IDF16
INTEL DEVELOPER FORUM

Deep Learning Frameworks and Optimization Paths on Intel® Architecture

Andres Rodriguez – Solutions Architect, Intel Data Center Group
Ravi Panchumarthy – Systems Engineer, Intel Data Center Group
Elvis Jones – Solutions Architect, Amazon Web Services

Agenda

- Deep learning and frameworks overview
- Optimizations in Intel® Math Kernel Library (Intel® MKL) and Intel® MKL-DNN
- Intel optimized frameworks on Amazon Web Services (AWS*)

Agenda

- **Deep learning and frameworks overview**
- Optimizations in Intel® Math Kernel Library (Intel® MKL) and Intel® MKL-DNN
- Intel optimized frameworks on Amazon Web Services (AWS*)

Classification

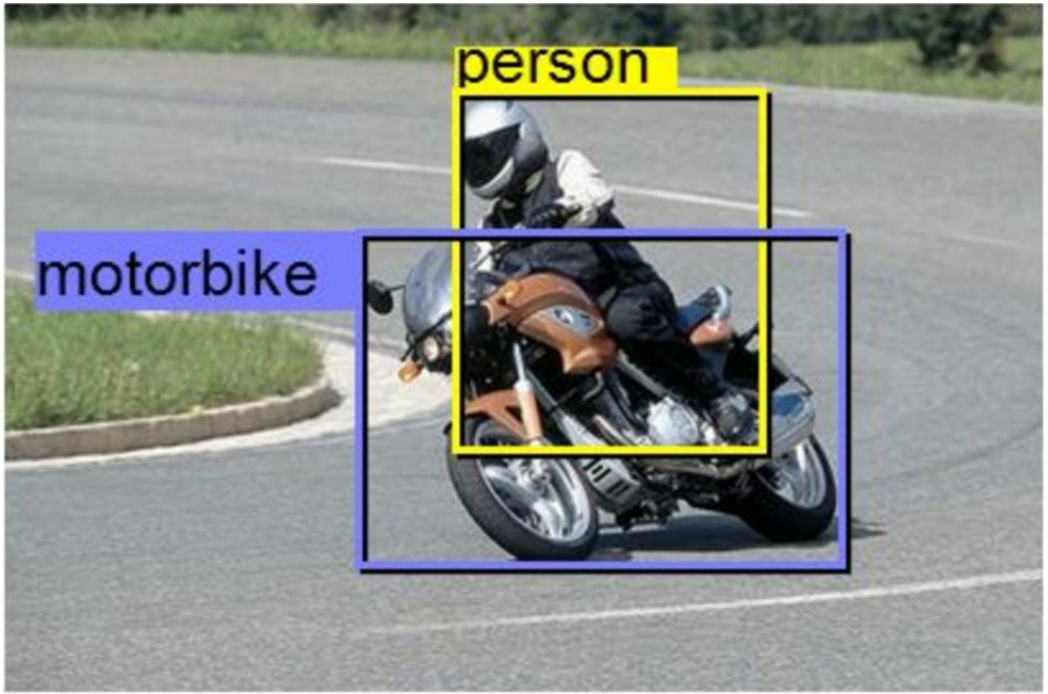
Label the image

- Person
- Motorcyclist
- Bike

https://people.eecs.berkeley.edu/~jhoffman/talks/lsda-baylearn2014.pdf

Detection

Detect and
label objects



https://people.eecs.berkeley.edu/~jhoffman/talks/lsda-baylearn2014.pdf

Semantic Segmentation

Label every pixel

https://people.eecs.berkeley.edu/~jhoffman/talks/lsda-baylearn2014.pdf

Natural Language Object Retrieval

a scene with three people

query='*man far right*'

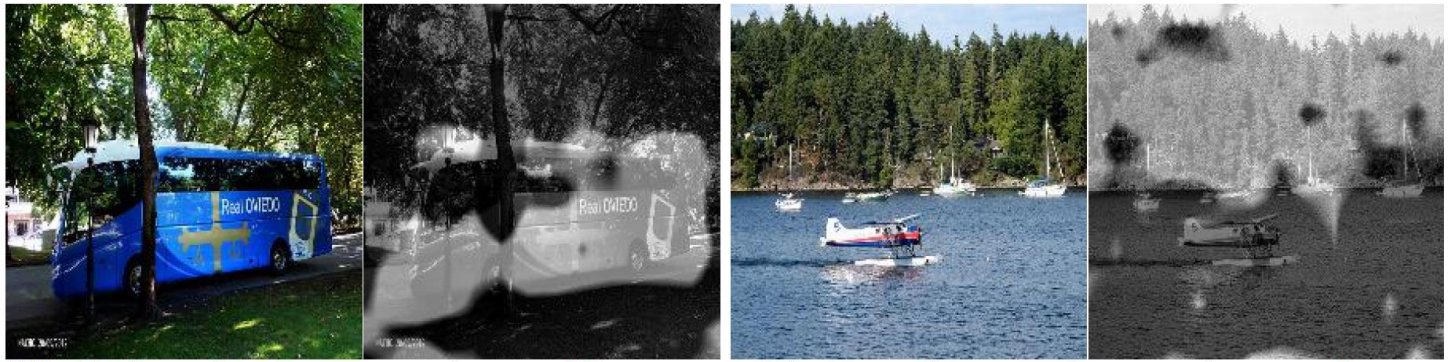
query='*left guy*'

query='*cyclist*'

http://arxiv.org/pdf/1511.04164v3.pdf

IDF16
INTEL DEVELOPER FORUM

Visual and Textual Question Answering



What is the main color on the bus ?

Answer: blue

What type of trees are in the background ?

Answer: pine

How many pink flags are there ?

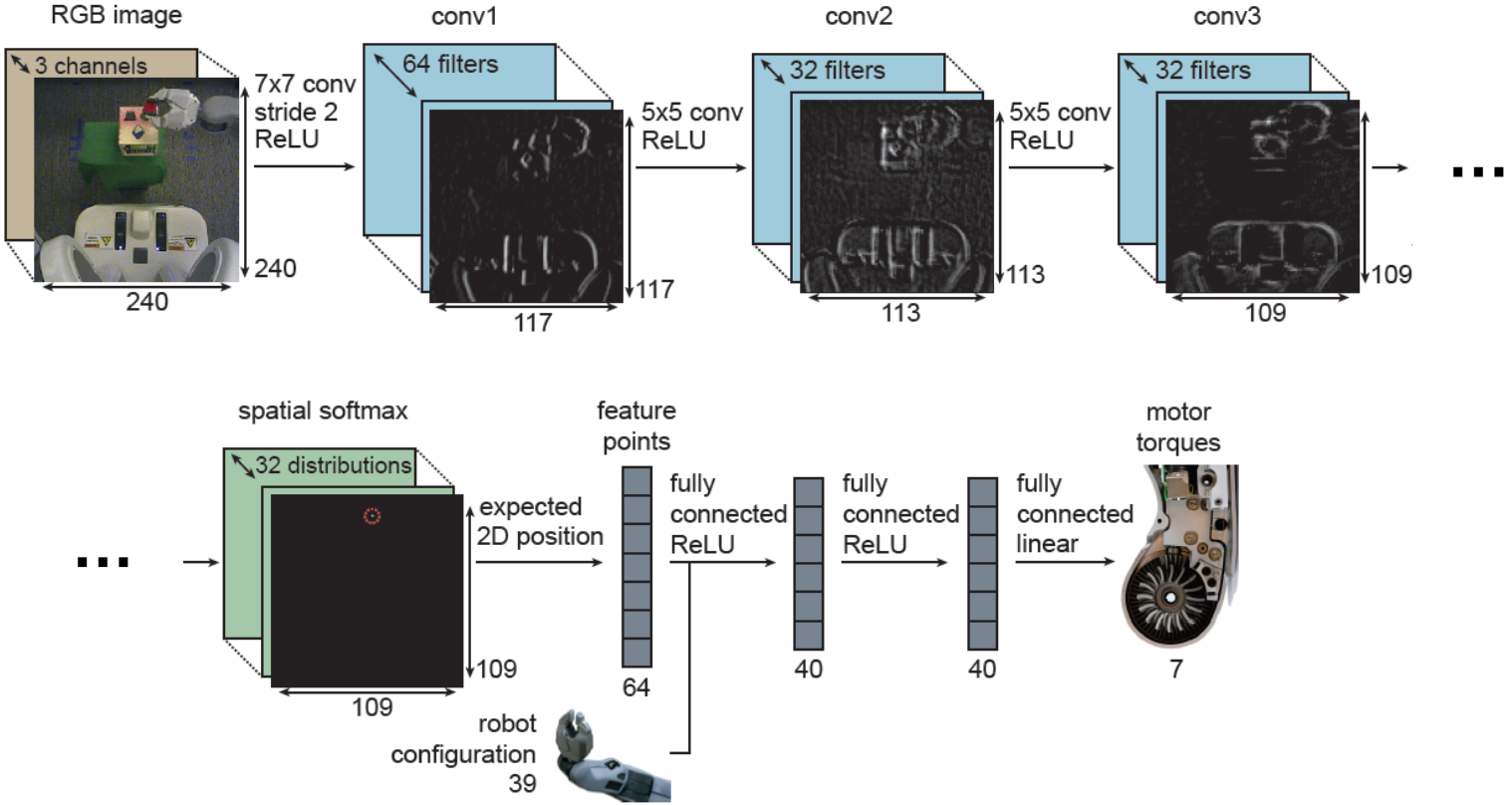
Answer: 2

Is this in the wild ?

