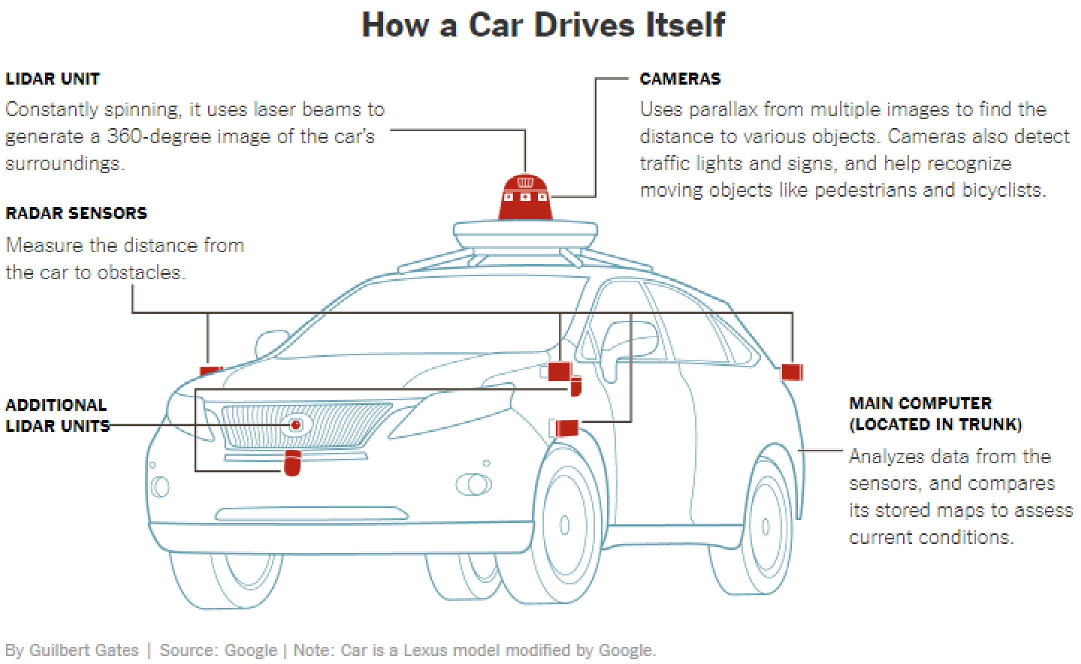
SPARK+AI
SUMMIT 2018

Accelerating Inference in the Data Center

Dr. Malini Bhandaru & Karol Zalewski
Contributors: Santiago Mok, Konrad Kurdej,
Sundar Nadathur, Alexander Kanevskiy, Ismo Puustinen
Intel

#HWCSAIS11

Autonomous Vehicles – R & D Data Pressure



How a Car Drives Itself

LIDAR UNIT

Constantly spinning, it uses laser beams to generate a 360-degree image of the car's surroundings.

RADAR SENSORS

Measure the distance from the car to obstacles.

CAMERAS

Uses parallax from multiple images to find the distance to various objects. Cameras also detect traffic lights and signs, and help recognize moving objects like pedestrians and bicyclists.

ADDITIONAL LIDAR UNITS

MAIN COMPUTER (LOCATED IN TRUNK)

Analyzes data from the sensors, and compares its stored maps to assess current conditions.

By Guilbert Gates | Source: Google | Note: Car is a Lexus model modified by Google.

Image Credit: https://clepa.eu/mediaroom/autonomous-vehicles-will-drive-change-auto-manufacturing-insurance/
https://ia.acs.org.au/article/2017/who-should-the-driverless-car-kill-.html

1-20 TB/car/hour
# cameras, resolution, other sensor arrays

Inference Everywhere Faster Please!

- Speed ground truth generation
  - Human improves upon automated
- Speed Privacy transformations
  - Face/license plate obscurring
- Speed simulation
  - Detect (edge-ish), Plan, Act

https://medium.com/@xslittlegrass/self-driving-car-in-a-simulator-with-a-tiny-neural-network-13d33b871234

Compute Continuum

CPUs
Flexibile,Slower

GPUs
FPGAs,
Movidius

ASICs
Fixed, Faster

Can Spark Leverage? Easily?

FPGA

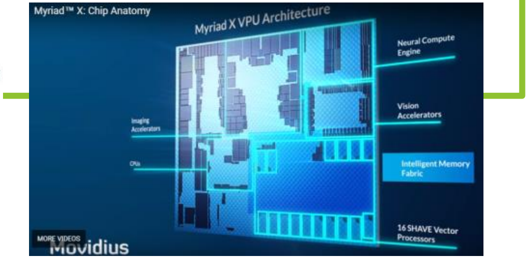
- Logic blocks, memory, security, variable sizes
- Programmable, OpenCL
- Fast but Expensive
- Applications: Networking, Telecommunication, Research, Machine Learning

