TEAS Working Group
Internet-Draft
Intended status: Informational
Expires: 23 April 2026

L. M. Contreras, Ed.
Telefonica
I. Bykov, Ed.
Ribbon Communications
K. G. Szarkowicz, Ed.
Juniper Networks
20 October 2025

5QI to DiffServ DSCP Mapping Example for Enforcement of 5G End-to-End Network Slice QoS
draft-cbs-teas-5qi-to-dscp-mapping-05

Abstract

5G End-to-End Network Slice QoS is an essential aspect of network slicing, as described in both IETF drafts and the 3GPP specifications. Network slicing allows for the creation of multiple logical networks on top of a shared physical infrastructure, tailored to support specific use cases or services. The primary goal of QoS in network slicing is to ensure that the specific performance requirements of each slice are met, including latency, reliability, and throughput.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 23 April 2026.

Table of Contents

1. Introduction

5G End-to-End Network Slice QoS is an essential aspect of network slicing, as described in both IETF drafts and the 3GPP specifications. Network slicing allows for the creation of multiple logical networks on top of a shared physical infrastructure, tailored to support specific use cases or services. The primary goal of QoS in network slicing is to ensure that the specific performance requirements of each slice are met, including latency, reliability, and throughput.

This document focuses specifically on the QoS aspect of overall network slice realization model. The primary goal of QoS in network slicing is to ensure that the specific performance requirements of each slice are met, including latency, reliability, and throughput. As such, this document provides an example of possible grouping of 5QI values to DSCP marking that can be used as one of the building block in overall network slice realization model to aid the enforcement of the 5G Network Slice end-to-end. The grouping described are provided for illustration purposes only, and should not be considered as deployment guidance. It is not intended to

influence the way in which external systems (e.g., 3GPP or O-RAN) identify their traffic types. At the time of grouping, different criteria can be followed according to the network operator interests.

The current draft explores the impact of 3GPP traffic mapped to 5QI, being marked with DSCP values, considering scenarios involving multiple slices as well as a single slice.

For details regarding the mapping of 3GPP Network Slices to Transport Network Slices, please refer to the Network Slice Realization document [I-D.ietf-teas-5g-ns-ip-mpls], which describes an overall Network Slice realization model for IP/MPLS networks with a focus on the Transport Network fulfilling 5G slicing connectivity service objectives, and the Network Slice Application document [I-D.ietf-teas-5g-network-slice-application], which describes the overall Network Slice relationship between 3GPP and Transport Network domains.

The support of L4S [RFC9331] is being introduced in 5G systems as an operational capability for IP ECN marking and remarking. This document is focused on DSCP to 5QI mapping, leaving L4S out of scope.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The following abbreviations are used in this document:

5GC: 5G Core Network

5QI: 5G QoS Identifier

QFI: QoS Flow Identifier

ARP: Allocation and Retention Priority

S-NSSAI: Single Network Slice Selection Assistance Information

RAN: Radio Access Network

TN: Transport Network

CN: Mobile Core Network

DSCP: Differentiated Services Code Point

3. 5G QoS

In the context of 5G, the 5QI is a scalar value used to differentiate QoS characteristics in the 5G System (5GS). It indicates the QoS that a specific data flow must receive. As mentioned in [TS-23.501], the 5QI to QoS mapping is provided by the 5G QoS profile, which includes parameters such as priority level, packet delay budget, packet error rate, etc.

[RFC9543] focuses on how network slices can be instantiated, managed, and monitored by utilizing existing IETF protocols and models. It introduces the concept of the IETF Network Slice Controller (NSC), which interacts with higher-level Network Management Systems (NMSs) and orchestrates network resources to create network slices. The NSC may interact with other network controllers (including Path Computation Element (PCE)), to manage and optimize the underlying network.

[I-D.ietf-teas-5g-ns-ip-mpls] discusses the mapping between the 5G QoS framework and the Differentiated Services (DiffServ) model. The DiffServ model uses the DSCP, a 6-bit field in the IPv4 or IPv6 packet header, to classify and prioritize traffic. The mapping between 5QI and DSCP enables the proper handling and forwarding of packets based on their corresponding QoS requirements.

