

# Video Analytics with NVIDIA stack

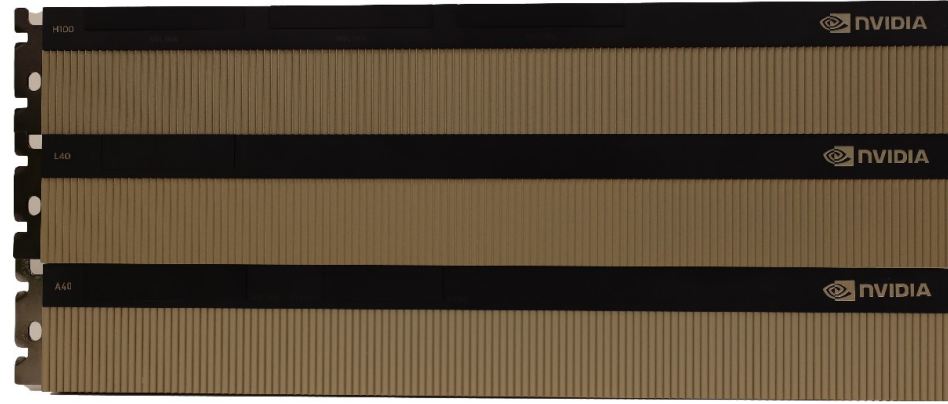
Milan Pultar



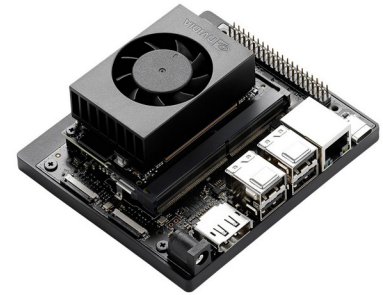
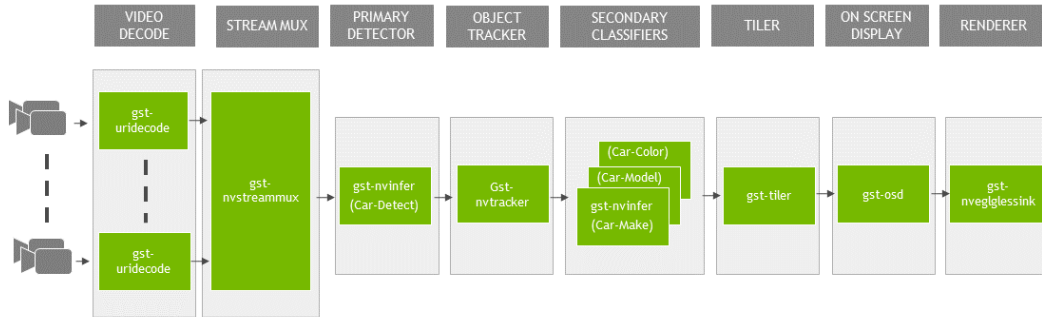
QUANTASOFT

