THE LINUX FOUNDATION
OPEN SOURCE SUMMIT

Linaro

Deep Learning Neural Network Acceleration at the Edge

Andrea Gallo
VP Segments and Strategic Initiatives

29-Aug-2018
Vancouver

THE LINUX FOUNDATION

LEADING COLLABORATION IN THE ARM ECOSYSTEM

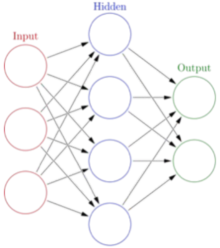
Disclaimer
All information in this session is public

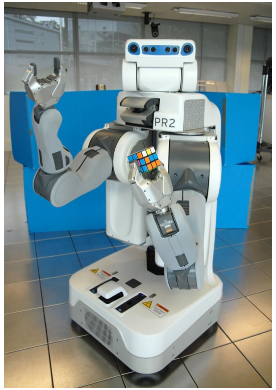
No confidential information has been disclosed from private communication between Linaro and Linaro members

URL’s to the original source are provided in each slide

Why Deep Learning?

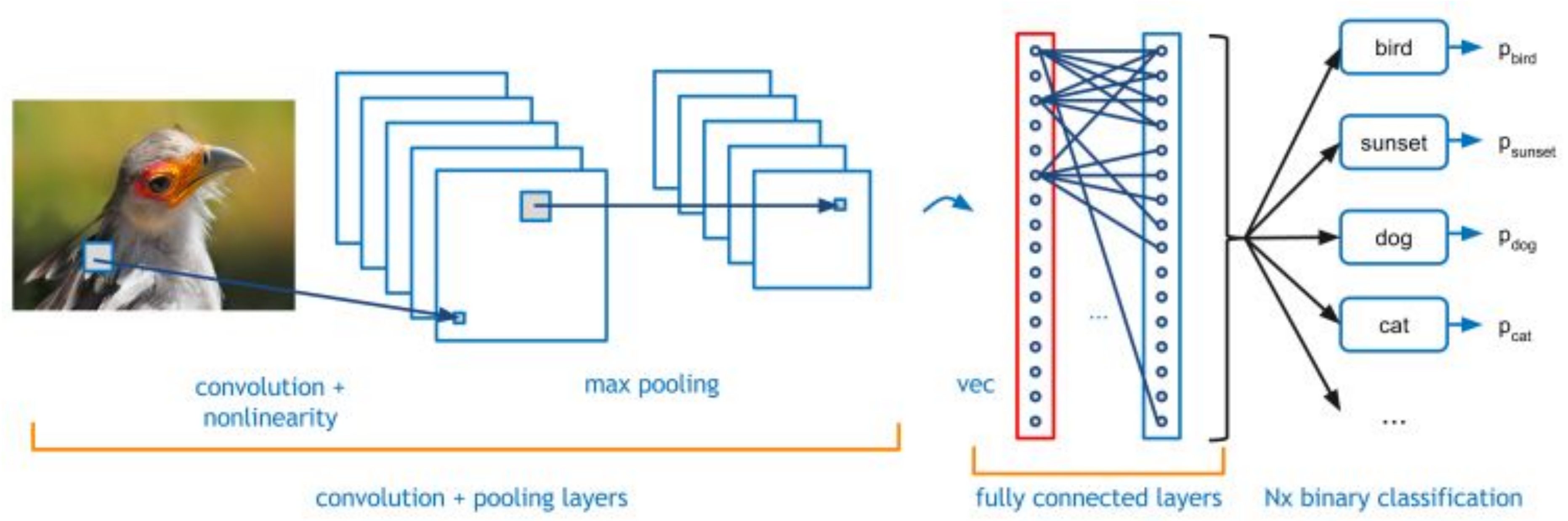
End-to-End Learning for Many Tasks





Slide from [DIY Deep Learning for Vision: a Hands-On Tutorial with Caffe]

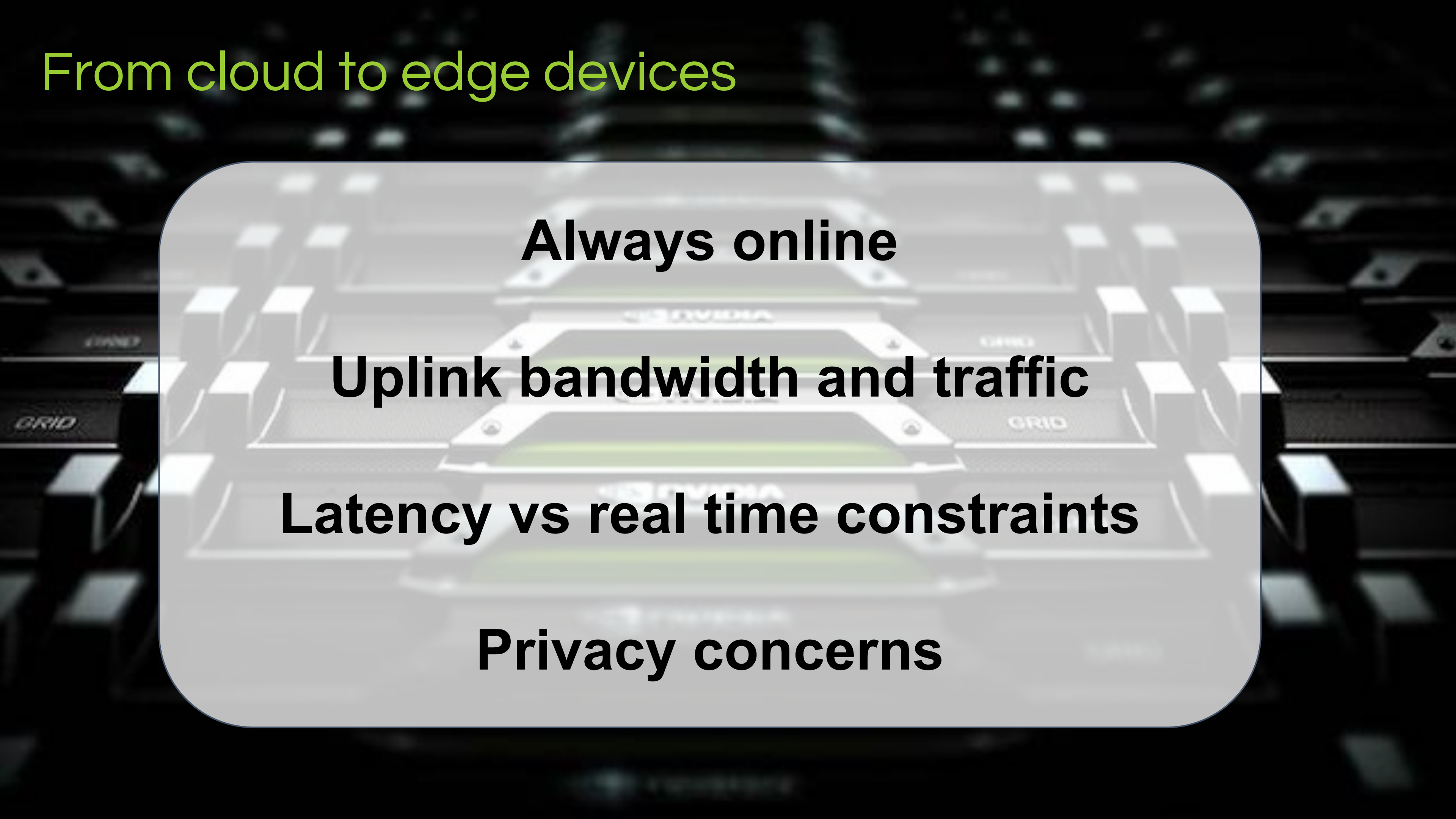
It’s complex!!!



Slide from DIY Deep Learning for Vision: a Hands-On Tutorial with Caffe

From cloud to edge devices

From cloud to edge devices



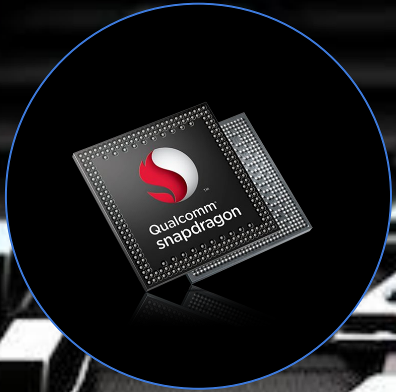
Always online

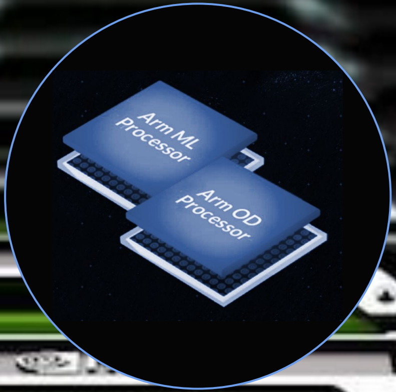
Uplink bandwidth and traffic

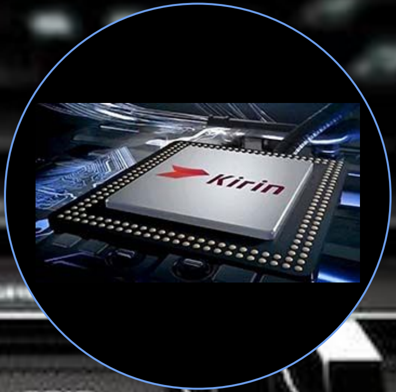
Latency vs real time constraints

Privacy concerns

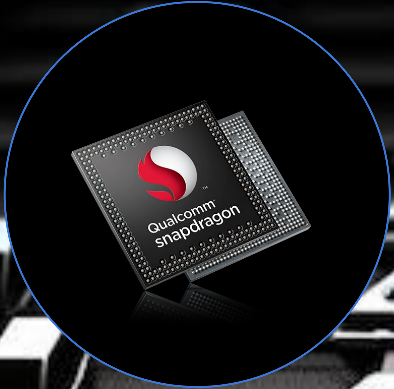
From cloud to edge devices

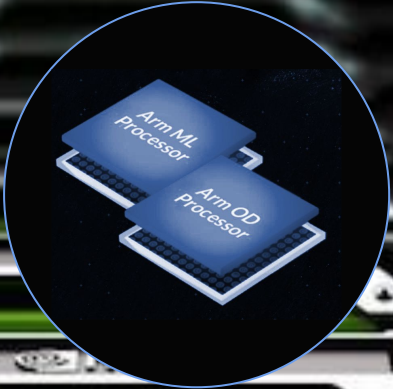


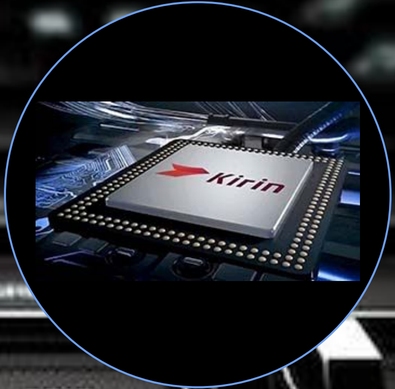


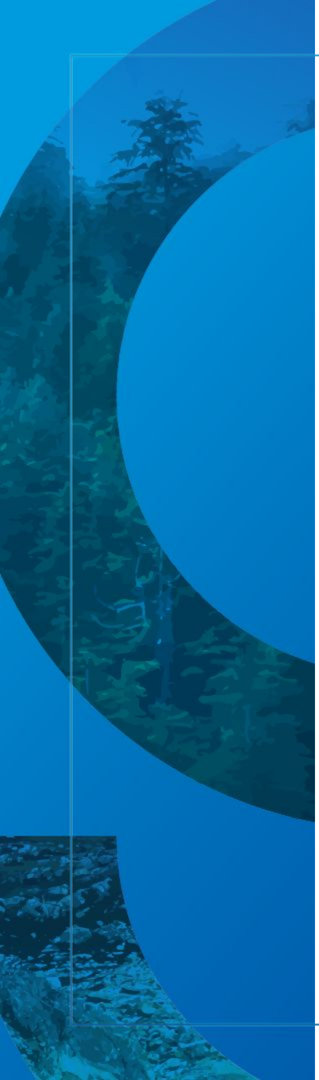


From cloud to edge devices









AI/ML Frameworks

TensorFlow and TensorFlow Lite

Developed in-house by the Google Brain team

- Started as DistBelief in 2011
- Evolved into TensorFlow with its first commit in November 2015
- V1.0.0 released on Feb 11, 2017

TensorFlow and TensorFlow Lite

Developed in-house by the Google Brain team

- Started as DistBelief in 2011
- Evolved into TensorFlow with its first commit in November 2015
- V1.0.0 released on Feb 11, 2017

TensorFlow can be built as

- TensorFlow for cloud and datacenters
- TensorFlow Lite for mobile devices
- TensorFlow.js for AI in web browsers

TensorFlow and TensorFlow Lite

Developed in-house by the Google Brain team

- Started as DistBelief in 2011
- Evolved into TensorFlow with its first commit in November 2015
- V1.0.0 released on Feb 11, 2017

TensorFlow can be built as

- TensorFlow for cloud and datacenters
- TensorFlow Lite for mobile devices
- TensorFlow.js for AI in web browsers

Support multiple accelerators

→ CUDA and TPU

→ Android NNAPI and NN HAL

→ WebGL

TensorFlow and TensorFlow Lite

Developed in-house by the Google Brain team

- Started as DistBelief in 2011
- Evolved into TensorFlow with its first commit in November 2015
- V1.0.0 released on Feb 11, 2017

