Xilinx ML Suite

Amay Churi

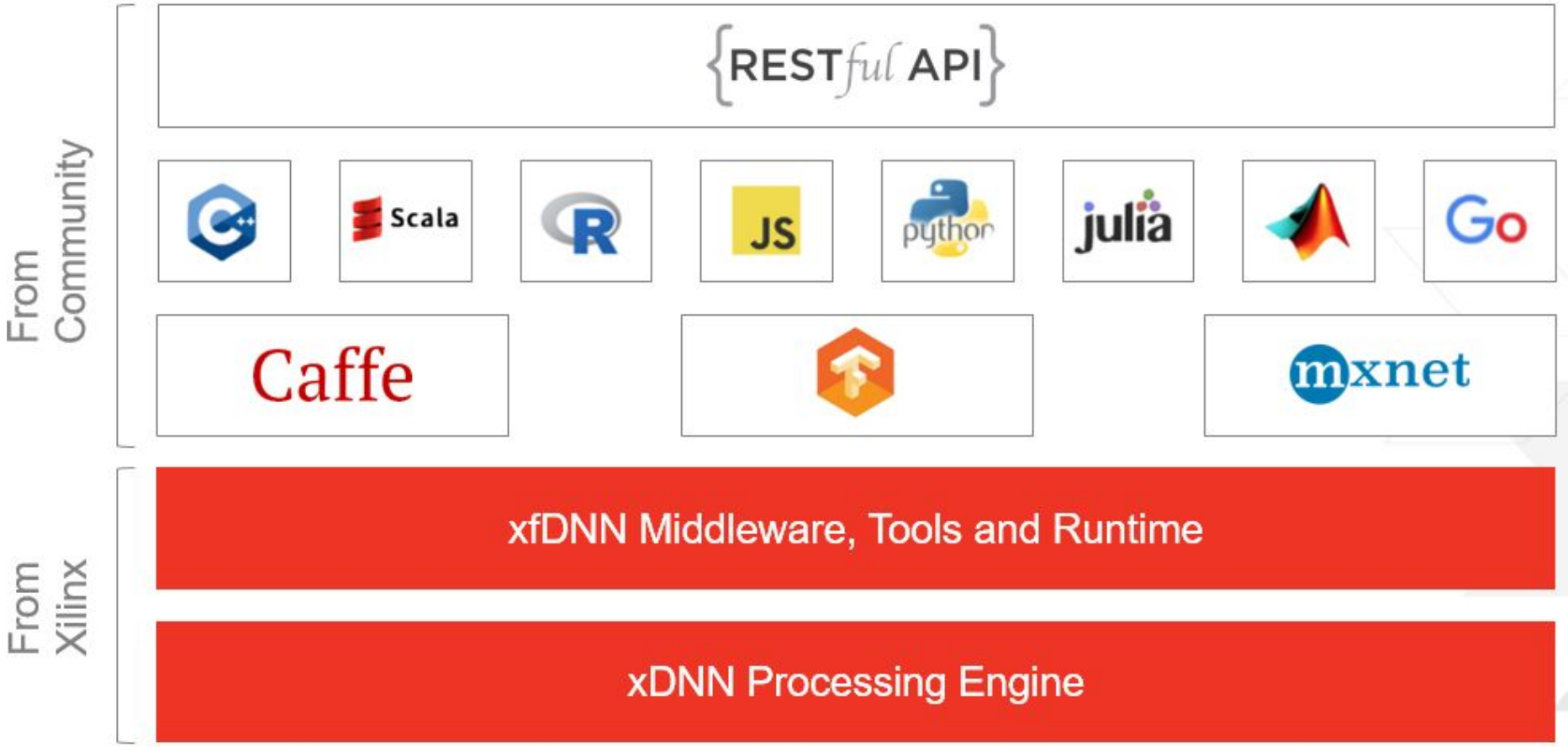
Suprajith Suresh Hakathur

Highlights:

- ML Suite delivers low-latency, high-throughput, and power-efficient machine learning inference for real world applications.
- 2018 Vision Product of the Year award for the best cloud technology at the Embedded Vision Summit.

ML Suite has:

- **xDNN IP** - High Performance general CNN processing engine.
- **xfDNN Middleware** - Software Library and Tools to Interface with ML Frameworks and optimize them for Real-time Inference.
  - xfDNN Compiler
  - xfDNN Quantizer
- **ML Framework** - Caffe, Tensorflow.


