The values ID1, ID2, etc. represent traffic classes that are within the service rate. The value V represents traffic requiring strict priority (e.g. voice).

- Flexible prioritization to match an organization’s existing QoS design
- Traffic shaped to not exceed total bit rate and service rate, reduce dropped packets
- If the bandwidth allotted to a traffic class is unused, it can be used by other classes
- The overall LTE shaped rate \( T \) (first level policy) defined on LTE egress from the ISR should be set to a value that is always achievable. If the shaped rate is above the actual achievable rate, QoS may not be engaged, and traffic requiring QoS (e.g. Voice over IP) may experience degradation (e.g. voice clipping, echo, silence). Determining the overall LTE shaped rate requires a repeatable process at each location, including peak days, times and conditions. Examples of methods to do this include:
  
  o For a location where LTE MPN is not yet deployed, bring an ISR (e.g. 819) and laptop running a test tool (e.g. JPERF), perform a site survey (optionally using the SMS app above), and run the test tool with a UDP traffic stream (may require a test tool at the main location to “catch” the stream, e.g. a PC also running JPERF with “server” radio button selected). With iterations at higher speeds, determine at what Mbps rate the JPERF catcher sees less than what the JPERF laptop behind the LTE ISR is sending. Also, a “show interface cellular” command (with appropriate interface number) can confirm the value seen by the ISR. Note that doing the latter, requires the JPERF test stream to run twice as long as the ISR LTE interface’s “load interval” (which can be set down to 30 seconds with the ISR interface configuration command “load-interval 30”).
  
  o For a location where LTE MPN is already deployed, confirm that a site survey has been completed. If LTE is the backup path, determine how to run the test across LTE. Run a test tool from that location with a UDP traffic stream (may require a test tool at the main location to “catch” the stream). With iterations at higher speeds, determine at what Mbps rate the JPERF catcher sees less than what the JPERF laptop behind the LTE ISR is sending. The ISR “show interface cellular” command (with appropriate interface number) will be required, as the LTE interface will account for real traffic also traversing LTE. The “load interval” caveat above is still pertinent. Note that if there is production traffic over LTE, the test tool speeds should be set to not unduly impact applications.
  
  o If the overall shaped rate \( T \) chosen is greater than the ISR’s achievable upstream LTE bit rate, packets may be dropped without QoS being engaged. The behavior and impact on users and applications would be the same as if no QoS was defined, e.g. best effort treatment for all traffic.

**Private Network Traffic Management – 1900/2900/3900/8x9 Design and Configuration**

**Typical Deployment:** The following scenario may be seen in a typical deployment. After an ISR installation with LTE site survey, downstream LTE performance may consistently exceed 10Mbps and upstream LTE total bit rate may consistently exceed 3Mbps. Leveraging Verizon PNTM with the 2Mbps service rate, below is a feasible logical and configuration model. It assumes 200Kbps of voice traffic (e.g. 5 simultaneous G.729 calls), a maximum of 1.8Mbps of important traffic (CS3/AF3x/CS4/AF4x), with 2 important traffic sub-classes (CS3/AF3x, CS4/AF4x) each of which is offered a minimum of 800Kbps and 600Kbps respectively. More granular policies are feasible.
Ingress marking will not be shown as that is not specific to the LTE QoS egress deployment. It is assumed.

The ISR QoS egress policies depicted in the diagram below implement the logical design shown above.

```
class-map match-any CS5-EF
  match ip precedence 5
  match ip precedence 3
  class-map match-any PREC-3
  match ip precedence 4
  class-map match-any PREC-4
  match ip precedence 3
  4
  match access-group NEMO

policy-map LTE-SHAPER
  class class-default
  shape average 3000000
  service-policy PNTM-B

interface Cellular0
  service-policy output LTE-SHAPER

policy-map PNTM-B
  class CS5-EF
  priority 200
  class PREC-34
  shape average 1800000
  bandwidth 1800
  service-policy SUB-CLASS-34
  class class-default
  fair-queue
  random-detect dscp-based

policy-map SUB-CLASS-34
  class PREC-3
  bandwidth 600
  class PREC-4
  bandwidth 800
```
The 2Mbps service rate assumed in the above scenario offers flexibility and a significant bit rate for important applications. The 2Mbps service rate can also offer value if the total upstream consistent throughput is as low as 2Mbps. However, not all of the service rate would be configured in this instance. The recommended practice (and Cisco ISR configuration default) is to limit traffic with preferential treatment to 75% of the total traffic sent. This assists applications with best-effort service in functioning properly, and can assist important applications inadvertently missed in the traffic classes offered preferential treatment to function properly.

