SEOUL

19.09.26

DEV DAY

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DEV DAY
SEOUL

19.09.26

모두를 위한 컴퓨터 비전 딥러닝 툴킷, GluonCV 따라하기

2-1. MXNet / Gluon Overview

김무현 데이터 사이언티스트
Amazon Machine Learning Solutions Lab

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our mission at AWS

Put machine learning in the hands of every developer

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customers running machine learning on AWS

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

The Amazon Machine Learning stack

AI SERVICES

(App developers with little knowledge of ML)

Vision

- Amazon Rekognition Image
- Amazon Rekognition Video

Speech

- Amazon Textract — NEW
- Amazon Polly
- Amazon Transcribe

Language

- Amazon Translate
- Amazon Comprehend

Chatbots

- Amazon Lex

Forecasting

- Amazon Forecast — NEW

Recommendations

- Amazon Personalize — NEW

ML SERVICES

(ML developers and data scientists)

Amazon SageMaker

- Ground Truth — NEW
- Notebooks
- Algorithms
- Marketplace — NEW
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning — NEW
- Training
- Optimization (Neo) — NEW
- Deployment
- Hosting

ML FRAMEWORKS & INFRASTRUCTURE

(ML researchers and academics)

Frameworks

- TensorFlow
- mxnet
- PYTORCH
- Chainer
- HOROVOD

Interfaces

- GLUON
- Keras

Infrastructure

- Amazon EC2 P3 & P3DN — NEW
- Amazon EC2 C5
- FPGAs
- AWS Greengrass
- Amazon Elastic Inference — NEW
- Amazon Inferentia — NEW

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

ML FRAMEWORKS &
INFRASTRUCTURE

(ML researchers
and academics)

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What you'll learn about today

History & Motivation

Being more productive as a developer

Getting more training done, in less time

Deployment: Easy, efficient and scalable

Portability and flexibility: Languages, devices, and fra meworks

Getting started and Resources

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

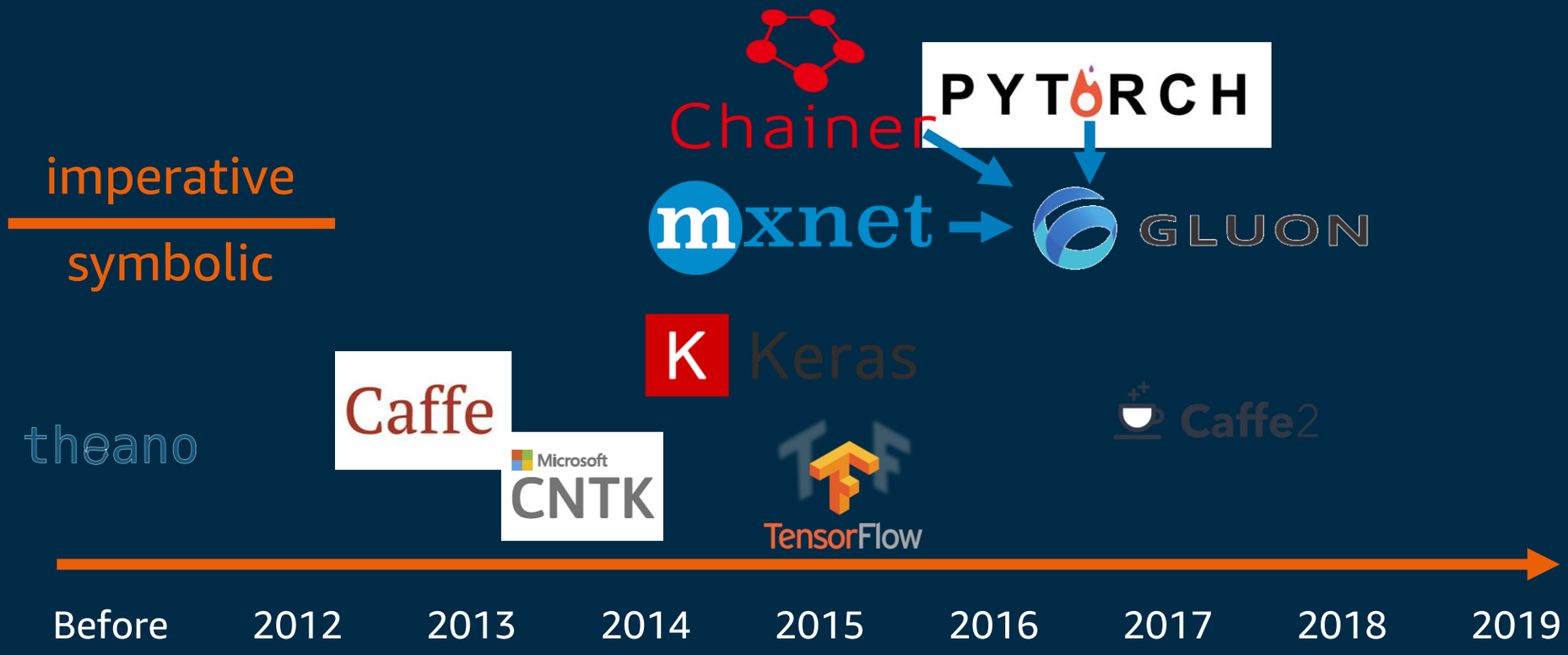
aws

DEV DAY

History

aws

History



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

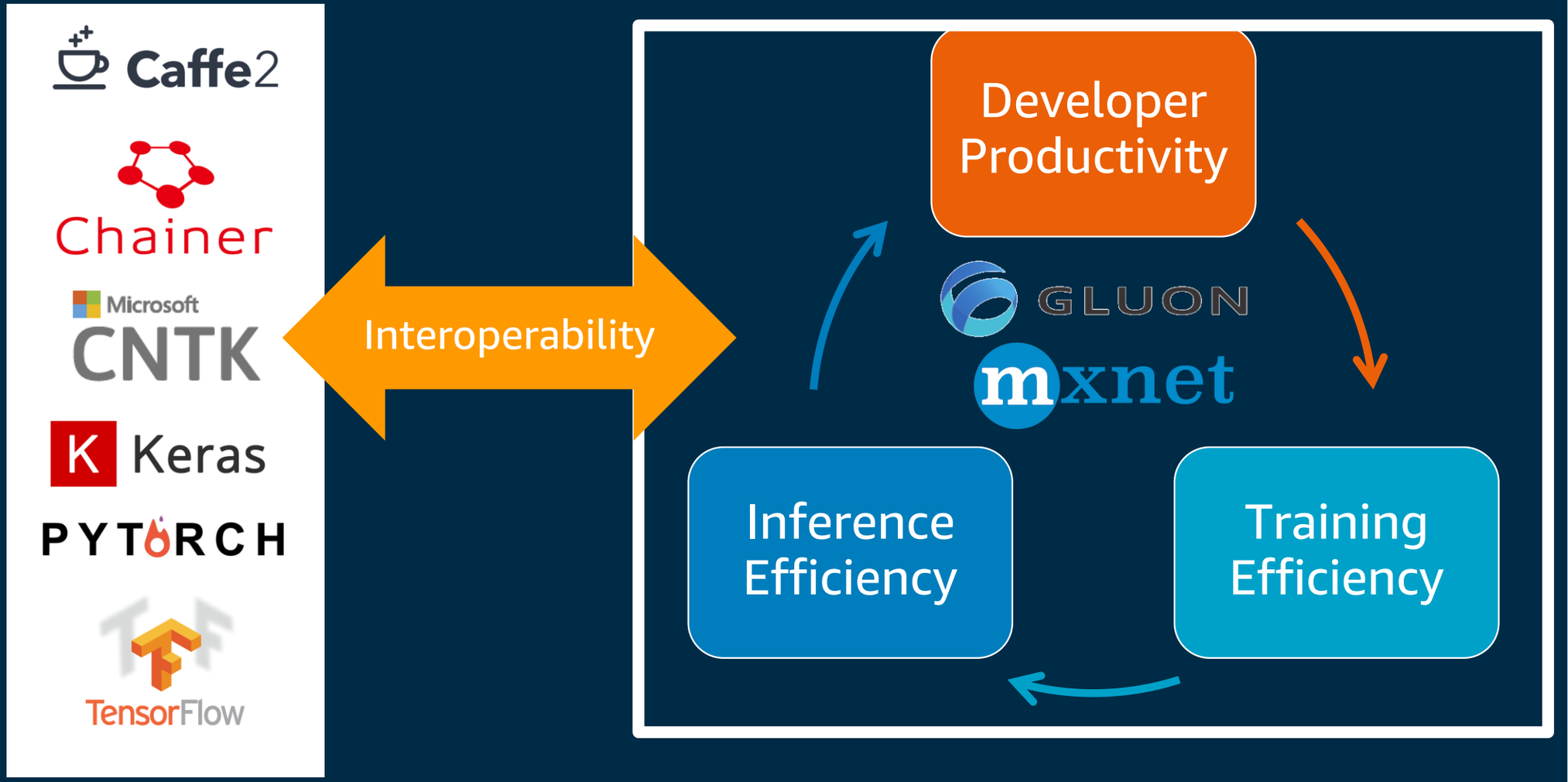
aws

DEV DAY

MXNet and Gluon: Goals

aws

Goals



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

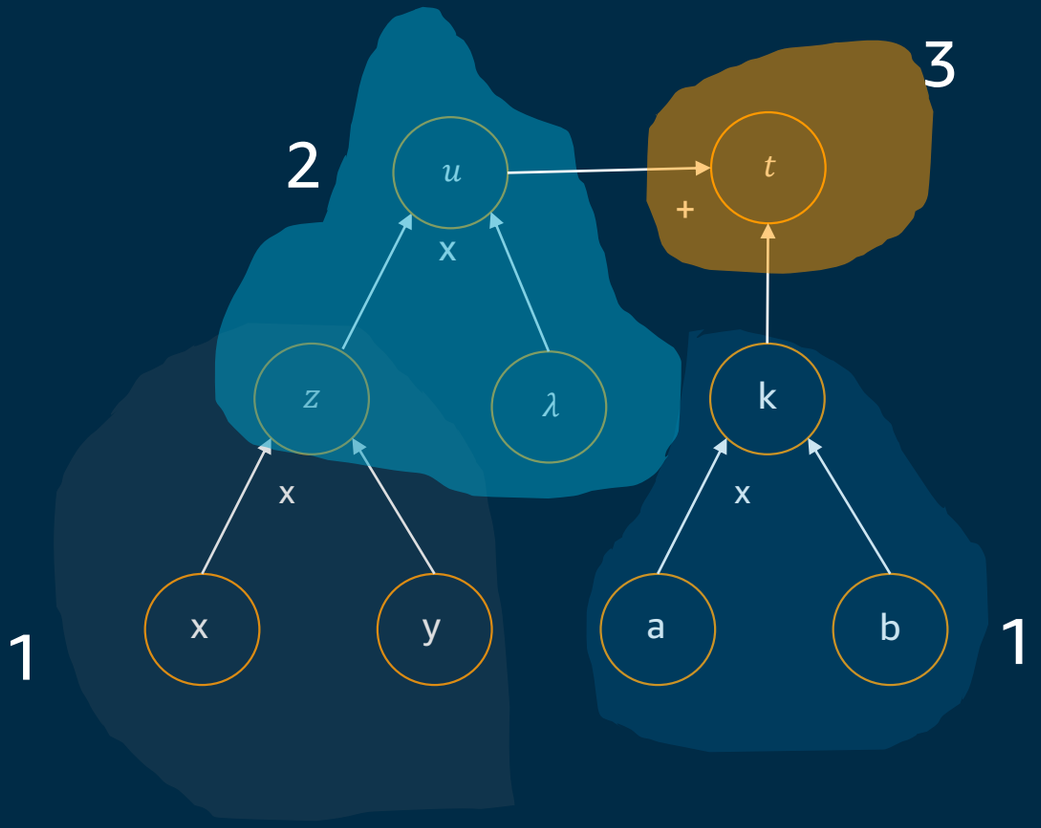
aws

Computational dependency graph

$$z = x \cdot y$$

$$k = a \cdot b$$

$$t = \lambda z + k$$



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

MXNet computational dependency graph

```
net = mx.sym.Variable('data')
net = mx.sym.FullyConnected(net, name='fc1', num_hidden=128)
net = mx.sym.Activation(net, name='relu1', act_type="relu")
net = mx.sym.FullyConnected(net, name='fc2', num_hidden=10)
net = mx.sym.SoftmaxOutput(net, name='softmax')
```