Answer: no

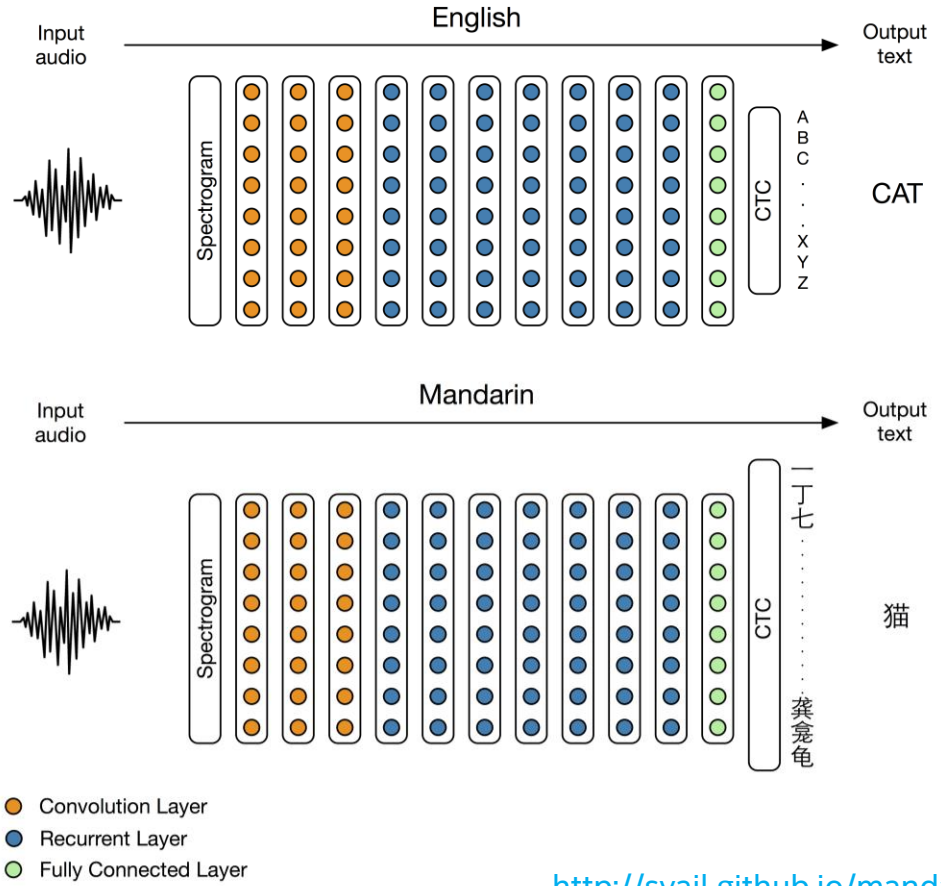
https://arxiv.org/pdf/1603.01417v1.pdfp://arxiv.org/abs/1603.01417

Visuomotor Control



https://arxiv.org/pdf/1504.00702v5.pdf

Speech Recognition



The same architecture is used for English and Mandarin Chinese speech recognition

http://svail.github.io/mandarin/

Q&A Natural Language Understanding

Question: Where was Mary before the Bedroom?
Answer: Cinema.

Facts	Episode 1	Episode 2	Episode 3
Yesterday Julie traveled to the school.			
Yesterday Marie went to the cinema.			
This morning Julie traveled to the kitchen.			
Bill went back to the cinema yesterday.			
Mary went to the bedroom this morning.			
Julie went back to the bedroom this afternoon.			
[done reading]			

https://arxiv.org/pdf/1506.07285.pdf

Personal Assistant

amazon echo

Microsoft

Hi. I'm Cortana.
Ask me a question!

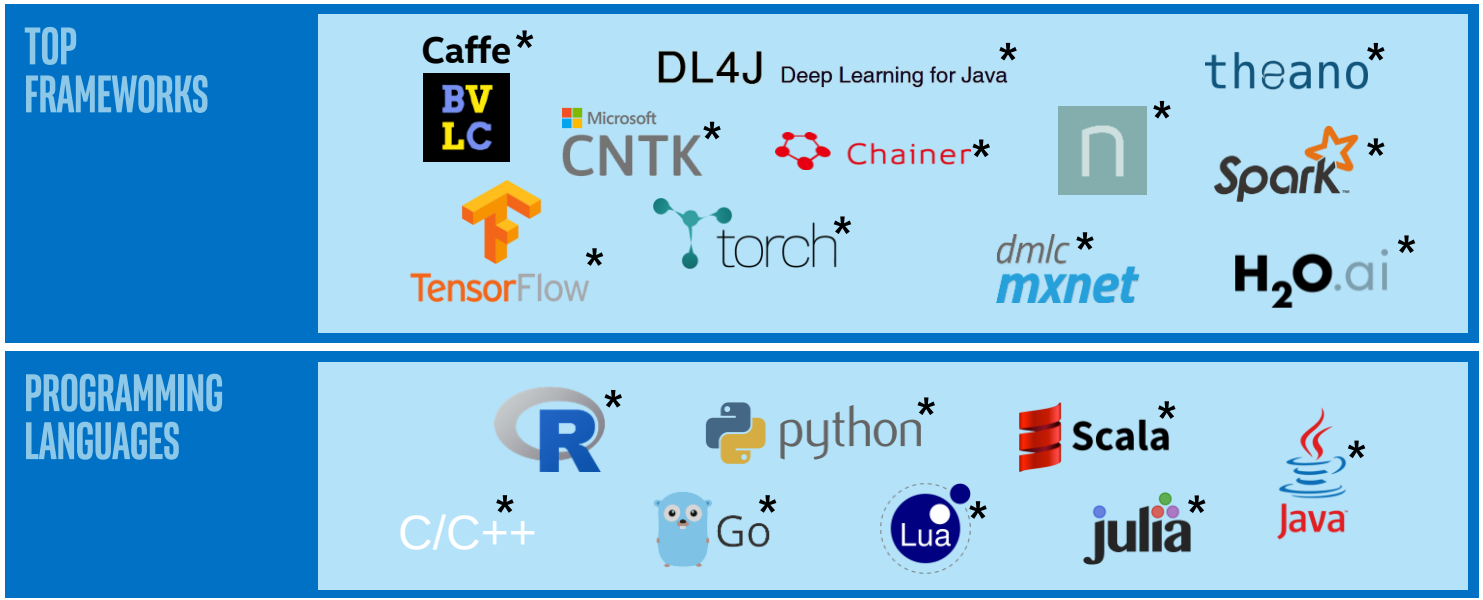
Siri

IDF16
INTEL DEVELOPER FORUM

DL Tools

Machine Learning

Autonomous computation methods that learn from experience (data)

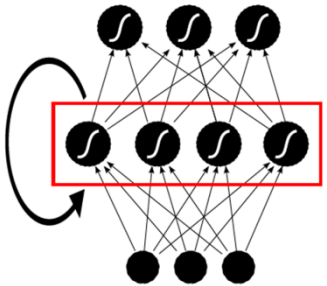


Agenda

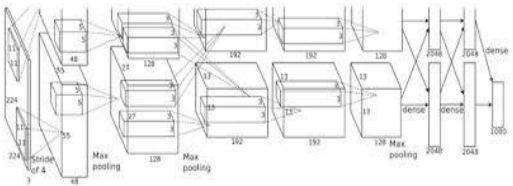
- Deep learning and frameworks overview
- **Optimizations in Intel® Math Kernel Library (Intel® MKL) and Intel® MKL-DNN**
- Intel optimized frameworks on Amazon Web Services (AWS*)

