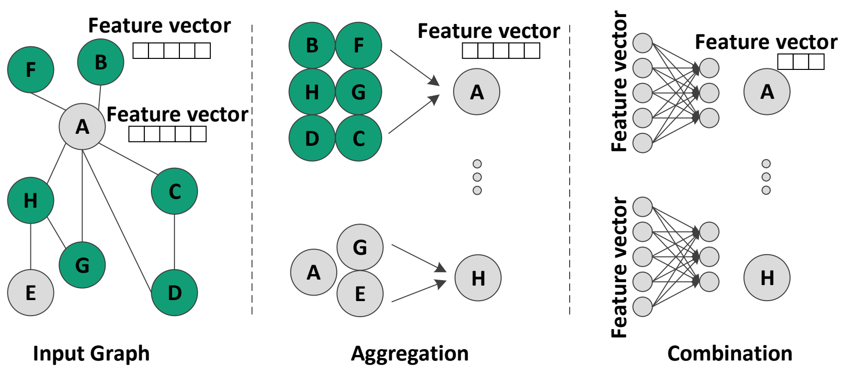
Li JJ, Wang K, Zheng H *et al*. GShuttle: Optimizing memory access efficiency for graph convolutional meural network accelerators. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 38(1): 115−127 Jan. 2023. DOI: 10.1007/s11390-023-2875-9

GShuttle: Optimizing Memory Access Efficiency for Graph Convolutional Neural Network Accelerators

Jia-Jun Li (李家军), Ke Wang (王可), Hao Zheng (郑皓)
and Ahmed Louri

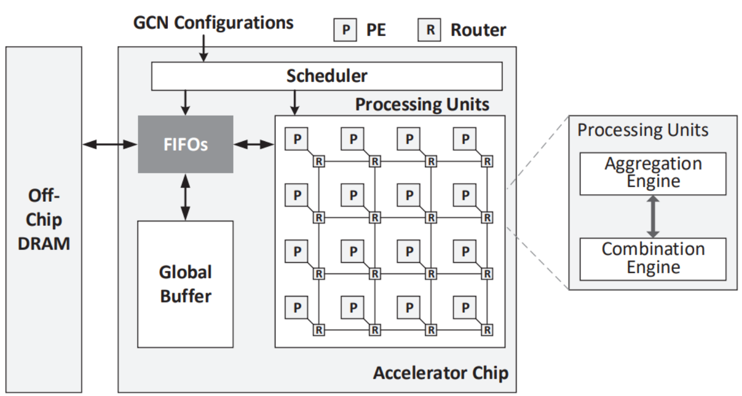
Research Objectives

GCNs are popular but poses a great challenge to the processing hardware



- **Graphs can be very large**
  - Billions of users in social networks
- **Stringent Latency requirements**
  - Power grid cascading failure prediction
- **High Throughput requirements**
  - E-commerce analysis on shopping season

GCN accelerators failed to exploit all the opportunity of data reuse



A typical GCN accelerator

Optimizing memory access efficiency is the key to energy efficiency for GCN accelerators

Research Method

- **Define the optimization problem and build analytical models for memory accesses**

$$
\begin{aligned}
\underset{\mathbb{X}}{Minimize} \quad & V = V_d(\mathcal{X}^t, \mathcal{X}^{oo}, \mathcal{X}^f) + \omega \cdot V_s(\mathcal{X}^u, \mathcal{X}^{oi}) \\
s.t. \quad & 0 < T_m \le M, \quad 0 < T_k \le K \\
& 0 < T_{n0} \le N, \quad 0 < T_{n1} \le N \\
& 0 < T_{c0} \le C, \quad 0 < T_{c1} \le C \\
& S_X + S_W + S_{B1} <= GLBsize \\
& S_A + S_O + S_{B2} <= GLBsize \\
& P_{n0} \times P_{c0} \times P_k \le \#PEs
\end{aligned}
$$

$$
\begin{cases}
S_X = \gamma_X \times T_{n0} \times T_k \\
S_W = T_k \times T_{c0} \\
S_{B1} = T_{n0} \times T_{c0} \\
S_{B2} = T_{n1} \times T_{c1} \\
S_A = \gamma_A \times T_m \times T_{n1} \\
S_O = T_m \times T_{c1}
\end{cases} \tag{4}
$$

