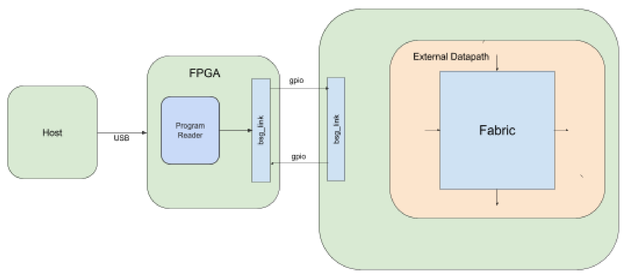
RTA : A Dual-Mode Fabric for Systolic and Reconfigurable Spatial Execution

STUDENTS: Ayush Kulkarni, Eric Liu, Fergus Xu, Rampranav Navendran, Sheehan Ganguly, Shwet Chitnis

Abstract

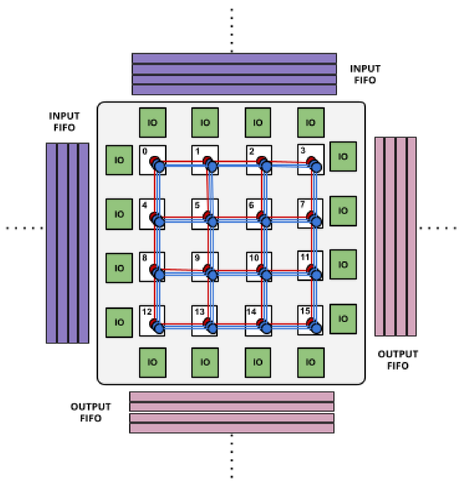
- RTA Is a reconfigurable tensor fabric that combines Google TPU MXU-style systolic execution for dense GEMM with CGRA-style spatial dataflow for non-linear kernels softmax, reductions, normalization, activations and elementwise transforms.
- By mapping both execution styles onto a shared 2D processing-element array, RTA reduces redundant hardware and intermediate data-movement, enabling a more area-efficient and flexible substrate for end-to-end ML acceleration.

System Overview

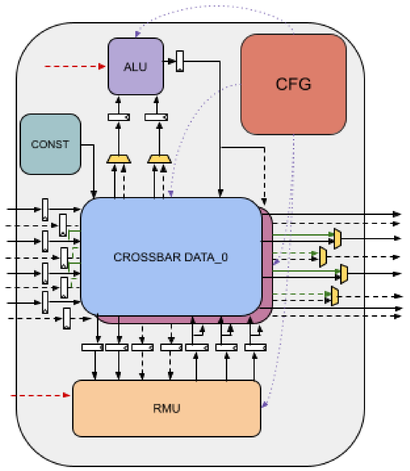


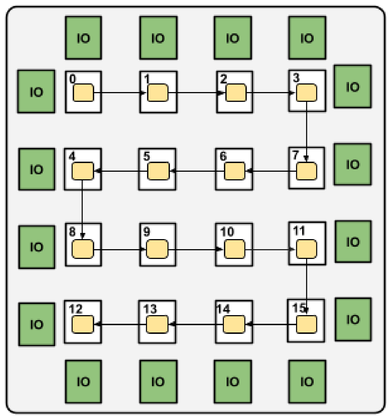
Core Fabric

- **4x4 reconfigurable dual-mode** compute fabric with statically configured crossbars.
- **Dual 8-bit planes** for flexible token delivery,
- **Singe 1-bit** predicate plane.
- **2048-bit** bitstream size.
- **50 MHz clock frequency.**
- **Signed INT4 Systolic Mode.** Inner-product based **Dot-4** output stationary accumulation.
- **Signed INT8 CGRA Mode.** Supports **ADD, SHIFT, MULTIPLY,** and **BOOLEAN** operation groups.



