International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Y.3113

(02/2021)

SERIES Y: GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS, NEXT-GENERATION NETWORKS, INTERNET OF THINGS AND SMART CITIES

Future networks

Requirements and framework for latency guarantee in large-scale networks including the IMT-2020 network

Recommendation ITU-T Y.3113

ITU-T Y-SERIES RECOMMENDATIONS

GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS, NEXT-GENERATION NETWORKS, INTERNET OF THINGS AND SMART CITIES

GLOBAL INFORMATION INFRASTRUCTURE	
General	Y.100–Y.199
Services, applications and middleware	Y.200–Y.299
Network aspects	Y.300–Y.399
Interfaces and protocols	Y.400–Y.499
Numbering, addressing and naming	Y.500–Y.599
Operation, administration and maintenance	Y.600–Y.699
Security	Y.700–Y.799
Performances	Y.800–Y.899
INTERNET PROTOCOL ASPECTS	
General	Y.1000–Y.1099
Services and applications	Y.1100–Y.1199
Architecture, access, network capabilities and resource management	Y.1200–Y.1299
Transport	Y.1300–Y.1399
Interworking	Y.1400–Y.1499
Quality of service and network performance	Y.1500–Y.1599
Signalling	Y.1600–Y.1699
Operation, administration and maintenance	Y.1700–Y.1799
Charging	Y.1800–Y.1899
IPTV over NGN	Y.1900–Y.1999
NEXT GENERATION NETWORKS	
Frameworks and functional architecture models	Y.2000–Y.2099
Quality of Service and performance	Y.2100–Y.2199
Service aspects: Service capabilities and service architecture	Y.2200–Y.2249
Service aspects: Interoperability of services and networks in NGN	Y.2250–Y.2299
Enhancements to NGN	Y.2300–Y.2399
Network management	Y.2400–Y.2499
Network control architectures and protocols	Y.2500–Y.2599
Packet-based Networks	Y.2600–Y.2699
Security	Y.2700–Y.2799
Generalized mobility	Y.2800–Y.2899
Carrier grade open environment	Y.2900–Y.2999
FUTURE NETWORKS	**Y.3000–Y.3499**
CLOUD COMPUTING	Y.3500–Y.3599
BIG DATA	Y.3600–Y.3799
QUANTUM KEY DISTRIBUTION NETWORKS	Y.3800–Y.3999
INTERNET OF THINGS AND SMART CITIES AND COMMUNITIES	
General	Y.4000–Y.4049
Definitions and terminologies	Y.4050–Y.4099
Requirements and use cases	Y.4100–Y.4249
Infrastructure, connectivity and networks	Y.4250–Y.4399
Frameworks, architectures and protocols	Y.4400–Y.4549
Services, applications, computation and data processing	Y.4550–Y.4699
Management, control and performance	Y.4700–Y.4799
Identification and security	Y.4800–Y.4899
Evaluation and assessment	Y.4900–Y.4999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T Y.3113

Requirements and framework for latency guarantee in large-scale networks including the IMT-2020 network

Summary

Recommendation ITU-T Y.3113 specifies requirements and a framework for effective and efficient solutions for latency guarantee and cooperation among heterogeneous quality of service (QoS) domains.

For a latency guarantee in multi-domain large-scale networks, it is necessary to clarify how the numerous data-plane functional entities should be arranged and operate in conjunction with each other. In order for the solution to be both effective and efficient, selecting an appropriate traffic granularity for network treatment is essential. The variety of granularity of flow aggregates (FAs), between flow and class, should be taken into consideration.

On the Internet or the IMT-2020 network, there are inevitably multiple domains with possibly different QoS architectures. Even with multiple heterogeneous domains, there should be an underlying unified resource reservation and admission control functions, while the data plane functions should be based on FAs and appropriate regulations in the middle of an end-to-end path.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T Y.3113	2021-02-13	13	11.1002/1000/14595

Keywords

Flow aggregate, IMT-2020, latency guarantee, large-scale network, quality of service.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

© ITU 2021

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

Recommendation ITU-T Y.3113

Requirements and framework for latency guarantee in large-scale networks including the IMT-2020 network

1 Scope

This Recommendation specifies requirements and a framework for latency guarantee in large-scale networks, including the IMT-2020 network, as follows:

– the requirements for achieving latency guarantee in large-scale networks including the IMT-2020 network;

– overall framework and functional entities, and their interworking to achieve latency guarantee, effectively and efficiently.

Routing and upper layer functions lie outside the scope of this Recommendation.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.2111] Recommendation ITU-T Y.2111 (2011), *Resource and admission control functions in next generation networks*.

[ITU-R M.1645] Recommendation ITU-R M.1645 (2003), *Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000*.

[ITU-R M.2083] Recommendation ITU-R M.2083 (2015), *IMT vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 customer premises equipment [b-ITU-T E.800]: Telecommunications equipment located at the customer installation on the customer side of the network interface.

3.1.2 service provider [b-ITU-T E.800]: An organization that provides services to users and customers.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 aggregation domain: A set of relay nodes in the path, for a flow, between the aggregation and segregation points.

3.2.2 domain: A set of relay nodes and end-hosts under a single administrative control or within a closed group of administrative control; these include campus wide networks, private wide area networks, and IMT-2020 networks.

NOTE – This definition references the description in the Introduction to [b-IETF RFC 8655].

3.2.3 IMT-2020: Systems, system components, and related technologies that provide far more enhanced capabilities than those described in [ITU-R M.1645].

NOTE 1 – [ITU-R M.1645] defines the framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000 for the radio access network.

NOTE 2 – Based on [ITU-R M.2083].

3.2.4 large-scale network: A network or a set of networks, whose longest end-to-end path includes 16 or more relay nodes.

3.2.5 relay node: A node supporting relay functionality that acts as an intermediary node, through which other nodes can pass their traffic (e.g., router, switch or gateway).

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

5G	fifth Generation
5QI	5G QoS Identifier
AN	Access Network
ATS	Asynchronous Traffic Shaping
CN	Core Network
CPE	Customer Premises Equipment
DiffServ	Differentiated Services
DL	Downlink
DN	Data Network
DRR	Deficit Round Robin
e2e	end to end
FA	Flow Aggregate
FIFO	First In, First Out
GBR	Guaranteed Bit Rate
GFBR	Guaranteed Flow Bit Rate
IntServ	Integrated Services
IP	Internet Protocol
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
IR	Interleaved Regulator
MDBV	Maximum Data Burst Value
MFBR	Maximum Flow Bit Rate
NGBR	Non-GBR
NG-RAN	Next Generation-Radio Access Network
PDB	Packet Delay Budget
PDU	Protocol Data Unit

PGPS Packetized Generalized Processor Sharing

QFI QoS Flow Identifier

QoS Quality of Service

RAN Radio Access Network

RSpec Request Specification

SDF Service Data Flow

SMF Session Management Function

TDM Time Division Multiplexing

TSpec Traffic Specification

UE User Equipment

UL Uplink

UPF User Plane Function

uRLLC ultra-Reliable Low Latency Communication

5 Conventions

None.

6 Introduction

The IMT-2020 network is expected to be able to support a variety of different services and applications. Among them, the ultra-reliable and low-latency communication (uRLLC) service requires an end-to-end (e2e) network latency bound [ITU-R M.2083]. A network latency guarantee is also one of the most important requirements for the 6G network [b-ITU-T Y-Suppl.66]. For example, it is required to have a 5 ms latency upper bound for the tactile Internet. Remote industrial management control and remote robotic surgery are example applications of the tactile internet. On the Internet, when measured in 1998, 90% of hosts were within fewer than 18 hops from the University of California at Los Angeles [b-Fei]. The Internet in 2030 should not be smaller than this measurement. As such, a large-scale network is defined to include 16 or more e2e hops in this Recommendation.

Different networks on the Internet including the IMT-2020 network, which is further divided into access network (AN), core network (CN) and data network (DN), have different quality of service (QoS) models and components. A QoS model includes the mechanisms, functions, description of QoS parameters information, and how that information should be treated and interpreted on the network. An integrated approach is necessary, which makes the various domains in the Internet work together as a whole.

