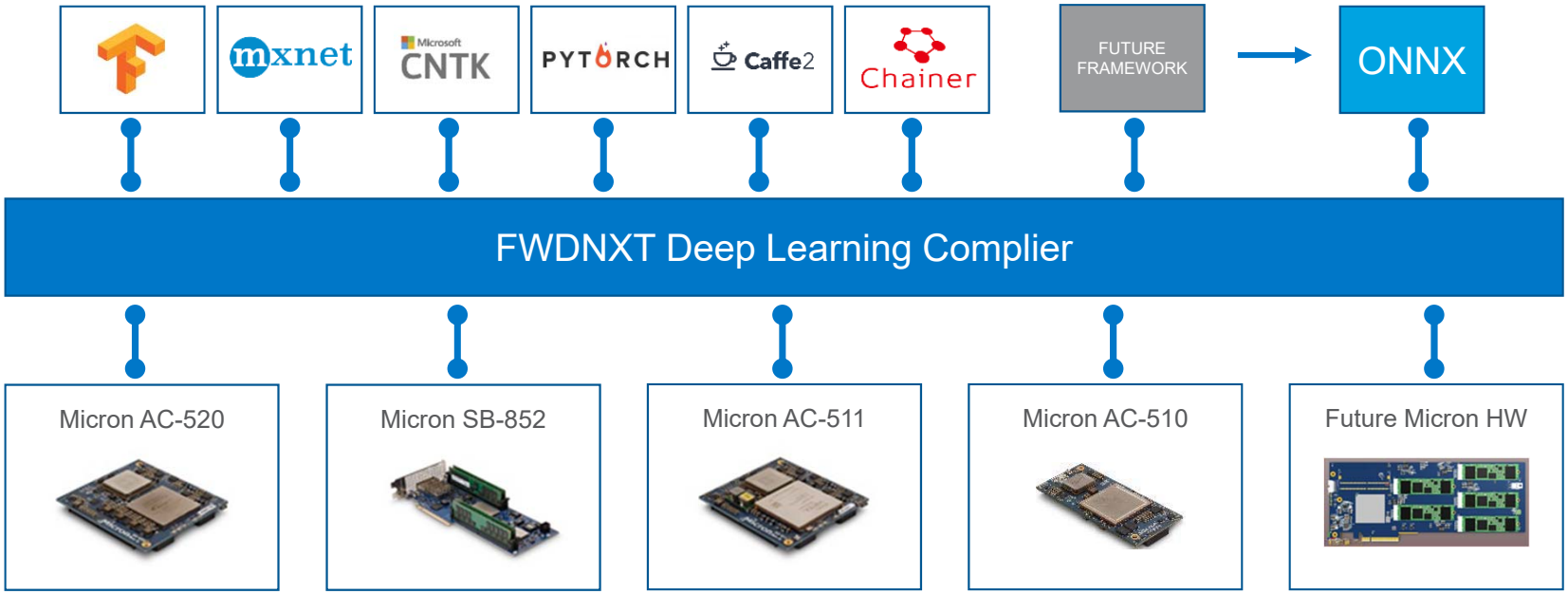
Machine Learning: Accelerating Inference

Mark Hur

Director, Advanced Computing Solutions

January 23, 2019

Acceleration for any Network and Framework



Product Line:

AC Series Modules and EX Series Backplanes

Series	Description	Model No	Description	
AC-Series	FPGA Compute Modules	AC-510	2GB HMC, KU060, Gen3 x8 PCIe	
		AC-511	2GB HMC, 16GB DDR4, VU7P/VU9P, Gen3 x8 PCIe	
		AC-520	2GB HMC, 16GB DDR4, GX1150, Gen3 x8 PCIe	
EX-Series	Carrier boards for AC Series	EX-700	Holds up to 6 AC Modules	
		EX-750	Holds up to 4 AC Modules	

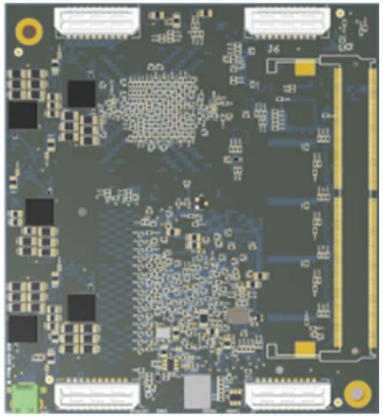
Product Line:
SB-Series and Framework

Series	Description	Model No	Description	
SB-Series	Single Board	SB-851	2 x QSFP28, VU9P, Gen3 PCIe x16	
		SB-853	16GB DDR4, 2 x QSFP28, VU9P, Gen3 PCIe x16	
		SB-852	2GB HMC, Up to 512GB DDR4, 2 x QSFP28, VU7P/VU9P, PCIe x16	
Software/Firmware	HDK	PicoFramework: Host side drivers, APIs, PCIe interface, memory controllers, DMA engine, over 10 years of HPC development effort		
	HMC Controller	Two versions: (1) Optimized for random access performance (2) Optimized for bandwidth		

