

Локальный AI-сервер - чек-лист подбора

Считайте edge-сервер в потоках вашей реальной модели, а не в TOPS из datasheet. Сначала бюджет декодирования, потом инференс, всегда - запас.

А. Считайте в потоках, не в TOPS

- ☐ Посчитайте камеры и их разрешение - 4K грузит GPU куда сильнее, чем 1080p.
- ☐ Сначала декодирование: уточните число аппаратных декодеров и потоков на декодер.
- ☐ Затем инференс: узнайте потоки-на-GPU для ВАШЕЙ модели, разрешения и fps.
- ☐ Считайте ~16-40 потоков лёгкого детектора на GPU; тяжёлые модели - много меньше.

В. Выберите железо

- ☐ Подберите класс GPU под задачу (массовый инференс-GPU или малый edge-модуль).
- ☐ Подтвердите число GPU на сервер и что в помещении хватит питания и охлаждения.
- ☐ Решите: своя сборка (NVIDIA-Certified) или готовый аплайанс - контроль против скорости.
- ☐ Подтвердите, что аналитика поддерживает ваши GPU и работает полностью локально.

С. Заложите отказ и рост

- ☐ Один сервер - единая точка отказа: заложите второй или путь резервирования.
- ☐ Подтвердите, что можно добавлять GPU или серверы по мере роста числа камер.
- ☐ Подтвердите, что VMS / аналитика объединяют несколько серверов в одну систему.
- ☐ Сравните разовую стоимость сервера с постоянным облачным egress, который он заменяет.

Д. Резидентность и комплаенс

- ☐ Уточните, должно ли видео остаться в здании или в стране (GDPR глава V).
- ☐ Для биометрии СНАЧАЛА подтвердите правовое основание (EU AI Act, GDPR ст. 9, BIPA).
- ☐ Подтвердите, что узнаваемое видео не покидает границу при требовании резидентности.
- ☐ Зафиксируйте срок хранения и кто имеет доступ к результату локальной аналитики.

Помните: аппаратные декодеры GPU заполняются раньше, чем его AI-вычисления, поэтому сервер, посчитанный только в TOPS, потянет меньше камер, чем кажется. Локально лучше, когда модель тяжёлая или кросс-камерная, данные должны остаться у вас, или egress дороже коробки. Инженерное руководство, не юридический совет. Источники: документация NVIDIA L4 / DeepStream / Jetson; GDPR Reg. (EU) 2016/679 гл. V и ст. 5; EU AI Act Reg. (EU) 2024/1689; Illinois BIPA 740 ILCS 14; ONVIF Profile M.