Legacy technologies, based on either microscopic flows or simple classes, are too complex to be implemented on a large-scale network, or ineffective in an arbitrary network topology. Time-synchronized approaches, which are similar to time division multiplexing (TDM), taken in IEEE Time Sensitive Network (TSN) task group (TG) are not scalable to IMT-2020 networks or the current Internet.

In order for the solution to be both effective and efficient, selecting appropriate traffic granularity for network treatment is essential. The various granularity levels of flow aggregates (FAs), between flow and class, should be taken into consideration, and used as a basis for QoS control.

Therefore, this Recommendation specifies requirements and a framework for a latency guarantee in large-scale networks.

7 Requirements

Req_1. It is required to be able to specify, by the customer premises equipment (CPE) or on behalf of the CPEs with insufficient signalling capabilities, the CPE flow destination and characteristics (e.g., average data rate and maximum burst size).

Req_2. It is recommended to be able to specify, by the CPE or on behalf of the CPEs with insufficient signalling capabilities, the CPE desirable e2e latency upper bound.

NOTE 1 – For detailed information about the QoS negotiation process, see [ITU-T Y.2111].

Req_3. It is required that a network be able to determine the latency upper bound within the network, of a traversing flow, with the flow destination and the characteristics specified.

Req_4. It is recommended that the means to provide the latency upper bound be: 1) implementable in CNs where there are millions of active flows at an output port; 2) applicable to an arbitrary network topology; 3) of minimal effect on the average latency and the throughput; and 4) scalable to a large-scale network.

Req_5. It is recommended that the latency upper bound be susceptible to negotiation, for a flow, between the CPE and the service provider.

NOTE 2 – The latency upper bound negotiation can be a two-way handshake, or a more complex process. The simplest negotiation is that of the integrated services (IntServ). A CPE with signalling capability specifies its flow request specification (RSpec) and traffic specification (TSpec). Based on these specifications, the service provider decides whether the CPE-requested upper bound can be met. If not, the provider is required to deny the admission. The latency bound negotiation can be more complex, in which a renegotiation process is included. If a first negotiation fails, the flow may restart with a new TSpec. For another example, the QoS provisioning is based on the network allowance. This means that an individual flow does not specify its latency bound requirement (RSpec in IntServ). Rather, as a flow determines its TSpec (e.g., its burst size and input rate), then based on the best e2e path among those can be provided, the feasible latency bound is calculated and notified to the flow. The flow decides whether to accept it.

Req_6. It is recommended that the dynamic latency upper bound be susceptible to negotiation.

NOTE 3 – Dynamic negotiation refers a process in which already accepted flows and their upper bounds are reconsidered as new flows that are coming into the networks. It is necessary because as a network state changes, the guaranteed latency bound of already accepted flows can be also changed.

Req_7. It is required that networks be able to handle FAs as control elements.

Req_8. It is required that networks be able to aggregate and segregate flows at any desired points in a network or equivalently within an aggregation domain.

Req_9. It is required that within an aggregation domain, flows be aggregated according to an aggregation domain-specific rule (e.g., a rule that flows with the same input and output ports of the domain are aggregated into a single FA).

Req_10. It is recommended that aggregation domains be susceptible to merger and division on demand.

Req_11. It is required that flow aggregation rules be susceptible to negotiation among the aggregation domains.

Req_12. It is recommended that the TSpec include: a maximum burst size; an average input rate; a peak rate; and a maximum packet size.

Req_13. It is recommended that best-effort service traffic not affect the latency bound of high-priority flows.

Req_14. It is required that a traffic regulation capability be provided at the boundary of aggregation domains.

8 Framework

This Recommendation specifies a framework to overcome the current limitations that are analysed in Appendix I. In the proposed framework, flows are aggregated according to their {input, output port} pair in an aggregation domain, and minimal interleaved regulators (IRs) per FA are implemented at the boundary of the aggregation domain. The minimal IRs at aggregation domain boundaries suppress the burst accumulations with latency upper bounds intact [b-Le Boudec]. The conditions under which a minimal IR does not increase the upper latency bound can be summarized as follows.

1) Every flow into the aggregation domain conforms to an arrival curve with parameters {average arrival rate, maximum burst size}.
2) The aggregation domain outputs all the packets first in, first out (FIFO).
3) The IR regulates every flow to reproduce the arrival characteristics at the ingress of the aggregation domain.
4) (Minimal IR) An IR transmits immediately when a packet at the head of queue meets the output condition. Such an IR is called a minimal IR.
5) An IR provides zero latency, including transmission latency, for packets satisfying the output condition. For example, if a packet comes in when the queue is empty, it can be cut-through if it is eligible to leave.

The following e2e latency guarantee framework is specified based on the conditions described in the foregoing.

– Flows are divided into high priority and low priority.
– Low priority flows are put in a single FIFO queue at the output port of all nodes and processed in strict priority mode with preemption.
– High-priority flows are handled as follows.
  A. Select an appropriately sized aggregation domain.
  B. The flows with the same {input, output port} in the aggregation domain are aggregated into a single FA.
  C. In a node, a fair queuing-based scheduling is performed per FA with a queue for each FA. This operation lets the aggregation domain for an FA become a FIFO system in condition 2).
  D. Install minimal IRs per FA at the boundaries of aggregation domains. See Figure 1.
  E. Only the flows conforming to the initial arrival characteristics or the flows from the IR are accepted into an aggregation domain.
  F. The aggregation domains with IRs are interconnected to form an entire network. See Appendix II for implementation practice.

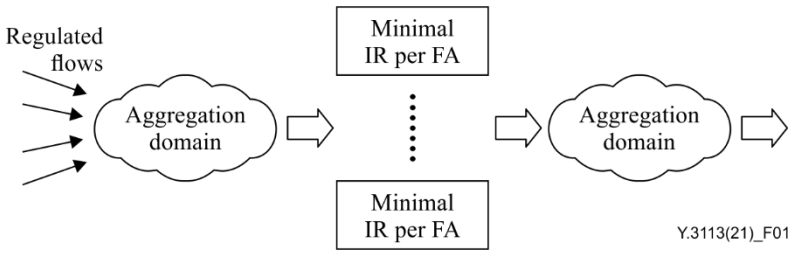

Figure 1 – Framework for latency guarantee

9 Security considerations

The QoS management of IMT-2020 network includes user equipment (UE), ANs, and CN that are subject to security and privacy measures. Sensitive information should be protected as a high priority in order to avoid leakage and unauthorized access. Security and privacy concerns should be aligned with the requirements specified in [b-ITU-T Y.2701] and [b-ITU-T Y.3101].

Appendix I

Analysis of existing solutions and the proposed framework

(This appendix does not form an integral part of this Recommendation.)

Flow-based approaches such as the IntServ framework are known to provide a latency bound guarantee when the following three conditions are met.

1) Every flow conforms to its predefined characteristics {the average input rate, maximum burst size}.
2) The sum of average input rates of flows in every output port does not exceed the link capacity of the port.
3) Every flow is guaranteed to receive a service rate not less than the input rate.

Conditions 1) and 2) make resource reservation and admission control mandatory. Condition 3) requires a scheduler appropriate to the flows . IntServ scheduling complexity is proportional to the number of flows, which grows to millions in CNs. This scheduling complexity prohibits the IntServ from being implemented in real networks.

Therefore, the differentiated services (DiffServ) framework has been proposed, which provides relative performance differentiation with 8 or 32 queues for each priority class. Flows belonging to a class are put into a queue. The queues are served with strict priority. Because of such a simplification, DiffServ has been adopted in the current Internet. However, the maximum burst of a flow increases proportionally to the sum of all maximum bursts of flows within a queue. When there is a cycle in a network, such as a mesh network, the maximum burst grows to infinity, as does the latency bound. The DiffServ framework does not provide a latency bound in a general topology network.

