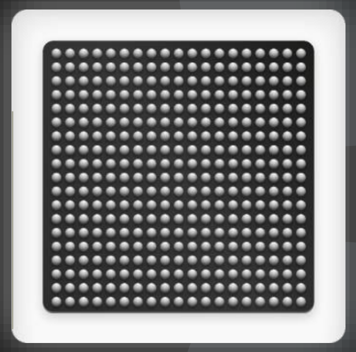
HAILO

Designing Powerful, Scalable & Cost-Efficient AI-Powered Video Management Systems

Webinar



Yaniv Iarovici & Gilad Nahor

9 May 2024

Agenda

- Part I – Introduction (10 min.):
  - → AI-powered VMS
  - → Hailo's advanced analytics solutions for VMS
- Part II – Integration (20 min.):
  - → Integrating Hailo-based AI analytics into VMS
  - → How to design a multi-stream pipeline
    - › Suggested steps, tools, tips and pitfalls
  - → Integration with VMS software, Network Optix example
  - → Next generation VMS capabilities
    - › Using CLIP model for free text searching on live video streams
- Part III – VMS demo (10 min.)
- Part IV – Q&A (15 min.)

Notes

- This webinar is being recorded, a link will be shared with all participants by email, and on Hailo's website
- The presentation will be shared with participants and will be available to download on Hailo's website
- Developer Zone access is required for accessing links to the documentation.
  To sign up click here

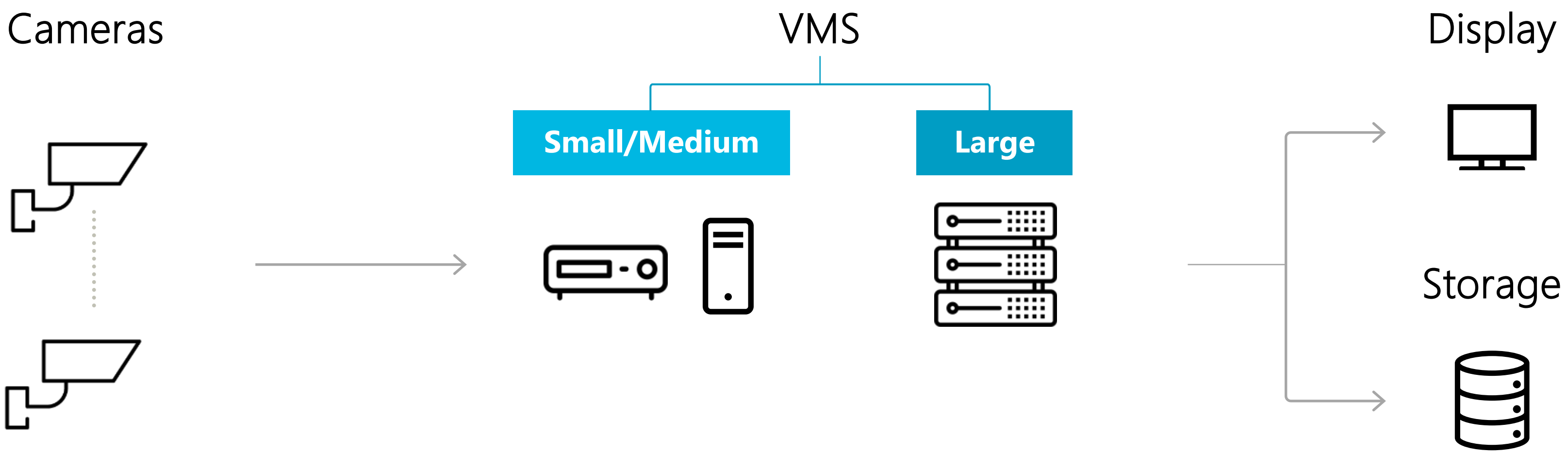
Introduction

AI-Powered VMS

HAILO

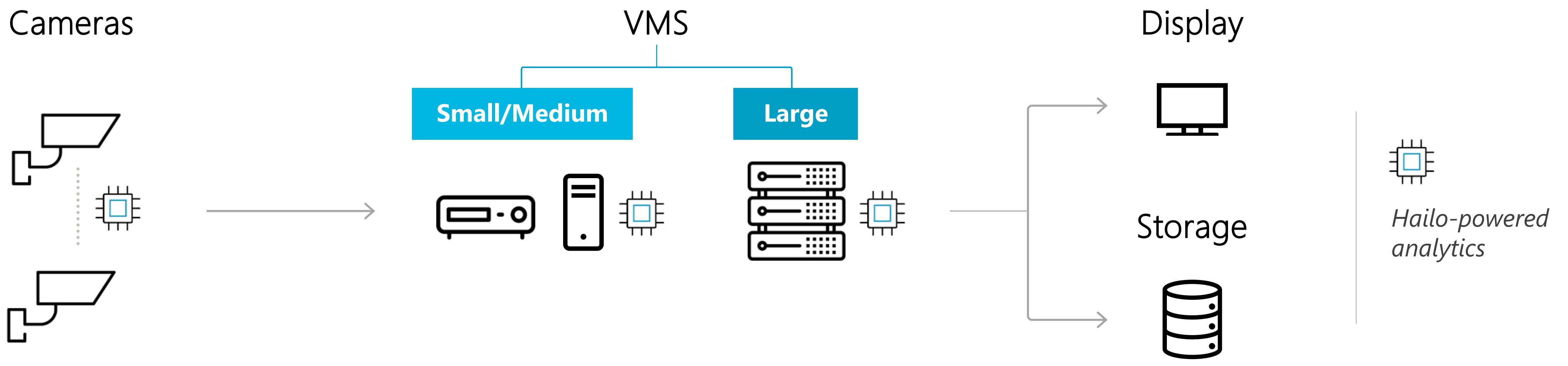
What is a VMS?

- Video Management Systems (VMS) handle multiple video channels at scale
- VMS handle streaming, storage, display, data indexing, monitoring and forensic data analysis, recording and fetching
- Monitoring challenge – using archaic tech and human operators



Advanced Analytics with AI / ML

- AI video analytics are being rapidly adopted by VMS
- Configurations are diverse, introducing analytics to the right components will maximize the benefits:
  - → Enhanced safety – spotting relevant ROIs & streams, enabling video history search, and many other apps
  - → Improved network utilization – streaming relevant events only
  - → Improved storage utilization – removing irrelevant content



Robust Ecosystem

OEM

Tailored Compute solutions

ODM

Customized HW solutions

aetina, Analog Technologies, BCM Tech, DELL, HPC Systems, Lanner, MiTAC, Unigen, sunix, TAURO Technologies

ISV

Analytic solutions across wide array of technologies

AllGoVision see. sense. secure, CVEDIA, GeoVision, GORILLA, innovatrics, Linker Vision, PARAVISION

