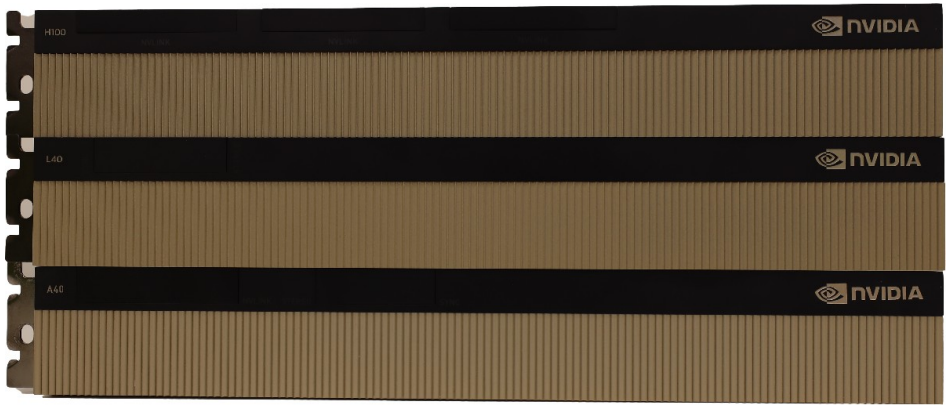
Video Analytics with NVIDIA stack

Milan Pultar

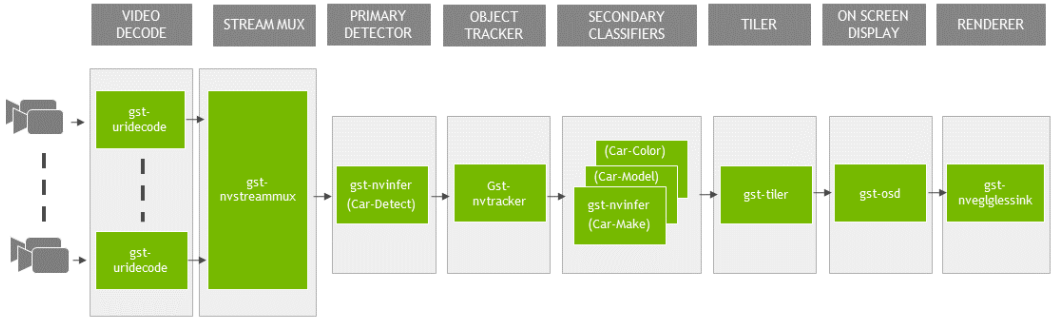
NVIDIA DeepStream SDK

- Based on GStreamer
- Configurable pipeline
- Built-in multi-stream object tracker
- GPU optimization (TensortRT, float16, int8)
- Support of edge devices (Jetson)