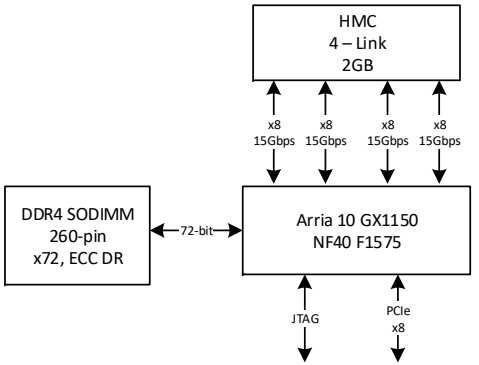
Micron AC-520 FPGA Module

Form Factor	Micron Double Wide Module 98.56 mm x 89.49 mm (L x W) Requires Micron EX-700 or EX-750 carrier board
Host Interface	8-lane PCI-Express Gen3
FPGA	Intel Arria 10 GX1150 F1575 NF40 Package -2 Speed, E Temp Other FPGA options available, please contact Micron
Hybrid Memory Cube	2GB 4 Link Package HMC Four Half Width Links (x8) @ 15Gb/s Bandwidth: 120GB/s
DDR4 SDRAM Memory	16GB ECC SODIMM PC4-2166
Application Development	Intel OpenCL SDK & HDL
Electrical	On module power derived from 12V on carrier board
Quality	Manfuactured to ISO9001 standards FCC, CE, KCC, VCCI, RoHS
Cooling	Standard double width passive heat sink 96 mm x 89 mm x 20mm (L x W x D)
System Monitoring	Power and Tempature Monitoring LEDs to monitor Init, Done, PowerGood OverTemp, TempWarn JTAG Header
Deliverables	AC-520 Module, JTAG cable and adapter Pico Framework Micron HMC Controller IP OpenCL BSP

Top View



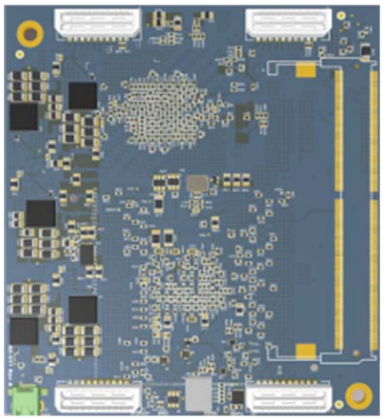
Bottom View

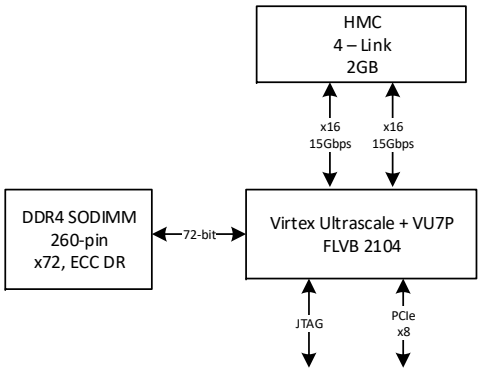


Micron AC-511 FPGA Module

Form Factor	Micron Double Wide Module 98.56 mm x 89.49 mm (L x W) Requires Micron EX-700 or EX-750 carrier board
Host Interface	8-lane PCI-Express Gen3
FPGA	Xilinx Ultrascale+ VU7P FLVB 2104 Package -2 Speed, E Temp Other FPGA options available, please contact Micron
Hybrid Memory Cube	2GB 4 Link Package HMC Two Full Width Links (x16) @ 15Gb/s Bandwidth: 120GB/s
DDR4 SDRAM Memory	16GB ECC SODIMM PC4-2166
Application Development	Xilinx SDA & HDL
Electrical	On module power derived from 12V on carrier board
Quality	Manfuactured to ISO9001 standards FCC, CE, KCC, VCCI, RoHS
Cooling	Standard double width passive heat sink 96 mm x 89 mm x 20mm (L x W x D)
System Monitoring	Power and Tempature Monitoring LEDs to monitor Init, Done, PowerGood OverTemp, TempWarn JTAG Header
Deliverables	AC-511 Module, JTAG cable and adapter Pico Framework Micron HMC Controller IP Xilinx OpenCL DSA 1 Year Warranty & online support

