SFO17-509

Deep Learning on ARM Platforms

- from the platform angle

Jammy Zhou - Linaro

Linaro
connect
San Francisco 2017

ENGINEERS
AND DEVICES
WORKING
TOGETHER

Agenda

- Deep learning basics
- Platform overview
- Gaps and challenges

Basic Elements

Algorithms & Models

CNN, R-CNN, RNN, LSTM, Feed Forward, BP, SGD, Deep Reinforcement learning, etc

AlexNet, VGG, GoogLeNet, ResNet, SqueezeNet, SyntaxNet, MobileNet, SSD, YOLO, etc

+

Computing Platforms

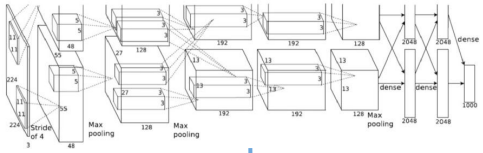
CPU, GPU, FPGA, TPU, etc

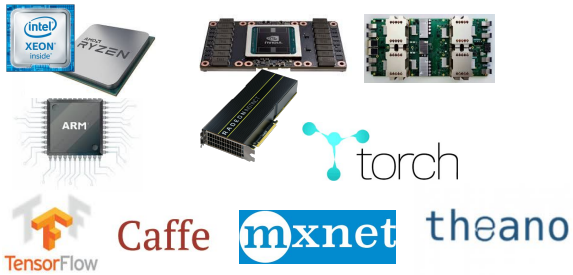
Deep learning software stack

+

Datasets

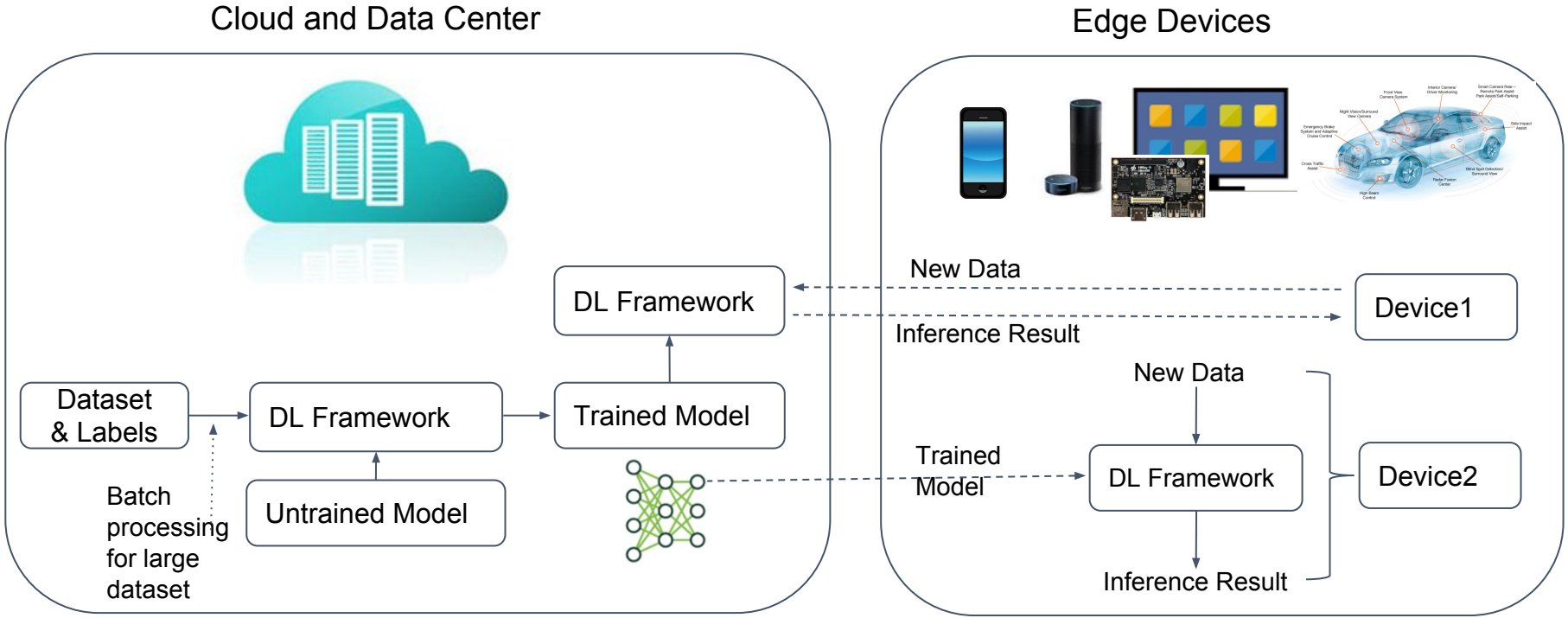
MNIST, ImageNet, CIFAR10, CIFAR100, AudioSet, DET, CLS-LOC, ActivityNet, etc