# NVIDIA DeepStream SDK

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



*NVIDIA Datacenter GPUs*



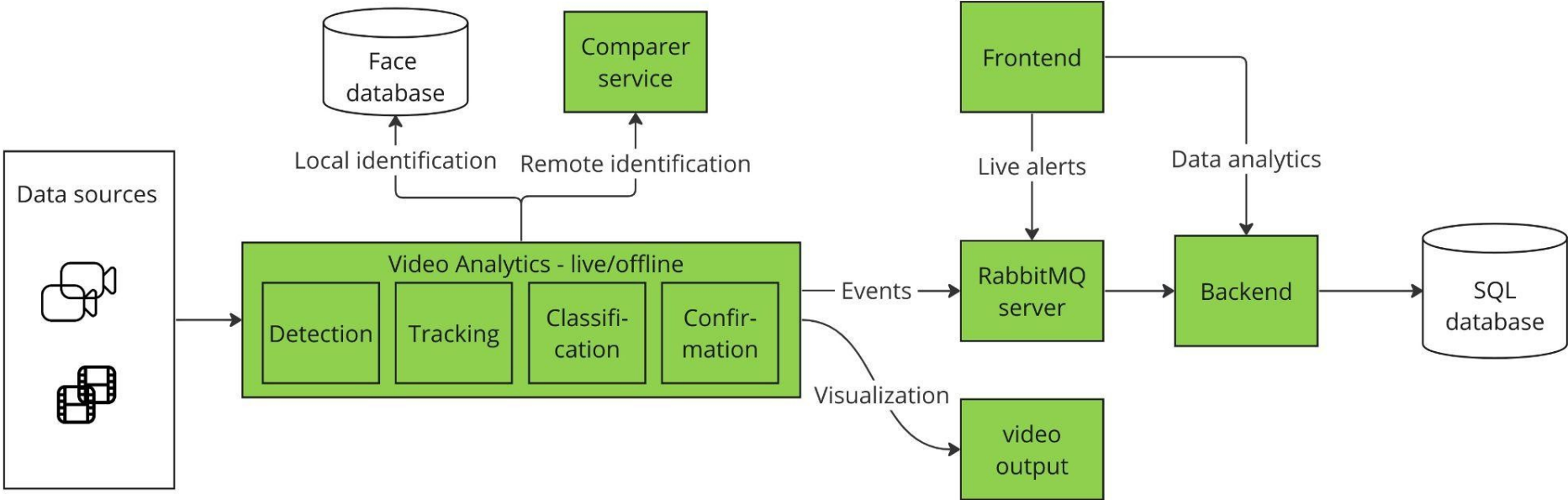
*NVIDIA Jetson Orin Nano*



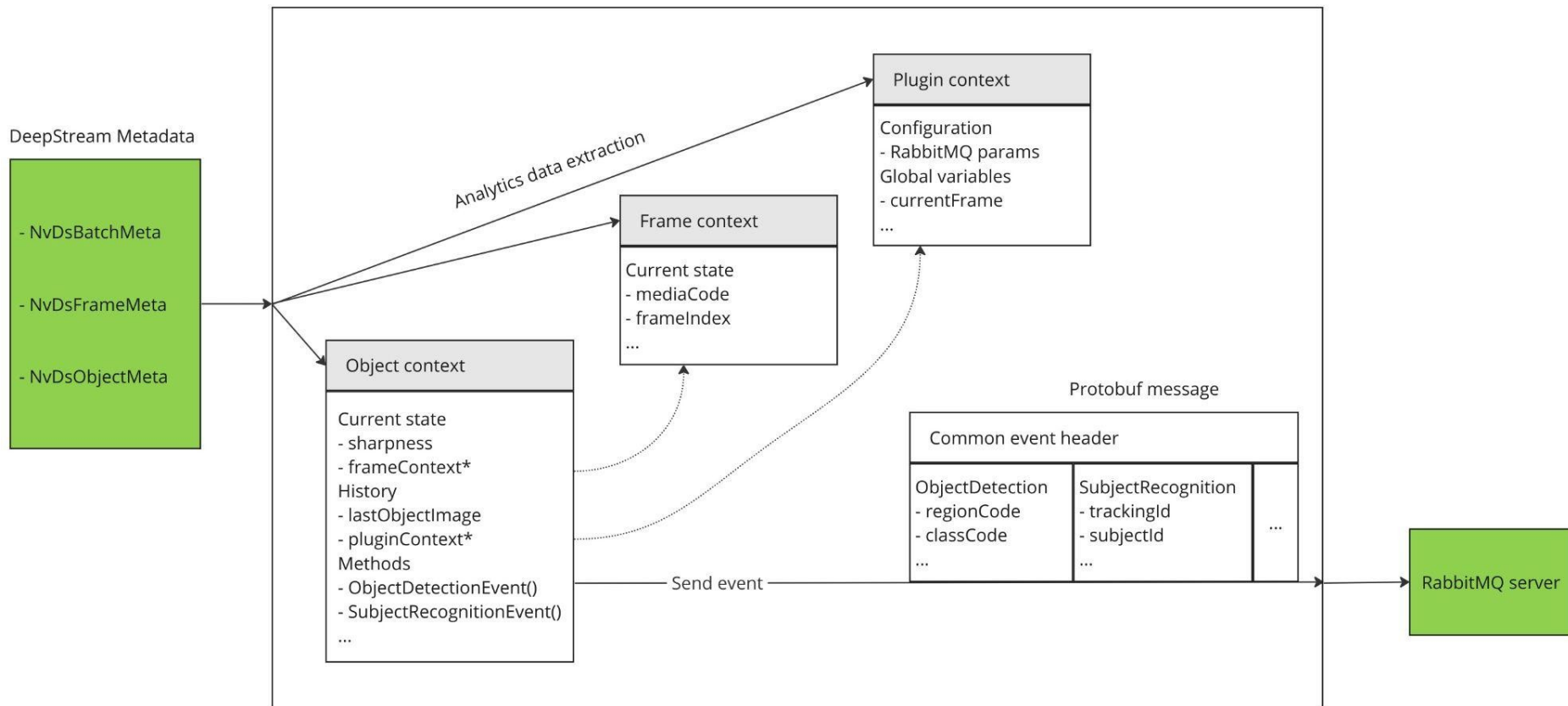
# NVIDIA DeepStream SDK plugins

- 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

Video analytics results

Reload list Show top N records: 100000 Filter records by selected analytic job  Live  Offline

