SIMD and Weight-Output reconfigurable 2D systolic array-based AI accelerator

Vanilla VGG16 on 8x8 Systolic Array

Architecture:

- 8x8 2D Systol ic Array
- Fully integrated with L0, OFIFO, and SFU

Model:

- Mapped VGG16 Layer-27 (Squeezed 8x8) w/o Bat chNorm
- Also Mapped Resnet for the test of Hadamard
- 2 bit / 4 bit weight and activation

VGG16 4bit without BN	92%
VGG16 4bit with BN(for weight combine)	92%
VGG16 2bit	89%
Resnet 4bit(for hadamard)	91%

```
...
Epoch [48/50] (LR: 0.00010) -> Test Accuracy: 91.79%
Epoch [49/50] (LR: 0.00010) -> Test Accuracy: 91.76%
Epoch [50/50] (LR: 0.00010) -> Test Accuracy: 91.62%
```

```
Epoch [500/500] Acc: 88.39%
Training Finished. Best Acc: 89.03%
```

FPGA Mapping Results(cyclonelV GX):

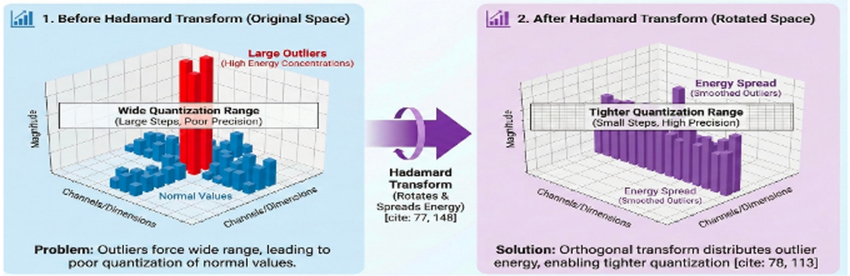
MAX Frequency	113.43MHz
Total Power	520.58mW
Efficiency	27.89GOPS/W

Power Analyzer Status	Successful - Mon Dec 1 00:03:43 2025
Quartus Prime Version	25.1std.0 Build 1129 10/21/2025 SC Lite Edition
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Power Models	Final
Total Thermal Power Dissipation	520.58 mW

	Fmax	Restricted Fmax	Clock Name	Note
1	113.43 MHz	113.43 MHz	clk	

Alpha 1:hadamard

- Hadamard is an orthogonal, lossless transform that redistributes activation/weight energy.
- By smoothing outliers, it can reduce quantization range and error in PTQ.
- With proper folding into weights, it introduces zero extra runtime cost
- On small networks (ResNet-20), improvements are modest and somewhat unstable, but the technique is promising for large models and LLMs where outliers dominate quantization behavior.



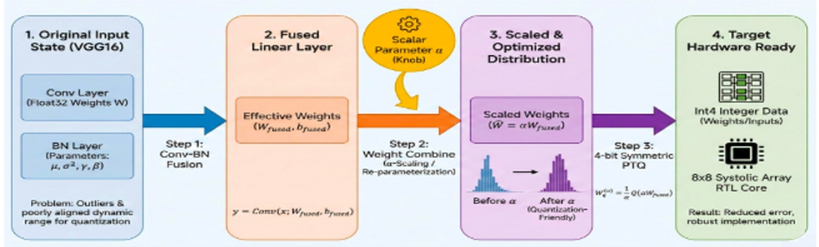

```
========== Hadamard vs Baseline on one Conv layer ==========
Layer: Conv2d(16, 16, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bia
s=False)
4-bit baseline MSE       : 6.491477e-03
4-bit Hadamard-domain MSE: 5.330979e-03
MSE ratio (Had / base)     : 0.821
============================================================
========== Hadamard vs Baseline on one Conv layer ==========
Layer: Conv2d(16, 16, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bia
s=False)
4-bit baseline MSE       : 5.493672e-03
4-bit Hadamard-domain MSE: 3.752234e-03
MSE ratio (Had / base)     : 0.683
============================================================
```

Alpha 2:weight combine and α-Scaling

- Weight combine uses a scalar α to re-parameterize Conv+BN weights and make them more quantization-friendly in VGG16.
- We fuse Conv+BN first, then apply α-scaling and 4-bit quantization on the fused kernel of the 8×8 target layer.

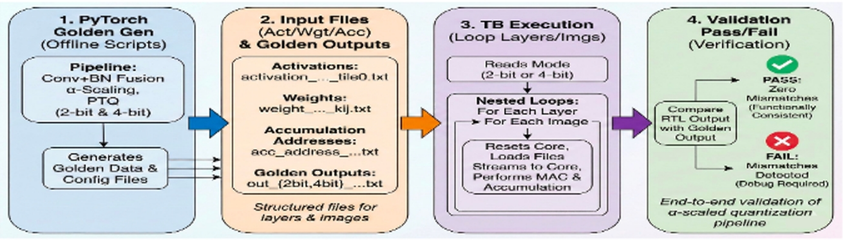
The method is:

– Easy to integrate with our 8×8 systolic hardware flow,
– Complementary to Hadamard-based outlier smoothing and other PTQ techniques.
– the combination of Conv and BN can reduce the time of moving data in hardware



Alpha 3:multi input and multi layer in testbench

- Multi-layer, multi-image testbench for the 8×8 systolic core.
- Supports both 2-bit and 4-bit modes through a single mode signal controlled by inst w.



Alpha4:2-bit and 4-bit reconfigurable with both Weight Stationary and Output Stationary

- Support calculating 2 partial sums of adjacent channels simultaneously in a single PE at OS mode for 2-bit activation.
- Reused mac unit in each PE for both WS and OS partial sum calculation, saving DSP resource compared to duplicated version.
- Improved port reuse. Both out_s and in_n are used for weight loading, partial sum loading and partial sum output.

Alpha5:Gating for inputs sampling

-Avoid register value(s) toggling when input(s) are zero value.
-Energy saving for model of high sparsity (e.g. pruned model).