Viaduct: организация инференса видеоаналитики

Зитцер Данил, ivideon

21 ноября 2023

О себе

Team Lead @ ivideon

Video analytics

@DanilZittser

ivideon

Ведущий поставщик
облачных видеорешений

13 лет
на мировом рынке

с 2012 года
обладаем экспертизой
внедрения решений
в портфели крупных компаний

250 тыс.
камер, подключенных
к системе

охват
+5 млн
клиентов в
130 странах мира

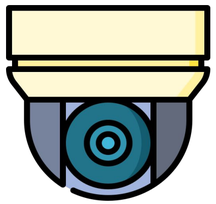
Наши решения видеоаналитики:

- Распознавание лиц, СКУД
- Распознавание автономеров
- Retail Analytics Box
- Подсчёт трафика посетителей
- Контроль очередей, отсутствия сотрудника на рабочем месте и др.
- Контроль кассовых операций

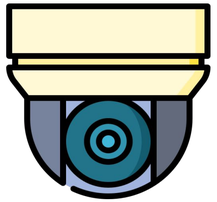
Как устроена обработка?

Ivideon Server → Cloud

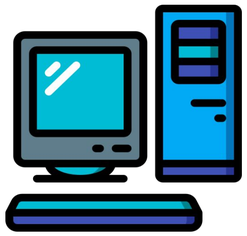
1



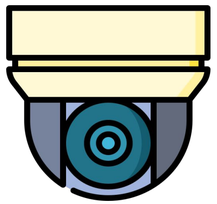
2



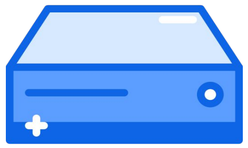
+



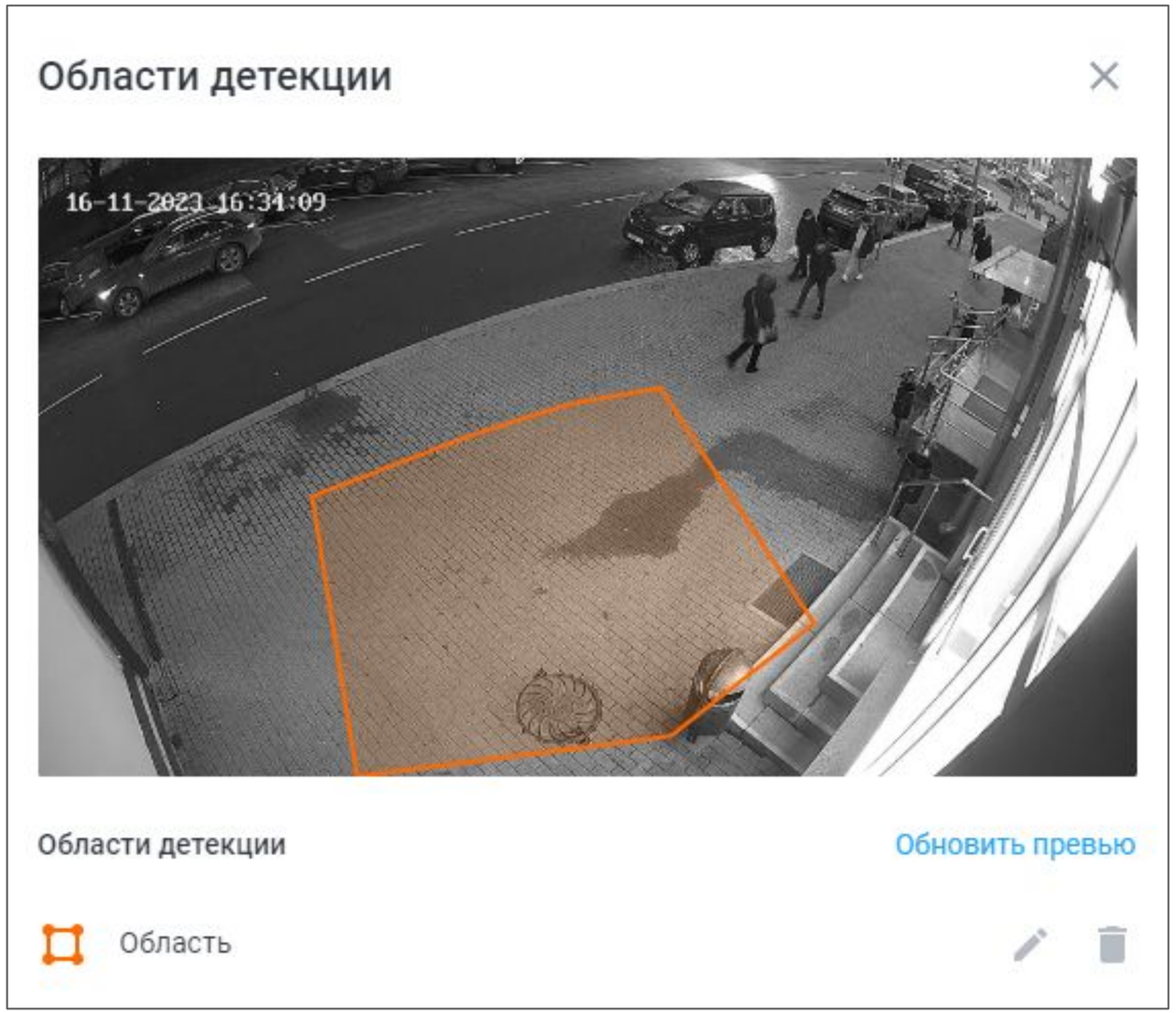
3



+

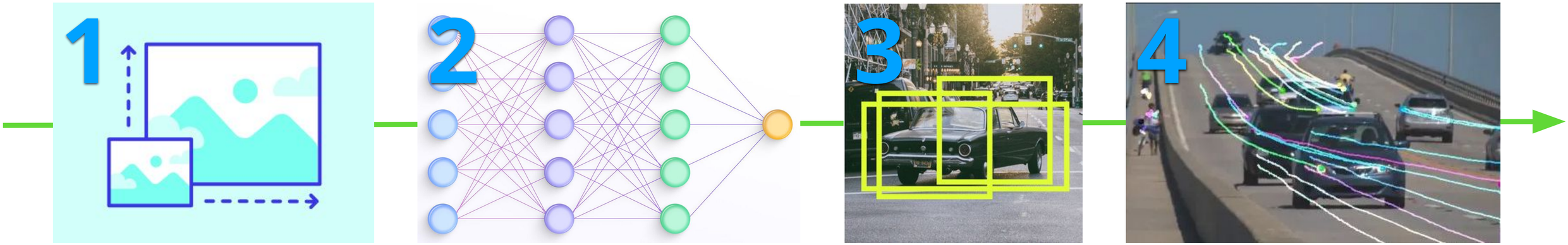


Core-аналитика



Детекция + Трекинг

Базовый пайплайн



Препроцессинг → Инференс (CPU / GPU) → Постпроцессинг → Эвристика (состояние + настройки)

Legacy-аналитика

H.265
HEVC
High Efficiency Video Coding

- C++
- YOLO V4 Tiny @ RTX A2000
- OpenCV DNN
- 17 каналов

- Python
- SSD MobileNet V2
- OpenVINO
- Сколько каналов?

Inference Engine

- Model Zoo
- CPU Performance
- Ecosystem

OpenVINO™

Синтетика

Flags:

