Real-time deep learning on video streams

by Eran Avidan
Senior software engineer

Advanced Analytics @ Intel

	Embed learning	HW validation	Product Dev	Sales	Industrial AI	Health
Value	Improve products power & performance	Cut product time to market	Reduce test costs and improve quality	Increase revenue	Reduce Manufacturing costs	Improve clinical trials outcome
HOW	Adaptive & personalized HW	Automated validation with context	For every unit test only what's needed	Autonomous accounts coverage	Proactive actuation to changes	Continuance monitoring at home

Vision: Put AI to work for human experts

Automating visual inspection

Video analysis for Visual inspections

Different defects may require different models

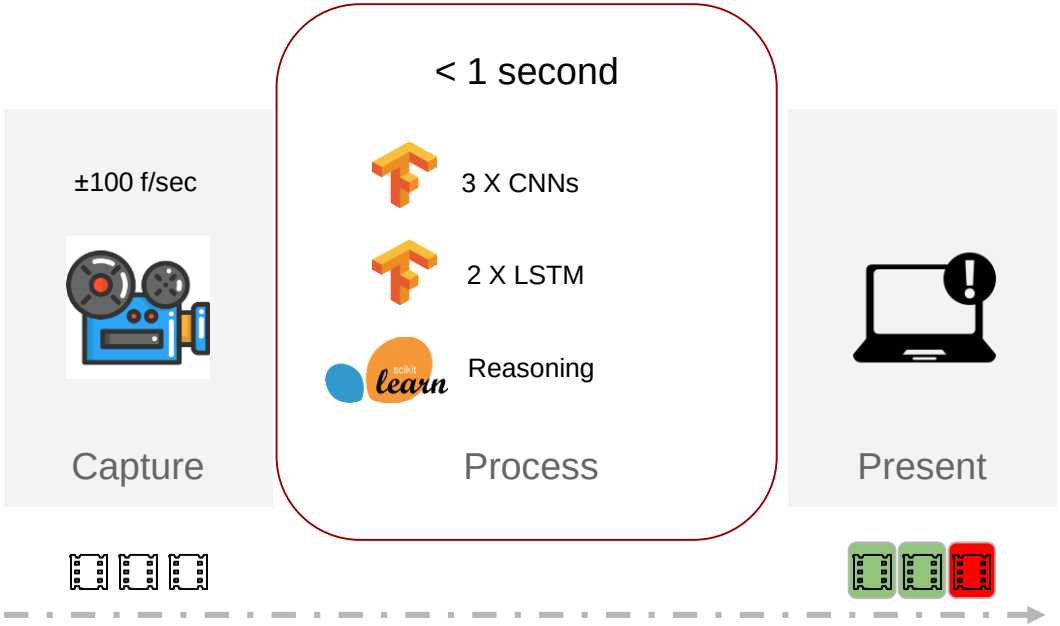
Anomaly behaviors lacks smoothness over time

