How OpenStack enables face recognition with GPUs and FPGAs

CERN OpenStack Day

May 2019

Erwan Gallen
@egallen

Red Hat

About your presenter: Erwan Gallen

Erwan Gallen
IRC: egallen
Twitter: @egallen
https://egallen.com
https://erwan.com

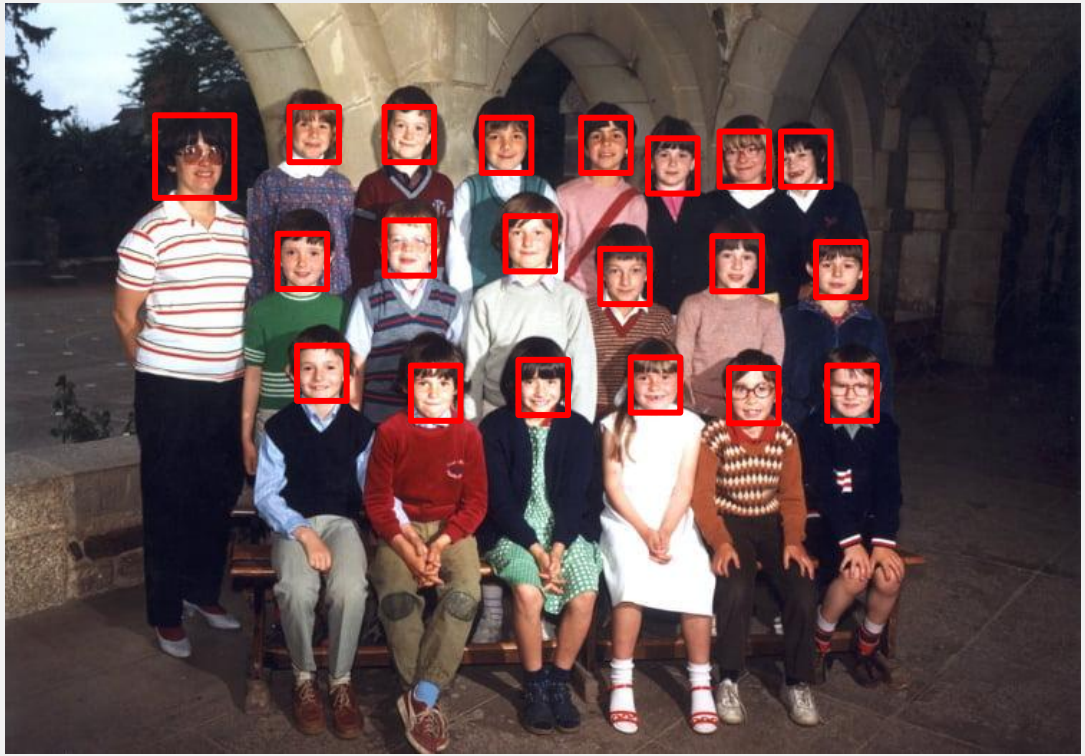
Product Manager @ Red Hat
Compute and HPC Red Hat Cloud Platforms
OpenStack French User Group

Agenda

- What is face recognition?
- Machine Learning algorithms for face recognition
- Hardware accelerators with OpenStack
  - From Edge to Data Center
  - GPUs
  - FPGAs

What is face recognition?

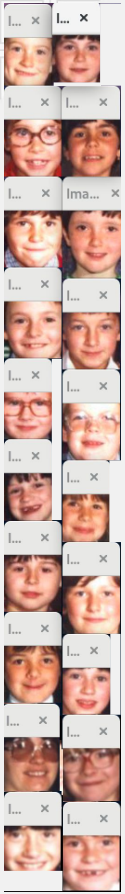
Face recognition is not only **face detection**



Localize faces in pictures

Subset of the face recognition

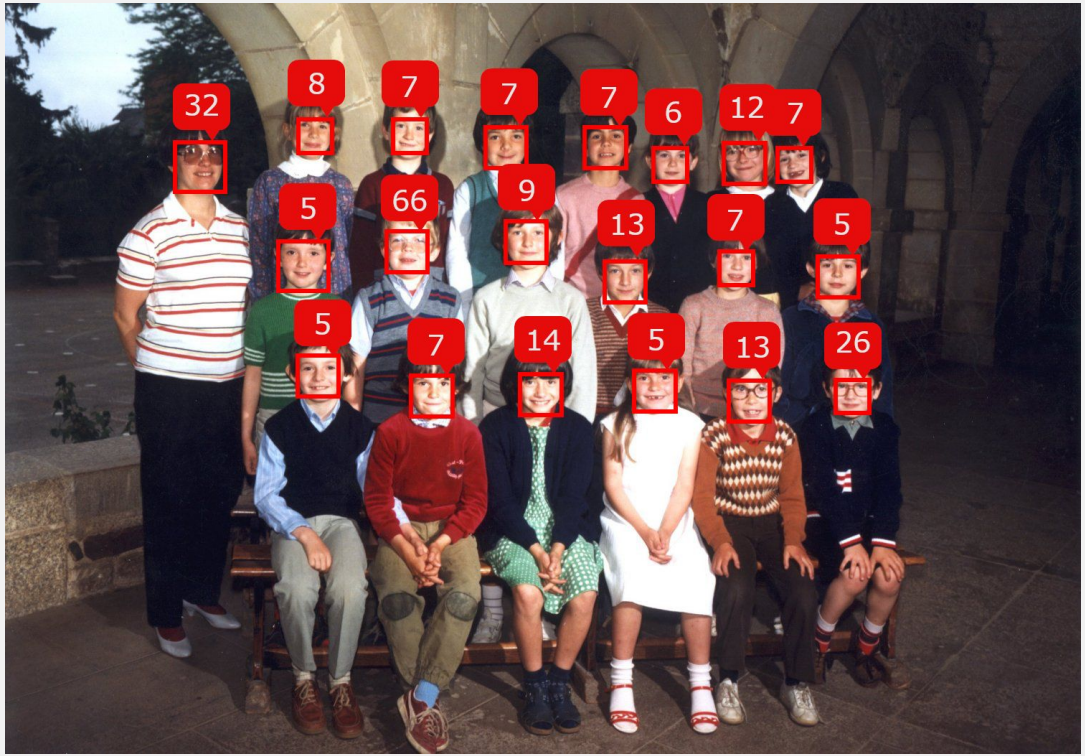
Face recognition is not only **face detection**



Testing face_recognition from Adam Geitgey https://github.com/ageitgey:

```
$ python find_faces_in_picture.py
I found 20 face(s) in this photograph.
A face is located at pixel location Top:  159, Left: 538, Bottom: 211, Right: 590
A face is located at pixel location Top:  488, Left: 405, Bottom: 550, Right: 467
A face is located at pixel location Top:  523, Left: 1006, Bottom: 585, Right: 1068
A face is located at pixel location Top:  199, Left: 1074, Bottom: 251, Right: 1126
A face is located at pixel location Top:  170, Left: 805, Bottom: 232, Right: 868
A face is located at pixel location Top:  357, Left: 833, Bottom: 419, Right: 895
A face is located at pixel location Top:  170, Left: 667, Bottom: 232, Right: 729
A face is located at pixel location Top:  522, Left: 1155, Bottom: 573, Right: 1207
...
```

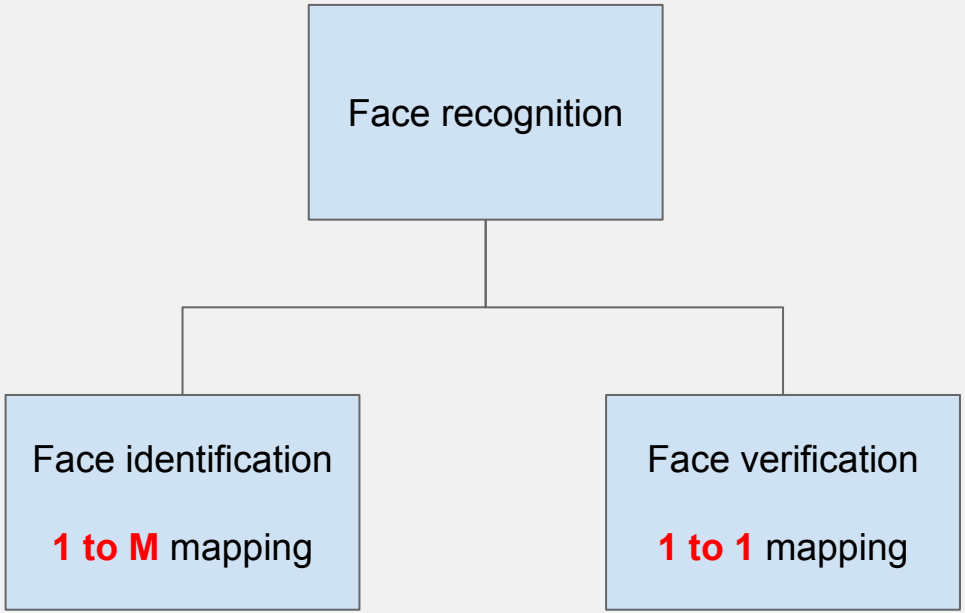
Face recognition is not **face classification**

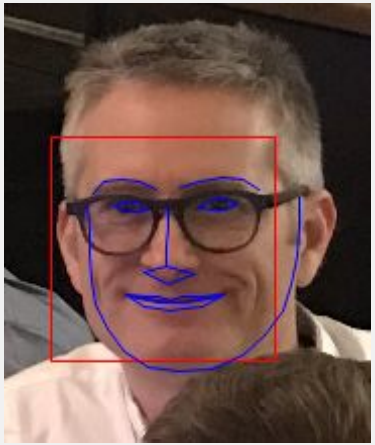


Emotion

Age (not perfect)

What is face recognition?





