N2N-BP²CIM: A 28nm 1018.1GOPS/mm² End-to-End Bit-Parallel, Bit-Parallel Computing-In-Memory DNN Processor with High-bandwidth 14T-SRAM Bitcell

[Placeholder for Author List]

[Placeholder for Affiliations]

The growing rate in deep neural network model size far exceeds the performance gain which technology scaling can provide, constantly widening the gap between software workloads and hardware computation capabilities. As a result, enhanced area efficiency for DNN processors is required to address this disparity. Despite the substantial energy efficiency improvement demonstrated by Computing-In-Memory (CIM) techniques, the area efficiency achieved by existing CIM processors still remains severely inadequate. This limited area efficiency results from the two main issues illustrated in Fig. 1. Firstly, most CIM designs utilize partially-activated CIM bitcells and adopt a low-bandwidth bit-serial (BS) input scheme. These methods lead to severe underutilization of parallelism and significantly compromise the achievable throughput performance, causing substantial degradation of area efficiency. Secondly, previous processor designs both struggle to handle diverse input topologies like conv1x1 and fully-connected (FC) layer with rigid input routes and incur additional data access for post-processing such as pooling and linear operation with SIMD. Consequently, this lack of adaptability to network structures seriously diminishes system-level throughput and efficiency.

To address the above challenges, this work proposes N2N-BP^{2}CIM, the first bit-parallel, bit-parallel (BP^2) end-to-end digital CIM processor for efficient DNN inference. N2N-BP^{2}CIM achieves peak system-level area and energy efficiency of 1018.1GOPS/mm^2 and 51.4TOPS/W, respectively, featuring the following design techniques: 1) A compact 14T SRAM high-bandwidth BPCIM bitcell realizing a 4x throughput improvement and boosting the area efficiency by 1.58x; 2) A reconfigurable compressor tree (RCT) for efficient accumulation of BP^2 partial products within a single cycle, achieving 2.39x higher area efficiency than conventional BSBP approaches; and 3) a Versatile Streaming Engine (VSE) that handles diverse input settings and streamlines output processing, resulting in a

55.12% reduction in end-to-end inference latency.

Fig. 2 shows the overall architecture of N2N-BP^{2}CIM. The processor consists of two 4.5KB BP^{2}CIM cores, a Precision-Flexible Accumulator (PFA), a VSE, and separated Global Memories for Instruction (GIM), Weight (GWM), and Activation (GAM). Each CIM core contains 18 0.25KB macros, which are organized into left and right 3x3 arrays similar to [4] to achieve the peak performance for the predominant conv3x3 operation in DNN applications. The PFA flexibly accumulates the MAC results produced by the cores according to the configured precision. The VSE consists of an Input Buffer (INBUF) for input transmission with high adaptability, a Near-Memory Pipeline (NMP) with efficient sequential post-processing, and a Short-Cut Buffer (SCB) that provides a short-cut path for NMP to eliminate any potential execution stall.

The proposed BP^{2}CIM macro illustrated in Fig. 2 is composed of 8 processing elements (PEs), with each PE housing 64 rows of 4-bit CIM clusters and 1 RCT. Each 4-bit CIM cluster, storing a 4-bit weight, consists of four 14T bit-parallel CIM bitcells, as depicted in Fig 3. These bitcells are arranged in a 2-by-2 manner with the RCT positioned vertically across them, optimizing routing length and minimizing routing efforts. During CIM operation, a total of 64 4-bit input activations IN[255:0] are transmitted to the macro each cycle. The proposed BPCIM bitcells then perform bitwise multiplications in a fully bit-parallel fashion to generate the corresponding 64-channel 4-bit x 4-bit outputs OUTB[1023:0]. Subsequently, the proposed RCT accumulates these BP^2 partial products, accomplishing a 64-dimension MAC computation in each PE within a cycle. The proposed 14T BPCIM bitcell consists of a 6T standard SRAM and four customized 2T NAND gates to realize efficient bit-parallel multiplication. During standard read/write operations, the stored weight is accessed by activating WL. While in CIM operation, a 4-bit input IN[3:0] is distributed to the four NAND gates to produce the inverted bitwise multiplication result OUTB[3:0]. The bitcell layout has a compact footprint of 1.512 um^2, only twice the size of a standard 6T SRAM. The resulting 14T BPCIM macro achieves a 4x throughput improvement with merely 2.53x area overhead, effectively boosting the area efficiency by 1.58x compared to the conventional BS CIM macros utilizing 8T-SRAM bitcells.