TensorFlow can be built as

- TensorFlow for cloud and datacenters
- TensorFlow Lite for mobile devices
- TensorFlow.js for AI in web browsers

Support multiple accelerators

→ CUDA (

→ Androi

→ WebGL

31,713 commits

1,624 contributors

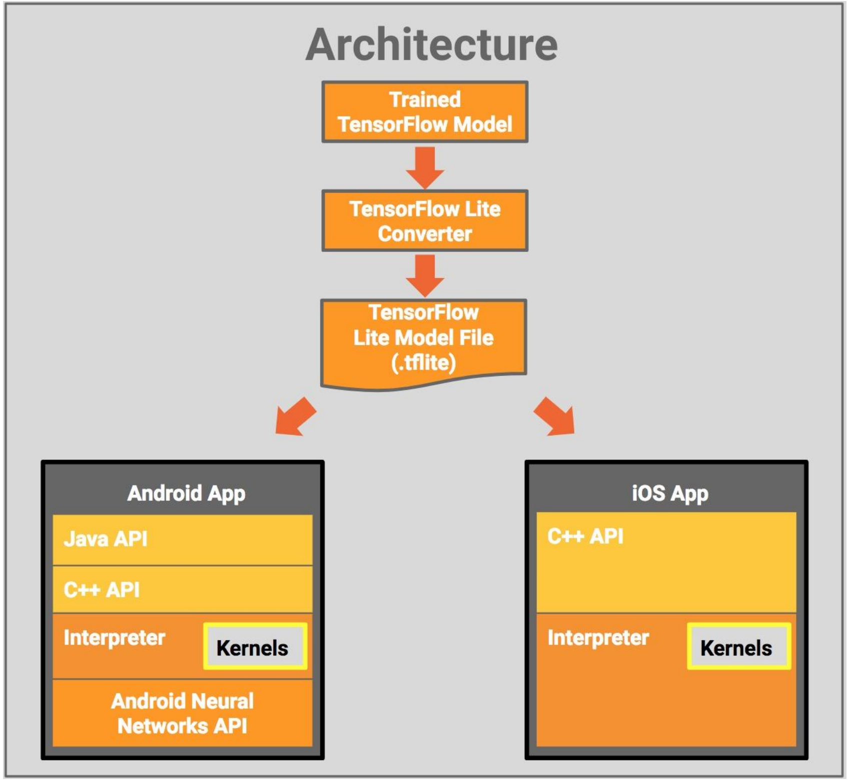
1,610,734 lines of code

456 years of effort

1st Commit Nov ‘15

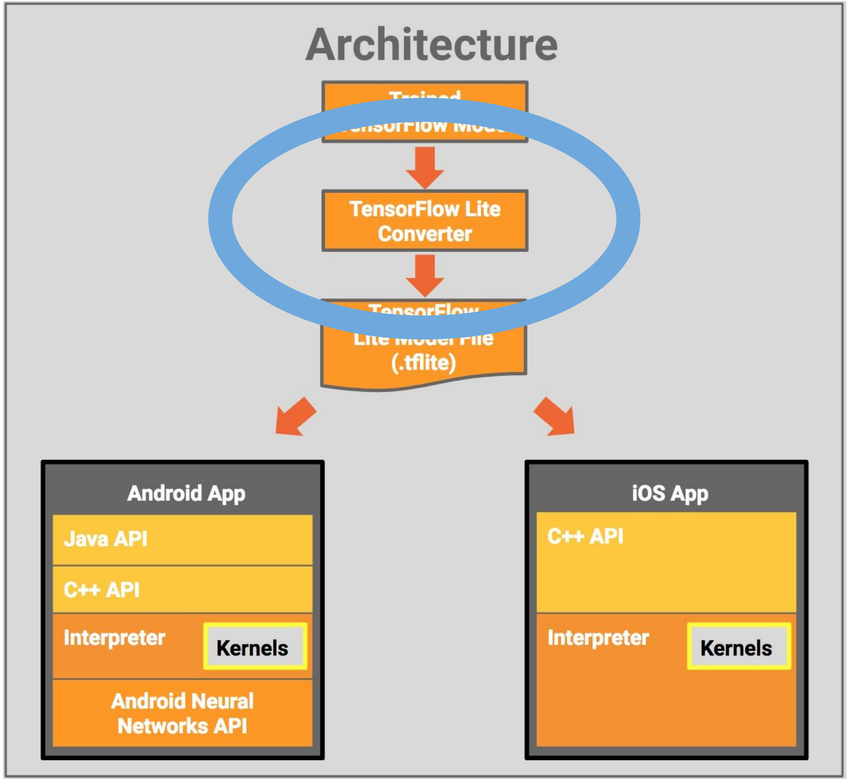
BLACKDUCK | Open Hub

From TensorFlow to TensorFlow Lite



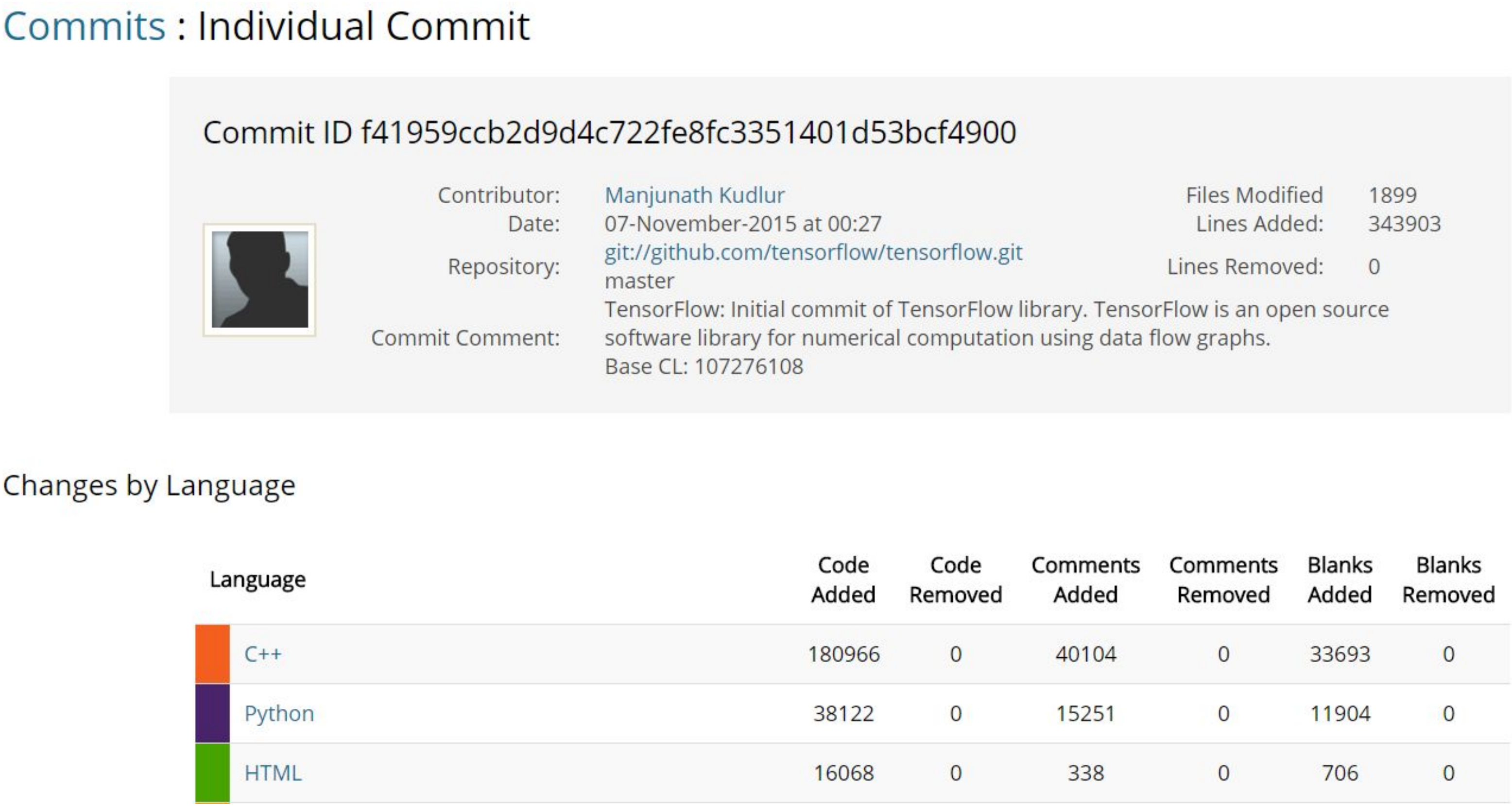
TensorFlow Lite uses FlatBuffers

From TensorFlow to TensorFlow Lite



TensorFlow Lite uses FlatBuffers

TensorFlow 1st Commit in November 2015



Commits : Individual Commit

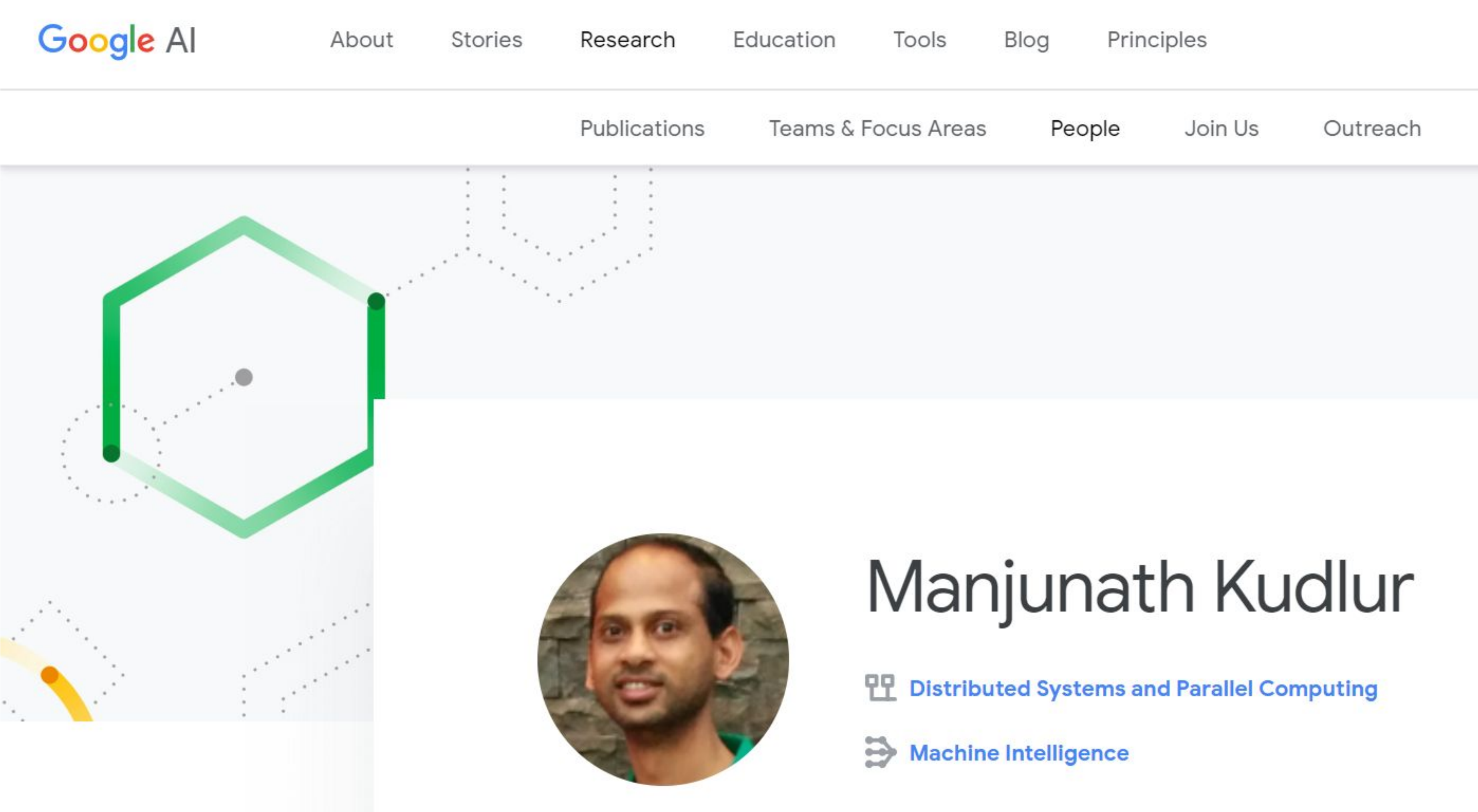
Commit ID f41959ccb2d9d4c722fe8fc3351401d53bcf4900

Contributor:	Manjunath Kudlur	Files Modified	1899
Date:	07-November-2015 at 00:27	Lines Added:	343903
Repository:	git://github.com/tensorflow/tensorflow.git master	Lines Removed:	0
Commit Comment:	TensorFlow: Initial commit of TensorFlow library. TensorFlow is an open source software library for numerical computation using data flow graphs. Base CL: 107276108		

Changes by Language

Language	Code Added	Code Removed	Comments Added	Comments Removed	Blanks Added	Blanks Removed
🟧 C++	180966	0	40104	0	33693	0
🟪 Python	38122	0	15251	0	11904	0
🟩 HTML	16068	0	338	0	706	0

