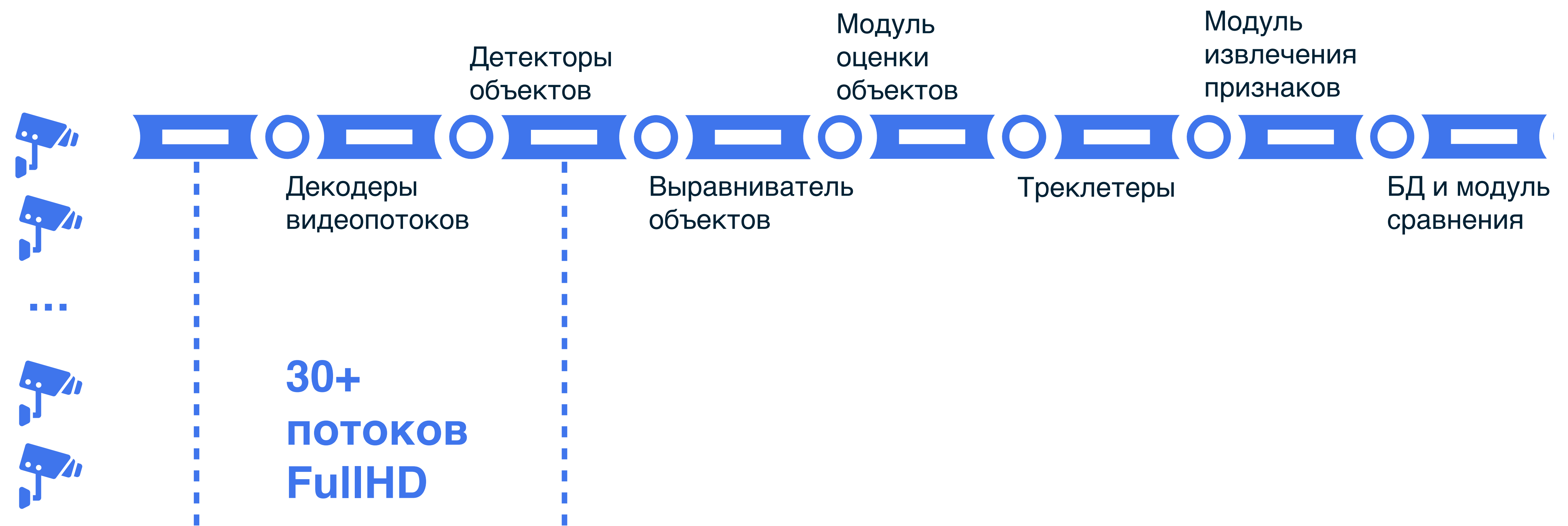


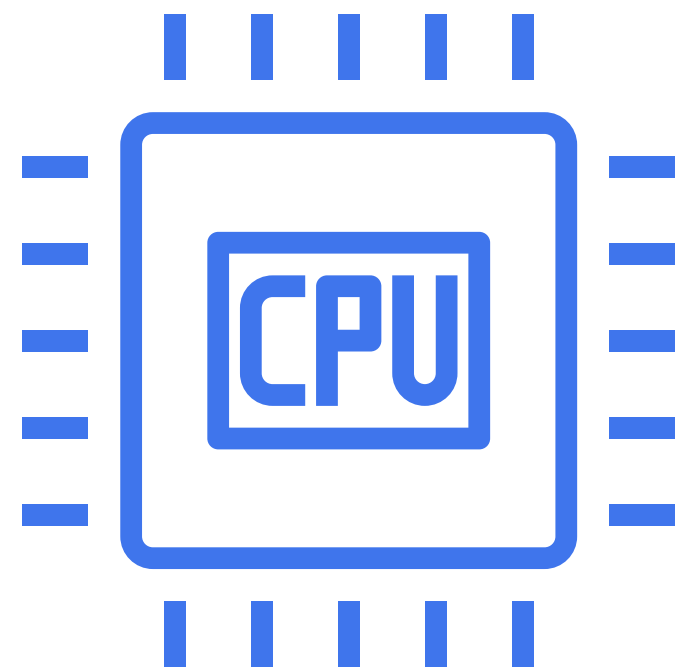
У Faceter есть интеллект  
и зрение: как мы  
детектируем лица  
и объекты с 30+ Full-HD  
камер на одном ускорителе



