

MGG: Accelerating Graph Neural Networks with Fine-grained Intra-kernel Communication-Computation Pipelining on Multi-GPU Platforms

Yuke Wang, Boyuan Feng, Zheng Wang,
*Tong Geng, ^Kevin Barker, ^Ang Li, Yufei Ding.

*: *University of Rochester*, ^: *Pacific Northwest National Laboratory*

University of California, Santa Barbara

Graphs are everywhere, and GNN is the key!



Social Networks



Financial Services

Power Grid

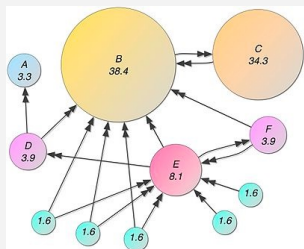


Molecular Biology

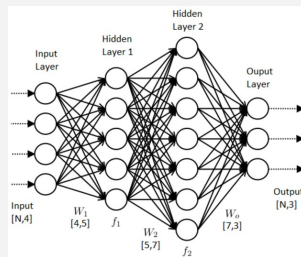


GNN: Graph Neural Networks

GNN Layer =

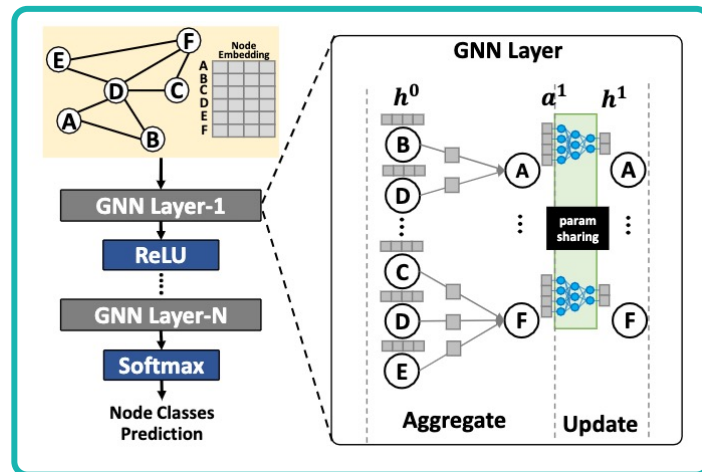


+



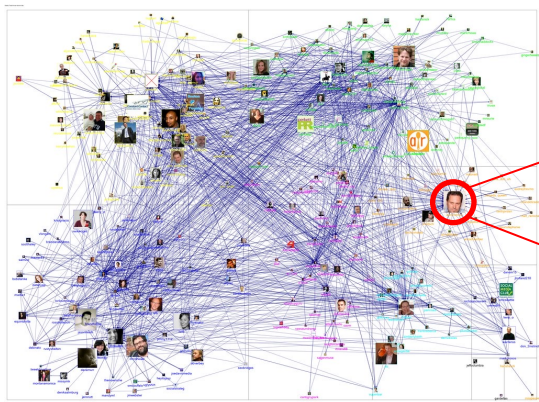
Operation view

Model view



GNNs are Scaling Up!

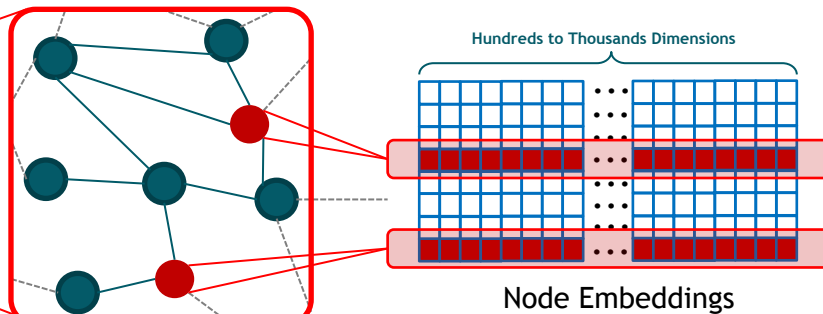
➤ 1. Graph Structural Scaling.



Rich structural information
(e.g., Community) for various
tasks (e.g., link prediction)

ogbn-paper100M has
111M nodes and 1.6B edges

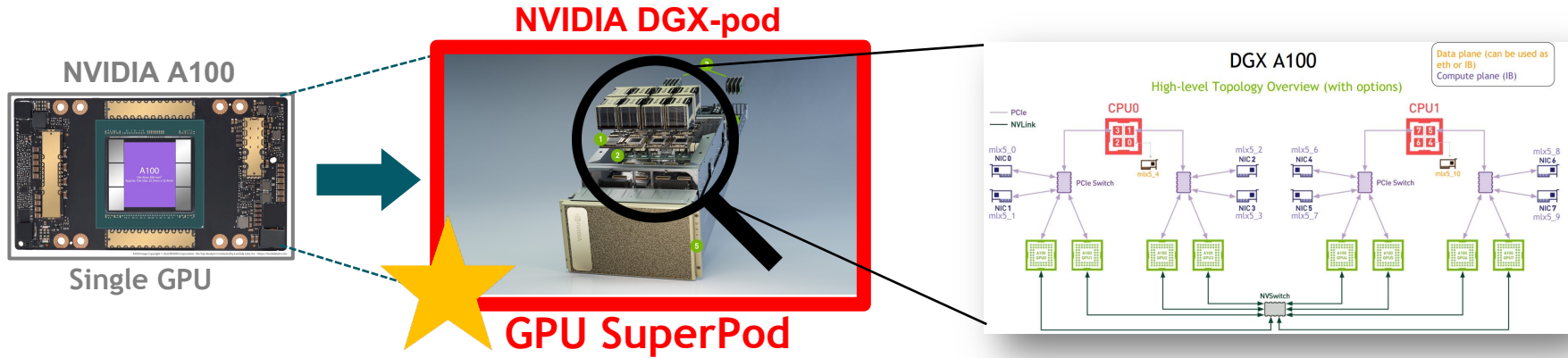
➤ 2. Graph/Node Embedding Scaling.



