Xilinx DNN Processor

An Inference Engine, Network Compiler + Runtime for Xilinx FPGAs

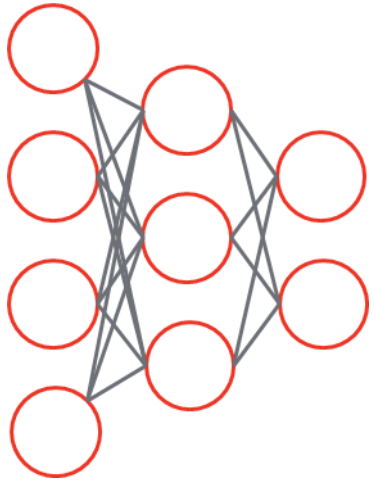
Rahul Nimaiyar, Brian Sun, Victor Wu, Thomas Branca, Yi Wang, Justin Oo, Elliott Delaye, Aaron Ng, Paolo D'Alberto, Sean Settle, Bill Teng, Manasa Bollavaram, Chaithanya Dudha, Hanh Hoang, Swati Gupta,, Alina Huang, Ephrem Wu, Samuel Bayliss, Phil James-Roxby, Ralph Wittig, Ashish Sirasao, Ravi Sunkavalli

21st August 2018

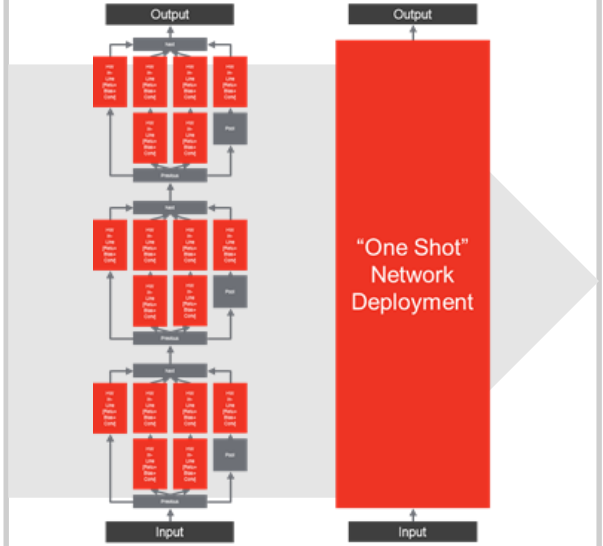
XILINX.

Xilinx Inference Engine – DNN Processor + Compiler

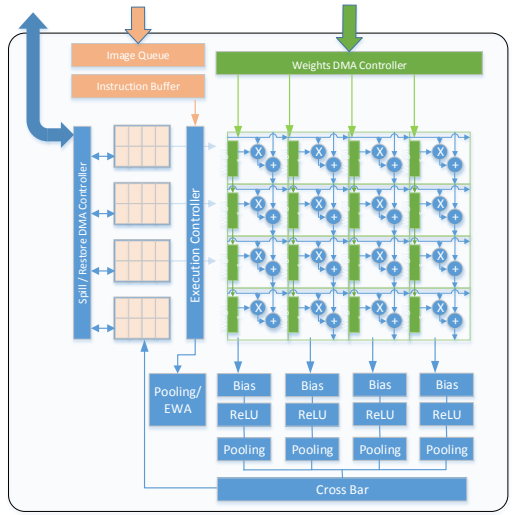
Trained Model



Compiler + Runtime



Xilinx DNN Processor



Low Latency, High Throughput – Batch 1

60-80% Efficiency

No FPGA Expertise Required

XILINX.

Virtex® UltraScale+™ VU9P FPGA

Virtex® UltraScale+™ FPGA VCU1525

Developer Board and Reference Design

- 16nm TSMC FF+ FPGA
- 2.5M System Logic Cells
- 6840 DSP Blocks (18x27 MACs)
- 382 Mbit On-Die SRAM
- 4 DDR4-2400 x72 Channels

- VU9P Virtex UltraScale+ FPGA
- 21 TOPS (INT8)
- 382 Mbit on-chip SRAM
- 64 GByte on-board DRAM
- 75W

XILINX

Motivations for Deep Learning on FPGA

Data Parallel

- 2D Array of MACs
- Flexible on-chip memory access
  - High Bandwidth, Multiple Access Ports

Data Reuse

- Near Memory Compute
- Programmable routing for data & filter reuse

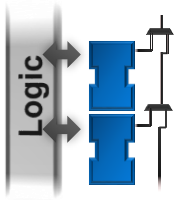
Compression & Sparsity

- Flexible Data Types
  - FP32/16, INT16/8/4/2, Binary/Ternary
- Sparsity friendly compute

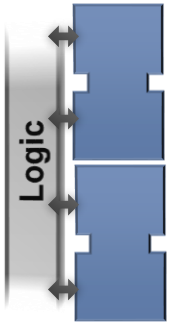
Virtex® UltraScale+™ Full Spectrum of Memory

Distributed RAM
(10s of megabits)

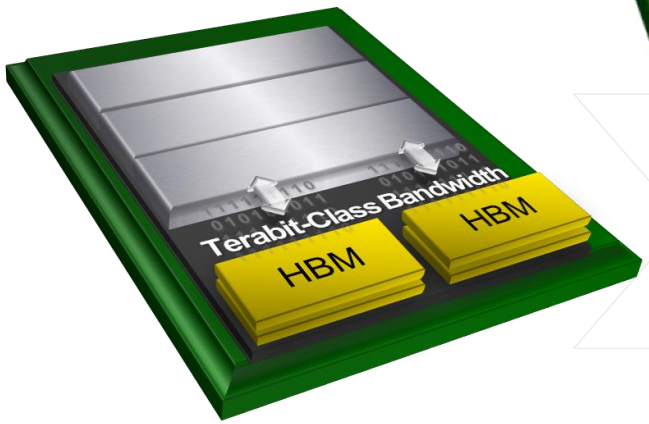
Block RAM
(10s of megabits)



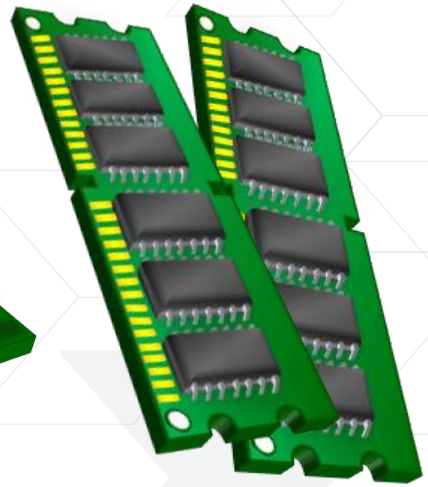
UltraRAM
(100s of Megabits)



High Bandwidth Memory
(Multi-Gigabyte)



External Memory
2666-DDR4
(Multi-Gigabyte)