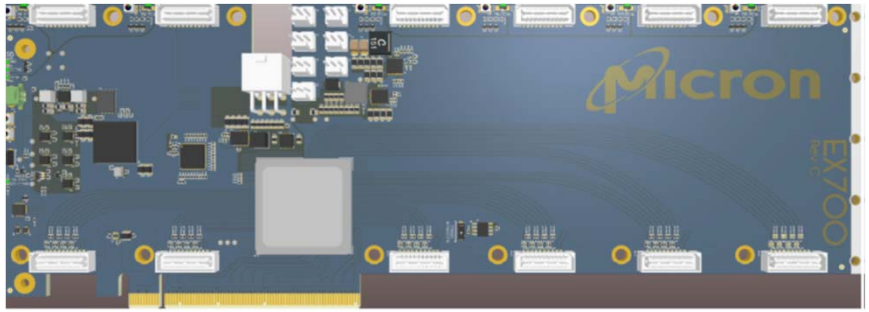
Top View


Bottom View



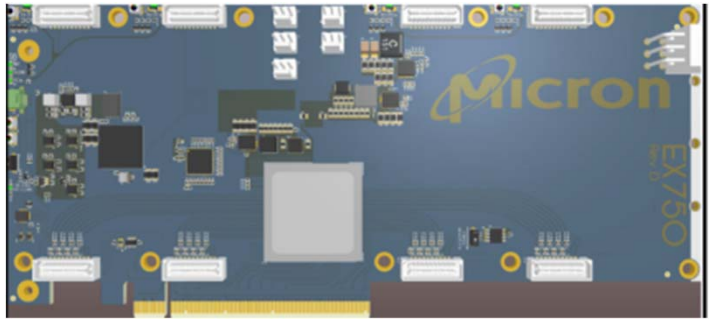
EX-Series Backplanes

EX-700



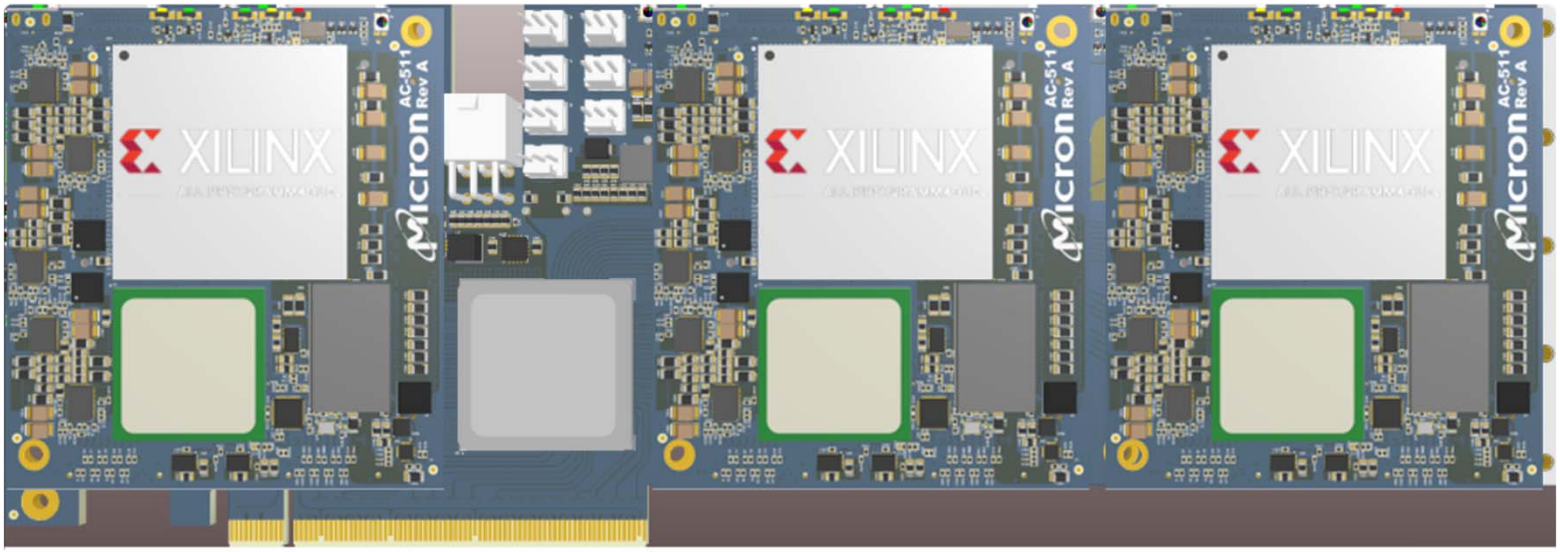
Form Factor	Full-height, Full length, double-width PCI Express Card 312 mm x 111.15 mm (L x H)
Host Interface	16-lane PCI-Express Gen3
PCI Express Switch	Avago PEX 8780 PCI Express Switch 8-lane PCI-Express Gen3 to each AC-Series Module
Electrical	On board power derived from 12V 6-pin AUX connector
Quality	Manfuactured to ISO9001 standards UL, FCC, CE, KCC, VCCI, RoHS
Cooling	Standard double width passive heat sink
Deliverables	EX-700 PCI-Express Carrier Board 1 Year Warranty & online support