Note: In the instance where less than 2.7Mbps total LTE upstream bit rate is consistently available, the service rate used by the ISR would be reduced to a maximum of 75% of the total LTE upstream bit rate. As an example, if there was 2Mbps total upstream consistent LTE bit rate, the service rate leveraged would be 1.5Mbps.

**Less-Than-Optimal Deployment Scenario, Configuration and Show Output**

The following scenario is a less-than optimal deployment, where a minimal upstream bit rate is achievable, and the lower service rate (512Kbps) is chosen for an LTE connection with QoS. The detailed configuration examples are based on this scenario, as testing with the configurations was highlighted to ensure proper service can be provided to key applications at these lower bit rates. The scenario is applicable down to 750Kbps of total consistently achievable upstream LTE bit rate.

The diagram below depicts the domain of LTE QoS (from the ISR to the LTE MPN anchor point). The line shown in red highlights the tunnel if the optional DMNR (native routing) MPN service is leveraged. The addressing shown is from the testing associated with the complete ISR configurations on the following pages.
The diagram below depicts the logical QoS egress policy, assuming 1.5Mbps consistently achievable upstream total LTE bit rate.

- Voice marked DSCP EF, Very important apps marked AF4x/CS4, Important apps marked AF3x/CS3
- Traffic shaped to not exceed total bit rate and service rate, reduce dropped packets
- The bandwidth for strict priority class (voice) is subtracted from the Service Rate
- The total for Classes CS3/AF3x/CA4/AF4x is shaped to 400Kbps (500Kbps – 100Kbps)

The ISR QoS egress policies depicted in the diagram below implements the logical design shown above.

class-map match-any CS5-EF
  match ip precedence 5
class-map match-any PREC-3
  match ip precedence 3
class-map match-any PREC-4
  match ip precedence 4
class-map match-any PREC-34
  match ip precedence 3 4
match access-group NEMO

policy-map LITE-SHAPER
  class class-default
  shape average 1500000
service-policy PNTM-A

policy-map PNTM-A
  class CS5-EF
    priority 100
  class PREC-34
    shape average 400000
    bandwidth 400
service-policy SUB-CLASS-34
  class class-default
    fair-queue
    random-detect dscp-based

policy-map SUB-CLASS-34
  class PREC-3
    bandwidth 150
  class PREC-4
    bandwidth 250

There are two example configurations that follow, based on scenario above. Note that commands that do not impact QoS and are optional or are included by default may not be included:

- LTE eHWIC (ISR 1900, 2900, 3900) and GRWIC (CGR2010)
ISR Configuration for LTE eHWIC (1900/2900/3900) – LTE QoS

!### IOS 15.5(3)M4 or later ###

service internal

hostname C1921-QOS-A

boot-start-marker

boot system flash:c1900-universalk9-mz.SPA.155-3.M4.bin

boot-end-marker

ip dhcp pool VLAN1

  network 10.250.2.0 255.255.255.0
  default-router 10.250.2.1
  dns-server 4.2.2.2

ip cef

username cisco privilege 15 secret 5 xxxxxxxxxxxxxx

chat-script LTE "" "AT!CALL1" TIMEOUT 20 "OK"

ip tcp mss 1460

! class-map match-any CS5-EF

  match ip precedence 5

class-map match-any PREC-3

  match ip precedence 3

class-map match-all MARK-CS3

  match access-group name CS3

class-map match-all MARK-CS4

  match access-group name CS4

class-map match-all MARK-CS5

  match access-group name CS5

class-map match-any PREC-4

  match ip precedence 4

class-map match-any CS3-4

  match ip precedence 3 4

  match access-group 105

! policy-map SUB-CLASS-34

  class PREC-3

    bandwidth 150

  class PREC-4

    bandwidth 200

policy-map PNTM-A

  class CS5-EF

    priority 100

  class CS3-4

    bandwidth 400

      shape average 400000

      service-policy SUB-CLASS-34

  class class-default

    fair-queue

    random-detect dscp-based

policy-map LTE-SHAPER

  class class-default

    shape average 1500000

    service-policy PNTM-A
policy-map BCPS-IN
  class MARK-CS4
    set dscp cs4
  class MARK-CS5
    set dscp cs5
  class MARK-CS3
    set dscp cs3