From Community
RESTful API
C++
Scala
R
JS
python
julia
Go
Caffe
mxnet
From Xilinx
xfDNN Middleware, Tools and Runtime
xDNN Processing Engine

ML Framework

ML Suite Supports following frameworks:

- Caffe
- Tensorflow
- Keras
- MXNet
- DarkNet

xfDNN Middleware

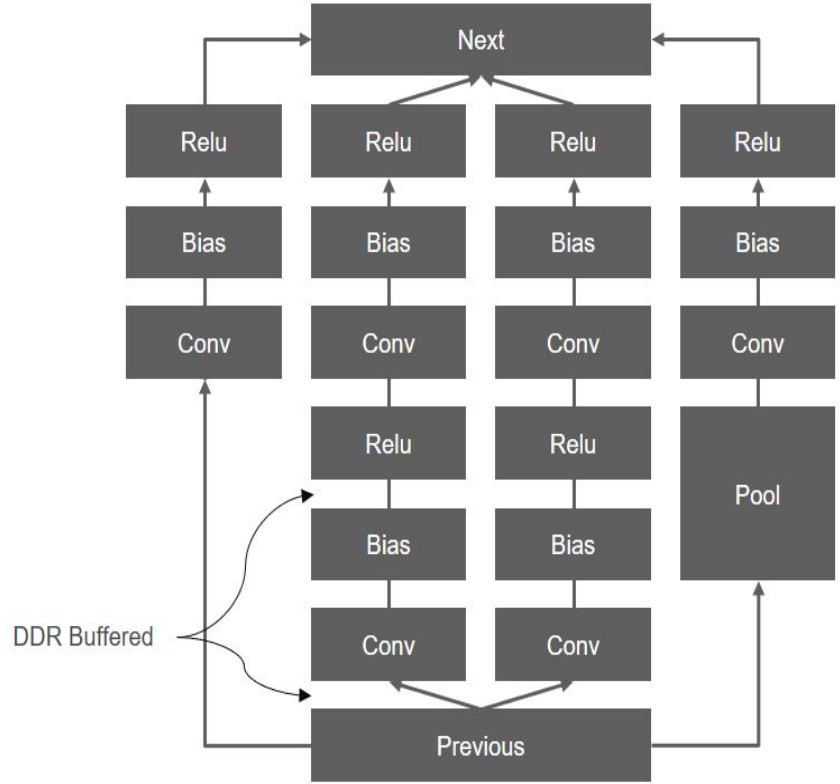
- High-performance software library with a well-defined API.
- Acts as a bridge between deep learning frameworks (Caffe, TF) and **xDNN IP** running on an FPGA.
- It requires a system running SDAccel reconfigurable acceleration stack compliant system.
  - SDAccel environment provides a compiler, a debugger and a profiler.
  - Supports standard OpenCL APIs.
- xfDNN Quantizer enables fast, high-precision calibration to lower precision deployments to INT8 and INT16.

xfDNN Middleware

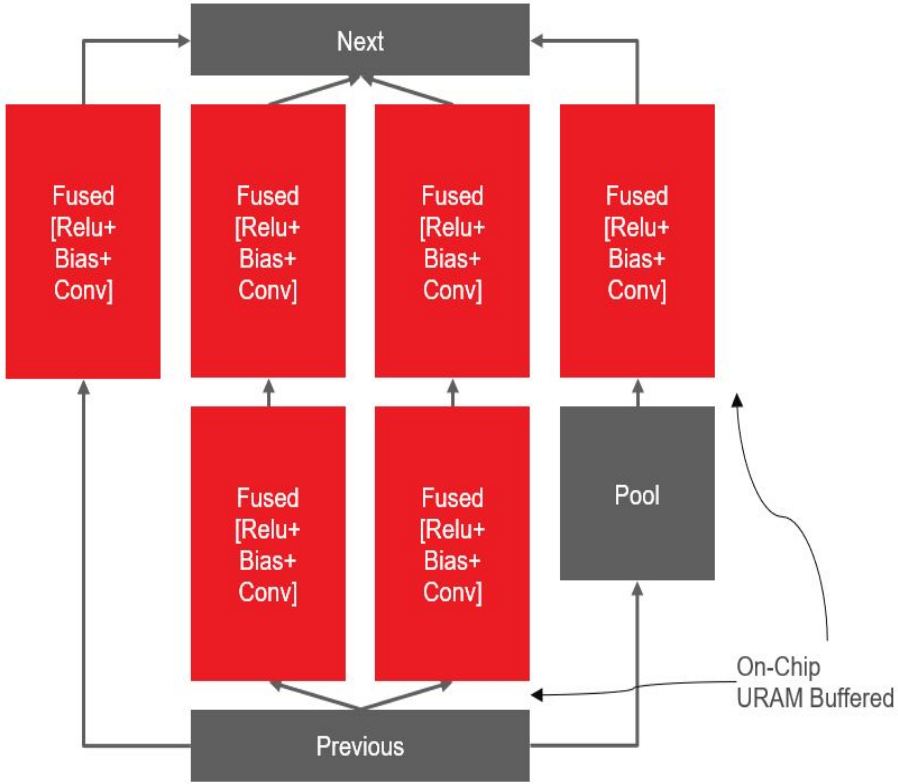
- Provides tools for network optimization by
  - Fusing layers
  - Optimizing memory dependencies in the network
  - Pre-scheduling the entire network removing CPU host control bottlenecks

