Xilinx ML Suite Overview
Jupyter Notebooks walkthrough

Jon Beckwith
SAE, Xilinx

XILINX.

Dad joke time

> What kind of snake does a math teacher own?

> > A pi-thon

XILINX.

Before we begin, FPGAs in the cloud

Cloud Based Acceleration

Alveo U200
- 18.6 Peak INT8 TOPs
- 77GB/s DDR Memory Bandwidth
- 31TB/s Internal SRAM Bandwidth

Alveo U250
- 33.3 Peak INT8 TOPs
- 77GB/s DDR Memory Bandwidth
- 38TB/s Internal SRAM Bandwidth

Alveo U280
- 24.5 Peak INT8 TOPs
- 460GB/s HBM2 Memory Bandwidth
- 30TB/s Internal SRAM Bandwidth

Accelerated Cloud Service Partners

Click to learn more

Alibaba Cloud   aws   Baidu

HUAWEI   NIMBIX   Tencent Cloud

Reconfigurable Acceleration in the Cloud

Faced with exponential growth in computing requirements and the inability for CPU technology to keep pace, cloud and data center architectures are moving toward accelerated computing. Accelerators compliment CPU-based architectures and deliver both performance and power efficiency.

FPGAs can deliver 10x acceleration across a broad set of applications and are reconfigurable to provide an ideal fit for the changing workloads of the modern data center.

With acceleration capabilities a full generation ahead of any other FPGA, Xilinx UltraScale™ and UltraScale+ FPGAs are empowering hardware and application developers in many of the world's largest and most innovative cloud computing services.

> **https://www.xilinx.com/products/design-tools/cloud-based-acceleration.html**

>> 4

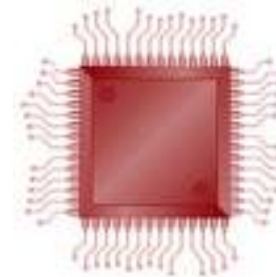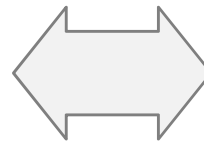XILINX.

Introducing Amazon EC2 F1

> **Amazon EC2 F1 is a compute instance with Xilinx FPGAs which can be programmed to create custom hardware accelerated applications**

> **F1 instances are easy to program and come with everything needed to develop, simulate, debug, and compile hardware accelerators**

> **Once a FPGA design is complete, it can be registered as an Amazon FPGA Image (AFI), and deployed to F1 instance in just a few clicks**
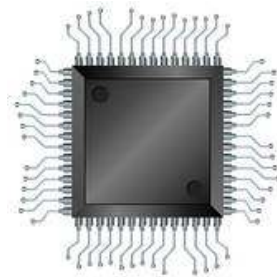
Host CPU

Xilinx FPGA

Compelling Applications for Cloud FPGA Acceleration

Genomics

Big Data Analytics

Financial Analytics

Security

Image and Video Processing

Machine Learning

XILINX.

