CERN openlab Machine Learning and Data Analytics Workshop

April 29th, 2016

Machine Learning and Data Analytics at Cisco

At the edge and core of the network

Enzo Fenoglio, Hugo Latapie, JP Vasseur, Nick Hall, Trevor Smith

Overview on…

- ❖Platform for Network Data Analytics
- ❖Deep Learning for Visual Analytics
- ❖Data Fusion for Edge Computing
- ❖Self-Learning Networks

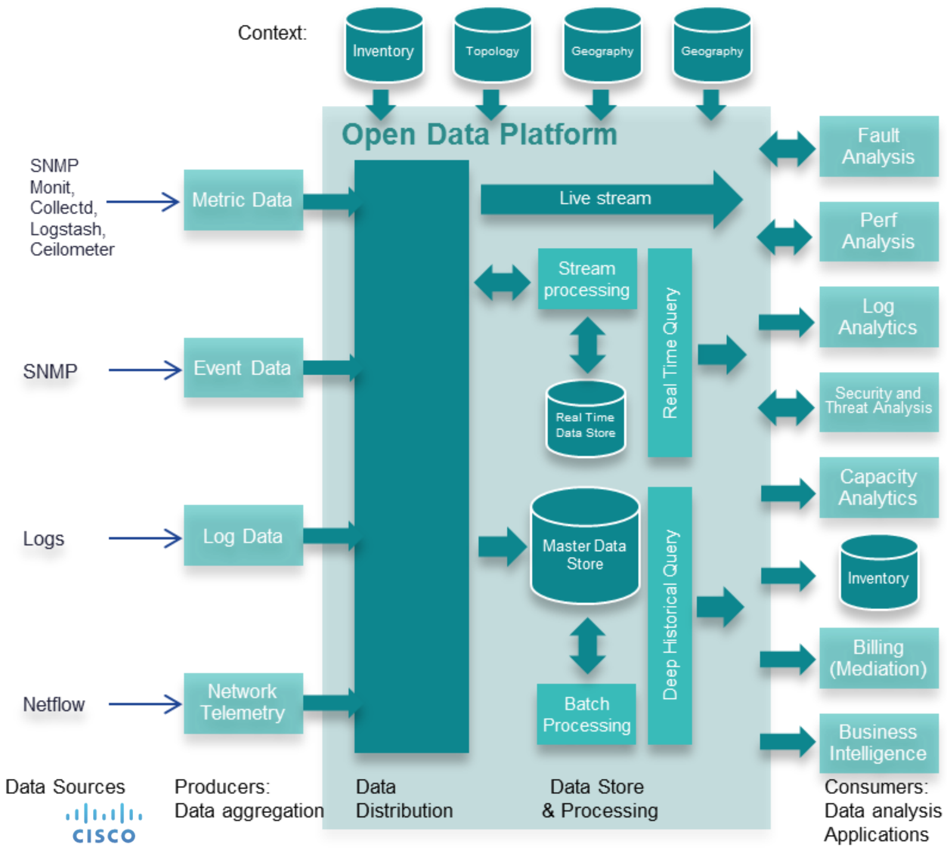
PaNDA

Platform for Network Data Analytics

Problem statement

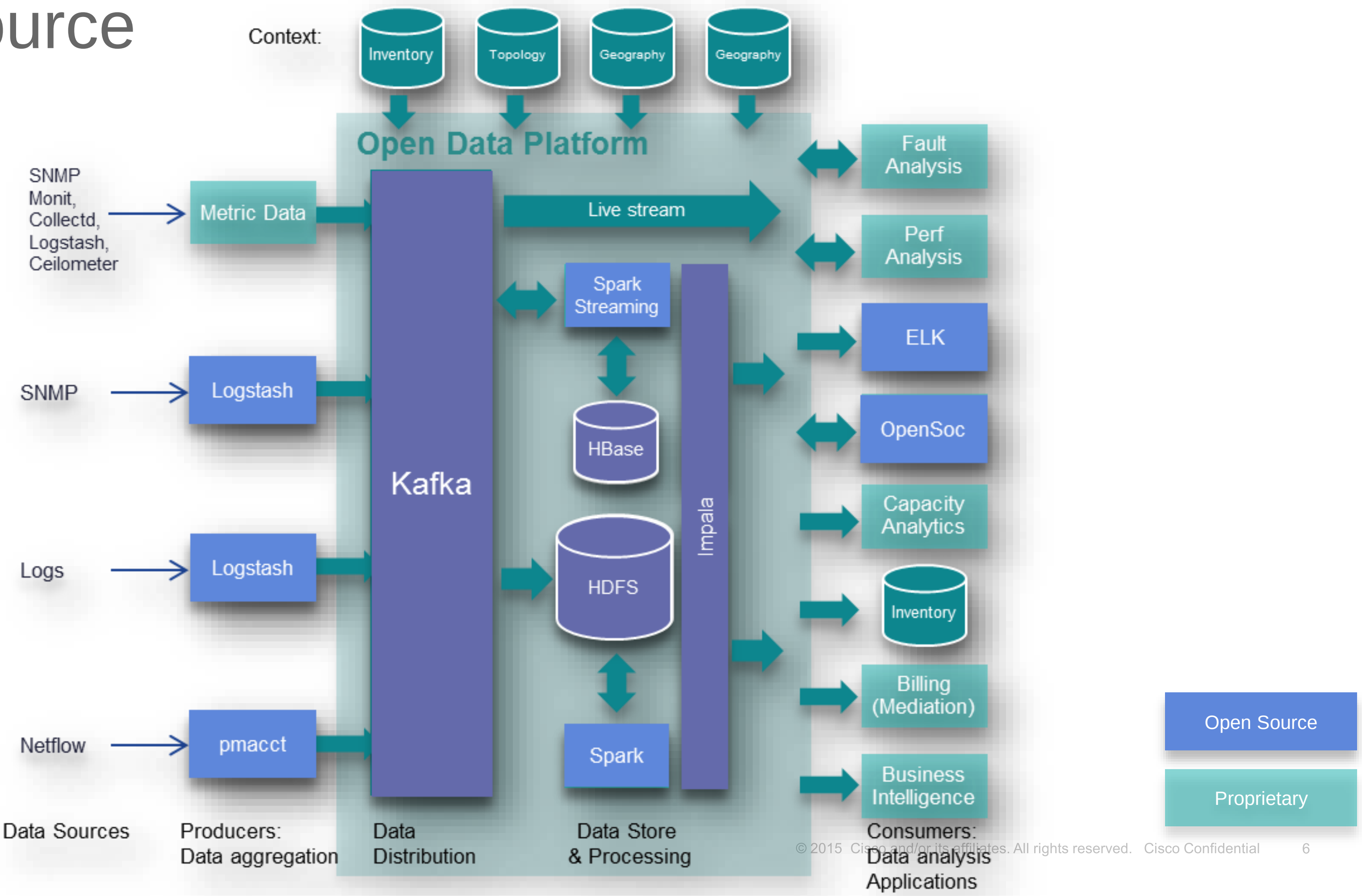
- NFV (network function virtualisation), SDN & IoT = dynamic, scalable and service-assured infrastructures on which to deploy
- These service topologies generate large quantities of system log, network flow and telemetry data
- Reduce the operational complexity for SP&Enterprise customers
  - Low-order : “Do I have service-impacting issues?”
  - Mid-order : “Will I have any service-impacting issues”
  - High-order : “Automate the control of my services”
- This, alongside existing workflows, tools and SLA’s, drives the need for an open, scalable analytics platform

Platform



- ❑ Lambda-based architecture: (batch + streaming)
- ❑ Collect data once : allow any analysis application to mine any data source
- ❑ Extensible : enable the rapid deployment of analysis functions
- ❑ Streaming (online) & batch (offline) machine intelligence
- ❑ Leverage the continual innovation in open-source data and ML community
- ❑ Contribute our work back to this community

Open-source


Context:
Inventory
Topology
Geography
Geography
Open Data Platform
SNMP
Monit,
Collectd,
Logstash,
Ceilometer
Metric Data
SNMP
Logstash
Logs
Logstash
Netflow
pmacct
Kafka
Live stream
Spark Streaming
HBase
HDFS
Spark
Impala
Fault Analysis
Perf Analysis
ELK
OpenSoc
Capacity Analytics
Inventory
Billing (Mediation)
Business Intelligence
Data Sources
Producers: Data aggregation
Data Distribution
Data Store & Processing
Consumers: Data analysis Applications
Open Source
Proprietary

Technology map

INSIGHT

Find out more about the [Insight Data Engineering Fellows Program] and [Data Labs]

Mouse over each box for more details.