IEEE 802.1 TSN [b-IEEE TSN] aims to guarantee a latency upper bound in a single domain network, therefore a small scale network only. The major part of the effort in TSN depends on the time synchronization function across the network, and the synchronized behaviour of the nodes. The asynchronous traffic shaping (ATS) [b-IEEE 802.1Qcr] technique presented in TSN employs a node with an output port with IRs for each input port, and a strict priority class-based FIFO system side by side [b-Specht]. An IR is a single queue system that examines the packet at the head of the queue and lets it leave as soon as it is qualified according to the regulation rule of the flow to which the packet belongs. The remainder of the packets in the queue may be delayed, even when they are already qualified. However, it is proven that a minimal IR does not increase the latency upper bound of the associated FIFO system, such as the class-based FIFO system employed in ATS [b-Le Boudec]. However, the ATS framework requires an IR for each input port at every output port of every node. Frequent regulation significantly affects statistical performance, such as average latency. It also implies increased implementation cost.

IntServ, DiffServ, TSN synchronous approaches and TSN ATS have their own shortcomings when employed in large scale multi-domain networks. A new framework is required, which is less complex than IntServ and has better statistical performance than ATS. The framework has to provide latency bounds in arbitrary topology networks. It also has to be scalable to a large-scale network.

The framework proposed in this Recommendation satisfies all the requirements mentioned in the foregoing. It has a lower scheduling complexity compared to flow-based frameworks, since the number of FAs in a scheduler is certainly less than the number of flows. Compared to the ATS, the reduced number of regulators of the suggested framework contributes to reduced cost, as well as improved average latency. A performance analysis with a simple exemplary network is provided in Appendix III. For further details, see [b-Joung]. It is shown that the proposed framework gives a better latency bound than the IntServ and the ATS.

Appendix II

Implementation practice of the framework

(This appendix does not form an integral part of this Recommendation.)

Figure II.1 depicts a practical implementation of the proposed framework. It contains an aggregation domain (called a "domain" in this appendix) and an ingress edge node of the adjacent domain. Only high-priority traffic is depicted. Every node fairly schedules the FAs, which are aggregated according to the input-output port of the domain. An FA in the domain is fed to the next domain, at whose ingress edge node the minimal IRs are implemented and the FA is regulated. If minimal IRs were located in the egress edge node, then the scheduled packets according to a fair scheduler should be redistributed into different IRs. The eligible times and corresponding transmission times of packets from different IRs would overlap, which causes violation of the non-zero latency condition. Exactly as in the ATS framework, it is assumed that zero latency may be provided with a switch module in a node, e.g., with an infinitely large bandwidth of the switch module.

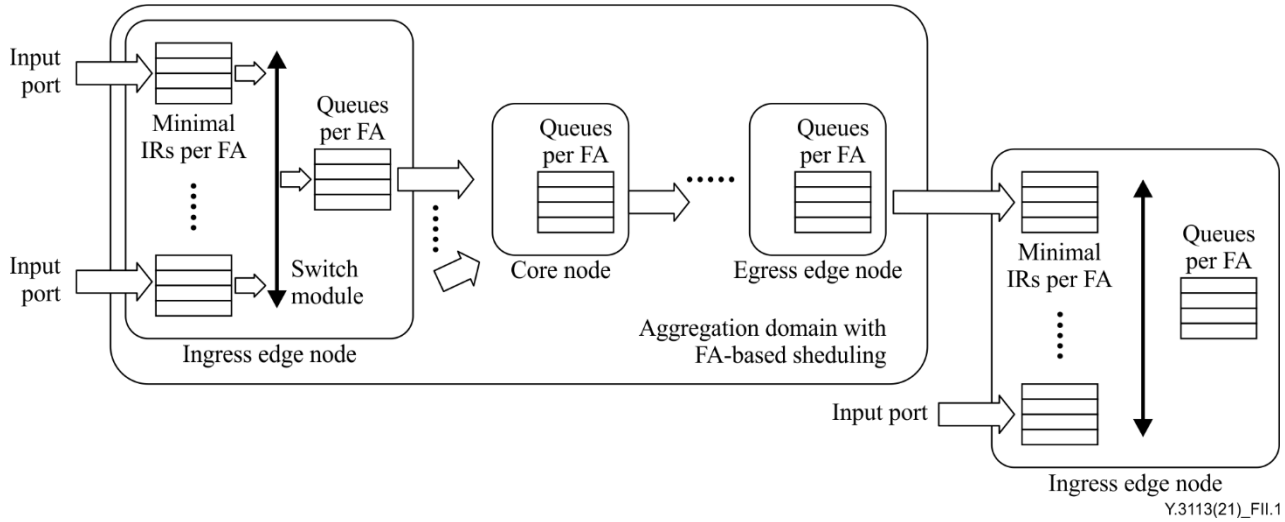

Figure II.1 – Example architecture of the nodes in the proposed latency guarantee framework

Appendix III

Latency bounds comparison of IntServ, TSN ATS, and the proposed framework in example networks

(This appendix does not form an integral part of this Recommendation.)

An example network for the description of the proposed framework is depicted in Figure III.1, in which minimal IRs are implemented between aggregation domains (called "domains" in this appendix). Assume the internetwork in Figure III.1 is perfectly symmetrical. A flow travels *d* domains, with identically *h* hops in a domain, making the total number of hops the flow travels *E=hd*. The critical design choice in this architecture would be the value of *h* (and thus *d*), given *E*.

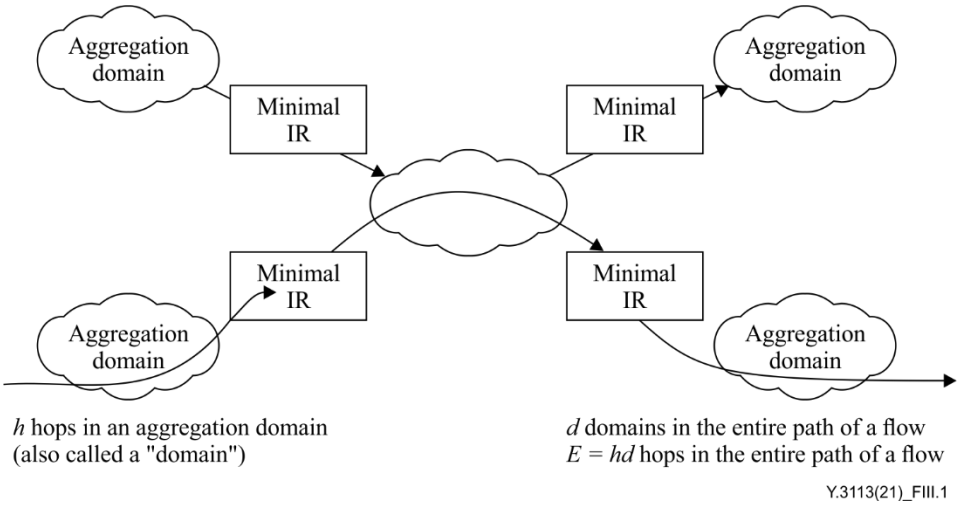

Figure III.1 – An example network architecture of the proposed framework

The numerical performance analysis of the proposed framework is presented. The symbols for the parameters frequently used in the analysis are given in Table III.1.