EX-750



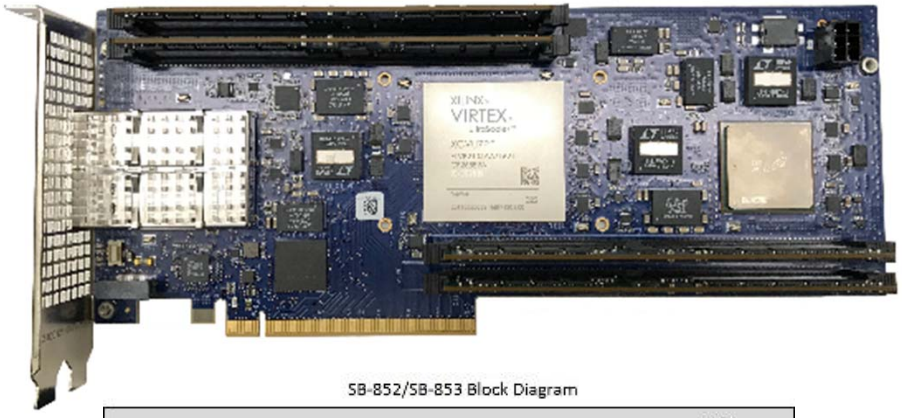
Form Factor	Full-height, double-width PCI Express Card (GPU Size) 247.65 mm x 111.15 mm (L x H)
Host Interface	16-lane PCI-Express Gen3
PCI Express Switch	Avago PEX 8780 PCI Express Switch 8-lane PCI-Express Gen3 to each AC-Series Module
Electrical	On board power derived from 12V 6-pin AUX connector
Quality	Manfuactured to ISO9001 standards UL, FCC, CE, KCC, VCCI, RoHS
Cooling	Standard double width passive heat sink
Deliverables	EX-750 PCI-Express Carrier Board 1 Year Warranty & online support

Three FPGAs in one PCIe Slot

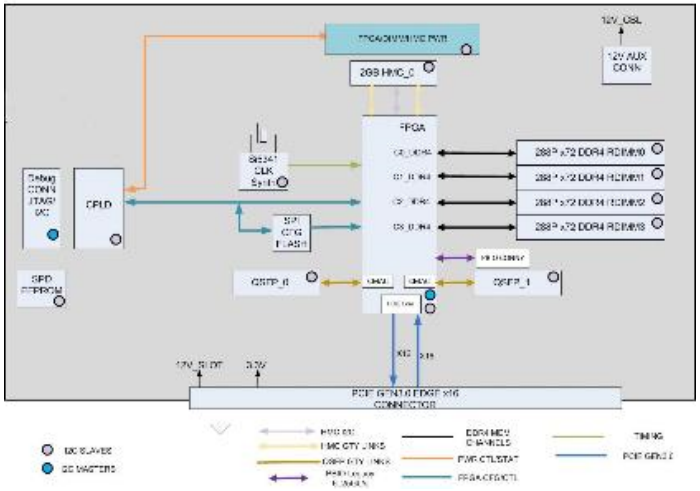


SB-Series Board: SB-852

Item	Description
Form Factor	Full-height, GPU-length (10.5 inch) PCIe Card
Host Interface	PCIe Gen3 x16
FPGA	Xilinx VU7P/VU9P (-2FLVB2104E)
HMC Memory	2GB, 2 x Full Width Link (120GB/s)
DDR Memory	4 Channels of DDR4, up to 128GB
Configuration FLASH	2Gb SPI
I/O	2 x QSFP28
System Clock	250MHz
PCIe Reference Clock	100MHz
Power Dissipation	150W Max
Quality Standards	IPC Class 2, ISO 9001
Environmental	RoHS, UL, FCC, CE, VCCI, KCC
Status LEDs	Power, Link, and FPGA status
Cooling	Standard double width passive heat sink
Deliverables	SB-852 board, PicoFramework 1 Year Warranty & Online Support

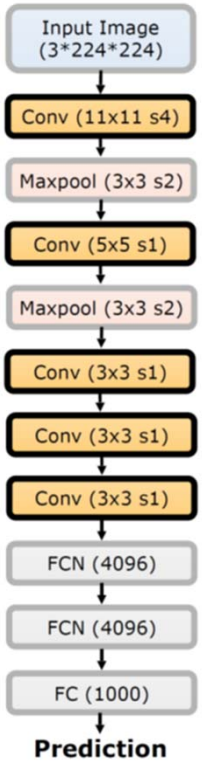


SB-852/SB-853 Block Diagram



Specifications Subject to Change

Deep Neural Networks



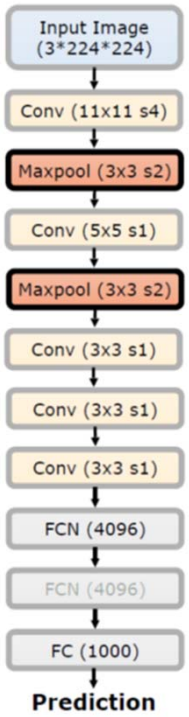
Convolutions:

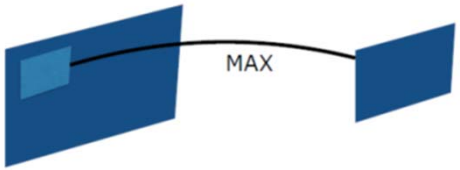
- Compute Intensive
- Embarrassingly Parallel
- >95% of the workload
- Mostly MAC Operations

Source: FWDNXT

Micron

Deep Neural Networks





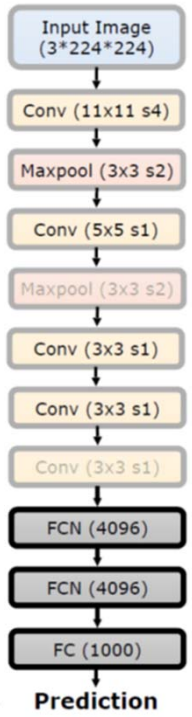
Max Pooling:

- Make up ~1% of workload
- Lesser parallelism to exploit
- Mostly comparisons

Source: FWDNXT

Micron

Deep Neural Networks



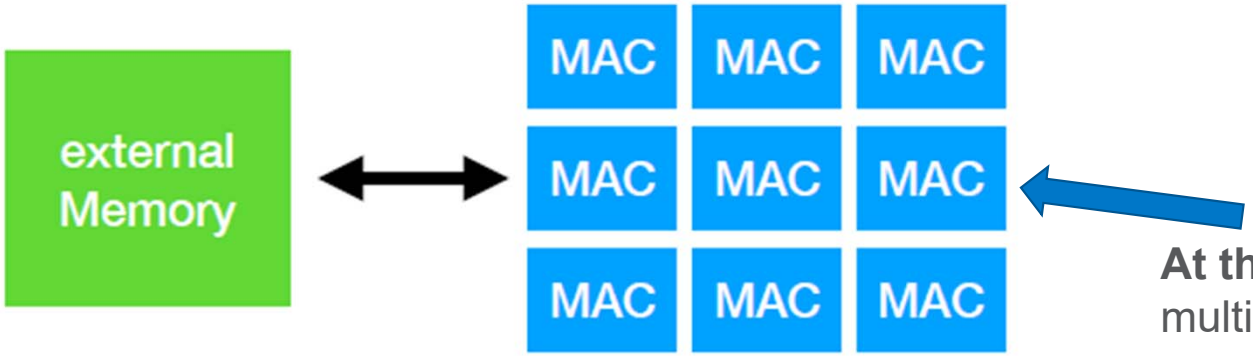
Fully Connected Layers:

- Tens of MB of weights
- No weights reuse
- Bandwidth Intensive
- Mostly MAC Ops

Source: FWDNXT

Micron

JBOM: Just a Bunch of MACs



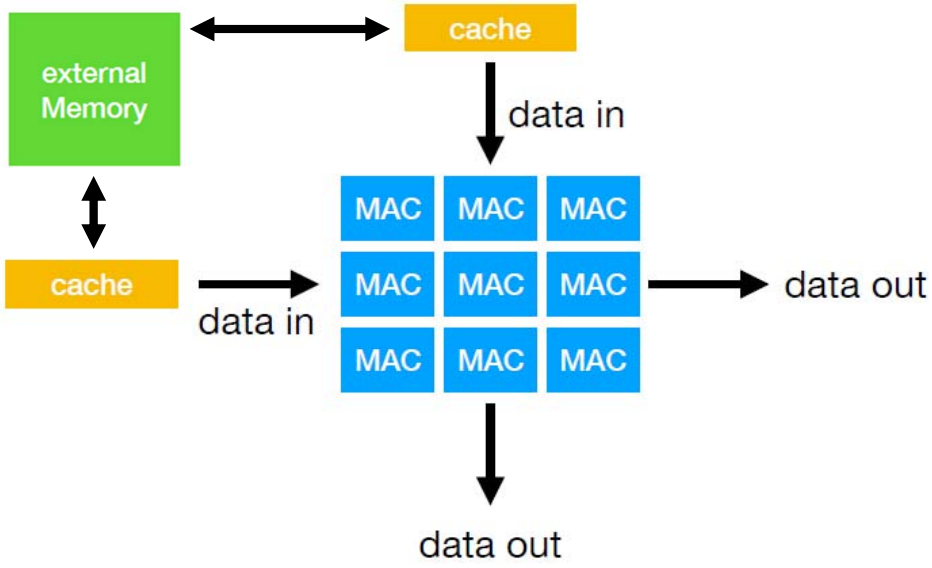
At the core:
multiply accumulate units
- embarrassingly parallel
- > 95% of the workload
- mostly MAC operations

- Goal: High computational efficiency
- Problem: Single-instruction, multiple data
- Two dominant architectures:
  1. Global data movement: **systolic arrays**
  2. Finer data movement: **cached arrays**