To achieve this mapping, the 5G system should have a pre-configured mapping table that associates each 5QI value with a specific DSCP value. When a User Plane Function (UPF) in the 5G system receives packets from a data flow with a specific 5QI, it will consult the mapping table and mark the packets with the appropriate DSCP value

before delivering the flow to the network. This marking allows the network to treat and forward the packets according to their QoS requirements based on the DiffServ model.

In summary, QoS in the context of network slicing ensures that each slice meets its specific performance requirements. The 5QI is used to differentiate QoS characteristics in 5G systems, and its mapping to DSCP enables the network to classify and prioritize traffic according to their QoS requirements based on the DiffServ model.

4. 5G user traffic classes types

4.1. Scope of the Transport Network

The 5G System leverages on the transport network to deliver the traffic flows and interconnect its components. The connectivity between the radio base station (i.e., gNB) and the UPF is tunneled using GTP. It is at the UPF where the GTP tunnel is terminated and where the different 5G flows can be handled according to its corresponding 5QI. Thus, traffic to and from other UPF or an external Data Network (DN) can be marked accordingly by means of corresponding DSCP values.

Assuming that both segments, i.e. gNB to UPF, and UPF to DN, can be implemented by means of an IETF Network Slice Service, this implies that forwarding of the 5G flows can be aware or not of the expected service QoS. [I-D.ietf-teas-5g-ns-ip-mpls] provides more details about 5QI-aware and -unaware connectivity models.

4.2. Example of grouping

In order to handle of the variety of 5QI (and/or QCI) types in the network, it is necessary to associate some 5QI values to the limited number of queue classes present in the network elements. An strategy to do so is to group different 5QI types in classes based on their main Service Level Objectives, nominally the corresponding expected latency, packet loss requirement and traffic type (i.e., guaranteed or non-guaranteed bit rate).

For example, the following grouping could be considered:

* 5QI/QCI Group 1: flows with 5QIs showing low latency (< 20 ms) and packet loss in the range 10^-4 to 10^-6, corresponding to 5QIs 80, 82, 83, 84, 85, 86.

* 5QI/QCI Group 2: flows with 5QIs showing moderate latency values (< 100 ms) with diverse packet loss levels, corresponding to 5QIs 3, 65, 69, 75, 79.

* 5QI/QCI Group 3: rest of 5QI of GBR type.

* 5QI/QCI Group 4: rest of 5QIs of non-GBR type.

As result, the following table shows the resulting grouping example in terms of concerned 5QI / QCI values. The table also considers fronthaul traffic as the highest priotiy one, being fronthaul not related to 5QI / QCIs.

Queue PQ	5QI	5QI Group	DSCP (DSCPXX)	Traffic flow example CPRI (RoE), eCPRI CU-P
NPQ-6	80	1	CS5 (DSCP40)	Low Latency eMBB,AR/VR
NPQ-6	82	1	EF (DSCP46)	Discrete Automation small packets
NPQ-6	83	1	EF (DSCP46)	Discrete Automation big packets
NPQ-6	84	1	EF (DSCP46)	Intelligent Transport Systems
NPQ-6	85	1	EF (DSCP46)	Electricity Distribution
NPQ-6	86	1	CS5 (DSCP40)	V2x Collision Avoidance
NPQ-3	3	2	AF41 (DSCP34)	Real Time Gaming, V2X
NPQ-3	65	2	AF42 (DSCP36)	Mission Critical PTT (MCPTT)
NPQ-3	69	2	AF43 (DSCP38)	Mission critical delay sensitive
NPQ-3	75	2	AF42 (DSCP36)	V2X messages over MBMS bearer
NPQ-3	79	2	AF41 (DSCP34)	V2x Messages
NPQ-2	1	3	AF32 (DSCP28)	Conversational Voice
NPQ-2	2	3	AF32 (DSCP28)	Conversational Video
NPQ-2	4	3	AF33 (DSCP30)	Non-Conversational Video
NPQ-2	66	3	AF31 (DSCP26)	Mission Critical PTT Voice
NPQ-2	67	3	AF31 (DSCP26)	Mission Critical Video UP
NPQ-2	87	3	AF32 (DSCP28)	Interactive Service Motion Track Data
NPQ-2	88	3	AF32 (DSCP28)	Int. Ser. AI/ML image recognition
NPQ-2	89	3	AF33 (DSCP30)	Visual content rendering small pck
NPQ-2	90	3	AF33 (DSCP30)	Visual content rendering big pck
NPQ-0	5	3	AF11 (DSCP10)	IMS Signalling
NPQ-0	6	3	AF11 (DSCP10)	TCP-Based signalling,buffered
NPQ-0	7	3	AF11 (DSCP10)	Voice, 100ms Video streaming, Gaming
NPQ-0	8	3	AF12 (DSCP12)	300ms Video streaming, Gaming
NPQ-0	9	3	AF12 (DSCP12)	300ms Video streaming, Gaming
NPQ-0	10	3	AF13 (DSCP14)	1100ms Video streaming, Gaming
NPQ-0	70	3	AF11 (DSCP10)	Mission critical Data

