

A 6.0 TOPS/W Reconfigurable AI-based Channel State Information Compression using Delta Encoding in Multi-Receiver Mobile Systems

Hana Kim^{1,2,*}, Zihan Xia^{1,*}, Yuchan Li¹, Suraj P N¹, Rishabh Kumar¹, Pranav Raj¹, *Member, IEEE*, Hyunseok Lee³, Junho Lee³, Jiyong Yoon³, Jungwon Lee³, Ji-Hoon Kim², and Mingu Kang¹, *Member, IEEE*
 mingu@ucsd.edu *Equal contribution

¹Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA, USA

²Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea

³Samsung Electronics Co., Gyeonggi-do, Republic of Korea

Abstract—This paper presents a low-power processor for AI-based channel state information (CSI) compression, for the first time. Exploiting the high similarity between signals from multiple receivers, the proposed delta encoding enhances sparsity and generates small-magnitude numbers, reducing energy consumption in computation and data movement. A customized computing paradigm, integrating mixed sign-magnitude (S&M) and two’s complement (2S/C) number representations, is developed to further optimize efficiency. The architecture also offers high reconfigurability, supporting diverse layer structures in state-of-the-art models for CSI compression, including hybrid convolution (CONV) and transformer blocks. Fabricated using a 65 nm process, the silicon prototype achieves a peak energy efficiency of 6.0 TOPS/W, highlighting its potential for mobile devices.

Index Terms—channel state information (CSI), delta encoding, MIMO system, sparsity, communication.

I. INTRODUCTION

In a MIMO system, channel state information (CSI) is acquired by the mobile user equipment (UE) during the channel training and fed back to the base station (BS) (Fig. 1). This CSI is utilized to enhance communication quality by adapting signal transmission and reception. However, such a process is hindered by feedback overhead due to the large amount of transmission data through bandwidth-limited channels. To mitigate this, the UE employs an AI model to compress the CSI data [1-2]. This work proposes an accelerator for CSI compression for mobile UE under stringent power and area budgets, leveraging STNet [3], a state-of-the-art model combining convolution (CONV) and transformer for high compression efficiency. Key challenges (Fig. 1) include: C1) increased computational complexity from executing AI models multiple times to process inputs from the UE’s four antennas, C2) diverse layer structures in STNet with greatly varying operand tensor shape (2–2048 dimensions), C3) massive data movement, particularly in the bulky fully-connected (FC) layer. Furthermore, while mobile devices often include four antennas, the number of active receivers can dynamically change depending on the device’s workload and power budget, necessitating flexible execution. To address these, we propose S1) a delta-channel processing (DCP) with mixed sign-magnitude (S&M) and two’s complement (2S/C) number systems that

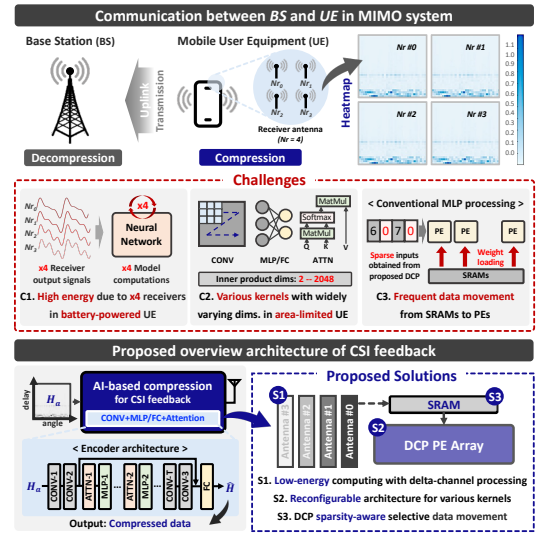


Fig. 1. AI-based channel state information (CSI) compression in a mobile user equipment (UE) with four antennas. Challenges vs. solutions in the proposed processor by exploiting high data similarity across four uplink receivers.

exploits the high similarity among the four receivers to obtain both sparse and small-magnitude signals for energy saving, S2) a highly reconfigurable architecture with specialized processing elements (PEs) to adapt to various layer structures, and S3) a selective data load to avoid redundant memory transactions via exploiting sparse signals introduced by DCP. The number of active receiver signals is also adjustable by configuring the batch size of DCP.

