Latest trends in AI and Machine Learning

Julien Simon
Global Evangelist, AI & Machine Learning, Amazon Web Services

@julsimon
https://medium.com/@julsimon

Agenda

A quick word about AWS

Cool stuff that our customers build

FPGAs on AWS

FPGAs for AI and Machine Learning

Getting started

Figure 1. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide

Source: Gartner (July 2019)

AWS Recognized as a Cloud Leader for the 9th Consecutive Year

Gartner, Magic Quadrant for Cloud Infrastructure as a Service, Worldwide, Raj Bala, Bob Gill, Dennis Smith, David Wright, July 2019. ID G00365830. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose. The Gartner logo is a trademark and service mark of Gartner, Inc., and/or its affiliates, and is used herein with permission. All rights reserved.

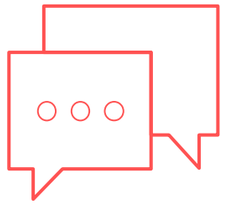
AWS Global Infrastructure

22 Regions (4 coming soon), 69 Availability Zones

AI is the centerpiece for digital transformation

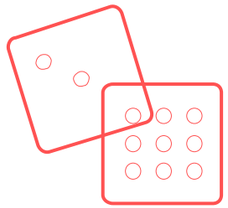
Customer experience

Business operations



Decision making

Innovation



Competitive advantage

40% of digital transformation initiatives supported by AI in 2019

More AI/ML happens on AWS than anywhere else

10,000+ active customers

https://aws.amazon.com/machine-learning/customers/

Nucleus Report, October 2019

96% of Deep Learning projects run in the cloud

89% of cloud-based Deep Learning projects run on AWS

85% of cloud-based TensorFlow projects run on AWS

83% of cloud-based PyTorch projects run on AWS

https://nucleusresearch.com/research/single/guidebook-deep-learning-on-aws/

Our stack

The AWS ML Stack

Broadest and deepest set of capabilities

AI Services

VISION			SPEECH		LANGUAGE		CHATBOTS	FORECASTING	RECOMMENDATIONS
AMAZON REKOGNITION IMAGE	AMAZON REKOGNITION VIDEO	AMAZON TEXTRACT	AMAZON POLLY	AMAZON TRANSCRIBE	AMAZON TRANSLATE	AMAZON COMPREHEND & AMAZON COMPREHEND MEDICAL	AMAZON LEX	AMAZON FORECAST	AMAZON PERSONALIZE

ML Services

Amazon SageMaker

Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting

ML Frameworks + Infrastructure

FRAMEWORKS	INTERFACES
TensorFlow, mxnet, PYTORCH	GLUON, Keras

INFRASTRUCTURE								
EC2 P3 & P3DN	EC2 G4 EC2 C5	FPGAs	AWS DL CONTAINERS & AMIs	AMAZON ELASTIC CONTAINER SERVICE	AMAZON ELASTIC KUBERNETES SERVICE	AWS IoT GREENGRASS	AMAZON ELASTIC INFERENCE	AWS INFERENTIA

Areas of Focus

1

Flexibility & Cost

More bang for your buck
Optimized frameworks

Save up to 90% on training
Managed Spot Instances

Save up to 80% on inference
Amazon Elastic Inference

2

Dat
a

Annotating data sets at scale
Amazon SageMaker Ground Truth

Reinforcement Learning
Amazon SageMaker RL

3

Ease of use

High level services
Call an API, get the job done

Off the shelf algorithms and models
AWS Marketplace for Machine Learning

AutoML
Amazon Personalize, Amazon Forecast

Cool stuff that our customers build

The AWS ML Stack

Broadest and deepest set of capabilities

ML Frameworks + Infrastructure

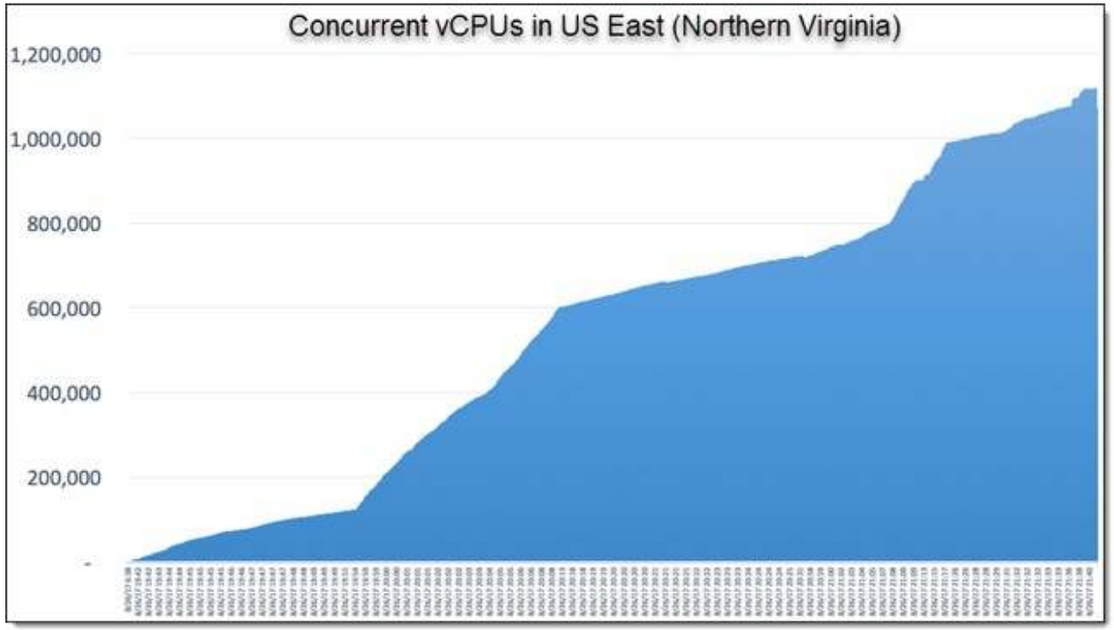
FRAMEWORKS	INTERFACES
TensorFlow, mxnet, PYTORCH	GLUON, Keras