IDF16
INTEL DEVELOPER FORUM

Diversity in Deep Networks



Recurrent NN



CNN – AlexNet*

GoogLeNet*

Variety in Network Topology

- Recurrent NNs common for NLP/ASR, DAG for GoogLeNet, Networks with memory…

But there are a few well defined building blocks

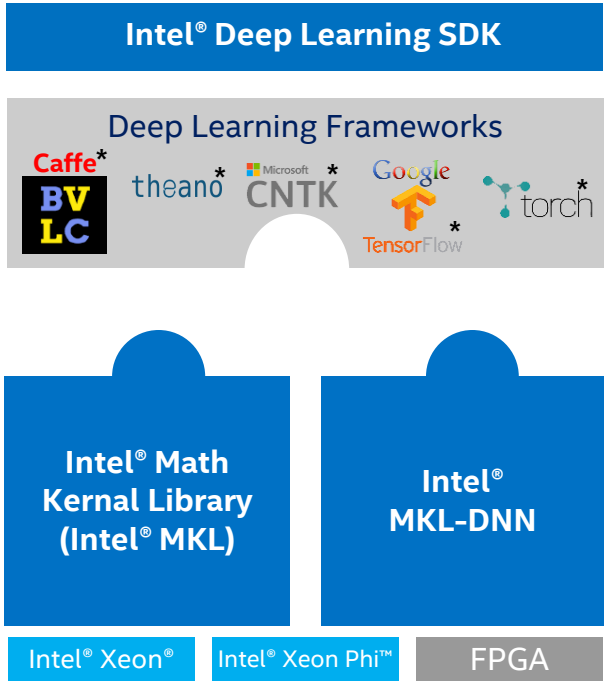
- Convolutions common for image recognition tasks
- GEMMs for recurrent network layers—could be sparse
- ReLU, tanh, softmax

IDF16
INTEL DEVELOPER FORUM

Intel® Math Kernel Library (Intel® MKL)

- Optimized AVX-2 and AVX-512 instructions
- Intel® Xeon® and Intel® Xeon Phi™ processors
- Supports all common layers types
- Coming soon: Winograd-based convolutions

Intel Deep Learning Software Stack



Intel Deep Learning SDK – free tools to accelerate design, training and deployment of deep networks

Targeted release: Q4' 2016

Intel® MKL-DNN – free open source DNN functions designed for max Intel HW performance and high-velocity integration with DL frameworks

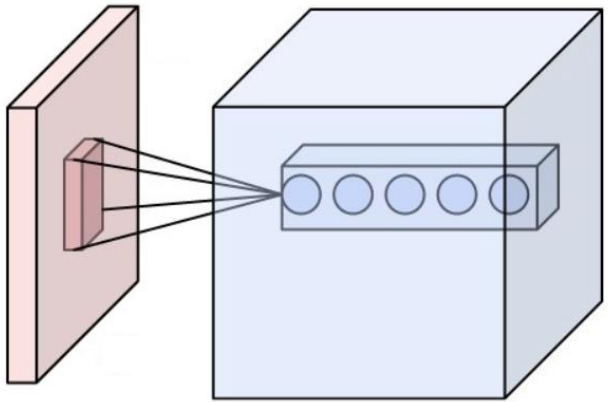
Targeted release: Q4' 2016 (APIs and preview Q3' 2016)

- Open source DNN functions included in MKL 2017
- IA optimizations contributed by community
- Binary GEMM functions
- Apache* 2 license

Intel libraries as path to bring optimized ML/DL frameworks to Intel hardware

IDF16 INTEL DEVELOPER FORUM

Naïve Convolution



```
1: for i0 ∈ 0, . . . , minibatch do
2:     for i1 ∈ 0, . . . , ifm do
3:         for i2 ∈ 0, . . . , ofm do
4:             for i3 ∈ 0, . . . , outh do
5:                 for i4 ∈ 0, . . . , outw do
6:                     for i5 ∈ 0, . . . , kh do
7:                         for i6 ∈ 0, . . . , kw do
8:                             output[i0, i1, i3, i4]+ =
9:                             input[i0, i1, i3 ∗ s + i5 − 1, i4 ∗ s + i6 −
                               1] ∗ wts[i1, i2, i5, i6]
```

https://en.wikipedia.org/wiki/Convolutional_neural_network

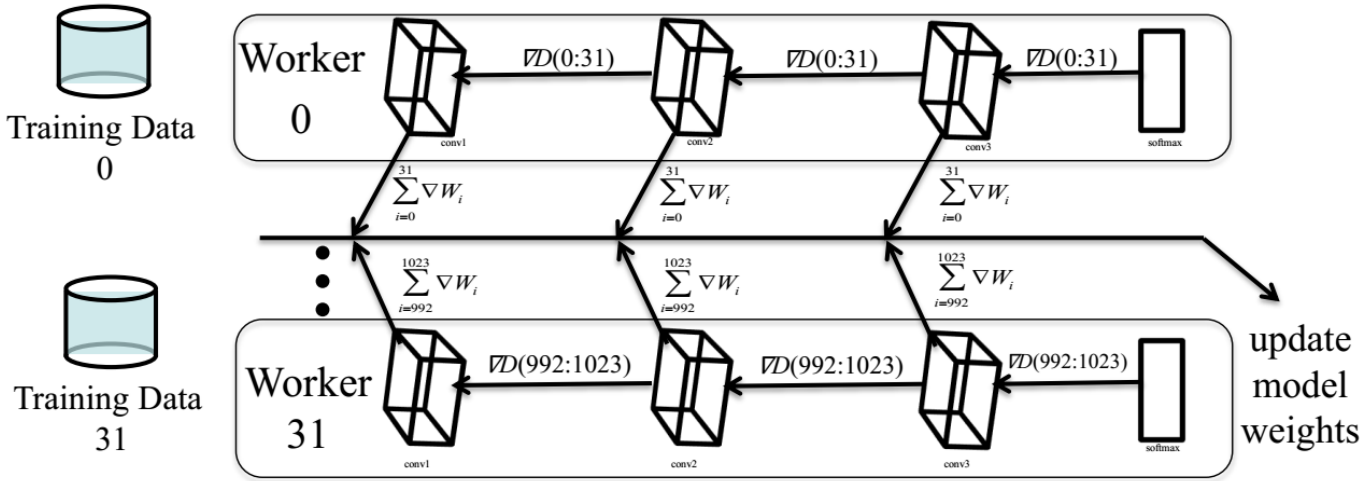
Cache Friendly Convolution

```
1:  for i0 ∈ 0,...,minibatch do
2:    for i1 ∈ 0,...,ifm/SW do
3:      for i2 ∈ 0,...,ofm/SW do
4:        for i3 ∈ 0,...,outh/RBh do
5:          for i4 ∈ 0,...,outw/RBw do
6:            for rbh ∈ 0,...,RBh do
7:              for rbw ∈ 0,...,RBw do
8:                reg = rbh * RBw + rbw
9:                outy = i3 * RBh + rbh
10:               outx = i4 * RBw + rbw
11:               vout[reg] =
                  LOAD(output[i0][i2][outy][outx])
12:             for i5 ∈ 0,...,SW do
13:               for i6 ∈ khstart,...,khend do
14:                 for i7 ∈ kwstart,...,kwend do
15:                   vwt = LOAD(wts[i1 * SW +
                      i5][i2][i6][i7][0])
16:                   for i8 ∈ 0,...,RBh do
16:                     for i8 ∈ 0,...,RBh do
17:                       for i9 ∈ 0,...,RBw do
18:                         reg = i8 * RBw + i9
19:                         outy = i3 * RBh + i8
20:                         outx = i4 * RBw + i9
21:                         inpy = outy * stride + i6 − 1
22:                         inpx = outx * stride + i7 − 1
23:                         vout[reg] =
                            VFMA(vout[reg],
                            bcast(input[i0][i1][outy][outx][0]))
                            , vwt)
24:   for rbh ∈ 0,...,RBh do
25:     for rbw ∈ 0,...,RBw do
26:       reg = rbh * RBw + rbw
27:       outy = i3 * RBh + rbh
28:       outx = i4 * RBw + rbw
29:       STORE(vout[reg], output[i0][i2][outy][outx])
```

