gSuite: A Flexible and Framework Independent Benchmark Suite for Graph Neural Network Inference on GPUs

T. Tekdoğan, S. Göktaş, A. Yilmazer-Metin

İTÜ İSTANBUL TEKNİK ÜNİVERSİTESİ 1773

aselsan

IEEE International Symposium on Workload Characterization (IISWC) 2022, Austin, TX, USA - November 6-8 2022

Outline

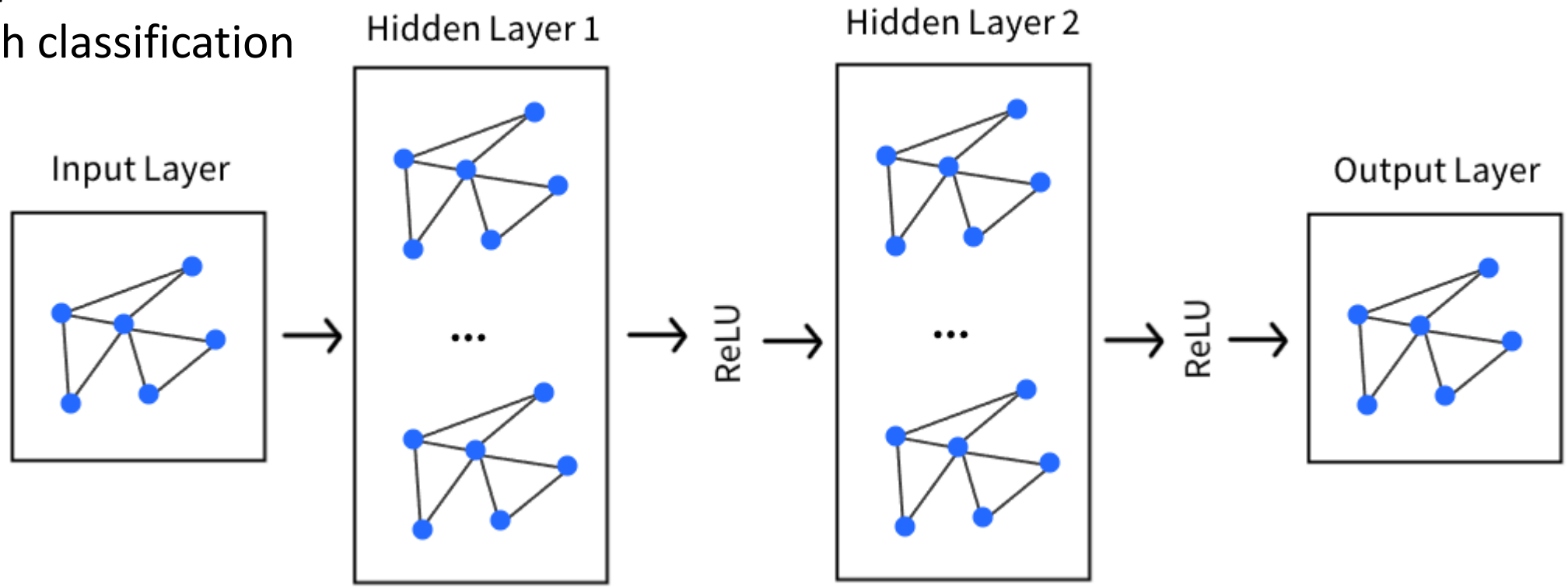
- Why Graphs?
- Why GNNs?
- Background
- Limitations
- Our Design Approach
- Evaluation
- Conclusion
- Future Work

Why Graphs?

- Able to Express
  - Non-euclidean data
  - Relationship among objects
  - Topology of networks