Fruitful Node/Graph-level
Properties for various tasks
(e.g., node classification)

Reddit graphs has
602 embedding dimension

DL Infrastructures are Scaling to Catch Up!



High Comp/Mem
Capacity

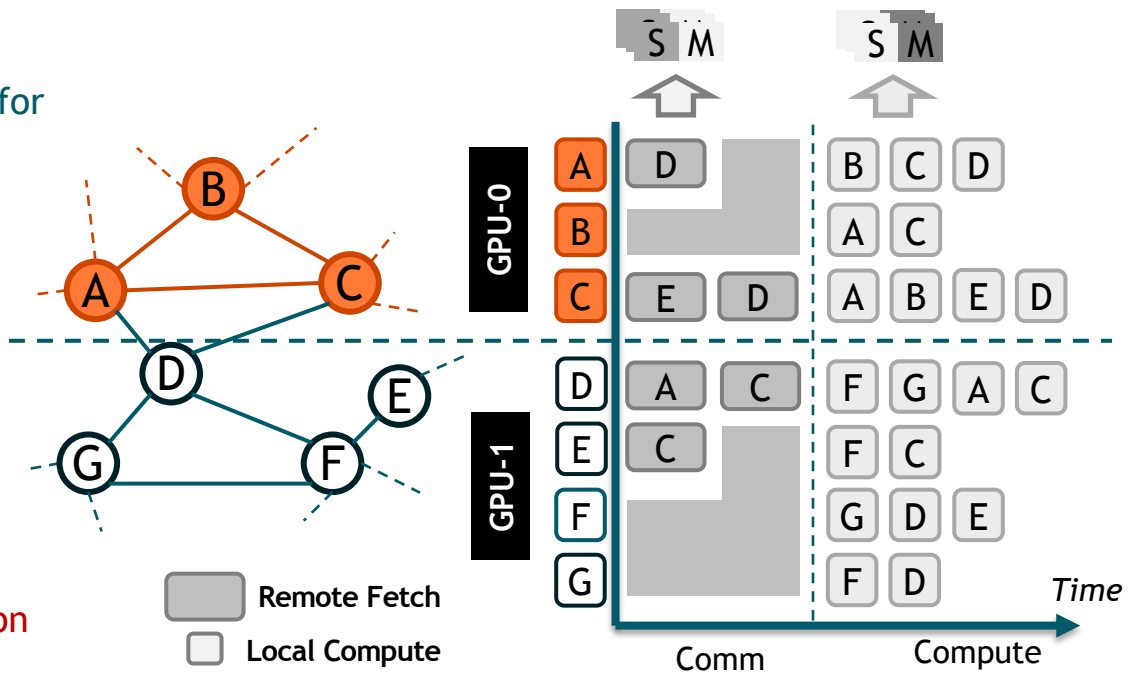
High Comm.
Bandwidth

Building Blocks
for Large Clusters

Powerful Multi-GPU Platforms Cannot Solve Everything!

➤ Single-GPU solution does not work well for multiple GPUs. (e.g., DGL):

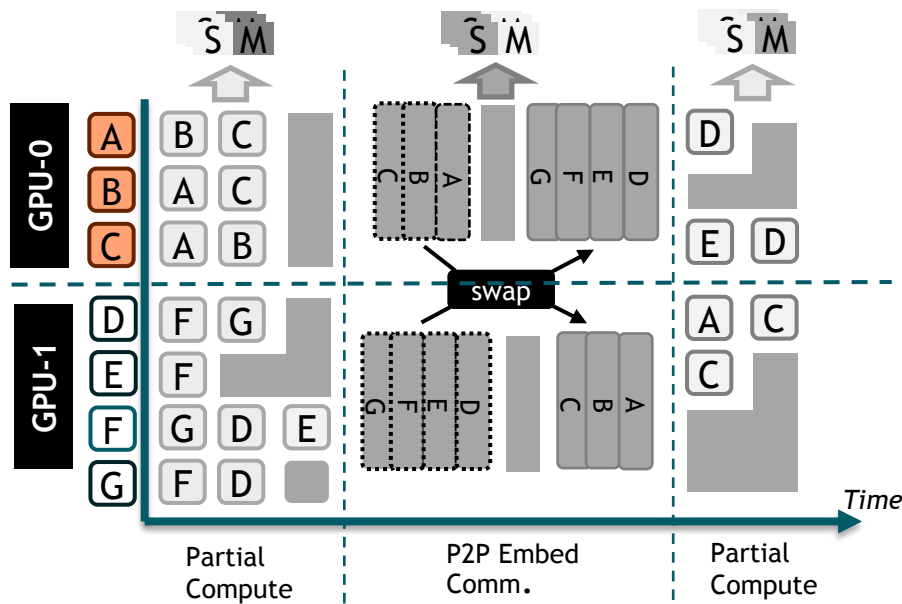
- Communication and computation in separated phases.
- Remote neighbor access are fine-grained and irregular.



