

Software and Hardware Co-optimization for Graph Neural Networks on FPGA

Ruiqi Chen, Advisor: Bruno da Silva, Co-Advisor: Kun Wang

ruiqi.chen@vub.be, Bruno.da.Silva@vub.be, kun.wang@ieee.org

Challenges 1 Bit-level

Precision

2-bit

INT8

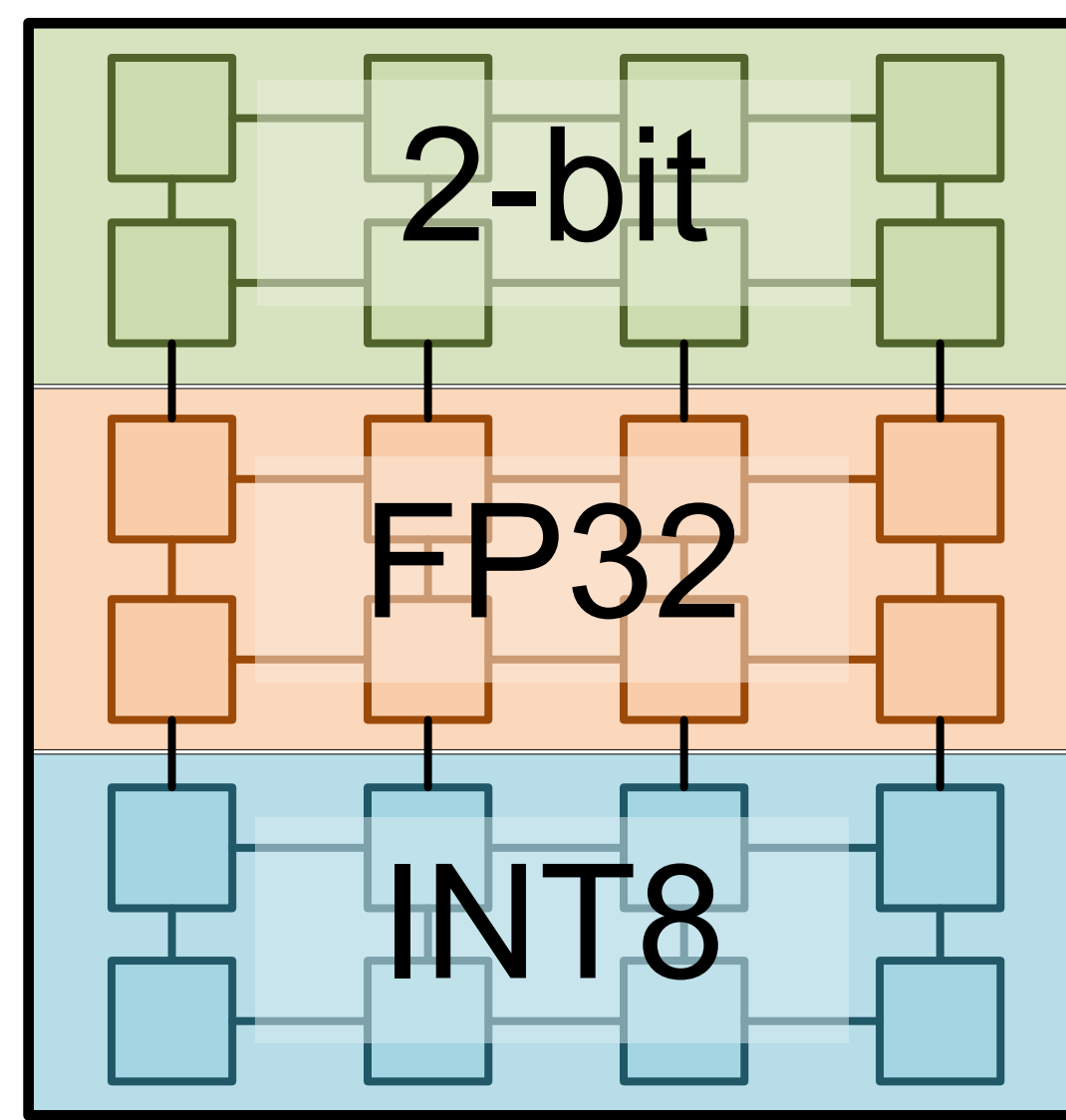
FP32

Computation_0

Computation_1

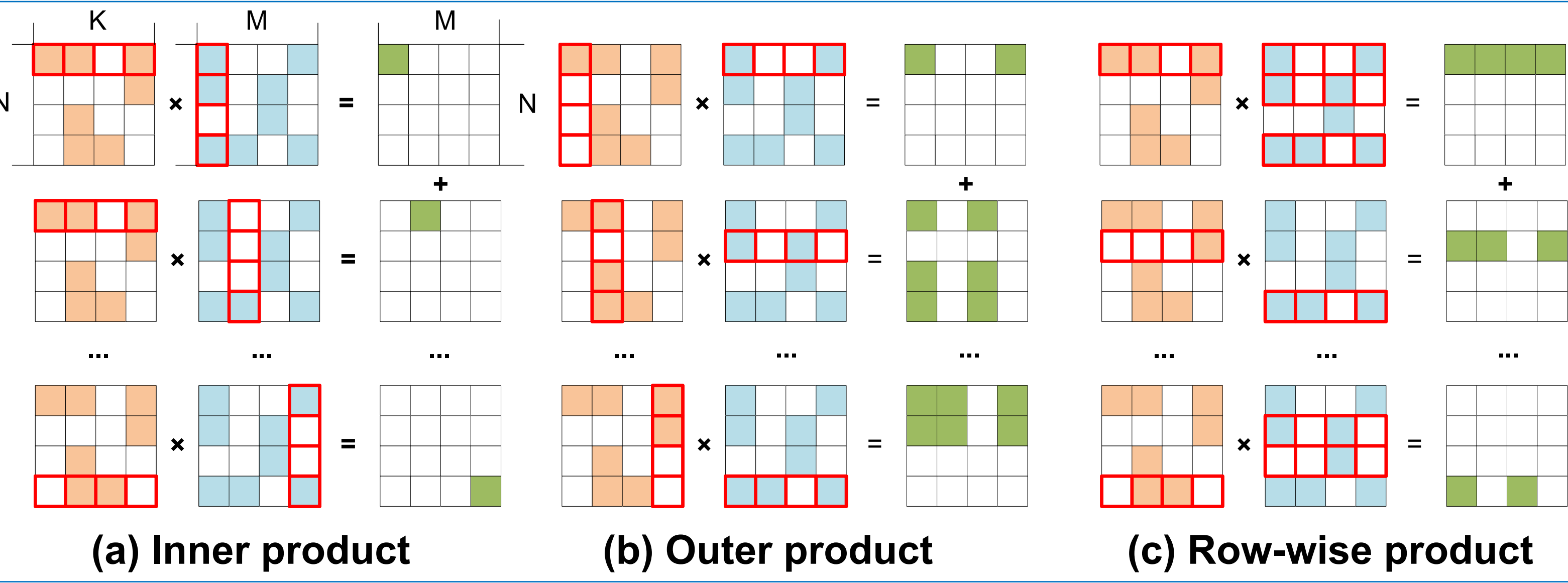
Computation_2

Model



Accelerator

Challenges 2 Data-structure-level



(a) Inner product

(b) Outer product

(c) Row-wise product

Challenges 3 Computation-level

Operator

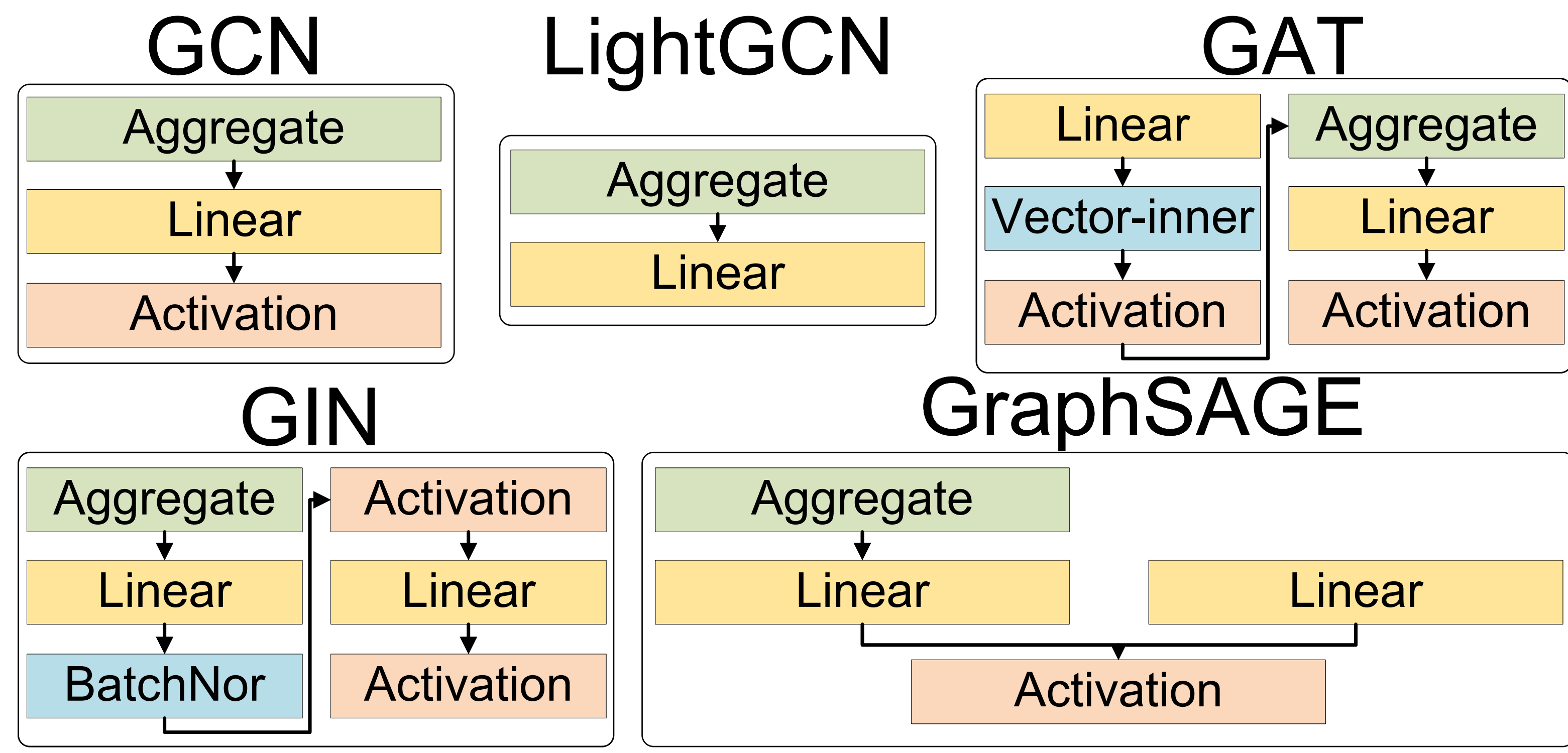
Aggregate
Activation

Linear
Others

Process

Aggregate
Linear
Others
Activation

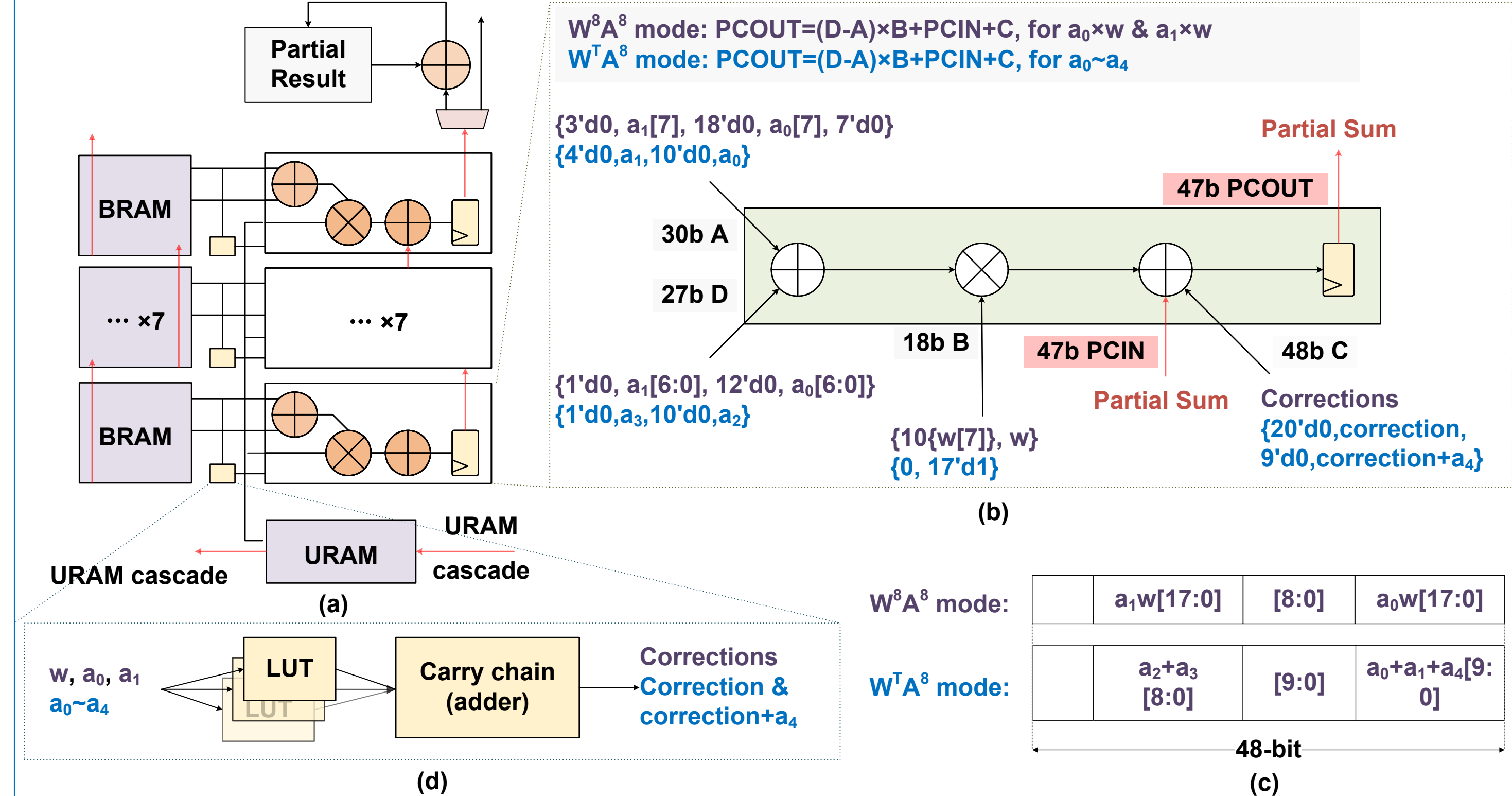
Challenges 4 Algorithms (Models)-level



Contributions

- **C1:** Mix quantization strategy with an **unified processing element (PE)** array design for GCN accelerator
- **C2:** **Optimizing sparse matrix encoding formats** and PE array designs for GNNs
- **C3:** Q-learning-based **layer-wise parameter optimization** and hardware-friendly nonlinear computation unit design
- **C4:** A highly flexible FPGA-based **overlay processor** for GNN models

C1: Unified PE array



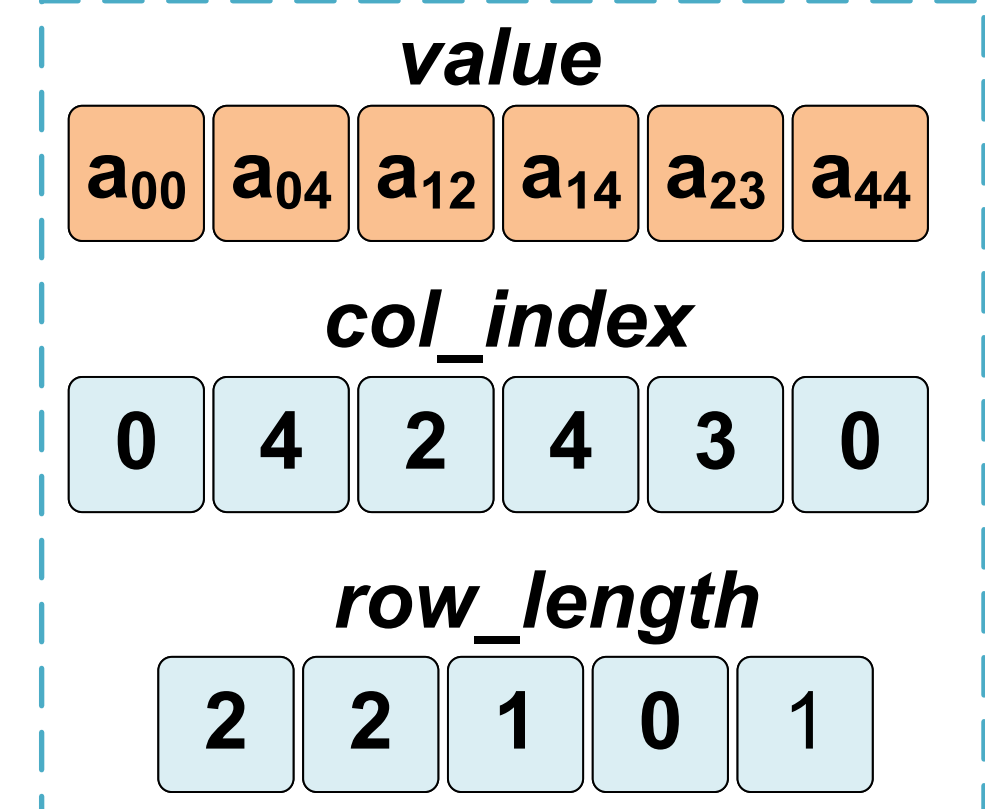
Conference Publication: DATE 2025

C2: Encoding format

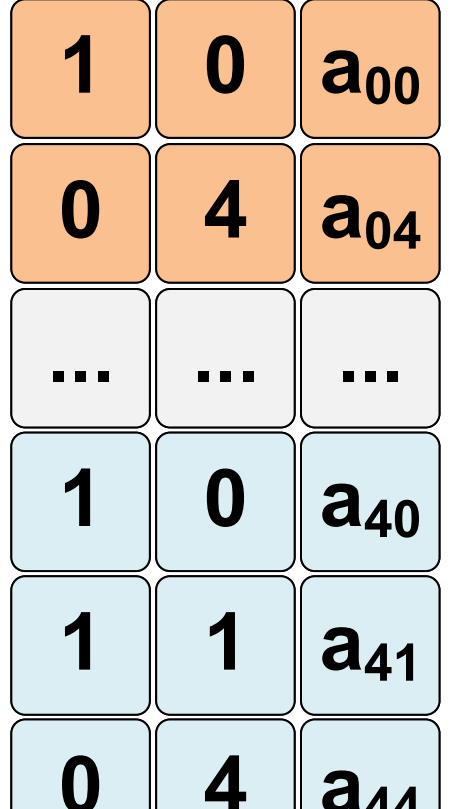
Sparse matrix



SCSR



OPCOO



Conference Publications: FCCM 2022, FPGA 2023, ISCAS 2023
Journal publication: ACM-TRETS

C3: layer-wise parameter optimization

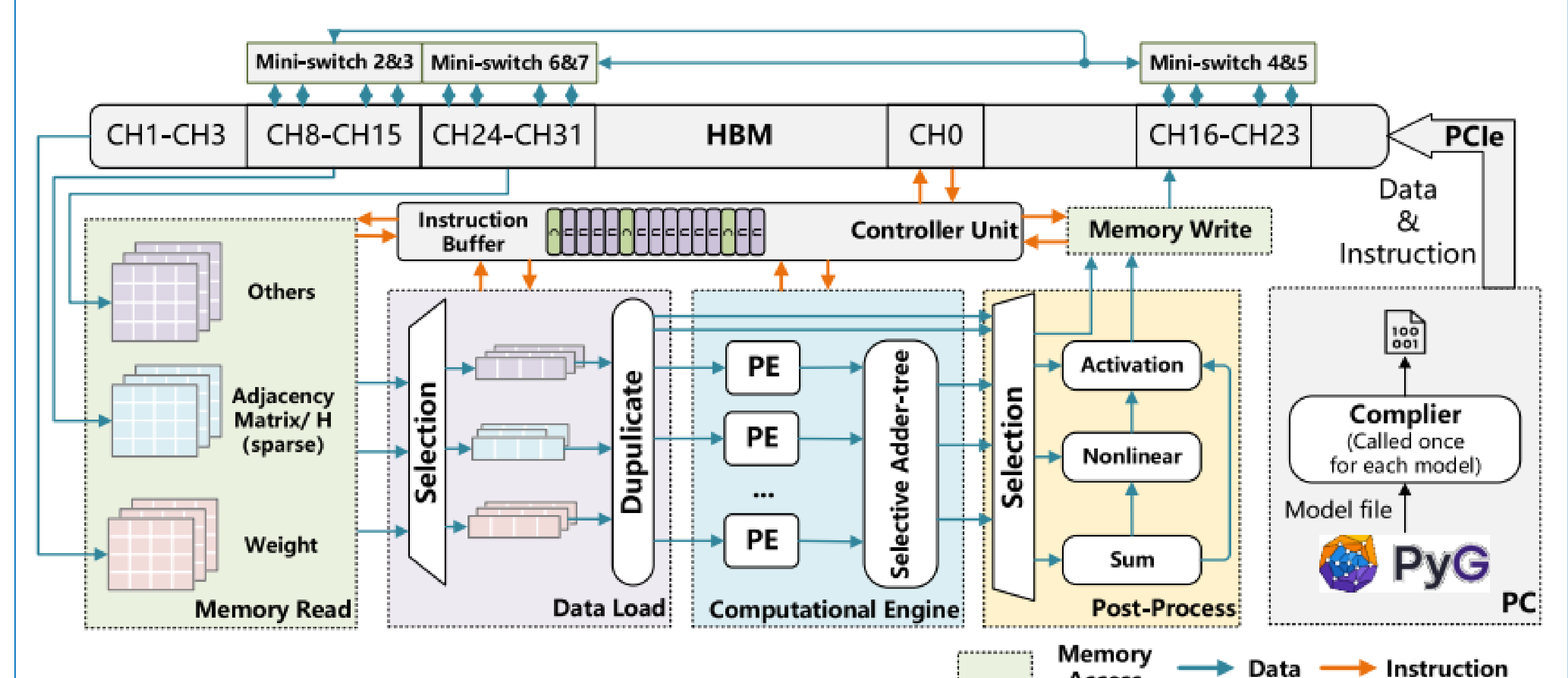
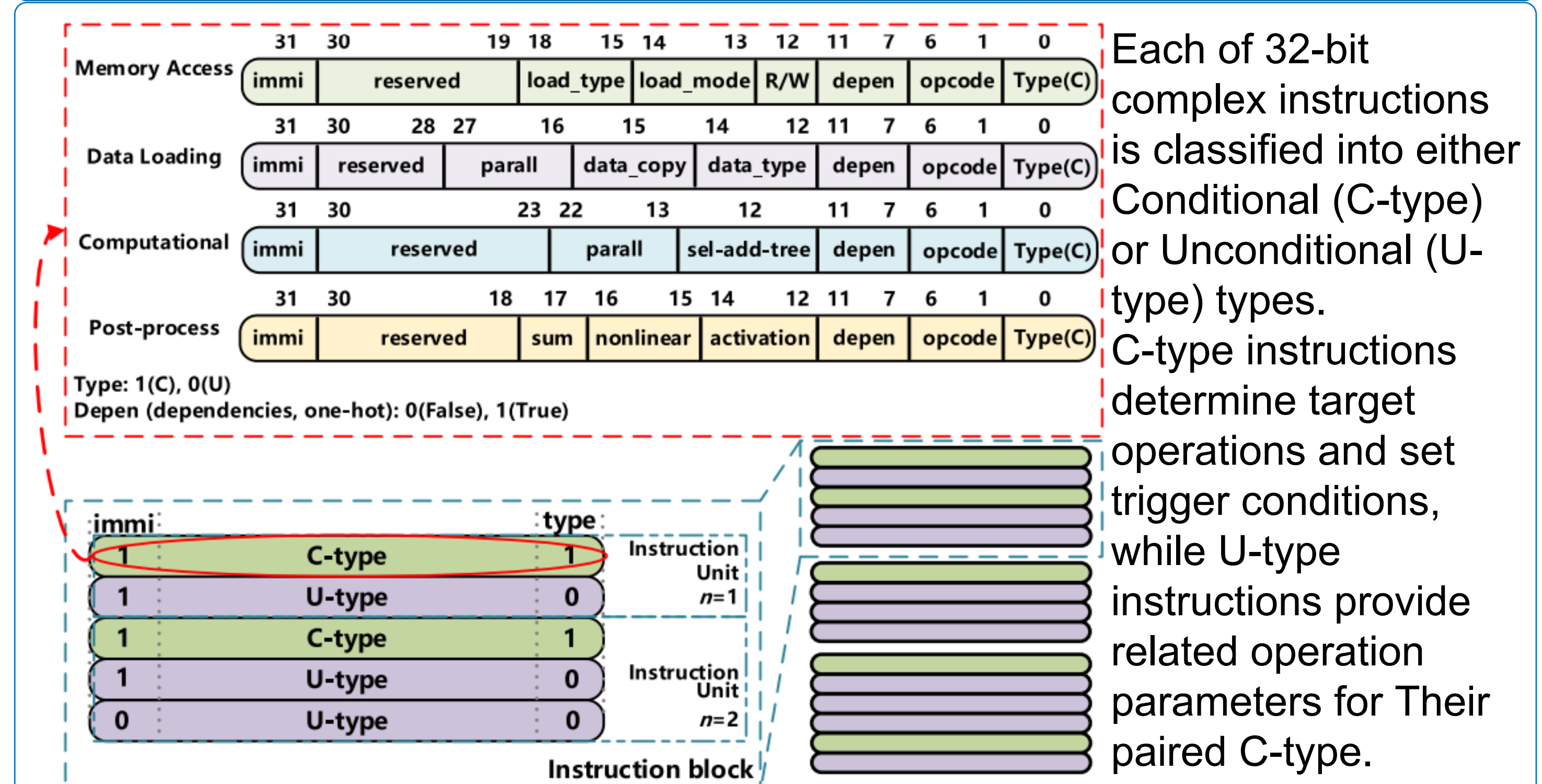
Constraint: $p_0 + p_1 + p_2 + \dots + p_k = 1, p_0 \neq 0, step = 0.1$

Generate parameter list: $L^i = (p_0^i, p_1^i, p_2^i, \dots, p_k^i)$

- 1: if $Result_{last} < Result_{current}$: then
- 2: $reward = 1$
- 3: else
- 4: $reward = -1$
- 5: end if
- 6: Q-table: $Q(s, a) = 0$; learning rate: $lr=0.1$; discount factor: $\gamma=0.9$;
- 7: $delta = abs(Result_{current} - Result_{last})$
- 8: while $delta < 0.001$ do
- 9: $LightGCN_train_function(p_0^i, p_1^i, p_2^i, \dots, p_k^i)$
- 10: $reward(s)$
- 11: end while
- 12: $Q(s, a) \leftarrow Q(s, a) + lr * \{reward + \gamma * max_{a'} [Q(s', a')] - Q(s, a)\}$
- 13: $s' \leftarrow s$; output Q-table, state
- 14: return optimized parameters

Conference Publication: DATE 2024

C4: Overlay processor for GCN



Conference Publications: FPGA 2023, FPL 2023
Journal Publication: ACM-TRETS

References

- R. Chen, J. Liu, S. Tang, Y. Liu, Y. Zhu, M. Ling, et al. "ATE-GCN: An FPGA-based Graph Convolutional Network Accelerator with Asymmetrical Ternary Quantization". In: 2025 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2025, pp. 1-7.
- R. Chen, Y. Ma, S. Zheng, S. Huang, C. Chen, J. Yu, et al. "Biological Activity Prediction of GPCR-targeting Ligands on Heterogeneous FPGA-based Accelerators". In: 2022 IEEE 30th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). IEEE, 2022, pp. 1-4.
- R. Chen, H. Zhang, S. Li, E. Tang, J. Yu, and K. Wang. "GraphOPU: A Highly Integrated FPGA-Based Overlay Processor for Graph Neural Networks". In: 2023 33rd International Conference on Field-Programmable Logic and Applications (FPL). IEEE, 2023, pp. 228-234.
- R. Chen, H. Zhang, Y. Ma, J. Chen, J. Yu, and K. Wang. "eSpMV: An Embedded-FPGA-based Hardware Accelerator for Symmetric Sparse Matrix-Vector Multiplication". In: 2023 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2023, pp. 1-5.
- R. Chen, H. Zhang, Y. Ma, E. Tang, S. Li, Y. Zhu, et al. "Graph-OPU: An FPGA-Based Overlay Processor for Graph Neural Networks". In: Proceedings of the 2023 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA). 2023, pp. 49-49.
- S. Li, R. Chen, E. Tang, Y. Liu, J. Yang, and K. Wang. "S-LGCN: Software-Hardware Co-Design for Accelerating LightGCN". In: 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2024, pp. 1-6.
- Y. Liu, R. Chen, S. Li, J. Yang, S. Li, and B. da Silva. "FPGA-Based Sparse Matrix Multiplication Accelerators: From State-of-the-Art to Future Opportunities". In: ACM Trans. Reconfigurable Technol. Syst., 17.4 (Nov. 2024).
- E. Tang, S. Li, R. Chen, H. Zhou, Y. Ma, H. Zhang, et al. "GraphOPU: A Highly Flexible FPGA-Based Overlay Processor for Graph Neural Networks". In: ACM Trans. Reconfigurable Technol. Syst., 17.4 (Nov. 2024).