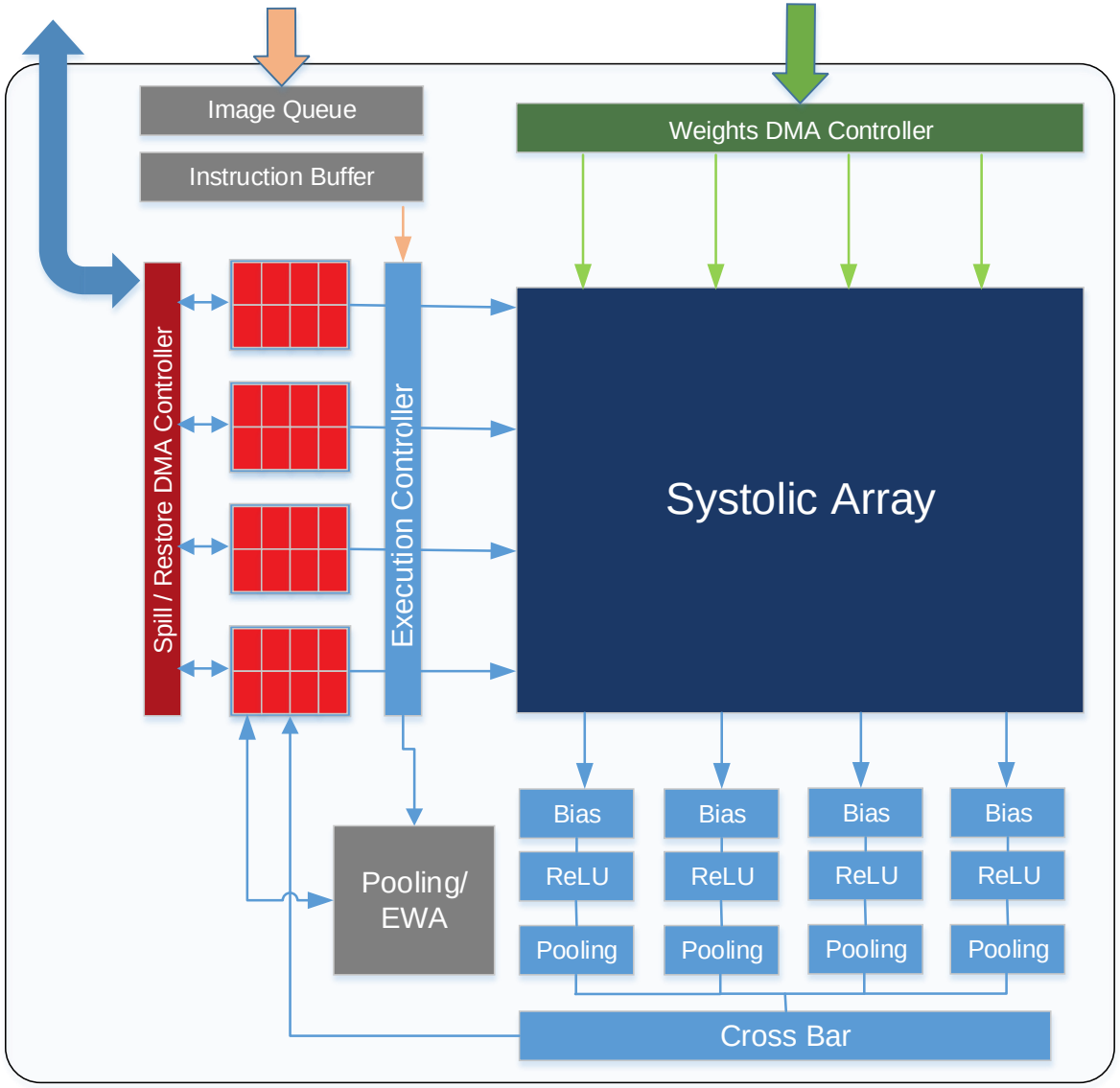


VU9P	36Mb *675Tb/s*	77Mb *216Tb/s*	270Mb *80Tb/s*	VU37P: 8GB, 460GB/s	64GB *85GB/s*

5 Tiers of Memory -> Build custom memory hierarchy.
500 Mb of On-chip Memory and Tb/s of On-chip Memory Bandwidth

XILINX.

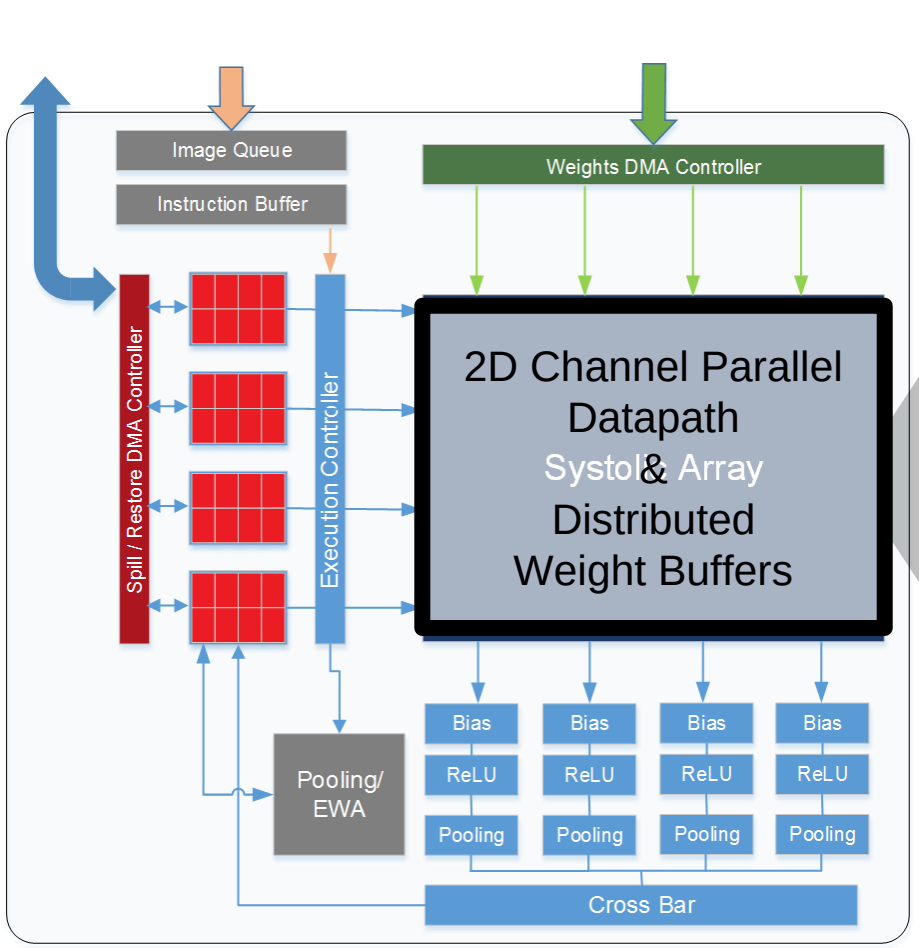
Xilinx DNN Processor (xDNN)