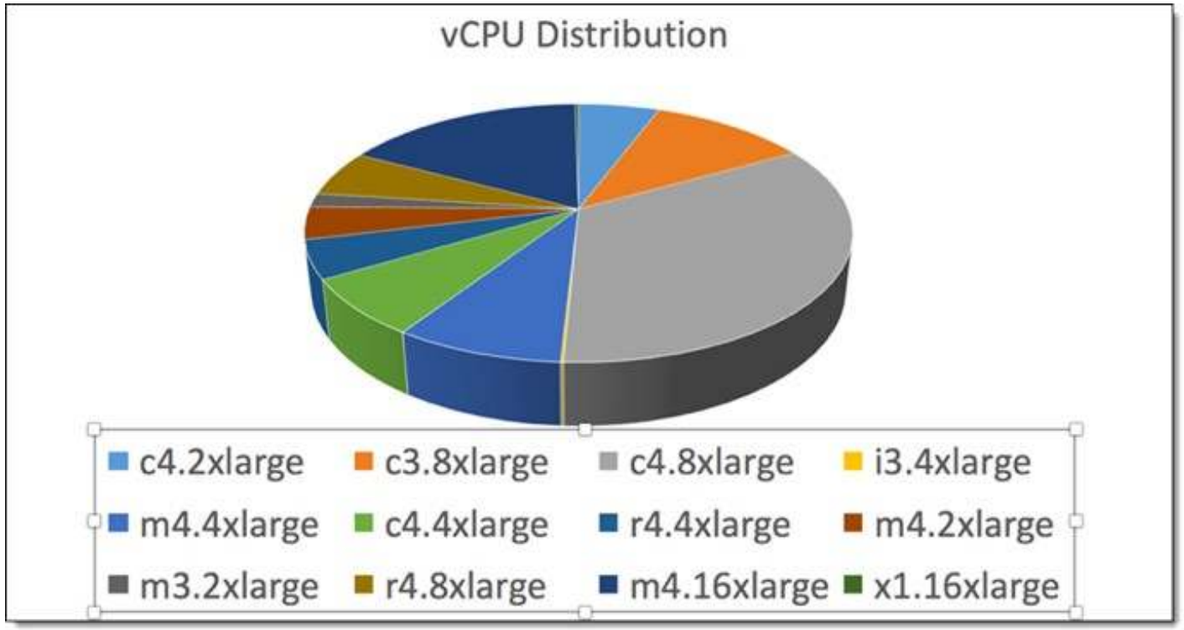
INFRASTRUCTURE								
EC2 P3 & P3DN	EC2 G4 EC2 C5	FPGAs	AWS DL CONTAINERS & AMIs	AMAZON ELASTIC CONTAINER SERVICE	AMAZON ELASTIC KUBERNETES SERVICE	AWS IoT GREENGRASS	AMAZON ELASTIC INFERENCE	AWS INFERENTIA

NLP project at Clemson University

1.1 million vCPUs in a single region

Optimized cost thanks to Spot Instances





https://aws.amazon.com/blogs/aws/natural-language-processing-at-clemson-university-1-1-million-vcpus-ec2-spot-instances/

Speeding up medical decisions

https://aws.amazon.com/solutions/case-studies/arterys/

ARTERYS
Medical Imaging Cloud AI

Anatomy contouring

As accurate as experts

15-20 seconds instead of 45-60 minutes

Autonomous Vehicles

https://www.tusimple.com
https://www.youtube.com/watch?v=VXSIq33WZoo

Level 4 autonomy

1,000-meter perception based on optical systems (day & night)

Billions of miles simulated on AWS

3 to 5 trips per day along three fixed routes in Arizona, with an average run of 200 miles

The AWS ML Stack

Broadest and deepest set of capabilities

ML Services

Amazon SageMaker							
Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting

Formula 1

https://aws.amazon.com/f1insights/

- 120 sensors per car
- 3GB and 1,500 data points per second
- 65 years of historical data