NVIDIA Datacenter GPUs

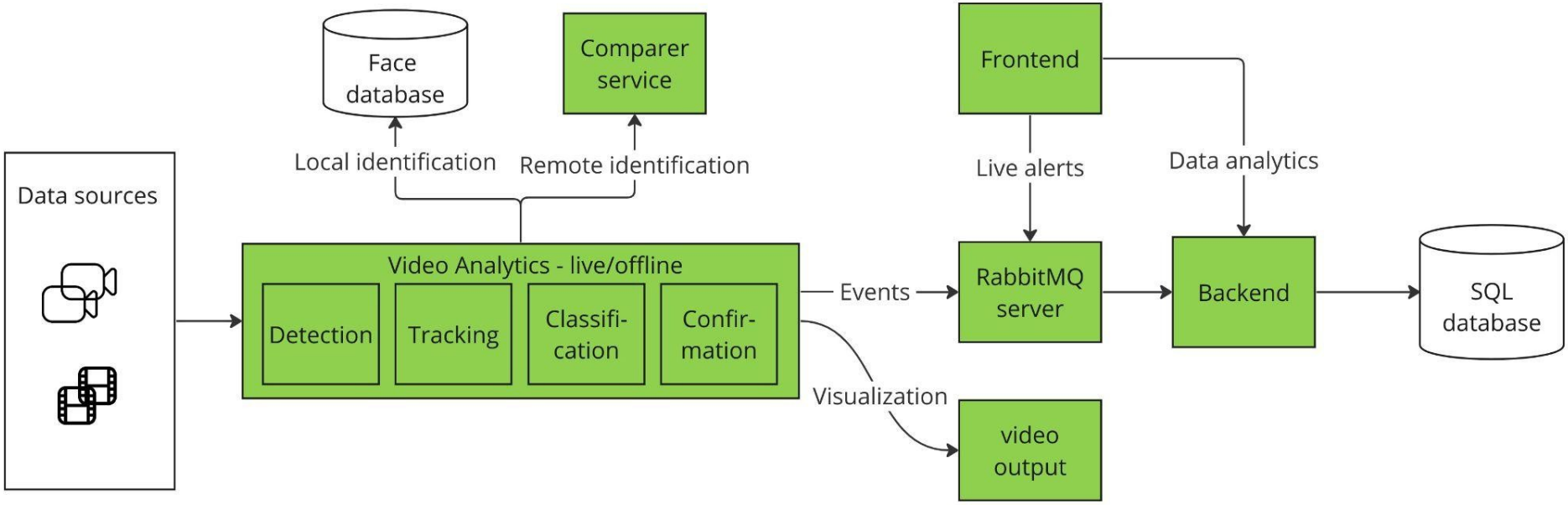
NVIDIA Jetson Orin Nano



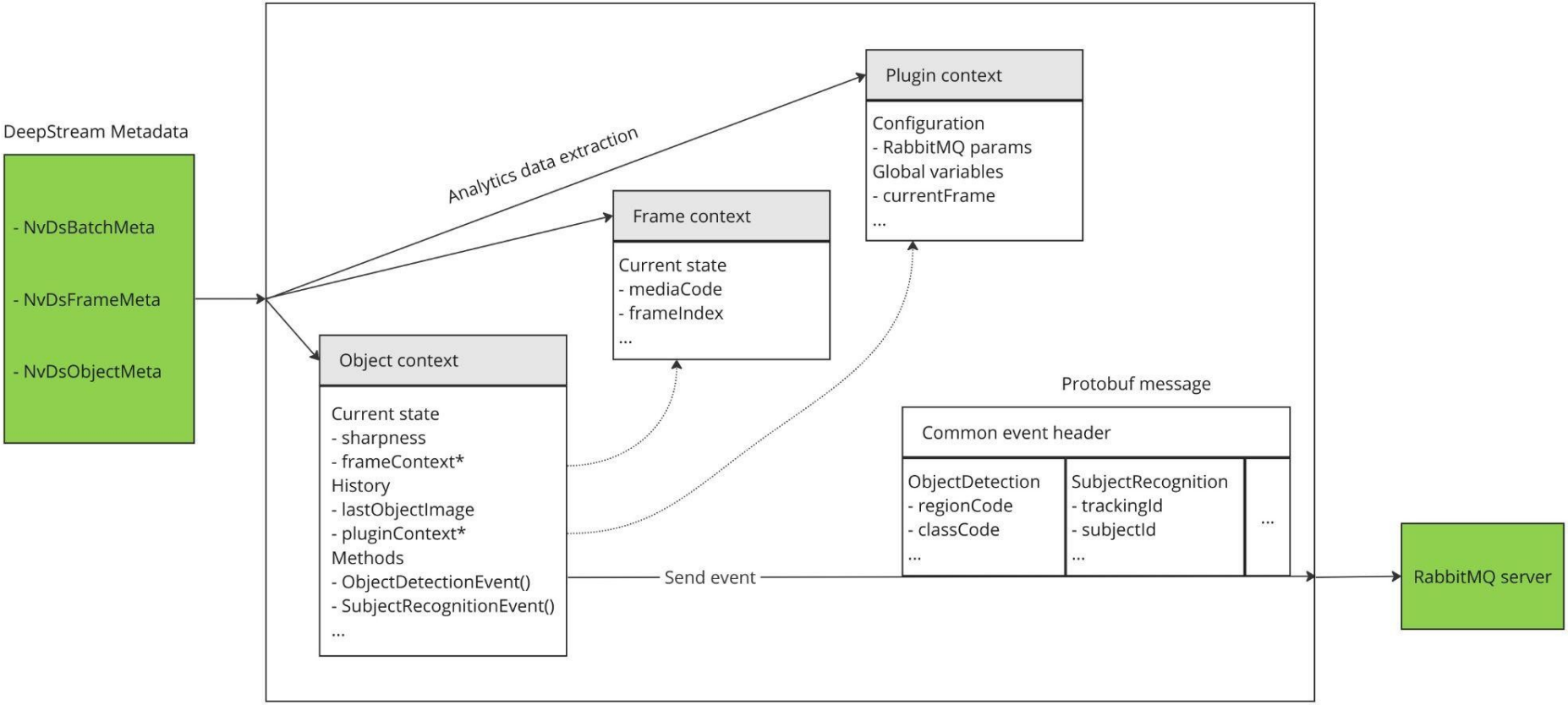
NVIDIA DeepStream SDK plugins

- TensorRT inference
- Video overlay drawing
- Hardware accelerated video encoder/decoder
- Message broker (MQTT, gRPC)
- Dewarper
- Speech synthesis
- GPU video decoder
- …