Figure 1: 5QI and (O)-RAN traffic grouping example

This strategy has been also proposed in [ORAN-WG9].

It should be noted that the grouping exercise above is just simply an example on how this methodology could be exploit by network operators at the time of handling traffic of different types entering the network. It is not intended to influence the way in which external systems (e.g., 3GPP or O-RAN) identify their traffic types. At the time of grouping, different criteria can be followed according to the network operator interests.

5. Co-existence of service traffic classes in multi-service networks

Service provider networks are nowadays typically multiservice. It means, they carry different categories of traffic, like, for example, business traffic, residential traffic, mobile traffic, and so on. Moreover, each category of the traffic might further have different flow types. Again, examples are residential voice (residential phone service implemented via VoIP - voice over IP), IPTV, best effort Internet, etc.

Therefore, it is expected that 5G mobile traffic, and other traffic might be mixed over the same transport infrastructure. Appropriate resource allocation and QoS strategy is required to ensure that SLOs for traffic with more demanding requirements are met. This is especially important during network failures and traffic rerouting. Such events should not negatively impact priority traffic (e.g. voice or mobile signaling), but may impact less important traffic (e.g. best effort Internet)

Typical router hardware has 8 queues. Thus, the large number of flows, with various SLO requirements must be squeezed into maximum 8 queues. In addition to 5G user plane 5QI grouping discussed in Section 4.2, other flows occurring in the network must be taken into account. Table 1 provides an example of typical flows - together with their very high level latency/jitter requirements - that can be observed in the multiservice transport network used to transport 4G/5G flows, and residential/bussines services.

Flow type	Per-hop latency	Per-hop jitter
CIPRI (RoE)	~1-20 μ s	~1-20 μ s
eCPRI CU-plane	~1-20 μ s	~1-20 μ s
OAM with aggressive timers	~1 ms	~1 ms
5QI/QCI Group 1	~1 ms	~1 ms
Low latency traffic	~1 ms	~1 ms
Network Control	~5 ms	~1-3 ms
4G/5G C-plane and M-plane	~5 ms	~1-3 ms
5QI/QCI Group 2	~5 ms	~1-3 ms
5QI/QCI Group 3	~10 ms	~5 ms
Guaranteed business traffic	~10 ms	~5 ms
5QI/QCI Group 4	~10-50 ms	~5-25 ms
Best effort	none	none

Figure 2: High-level latency estimations

Note: Per-hop latency includes all latency contributors of the transport node, which includes frame transmission delay, self-queueing delay, queuing delay, store-and-forward delay, etc. Values specified in the table are very raw, high-level sample estimations. Exact per-hop requirements depend on the overall network budget, number of hops, budget allocated to fibers, etc. The table intends to emphasize only relative order of magnitude for per-hop latency/jitter to illustrate the process of assigning traffic to QoS queues.

Both Common Public Radio Interface (CPRI), transmitted in Ethernet frames using Radio over Ethernet (RoE) encapsulation, as well as eCPRI Control and User plane (CU-plane), which uses Ethernet frames or IP packets, have very strict latency/jitter requirements, expressed in microseconds.

Next are low latency (lower miliseconds) flows, like Operations, Administration and Maintenance (OAM) with aggressive (miliseconds) timers. Typical examples here are single-hop Bidirectional Forwardig Detection (BFD) sessions with, e.g., 3x10 miliseconds (or lower) end-to-end timers to monitor reachability between directly connected IGP neighbors, or, CFM (Connectivity Fault Management) frames, again with few miliseconds timers, monitoring direct connections. 5QI/QCI Group 1, as well as residential/business low latency traffic has similar latency requirements.