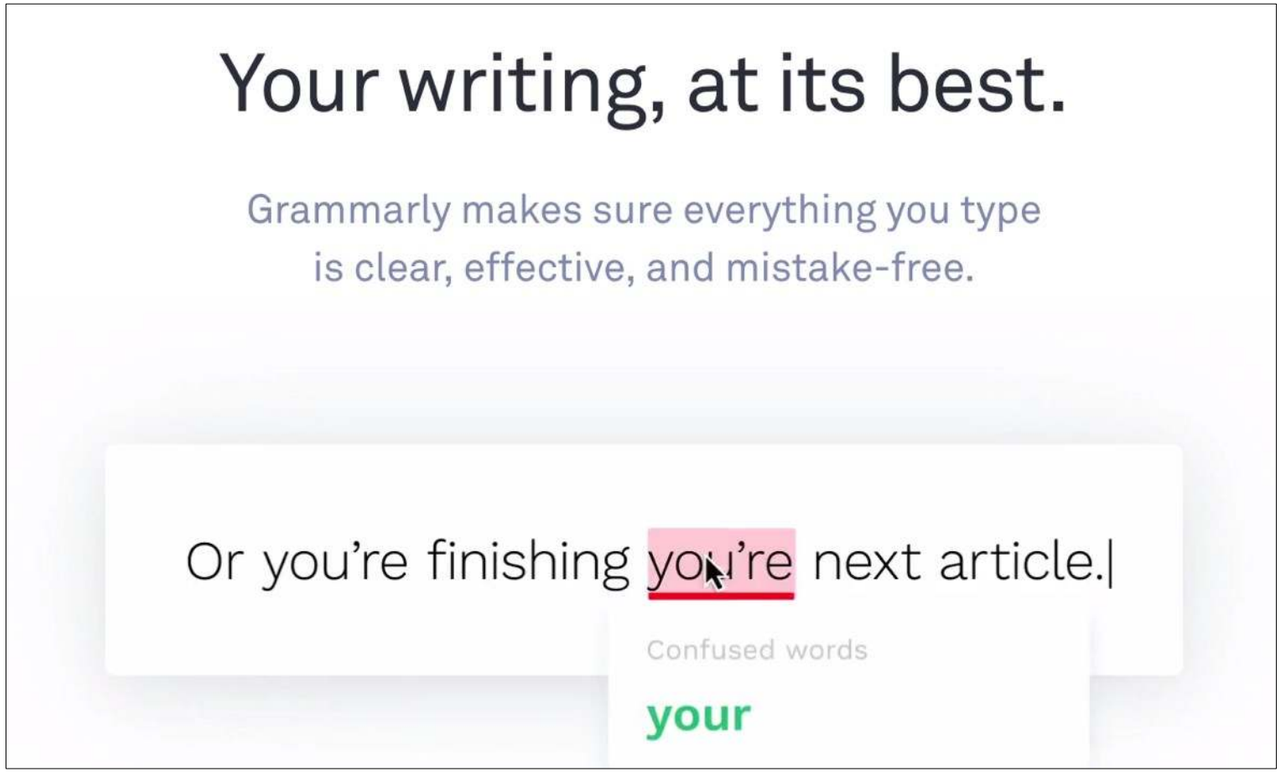
Overtake probability
Car performance
Pitstop advantage