VMS Vendor

Video management platform, incorporating storage, network & analytics

AVIGILON, Genetec™, Hanwha, milestone, NetworkOptix

System Builder / VAR

Aggregate technologies, offer solutions to system integrators

VELASEA

System Integrator

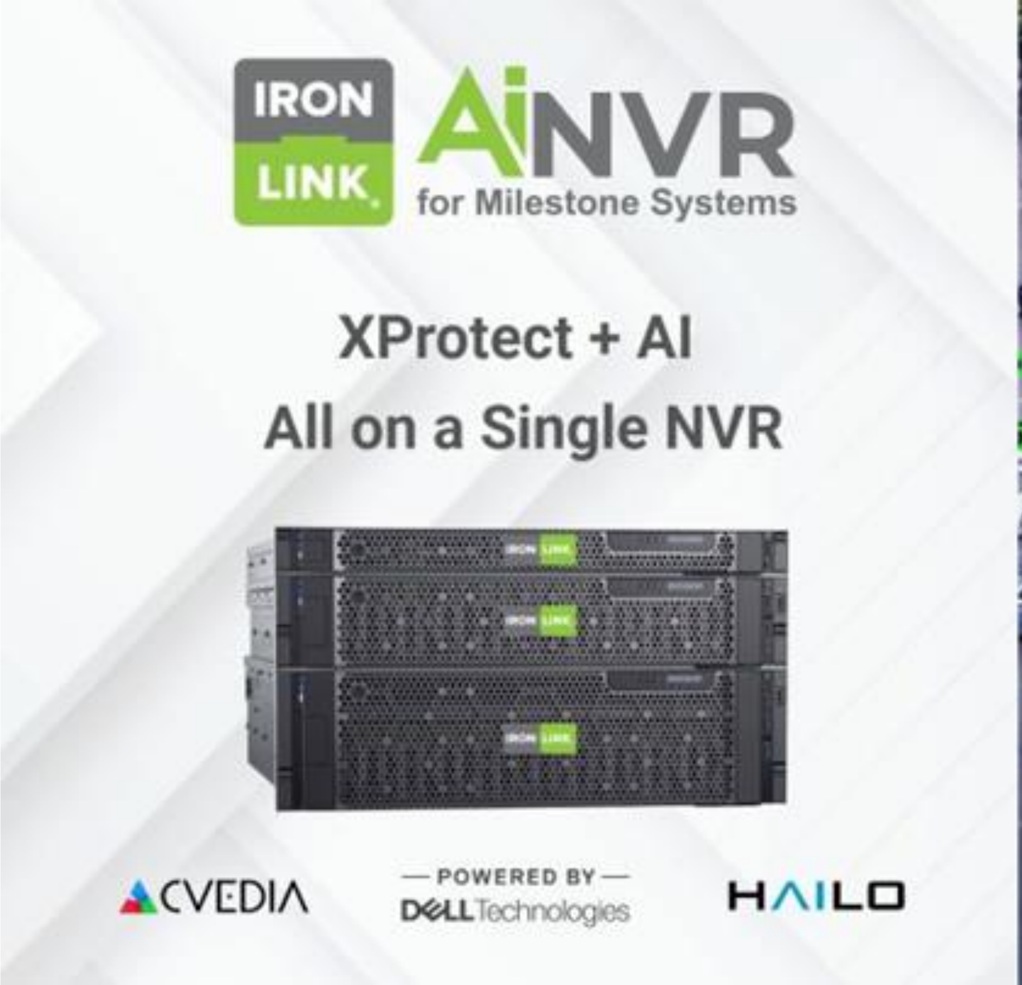
Direct support for end customers

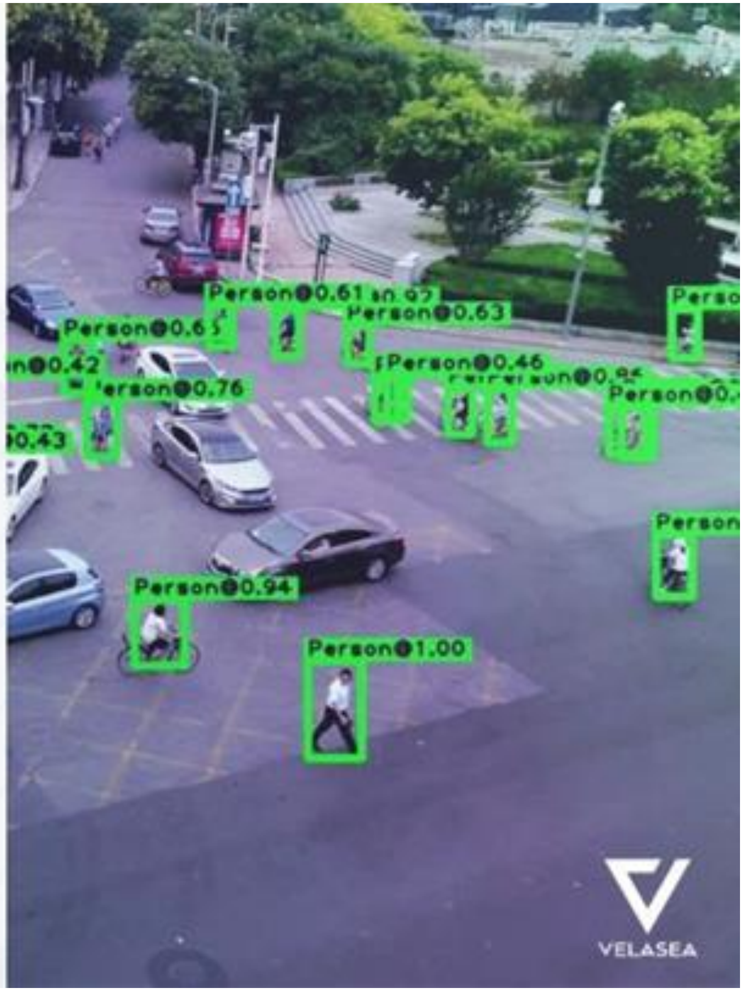
End Customer

Broad spectrum of applications:

Banking & finance; Healthcare;
Manufacturing; Retail;
Smart building; Smart city;
Transport/Logistics & utilities;
and many more…

Ecosystem – End-to-End System Example



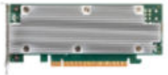
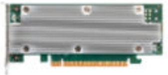
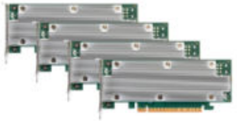


Hailo's Advanced Analytics Solutions for VMS

HAILO

Scalable Solutions up to 200 Channels

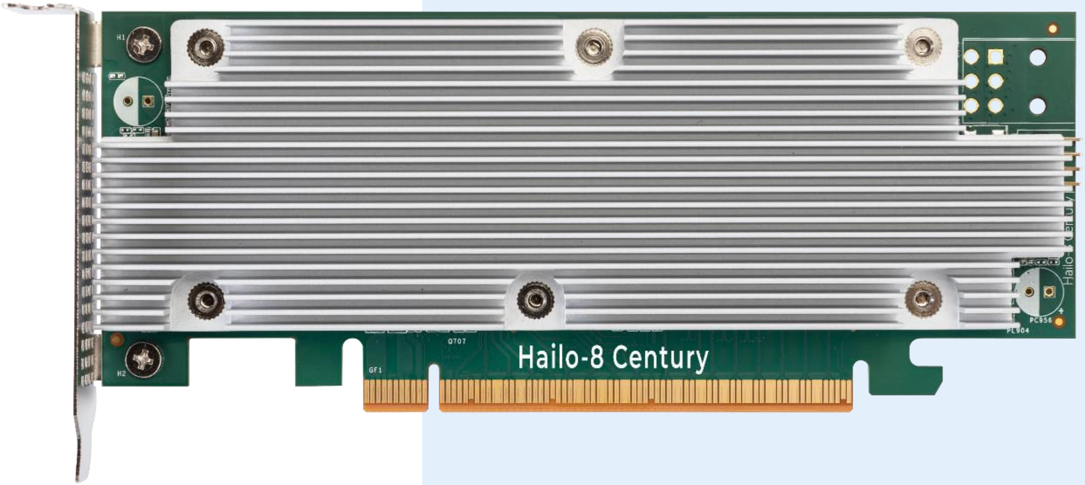
Small ← → Large

Form factor				
# of video channels (FHD @ 25 FPS)	16-32	16-32	Up to 100	Up to 200
AI capacity (TOPS)	26-52 	52-78 	104-208 	104-208 
CPU	arm NEOVERSE intel CELERON intel CORE i5 intel CORE i7 intel CORE i3		intel XEON AMD EPYC	