Click on each technology for resources to get started.

Last updated Oct 26th, 2015
You can also find the previous version [here].

Ingestion
1. Kafka
2. Logstash
3. RabbitMQ
4. Fluentd
5. AWS Kinesis

File Format
1. Avro
2. ProtoBuf
3. Thrift
4. Parquet
5. ORC Files

File System
1. HDFS
2. AWS S3
3. Azure
4. Tachyon
5. Ceph

Batch Processing
1. Spark
2. Hadoop MapReduce
3. AWS EMR
4. Flink
5. Tez

High-Level MR
1. Pig
2. Cascading
3. Hadoop Streaming
4. Cascalog

Batch ML
1. H2O
2. Mahout
3. Spark MLlib
4. FlinkML

Batch Graph
1. GraphLab
2. Giraph
3. Spark GraphX
4. Hama

Batch SQL
1. Hive
2. Presto
3. Drill
4. Impala

Stream Processing
1. Storm
2. Spark Streaming
3. AWS Lambda
4. Samza
5. Flink

General management tools for data pipelines

Cluster Management
1. Docker
2. Zookeeper
3. YARN
4. Mesos
5. Hue

Scheduling/Monitoring
1. Luigi
2. Airflow
3. Nagios
4. Graphite
5. Azkaban
6. Zabbix

Data Store
Transactions
Analytics
Uptime Critical
Search
Graph
Geospatial
Time Series
Cache

Web Framework
1. Ruby on Rails
2. Node.js
3. Django
4. AngularJS
5. Flask

Data Visualization
1. D3
2. Tableau
3. Leaflet
4. Highcharts
5. Kibana

Platform Orchestration
1. Salt

Transactions
1. MySQL
2. Oracle
3. PostgreSQL

Analytics
1. AWS Redshift
2. Vertica
3. HBase

Uptime Critical
1. Cassandra
2. Riak
3. AWS DynamoDB

Search
1. Elasticsearch
2. Solr
3. MongoDB

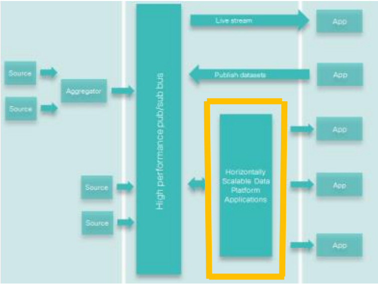
Graph
1. Neo4j
2. OrientDB
3. ArangoDB

Geospatial
1. CouchDB
2. PostGIS
3. Elasticsearch

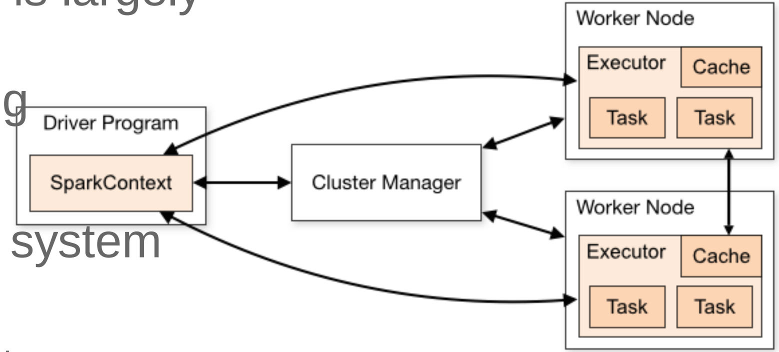
Time Series
1. InfluxDB
2. Cassandra
3. Druid
4. OpenTSDB

Cache
1. Redis
2. Memcached
3. Hazelcast

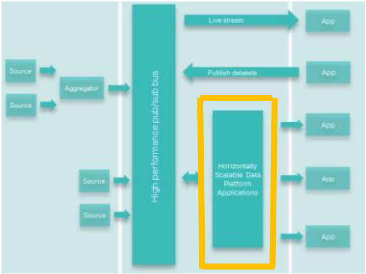
Massively-parallel batch processing



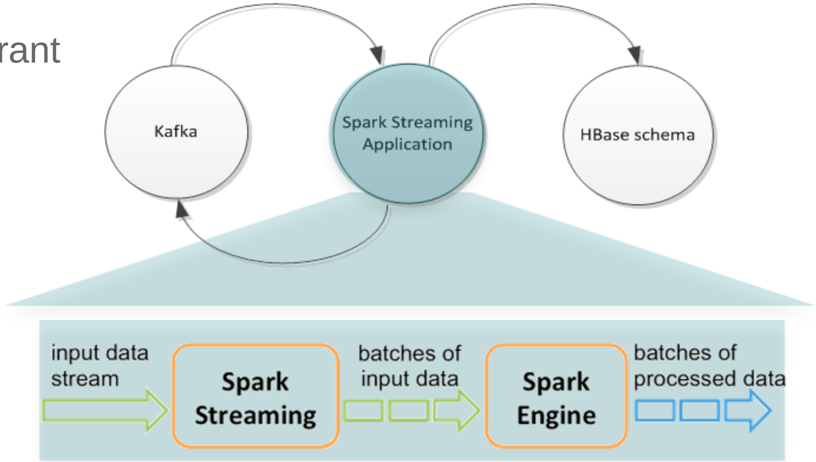
- Support for applications that deliver computations over very large datasets with highly heterogeneous structure
  - Addresses data volume & variety
- **Apache Spark** (spark.apache.org)
  - Framework and engine for distributed, large scale data processing
  - Many times faster than MapReduce, which it is largely replacing in industry
  - Also provides engine behind Spark Streaming
- **HDFS** (hadoop.apache.org)
  - Fault tolerant and self-healing distributed file system
  - Large-scale data processing workloads
  - Focus on scalability, flexibility and throughput
  - Proven deployments of >100PB