A Stratix IV FPGA from Altera

A Spartan FPGA from Xilinx

Movidus Chip

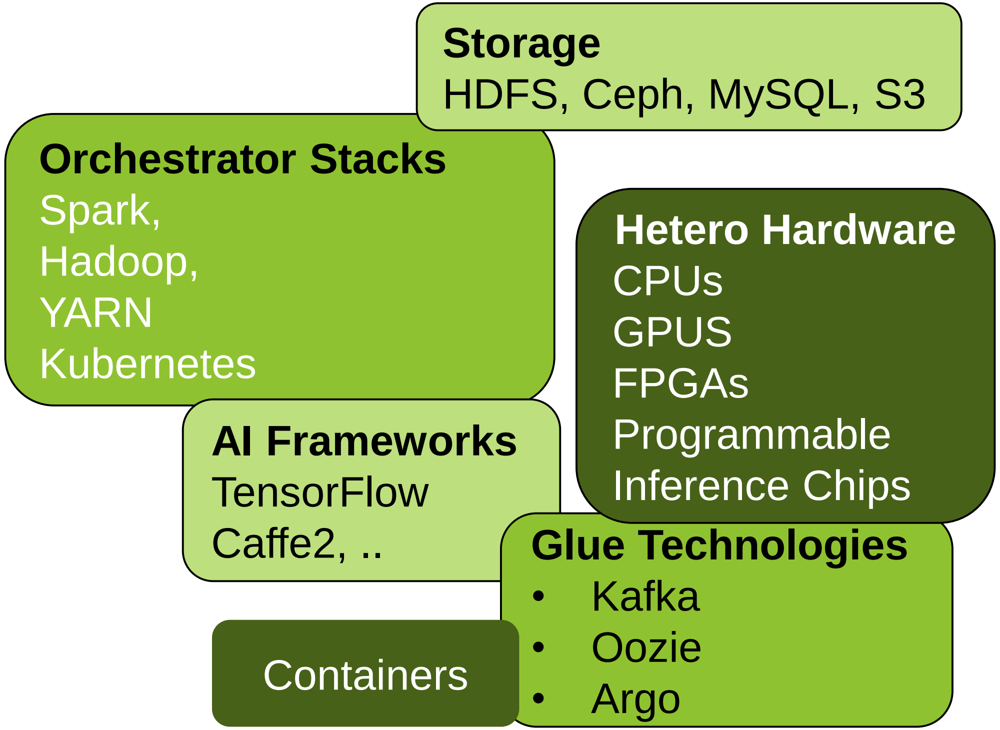
- Programmable, SDK
- Low Power
- Tuned for image processing
- Fast, Inexpensive
- Applications: Drones, Cameras, Augmented Reality



Data Center Platform

Drivers

- Fungible
- Dynamic
- Resilient
- Easy to Use
- Fast



Storage
HDFS, Ceph, MySQL, S3

Orchestrator Stacks
Spark,
Hadoop,
YARN
Kubernetes

Hetero Hardware
CPUs
GPUS
FPGAs
Programmable
Inference Chips

AI Frameworks
TensorFlow
Caffe2, ..

