

Вести OTT-платформу вживую — две задачи: дашборд информирует, алерт прерывает. Следите за четырьмя «золотыми» сигналами, будьте человека только по симптомам, что чувствует зритель, соразмеряйте алерты со скоростью расхода бюджета ошибок и подкрепляйте всё разумной моделью on-call. Правило под всем: пейджер, что звонит по любому поводу, — пейджер, который никто не читает.

1 · ЧЕТЫРЕ «ЗОЛОТЫХ» СИГНАЛА (СЛЕДИТЕ)

- ☐ **Latency** — старт, манифест, доставка сегмента; берите p99 и латентность сбоя отдельно.
- ☐ **Traffic** — одновременные зрители, запросы/сек, egress; конкурентность — пульс события.
- ☐ **Errors** — сбои воспроизведения, CDN 4xx/5xx, ошибки лицензий; берите долю, не счёт.
- ☐ **Saturation** — запас origin, транскодера, cache-hit, лиц. сервера; хвост латентности предупреждает первым.

2 · АЛЕРТ ПО СИМПТОМАМ, НЕ ПО ПРИЧИНАМ

- ☐ **Симптом, не причина** — пейдж на «сбои воспр. в Германии», а не на каждый внутренний 503.
- ☐ **Пройдите 4 теста** — срочно, действенно, видно зрителю и не дублирует чужой пейдж.
- ☐ **Верное представление** — перцентиль, не среднее; доля, не счёт; скорость, не текущее значение.
- ☐ **Убейте alert fatigue** — несрочные причины на дашборд; оставьте лишь пейджи, ради которых стоит проснуться.

ЧТО НА САМОМ ДЕЛЕ ДАЁТ 99,9% — И ПОЧЕМУ ВНУТРИ СТРОЖЕ

SLA «99,9% аптайма» звучит надёжно, но допускает реальный простой: около 43 минут в месяц, или ~8,8 часа в год — и ничто не мешает этим минутам прийти на ваше крупнейшее живое событие. Каждая лишняя «девятка» резко срезает допуск: 99,95% — около 22 минут в месяц, 99,99% — около 4,4 минуты, 99,999% — около 26 секунд — и каждая девятка стоит непропорционально дорожке в постройке и эксплуатации, так что большинству OTT-сервисов телеком-уровень не нужен и не по карману. Не менее важно, что SLA реально выплачивает при нарушении. Типичное средство CDN — сервисный кредит: например, Amazon CloudFront возвращает 10% счёта за месячный аптайм 99%–99,9% и 25% ниже 99% — это доля того, что вы заплатили вендору, а не выручка от подписок и рекламы и не репутация, потерянные, пока зрители смотрели на спиннер. Относитесь к SLA как к инструменту биллингового риска, а не к страховке. Практическая позиция: выберите внутренний SLO, который бизнесу действительно нужен, задайте клиентский SLA слабее этого SLO, чтобы ваш сигнал сработал раньше нарушения обещания, держите более одного CDN, чтобы плохая ночь одного вендора была переключением, а не простоем, и читайте каждый SLA вендора на исключения и расчёт кредита до подписания. А числа подкрепите людьми: устойчивой ротацией on-call, ясными уровнями severity, проверенными runbook, чтобы ответ в 3 ночи был процедурой, а не импровизацией, и безвиновными postmortem, превращающими каждый инцидент в постоянное исправление, а не в поиск виноватого.

3 · SLO → БЮДЖЕТ ОШИБОК → BURN RATE

- ☐ **Задайте SLO** — напр. 99,9% успешных; держите клиентский SLA слабее, чтобы сработать первым.
- ☐ **Бюджет ошибок** — $100\% - 99,9\% = 0,1\%$ допустимо сбоя; «не падать» превращается в число.
- ☐ **Пейдж на быстрый burn** — $1\text{ч burn} \geq 14,4\text{х}$ (бюджет за ~2 дня) будит on-call сейчас.
- ☐ **Ticket / review** — $6\text{ч burn} \geq 6\text{х}$ — тикет; $3\text{д burn} \geq 1\text{х}$ — еженедельный обзор.

4 · SLA И МОДЕЛЬ ON-CALL

- ☐ **Знайте downtime** — $99,9\% = \sim 43\text{ мин/мес}$; $99,99\% = \sim 4,4\text{ мин/мес}$. Минуты могут выпасть на премьеру.
- ☐ **Читайте средство** — кредит CDN (напр. CloudFront 10%/25%) возвращает счёт, не потерянную выручку.
- ☐ **Ротация on-call** — primary + secondary, короткие смены; следите за нагрузкой пейджера.
- ☐ **Severity + postmortem** — Sev-1/2/3 задаёт ответ; безвиновный разбор чинит систему, а не ищет виноватых.