WebRTC RTP Payload Format (Opus, VP8)

Core component of WebRTC is to manage the encoder, to transforms captured raw audio/video as RTP packets, and sent them out, this is on the sender side; on the receiver side, it to mange the decoder, to reconstruct the media bitstream (encoded) from received RTP packets, decode, and play them back.

WebRTC uses RTP for streaming media signals over the network.

Opus, VP8 is used as the codec by EVC15, the following describes the key concept how the Opus/VP8 bitstream is encapsulated in RTP payload.

Audio Opus

Ref: http://datatracker.ietf.org/doc/draft-ietf-payload-rtp-opus/

Codec Overview

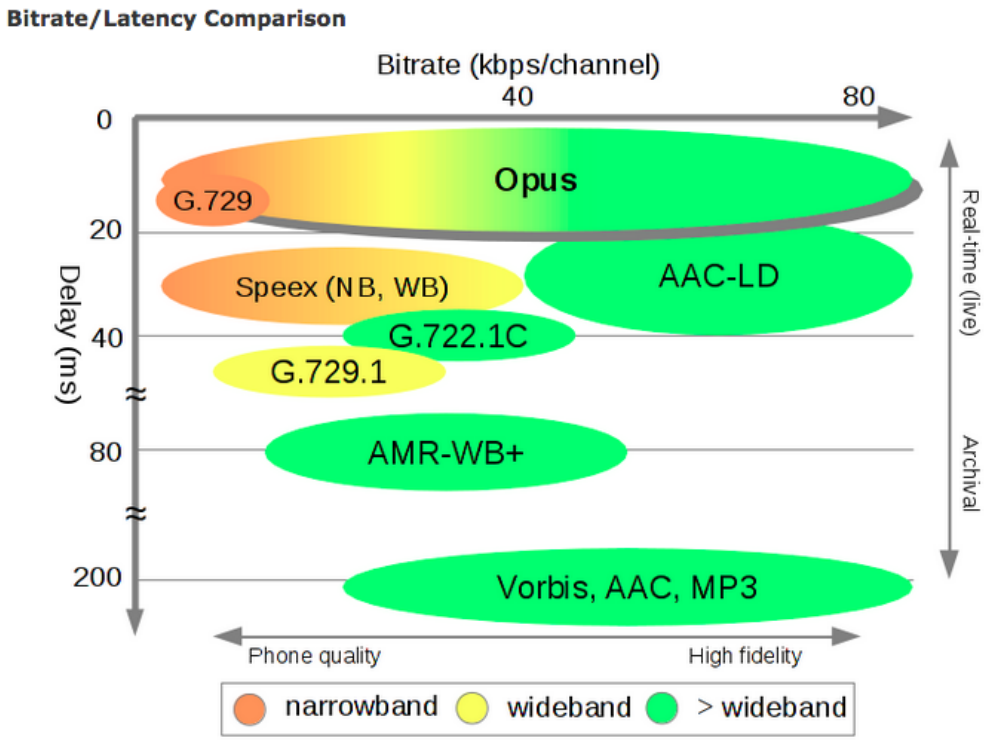
5 audio bandwidth

Opus supports 5 different audio bandwidth

Abbreviation	Name	Audio Bandwidth (Hz)	Sampling Rate (Hz)
NB	Narrowband	0 - 4000	8000
MB	Mediumband	0 - 6000	12000
WB	Wideband	0 - 8000	16000
SWB	Super-wideband	0 - 12000	24000
FB	Fullband	0 - 20000	48000

2 Encoder

Opus has 2 encode mode to adaptor for voice(speech) and audio(music) scenarios, based on SIKL and CELT, which by design helps Opus achieving better performance on bitrate control, and latency.



6 Frame Size

The Opus encoder can output encoded frames representing 2.5, 5, 10, 20, 40, or 60 ms of speech or audio data.