Hailo-8 Century High Performance PCIe Cards

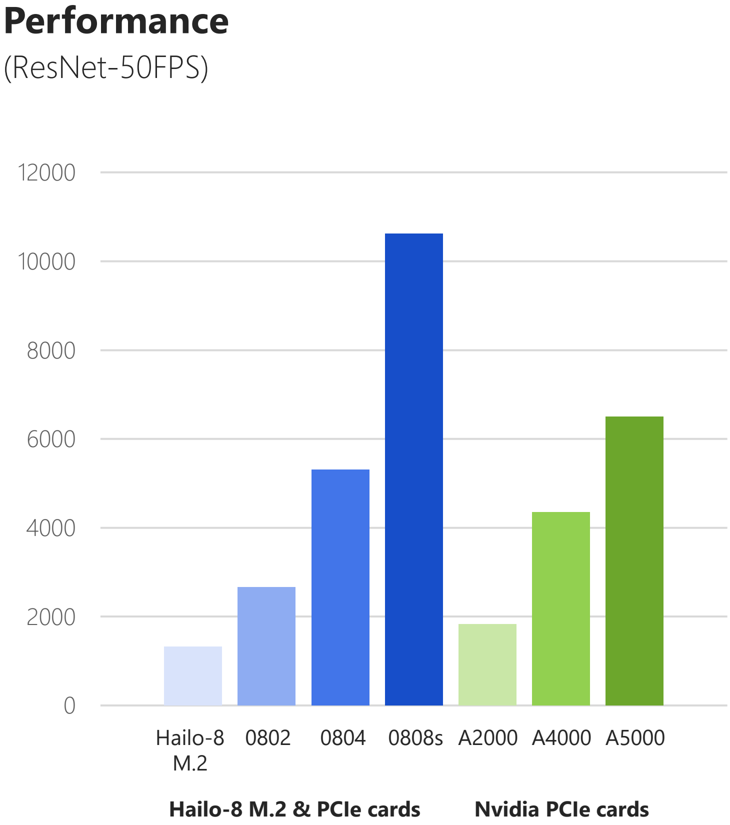
Key Features & Benefits

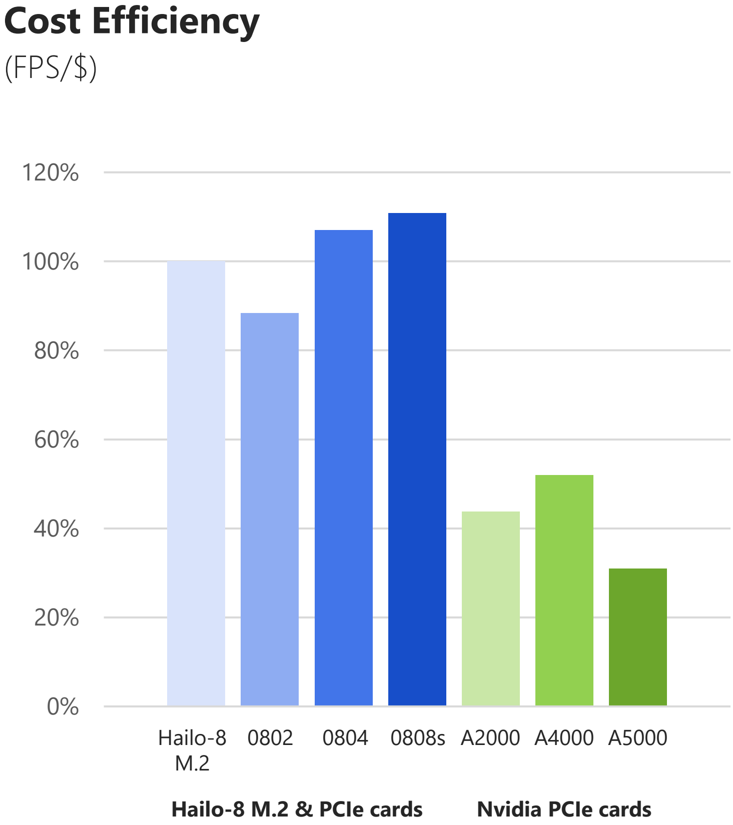
- Delivering 52-208 TOPS
- Best-in-class power efficiency at 400 FPS/W ResNet50
- Highest cost-efficiency (FPS/$)
  → Starting at $249
- Supporting temperature range of -40°C to 85°C
- Passively cooled

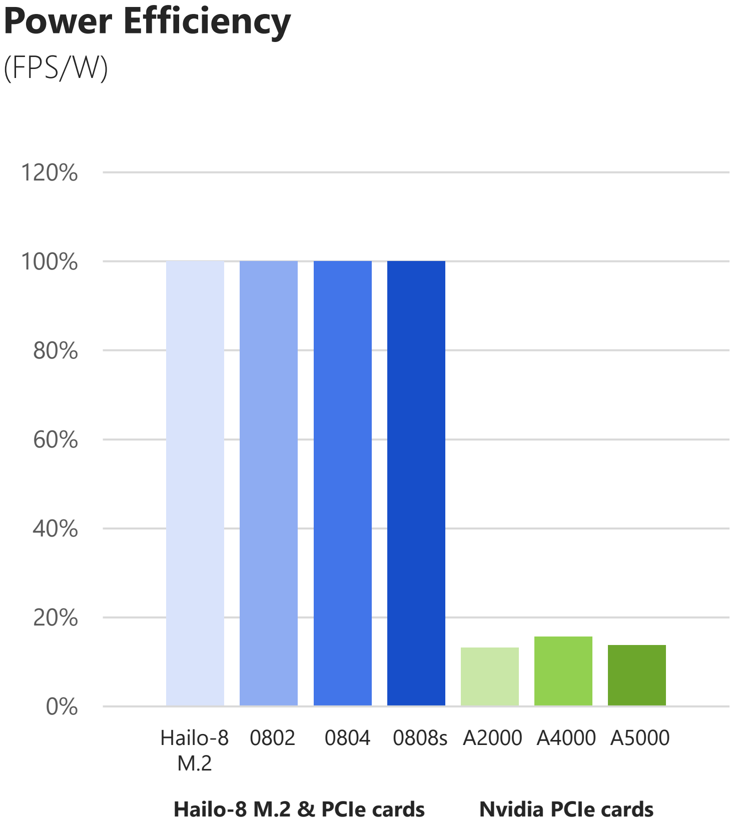


Superior AI Performance

Hailo-8 M.2 & Century vs. Nvidia GPU PCIe Cards







Disruptive Cost-Efficiency

Higher density, low power, cost optimized AI solutions, compared to GPU-based systems.
Enable smaller form factor & lower TCO with HHHL PCIe cards & M.2 modules.

	Small/Medium VMS system up to 32 Channels		**Large VMS System** up to 100 channels		**Large VMS System** up to 200 channels	
AI Component	GPU	Hailo Century 0802/03 M.2 modules	GPU	Hailo Century 0804/0808S	GPU	Hailo Century 0808S
Form Factor	1U	SFF	2U	1U	3U	2U
Typical System MSRP	$10,000	$2,500	$20,000	$5,000	$30,000	$10,000

Up to 75% cost saving!

Hailo Solutions for VMS

Powerful, scalable & efficient AI offering

Cost Efficient

Unrivalled AI compute power per $

Scalable & Versatile

Wide range of form factors ranging from 13-208 TOPS

Easily Integrated

Comprehensive & field-proven software suite

Durable

Industrial grade, passive cooling

Real-Time Insights

Low latency and higher frame rates enable detection and search across multiple video streams

High Accuracy

Low rate of false alarms and mis-detections

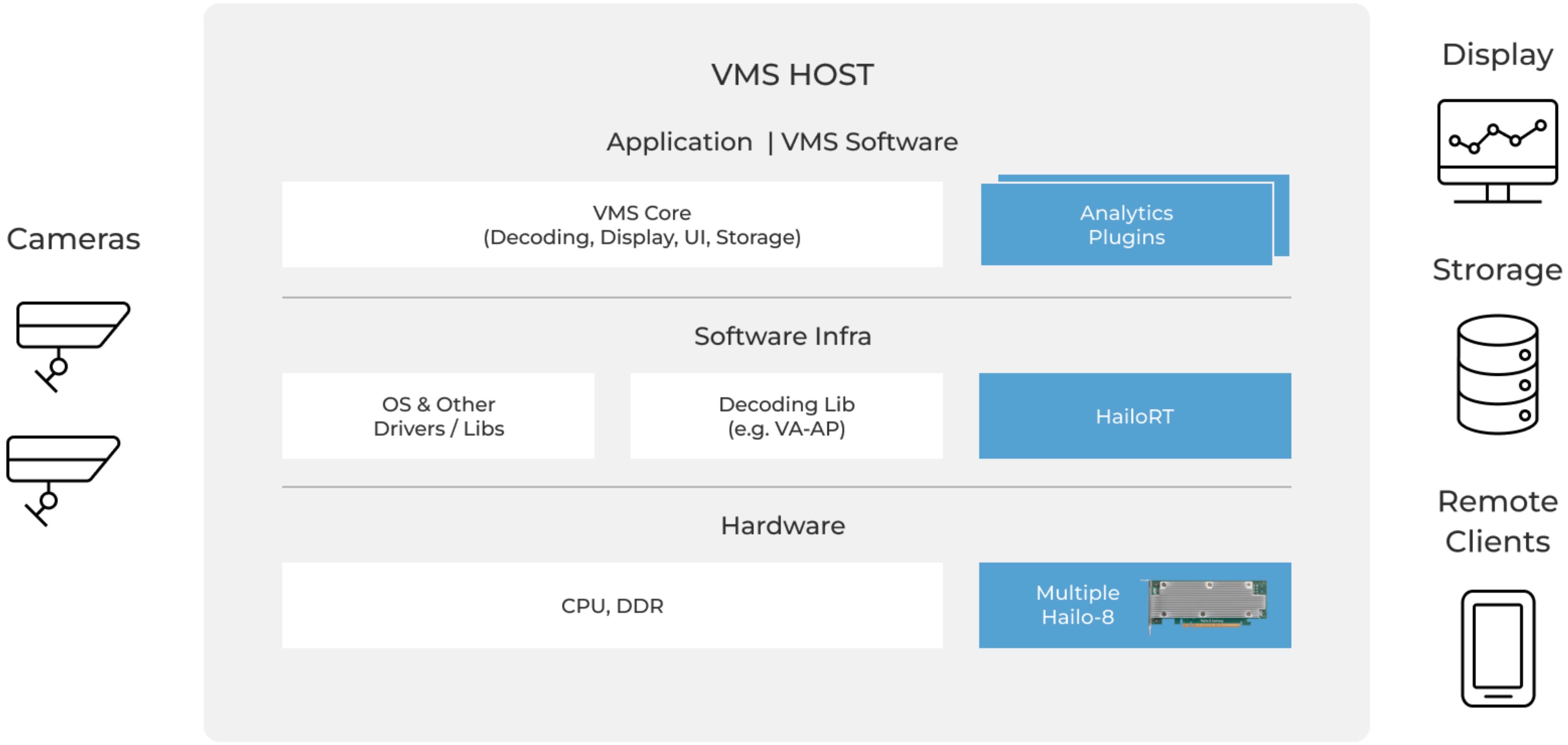
Cutting Edge Analytics

Industry transforming, advanced models and pipelines, including GenAI workloads

Integrating Hailo-Based AI Analytics into VMS

HAILO

VMS Architecture



Cameras

VMS HOST

Application | VMS Software

VMS Core
(Decoding, Display, UI, Storage)