Faces Conf.: 0.250 Embeddings  Only identified Min. score: 50.000  Positions 1.11.2023

Persons Conf.: 0.500 Vehicles Conf.: 0.500  License plates Conf.: 0.000 Baggage Conf.: 0.950 Guns Conf.: 0.950

Objects Faces Subjects Subject groups Persons Vehicles License plates Baggage Guns


Export images to disk

group by area Drag a field here to group by that field


	Case name	Media name	Subject	Created
15	8AK 9354	qs-04-rampa	8AK9354	09/01/2024 07:03:06
16	8AK 9354	qs-04-rampa	8AK9354	09/01/2024 07:03:13
17	8AK 9354	qs-04-rampa	8AK9354	09/01/2024 08:38:06
18	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:39:19
19	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:39:22
20	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:39:28
21	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:39:20
22	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:39:25
23	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:39:34
24	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:56:07
25	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:56:12
26	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:56:15
27	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:56:09
28	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:56:14
29	8AK 9354	qs-04-rampa	8AK9354	12/01/2024 13:56:19
30	8AK 9354	qs-04-rampa	8AK9354	18/01/2024 10:06:38
31	8AK 9354	qs-04-rampa	8AK9354	18/01/2024 10:06:44

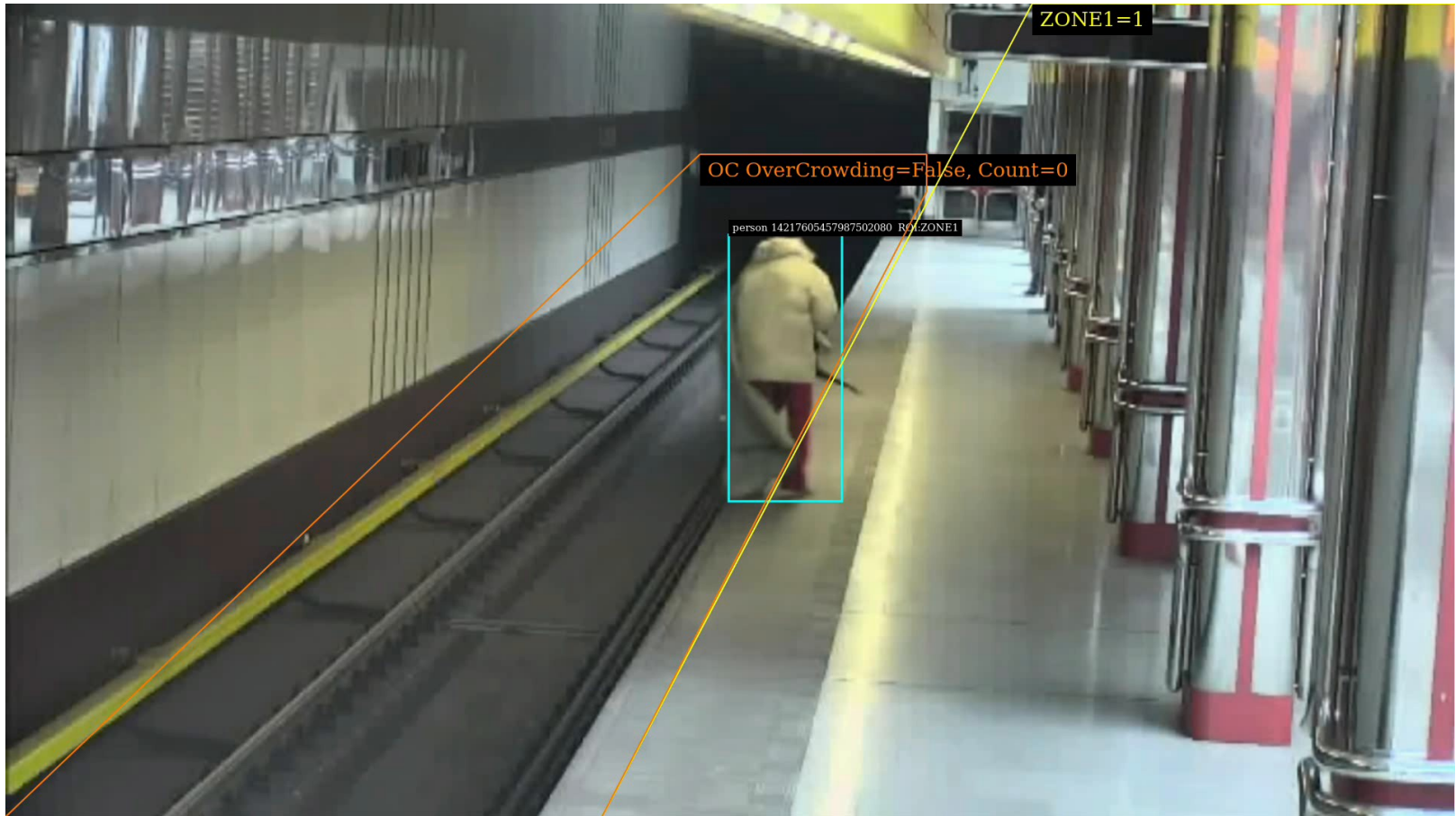
Object detail

Play Identify by ABIS Export to ABIS Enroll to ABIS



Frame image





[https://drive.google.com/drive/folders/11rxgDm9qdCGEB1sttO\\_-YdaMP\\_bGxLwA?usp=sharing](https://drive.google.com/drive/folders/11rxgDm9qdCGEB1sttO_-YdaMP_bGxLwA?usp=sharing)