During the accumulation process, conventional BS CIM designs usually employ adder trees with shift-and-add modules, necessitating bulky flip-flops (FF) for accumulation while suffering from lengthy processing time due to repetitive iterations. Fig. 4 presents the structure of the proposed FF-free RCT, which accumulates all BP^2 partial products generated by CIM bitcells within a single cycle while providing precision and sign flexibility for both input and weight. The RCT has two operating phases: Sign Unification Phase and Compression Phase. In the Sign Unification Phase, the Baugh-Wooley algorithm calibrates the results for signed formats by inverting specific partial products and introducing a pre-calculated compensation vector, offering comprehensive sign representation flexibility. In the Compression Phase, the 4x4 64-channel partial product vectors are serially compressed by four local compressors and one global compressor into a 14-bit global sum vector X_{sum} and a 12-bit global carry vector X_{carry}. The two vectors are subsequently summed up by a 12-bit Ripple Carry Adder (RCA) to generate the final 14-bit partial sum. The Three-Dimensional-Minimization (TDM) algorithm [6] is adopted in both the local and global compressor designs to minimize the delay and area overhead, effectively reducing 15.72% delay and 14.92% area compared to traditional RCA-based design. The proposed BPCIM bitcell and RCT are tailored for efficient 4-bit x 4-bit MAC operations, but can readily support higher input and weight precision by dynamically adapting to the assigned precision through the use of the PFA. The higher-precision input activations can be sequentially processed at a rate of 4 bits per cycle while the weights are spatially expanded into the two CIM macro arrays, with the left array storing the MSB 4 bits and the right one storing the LSB 4 bits of 8-bit weights. The proposed BP^2 RCT significantly improves the area efficiency by 2.39x compared to traditional BSBP method, while the PFA maintains the input and weight precision flexibility.

Conventional end-to-end CIM processors struggle to consistently deliver high performance across varying network structures due to their rigid input scheme and the SIMD-style post-processing. In contrast, the proposed VSE, composed of an INBUF, an NMP, and a SCB, realizes a well-rounded on-chip data communication fabric that offers diverse input streaming and efficient output processing, as depicted in Fig. 5. The INBUF, consisting of a 4x3 buffer array and a flexible streaming pathway, functions as an intermediary pipeline stage to receive input activations

fetched from GAM and efficiently stream them into the 3x3 CIM macro arrays. Besides forming a horizontal pipeline to fully activate all CIM macros for conv3x3, the incorporation of INBUF also maximizes the data reuse and hardware utilization via its flexible streaming pathway among diverse input topologies, such as FC layer and conv1x1. This effectively eliminates the need for weight re-access and achieves latency reduction by 3.09x for non-conv3x3 operations compared to the conventional activation-rotator-based approach [4]. In addition, the proposed INBUF supports the pooling-friendly algorithm introduced in [3], which significantly simplifies the pooling logic required in NMP and further reduces its size by 1.82x. The NMP sequentially performs essential post-processing tasks such as activation function, pooling, and quantization. Unlike the SIMD approach that demands frequent bidirectional data transfers between memory and processing blocks, resulting in system-level bandwidth bottlenecks and increased power consumption [4][5], the proposed NMP preserves a unidirectional processing flow without incurring superfluous data access. Furthermore, the SCB is placed adjacent to NMP to provide a short-cut path for operations entailing additional influx of data, preventing the one-way NMP flow from stalling by extra data fetching. It efficiently handles the activations with high channel depths and supports the widely used residual connection operations. The VSE, comprising INBUF, NMP, and SCB, effectively contributes to a 55.12% reduction in end-to-end execution time of CIM processor.

Figure 6 presents the measurement results for N2N-BP^{2}CIM chip fabricated using a 28nm CMOS technology. The operating frequencies can range from 20 to 180MHz, powered by supply voltage from 0.60 to 1.15V. This chip reaches an area efficiency of 1018.1 GOPS/mm^2 at 180MHz. The measured top energy efficiency of 51.34 TOPS/W is observed at 20MHz with a 0.6V supply. Compared with the state-of-the-art SRAM CIM end-to-end processors, N2N-BP^{2}CIM demonstrates an improvement in area efficiency by more than 2.28x. Furthermore, higher levels of energy efficiency and network adaptability are accomplished without compromising any computational accuracy. Fig. 7 shows the die photo as well as the chip summary.