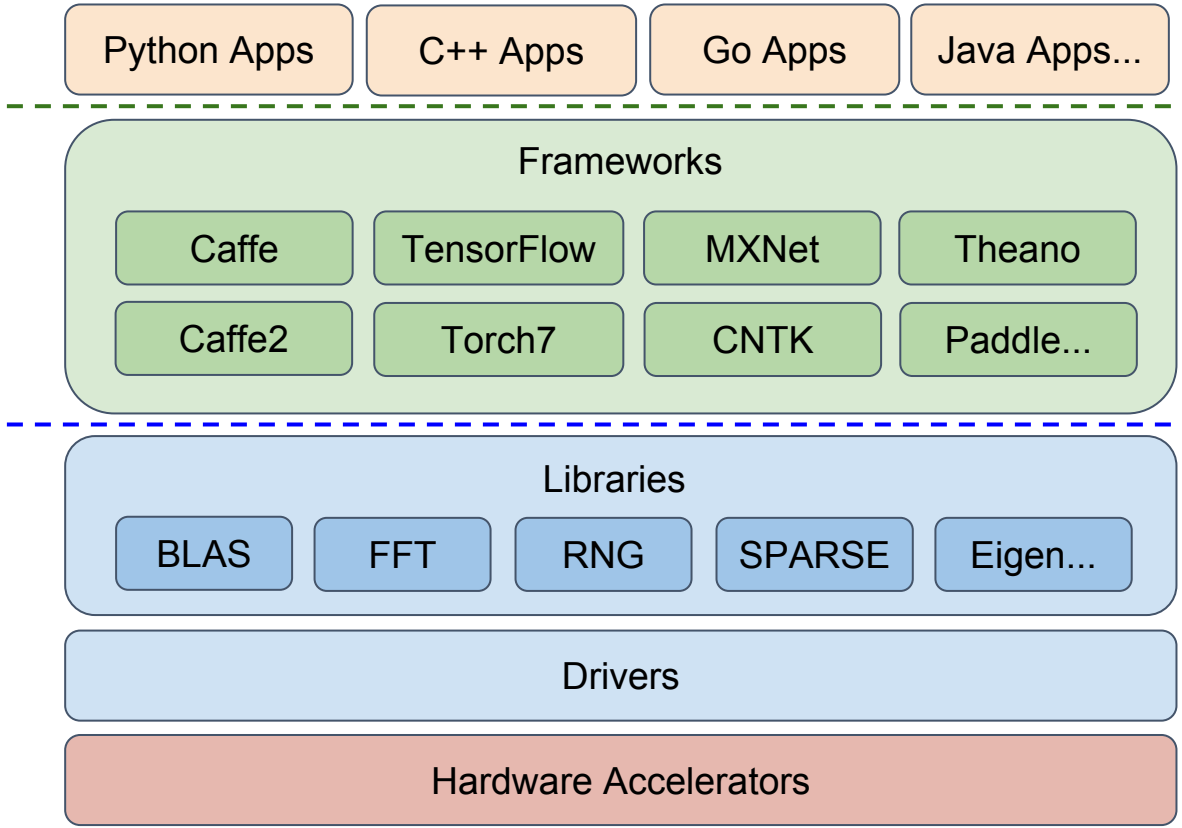
Deep Learning

Training & Inference



- **Training** on cloud & data center (desktop workstation can be used as well for development)
- **Inference** on cloud & data center or edge device

