SPARK+AI
SUMMIT 2019

Build. Unify. Scale.

WIFI SSID: SparkAISummit | Password: UnifiedAnalytics

ORGANIZED BY
databricks

SparkWeaver: Accelerating Real-time DNN Applications with Spark and DnnWeaver

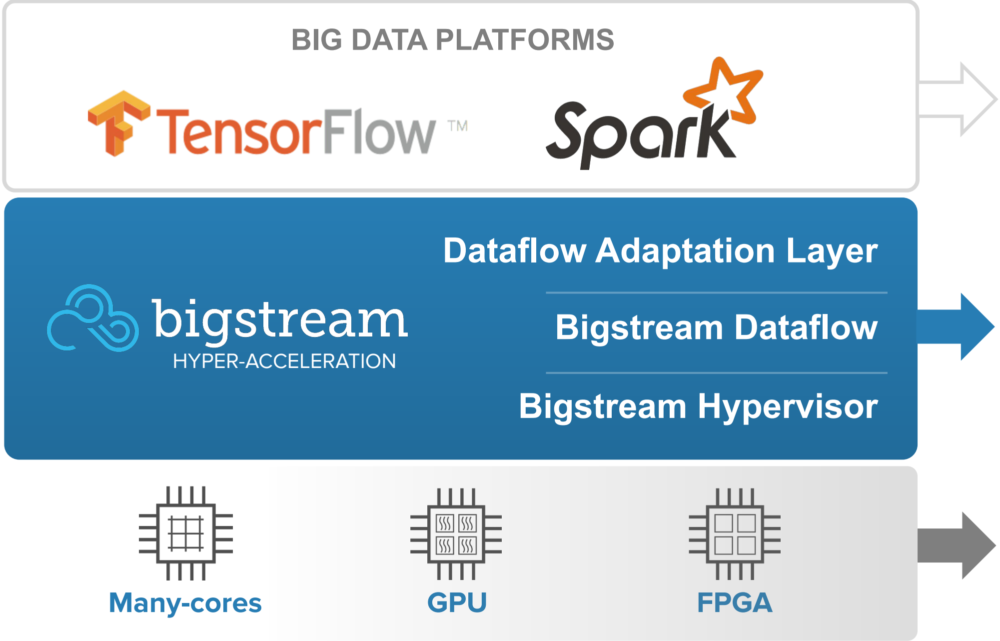
Behnam Robatmili, Jongse Park, and Blake Skinner
Bigstream Solutions