Quantasoft IdentLock IVA software architecture



Quantasoft IdentLock IVA Events plugin



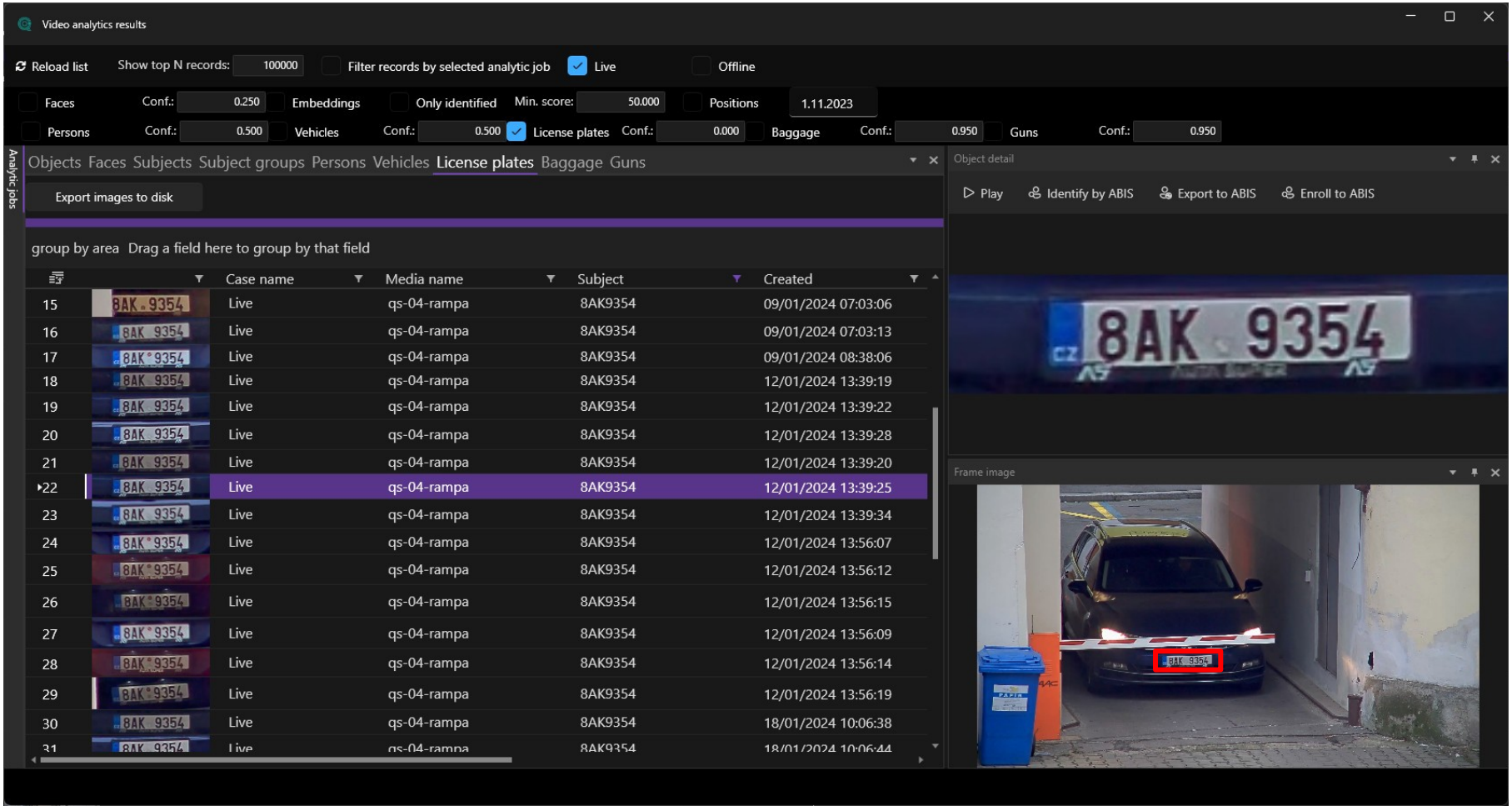
Compiling Protobuf definitions any language

- https://github.com/quantasoftcz/protoc-polyglot (Milan Pultar, Hugo Kunák)
- Supported languages: Python, JavaScript, C++, Rust, Java, Go, Objective-C, PHP, Ruby, C#

```
alias DOCKER_RUN='docker run --rm -v $(pwd)/output:/data/output -v $(pwd)/protos:/data/protos protocpolyglot/protoc-polyglot'
```

