A 65nm Floating-Point Processor Supporting Parallel Digital Sparsity & Intensive In-Memory Computing with Adjustable Precision for Convolutional Neural Network

Group : A513 Member : Pei-En Liu, Yu-Da Lu, Guo-Zhang Liao Advisor : Prof. Meng-Fan (Marvin) Chang

I. ABSTRACT

Floating-point CIM systems must balance full IEEE-754 precision with enough speed, yet existing approaches—wide fixed-point alignment, bank-local FP logic, or exponent/mantissa splits—either introduce serial bottlenecks or fail to fully exploit parallelism. Paper [1] observes that CNN exponents follow a near-Gaussian distribution and routes the dense "intensive" majority through a parallel CIM core (with on-the-fly alignment) and the sparse minority through a digital core with specialized encoding.

Building on strategy in paper [1], we employ five core mechanisms—parallel processing, CIM pipelining, round-robin banking, custom sparse encoding & refill, and adjustable precision—to realize a high-speed, energy-optimized architecture that maintains full IEEE-754 accuracy.

II. RESEARCH METHODOLOGY

(1) Parallel Processing: We align the intensive and sparse computation streams so neither path dominate the other, eliminating critical bottlenecks and ensuring both cores run at full capacity (Fig. 1).
(2) CIM Pipelining: Each MAC unit immediately begins processing the next activation batch as soon as it finishes its current task, keeping all units active throughout convolution and maximizing throughput (Fig. 2).
(3) Round-Robin Banking: Sparse entries are striped across multiple single-port SRAM banks; broadcasting read and write addresses to all banks enables true concurrent access and dramatically reduces fetch latency (Fig. 3).
(4) Custom Sparse Encoding & Refill: Every encoding format embeds bias-lookup indices and reference address directly within each 32-bit entry, removing extra decode stages
(5) Adjustable Precision: By disabling a few least-significant multiplier bits in the intensive path, we shorten multiply cycles and cut energy per layer with negligible impact on overall accuracy.