Mode	fs	2.5	5	10	20	40	60
ts incr	all	120	240	480	960	1920	2880
voice	NB/MB/WB/SWB/FB	x	x	o	o	o	o
audio	NB/WB/SWB/FB	o	o	o	o	x	x

Table 2: Supported Opus frame sizes and timestamp increments marked with an o. Unsupported marked with an x.

Payload Parameters

960 timestamp interval

WebRTC RTP Opus payload uses 20ms frame size, 48kHz Sampling rate, which is saying each RTP packet represents a 20ms signal containing 960 (48000 / 1000 / 20) sampled signal points.

```
Opus supports 5 different audio bandwidths, which can be adjusted
during a stream.  The RTP timestamp is incremented with a 48000 Hz
clock rate for all modes of Opus and all sampling rates.  The unit
for the timestamp is samples per single (mono) channel.  The RTP
timestamp corresponds to the sample time of the first encoded sample
in the encoded frame.  For data encoded with sampling rates other
than 48000 Hz, the sampling rate has to be adjusted to 48000 Hz.
```

This explains, in Wireshark, we see the RTP packet timestamp increases by 960.

Filter: rtp.p_type == 96 and ip.dst == 10.0.0.2 | Expression... Clear Apply Save

No.	Time	Source	Src.port	Destination	Dest.port	Protocol	Length	Info
61563	3.376445000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36674, Time=9836160
61563	3.406031000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36675, Time=9837120
61563	3.437237000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36676, Time=9838080
61563	3.437352000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36677, Time=9839040
61563	3.468354000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36678, Time=9840000
61563	3.499594000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36679, Time=9840960
61563	3.500126000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36680, Time=9841920
61563	3.530854000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36681, Time=9842880
61563	3.562232000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36682, Time=9843840
61563	3.562417000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36683, Time=9844800
61563	3.594294000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36684, Time=9845760
61563	3.625161000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36685, Time=9846720
61563	3.625741000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36686, Time=9847680
61563	3.660307000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36687, Time=9848640
61563	3.660462000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36688, Time=9849600
61563	3.690311000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36689, Time=9850560
61563	3.720261000	222.73.107..	19632	10.0.0.2	61563	RTP	67	PT=DynamicRTP-Type-96, SSRC=0x1BE11C4F, Seq=36690, Time=9851520

```
▷ Frame 481: 67 bytes on wire (536 bits), 67 bytes captured (536 bits) on interface 0
▷ Ethernet II, Src: Netgear_ad:9d:2c (a0:21:b7:ad:9d:2c), Dst: Apple_23:05:20 (28:37:37:23:05:20)
▷ Internet Protocol Version 4, Src: 222.73.107.149 (222.73.107.149), Dst: 10.0.0.2 (10.0.0.2)
▷ User Datagram Protocol, Src Port: 19632 (19632), Dst Port: 61563 (61563)
▽ Real-Time Transport Protocol
    10.. .... = Version: RFC 1889 Version (2)
    ..0. .... = Padding: False
    ...0 .... = Extension: False
    .... 0000 = Contributing source identifiers count: 0
    0... .... = Marker: False
    Payload type: DynamicRTP-Type-96 (96)
    Sequence number: 36679
    Timestamp: 9840960
    Synchronization Source identifier: 0x1be11c4f (467737679)
    Payload: 7940b43fc7722fd5e409fac529
```

Opus supports variable bitrate (VBR) from 6 kb/s to 510 kb/s. The bitrate can be changed dynamically within that range.

```
For a frame size of 20 ms, these are the bitrate "sweet spots" for
Opus in various configurations:

o  8-12 kb/s for NB speech,
o  16-20 kb/s for WB speech,
o  28-40 kb/s for FB speech,
o  48-64 kb/s for FB mono music, and
o  64-128 kb/s for FB stereo music.
```

Following is the audio media bitrate usage for EVC15 (iftop output), with 2s, 10s, 40s average.

```
10.0.0.2:63908          => 222.73.107.149:53294          54.0Kb  54.2Kb  54.3Kb
                        <=                               23.2Kb  23.4Kb  23.4Kb
```