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Training

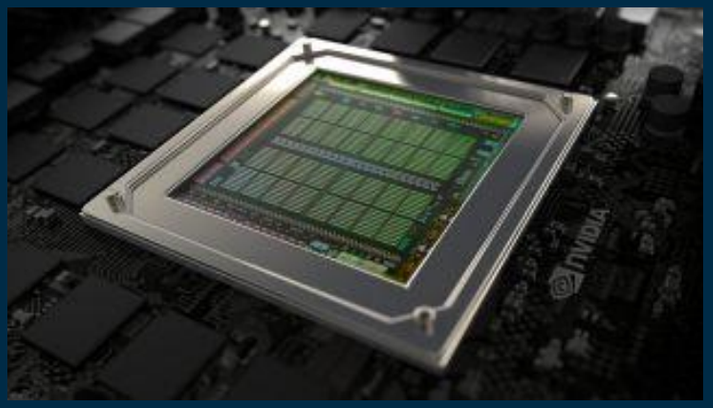
```
import logging

logging.getLogger().setLevel(logging.DEBUG)  # logging to stdout

# create a trainable module on compute context

mlp_model = mx.mod.Module(symbol=mlp, context=ctx)

mlp_model.fit(train_iter,

              eval_data=val_iter,

              optimizer='sgd',

              optimizer_params={'learning_rate':0.1},

              eval_metric='acc',

              batch_end_callback = mx.callback.Speedometer(batch_size, 100),

              num_epoch=10)
```

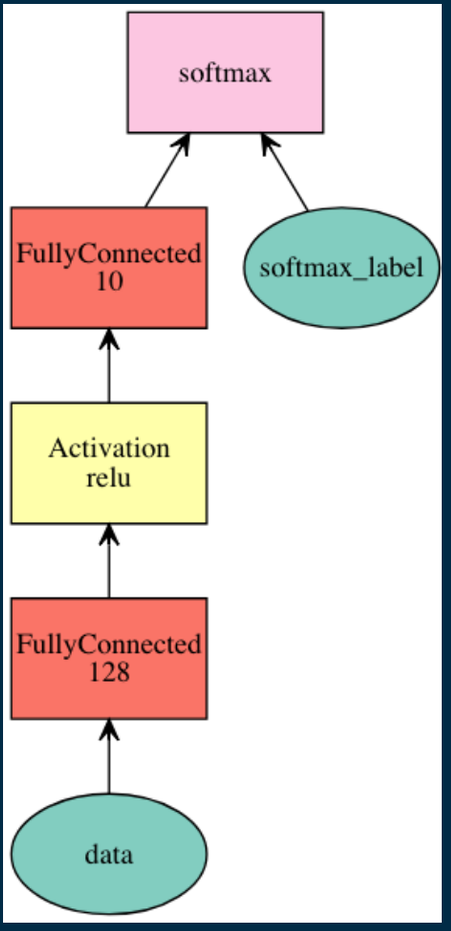
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

MXNet computational dependency graph

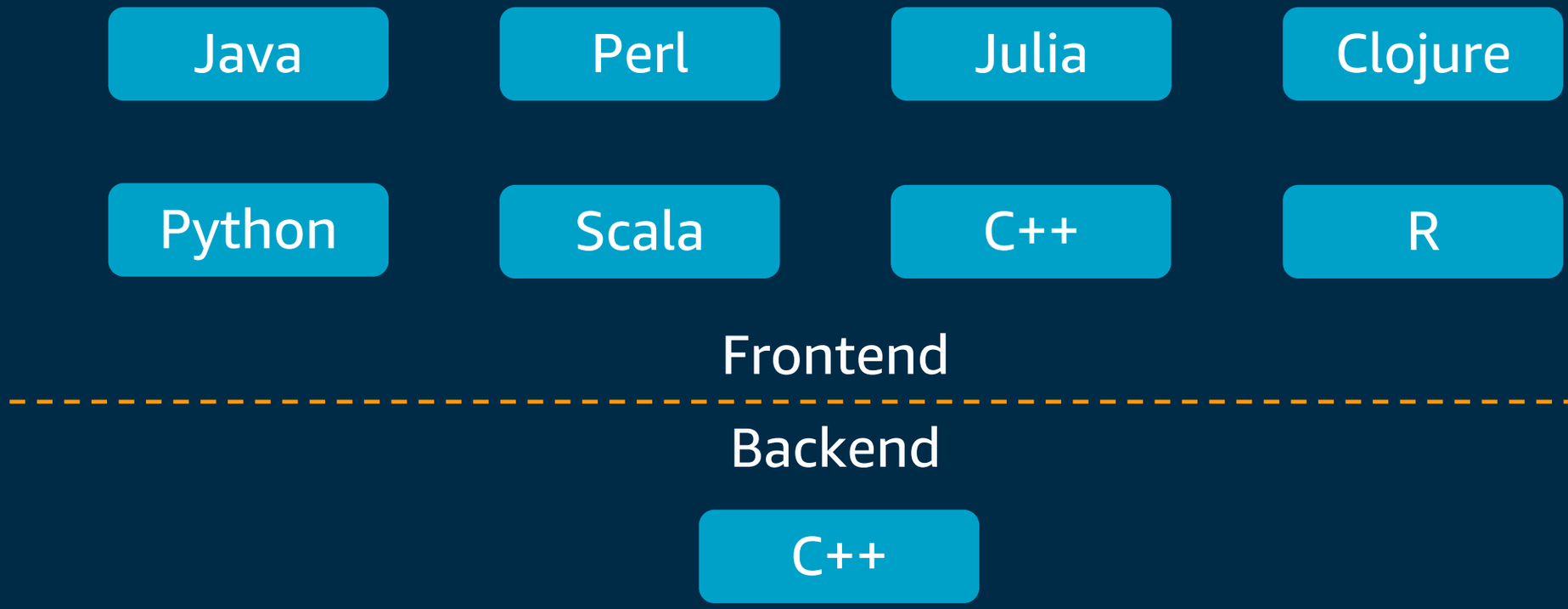
```
net = mx.sym.Variable('data')
net = mx.sym.FullyConnected(net, name='fc1', num_hidden=64)
net = mx.sym.Activation(net, name='relu1', act_type="relu")
net = mx.sym.FullyConnected(net, name='fc2', num_hidden=10)
net = mx.sym.SoftmaxOutput(net, name='softmax')
```





© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Multi-language support



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DEV DAY

Developer Productivity

aws

Simple, Easy-to-Understand Code

Flexible, Imperative Structure

Dynamic Graphs

High Performance

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Network definition in Gluon

```
net = gluon.nn.HybridSequential()

with net.name_scope():

    net.add(gluon.nn.Dense(units=64, activation='relu'))

    net.add(gluon.nn.Dense(units=10))

softmax_cross_entropy = gluon.loss.SoftmaxCrossEntropyLoss()

net.initialize(mx.init.Xavier(magnitude=2.24), ctx=ctx, force_rei
nit=True)

trainer = gluon.Trainer(net.collect_params(), 'sgd', {'learning_r
ate': 0.02})
```

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Training in Gluon

```
smoothing_constant = .01
for e in range(10):
    cumulative_loss = 0
    for i, (data, label) in enumerate(train_data):
        data = data.as_in_context(model_ctx).reshape((-1, 784))
        label = label.as_in_context(model_ctx)
        with autograd.record():
            output = net(data)
            loss = softmax_cross_entropy(output, label)
        loss.backward()
        trainer.step(data.shape[0])
```