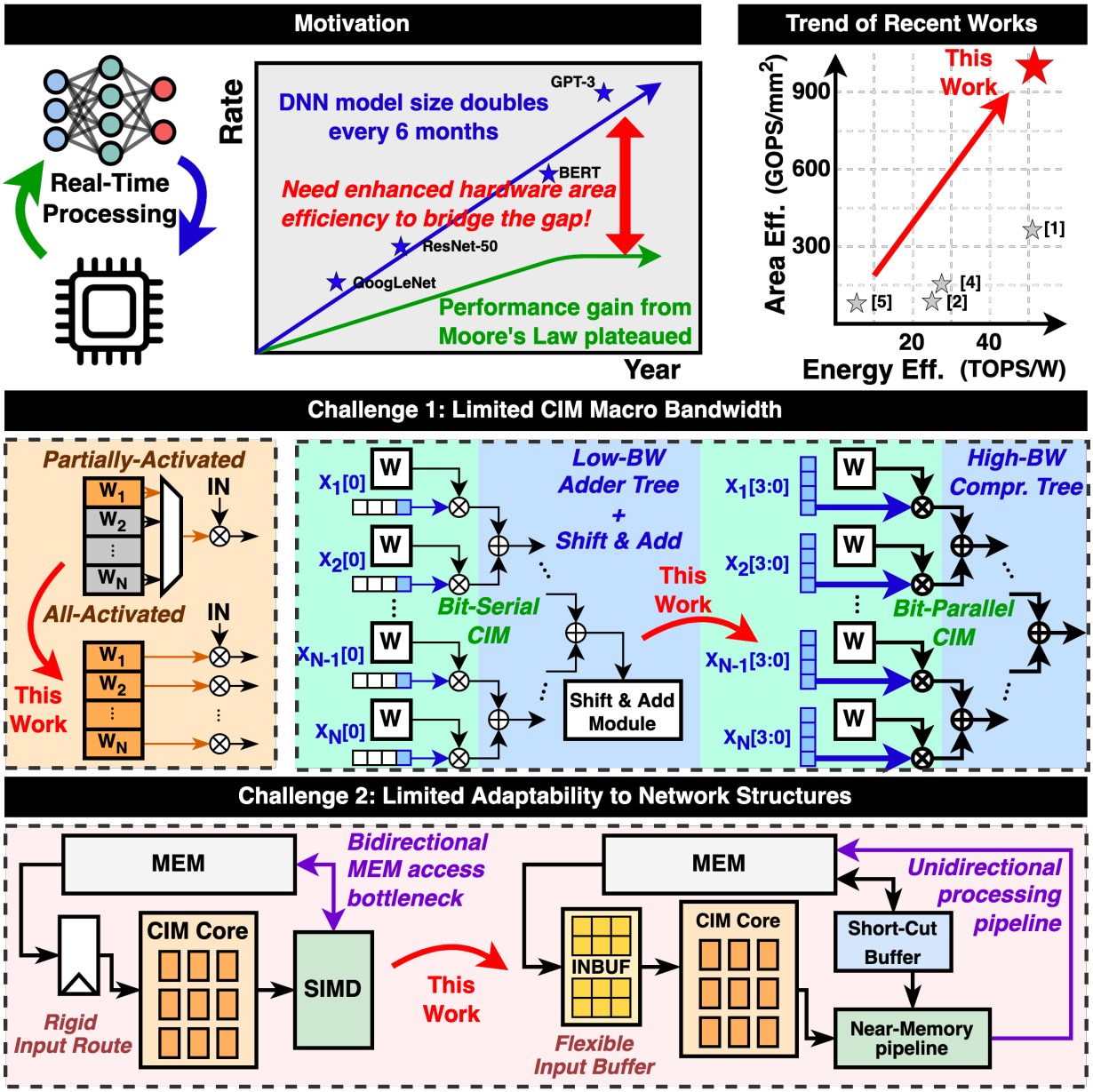
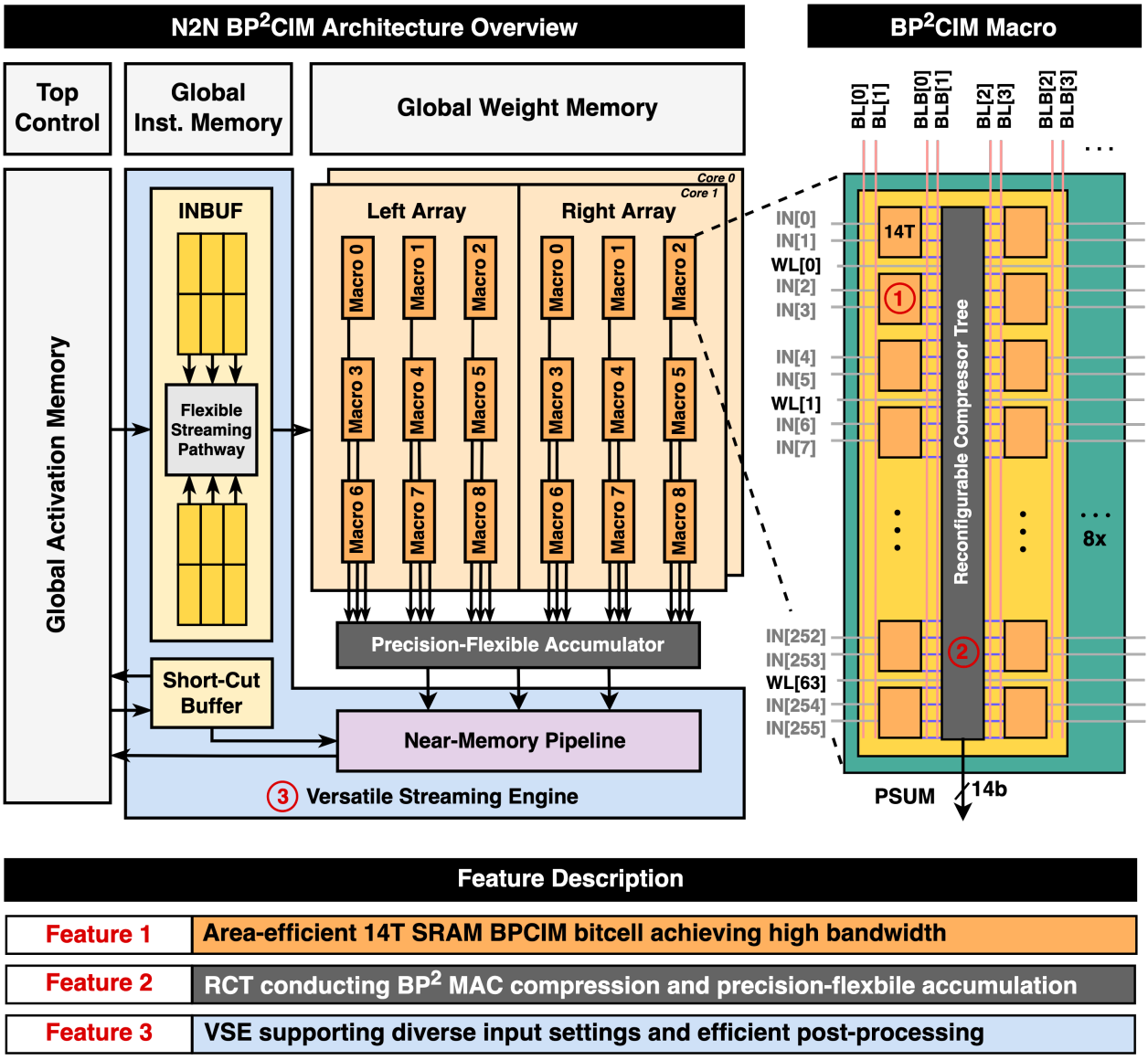


