Шпаргалка: бюджет задержки real-time AI

Fora Soft Learn · AI for Video Engineering

1 — Цель (mouth-to-ear задержка, ITU-T G.114)

```
< 150 мс = прозрачно  |  150-400 мс = заметно, но рабоче  |  > 400 мс = недопустимо
Разговорный AI проектируйте под < 200 мс (разрыв реплик); цельтесь в 100 мс — для запаса.
```

2 — Фиксированный конвейер (вы им НЕ управляете — сложите первым)

Захват (камера + микрофон)	3-10 мс
Видео кодек / декод (железо, каждый)	~10 мс
Аудиокодек — Opus, RFC 6716 (по умолч.)	26.5 мс
Распространение по сети	~10 мс на 1000 км RTT
TURN-релей / SFU-форвардинг	10-30 / 5-20 мс
Джиттер-буфер (real-time, адаптивный)	10-50 мс
Рендер на экран (60 Гц)	8-16 мс

3 — Ориентиры задержки AI-инференса (2026, быстро меняются)

Сегментация на устройстве — MediaPipe (GPU)	< 3 мс / кадр
Первые слова ASR — Deepgram Nova-3	~150 мс интерим, < 300 мс
Первый звук TTS — ElevenLabs Flash / Cartesia Sonic	~75-90 мс
Голос-в-голос — OpenAI Realtime	~300-500 мс (НЕ влезает вживую)

4 — Налог скорости света (география, а не скорость модели)

```
Свет в волокне ~200 000 км/с -> ~10 мс RTT на 1000 км. Лондон <-> С. Вирджиния ~60 мс RTT
ещё до роутеров. Далёкий облачный GPU тратит весь бюджет на одну только сеть.
```

5 — Правило размещения + прикидка на салфетке

- [ ] Сначала сложите фиксированную часть. Региональный звонок: 60 (сеть) + 26.5 (Opus) + 40 (джиттер) + ~35 (захват/кодек/декод/рендер) = ~161 мс.
- [ ] Остаток на AI = цель − фиксированная сумма. Против 200 мс это ~39 мс. Модель должна влезть в НЕГО — иначе функция не real-time.
- [ ] Мгновенно + малая модель -> НА УСТРОЙСТВЕ (без сети). Вживую + большая -> EDGE (< 20 мс).
- [ ] Терпит 1-3 с -> ОБЛАКО (саммари, аналитика, копилоты). Не воюйте с физикой ради этого.
- [ ] Тестируйте на той сетевой дистанции, что у реальных пользователей — не рядом с сервером.