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Imperative API

Debuggable

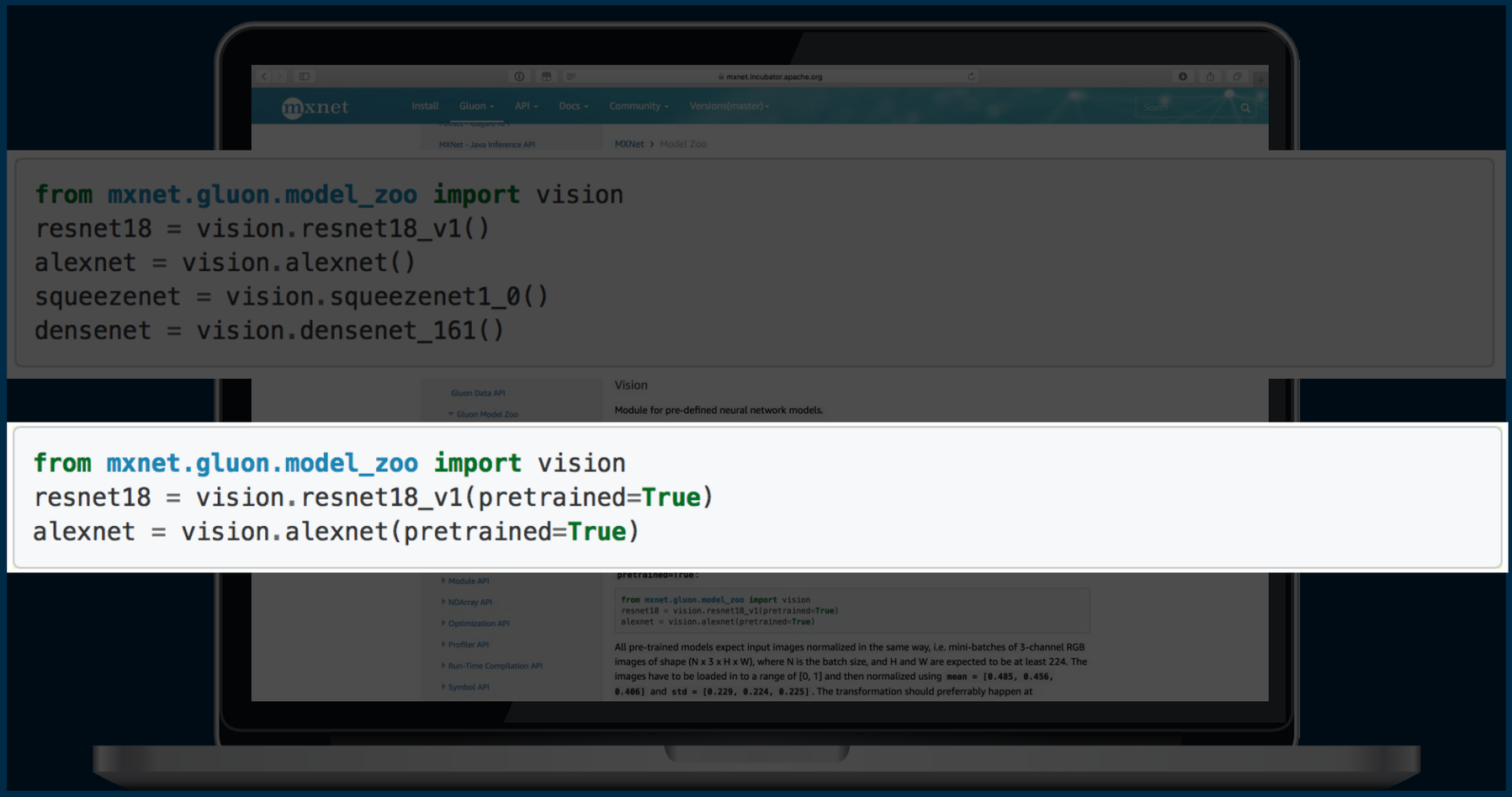
Flexible

Scalable

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

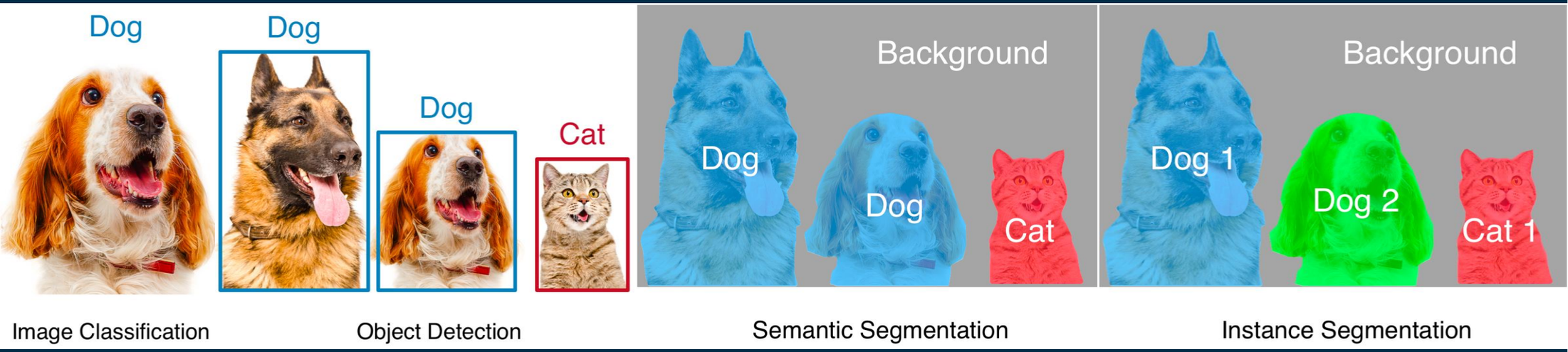
aws

Gluon Model Zoo



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

GluonCV: a deep learning toolkit for computer vision



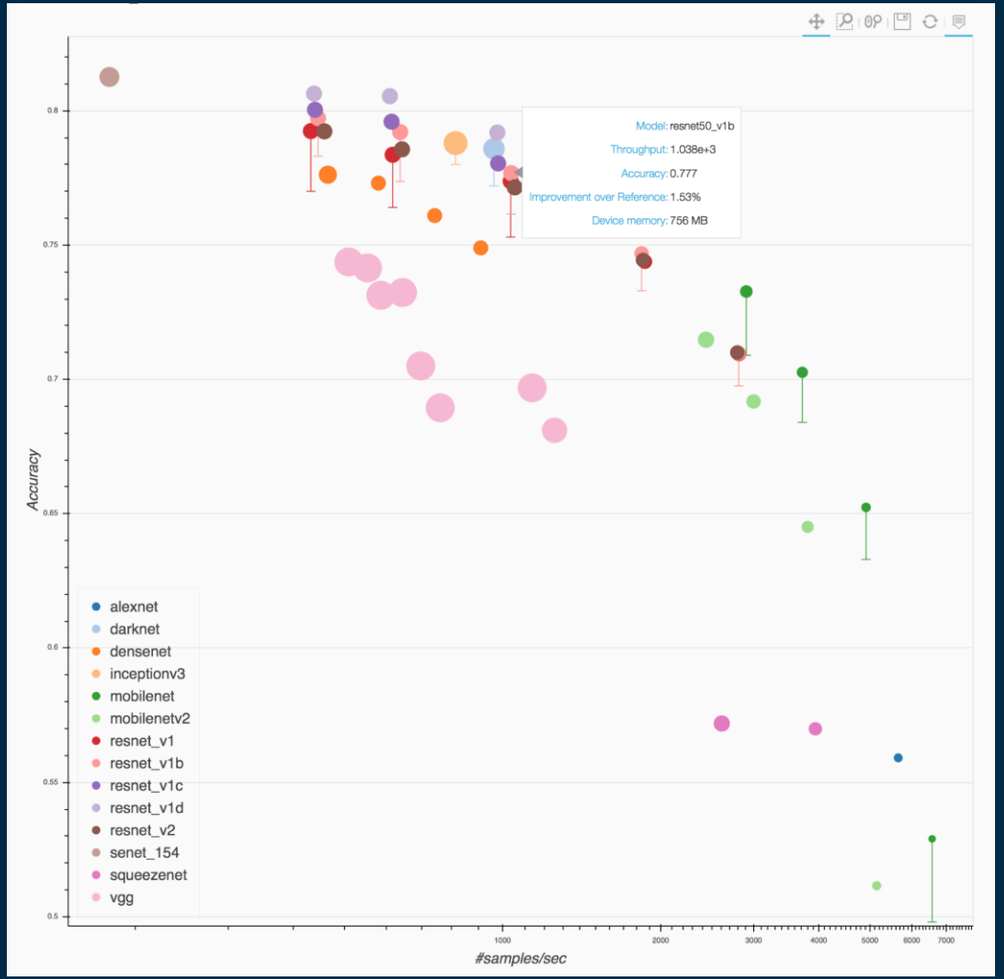
50+ Pre-trained models, with training scripts, datasets, tutorials

https://gluon-cv.mxnet.io

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

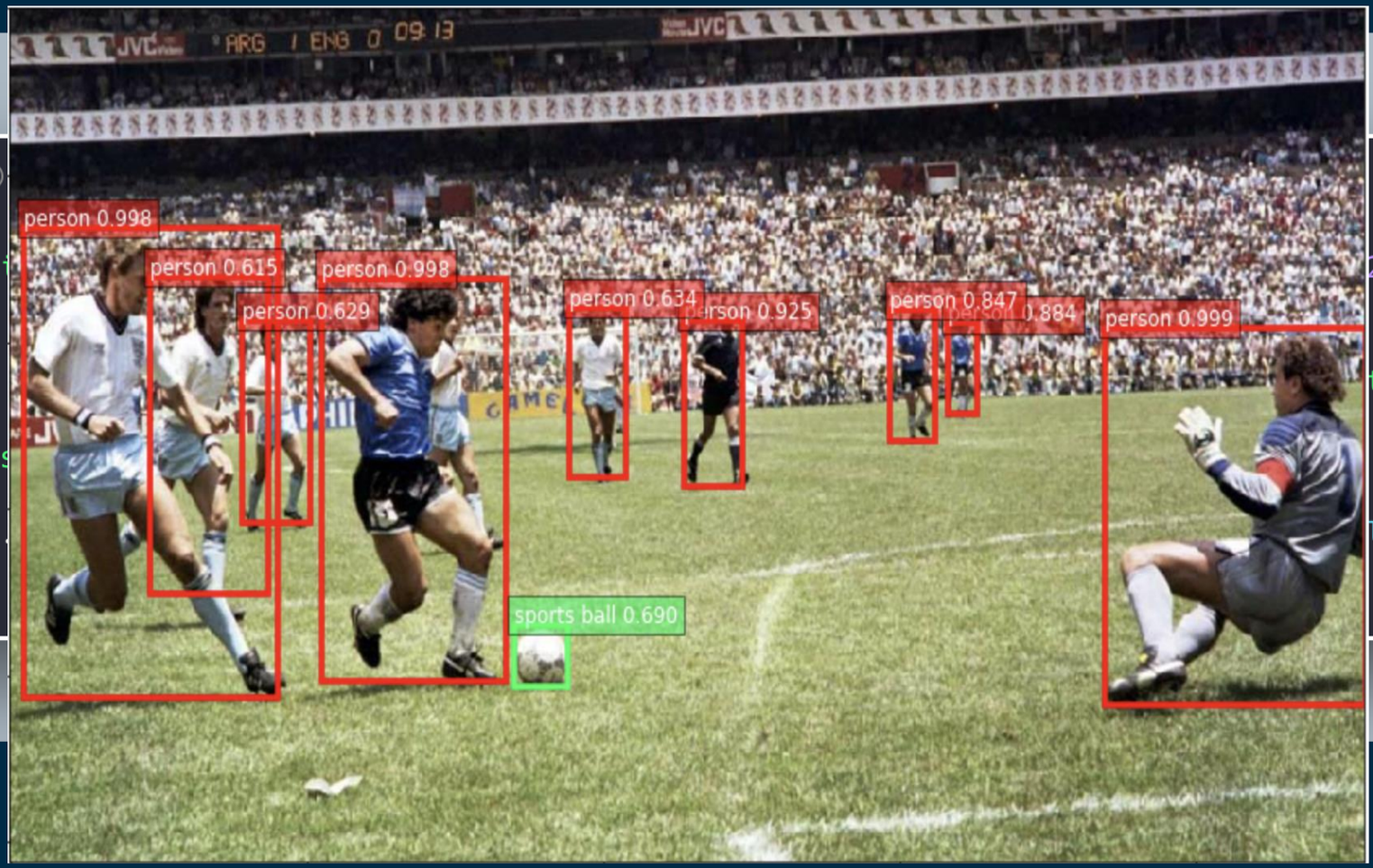
aws

GluonCV pre-trained models



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

GluonCV example code



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Chick-fil-A keeps waffle fries fresh using MXNet

- Track waffle fry freshness
- Identify fries that have exceeded hold time
- Gluon Computer vision model for object detection and tracking
- A team of students with no ML expertise
- 12 months from no ML knowledge to completion

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

GluonNLP: a deep learning toolkit for natural language processing

- 300+ word embedding pre-trained models
- 5 language models
- Neural Machine Translation (Google NMT, Transformer)
- Flexible data pipeline tools
- Public datasets
- NLP examples, e.g. sentiment analysis