Improving how we write

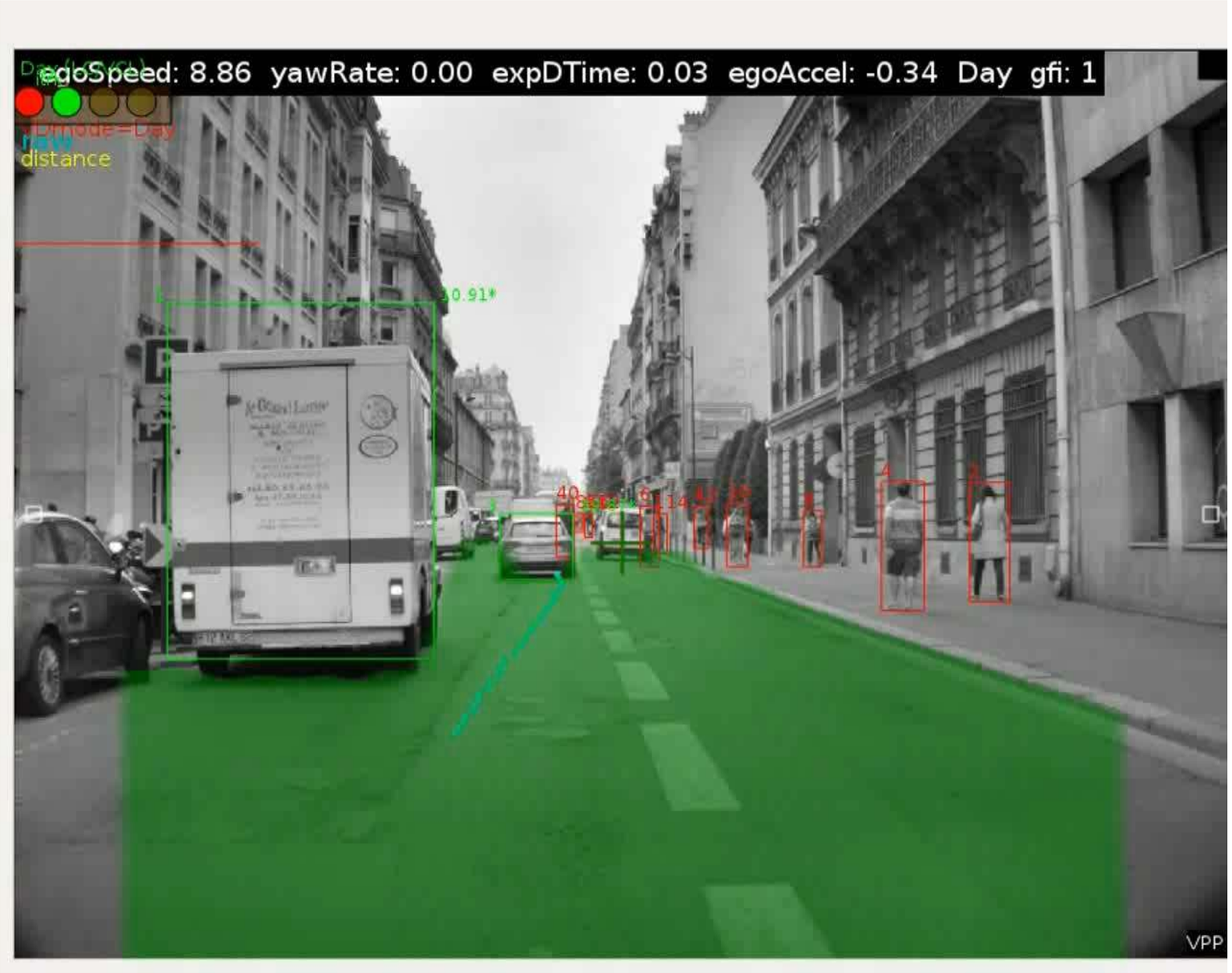
grammarly

"Amazon SageMaker makes it possible for us to develop our TensorFlow models in a distributed training environment (…)

We can run inference on SageMaker itself, or if we need just the model, we download it from S3 and run inference of our mobile device implementations for iOS and Android customers."



Advanced Driver Assistance Systems



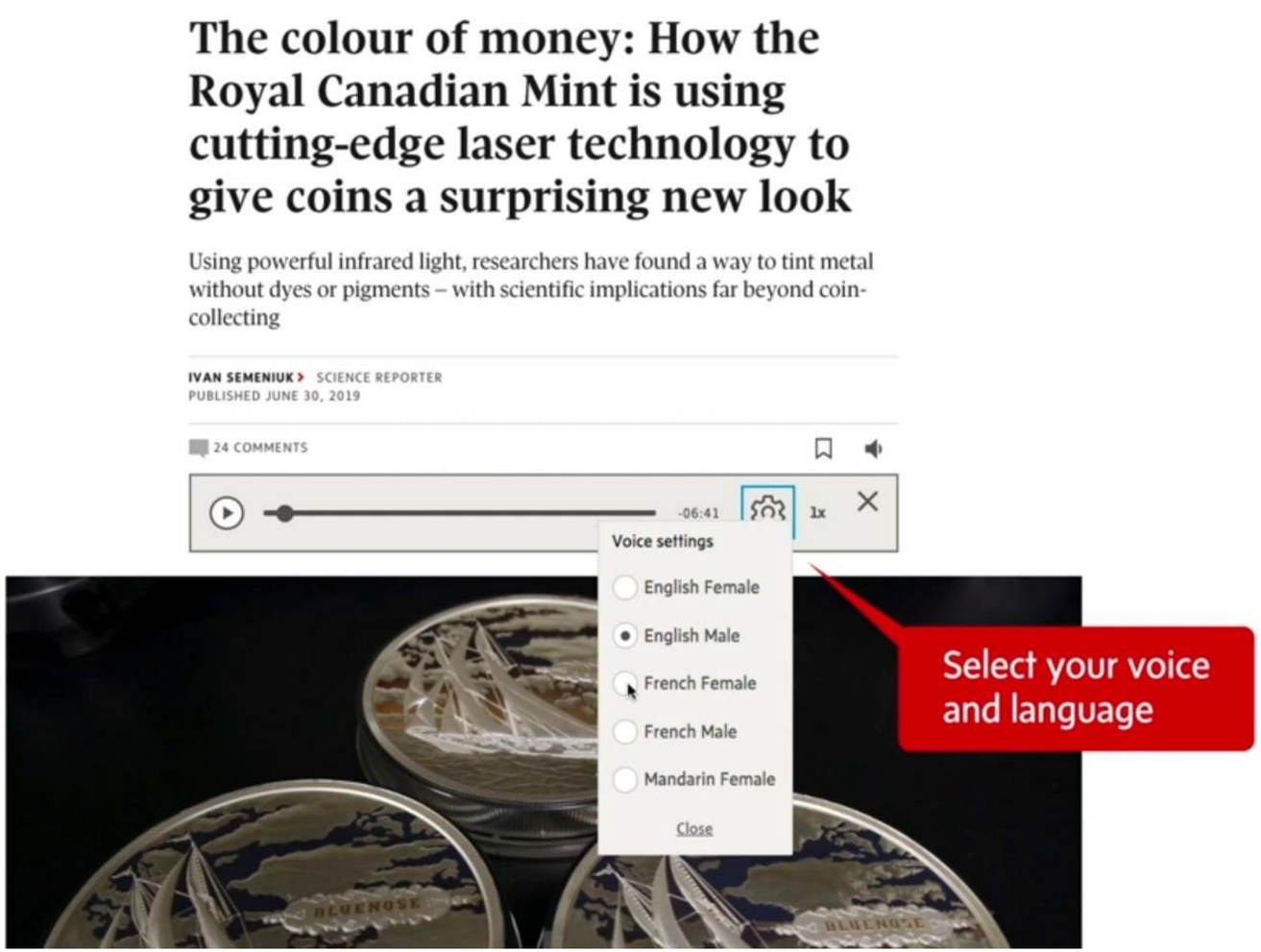
The AWS ML Stack

Broadest and deepest set of capabilities

AI Services

VISION			SPEECH		LANGUAGE		CHATBOTS	FORECASTING	RECOMMENDATIONS
AMAZON REKOGNITION IMAGE	AMAZON REKOGNITION VIDEO	AMAZON TEXTRACT	AMAZON POLLY	AMAZON TRANSCRIBE	AMAZON TRANSLATE	AMAZON COMPREHEND & AMAZON COMPREHEND MEDICAL	AMAZON LEX	AMAZON FORECAST	AMAZON PERSONALIZE