xfDNN Compiler

- Compiler interfaces with ML Frameworks to read deep learning networks (Graph).
- Cleanup to produce unified dataflow graph.
- Basic optimizations, node merging, memory optimization.
- DDR static vs dynamic scheduling.
- Partitioning, Parallelism.



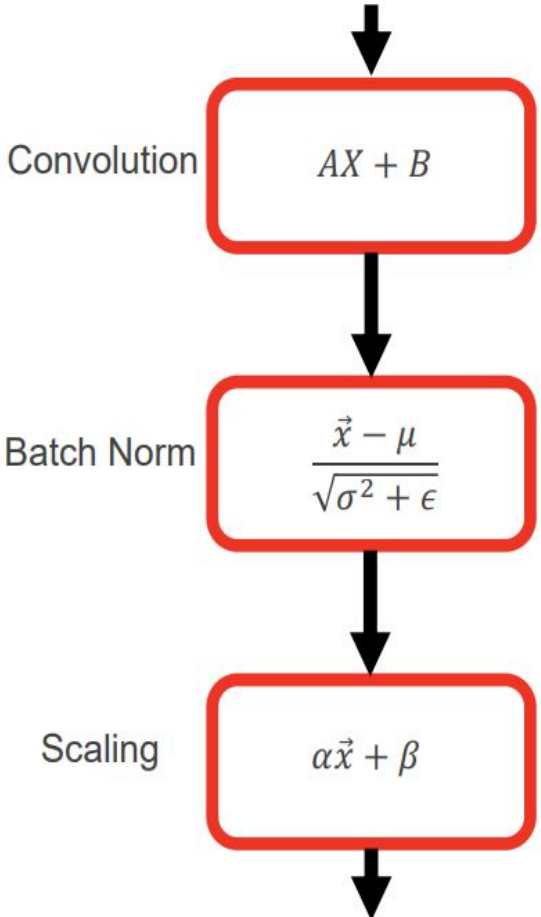
Unoptimized Model

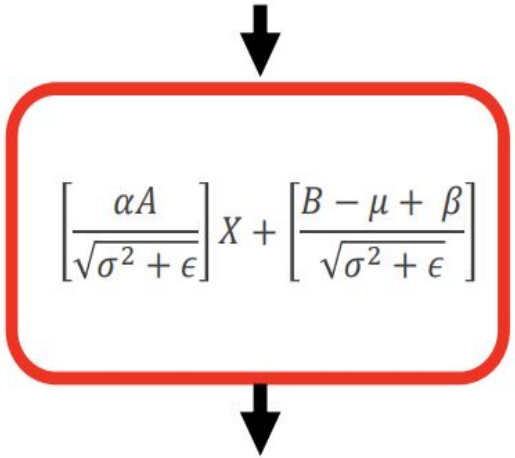


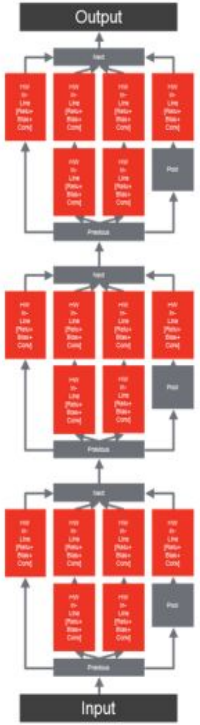
xfDNN Intelligently Fused layers
Streaming optimized for URAM