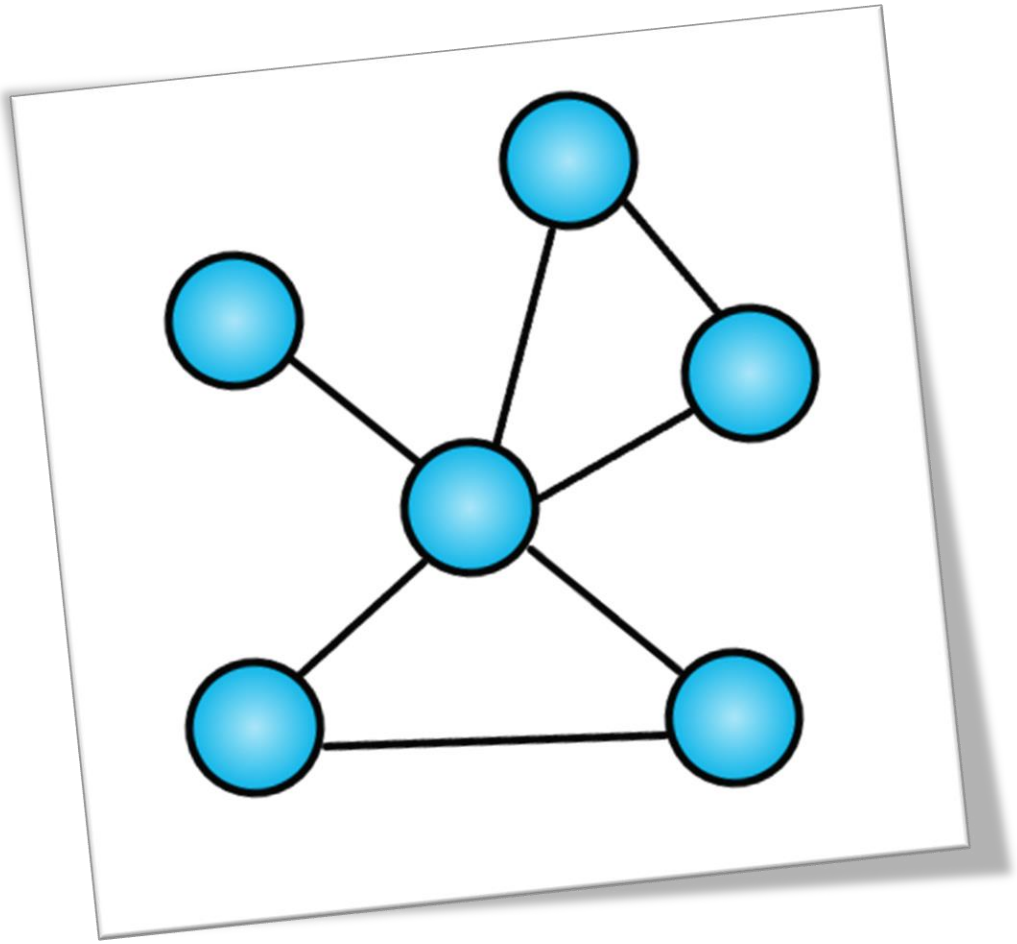
Why Graph Neural Networks?

- Can learn from graph data
- Many tasks are included:
  - Node prediction
  - Link prediction
  - Graph classification



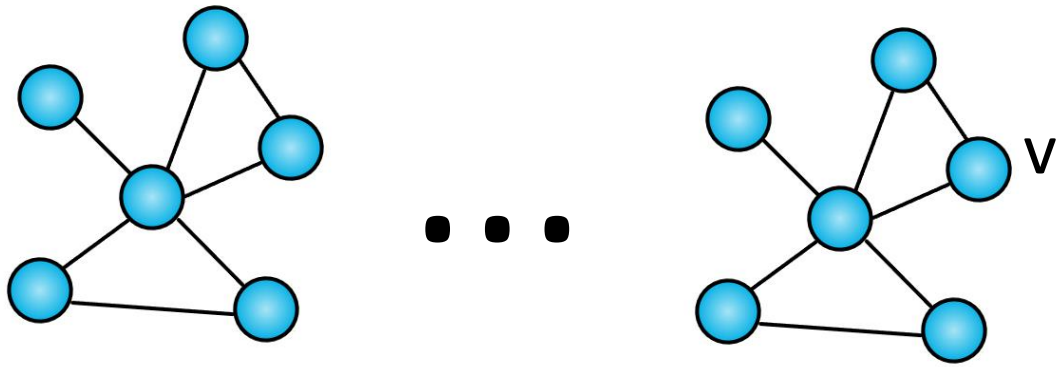
Background

- Graphs
  - Graph $\mathcal{G} = \{\mathcal{V},\mathcal{E}\}$
  - Nodes $\mathcal{V} = \{v1,v2,...,vn\}$
  - Edges $\mathcal{E} = \{e1,e2,...,em\}$
  - Number of nodes n
  - Number of edges m



Background

- Graphs
  - Current GNN Layer k ϵ [1,L]
  - Feature Representation $h_v^{(k)}$
  - Edge Representation $g_e^{(k)}$
  - Neighbour nodes of node v, N(v)



Background

Computation of GNNs

- Consists of two stages:
  - **Inference**
  - Training
- Two phases of Inference:
  - Aggregation (message)
  - Combination (update)

Background

- Frameworks
  - PyTorch Geometric (PyG) 
  - Deep Graph Library (DGL) 

  - Graph Nets 
  - Spektral 

Background

- GNN Models
  - Graph Convolutional Networks (GCNs)
  - Graph Isomorphishm Networks (GINs)
  - GraphSAGE (SAGs)
  - Spatio-Temporal GCNs (STGCNs)
  - ...

