WebRTC topology decision sheet

Mesh vs MCU vs SFU — pick in under a minute. 2026 numbers.

Mesh

Server media role: none.

Client upload: (N−1) × bitrate.

Client download: (N−1) × bitrate.

Practical ceiling: 4–5 participants.

Added latency: ~0 ms.

Best for: 2-party calls, demos, privacy-sensitive one-on-ones.

MCU

Server media role: decode + mix + re-encode.

Client upload: 1 × bitrate.

Client download: 1 × bitrate.

Server cost: 10–50× SFU.

Added latency: 200–400 ms.

Best for: composite output for broadcast, recording, SIP/legacy.

SFU

Server media role: forward only.

Client upload: 1 × bitrate.

Client download: (N−1) × bitrate.

Per node: 500–1,000 video participants.

Added latency: 100–200 ms.

Best for: almost every group call in 2026.

Bandwidth math — 720p at 1.5 Mbps per stream

N	Mesh up (Mbps)	Mesh down (Mbps)	MCU up/down	SFU up / down
2	1.5	1.5	1.5 / 1.5	1.5 / 1.5
3	3.0	3.0	1.5 / 1.5	1.5 / 3.0
4	4.5	4.5	1.5 / 1.5	1.5 / 4.5
5	6.0	6.0	1.5 / 1.5	1.5 / 6.0
6	7.5	7.5	1.5 / 1.5	1.5 / 7.5
7	9.0	9.0	1.5 / 1.5	1.5 / 9.0
8	10.5	10.5	1.5 / 1.5	1.5 / 10.5

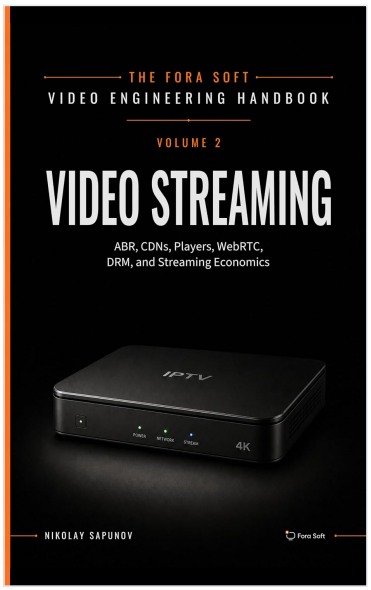
Orange = exceeds typical home upload ceiling (~10 Mbps). At N=7 Mesh becomes unusable.

Decision steps

1. Is the call always two-party? Mesh. No server in the media path. Lowest latency. Cleanest privacy story for telemedicine and one-on-one consultations.
2. Need a single composite stream for broadcast, recording archive, SIP/legacy endpoint? MCU. Accept the 10–50× cost and the 200–400 ms latency — you are paying for the mixer.
3. Otherwise: SFU. Cheap server, full layout flexibility, scales to ~500–1,000 per node.
4. Over 500 concurrent in one room, or audience spread across regions? Cascading SFUs. Multiple nodes, optimised backbone, Redis-coordinated. Scales to millions.

Five pitfalls to avoid

- Assuming "P2P = free". 15–30% of Mesh calls relay through TURN; you pay for that bandwidth.
- Building a Mesh that scales to 8+ participants by adjusting client buffers. The upload ceiling is hardware; no buffer fixes it.
- Choosing MCU because "it's simpler for the client". Modern clients decode N−1 streams without trouble. You are paying for elegance you don't need.
- Running one SFU node and assuming it scales globally. One node is one region; latency past 150 ms across an ocean degrades every call.
- Forgetting simulcast/SVC. Without it the SFU sends 1080p to a phone on 4G and the phone freezes.



GO DEEPER · VOLUME 3 OF 7

Video Streaming: A Complete Guide to Video Delivery

From ABR and CDNs to Players, WebRTC, DRM, and the Economics of Streaming

By Nikolay Sapunov, CEO at Fora Soft · 1,182 pages · Kindle & paperback on Amazon

Read it on Amazon → amazon.com/dp/B0HH76ZYWC

This sheet compares Mesh, MCU and SFU with bandwidth math and a decision path. Volume 3 covers WebRTC and real-time delivery in depth, plus CDN and caching for the broadcast tail and the economics of delivery that decide server cost per call.

WHAT YOU WILL FIND INSIDE

- WebRTC and real-time delivery
- CDN and caching
- Economics of delivery
- ABR and simulcast layers

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

Not sure whether your calls need Mesh, an MCU or an SFU?

We will size bandwidth and server cost for your room sizes and regions and recommend the topology before you write code. 250+ video products shipped since 2005.

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