Platform Overview



Heterogeneous Accelerators

GPU

- Nvidia
  - Tesla, Quadro, GeForce, and etc
  - Drive PX for self driving
  - Jetson TK1/TX1/TX2 for embedded
- AMD
  - Radeon Instinct, Radeon Pro
  - Embedded G series
- Mobile GPUs
  - ARM Mali
  - Qualcomm Adreno
  - Imagination PowerVR
  - VeriSilicon Vivante

ASIC

- Google TPU
- Intel Nervana
- Nvidia Xavier DLA
- Horizon Robotics (地平线) BPU
- Cambricon (寒武纪) NPU
- Fujitsu DLU
- GraphCore IPU

FPGA

- Intel/Altera Stratix, Arria, Cyclone
- Xilinx Ultrascale, Ultrascale+
- Deephi Tech (深鉴科技) DPU (Xilinx based)
- Baidu XPU (Xilinx based)

DSP

- Qualcomm Hexagon
- CEVA XM
- TI C66x
- Cadence Tensilica Vision

Interface & Libraries

- Nvidia CUDA
  - Proprietary (including the driver stack)
  - Widely supported by the DL frameworks and the industry
- AMD HIP
  - Fully open source together with the ROCm driver stack
  - Support both NVCC (for CUDA) and HCC compiler backends
- Khronos Standards
  - OpenCL
    - Support not only GPU, but also FPGA, DSP and other accelerators
  - OpenVX with Neural Network Extension
    - Acceleration for image processing and computer vision applications
  - Vulkan
    - Designed for both graphics and compute, mainly used for graphics now
- Nvidia
  - cuDNN, cuBLAS, cuSPARSE, etc
- AMD
  - MIOpen, hipBLAS/rocBLAS, hcRNG, hcFFT/rocFFT, etc
- OpenCL
  - AMD clBLAS, clFFT, clRNG, clSPARSE
  - ARM Compute Library, requires:
    - cl_arm_non_uniform_work_group_size
    - cl_khr_fp16
  - Qualcomm Neural Processing Engine
- Others
  - Eigen - C++ template library for LA
  - OpenBLAS, NNPack, MKL, etc
  - OpenCV - Open source CV library
    - Have Halide based DNN module
  - Xilinx DNN and GEMM libraries