#UnifiedAnalytics #SparkAISummit

SPARK+AI SUMMIT 2019

bigstream

A little about Bigstream



Zero code change

Cross platform

Intelligent, automatic computation slicing

Cross acceleration hardware

2X to 30X acceleration

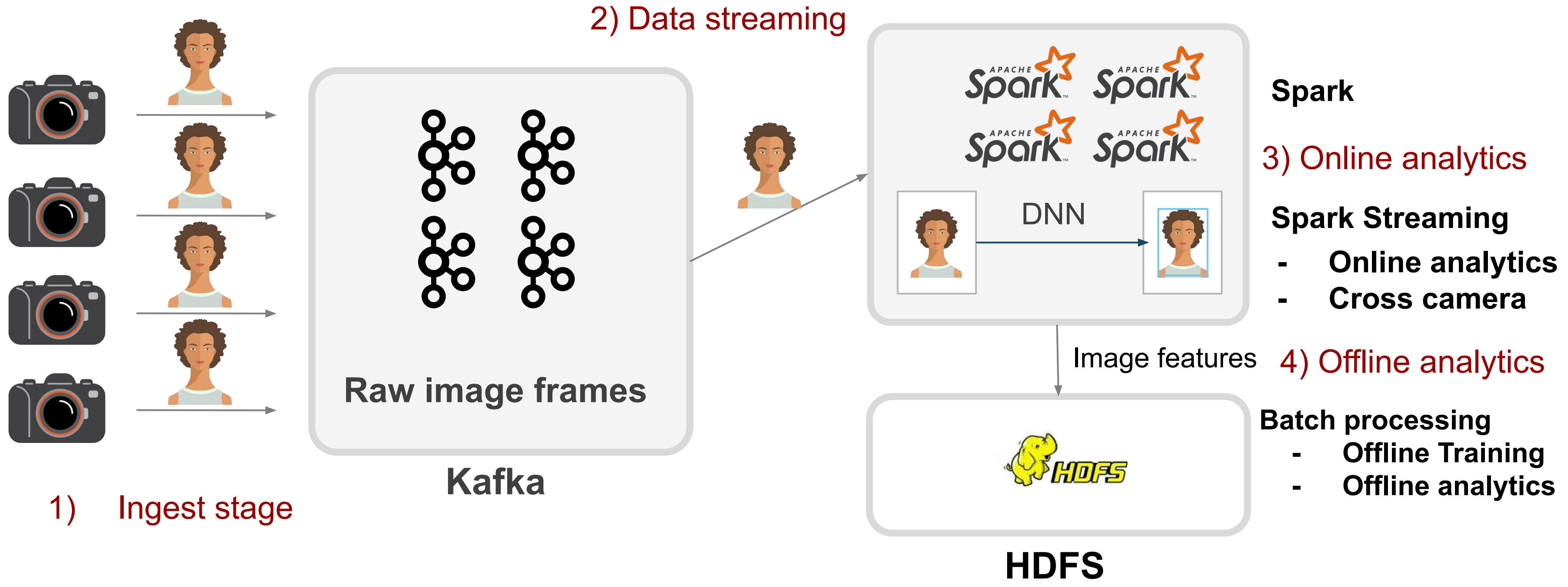
Ingest Bottleneck in Big Data

Applications with Ingest Bottleneck

- **Many big data applications**
  - Lots of raw data
  - Video surveillance
    - Industrial camera market is projected to increase 2.3x by 2024 [1]
    - For a 4k camera in 60fps, the amount of data per hour is 5.2 TB (1TB for 10fps)
  - Voice recognition
  - Fraud detection