Unique problems

Using available equipment for analysis

Centrally serving deep learning models with focus on **low latency**

For instance



Using the **lowest possible** inference time of each model

Challenges

Real time prediction serving challenges

Serving complex models (CNN) is part of the critical path

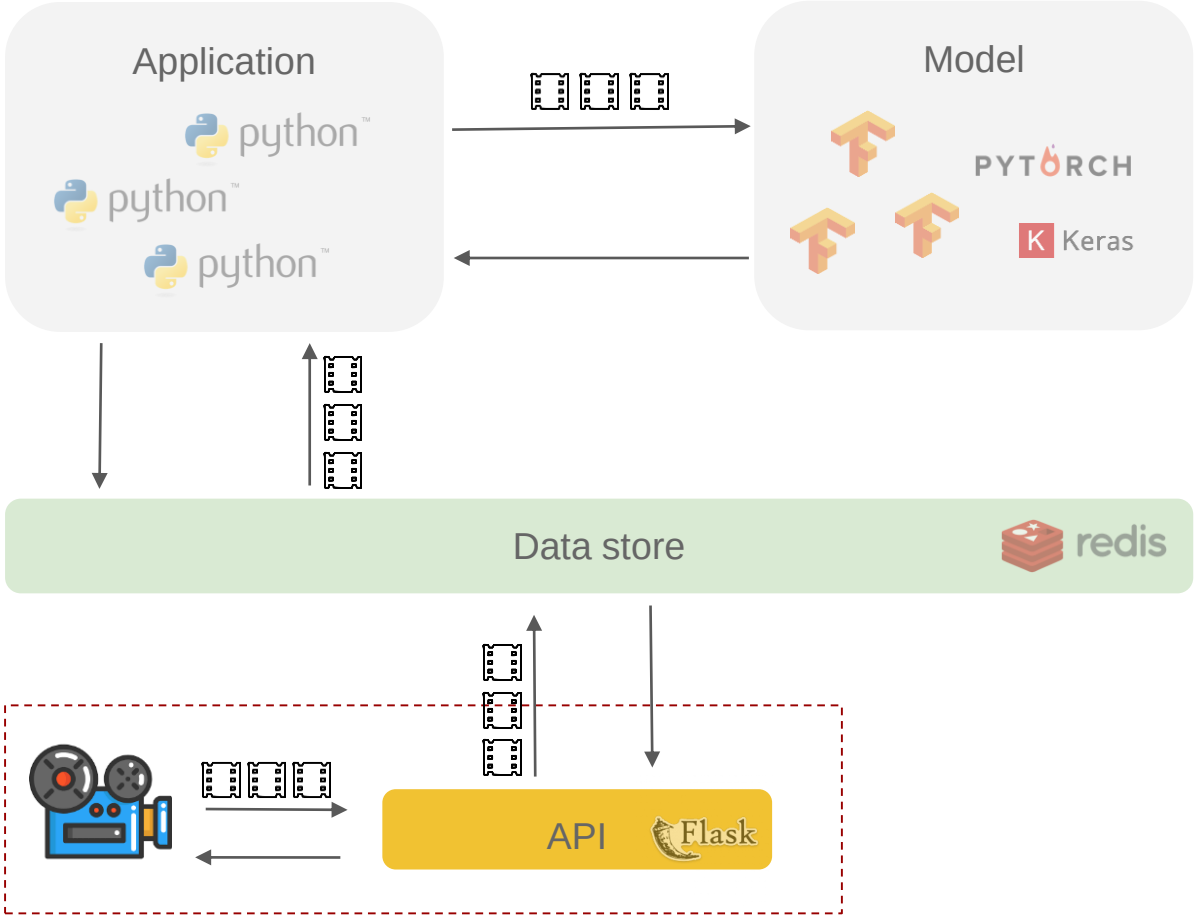
Fast turnaround

Batching to maximize throughput increases latency

Variety of models and frameworks

Archelon

A scalable, fault tolerant and fully asynchronous serving system for video streams