Glue Technologies

- Kafka
- Oozie
- Argo

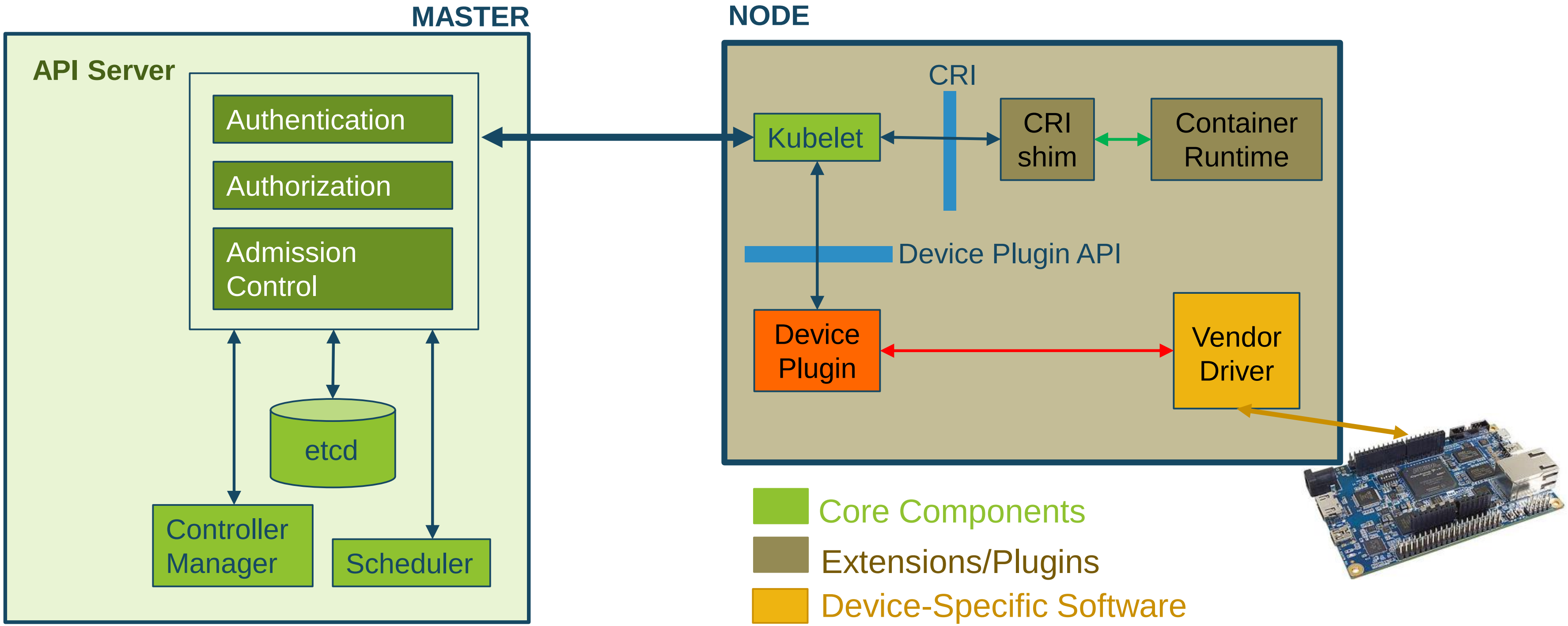
Containers

Environment

- Kubernetes – resilient, auto scaling, easy to use
- Spark – big data in memory processing, possible data locality

Kubernetes Device Plugin

Enables use of new Resources



🟩 Core Components

🟫 Extensions/Plugins

🟨 Device-Specific Software

🟧 New Work for Device Plugins

https://kubernetes.io/docs/concepts/extend-kubernetes/compute-storage-net/device-plugins/

Experiment

- SqueezeNet 1.1
- gRPC calls

  3-4 ms
- Data pre-Processing

  16-30 ms

FPGA Inference

- Model Size
- FPGA Size
- Trade-off
  - Model accuracy, speed
  - Compile to target hardware

Supported
Deep Learning Topologies

- AlexNet
- GoogleNet v1
- VGG-16 & VCG-19
- SqueezeNet 1.0 & 1.1
- ResNet-18
- SqueezeNet-based variant of SSD
- GoogleNet-based variant of SSD
- VGG-based variant of SSD

Movidius USB Learnings & Workarounds

USB
- Access to host network (isolation loss)
  `--net=host`
- Visibilibility into Device Manager events in Docker environment
  `libusb`
- Privilege Escalation (insecure)
  `--privileged`
- Access to Virtual File System to access USB device from within container
  `-v /dev:/dev`

- No Python support
  - loss of data locality
- Model – as-a-service

Common Paradigm: TensorFlow Serving

- Movidius NCSDk2 – resolves some issues
- Feedback to Movidius team
- Service running on bare metal
- Movidius PCIe device coming soon!
  USB related issues moot

Movidius Next

- SDK2 just released
  - Up to 10 models may co-exist on one device,
  - FIFO queue,
  - 32 bit floating point
- Chip-2 Coming soon – at least an order of magnitude faster

  https://developer.movidius.com/start
  https://github.com/movidius/ncsdk
  https://github.com/kzzalews/sparkaisummit_movidius

Results

	CPU	FPGA	Movidius
Software Tools		CentOS 7.4 Intel Acceleration StackStack 1.0 Intel OpenVINO Toolkit with FPGA Support	SDK 1
Hardware	CPU: Intel Xeon CPU E5-1650 v2 @ 3.50GHz	FPGA: Arria 10 GX (1150K Logic elements, 8GB DDR4, PCIe Gen3)	Movidius
Inference Time/image	7.5 ms	3.2 ms	34 ms

Demo

https://videoportal.intel.com/media/0_selfn06l

Future Work

- Kubernetes Device Manager support for Movidius
- Explore native Spark support for Movidius
- Kubernetes/Spark Scheduler Enhancements
  - Wait for HW or launch anywhere?
  - Speed, power, and latency implications
  - Targeted models

Conclusion

- FPGA support more mature
- Give Movidius a try, delightful at its price point!!

  https://developer.movidius.com/start

  https://github.com/movidius/ncsdk

  https://github.com/kzzalews/sparkaisummit_movidius

References

Kubernetes Device Plugin:

- https://kubernetes.io/docs/tasks/manage-gpus/scheduling-gpus/
- https://kubernetes.io/docs/concepts/cluster-administration/device-plugins/
- https://github.com/kubernetes/community/blob/master/contributors/design-proposals/resource-management/device-plugin.md

FPGAs and the Movidius Chip

- https://venturebeat.com/2018/02/27/intel-makes-it-easier-to-bring-movidius-ai-accelerator-chip-into-production/
- https://newsroom.intel.com/editorials/introducing-myriad-x-unleashing-ai-at-the-edge/
- https://www.altera.com/products/fpga/stratix-series/stratix-10/overview.html
- https://medium.com/@xslittlegrass/self-driving-car-in-a-simulator-with-a-tiny-neural-network-13d33b871234

SparkCL: A Unified Programming Framework for Accelerators on Heterogeneous Clusters:

- https://arxiv.org/ftp/arxiv/papers/1505/1505.01120.pdf

Thank You!

Karol.Zalewski@intel.com
Malini.K.Bhandaru@intel.com