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

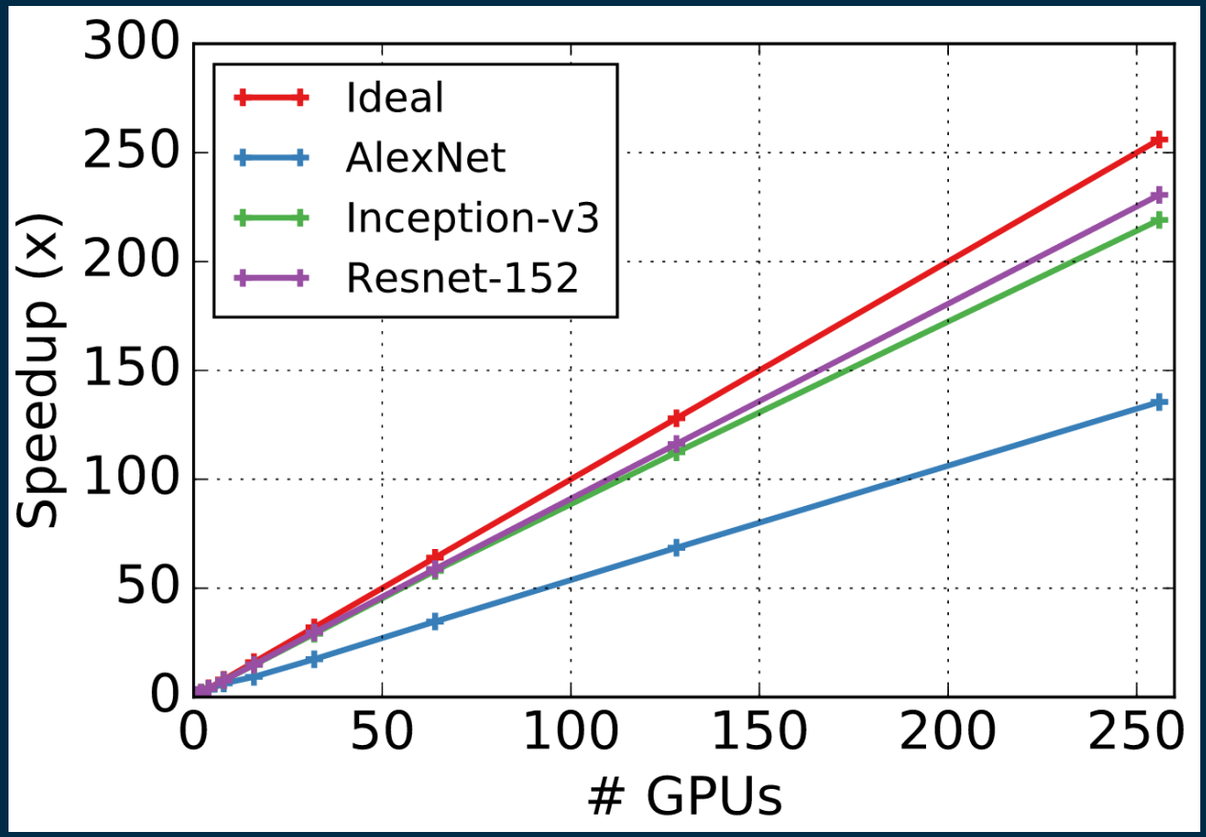
aws

DEV DAY

Training Efficiency

aws

Training efficiency – 92%



https://mxnet.incubator.apache.org/tutorials/vision/large_scale_classification.html

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

TuSimple uses MXNet for autonomous vehicles

- Deep Learning algorithms built with MXNet
- Computer vision and driving simulation
- MXNet teaches computers to Recognize and track objects
- Avoid collisions and prioritize safety
- Largest simulation of its kind
- Simulated a billion miles of road driving with wide range of variables and driving conditions

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

DEV DAY

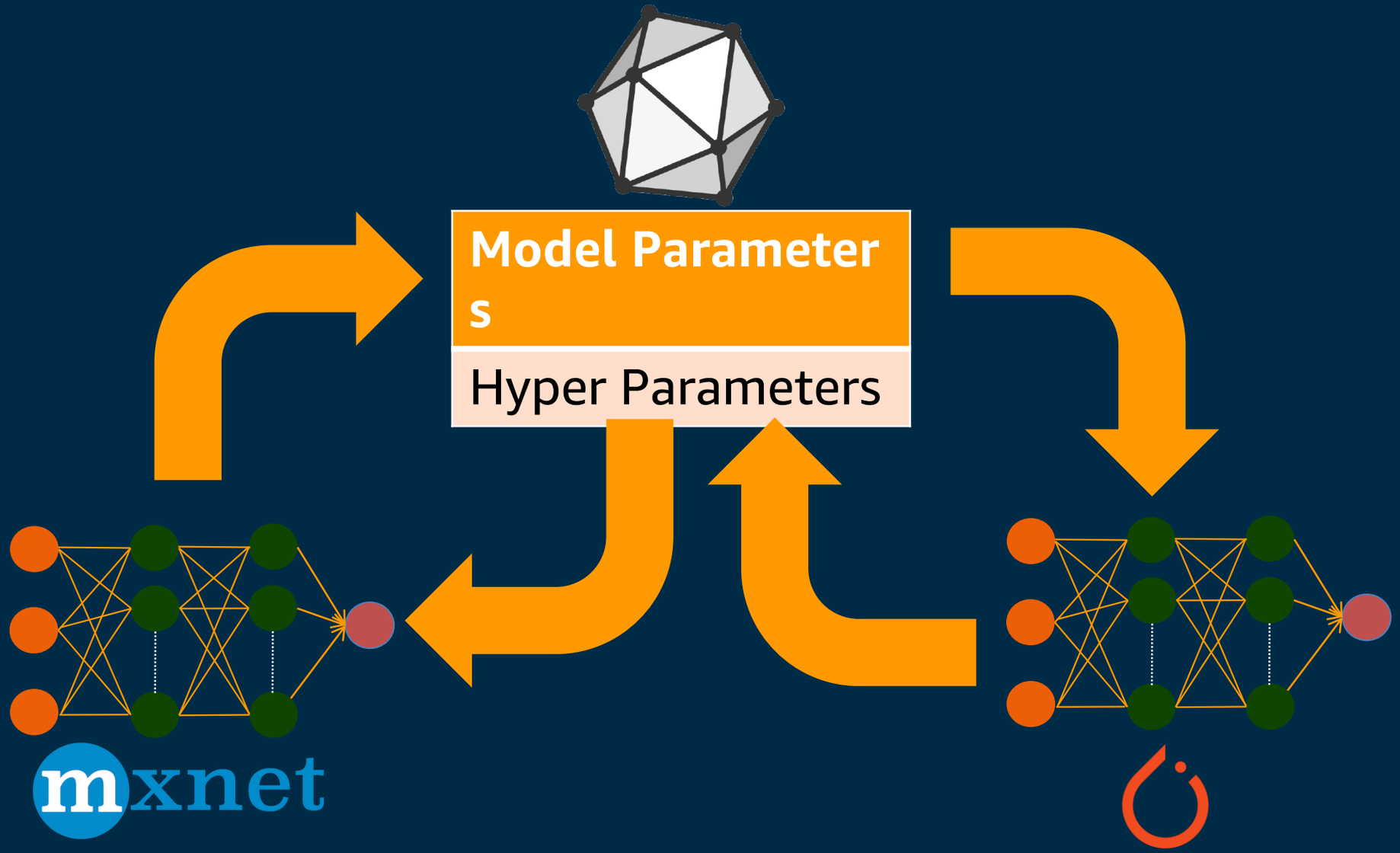
Interoperability

aws

ONNX

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Portability with ONNX



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

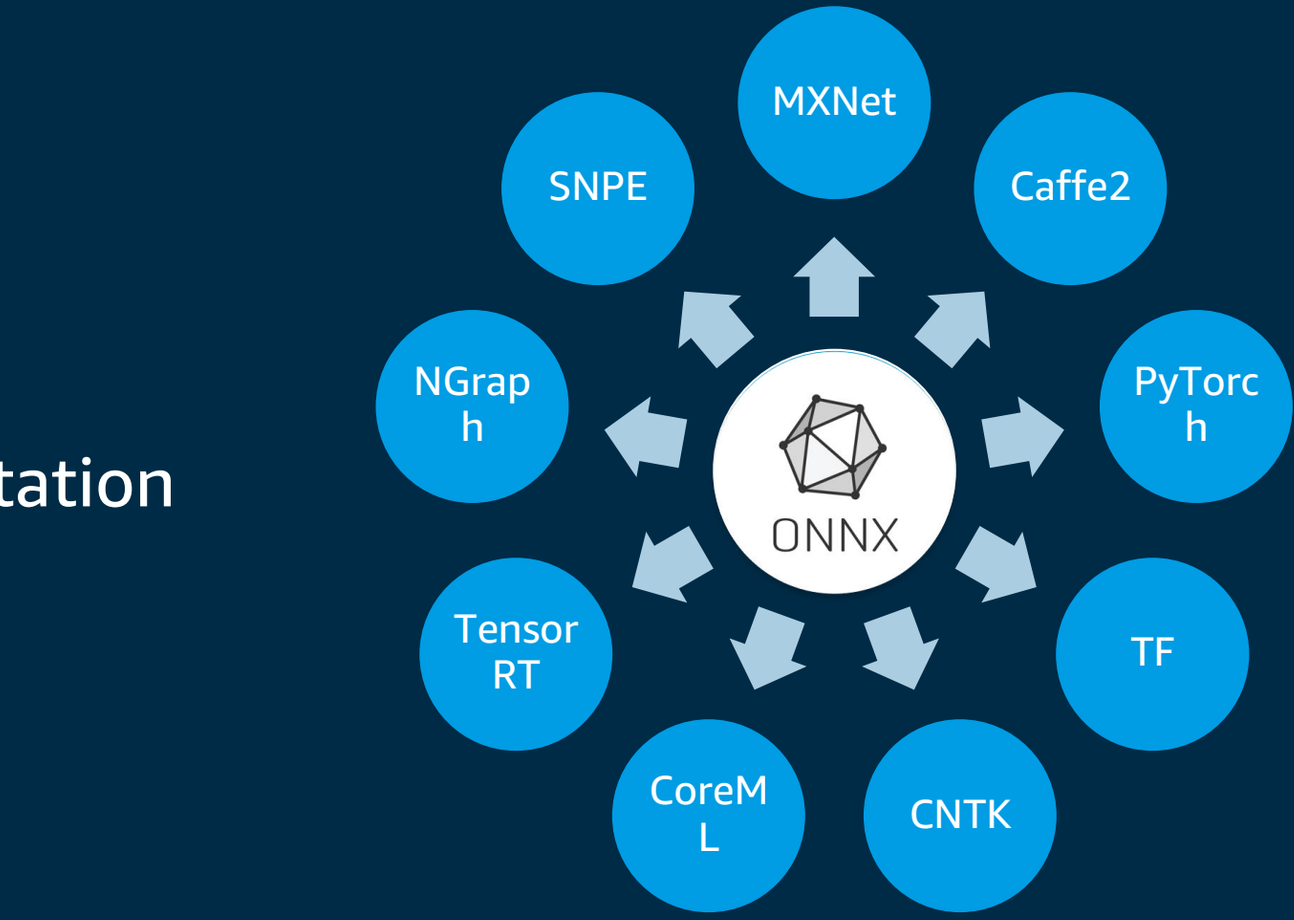
Open Neural Network eXchange - Overview

Many Frameworks

Many Platforms

ONNX: Common Intermediate Representation (IR)