2020-12-21 09:35:09





Person = 0 Vehicle = 1

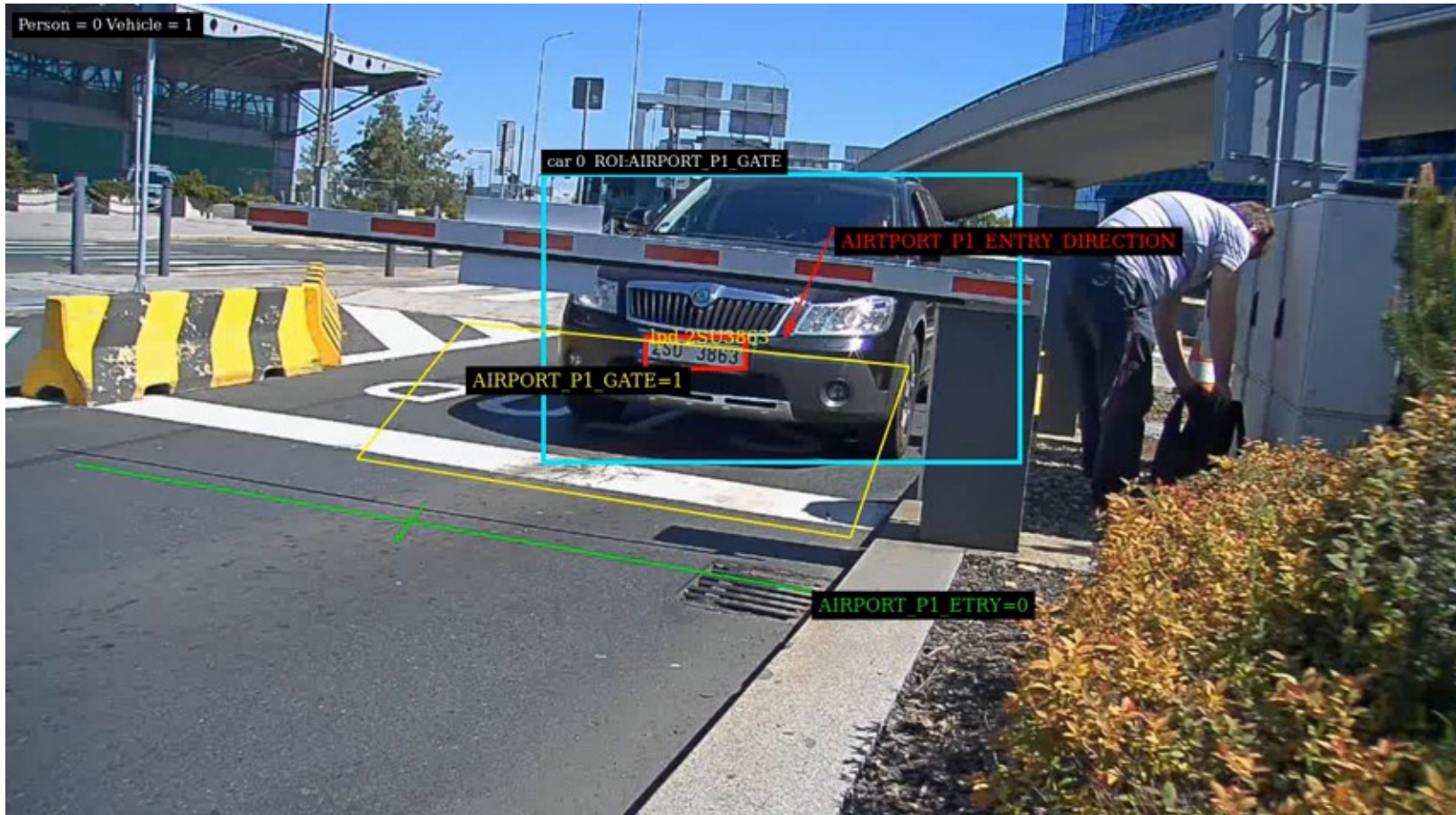
car 0 ROI: AIRPORT\_P1\_GATE

AIRPORT\_P1\_ENTRY\_DIRECTION