1. https://www.gminsights.com/industry-analysis/ip-camera-market

Traditional Architecture does not Scale

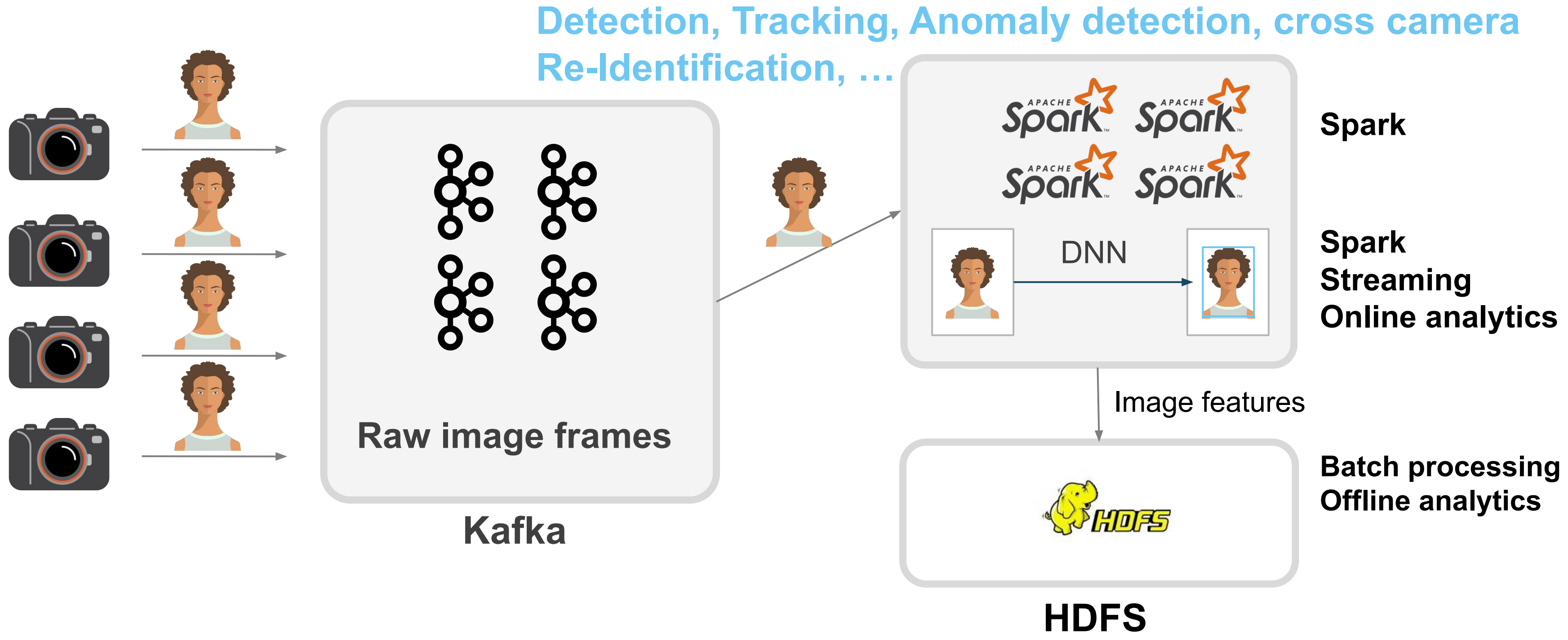


Use Cases

- How many people went from the shoe department to the jewelry department?
- How many people were observed walking around the entire building on a given day?

Requires cross-camera online and offline analytics

Traditional Architecture does not Scale



Semantic Compression with DNNs

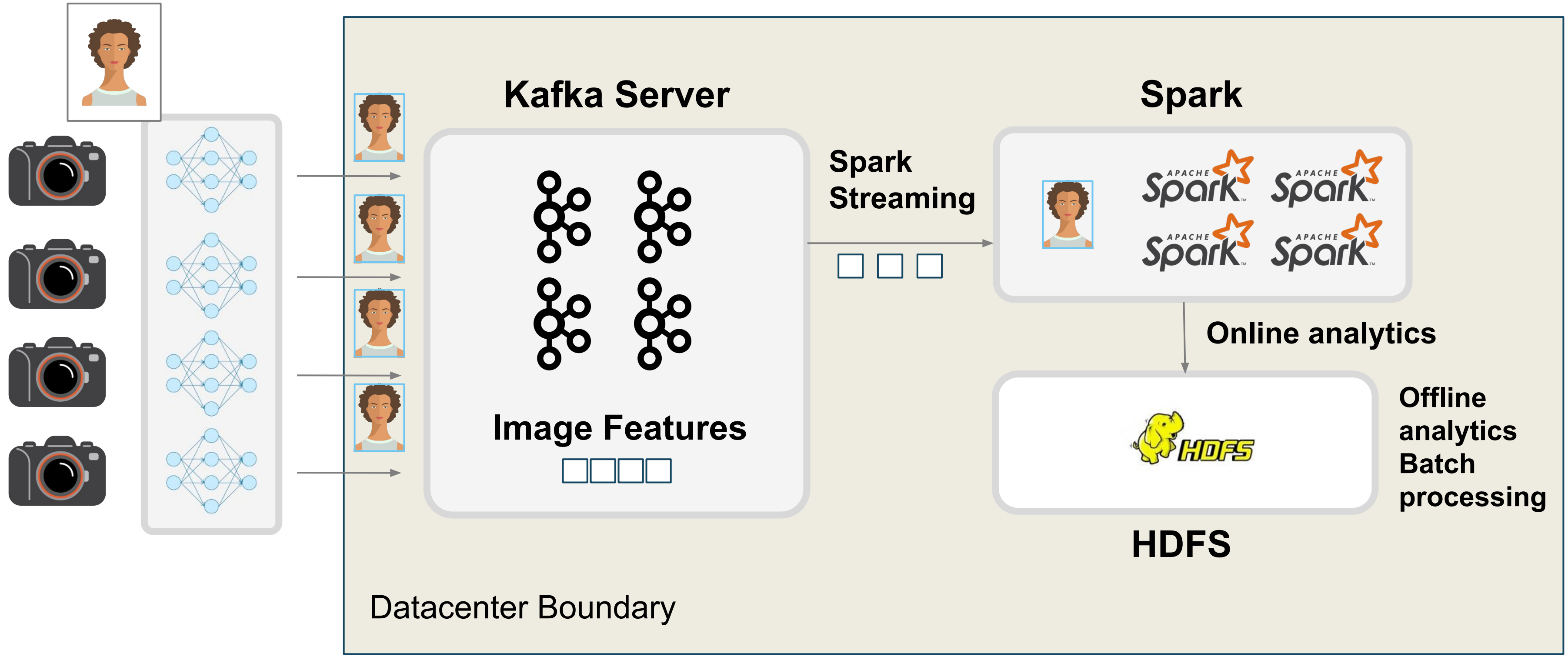
- **DNNs can be used for compression**
  - Converting raw data into condensed, semantic data
  - For video analytics, we observed a ~5x compression rate