Table III.1 – Mathematical Symbols used in the analysis

Symbol	Quantity
L_i	Maximum packet length of flow i
$L_{\max}$	Maximum packet length of all the flows in a scheduler
r	Link capacity
ρ_i	Arrival rate of flow i
σ_i	Maximum burst size of flow i
φ_i	Quantum value assigned for flow i
Θ_i^S	Latency of flow $\boldsymbol{i}$ at scheduler S
h	Number of hops in a network
n	Number of flows in a flow aggregate
p	Number of ports in a node
d	Number of networks in a flow path

If a flow i traverses only the latency rate (LR) schedulers S_j in its path (with a total of k LR schedulers), then the e2e latency experienced by the packets in the flow is bounded by Inequality III.1 [b-Stiliadis].

$$D_i \leq \frac{\sigma_i - L_i}{\rho_i} + \sum_{j=1}^{k} \Theta_j{}^{S_j} \quad \text{(III.1)}$$

The *latency* (written in italic) defined in the LR server context differs from the term used elsewhere in this Recommendation. Latency in non-italic is synonymous with delay. The *latency* of a scheduler in an LR server can be interpreted to be a maximum time a flow may have to wait, from the start of a busy period, to be served with its allocated service rate. The packetized generalized processor sharing (PGPS) is an ideal but complex LR scheduler. PGPS *latency* is given by Expression III.2 [b-Stiliadis].

$$\Theta_i{}^{\mathrm{PGPS}} = \frac{L_i}{\rho_i} + \frac{L_{\max}}{r} \quad \text{(III.2)}$$

The deficit round robin (DRR) is the representative round-robin LR scheduler with reduced complexity. The *latency* of a DRR scheduler, if quantum values are smaller than the maximum packet length, is given by Expression III.3 [b-Lenzini].

$$\Theta_i{}^{\mathrm{DRR}} = \frac{1}{r}\left[(F - \varphi_{\mathrm{i}})\left(1 + \frac{L_i}{\varphi_i}\right) + \sum_{n=1}^{N} L_n\right] \quad \text{(III.3)}$$

where F is the sum of all quantum values (φ_i) of active flows in the scheduler, and N is the number of active flows. Quantum refers to the amount of data serviced at one time, which is determined in proportion to the service rate allocated to each flow [b-Shreedhar].

A FIFO scheduler is also an LR scheduler with the *latency* given by Expression III.4.

$$\Theta_i{}^{\mathrm{FIFO}} = \frac{1}{r}\left[\sum_{n=1}^{N} \sigma_n\right] \quad \text{(III.4)}$$

where N is the number of active flows.

Consider an aggregation domain in which all the flows have the same characteristics and have to pass the same number of hops h to depart, as shown in Figure III.2. Every node has p inputs and p output ports. A total of np^h flows are input into an ingress port, and among them np^{h-1} flows are output from the same port used for egress. On the second node, np^{h-2} of these flows are output from the same egress port, and on the last node, $np^{h-h} = n$ of the flows exit from the same egress port. Therefore, there are n flows having the same pair of {input and output ports} in the domain. Suppose this input/output (I/O) pattern occurs on all nodes.

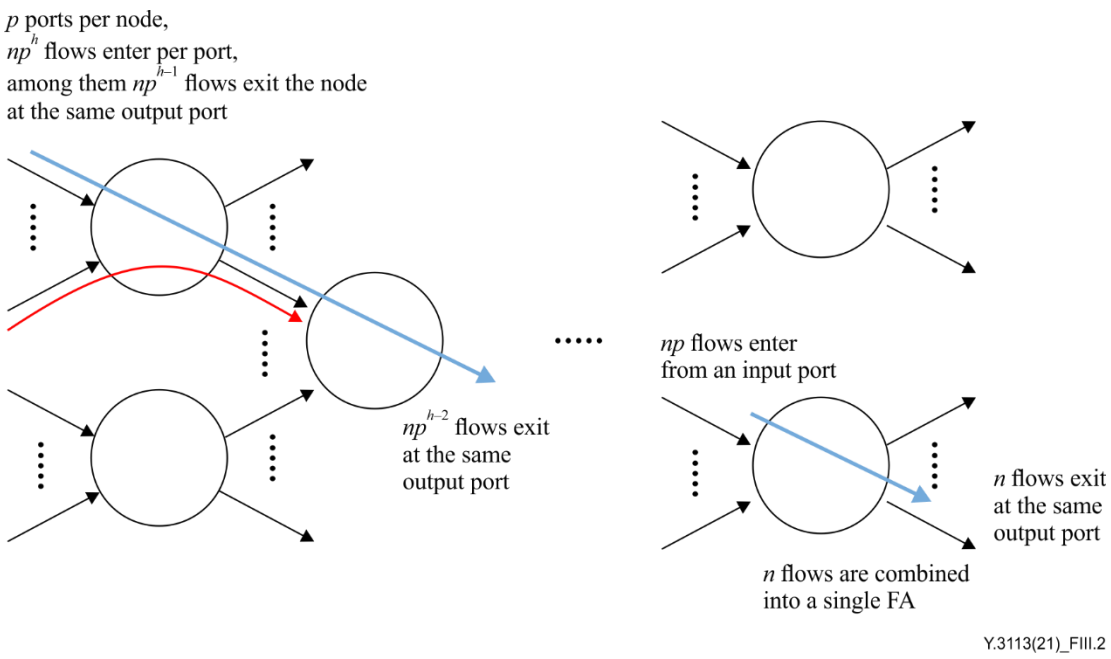

Figure III.2 – Flow aggregation architecture within an aggregation domain of the example network

From Expressions III.1 and III.2, the domain latency of the flow-based framework with the PGPS schedulers is given by

$$D_i^{F\_PGPS} \le \sum_{j=1}^{k} \Theta_j^{F\_PGPS} = \frac{h(np^h+1)}{r/L}. \quad \text{(III.5)}$$

Assume for simplicity that $\sigma_i = L_i = L_{\max} = L$ and $\rho_i = r/np^h$, for all i. Similarly, for an FA-based framework with the PGPS, there are p^h FAs in an output port, and the maximum burst of an FA is nL, therefore

$$D_i^{FA\_PGPS} \le \frac{nL-L}{\rho_i} + \sum_{j=1}^{k} \Theta_j^{FA\_PGPS} = \frac{(h+n-1)p^h+h}{r/L}. \quad \text{(III.6)}$$

Consider a case where the domain in Figure III.2 employs the ATS framework, i.e., FIFO and minimal IRs at every node. In this case Expression III.1 is applied to a single node, since the minimal IR is not an LR scheduler. Assume for simplicity that $\sigma_i = L_i = L_{\max} = L$ again. We have np^h flows in an output port, whose burst sizes are all L. The burst size of the aggregated flows at the FIFO scheduler is therefore np^hL. Using Expressions III.1 and III.4) gives

$$D_i^{FIFO} \le \frac{np^hL - L}{\rho_i} + \sum_{j=1}^{k} \Theta_j^{FIFO} = \frac{2np^h - 1}{r/L}$$

since we assume $\rho_i = r$ for the highest priority FIFO scheduler. We have h such nodes in a domain, therefore the latency bound of the ATS domain is given by

$$D_i^{ATS} \le \frac{h(2np^h-1)}{r/L}. \quad \text{(III.7)}$$

To elaborate the internetwork in Figure III.1, the number of flows entering a port is denoted, F, and represented by np^h, as in Figure III.2. The internetwork can now be depicted as in Figure III.3.

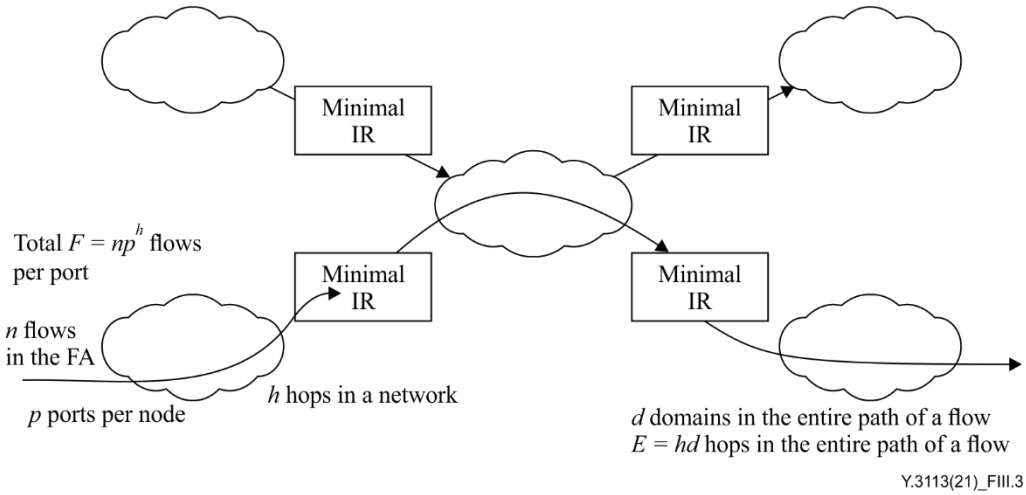

Figure III.3 – The example network, with parameters for the latency bound calculation of the flow under observation

Now consider the e2e latency bound of the internetwork, with fixed values of E and F. Now determine the latency bound while h (and thus n and d) varies. A larger value of h means smaller values of d and n, and a lower number of minimal IRs. If $h = E$, then there is no minimal IR. If $n = 1$, then there is no flow aggregation, which is similar to IntServ framework. A smaller value of h means a smaller domain size and more minimal IRs. If $h = 1$, then IR resides at every node, which is similar to the ATS framework.

