LVC21-105

Machine Learning on Arm Servers – An update

Linaro Virtual Connect 2021

Ashok Bhat, Sr Product Manager, Arm

March 2021

Agenda

TensorFlow and PyTorch on Arm Servers for on-CPU inference

- What is the end goal?
- What is available today for end users?
- What is being worked on to improve usability and performance?
- How to get involved?

Not being covered

- Training use-case
- Machine learning using Arm + GPU
- Benchmarks and performance comparison

On-CPU Machine Learning (Inference)

Goal: Easy to use, best-in-class performance, ML inference solution on Arm servers using ML specific CPU features

Easy to use	Wide variety of inference workloads	Using Arm architecture features	On latest Arm based hardware
Container images and Python Packages	Image classification	INT8, Bfloat16, FP32	Arm Neoverse N1, V1 (Zeus), N2 (Perseus)
Popular ML frameworks support Arm as a first-class citizen	Object detection	Matrix Multiplier Extension	Fujitsu A64FX
	…	SVE/2	…
		…	

AArch64 packages and images

For machine learning users on Arm servers

Ready-to-use Python Packages

Goal: Readily available TensorFlow and PyTorch packages from standard repositories

Current status

- TensorFlow 1.15 and 2.4 package snapshots available
  - https://snapshots.linaro.org/ldcg/python/tensorflow/latest/
- PyTorch official builds available
  - https://download.pytorch.org/whl/torch_stable.html
  - https://download.pytorch.org/whl/nightly/cpu/torch_nightly.html

Next steps

- TensorFlow - Work with upstream to provide AArch64 packages

Docker recipes

Goal: Recipe to build your own Docker images

Current status

- TensorFlow
  - Versions – 1.15 and 2.3
    - Configurations – Eigen backend, oneDNN(ACL)
    - https://github.com/ARM-software/Tool-Solutions/tree/master/docker/tensorflow-aarch64
- PyTorch
  - Versions – 1.6
    - Configurations – OpenBLAS backend, oneDNN (ACL)
    - https://github.com/ARM-software/Tool-Solutions/tree/master/docker/pytorch-aarch64

Next steps

- Upgrade recipes to newer versions

Docker images

Goal: Readily available docker images on par with other architectures

Current status

- Images for Arm Neoverse N1 is available in a Linaro Docker Hub
  - TensorFlow 2.3 with Eigen, oneDNN (ACL)
    - https://hub.docker.com/r/linaro/tensorflow-arm-neoverse-n1
  - PyTorch 1.6 with OpenBLAS, oneDNN(ACL)
    - https://hub.docker.com/r/linaro/pytorch-arm-neoverse-n1

Next steps

- Upgrade images to newer versions
- Work with upstream to provide images in the standard repositories

arm

Best-in-class performance using ML-specific Arm features

For machine learning users on Arm servers

Key open source projects

TensorFlow

PyTorch

OpenBLAS

Eigen

oneDNN

Arm Compute Library

Key open source projects for ML on Servers

Frameworks

- ML Framework - TensorFlow
  - Popular open source ML framework
  - Has multiple backends on x86 – Eigen GEBP, oneDNN (with BLAS, direct kernels, JIT)
- ML Framework - PyTorch
  - Popular ML framework
  - Has multiple backends on x86 – NNPACK, FBGEMM, OpenBLAS, oneDNN (with BLAS, direct kernels, JIT)
  - Has multiple backends on Arm – QNNPACK, OpenBLAS