Source: FWDNXT

Micron

Systolic Array



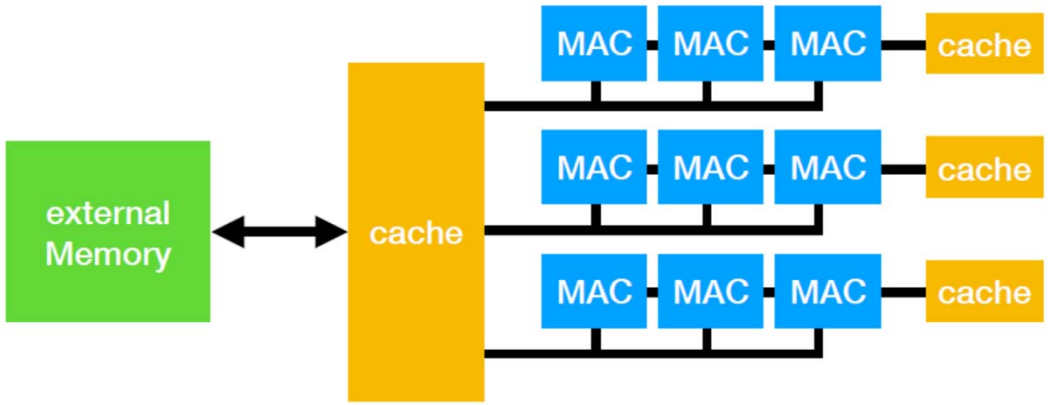
- Compute nodes arranged in pipelined grid layout
- Data flows through it, in steps synchronized in columns and rows

- **Advantage:** Simple HW design: pipelined MAC array
- **Disadvantage:** Harder to map algorithms efficiently, low HW utilization

Source: FWDNXT

Micron

Cached Array



- Compute nodes divided into subgroups which are individually controlled
- System runs by executing a stream of instructions, sequenced by program counter

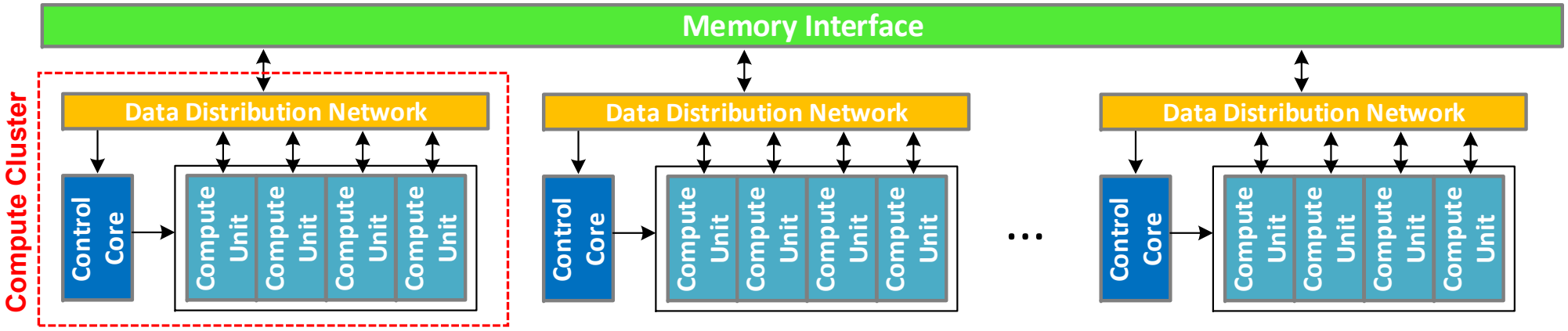
- **Advantages:** Very high utilization, can turn off unused portions
- **Disadvantages:** More complex HW design

