Audio Architecture - Field Guide

P2P vs MCU vs SFU for group-call audio, the RFC 6464 trick, and how to pick - on one page.

P2P vs MCU vs SFU at a glance

Criterion	P2P	MCU	SFU
Server in audio path	None	Yes, heavy	Yes, light
Server decodes audio	n/a	Every stream	Never
Server CPU per room	Zero	Very high	Low
Client downloads	1 per person	1 (the mix)	1 per speaker
Added latency	Lowest	Highest	Low
Scales to big rooms	No	CPU-limited	Yes (default)
Single-file recording	No	Easy	Needs egress
Dumb receivers (PSTN)	No	Yes	No

The SFU trick: forward only the talkers (RFC 6464)

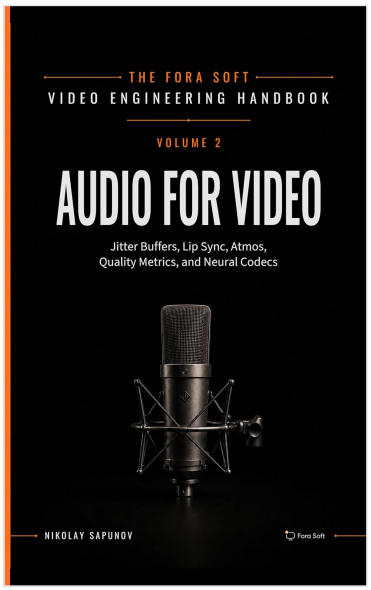
What it is	Each sender stamps every audio packet with its loudness, 0-127 in -dBov.
Where it lives	In an RTP header extension (RFC 6464 / RFC 5285) - readable without decode.
Why it matters	The SFU ranks speakers from the number alone, then forwards only the loud few.
The V bit	One bit flags whether the sender thinks the packet holds real voice.

What real engines do

mediasoup	SFU. ActiveSpeakerObserver uses RFC 6464 levels (Volfin-Cohen). No server decode.
LiveKit	SFU. Forwards Opus untouched; recording = subscribe + decode + mix in egress.
Jitsi	SFU. Dominant-speaker + Last-N forwards only the most relevant streams.
Opus DTX	Drops a quiet mic from ~24-32 kbit/s to ~1 kbit/s; can confuse level readers.

Remember

- [ ] Two people, no recording need: use P2P - direct path, lowest latency, zero server cost.
- [ ] Above four people, or once you add video: put a server in the middle, almost always an SFU.
- [ ] SFU = forward, never decode. MCU = decode, mix into one N-1 stream, re-encode. That sets the cost.
- [ ] An MCU sends each person everyone's voice EXCEPT their own (N-1 mix) to stop feedback.
- [ ] SFU recording is NOT free: no mix exists, so budget a separate egress service that mixes.
- [ ] Use an MCU when a receiver can only play one stream (PSTN dial-in, set-top, gateway).
- [ ] Active-speaker UI flickers when people go quiet? Suspect Opus DTX gaps, not the algorithm.
- [ ] Hybrids win: start P2P for two, promote to SFU on the 3rd join; add a small MCU for PSTN legs.



GO DEEPER · VOLUME 5 OF 7

Audio for Video: A Complete Guide to Sound in Video Products

From Jitter Buffers and Lip Sync to Atmos, Quality Metrics, and Neural Codecs

By Nikolay Sapunov, CEO at Fora Soft · 1,288 pages · Kindle & paperback on Amazon

Read it on Amazon → amazon.com/dp/B0HH6XPY46

This field guide compares P2P, MCU and SFU for group-call audio and shows how RFC 6464 levels let an SFU forward only the talkers. Volume 5 goes further into what happens to call audio on the network: jitter buffers and NetEq, loss concealment, FEC and RED, and bitrate and topology decisions.

WHAT YOU WILL FIND INSIDE

- Topology decisions for group-call audio
- Loss concealment, FEC and RED
- Jitter buffers and NetEq in practice
- Recording and transcription of calls

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI
Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF
Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery
Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing
Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs
Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video
Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Choosing P2P, SFU or MCU for your group-call audio?

We will size the topology for your room sizes and recording needs and estimate server cost per room. 250+ video products shipped since 2005.

Book a 30-min call → **WhatsApp →** **Email us →**