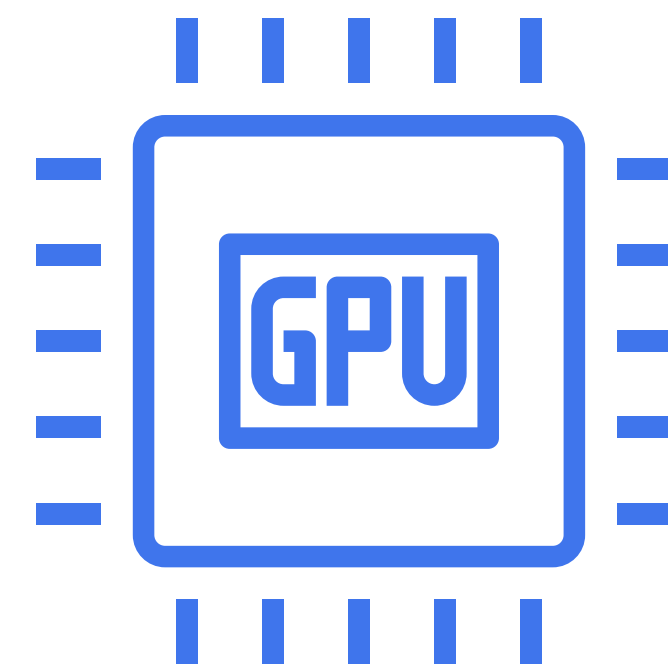
[faceter.cam](https://faceter.cam)

# 1 Конвейер аналитики Faceter





- 2 - 32 вычислительных ядра - “Профессора”
- Высокая тактовая частота
- Расширения MMX, SSE, AVX - ассистенты профессоров
- Универсальность по отношению к исполняемым задачам



- 1000 - 5000 вычислительных ядер - “Студенты”
- Более низкая тактовая частота
- Массивный параллелизм с большими объемами данных



## Модуль декодирования видеопотоков



- Использование специализированных ядер GPU для декодирования
- Постобработка (конвертация из формата YUV в RGB) на GPU
- Результат декодирования - фреймы, пригодные для анализа, находятся в памяти GPU
- Возможен доступ к данным из другого процесса с использованием механизма IPC (inter-process communication)
- 2 режима работы: в реальном времени (работа с real-time потоком) и “так быстро, как могу” - обработка архива видео, например, обработка 10 минутного ролика занимает около 25 сек. включая детекцию лиц.



## 4 Модуль детекции лиц

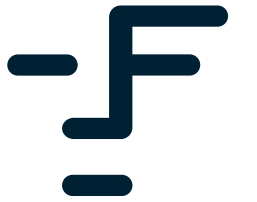


- Препроцессинг на GPU (ресайзинг, вычитание mean, конвертация в float матрицу) в параллельном потоке
- Облегченная SSD архитектура нейронной сети на GPU
- Постпроцессинг (расчет якорных боксов, алгоритм NMS, применение порогов)

**2**мс/кадр







## Декодер

- Результирующие фреймы находятся в памяти GPU

## Детектор

- Входной фрейм для детекции должен быть на GPU

### Проблема

Декодер и детектор - 2 разных процесса и не имеют прямого доступа к памяти друг друга