Weakness:

- High individual neighbor access cost.
- GPU idleness between the computation and communication phases.

Traditional Distributed Graph Solutions Do Not Work Well!



➤ Schedule Transformation for Dense Communication (e.g., NeuGraph, P3, ROC):

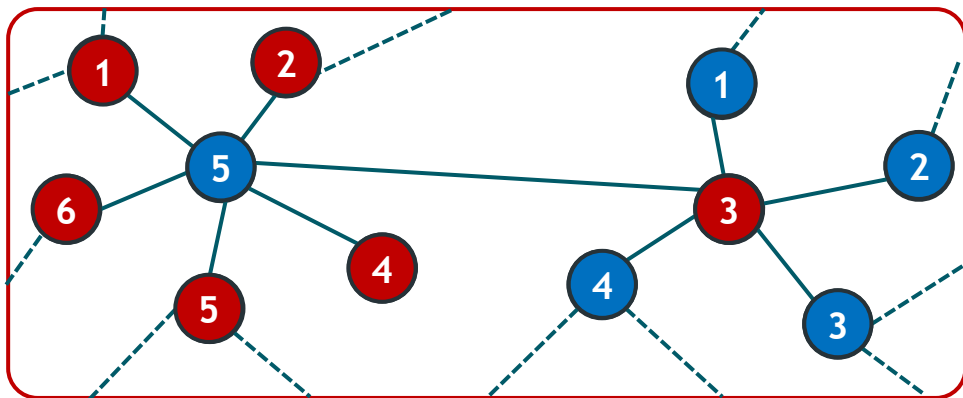
- Neighbor aggregation is divided into multiple rounds.
- Neighbor movement are dense, regular and coarse-grained.
- Neighbor access in each round of aggregation is all local.

Weakness:

- Additional algorithmic modification.
- Redundant data movements.
- Decreased computation efficiency,

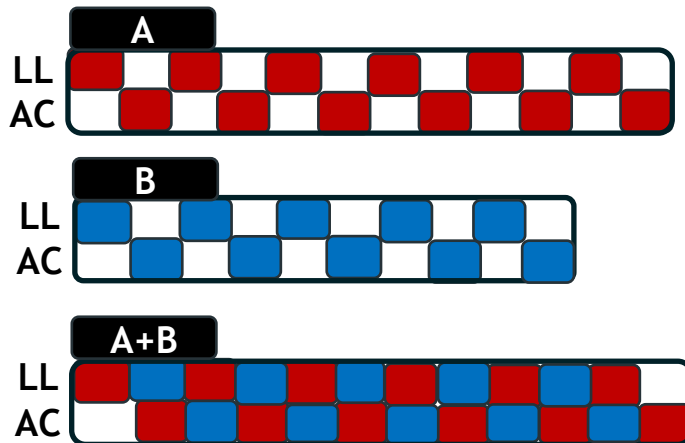
- Ma, Lingxiao, et al. "NeuGraph: Parallel Deep Neural Network Computation on Large Graphs." *USENIX Annual Technical Conference*. 2019.
- Gandhi, Swapnil, and Anand Padmanabha Iyer. "P3: Distributed Deep Graph Learning at Scale." *OSDI*. 2021.
- Jia, Zhihao, et al. "Improving the accuracy, scalability, and performance of graph neural networks with roc." *Proceedings of Machine Learning and Systems 2* (2020).

Algorithm-Observation: Opportunity for Fine-grained Pipelining



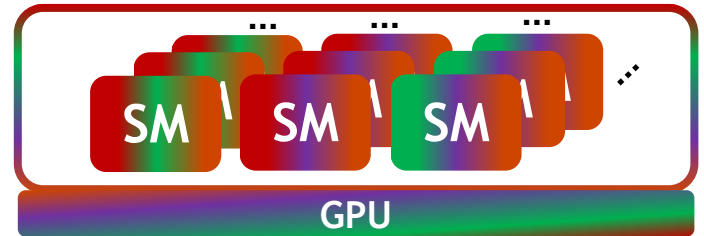
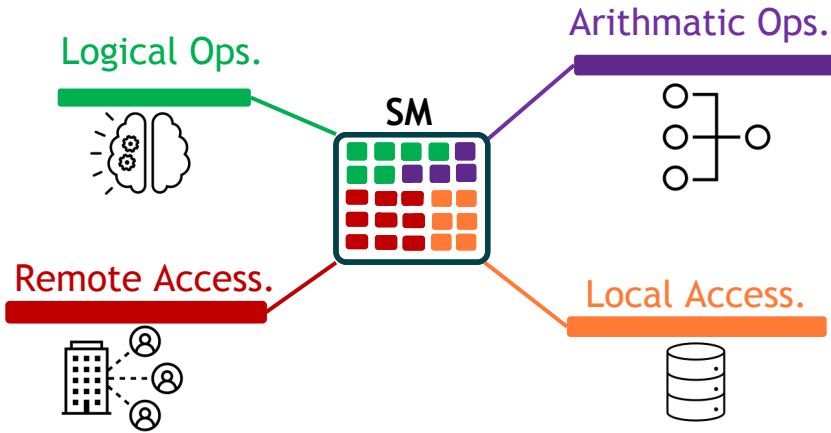
Fine-grained neighbor aggregation dependency.