Automatic Translation and Text To Speech



https://www.theglobeandmail.com/inside-the-globe/article-new-to-the-globe-listen-to-articles-in-english-french-or-mandarin/

Domino's

Personalizing customer experiences

Domino's uses Amazon Personalize to customize and scale relevant marketing communications to customers based on time, context, and content, thereby improving and enhancing their experience with the Domino's brand.

FPGAs on AWS

The AWS ML Stack

Broadest and deepest set of capabilities

AI Services

VISION			SPEECH		LANGUAGE		CHATBOTS	FORECASTING	RECOMMENDATIONS
AMAZON REKOGNITION IMAGE	AMAZON REKOGNITION VIDEO	AMAZON TEXTRACT	AMAZON POLLY	AMAZON TRANSCRIBE	AMAZON TRANSLATE	AMAZON COMPREHEND & AMAZON COMPREHEND MEDICAL	AMAZON LEX	AMAZON FORECAST	AMAZON PERSONALIZE

ML Services

Amazon SageMaker

Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting

ML Frameworks + Infrastructure

FRAMEWORKS	INTERFACES
TensorFlow, mxnet, PYTORCH	GLUON, Keras

INFRASTRUCTURE								
EC2 P3 & P3DN	EC2 G4 EC2 C5	FPGAs	AWS DL CONTAINERS & AMIs	AMAZON ELASTIC CONTAINER SERVICE	AMAZON ELASTIC KUBERNETES SERVICE	AWS IoT GREENGRASS	AMAZON ELASTIC INFERENCE	AWS INFERENTIA

FPGAs on AWS

- Financial computing
- Genomics
- Engineering simulations
- Image and video processing
- Big data and machine learning
- Security
- Compression
- …and more

Three Ways to Use FPGAs on AWS

1 Hardware Developers

Use F1 Hardware Development Kit (HDK) to develop and deploy custom FPGA accelerations using Verilog and VHDL

2 Software Developers

Use OpenCL to build custom accelerations from C/C++ code

3 AWS Users

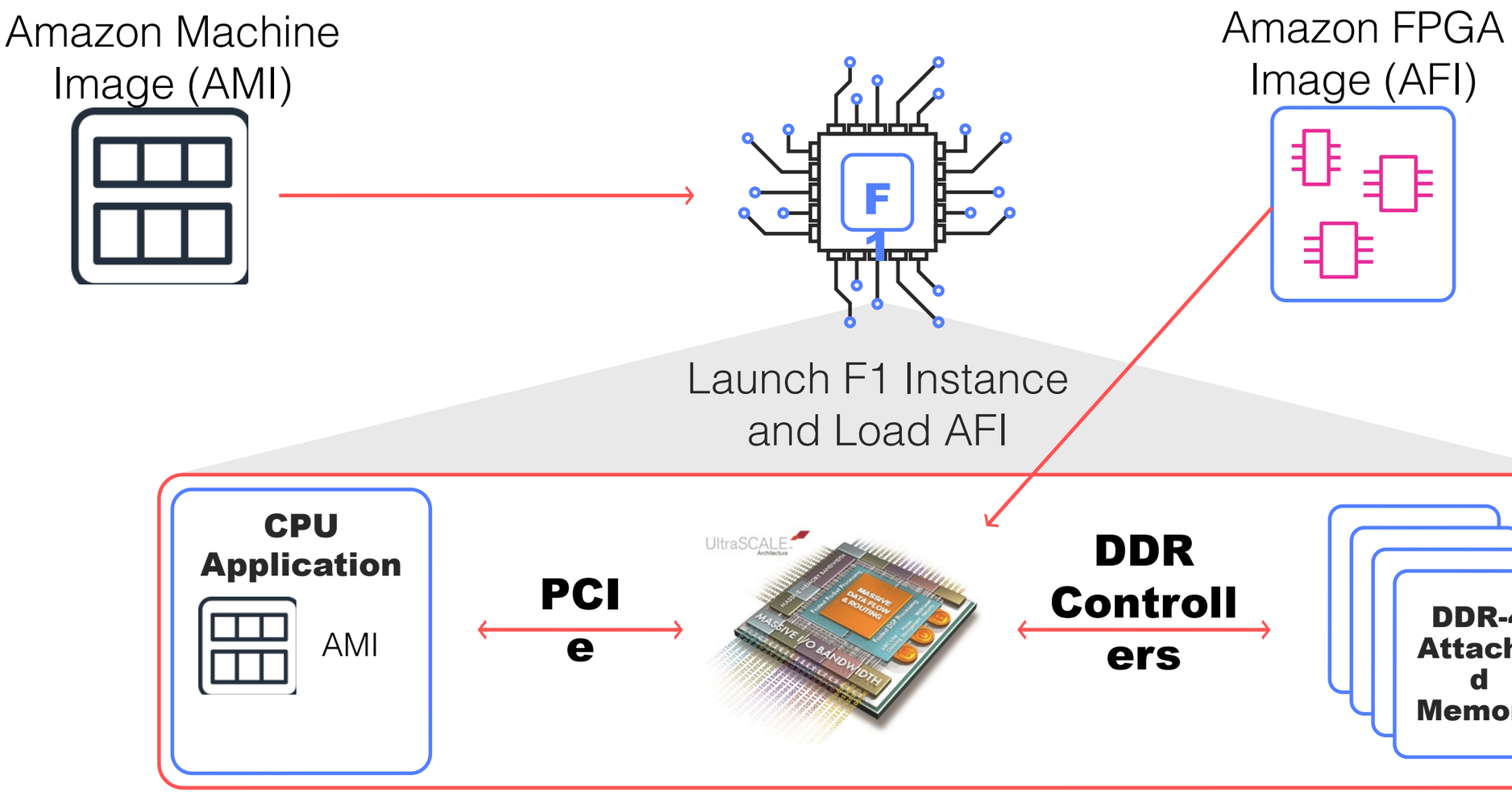
Use pre-built and ready to use accelerations available in AWS Marketplace