Figure 1: Motivation of N2N-BP^{2}CIM and design challenges of CIM processors.



Feature Description	
Feature 1	Area-efficient 14T SRAM BPCIM bitcell achieving high bandwidth
Feature 2	RCT conducting BP^2 MAC compression and precision-flexbile accumulation
Feature 3	VSE supporting diverse input settings and efficient post-processing

Figure 2: Architecture overview of N2N-BP^{2}CIM.

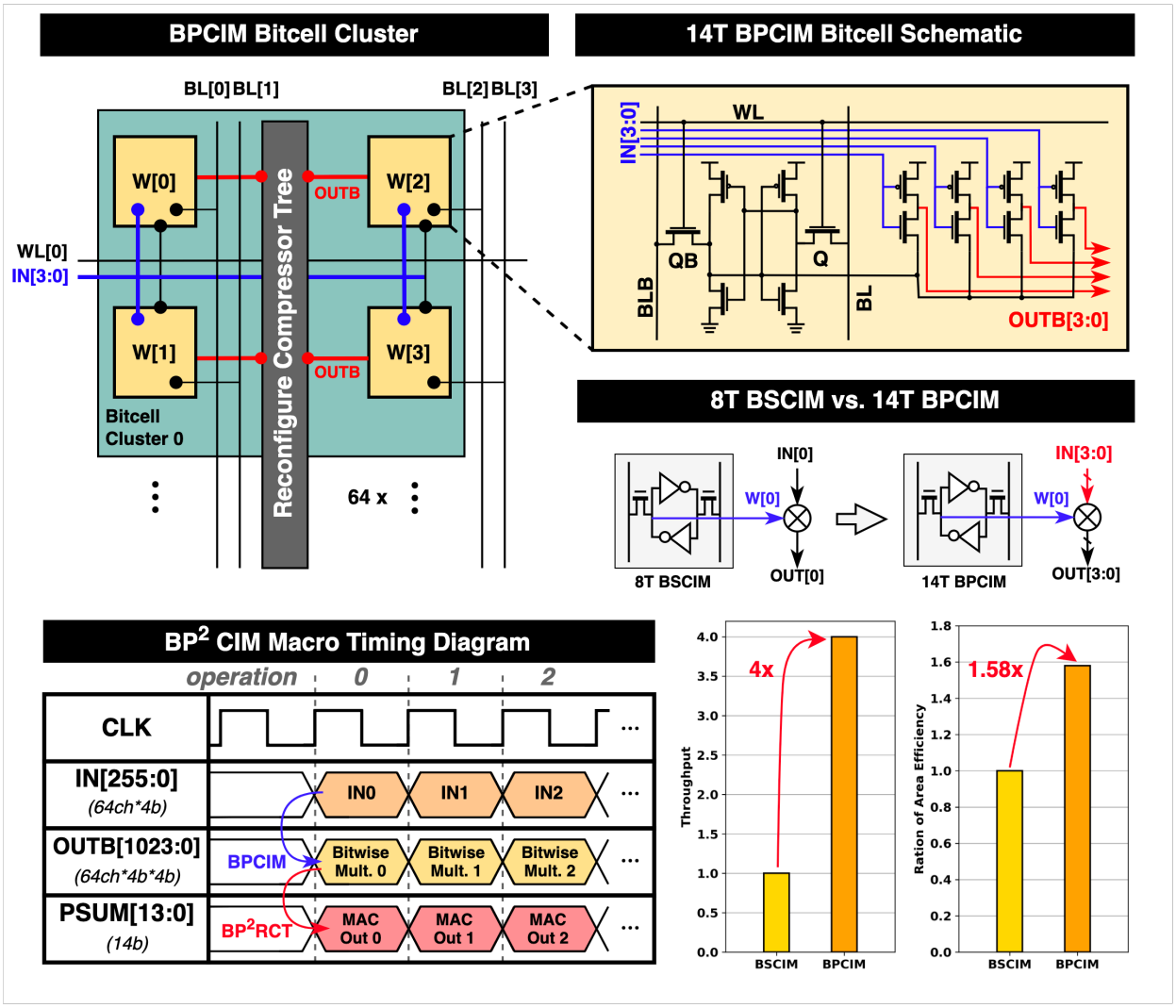


Figure 3: The structure of BP^{2}CIM macro and the schematic of the 14T SRAM BPCIM bitcell.