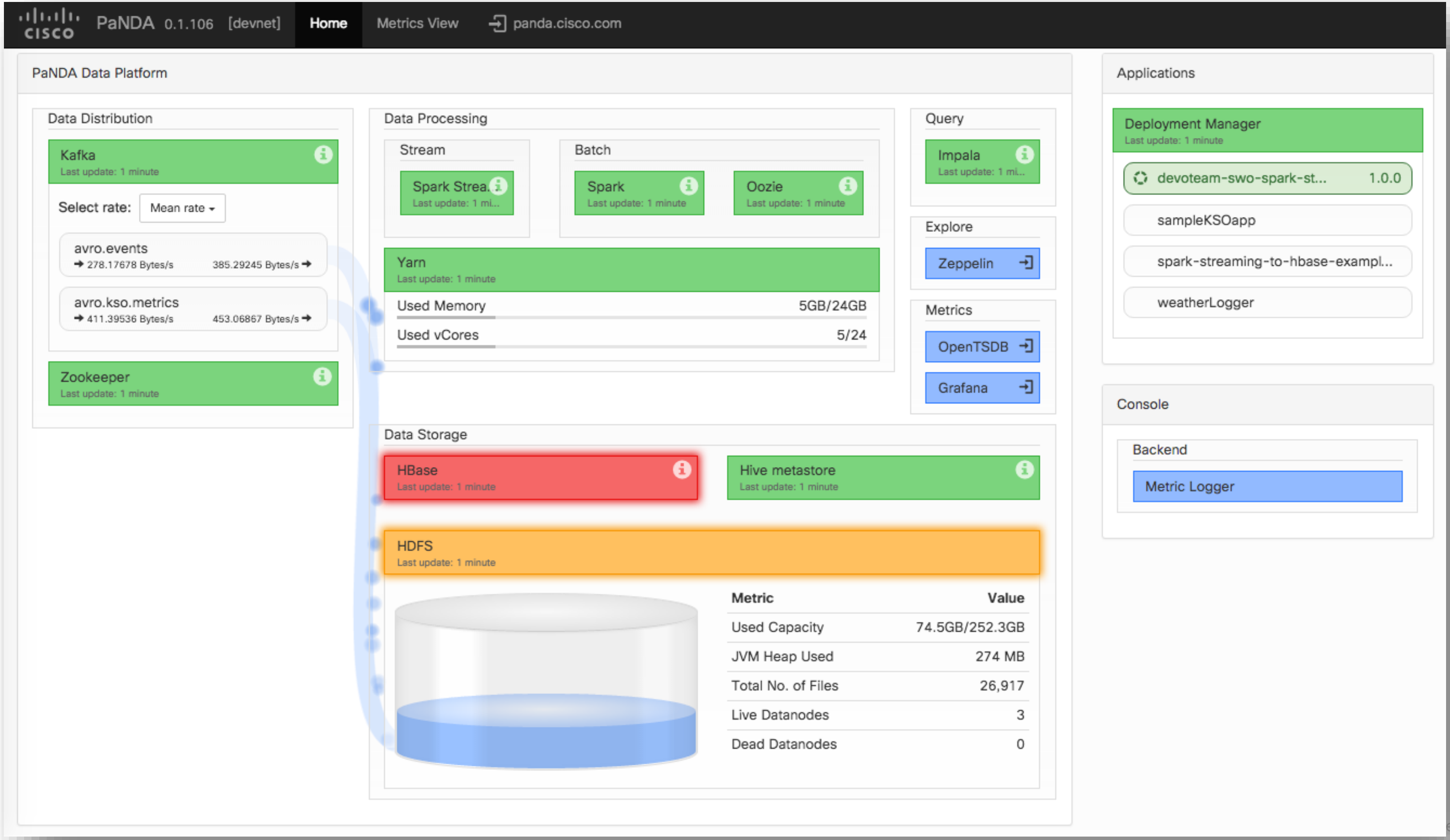
Stream processing



- Support for applications that need to deliver computations over data in near real time (e.g. 1s)
  - Addresses data velocity
- **Apache Spark Streaming** (spark.apache.org)
  - Framework and engine for distributed, scalable fault-tolerant streaming applications
  - Micro-batch orientation
  - Consume/produce to/from Kafka
- **Apache HBase** (hbase.apache.org)
  - Distributed, scalable data store
  - Designed for fast, random access to very large data sets
    - e.g. billions of rows and millions of columns
  - Persists results of streaming computation in optimized schema



Platform management



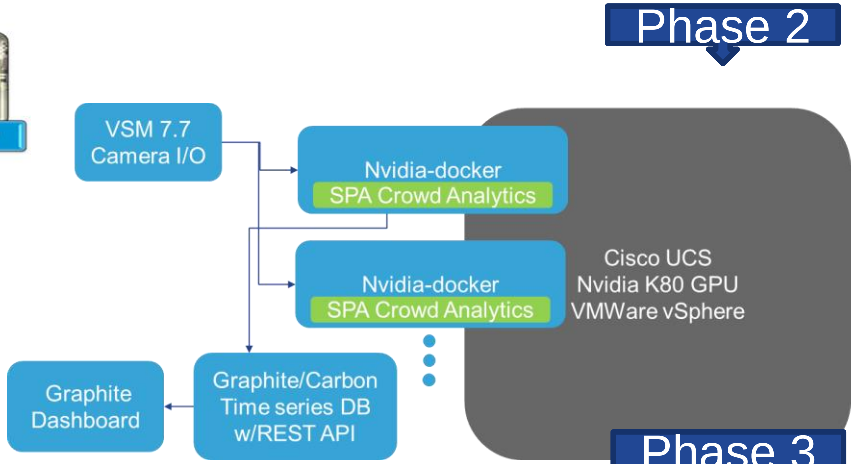
panda.cisco.com

Deep Learning for Visual Analytics

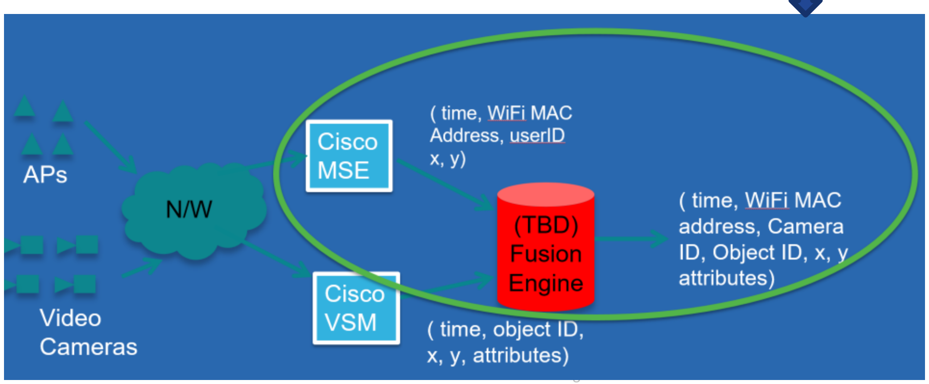
Spatial Predictive Analytics: Problem to solve

- Cisco customers ask for accurate people counting, location, tracking, and more advanced crowd analytics
- There are already practical solutions for the sparse target case but **not** for the dense target case
- Outstanding crowd analytics using only CV for real-world PoC is difficult. Clutter and occlusions add even more complexities.

