ETSI TS 126 256 V18.0.0 (2024-05)

LTE;
5G;
Codec for Immersive Voice and Audio Services
- Jitter Buffer Management
(3GPP TS 26.256 version 18.0.0 Release 18)

Reference
DTS/TSGS-0426256vi00

Keywords
5G,LTE

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Legal Notice

This Technical Specification (TS) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under https://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Specification has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
  - 1 presented to TSG for information;
  - 2 presented to TSG for approval;
  - 3 or greater indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the document.

In the present document, modal verbs have the following meanings:

shall indicates a mandatory requirement to do something

shall not indicates an interdiction (prohibition) to do something

The constructions "shall" and "shall not" are confined to the context of normative provisions, and do not appear in Technical Reports.

The constructions "must" and "must not" are not used as substitutes for "shall" and "shall not". Their use is avoided insofar as possible, and they are not used in a normative context except in a direct citation from an external, referenced, non-3GPP document, or so as to maintain continuity of style when extending or modifying the provisions of such a referenced document.

should indicates a recommendation to do something

should not indicates a recommendation not to do something

may indicates permission to do something

need not indicates permission not to do something

The construction "may not" is ambiguous and is not used in normative elements. The unambiguous constructions "might not" or "shall not" are used instead, depending upon the meaning intended.

can indicates that something is possible

cannot indicates that something is impossible

The constructions "can" and "cannot" are not substitutes for "may" and "need not".

will indicates that something is certain or expected to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

will not indicates that something is certain or expected not to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

might indicates a likelihood that something will happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

might not indicates a likelihood that something will not happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

In addition:

is (or any other verb in the indicative mood) indicates a statement of fact

is not (or any other negative verb in the indicative mood) indicates a statement of fact

The constructions "is" and "is not" do not indicate requirements.

1 Scope

The present document defines the Jitter Buffer Management (JBM) solution for the Immersive Voice and Audio Services (IVAS) codec [2]. Jitter Buffers are required in packet-based communications, such as 3GPP MTSI [7], to smooth the inter-arrival jitter of incoming media packets for uninterrupted playout.

The procedure of the present document is recommended for implementation in all network entities and UEs supporting the IVAS codec; procedures described in [4] and used in this document, such as multi-channel time-scale modification, metadata adaptation and rendering are mandatory for implementations in all network entities and UEs supporting the IVAS codec.

The present document does not describe the C code of this procedure. For a description of the floating-point C code implementation see [3].

In the case of discrepancy between the Jitter Buffer Management described in the present document and its C code specification contained in [3], the procedure defined by [3] prevails.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TS 26.250: "Codec for Immersive Voice and Audio Services – General Overview".

[3] 3GPP TS 26.258: "Codec for Immersive Voice and Audio Services; ANSI C code (floating-point)".

[4] 3GPP TS 26.253: "Codec for Immersive Voice and Audio Services - Detailed Algorithmic Description incl. RTP payload format and SDP parameter definitions".

[5] 3GPP TS 26.441: "Codec for Enhanced Voice Services (EVS); General overview".

[6] 3GPP TS 26.171: "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description".

[7] 3GPP TS 26.114: "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction"

[8] 3GPP TS 26.448: "Codec for Enhanced Voice Services - Jitter Buffer Management"

[9] 3GPP TS 26.131: "Terminal acoustic characteristics for telephony; Requirements"

[10] 3GPP TS 26.261, "Terminal audio quality performance requirements for immersive audio services"

3 Definitions of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [1].

3.2 Symbols

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [1].

AMR	Adaptive Multi-Rate (Codec)
AMR-WB	Adapative Multi-Rate Wideband (Codec)
DTX	Discontinuous Transmission
EVS	Enhanced Voice Services (Codec)
IVAS	Immersive Voice and Audio Services
JBM	Jitter Buffer Management
RTP	Real-Time Transmission Protocol
TSM	Time-Scale Modification

3.4 Mathematical Expressions

4 General

4.1 Introduction

The Jitter Buffer Management solution specified in this document extends the IVAS decoder with a mechanism to cope with the effects of packet-based communication over wireless transmission channels, i.e. buffering packets with different inter-arrival jitter and triggering of adapatation mechanisms to ensure low-delay communications.