https://arxiv.org/pdf/1602.06709v1.pdf

Caffe* Optimized for Intel® Architecture

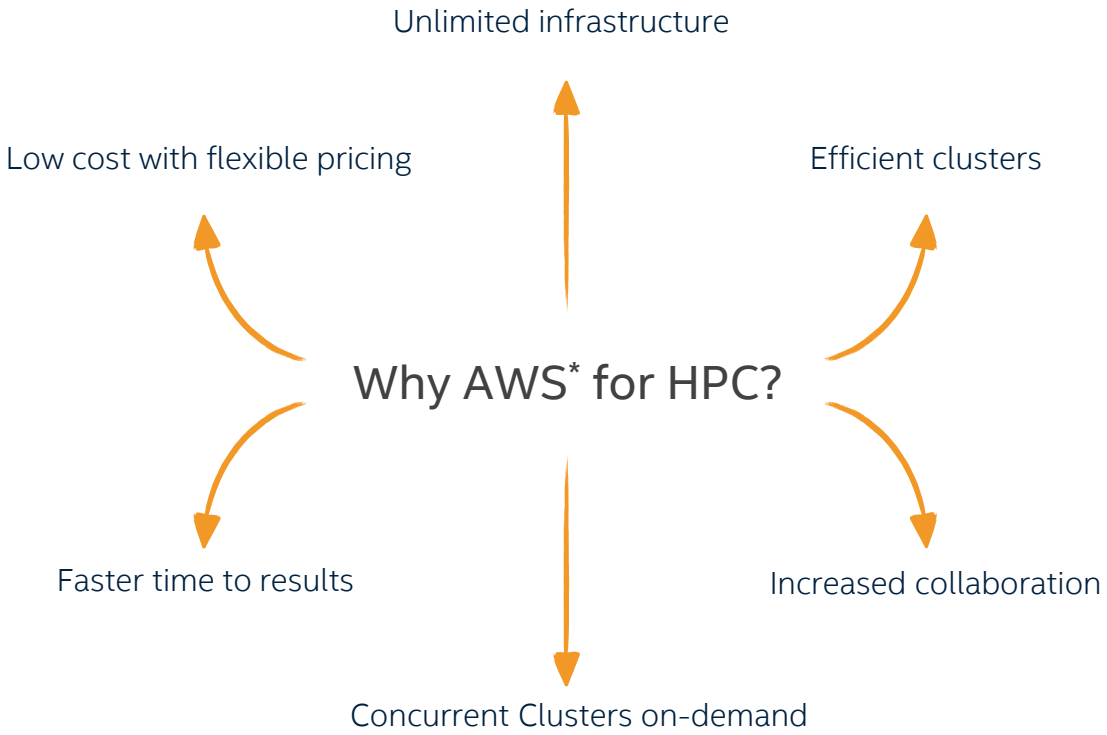
- All the goodness of BVLC Caffe* +
  - Integrated with Intel® Math Kernel Library (Intel® MKL) 2017
  - Multi-node distributed training



Forrest Iandola, et al., "Scaling DNN Training on Intel Platforms." 2016

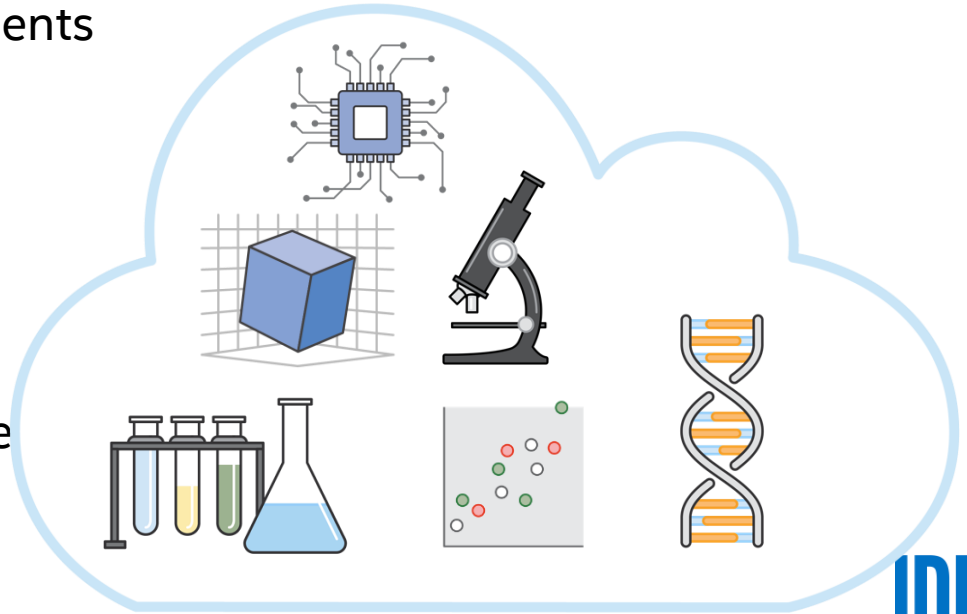
Agenda

- Deep learning and frameworks overview
- Optimizations in Intel® Math Kernel Library (Intel® MKL) and Intel® MKL-DNN
- **Intel optimized frameworks on Amazon Web Services (AWS*)**


Unlimited infrastructure
Low cost with flexible pricing
Efficient clusters
Why AWS* for HPC?
Faster time to results
Increased collaboration
Concurrent Clusters on-demand

How is Amazon Web Services (AWS*) Used for HPC?

- High Performance Computing (HPC) for Engineering and Simulation
- High Throughput Computing (HTC) for Data-Intensive Analytics
- Collaborative Research Environments
- Monte-Carlo Simulations
- Data Visualization
- Hybrid Supercomputing centers
- Citizen Science
- Engineering/Science-as-a-Service
- Internet of Things (IOT)
- Serverless Computing



Goal: A simplified, repeatable, process

Goal = a simplified, repeatable, process for creating ML/DL clusters

CloudFormation* CfnCluster

Amazon Web Services (AWS*) CloudFormation*

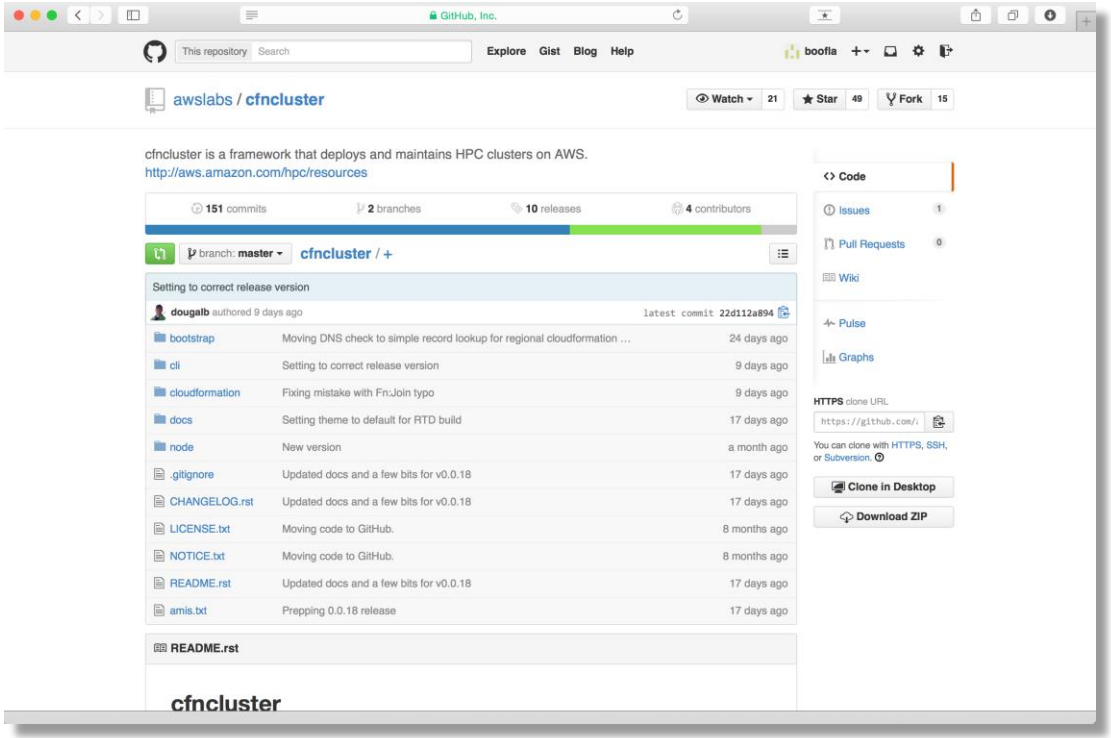
CloudFormation

- Fundamental service in AWS* used for automating deployment and configuration of resources

- CloudFormation* Template 
  - JSON-formatted document which describes a configuration to be deployed in an AWS account
  - When deployed, refers to a “stack” of resources
  - Not a “script”, a *document*