Spatial Predictive Analytics Overview



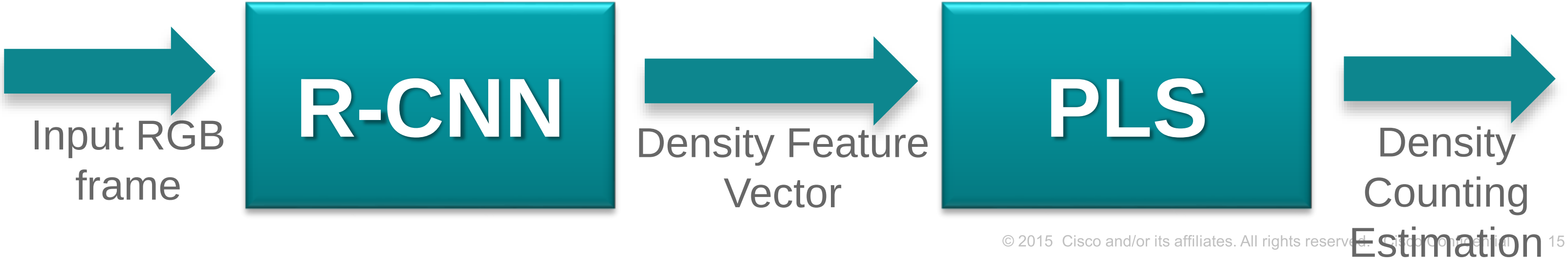
- General DL pipeline based on **docker** overlay networking and **swarm** with support for ~10^3 nodes and ~10^4 containers
- High performance video pipeline based on **gstreamer**



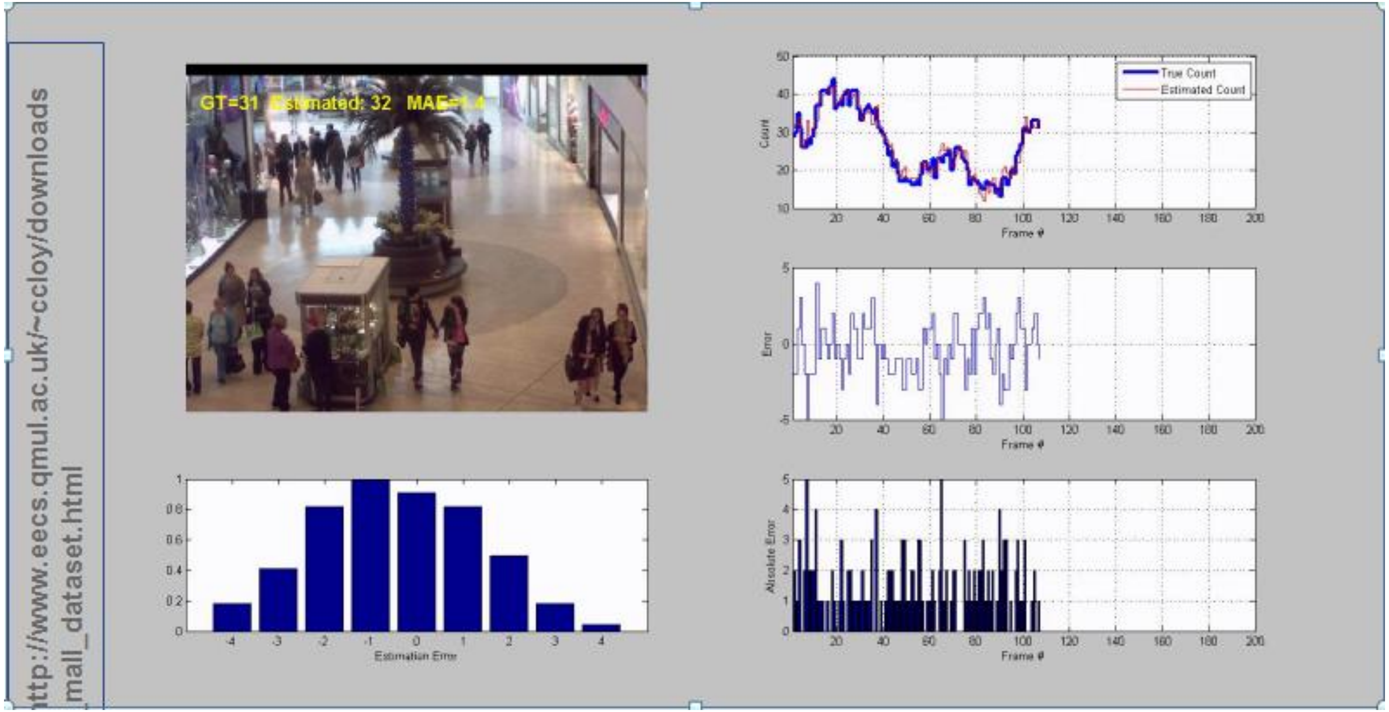
cisco

Spatial Predictive Analytics – Crowd Counting

- We aim at counting individuals in crowded environments
- Traditionally the number of people in region of interest is inferred by (1) person-counting sensors, (2) special purpose top view cameras
- We use Region-based Convolutional architecture (R-CNN) and multivariate regression (PLS) for density counting estimation



	MAE	MSE
Chen [2] (BMVC2012)	3.59	19.0
Chen [3] (CVPR2013)	3.43	17.7
PLS (with original HLAC)	3.82	22.3
SVR (with proposed feature)	3.41	18.5
Proposed method	1.8	7.5



Spatial Predictive Analytics – multi-tracking

- We aim at tracking multiple person once localized in a frame
  1. estimate the spatio-temporal position of each person in each frame using a DL pipeline: CNN (GoogLeNet)+LSTM
  2. assign UIDs to formed trajectories: Kalman filter + Jonker-Volgenant to solve the assignment problem

Training

- ❑ **Ground-Truth generation is a pain**

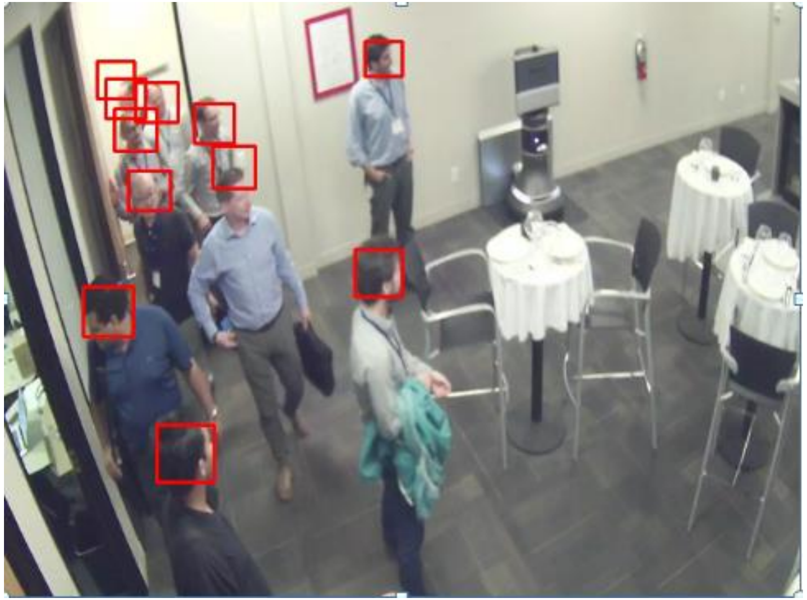
- ❑ **Transfer Learning : use previously trained model to perform inductive learning**

- ❑ **Semi-supervised learning : partially labelled dataset reduces the training iterations and improves accuracy**