BLACKDUCK | Open Hub



Manjunath Kudlur

Distributed Systems and Parallel Computing

Machine Intelligence

Caffe

- Made with expression, speed, and modularity in mind
- Developed by Berkeley AI Research (BAIR) and by community contributors
  - **Yangqing Jia** created the project during his PhD at UC Berkeley
  - Caffe is released under the BSD 2-Clause license
- Focus has been vision, but also handles sequences, reinforcement learning, speech + text
- Tools, reference models, demos, and recipes → Caffe Zoo
- Seamless switch between CPU and GPU

4,137 commits

314 contributors

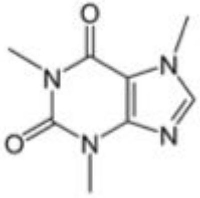
76,076 lines of code

19 years of effort

1st commit in Sept'13

15,000+ forks

BLACKDUCK | Open Hub



caffe.berkeleyvision.org

github.com/BVLC/caffe

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Caffe2

Facebook Open Source

Caffe2 improves Caffe 1.0 in a series of directions

- First-class support for large-scale distributed training
- Mobile deployment
- New hardware support (in addition to CPU and CUDA)
- Flexibility for future directions such as quantized computation
- Stress tested by the vast scale of Facebook applications
- Examples and pre-trained models available from the Caffe2 Zoo
- Running on mobile devices with Android and iOS
  - Step-by-step tutorial with camera demo
- Caffe1 models do not run with Caffe2
  - Converter tool available

3,678 commits

332 contributors

275,560 lines of code

73 years of effort

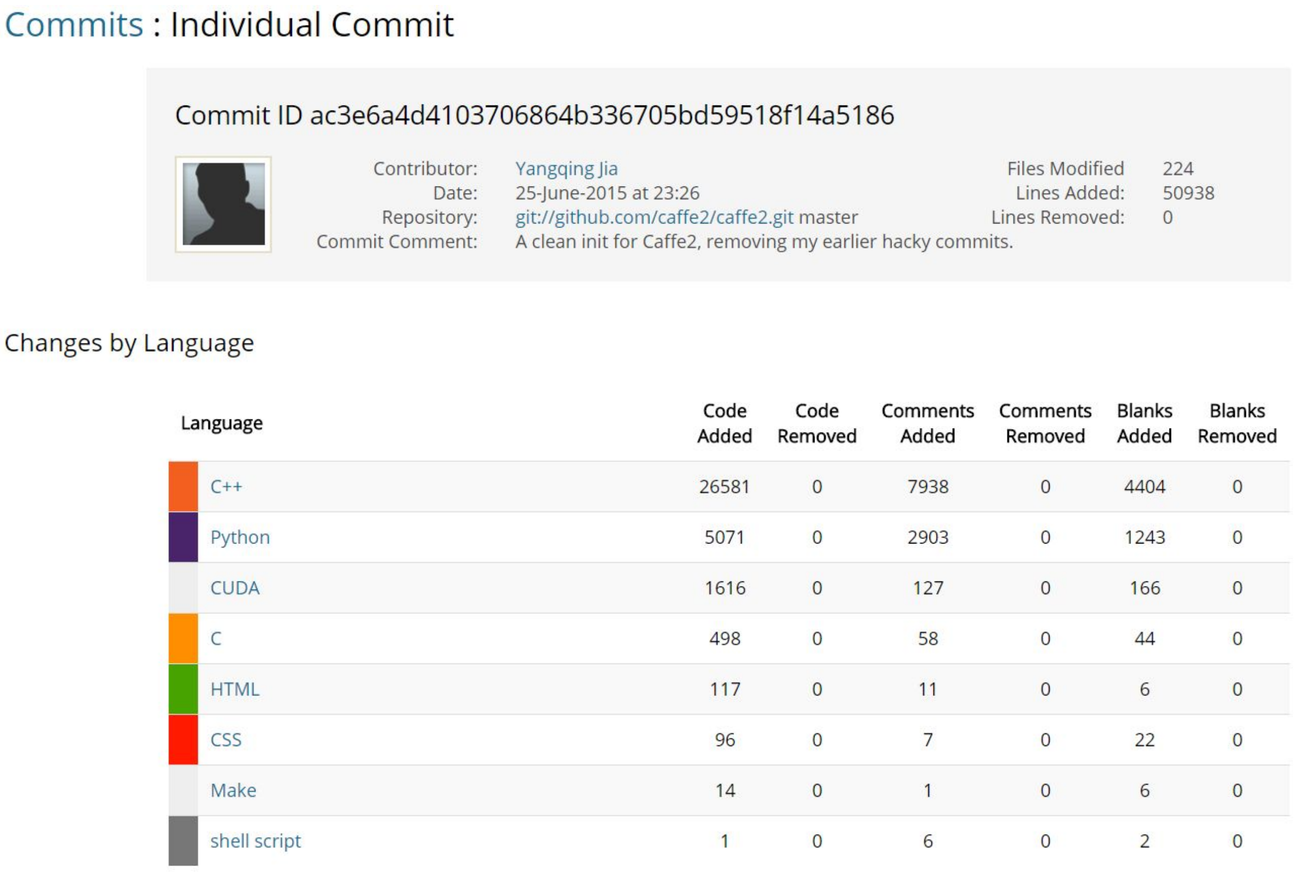
1st commit in June ‘15

BLACKDUCK | Open Hub

Caffe2 1st commit in June 2015

Facebook Open Source

Caffe2



Commits : Individual Commit

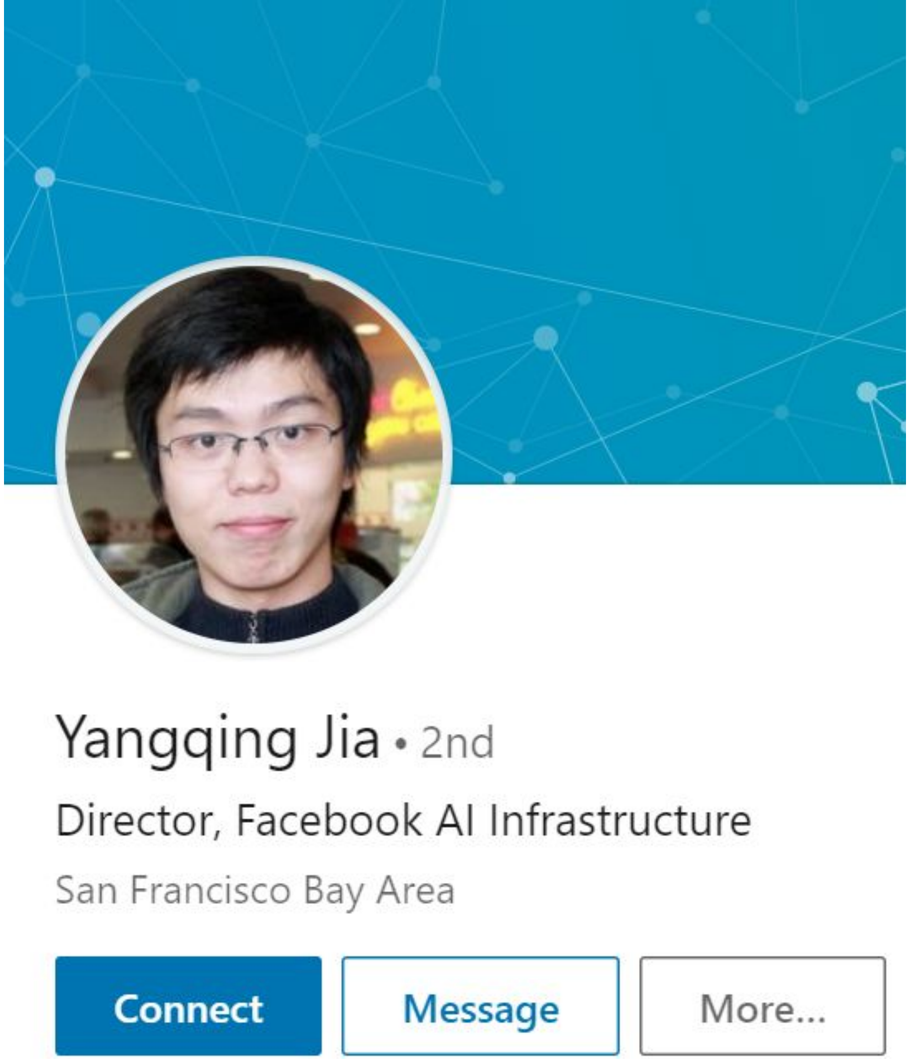
Commit ID ac3e6a4d4103706864b336705bd59518f14a5186

Contributor:	Yangqing Jia	Files Modified	224
Date:	25-June-2015 at 23:26	Lines Added:	50938
Repository:	git://github.com/caffe2/caffe2.git master	Lines Removed:	0
Commit Comment:	A clean init for Caffe2, removing my earlier hacky commits.		

Changes by Language

Language	Code Added	Code Removed	Comments Added	Comments Removed	Blanks Added	Blanks Removed
C++	26581	0	7938	0	4404	0
Python	5071	0	2903	0	1243	0
CUDA	1616	0	127	0	166	0
C	498	0	58	0	44	0
HTML	117	0	11	0	6	0
CSS	96	0	7	0	22	0
Make	14	0	1	0	6	0
shell script	1	0	6	0	2	0

BLACKDUCK | Open Hub



Yangqing Jia • 2nd

Director, Facebook AI Infrastructure

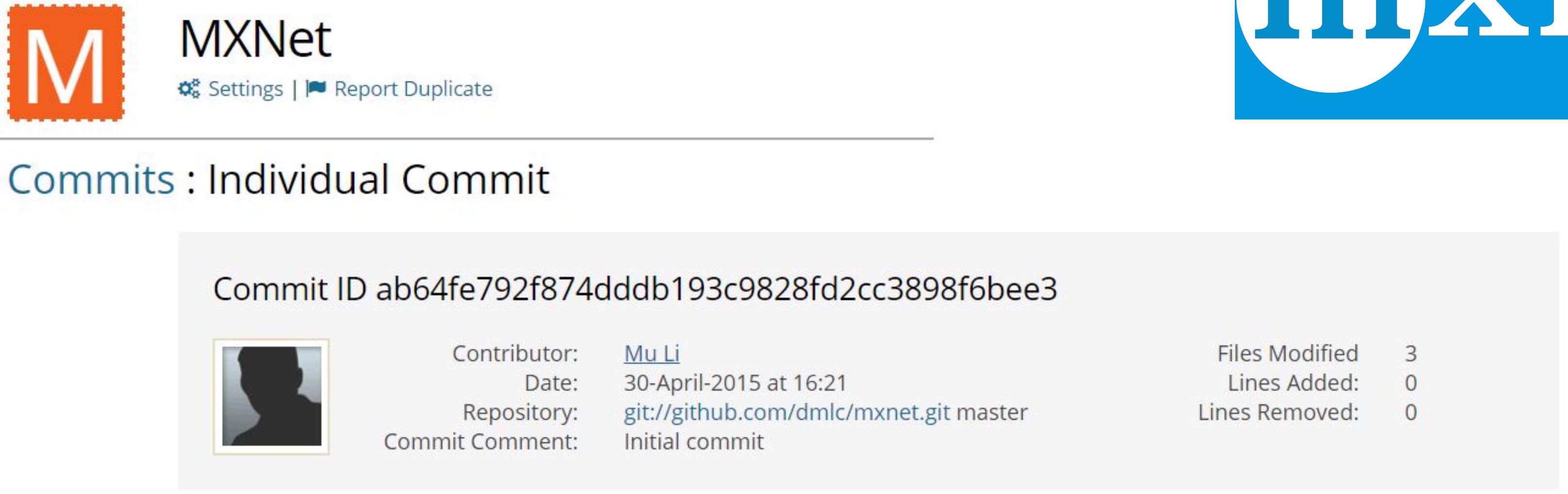
San Francisco Bay Area

MXNet is a multi-language machine learning (ML) library to ease the development of ML algorithms, especially for deep neural networks. MXNet is computation and memory efficient and runs on various heterogeneous systems, ranging from mobile devices to distributed GPU clusters.

Currently, MXNet is supported by Intel, Dato, Baidu, Microsoft, Wolfram Research, and research institutions such as Carnegie Mellon, MIT, the University of Washington, and the Hong Kong University of Science and Technology.

