Real-Time AI Conferencing - Build Blueprint

Fora Soft Learn · AI for Video Engineering · Capstone 10.1

1 - The reference stack: one SFU + the agent pattern

Media plane	Clients -> SFU media server -> clients. WebRTC + Opus. Forwards, never transcodes.
AI plane	One agent joins each room as a participant: STT -> captions, translation, notes.
Edges + control	Token server (auth) - TURN/coturn (hard NAT) - app backend (transcripts).

2 - Build vs buy (2026): adopt infra, buy models, build the product

```
ADOPT  SFU (LiveKit / mediasoup) · agent framework (LiveKit Agents) · TURN (coturn)
BUY    speech-to-text (Deepgram / AssemblyAI) · translation (frontier LLM / MT API)
BUILD  on-device blur (MediaPipe + WebGPU) · token server · app backend · the UX
NOTE   Krisp went paid 2026-05-01; RNNoise unmaintained since 2024; DeepFilterNet free.
```

3 - Two budgets that run in parallel (with the arithmetic)

```
INTERACTIVE MEDIA  one-way ~200 ms.  If network = 120 ms, only ~80 ms is left for
                   capture + on-device AI + encode + playout. Keep on-device models small.
AI VOICE LOOP      ~800 ms.  VAD 50 + STT 150 + LLM 400 + TTS 150 + net 50 = 800 ms.
                   Stream every stage; use a learned turn detector, not a silence timer.
```

4 - Scaling tiers and the cost rule

```
<100 -> single SFU node.  100-1K -> simulcast + SVC (500-800 video/node).
1K+  -> cascaded SFUs across regions (a relay link is just another participant).
COST  on-device = $0 to operator · SFU media ~$0.0005/participant-min · agent ~$0.01/min.
      25-person hour, all 5 features: ~$0 + ~$0.75 + ~$1.20 = under $2.00 (managed).
      Self-host the SFU past ~500 sustained concurrent sessions.
```

5 - Build order: each milestone ships a working product

```
1 Plain call (SFU+token+TURN)  2 On-device blur+denoise  3 Caption agent (streaming STT)
4 Translation + notes (same agent, same transcript)  5 Moderation/avatars + AI Act notice.
```

6 - Go-live checklist (engineering context, not legal advice)

- [ ] Plain call is solid first: audio/video + reconnect on the devices your users actually have.
- [ ] Per-person AI (blur, denoise) runs on the device - never in the cloud per participant.
- [ ] Exactly ONE noise suppressor on raw audio, before Opus; disable the browser built-in if you add yours.
- [ ] SFU never transcodes; moderation runs on a fan-out copy, not inline on the media server.
- [ ] TURN relay deployed with time-limited credentials; ~15-20% of networks need it to connect.
- [ ] Per-session observability (connect time, packet loss, agent latency) wired in from milestone 1.
- [ ] Disclose AI-generated or altered voice/face to participants - EU AI Act Article 50 (from 2026-08-02).