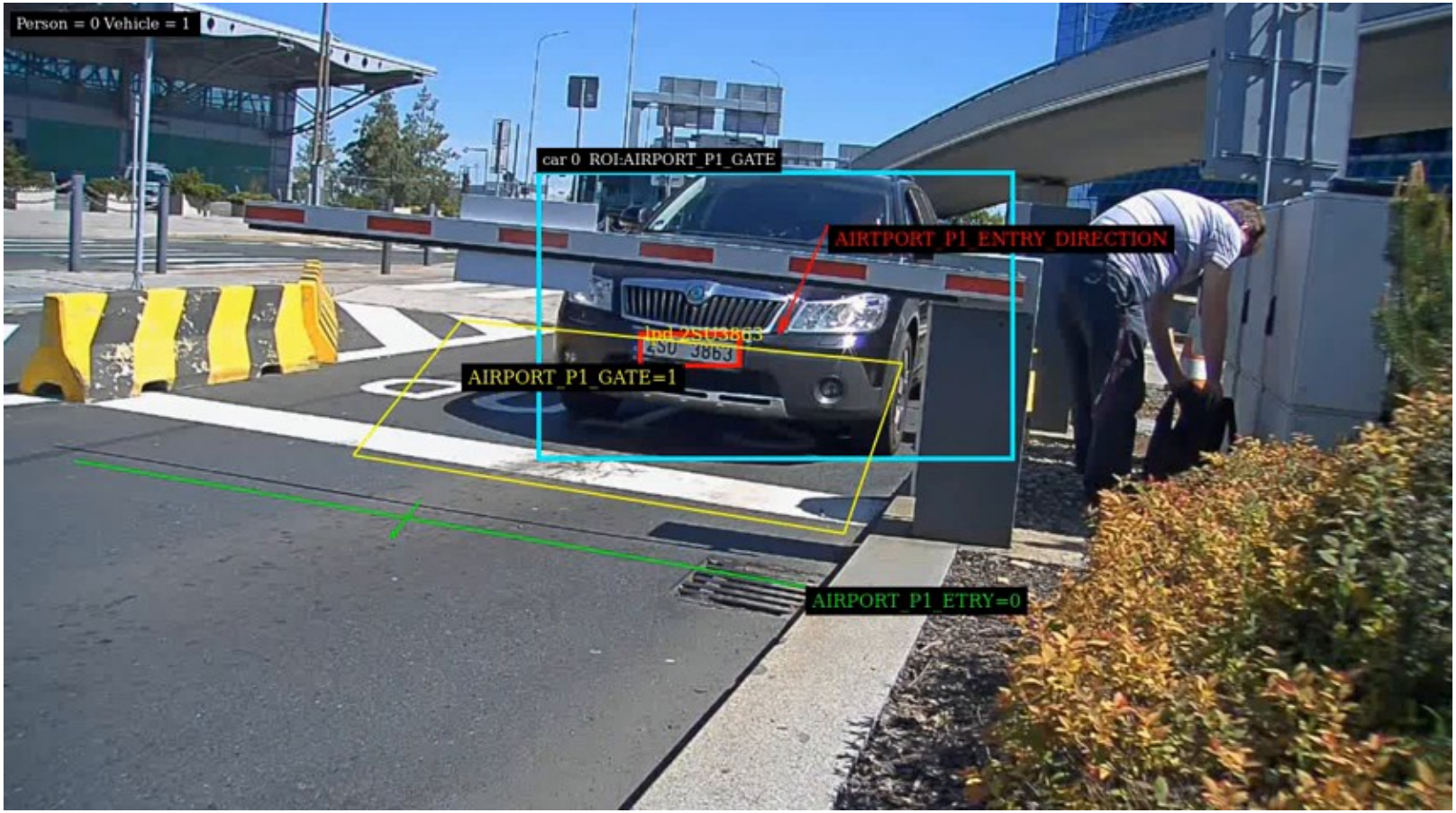
- Compilation
  - DOCKER_RUN python/cli.py protoc
  - DOCKER_RUN cpp/cli.py protoc
  - DOCKER_RUN js/cli.py protoc
- Generate documentation
  - DOCKER_RUN cli.py doc

License plate recognition



https://drive.google.com/drive/folders/11rxgDm9qdCGEBlsttO_-YdaMP_bGxLwA?usp=sharing