interface Loopback100
  ip address 10.250.0.1 255.255.255.255

interface Embedded-Service-Engine0/0
  no ip address
  shutdown

interface GigabitEthernet0/0
  description $ETH-LAN$$ETH-SW-LAUNCH$$INTF-INFO-GE 0/0$
  ip address 10.10.10.1 255.255.255.248
  duplex auto
  speed auto

interface GigabitEthernet0/1
  ip address 10.10.29.12 255.255.255.0
  duplex auto
  speed auto

  ### Layer 2 Ethernet switch port default configuration removed ###

interface Cellular0/0/0
  ip address negotiated
  ip mobile router-service roam
  ip mobile router-service collocated ccoa-only
  ip virtual-reassembly in
  encapsulation slip
  load-interval 30
  dialer in-band
  dialer idle-timeout 0
  dialer string LTE
  dialer watch-group 1
  pulse-time 0
  service-policy output LTE-SHAPER

interface Vlan1
  ip address 10.250.1.1 255.255.255.0
  ip virtual-reassembly in
  load-interval 30
  service-policy input BCPS-IN

router mobile

  ip mobile secure home-agent 66.174.X.Y spi decimal 256 key ascii VzWNeMo
  algorithm hmac-md5
  ip mobile router
    address 1.2.3.4 255.255.255.0
    collocated single-tunnel
    home-agent 66.174.X.Y
mobile-network Loopback100
mobile-network Vlan1
register retransmit initial 2000 maximum 2000 retry 2
register extend expire 40 retry 10 interval 4
register lifetime 181
reverse-tunnel
tunnel mode gre
no multi-path
!
ip access-list extended CS3
   permit ip host 10.250.1.12 any
ip access-list extended CS4
   permit ip host 10.250.1.10 any
ip access-list extended CS5
   permit ip host 10.250.1.11 any
!
access-list 105 permit udp any any eq mobile-ip
!
alias exec spc show policy-map interface cell 0/0/0 out
alias exec sic show interface cell 0/0/0
alias exec siv show interface vlan1
alias exec spv show policy-map interface vlan1 in
!
line con 0
   exec-timeout 0 0
   login local
line aux 0
line 2
   no activation-character
   no exec
   transport preferred none
   transport output lat pad telnet rlogin lapb-ta mop udptn v120 ssh
   stopbits 1
line 0/0/0
   script dialer LTE
   modem InOut
   no exec
   transport input telnet
line vty 0 4
   exec-timeout 0 0
   privilege level 15
   login local
   transport input telnet ssh
!

*** Note that the ingress QoS policy on VLAN1 (BCPS-IN) is configured to re-mark traffic for QoS lab testing. For an organization’s production configuration, there may be no policy (marking may be done by trusted devices or LAN switches) or a policy that integrates with an organization’s existing QoS design.
ISR Configuration for LTE 8x9 – LTE QoS