LL: Load Local Neighbor
AC: Aggregate Computation

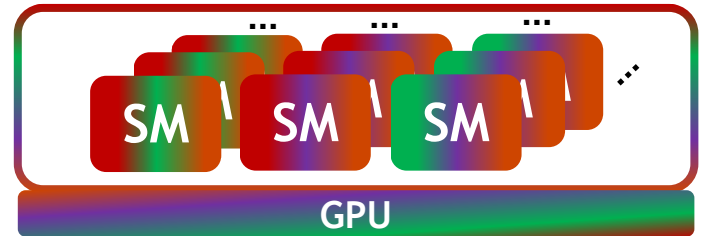
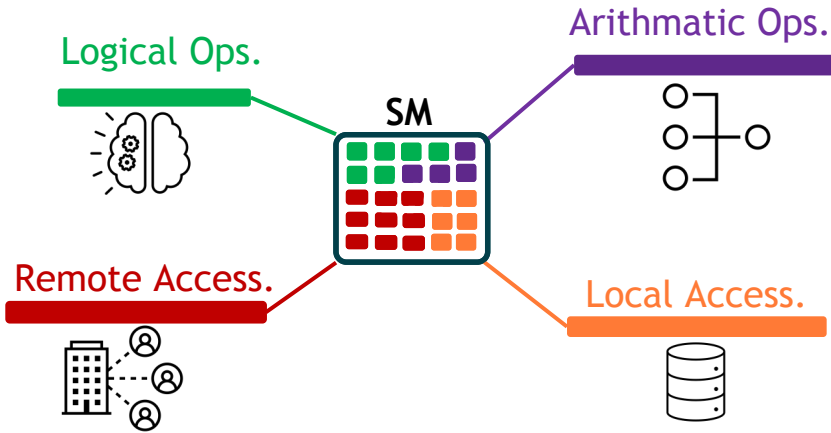


New opportunities: we can amortize communication costs by **fine-grained overlapping** neighbor aggregation from different nodes.

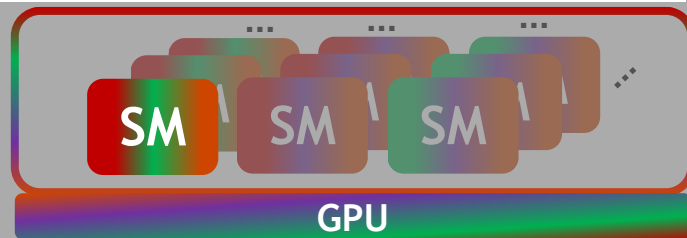
Hardware-Observation: SM Multiplexing for Operation Overlapping



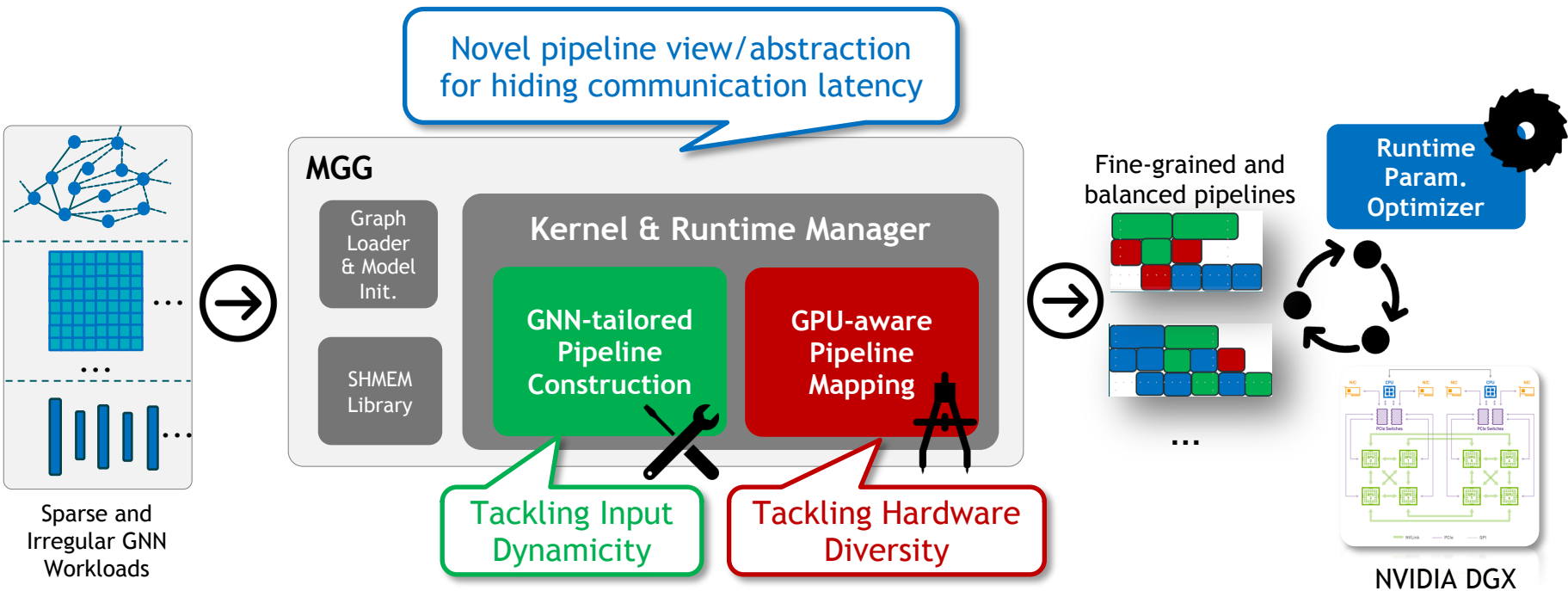
Hardware-Observation: SM Multiplexing for Operation Overlapping