Gluon API, examples, tutorials and pre-trained models from the Gluon model zoo

mxnet 1st Commit in April 2015



MXNet

Settings | Report Duplicate

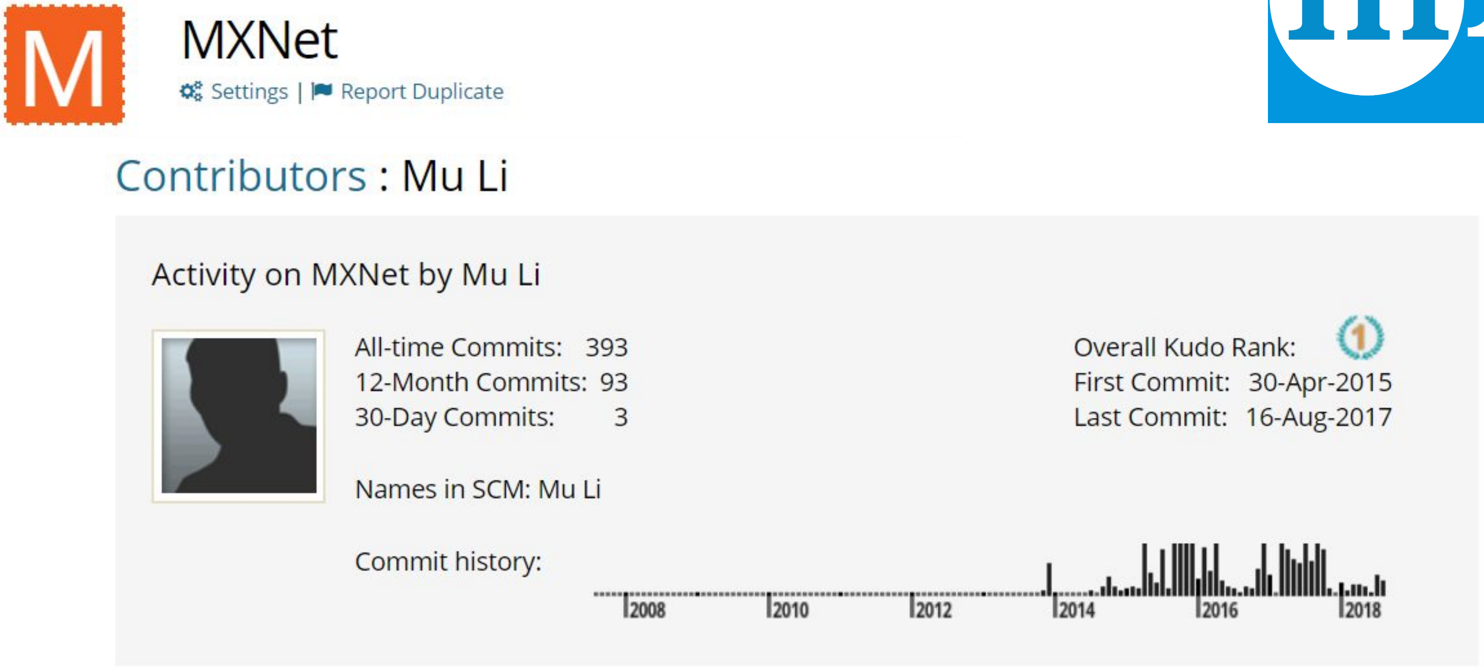
Commits : Individual Commit

Commit ID ab64fe792f874dddb193c9828fd2cc3898f6bee3

Contributor:	Mu Li	Files Modified	3
Date:	30-April-2015 at 16:21	Lines Added:	0
Repository:	git://github.com/dmlc/mxnet.git master	Lines Removed:	0
Commit Comment:	Initial commit		

BLACKDUCK | Open Hub

mxnet 1st Commit in April 2015



MXNet

Settings | Report Duplicate

Contributors : Mu Li

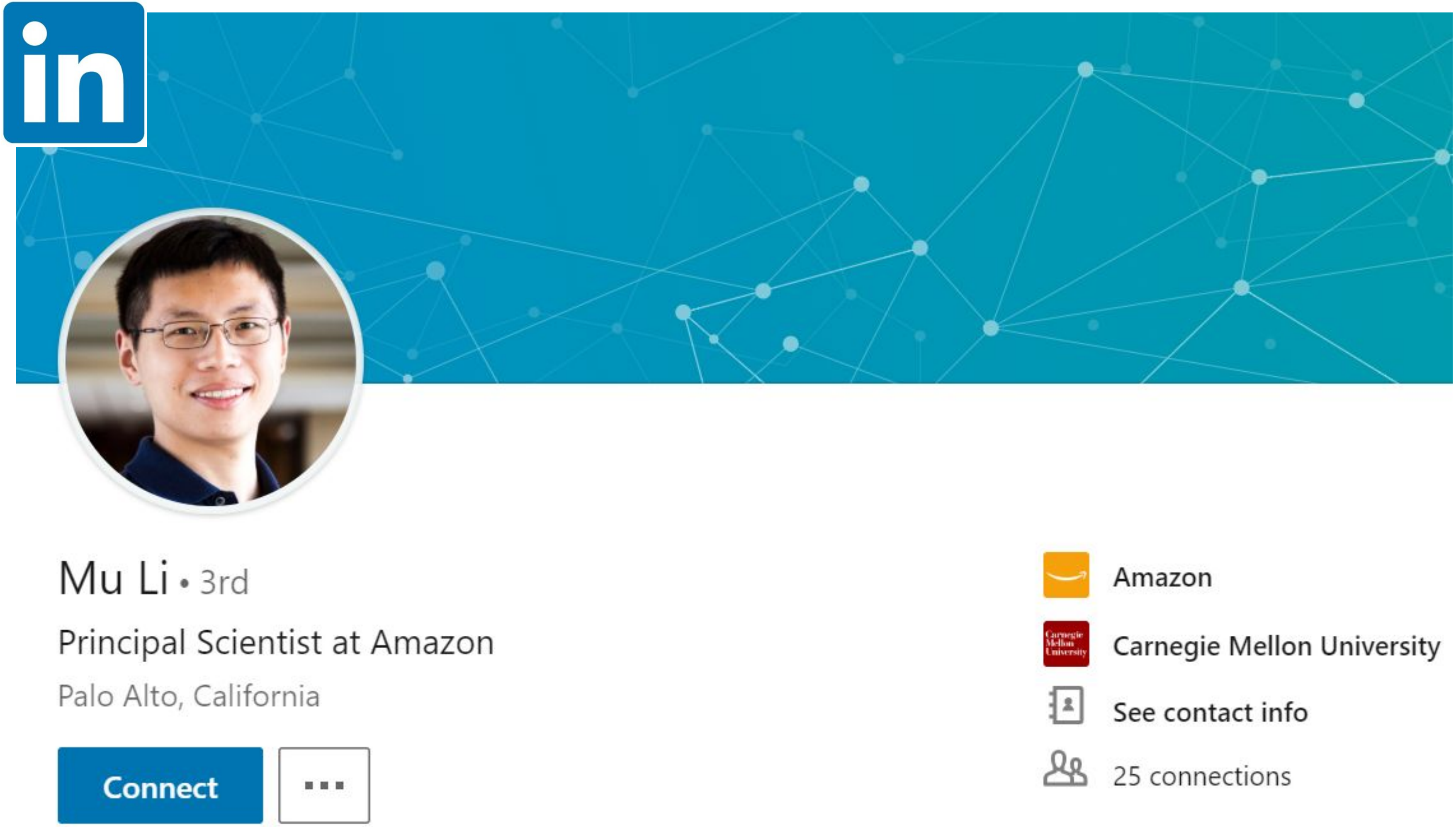
Activity on MXNet by Mu Li

All-time Commits:	393
12-Month Commits:	93
30-Day Commits:	3

Overall Kudo Rank:	1
First Commit:	30-Apr-2015
Last Commit:	16-Aug-2017

Names in SCM: Mu Li

Commit history:



Mu Li • 3rd

Principal Scientist at Amazon

Palo Alto, California

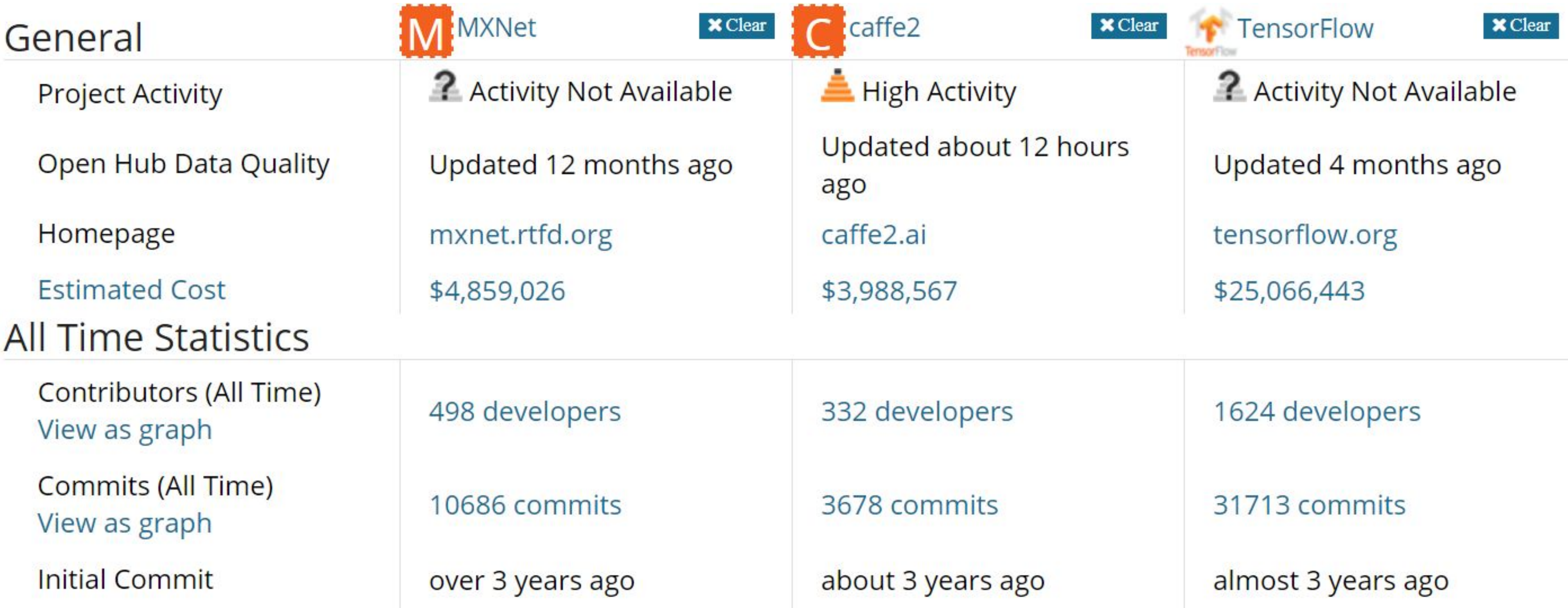
Amazon

Carnegie Mellon University

See contact info

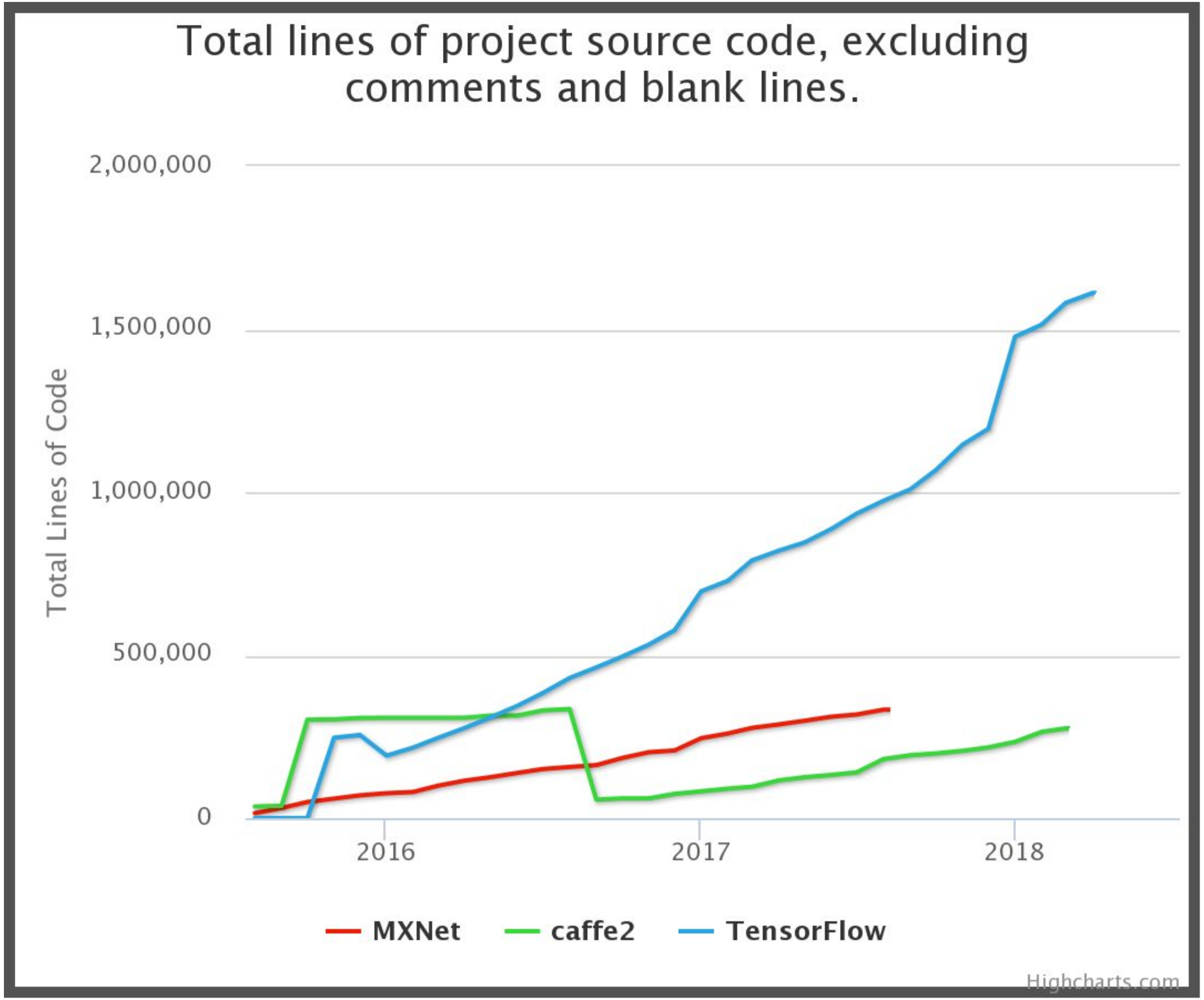
25 connections

Deep Learning framework comparison



General	M MXNet ✖Clear	C caffe2 ✖Clear	TensorFlow ✖Clear
Project Activity	? Activity Not Available	High Activity	? Activity Not Available
Open Hub Data Quality	Updated 12 months ago	Updated about 12 hours ago	Updated 4 months ago
Homepage	mxnet.rtfd.org	caffe2.ai	tensorflow.org
Estimated Cost	$4,859,026	$3,988,567	$25,066,443
All Time Statistics			
Contributors (All Time) View as graph	498 developers	332 developers	1624 developers
Commits (All Time) View as graph	10686 commits	3678 commits	31713 commits
Initial Commit	over 3 years ago	about 3 years ago	almost 3 years ago