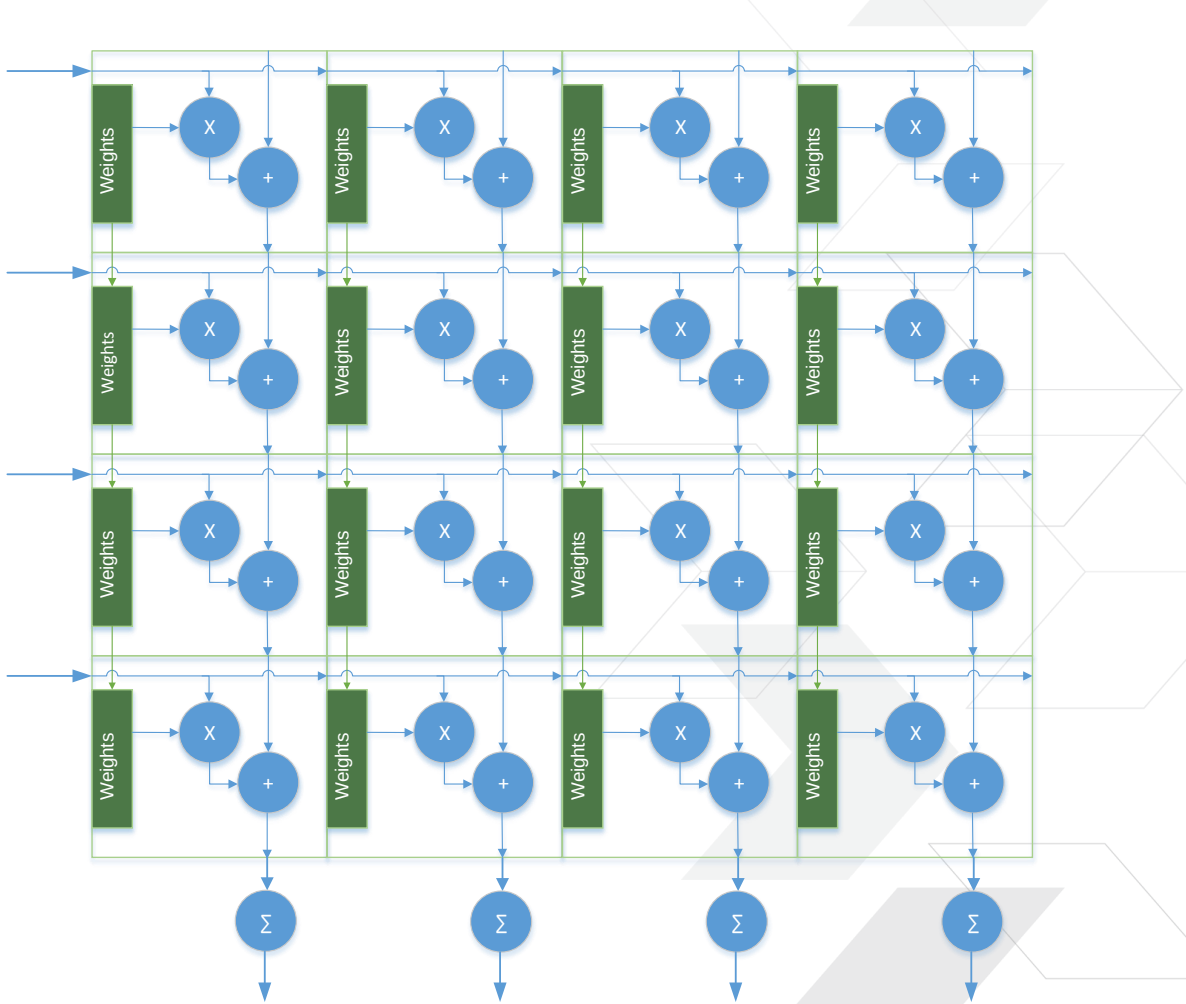


- Configurable Overlay Processor
- DNN Specific Instruction Set
  - Convolution, Max Pool etc.
- Any Network, Any Image Size
- High Frequency & High Compute Efficiency
- Compile and run new networks

XILINX.

xDNN – Channel Parallel Systolic Array

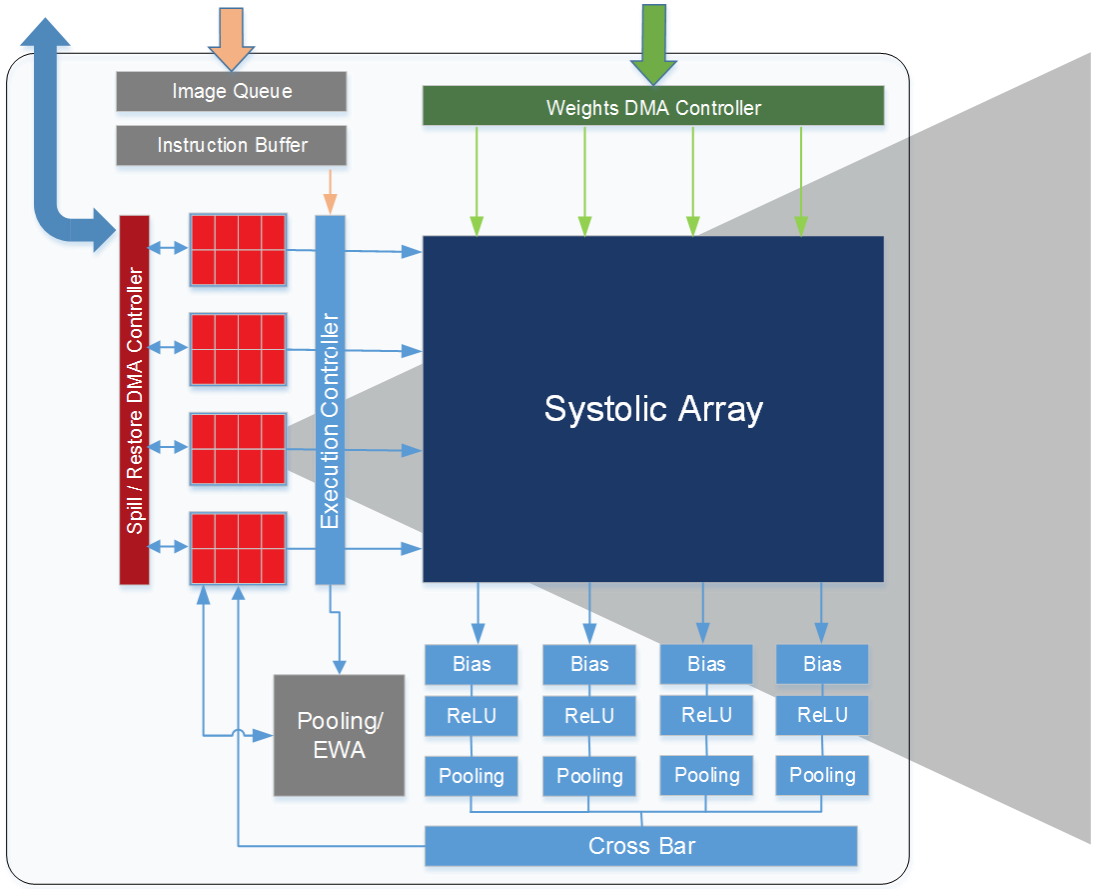


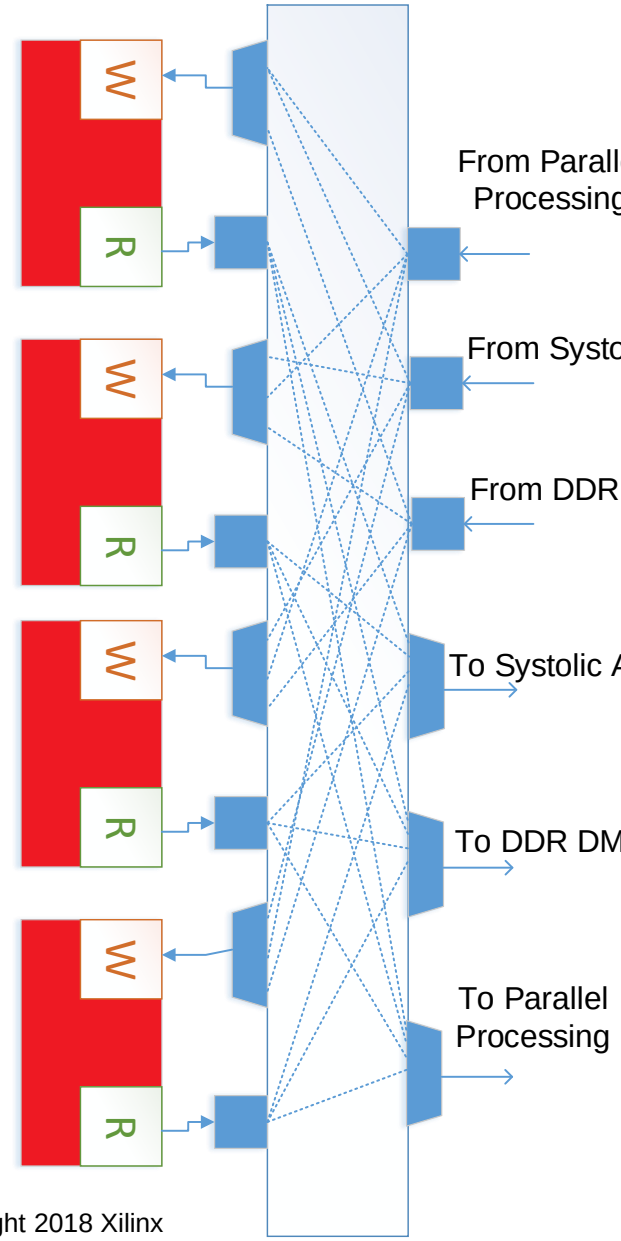


Micro-Architecture Optimized for underlying Ultrascale+ FPGA Fabric

XILINX.