Hardware-Observation: SM Multiplexing for Operation Overlapping



MGG Overview

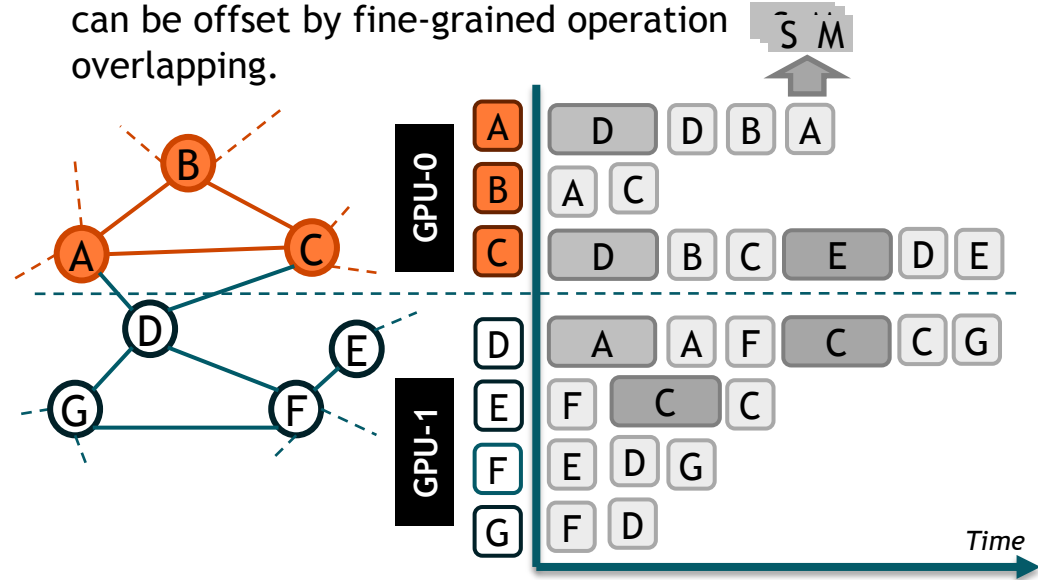


Contribution-1: Pipeline View and Abstraction.

Neighbor
Access
Latency



Key insight: Communication overhead can be offset by fine-grained operation overlapping.