Ind: 25195363  
SU 3863

AIRPORT\_P1\_GATE=1

AIRPORT\_P1\_ENTRY=0

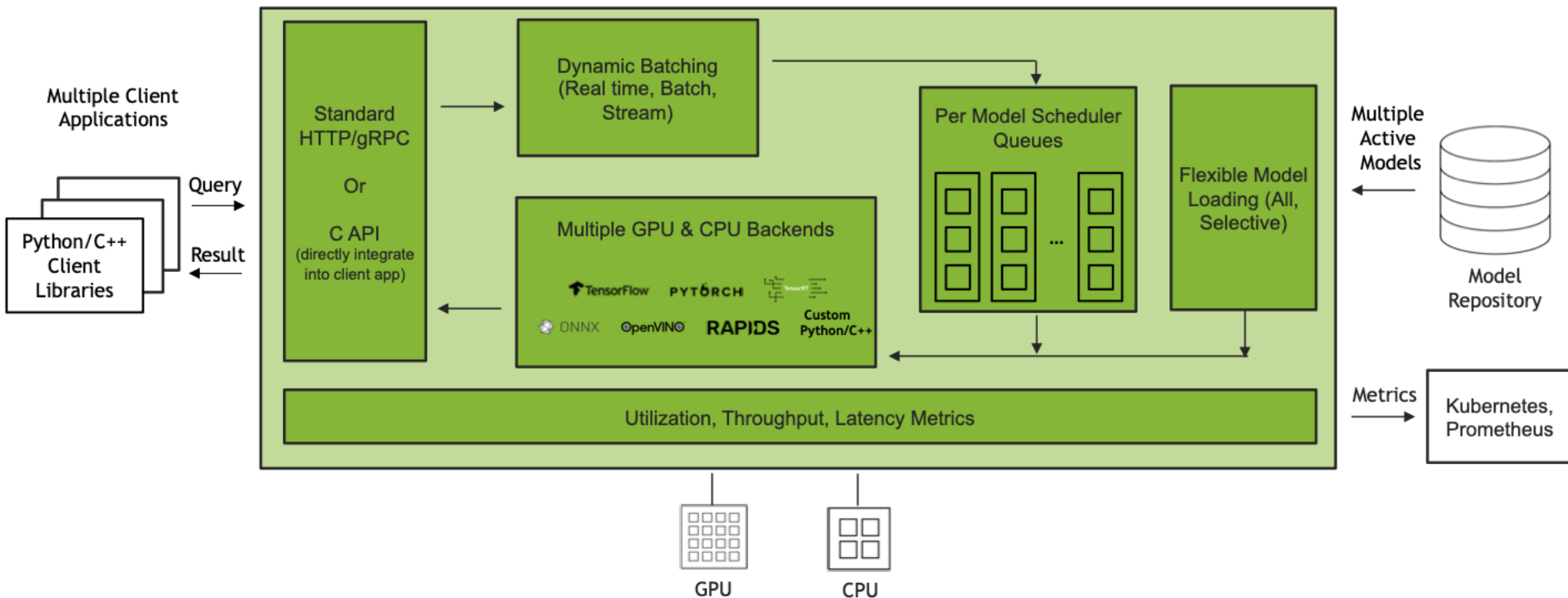