F1 instances: Optimized for Hardware Accelerators

- Xilinx 16nm UltraScale+ VU9P FPGA – Up to 8 FPGAs per instance
  - Each FPGA includes 64 GiB DDR4 ECC
  - Up to full bandwidth PCIe x16
  - 2.5 Million logic elements and 6,800 Digital Signal Processing engines
- 3 different instance sizes with up to 32 cores (64 vCPUs) per instance
- 32:1 memory to core ratio and up to 1 TB of RAM
- Includes local NVME storage

Model	FPGA	vCPU	Memory (GiB)	Instance storage (GiB)	Networking performance	EBS bandwidth
f1.2xlarge	1	8	122	1 x 470 NVMe SSD	Up to 10 Gbps	1,7 Gbps
f1.4xlarge	2	16	244	1 x 940 NVMe SSD	Up to 10 Gbps	3,5 Gbps
f1.16xlarge	8	64	976	4 x 940 NVMe SSD	25 Gbps	14 Gbps

FPGA acceleration using F1



An F1 instance can have any number of AFIs

An AFI can be loaded into the FPGA in seconds

Developer Tools

aws marketplace

FPGA Developer AMI

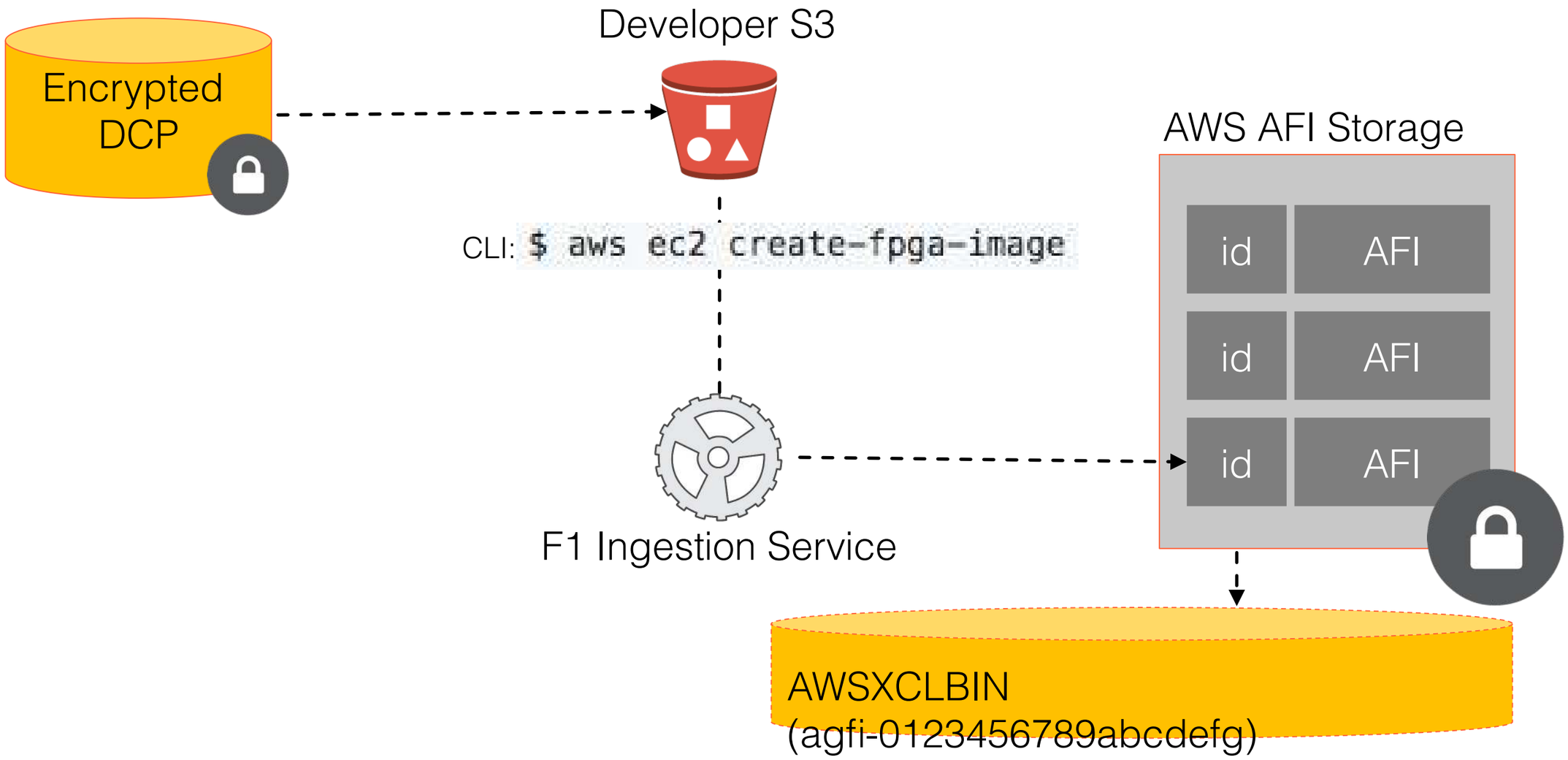
By: Amazon Web Services Latest Version: v1.7.0

