

W Precision-Aware Communication in CGRAs

UNIVERSITY of WASHINGTON

Shwet Chitnis*, Fergus Xu*, Ayush Kulkarni*, Jiayi Wang*, Jingqun Zhang*, Arjun Raje†, Ang Li*

*University of Washington, †Carnegie Mellon University

Abstract

Mixed-precision machine-learning kernels increasingly stress CGRA interconnects because communicated values span byte- to word-scale precisions, while many CGRAs route fixed-width tokens under a uniform abstraction. This mismatch imposes a precision tax: narrow operands underutilize wide links, while wide values on narrow fabrics require fragmentation across multiple transfers. We present Fringe, a precision-aware CGRA communication substrate that treats precision as a first-class routing resource. Fringe exposes compiler-scheduled routing planes at different widths (8b and 32b data paths plus a 1b predicate path), enabling the mapper to allocate dependences onto precision-matched physical links without dynamic packing/unpacking hardware or loss of compile-time determinism.

Precision Tax

Fixed-width routing forces 8b values to occupy 32b links, increasing contention and blocking unrelated traffic.

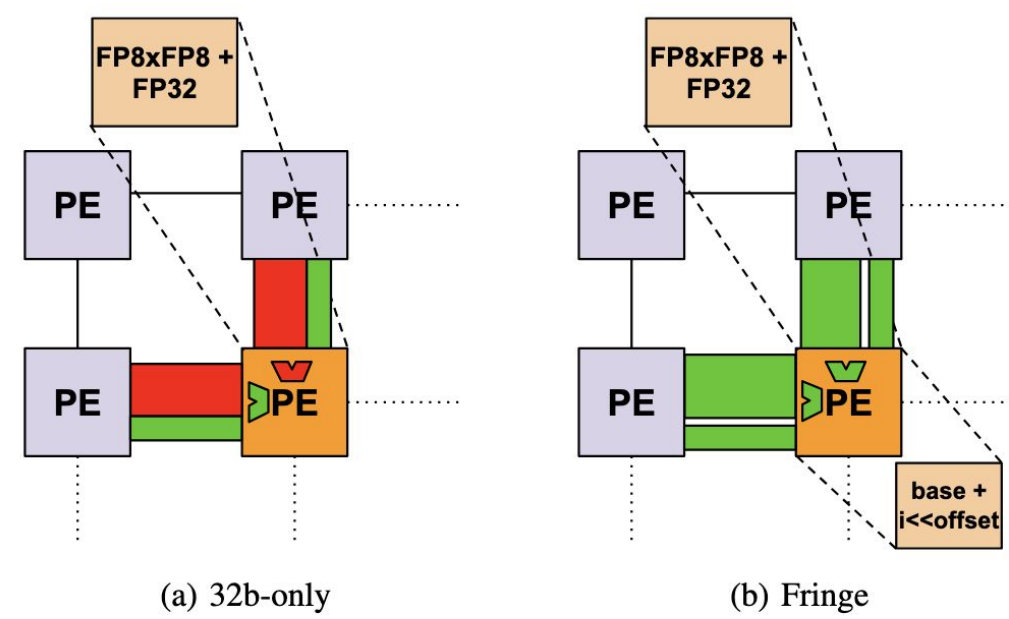
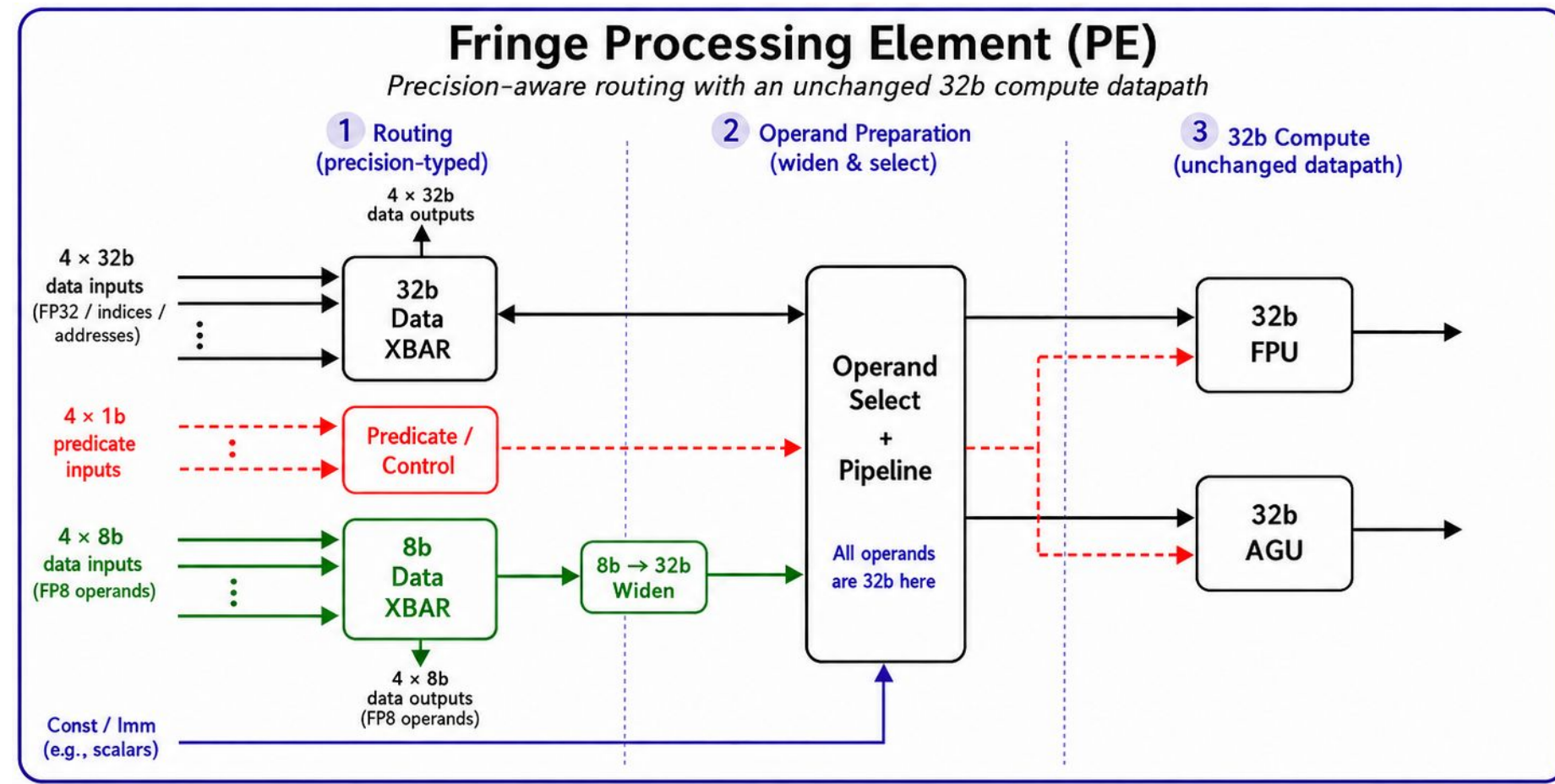
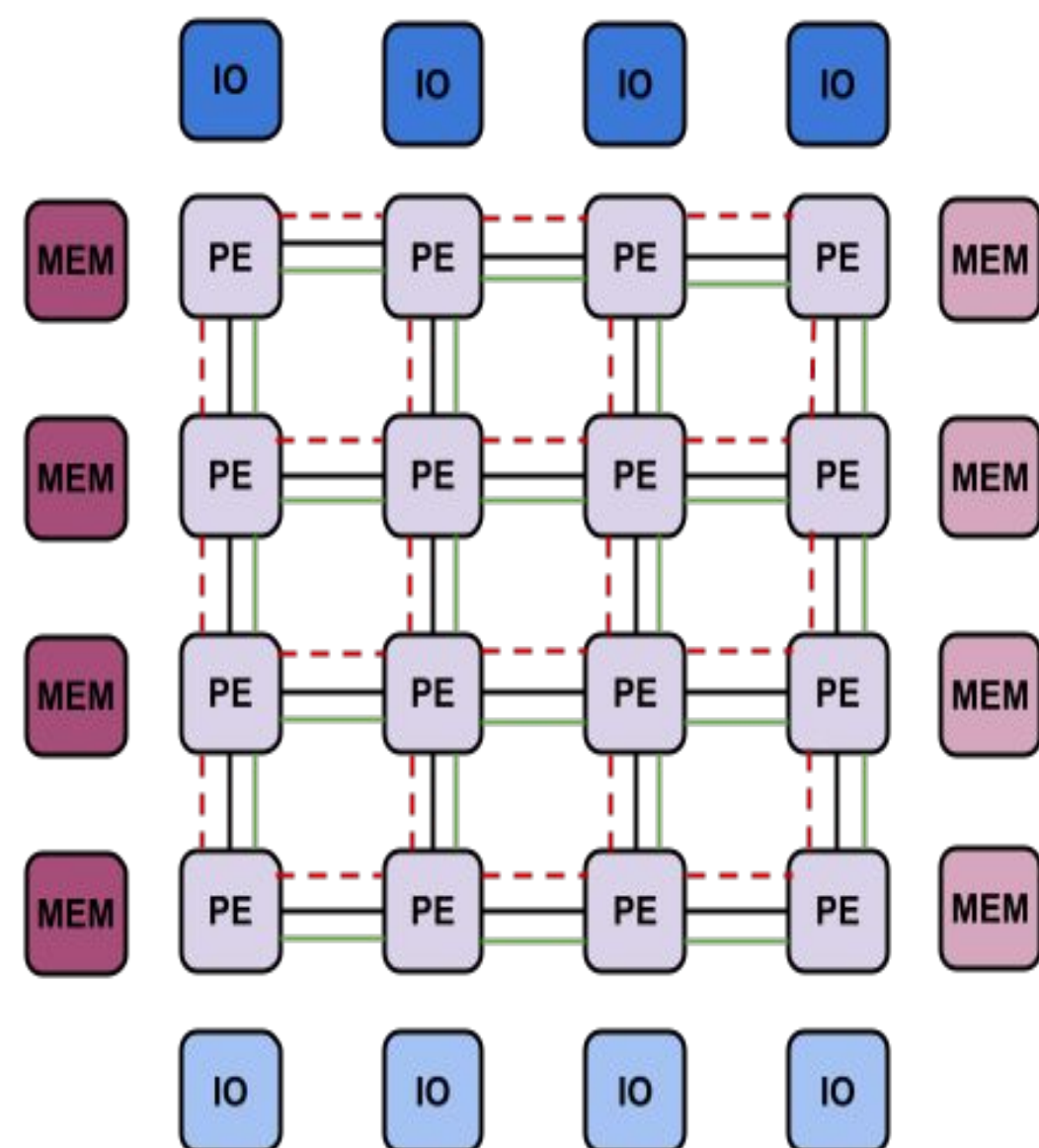
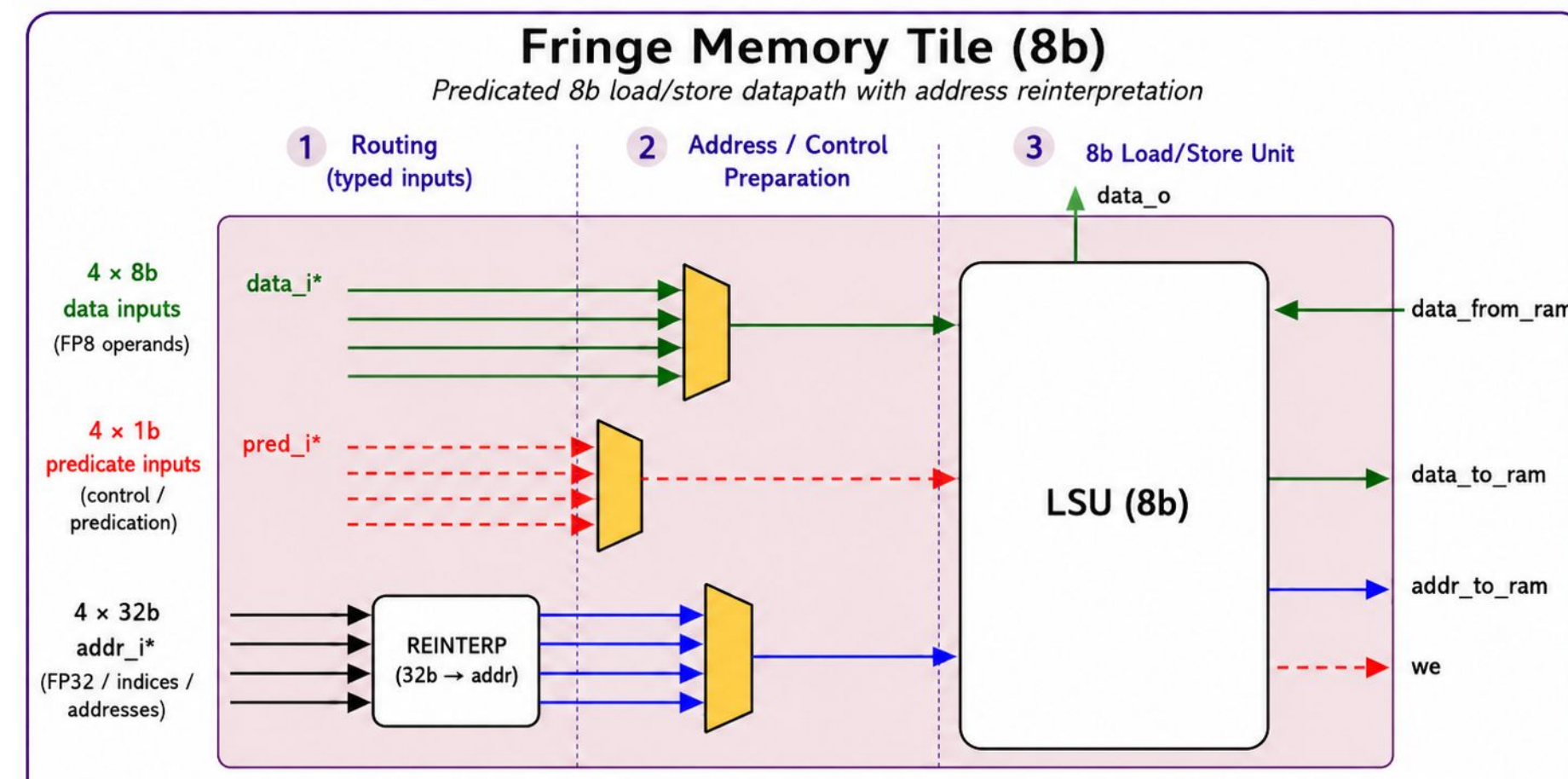