Traffic with medium latency requirements is network control (OSPF, IS-IS, BGP, LDP, PTP aware-mode, ...), mobile control and management plane (C-plane, M-plane), 5QI/QCI Group 2 traffic, as well as OAM with relaxed (100ms to seconds) timers. Typical example of OAM with relaxed timers are multi-hop BFD packets, with e.g., 3x100 miliseconds (or higher) end-to-end timers to monitor reachability of multi-hop BGP sessions. Alos, worth to note is, that only PTP with

physical layer time stamping is recommended for 5G applications, as PTP without physical layer time stamping accommodates to much jitter on the end-to-end path between grand master and the client. Jitter of PTP packets with physical layer time stamping is properly accounted based on time stamps, without the need to treat PTP as strict priority traffic. However, QoS features should ensure that PTP packets are not dropped during congestion.

Traffic sustaining higher latency is guaranteed business traffic, as well 5QI/QCI Group 3 traffic.

And, finally, 5QI/QCI Group 4 and other best effort traffic does not have any specific latency requirements - it is simply served as best effort, if the resources are still available after serving higher priority traffic flows discussed earlier.

Depending on the hardware support, there are many QoS models available in the transport nodes. It is out-of-scope for this document to discuss traffic flow mappings to QoS queues in all possible QoS models. However, examples of two most common models are reviewed for reference.

5.1. QoS model with single priority queue

In this model, one of the queues is a priority queue, and remaining queues are non-priority queues. Non-priority queues are served only, if the priority queue is empty, which gives strict precedence to priority queue. Non-priority queues are served in a round robin (RR) fashion. Depending on the queueing implementation this can be plain round robin, or weighted round robing (WRR), where non-priority queue with higher weight is served more frequently than non-priority queue with lower weight. This results in lower congestion probability for the queue with higher weight. More advanced scheduling schemes for non-priority queues include weighted deficit round robin (WDRR), or weighted modified deficit round robin (WMDRR). It is out of scope for this document to discuss all possible queue scheduling algoritms. However, the reader is encouraged to read [RFC7806] for more information.

In single priority queue model, example flow to queue mapping is outlined in Figure 3.