- **Cached Array more power efficient than Systolic Array**

Source: FWDNXT

Micron

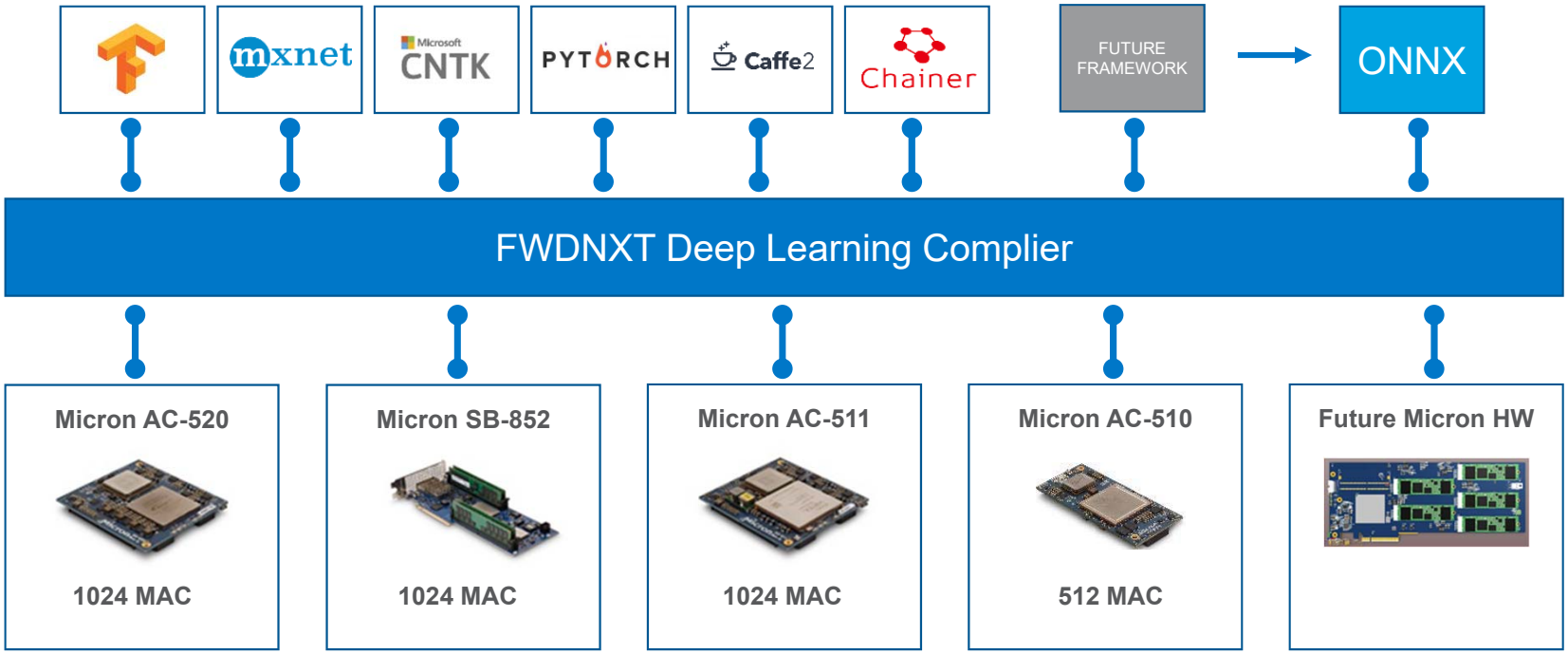
Inference Engine Architecture



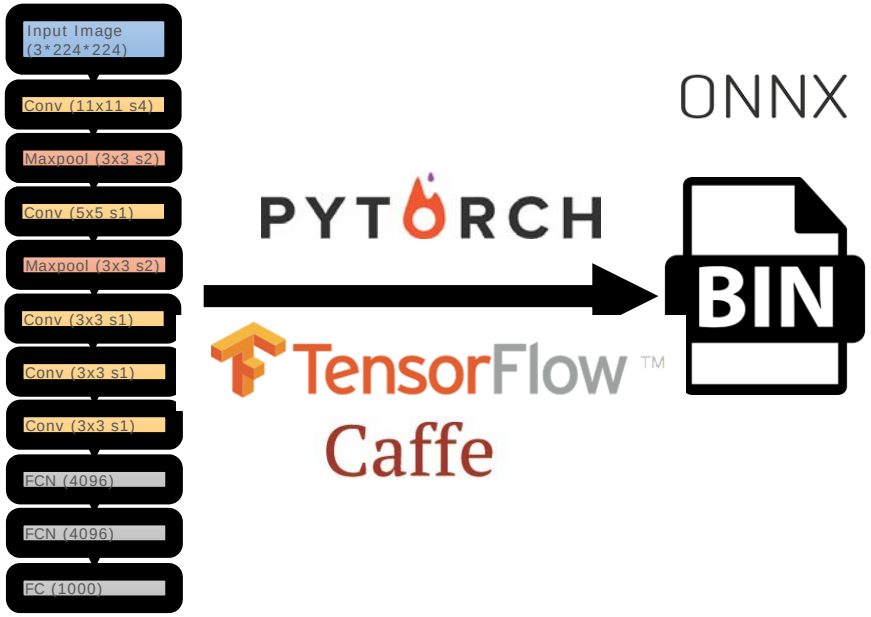
- Modular architecture in the form of independent “compute clusters”
- A cluster consists of:
  - A control core: RISC-like instruction pipeline
  - 4x compute units: engines for processing model’s layers
  - A data distribution network: local interface to memory
- Top level memory interface is outside the compute cluster for ease of porting to different memory types

Source: FWDNXT

Micron


mxnet
Microsoft CNTK
PYTORCH
Caffe2
Chainer
FUTURE FRAMEWORK
ONNX
FWDNXT Deep Learning Complier
Micron AC-520
1024 MAC
Micron SB-852
1024 MAC
Micron AC-511
1024 MAC
Micron AC-510
512 MAC
Future Micron HW