Analytics
Plugins

Software Infra

OS & Other
Drivers / Libs

Decoding Lib
(e.g. VA-AP)

HailoRT

Hardware

CPU, DDR

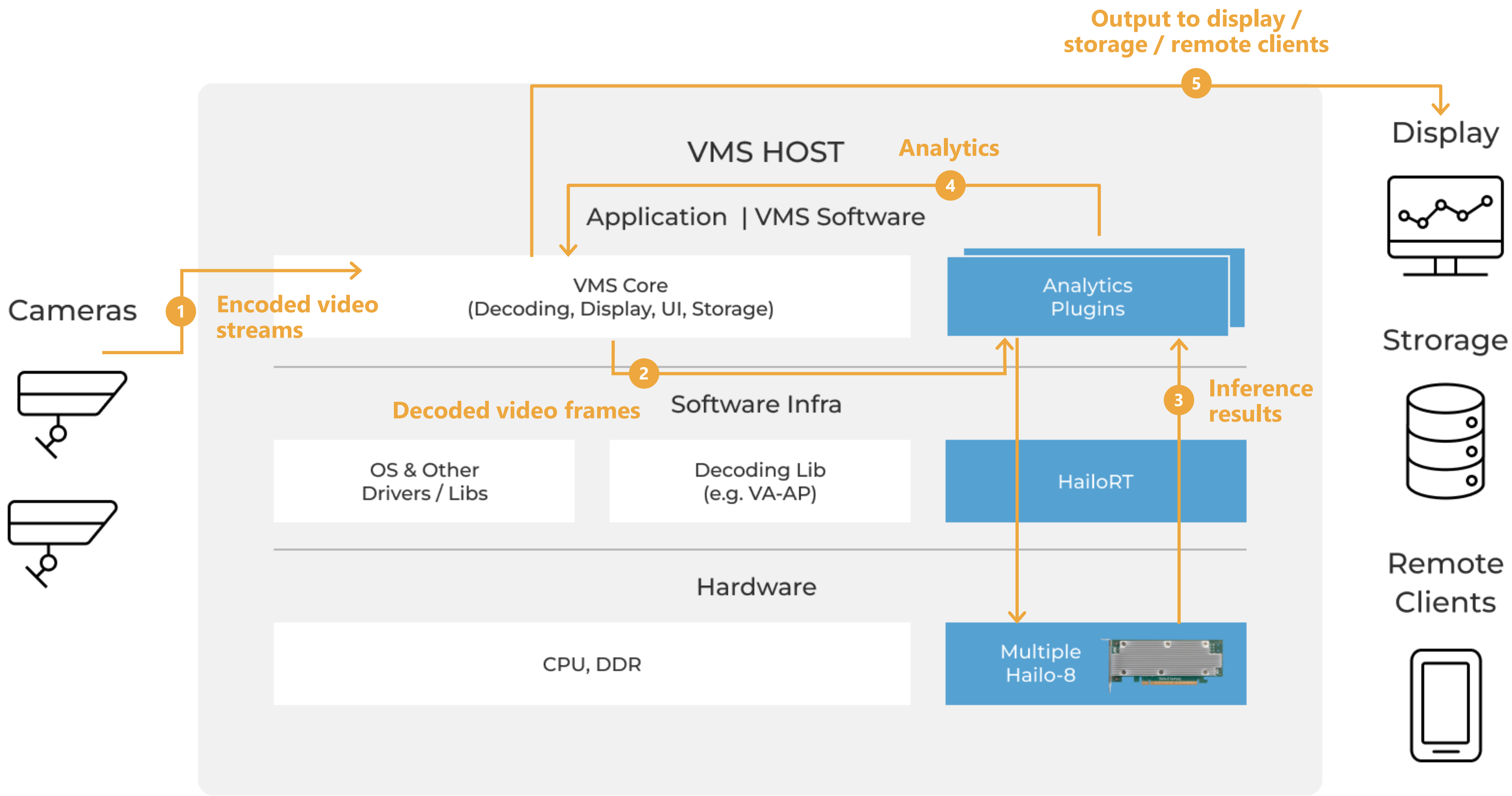
Multiple
Hailo-8

Display

Stroage

Remote
Clients

VMS Architecture



How to Select Your AI-Powered VMS Platform

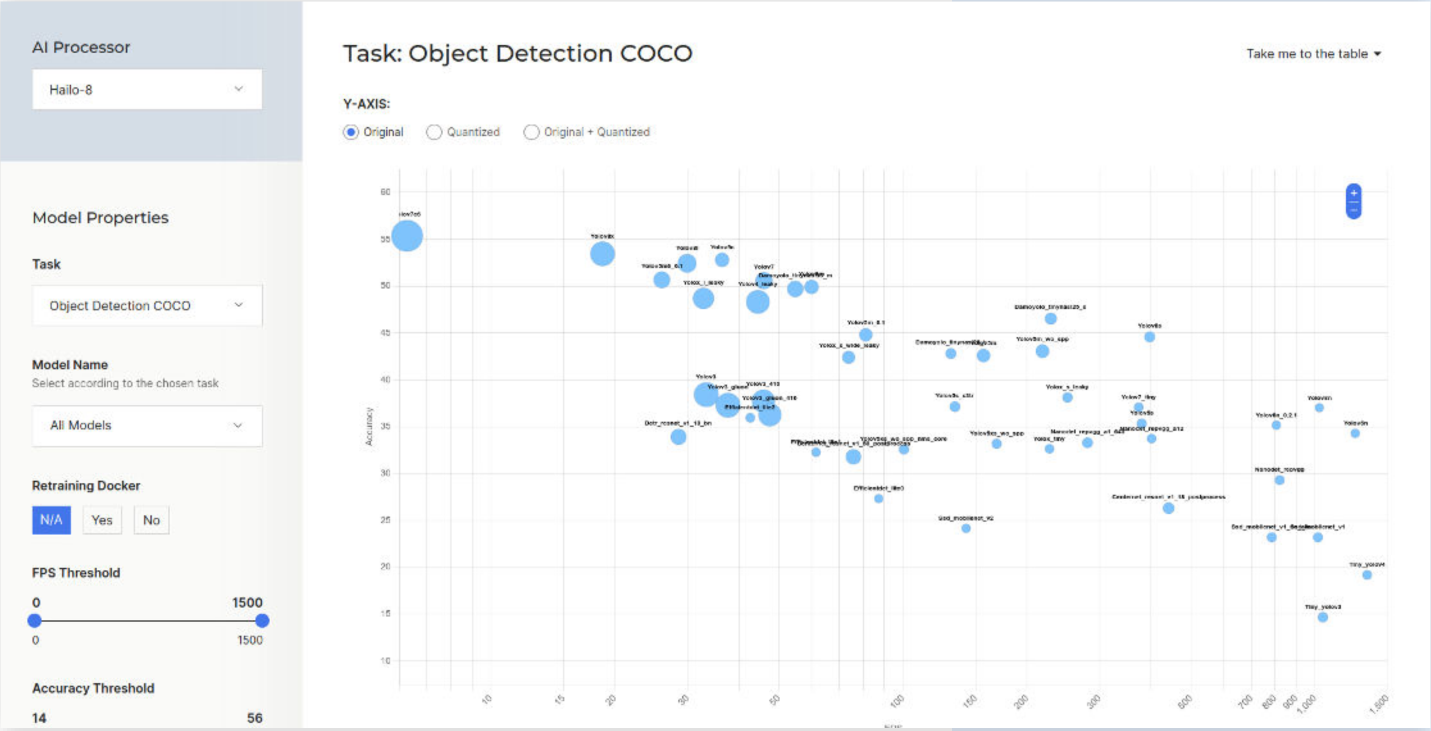
When integrating to a VMS platform, we need to check the following parameters:

1. How is the analytics plugin called?
   - → Blocking / non-blocking? (Prefer non-blocking to get best performance)
   - → Does the number of streams is pre-configured or can be changed online.
2. Which data should be sent and received by the plugin?
   - → RGB / encoded input / Read directly from RTSP
   - → Does the plugin need to track detections?
   - → Is the plugin in-charge of drawing / display?
   - → Does the frame need to be sent back?