BLACKDUCK | Open Hub



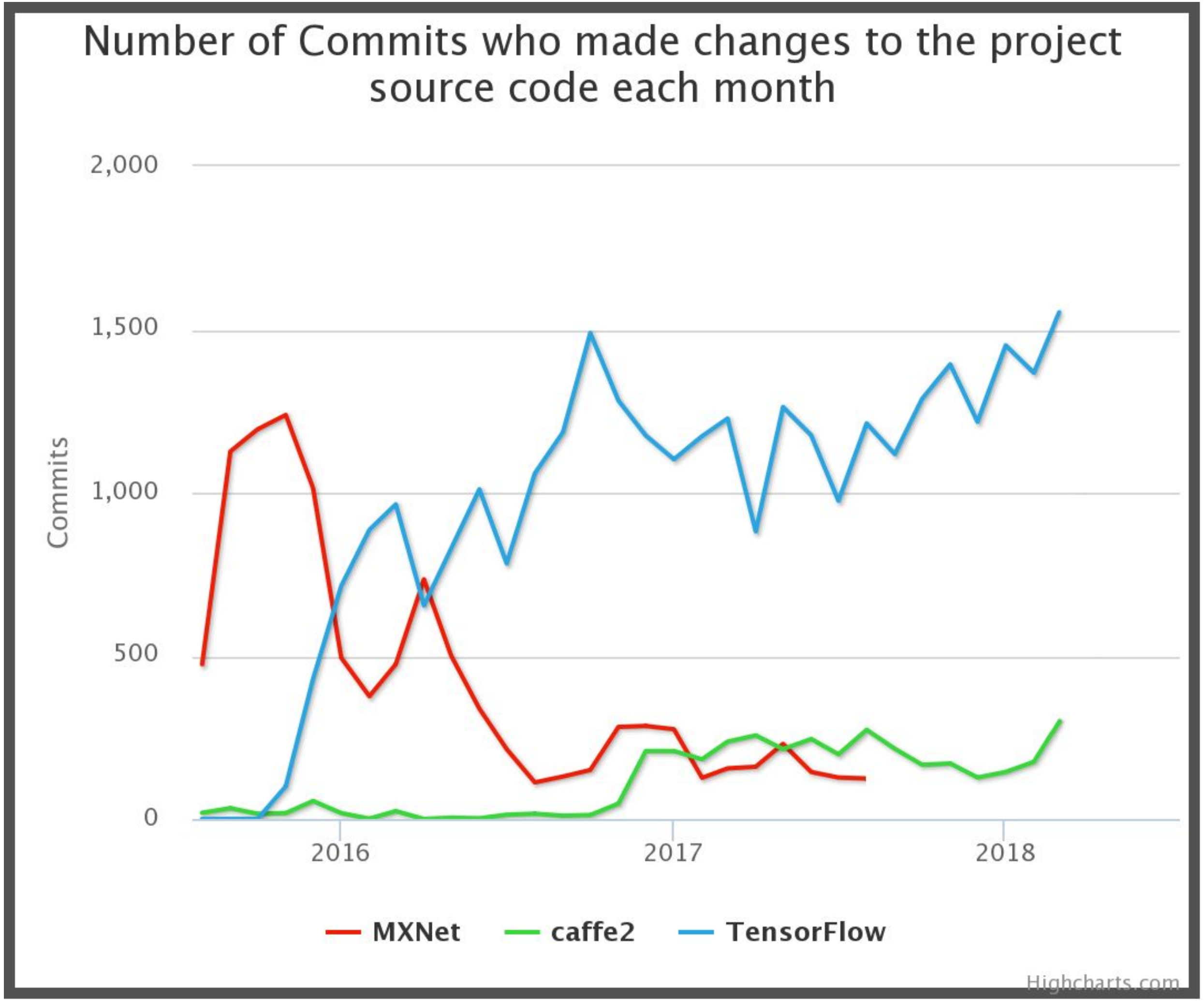
Total lines of project source code, excluding comments and blank lines.

mxnet

Caffe2

TensorFlow

BLACKDUCK | Open Hub



mxnet

Caffe2

TensorFlow

BLACKDUCK | Open Hub

Linaro

LEADING COLLABORATION
IN THE ARM ECOSYSTEM

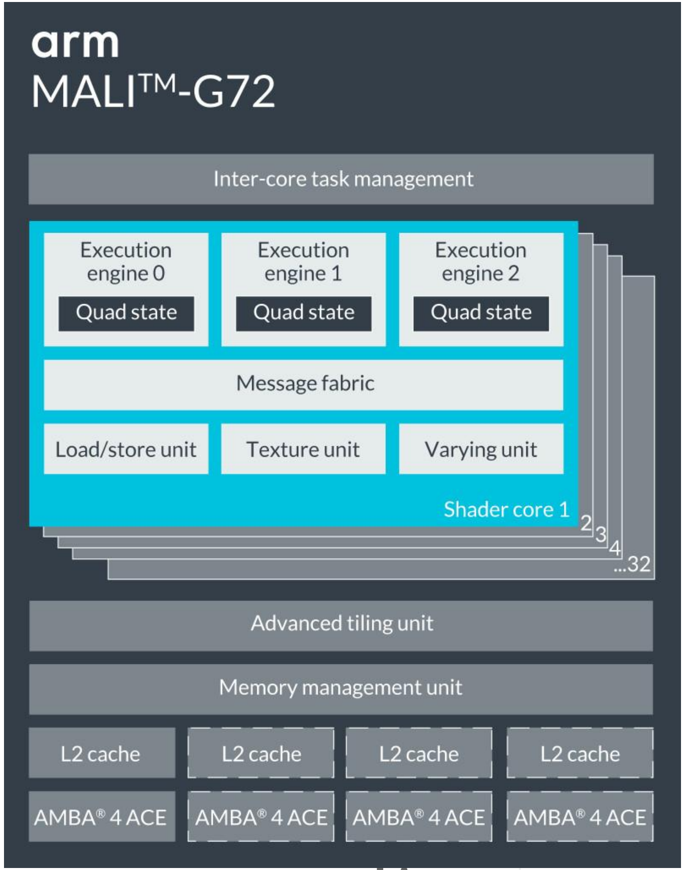
Observations

- Each cloud player has its own deep learning framework
- Each AI framework has its own entire ecosystem of formats, tools, model store
- Each AI framework represents a significant investment
- Scaling and acceleration are fundamental to performance

NN accelerators and software solutions

Arm Mali-G72

Arm Mali-G72 is the second generation Bifrost-based GPU for High Performance products. Benefitting from advanced technologies such as claused shaders and full system coherency, Mali-G72 adds increased tile buffer memory supporting up to 16 x Multi-Sample Anti-Aliasing at minimal performance cost. Arithmetic optimizations tailored to complex Machine Learning and High Fidelity Mobile Gaming use cases provide 25% higher energy efficiency, 20% better performance density and 40% greater overall performance than devices based on previous generation Bifrost GPU.



https://developer.arm.com/products/graphics-and-multimedia/mali-gpus/mali-g72-gpu

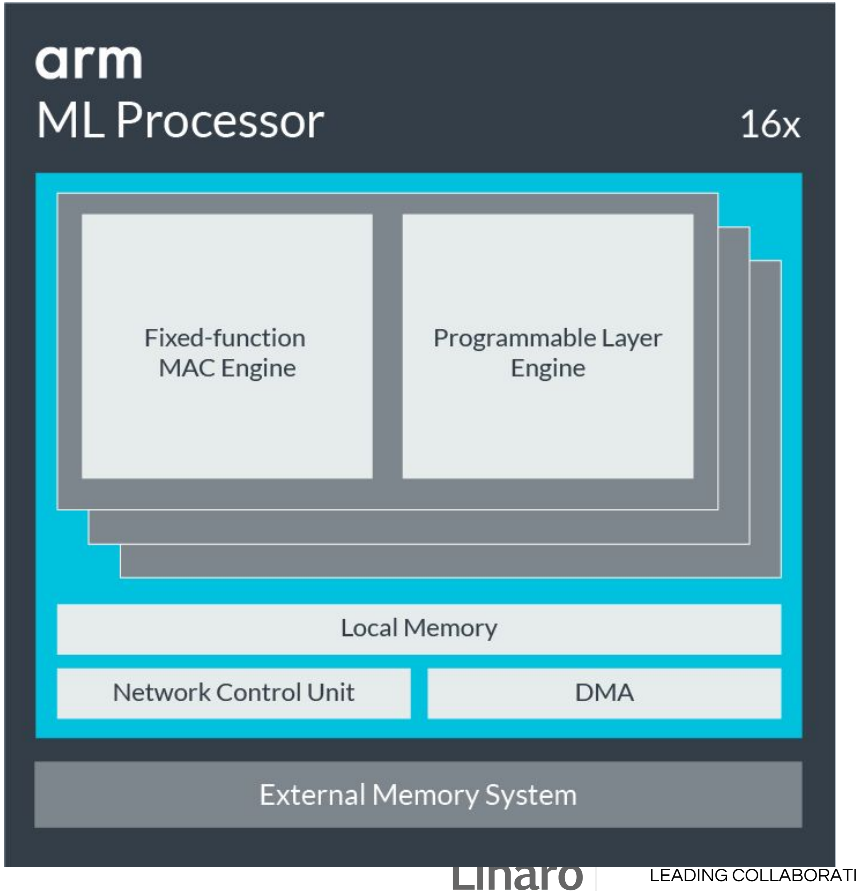
Arm ML processor

The Arm Machine Learning processor is an optimized, ground-up design for machine learning acceleration, targeting mobile and adjacent markets:

- optimized fixed-function engines for best-in-class performance
- additional programmable layer engines support the execution of non-convolution layers, and the implementation of selected primitives and operators

The network control unit manages the overall execution and traversal of the network and the DMA moves data in and out of the main memory.

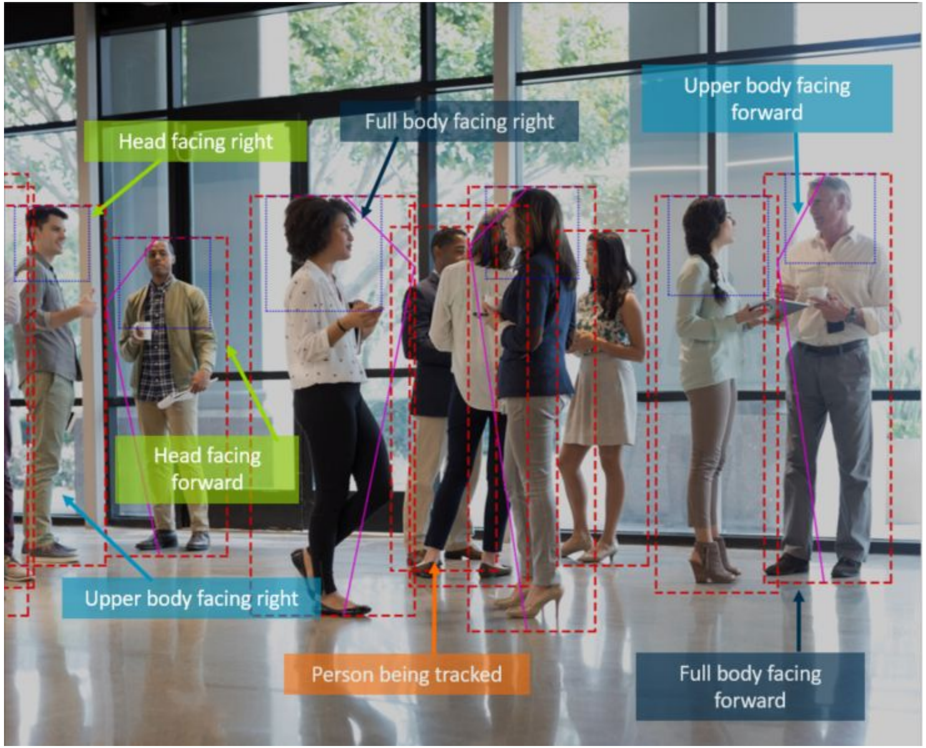
Onboard memory allows central storage for weights and feature maps



https://developer.arm.com/products/processors/machine-learning/arm-ml-processor

Arm OD processor

- Detects object in real time with Full HD at 60fps.
- Object sizes from 50x60 pixels to full screen.
- Virtually unlimited objects detected per frame.
- Detailed people model provides rich metadata and allows detection of direction, trajectory, pose and gesture.
- Advanced software running on accompanying application processor allows for higher-level behaviour to be determined, including sophisticated inter-frame tracking.
- Additional software libraries enable higher-level, on-device features, such as face recognition.