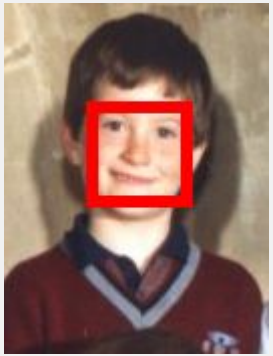
- Identifies persons on face images or video frames like humans
- Extract features from an input face image
- Compare them to labeled features in a database

What is face recognition? **Face identification**

Who this person is?
Identify faces

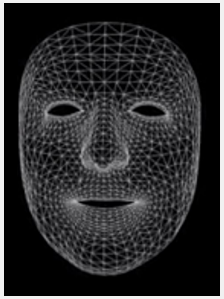
Red Hat

What is face recognition? **Face verification**



=

Anti-spoofing,
Liveness



Is that you?

Access granted to the classroom

Why face recognition?

- Social media
- Photo management software
- Retail security
- Smartphone security
- Airports
- Company security

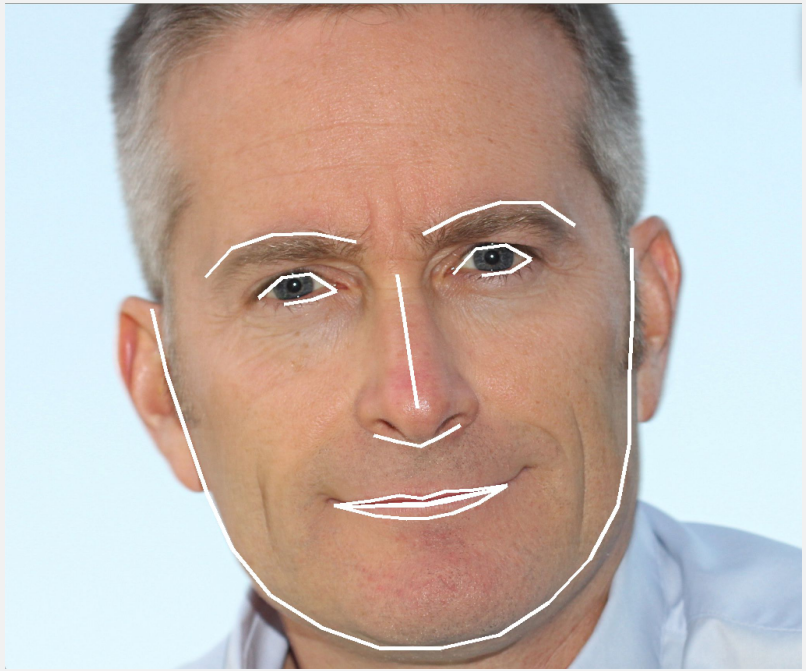
  ...

Deep face recognition

Face recognition testing pipeline

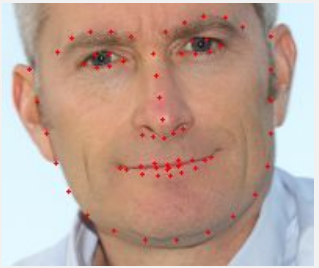
Step 1: Locate faces

Step 2: Identify facial features



68 face landmarks estimation:

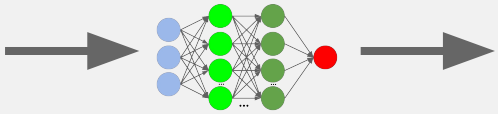
- Eyes
- Eyebrows
- Nose
- Mouth
- Jawline



Step 3: Align face, data normalization

Step 4: Extract embeddings

12 million pixels x 3 RGB
=64MP



Embedding model

```
[-0.09794946  0.1750353  -0.00183518 -0.06969877 -0.18536897  0.04571179
 -0.12051301 -0.01134866  0.12275194 -0.05641487  0.20879458  0.06214489
 -0.22761621  0.0086692  -0.07488536  0.12529071 -0.25272095 -0.11866099
 -0.15493473 -0.1316791  -0.04497606  0.12636296 -0.00205732  0.02656962
 -0.04635093 -0.30405098 -0.05450211 -0.01051785  0.21232744 -0.13787782
 -0.03417146  0.0266749  -0.18036658 -0.02583698  0.04969187 -0.04712777
 -0.02467611 -0.10555979  0.23597333  0.10794001 -0.11396772 -0.0573921
  0.09248708  0.361954    0.22029321  0.01929749  0.03292914 -0.01291857
  0.0557532  -0.24084231  0.06511801  0.23539528  0.119146    0.03241641
  0.04266791 -0.22245112 -0.00418036  0.16938339 -0.16139159  0.13203897
  0.07769118 -0.1201788  -0.0789692  -0.0477482   0.21210161  0.10167966
 -0.07848857 -0.24633884  0.18544734 -0.12879251 -0.04530688  0.09393331
 -0.10176498 -0.11792129 -0.26475698 -0.02540269  0.39102545  0.15318437
 -0.11986607  0.01925911 -0.06814814 -0.10627415  0.05448525  0.05113249
 -0.1767479  -0.11013202 -0.1219693   0.08241013  0.22501558  0.06293185
 -0.07381546  0.15070631  0.050916   -0.11006542  0.03469372  0.0759664
 -0.10799524 -0.0693723  -0.08497648 -0.01509521 -0.0022208  -0.17947954
 -0.04660298  0.12837186 -0.16561005  0.11340192  0.02368122 -0.04795649
 -0.05999342 -0.05129085 -0.10749085  0.00346414  0.21041188 -0.2466156
  0.21759526  0.20821606  0.09382657  0.15210913  0.06711852  0.06401221
 -0.00497382 -0.06777511 -0.09036514 -0.15843755 -0.06740897 -0.10087758
  0.07872257  0.05411601]
```

128 dimensional unit hypersphere

Step 5: Compare embeddings

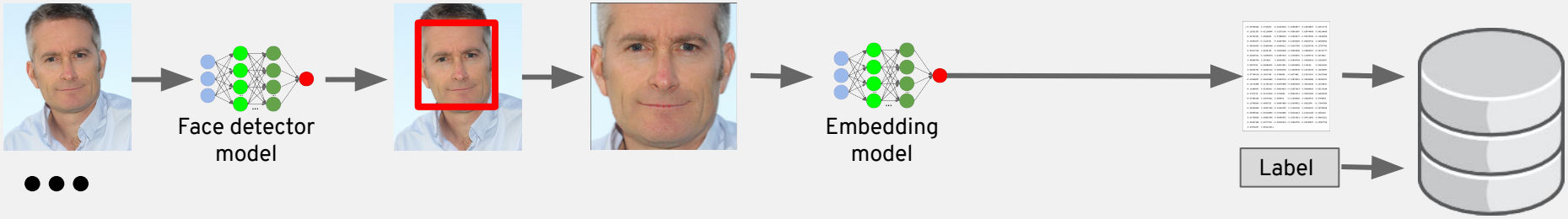
Face match if the euclidean distance between the face vectors is 0.6 or less (face_recognition).