Application
python
python
python
Model
PYTORCH
Keras
Data store
redis
API
Flask

The Application

Asynchronous Inference Unit

Always ON

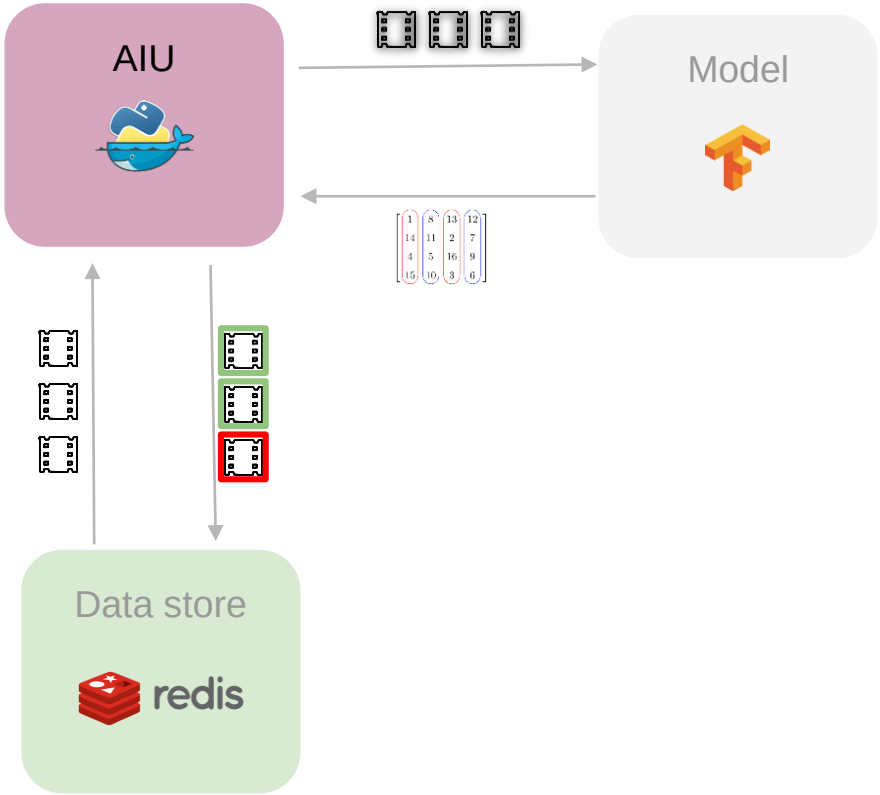
Constant resource utilization

Dynamic Batching

Stateless

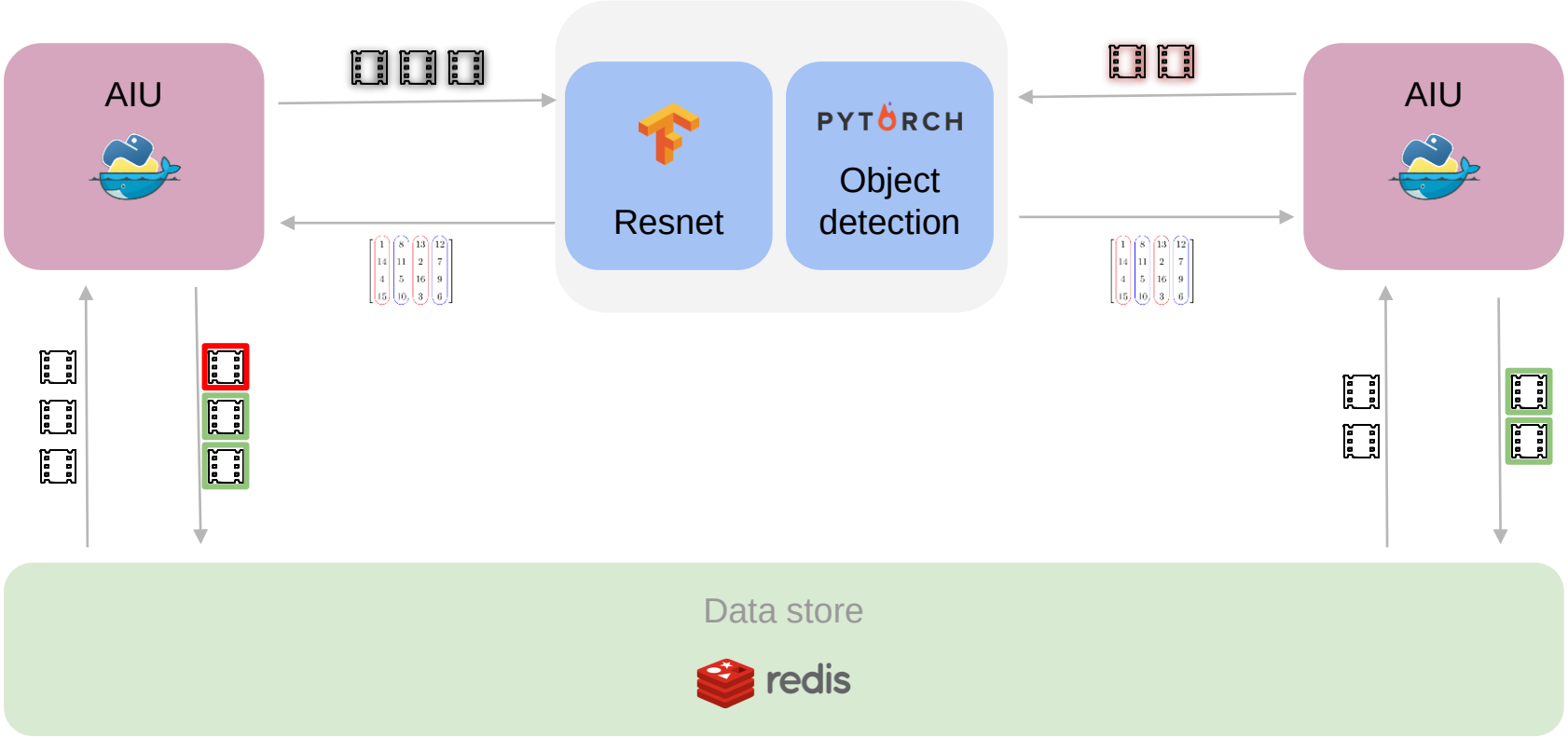
Scalable

Pre-Post processing