It is used in conjunction with the IVAS decoder (described in [4] in detail), which can also decode EVS [5] and AMR-WB [6]. The described solution is based on [8], which has been optimized for the Multimedia Telephony Service for IMS (MTSI) and fulfils the requirements for delay and jitter-induced concealment operations set in [7]. Main differences to [8] are the support of immersive media formats and a corresponding time-warping scheme operating within the decoder.

4.2 Packet-based communications

In packet-based communications, packets arrive at the terminal with random jitters in their arrival time. Packets may also arrive out of order. Since the decoder expects to be fed a speech packet in a regular interval (for 3GPP speech codecs this is every 20 milliseconds) to output speech samples in periodic blocks, a de-jitter buffer is required to absorb the jitter in the packet arrival time. The larger the size of the de-jitter buffer, the better its ability to absorb the jitter in the arrival time and consequently fewer late arriving packets are discarded. Voice communications is also a delay critical system and therefore it becomes essential to keep the end to end delay as low as possible so that a two way conversation can be sustained.

The defined adaptive Jitter Buffer Management (JBM) solution reflects the above mentioned trade-offs. While attempting to minimize packet losses, the JBM algorithm in the receiver also keeps track of the delay in packet delivery

as a result of the buffering. The JBM solution suitably adjusts the depth of the de-jitter buffer in order to achieve the trade-off between delay and late losses.

4.3 IVAS Receiver architecture overview

An IVAS receiver for MTSI-based communication is built on top of the IVAS Jitter Buffer Management solution. It follows the same principles as specified in clause 4.3 in [8] for the EVS Jitter Buffer Management solution. The received IVAS frames, contained in RTP packets, are depacketized and fed to the Jitter Buffer Management (JBM). The JBM smoothes the inter-arrival jitter of incoming packets for uninterrupted playout of the decoded EVS frames at the Acoustic Frontend of the terminal.

Figure 1 in [8] illustrates the architecture and data flow of the receiver side of an EVS terminal. The example architecture for EVS is also applicable to IVAS to outline the integration of the JBM in a terminal. This specification defines the JBM module and its interfaces to the RTP Depacker, the IVAS Decoder [4], and the Acoustic Frontend [9] and [10]. The modules for Modem and Acoustic Frontend are outside the scope of the present document. The implementation of the RTP Depacker is outlined in [8] and also applicable for IVAS.

Real-time implementations of this architecture typically use independent processing threads for reacting on arriving RTP packets from the modem and for requesting PCM data for the Acoustic Frontend. Arriving packets are typically handled by listening for packets received on the network socket related to the RTP session. Incoming packets are pushed into the RTP Depacker module which extracts the frames contained in an RTP packet. These frame are then pushed into the JBM where the statistics are updated and the frames are stored for later decoding and playout. The Acoustic Frontend contains the audio interface which, concurrently to the push operation of IVAS frames, pulls PCM buffers from the JBM. The JBM is therefore required to provide PCM buffers, which are normally generated by decoding IVAS frames by the IVAS decoder or by other means to allow uninterrupted playout. Although the JBM is described for a multi-threaded architecture it does not specify thread-safe data structures due to the dependency on a particular implementation.

Note that the JBM does not directly forward frames from the RTP Depacker to the IVAS decoder but instead uses frame-based adaptation to smooth the network jitter. In addition signal-based adaptation is executed on the decoded PCM buffers, described in detail in clause 6.2.7.3 [4] before they are pulled by the Acoustic Frontend. The corresponding algorithms are described in the following clauses.

5 Jitter Buffer Management

5.1 Overview

Jitter Buffer Management (JBM) includes the jitter estimation, control and jitter buffer adaptation algorithm to manage the inter-arrival jitter of the incoming packet stream.