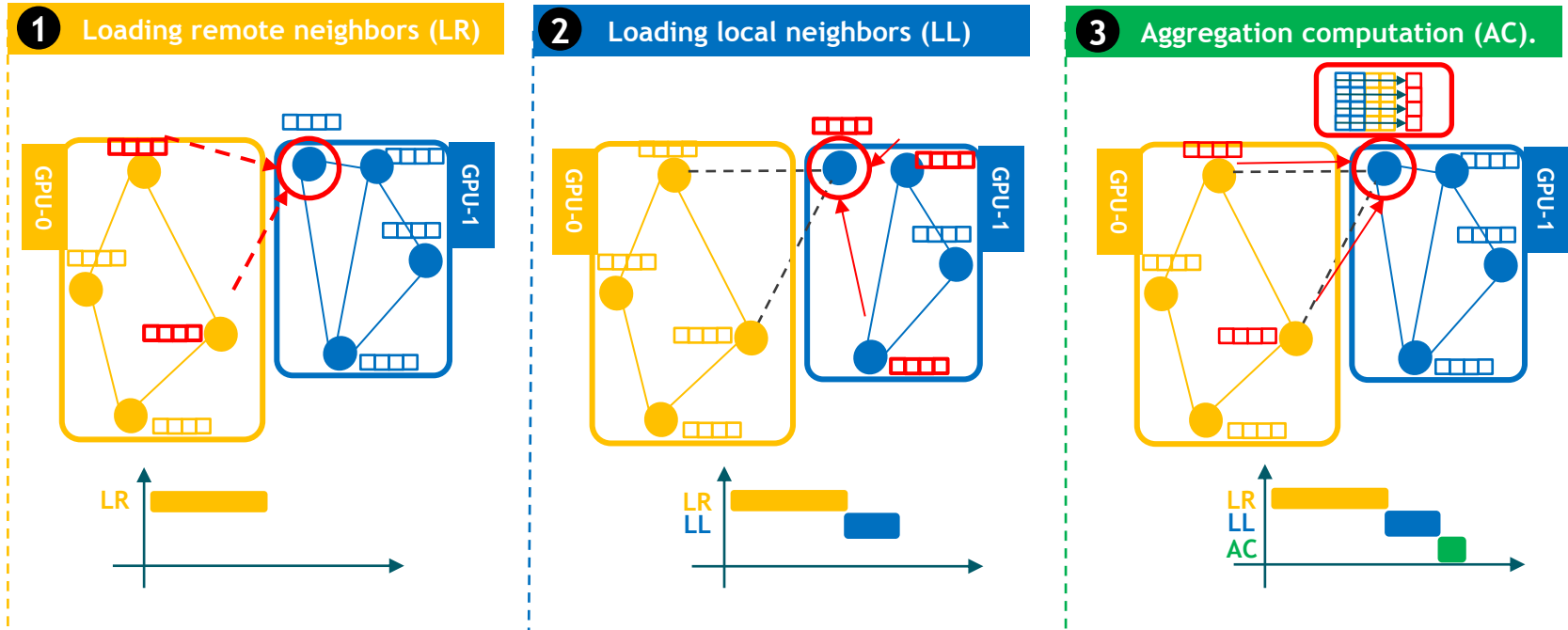


Challenges:

- Communication overweight the local computations/access and dominate the execution.
- Communication exacerbate the workload imbalance.

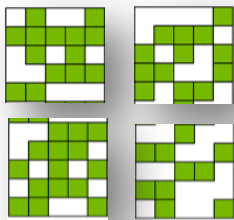
- **The idle cycles** of GPUs communication can be fulfilled by other local computing.
- Multi-GPU GNN workload can be abstracted as a **fine-grained dynamic** software pipeline.

A three-stage dynamic software pipeline.



Contribution-2: Pipeline-aware Workload Management.

Input
Dynamycity



Challenge: Input diversity (e.g., graph size/sparsity) would affect the pipeline efficiency (e.g., bubble ratio).

Key Insight: Pipeline can be tailored for maximizing efficiency based on diverse GNN inputs properties (e.g., #nodes, #edges and node degree).

Heterogeneity & Granularity-aware
pipeline enhancement.

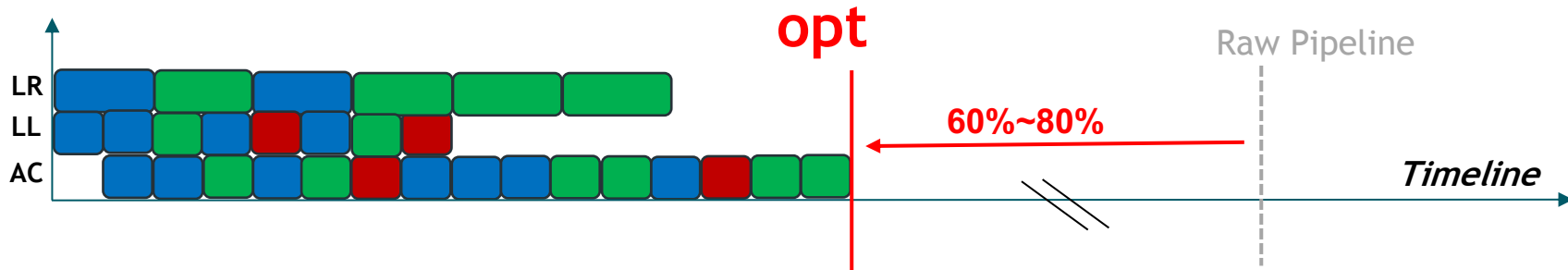
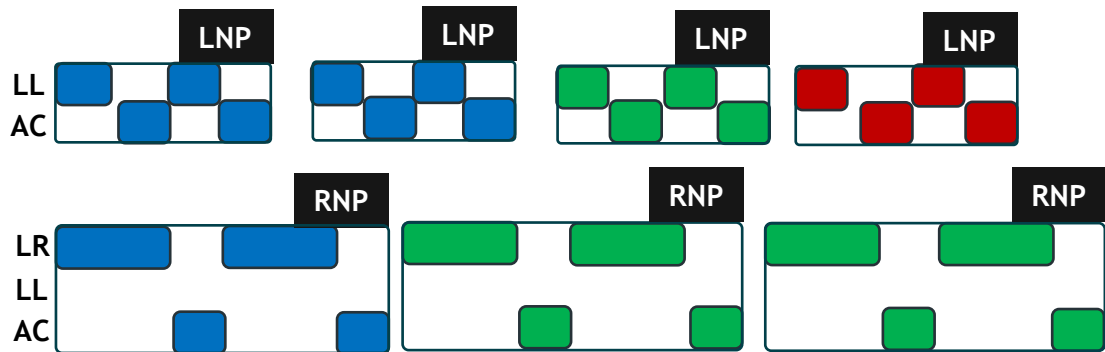
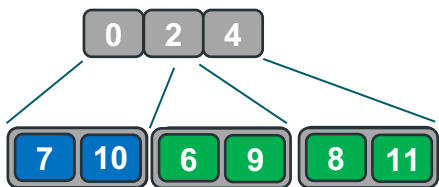
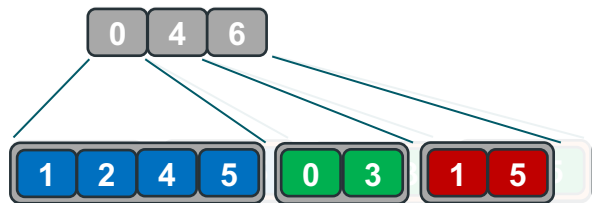