Consider the e2e latency bounds of the IntServ, ATS, and the proposed framework. From the relationships between the variables p, F, E, namely $np^h = F$, $hd = E$, can be obtained $d = E/h$, $n = F/p^h$. First, for IntServ, since it has the "pay burst only once" property, from Expression III.5,

$$T_i^{\text{F\_PGPS}} \leq \frac{h(np^h+1)}{r/dL}. \tag{III.8}$$

Second, for the proposed framework, from Expression III.6,

$$T_i^{\text{FA\_PGPS}} \leq \frac{(h+n-1)p^h+h}{r/dL}. \tag{III.9}$$

Similarly, from Expression III.7,

$$T_i^{\text{ATS}} \leq \frac{h(2np^h-1)}{r/dL} \tag{III.10}$$

Now consider a case where $p = 2$, $E = 16$, $F = 65\,536 = 2^{16}$, $r = 1$ Gbit/s, $L = 10$ Kbit. The right hand side (RHS) of Expression (III.9) gives the red curve in Figure III.4. The blue line represents the value of the RHS of Expression (III.8), IntServ, which is 10.486. The value of the RHS of Expression (III.10), ATS (not shown), is 20.97.

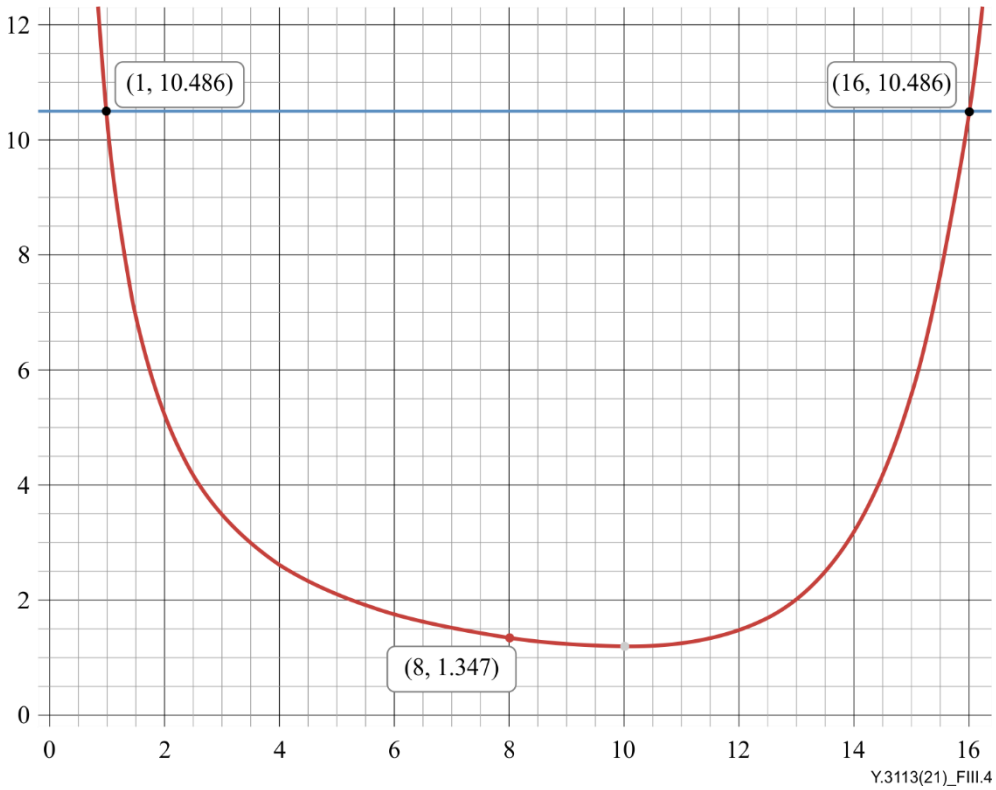

Figure III.4 – Latency bound of the proposed framework – Plot 1

The red curve is plotted with $p = 2$, $E = 16$, $F = 2^{16} = 65\,536$, $r = 1$ Gbit/s, $L = 10$ Kbit, with varying h. When $h = 8$, the latency bound is 1.347 s. When $h = 1$ and $h = 16$, it is 10.486 s, which is identical to the latency bound given with IntServ. The latency bound of the ATS framework is 20.97 s. The red curve reaches a minimum near $h = 10$

Note that all the possible choices of h and d give lower than or equal latency bounds compared to those of IntServ. They are always smaller than those of ATS. The optimal choice in this case is $\{h = 8, d = 2\}$, which gives 1.347 s of latency bound, which is almost eight times better than IntServ, and 16 times better than ATS.

This result is remarkable. By only dividing a path into two parts, putting a minimum IR in the path, and aggregating flows accordingly, the latency bound can be reduced by as much as to 1/8, compared to that of IntServ. The scheduler complexity is reduced by the order of 2^8.

Next, consider a network with $p = 8$, $E = 4$, $F = 8^4 = 4\,096$, $r = 1$ Gbit/s, $L = 10$ kbit/s. This set of parameters represents a smaller sized network with nodes with more ports. The RHS of Expression III.8 gives the red curve in Figure III.5. The blue line represents the value of the RHS of Expression III.9, IntServ, which is 0.164. The value of the RHS of Expression III.10, the latency bound of ATS (not shown), is 0.328.

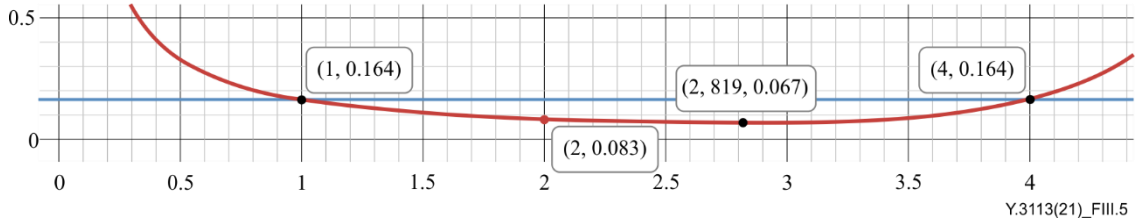

Figure III.5 – Latency bound of the proposed framework – Plot 2

The red curve is plotted with $p = 8$, $E = 4$, $F = 8^4 = 4\,096$, $r = 1$ Gbit/s, $L = 10$ Kbit/s, with varying h. When $h = 2$, the latency bound is 0.083 s. When $h = 1$ and $h = 4$, it is 0.164 s, which is identical to the latency bound given with IntServ. The latency bound of the ATS framework is 0.328 s. The red curve reaches the minimum with $h = 2.819$

The possible sets of parameters in this case are $\{h = 1, d = 4, n = 512 = 8^3\}$, $\{h = 2, d = 2, n = 64\}$, and $\{h = 4, d = 1, n = 1\}$. The optimum choice for the value of h is 2. It means that even in a small network with an e2e hop count of 4, dividing the path into two, aggregating flows accordingly, and inserting a minimal IR per FA would produce a latency bound that is only half that of IntServ. The scheduler complexity is also reduced by the order of 8^2. Table III.2 summarizes the latency bounds of three frameworks, with two network scenarios.