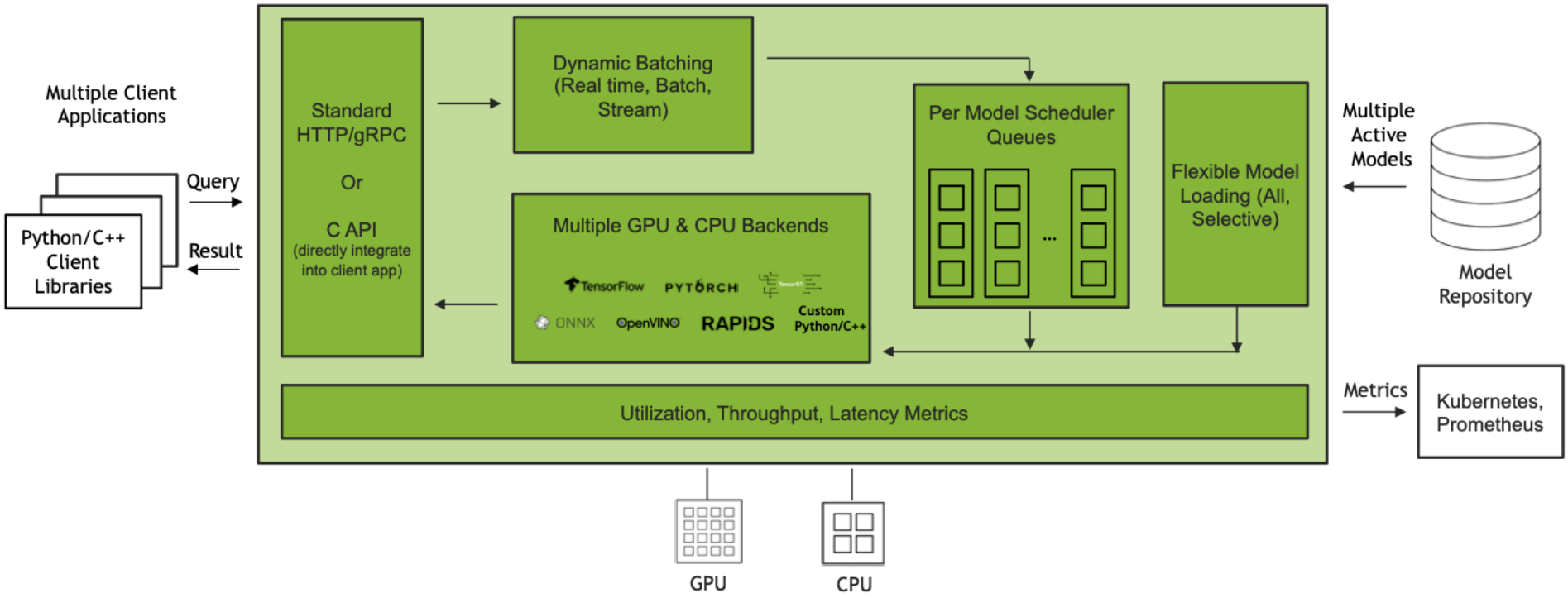

2020-12-21 09:35:09
face 15947510880569131017


Person = 0 Vehicle = 1
car 0 ROI:AIRPORT_P1_GATE
AIRTPORT_P1_ENTRY_DIRECTION
AIRPORT_P1_GATE=1
AIRPORT_P1_ETRY=0

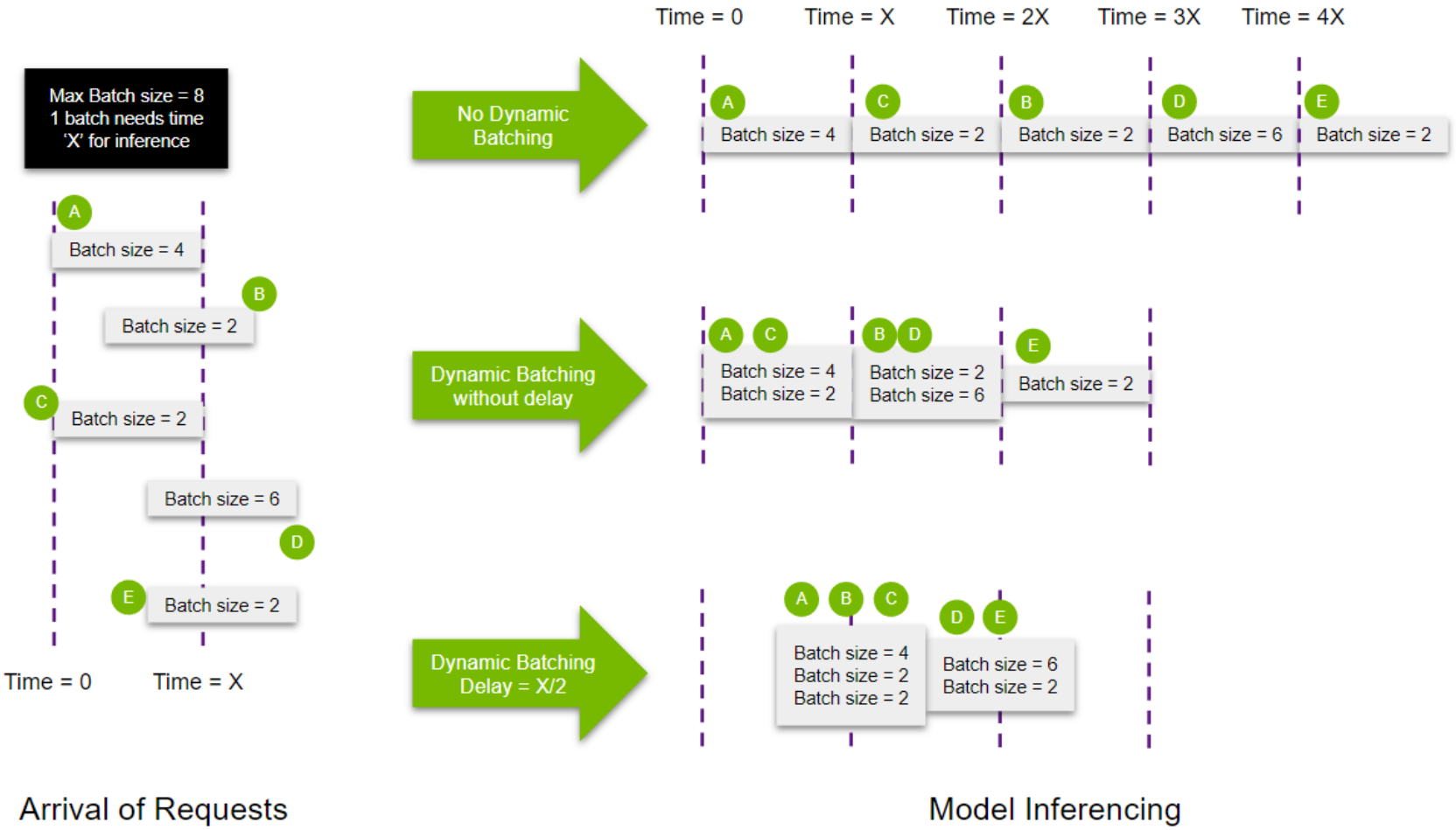
- Simulated assault scenario:
- https://drive.google.com/file/d/1f-kYFMwU1Shgp2yJ7DvmqWoFZ2KwRi3F/view?usp=sharing