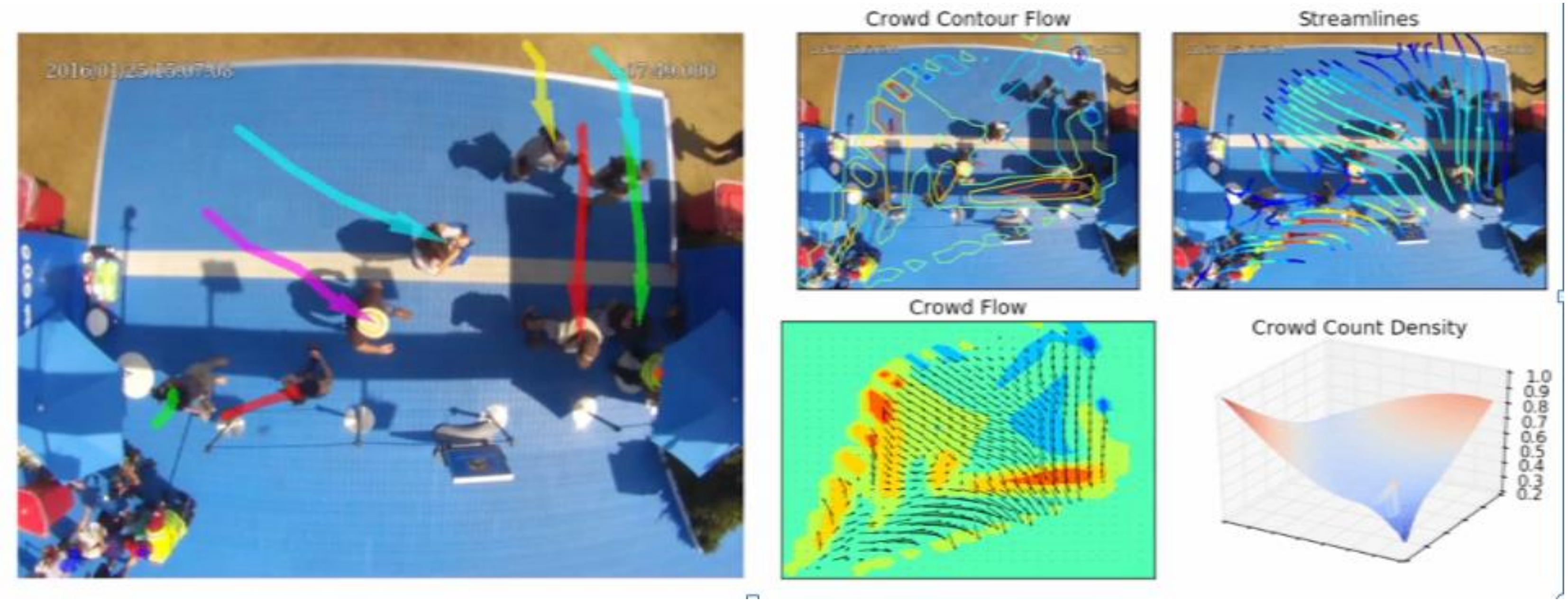
Spatial Predictive Analytics -- Location

Open Source Retail Mall Dataset

Cisco Lab in San Jose



Spatial Predictive Analytics -- Tracking

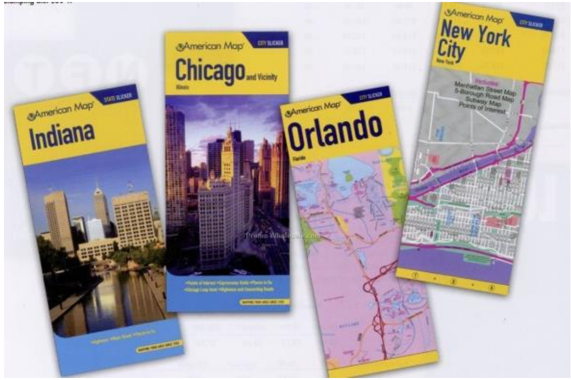


Near field DL pipeline

Far field DL pipeline

Data Fusion for Edge Computing

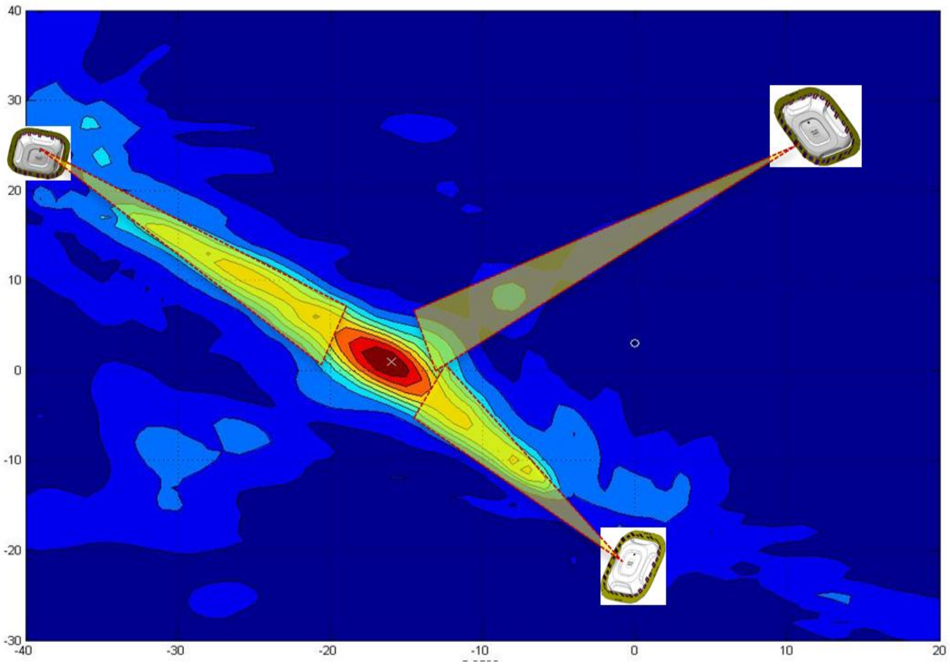
Evolution of location use-cases



Healthcare	**Asset management & wayfinding**
Retail	**Engage shopper in aisle & deliver proximity-based offers**
Museum	**Enabling the Digital docent**
Office	**Wayfinding & workspace optimization**