Heterogeneity & Granularity-aware Pipeline Enhancement.



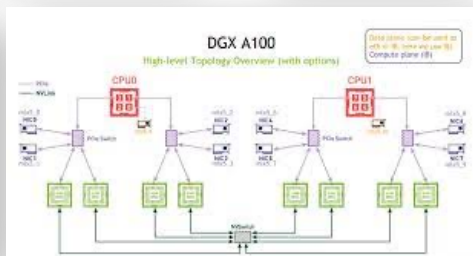
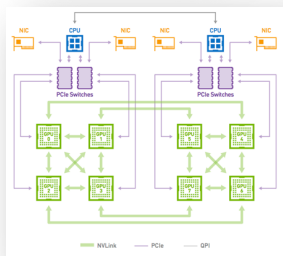
Facilitate a more balanced workload distribution among pipeline stages

LNP: Local Neighbor Partition
RNP: Remote Neighbor Partition



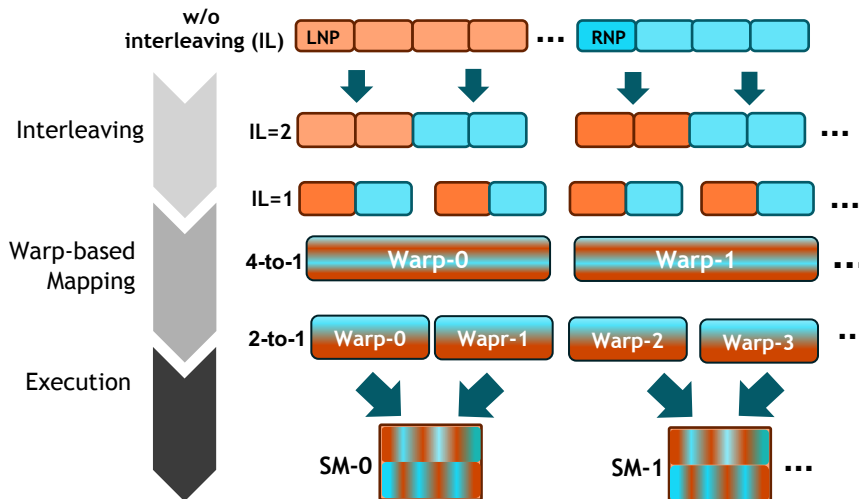
Contribution-3: GPU-Aware Pipeline Mapping.

Hardware Diversity



Challenges: Hardware diversity (e.g., Comp/Comm Speed) would affect the pipeline execution performance (e.g., SM utilization and occupancy).

Key insight: Dynamically configurable pipeline-workload-to-SM mapping can maximize pipeline execution performance.



Intelligent Runtime Design.

Specialized Memory Design & Optimization.

Evaluation

Dataset	#Vertex	#Edge	#Dim	#Class
reddit(RDD) [45]	232,965	114,615,892	602	41
enwiki-2013(ENWIKI) [23]	4,203,323	202,623,226	300	12
it-2004 (IT04) [10]	41,291,594	1,150,725,437	256	64
ogbn-paper100M(PAPER) [12]	111,059,956	1,615,685,872	128	64
ogbn-products(PROD) [17]	2,449,029	61,859,140	100	47
ogbn-proteins(PROT) [17]	132,534	39,561,252	8	112
com-orkut(ORKT) [23]	3,072,441	117,185,083	128	32

➤ GNN Models.

- ❖ Graph Convolutional Network (GCN):
2 layers with 16 hidden dimensions.
- ❖ Graph Isomorphism Network (GIN):
5 layers with 64 hidden dimensions.

➤ Evaluation Platform.

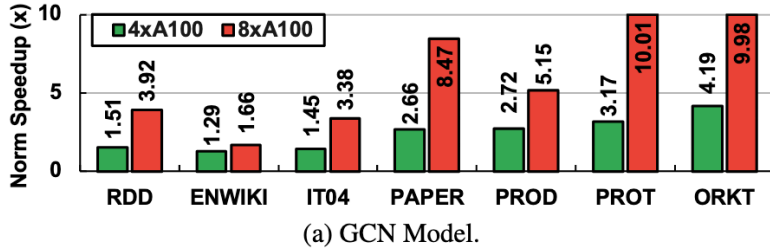
- ❖ NVIDIA DGX-A100 with dual AMD Rome 7742 processors (each with 64 cores, 2.25 GHz), 1TB host memory, and 8x(A100-40GB) connected via NVSwitch.