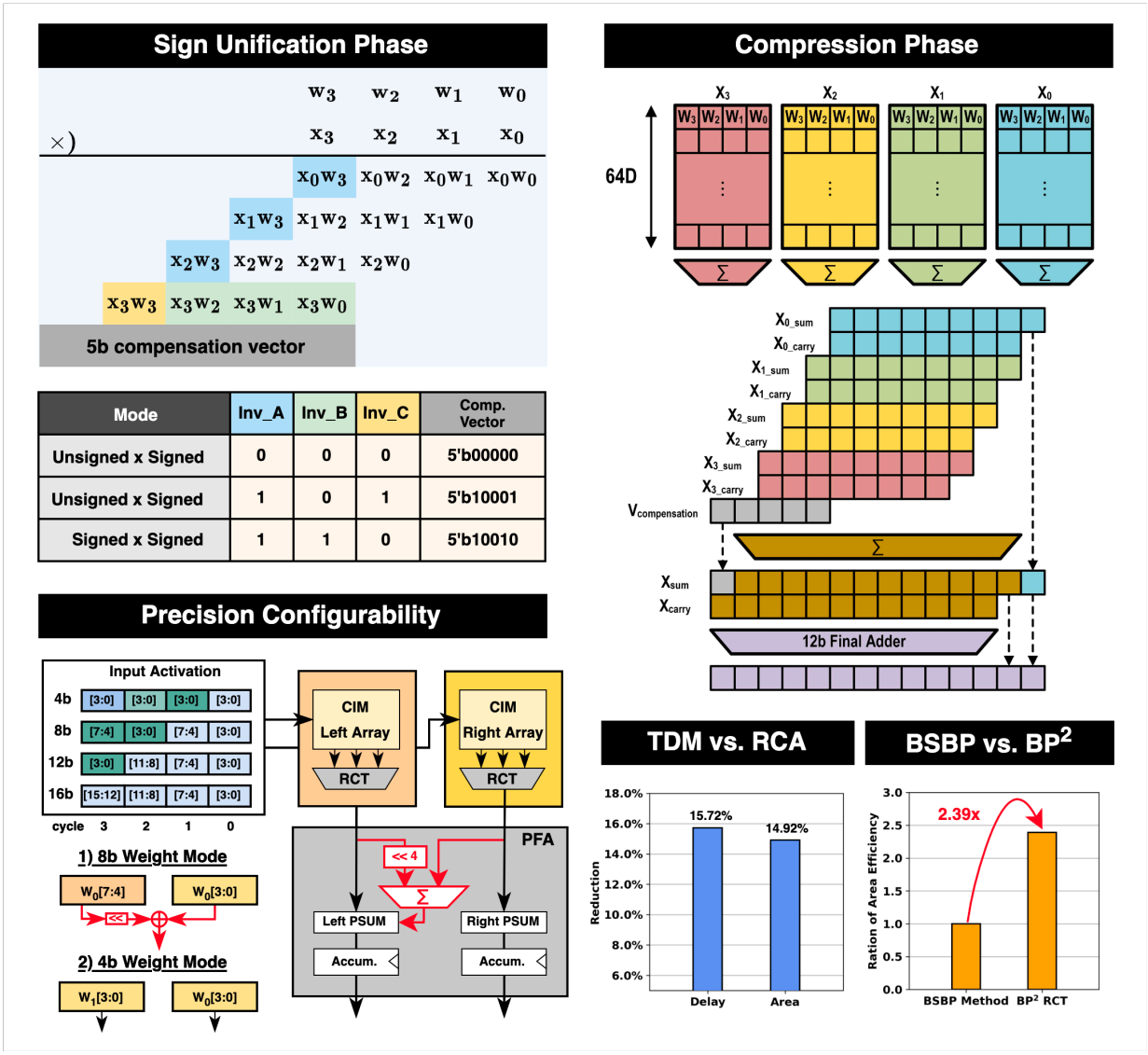


Figure 4: Configuration of the BP^2 RCT and its precision flexibility using PFA.