Gitub:
aws/aws-fpga
SDK & HDK

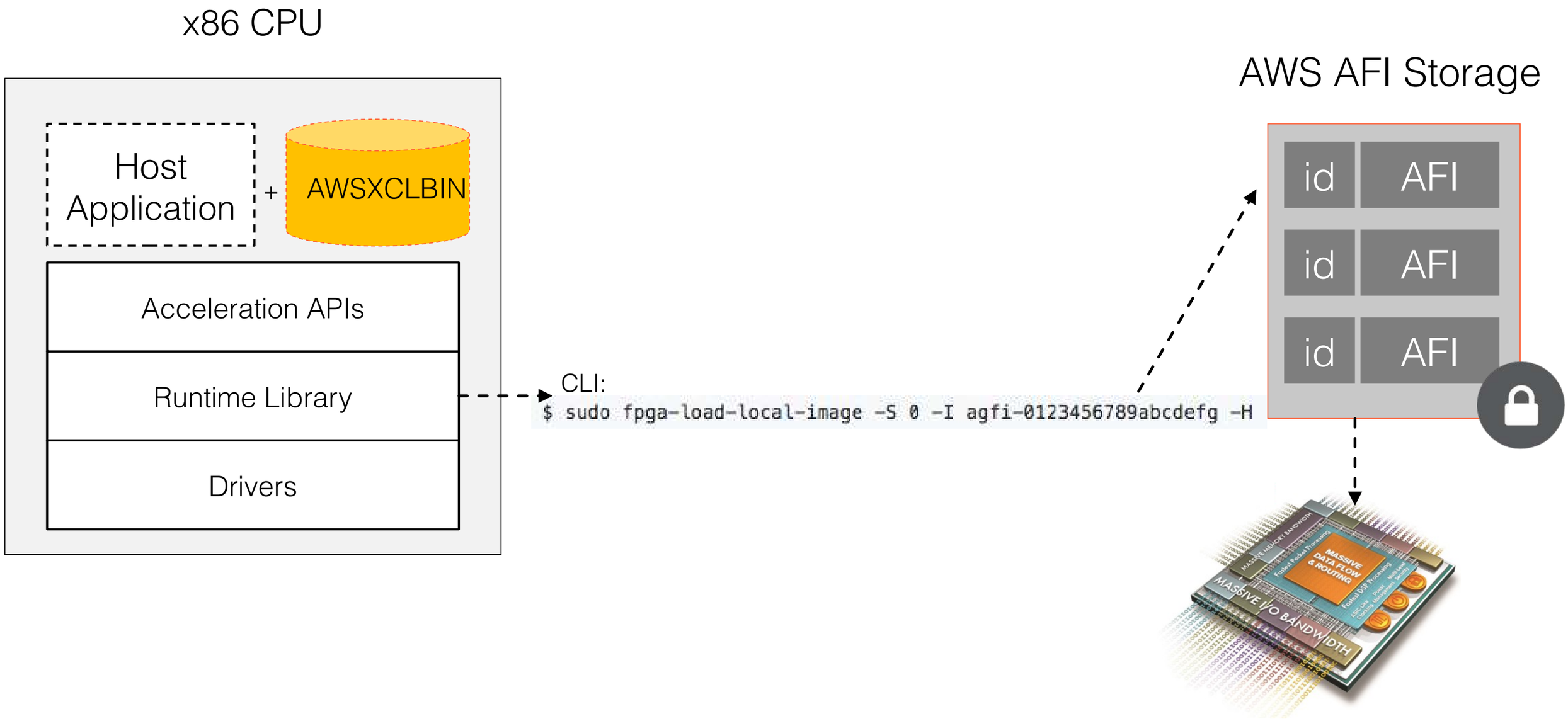
```
aws ec2 help
```

Tool	Development/Runtime	Tool location	Description
SDx 2017.4, 2018.2, 2018.3 & 2019.1	Development	FPGA Developer AMI	Used for Software Defined Accelerator Development
Vivado 2017.4, 2018.2, 2018.3 & 2019.1	Development	FPGA Developer AMI	Used for Hardware Accelerator Development
FPGA AFI Management Tools	Runtime	SDK - fpga_mgmt_tools	Command-line tools used for FPGA management while running on the F1 instance
Virtual JTAG	Development (Debug)	FPGA Developer AMI	Runtime debug waveform
wait_for_afi	Development	wait_for_afi.py	Helper script that notifies via email on AFI generation completion
notify_via_sns	Development	notify_via_sns.py	Notifies developer when design build process completes
AFI Administration	Development	Copy, Delete, Describe, Attributes	AWS CLI EC2 commands for managing your AFIs

Building Amazon FPGA Image (AFI)



Loading Amazon FPGA Image (AFI)



Quick demo

FPGAs for Machine Learning

Why FPGAs for Machine Learning

GPUs are great for training, but what about inference?