xDNN Processor – Tensor Memory





Channel Parallel Concurrent Access

From Parallel Processing

From Systolic Array

From DDR DMA

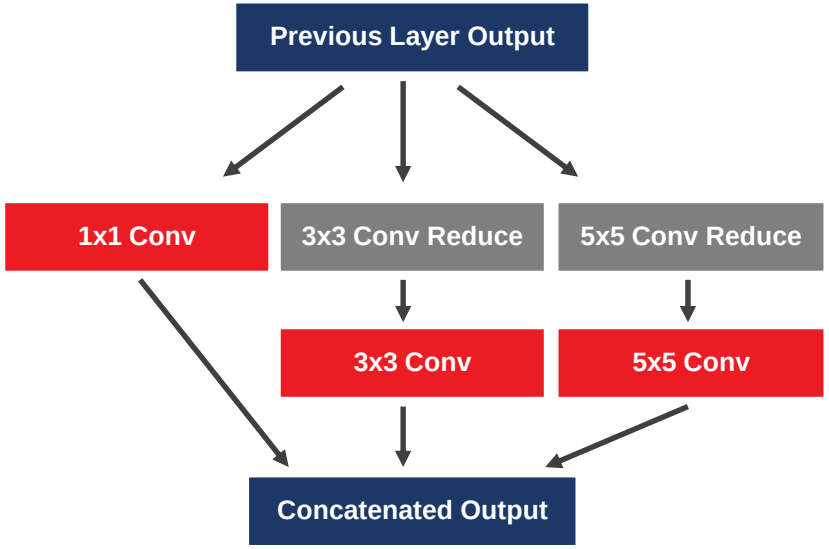
To Systolic Array

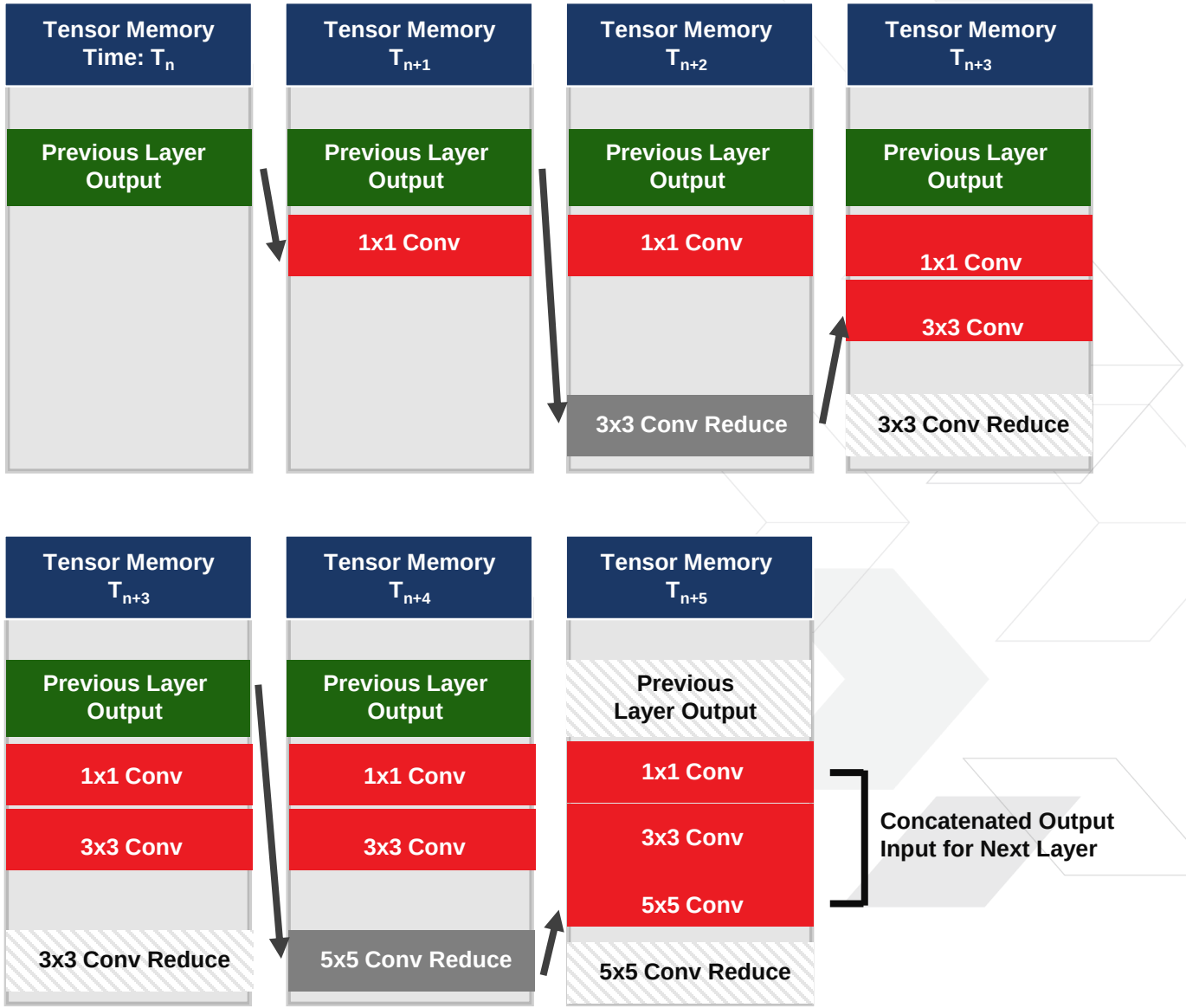
To DDR DMA

To Parallel Processing

XILINX

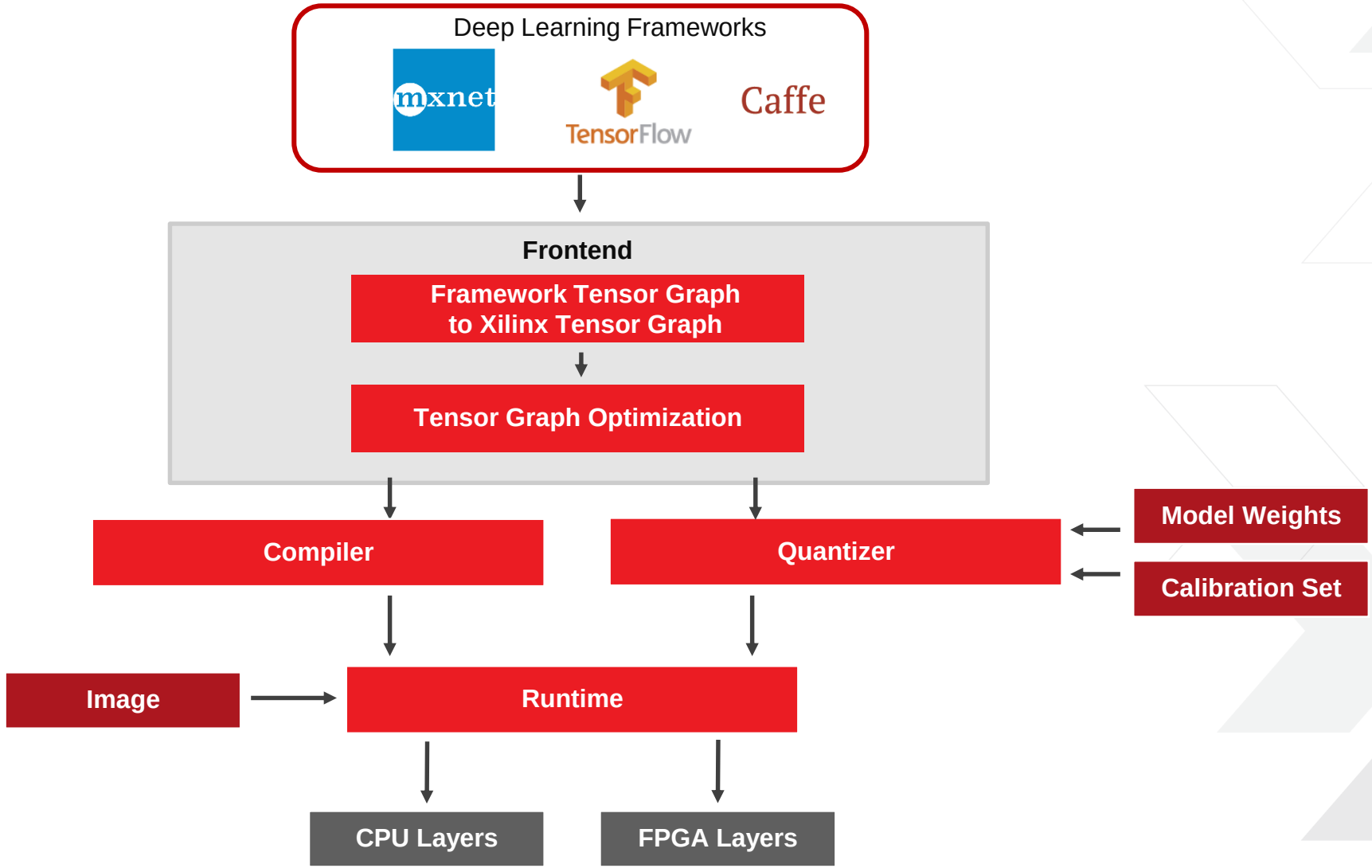
Efficient Memory Utilization





XILINX.