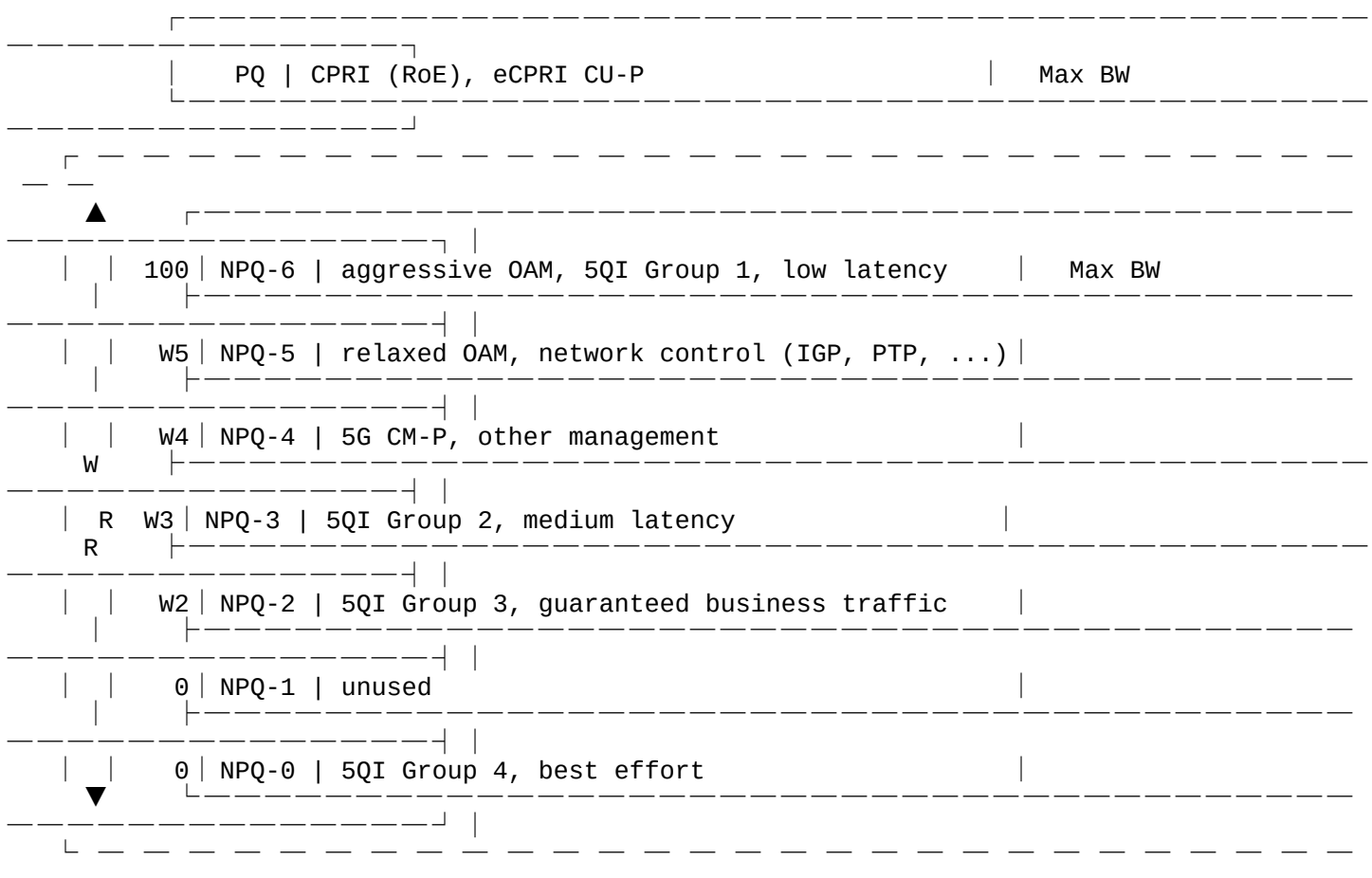

Figure 3: Flow mapping with single priority queue

Note: The numbers and flow grouping indicated in Figure 3 are provided for illustration purposes only and should not be considered as deployment guidance.

Priority queue is used to serve strict priority traffic, with microseconds latency requirements. Therefore, CPRI/RoE and eCPRI control and user plane is mapped to priority queue. This queue is always served before non-priority queues, and only when this queue is empty, non-priority queues are served. This has two implications:

* the latency of packets served via priority queue is lower (lowest possible in given hardware platform), compared to latency of the packets served by non-priority queue

* priority queue can starve non-priority queues, if the traffic volume served by priority queue reaches link capacity.

The first characteristic of priority scheduling is anticipated. However, the second characteristics might cause full drops in non-priority queues. Therefore, when priority queue is used, following two measures must be considered:

* network capacity must be dimensioned in such a way, so that expected maximum CPRI/eCPRI traffic volume does not take entire link capacity. For example, good practice is to dimension the

network so that expected maximum CPRI/eCPRI traffic volume do not exceed certain percentage of link capacity, and perform network upgrade, if the limit is crossed.

* priority queue is policed/rate-limited to the expected maximum CPRI/eCPRI traffic volume plus some small (10-20%) additional threshold (Max BW in Figure 3)

With these measures CPRI/eCPRI traffic can be served without drops and extra latency, while some capacity resources on the link are guaranteed for non-priority traffic.

Non-priority queues are served in WRR (or some sort of more advanced weighted scheduling) manner. Traffic with low latency (miliseconds) range should be served via non-priority queue with considerably (order of magnitude) higher weight comparing to other non-priority queues. This causes very frequent queue servicing, which minimizes the delay of the packets served via this queue, as packets do not need to stay to long in the queue. This is the scheduling behavior similar to priority scheduling, therefore policing/rate-limiting of this queue is strongly recommended to avoid nearly starvation of other non-priority queues.

Remaining traffic flows might be distributed across remaining non-priority queues, grouping the flows with similar characteristics in the same queue, and providing weights based on network dimensioning, taking into account expected traffic volumes. Queue buffer sizes in all cases must be aligned to maximum latency requirements of the traffic flows assigned to the queue. Non-priority queue for the best effort traffic should have lowest possible weight, so that it is served only in the case there is no packet waiting in any other queue.