Euclidean distance between faces:

- If < 0.6 : it's a match
- If > 0.6 : it's not a match

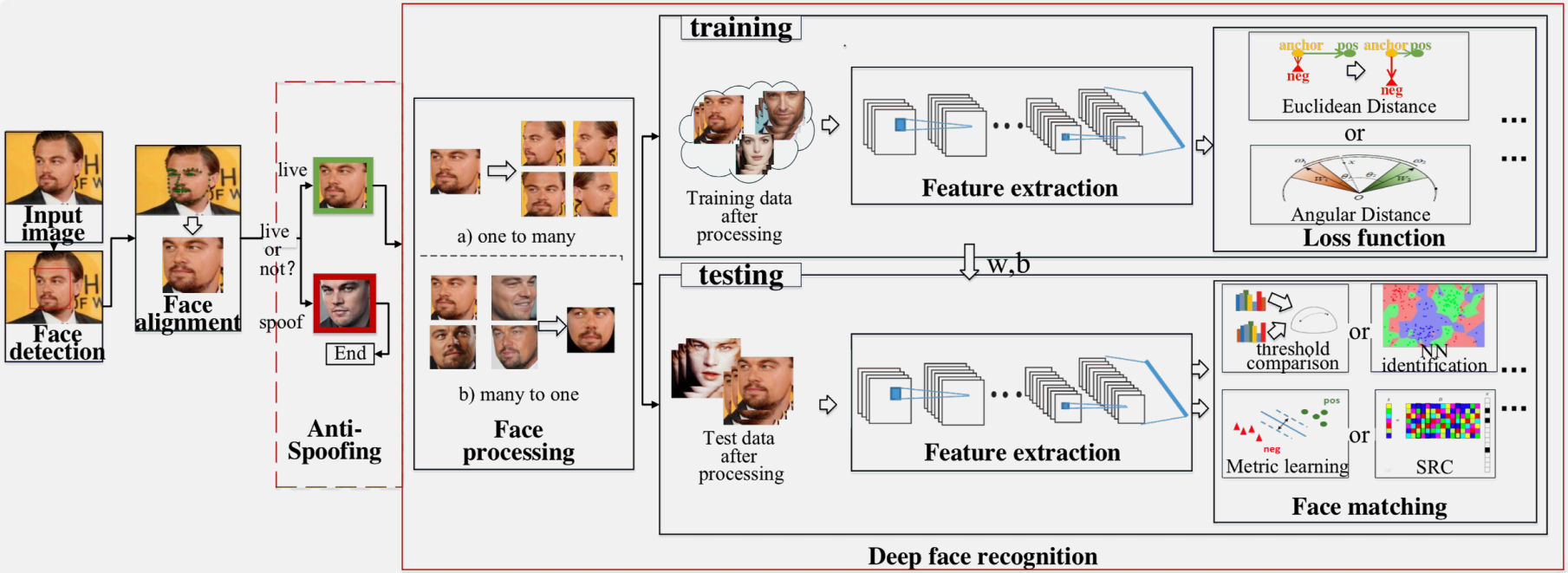
Face recognition pipeline

Extraction



Detection

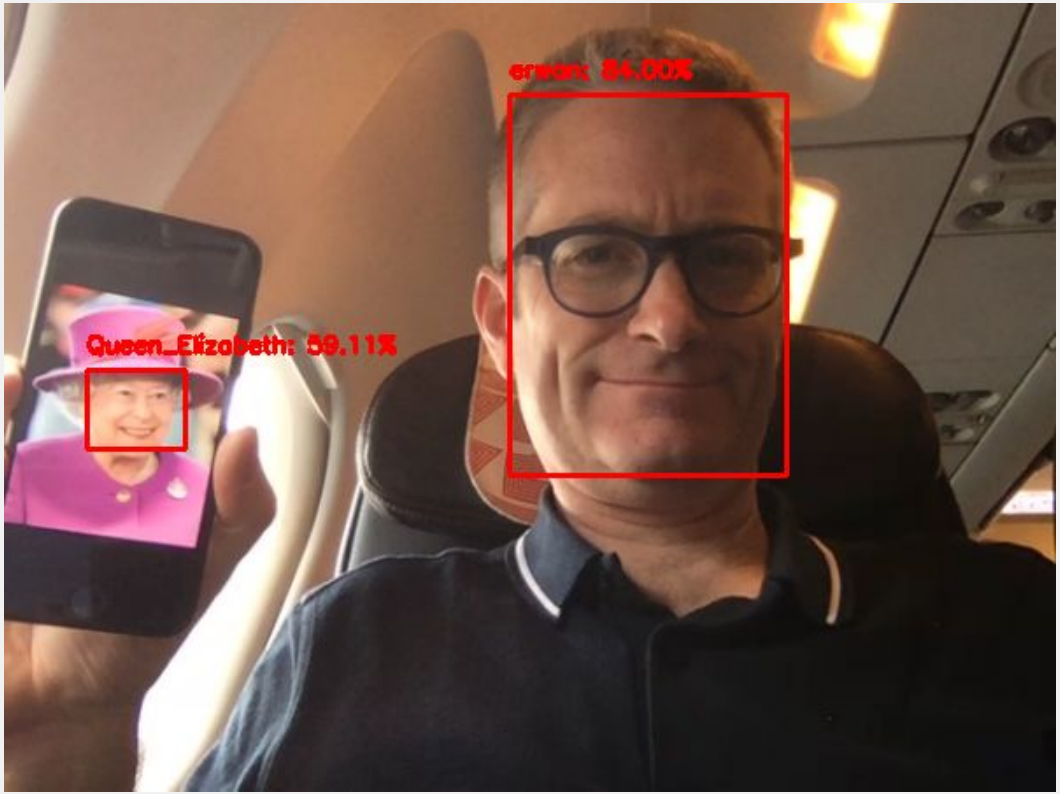
Full Face Recognition pipeline



Source: “Deep Face Recognition: A Survey”, Mei Wang, Weihong Deng

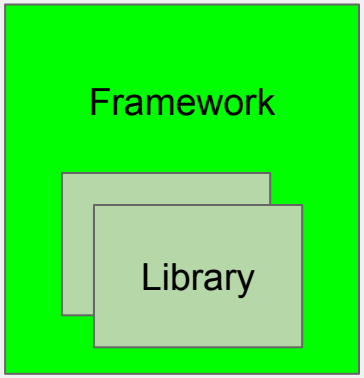
Face recognition demo using Python

Face recognition demo using Python

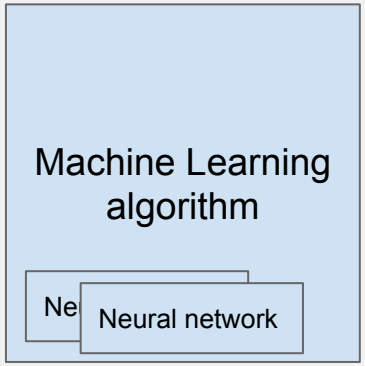


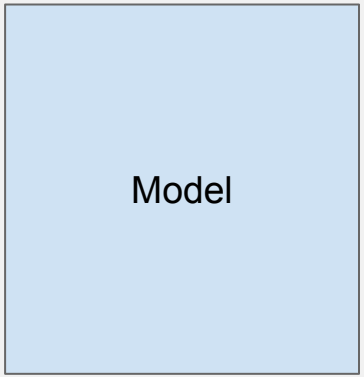
Machine Learning algorithms

Glossary

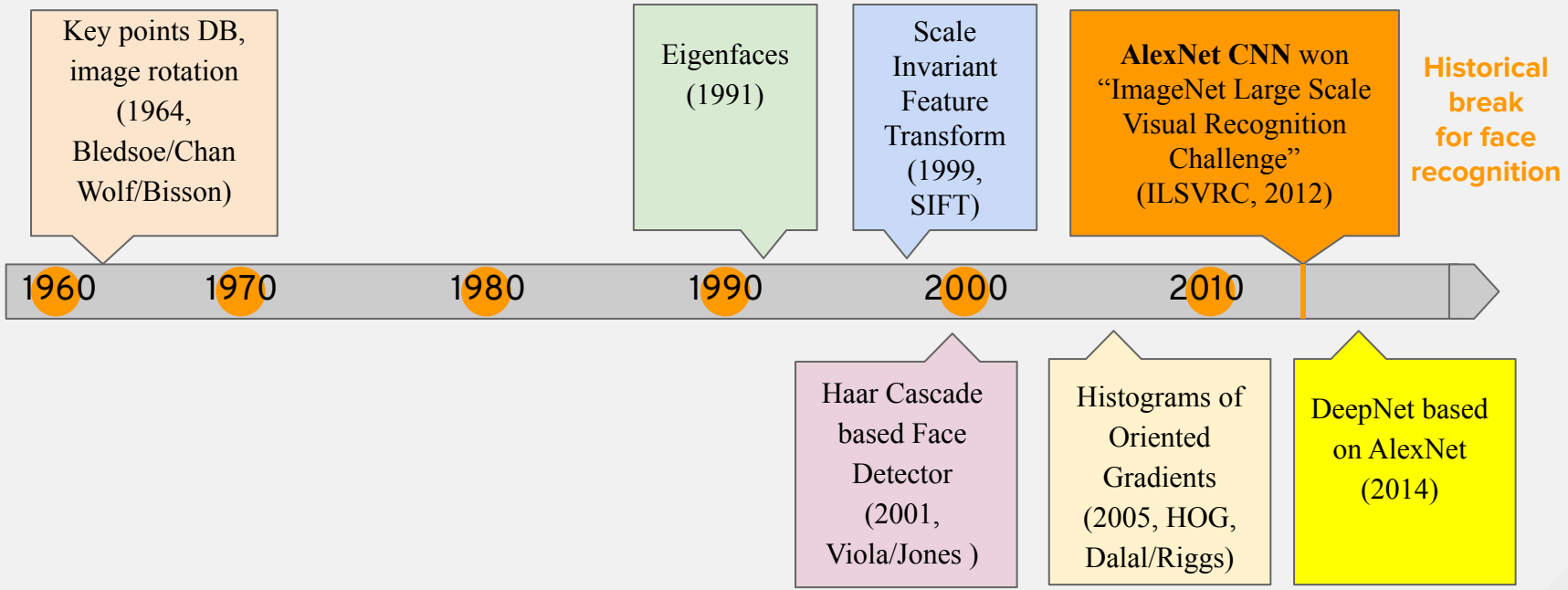

Framework
Library


Datasets


Machine Learning algorithm
Ne
Neural network

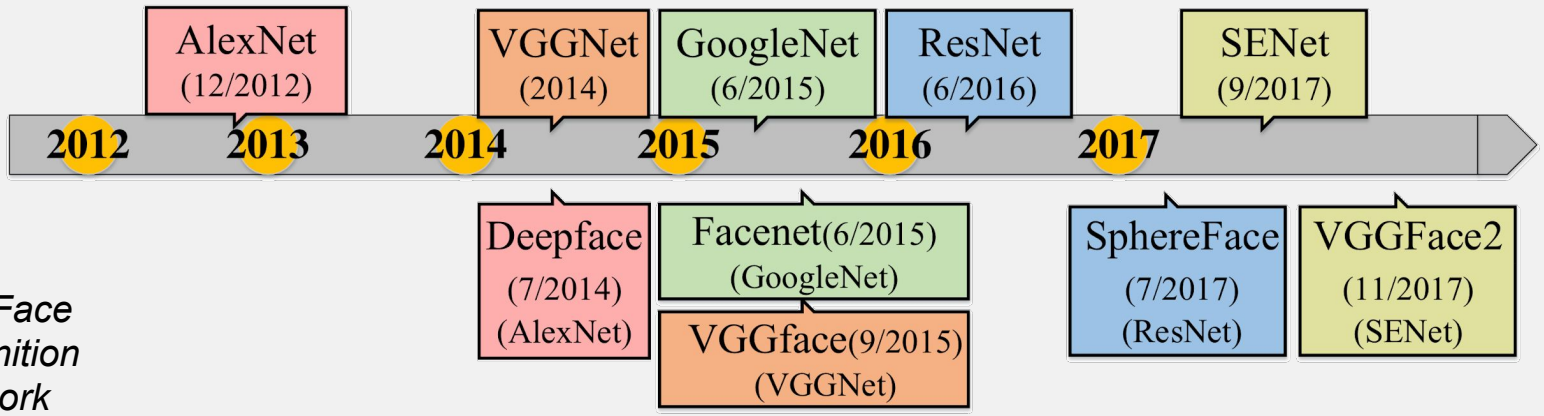

Model

Face processing history



Machine learning network architectures

Object Classification Network architectures



