

Просчитайте арифметику конкурентности для своей аудитории и сверьте архитектуру с рычагами, сгибающими кривую затрат. Инженерное руководство — egress CDN тарифицируется ступенчато, зависит от коммита и считается по 95-му перцентилю, поэтому уточняйте свою ставку и фиксируйте дату.

1 · АРИФМЕТИКА — заполните под свой пик (конкурентность — главное число) 3 · ДВА ПЛАНА, ДВА ПЛАНИРОВАНИЯ (считайте отдельно — частая ошибка)

- ☐ **Пиковая конкурентность.** Максимум одновременных потоков, который вы ждёте — НЕ зарегистрированные, НЕ среднее. Под него и считайте всё.

- ☐ **Битрейт на поток.** ~4 Mbps для 1080p на современном кодеке. Ниже с HEVC/AV1. _____ Mbps
- ☐ **Пропускная способность = конкурентность × битрейт.** 1 000 000 × 4 Mbps = 4 Tbps. Ваше число: _____
- ☐ **Egress на зритель-час.** битрейт × 3 600 с ÷ 8 = ~1,8 ГБ/ч при 4 Mbps. Стоимость в час = конкурентность × 1,8 ГБ × ваша \$/ГБ. При \$0,04/ГБ и 1 млн зрителей ≈ \$72 000/ч.

2 · РЫЧАГИ, КОТОРЫЕ СГИБАЮТ КРИВУЮ (отметьте, что уже есть в проекте)

- ☐ Cache-hit ratio (offload origin) выше 85% — а для live лучше за 90%.
- ☐ Включён origin shielding: десятки промахов на эджах схлопываются в ~1 запрос к origin на объект.
- ☐ Доставка переносима между несколькими CDN (устойчивость + региональный охват).
- ☐ Content steering (ETSI TS 103 998 / HLS EXT-X-CONTENT-STEERING) переключает зрителей между CDN на лету — без кастомного кода плеера.

ОДНО ПРАВИЛО МАСШТАБИРОВАНИЯ

Платформу ОТТ задаёт конкурентность, а не размер каталога. Видеопуть (data plane) растёт с полосой и укрощается cache offload — высоким cache-hit, origin shielding и несколькими CDN с content steering. Путь «решать и записывать» (control plane) растёт с запросами в секунду и укрощается прогревом заранее под кривую прихода, ведь thundering herd live-премьеры приходит в одни и те же шестьдесят секунд. Считайте каждый план отдельно, проектируйте под отказ и нагружайте на пик до события — никогда не в проде в самую важную ночь.

- ☐ **Data plane** считается в полосе (Tbps, egress ГБ): кодирование, упаковка, origin, CDN, плеер. Рычаг: cache offload.
- ☐ **Control plane** считается в запросах/с: вход, права, биллинг, реклама, аналитика. Рычаг: автоскейлинг.
- ☐ Права — единый сервис, который вызывает каждый плеер, а не копия логики в каждом приложении.
- ☐ Low-latency live (LL-HLS / LL-DASH) умножает нагрузку мелкими запросами — заложите её до включения.

4 · ПЕРЕД ГЛАВНОЙ НОЧЬЮ (ворота готовности)

- ☐ Спрогнозировали ли ПИКОВУЮ конкурентность отдельно для каталога и для live-всплеска?
- ☐ Прогрет ли control plane ЗАРАНЕЕ под кривую прихода (предиктивный автоскейлинг)?
- ☐ Деградируем ли мягко — комната ожидания, retry/backoff — вместо жёсткого отказа?
- ☐ Нагрузили ли всю платформу на целевой пик в репетиции, а не в проде?
- ☐ Есть ли субсекундный автофейловер между CDN, если один деградирует на всплеске?