- Open source
- Community driven
- Simple

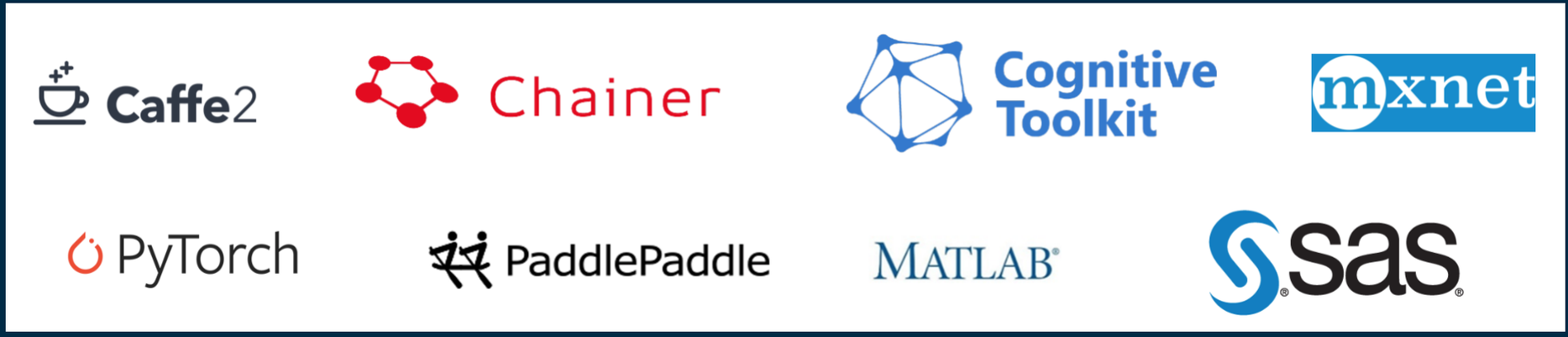


© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Supported tools

Frameworks:



Converters:

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Supported runtimes

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Supported compilers and visualizers

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Keras – MXNet

https://github.com/awslabs/keras-apache-mxnet

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Keras – Apache MXNet

- Deep Learning for Humans
- 2nd most popular Deep Learning framework
- Now Keras users can leverage MXNet's performance

```
pip install mxnet-(mkl|cu92)
pip install keras-mxnet
---
~/.keras/keras.json
backend: mxnet
image_data_format: channels_first
---
```

```
from keras.models import Sequential
model = Sequential()
from keras.layers import Dense
model.add(Dense(units=64, activation='relu', input_dim=100))
model.add(Dense(units=10, activation='softmax'))
model.compile(loss='categorical_crossentropy',
              optimizer='sgd',
              metrics=['accuracy'])
model.fit(x_train, y_train, epochs=5, batch_size=32)
model.train_on_batch(x_batch, y_batch)
loss_and_metrics = model.evaluate(x_test, y_test, batch_size=128)
classes = model.predict(x_test, batch_size=128)
```

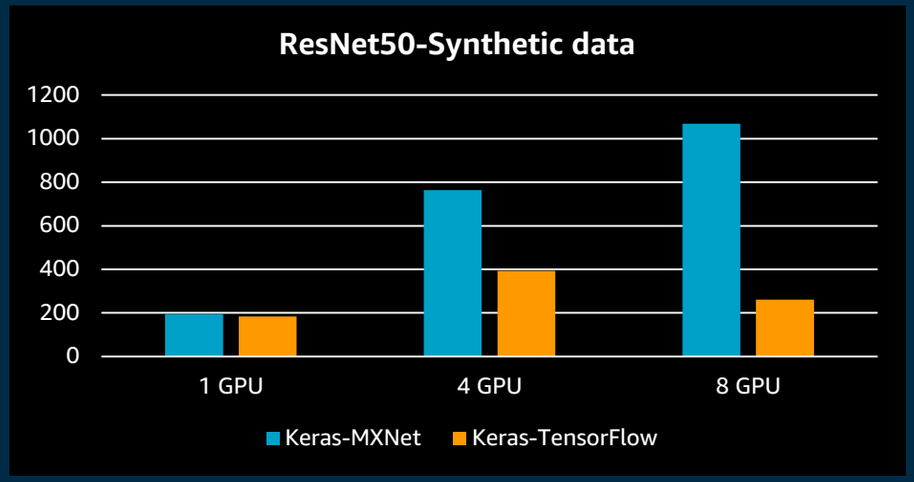
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Keras benchmarks

Setup: https://github.com/awslabs/keras-apache-mxnet/tree/master/benchmark

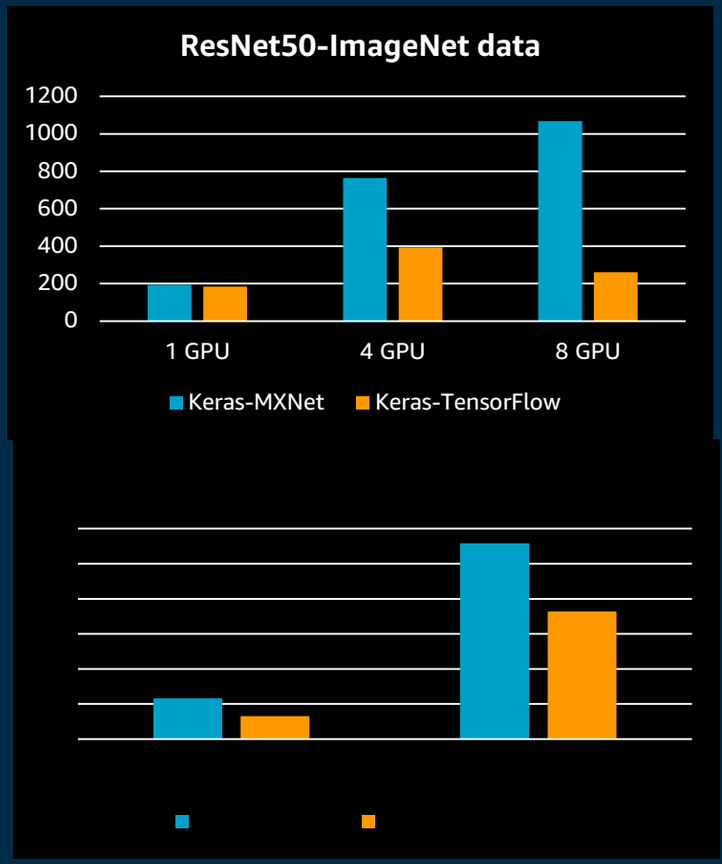
	Training	Inference
Instance	P3.8x Large, P3.16x Large	C5.xLarge, C4.8xLarge
Network	ResNet50v1	ResNet50v1
Batch size	32 * Num of GPUs	32
Image size	3*256*256	3*256*256



GPUs	Keras-MXNet [Image/sec]	Keras-TensorFlow [Image/sec]	Speed Up
1	194	184	1.05
4	764	393	1.94
8	1068	261	4.09

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Keras benchmarks



GPUs	Keras-MXNet	Keras-TensorFlow	Speed Up
1	135	52	2.59
4	536	162	3.30
8	722	211	3.42

Instance	Keras-MXNet	Keras-TensorFlow	Speed Up
C5.X Large	5.79	3.27	1.782
C5.8X Large	27.9	18.2	1.53