1. https://www.gminsights.com/industry-analysis/ip-camera-market

Large Scale Image Processing

- **Deep learning on traditional big data clusters presents many challenges**
  - Computationally intensive
    - Adds pressure to the entire ETL toolchain
  - Traditional CPUs are not ideal for evaluating DNN models
  - Doing many levels of DL processing on every input frame requires
    - Storing a lot of raw data
    - Storing and managing all interim data

DNN Optimized Ingest



Challenges with DNNs

- Computationally expensive
- Require a lot of data and energy

DNNs with FPGA

- **FPGAs are a good candidate**
  - Faster than CPUs
  - More power efficient than GPUs
  - More programmable than ASICs
- **Programmability**
  - Need HDL

Solution: DNN+FPGA for Ingest

- **Ingest only the data you need**
  - Run DNNs on the edge
  - Condensed, meaningful features instead of raw, largely meaningless data
- **Accelerate with FPGAs**
  - Power efficient
  - Can be deployed with minimal infrastructure
  - Using DNNWEAVER technology for programmability
    - Compiler and full stack for automatic DNN acceleration

DnnWeaver

DnnWeaver

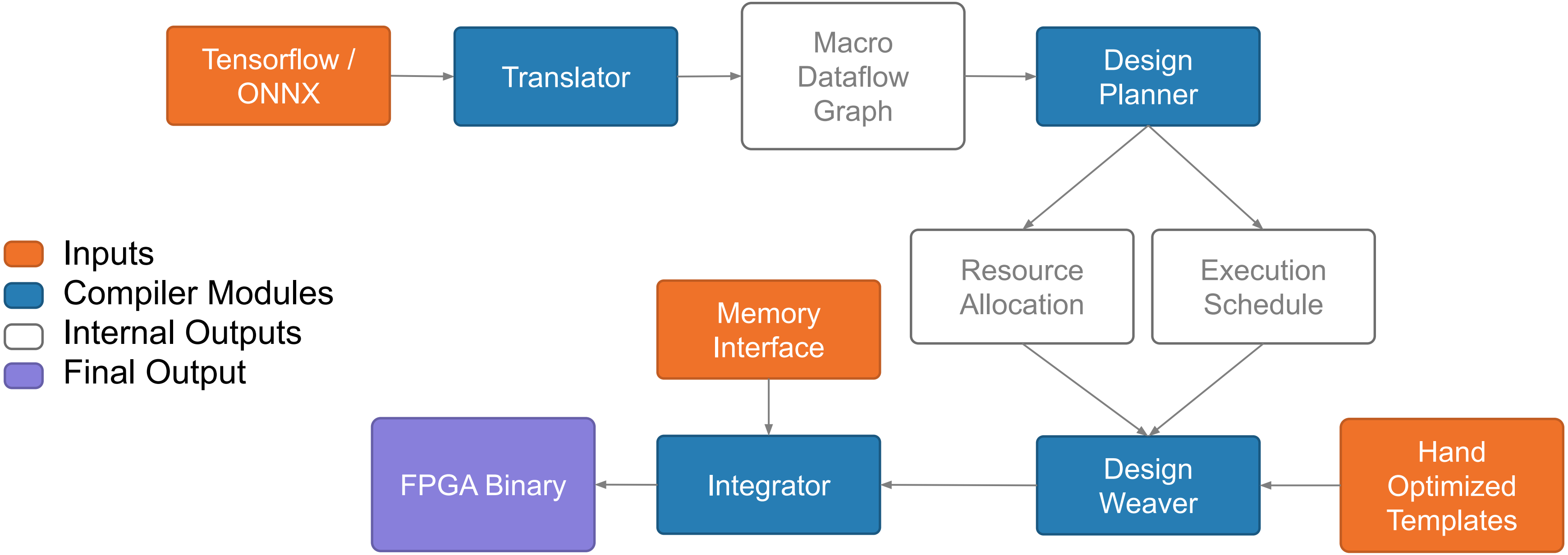
- **Ease DNN Deployment to FPGAs**
  - Tensorflow and ONNX
  - No code changes
  - No hardware expertise needed