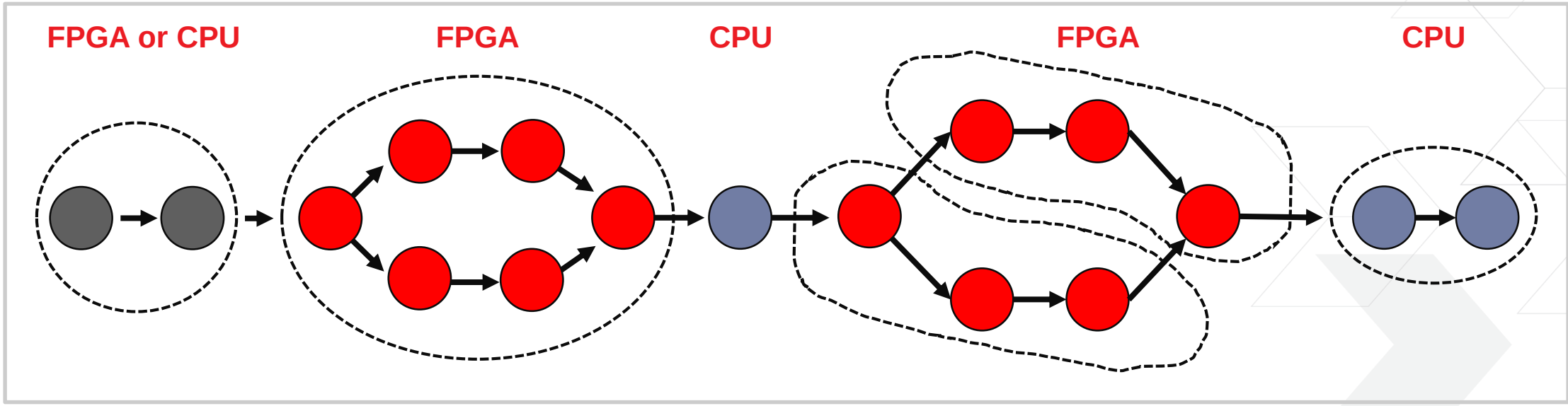
xDNN Compiler + Runtime



https://github.com/Xilinx/ml-suite

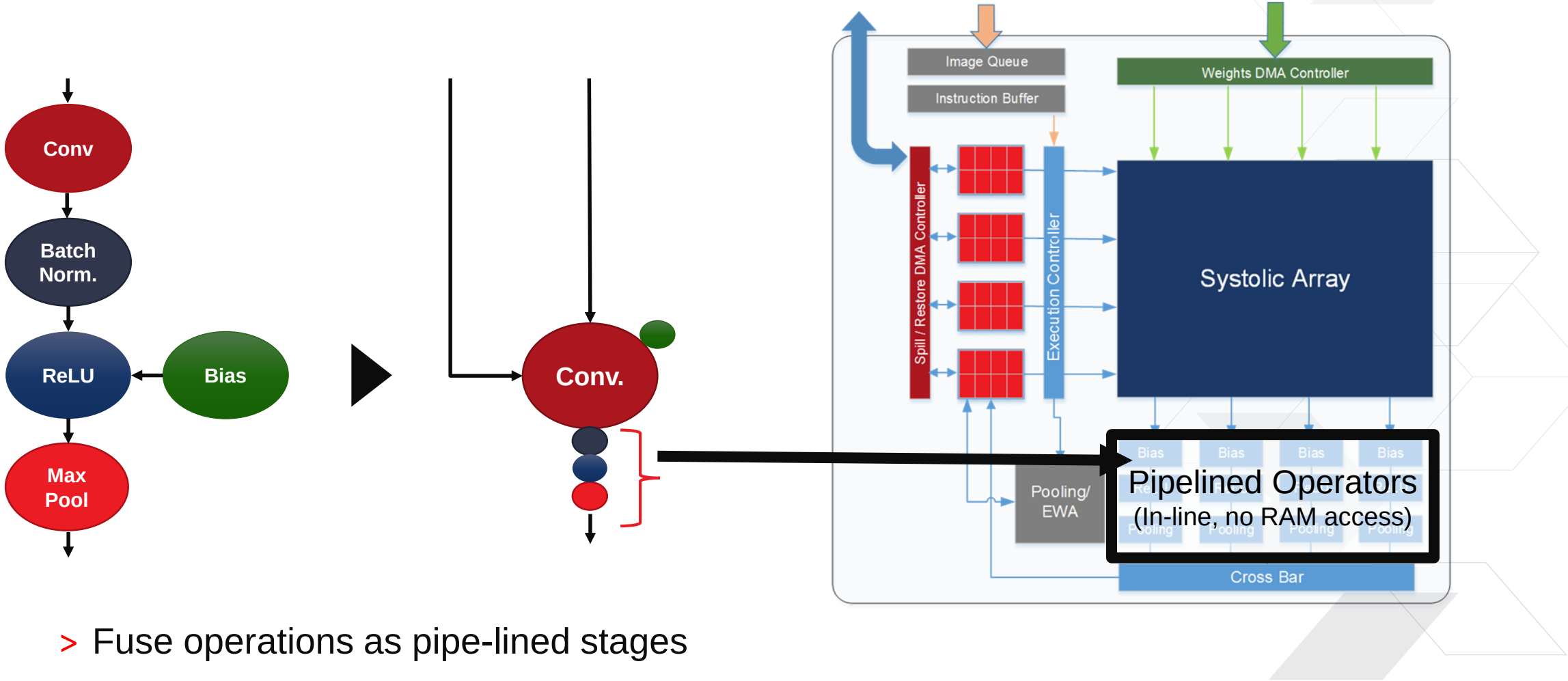
Graph Partitioning

- One time loading of sub-graph instructions
- Data Flow Execution



1 Core -> Multi-Core -> Multi-Chip

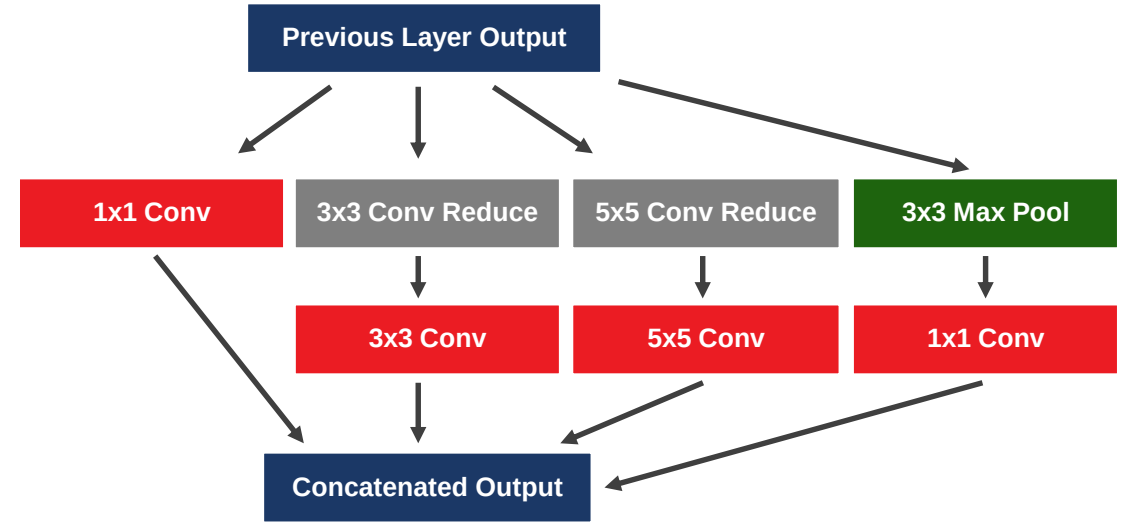