How to Design a Multi-Stream Pipeline

Define a prototype pipeline required for your application:

- Which tasks are required?
- Are there dependencies between networks?
- What are resolutions and formats?
- Define video processing requirements (decoding, encoding, resize, crop, format conversion)
- Select networks - [Model Explorer]()
- Test required networks and expected bandwidth using the [hailortcli run2 tool]()



AI Processor: Hailo-8

Task: Object Detection COCO

Y-AXIS: ● Original ○ Quantized ○ Original + Quantized

Model Properties

Task: Object Detection COCO

Model Name
Select according to the chosen task
All Models

Retraining Docker: N/A | Yes | No

FPS Threshold: 0 – 1500

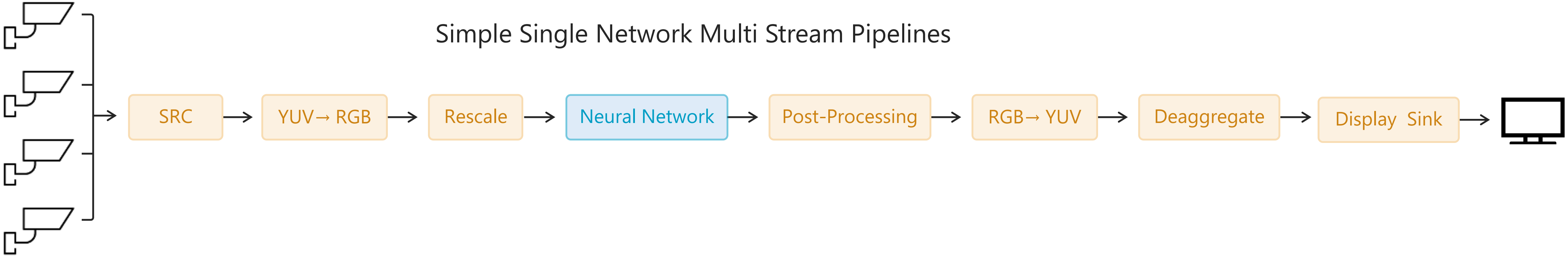
Accuracy Threshold: 14 – 56

Multi Stream Pipeline Implementation Options

🟨 Elements running on host

🟦 Elements offloaded to Hailo

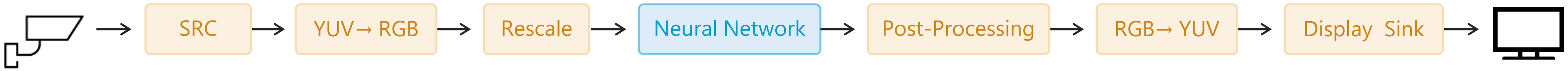
Multi stream pipeline (all streams are aggregated to one pipeline)



Simple Single Network Multi Stream Pipelines

Multiple single stream pipelines (the same pipeline is duplicated per stream)

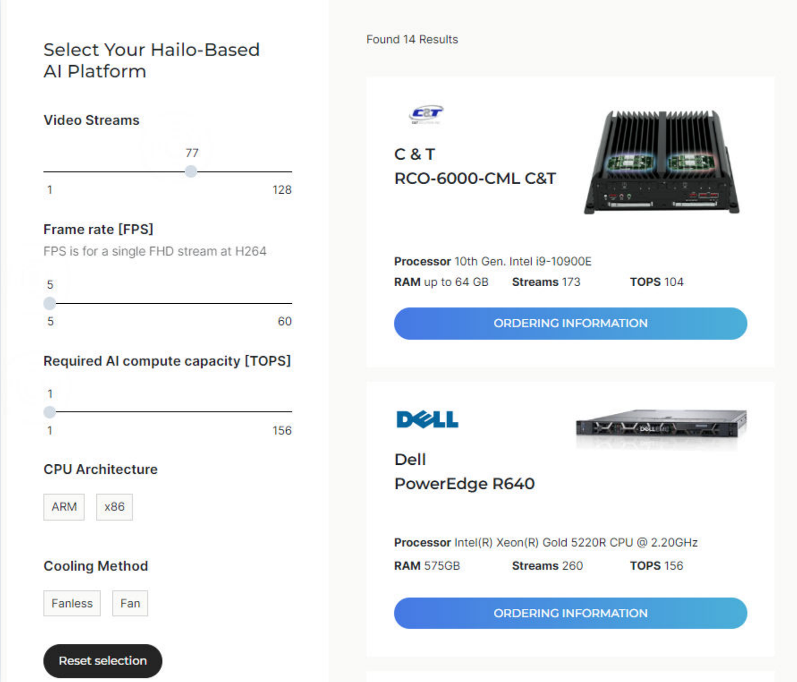
Simple Single Network Pipelines



Fine Tune for Performance – Hardware

- Which HW platform are you using?
  - → Which tasks can be HW accelerated?
  - → Use VAPPI, ISP, HW encoder / decoder where possible
- How many Hailo devices do you have?
  - → Define how to allocate networks to devices
  - → Experiment with batch size, scheduler priorities, timeouts, etc.

Platform Selection Guide



Select Your Hailo-Based AI Platform

Video Streams 77 (1–128)

Frame rate [FPS] 5 (5–60)

FPS is for a single FHD stream at H264

Required AI compute capacity [TOPS] 1 (1–156)

CPU Architecture: ARM, x86

Cooling Method: Fanless, Fan

Found 14 Results

C & T RCO-6000-CML C&T

Processor 10th Gen. Intel i9-10900E
RAM up to 64 GB **Streams** 173 **TOPS** 104

Dell PowerEdge R640

Processor Intel(R) Xeon(R) Gold 5220R CPU @ 2.20GHz
RAM 575GB **Streams** 260 **TOPS** 156

Find your platform here

Fine Tune for Performance – hailortcli run2

- See documentation: Multiple HEF Inference

For each vdevice, you can control:

- Device count
- batch size
- Framerate
- Scheduler threshold, timeout and priority

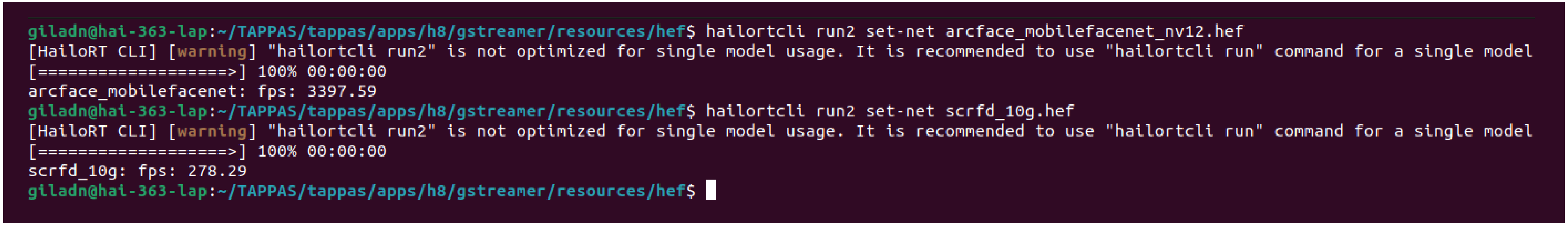
Note: To run more than one "vdevice" use multiple hailortcli run2 processes.

Fine Tune for Performance – Example

The task is "face recognition", it is implemented by running 2 cascaded networks:

- Face detection and landmark network: scrfd_10g.hef
- Face recognition network: arcface_mobilefacenet_nv12.hef

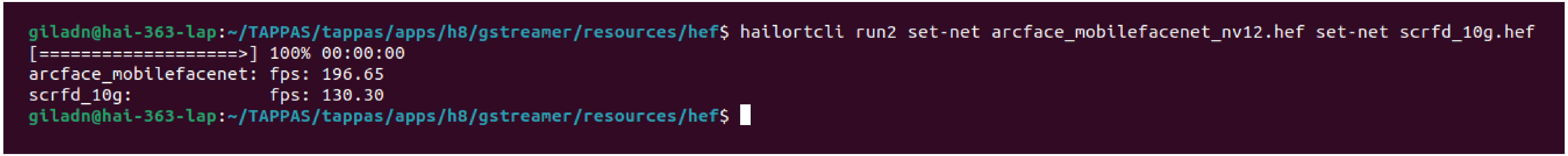
1. Check the maximum performance of both networks:



```
giladn@hai-363-lap:~/TAPPAS/tappas/apps/h8/gstreamer/resources/hef$ hailortcli run2 set-net arcface_mobilefacenet_nv12.hef
[HailoRT CLI] [warning] "hailortcli run2" is not optimized for single model usage. It is recommended to use "hailortcli run" command for a single model
[===================>] 100% 00:00:00
arcface_mobilefacenet: fps: 3397.59
giladn@hai-363-lap:~/TAPPAS/tappas/apps/h8/gstreamer/resources/hef$ hailortcli run2 set-net scrfd_10g.hef
[HailoRT CLI] [warning] "hailortcli run2" is not optimized for single model usage. It is recommended to use "hailortcli run" command for a single model
[===================>] 100% 00:00:00
scrfd_10g: fps: 278.29
giladn@hai-363-lap:~/TAPPAS/tappas/apps/h8/gstreamer/resources/hef$ 
```

Fine Tune for Performance – Example (Cont.)