The IVAS Jitter Buffer Management allows for fine grain adjustment of the play out delay by generating time scale modified (TSM) versions of a multi-channels signal, i.e. provide decoded frames that are longer or short in duration than the default frame length. The IVAS JBM splits the decoding and reconstruction/rendering into the steps transport channel and metadata decoding, the multi-channel time scale modification of the transport channels, resulting in a time scale modified version of the transport channels with specific duration, the adaption of the metadata and other rendering/reconstruction parameters to the time scale modified duration of the IVAS frame, and the reconstruction and rendering adapted to the new time scale modified frame length. The IVAS JBM decoding process performs a number of processing steps to provide the processed (output) signal based on the input audio signal representation (the encoded IVAS frame), where the time scale modification is performed on the intermediate audio signals, i.e. the transport channels which are generated by the first processing step of decoding the transport channels and meta data, and performs the second processing, i.e. the reconstruction/rendering of the output signal based on the time scaled intermediate audio signals. The reconstruction and rendering is adapted to the time scale modification as are the parameters needed for the reconstruction, and the meta data needed for the reconstruction.

The entire solution for IVAS consists of the following components, as depicted in Figure 1:

- RTP Depacker (see clause 5.2) to analyse the incoming RTP packet stream and to extract the EVS speech frames along with meta data to estimate the network jitter

- De-jitter Buffer (clause 5.6) to store the extracted IVAS speech frames before decoding and to perform frame-based adaptation
- IVAS Transport Channel and Metadata Decoding [4] (clause 6) for decoding the received IVAS frames to PCM data
- Playout Delay Estimation Module (clause 5.3.5) to provide information on the current playout delay due to JBM
- Network Jitter Analysis (clause 5.3) for estimating the packet inter-arrival jitter and target playout delay
- Adaptation Control Logic (clause 5.4) to decide on actions for changing the playout delay based on the target playout delay
- Multi-Channel Time-Scale Modification (clause 5.4.3) and clause 5.6 of the present document to perform signal-based adaptation for changing the playout delay
- Metadata Adaptation and processing parameter adaption [4] (clause 6.2.7.2, and clauses 6.3.7, 6-4-11, 6.5.4, 6.6.7, 6.7.8, 6.8.5, 6.10, and 6.9.8) and clause 5.6 of the present document to perform adapation for meta data for fitting to time-warped signals
- Reconstruction and Rendering [4] (clauses 6 and 7) and clause 5.5 of the present document to convert transport-channel and meta data to the reconstructed output channels

