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

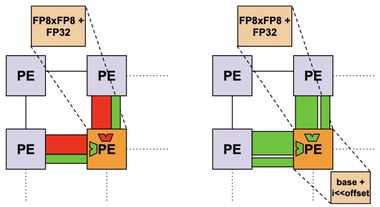
PNCEL

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

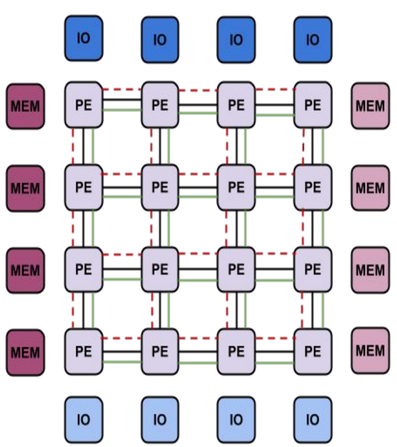
Fixed-width routing forces 8b values to occupy 32b links, increasing contention and blocking unrelated traffic.

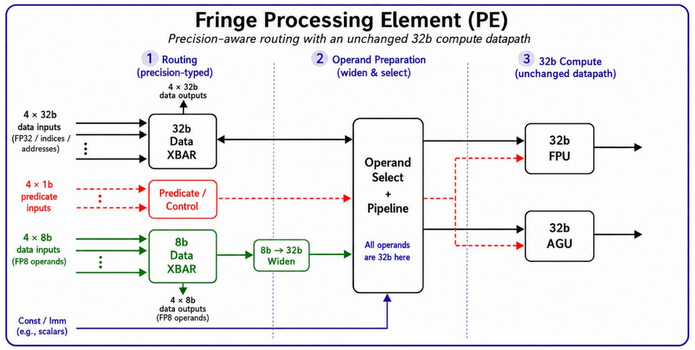


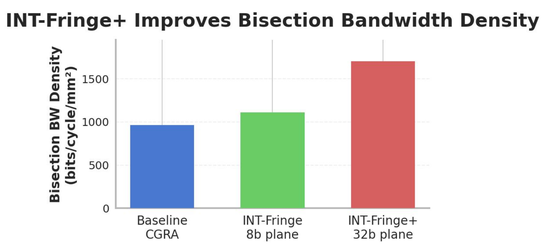
(a) 32b-only (b) Fringe

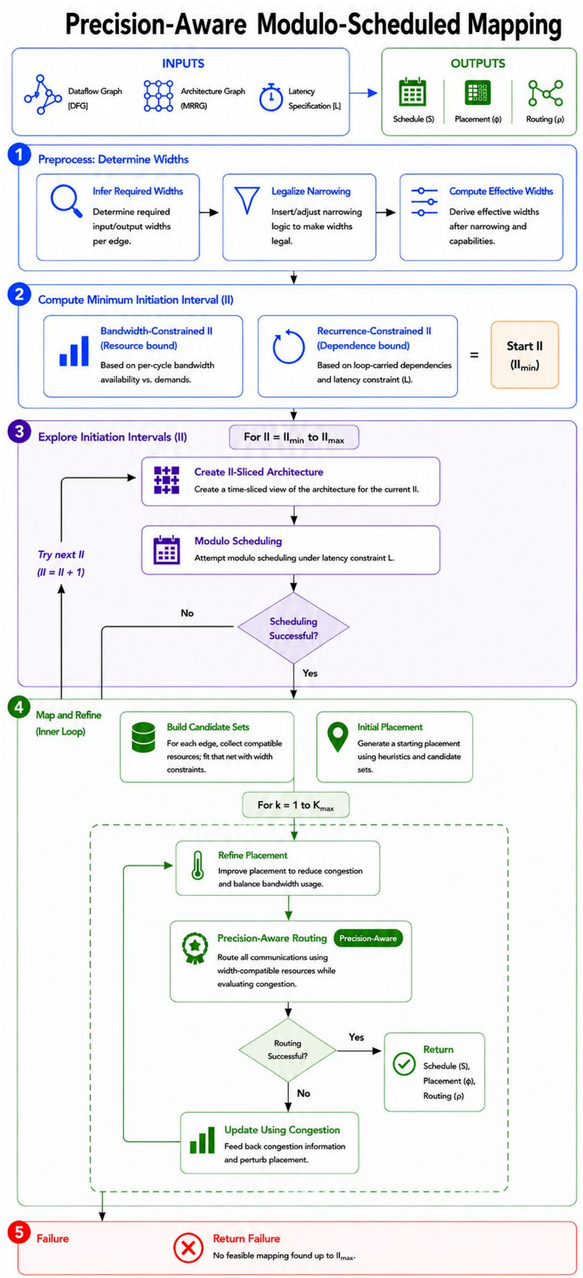
Fig. 1. Fringe routes narrow and wide dependences on separate precision-matched planes.

Fringe

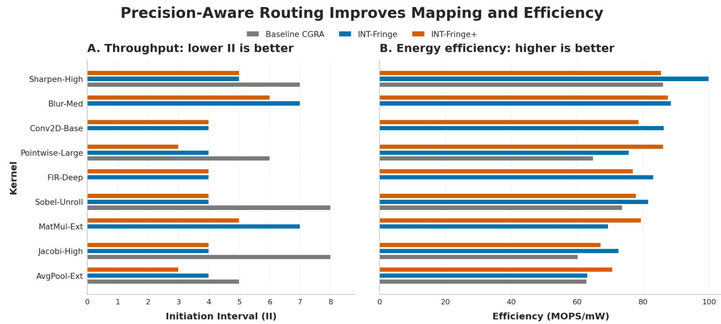








Results



Key Takeaways

Lower cost than duplicating 32b routing
+8.4% area, +14.4% static power

Up to 2× 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