Network Optimization by fusing layers

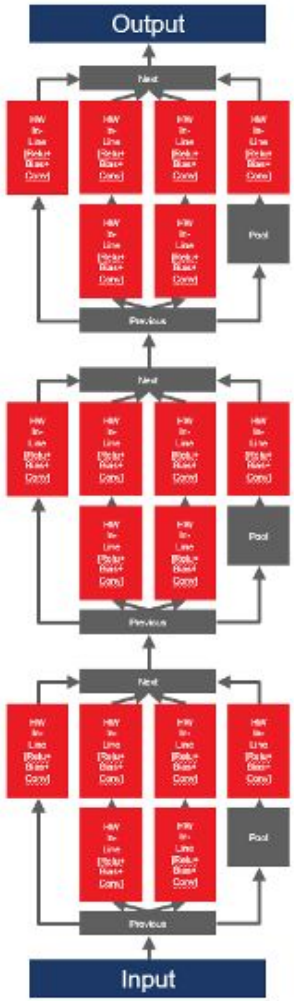
Merging Layers: Convolution + BN + Scaling





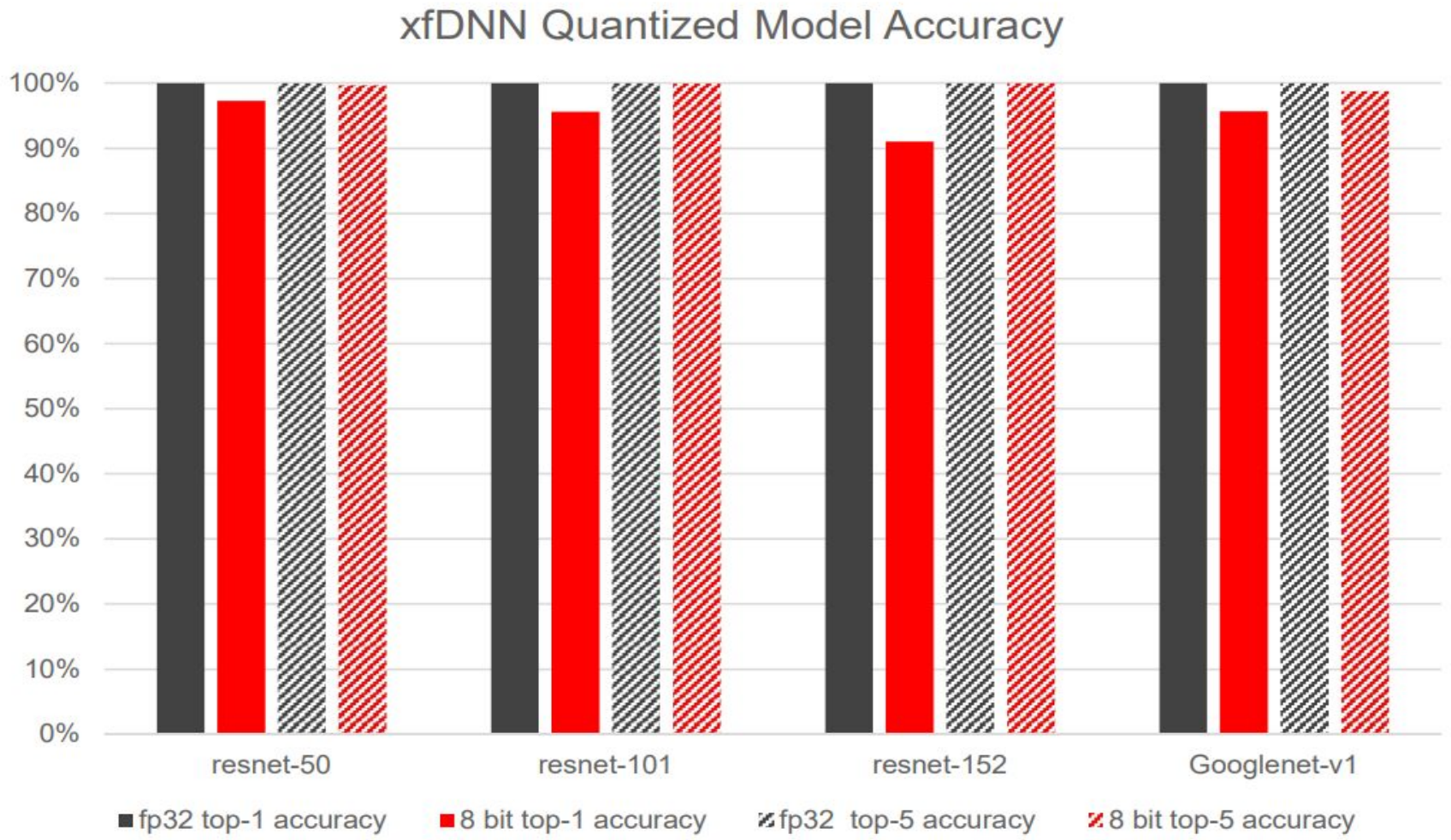


After Completion of optimization per layer, entire network in optimized for deployment in “**One-shot**“ execution flow.