NVIDIA TRITON INFERENCE SERVER ARCHITECTURE

Open-Source Software For Scalable, Simplified Inference Serving

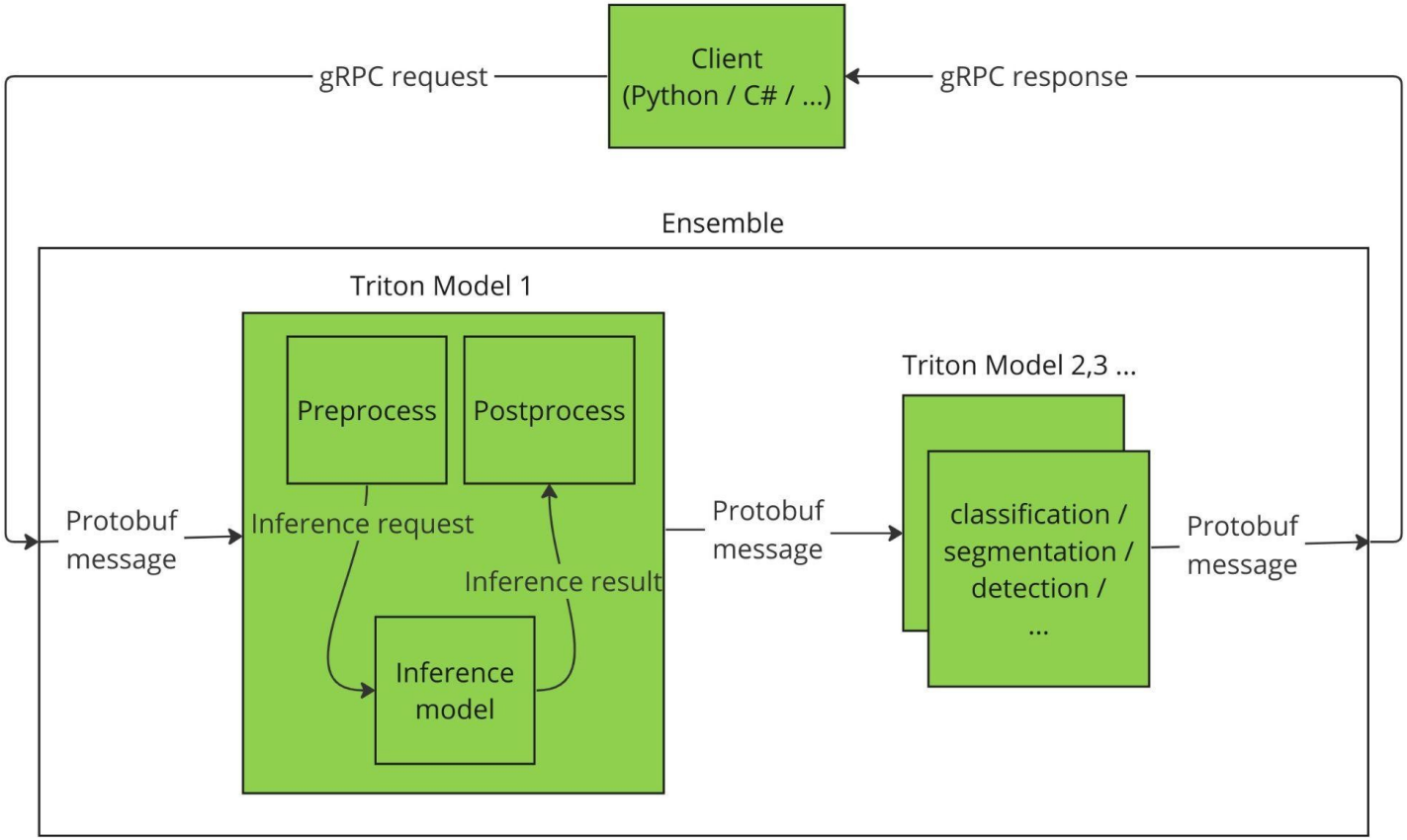


Source: https://aws.amazon.com/blogs/machine-learning/deploy-fast-and-scalable-ai-with-nvidia-triton-inference-server-in-amazon-sagemaker/