Key open source projects for ML on Servers

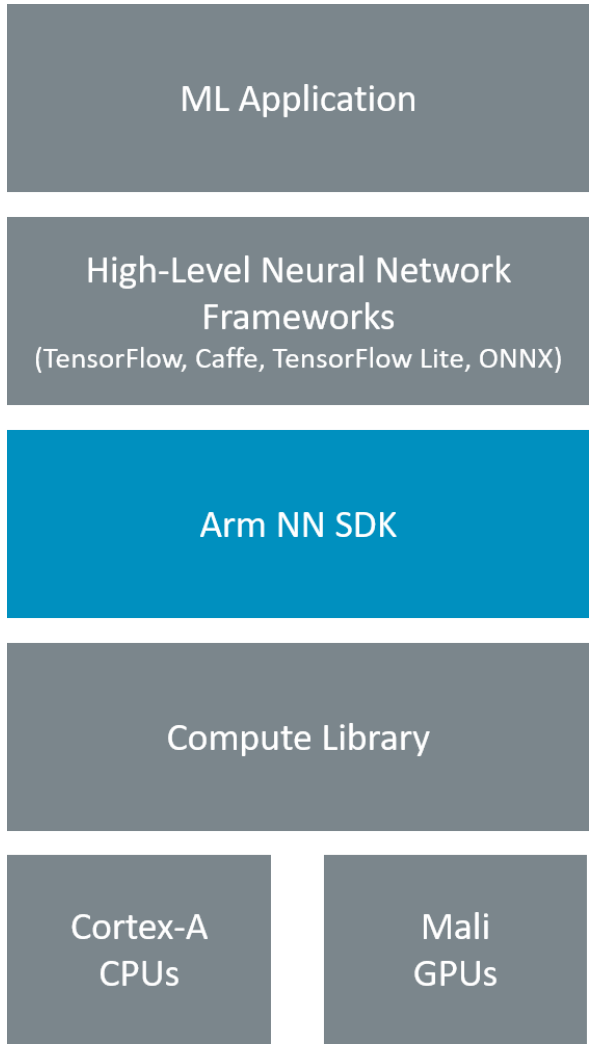
Libraries

- Library - Eigen
  - Eigen is a C++ template library for linear algebra: vectors, matrices, and related algorithms
  - TensorFlow heavily uses Eigen to represent internal data structures and their operations.
  - Eigen's GEBP kernel is used as a default CPU backend for FP32 contraction kernel
- Library - oneDNN
  - Intel's ML acceleration open-source library – Integrated with all major frameworks
  - Experimental support for AArch64
- Library - Arm Compute Library
  - Open source ML acceleration library for Arm used in edge/mobile use-cases
  - Contains high level operators
- Library - OpenBLAS
  - Most common open source BLAS backend