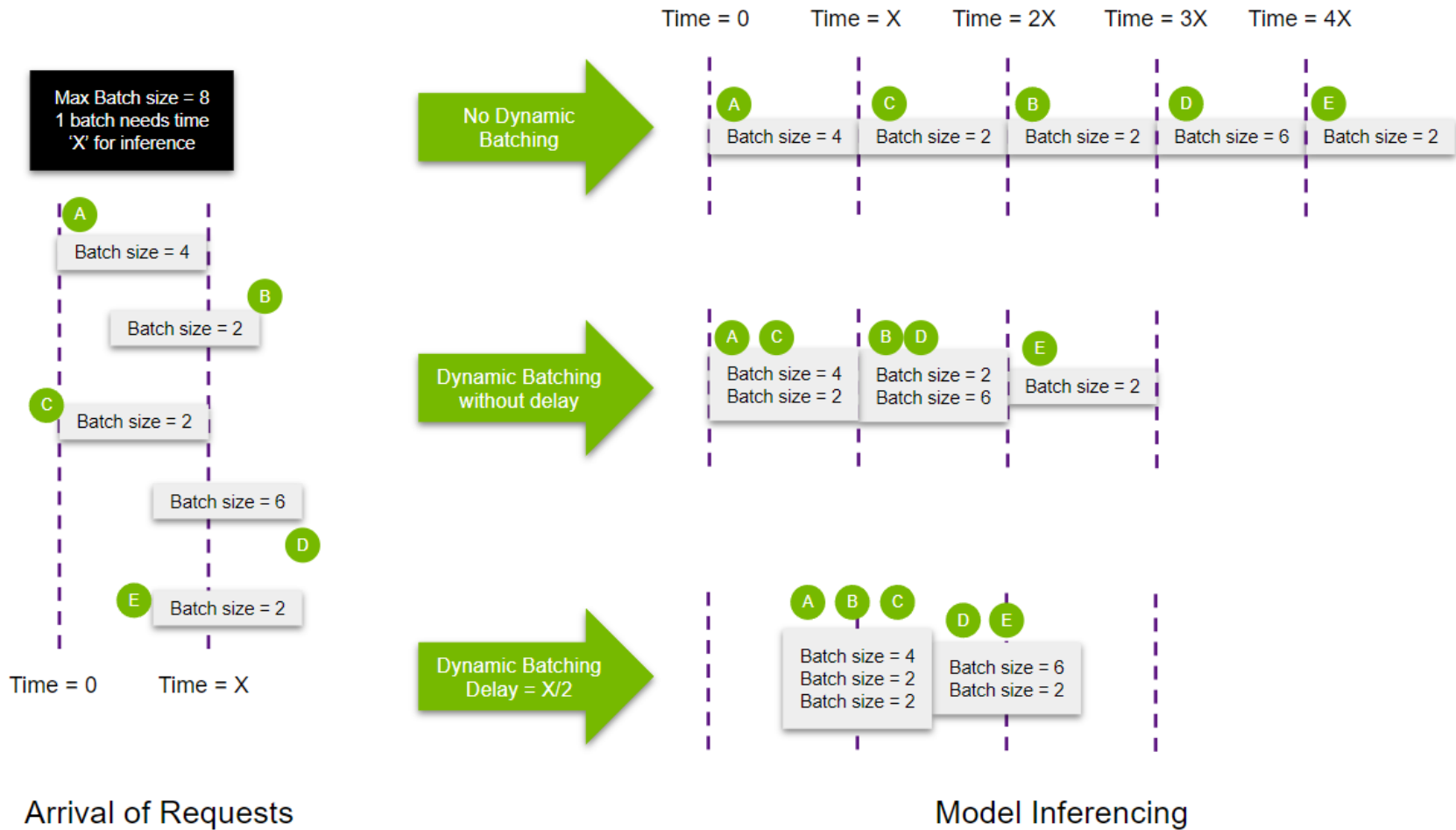


- Simulated assault scenario:
- <https://drive.google.com/file/d/1f-kYFMwU1Shgp2yJ7DvmqWoFZ2KwRi3F/view?usp=sharing>