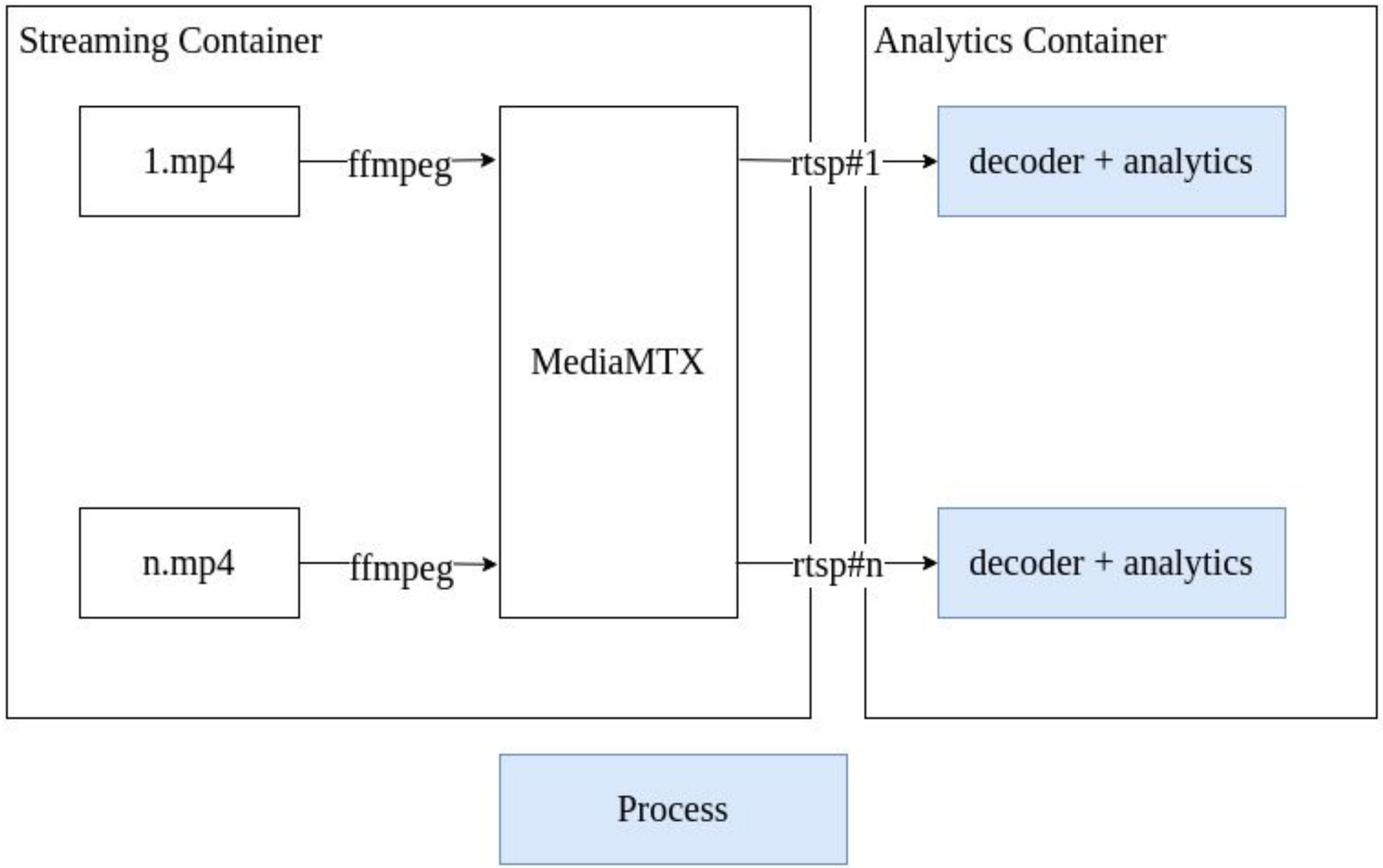
- AVX-512
- AVX512_VNNI ❗

Intel(R) Xeon(R) Silver 4314
CPU @ 2.40GHz, 64 core

Accuracy	API	Hint	FPS	CPU, %
FP32	Sync	Latency	165	1220
		Throughput	68	350
	Async	Latency	300	2423
		Throughput	376	2850
FP16	Sync	Latency	296	1582
		Throughput	117	391
	Async	Latency	353	2811
		Throughput	409	3124
INT8	Sync	Latency	718	1591
		Throughput	322	397
	Async	Latency	1389	3198
		Throughput	1899	3172

```
benchmark_app -m model.xml
```