Table III.2 – Latency bounds comparison of the IntServ, the ATS, and the proposed framework

Frameworks	**Network parameters**	
	Core network scenario ($p = 2$, $E = 16$, $F = 2^{16} = 65\,536$, $r = 1$ Gbit/s, $L = 10$ Kbit/s)	**Local network scenario** ($p = 8$, $E = 4$, $F = 8^4 = 4\,096$, $r = 1$ Gbit/s, $L = 10$ Kbit/s)
IntServ	10.486 s	0.164 s
ATS	20.97 s	0.328 s
Proposed framework	1.347 s with $h = 8$	0.083 s with $h = 2$

The proposed framework with minimal IRs and FA scheduling can be seen as a generalized framework that embraces the IntServ and the TSN ATS framework at its extreme implementation cases, as Table III.3 suggests. At one extreme the domain for flow aggregation encompasses only a single node, then IRs are in between every node, which is similar to the TSN ATS framework except that ATS uses a class-based FIFO scheduler. At the other extreme, the domain for flow aggregation encompasses the whole internetwork that does not need any IR, which is similar to the IntServ framework. The difference in this case is that the proposed framework aggregates flows according to the input and output ports of a domain.

Table III.3 – Categorization of three frameworks based on the IR locations and the scheduler used

IR locations	**Scheduler**		
	Flow-based	**Based on FA with {input, output port} of a domain**	**FIFO**
Zero IR	IntServ	Proposed framework	
IR between domains			
IR at every node			ATS

The major complexity of the three frameworks comes from the scheduler. In this regard, ATS has the advantage. The proposed framework shows smaller or equal complexity to that of IntServ. The IR also contributes to the complexity, but it is negligible since it maintains a single queue. The IR still

has to maintain and update every flow state. The drawback of the IR resides in the average latency. It is conjectured that more IRs produce larger average latency. This is for further study by analysis or simulation. The number of IRs required for the ATS framework is proportional to the square of the port numbers of all nodes in a network. The number of IRs required for the proposed framework is proportional to the square of the number of ports of all the edge nodes, which is always less than that in ATS. Therefore, the proposed framework is expected to enjoy less complexity than IntServ and a smaller average latency than the ATS, with a smaller latency bound than both.

Appendix IV

Gap analysis of the requirements and framework for the IMT-2020 network

(This appendix does not form an integral part of this Recommendation.)

IV.1 Overview of the IMT-2020 network architecture and its QoS framework

The IMT-2020 network is a complex internetwork of different architectures and purposes. IMT-2020 network components are divided into UE, AN, CN, and DN. In Figure IV.1, functional entities other than UE, AN, and DN are included in the CN. The IMT-2020 network emphasizes three features: control and user plane separation (distributed softwarization); AN and CN independence; and network slicing. The CN has multiple user plane functions (UPFs), as data plane components, physically separated and distributed, with other control plane functions virtually distributed with the network slicing.

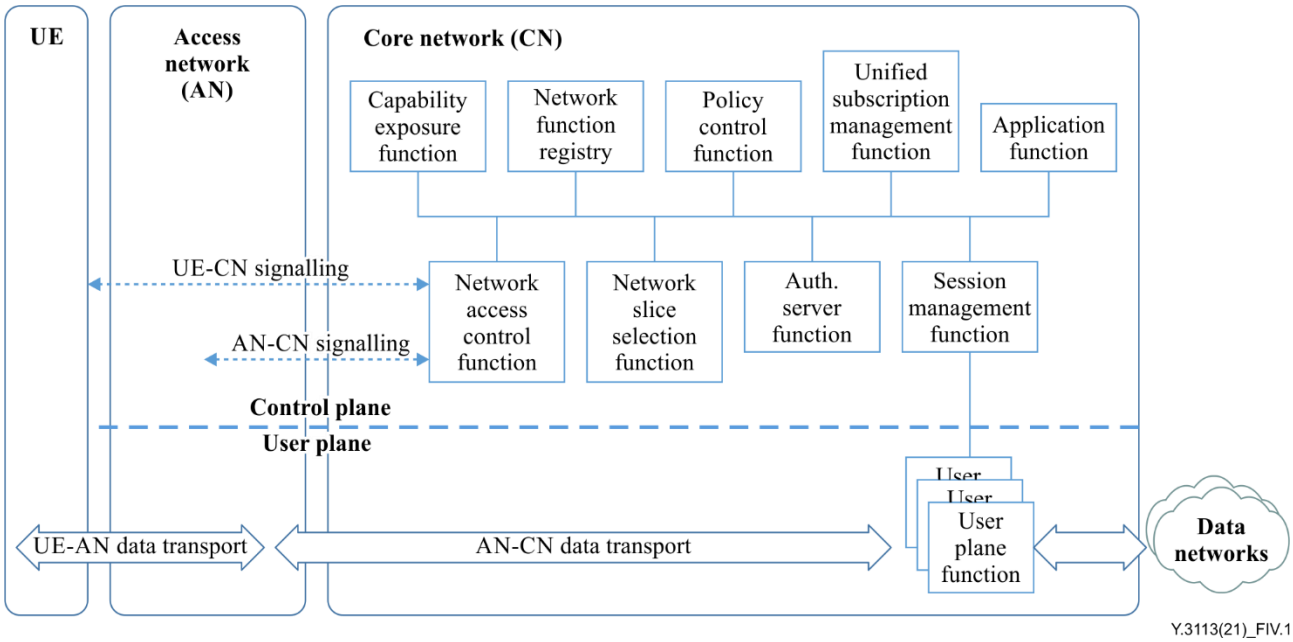

Figure IV.1 – Framework of the IMT-2020 network [b-ITU-T Y.3102]

In this Recommendation, it is assumed that a single UPF governs an aggregation domain that is a part of a network slice and also a part of a CN, with a physically identifiable boundary. The UPF performs per-flow QoS handling, including transport level packet marking for uplink (UL) and downlink (DL), rate limiting and reflective QoS (DiffServ code point) marking on the DL [b-ETSI TS 123 501]. A domain may encompass multiple routers or switches. By default, within a UPF domain relay, nodes keep the same scheduling, queueing and regulation policy.

Note that a session management function (SMF) belongs to a single network slice, and a UPF belongs to a single SMF, therefore a UPF belongs to a single network slice [b-ETSI TS 128 530].

Figure IV.2 depicts the QoS flow classification principle in the IMT-2020 network. QoS flow is the finest granularity for QoS management in the IMT-2020 core network. A QoS flow can either guarantee the bit rate or not, i.e., guaranteed bit rate (GBR) QoS flow or non-GBR (NGBR) QoS flow. With the QoS flow concept, granularity for QoS treatment (per flow) becomes different from tunnelling (per protocol data unit (PDU) session, basically), and it enables more flexible QoS control.

The terms and concepts defined in the 3GPP 5G system are mapped to the terms used in this Recommendation in Table IV.1.

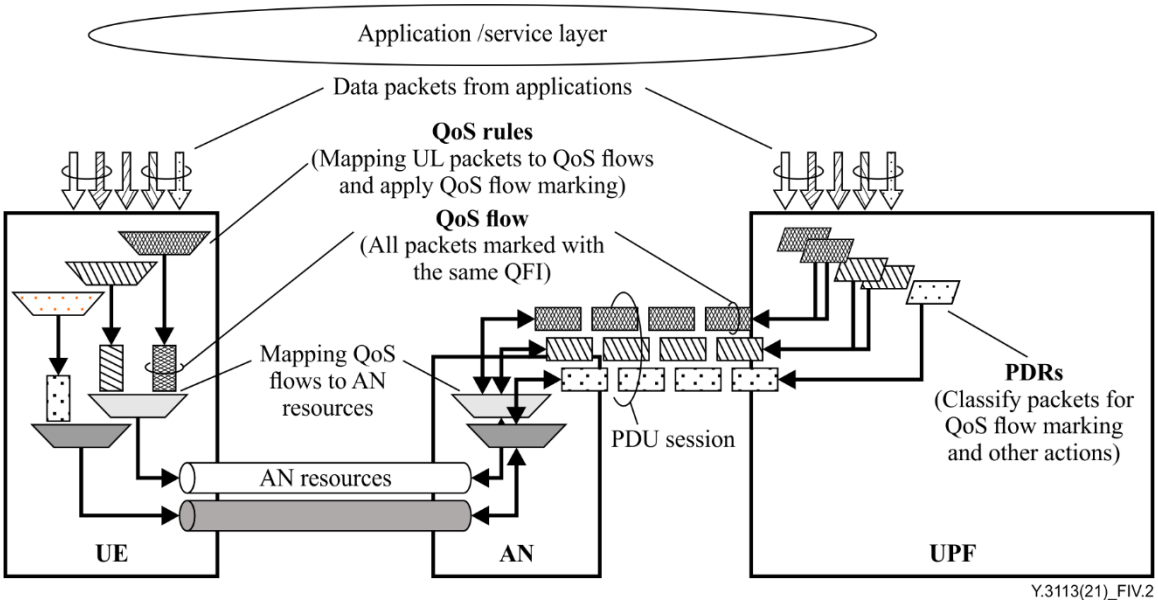

Figure IV.2 – The principle for classification and User plane marking for QoS flows [b-ETSI TS 123 501]