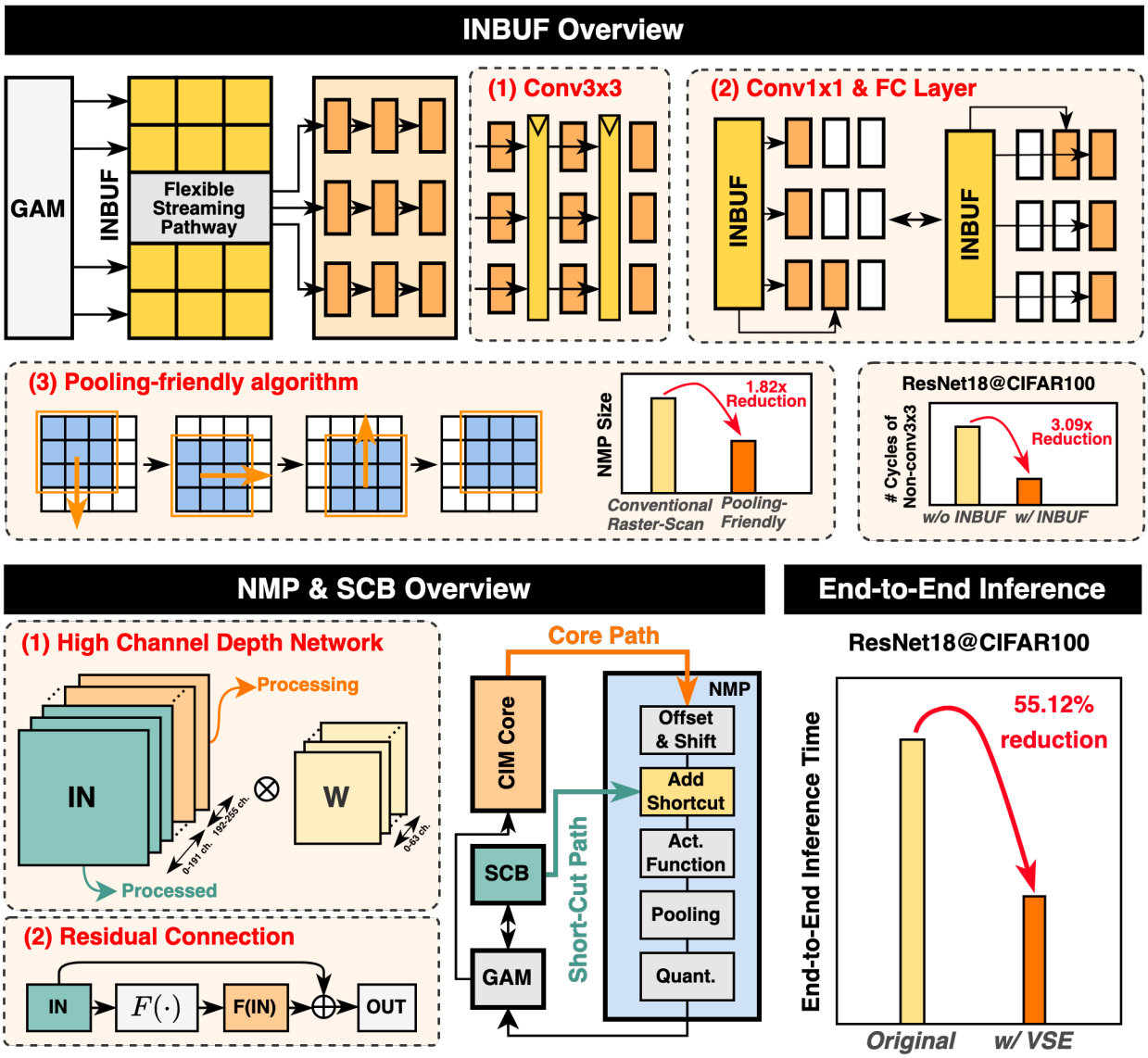
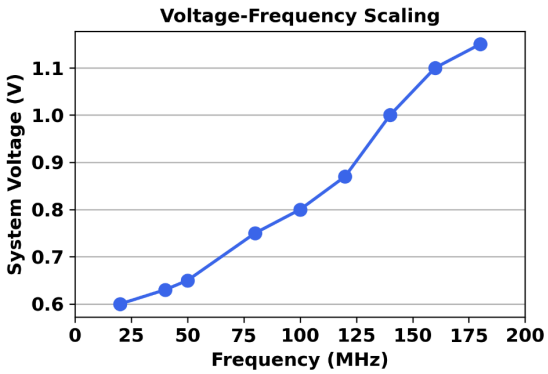
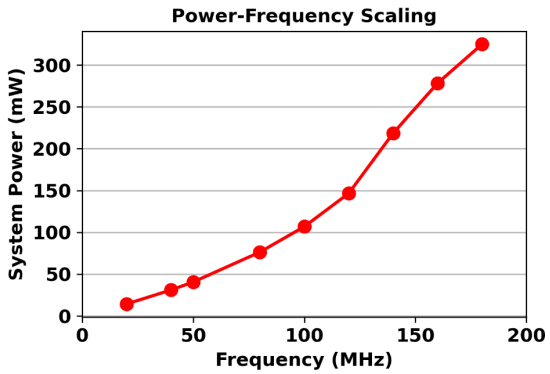