Fusing of Operations

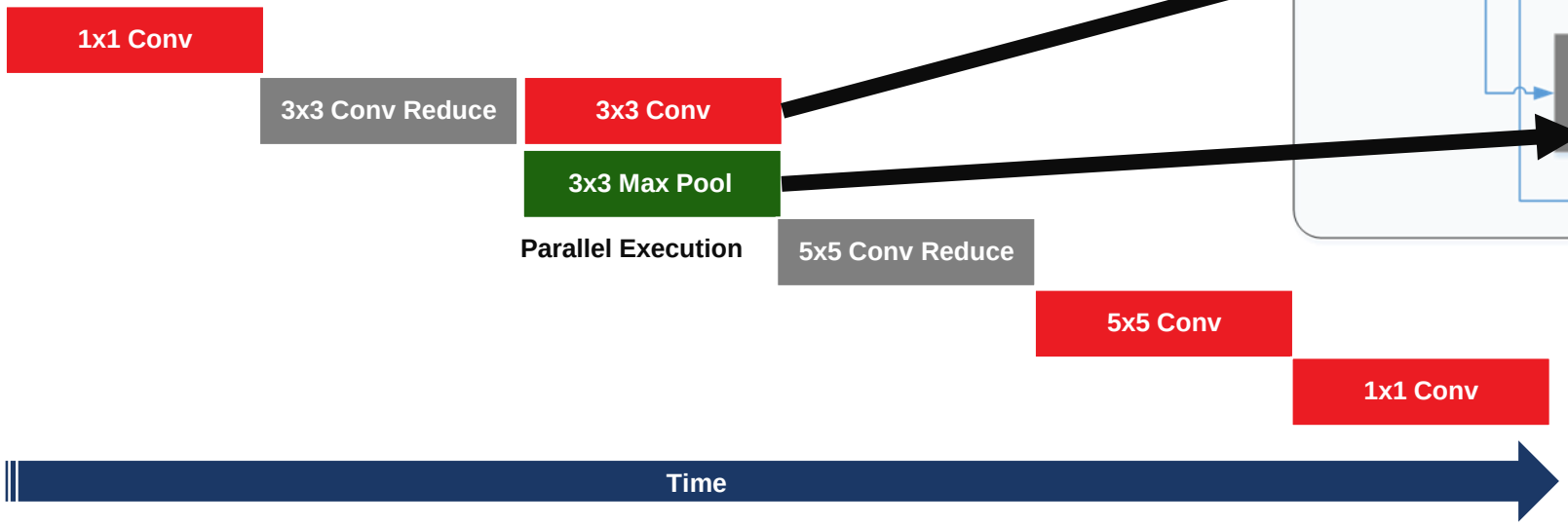


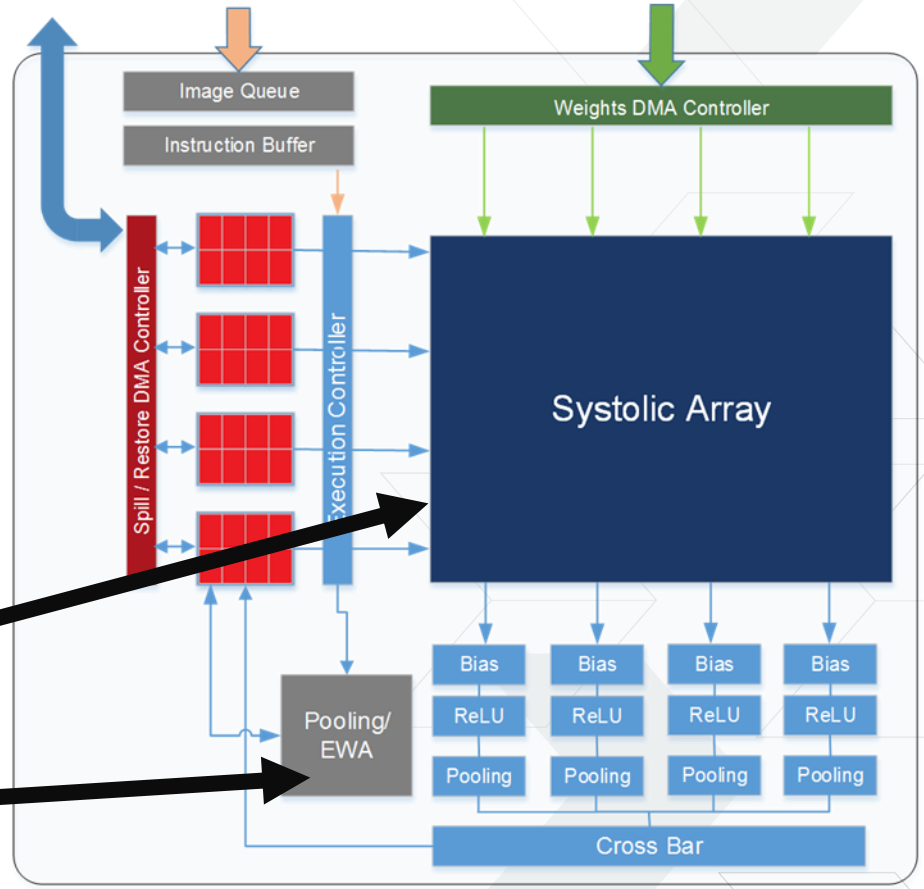
- Fuse operations as pipe-lined stages
- Access activations only once

XILINX.

Instruction Level Parallelism





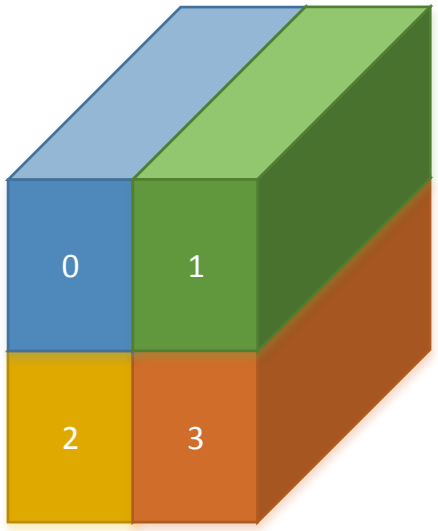


XILINX.