2. Naïve test, try to run both networks:



```
giladn@hai-363-lap:~/TAPPAS/tappas/apps/h8/gstreamer/resources/hef$ hailortcli run2 set-net arcface_mobilefacenet_nv12.hef set-net scrfd_10g.hef
[===================>] 100% 00:00:00
arcface_mobilefacenet: fps: 196.65
scrfd_10g:             fps: 130.30
giladn@hai-363-lap:~/TAPPAS/tappas/apps/h8/gstreamer/resources/hef$ 
```

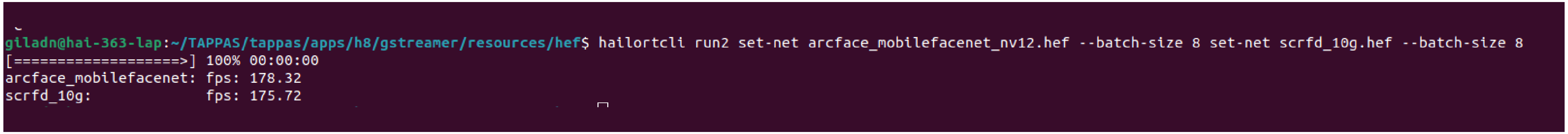
Fine Tune for Performance – Batching

- What is "batching" good for?
- Can we keep increasing batch size for better performance? No…
  - → Bigger batch will require more memory.
  - → Limited by PCIe page size
  - → [HailoRT] [warning] Desc page size value (1024) is not optimal for performance.
- Increasing batch size can increase FPS but it will also increase latency.

Fine Tune for Performance – Example (Cont.)

3. Ramp up the batch size

```
hailortcli run2 \
set-net arcface_mobilefacenet_nv12.hef --batch-size 8 \
set-net scrfd_10g.hef --batch-size 8
scrfd_10g:              fps: 175.93
arcface_mobilefacenet: fps: 177.33
```

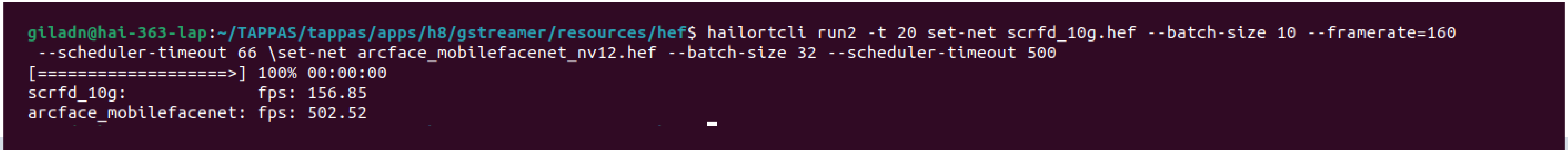


```
giladn@hai-363-lap:~/TAPPAS/tappas/apps/h8/gstreamer/resources/hef$ hailortcli run2 set-net arcface_mobilefacenet_nv12.hef --batch-size 8 set-net scrfd_10g.hef --batch-size 8
[===================>] 100% 00:00:00
arcface_mobilefacenet: fps: 178.32
scrfd_10g:             fps: 175.72
```

Fine Tune for Performance – Example (Cont.)

4. Fine tune the schedular
   → In our test case we need to run face detection on each frame and send the detected faces to the recognition network.
   → In this example we will run detection at 15 fps for 10 streams. i.e., 150 fps.

```
hailortcli run2 -t 20 \
set-net scrfd_10g.hef --batch-size 10 --framerate=160 --scheduler-timeout 66 \
set-net arcface_mobilefacenet_nv12.hef --batch-size 32 --scheduler-timeout 500
arcface_mobilefacenet: fps: 501.96
scrfd_10g:             fps: 156.86
```



```
giladn@hai-363-lap:~/TAPPAS/tappas/apps/h8/gstreamer/resources/hef$ hailortcli run2 -t 20 set-net scrfd_10g.hef --batch-size 10 --framerate=160
 --scheduler-timeout 66 \set-net arcface_mobilefacenet_nv12.hef --batch-size 32 --scheduler-timeout 500
[===================>] 100% 00:00:00
scrfd_10g:             fps: 156.85
arcface_mobilefacenet: fps: 502.52
```

Fine Tune for Performance – Optimization

Minimize costly video operations:

- Use secondary stream from Camera / ISP. (Set fps and resolution from source)
- Make sure zero copy and in-place editing is used when possible
- Use the lowest resolution stream possible
- Keep original high-resolution stream for cropping and display
- Experiment… (HW accelerators, different formats, order of operations)
  → For example, resize->conversion might be quicker than conversion ->resize

Fine Tune for Performance

- Why and when use Hailo's on chip conversions?
  - → Fixed resize, NV12, YUY2, RGBx inputs conversions
  - → Many post-processing functions are supported by the Hailo Dataflow Compiler & HailoRT (yolov5, yolov8, yolox, SSD, …)
  - → See: Model_Optimization_Tutorial
- Use Async API (see Async API example)
- Run performance tests and debug tools to find bottlenecks
- Rinse and repeat….

Integration with VMS Software – Nx Platform Example

- Network Optix Plugin requirements
  - → Support arbitrary number of streams
  - → Frames are decoded by the server and handled by a callback
  - → Plugin returns metadata to the server with tracking ID and frame timestamp
  - → Drawing is done by the server, no need to send back the video frame
- Integration was done using a per stream pipeline
- Used Gstreamer pipeline with "appsrc" plugin to send data to pipeline (With timestamp)
- We used Async API triggering a callback function for each processed frame which sent the metadata + timestamp back to server

Next generation VMS capabilities

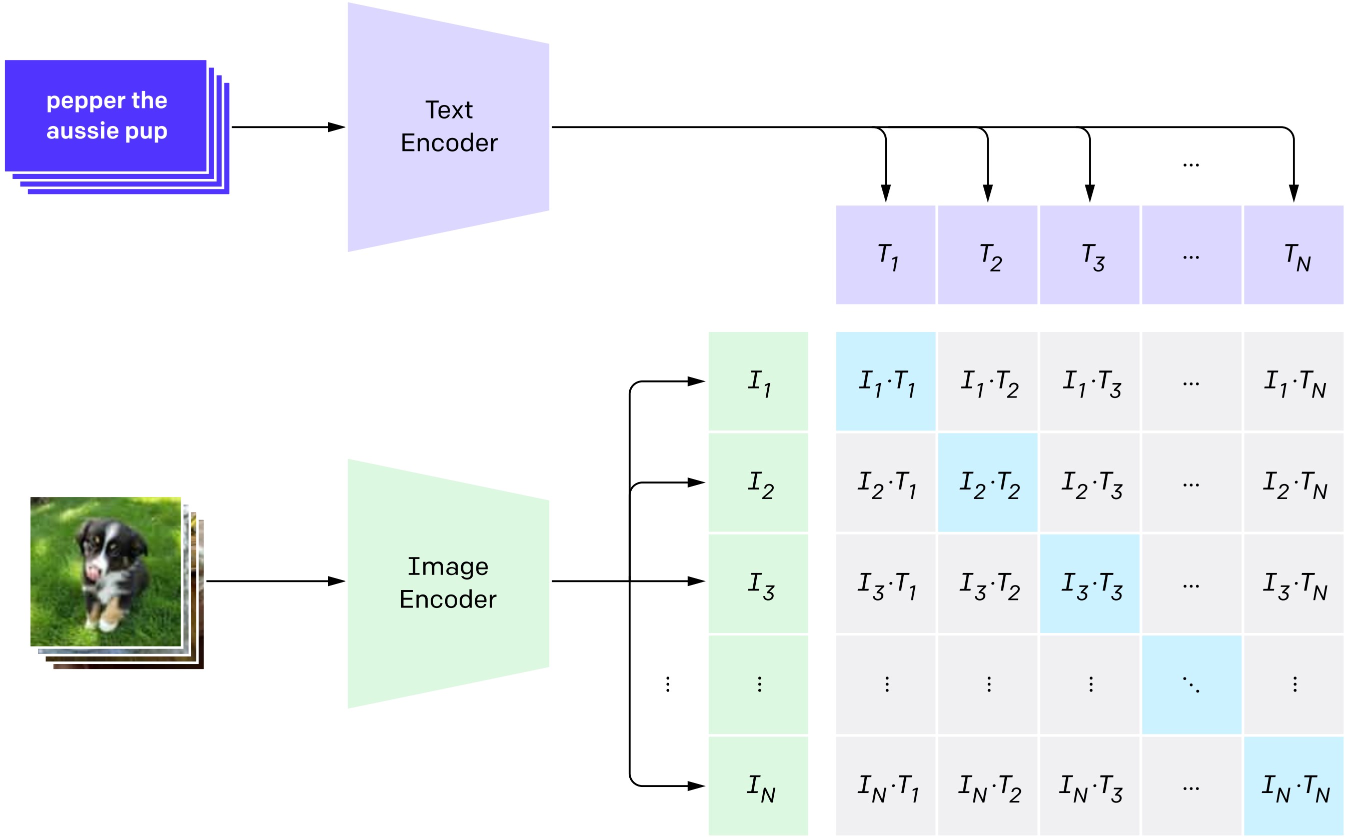
Using CLIP for zero shot free text searching on live video streams

HAILO

What is CLIP?