Figure 5: Architecture of VSE and the functionality of INBUF, NMP, and SCB.



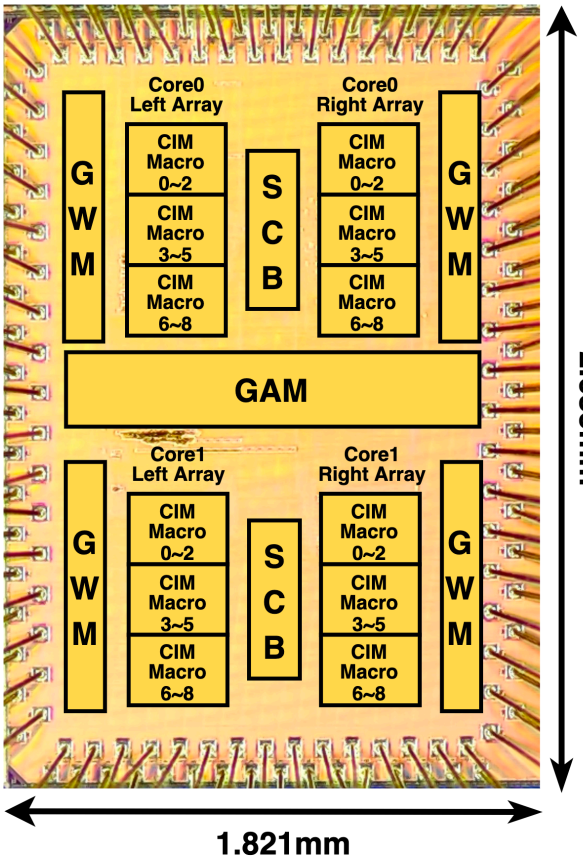


Comparison with Recent End-to-End SRAM CIM Processor for DNN

	ISSCC22[2]	VLSI23[5]	JSSC23[4]	ISSCC23[1]	**This Work**
Technology	28nm	55nm	28nm	28nm	**28nm**
Computation Type	BP^2 Digital COMB	BS^2 Digital CIM	Low-Precision Analog CIM	BS^2 Digital CIM	**BP^2 Digital CIM**
Lossless Computation	Yes	Yes	No	No	**Yes**
Area (mm^2)	8.70	12.87	20.89	4.54	**4.89**
Activation (b)	4	1-8	1-2	4/8, FP/BP16	**4/8/12/16**
Weight (b)	3	2/4/8	1-2	4/8, FP/BP16	**4/8**
System Area Eff. ($GOPS/mm^2$) [a]	83.62	79.25	146.5	361.2	**1018.1** [b]
System Energy Eff. (TOPS/W) [a]	24.7	5.62	27.3	51.0	**51.4** [b]

a: Normalized to 4bx4b operation. One MAC is counted as two operations. Throughput is assessed according to its peak value of the actual computation capacity, without the sparsity gain from operation skipping.
b: Best energy efficiency point is measured under 20MHz with a 0.6V supply; best area efficiency point is measured under 180MHz with a 1.15V supply.