Source: https://github.com/triton-inference-server/tutorials/tree/main/Conceptual_Guide/Part_2-improving_resource_utilization#concurrent-model-execution

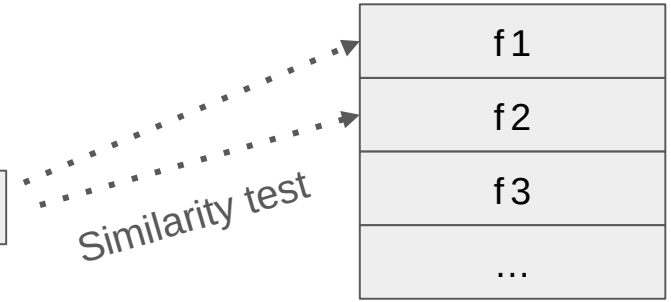
Quantasoft IdentLock IIA architecture



Person identification

- 5 face keypoints
- Face embedding

$f = [f_1, f_2, f_3, \dots f_n]$



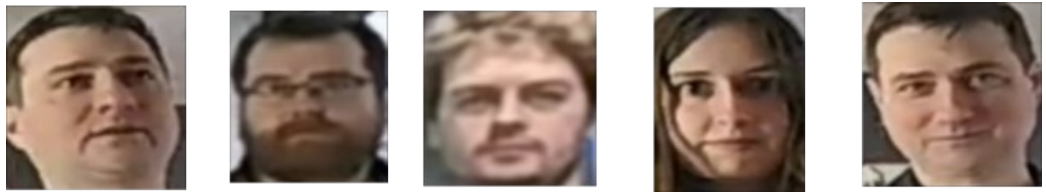
Subject database

- Person embedding

$p = [p_1, p_2, p_3, \dots p_m]$

Face attributes

- Dedicated models:
  - Gender: male/female
  - Age category: (0, 3), (4, 12), (13, 19), …
  - Liveness
  - Quality
- Shared model for 40 classes:
  - Big lips
  - Blond hair
  - Straight hair
  - Eyeglasses
  - Wearing lipstick
  - …