➤ Baseline.

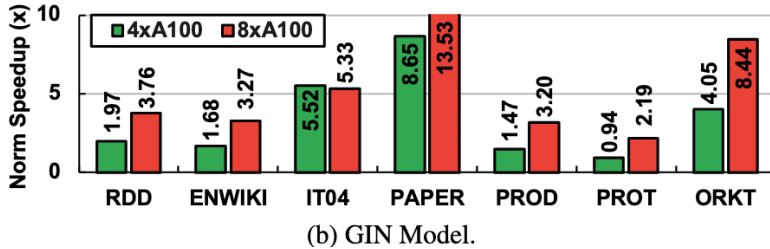
- ❖ Deep Graph Library [ICLR'19].
- ❖ Unified Memory [NVIDIA].
- ❖ ROC [MLSys'20].

Evaluation: Overall Performance

- Compare with DGL.
(Separated Communication and Computation)



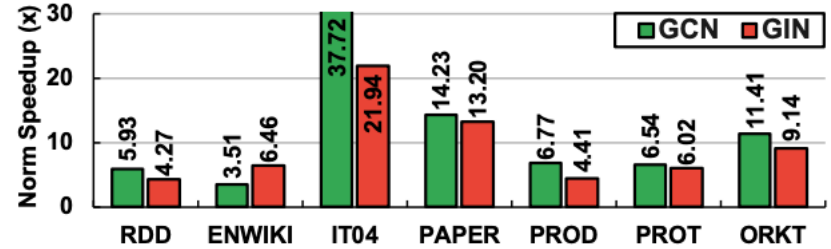
(a) GCN Model.



(b) GIN Model.

Averaged 4.41x speedup in comparison with DGL on GCN and GIN.

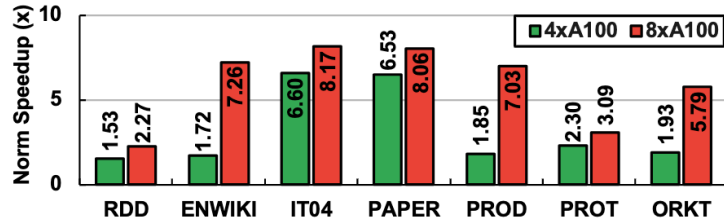
- Compare with ROC.
(Schedule Transformation for Dense Communication)



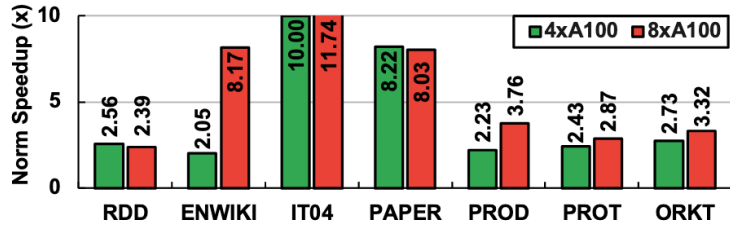
Averaged 10.83x speedup in comparison with ROC (8xA100) on GCN and GIN.

Evaluation: Additional Comparisons

- Compare with MGG-UVM.
(Pipeline with Coarse-grained Communication)



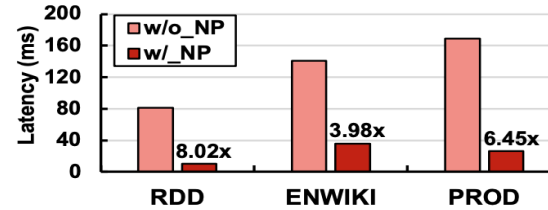
(a) GCN Model.



(b) GIN Model.

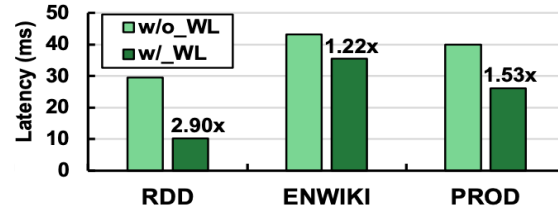
Averaged 4.81x speedup in comparison with MGG-UVM on GCN and GIN.

- Neighbor Partitioning (input Dynamicity)



Averaged 2.24x with NP

- Workload Interleaving (Hardware Diversity)



Averaged 1.88x with WL

Contribution Summary

✓ Exploiting the joint optimization of the communication and computation.



A novel and unique multi-stage pipeline view/abstraction

✓ Capitalizing pipelining benefits for input dynamicity.



GNN-tailored pipeline construction

✓ Enhance pipeline efficiency for diverse hardware.



GPU-aware pipeline mapping



Thank You

Q & A



Contact: yuke_wang@cs.ucsb.edu



Code: <https://github.com/YukeWang96/MGG-OSDI23-AE.git>

Acknowledgement

1. U.S. DOE Office of Science, Office of Advanced Scientific Computing Research, under Award 66150: "CENATE - Center for Advanced Architecture Evaluation". The Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830.
2. NSF Award 2124039 and Cloudbank.
3. Amazon Faculty Research Award 2021 for Yufei Ding.
4. NVIDIA Graduate Fellowship 2022-2023 for Yuke Wang.



U.S. DEPARTMENT OF
ENERGY