Результаты эксперимента



12-13 каналов

Load Average > 150

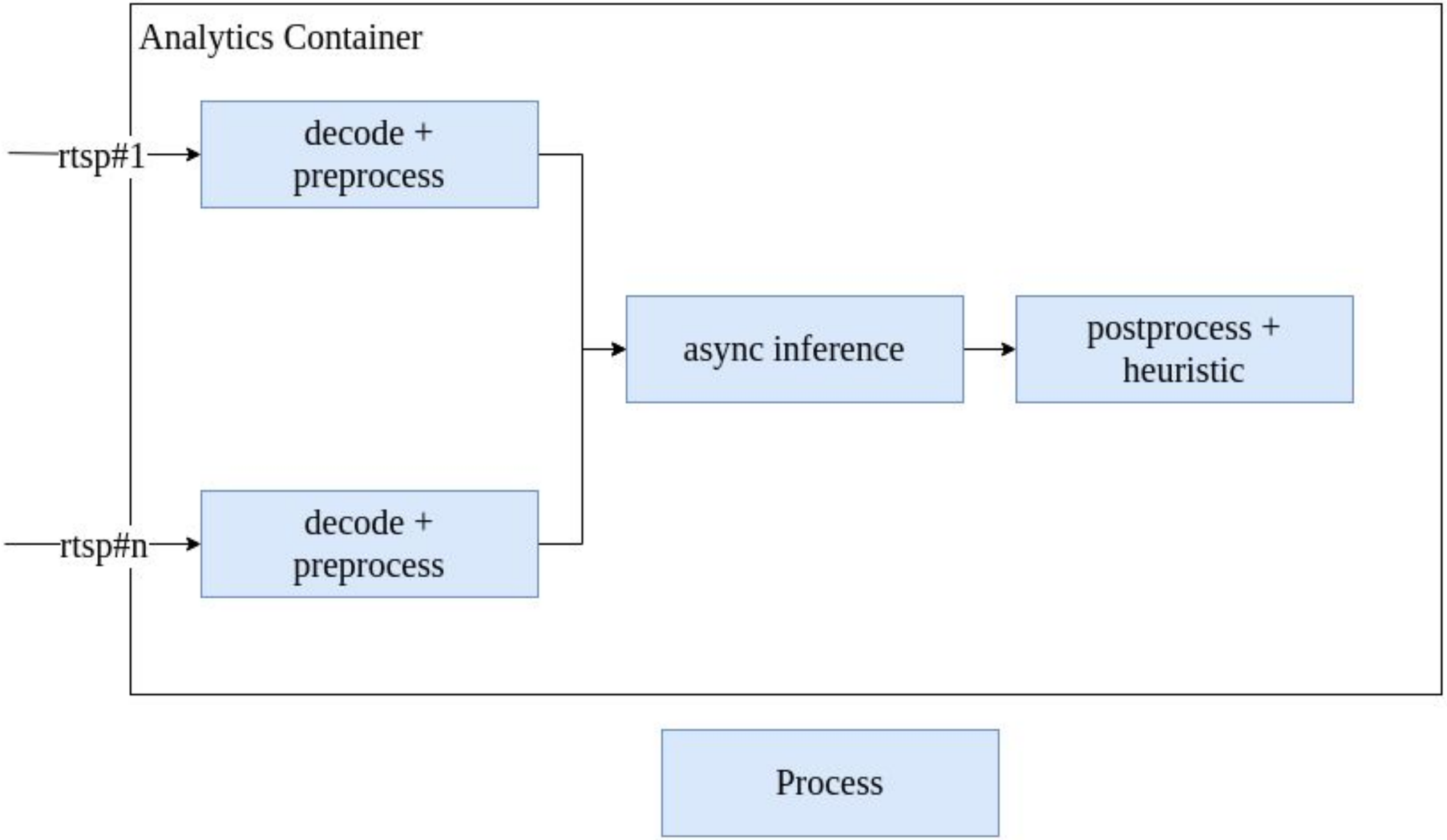
Inference Request

- Intel® Xeon® Silver 4316 - Q2'21 - 40 InferRequests
- Intel® Xeon® Platinum 8380 - Q2'21 - 80
- Intel® Xeon® Platinum 8490H - Q1'23 - 120