ML Suite

Accelerating AI Inference into Your Cloud Applications

Deep Learning Applications

Featuring the Most Powerful
FPGA in the Cloud

**Virtex® Ultrascale+™
VU9P**

**Zynq® Ultrascale+™
MPSoC**

Cloud

On Premises

Edge
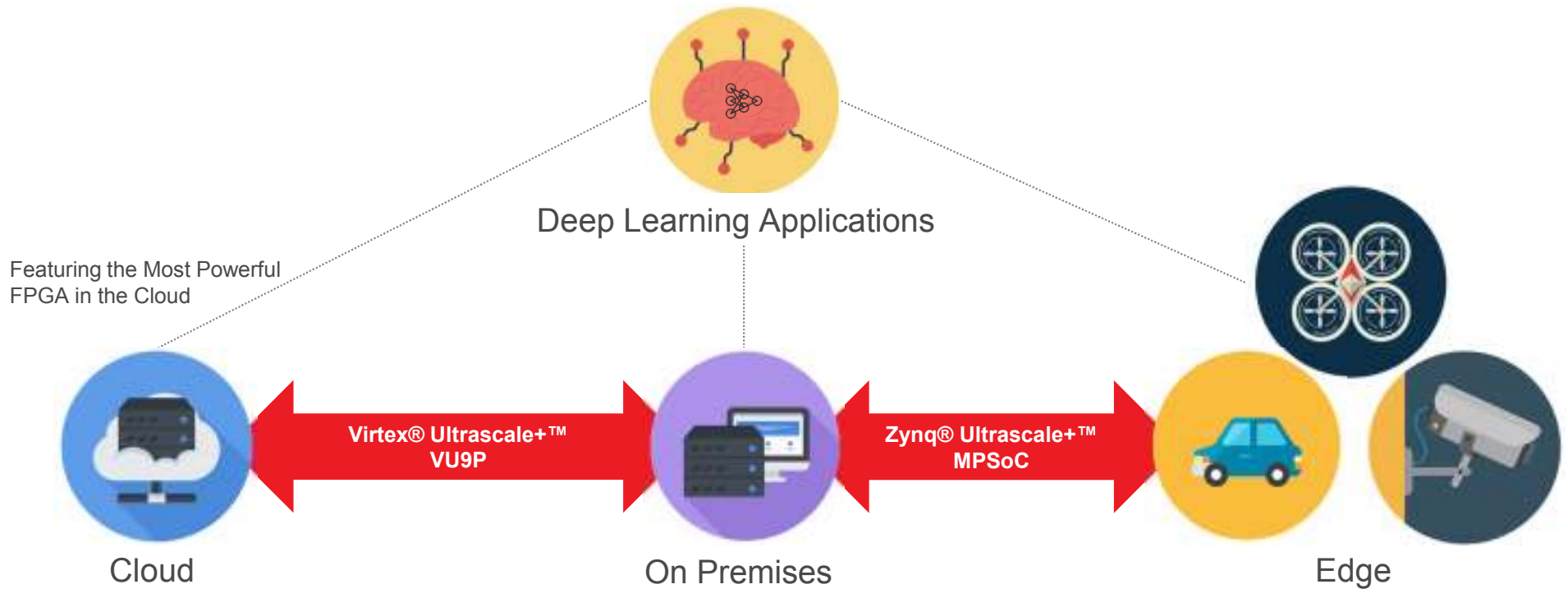
Want to try the out Xilinx ML Suite ?

➤ https://github.com/Xilinx/ml-suite

Xilinx ML Suite - AWS Marketplace

❯ ML Suite

≫ Supported Frameworks:
  – Caffe
  – MxNet
  – Tensorflow
  – Keras
  – Python Support
  – Darknet

≫ Jupyter Notebooks available:
  – Image Classification with Caffe
  – Using the xfDNN Compiler w/ a Caffe Model
  – Using the xfDNN Quantizer w/ a Caffe Model

≫ Pre-trained Models
  – Caffe 8/16-bit
    – GoogLeNet v1
    – ResNet50
    – Flowers102
    – Places365
  – Python 8/16-bit
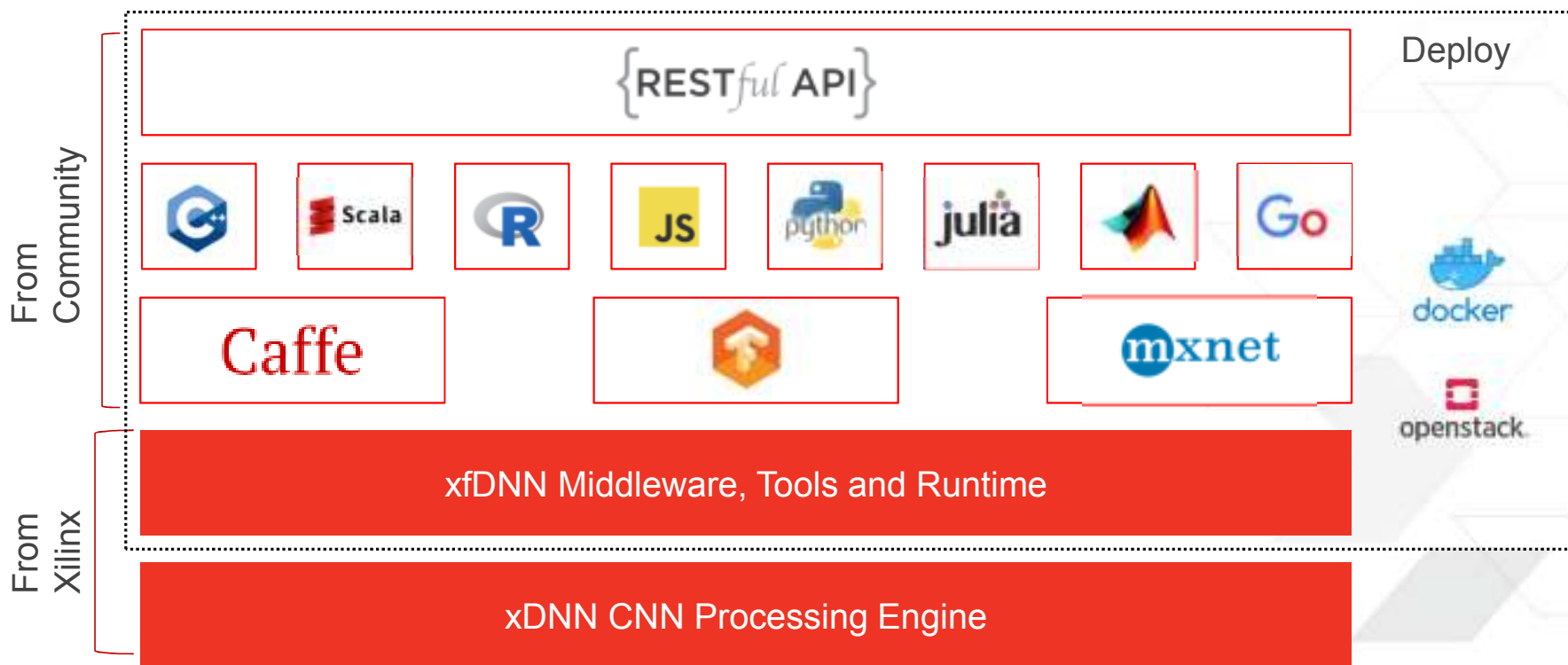   – Yolov2
  – MxNet 8/16-bit
    – GoogLeNet v1