### Решение

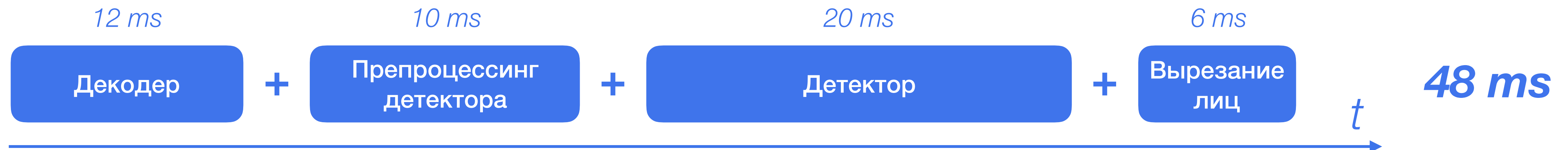
Механизм IPC (inter-process communication):

- В процессе декодера получаем IPC-handle участка памяти с фреймом
- Передаем IPC-handle в процесс детектора
- Используя IPC-handle получаем доступ к исходному участку памяти в контексте процесса детектора

# Распараллеливание процессов



Последовательное выполнение этапов (батч 10 кадров)



Параллельное выполнение задач (батч 10 кадров)



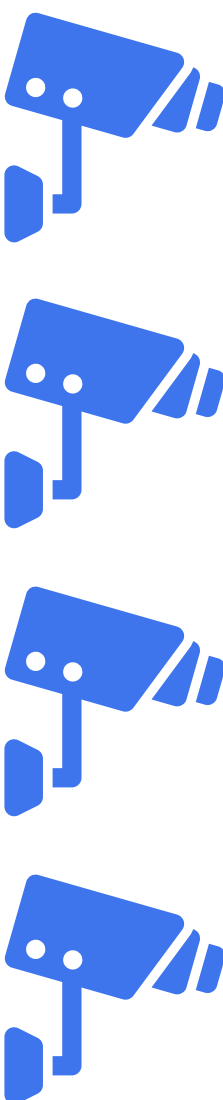


## Менеджер подключений

- Создает декодеры и детекторы в соответствии с таблицей связей
- Перезапускает декодеры при обрыве соединения
- Открывает/закрывает IPC каналы
- Балансирует нагрузку



### Декодеры



- Декодируют фреймы
- Отправляют сообщения с id стрима, ipc-handle, прочей служебной информацией о декодированных кадрах
- Сообщают менеджеру подключений о обрыве соединений, состоянии буферов кадров



### IPC каналы

#### Лица

- Передает broadcast /unicast сообщения между декодерами и детекторами

#### Авто. номера

- Передает broadcast /unicast сообщения между декодерами и детекторами

...



### Детекторы

- Получают сообщения о декодированном кадре
- На основе ipc-handle получают доступ к декодированным кадрам
- Производят препроцессинг кадров перед подачей в нейронную сеть
- Детектируют объекты
- Отправляют данные о детекции на следующий этап конвейера





- Увеличение производительности в 6 раз, по сравнению с первоначальным вариантом
- Гибкая балансировка нагрузки, в том числе между устройствами GPU
- Относительно небольшая загрузка CPU

**31 FullHD  
ВИДЕОПОТОК  
декодирование и  
детекция в real-time**

## Ключевые моменты решения

- Максимальное использование аппаратной ускорения и GPU вычислений
- Исключение копирования больших объемов данных между модулями, с использованием механизма IPC как для CPU данных, так и для GPU данных
- Максимальное распараллеливание процессов

Тестирующая конфигурация

- CPU: Intel Core i7
- GPU: Nvidia GTX 1080Ti

Тестируемый видеопоток

- Разрешение 1920x1080
- 20 fps
- Одно лицо в каждом кадре



В пике загрузка ядер GPU и декодинга: 80-85%

- Оптимизация выделения памяти под данные (кэш фреймов, расшаривание между процессами единого блока памяти с указанием смещения для каждого фрейма)
- Оптимизация текущего "узкого места" (вынесение постпроцессинга в отдельный поток, использование 16bits float в нейросети)
- "Спящий режим" - пониженная частота обработки кадров, при отсутствии детекций
- Распределенные вычислительные мощности на внешних модулях на стороне клиента (nVidia Jetson, Intel Neural Compute Stick, Google Coral Dev Board)



Спасибо

Будем рады ответить  
на любые ваши вопросы

[faceter.cam](https://faceter.cam)