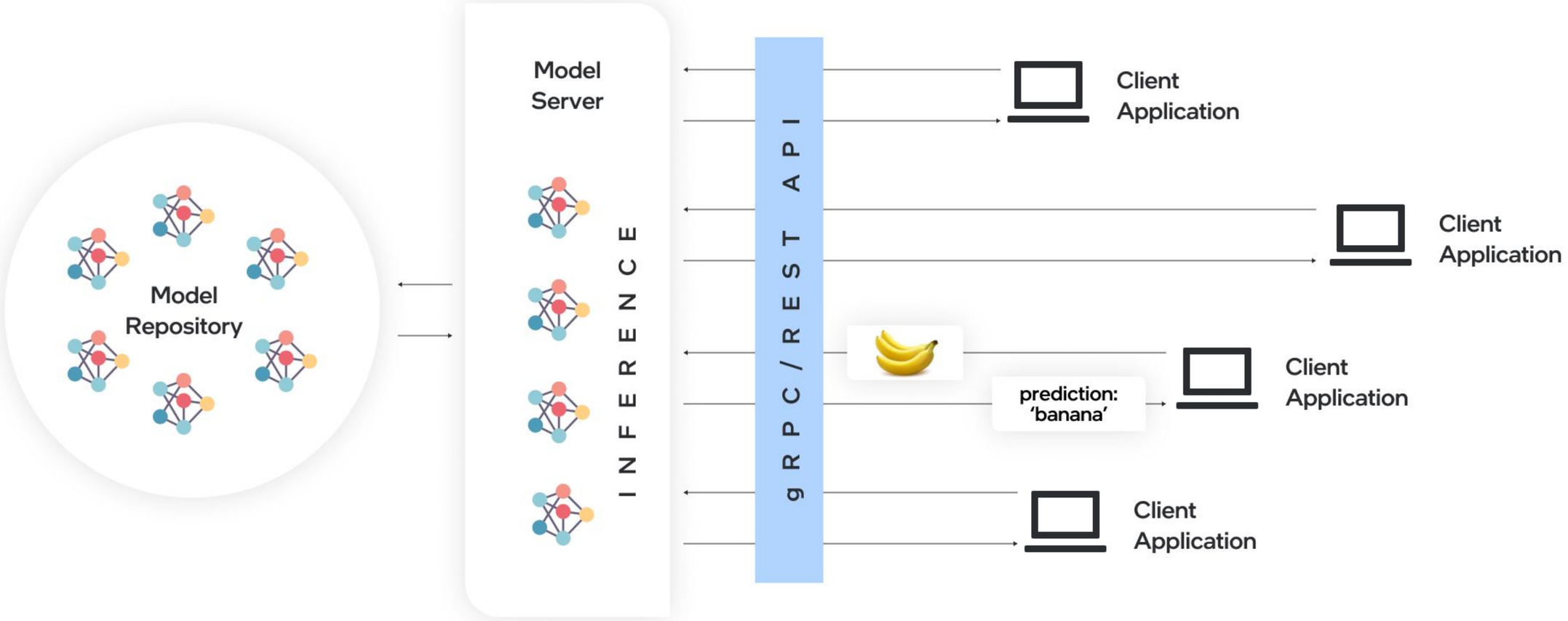
- Core i9-12900TE - Q1'22 - 10
- Core i5-13600K - Q4'22 - 14
- Core i9-13900K - Q4'22 - 24

* https://docs.openvino.ai/2023.1/_static/benchmarks_files/OV-2023.1-Platform_list.pdf

Желаемая схема



OpenVINO™ Model Server



А что ещё?

avito-tech/
aqueduct

Framework for create performance-efficient prediction

8 Contributors

38 Used by

151 Stars

14 Forks

Viaduct

Handler

```python
class Handler():
    """Интерфейс компонента аналитики."""

    def on_start(self) -> None:
        """Например, сюда помещаем:
        - загрузку весов модели;
        - инициализацию состояний.
        """
        pass

    @abstractmethod
    def handle(self, *args, **kwargs) -> Any:
        raise NotImplementedError

    def on_exit(self) -> None:
        pass
```