Table IV.1 – Mapping of 3GPP 5G terms and concepts to terms used in this Recommendation

Terms used in this Recommendation	Terms defined in 3GPP 5G System TS	Definitions in [b-ETSI TS 123 501] and [b-ETSI TS 124 501]
Flow	5G QoS flow	All traffic mapped to the same 5G QoS flow receive the same forwarding treatment (e.g., scheduling policy, queue management policy, rate shaping policy and radio link control configuration). QoS rules are used for identification of a QoS flow. QoS rules are basically a set of packet filters that are constructed with Ethernet, Internet protocol (IP), user datagram protocol/transmission control protocol header fields.
–	PDU session	Association between the UE and a DN that provides a PDU connectivity service, which is a service that provides exchange of PDUs between a UE and a DN. Generally a PDU session can be mapped to one or more QoS flows (see Note).
Flow latency bound requirement (RSpec)	5G QoS identifier (5QI)	A scalar that is used as a reference to a specific QoS forwarding behaviour (e.g., packet loss rate, packet delay budget (PDB)) to be provided to a 5G QoS flow.
High and low priority flow	GBR and NGBR QoS flow	GBR flow is a QoS flow using the GBR resource type or the delay-critical GBR resource type and requiring guaranteed flow bit rate. NGBR flow is a QoS flow using the NGBR resource type and not requiring guaranteed flow bit rate.
	Priority	5QI is also associated with priority.
Latency upper bound	PDB	The PDB determines an upper bound for the time that a packet may be delayed between the UE and the UPF that terminates the interface between the DN and the UPF (N6).

Table IV.1 – Mapping of 3GPP 5G terms and concepts to terms used in this Recommendation

Terms used in this Recommendation	Terms defined in 3GPP 5G System TS	Definitions in [b-ETSI TS 123 501] and [b-ETSI TS 124 501]
Average input rate	Guaranteed flow bit rate (GFBR) and maximum flow bit rate (MFBR)	The GFBR denotes the bit rate that is guaranteed to be provided by the network to the QoS flow over the averaging time window. The MFBR limits the bit rate to the highest bit rate that is expected by the QoS flow (e.g., excess traffic may get discarded or delayed by a rate shaping or policing function at the UE, radio access network (RAN), UPF). Bit rates above the GFBR value and up to the MFBR value, may be provided with relative priority determined by the priority level of the QoS flows.
Maximum burst size	Maximum data burst value (MDBV)	MDBV denotes the largest amount of data that the 5G-AN is required to serve within a period of 5G-AN PDB (i.e., the 5G AN part of the PDB).
NOTE – In a PDU session of Internet Protocol version 4 (IPv4), Internet Protocol version 6 (IPv6), IPv4v6 and Ethernet PDU session type, the non-access stratum protocol enables different forwarding treatments of UL user data packets in one or more QoS flows based on signalled QoS rules, derived QoS rules or any combination of them. In an unstructured PDU session type, all UL user data packets are associated with the same QoS flow (clause 6.2.5.1.1 of [b-ETSI TS 124 501]).		

IV.2 Gap analysis of the requirements for the IMT-2020 network

The 5G QoS model is based on QoS flows (clause 5.7.1.1 of [b-ETSI TS 123 501]). The 5G QoS model supports both GBR QoS flows and NGBR QoS flows. The QoS flow is the finest granularity of QoS differentiation in the PDU session. A QoS flow identifier (QFI) is used to identify a QoS flow in the 5G system. User plane traffic with the same QFI within a PDU session receives the same traffic forwarding treatment (e.g., scheduling, admission threshold). The QFI is carried in an encapsulation header on the interface between the RAN and the initial UPF (N3) (and the interface between the intermediate UPF and the UPF session anchor (N9)) i.e., without any changes to the e2e packet header. A QFI shall be used for all PDU session types. The QFI shall be unique within a PDU session. The QFI may be dynamically assigned or may be equal to the 5QI, which is a static number assigned to a flow based on its characteristics. For example, the value 80 is assigned as the standardized 5QI to low latency eMBB applications augmented reality applications with the specified 10 ms PDB.

For GBR QoS flows only, the following additional QoS parameters exist (clause 5.7.2.5 of [b-ETSI TS 123 501]):

– GFBR – UL and DL;

– MFBR – UL and DL.

The GFBR denotes the bit rate that is guaranteed to be provided by the network to the QoS flow over the averaging time window. The MFBR limits the bit rate to the highest bit rate that is expected by the QoS flow (e.g., excess traffic may get discarded or delayed by a rate shaping or policing function at the UE, RAN, UPF).

Each GBR QoS flow with a delay-critical resource type shall be associated with an MDBV (clause 5.7.3.7 of [b-ETSI TS 123 501]). MDBV denotes the largest amount of data that the 5G-AN is required to serve within a 5G-AN PDB period. Every standardized 5QI (of delay-critical GBR resource type) is associated with a default value for the MDBV (specified in the QoS characteristics of Table 5.7.4-1 of [b-ETSI TS 123 501]). The MDBV may also be signalled together with a standardized 5QI to the (R)AN, and if it is received, it shall be used instead of the default value. The

MDBV may also be signalled together with a pre-configured 5QI to the (R)AN, and if it is received, it shall be used instead of the pre-configured value.

With these QoS parameters, the SMF binds service data flows (SDFs) to QoS flows based on the QoS and service requirements (clause 5.7.1.5 of [b-ETSI TS 123 501]). The SMF assigns the QFI for a new QoS flow and derives its QoS profile, corresponding UPF instructions and QoS rule(s) from the PCC rules and other information provided by the PCF. A mapping from IP flow to SDF and further down to 5G QoS flow may not be one-to-one, however.

The following list of requirements is specified in clause 7, as well as the gap analysis of each requirement for the IMT-2020 network.

Req_1. It is required to be able to specify, by the CPE or on behalf of the CPEs with insufficient signalling capabilities, the CPE flow destination and characteristics (e.g., average data rate and maximum burst size).

Analysis: The SMF binds SDF to QoS flows with MFBR and MDBV, which correspond to the average data rate and the maximum burst size in requirement 1. Therefore, a standard IMT-2020 network meets requirement 1 as it is.

Req_2. It is recommended to be able to specify, by the CPE or on behalf of the CPEs with insufficient signalling capabilities, the CPE desirable e2e latency upper bound.

Analysis: UE or an application in a piece of UE in IMT-2020 is able to specify its PDB if the flow is GBR QoS flow. Requirement 2 is met in the current standard IMT-2020 network.

Req_3. It is required that a network be able to determine the latency upper bound within the network, of a traversing flow, with the flow destination and the characteristics specified.

Analysis: A network, regardless of the AN or CN of the IMT-2020 system, or independent DN system, is required to be able to provide and announce the latency upper bound (or PDB) in a network. It is required to add a new latency bound decision function to IMT-2020 networks.

Req_4. It is recommended that the means to provide the latency upper bound be: 1) implementable in CNs where there are millions of active flows at an output port; 2) applicable to an arbitrary network topology; 3) of minimal effect on the average latency and the throughput; and 4) scalable to a large-scale network.