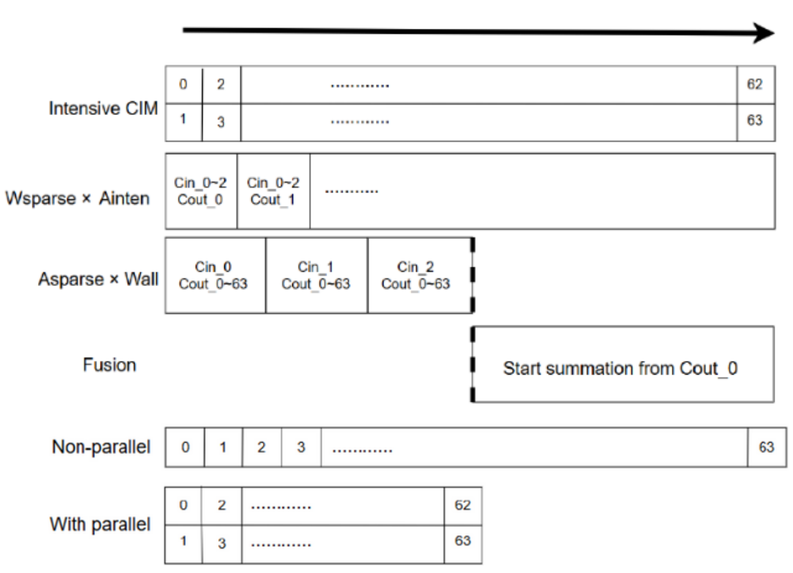

Fig. 1 Inten. Parallel Processing

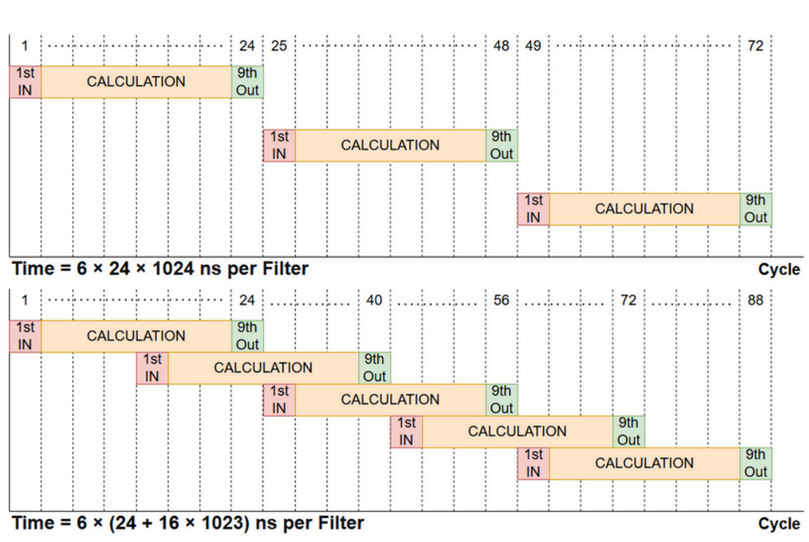

Fig. 2 Timeline With and Without CIM Pipelining

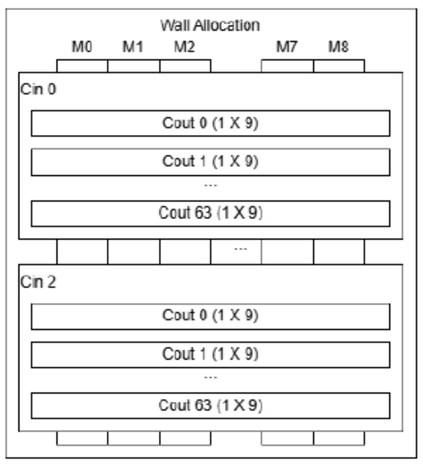

Fig. 3 Round-Robin Banking

III. EXPERIMENTAL RESULTS

Our architecture preserves full IEEE-754 accuracy while delivering substantive improvements: parallel processing for odd/even output channel in intensive core cuts total layer time by 50% ; CIM pipelining keeps every MAC unit fully engaged, yielding a 1.5× overall speedup over non-pipelined designs ; round-robin banking reduces fetch latency from tens of cycle to one.

Within the above framewrok, adjustable precision further trims per-layer energy by 14% (Fig. 5) with 99.95 % accuracy (Fig. 4), cutting per-layer latency by 56.25 % along with parallel processing (Fig. 6).

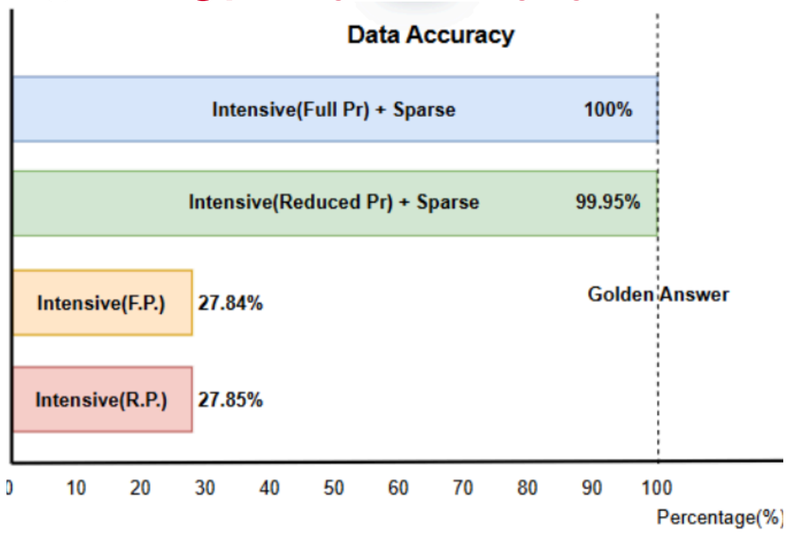

Fig. 4 Data Accuracy

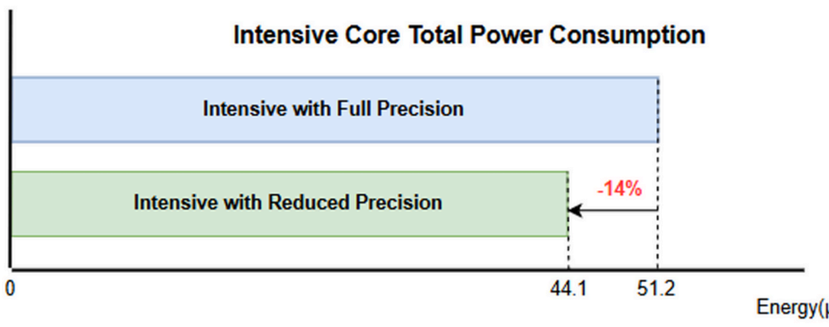

Fig. 5 Inten. Power Consumption

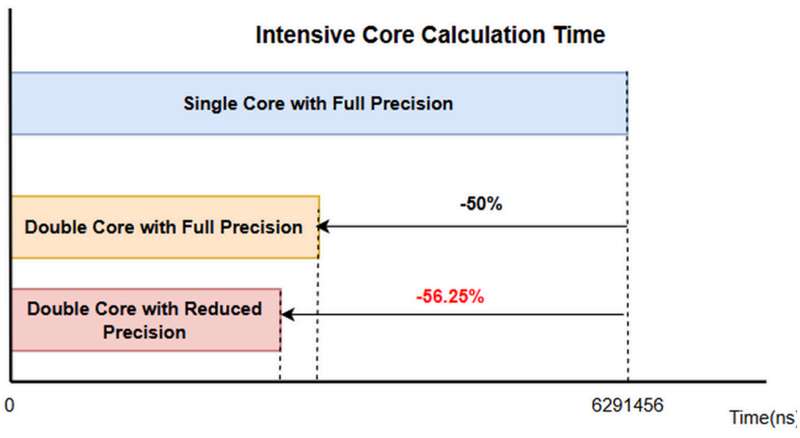

Fig. 6 Inten. Calculation Time

[1] J. Yue, C.-C. Li, Y.-H. Lai, W.-H. Liao, C.-H. Wang, T.-H. Yang, C.-C. Wu, Y.-M. Huang, and T.-W. Lu, "A 28 nm 16.9–300 TOPS/W computing-in-memory processor supporting floating-point NN inference/training with intensive-CIM sparse-digital architecture," IEEE Journal of Solid-State Circuits, vol. 59, no. 8, pp. 2630–2643, Aug. 2024.