Deep Face Recognition Network architectures

Source: “Deep Face Recognition: A Survey”, Mei Wang, Weihong Deng

Datasets

	Training Set (face images)	Identities
Labeled Faces in the Wild (LFW)	13 233	5 749
Megaface (Creative Commons Flickr)	4.7 Million	672K
CelebA Dataset	202 599	10 177




Need variety of images: various angles, lightings, clothes, wearing accessories, ...

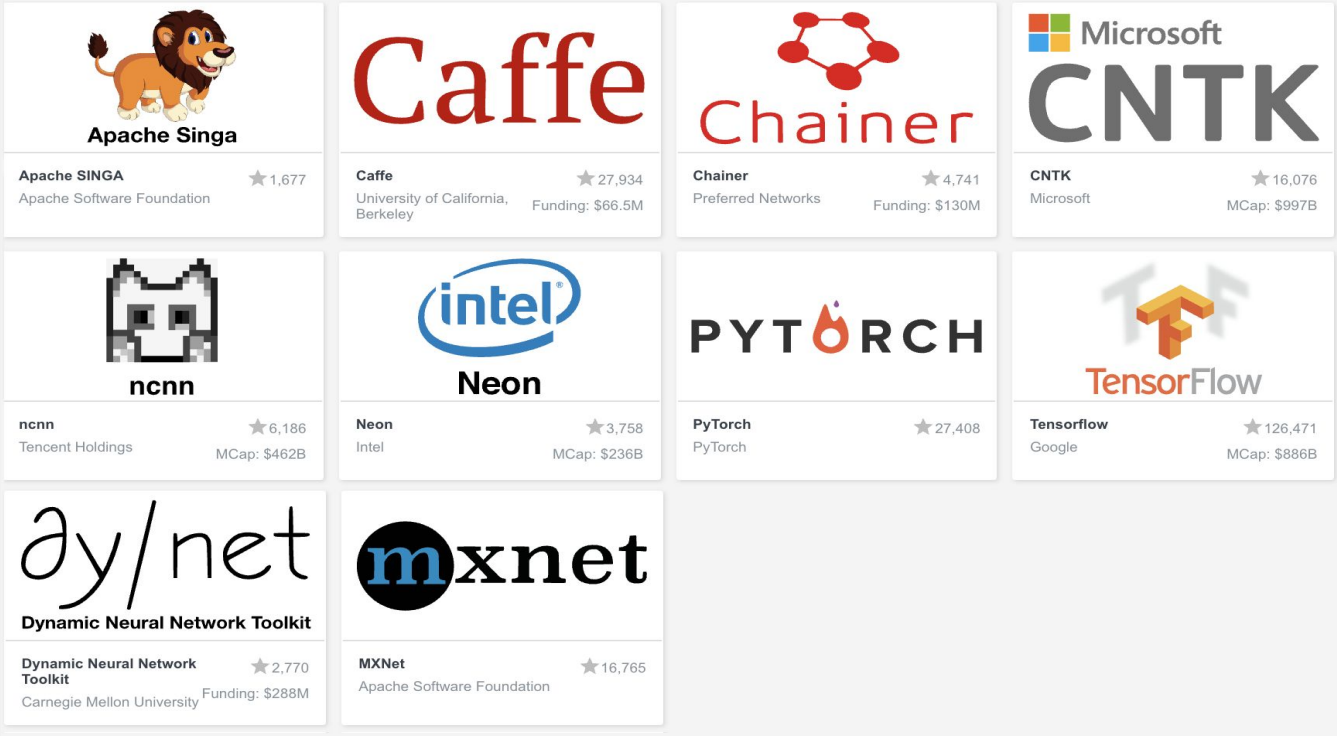
Pictures from: https://unsplash.com

Machine learning accuracy

When	Method	Accuracy	Architecture	Dataset	Training Set (face images)	Identities
2014	Deepface	97.35%	Alexnet	Facebook	4.4M	5 749
2015	Facenet	99.63%	Google Net-24	Google	500M	10M
2015	Baidu	99.67%	CNN-9	Baidu	1.2M	18K
2017	SphereFace	99.42%	ResNet-64	CASIA-webface	0.49M	10K
2018	Arcface	99.83%	ResNet-100	MS-Celeb-1M	3.5M	31K