II. PROPOSED ARCHITECTURE FOR CSI COMPRESSION

A. Overall Architecture

Fig. 2 illustrates the overall architecture, which consists of: 1) a PE array comprising 32 DCP PEs, each equipped with a 16-tap dot-product unit of mixed number formats. Each PE can be independently disabled with clock gating for selective use; 2) a vertically placed SRAM (V-SRAM) and a horizontally placed SRAM (H-SRAM), which alternately serve as input and output buffers; 3) a special function processor (SFP), which includes 32-way special function units and a column-wise reduction unit (CRU). The special function unit supports delta

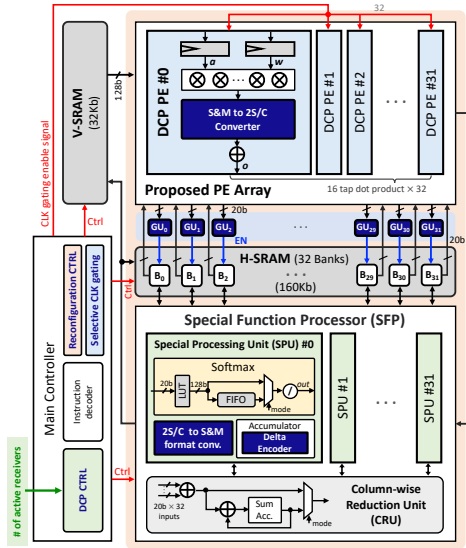


Fig. 2. Overview of proposed processor with delta-channel processing (DCP) with high reconfigurability.

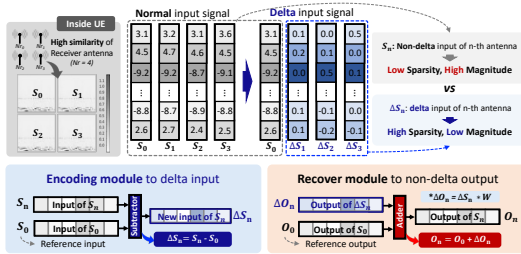


Fig. 3. Delta encoding for signals from four receivers and their recovery of outputs back to non-delta representation.

signal encoding, format conversion, and softmax operation with look-up-tables for exponential function evaluations; 4) a centralized controller coordinating the sequencing of instructions.

B. Delta Encoding for Signals from Four Receivers

Fig. 3 illustrates the delta encoding operation to compute $\Delta S_n = S_n - S_0$, where one of the received signals S_0 is used as the reference. S_n is the original values from four receivers, and ΔS_n is generated in the SFP. By exploiting the similarity across four receivers, a significant portion of signal values (39.5–55.7%) become exact zero or near-zero values (mean: 0.73–1.36). The obtained differential signals are then fed into subsequent layers, enabling low-power operation due to the sparsity and low magnitude. Finally, the output ΔO_n computed from ΔS_n is added to O_0 to reconstruct O_n by the linearity of neural network operations. Note the delta encoding-based processing guarantees the correct results without approximation, and the number of signals involved in delta encoding can be flexibly adjusted from 1 to 4 depending on the resource budget of the mobile device.

C. Processing Element (PE) with Mixed Number System

Despite the sparsity benefits, the small magnitudes alone do not reduce power. This is because, in the typical 2S/C number system, consecutive signed inputs with small magnitude can

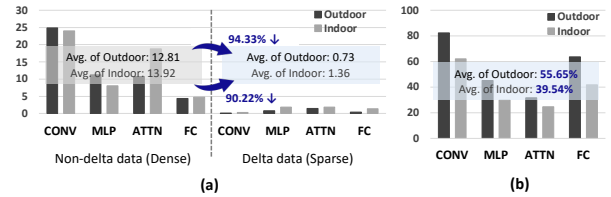


Fig. 4. Impact on inputs of each layer by delta encoding for two representative channel scenarios of indoor and outdoor from Sionna library [4]: (a) magnitude reduction of input signal, (b) sparsity of delta-encoded input ΔS_n .

