Weight and Output Stationary Reconfigurable 2D Systolic Array-Based AI Accelerator and Mapping on Cyclone IV GX

Team 2

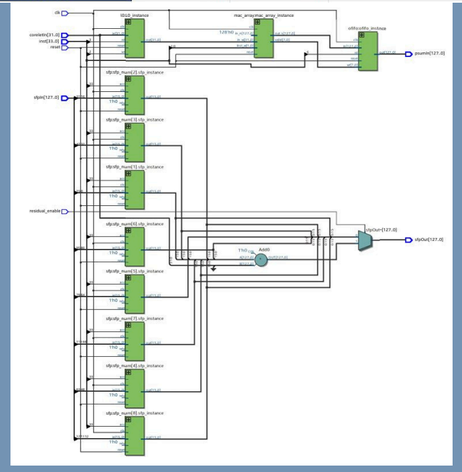
Jason Li, Benjamin Nguyen, Christopher Hughes, Chihao Yu, Lingxiao Li

Motivation

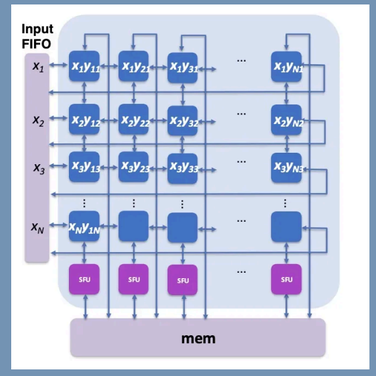
As Convolutional Neural Networks (CNNs) excel in computer vision, there's a race to improve their training efficiency. Nowadays, people are increasingly concerned about the high power consumption of AI chips, which not only raises costs but also impacts the environment. Our goal is to design a low-power, efficient CNN accelerator chip to address these issues, reducing both the cost and carbon footprint of AI training.

Mapping on FPGA Cyclone IV GX

Metric	Cyclone IV	MAX10	MAX10+residual+clkgating
Fmax (MHz)	130	110	112
TOPS/s	0.01664	0.01408	0.014336
GOPs/s	16.64	14.08	14.33
TOPS/W	0.049	0.106	0.107
Total Logic Element	17063	16996	16764
Total Registers	12098	12098	11978
Core Dynamic Power(mW)	32.11	24.34	26.15
Total Power(mW)	335	132	134



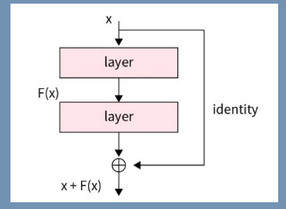
2D Systolic Array

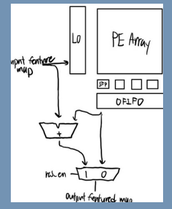


Alpha 1: FPGA Optimization

- Mapped corelet on Intel MAX 10 (10M50DCF672C8G) FPGA, reducing dynamic power consumption from 32 mW (Cyclone IV EP4CGX150) to 24 mW.
- Achieved 25% power savings without changing code, demonstrating that a more cost-effective FPGA can enhance performance.

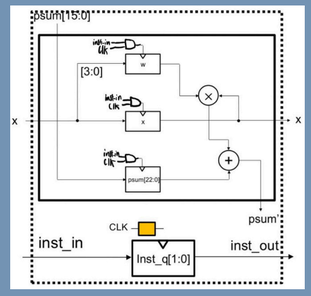
Alpha 2: Residual Connections





- Integrated residual connections into the corelet hardware, enabling support for ResNet models in addition to VGG.
- Addresses vanishing gradient problem with skip connections, improving gradient flow through deep networks.
- Enhances hardware versatility by enabling a seamless transition between AI model architectures

Alpha 3: Clock Gating



- Implemented clock gating for the MAC tile, activating the clock only during kernel loading or execution.
- Reduces power consumption by disabling registers when not in use.

Alpha 4: Pruning for Efficiency

- Planned unstructured pruning of the VGG model, targeting 80% sparsity.
- This will significantly lower toggle rates and power consumption, with clock gating activated for zero weights, eliminating unnecessary latching