≫ xfDNN Tools
  – Compiler
  – Quantizer

https://aws.amazon.com/marketplace/pp/B077FM2JNS?qid=1544477354556&sr=0-2&ref_=srh_res_product_title
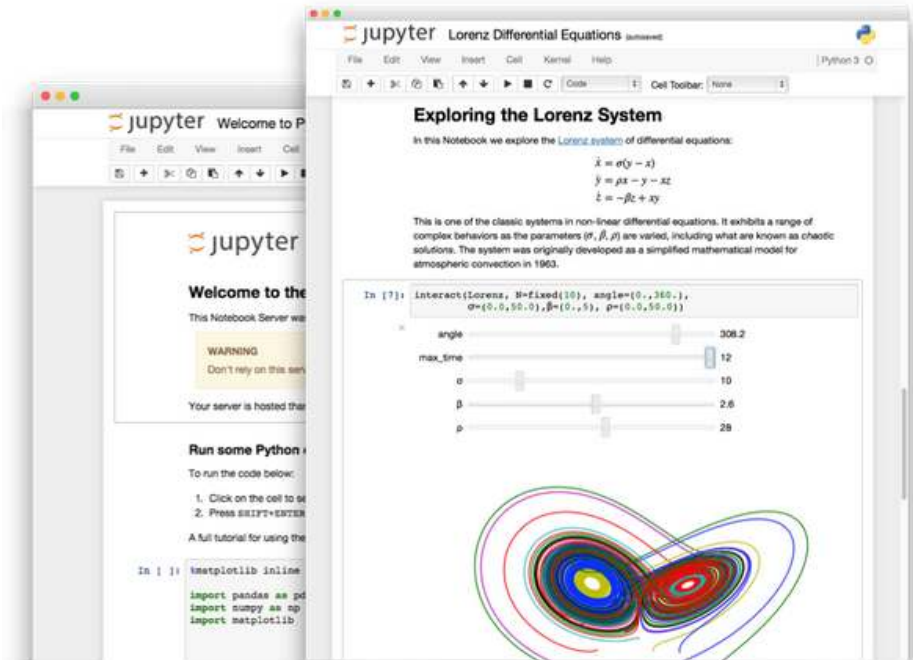
Seamless Deployment with Open Source Software

Deploy

From Community

{RESTful API}

C++ Scala R JS python julia MATLAB Go

Caffe TensorFlow mxnet

docker

openstack

From Xilinx

xfDNN Middleware, Tools and Runtime

xDNN CNN Processing Engine

XILINX.

Jupyter Notebooks Walkthrough

Jupyter Notebooks

ML Suite AMIs launch Jupyter Notebooks at boot
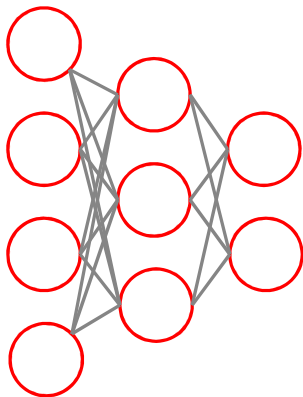
http://<aws_ip_addr>:8888

Demo

XILINX.

xfDNN Inference Toolbox
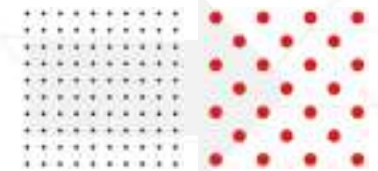
Graph Compiler

- Python tools to quickly compile networks from common Frameworks – Caffe, MxNet and Tensorflow

Network Optimization

- Automatic network optimizations for lower latency by fusing layers and buffering on-chip memory

xfDNN Quantizer

- Quickly reduce precision of trained models for deployment
- Maintains 32bit accuracy at 8 bit within 2%

€ XILINX.

Adaptable.
Intelligent.