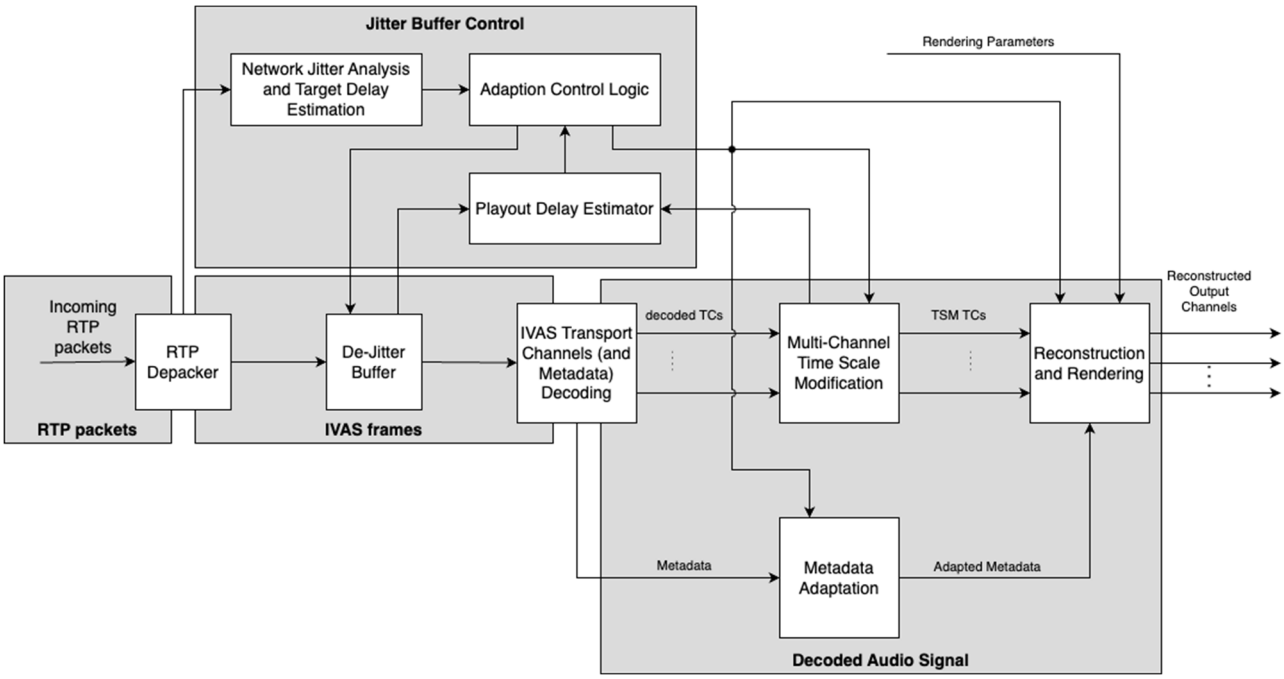

Figure 1: Modules of the IVAS Jitter Buffer Management Solution

5.2 Depacketization of RTP packets (informative)

The RTP Depacker module of the JBM performs the depacketization of the incoming RTP packet stream. The operation is further described in clause 5.2 of [8]. Packetization rules for IVAS are defined in Annex A of [4].

5.3 Network Jitter Analysis and Delay Estimation

The network jitter analysis and delay estimation for IVAS is identical to EVS, which is specified in [8], clause 5.3.

5.4 Adaptation Control Logic

5.4.1 Control Logic

The control logic is identical to EVS. The operation is described in clause 5.4.1 of [8].

5.4.2 Frame-based adaptation

5.4.2.1 General

Adaptation on the frame level is performed on coded speech frames, i.e. with a granularity of one speech frame of 20 ms duration. Inserting or deleting speech frames results in adaptation with higher distortion but allows faster buffer adaptation than signal-based adaptation. Inserting or deleting NO_DATA frames during DTX allows fast adaption while minimizing distortion.

5.4.2.2 Insertion of Concealed Frames

Insertion of concealed frames identical to EVS. The operation is described in clause 5.4.2.2 of [8].

5.4.2.3 Frame Dropping

Frame dropping is identical to EVS. The operation is described in clause 5.4.2.3 of [8].

5.4.2.4 Comfort Noise Insertion in DTX

Comfort Noise insertion in DTX is identical to EVS. The operation is described in clause 5.4.2.4 of [8].

5.4.2.5 Comfort Noise Deletion in DTX

Comfort Noise deletion in DTX is identical to EVS. The operation is described in clause 5.4.2.5 of [8].

5.4.3 Signal-based adaptation

To alter the playout delay the decoder is able to generate time-warped signals. This allows increasing the number of samples for increasing the playout delay or reducing the number of samples to reduce the playout delay. The time-warping is performed by the decoder and its basic operation is described in [8], clause 5.4.3. IVAS extends this time-scale modification to work on multi-channel signals. This is specified in [4], clause 6.2.7.3.

The meta data and processing parameters are adapted to the achieved time-scale modification according to [4] (clause 6.2.7.2, and clauses 6.3.7, 6-4-11, 6.5.4, 6.6.7, 6.7.8, 6.8.5, 6.10, and 6.9.8).

5.5 Reconstructed signal output

5.5.1 General

The FIFO Receiver Output Buffer of [8] clause 5.5 is replaced by the structure outlined in Figure 2. A pull from the acoustic front end renders either the number of samples requested if enough samples to be rendered are available or as many samples as are still available to be rendered according to the transport channel buffer. If the number of samples rendered is not enough to satisfy the pull request a new frame is decoded and TSM and meta data adaption are applied and enough samples are reconstructed and rendered from this frame to satisfy the pull request. The Receiver Output Buffer may either be omitted or the size has to be large enough to store enough samples for one pull call from the frontend.