Trained on image, image caption pairs.

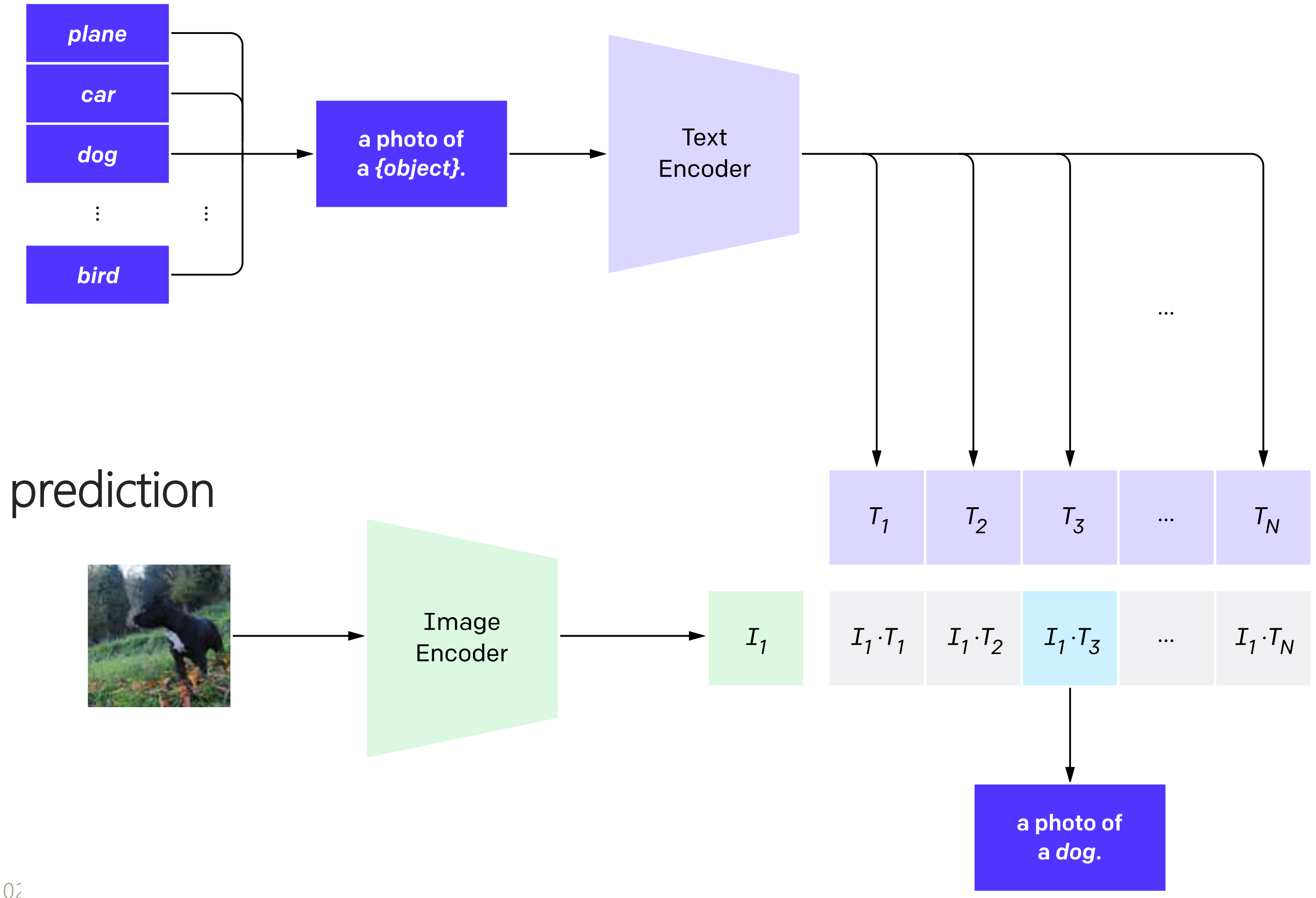
Takes inputs from text and image domains and generate a vector in a shared latent space.



	T_1	T_2	T_3	…	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	…	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	…	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	…	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	…	$I_N \cdot T_N$