Automatic Intra Layer Tiling

- Tile when Feature Map size exceeds on-chip memory
- Work on full feature map depth

Any Feature Map Size



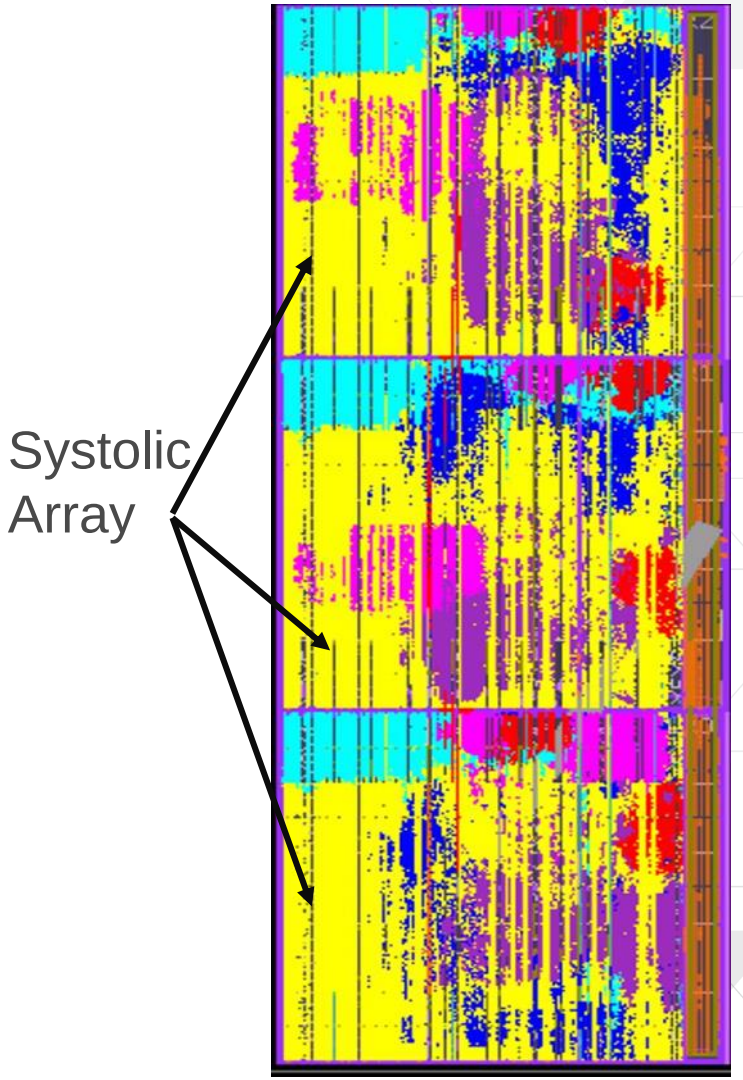
XILINX.

xDNN Key Functions

Feature	Details
Convolution	NxM, Stride 1,2,4,8 N,M=1-15
Max Pool	NxM, Stride 1,2,4,8 N,M=1-15
Avg Pool	NxM, Stride 1,2,4,8 N,M=1-15
Dilated Convolution	Factor 1,2,4
De-Convolution	NxM, Stride 1,2,4,8 N,M=1-15
Up-sampling	Factor 2,4,8,16
Activation	ReLU, pReLU
Elementwise Addition	Any – Square, Rectangular
Precision	Int8, Int16
Network	Classification: e.g. ResNet Object Detection: e.g. YOLO v2, Segmentation: e.g. MaskRCNN

xDNN Implementation on VU9P

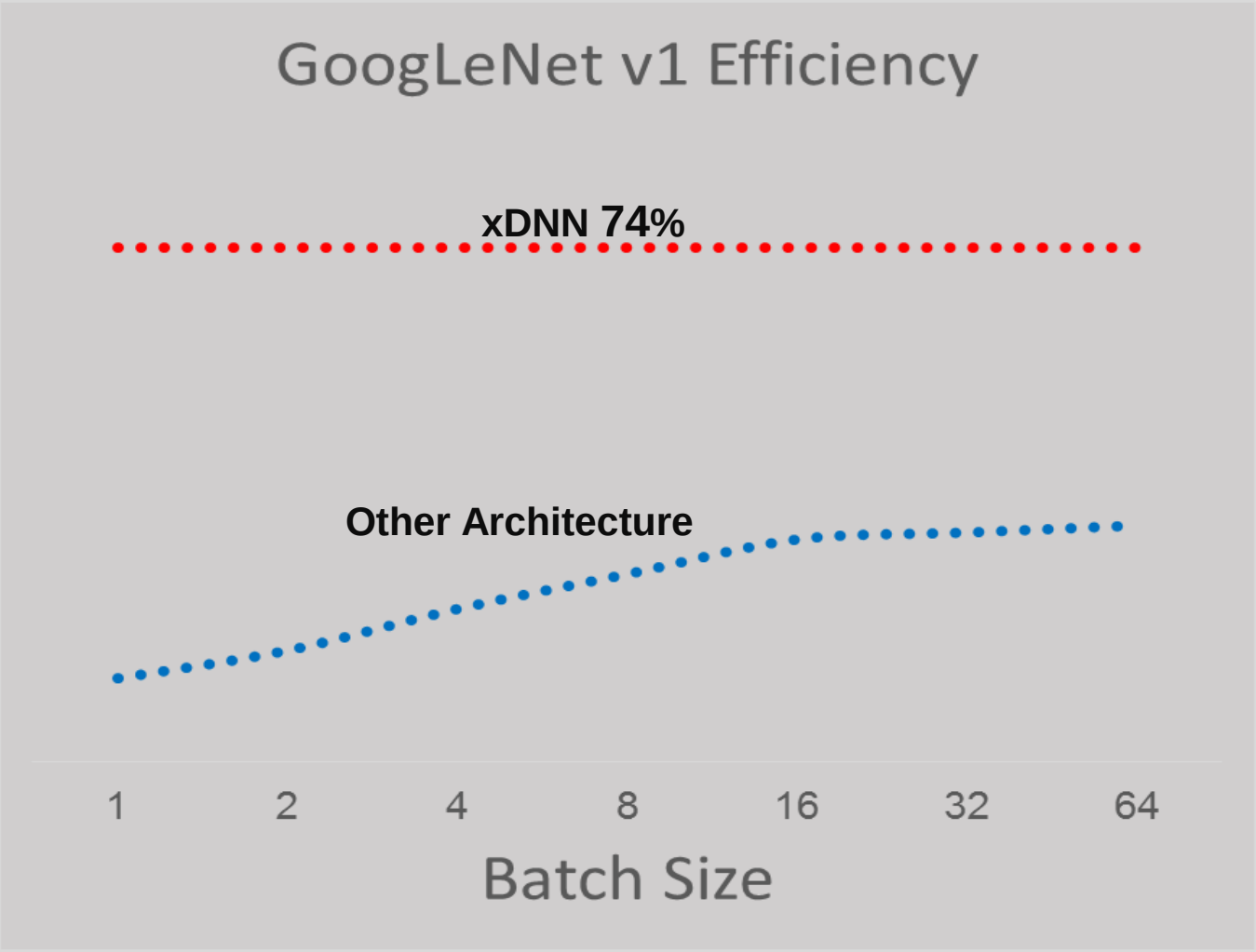
Resource	Count	VU9P Utilization
LUTs	612k	52%
DSPs	5493	80%
BRAM	228	38%
URAM	864	92%