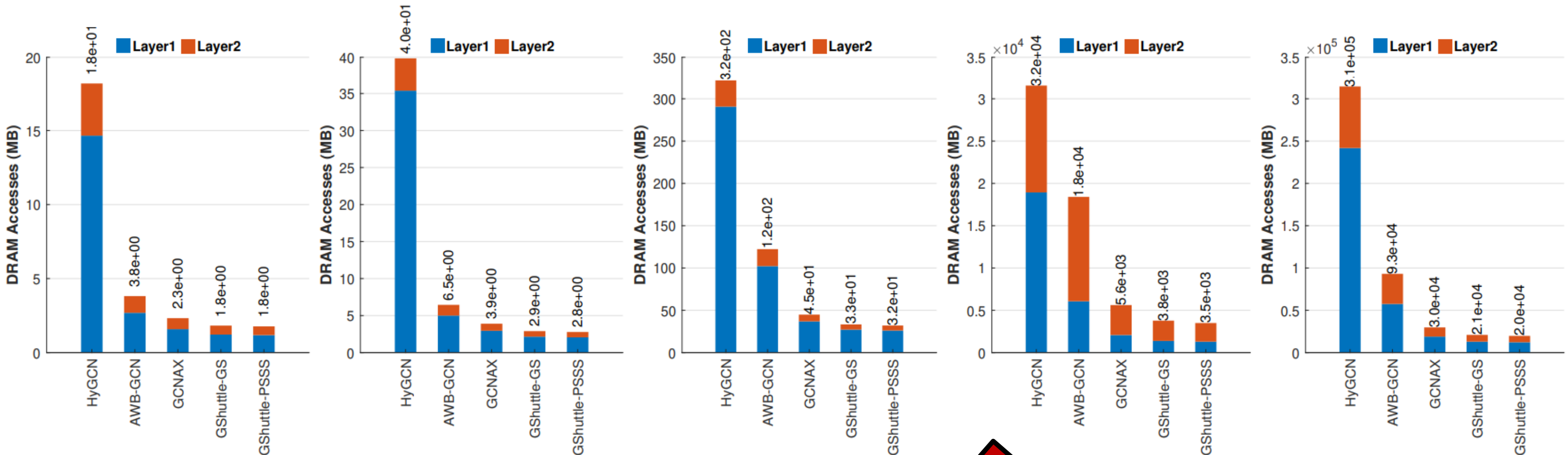
$$
\begin{cases}
\alpha_X = \alpha_W = \dfrac{N}{T_{n0}} \times \dfrac{C}{T_{c0}} \times \dfrac{K}{T_k} \\
\alpha_{B1} = \dfrac{N}{T_{n0}} \times \dfrac{C}{T_{c0}} \\
\alpha_{B2} = \alpha_A = \dfrac{M}{T_m} \times \dfrac{C}{T_{c1}} \times \dfrac{N}{T_{n1}} \\
\alpha_O = \dfrac{M}{T_m} \times \dfrac{C}{T_{c1}}
\end{cases} \tag{5}
$$

- **Develop two algorithms to solve the problem**

Table 2. The greedy search algorithm to determine the design variables.

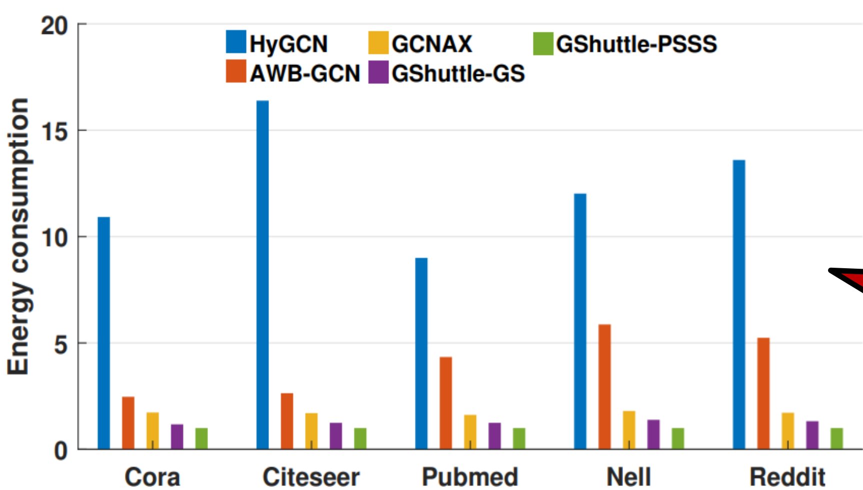
Conditions	Loop Fusion	Inter-tiling Loop Order	Tile Size Setting Priority
$N \cdot C \ge GLB_{size}$	No	$n_0 \to c_0 \to k$, $m \to c_1 \to n_1$	① T_{n0}, Tm ② T_{c0}, T_{c1} ③ T_{n1}, T_k
$N \cdot C < GLB_{size}$	Yes	$n_0 \to c_0 \to k \to m$	① T_{n0}, T_{n1} ② T_{c0}, T_{c1} ③ T_m, T_k

Research Results



(a) Cora (b) Citeseer (c) Pubmed (d) Nell (e) Reddit

GShuttle reduces DRAM accesses up to a factor of 1.7X compared to the SOTA approaches



GShuttle improves energy efficiency by 1.4X compared to the SOTA approaches

Research Conclusions

- GShuttle can find the optimal design variables of GCN dataflow under certain design constraints.
- The results show that GShuttle could significantly reduce the number of DRAM and SRAM accesses for GCN Accelerators.
- we expect that GShuttle can be applied to many existing GCN accelerators such as HyGCN and AWB-GCN.