Figure 6: Measurement results and comparison to state-of-the-art SRAM CIM processor works



Chip Summary	
Technology	28nm CMOS
Chip Area	4.89mm^2
CIM Core Area	0.99mm^2
CIM Storage	9KB
Activation	4/8/12/16b
Weight	4/8b
Voltage	0.6-1.15V
Frequency	20-180MHz
System Power[a]	14.34mW
System Area Efficiency[b]	1018.1GOPS/mm^2
System Energy Efficiency[a]	51.4TOPS/W

a: Measured at 20MHz with a 0.6V supply
b: Measured at 180MHz with a 1.15V supply

Figure 7: Die photo and chip summary of N2N-BP^{2}CIM.

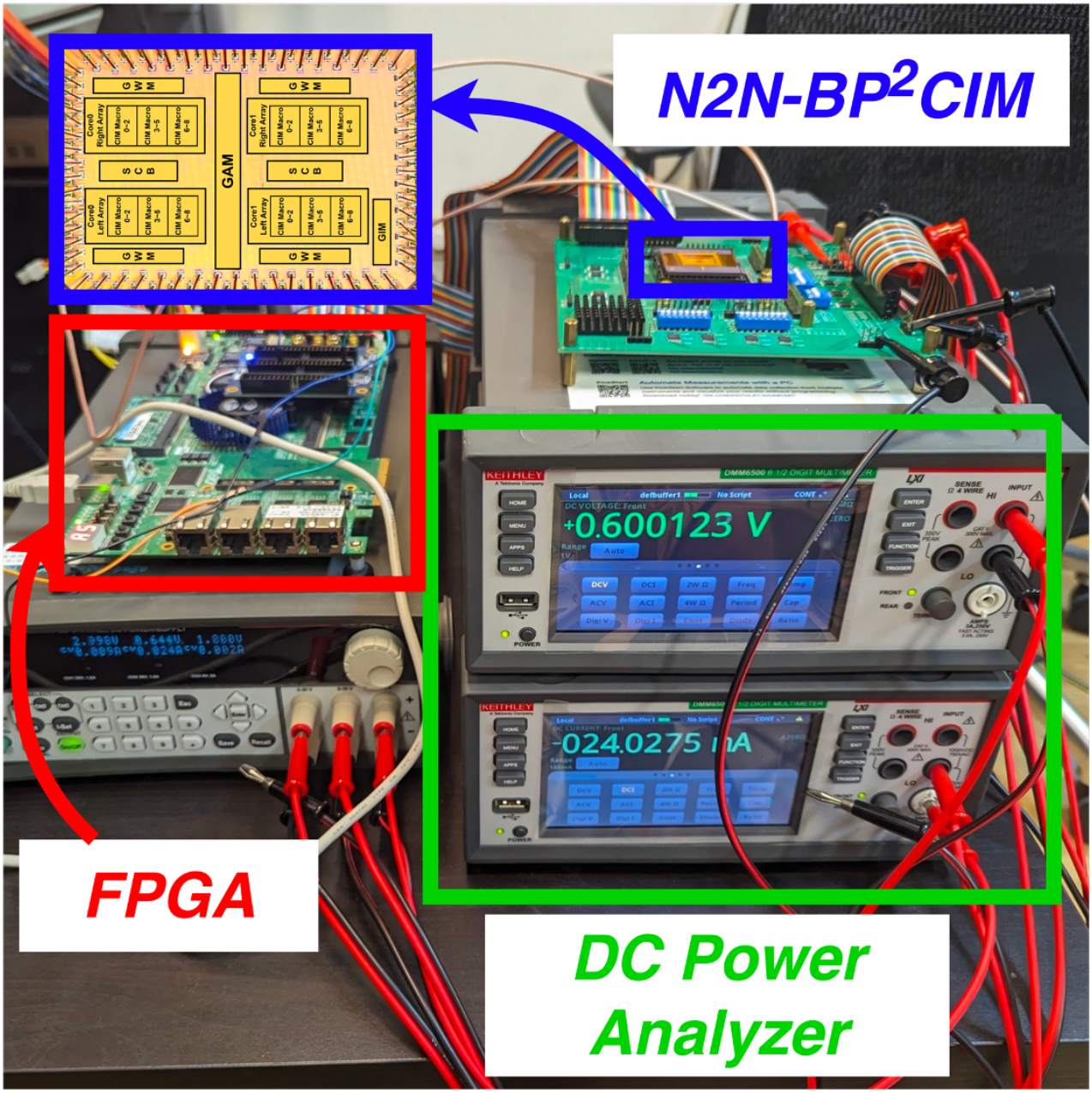


Figure S1: The measurement environment of N2N-BP^{2}CIM.