Задание flow

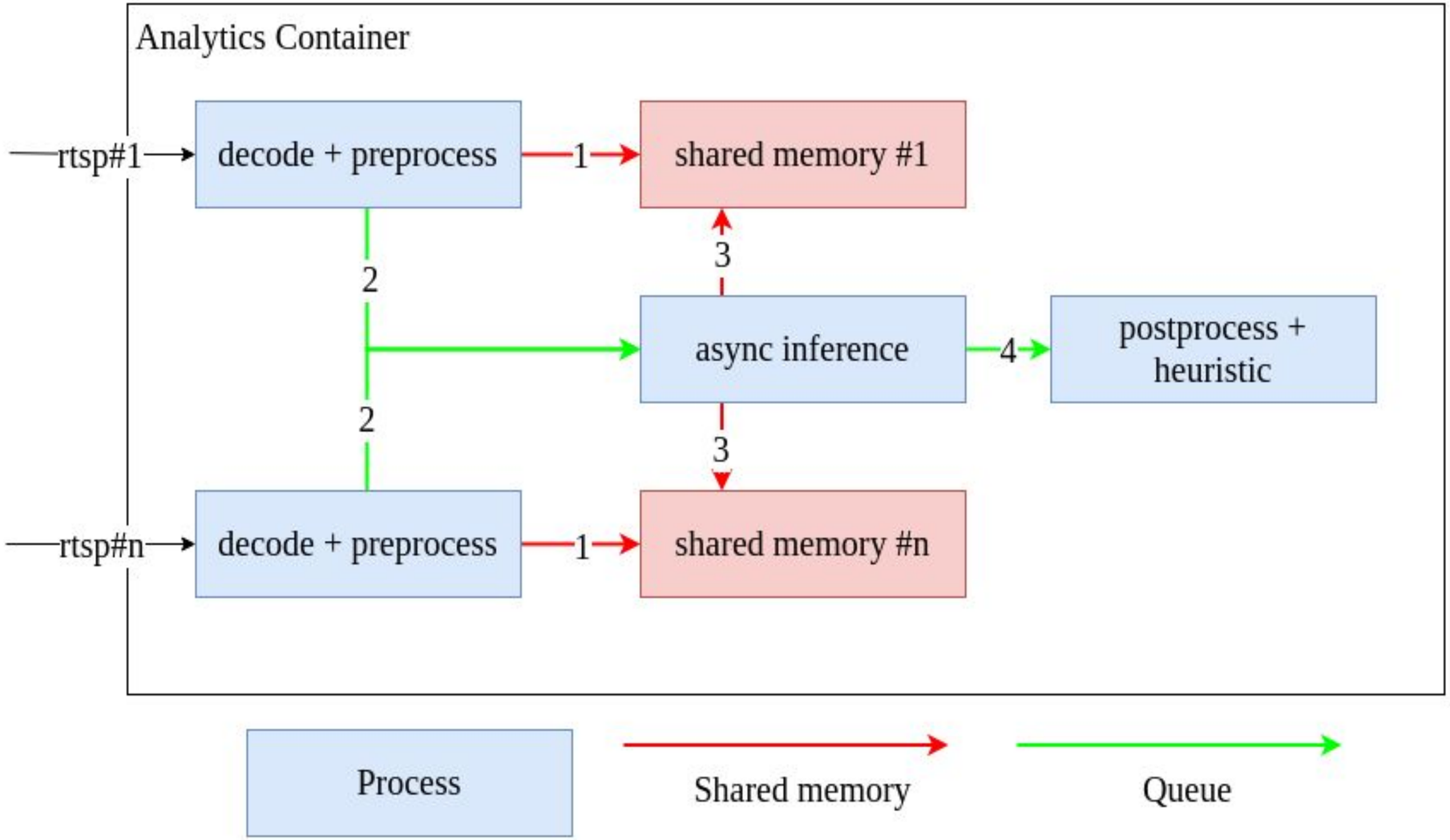
```
flow:
- flow_step:
    task_handlers:
    - preprocessing
    - shared_memory_writer  # service handler
    - batcher  # service handler
- flow_step:
    task_handlers:
    - shared_memory_reader  # service handler
    - inference
- flow_step:
    task_handlers:
    - postprocessing
    - heuristic:
        counters_config: null  # overwritten
```

Интеграция

```
1  def decoder(analytics):
2    analytics.init_stream(stream_id, stream_settings)  3
3
4    while True:
5      frame = decode_frame()
6      analytics.process_frame(stream_id, frame)  4
7      events = analytics.get_events(stream_id)  5
8
9
10 def main():
11   analytics = analytics_factory(path_to_config)  1
12
13   for i in num_workers:
14     Process(target=decoder, args=(analytics, )).start()
15                                                          2
16   ...
17
18   analytics.stop()  6
```

Схема

16 каналов, LA < 60



Возможности

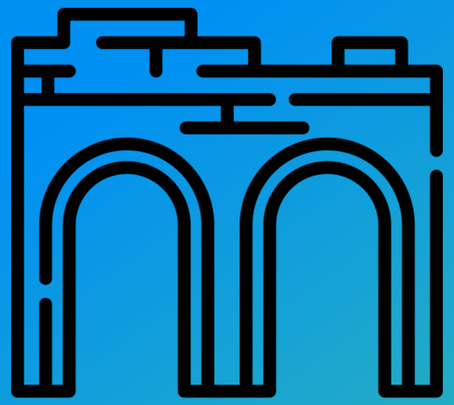
- ✅ flow аналитики через yaml-файл
- ✅ "ленивое" инстанцирование обработчиков
- ✅ не зависит от Inference Engine
- ✅ трансфер кадров через shared memory
- ✅ батчирование

Итоги

	Архитектура	Девайс	Кол-во каналов
Legacy	YOLO	RTX A2000 💸💸	17
Viaduct	SSD	Xeon Silver 4314 💸💸💸	16
Viaduct	YOLO	RTX A2000 💸💸	18

Выводы

- оптимизация:
  - архитектура модели
  - выполнение сети
  - организация flow
- CPU ≤ GPU
- “синтетика” только приблизительный ориентир
  - 1900 FPS в синтетике vs 384 FPS в реальном кейсе
- GIL 😞



avito-tech / aqueduct

ivideon / viaduct