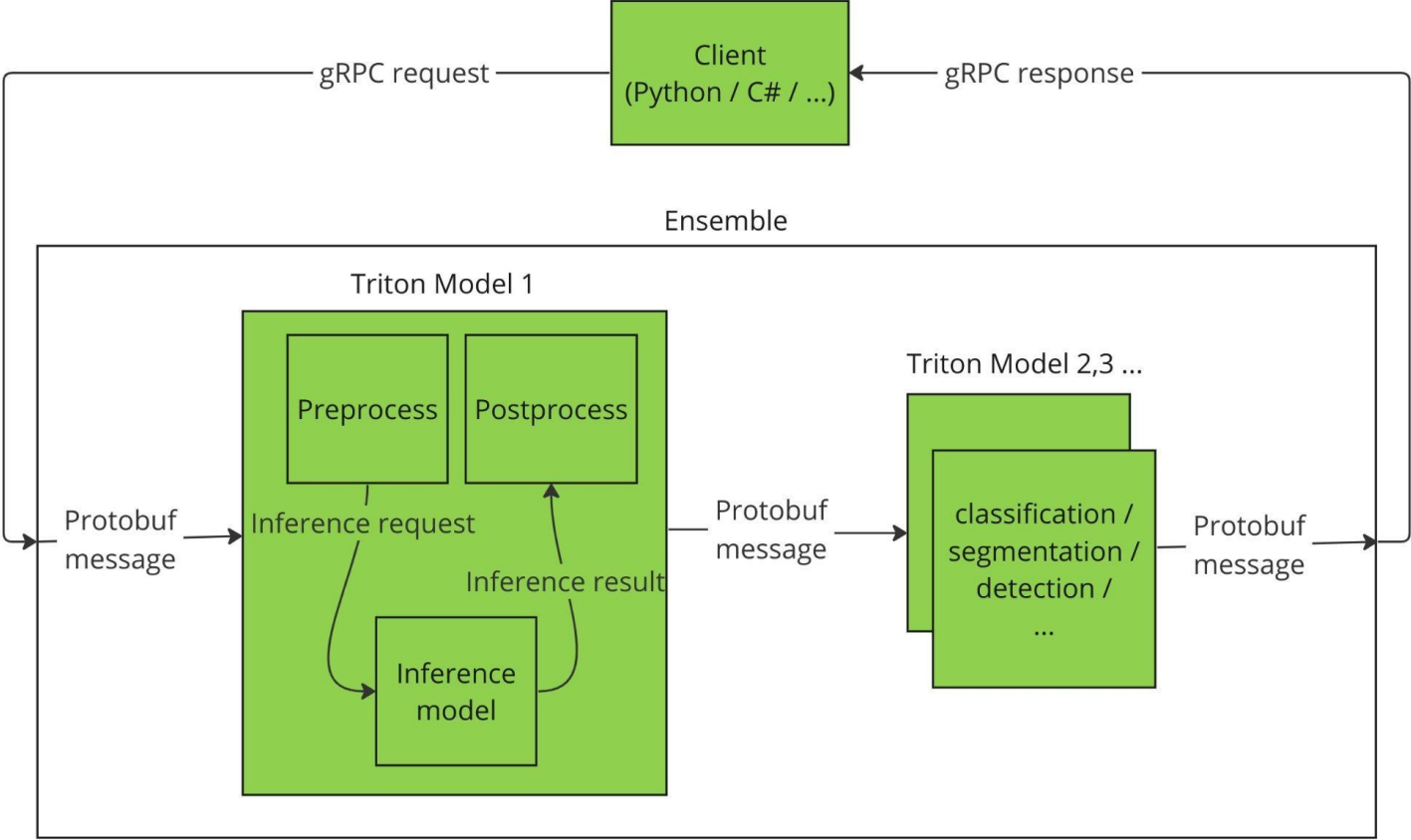
# NVIDIA TRITON INFERENCE SERVER ARCHITECTURE

Open-Source Software For Scalable, Simplified Inference Serving





# Quantasoft IdentLock IIA architecture



# Person identification

- 5 face keypoints
- Face embedding

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

Similarity test

f1
f2
f3
...