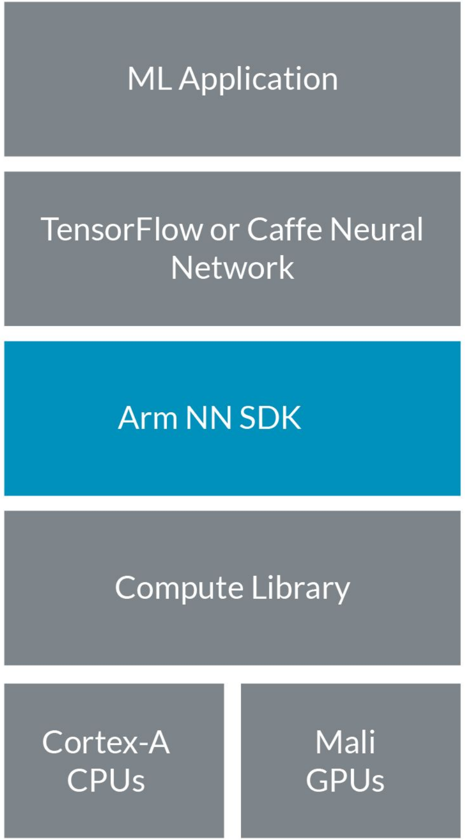
https://developer.arm.com/products/processors/machine-learning/arm-od-processor

Arm NN

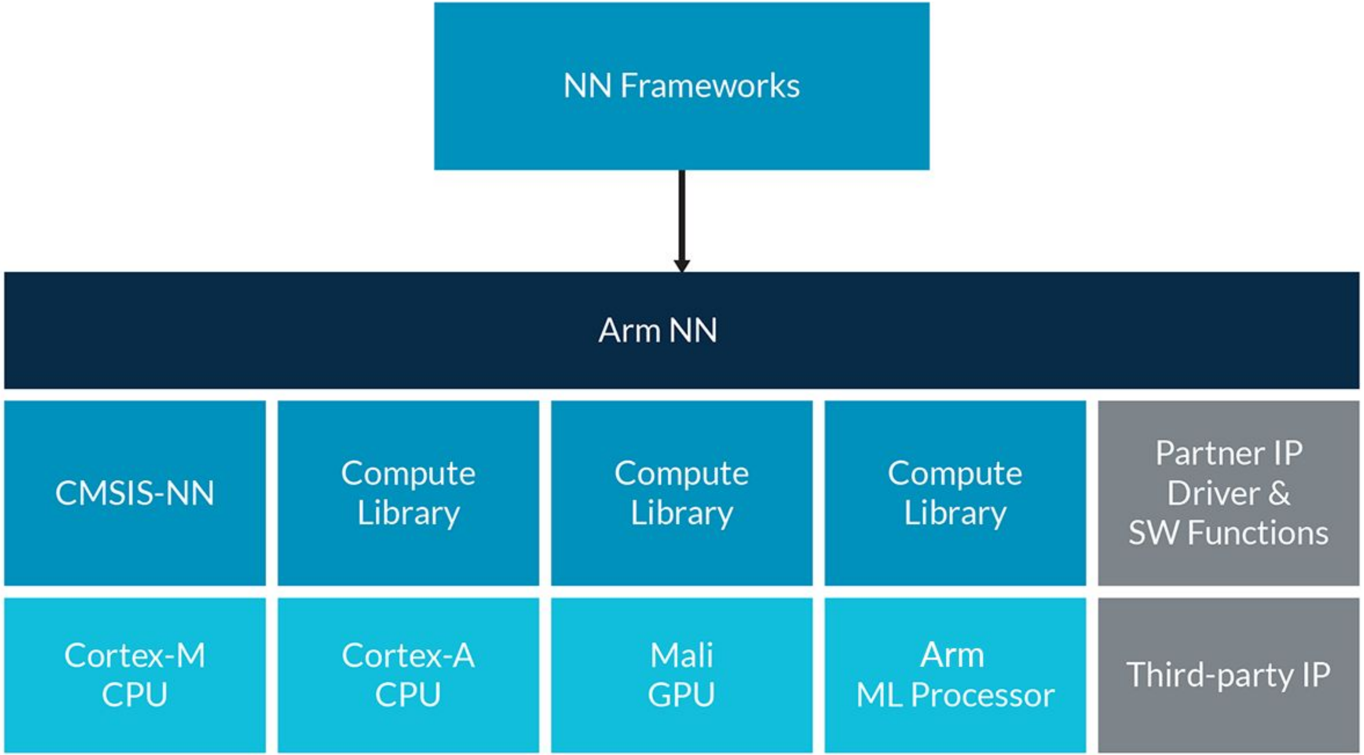
Arm NN SDK is a set of open-source Linux software and tools that enables machine learning workloads on power-efficient devices. It provides a bridge between existing neural network frameworks and power-efficient Arm Cortex CPUs, Arm Mali GPUs or the Arm Machine Learning processor.

Arm NN SDK utilizes the Compute Library to target programmable cores, such as Cortex-A CPUs and Mali GPUs, as efficiently as possible. It includes support for the Arm Machine Learning processor and, via CMSIS-NN, support for Cortex-M CPUs.

https://developer.arm.com/products/processors/machine-learning/arm-nn

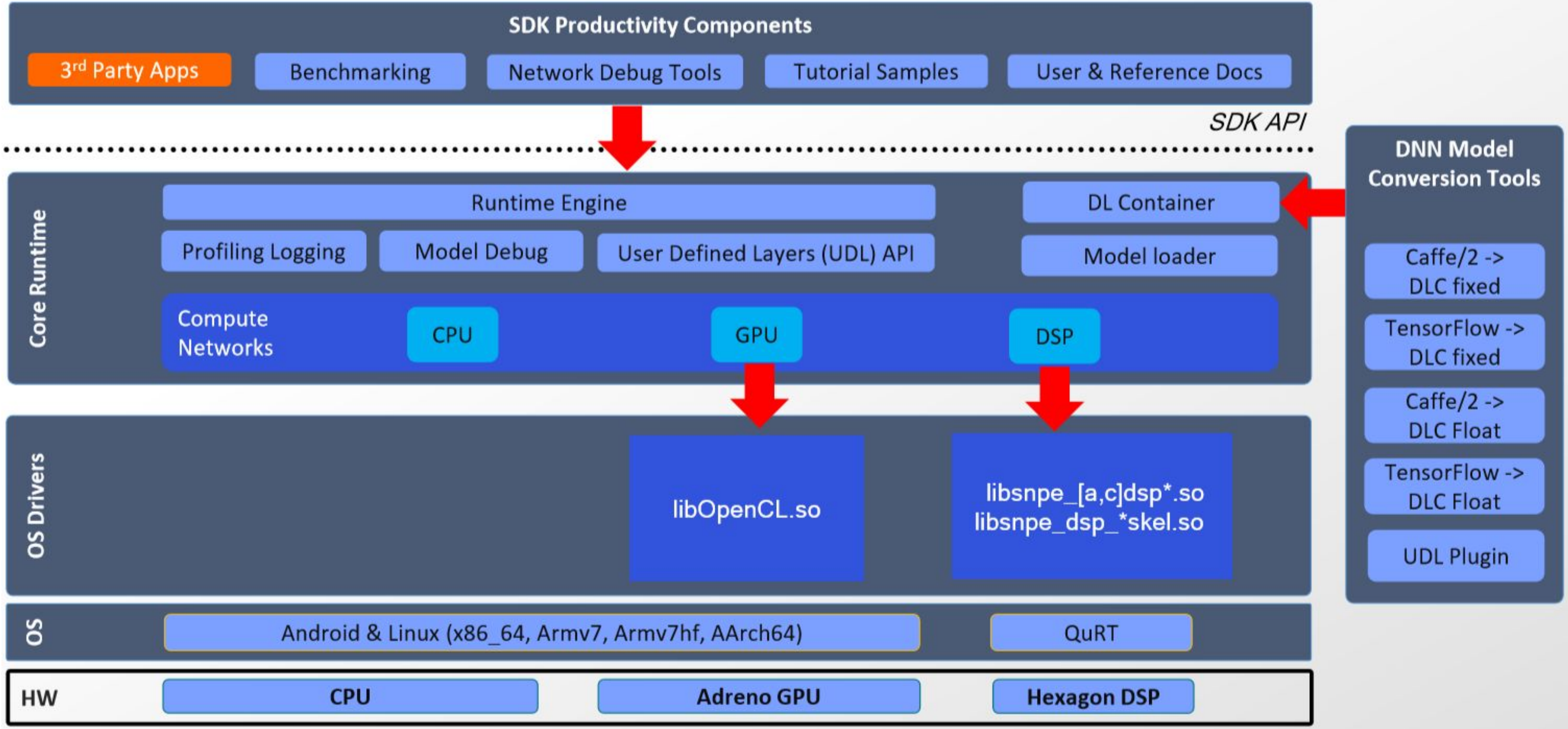


Arm NN



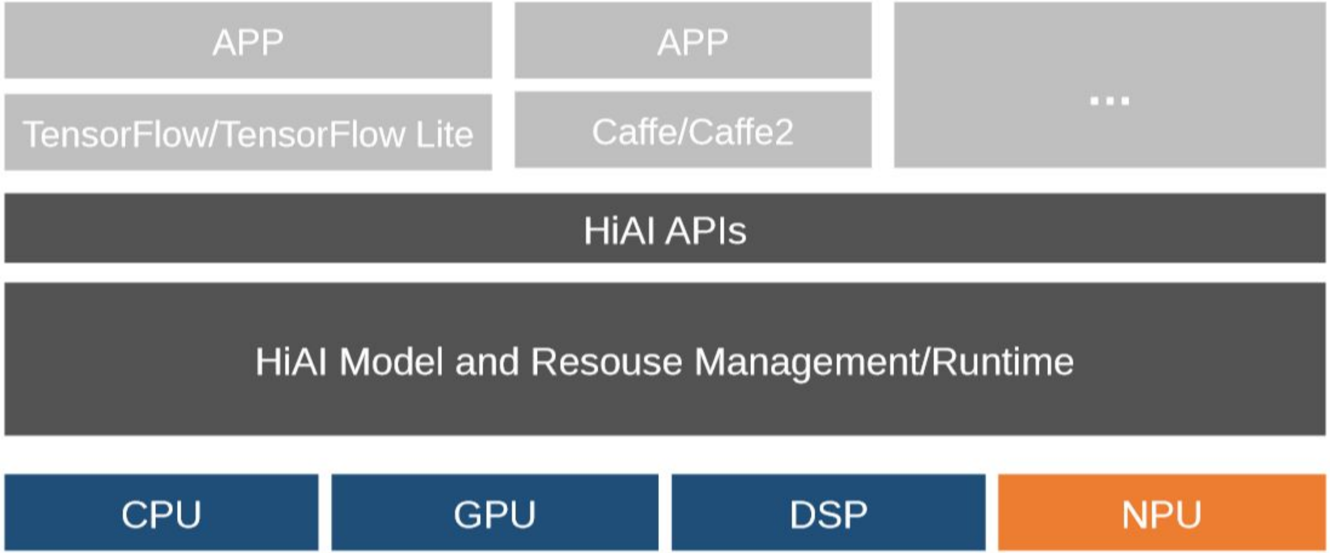
https://developer.arm.com/products/processors/machine-learning/arm-nn

