

Карта архитектуры гибридного стека

Одна страница. Выберите гибридный шаблон под продукт и соберите стек.

1. Четыре канонических шаблона

Шаблон А — интерактивная вершина, broadcast-хвост

Live-шопинг, аукционы, town hall, watch-party.

Интерактивно WebRTC или WHEP для ведущего и покупателей в кадре (≤ 50 паблишеров).	Broadcast LL-HLS или LL-DASH поверх CMAF для пассивной аудитории (1k–10M).
---	---

Почему этот шаблон. Лаг поток-чат больше 10 с убивает конверсию; чистый WebRTC на 50k зрителях — это 3–10х от LL-HLS на массовом HTTP-CDN.

Шаблон В — broadcast-основа, preview модератора

Premium-спорт, live-новости, регулируемый broadcast.

Интерактивно Приватный WHEP-эндпоинт для пультавой (10–100 зрителей).	Broadcast LL-HLS или LL-DASH для публичного зрительского тарифа (миллионы).
--	--

Почему этот шаблон. Обычный HLS добавил бы 2–3 с слепого пятна для продюсера; приватный WHEP-preview убирает это для ограниченной

Шаблон С — реалтайм-ядро, запись-хвост

Видеосвязь, телемедицина, e-learning.

Интерактивно WebRTC SFU полностью (каждый живой участник интерактивен).	Broadcast HLS-запись для VOD-воспроизведения после сессии.
--	---

Почему этот шаблон. WebRTC — единственный вариант для живой встречи; повтор на массовом HTTP примерно в 100х дешевле живого SFU.

Шаблон D — MoQ A/B параллельно с LL-HLS

Live с прицелом в будущее (спорт, ставки, реалтайм).

Интерактивно Опт-ин доля на Media over QUIC (draft-ietf-moq-transport-17, янв. 2026).	Broadcast LL-HLS поверх CMAF — несущий путь для остальной аудитории.
--	---

Почему этот шаблон. MoQ ещё Internet-Draft, есть риск несовместимости; параллельный A/B растит операционные мышцы, пока LL-HLS держит нагрузку.

2. Стыковка внутри origin — упаковать один раз, доставить многократно

Главная инженерия гибридного стека — посередине, в момент, когда контрибьюция-поток становится одним или несколькими потоками доставки. Дисциплина: транскодировать входящий RTMP / SRT / WHIP один раз в источник CMAF (ISO/IEC 23000-19) fragmented MP4. С него отдавать HLS-манифест, DASH-манифест и (если нужно) WHEP-egress эндпоинт. Байты аудио и видео идентичны на каждом отрезке — различаются только формат манифеста, транспорт и фрейминг.

3. Деление счёта по роли, а не по аудитории

Чистый WebRTC для всех
50 021 одновременных SFU-сессий по 4 Mbps.
Часовая стоимость WebRTC-сервера: ~ \$3 000–9 000.
Задержка для всех: суб-секунда (избыточно для 50 000 в чате).

Гибридный WebRTC + LL-HLS
21 SFU-сессия + 50 000 кэшируемых LL-HLS зрителей.
Стоимость: ~ \$900 (LL-HLS CDN) + \$10 (SFU на 21 чел.) = ~ \$910.
Задержка по роли: суб-секунда у 21 в кадре, 2–4 с у остальных.

Соотношение: ~ 3–10 раз в пользу гибридного стека при том же продукт-опыте.

4. Переход с чистого WebRTC на гибридный стек

Шаг 1 — разделить аудиторию. Найти минимальный набор, кому реально нужна суб-секунда (ведущий, покупатели в кадре, модераторы, продюсер).

Шаг 2 — добавить CMAF + LL-HLS поверхность. На origin подписаться на дорожку ведущего из SFU и перекодировать в CMAF fragmented MP4. Отдавать LL-HLS-манифест за массовым HTTP-CDN.

Шаг 3 — оставить чат реалтаймовым. WebSocket-сайдкар. Чат не на медиа-пути; не давайте ему WebRTC-бюджет задержки.

Шаг 4 — оставить WebRTC для интерактивного кольца. SFU не удаляйте. Ведущий и покупатели в кадре по-прежнему публикуют и подписываются через него.

Шаг 5 — измерить счёт. SFU теперь обслуживает 20–50 одновременных паблишеров и сабскрайберов вместо 50 000; LL-HLS обслуживает 50 000 на массовом HTTP. Снижение egress-расходов в 6–10 раз при той же аудитории и том же продукт-опыте