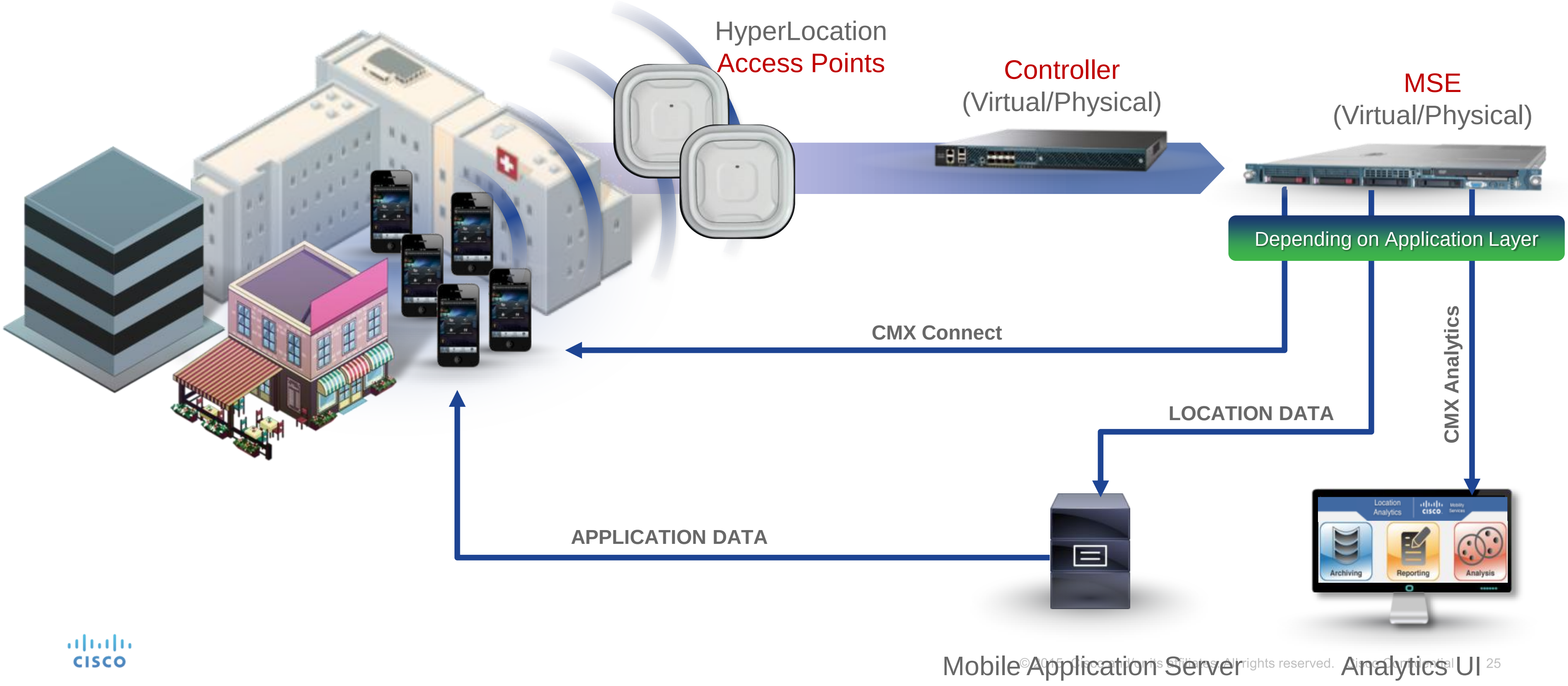
Technology – network based location

- **State of the art:**
  - 5-7m accuracy
    - Multi-lateration on WiFi Client based on RSSI at multiple APs
- **Cisco-Hyperlocation:**
  - 1m accuracy
    - Increase accuracy & reliability
    - AP connected clients
    - Add Angle-of-Arrival in addition to RSSI



How Hyperlocation Works

Built on Cisco Unified Access

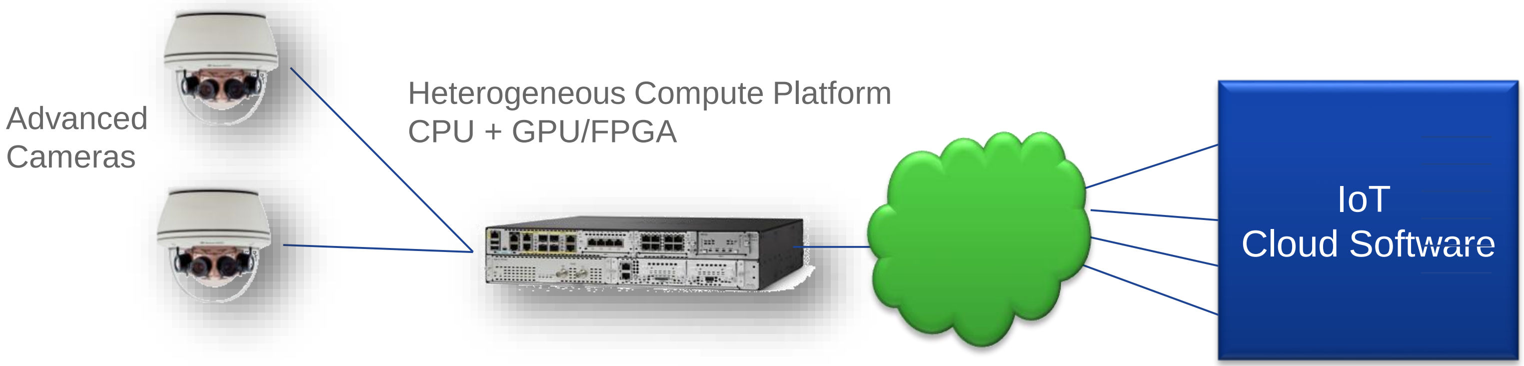


Machine Learning from End Point to Cloud

An intersection of technology and trends:

- Recent camera sensor technologies offer advanced capabilities
- Advanced cameras + analytics create powerful IoT sensors
- Visual analytics and data fusion fit naturally into the fog architecture

The industry is moving towards scalable, flexible, distributed analytics platforms



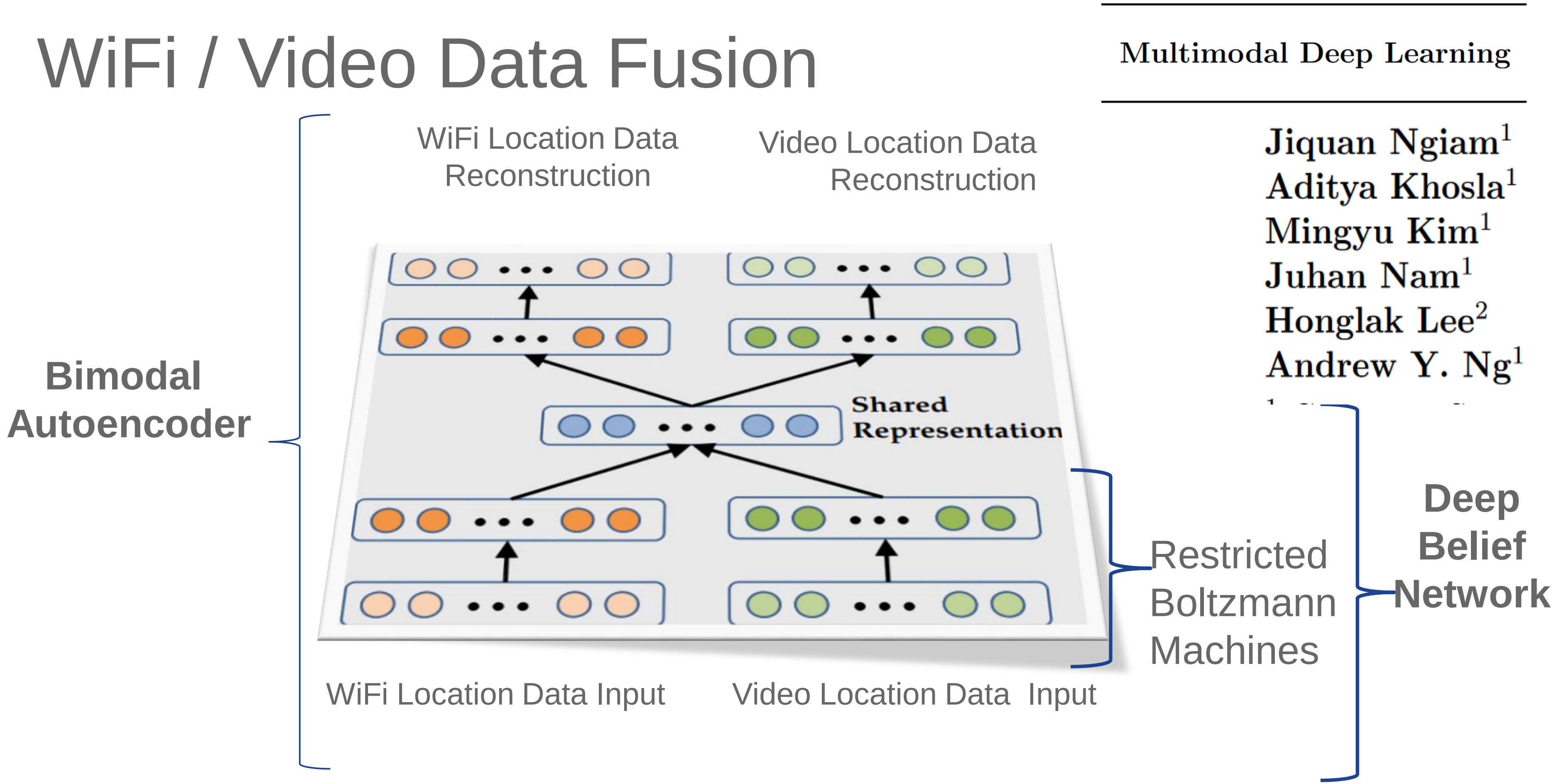
Cisco Data Center Portfolio and NVidia

- **Virtual workstation for high-end graphics applications**
  - Cisco UCS C240 M4 Rack servers
    - Support NVIDIA GRID 1.0 and 2.0 with K1/K2 and M60 cards
    - Support for Magma ExpressBox for higher density
  - Cisco UCS B200 M4 Blade servers
    - Recently introduced NVIDIA M6 MXM support on Blade server