Source: “Deep Face Recognition: A Survey”, Mei Wang, Weihong Deng

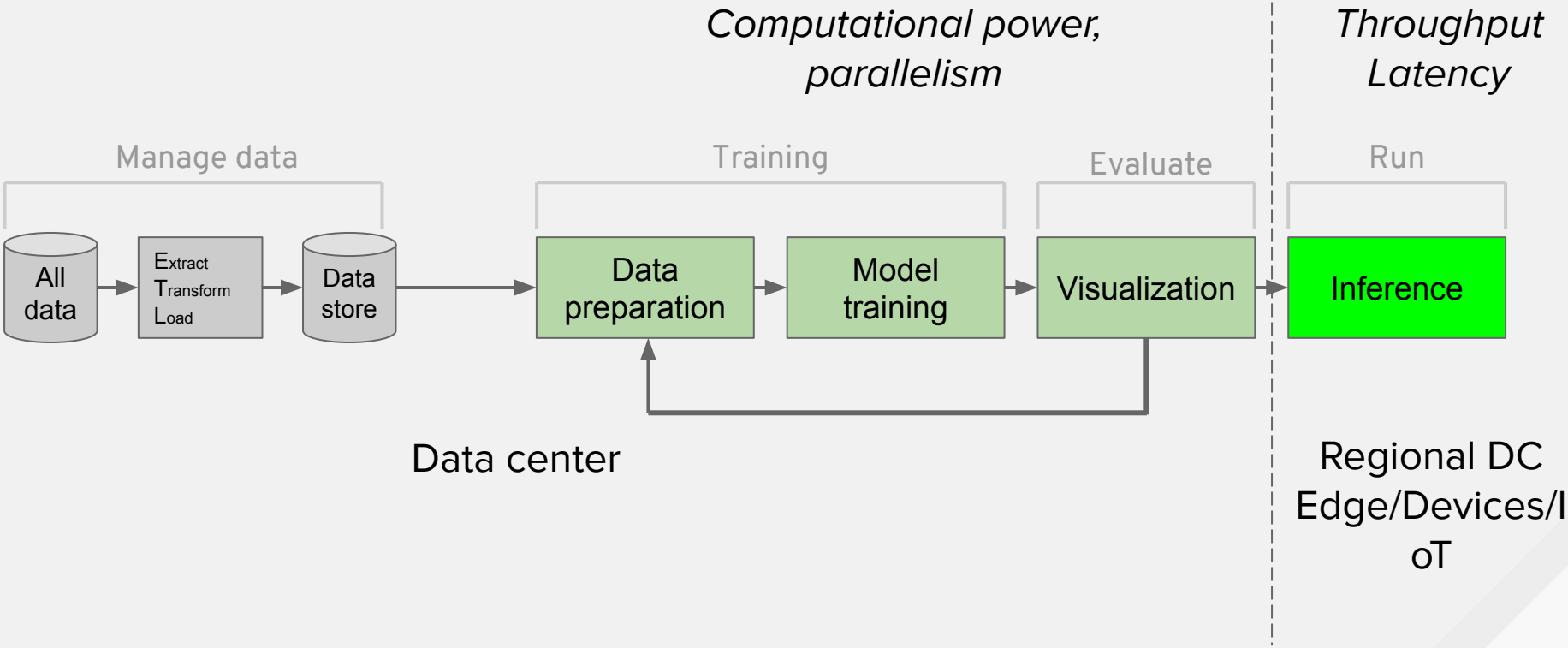
Deep Learning Frameworks



Source: https://landscape.lfdl.io/grouping

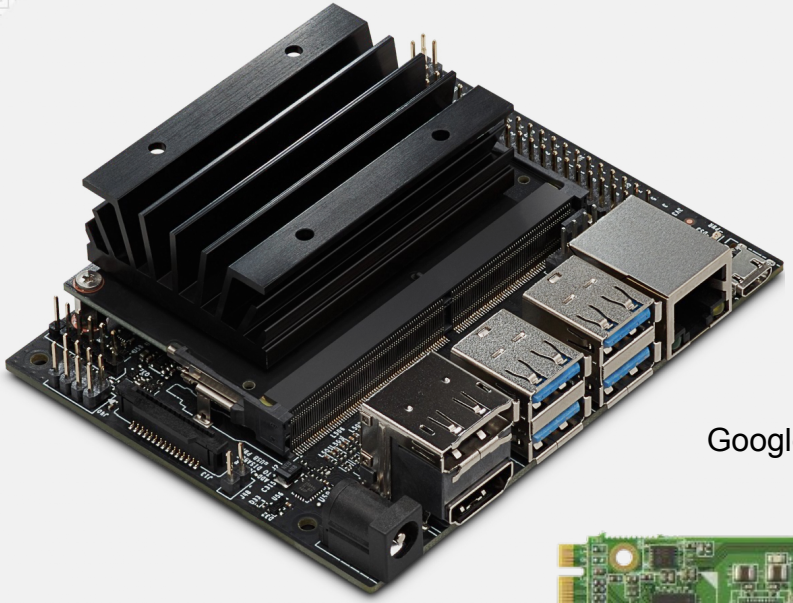
Hardware accelerators

Analytics pipelines, from training to inference



Red Hat

Hardware accelerators for devices

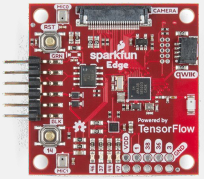


GPU NVIDIA Maxwell
NVIDIA Jetson Nano

TPU Google Coral
Google Edge TPU ML accelerator coprocessor

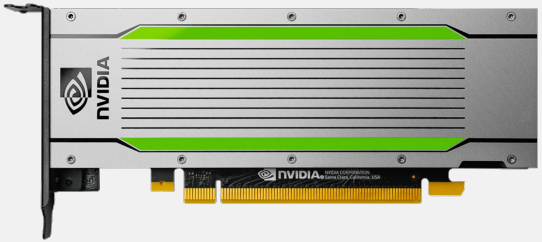
VPU Intel Myriad X (x1)
Intel Movidius Neural Compute Stick 2

VPU Intel Myriad X (x2)
UP AI Core XM 2280 (M.2)

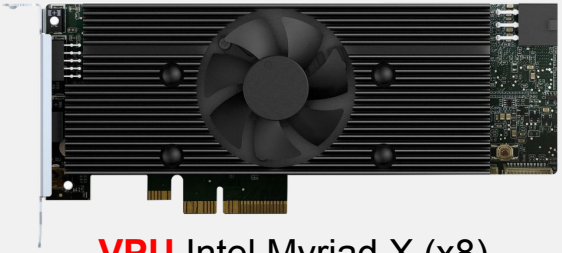


MCU Apollo3 Blue
SparkFun Edge Development Board

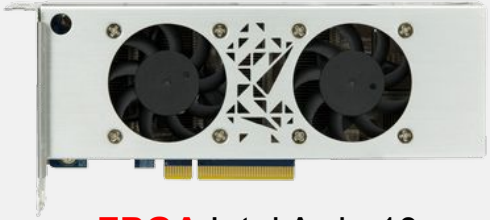
Hardware accelerators for edge



GPU NVIDIA Turing
NVIDIA Tesla T4



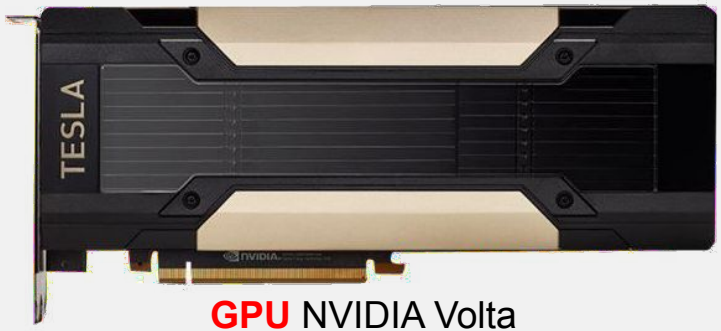
VPU Intel Myriad X (x8)
IEI Mustang-V100-MX8-R10



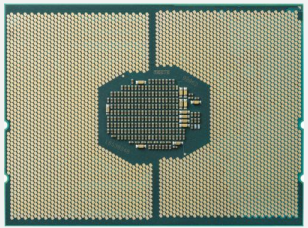
FPGA Intel Arria 10
IEI Mustang-F100-A10

Red Hat

Hardware accelerators for data centers

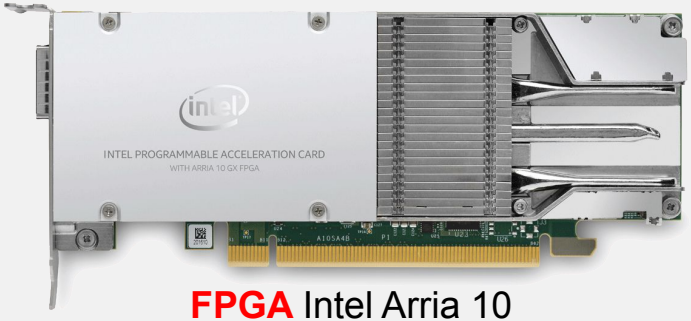


GPU NVIDIA Volta
NVIDIA Tesla V100



CPU Intel Xeon Cascade Lake
Deep learning boost
(AVX-512 Vector Neural Network Instructions, Brain floating-point format)