800MHz – 90% of device Fmax

- 3 Large 96x16 PEs– 1 in each SLR
- Systolic Array at 800 MHz; Rest of logic at 400 MHz

xDNN Compute Efficiency

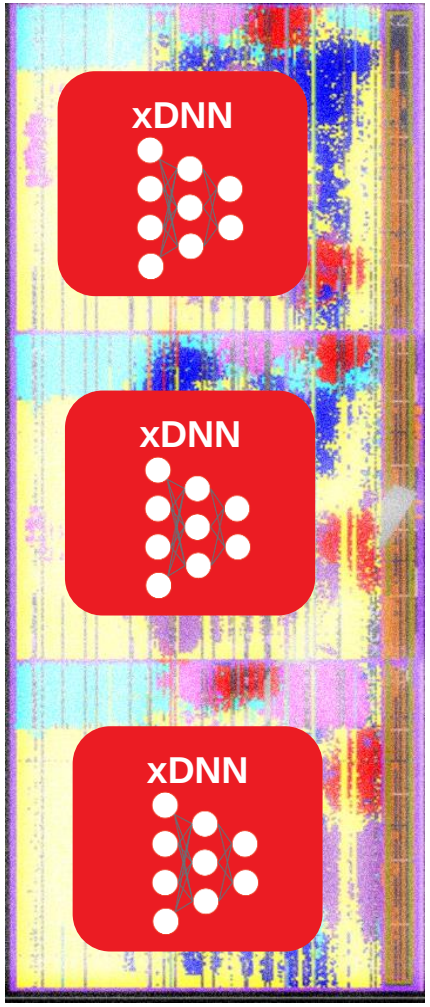


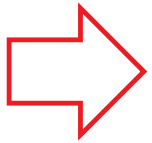
60-80% Efficiency Across Networks

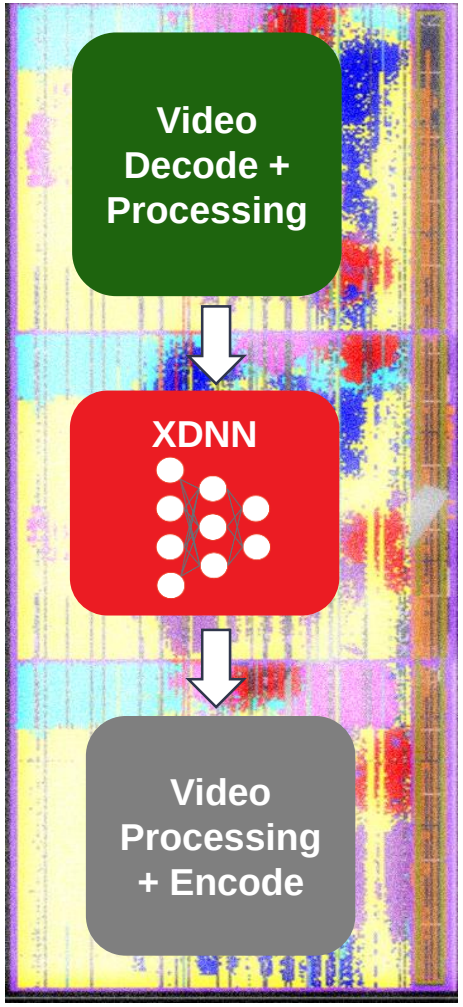
Full Benchmark Data at Xilinx Developers Forum – Oct' 1, 2018

XILINX.

Custom Deep Learning Pipeline





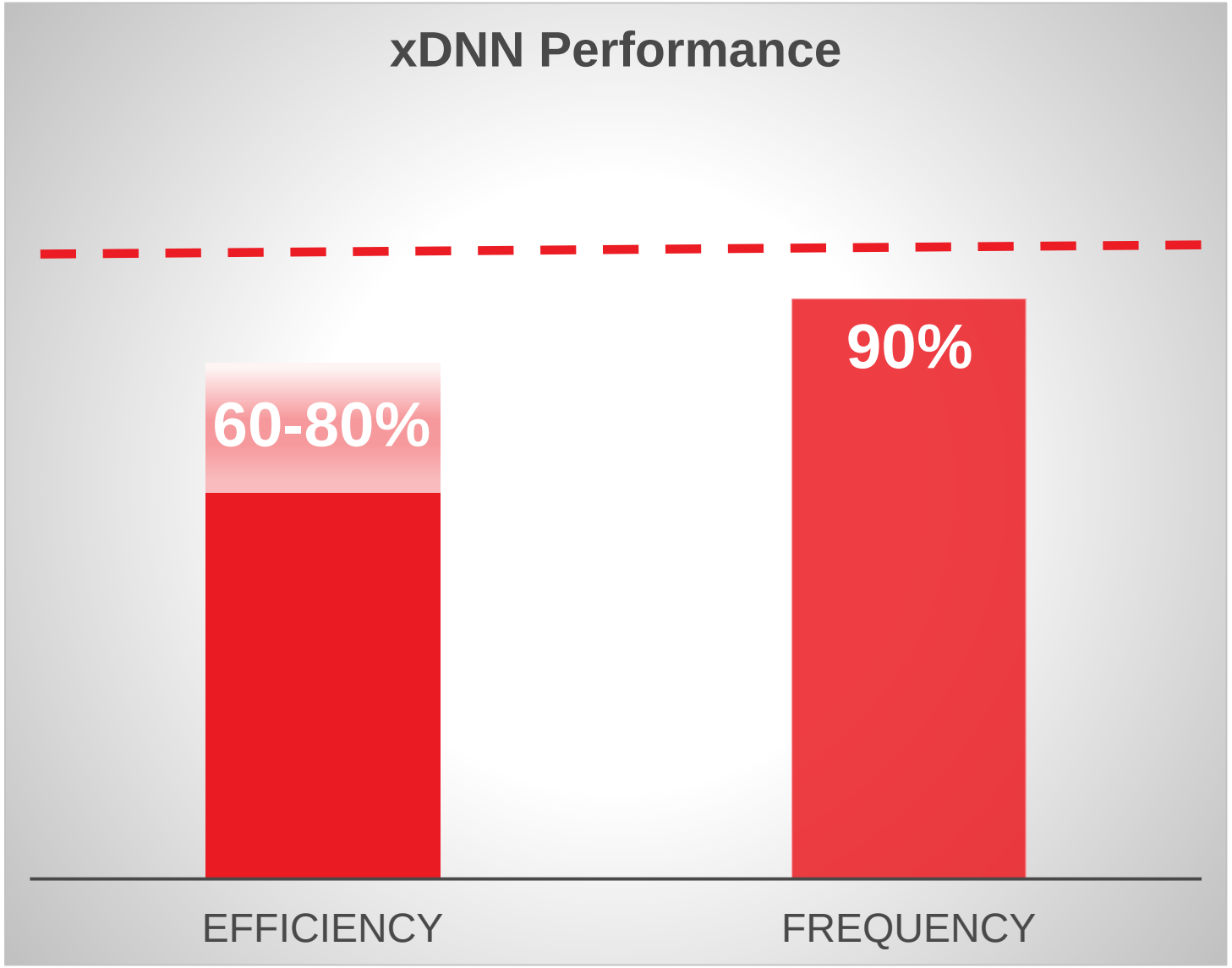


- Video + ML
- Genomics + ML
- Risk Modelling + ML
- Database + ML
- Network IPS + ML
- Storage + ML

Integrate Custom Applications with xDNN. Lower end-to-end latency

XILINX.

Xilinx DNN Processor - Summary



- No FPGA Expertise Needed
- Compile & Run Trained Networks
- Deploy on AWS F1, other Cloud Platforms
- Deploy in PCIe Cards

XILINX.

Additional Information

Connect • Learn • Share

XDF connects software developers and system designers to the deep expertise of Xilinx engineers, partners, and industry leaders.

Silicon Valley October 1-2
Beijing October 6th
Frankfurt December 10

Learn More www.xilinx.com/xdf