### IOS 15.5(3)M4 or later ###

service internal
hostname C819VZ-XL
!
boot-start-marker
boot system flash:c800-universalk9-mz.SPA. 155-3.M4.bin
boot-end-marker
!
ip dhcp pool VLAN1
  network 10.250.2.0 255.255.255.0
  default-router 10.250.2.1
  dns-server 4.2.2.2
!
chat-script lte "" "AT!CALL" TIMEOUT 20 "OK"
!
username verizon privilege 15 password 0 xxxxxxxxxx
!
class-map match-any CS5-EF
  match ip precedence 5
class-map match-any PREC-3
  match ip precedence 3
class-map match-any MARK-CS3
  match access-group name CS3
class-map match-any MARK-CS4
  match access-group name CS4
class-map match-any MARK-CS5
  match access-group name CS5
class-map match-any PREC-4
  match ip precedence 4
class-map match-any CS3-4
  match ip precedence 3 4
  match access-group 105
!
policy-map SUB-CLASS-34
  class PREC-3
    bandwidth 150
  class PREC-4
    bandwidth 200
policy-map PNTM-A
  class CS5-EF
    priority 100
class class-default
  bandwidth 400
  shape average 400000
  service-policy SUB-CLASS-34
class class-default
  fair-queue
  random-detect dscp-based
policy-map LTE-SHAPER
  class class-default
    shape average 1500000
    service-policy PNTM-A
policy-map BCPS-IN
  class MARK-CS4
    set dscp cs4
  class MARK-CS5
    set dscp cs5
  class MARK-CS3
    set dscp cs3
!
interface Loopback100
  ip address 192.168.249.2 255.255.255.255
!
interface Loopback1234
  description ### NEMO Router Home Address. Dummy non-Routable IP ###
  ip address 1.2.3.4 255.255.255.255
!
interface Cellular0
  ip address negotiated
  no ip unreachables
  ip accounting precedence input
  ip flow ingress
  ip flow egress
  ip mobile router-service roam
  ip mobile router-service collocated ccoa-only
  encapsulation slip
  load-interval 30
  dialer in-band
  dialer idle-timeout 0
  dialer string lte
  dialer watch-group 1
  async mode interactive
  pulse-time 0
  service-policy output LTE-SHAPER
!
interface FastEthernet0
  no ip address
!
interface FastEthernet1
  no ip address
!
interface FastEthernet2
  no ip address
!
interface FastEthernet3
  no ip address
!
interface GigabitEthernet0
  ip address 10.10.10.2 255.255.255.0
  no ip split-horizon
  shutdown
duplex auto
  speed auto
  no keepalive
!
interface Vlan1
  ip address 10.250.2.1 255.255.255.0
  load-interval 30
service-policy input BCPS-IN
!
router mobile
!
ip forward-protocol nd
!
ip mobile secure home-agent 66.174.251.2 spi decimal 256 key ascii VzWNeMo
algorithm hmac-md5
ip mobile router
    address 1.2.3.4 255.255.255.0
    collocated single-tunnel
    home-agent 66.174.251.2
mobile-network Vlan1
    mobile-network Loopback100
    non-connected-network 192.168.99.0 255.255.255.0
    register retransmit initial 2000 maximum 2000 retry 2
    register extend expire 40 retry 10 interval 4
    register lifetime 181
reverse-tunnel
tunnel mode gre
no multi-path
!
ip access-list extended CS3
    permit ip host 10.250.2.10 any
ip access-list extended CS4
    permit ip host 10.250.2.11 any
ip access-list extended CS5
    permit ip host 10.250.2.12 any
!
dialer watch-list 1 ip 5.6.7.8 0.0.0.0

dialer watch-list 1 delay route-check initial 60

dialer watch-list 1 delay connect 1
!
access-list 105 permit udp any any eq mobile-ip
!
alias exec spc show policy-map interface cell 0 out
alias exec sic show interface cell 0
alias exec siv show interface vlan1
alias exec spv show policy-map interface vlan1 in
!
line con 0
    exec-timeout 0 0
line 2
    no activation-character
    no exec
    transport input all
line 3
    script dialer lte
    modem InOut
    no exec
    transport input telnet
line vty 0 4
    login local
    transport input all
!
end
*** Note that the ingress QoS policy on VLAN1 (BCPS-IN) is configured to re-mark traffic for QoS lab testing. For an organization's production configuration, there may be no policy (marking may be done by trusted devices or LAN switches) or a policy that integrates with an organization's existing QoS design.
Operation and Show Commands

ISR 819 QoS policy:
- 3 level, total shaper to 1.5 Mbps, 0.5 Mbps service rate,
- 100 Kbps strict priority (EF), 400 Kbps AF3x/AF4x/CS3/CS4)
- At least 250 Kbps for CS4/AF4x, 150 Kbps for CS3/AF3x