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DEV DAY

Inference

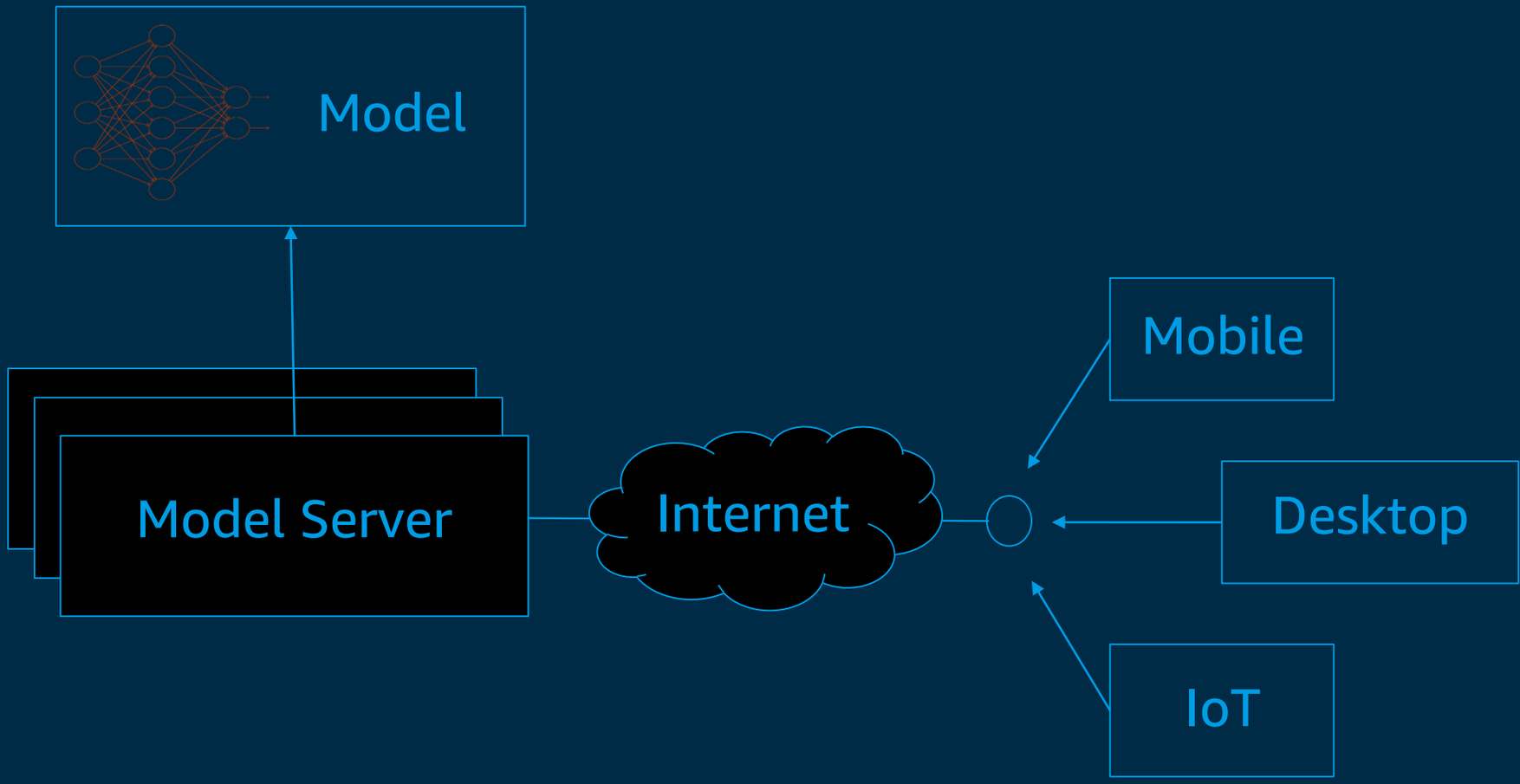
aws

Model Server

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

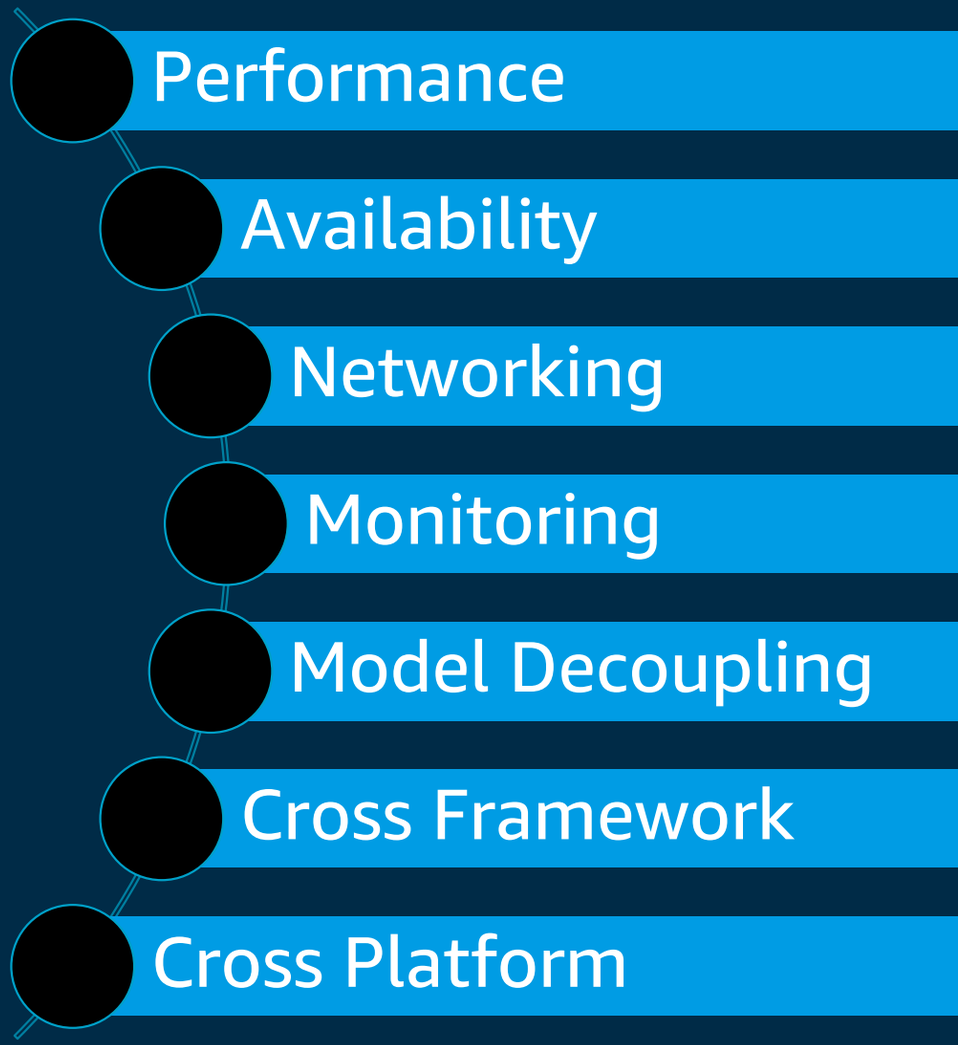
What does a model server look like?

So what does a deployed model looks like?



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Undifferentiated Heavy Lifting of Model Serving



Model Archive

Trained Network

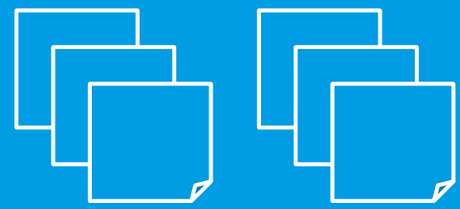
Model Signature

Custom Code

Auxiliary Assets

Model Export CLI

Model Archive

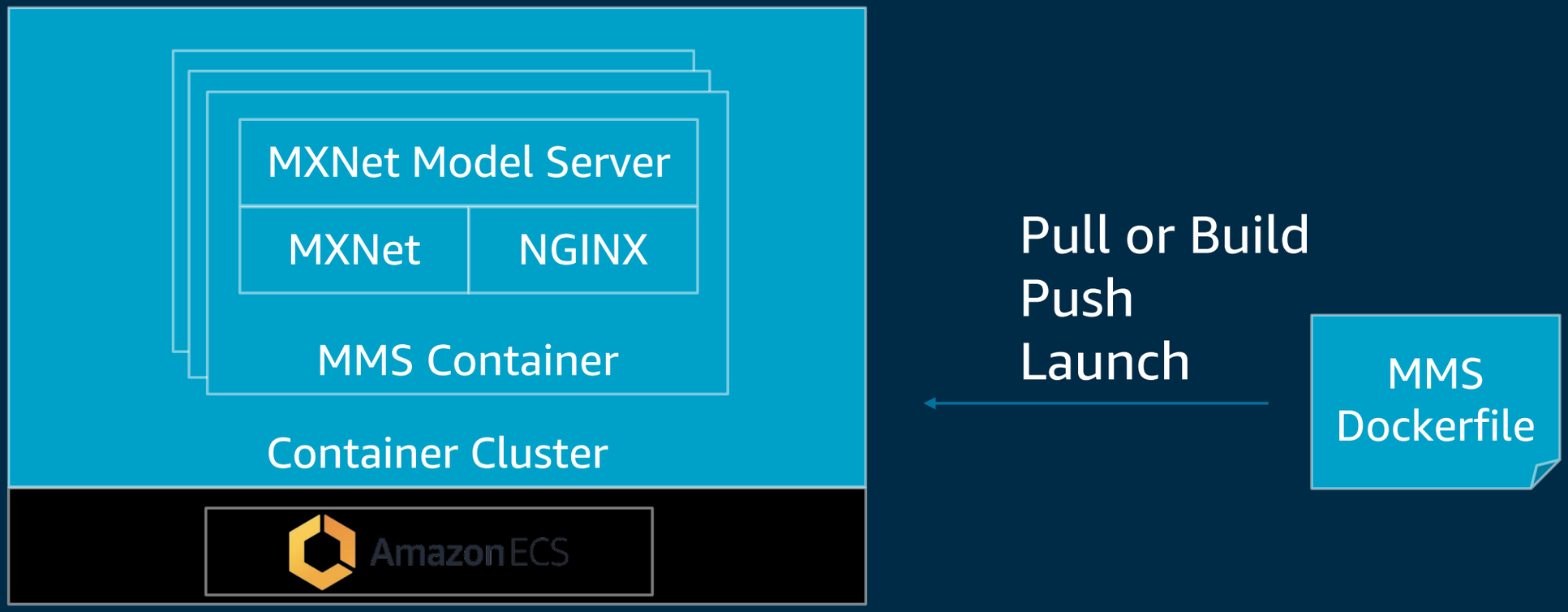


© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Containerization

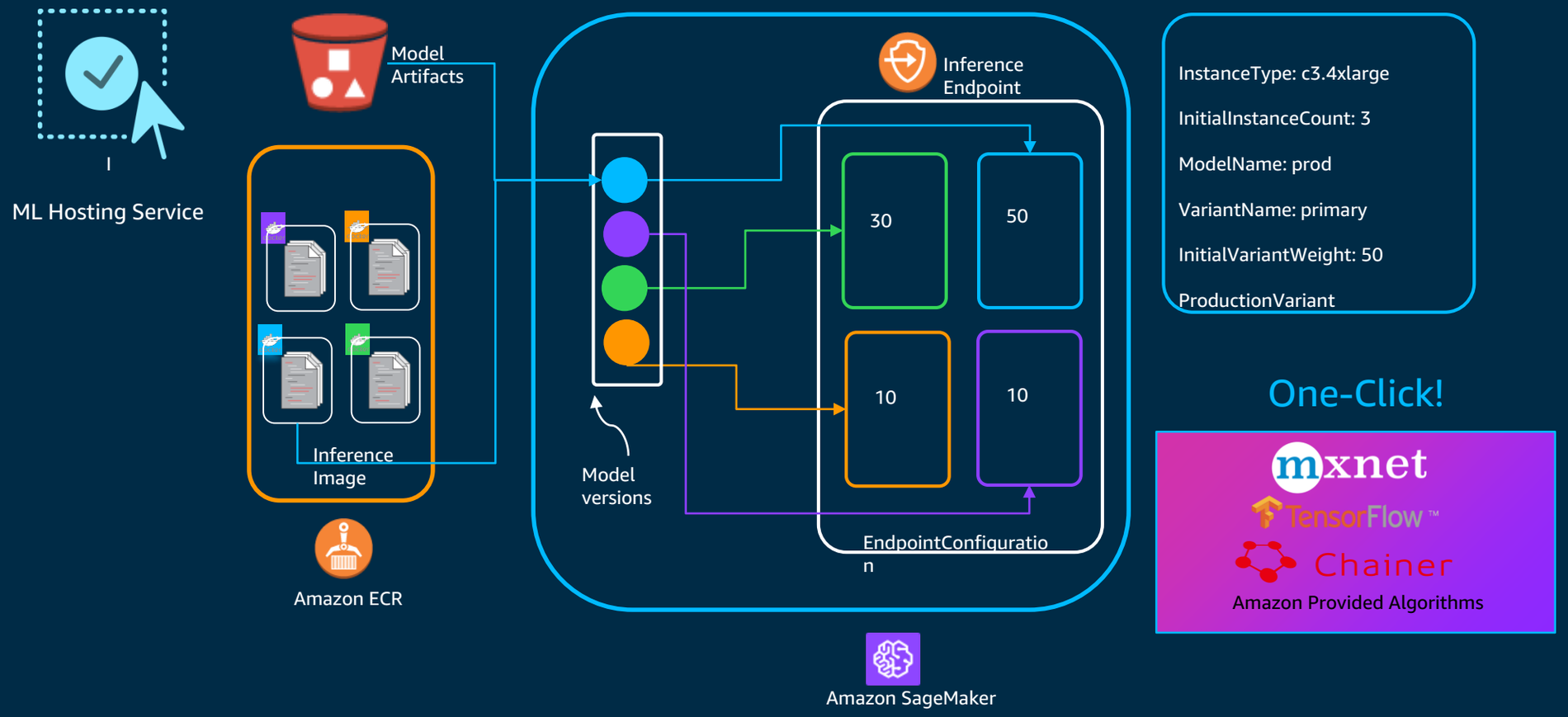
Lightweight virtualization, isolation, runs anywhere