5.2. QoS model with multiple priority queues

In this model, there are multiple priority queues, serviced strictly in priority order. Remaining, non-priority queues, are serviced in WRR (or some enhanced version of WRR) manner. Example flow to queue mapping using multiple priority QoS model is outlined in Figure 4.

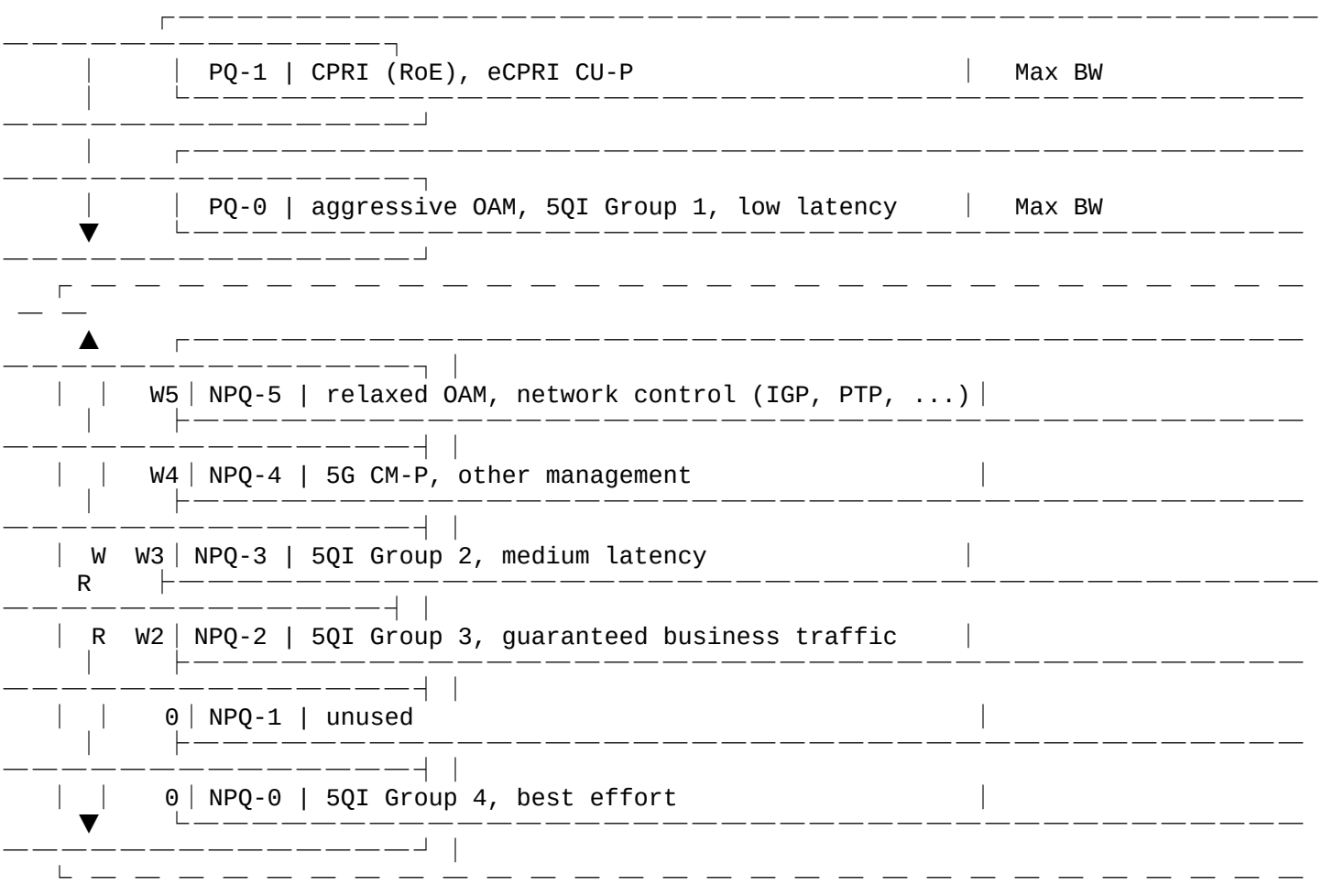

Figure 4: Flow mapping with multiple priority queues

Note: The numbers and flow grouping indicated in Figure 3 are provided for illustration purposes only and should not be considered as deployment guidance.