Snapdragon NPE SW Diagram



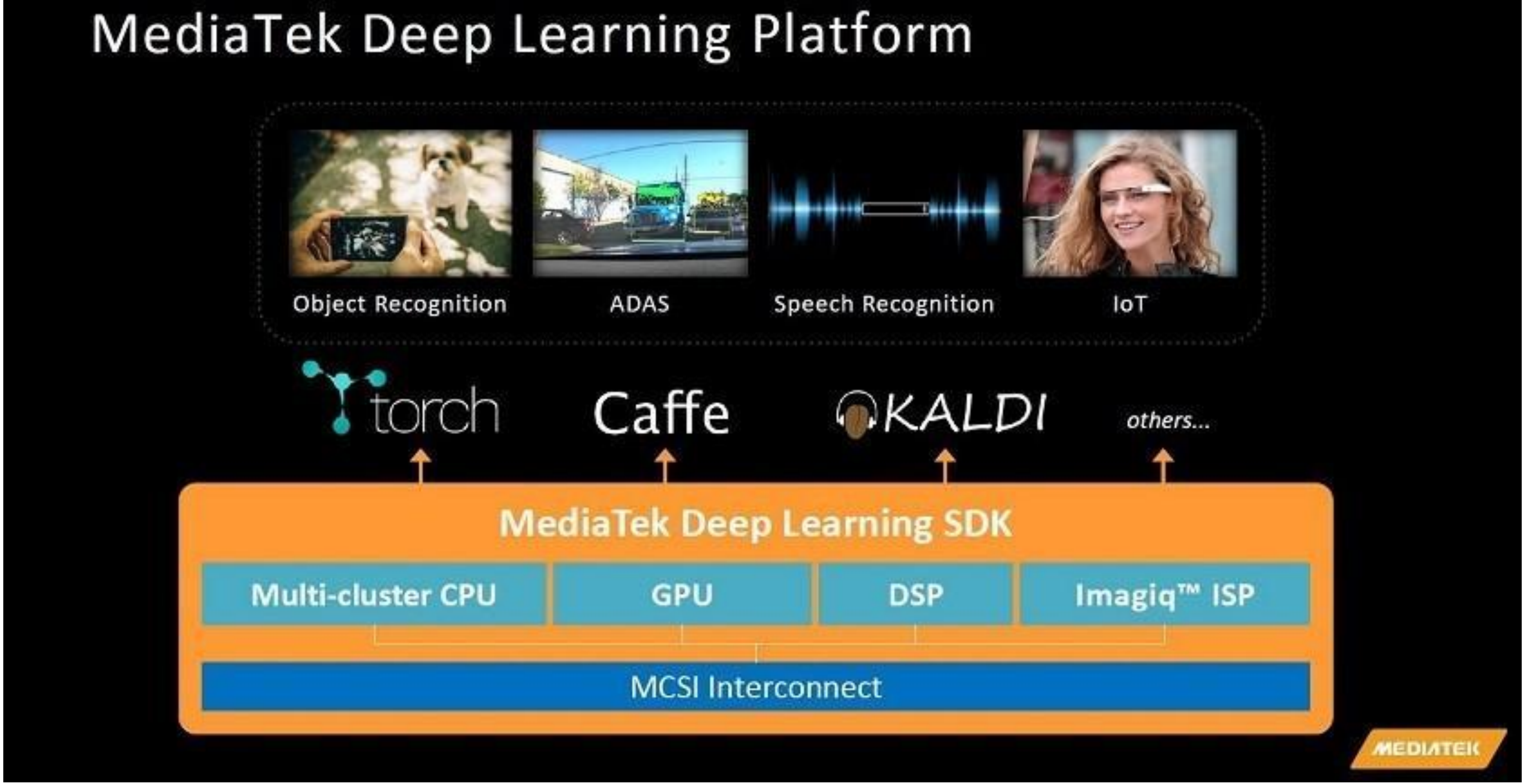
https://connect.linaro.org/resources/hkg18/hkg18-306/

HISILICON


APP
APP
TensorFlow/TensorFlow Lite
Caffe/Caffe2
...
HiAI APIs
HiAI Model and Resouse Management/Runtime
CPU
GPU
DSP
NPU

- 99 operators
- Caffe, TensorFlow, TensorFlow Lite, Huawei HiAI SDK, Android NN
- Converter tools from AI models to serialized offline model

https://connect.linaro.org/resources/hkg18/hkg18-302/



MediaTek Deep Learning Platform

Object Recognition

ADAS

Speech Recognition

IoT

torch

Caffe

KALDI

others...

MediaTek Deep Learning SDK

Multi-cluster CPU

GPU

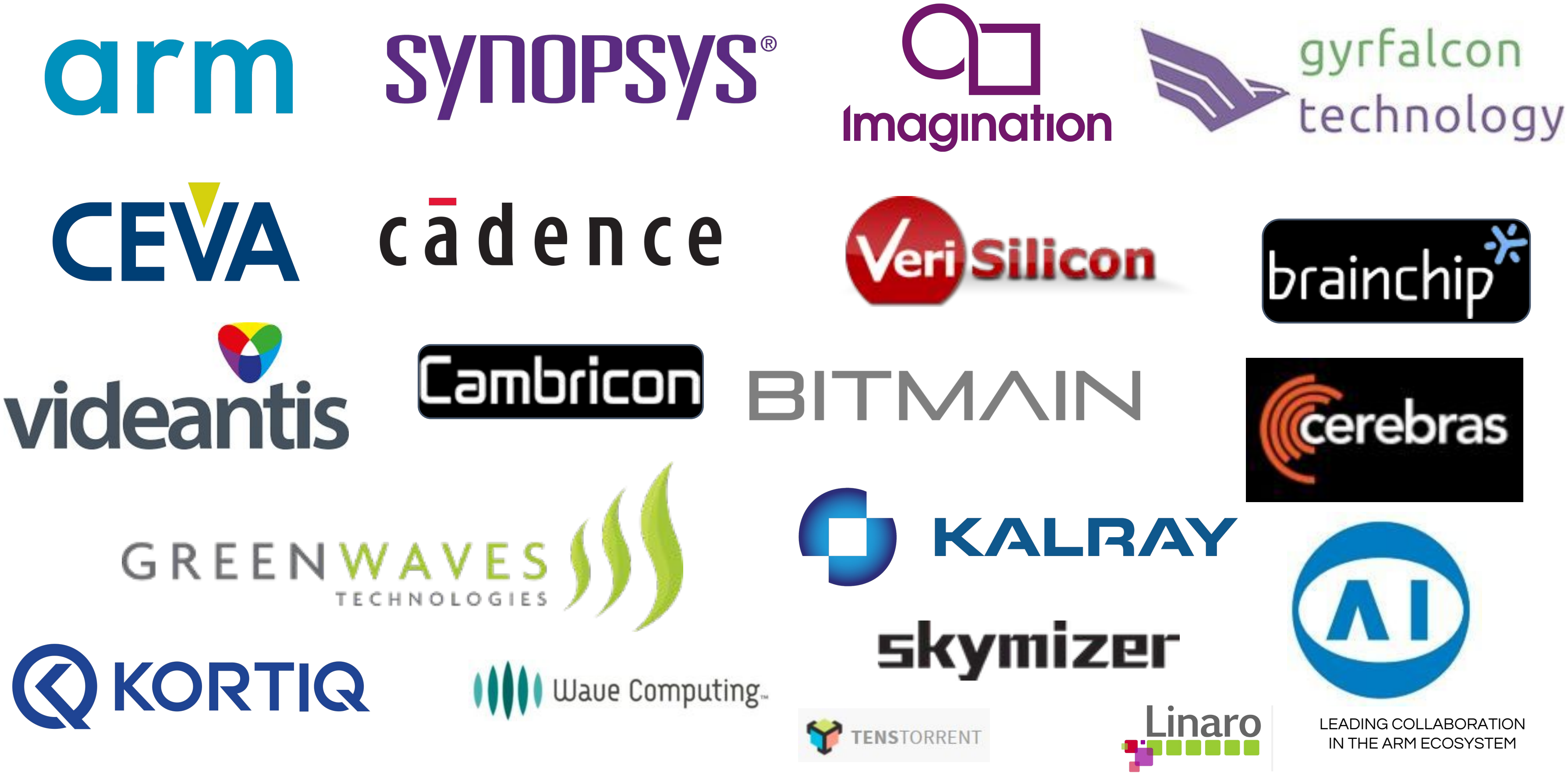
DSP

Imagiq™ ISP

MCSI Interconnect

https://www.forbes.com/sites/tiriasresearch/2017/03/31/mediatek-brings-neural-networks-to-devices/#6468bd5f3eac

An ecosystem of 3rd parties providing NN IP and tools



Observations

- Complete offload vs heterogenous computing
- Shared memory vs sub-system memories and DMA
- Fixed operators and software fallback
- Graph split vs cost of context switch
- Serialized models and converter tools

- Forked and accelerated inference engine for each NN IP and each framework
  → high total cost of ownership
  → delayed rebases and updates
  → delayed security fixes

Call to Action

Linaro Collaboration

Members fund Linaro and drive work through engineering steering committees

Member and Linaro engineers collaborate to develop work once, for all

Linaro delivers output to members, into open source projects, and into the community

Now ~25 members, up from 6 in 2010

Over 300 OSS engineers globally, including 140 Linaro staff

Core Members

arm

HISILICON

QUALCOMM INNOVATION CENTER, INC. QuIC

Club Members

Google

socionext

ST life.augmented

Texas Instruments

UNISOC

ZTE

Group Members

CAVIUM

CISCO

COMCAST

CYPRESS EMBEDDED IN TOMORROW

ENEA

FUJITSU

HXT 华芯通半导体

KYLIN 银河麒麟

NOKIA

NXP

QUALCOMM QUALCOMM DATACENTER TECHNOLOGIES, INC.

redhat

SAMSUNG

XILINX ALL PROGRAMMABLE

Community Members

IBM

THE LINUX FOUNDATION

Linaro works Upstream

Delivering high value collaboration

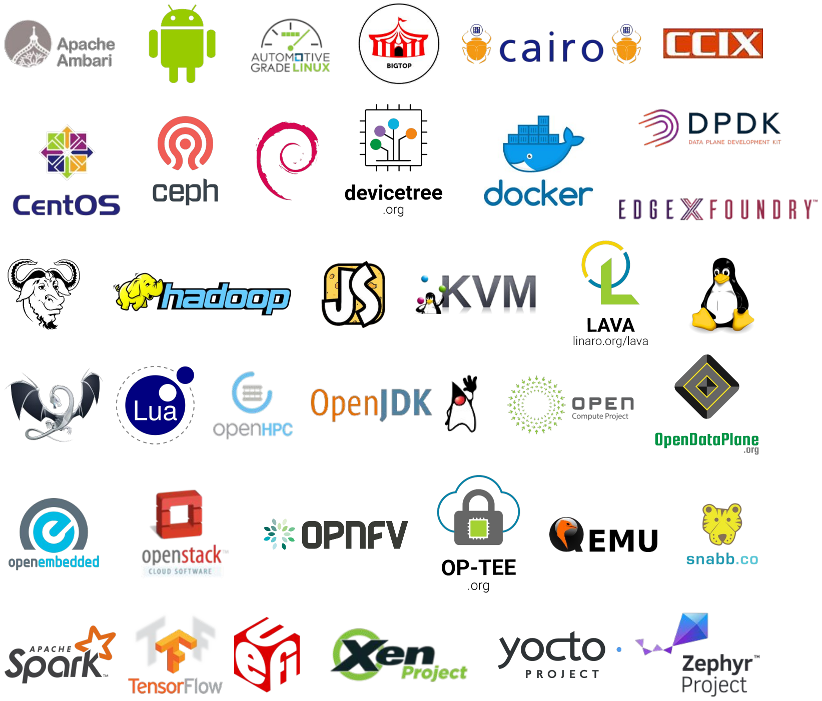
Top 5 company contributor to Linux and Zephyr kernels

Contributor to >70 open source projects; many maintained by Linaro engineers

	Company	4.8-4.13 Changesets	%
1	Intel	10,833	13.1%
2	Red Hat	5,965	7.2%
3	Linaro	4,636	5.6%

Source: 2017 Linux Kernel Development Report, Linux Foundation

Selected projects Linaro contributes to



ONNX

Open Neural Network Exchange (ONNX) provides an open source format for AI models. It defines an extensible computation graph model, as well as definitions of built-in operators and standard data types. Initial focus on inference.

Apache

amazon web services™

Microsoft

CNTK

mxnet

Caffe2

torch

Linaro

LEADING COLLABORATION
IN THE ARM ECOSYSTEM

ONNX

Open Neural Network Exchange (ONNX)

An open source format for AI models

An extensible computation graph model

Definitions of built-in operators and standard data types

Initial focus on inference

Apache

amazon web services™

Microsoft

CNTK mxnet Caffe2 torch

Linaro LEADING COLLABORATION IN THE ARM ECOSYSTEM

ONNX

ONNX Interface for Framework Integration (ONNXIFI)

Standardized interface for neural network inference on special-purpose accelerators, CPUs, GPUs, DSPs, and FPGAs