- **Open source implementation based on original paper[1]**

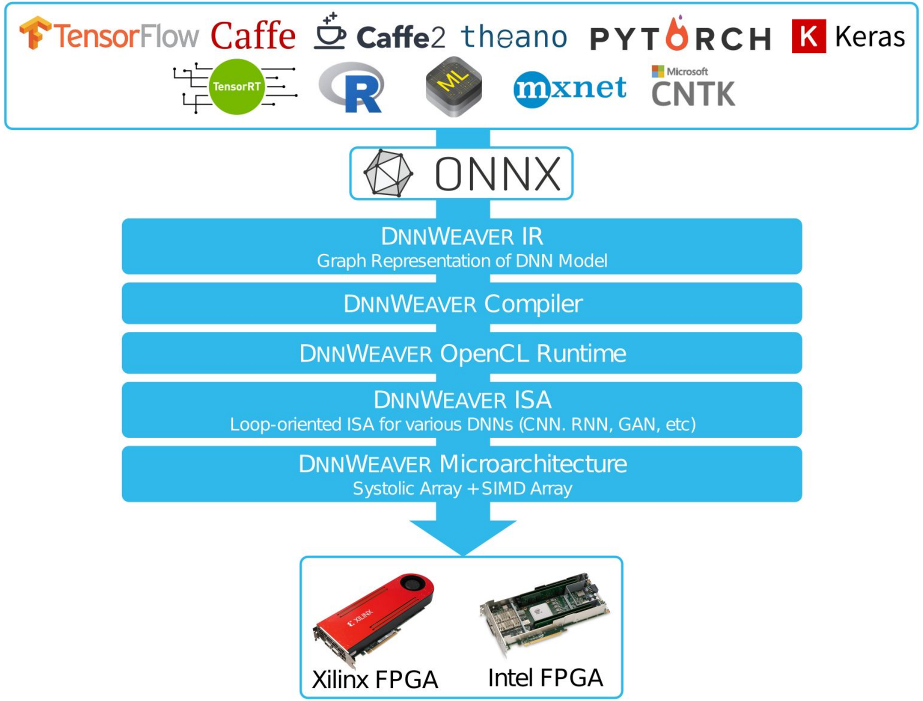
- **Enterprise version under development by Bigstream**