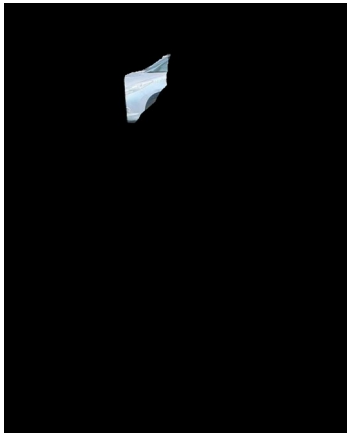


Live

Spoofed

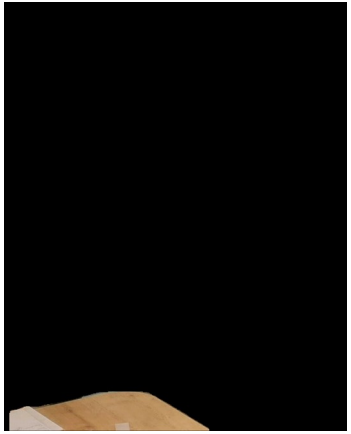
Object segmentation

Car

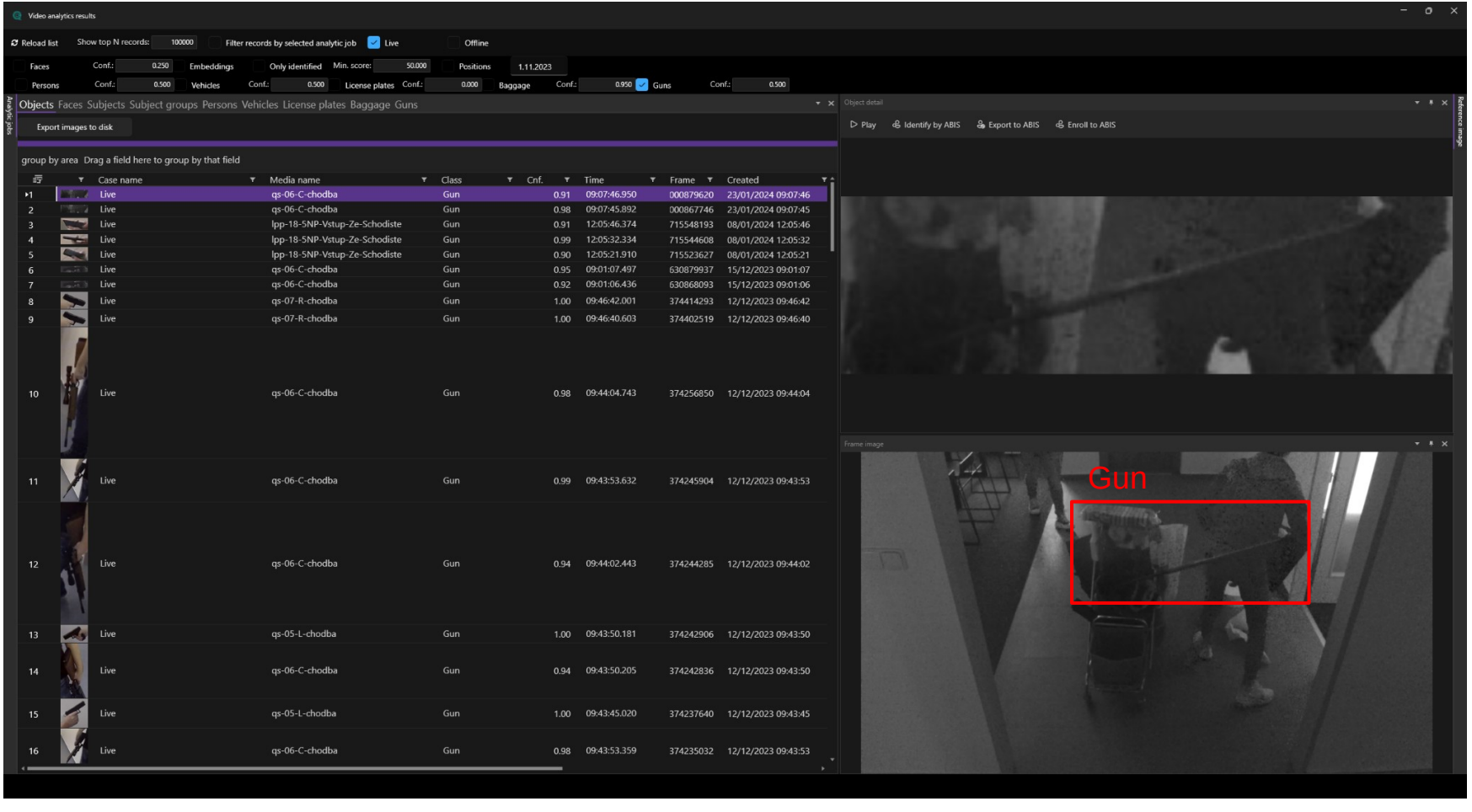


Person

Table



But remember, no model is perfect



Performance test - equipment provided by M Computers

32-core CPU, 192 GB RAM, NVME storage, 10 GbE networking, NVIDIA H100, L40, A40

Performance test - vector search

- Search time for vector size of 128 and 512

	size = 1M vectors	10M	40M	80M
NVIDIA H100 80 GB PCIe	2ms, 4ms	12ms, 32ms	46ms, 128ms	93ms, -
L40 48 GB	2ms, 8ms	21ms, 86ms	87ms, -	-, -
3090Ti 24GB	2ms, 14ms	23ms, 139ms	93ms, -	-, -
1080Ti 12GB	11ms, 39ms	-, -	-, -	-, -

Performance test - AI-Benchmark - Total results

- AI-Benchmark - https://ai-benchmark.com/tests.html

	Total AI score	Inference score	Training score
NVIDIA H100 80 GB PCIe	90 614	43 211	47 403
L40 48 GB	73 423	33 498	39 925
A40 48 GB	48 771	23 394	25 377

Performance test - AI-Benchmark - Individual results

- Individual tests - inference time per batch of inputs

	U-Net (bs = 4)	ResNet-V2-50 (bs = 10)	LSTM-Sentiment (bs = 100)
NVIDIA H100 80 GB PCIe	42.9 ± 0.6 ms	11.2 ± 0.5 ms	336 ± 17 ms
L40 48 GB	79.1 ± 0.5 ms	11.8 ± 0.9 ms	314 ± 8 ms
A40 48 GB	111.6 ± 0.5 ms	20.1 ± 0.5 ms	317 ± 10 ms

Performance test - training time - ArcFace

	Dataset: MS1MV3 Model: ResNet-50	Dataset: Glint360K Model: ResNet-100
NVIDIA H100 80 GB PCIe	15.50 h	84.34 h (3 d 12.34 h)
L40 48 GB	27.00 h (1 d 3.00 h)	192.30 h (6 d 18.30 h)
A40 48 GB	33.42 h (1 d 9.42 h)	-

Takeaways

- DeepStream has a wide range of tools for video analytics
- It supports small edge devices as well as large GPUs
- New features - like more sophisticated event processing - can be extended via custom plugins

- Triton Inference Server is a useful tool for image analytics
- Models can be joined into ensembles to include pre/post processing
- Nested model scheme improves code readability and versatility of ensembles
- Complex data can be sent in Protobuf to enable modularity of ensembles

Thank you