Fig. 1. Fringe routes narrow and wide dependences on separate precision-matched planes.

Fringe

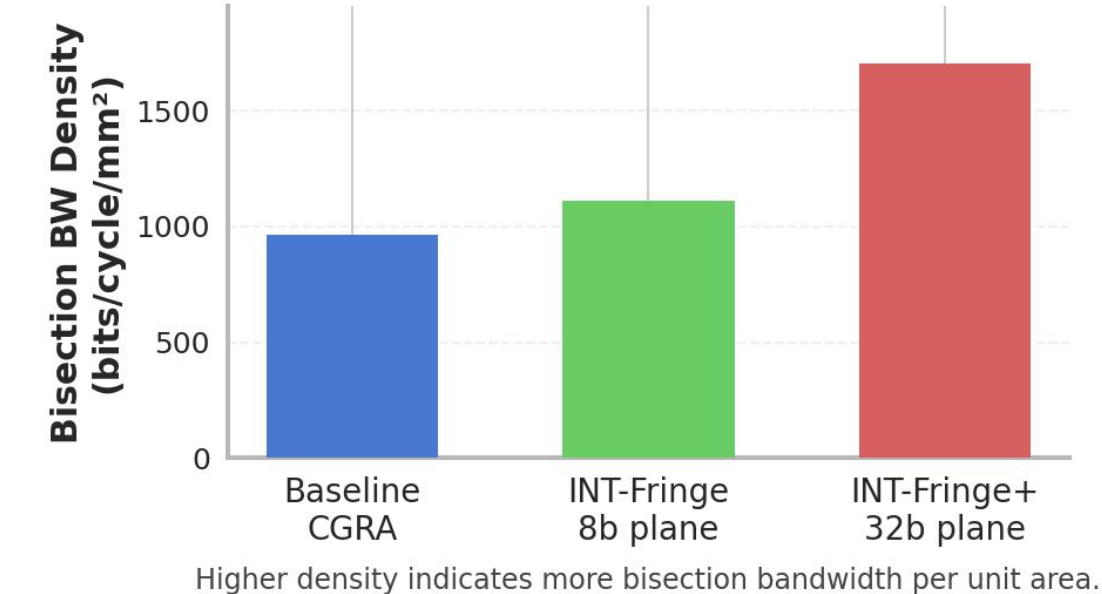


- Separate 32b and 8b crossbars reduce contention and bandwidth waste.
 - FP8 operands stay 8b across the fabric and widen only near the PE.
 - Operand Select + Pipeline feeds both FPU and AGU with 32b operands.
 - Predicate (1b) plane provides control/predication to the compute units.
 - Compute datapath remains 32b and is unchanged.
- Fringe PE tile.** Fringe exposes precision-typed communication paths inside each tile: FP8 operands route through the 8b data plane and widen only near the operand-select stage, while FP32 values, indices, addresses, and accumulators remain on the 32b data plane. The 32b compute datapath is unchanged, and all 32b results share the same 32b output network. 8b outputs (FP8 operands) are produced directly by the 8b XBAR.