1: https://github.com/hsharma35/dnnweaver2

End-to-end DNN acceleration

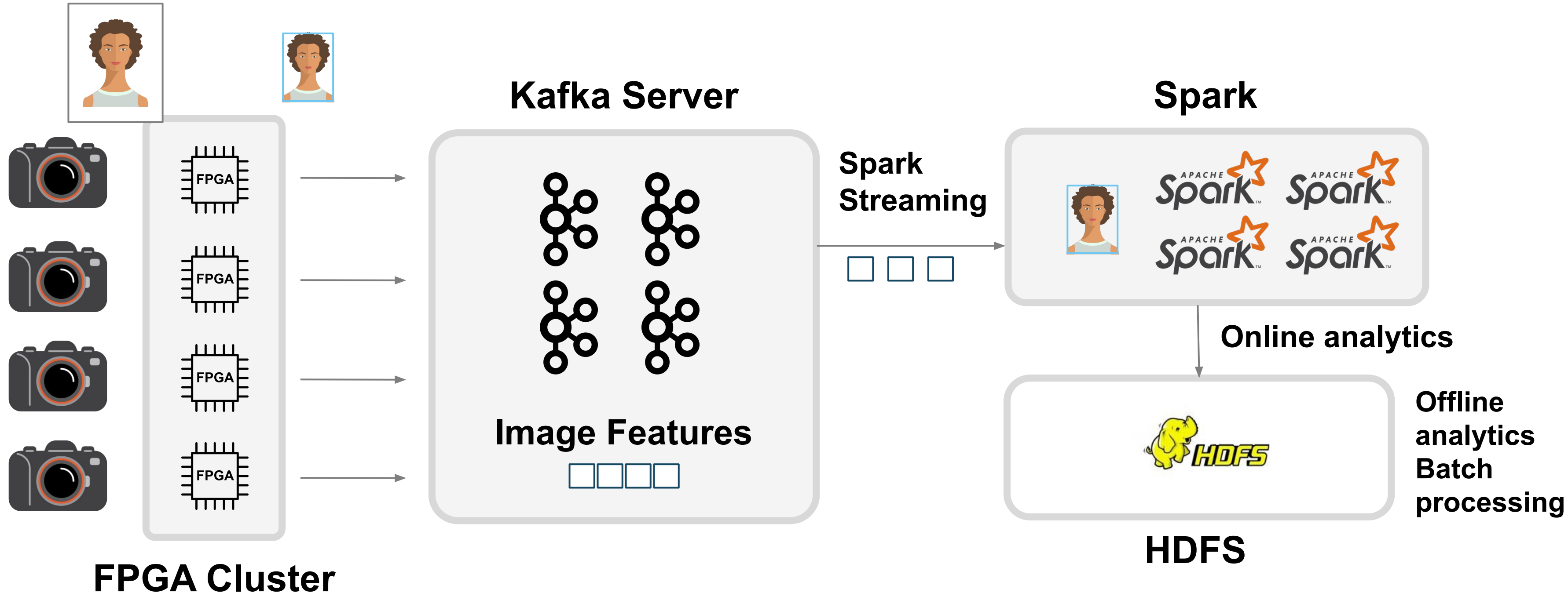


DNNWEAVER Compute Stack



SparkWeaver

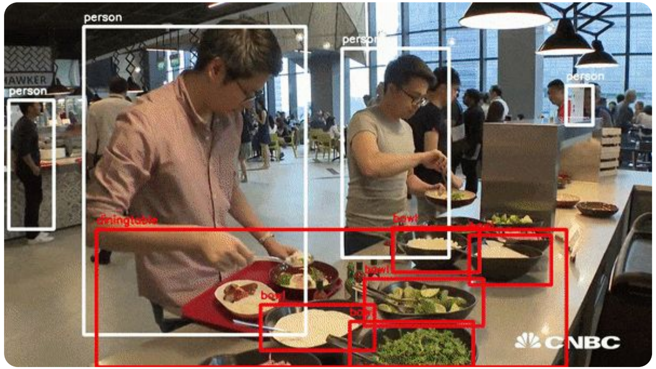
SparkWeaver Architecture



Detection and Tracking with YOLO[1] and Deep SORT[3]

YOLO → Bounding Boxes → Deep SORT → Tagged Bounding Boxes

[2]



[3]

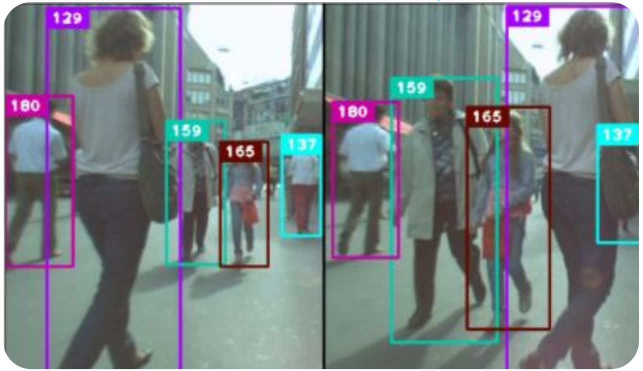
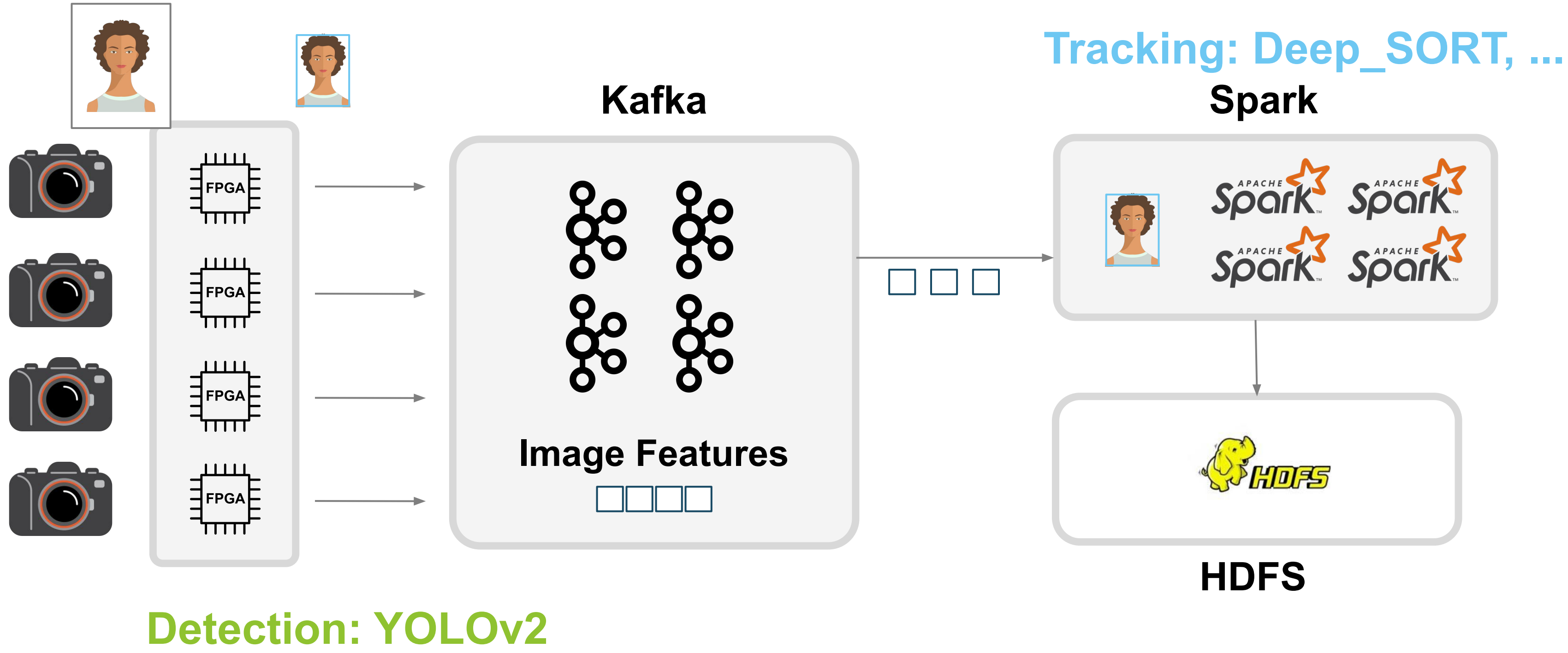