The main difference comparing to the previous example is the 2nd priority queue (PQ-0), dedicated to low latency flows, like OAM with aggressive timers, or 5GI Group 1 flows. PQ-0 queue is only served, when the PQ-1 queue is empty. Thus, while both PQ-1 and PQ-0 queues are used to serve traffic with low latency requirements, traffic served via PQ-1 will observe smaller latency compared to traffic served via PQ-0. As already discussed previously, rate-limiter/ policer should be used on both priority queues to avoid complete starvation of non-priority queues.

Acknowledgments

The authors thank Ruediger Geib and John Kaippallimalil for the comments received that helped to improve and clarify the document.

The contribution of L.M. Contreras has been partially funded by the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union - NextGenerationEU under the projects 6GBLUR-smart (Ref. TSI-063000-2021-56) and 6GBLUR-joint (Ref. TSI-063000-2021-57), and by the European Commission under the Horizon Europe SNS-JU project UNITY-6G (grant agreement 101192650).

References

Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

Informative References

[I-D.ietf-teas-5g-network-slice-application] Geng, X., Contreras, L. M., Rokui, R., Dong, J., and I. Bykov, "IETF Network Slice Application in 3GPP 5G End-to-End Network Slice", Work in Progress, Internet-Draft, draft-ietf-teas-5g-network-slice-application-05, 7 July 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-teas-5g-network-slice-application-05>.

[I-D.ietf-teas-5g-ns-ip-mpls] Szarkowicz, K. G., Roberts, R., Lucek, J., Boucadair, M., and L. M. Contreras, "A Realization of Network Slices for 5G Networks Using Current IP/MPLS Technologies", Work in Progress, Internet-Draft, draft-ietf-teas-5g-ns-ip-mpls-18, 3 April 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-teas-5g-ns-ip-mpls-18>.

[ORAN-WG9] "O-RAN Xhaul Packet Switched Architectures and Solutions", 28 February 2024, <https://orandownloadsweb.azurewebsites.net/specifications>.

[RFC7806] Baker, F. and R. Pan, "On Queuing, Marking, and Dropping", RFC 7806, DOI 10.17487/RFC7806, April 2016, <https://www.rfc-editor.org/rfc/rfc7806>.

[RFC9331] De Schepper, K. and B. Briscoe, Ed., "The Explicit Congestion Notification (ECN) Protocol for Low Latency, Low Loss, and Scalable Throughput (L4S)", RFC 9331, DOI 10.17487/RFC9331, January 2023, <https://www.rfc-editor.org/rfc/rfc9331>.

[RFC9543] Farrel, A., Ed., Drake, J., Ed., Rokui, R., Homma, S., Makhijani, K., Contreras, L., and J. Tantsura, "A Framework for Network Slices in Networks Built from IETF Technologies", RFC 9543, DOI 10.17487/RFC9543, March 2024, <https://www.rfc-editor.org/rfc/rfc9543>.

[TS-23.203] "3GPP TS-29.203 Policy and charging control architecture", 3 April 2024, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=810>.

[TS-23.207] "3GPP TS 23.207 End-to-end Quality of Service (QoS) concept and architecture", 25 March 2022, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=814>.

[TS-23.501] "3GPP TS 23.501: System architecture for the 5G System (5GS)", 25 March 2022, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>.

[TS-23.502] "3GPP TS 23.502: Technical Specification Group Services and System Aspects; Procedures for the 5G System (5GS);Stage 2 (Release 19)", 26 June 2024, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3145>.

[TS-29.213] "3GPP TS 29.213 Policy and Charging Control signalling flows and Quality of Service (QoS) parameter mapping", 21 March 2022, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1673>.

[TS-29.513]
"3GPP TS-29.513 5G System; Policy and Charging Control signalling flows and QoS parameter mapping; Stage 3", 7 June 2023, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3354>.

Authors' Addresses

Luis M. Contreras (editor)
Telefonica
Email: luismiguel.contrerasmurillo@telefonica.com

Ivan Bykov (editor)
Ribbon Communications
Email: Ivan.Bykov@rbbn.com

Krzysztof G. Szarkowicz (editor)
Juniper Networks
Email: kszarkowicz@juniper.net