Graph Partitioning for Near Memory Processing
Chenfeng Zhao  Roger D. Chamberlain  Xuan Zhang
McKelvey School of Engineering, Washington University in St. Louis

Background and Motivation

Graph processing applications have a high memory bandwidth requirement.

Near-Memory Processing (NMP) architectures based on multiple 3D memory cubes are proposed to accelerate parallel graph processing applications.

However, cross-cube communication is a system bottleneck, taking a significant portion of execution time (12%-78%) and energy consumption (12%-73%).

SuperCut

In this work, we propose SuperCut, a co-design framework for near-memory graph processing.

Programming Model

We propose a three-phase programming model supporting the partitioning algorithms.

It explicitly handles computation and communication via user-defined functions.

Input: The SuperCut graph \( H \) and original graph \( G \)

Output: Results of graph processing applications

1. for each original vertex \( v_{org} \in G \) do
2. \( \text{gather}_{\text{combine}}(v_{org}) \)
3. end do
4. for each destination-cut vertex \( v_{dc} \in H \) do
5. \( \text{update} \leftarrow \text{gather}_{\text{combine}}(v_{dc}) \)
6. \( \text{scatter}(\text{update}) \)
7. end do
8. for each original vertex \( v_{org} \) and replica \( v_r \) do
9. \( \text{apply}(v_{org}); \text{apply}(v_r) \)
10. end do

Graph Partitioning algorithms

SuperCut adopts a set of partitioning algorithms to preprocess graph datasets, including mixed-cut partitioning, a stochastic-and-heuristic-based optimization algorithm and partial graph partitioning.


Evaluation Results

Case study: Average Teenage Follower (ATF) and Breadth-First Search (BFS) on 5 graph datasets.

Energy Evaluation: SuperCut achieves 1.09x to 2.6x total energy reduction relative to the state-of-the-art.

Performance Evaluation: SuperCut achieves 1.12x to 2.6x speedup relative to the state-of-the-art.