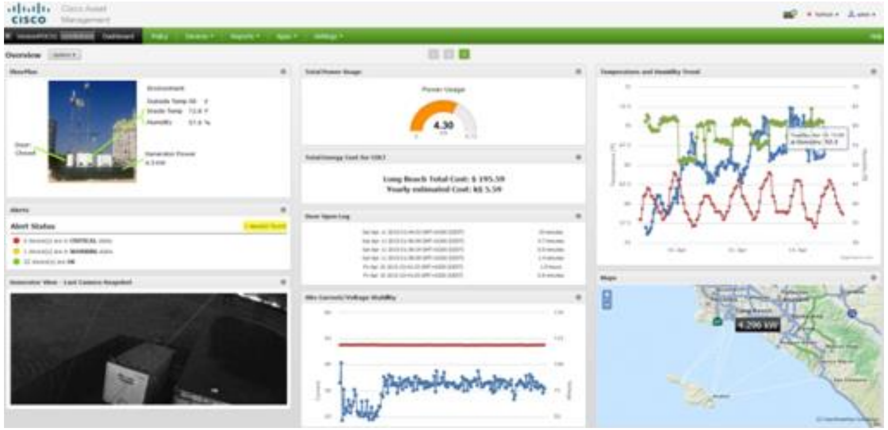
- **Cisco HyperFlex –2nd generation HyperConverged platform**
    - Phase 1: K1/K2 support
    - Phase 2: M6/M60 support

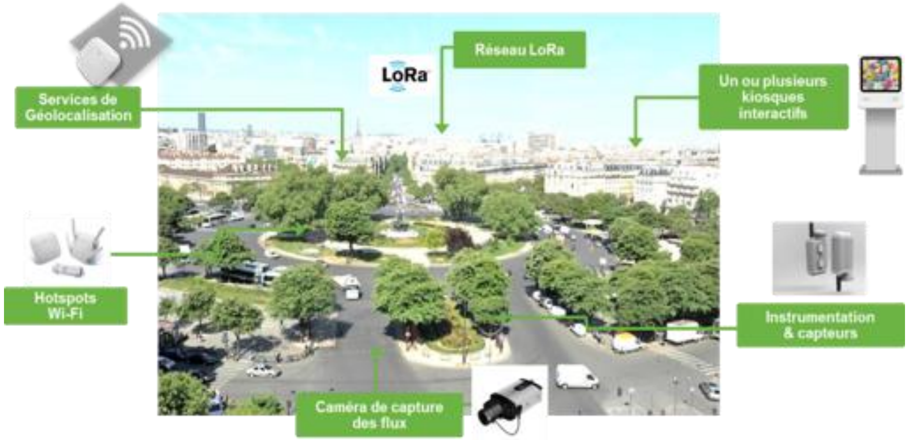
- **Deep Learning and HPC**
  - Cisco UCS C240 Rack servers with TESLA K80

WiFi / Video Data Fusion



Smart Cities: Innovative pilot proposed by the City of Paris





1. **Optimization of energy management in public buildings**

2. **Better Understanding of the public space**

CISCO — CDP, MERAKI wifi, CMX, LoRaWAN, cameras, video analytics, Energy / Asset Management

VINCI ENERGIES — axians — citeos

relayr bring things to life — BH technologies - éclairage — Bruitparif Observatoire du bruit — placemeter — JCDecaux — actility Making Things Smarter — Breezometer — OpenDataSoft

DataCity PROGRAMME D'EXPÉRIMENTATION POUR IMAGINER NOS VILLES DE DEMAIN — NUMA — qucit

Self-Learning Networks

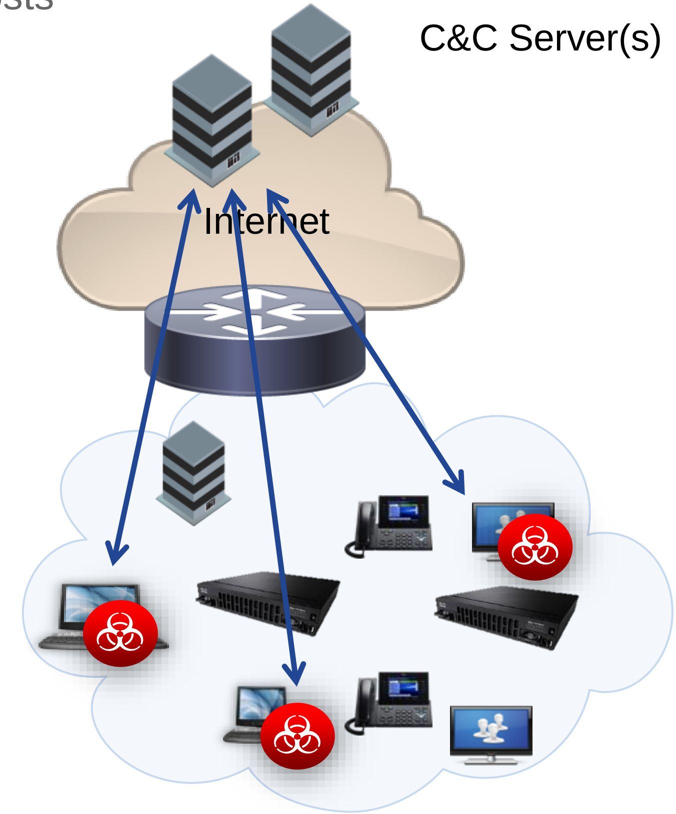
What Self Learning Networks is About ...

- SLN is fundamentally a hyper-distributed analytics platform at the edge
- Putting together analytics and networking
  - Goldmine of untouched data on networking gear (***sensing***)
  - Network learns and computes models on premise (***analytics***)
  - The Network adapts, modifies its behavior (***control***)
- SLN *for* Security: attacks are incredibly sophisticated and targeted, ex-filtration of data being a major concern, requiring a next-generation approach.
- True Technology disruption