Dynamic discovery of available backends and supported ONNX operators

Initialize and deinitialize backends

Specify memory locations and metadata

Run an ONNX graph

f

Apache

amazon web services™

Microsoft

CNTK

mxnet

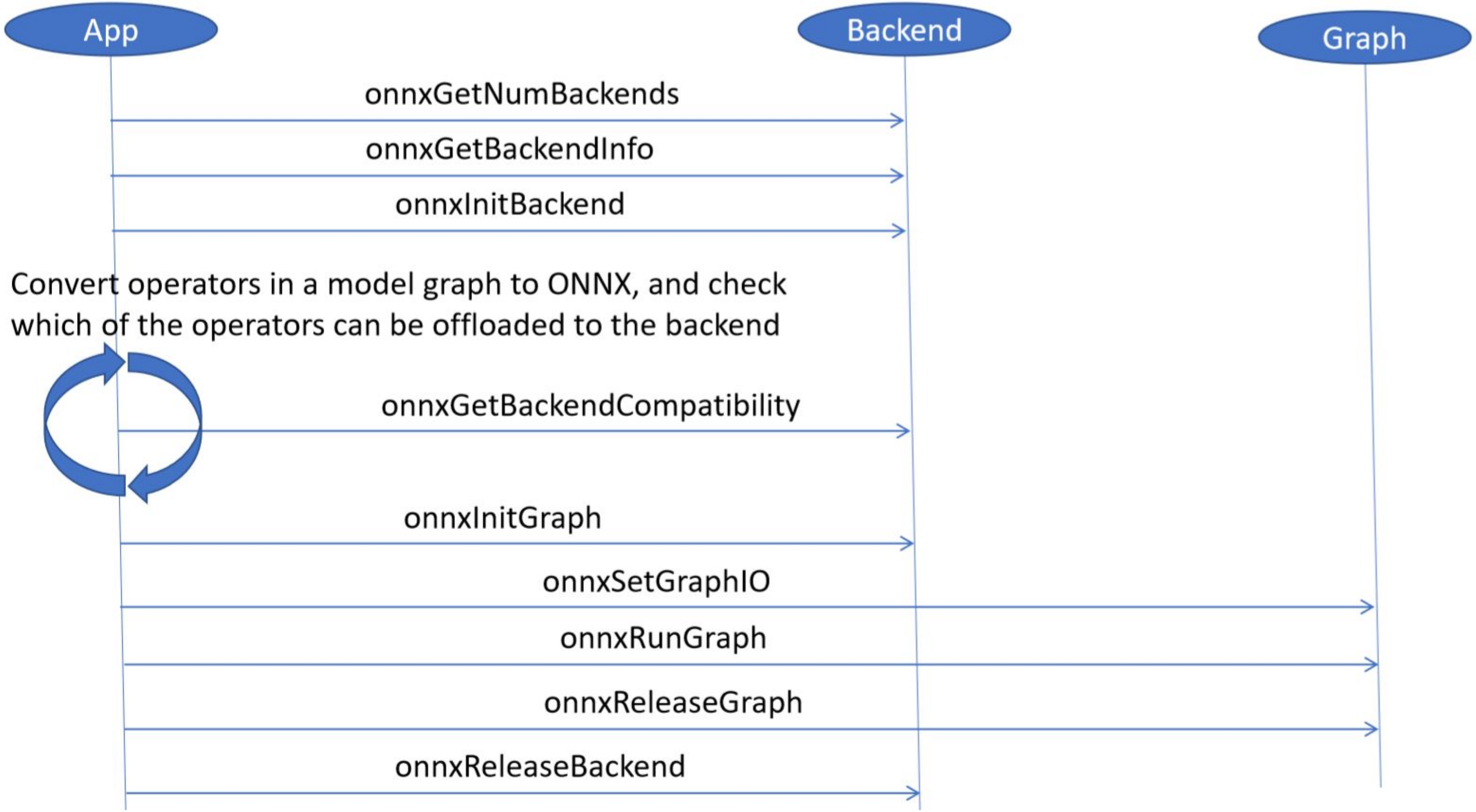
Caffe2

torch

Linaro

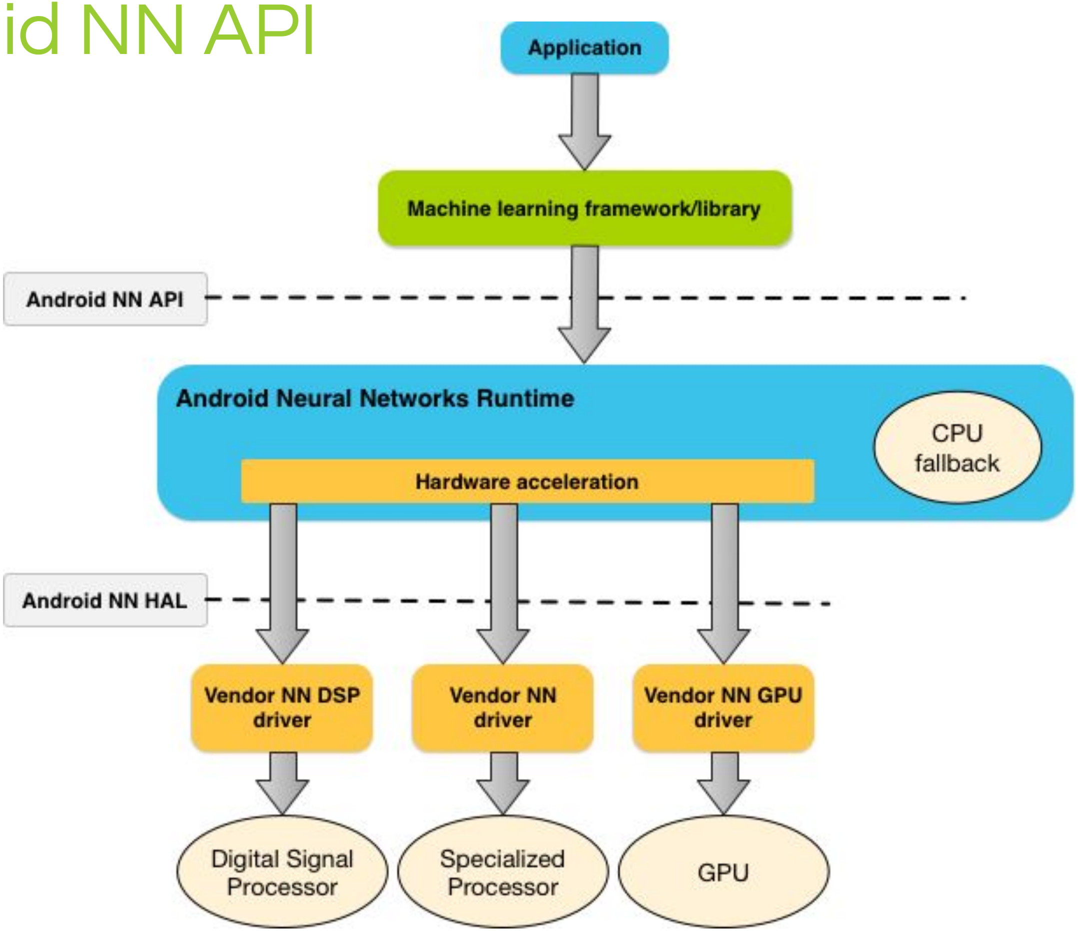
LEADING COLLABORATION
IN THE ARM ECOSYSTEM

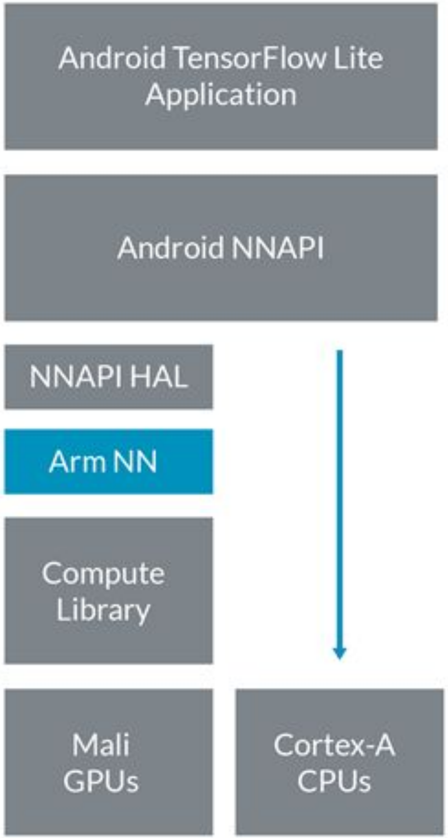
ONNXIFI API Call Flow



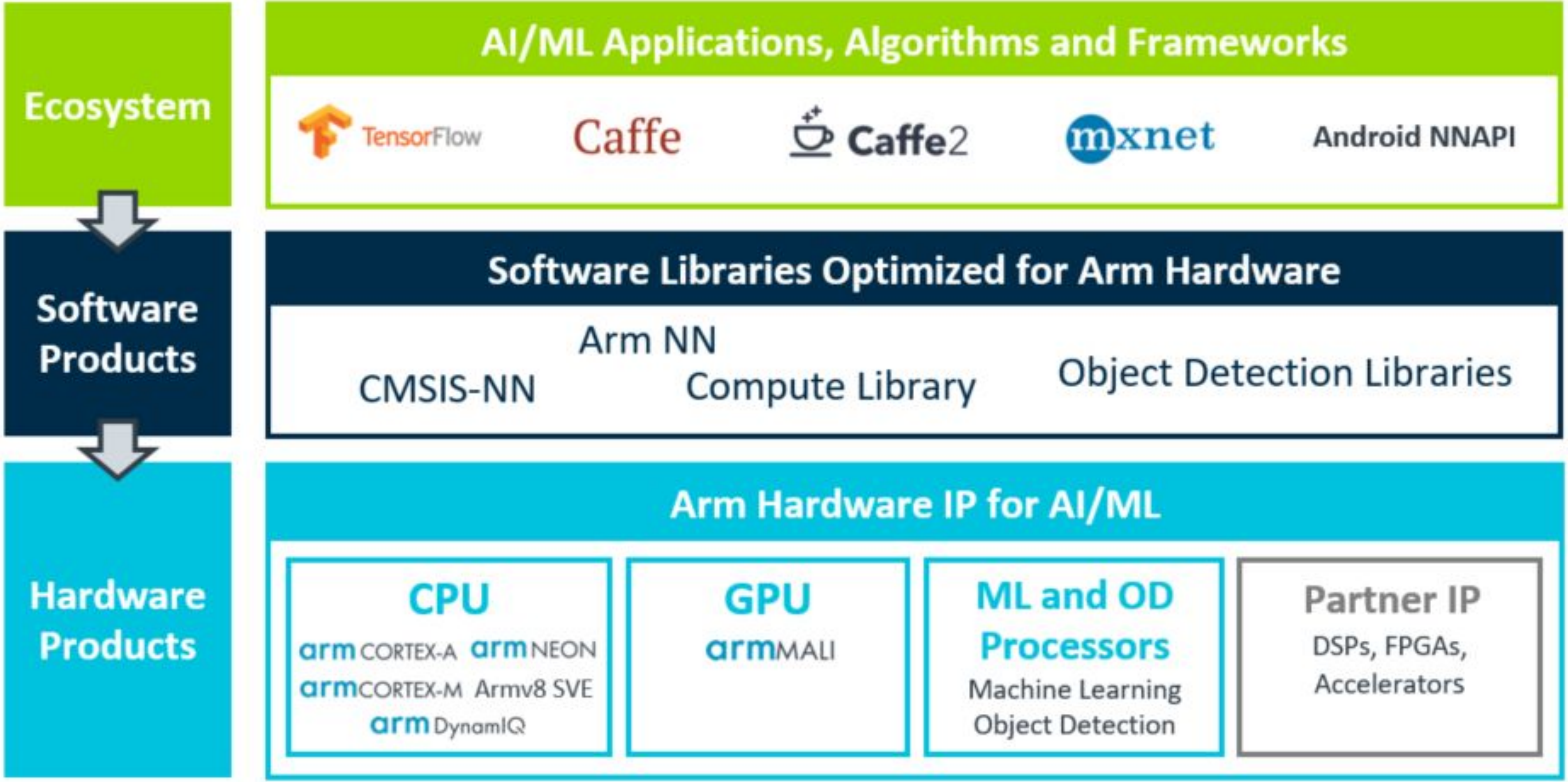
Convert operators in a model graph to ONNX, and check which of the operators can be offloaded to the backend

Android NN API





https://developer.android.com/ndk/guides/neuralnetworks/



Ecosystem

AI/ML Applications, Algorithms and Frameworks

TensorFlow Caffe Caffe2 mxnet Android NNAPI

Software Products

Software Libraries Optimized for Arm Hardware

Arm NN

CMSIS-NN Compute Library Object Detection Libraries

Hardware Products

Arm Hardware IP for AI/ML

CPU
arm CORTEX-A arm NEON
arm CORTEX-M Armv8 SVE
arm DynamIQ

GPU
arm MALI

ML and OD Processors
Machine Learning
Object Detection

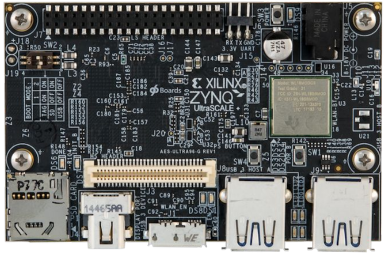
Partner IP
DSPs, FPGAs,
Accelerators

https://developer.arm.com/products/processors/machine-learning/arm-nn

Areas of Collaboration

- Common model description format and APIs to the back end
- Common optimized runtime inference engine for Arm-based SoC
- Dynamic plug-in framework to support multiple 3rd party NPU, CPU, GPU, DSP
- CI loops on reference development boards to measure accuracy, performance speed up and regression testing

LAVA



Linaro
connect
Hong Kong 2018

Discussions started last March

AI/ML Resources from HKG18

HKG18-417 - OpenCL support by NNVM & TVM

HKG18-413 - AI and Machine Learning BoF

HKG18-405 - Accelerating Neural Networks with...

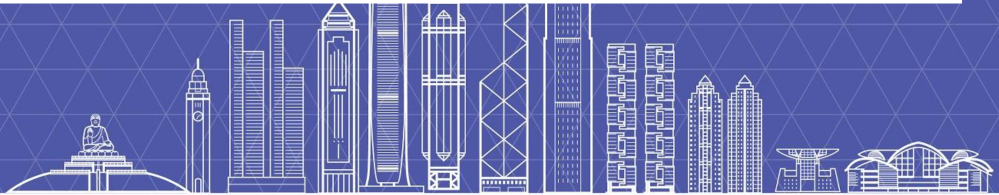
HKG18-312 - CMSIS-NN

HKG18-306 - Overview of Qualcomm SNPE

HKG18-304 - Scalable AI server

HKG18-302 - Huawei HiAI : Unlock The Future

HKG18-200K2 - Keynote: Accelerating AI from Cloud to Edge



Join us at the

AI and Neural Networks on Arm Summit

At Linaro Connect Vancouver 2018

Wednesday 19 September - Hyatt Regency Vancouver,
655 Burrard Street, V6C 2R7

$45 to attend the summit only

REGISTER HERE

https://connect.linaro.org/ai-neural-networks-arm-summit/

Linaro