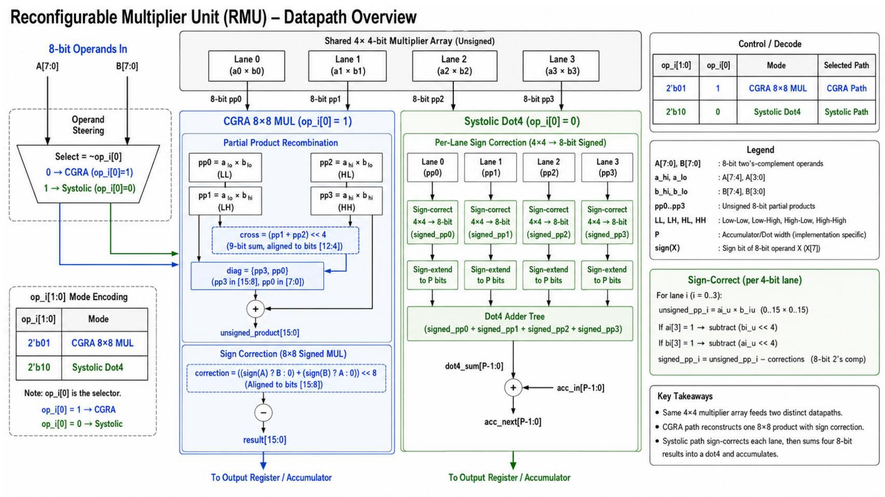
Processing Element & Configuration Protocol





Reconfigurable Multiply Unit

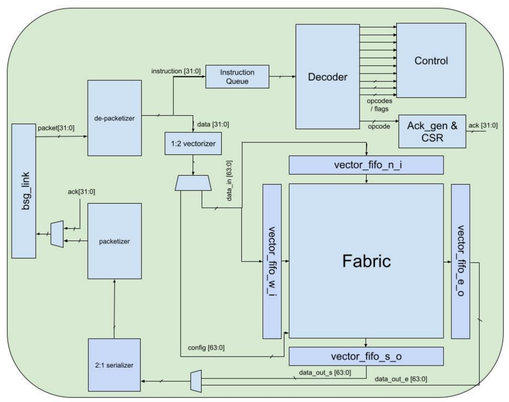
- A **hybrid** multiply unit that performs a **4-way inner-product accumulate** or a s**igned INT8 multiply** with a **16-bit** accumulation.



Chip Logical Protocol

- **ISA:** lightweight instruction set for minimal control logic overhead outside of core
- **Packet Format**: header-driven packet network, dynamic payload length, wormhole routing
- **Handshaking**: mix of ready/valid and valid/yumi with backpressure-based flow control
- **FSMs**: For input packet routing, vectorizing data, output data packetizing

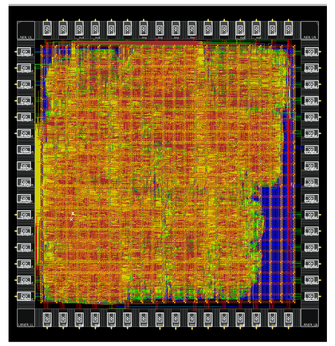
External Datapath Diagram



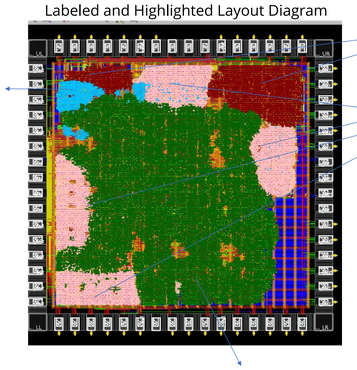
Physical Implementation

- HAMMER → Synthesis → Innovus Place & Route → Calibre DRC/LVS → Tempus STA
- **2.117 mm²** total placed area : **SoC Logic – 1.504 mm²** (71.1%), **pad ring – 0.442 mm²** (20.9%), **bsg_link – 0.168 mm²** (7.9%).
- Within **SoC Logic: CGRA – 1.087 mm²** (72.2%), **Avg PE Area – 62.4K um², FIFO Ring – 0.358 mm²** (23.8%), **External Interface Logic – 49.9K um²** (3.3%).
- **WNS : +1.043 ns** on **50 MHz** clock generated off-chip (from the FPGA), hold violations fixed.

Layout Diagram



Labeled and Highlighted Layout Diagram



ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISERS: Professor Ang Li
TA: Jiayi Wang, Jingqun Zhang
SPONSORS: TSMC, Apple