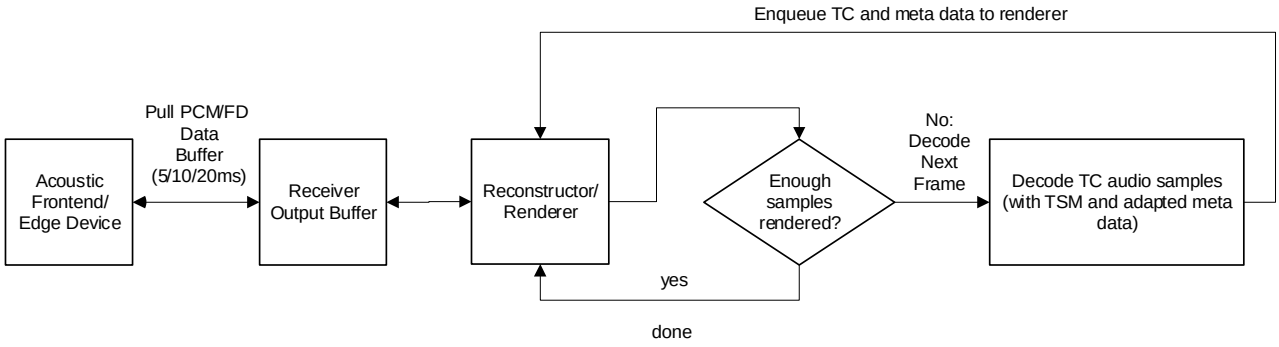

Figure 2: Reconstructed Signal Output

5.5.2 Interaction with Decoder Transport Channel Buffer

The transport channel buffer according to clause 6.2.7.2 [4] shall keep track of the number of already rendered samples, the number of samples still available for rendering the current IVAS frame, and the number of residual samples, i.e. transport channel samples that can not be rendered in the current frame, and provides this information to the De-Jitter Buffer for the playout delay estimation. The variable b_k used in clause 5.3.5 [8] Eq. 11 clause is now instead of the duration of samples buffered in the Receiver Output Buffer module at playout time k the duration of the samples still available for rendering and the residual samples combined and expressed in milliseconds.

5.5.3 Residual Samples Handling

The reconstruction and renderer has a time resolution that is smaller than the time resolution of the PCM data, that is the smallest portion that can be output by the reconstruction and rendering contains a multiple number of PCM samples per output channel. The signal based adaption results in time scale modified frames that may not fit into this time resolution. Any spill, that is any residual samples that do not fit into the time resolution, is handled by the transport channel buffer management according to clause 6.2.7.2 [4].

5.6 De-Jitter Buffer

The De-Jitter Buffer is identical to EVS. The operation is described in clause 5.6 of [8].

6 Decoder interaction

6.1 General

This JBM solution for the IVAS codec may also be used for EVS. The usage with EVS is described in clause 6.10 [4].

6.2 Decoder Requirements

The defined JBM depends on the decoder processing function to create an uncompressed PCM frame from a coded frame. The JBM requires that the number of channels and sampling rate in use is known in order to initialize the signal-based adaptation. The JBM also requires the presence of PLC functionality to create a PCM frame on demand without a coded frame being available as input for the decoder for missing or lost frames.

The JBM will make use of DTX for playout adaptation during inactive periods when noise is generated by the CNG. It has however its own functionality integrated for playout adaptation of active signals if the codec does not currently use or support DTX, as well as for adaptation during a long period of active signal. To use DTX the RTP Depacker needs to determine if a frame is an SID frame or an active frame and provide that information to the JBM together with the respective frame. During DTX the JBM may alter the periods between SID frames or between the last SID frame and the first active frame to use the CNG of the decoder to create additional PCM buffers with comfort noise or to omit comfort noise frames.

The JBM expects that the decoder outputs PCM frames of arbitrary duration and a fixed audio sampling rate set at initialization. The JBM expects that the decoder has an internal transport channel buffer that handles the buffering of the time scale modified PCM frames either for direct output or for reconstruction and rendering. If the codec supports bandwidth switching, a resampling functionality is required in the decoder to provide PCM frames at the set sampling rate.

Annex <A> (informative): Change history

Change history							
Date	**Meeting**	**TDoc**	**CR**	**Rev**	**Cat**	**Subject/Comment**	**New version**
09-2023	-					Starting Point - internal only	0.0.1
11-2023	SA4#126	SA4-231833				Initial presentation to 3GPP SA4	0.0.2
12-2023	SA#100					Presented to 3GPP SA Plenary for Information	1.0.0
01-2024	SA4#127	S4-240160				Additions referencing TS26.253	1.0.1
02-2024	SA4#127	S4-240481				Agreement in 3GPP SA4	1.1.0
03-2024	SA#103	SP-240034				Version 2.0.0 created by MCC	2.0.0
03-2024						Version 18.0.0 created by MCC	18.0.0

History

Document history		
V18.0.0	May 2024	Publication