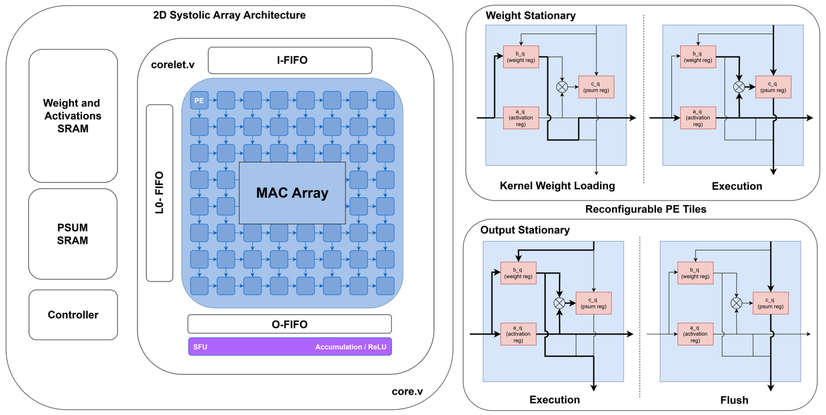
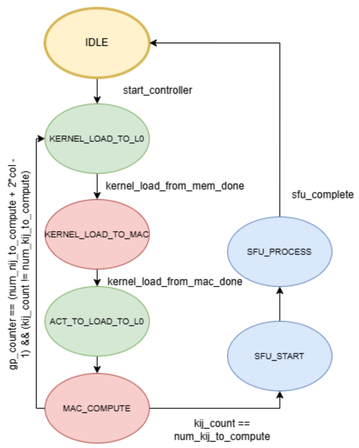
Reconfigurable Weight and Output Stationary SIMD 2D Systolic Array AI Accelerator on Cyclone IV GX

Aastha Shah, Ali Alabiad, Mohamed Ibrahim, Nathan Chao, Soham Karkhanis, Venkateshwaran Sivaramakrishnan

Core Architecture

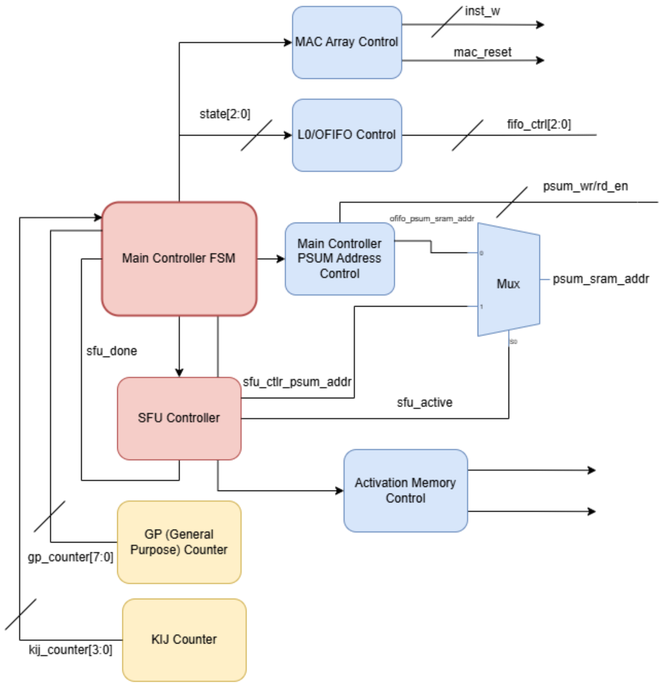


Alpha 1: Weight Stationary Controller

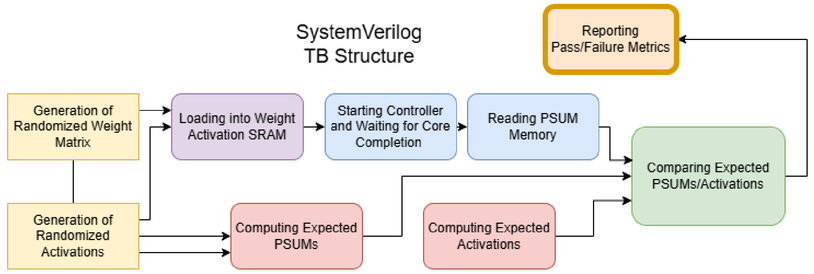


Function of the Controller:

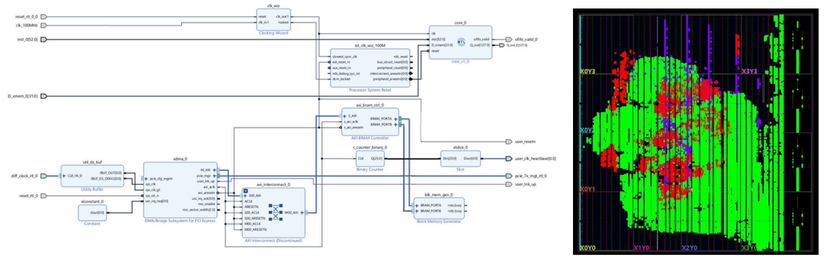
- Performs automated multiplication through pulsing controller start signal.
- Performs automated loading from SRAM into L0 and L0 into MAC.
- Performs automated storage of OFIFO entries into PSUM by controlling PSUM memory signals.
- Features pipelined SFU controller to perform accumulation of PSUM values in a fully automated fashion.