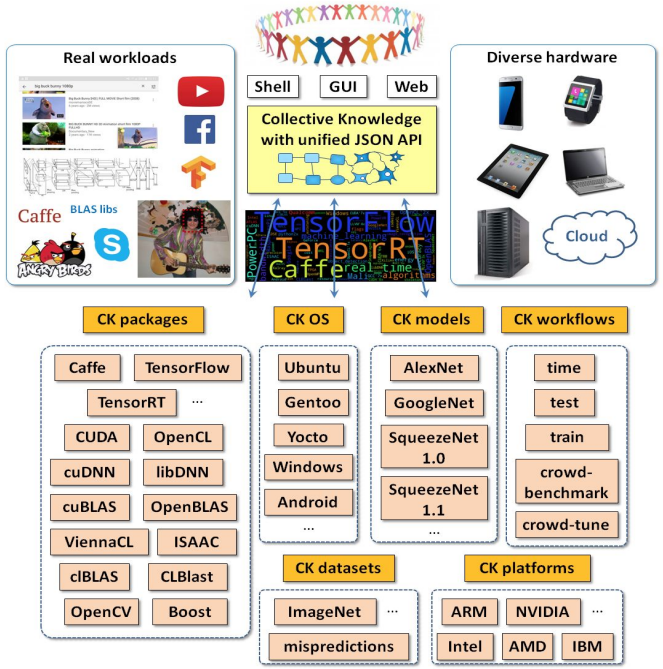
Frameworks

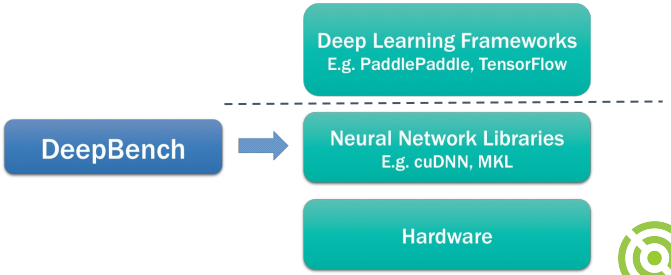
	Vendor	Languages	CUDA	HIP	OpenCL	ACL	NPE[1]
Caffe	BVLC	C++, Python, Matlab	Yes	Yes [2]	Yes [2]	Yes [2]	Yes
Caffe2	Facebook	C++, Python	Yes	Yes [2]	No	No	Yes
TensorFlow	Google	Python, C++, JAVA, Go, etc	Yes	No [3]	Yes [2]	No	Yes
MXNet	Apache	Python, R, C++, Perl, etc	Yes	Yes [2]	No	Yes [2]	No
Torch	Community	Lua, Python, Matlab	Yes	Yes [2]	Yes [2]	No	No
Theano	Montreal	Python	Yes	No	Yes [4]	No	No
CNTK	Microsoft	Python, C++, C#, .Net, etc	Yes	No [3]	No	No	No
Paddle	Baidu	Python, Go	Yes	No	No	No	No

[1] The trained models are converted to DLC format for execution
[2] Out-of-tree support, by vendors and community
[3] Under development
[4] Seems incomplete and buggy

Benchmark and Testing

- Collective Knowledge SDK for AI
  - Automate benchmarking, optimization, co-design the whole deep learning stack to satisfy different requirements (e.g, execution time, accuracy, power consumption, memory usage, etc)
  - Have portable and customizable workflow framework
  - CK modules with open & unified JSON APIs are used to abstract access to changing SW & HW
  - Image classification and compiler flag prediction are supported now
- DeepBench from Baidu
  - Benchmark basic low-level operations on different hardwares for deep learning
- Convnet-benchmarks