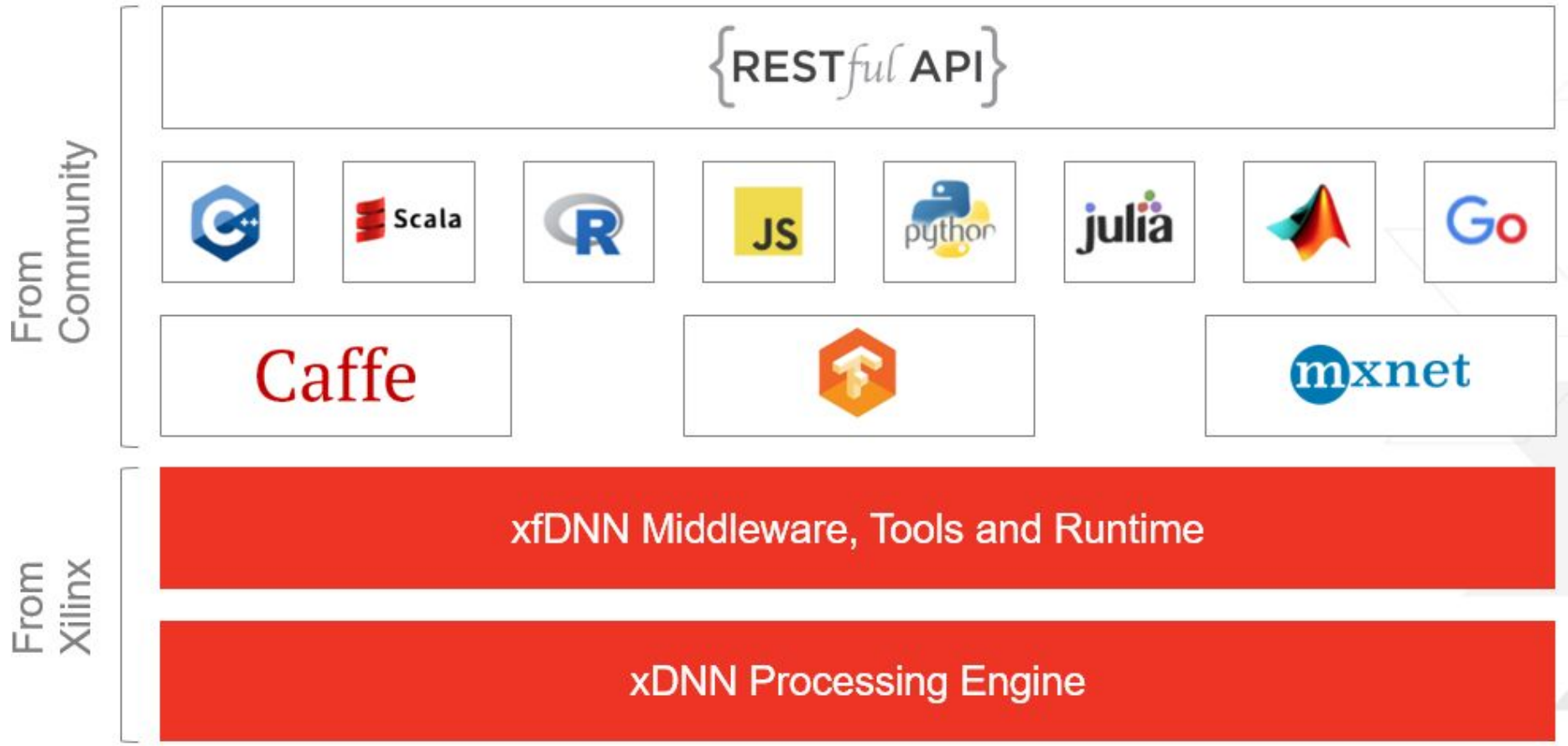
xfDNN Quantizer

- Performs a technique of quantization known as recalibration.
- Allows you to maintain the accuracy of the high precision model.
- It calculates the dynamic range of the model and produces scaling parameters recorded in a json file.
- Does not require full retraining of the model.
- These parameters are used by the xDNN overlay during execution of the network/model.


xfDNN Quantized Model Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
resnet-50
resnet-101
resnet-152
Googlenet-v1
fp32 top-1 accuracy
8 bit top-1 accuracy
fp32 top-5 accuracy
8 bit top-5 accuracy