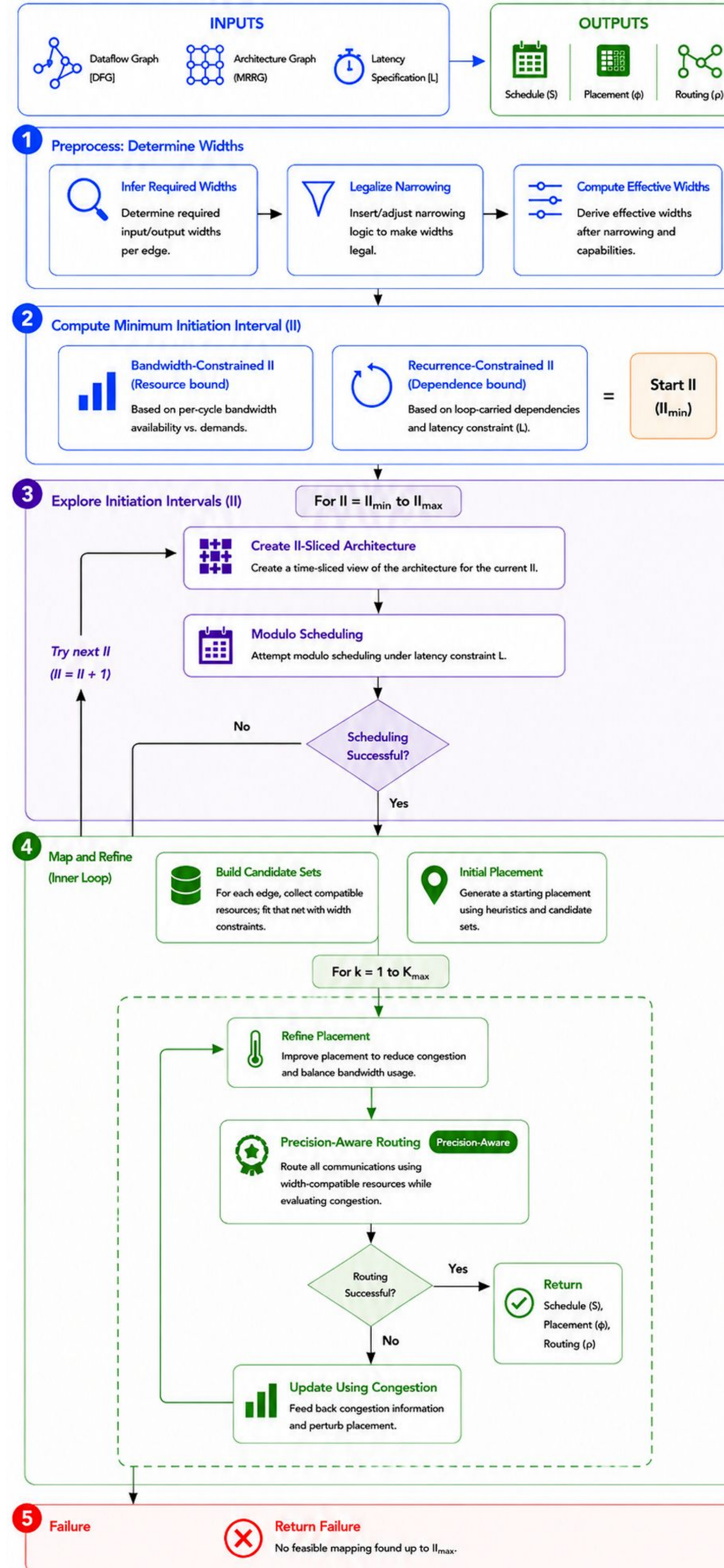


- Separate input selection for data, predicate, and address streams.
 - 32b addresses are reinterpreted before entering the 8b LSU.
 - LSU handles predicated 8b loads and stores.
 - data_from_ram and data_to_ram remain 8b.
 - Write enable is carried on the 1b predicate/control path.
- Fringe MEM tile.** The 8b memory tile accepts 8b data inputs, 1b predicate/control inputs, and 32b address inputs. Addresses are reinterpreted before entering the 8b LSU, which interfaces to memory through data_from_ram, data_to_ram, addr_to_ram, and we, and produces data_o.