Traffic sent through 819:
- ~90 Kbps voice, ~300 Kbps CS3, ~1.2Mbps best effort

Traffic sent to ISR 819 VLAN Interface:
C819VZ-XL#siv
("siv" is a configured alias that expands to "show int vlan1")
Vlan1 is up, line protocol is up
Hardware is EtherSVI, address is bcc4.93da.3448 (bia bcc4.93da.3448)
Internet address is 10.250.2.1/24
MTU 1500 bytes, BW 100000 Kbit/sec, DLY 100 usec,
reliability 255/255, txload 1/255, rxload 4/255
Encapsulation ARPA, loopback not set
Keepalive not supported
ARP type: ARPA, ARP Timeout 04:00:00
Last input 00:00:08, output never, output hang never
Last clearing of "show interface" counters never
Input queue: 0/75/0/0 (size/max/drops/flushes); Total output drops: 0
Queueing strategy: fifo
Output queue: 0/40 (size/max)
30 second input rate 1602000 bits/sec, 194 packets/sec
30 second output rate 85000 bits/sec, 50 packets/sec
73227 packets input, 37647211 bytes, 0 no buffer
Received 1006 broadcasts (27 IP multicasst)
0 runts, 0 giants, 0 throttles
0 input errors, 0 CRC, 0 frame, 0 overrun, 0 ignored
52232 packets output, 12161738 bytes, 0 underruns
0 output errors, 1 interface resets
424 unknown protocol drops
0 output buffer failures, 0 output buffers swapped out

Traffic sent by ISR 819 LTE Interface:
C819VZ-XL#sic
("sic" is a configured alias that expands to "show int cell0")
Cellular0 is up, line protocol is up
Hardware is Cellular, Modem inserted
Internet address is 10.14.11.104/32
MTU 1500 bytes, BW 50000 Kbit/sec, DLY 100000 usec,
reliability 255/255, txload 7/255, rxload 1/255
Encapsulation SLIP, loopback not set
Keepalive not supported
Last input 00:00:00, output never, output hang never
Last clearing of "show interface" counters never
Input queue: 0/75/0/0 (size/max/drops/flushes); Total output drops: 1523
Queueing strategy: Class-based queuing
Output queue: 41/1000/1499 (size/max total/drops)
30 second input rate 89000 bits/sec, 50 packets/sec
30 second output rate 1475000 bits/sec, 180 packets/sec
52636 packets input, 12734332 bytes, 0 no buffer
Received 0 broadcasts (0 IP multicasst)
0 runts, 0 giants, 0 throttles
0 input errors, 0 CRC, 0 frame, 0 overrun, 0 ignored, 0 abort
70294 packets output, 36627882 bytes, 0 underruns
0 output errors, 0 collisions, 1 interface resets
0 unknown protocol drops
0 output buffer failures, 0 output buffers swapped out
0 carrier transitions
DCD=up  DSR=up  DTR=up  RTS=up  CTS=up

Traffic sent by 819 LTE Interface: QoS Specifics
C819VZ-XL#spc (“spc” is an alias that expands to “show policy-map int cell0 out”
Cellular0

**Service-policy output: LTE-SHAPER**

Class-map: class-default (match-any)
72629 packets, 3962323 bytes
30 second offered rate 1643000 bps, drop rate 148000 bps
Match: any
Queueing
queue limit 64 packets
(queue depth/total drops/no-buffer drops) 42/1583/0
(pkts output/bytes output) 71046/37498130
shape (average) cir 1500000, bc 6000, be 6000
target shape rate 1500000

**Service-policy : PNTM-A**

queue stats for all priority classes:
Queueing
queue limit 64 packets
(queue depth/total drops/no-buffer drops) 0/0/0
(pkts output/bytes output) 49592/11108608

**Class-map: CS5-EF (match-any)**
49592 packets, 11108608 bytes
30 second offered rate 89000 bps, drop rate 0000 bps
Match: ip precedence 5
49592 packets, 11108608 bytes
30 second rate 89000 bps
Priority: 100 kbps, burst bytes 2500, b/w exceed drops: 0

**Class-map: CS3-4 (match-any)**
4321 packets, 5463327 bytes
30 second offered rate 311000 bps, drop rate 0000 bps
Match: ip precedence 3 4
4225 packets, 5449119 bytes
30 second rate 311000 bps
Match: access-group 105
96 packets, 14208 bytes
30 second rate 0 bps
Queueing
queue limit 64 packets
(queue depth/total drops/no-buffer drops) 0/0/0
(pkts output/bytes output) 4321/5463327
bandwidth 400 kbps
shape (average) cir 400000, bc 1600, be 1600
target shape rate 400000

**Service-policy : SUB-CLASS-34**

**Class-map: PREC-3 (match-any)**
4225 packets, 5449119 bytes
30 second offered rate 311000 bps, drop rate 0000 bps
Match: ip precedence 3
4225 packets, 5449119 bytes
30 second rate 311000 bps
Queueing
queue limit 64 packets
(queue depth/total drops/no-buffer drops) 0/0/0
(pkts output/bytes output) 4225/5449119
bandwidth 150 kbps