IDF16
INTEL DEVELOPER FORUM

CfnCluster



#cfncluster

https://aws.amazon.com/hpc/cfncluster

https://github.com/awslabs/cfncluster

CfnCluster

- Made public 2014-06-10, as cfncluster-0.0.5
- Amazon* Software License - https://aws.amazon.com/asl/
- Latest version is cfncluster-1.2.1, released on 2016-03-24

CfnCluster (cont)

- Architecture based on https://en.wikipedia.org/wiki/Beowulf_cluster
- OSes supported Amazon* Linux*, CentOS* 6 & 7, and Ubuntu* 1404
- Supports SGE, Openlava, Torque, and SLURM
- Extensible via Chef and pre/post-install scripts

Intel Optimized Frameworks on Amazon Web Services (AWS*)

Steps:

1. Launch a CloudFormation* template
2. Edit the CfnCluster configuration file
3. Use “cfncluster” to launch a cluster
4. Run some ML/DL examples

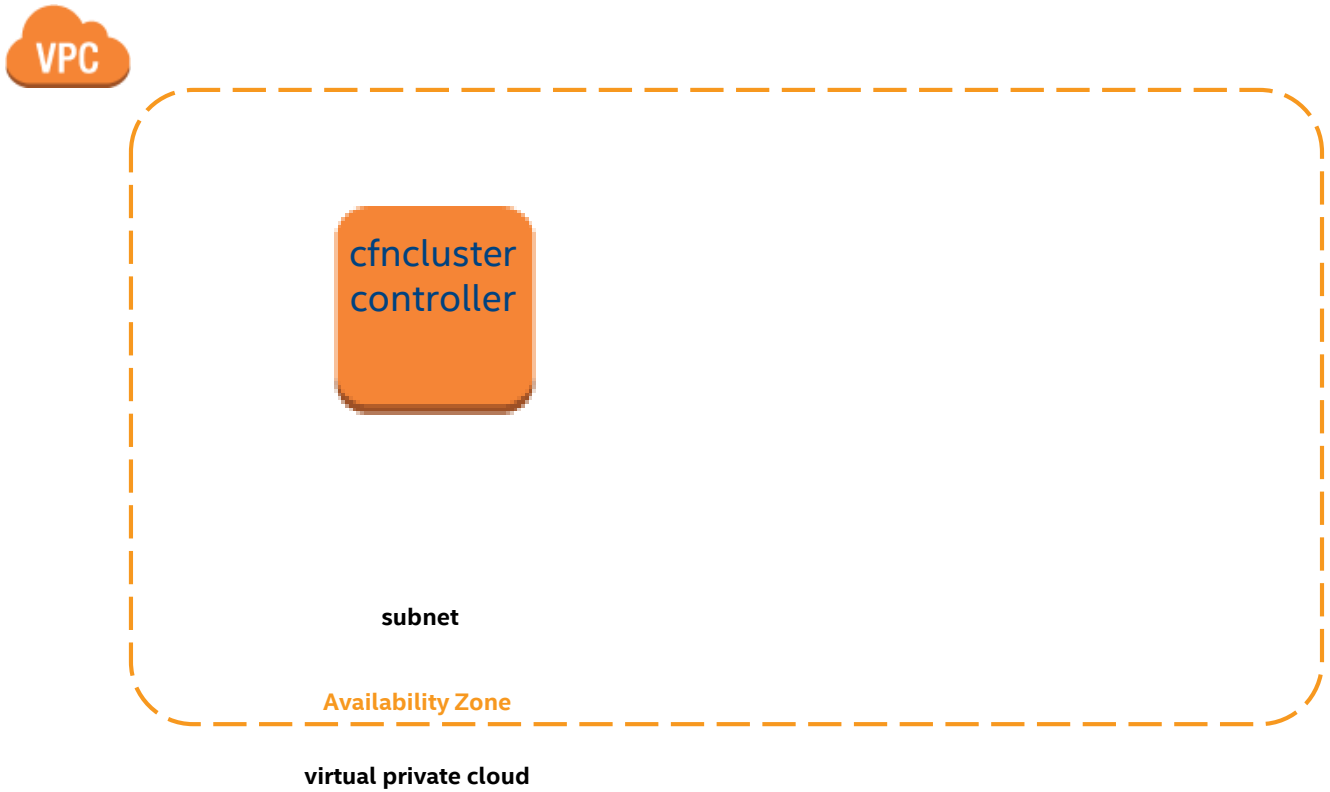
Intel Optimized Frameworks on Amazon Web Services (AWS*)

Steps:

1. **Launch a CloudFormation template**
2. Edit the CfnCluster configuration file
3. Use “cfncluster” to launch a cluster
4. Run some ML/DL examples

The CloudFormation stack

CloudFormation outputs:



Intel Optimized Frameworks on Amazon Web Services (AWS*)

Steps:

1. Launch a CloudFormation* template
2. **Edit the CfnCluster configuration file**
3. Use “cfncluster” to launch a cluster
4. Run some ML/DL examples

cfncluster configuration file details

```
[aws]
# The AWS region to run the cluster (i.e.: us-east-1, us-west-1, us-west-2, etc)
aws_region_name = us-west-2
# Set the following to the aws keys. If not defined, cnfcluster will attempt to use:
#   a) environment variables
# or
#   b) an EC2 IAM role
aws_access_key_id = AWS_ACCESS_KEY
aws_secret_access_key = AWS_SECRET_ACCESS_KEY

[cluster default]
#create a name for your VPC
vpc_settings = NAME_OF_THE_VPC
key_name = NAME_OF_THE_SSH_KEY
# Override path to cloudformation in S3
# (defaults to https://s3.amazonaws.com/cfncluster-<aws_region_name>/templates/cfncluster-<version>.cfn.json)
# template_url = https://s3.amazonaws.com/cfncluster-us-east-1/templates/cfncluster.cfn.json
# the compute instance type
# (defaults to t2.micro)
compute_instance_type = c4.8xlarge
# the master instance type
# (defaults to t2.micro)
master_instance_type = c4.large
# Initial number of EC2 instances to launch as compute nodes in the cluster.
# (defaults to 2 for default template)
initial_queue_size = 2
# Maximum number of EC2 instances that can be launched in the cluster.
# (defaults to 10 for the default template)
max_queue_size = 2
# ID of a Custom AMI, to use instead of published AMI's
```

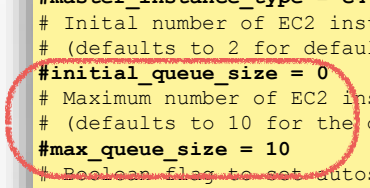
~/.cfncluster/config

Config options to explore …

Many options, but the most interesting ones immediately are:

```
# (defaults to t2.micro for default template)
compute_instance_type = c4.2xlarge
# Master Server EC2 instance type
# (defaults to t2.micro for default template
#master_instance_type = c4.4xlarge
# Inital number of EC2 instances to launch as com                    he cluster.
# (defaults to 2 for default template)
#initial_queue_size = 0
# Maximum number of EC2 instanc                                      .
# (defaults to 10 for the default template)
#max_queue_size = 10
# Boolean flag to set autoscaling group to maintain                  d scale back
# (defaults to false for the default template)
#maintain_initial_size = true
# Cluster scheduler
# (defaults to sge for the defau
scheduler = sge
# Type of cluster to launch i.e.
# (defaults to ondemand for the defau                 plate)
cluster_type = spot
# Spot price for the ComputeFleet
spot_price = 0.50

# Cluster placement group. This placement group must already exist.
# (defaults to NONE for the default template)
#placement_group = NONE
```



Min & Max size of your cluster.