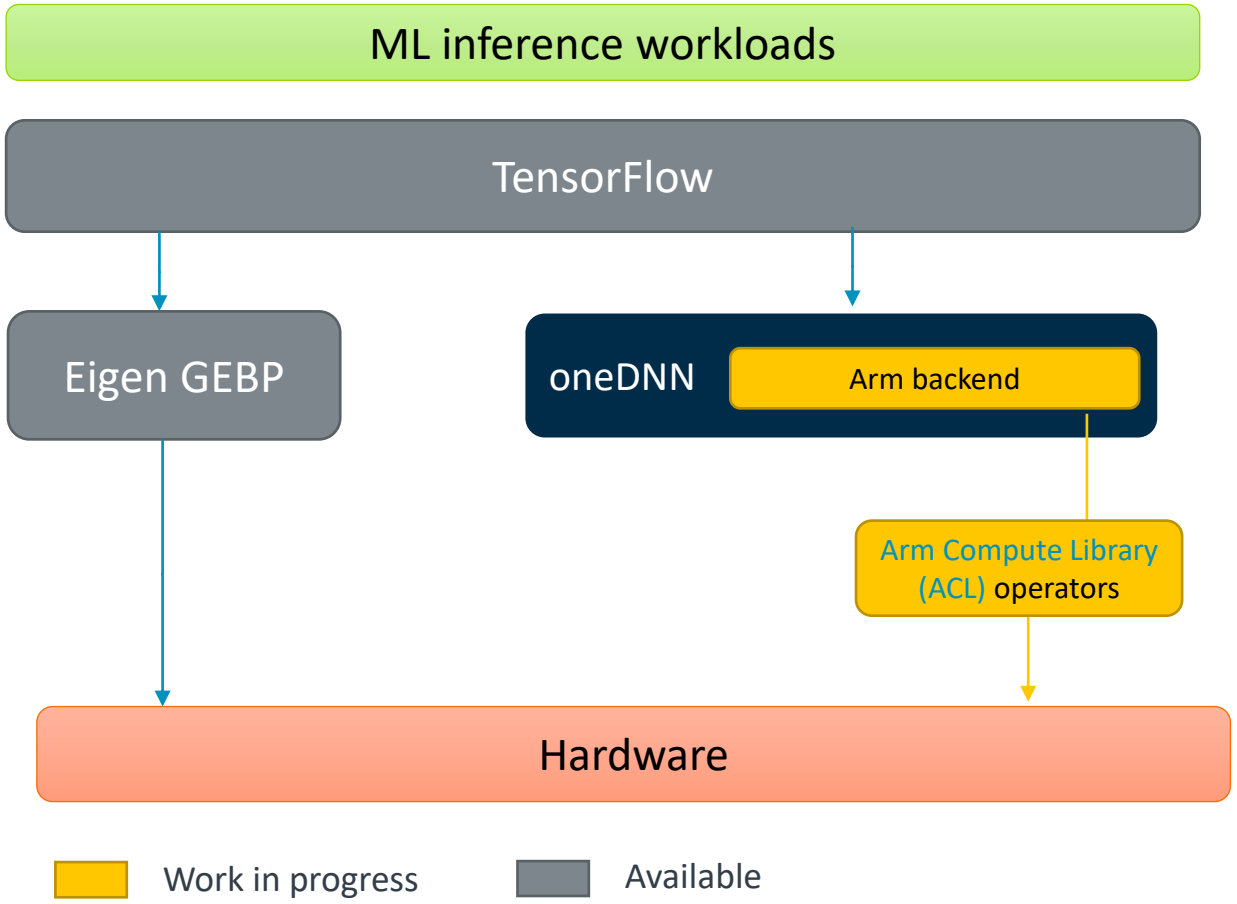
Arm Compute Library

A software library for computer vision and machine learning

- Collection of low-level functions
  - Optimized for Arm CPU and GPU architectures
  - Targeted at image processing, computer vision, and machine learning.
- Available free of charge under a permissive MIT open source license.
- Used to accelerate ArmNN (Arm's inference engine for CPUs, GPUs and NPUs)