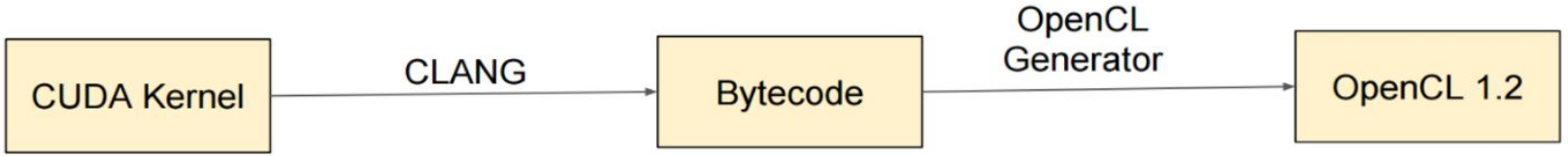


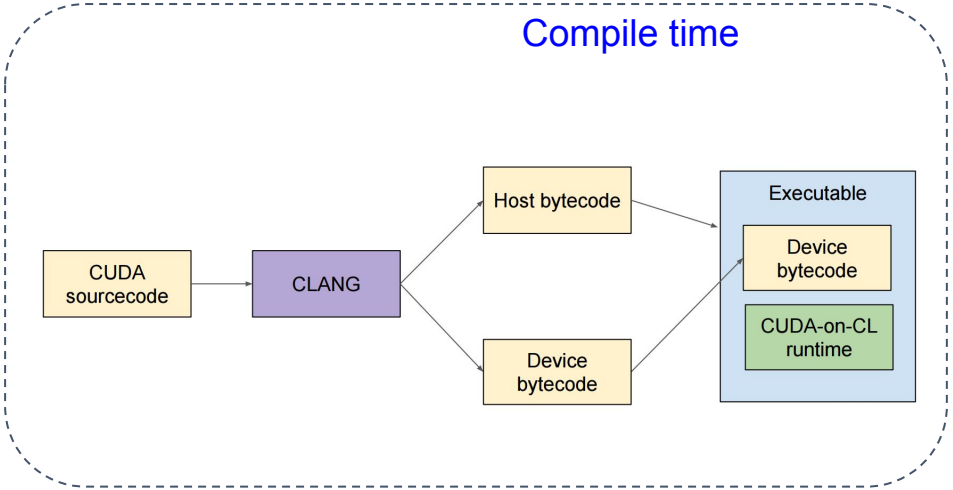


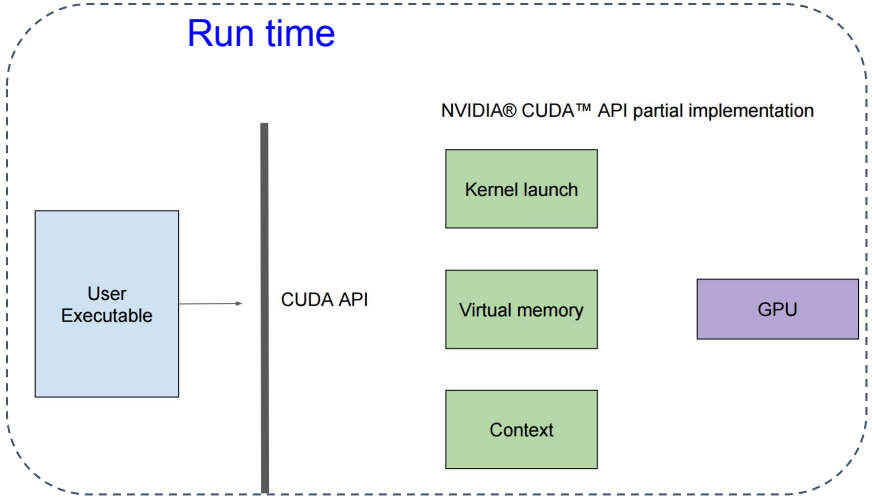
Gaps and Challenges - Arch Agnostic

- Vendor specific APIs are used for heterogeneous accelerators
  - Nvidia CUDA, AMD HIP, Google TPU StreamExecutor, and etc
  - Can OpenCL or something else be a competitive solution to rule the other world?
- OpenCL support is quited limited so far
  - There is no official OpenCL acceleration support by default for almost all major DL frameworks, although some of them have out-of-tree experimental support.
    - Can we get OpenCL support in upstream DL frameworks to reduce maintain effort?
    - Shall we enable and optimize OpenCL support for more DL frameworks?
  - Can Coriander be a solution to use existing CUDA backend of DL frameworks on OpenCL?
    - AMD HIP has similar mechanism to support CUDA on HIP
    - Coriander is still a research project with limited DL framework and HW support

Coriander - CUDA on OpenCL







http://www.iwocl.org/wp-content/uploads/iwocl2017-hugh-perkins-cuda-cl.pdf

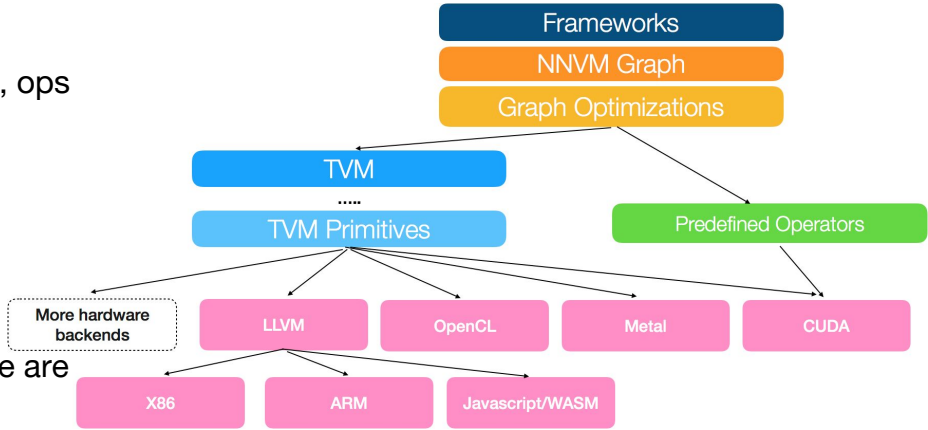
Gaps and Challenges - Arch Agnostic (Cont.)

