WebRTC Scale-Up Pre-Flight Checklist

30 items to verify before scaling a WebRTC product — single SFU, LastN, cascade, bridge, AI agents

1. Single-SFU baseline

- [ ] Per-node concurrent-participant ceiling measured under realistic load — not vendor-claimed.
- [ ] Egress NIC and CPU instrumented and dashboarded — alerts at 70% utilisation.
- [ ] DTLS handshake failure rate alerted on (cert mismatch, version, cipher).
- [ ] Process supervision: SFU restarts cleanly on crash, calls drain on shutdown.
- [ ] RTCP feedback loops and ICE consent-freshness load measured at peak.

2. LastN, simulcast, SVC

- [ ] LastN forwarding enabled by default for conferences past ~12 participants.
- [ ] Simulcast layers (typically 3) published by every video sender — verify in chrome://webrtc-internals.
- [ ] Dominant-speaker switching tested for stability — no rapid flapping under crosstalk.
- [ ] Audio-level RTP header extension (RFC 6464) negotiated on every publisher.
- [ ] SVC (VP9 or AV1) considered for the next generation; H.264 simulcast acceptable in 2026.

3. Cascade design

- [ ] Regions chosen by user density — typical: us-east-1, eu-central-1, ap-northeast-1, ap-southeast-1.
- [ ] Room-to-node routing tested: participant locality vs room affinity policy documented.
- [ ] Inter-node transport runs on private backbone, not public internet.
- [ ] Inter-node auth: mutual TLS or HMAC pre-shared key — rotation documented.
- [ ] Mesh failure drill: one node killed in staging; remaining nodes pick up the slack within 10 s.

4. Coordination layer

- [ ] Quorum-backed registry (Redis cluster, etcd, Hazelcast) — not a single Redis box.
- [ ] Room-roster propagation across nodes: sub-second target (50–200 ms typical).
- [ ] Coordination-layer outage drill: cluster loses a member; room state survives.
- [ ] Registry sized for the same blast radius as the SFU fleet — same SLA, same ops budget.
- [ ] Per-room state size capped; large rooms shard their roster across keys.

5. Broadcast bridge (when audience > 5K)

- [ ] LL-HLS / chunked CMAF packager subscribes to the SFU as a participant.
- [ ] Active-speaker layout composition tested — clean cuts, no flashing.
- [ ] CDN: at least one production CDN (Cloudflare Stream, Fastly, CloudFront, Akamai).
- [ ] Broadcast latency budget documented: 2–5 s glass-to-glass acceptable.
- [ ] Reactions / chat path: WebSocket loopback to the interactive tier.

6. AI agent worker fleet

- [ ] Framework chosen: LiveKit Agents 1.5+ or Pipecat 1.0+ (or commercial equivalent).
- [ ] Agent worker co-located with regional SFU — same AZ, ideally same VPC.
- [ ] End-of-speech to start-of-response measured at p50 and p95; target <800 ms p50.
- [ ] VAD → STT → LLM → TTS pipeline instrumented per stage; bottleneck stage identified.
- [ ] Interruption handling tested: TTS halts when user starts speaking; partial transcript committed.