CLIP Usage Example

2. Create dataset classifier from label text



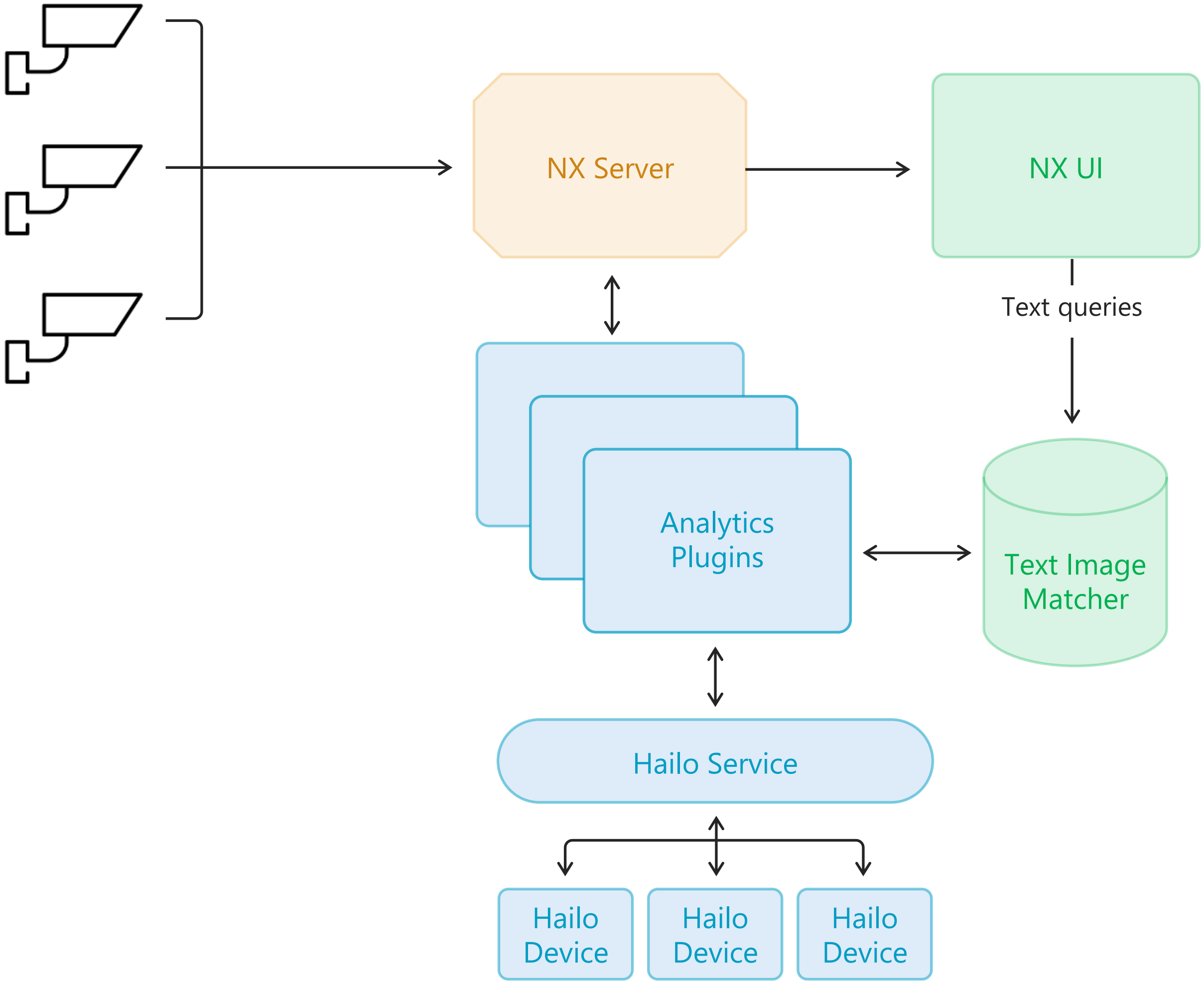
3. Use for zero-shot prediction

Next Generation VMS Capabilities

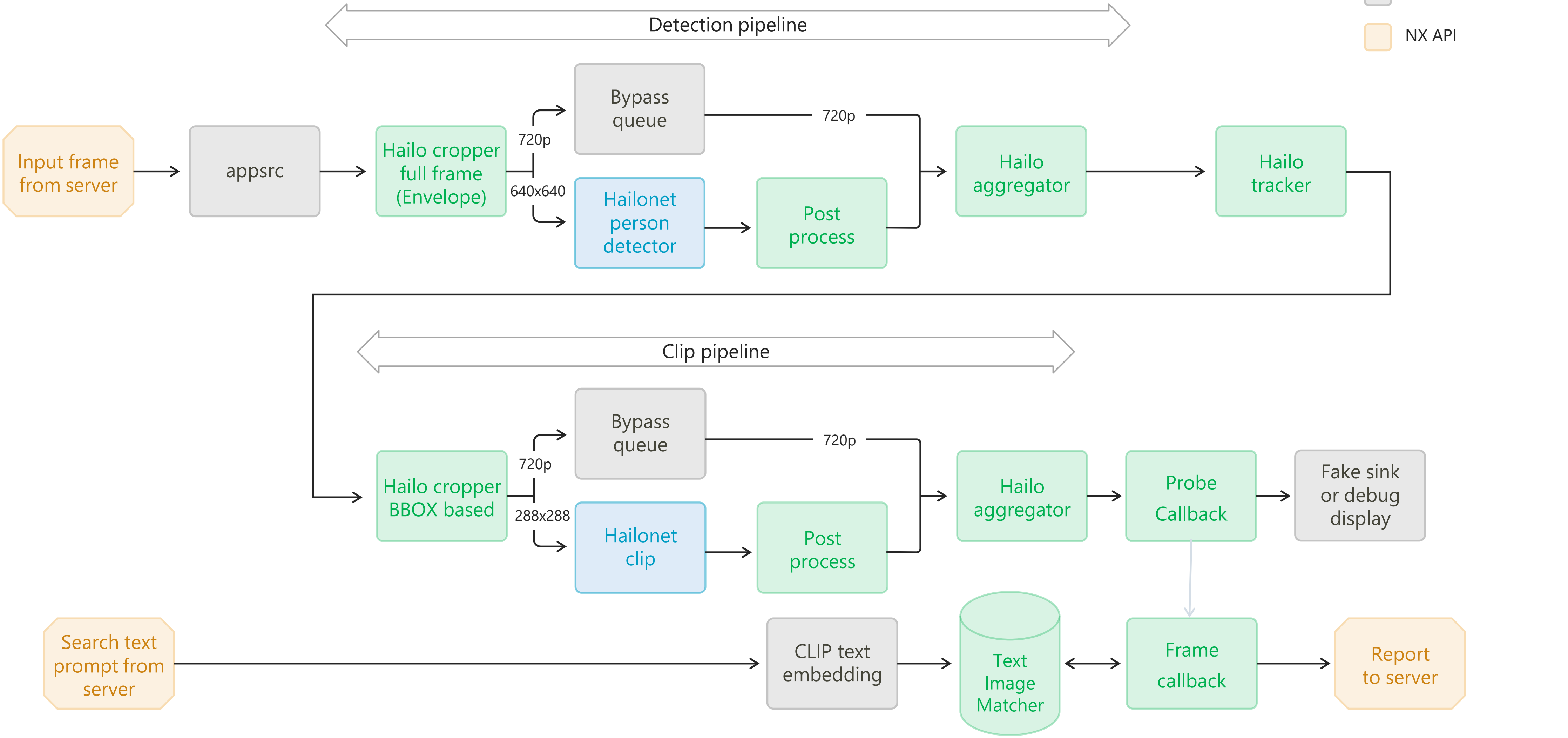
Why integrate CLIP into a VMS application?

- Natural language queries (Zero Shot)
- LLM based network allows for better generalization and scene understanding.
- Flexibility across domains (No specific domain training)
- Online actions:
  - → Relevant stream highlighting.
  - → Automated actions, Start recording, call security, set alarm
- Offline features:
  - → Efficient data management (retention policies and quality)
  - → Search for "new prompts" in available metadata

Integration with VMS Software – Nx Example



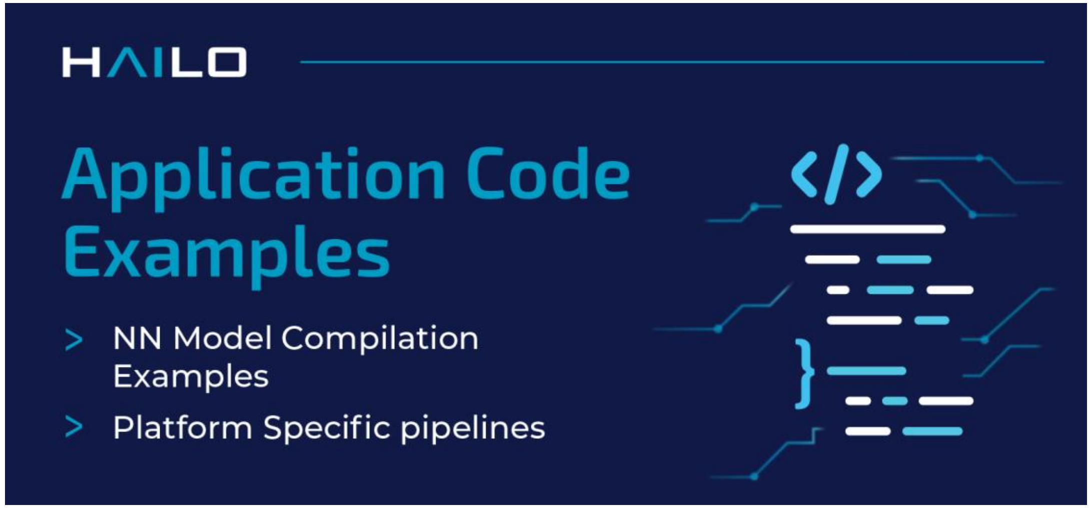
CLIP Pipeline Overview



NX VMS CLIP Demo

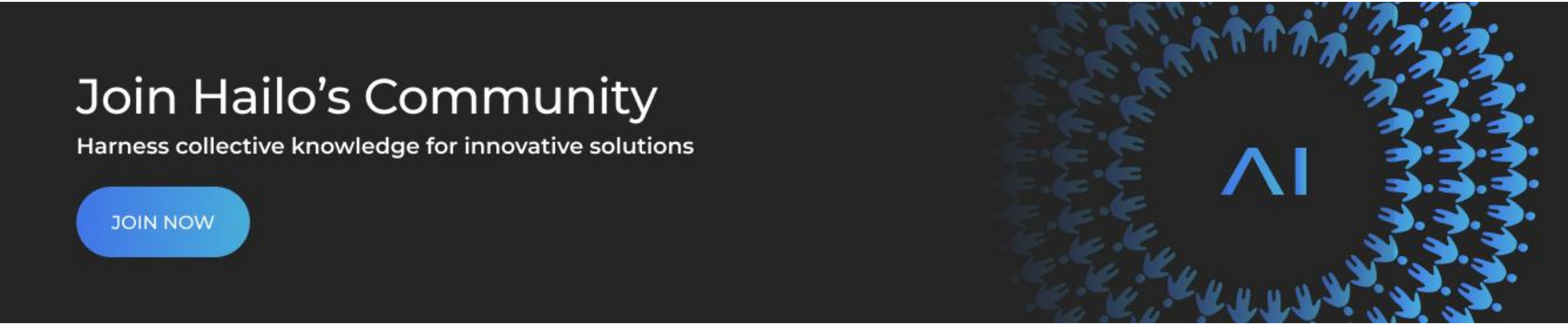
Open-Source CLIP Application

- Hailo is committed to the open-source community.
- Check out our CLIP app on GitHub Hailo-Application-Code-Examples
    - Find the application under runtime/gstreamer/hailo_clip
- Also available on Raspberry Pi 5 soon…


HAILO
Application Code Examples
> NN Model Compilation Examples
> Platform Specific pipelines

Hailo Community

- Hailo is launching a developer community https://community.hailo.ai/
- Pre-launch access to webinar participants.
- Sign in using your developer community credentials.
- Official link from Hailo developer zone will be added soon.



Summary – Hailo Solutions for VMS

Cost Efficient

Up to 75% cost saving on VMS hardware

Scalable & Versatile

Up to 200 channels of powerful AI analytics

Cutting Edge Analytics

Advanced models and pipelines, for accurate, zero shot search and indexing

Easily Integrated

field-proven integration with leading vendors

Q&A

HAILO

HAILO

Empowering Intelligence

THANK YOU

🌐 hailo.ai ✉ contact@hailo.ai