- The DL frameworks are fragmented
  - It is not easy to add new accelerator support, and it has to be done for each framework
    - Unified graph IR and tensor IR support for DL frameworks will be helpful
      - See next page for several major graph IR implementations
  - There is no good interoperability between different DL frameworks
    - Operator and tensor sharing
      - DLPack proposal seems promising, adopted by MXNet, PyTorch, Caffe2, tiny-dnn
    - Model translations between different DL frameworks
  - Applications need unified APIs for portability and flexibility
    - Keras: a Python DL library supporting TensorFlow, CNTK, Theano and MXNet backends
    - TensorFlow Lite and Neural Network API support on Android
      - To support pre-trained models with other DL frameworks, translation will be required

Computation Graph IR & Tensor IR

- Google XLA
  - A domain-specific compiler, which makes it easier to add custom backends, LLVM backends can be reused
  - JIT to analyse TF graph, fuse operators at runtime
  - AOT to support TF models on memory restricted devices - helpful for ARM devices
- MXNet NNVM & TVM
  - NNVM-Fusion plugin for GPU kernel optimization (e.g, ops fusion, runtime compilation)
  - TinyFlow: TF Front-End + NNVM + Torch7 Back-End
  - TVM: a tensor IR/DSL stack
    - It is DLpack compatible, and reuses Halide IR
- Intel/Nervana NGraph
  - Neon is supported, integrations with TF XLA and Caffe are ongoing
- ONNX (Open Neural Network Exchange)
  - Initiated by Facebook and Microsoft
  - Caffe2, PyTorch and CNTK will support it

Who will become the 'LLVM' for DL?



Gaps and Challenges - ARM Specific

- There are seldom ARM boards with good PCIe capability in the community
  - AMD ROCm requires PCIe Gen3 x8 or x16 with PCIe Atomics
  - High-end FPGA cards requires PCIe Gen3 x16 or Gen4 x8 (e.g, Virtex Ultrascale+ from Xilinx)
- Discrete graphics driver support on ARM platforms is quite limited
  - Nvidia Proprietary Linux GPU driver only supports Aarch32
  - AMD ROCm has support for Cavium ThunderX, but no generic ARM support
  - GPU virtualization support is also needed by ARM data centers
- There is no good OpenCL driver distribution for mobile GPUs
  - OpenCL support for 96Boards needs improving
- Optimizations for edge devices
  - Memory footprint optimization for DL frameworks
    - e.g, quantization and low precision inference support being done by Google for TensorFlow
  - Power efficiency with different accelerators
- Anything else?

Linaro
connect
San Francisco 2017

ENGINEERS
AND DEVICES
WORKING
TOGETHER

Backup

96Boards with OpenCL support

96Boards	Accelerators	OpenCL version	Driver Public Availability
Qualcomm DB410c	Adreno 306, Hexagon QDSP6 V5	OpenCL 1.1 EP	Yes for Android?
Bubblegum-96(*)	PowerVR G6230	OpenCL 1.2 EP	Yes for Linux
Mediatek X20	Mali T880 MP4	OpenCL 1.2	TBD
Hisilicon Hikey960	Mali G71 MP8	OpenCL 2.0	Ongoing for Android
Hisilicon Poplar	Mali T720	OpenCL 1.2	TBD
Qualcomm SD600eval	Adreno 320, Hexagon QDSP6 V4	OpenCL 1.1	TBD
Socionext F-Cue	Mali T624	OpenCL 1.1	TBD
Mavell Andromeda	Vivante GC7000UL	OpenCL 1.2	TBD
MStar Kava	Mali T820	OpenCL 1.2	TBD
Altera Chameleon96	Cyclone V SoC FPGA	OpenCL 1.0 EP	TBD

* ACL cannot be used on Bubblegum-96 for missing required OpenCL extensions

Linaro connect
San Francisco 2017

Thank You

#SFO17

SFO17 keynotes and videos on: connect.linaro.org

For further information: www.linaro.org