SDP Parameters

For full list please check:

https://tools.ietf.org/html/draft-ietf-payload-rtp-opus-11#section-6.1

Important parameters:

```
maxaveragebitrate:  specifies the maximum average receive bitrate of
   a session in bits per second (b/s).  The actual value of the
   bitrate can vary, as it is dependent on the characteristics of the
   media in a packet.  Note that the maximum average bitrate MAY be
   modified dynamically during a session.  Any positive integer is
   allowed, but values outside the range 6000 to 510000 SHOULD be
   ignored.  If no value is specified, the maximum value specified in
   Section 3.1.1 for the corresponding mode of Opus and corresponding
   maxplaybackrate is the default.

useinbandfec:  specifies that the decoder has the capability to take
   advantage of the Opus in-band FEC.  Possible values are 1 and 0.
```

Video VP8

The following describes the key conecpt how the encoded VP8 bitstream is encapsulated in RTP.

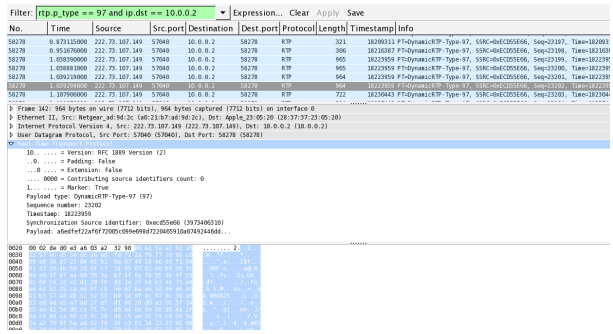
RTP Header

Conceptual Bit Diagram

```
The general RTP payload format for VP8 is depicted below.

 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|V=2|P|X|  CC   |M|     PT      |        sequence number        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             timestamp                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           synchronization source (SSRC) identifier            |
+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
|            contributing source (CSRC) identifiers             |
|                             ....                              |
+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
|            VP8 payload descriptor (integer #bytes)            |
:                                                               :
|                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               : VP8 payload header (3 octets) |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| VP8 pyld hdr  :                                               |
+-+-+-+-+-+-+-+-+                                               |
:                    Bytes 4..N of VP8 payload                  :
|                                                               |
|                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               :    OPTIONAL RTP padding       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Wireshark Sample



There is no "contributing source" in the sample on the right side.

- Version (V): 2
- Padding (P): 0
- Extension (X): 0
- Marker bit (M): ALWAYS be set for the very last packet of each encoded frame in line with the normal use of the M bit in video f
- Payload Type (PT): DynamicRTP-Type-97
- Timestamp: The RTP timestamp indicates the time when the frame was sampled at a clock rate of 90 kHz. (VP8/90000)
- Sequence number: The sequence numbers are monotonically increasing and set as packets are sent.

Frame Reconstruct Algorithm

1. Collect all packets with a given RTP timestamp.
2. Go through packets in order, sorted by sequence numbers, if packets are missing, send NACK as defined in [RFC4585] or decode with missing partitions.
3. A frame is complete if the frame has no missing sequence numbers, the first packet in the frame contains S=1 with partId=0 and the last packet in the frame has the marker bit set.

Payload Parameters

m=video 49170 RTP/AVPF 97
a=rtpmap:97 VP8/90000
a=fmtp:97 max-fr=30; max-fs=3600;

```
max-fr:  The value of max-fr is an integer indicating the maximum
   frame rate in units of frames per second that the decoder is
   capable of decoding.

max-fs:  The value of max-fs is an integer indicating the maximum
   frame size in units of macroblocks that the decoder is capable
   of decoding.

   The decoder is capable of decoding this frame size as long as
   the width and height of the frame in macroblocks are less than
   int(sqrt(max-fs * 8)) - for instance, a max-fs of 1200 (capable
   of supporting 640x480 resolution) will support widths and
   heights up to 1552 pixels (97 macroblocks).
```