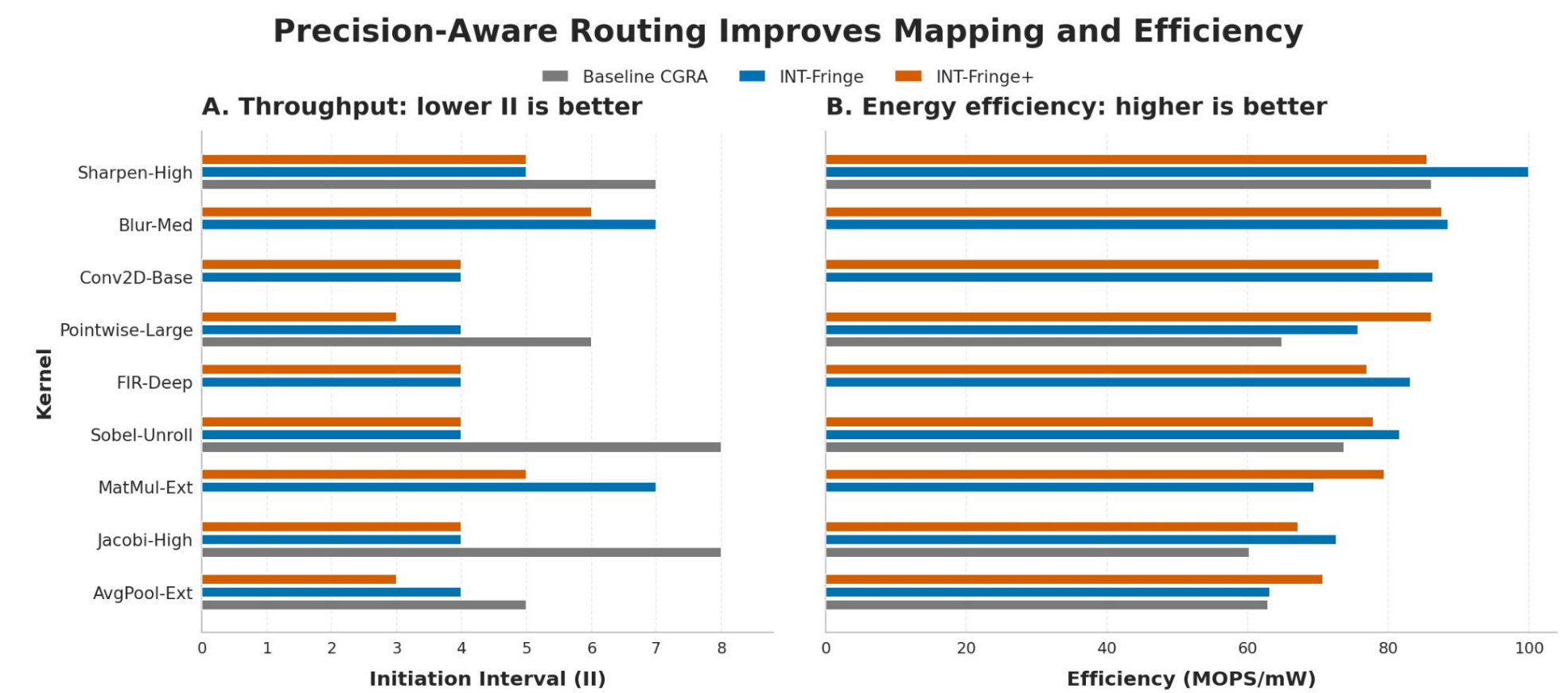
INT-Fringe+ Improves Bisection Bandwidth Density



Precision-Aware Modulo-Scheduled Mapping



Results



Key Takeaways

- Lower cost than duplicating 32b routing**
+8.4% area, +14.4% static power
- Up to 2x higher throughput**
vs. single-network baseline
- Best energy efficiency in 6/9 cases**
on routing-stressed streaming kernels
- Fixed-width CGRA links impose a precision tax**
8b operands consume 32b routes, wasting bandwidth and switching energy.
- Fringe adds a compiler-visible 8b routing plane**
It complements the baseline 32b data network for precision-matched communication.
- Static scheduling is preserved**
No dynamic packing, run-time arbitration, or loss of compile-time determinism.
- Bitwidth-aware mapping is the key enabler**
Precision informs lower bounds, placement, and routing decisions.
- Benefits depend on traffic mix**
Greatest gains appear when dense 8b streaming dominates; recurrence-heavy kernels may prefer an extra 32b plane.

Implemented in TSMC 28 nm at 700 MHz