Deep Learning Workflow



Neural Network

Training Frameworks

Network Description

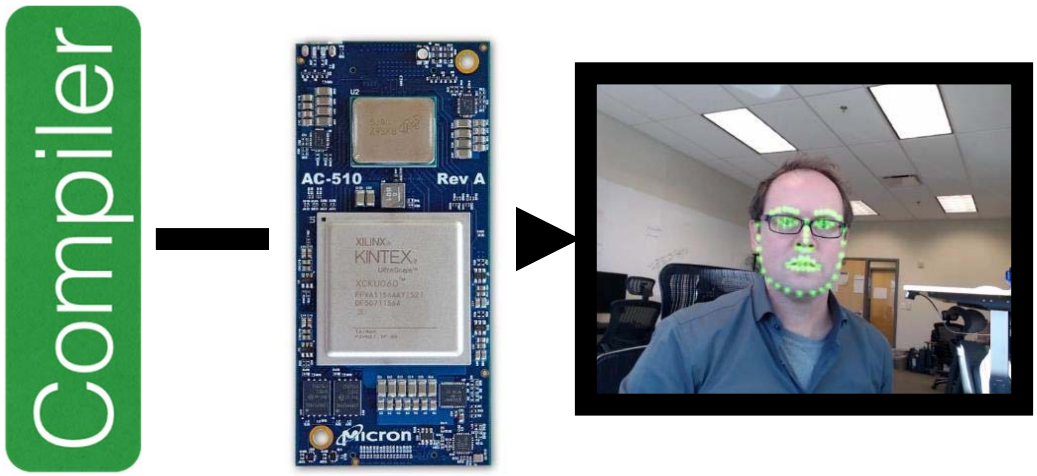
Steps:

1. Select neural network model
2. Train in your favorite Deep learning Framework
3. Obtain trained model and output as ONNX interchange format (32 bit)

Source: FWDNXT

Micron

Deep Learning Workflow



Compiler Hardware Deployment

Steps:

4. Input trained model to compiler
5. Compiler converts into machine code - do not have to code HW
6. Change a few lines of your application to target the hardware!
7. Done!

Source: FWDNXT

Micron

sample deployment code

import

instantiate

run

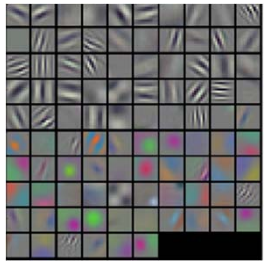
```
1 import snowflake
5 import PIL
6 from PIL import Image
7 import numpy as np
8
9 from argparse import ArgumentParser
10 # argument Checking
11 parser = ArgumentParser(description="FWDNXT Demonstration")
12 ...
19 args = parser.parse_args()
20
21 #Load image into a numpy array
22 img = Image.open(args.image)
24 #Resize it to the size expected by the network
25 img = img.resize((args.res[2], args.res[1]), resample=PIL.Image.BILINEAR)
27 #Convert to numpy float
28 img = np.array(img).astype(np.float32) / 255
30 #Transpose to plane-major, as required by our API
31 img = np.ascontiguousarray(img.transpose(2,0,1))
32
33 #Normalize images
34 stat_mean = list([0.485, 0.456, 0.406])
35 stat_std = list([0.229, 0.224, 0.225])
36 for i in range(3):
37     img[i] = (img[i] - stat_mean[i])/stat_std[i]
38
40 sf = snowflake.Snowflake()
                                    gs.res[1], args.res[2], args.res[0]), args.model_path, '')
45
46 #Create the storage for the result and run one inference
                                    e=np.float32)
48 sf.Run(img, result)
50 #Convert to numpy and print top-5
51 idxs = (-result).argsort()
54 print('------------ Snowflake results ------------')
55 for i in range(5):
62     print(idxs[i], result[idxs[i]])
```

Source: FWDNXT

Accelerating Any ML Algorithm

1. Use any Network

Neural network model file (weights + architecture)



CNN
RNN
LSTM

2. Use any Framework

PyTorch / Torch
TensorFlow
Caffe
others…

PYTORCH

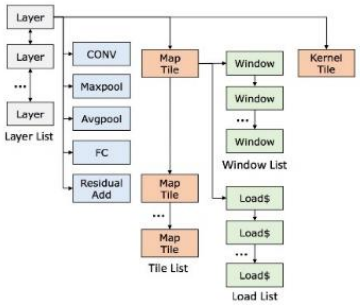
TensorFlow

Caffe2

mxnet

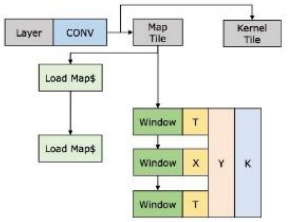
3. Compiler

Compiler data structure



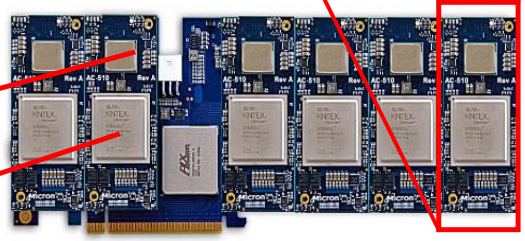
4. Machine code

Machine code for any neural network layers



5. ML runs on cores in Micron HW

512 MACs



High Performance Memory

FPGAs

Micron