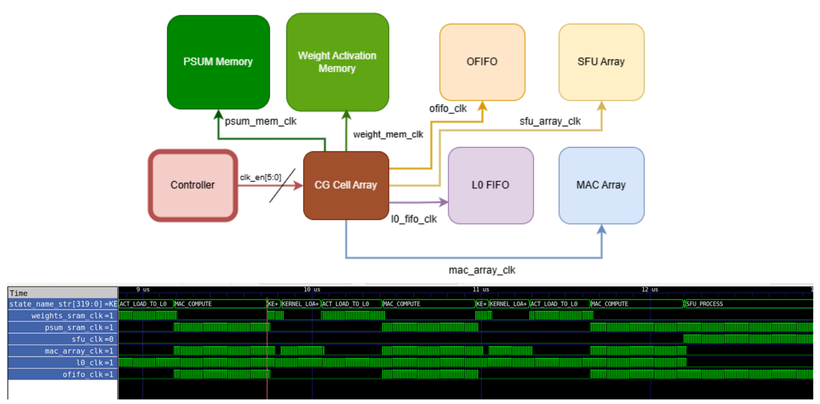
Alpha 2: Enhanced Verification



Alpha 3: PCIe IP Integration

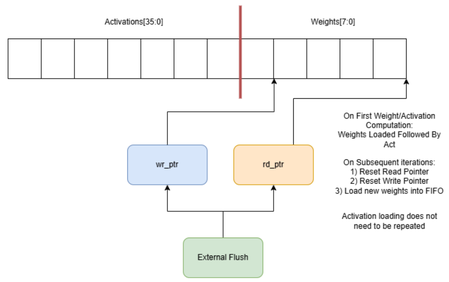


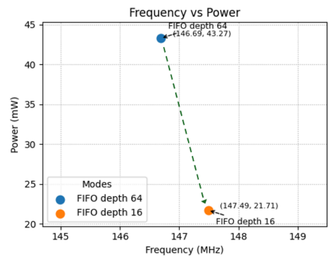
Alpha 4: Power Saving - Clock and Data Gating



Work In Progress Hardware Alphas:

1. **Run Length Encoding** - Decoder Module
2. **Latency reduction techniques:**
   - Modified FIFO loading
   - Filtering of useful PSUM entries
   - Optimizing FIFO lengths to reduce power consumption and optimize Fmax
3. **ResNet** hardware implementation





Software Alpha 1: Hybrid Scheduler

Hybrid Scheduler consists of an increasing linear rate scheduler followed by Cosine Annealing. The increase in learning rate allows for rapid learning and the Cosine Annealing helps the model avoid local minimas leading to better generalizations. This was used as part of the baseline model.

Software Alpha 2: 4-2bit model compression

Retraining a model with 4-bit activations on 2-bit quantization provides a more accurate model than training the model from scratch. This was used as part of the baseline model.

Software Alpha 3: Orchid - Optimizer using Orthogonalization

For weight matrices of two or more dimensions, we orthogonalize the momentum matrix using Newton Schulz approximation allowing for a more stable, better convergent training while decreasing the memory needed to update parameters. This led to better performances during pruning.

Software Alpha 4: Mixed Pruning

We utilized a mixture of structured and unstructured pruning on select layers achieving up to 70% sparsity while preserving the accuracy.

Software Alpha 5: Quantization using Rotation Matrices

We utilize Learned Step Quantization and Hadamard Matrices on the activations to preserve outliers in the data. Currently functional but requires improvements.

Model Sparsity and Performance

Model	Avg. Sparsity (%)	Test Acc. (%)
4-bit (Base)	23.36	92.31
4-bit (Base w/out hybrid sched)	35.4	86.58
4-bit (w/Orchid and hybrid sched)	42.35	94.15
4-bit (Pruned) (w/Orchid and hybrid sched)	70.01	90.81
4-bit (Pruned) (w/SGD and hybrid sched)	50.00	91.76
2-bit (Model Compression)	26.50	89.92
2-bit (Pretrain)	38.73	84.92