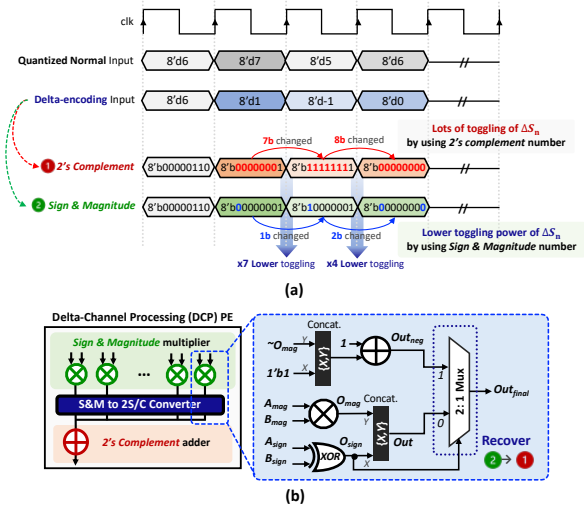


Fig. 5. (a) Toggling behavior comparison between 2S/C and S&M number systems, (b) proposed DCP processing element (PE) with a mixed S&M and 2S/C number representations.

still cause high toggling of internal electrical nodes. As shown in Fig. 5, where 1 to -1 transition results in 7-bit toggling, leading to substantial energy consumption. To address this issue, we exploit a mixed-number representation. The multiplication is performed using the S&M number to minimize toggling for small-magnitude signed values, e.g., transitions across -1, 1, and 0 result in only a single-bit toggle. However, the accumulation is highly inefficient in the S&M system due to a required magnitude comparison between two operands for each addition. To overcome this, we introduce a converter that transforms the products in S&M to 2S/C numbers for the processing of the accumulation in 2S/C number system. Despite the conversion, the total power is reduced by 45.5% (see Section III).

D. Proposed Selective Data Loading

FC layers tend to incur high data movement costs due to limited weight reuse despite their low arithmetic intensity. In particular, the final FC layer in STNet with large input and output dimensions (e.g., 2048/512) causes significant memory access overhead in mobile devices. To address this, we employ selective data loading (Fig. 6) by exploiting the sparsity from the delta encoding. When the input stationed in the PE's register is zero, the corresponding column's H-SRAM bank is gated to prevent weight loading, as the output will remain zero regardless of the weight. This approach synergizes with delta encoding with higher sparsity, reducing energy by 1.54×

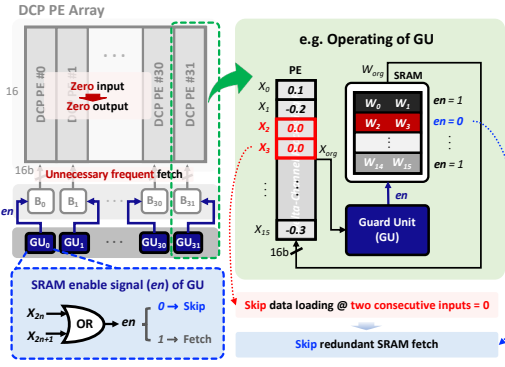


Fig. 6. Proposed selective data loading to minimize the data movement and memory access costs by exploiting high sparsity from delta encoding.

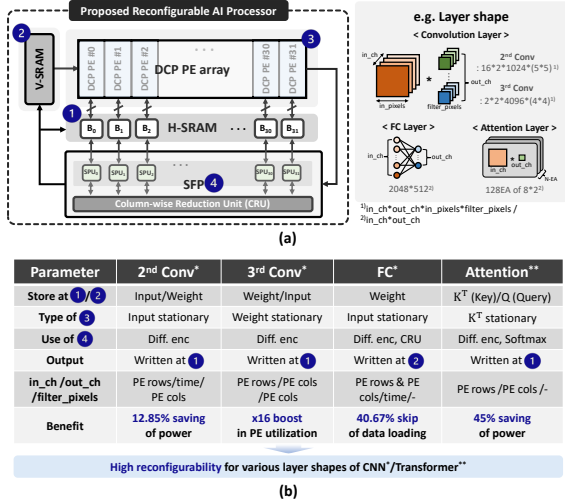


Fig. 7. Support for widely varying layer structures in CSI compression model (a) reconfigurable architecture, (b) mapping strategy.

E. Reconfiguration for Widely Varying Layer Structures

STNet employs various types of layers with wide channel size variations (e.g., from 2 to 2048). A brute-force application of conventional weight-stationary mapping could significantly increase latency and degrade hardware utilization, motivating the need for highly reconfigurable hardware and optimized mapping strategies tailored to each stage.

Fig. 7 illustrates the reconfiguration for several representative layers, indicating the data placement in H-SRAM ① and V-SRAM ② for each layer, the choice of data-stationary in the PE array registers ③ during execution, and the operation type of the SFP ④. In the 2nd CONV layer, input activations are stored in H-SRAM and loaded into the PE registers to be stationed. This exploits the high-sparsity delta-encoded inputs stationed for consecutive cycles, achieving 12.6% power savings compared to the conventional weight stationary mapping. During execution, weights flow from V-SRAM to the PE array column by column with output written in H-SRAM. In the 3rd CONV, having only two input and output channels leads to poor PE utilization. To mitigate this, 16 filters’ weights for two output channels are distributed and stationed across 32 DCP PE columns’ registers, improving the PE utilization by 16 \times . In the FC layer, input stationary is exploited to minimize data movement from H-SRAM by

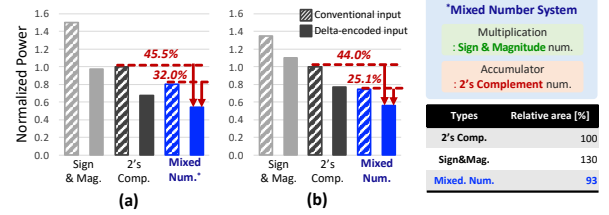


Fig. 8. PE power consumption with three different number representations (obtained via post-layout simulations). (a) 2nd Conv layer, (b) FC layer.

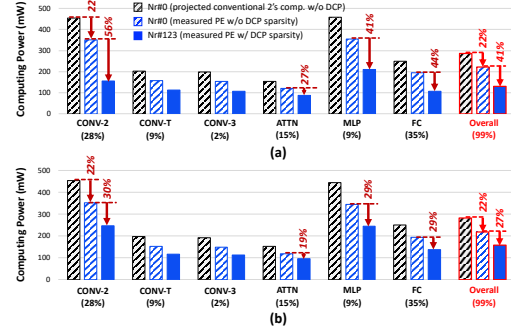


Fig. 9. Measured DCP benefits for various layers. The percentage on the X-axis indicates each operator’s contribution to the total number of FLOPs. (a) outdoor and (b) indoor datasets.

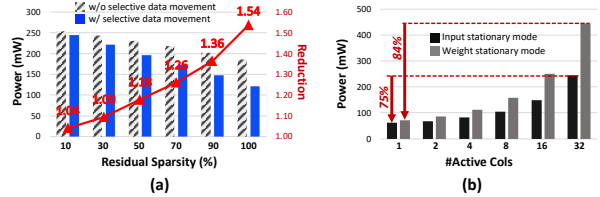


Fig. 10. Measured results for (a) the benefit from sparsity-aware selective data movement, (b) power scalability by PE gating.

SRAM gating whenever the stationed delta-encoded input in each DCP PE is zero, as shown in Fig. 6. The outputs from all columns are accumulated in the CRU before being written to V-SRAM. The attention layer has tiny dimensions of only 2–8, executing the limited number of DCP PEs. The power of such layers can be dramatically reduced by PE column-wise disabling, leading to 45% savings.

III. MEASURED RESULTS AND EXPERIMENTS

Fig. 8 shows the post-layout power analysis of the proposed DCP PE with 8-bit input operands and 24-bit output precisions compared to the conventional PE for sparse and non-sparse weight and input activation cases. These results show that the DCP PE achieves significant power savings (44–45.5%) by leveraging the mixed-number system in synergy with delta-encoded inputs, which introduce both high sparsity and small-magnitude values. Interestingly, despite the conversion overhead, the DCP PE occupies less area than both the S/M and 2S/C PEs by combining the simpler multiplier design of the S/M format with the more efficient adder tree structure of the 2S/C representation.

We measured the power using a realistic channel dataset generated with NVIDIA’s Sionna library [4], which produces MIMO channel conditions in both indoor and outdoor environments with 1024 sub-carriers, where square area

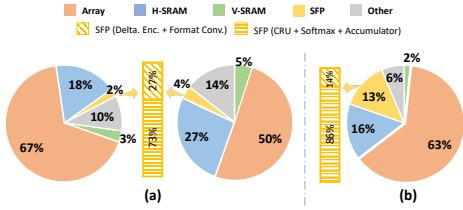


Fig. 11. Measured breakdowns of overall (a) power and (b) area.

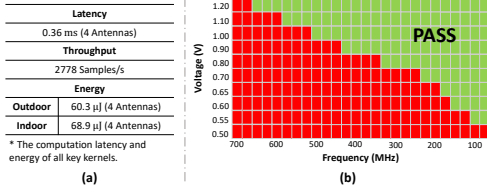


Fig. 12. (a) Application-level specification and (b) Shmoo plot.

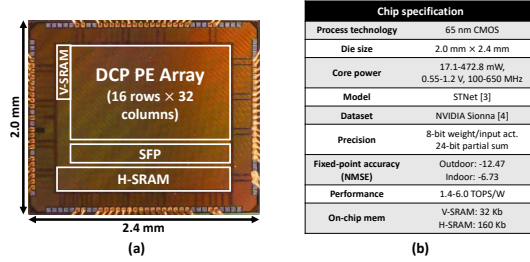


Fig. 13. Prototype chip: (a) micrograph fabricated in a 65 nm process technology, and (b) specifications.

sizes/frequency of $20 \text{ m}^2 / 5.3 \text{ GHz}$ and $400 \text{ m}^2 / 300 \text{ MHz}$ are employed, respectively. Fig. 9 and Fig. 10 show the measured power of a chip fabricated in a 65 nm process for various layers. As shown in Fig. 9, the overall power reduction across key layers is 22% compared to the 2S/C system and an additional 41% for the delta-encoded inputs. The 2nd and 3rd CONV layers take the dominant portion (83%) of the overall delay before the proposed optimizations are applied. The proposed reconfigurability in Fig. 7 delivers 87.6% / 93.4% delay reduction for 2nd / 3rd CONV layers, leading to overall 76.5% delay reduction. Fig. 10 shows that the selective data movement and column-wise disabling techniques achieve up to 1.54 \times and 84% power reductions.

The power breakdown in Fig. 11 indicates the overhead from delta encoding and format conversion is <1.1% in power and <1.9% in area. As shown in Fig. 12, the shmoo plot indicates the chip operates at 500 MHz at 1 V. Fig. 13 shows the die photo with an area of $2.0 \text{ mm} \times 2.4 \text{ mm}$ and a summary of specifications. Note that the proposed DCP is not an approximate computing technique. Therefore, it maintains the same fixed-point accuracy as the software implementation, achieving normalized mean squared error (NMSE, which is the widely used metric for the performance of CSI) of -12.5 and -6.7 for the outdoor and indoor datasets, respectively.

To the best of our knowledge, this work is the first CSI processor. We compared ours with prior works on reconfigurable sparsity-adaptive accelerators in Fig. 14. This work achieves the highest energy efficiency and competitive area efficiency by leveraging the DCP-based sparsity and the proposed DCP PE for the mixed-S&M and 2S/C number system.

	This Work	ISSCC'23 [5]	JSSC'24 [6]	VLSI'22 [7]
Sparsity	Delta-Channel Sparsity	Bit Sparsity by LSB Truncation	Sparsity-aware Freq. Boost	Unstructured Sparsity
Number System	Mixed sign-mag. & 2's comp.	Sign-mag.	2's comp.	2's comp.
Supported Kernels	CONV, FC/MLP, ATTN	CONV, FC/MLP	CONV, FC/MLP, ATTN	CONV, FC/MLP, ATTN
Tech. Node (nm)	65	28	65	5
Precision (bits)	INT8	INT8 (Approx.)	INT8	INT8/INT4
Area (mm^2)	4.80	2.18	6.40	0.153
Voltage (V)	0.55 – 1.20	0.65 – 0.90	0.90 – 1.20	0.46 – 1.05
Frequency (MHz)	100 – 650	55 – 285	600 – 1200	152 – 1760
Power (mW)	17 – 473	N.A.	400 – 1675 ⁽¹⁾	N.A.
Performance (TOPS) [scaled to 65nm]	0.10 – 0.66	0.08 – 0.44 [0.03 – 0.22] ⁽²⁾	0.61 – 1.23	1.8 ⁽²⁾ (1.05W) [0.19] ⁽¹⁾
Energy Effi. (TOPS/W) [scaled to 65nm]	1.41 – 6.00	4.20 – 8.09 [0.74 – 2.26] ⁽²⁾	0.6 – 1.0	39.1 ⁽²⁾ (0.46W) [0.92] ⁽¹⁾
Area Effi. (GOPS/ mm^2) [scaled to 65nm]	20.8 – 125.0	36.7 – 201.8 [9.3 – 51.1] ⁽¹⁾	95.3 – 192.1	11.7 ⁽²⁾ [30.35] ⁽¹⁾

- (1) Scaled to 65nm by the scaling rule [8].
(2) Peak performance and energy efficiency at INT8.
(3) Computing core power only with SRAM power excluded.

Fig. 14. Comparison with the state-of-the-art accelerators.

IV. CONCLUSION

This paper presents a low-power processor for AI-based channel state information (CSI) compression. It employs a selective data movement scheme and a specialized processing element (PE) to efficiently handle high-sparsity, low-magnitude signals generated through delta encoding, which leverages the strong similarity among signals received from multiple antennas. The architecture also features high reconfigurability, supporting diverse layer structures found in state-of-the-art CSI compression models. The silicon prototype achieves a peak energy / area efficiency of 6.0 TOPS/W and 125.0 GOPS/ mm^2 in a 65 nm process, paving the way toward ultra-low-power CSI in battery-operated user equipment.

ACKNOWLEDGMENT

This work was supported by Samsung Electronics Co., Ltd.

REFERENCES

- [1] Guo, J., Wen, C. K., Jin, S., & Li, G. Y. (2022). Overview of deep learning-based CSI feedback in massive MIMO systems. *IEEE Transactions on Communications*, 70(12), 8017-8045.
- [2] Burghal, D., Li, Y., Madadi, P., Hu, Y., Jeon, J., Cho, J., et al. (2023). Enhanced AI-based CSI prediction solutions for massive MIMO in 5G and 6G systems. *IEEE Access*, 11, 117810-117825.
- [3] Mourya, S., Amuru, S., & Kuchi, K. K. (2022). A spatially separable attention mechanism for massive MIMO CSI feedback. *IEEE Wireless Communications Letters*, 12(1), 40-44.
- [4] Hoydis, J., Cammerer, S., Aoudia, F. A., Vem, A., Binder, N., Marcus, G., & Keller, A. (2022). Sionna: An open-source library for next-generation physical layer research. *arXiv preprint arXiv:2203.11854*.
- [5] An, H., Chen, Y., Fan, Z., Zhang, Q., Abillama, P., Kim, H. S., et al. (2023, February). 29.3 an 8.09 tops/W neural engine leveraging bit-sparsified sign-magnitude multiplications and dual adder trees. *IEEE International Solid-State Circuits Conference* (pp. 422-424). IEEE.
- [6] Sundaram, S. S., Khodke, Y., Li, Y., Jang, S. J., Lee, S. S., & Kang, M. (2023). FreFlex: A high-performance processor for convolution and attention computations via sparsity-adaptive dynamic frequency boosting. *IEEE Journal of Solid-State Circuits*, 59(3), 855-866.
- [7] Keller, B., Venkatesan, R., Dai, S., Tell, S. G., Zimmer, B., Dally, W. J., et al. (2022, June). A 17–95.6 TOPS/W deep learning inference accelerator with per-vector scaled 4-bit quantization for transformers in 5nm. *IEEE Symposium on VLSI Technology and Circuits* (pp. 16-17).
- [8] Stillmaker, A., & Baas, B. (2017). Scaling equations for the accurate prediction of CMOS device performance from 180nm to 7nm. *Integration*, 58, 74-81.