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

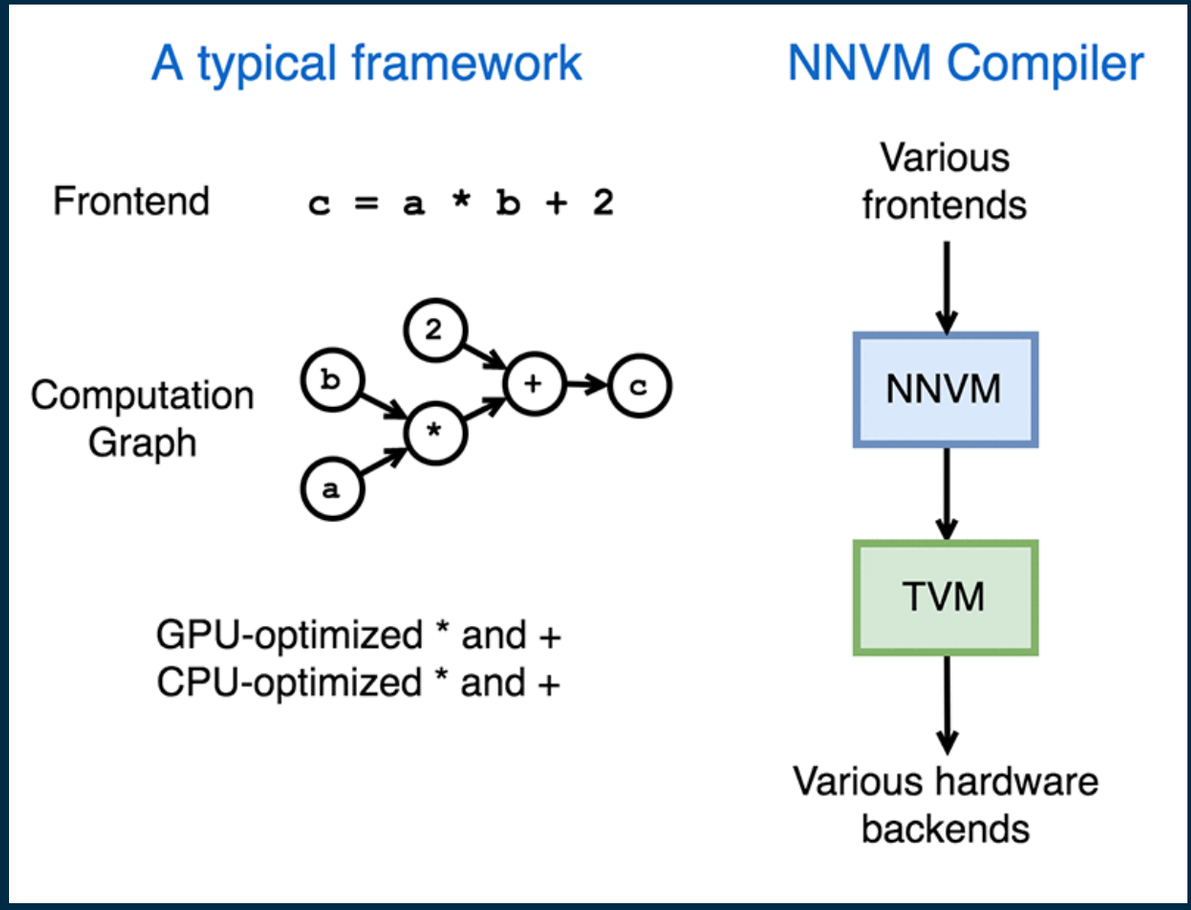
Deployment with Amazon Sage Maker



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

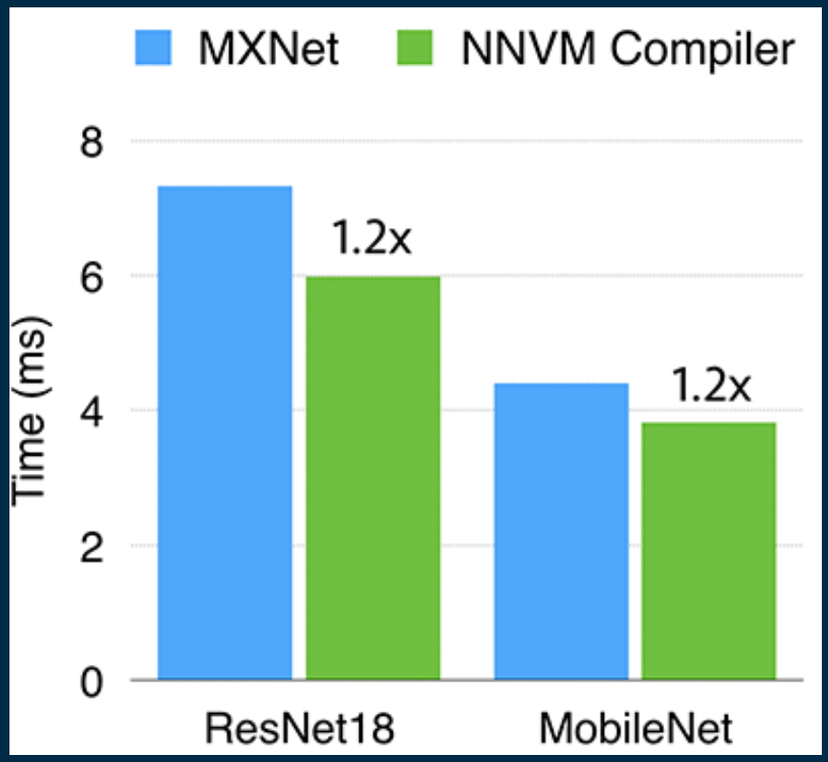
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

TVM and NNVM

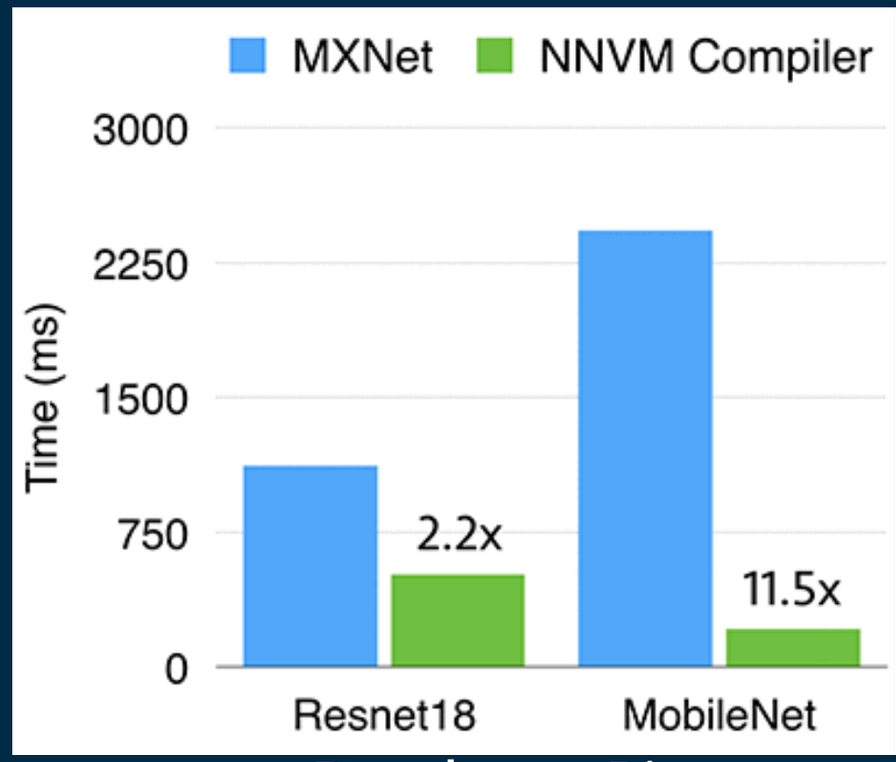


© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inference efficiency – TVM/NNVM



NVIDIA GPU



Raspberry Pi

https://aws.amazon.com/blogs/machine-learning/introducing-nnvm-compiler-a-new-open-end-to-end-compiler-for-ai-frameworks/

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Amazon SageMaker Neo

Train once, run anywhere with 2x the performance

Get accuracy and performance

Automatic optimization

Broad framework support

Broad hardware support

KEY FEATURES

Open-source Neo-AI device runtime and compiler under the Apache software license; 1/10th the size of original frameworks

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Deep Learning acceleration

CUDA & CuDNN

pip install mxnet-cu92

TensorRT

pip install mxnet-tensorrt-cu92

MKL, MKLML & MKLDNN

e.g. pip install mxnet-mkl

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

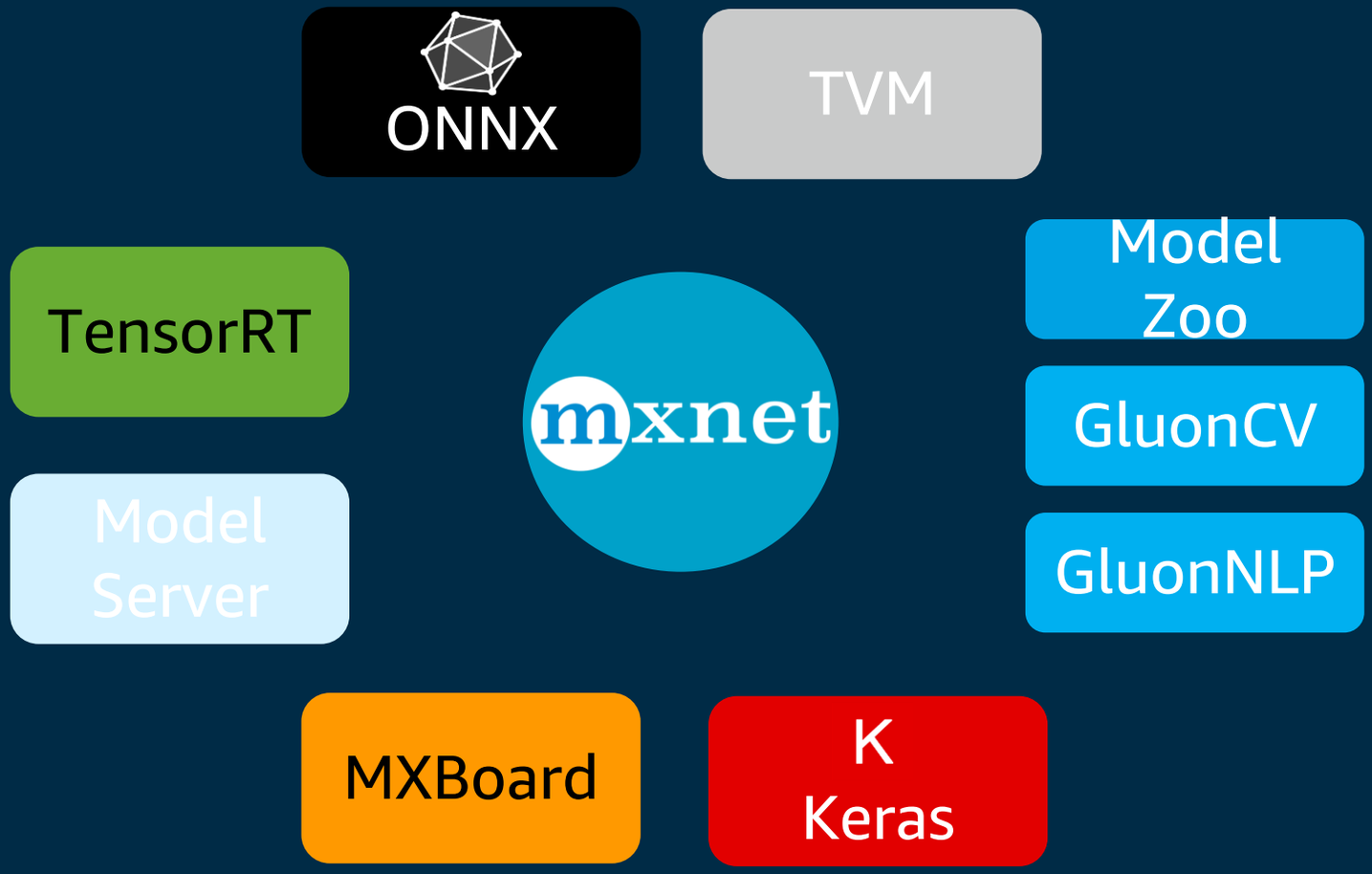
aws

DEV DAY

Now it’s Your Turn!