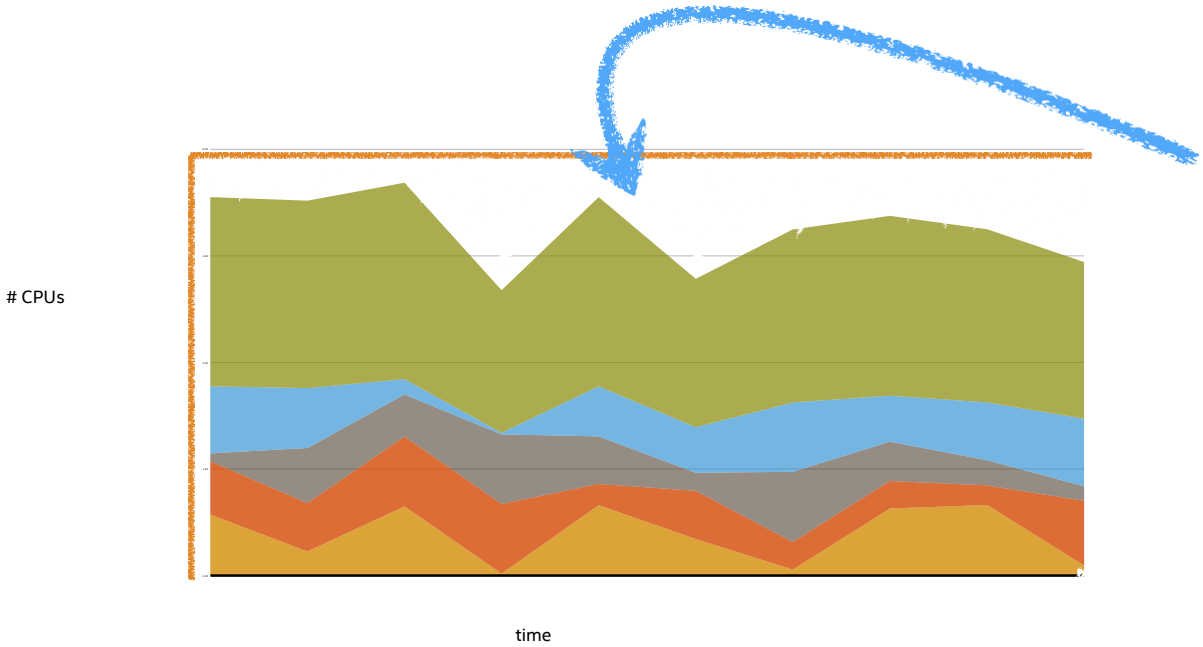
Whether to fall back when things get quiet

Also can use 'openlava' or 'torque'

Explore the SPOT market if you want to save money :-)

A placement group will provision your instances very close to each other on the network.

Amazon Web Services (AWS*) Spot Market



Spot Market

Our ultimate space filler.

Spot Instances allow you to name your own price for spare AWS* computing capacity.

Great for workloads that aren't time sensitive, and especially popular in research (hint: it's really cheap).

IDF16
INTEL DEVELOPER FORUM

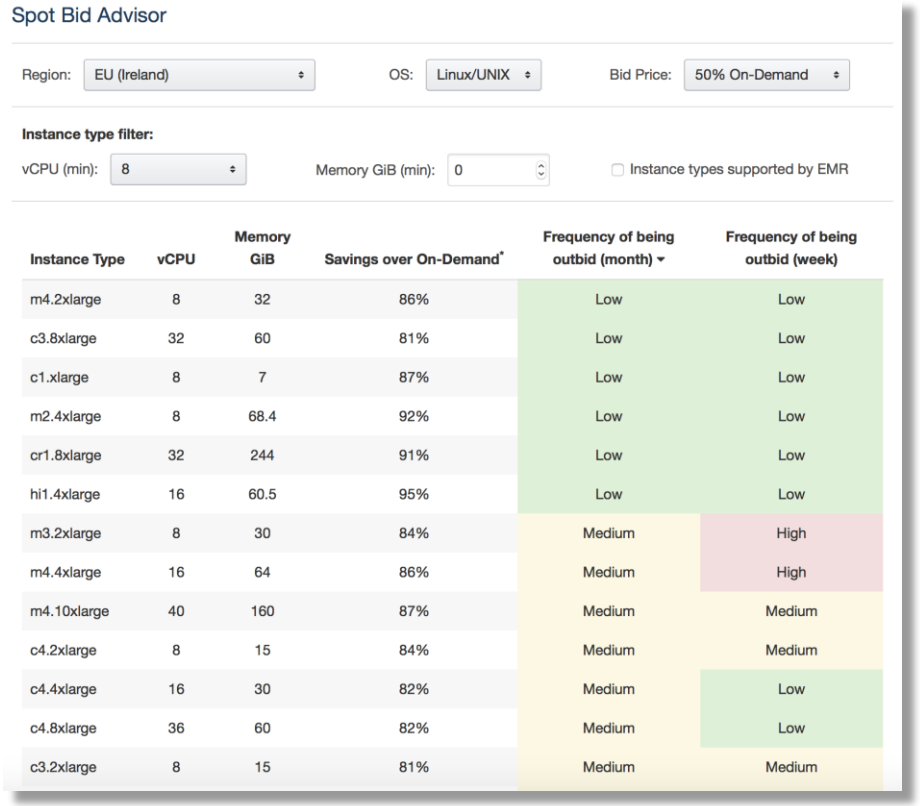
Spot Market – playing the [good] odds

Spot Bid Advisor

The Spot Bid Advisor analyzes Spot price history to help you determine a bid price that suits your needs.

You should weigh your application's tolerance for interruption and your cost saving goals when selecting a Spot instance and bid price.

The lower your frequency of being outbid, the longer your Spot instances are likely to run without interruption.



Instance Type	vCPU	Memory GiB	Savings over On-Demand*	Frequency of being outbid (month)	Frequency of being outbid (week)
m4.2xlarge	8	32	86%	Low	Low
c3.8xlarge	32	60	81%	Low	Low
c1.xlarge	8	7	87%	Low	Low
m2.4xlarge	8	68.4	92%	Low	Low
cr1.8xlarge	32	244	91%	Low	Low
hi1.4xlarge	16	60.5	95%	Low	Low
m3.2xlarge	8	30	84%	Medium	High
m4.4xlarge	16	64	86%	Medium	High
m4.10xlarge	40	160	87%	Medium	Medium
c4.2xlarge	8	15	84%	Medium	Medium
c4.4xlarge	16	30	82%	Medium	Low
c4.8xlarge	36	60	82%	Medium	Low
c3.2xlarge	8	15	81%	Medium	Medium

Intel Optimized Frameworks on AWS*

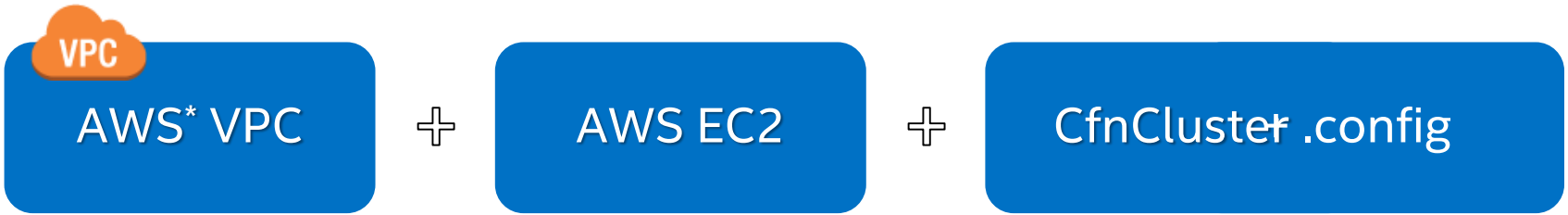
Steps:

1. Launch a CloudFormation* template
2. Edit the CfnCluster configuration file
3. **Use “cfncluster” to launch a cluster**
4. Run some ML/DL examples