FPGA Intel Stratix 10
Intel FPGA PAC D5005



FPGA Intel Arria 10
Intel PAC with Intel Arria 10 GX FPGA

Red Hat

Hardware accelerators for Machine Learning

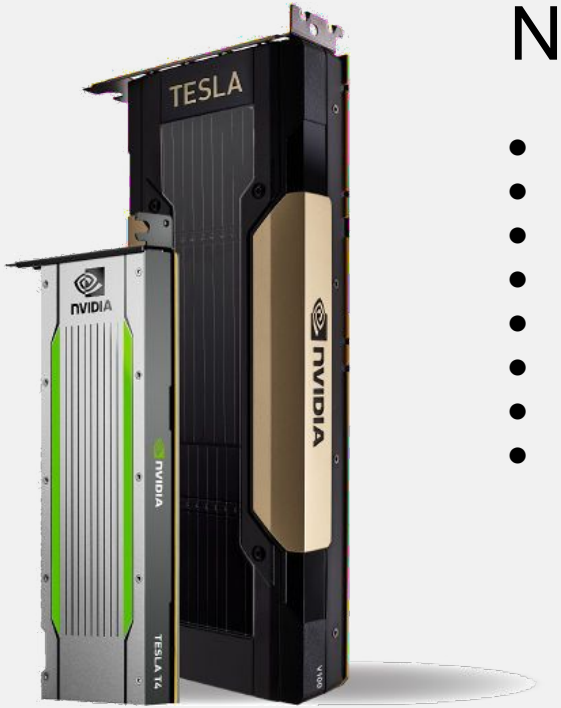
	Data Center	Edge	Device/IoT
CPU	Intel Xeon (AVX512VNNI) AMD Ryzen/Epyc	Xeon, Skylake D	ARM
GPU	NVIDIA Tesla V100 NVIDIA RTX6000, RTX8000 DGX-1 (8 x V100), DGX-2 (16 x V100) AMD Radeon Instinct MI25/MI60	NVIDIA Tesla T4	NVIDIA Jetson Nano NVIDIA Jetson TX1 NVIDIA Jetson TX2
FPGA	Intel FPGA PAC D5005 (Stratix 10) Intel PAC with Intel Arria 10 GX FPGA	IEI Mustang-F100-A10	
ASIC	IEI Mustang-V100-MX8	UP AI Core XM 2280	Intel Movidius NCS Google Coral TPU

GPUs

NVIDIA Tesla GPU accelerators

NVIDIA Tesla T4

- 70W
- “Turing” architecture
- Inferencing
- Light training
- Not only for Edge
- CUDA cores: 2560
- Tensor cores: 320



NVIDIA Tesla V100

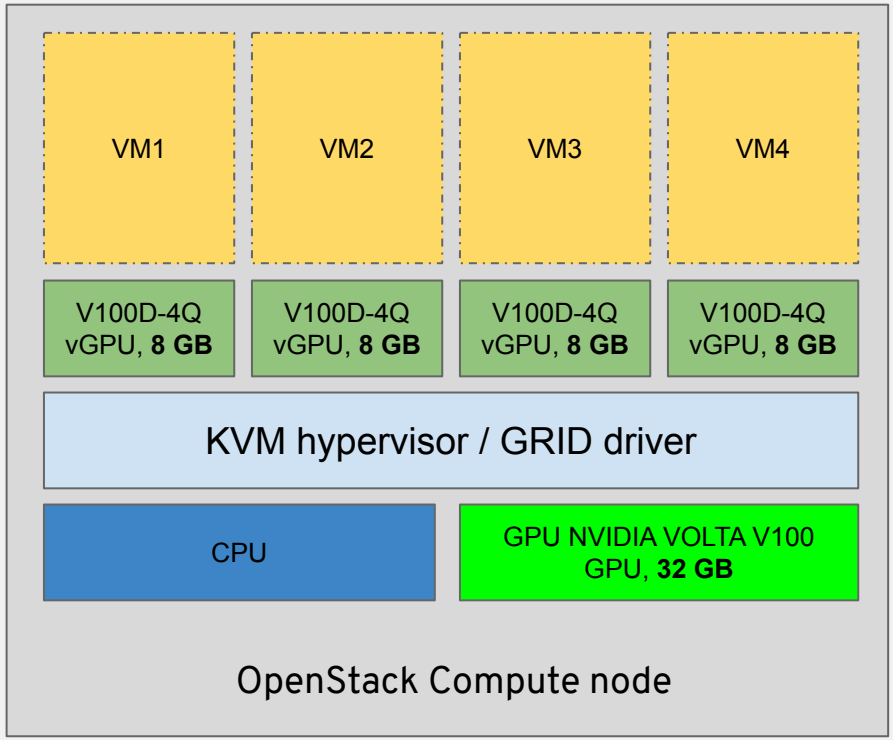
- 250W
- “Volta” architecture
- Training
- Inferencing
- Fanless
- Better to use OEM server
- CUDA cores: 5120
- Tensor cores: 640

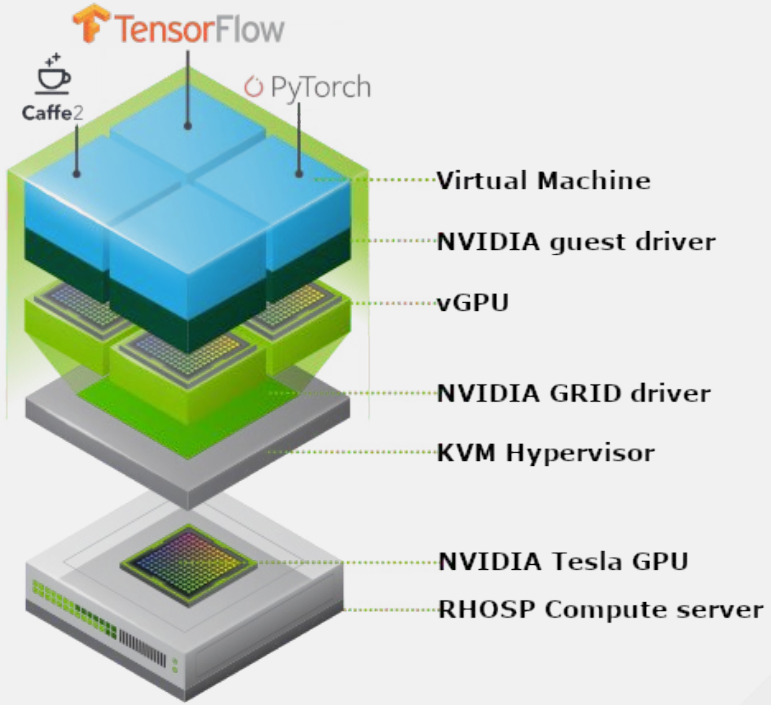
GPU Inference 47X higher throughput than CPU

Dlib with CPU, from 5FPS (240p) to 1FPS (1080p)

Workload: ResNet-50 | CPU: 1X Xeon E5-2690v4 @ 2.6 GHz | GPU: Add 1X Tesla P100 or V100

NVIDIA vGPU with GRID driver





Source: [NVIDIA software documentation]

vGPU Red Hat OpenStack Platform configuration

- Define a new vGPU compute node role in OSP director

```
parameter_defaults:
  ComputeExtraConfig:
    nova::compute::vgpu::enabled_vgpu_types:
      - nvidia-35
```

- Create a new overcloud image
- Deploy

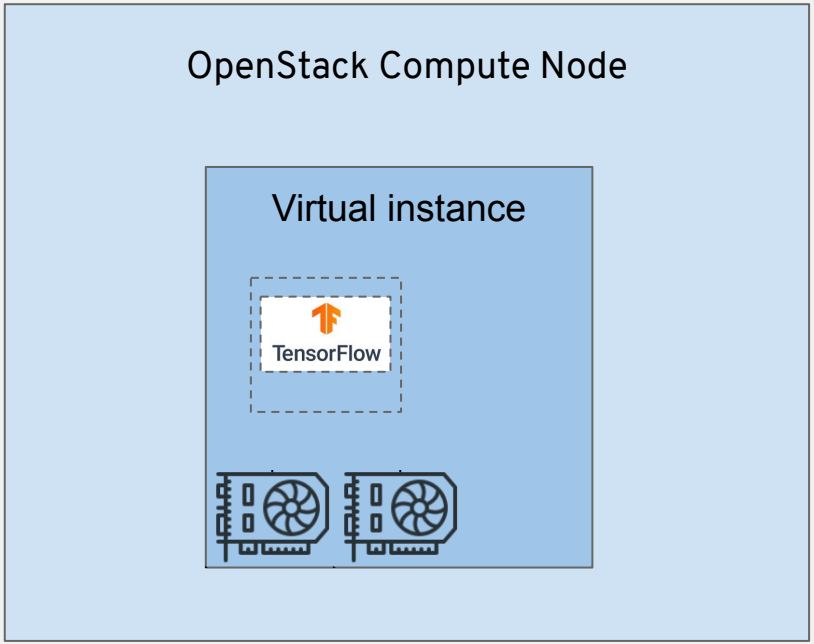
- Make sure the guest image includes the NVIDIA driver
- Create flavor(s) on demand

```
$ openstack flavor create --vcpus 4 --ram 4096 --disk 20 vgpu
$ openstack flavor set vgpu --property "resources:VGPU=1"
```