aws

Apache MXNet ecosystem

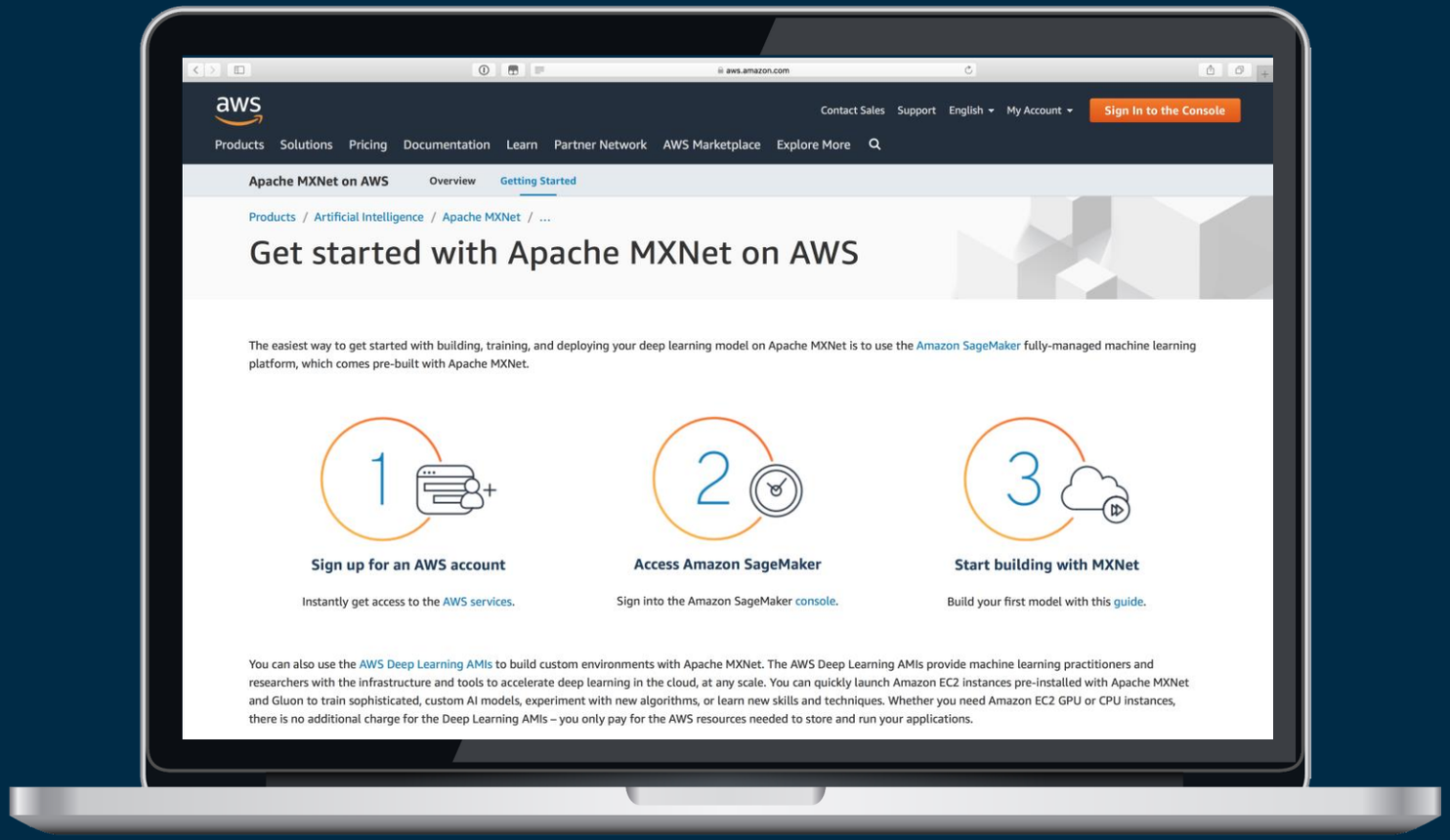


© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Apache MXNet customers

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Getting started with Apache MXNet on AWS



https://aws.amazon.com/mxnet/get-started/

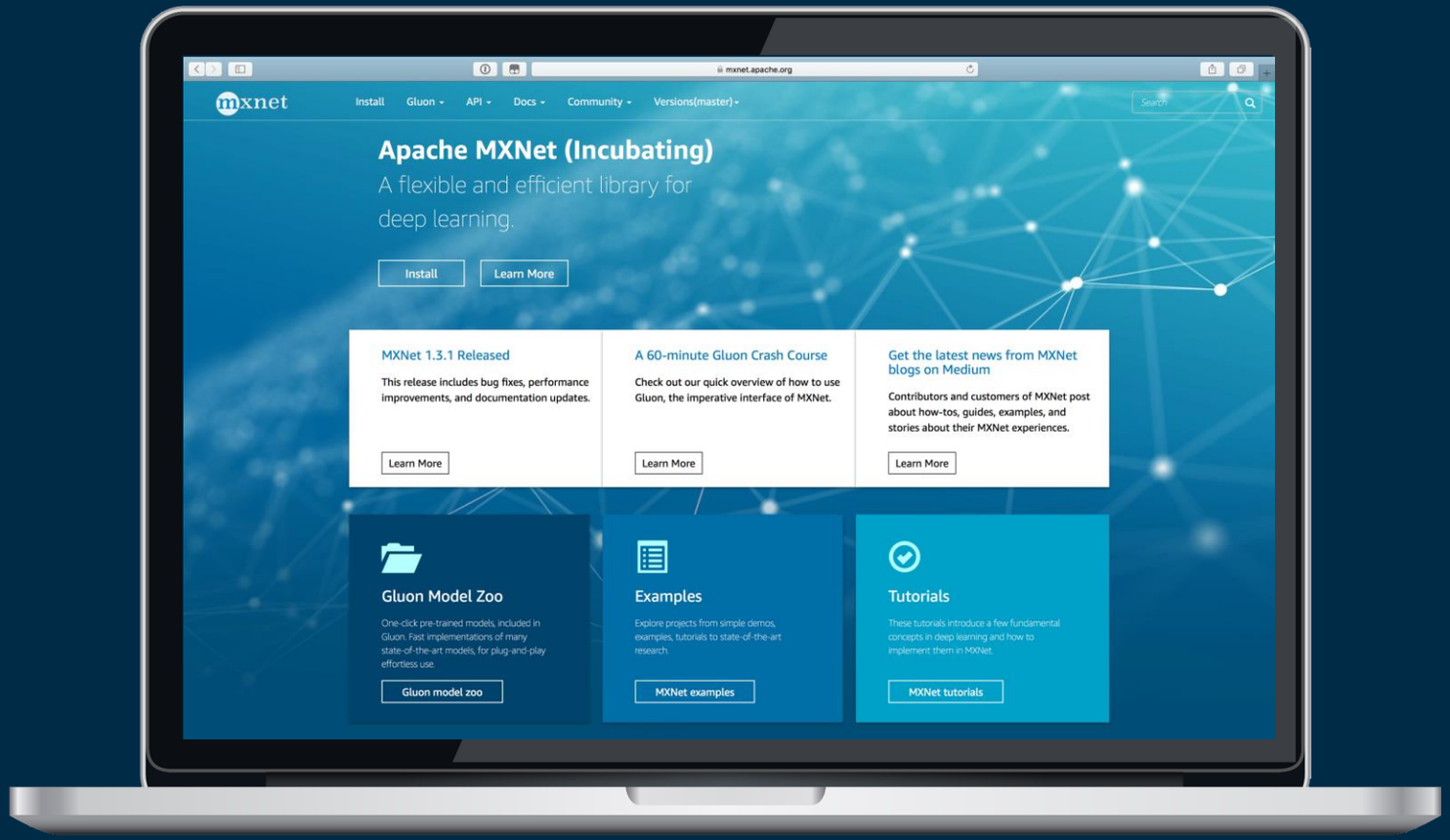
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Using Apache MXNet with AWS ML services

- Amazon SageMaker: aws.amazon.com/sagemaker
- Amazon SageMaker Neo: aws.amazon.com/sagemaker/neo
- Amazon SageMaker Reinforcement Learning: aws.amazon.com/about-aws/whats-new/2018/11/amazon-sagemaker-announces-support-for-reinforcement-learning
- Amazon Elastic Inference: aws.amazon.com/machine-learning/elastic-inference
- AWS IoT Greengrass ML Inference: aws.amazon.com/greengrass/ml
- Dynamic Training with Apache MXNet on AWS: https://aws.amazon.com/about-aws/whats-new/2018/11/introducing-dynamic-training-with-apache-mxnet/

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Project home page



https://mxnet.apache.org

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Staying in touch

- GitHub: github.com/apache/incubator-mxnet
- Discussion forum: discuss.mxnet.io
- Blog: medium.com/apache-mxnet
- SlideShare: slideshare.net/apachemxnet
- Twitter: @ApacheMXNet
- YouTube : youtube.com/apachemxnet
- Reddit: r/mxnet/
- Meetup: meetup.com/pro/deep-learning-with-apache-mxnet

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Contact us!

mxnet-info@amazon.com

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Summary

- Efficient distributed training
- Portability
- Efficient Inference
- Inference on Edge

- Easy Coding
- Easy Debugging
- Toolkits for Rapid Prototyping

Amazon SageMaker

- End-2-End Plarform
- Zero Setup
- Distributed Training
- AB/Testing
- Scalable Endpoints
- Automatic Model Tuning

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DEV DAY

Thank you!

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

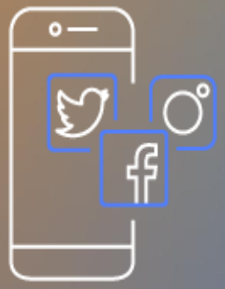
DEV DAY
SEOUL

19.09.26

여러분의 피드백을 기다립니다!

강연 평가 및 설문 조사
QR 코드를 통해 AWS DEV DAY SEOUL에 대한 여러분의 의견을 공유해주세요.
강연 평가 및 설문 조사에 참여해 주신 분께는 등록데스크에서 특별한 기념품을 드립니다.

강연 영상
AWS DEV DAY SEOUL 강연 영상은 행사 종료 후 메일로 공유드릴 예정입니다.



#AWSDEVDAYSEOUL
소셜미디어에 행사 참여 소감을 공유해주세요!

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.