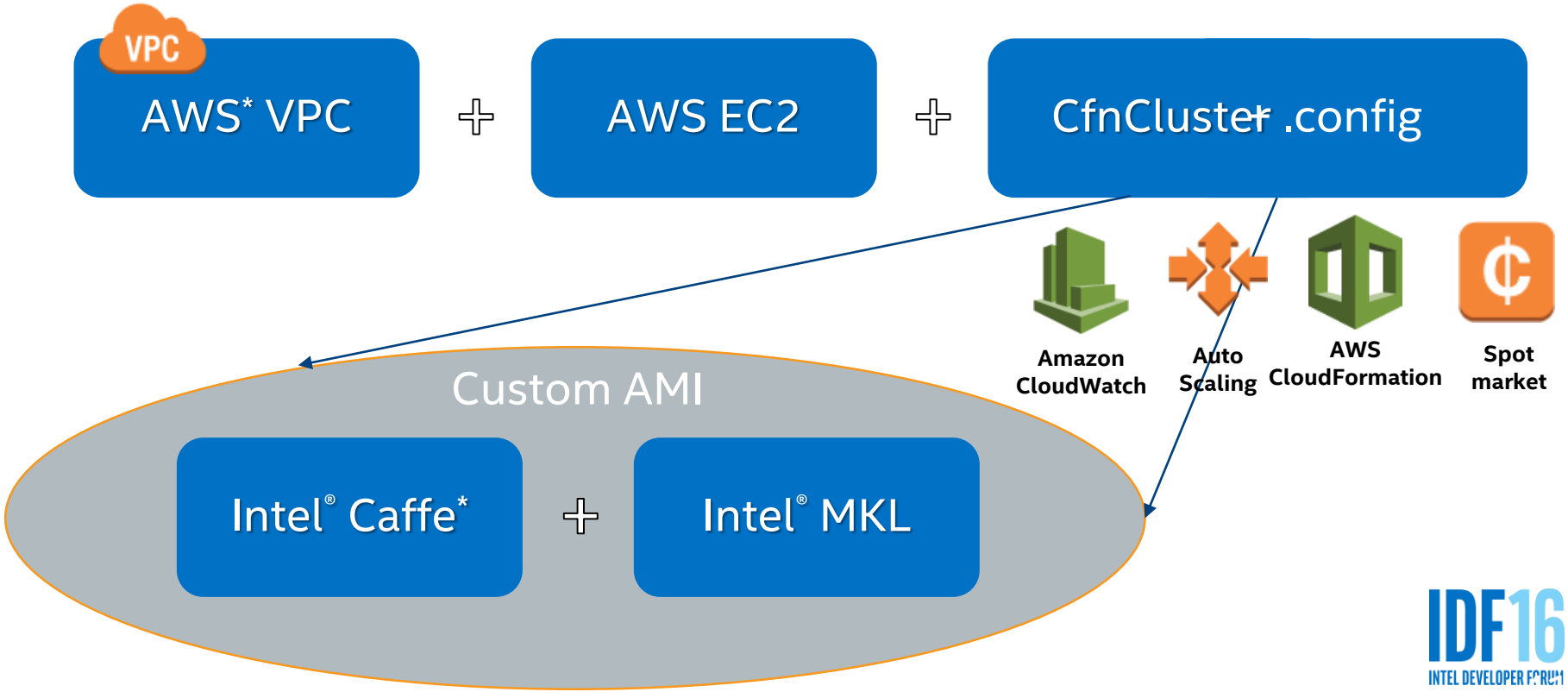
Components of our stack

Custom AMI

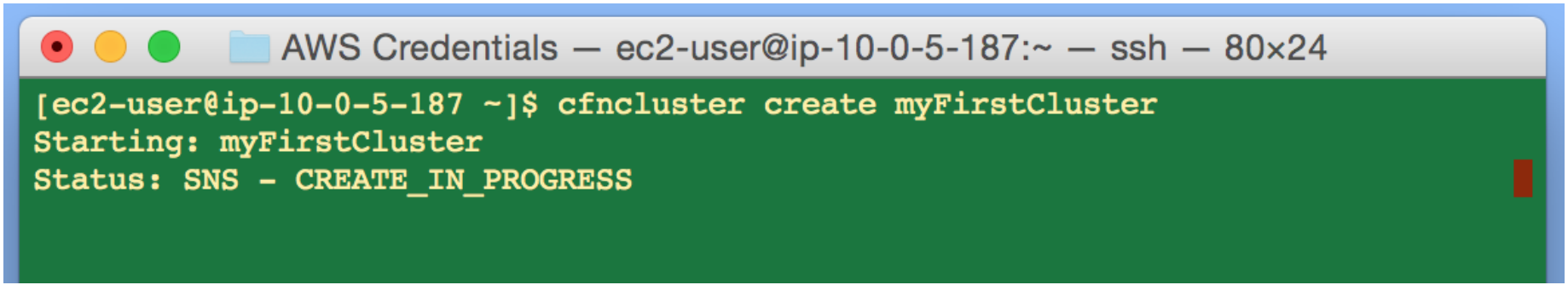
Components of our stack

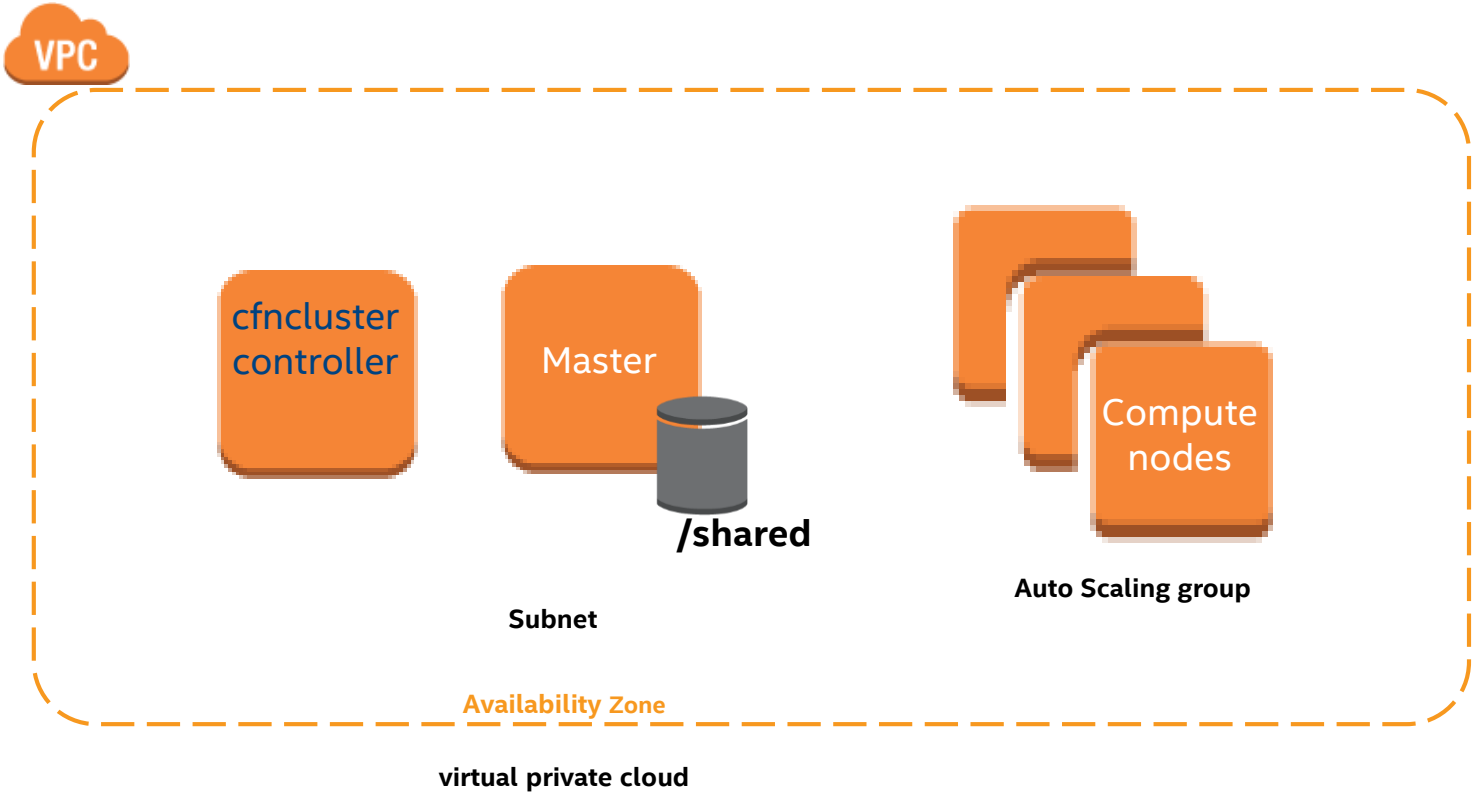


Components of our stack



cfncluster command




VPC
cfncluster controller
Master
/shared
Compute nodes
Auto Scaling group
Subnet
Availability Zone
virtual private cloud

Intel Optimized Frameworks on Amazon Web Services (AWS*)

Steps:

1. Launch a CloudFormation* template
2. Edit the CfnCluster configuration file
3. **Use “cfncluster” to launch a cluster**
4. Run some ML/DL examples

System-wide Upgrade from Intel® Xeon® v2 and Intel® Xeon® v3

```
$ ed ~/.cfncluster/config
/compute_instance_type/
compute_instance_type = c3.8xLarge
s/c3/c4/p
compute_instance_type = c4.8xLarge
w
949
$ cfncluster update boof-cluster
```

Yes, really :-)

This cluster intentionally left blank.

Your cluster is ephemeral.