*Subject database*

- Person embedding

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



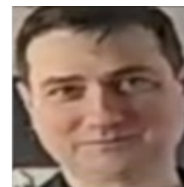
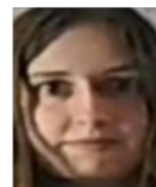
# Face attributes

- Dedicated models:

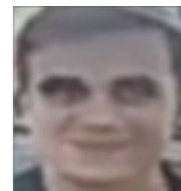
- Gender: male/female
- Age category: (0, 3), (4, 12), (13, 19), ...
- Liveness
- Quality

- Shared model for 40 classes:

- Big lips
- Blond hair
- Straight hair
- Eyeglasses
- Wearing lipstick
- ...



Live

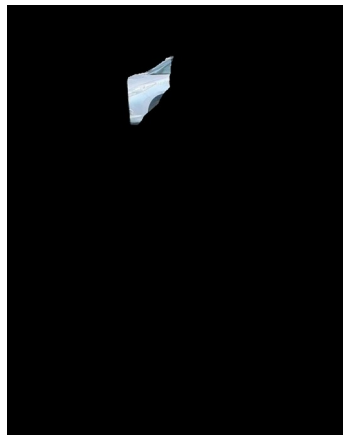


Spoofed

# Object segmentation



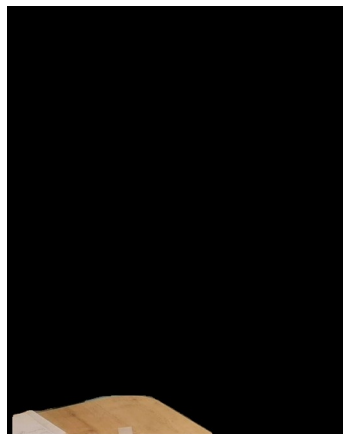
Car



Person



Table



# But remember, no model is perfect

Video analytics results

Reload list Show top N records: 100000 Filter records by selected analytic job  Live  Offline

Faces Conf.: 0.250 Embeddings Only identified Min. score: 50.000 Positions 1.11.2023

Persons Conf.: 0.500 Vehicles Conf.: 0.500 License plates Conf.: 0.000 Baggage Conf.: 0.950  Guns Conf.: 0.500

Objects Faces Subjects Subject groups Persons Vehicles License plates Baggage Guns

Export images to disk

group by area Drag a field here to group by that field

#	Case name	Media name	Class	Cnf.	Time	Frame	Created
#1	Live	qs-06-C-chodba	Gun	0.91	09:07:46.950	000879620	23/01/2024 09:07:46
2	Live	qs-06-C-chodba	Gun	0.98	09:07:45.892	000867746	23/01/2024 09:07:45
3	Live	lpp-18-SNP-Vstup-Ze-Schodiste	Gun	0.91	12:05:46.374	715548193	08/01/2024 12:05:46
4	Live	lpp-18-SNP-Vstup-Ze-Schodiste	Gun	0.99	12:05:32.334	715544608	08/01/2024 12:05:32
5	Live	lpp-18-SNP-Vstup-Ze-Schodiste	Gun	0.90	12:05:21.910	715523627	08/01/2024 12:05:21
6	Live	qs-06-C-chodba	Gun	0.95	09:01:07.497	630879937	15/12/2023 09:01:07
7	Live	qs-06-C-chodba	Gun	0.92	09:01:06.436	630868093	15/12/2023 09:01:06
8	Live	qs-07-R-chodba	Gun	1.00	09:46:42.001	374414293	12/12/2023 09:46:42
9	Live	qs-07-R-chodba	Gun	1.00	09:46:40.603	374402519	12/12/2023 09:46:40
10	Live	qs-06-C-chodba	Gun	0.98	09:44:04.743	374256850	12/12/2023 09:44:04
11	Live	qs-06-C-chodba	Gun	0.99	09:43:53.632	374245904	12/12/2023 09:43:53
12	Live	qs-06-C-chodba	Gun	0.94	09:44:02.443	374244285	12/12/2023 09:44:02
13	Live	qs-05-L-chodba	Gun	1.00	09:43:50.181	374242906	12/12/2023 09:43:50
14	Live	qs-06-C-chodba	Gun	0.94	09:43:50.205	374242836	12/12/2023 09:43:50
15	Live	qs-05-L-chodba	Gun	1.00	09:43:45.020	374237640	12/12/2023 09:43:45
16	Live	qs-06-C-chodba	Gun	0.98	09:43:53.359	374235032	12/12/2023 09:43:53

Object detail

Play Identify by ABIS Export to ABIS Enroll to ABIS

Frame image



# 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](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