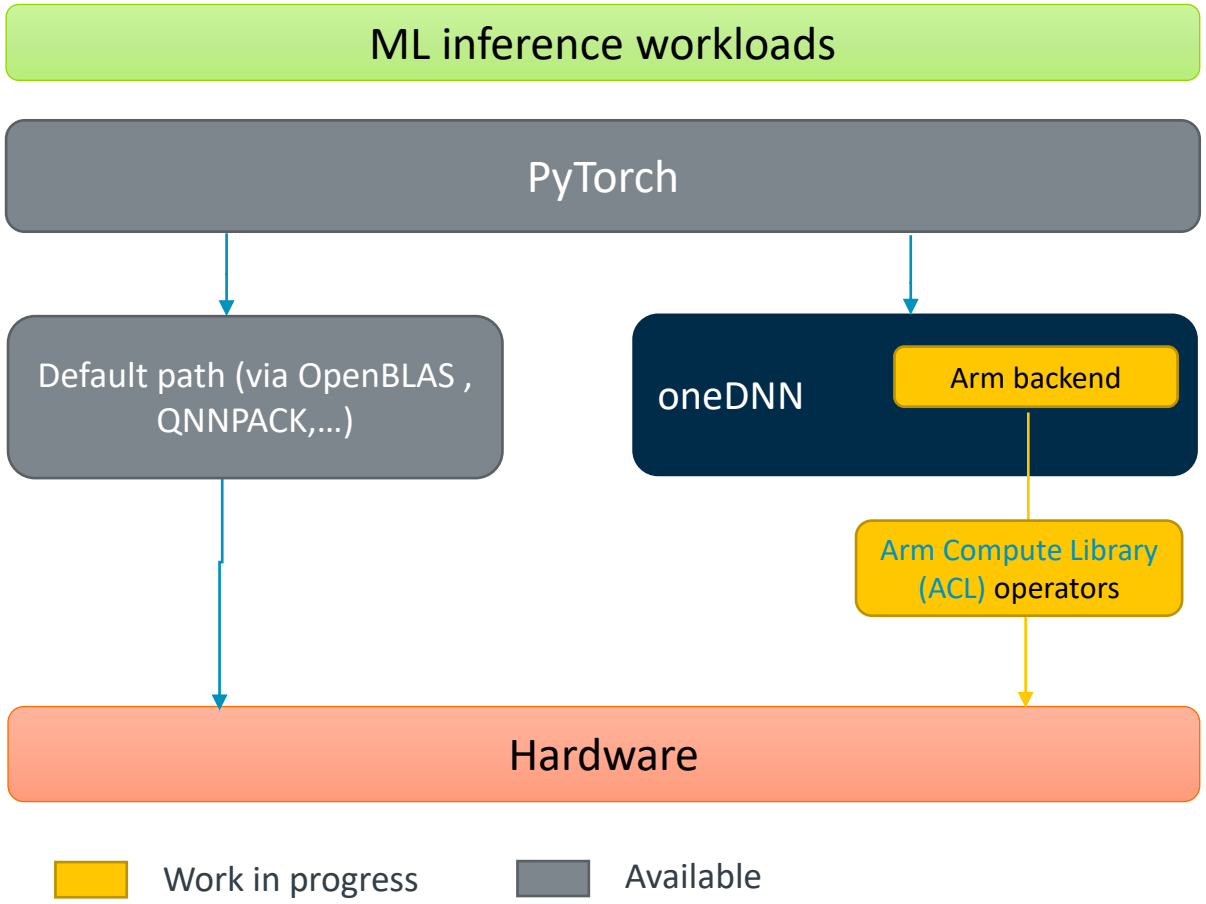
TensorFlow software stack on Arm – Status and Plan



Timeline	Library options
2.4 (Dec 20)	Eigen (FP32)
2.5* (Apr, 21)	Eigen (+SVE) oneDNN (ACL) (FP32)
2.6* (Q2, 21)	Eigen (FP32, Neon/SVE) oneDNN (ACL) (FP32)
2.7* (Q3, 21)	Eigen (FP32, Neon/SVE) oneDNN (ACL) (+BF16)

* Future release information (version and date) is Arm's estimate based on previous releases.

PyTorch software stack on Arm – Status and Plan



Timeline	Library options
1.7 (Oct 20)	OpenBLAS (FP32)
1.8 (Mar, 21)	OpenBLAS (FP32)
1.9* (Q2, CY21)	OpenBLAS (FP32) oneDNN (ACL) (FP32)
Future	oneDNN (ACL)(+BF16)

* Future release information (version and date) is Arm's estimate based on previous releases.

Data type support

Status and plan

Data type	TensorFlow Default	TensorFlow ACL (via oneDNN)	PyTorch Default	PyTorch ACL (via oneDNN)
FP32 type	Yes (Eigen)	Q1 CY21	Yes (OpenBLAS)	Q2 CY21
INT8 type	Yes (TFLite)	Q2 CY21	Yes (QNNPACK)	
BF16 type		*Q3 CY21		*Q3 CY21

* Usage of BF16 for accelerate FP32 models

Not planned

Wrap Up

Get involved in Machine Learning on Arm

Try	Try the Docker recipes/images to run TensorFlow and PyTorch on AArch64
Learn	Learn about the Arm Compute Library
Provide Feedback	Provide feedback on performance for your applications on Arm machines
Get involved	Get involved in the open-source development of ML inference on Arm Weekly public meeting to get involved at: https://bit.ly/arm-server-ml

arm

Thank You
Danke
Gracias
谢谢
ありがとう
Asante
Merci
감사합니다
धन्यवाद
Kiitos
شكرًا
ধন্যবাদ
תודה