Analysis: This requirement is the key for the successful deployment of latency critical GBR applications in a 5G system. This requirement can be met with the framework proposed in this Recommendation. It is required that IMT-2020 networks to follow the framework proposed in this Recommendation.

Req_5. It is recommended that the latency upper bound be susceptible to negotiation, for a flow, between the CPE and the service provider.

Req_6. It is recommended that the dynamic latency upper bound be susceptible to negotiation.

Analysis of Req_5 and Req_6: The basic approach in an IMT-2020 network to the admission decision is based on the pre-defined QoS profile of a flow and resource availability. As such, the negotiation is as simple as determining a QoS profile. On the other hand, the QoS parameter notification control in an IMT-2020 network indicates whether notifications are requested from the next generation-radio access network (NG-RAN) when the GFBR can no longer (or can again) be guaranteed for a QoS flow during its lifetime. Notification control may be used for a GBR QoS flow if the application traffic is able to adapt to the change in the QoS (clause 5.7.2.4.1 of [b-ETSI TS 123 501]). A similar admission control procedure is also specified. However, notification control focuses on re-admission control according to dynamic network environment changes. The means to determine whether to admit is not clearly specified in [b-ETSI TS 123 501] or [b-ETSI TS 123 503], but use of the performance measurements is suggested. It is recommended to add a new negotiation functionality to IMT-2020 networks.

Req_7. It is required that networks be able to handle FAs as control elements.

Req_8. It is required that networks be able to aggregate and segregate flows at any desired points in a network or equivalently within an aggregation domain.

Req_9. It is required that within an aggregation domain, flows be aggregated according to an aggregation domain-specific rule (e.g., a rule that flows with the same input and output ports of the domain are aggregated into a single FA).

Req_10. It is recommended that aggregation domains be susceptible to merger and division on demand.

Req_11. It is required that flow aggregation rules be susceptible to negotiation among the aggregation domains.

Analysis of Req_7 and Req_11: Requirements 7 to 11 are about flow aggregation. Currently the bearer service in IMT-2020 can be thought as a flow aggregation. A more flexible aggregation function in the IMT-2020 network is required.

Req_12. It is recommended that the TSpec include: a maximum burst size; an average input rate; a peak rate; and a maximum packet size.

Analysis: The current IMT-2020 standard specifies the MFBR and the MDBV to be QoS parameters, which correspond to the input rate and the maximum burst size. The maximum packet size is rather a network-specific parameter not specified for QoS control in a IMT-2020 network. It is recommended that the maximum packet size be a part of the TSpec in IMT-2020.

Req_13. It is recommended that best-effort service traffic does not affect the latency bound of high-priority flows.

Req_14. It is required that a traffic regulation capability be provided at the boundary of aggregation domains.

Analysis: It is recommended that a pre-emption capability on NGBR or low-priority traffic be implemented in IMT-2020 networks. It is also required that a regulation function (e.g., minimal IR per FA) be implemented in IMT-2020 networks.

IV.3 Gap analysis of the framework for the IMT-2020 network

The framework in this Recommendation depends on three essential elements: flow aggregation; aggregation domain; and minimal IR. The flow in this Recommendation can be specified to be a set of packets sharing network start and end points, as well as traffic characteristics such as the latency bound. A QoS flow in the IMT-2020 network, identified by a QFI, may be an aggregation of SDFs. An SDF may also be an aggregation of IP flows. A proper mapping from IP flow to QoS flow can satisfy the specification of the flow in this Recommendation. QoS flows can use the bearer service in an IMT-2020 network. A bearer in this regard can be seen as an FA. Again a careful initiation and a termination of a bearer can be seen as flow aggregation and also segregation. However, the initiation and termination points of a bearer service are rather fixed according to the network type, i.e., whether it is access or core network. It is recommended that an aggregation domain be arbitrarily set to include multiple networks or multiple UDFs. Implementation of the minimal IRs per FA at the boundaries of aggregation domains is also required.

Bibliography

[b-ITU-T E.800] Recommendation ITU-T E.800 (2008), *Definitions of terms related to quality of service*.

[b-ITU-T Y.2701] Recommendation ITU-T Y.2701 (2007), *Security requirements for NGN release 1*.

[b-ITU-T Y.3101] Recommendation ITU-T Y.3101 (2018), *Requirements of the IMT-2020 network.*

[b-ITU-T Y.3102] Recommendation ITU-T Y.3102 (2018), *Framework of the IMT-2020 network.*

[b-ITU-T Y-Suppl.66] ITU-T Y-series Recommendations – Supplement 66, (2020), *ITU-T Y.3000-series Network 2030 services*: *Capabilities, performance and design of new communication services for the Network 2030 applications*.

[b-ETSI TS 123 501] Technical Specification ETSI TS 123 501 V16.6.0 (2020), *5G; System architecture for the 5G System (5GS) (3GPP TS 23.501 version 16.6.0 Release 16).*

[b-ETSI TS 123 503] Technical Specification ETSI TS 123 503 V16.5.0 (2020), *5G; Policy and charging control framework for the 5G System (5GS); Stage 2 (3GPP TS 23.503 version 16.5.0 Release 16).*

[b-ETSI TS 124 501] Technical Specification ETSI TS 124 501 V16.5.1 (2020), *5G; Non-access-stratum (NAS) protocol for 5G system (5GS); Stage 3 (3GPP TS 24.501 version 16.5.1 Release 16).*

[b-ETSI TS 128 530] Technical Specification ETSI TS 128 530 V16.2.0 (2020), *5G; Management and orchestration; Concepts, use cases and requirements (3GPP TS 28.530 version 16.2.0 Release 16).*

[b-IEEE 802.1Qcr] IEEE 802.1Qcr-2020, *IEEE Standard for local and metropolitan area networks – Bridges and bridged networks – Amendment 34: Asynchronous traffic shaping*.

[b-IEEE TSN] IEEE 802.1 Working Group (2017). *Time-Sensitive Networking Task Group*. Available [viewed 2021-03-18] at: http://www.ieee802.org/1/pages/tsn.html

[b-IETF RFC 8655] IETF RFC 8655 (2019), *Deterministic networking architecture*.

[b-Fei] Fei A., Pei G., Liu R., Zhang L. (1998). Measurements on delay and hop-count of the Internet. In: *IEEE GLOBECOM*. Available [viewed 2021-03-18] at: http://web.cs.ucla.edu/~lixia/papers/98Globcom.pdf.

[b-Joung] Joung, J. (2020). Framework for delay guarantee in multi-domain networks based on interleaved regulators. *Electronics*, **9**, 436, 11 pp. Available [viewed 2021-03-18] from: https://www.mdpi.com/2079-9292/9/3/436.

[b-Le Boudec] Le Boudec, J. (2018). A theory of traffic regulators for deterministic networks with application to interleaved regulators. In: *IEEE/ACM Trans. Networking*, Vol. 26, No. 6, pp. 2721-2733. doi: 10.1109/TNET.2018.2875191.

[b-Lenzini] Lenzini, L., Mingozzi, E., Stea, G. (2004). Tradeoffs between low complexity, low latency, and fairness with deficit round-robin schedulers. In: *IEEE/ACM Trans. Networking*, vol. 12, no. 4, pp. 681-693. doi: 10.1109/TNET.2004.833131.

[b-Shreedhar] Shreedhar, M., Varghese, G. (1996). Efficient fair queueing using deficit round-robin. In: *IEEE/ACM Trans. Networking*, Vol. 4, No. 3, pp. 375-385 doi: 10.1109/90.502236.

[b-Specht] Specht, J., Samii, S. (2016). Urgency-based scheduler for time-sensitive switched Ethernet networks. In: *Proc. 28th Euromicro Conference on Real-Time Systems (ECRTS)*, Toulouse, pp. 75-85. doi: 10.1109/ECRTS.2016.27.

[b-Stiliadis] Stiliadis, D., Varma A. (1998). Latency-rate servers: A general model for analysis of traffic scheduling algorithms. In: *IEEE/ACM Transactions on Networking*, vol. 6, no. 5, pp. 611-624. doi: 10.1109/90.731196.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	**Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities**
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2021