Red Hat documentation:
https://access.redhat.com/documentation/en-us/red_hat_openstack_platform/14/html/instances_and_images_guide/ch-virtual-gpu

GPU training demo

TensorFlow training performance



TensorFlow

```
TensorFlow:   1.13
Model:        resnet50
Dataset:      imagenet
Mode:         training
Accelerator:  NVIDIA GPU
OS:           RHEL 7.6
```


Red Hat

CPU

```
Step            Img/sec                    total_loss
1   images/sec: 2.6 +/- 0.0 (jitter = 0.0)   8.108
10  images/sec: 2.7 +/- 0.0 (jitter = 0.0)   8.122
20  images/sec: 2.6 +/- 0.0 (jitter = 0.0)   7.983
...
----------------------------------------------------
total images/sec: 2.6
```

GPU

```
Step            Img/sec                    total_loss
1   images/sec: 157.9 +/- 0.0 (jitter = 0.0) 8.011
10  images/sec: 158.2 +/- 0.3 (jitter = 0.5) 7.732
20  images/sec: 158.2 +/- 0.2 (jitter = 0.6) 7.686
...
----------------------------------------------------
total images/sec: 157.85
```

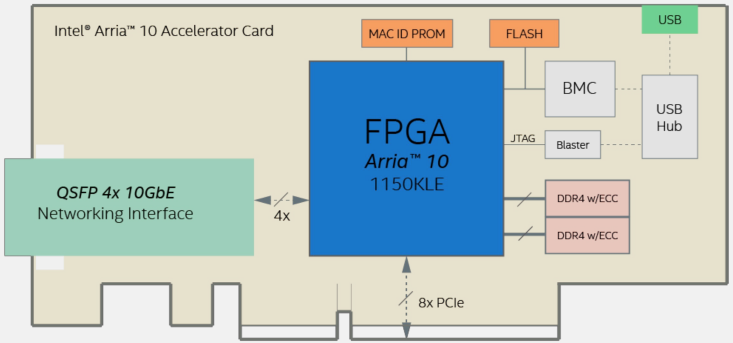
TensorFlow

```
TensorFlow:  1.13
Model:       resnet50
Dataset:     imagenet
Mode:        training
Accelerator: NVIDIA GPU
OS:          RHEL 7.6
```

Code: https://github.com/tensorflow/benchmarks

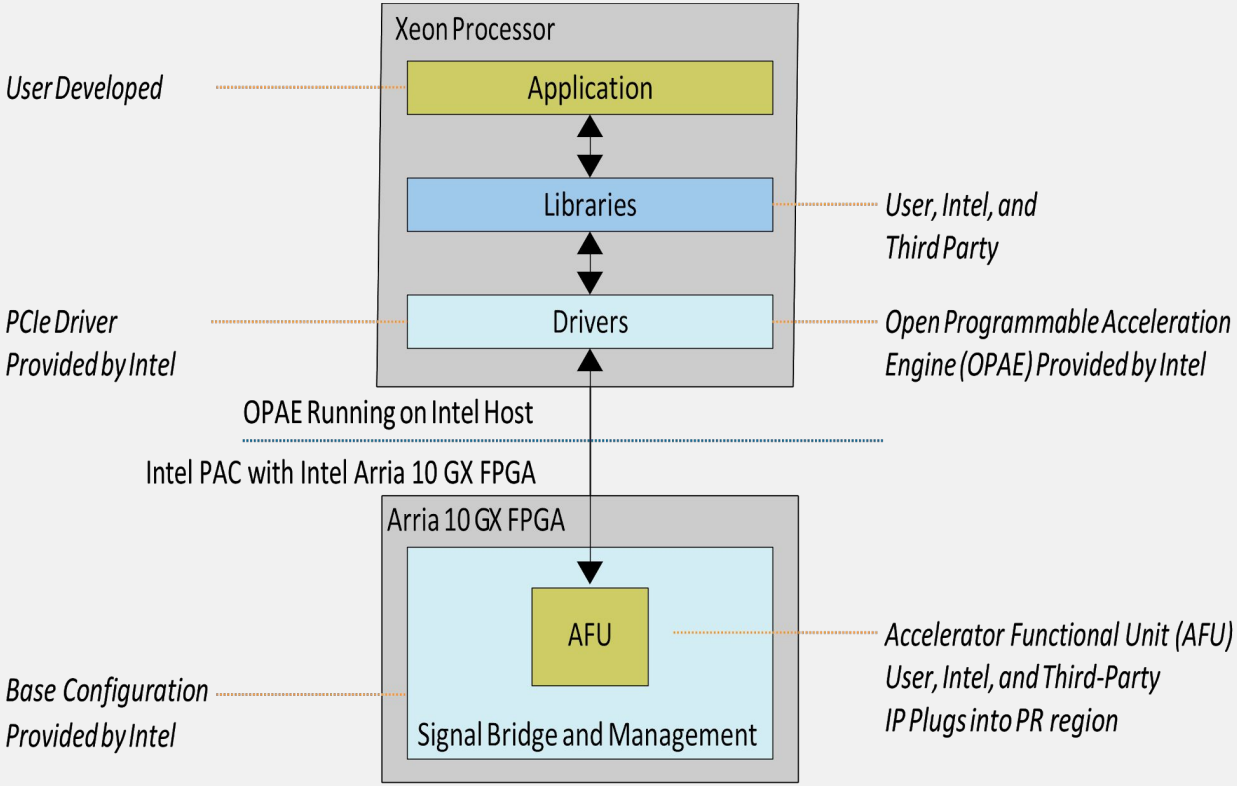
Intel Programmable Acceleration Card FPGA

Intel Programmable Acceleration Card



- Intel PAC with Intel Arria 10 GX FPGA
- Arria® 10 GX FPGA
  - Passively cooled
  - 1150K logic elements available
  - 53 Mb of embedded memory
- On-board Memory
  - 8 Gbytes DDR4 Memory Banks with ECC
  - 1Gb Mbit (128 MB) Flash
- Interfaces
  - PCIe x8 Gen3 electrical, x16 mechanical
  - 1x QSFP28 with 4x 10GbE support

Open Programmable Acceleration Engine (OPAE)



Open Programmable Acceleration Engine (OPAE)

- Consistent API across product generations and platforms abstraction for hardware specific FPGA resource details
- Designed for minimal software overhead and latency
  - Lightweight user-space library (libfpga)
- Open ecosystem for industry and developer community
  - License: FPGA API (BSD), FPGA driver (GPLv2)
- Supports both virtual machines and bare metal platforms

Intel Programmable Acceleration Card: fpgainfo

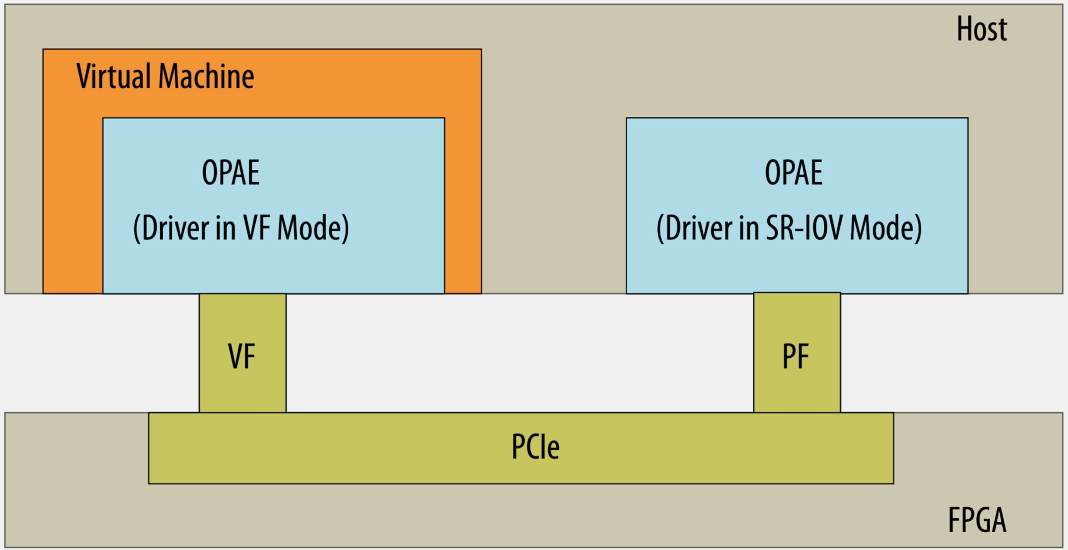
```
[stack@overcloud-computefpga-0 ~]$ fpgainfo fme
Board Management Controller, microcontroller FW version 26889
Last Power Down Cause: POK_CORE
Last Reset Cause: None
//****** FME ******//
Object Id                       : 0xEE00000
PCIe s:b:d:f                    : 0000:5E:00:0
Device Id                       : 0x09C4
Socket Id                       : 0x00
Ports Num                       : 01
Bitstream Id                    : 0x123000200000185
Bitstream Version               : 0x557D00030201
Pr Interface Id                 : 69528db6-eb31-577a-8c36-68f9faa081f6
```