31

Botnets and Data Ex-Filtration Techniques

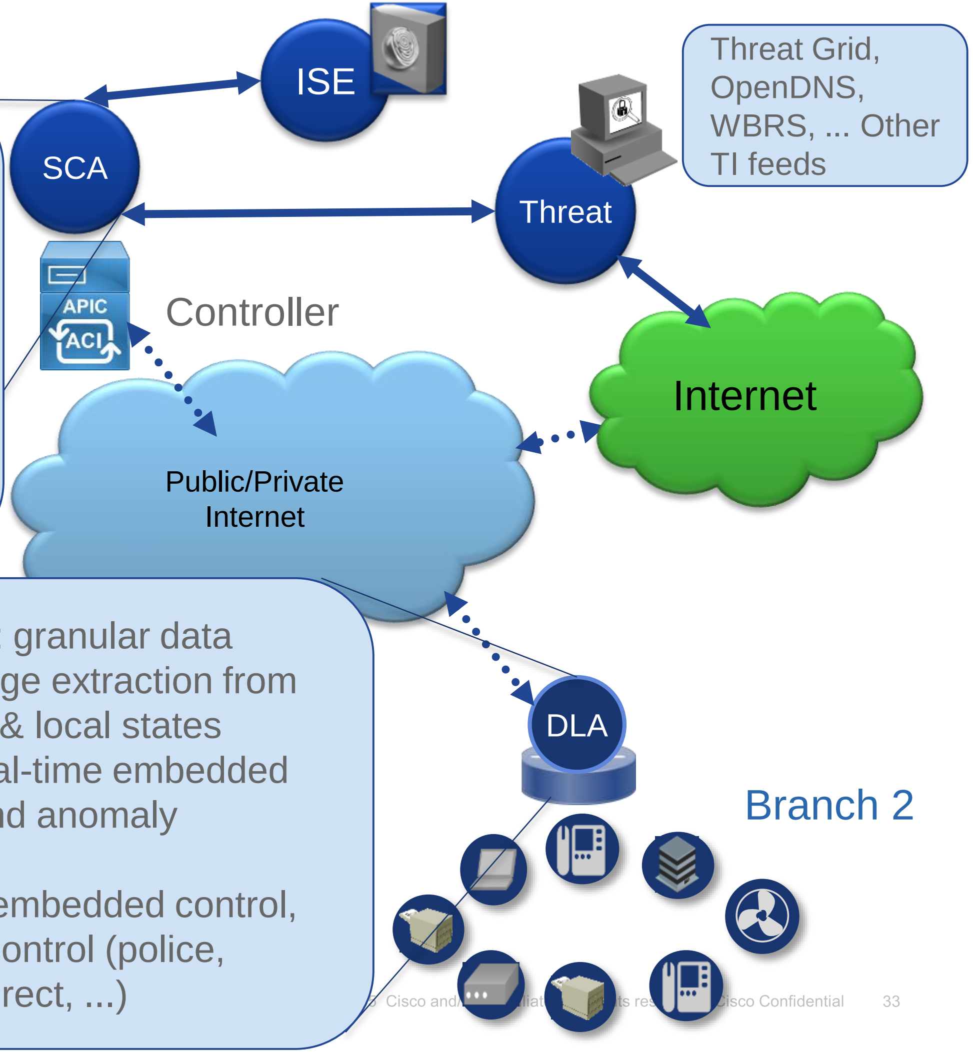
- Size can range from thousands to millions of compromised hosts
- Botnet can cause DDoS & other malicious traffic (spam, ...) to originate from the inside of the corporate network
- C&C (C2) servers become increasingly evasive
  - Fast Flux Service Networks (FFSN), single or double Flux
  - DGA-based malware (Domain Generation Algorithms)
  - DNS/NTP Tunneling
  - Peer-to-Peer (P2P) protocols
  - Anonymized services (Tor)
  - Steganography, potentially combined with Cryptography
  - Social media updates or email messages
  - Mixed protocols ....
  - Timing Channels



SLN Architecture

SCA

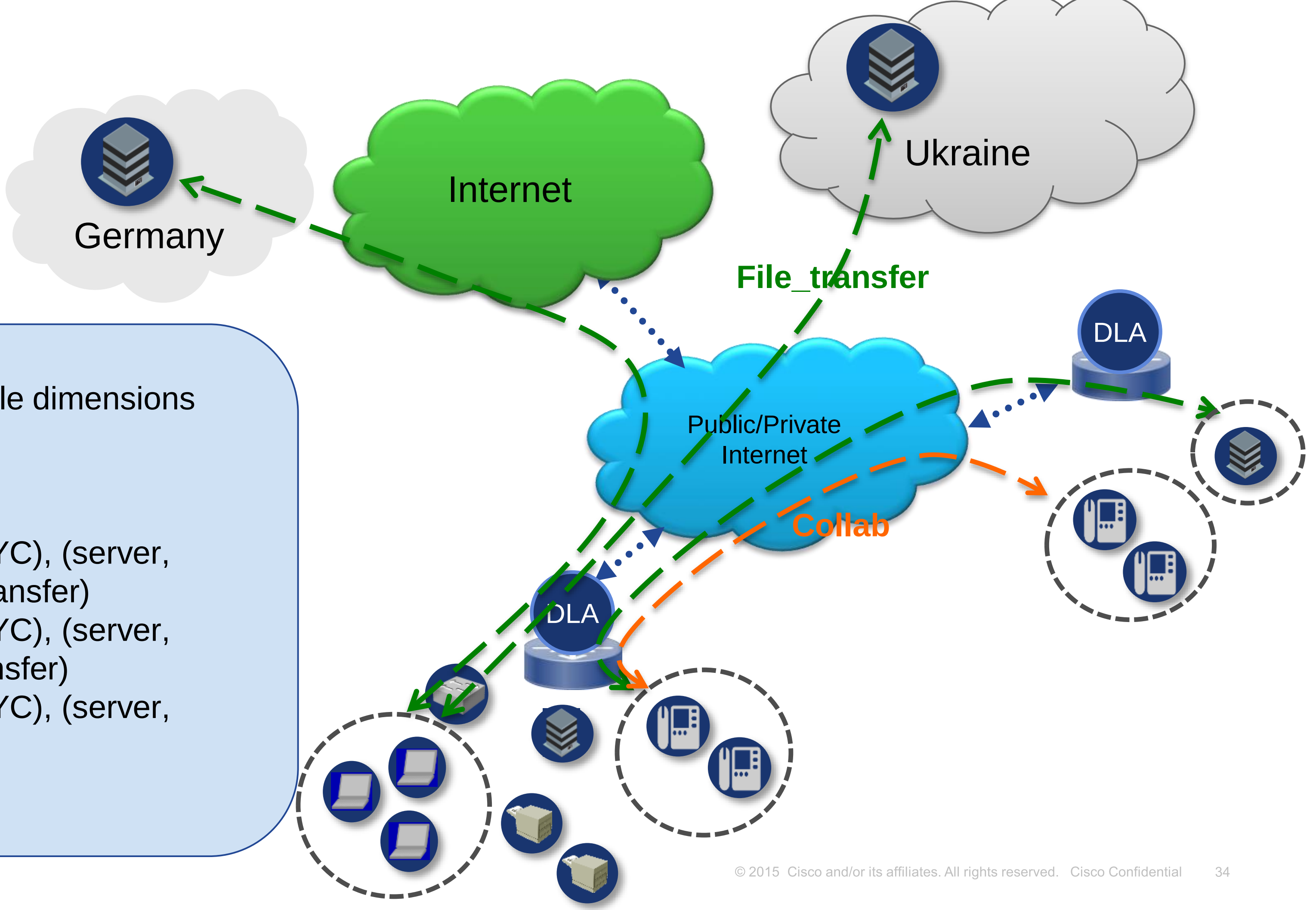
- Orchestration of Distributed Learning Agents (DLAs)
  Advanced Visualization of anomalies
  Centralized policy for mitigation Interaction with other security components.
  North bound API to SIEM/Database (e.g.Splunk) using CEF/CIM format
- Evaluation of anomaly relevance



DLA

- **Sensing (knowledge)**: granular data collection with knowledge extraction from control and data plane & local states
  Machine Learning: real-time embedded behavioral modeling and anomaly detection
- **Control**: autonomous embedded control, advanced networking control (police, shaper, recoloring, redirect, ...)

ML Models



Models use multiple dimensions and hierarchy

Example:
M1: {(windows; NYC), (server, Germany), File_transfer)
M2: {(windows; NYC), (server, Ukraine), File_transfer)
M3: {(windows; NYC), (server, Germany), Shell)
....