xDNN IP

- xDNN IP cores are high performance general CNN processing engines.
- Accepts a wide range of CNN networks and models.
- There are two configurations available (28x32 and 56x32 DSP Array) .
- The 28x32 configuration, also referred to as medium, is optimized for higher throughput.
- The 56x32 kernel is optimized for larger models and delivers lower latency.


From Community
{RESTful API}
Scala
JS
python
julia
Go
Caffe
mxnet
From Xilinx
xfDNN Middleware, Tools and Runtime
xDNN Processing Engine

Adaptable

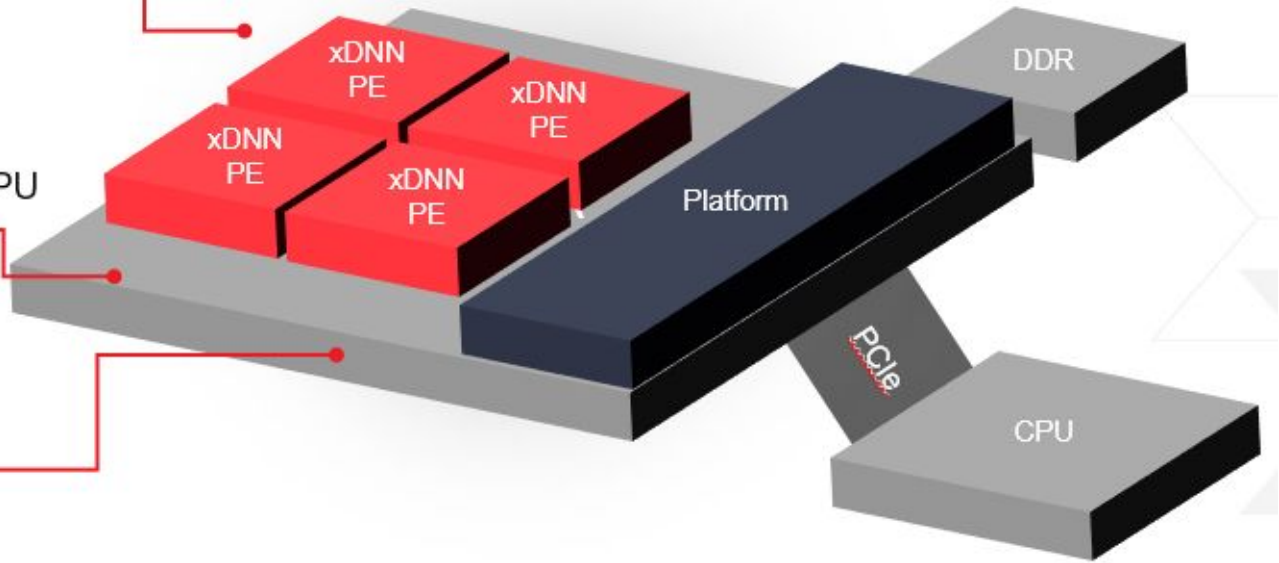
> AI algorithms are changing rapidly
> Adjacent acceleration opportunities