Throughput vs. latency: pick one?

- Using batches increases latency
- Using single samples degrades throughput

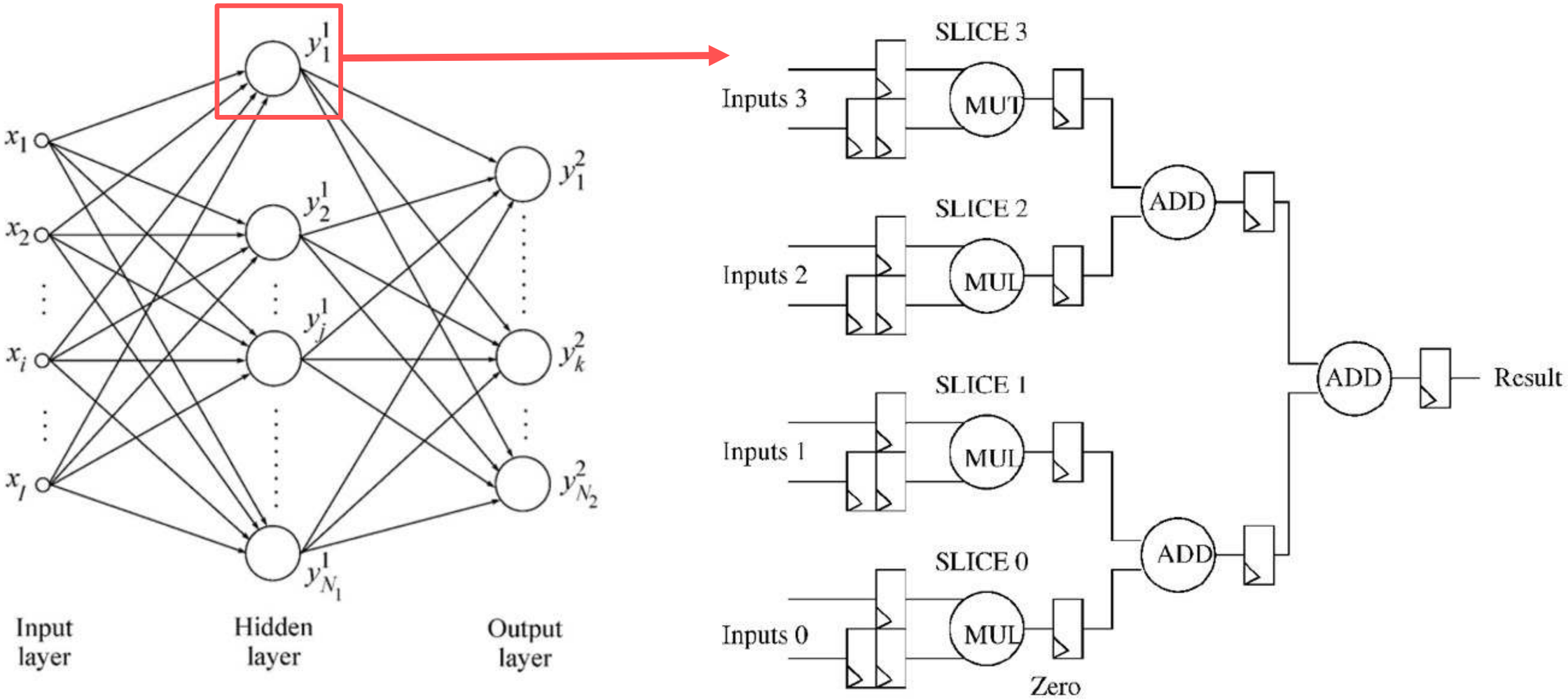
Power and memory requirements

- Floating-point operations are power-hungry
- Floating-point parameters need more DRAM, which is power-hungry too

Neural networks can be implemented efficiently on FPGA

Using custom logic to Multiply and Accumulate

Source: « FPGA Implementations of Neural Networks », Springer, 2006



Smaller weights → fewer gates, less data to feed to the FPGA

Optimizing models for FPGAs

Quantization: using integer weights

- 8/4/2-bit integers instead of 32-bit floats
- Reduces power consumption
- Simplifies the logic needed to implement the model
- Reduces memory usage

Pruning: removing useless connections

- Increases computation speed
- Reduces memory usage

Compression: encoding weights

- Reduces model size

On-chip SRAM becomes a viable option

More power-efficient than DRAM

Faster than off-chip DRAM

Machine Learning AFIs on AWS Marketplace

Accelerated ML Suite: Logistic Regression, K-Means, ALS

VGG16 convolutions layers

Torch for LuaJIT

Binarized Neural Networks

ML Suite https://github.com/Xilinx/ml-suite

Getting started

https://ml.aws

https://aws.amazon.com/sagemaker

https://aws.amazon.com/ec2/instance-types/f1/

https://medium.com/@julsimon/building-fpga-applications-on-aws-and-yes-for-deep-learning-too-643097257192

Thank you!

Julien Simon
Global Evangelist, AI & Machine Learning

@julsimon
https://medium.com/@julsimon