You've created a disposable cluster. But it's 100% recyclable

It's worth noting that anything you put into this cluster will vaporize when you issue the command

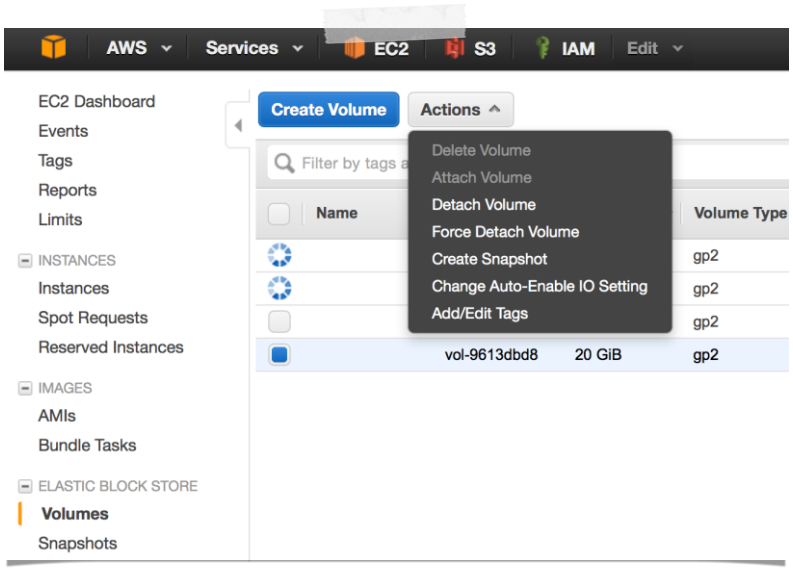
```
$ cfncluster delete <your cluster name>
```

... which might not be what you first expect.

It's easy to save your data and pick up from where you left off later

Before you delete your cluster, take a snapshot of the EBS (block storage) volume that you used for your /shared filesystem using the AWS* EC2 console (see the pic on the right)

The EBC volume you care most about is the one attached to the headnode instance (hint: it's probably the largest one)



Call to action

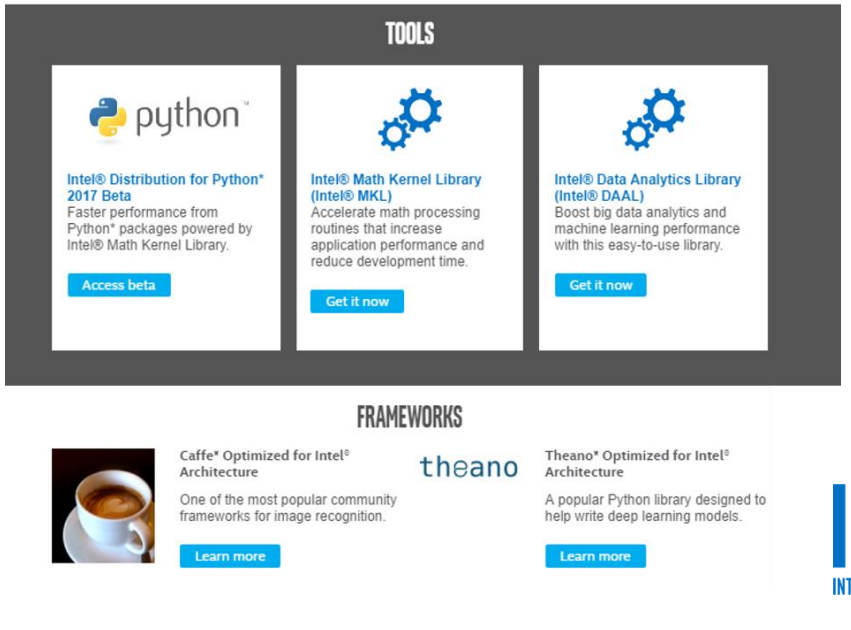
- Use Intel tools and optimized frameworks for DL workloads
  - https://software.intel.com/en-us/articles/training-and-deploying-deep-learning-networks-with-caffe-optimized-for-intel-architecture
  - https://software.intel.com/en-us/deep-learning-sdk
  - https://github.com/intelcaffe/caffe
  - https://github.com/intel/theano (other frameworks will be coming soon)
- Use multinode distributed training to reduce time-to-train
- Take advantage of AWS* CloudFormation* cluster
  - https://software.intel.com/en-us/articles/aws-cloudformation

Summary and Next Steps

- Deep learning usages are expanding
- Intel and AWS* partnered to provide cloud users easy access to Intel optimized deep learning frameworks
- Reduce time to train by using Intel optimized frameworks and distribute the training workload across various nodes

Check out Intel® ML Developer Zone
software.intel.com/machine-learning

- One developer portal for all Intel ML/DL tools, frameworks, training and support
- Download Caffe* Optimized for IA, Intel® MKL, Intel® DAAL, Intel® Distribution for Python*
- Community forums, articles, samples, tutorials and documentation



IDF16
INTEL DEVELOPER FORUM

Technical Sessions in Analytics Track

Tuesday, August 16, 2016

11:00 AM – 12:00 PM *ANATS01* — Deep Learning Frameworks and Optimization Paths on Intel® Architecture ***Level 2 Room 2001***

11:00 AM – 12:00 PM *SOFTS01* — Accelerating Machine Learning on Apache Spark* ***Level 2 Room 2006***

1:15 PM – 2:15 PM *ANATS03* — Enabling an End-to-End Architecture for Autonomous Cars ***Level 2 Room 2001***

2:30 PM – 3:30 PM *ANABI01* — Advanced Analytics – Trends, Challenges, Opportunities ***Level 2 Room 2016 Tech & Business Insight***

2:30 PM – 3:30 PM *ANATS05* — How to Parallelize Neural Networks (xNNs) for Intel® Xeon Phi™ ***Level 2 Room 2001***

4:00 PM – 5:00 PM *ANATI01* — Scaling to Meet the Growing Needs of Artificial Intelligence (AI) ***Level 2 Room 2016 Tech & Business Insight***

4:00 PM – 5:00 PM *ANATS07* — Innovative Use of Analytics and Machine Learning: Security, Network Function Virtualization (NFV), and Optimized Infrastructure ***Level 2 Room 2001***

Wednesday, August 17, 2016

11:00 AM – 12:00 PM *ANATS02* — Apache Spark* in Enterprise Analytics ***Level 2 Room 2002***

1:15 PM – 2:15 PM *ANATS04* — End-to-End Analytics Solutions with Trusted Analytics Platform ***Level 2 Room 2001***

2:30 PM – 3:30 PM *ANATS06* — The Complete Toolset for Accelerating Analytics – From Optimized System Architecture to Accelerators ***Level 2 Room 2001***

2:30 PM – 3:30 PM *IOTTS04* — IoT for Intervention During Equipment Failure ***Level 2 Room 2008***

4:00 PM – 5:00 PM *ANATS08* — Open Source Solutions for Network Intelligence ***Level 2 Room 2001***

IDF16
INTEL DEVELOPER FORUM

Experience Data Center Innovation & Get Engaged

Tech Showcase (1st Floor): see interactive demos & meet experts!

- **Intel Pavilion :** Cloud/SDI, Analytics, 5G technology demos
- **Data Center communities:** Artificial Intelligence, Intel Builders, NVMe, Memory
- Take part in the **Data Center Solutions Tour for opportunity to win a drone!**
- Meet experts in **Intel Builders community for a chance to win an Intel NUC Kit!**
- **Live Data Center Q&A with Experts: 4-7 pm Wed in Networking Plaza**

Data Center Experience Zone and Pentathlon (Floor 2 concourse)

Immerse yourself in data center-driven experiences in music, sports and more

Compete in five data center-related challenges based on Intel technologies to **win a T-shirt and qualify to win awesome fabulous prizes!**

- **Tuesday:** Rack Stack Challenge & Snap Telemetry Challenge, 11 a.m. – 5 p.m.
- **Wednesday:** Silicon Photonics Challenge & Net Boss Challenge, 11 a.m. – 5 p.m.
- **Thursday:** Masters Challenge @ 10:30 a.m., featuring top 3 scorers from Day 1 & 2 challenges

Follow us on Twitter at #IntelDCZone

IDF16
INTEL DEVELOPER FORUM

INTEL® HPC DEVELOPER CONFERENCE 2016
NOVEMBER 12 - 13, 2016 - SALT LAKE CITY

Going to SC16? Make Plans to Join Us!

The one and a half-day event brings together architecture experts and HPC industry leaders to discuss, share and highlight the latest in supercomputing.

Learn best practices to reach peak performance

Discuss code modernization in HPC

Hear the latest industry-specific approaches for tackling real-life challenges in parallel programming

Geared to HPC Developers and Covers Such Topics As:

Programming for the newest Intel® Xeon Phi™ processor
Data Analytics/Machine Learning • Software Visualization • Vectorization

For More Information and To Register: hpcdevcon.intel.com

IDF16
INTEL DEVELOPER FORUM

Legal Notices and Disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.
- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.
- Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.
- This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.
- All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
-
- *Other names and brands may be claimed as the property of others.

IDF16
INTEL DEVELOPER FORUM