Intel Programmable Acceleration Card: fpgainfo

```
[stack@overcloud-computefpga-0 ~]$ fpgainfo power
Board Management Controller, microcontroller FW
version 26889
Last Power Down Cause: POK_CORE
Last Reset Cause: None
//****** POWER ******//
Object Id                     : 0xF100000
PCIe s:b:d:f                  : 0000:02:00:0
Device Id                     : 0x09C4
Socket Id                     : 0x00
Ports Num                     : 01
Bitstream Id                  : 0x123000200000185
Bitstream Version             : 0x30201
Pr Interface Id               :
69528db6-eb31-577a-8c36-68f9faa081f6
( 0) Total Input Power        : 37.50 Watts
( 1) PCIe 12V Current         : 3.15 Amps
( 2) PCIe 12V Voltage         : 11.60 Volts
( 3) 1.2V Voltage             : 1.21 Volts
( 4) 1.2V Current             : 2.66 Amps
( 5) 1.8V Voltage             : 1.83 Volts
( 6) 1.8V Current             : 3.09 Amps
( 7) 3.3V Mgmt Voltage        : 3.33 Volts
( 8) 3.3V Current             : 0.72 Amps
( 9) FPGA Core Voltage        : 0.91 Volts
(10) FPGA Core Current        : 19.30 Amps
(13) QSFP P3V3                : No reading
(reading state unavailable)
(16) Core Supply Temp Input   : 0.45 Volts
(17) VCCR Voltage             : 1.05 Volts
(18) VCCT Voltage             : 1.04 Volts
(19) VCCR Current             : 1.14 Amps
(20) VCCT Current             : 0.33 Amps
(21) VPP Voltage              : 2.54 Volts
(22) VTT Voltage              : 0.59 Volts
```

Running the OPAE in a Virtualized Environment

In SR-IOV mode, a host processor uses a physical function (PF) to access management functions. A virtual machine (VM) uses a virtual function (VF) to access the AFU.



PF: Physical Function
VF: Virtual Function

Intel Programmable Acceleration Card

- SR-IOV mode

```
[stack@baremetalhost ~]$ lspci -nn | grep accelerators
5e:00.0 Processing accelerators [1200]: Intel Corporation Device [8086:09c4]
```

- Virtualized mode

```
[stack@overcloud-computefpga-0 ~]$ lspci -nn | grep accelerators
5e:00.0 Processing accelerators [1200]: Intel Corporation Device [8086:09c4]
5e:00.1 Processing accelerators [1200]: Intel Corporation Device [8086:09c5]

[stack@instancefpga-0 ~]$ lspci -nn | grep accelerators
00:09.0 Processing accelerators [1200]: Intel Corporation Device [8086:09c5]
```

Face detection demo
Intel PAC with Intel Arria 10 GX FPGA

Red Hat

OpenVINO™

```
OpenVINO:     2019.1.144
Model:        face-detection-adas-0001
Mode:         Inference
Framework:    Caffe
Accelerator:  Intel PAC with Intel Arria 10 GX FPGA
Devices:      FPGA+CPU
OS:           RHEL 7.6
```

- fpgainfo power without load

```
( 0) Total Input Power        : 38.50 Watts
```

- fpgainfo power with face detection

```
( 0) Total Input Power        : 44.00 Watts
```


Red Hat

```
[egallen@overcloud-computefpga-0 ~]$ ./fpga-face-detection.sh
[ INFO ] InferenceEngine:
       API version ............ 1.6
       Build .................. custom_releases/2019/R1.1_28dfbfdd28954c4dfd2f94403dd8dfc1f411038b
[ INFO ] Parsing input parameters
[ INFO ]     Detection model:           /home/egallen/face-detection-adas-0001_FP32/face-detection-adas-0001.xml
[ INFO ]     Detection threshold:       0.5
[ INFO ]     Utilizing device:          HETERO:FPGA,CPU
[ INFO ]     CPU extension library:
/home/egallen/inference_engine_samples_build/intel64/Release/lib/libcpu_extension.so
[ INFO ]     Batch size:                1
[ INFO ]     Number of infer requests:  5
[ INFO ]     Number of input web cams:  0
[ INFO ] Model   path: /home/egallen/face-detection-adas-0001_FP32/face-detection-adas-0001.xml
[ INFO ] Weights path: /home/egallen/face-detection-adas-0001_FP32/face-detection-adas-0001.bin
[ INFO ] Files were added: 1
[ INFO ]     /tmp/erwan-geneva1.mp4
[ INFO ]     Number of input web cams:    0
[ INFO ]     Number of input video files: 1
[ INFO ]     Duplication multiplayer:     0
[ INFO ]     Number of input channels:    1
[ INFO ] Trying to open input video ...
To close the application, press 'CTRL+C' or any key with focus on the output window
```

References

References

Red Hat OpenStack Platform, https://www.redhat.com/fr/technologies/linux-platforms/openstack-platform

NVIDIA GRID, https://www.nvidia.com/en-us/data-center/graphics-cards-for-virtualization

Intel Acceleration Hub, https://www.intel.com/content/www/us/en/programmable/solutions/acceleration-hub/overview.html

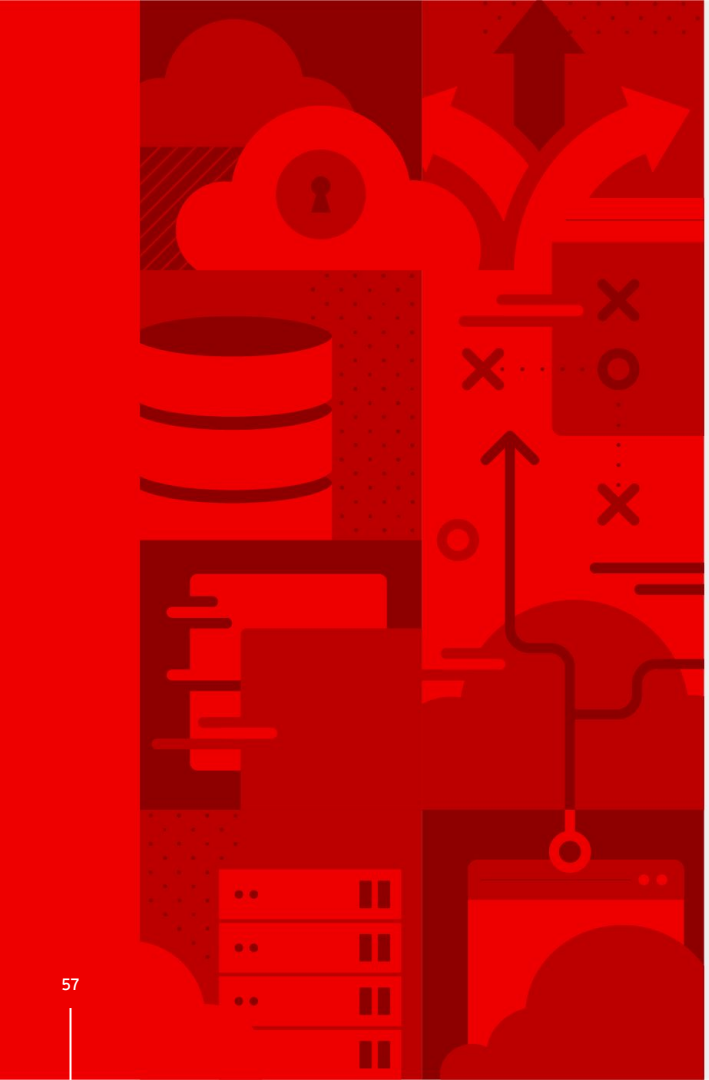
Intel OpenVINO, https://software.intel.com/en-us/openvino-toolkit

Thanks to:

Davis King, Dlib, http://dlib.net/

Adam Geitgey, face_recognition, https://github.com/ageitgey/face_recognition

Adrian Rosebrock, https://pyimagesearch.com



Thank you

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat