Efficient Message Passing Architecture for GCN Training onHBM-based FPGAs with Orthogonal Topology On-Chip Networks

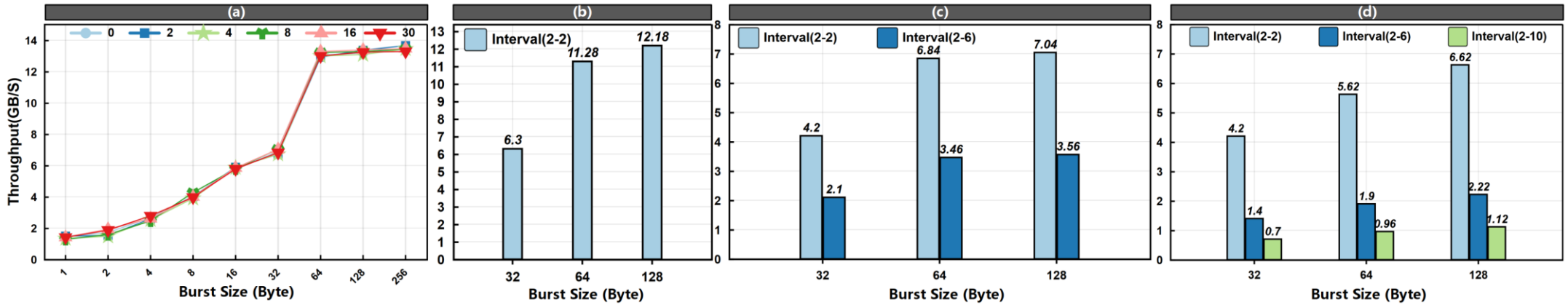
Qizhe WU, Letian ZHAO, Huawen LIANG, Jinyi ZHOU, Xiaotian WANG, Xi JIN

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41218-2

Problems & Ideas

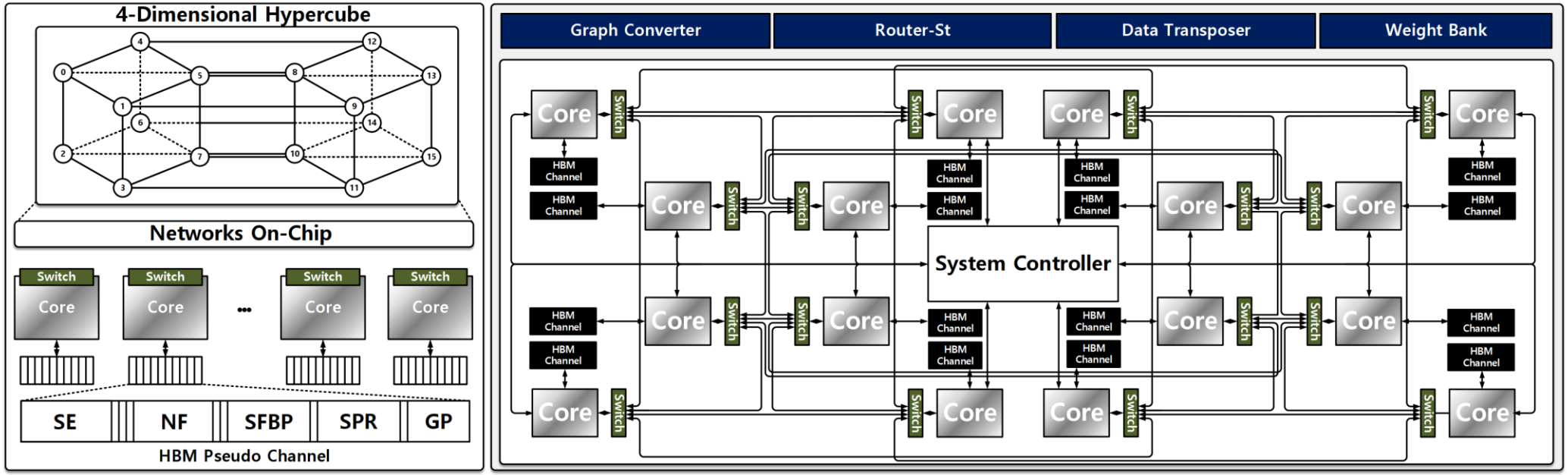
GCNs are widely adopted in graph-based deep learning for applications like recommendation systems and transportation networks, yet their training efficiency faces four key challenges :

- **Layer-State Retention Overhead**
- **Execution Order Sensitivity**
- **Shared Memory Contention** in SMP/UMA architectures caused by concurrent random accesses during aggregation, requiring optimized memory protocols; and
- **HBM Access Pattern Conflict (**as show in Fig.1**)** where sequential long-burst (combination phase) and random short-burst (aggregation phase) accesses create pseudo-channel contention, necessitating architectural solutions to resolve bandwidth scalability limitations. These bottlenecks highlight the need for hardware-algorithm co-optimization to advance GNN training efficiency.



Main Contributions

- We propose a dedicated multi-core GNN training accelerator for high-performance message passing on HBM-based FPGAs with NUMA-based memory access. The on-chip networks of the accelerator adopt a strictly orthogonal hypercube topology, and design a highly concurrent routing mechanism for GNN applications.
- We redesigned the dataflow of GNN backpropagation in the FPGA. It can decrease the off-chip memory requirements during training while reducing additional matrix transposes.
- We evaluate the performance on Xilinx UltraScale+ VCU128. In comparison to the state-of-the-art GNN training architecture HP-GNN we achieved a improvement of x1.03~1.81.



(a) Overall architecture *(b) Architectural detail*

SE : Subgraph Edge **NF : Node Feature** **SFBP : Save For Back Propagation** **SPR : Subgraph Partial Results** **GP : Global Parameter**