Cost-Optimized Voice AI Infrastructure (1.0–1.2s Latency)

Target

- Concurrency: 30 simultaneous calls
- Latency: Avg 0.9–1.1 s, p95 ≤ 1.2 s
- Stack: faster-whisper (STT), VeenaTTS (TTS), LiveKit, LLaMA-70B (Async)

Key Optimization Strategy

- STT and TTS share a single GPU to reduce cost
- Strict GPU semaphore limits prevent contention
- LLM runs asynchronously and is not in the audio critical path

Server 1 – STT + TTS (Shared GPU)

GPU: A40 (48GB)

CPU: 16 vCPU | RAM: 64GB

Shared GPU Concurrency: 2

Models: faster-whisper (medium), VeenaTTS

Latency Contribution: 550–850 ms (including queueing)

Server 2 – LiveKit Media

CPU-only: 8–12 vCPU | 24–32GB RAM

Latency Contribution: 20–40 ms

Server 3 – LiveKit SIP

CPU-only: 6–8 vCPU | 16GB RAM

Latency Contribution: 30–50 ms

Server 4 – LLaMA-70B (Async Inference + Fine-Tuning)

2× A300 40GB or 1× A100 80GB

CPU: 32 vCPU | RAM: 128GB

Latency: 400–800 ms (parallel, non-blocking)

End-to-End Latency (Perceived)

LiveKit + SIP: 50–70 ms

STT + TTS (shared GPU): 550–850 ms

LLM: Parallel (hidden)

Total Avg: 900 ms – 1.1 s

p95 Worst Case: ≤ **1.2 s**

Trade-offs

- Slight latency jitter under bursts
- Limited headroom beyond 5 concurrent calls
- Requires careful GPU concurrency control

Summary

This architecture achieves acceptable 1.0–1.2s latency at significantly reduced cost by consolidating STT and TTS onto a single A40 GPU while keeping LLM inference asynchronous.