Class-map: PREC-4 (match-any)
0 packets, 0 bytes
30 second offered rate 0000 bps, drop rate 0000 bps
Match: ip precedence 4
0 packets, 0 bytes
30 second rate 0 bps
Queueing
queue limit 64 packets
(queue depth/total drops/no-buffer drops) 0/0/0
(pkts output/bytes output) 0/0
bandwidth 200 kbps

Class-map: class-default (match-any)
96 packets, 14208 bytes
30 second offered rate 0000 bps, drop rate 0000 bps
Match: any

queue limit 64 packets
(queue depth/total drops/no-buffer drops) 0/0/0
(pkts output/bytes output) 0/0

Class-map: class-default (match-any)
18716 packets, 23051288 bytes
30 second offered rate 1242000 bps, drop rate 149000 bps
Match: any
Queueing
queue limit 64 packets
(queue depth/total drops/no-buffer drops/flowdrops) 40/1583/0/0
(pkts output/bytes output) 17133/20926195
Fair-queue: per-flow queue limit 16 packets
Exp-weight-constant: 9 (1/512)
Mean queue depth: 40 packets

<table>
<thead>
<tr>
<th>dscp</th>
<th>Transmitted</th>
<th>Random drop</th>
<th>Tail/Flow drop</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mark</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pkts/bytes</td>
<td>pkts/bytes</td>
<td>pkts/bytes</td>
<td>thresh</td>
<td>thresh</td>
<td>prob</td>
</tr>
<tr>
<td>default</td>
<td>16920/20904895</td>
<td>1120/1505437</td>
<td>463/619656</td>
<td>20</td>
<td>40</td>
<td>1/10</td>
</tr>
<tr>
<td>cs6</td>
<td>213/21300</td>
<td>0/0</td>
<td>0/0</td>
<td>32</td>
<td>40</td>
<td>1/10</td>
</tr>
</tbody>
</table>
ISR 1000 QoS policy (Similar to ISR 4000 series):
- 2 level, shape priority traffic to 2 Mbps service rate,
- 200 Kbps strict priority (EF), 800 Kbps AF3x/CS3, 1 Mbps for AF4x/CS4

1101VZ#sh run | s policy-map
policy-map SUB-CLASS-345
  class PREC-5
    priority level 1 percent 10
  class PREC-3
    bandwidth percent 40
  class PREC-4
    bandwidth percent 50

policy-map PNTM-A
  class PREC-345
    shape average 2000000
    bandwidth remaining ratio 3
    service-policy SUB-CLASS-345
  class class-default
    fair-queue
    random-detect dscp-based
    bandwidth remaining ratio 1

1101VZ### cell 0/2/1 due to unit configured for VZW split data routing
1101VZ#sh run int cell 0/2/1 | i pol
service-policy output PNTM-A

1101VZ#sh policy-map interface cell 0/2/1 out
Cellular0/2/1
Service-policy output: PNTM-A

Class-map: PREC-345 (match-any)
  0 packets, 0 bytes
  30 second offered rate 0000 bps, drop rate 0000 bps
  Match: ip precedence 3 4 5
  Queueing
  queue limit 64 packets
  (queue depth/total drops/no-buffer drops) 0/0/0
  (pkts output/bytes output) 0/0
  shape (average) cir 2000000, bc 8000, be 8000
  target shape rate 2000000
  bandwidth remaining ratio 3

Service-policy : SUB-CLASS-345

queue stats for all priority classes:
  Queueing
  priority level 1
  queue limit 512 packets
  (queue depth/total drops/no-buffer drops) 0/0/0
  (pkts output/bytes output) 0/0

Class-map: PREC-5 (match-any)
  0 packets, 0 bytes
30 second offered rate 0000 bps, drop rate 0000 bps
Match: ip precedence 5
Priority: 10% (200 kbps), burst bytes 5000, b/w exceed drops: 0

Priority Level: 1

Class-map: PREC-3 (match-any)
0 packets, 0 bytes
30 second offered rate 0000 bps, drop rate 0000 bps
Match: ip precedence 3
Queueing
queue limit 64 packets
(queue depth/total drops/no-buffer drops) 0/0/0
(pkt output/bytes output) 0/0
bandwidth 40% (800 kbps)