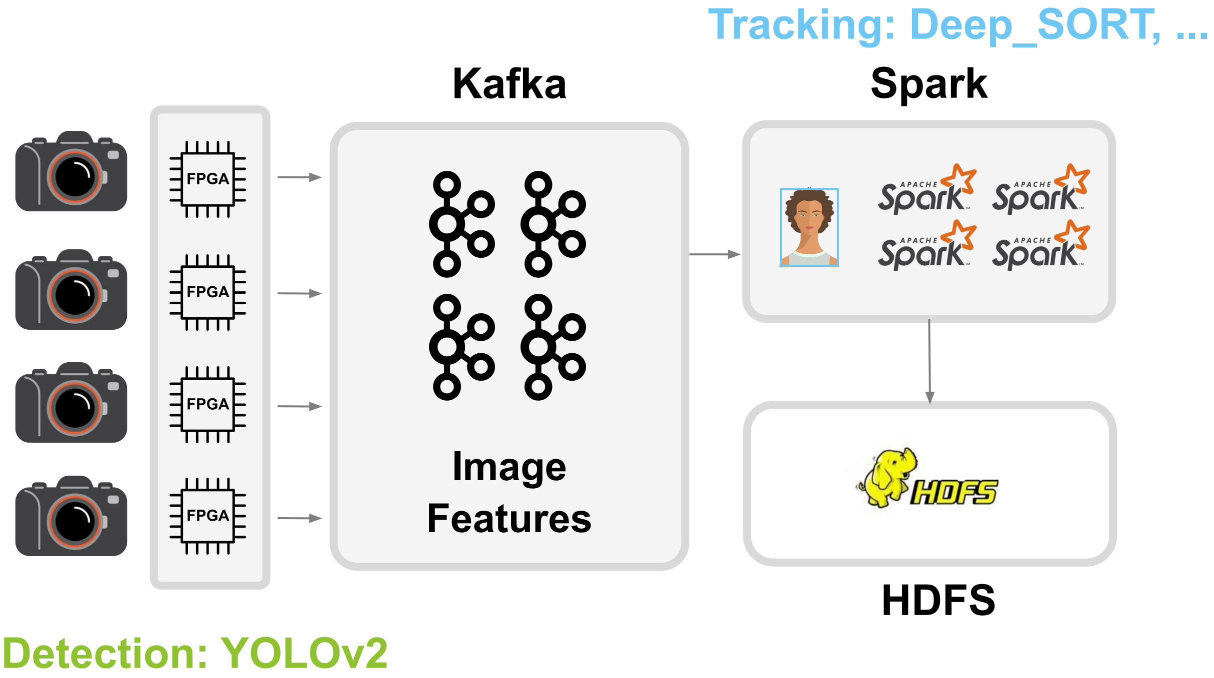
Image sources: [1] Redmon et al. “You Only Look Once, Unified, Real-Time Object Detection”
[2] https://github.com/thtrieu/darkflow
[3] Wojke et. al. “Simple Online and Real Time Tracking with a Deep Association Metric”