Additional benefits

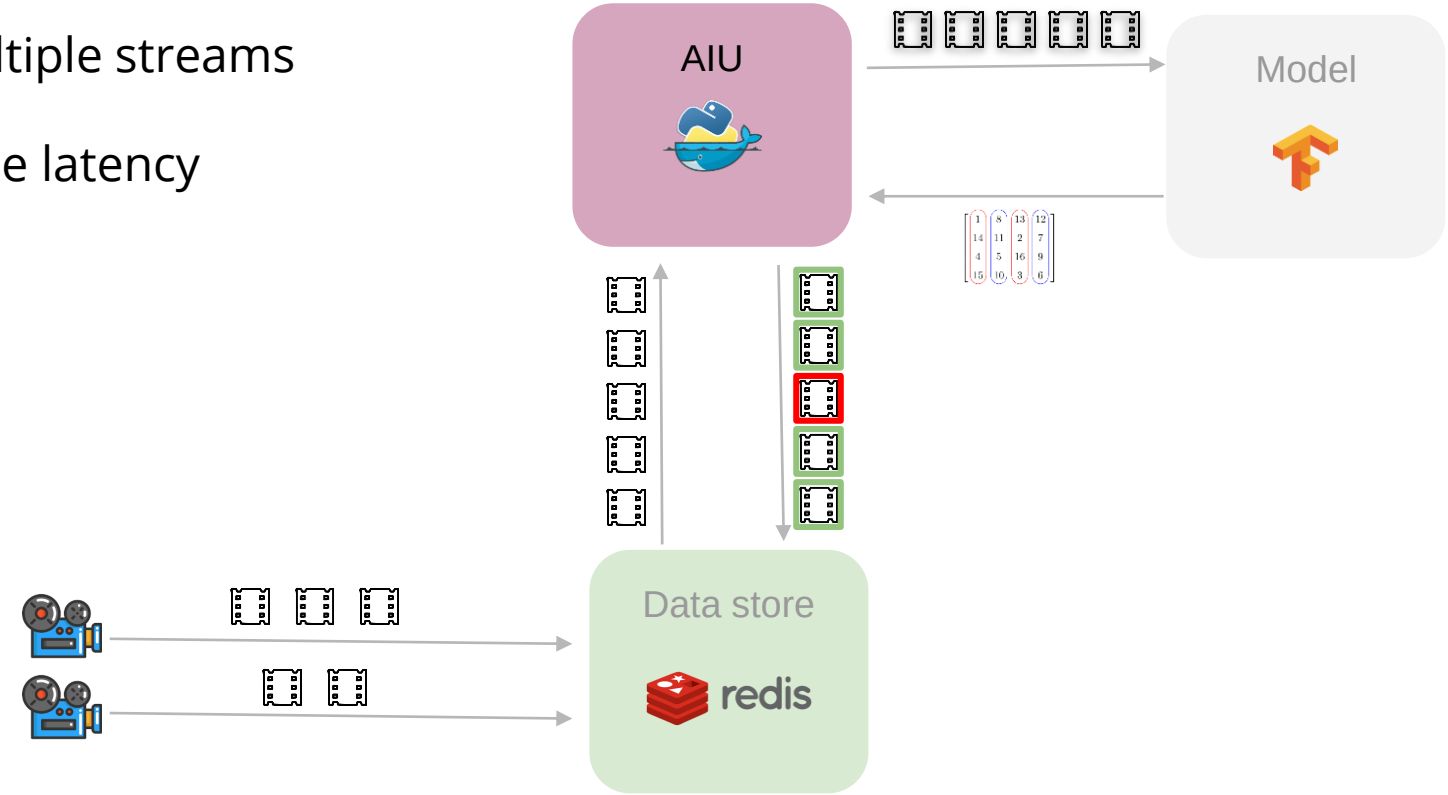
Separation of concerns



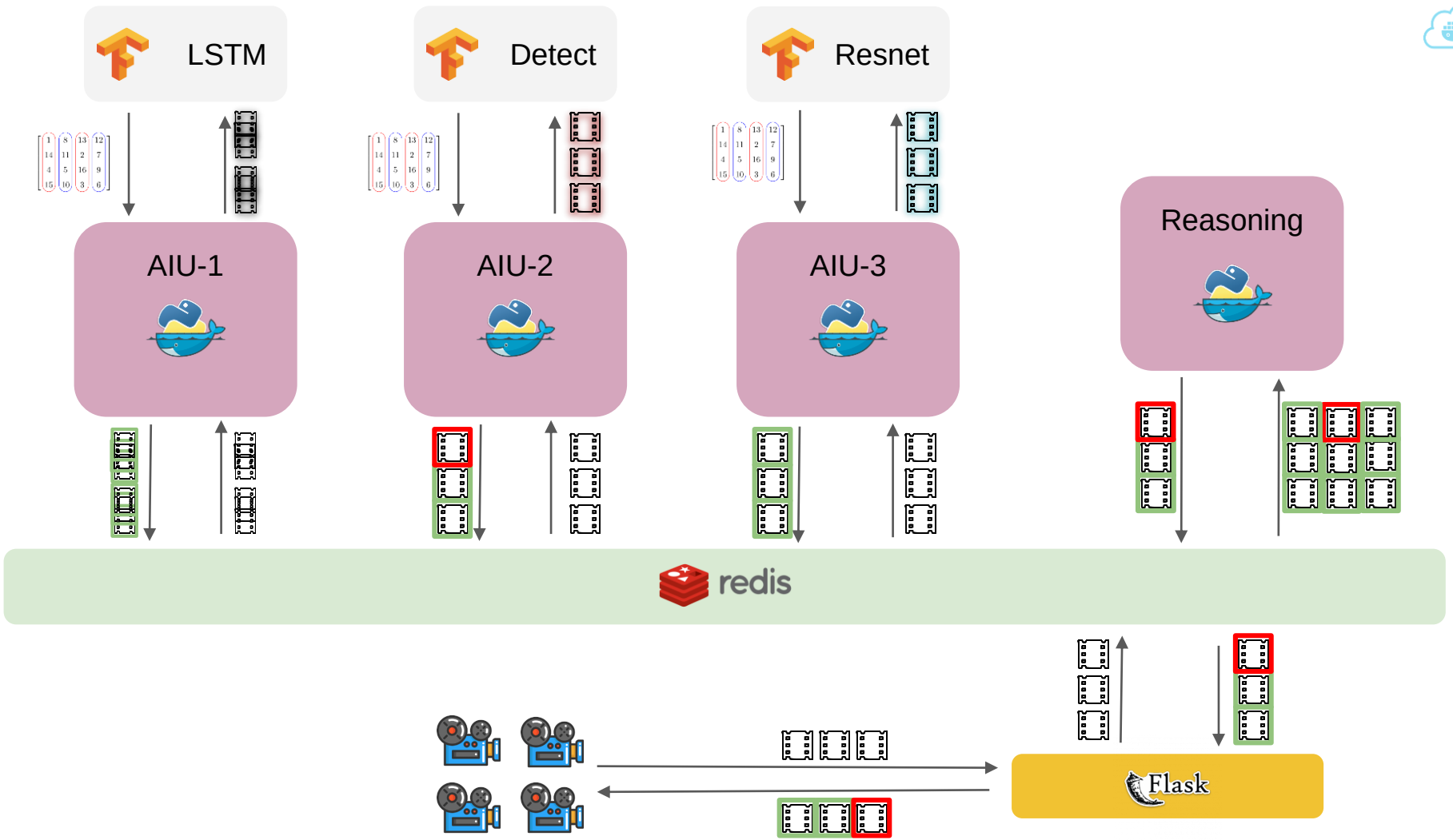
Better resource utilization

Process of multiple streams

Batch to reduce latency



Data flow


LSTM
Detect
Resnet
AIU-1
AIU-2
AIU-3
Reasoning
redis
Flask

General suggestions to reduce latency

Receive a smaller image if possible

If not, resize the image on arrival

Separate presentation from prediction

Choose processing resources according to need

Implement the best serving method for each model

```
exit(0)
```

Real-time deep learning on video streams

by Eran Avidan
Senior software engineer