Class-map: PREC-4 (match-any)
0 packets, 0 bytes
30 second offered rate 0000 bps, drop rate 0000 bps
Match: ip precedence 4
Match: access-group name PRO-NEMO
Queueing
queue limit 64 packets
(queue depth/total drops/no-buffer drops) 0/0/0
(pkt output/bytes output) 0/0
bandwidth 50% (1000 kbps)

Class-map: class-default (match-any)
0 packets, 0 bytes
30 second offered rate 0000 bps, drop rate 0000 bps
Match: any
queue limit 64 packets
(queue depth/total drops/no-buffer drops) 0/0/0
(pkt output/bytes output) 0/0

Class-map: class-default (match-any)
0 packets, 0 bytes
30 second offered rate 0000 bps, drop rate 0000 bps
Match: any
Queueing
queue limit 208 packets
(queue depth/total drops/no-buffer drops/flowdrops) 0/0/0/0
(pkt output/bytes output) 0/0
Fair-queue: per-flow queue limit 52 packets
Exp-weight-constant: 9 (1/512)
Mean queue depth: 0 packets
dscp Transmitted Random drop Tail/Flow drop Minimum
pkts/bytes pkts/bytes pkts/bytes thresh

bandwidth remaining ratio 1

1101VZ#
1101VZ# sh run | s policy-map
policy-map SUB-CLASS-345
class PREC-5
  priority level 1 percent 10
class PREC-3
class PREC-4
  bandwidth percent 50
policy-map PNTM-A
class PREC-345
  bandwidth average 2000000
  bandwidth remaining ratio 3
  service-policy SUB-CLASS-345
  class class-default
    fair-queue
    random-detect dscp-based
    bandwidth remaining ratio 1
alias exec spc show policy-map int cell 0/2/0 out
alias exec spv show policy-map int vlan 1 in
1101VZ#

1101VZ# sh run | s class-map
class-map match-any PREC-3
  match ip precedence 3
class-map match-any PREC-345
  match ip precedence 3 4 5
class-map match-any MARK-CS3
  match access-group name DEV-CS3
class-map match-any PREC-5
  match ip precedence 5
class-map match-any MARK-CS4
  match access-group name DEV-CS4
class-map match-any PREC-4
  match ip precedence 4
  match access-group name PRO-NEMO
class-map match-any MARK-CS5
  match access-group name DEV-CS5
1101VZ#

1101VZ# sh run | s access-list
ip access-list extended DEV-CS3
  permit ip host 10.250.1.12 any
ip access-list extended DEV-CS4
  permit ip host 10.250.1.11 any
ip access-list extended DEV-CS5
  permit ip host 10.250.1.10 any
ip access-list extended PRO-NEMO
  permit udp any any eq mobile-ip
access-list 100 permit ip 0.0.0.0 254.255.255.255 any
1101VZ#
Frequently Asked Questions

Q) If encryption is enabled (GETVPN, DMVPN, etc.) with QoS, QoS reorders packets. Can packets be dropped by the anti-replay algorithm in the crypto function?

A) While this is possible, it is not seen in most deployments. The anti-replay window (if enabled) defaults to a 64 packet sliding window, which is normally wide enough to accommodate this packet reordering. The packets most likely to be affected are those in the default/best-effort class(es). In the unlikely case this is suspected, the following 2 documents discuss the issue and how to expand the anti-replay sliding window.


Q) Why is a site survey so important for LTE QoS?

A) Site surveys are important for all LTE deployments. The radio signal reception strength and quality have a significant impact on the modulation chosen between the LTE modem and network, which directly impacts the throughput. Site surveys are available as a professional service, or can be done using a smart phone with texting, using the no-charge ISR app referenced under “LTE QoS Planning Guidelines” above.

Q) Can I use both precedence and DSCP values in QoS policies?

A) Yes. Precedence only takes into consideration the 3 most significant bits in the ToS byte. DSCP takes into consideration the 6 most significant bits. If a class map matches precedence 3 for example, it would match any IP packet with precedence 3, CS3, AF31, AF32, AF33. This can be useful for example in the 2nd level policy, making the configuration shorter. Similarly, precedence 5 would match on both DSCP CS5 and EF. DSCP is used to provide more granular classification, as is used the above example for the 3rd level policies.