SparkWeaver Architecture



SparkWeaver Architecture

- Multiple cameras stream video to an FPGA cluster
- FPGA clusters implement YOLOv2 with DNNWEAVER
- YOLO's image features are streamed to a Kafka server
- Features are aggregated by Spark and written to HDFS
- Detection performed on the fog layer, tracking on the cluster



Single Node Max FPS

Benchmark	FPS
Traditional Architecture	7.3
Detection only	10
Tracking only	12.8
YOLO on DnnWeaver	13.2
SparkWeaver	12.8

*Dependent on the number of people in a frame

Next Release Max FPS (projected)

Benchmark	FPS
SparkWeaver	46.1

*Dependent on the number of people in a frame

Single Node Compression Rate

- **5.5x (82%)**
  - Deep_SORT tracking only needs the pixels within the bounding boxes, and their locations

Demo

Streaming and Batch Analytic Operations

Streaming Analysis

- **Person Re-Identification[1]**
  - Multiple solution (hot area of research)
  - Some solutions use pre-trained DNNs[2]
    - Generate a feature vector and apply similarity check on vectors

- **Anomaly Detection[3]**
  - Suspicious events/threats

1. Zheng et al “Person Re-identification: Past, Present, and Future”
2. Hermans et al “In Defense of the Triplet Loss for Person Re-Identification”
3. https://databricks.com/blog/2018/09/13/identify-suspicious-behavior-in-video-with-databricks-runtime-for-machine-learning.html

Use Case 1

How many people went from department X to department Y?

```
people_present
    person_id:  INT,
    camera_id:  INT,
    enter:  TIMESTAMP,
    exit:  TIMESTAMP
```

```
select * from people_present where camera_id == X as peopleX

select * from people_present where camera_id == Y as peopleY

count(select person_id from peopleX inner join peopleY where
        WITHIN_THRESHOLD(peopleX.exit, peopleY.enter))
```

Use Case 2

How many people were observed walking around the entire building on a given day?

```
unique_people
    id:  INT
```

select id **from** unique_people **where**
count(**select** * **from** people_present **where** camera_id == CAMERA1) > 1 **and**
count(**select** * **from** people_present **where** camera_id == CAMERA2) > 1 **and**
count(**select** * **from** people_present **where** camera_id == CAMERA3) > 1 **and**
...

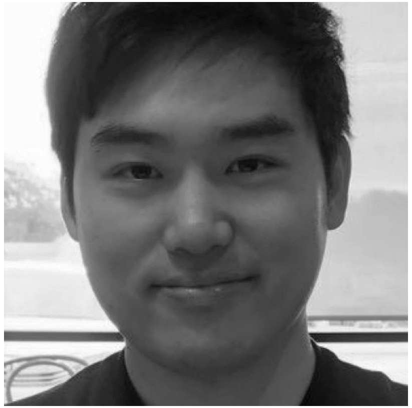
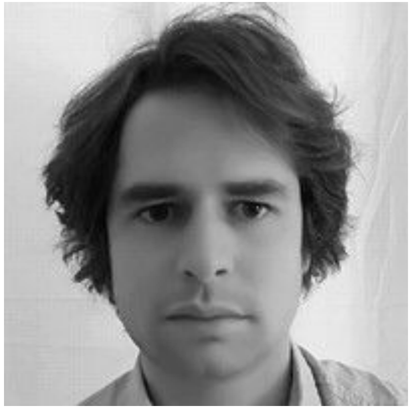
Conclusion

- **DNN-optimized ingest**
  - Smart compression
  - Use network resources on dense, highly meaningful data rather than sparse, raw data

Conclusion

- **FPGAs are well suited to DNN acceleration edge computing**
  - Highly parallel
  - Power efficient
  - Can be deployed with minimal resources
  - DNNWEAVER can compile Tensorflow/ONNX to FPGA
- **Not just DNNs, also infrastructure**
  - Online and offline analytics
  - Moving data

About the Authors

behnam@bigstream.co
jongse@bigstream.co
blake@bigstream.co

SPARK+AI
SUMMIT 2019

DON'T FORGET TO RATE AND REVIEW THE SESSIONS

SEARCH SPARK + AI SUMMIT

SPARK+AI
SUMMIT 2019

bigstream