Realtime

> 10x Low latency than CPU and GPU
> Data flow processing

Efficient

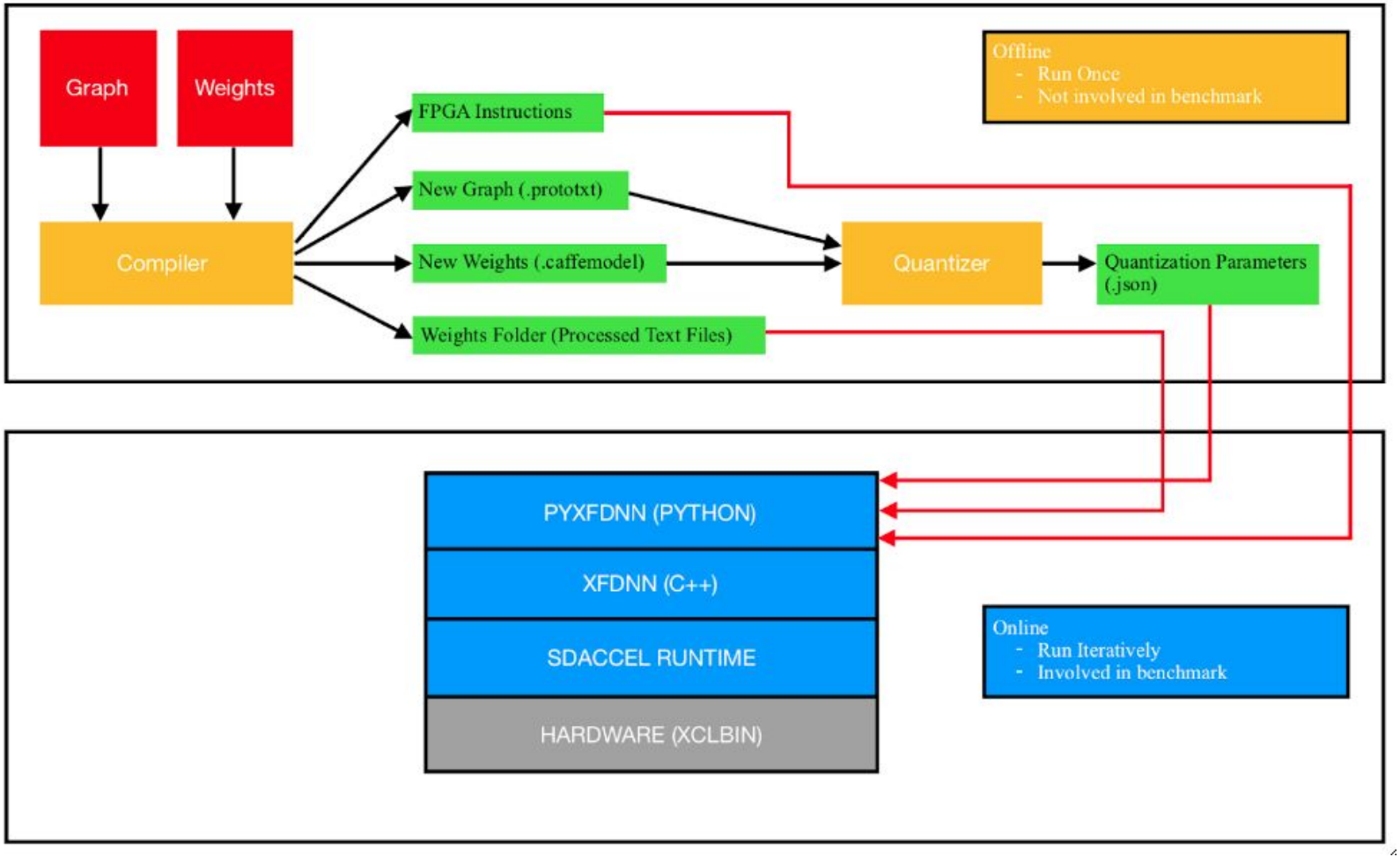
> Performance/watt
> Low Power



xDNN Overlay

- xDNN provides Overlay to combine multiple xDNN IP kernels.
- An overlay is an FPGA binary with multiple xDNN IP kernels.
- It helps in necessary connectivity for on board DDR channels.

Xilinx Caffe Flow



- The final layers of the network (Fully connected, Softmax) are run on the CPU, as those layers are not supported by the FPGA

Thank You.