Limitations

Study Name	GNN Models	Frameworks	Datasets	Extendibility	GNN Scope
Pytorch Geometric [18]	GCN, SAG, GIN, RGCN, ...	Pytorch	Cora, CiteSeer, Pubmed, MUTAG, PROTEINS, ...	Yes	Both
Deep Graph Library [61]	GCN, GAT, SAG, GIN, SGC, ...	Pytorch, MXNet, Tensorflow	REDDIT, ARXIV, PROTEINS, ...	Yes	Both
GCN-GPU Characterization [66]	GCN, GIN, SAG	PyG	Cora, CiteSeer, Pubmed, Reddit, LiveJournal	No	Inference
GNN-GPU Characterization [72]	GCN, GAT, GGNN,	PyG, DGL	Cora, CiteSeer, Pubmed, AIFB, MUTAG, BGS	No	Inference
GNNMark [4]	PinSAGE, STGCN, DGCN, GW, KGNN, ARGA, TLSTM	PyG	Cora, CiteSeer, Pubmed, NWP, MVL, LA, PEMS	No	Training
HyGCN [67]	GCN, SAG, GIN	PyG	IMDB, Cora, Citeseer, Colab, Pubmed, Reddit	No	Inference
GRIP [37]	GCN, GIN, G-GCN, SAG	GReTa	Pokec, YouTube LiveJournal, Reddit	No	Inference
gSuite	**GCN, GIN, SAG**	**None**	**Cora, Citeseer, Pubmed Reddit, LiveJournal**	**Yes**	**Inference**

Our Design Approach

- Flexibility
- Extendability
- Independence

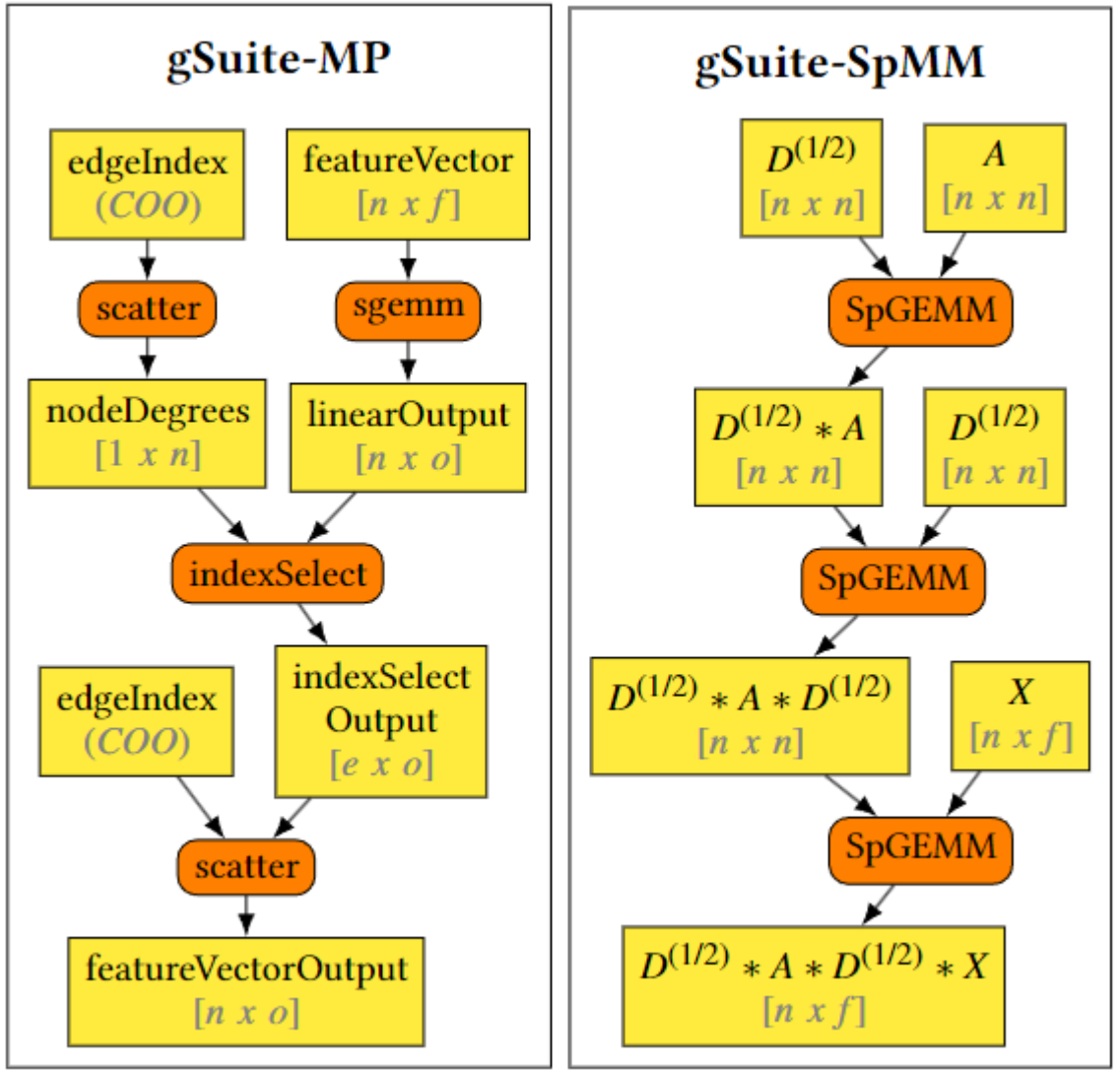
Our Design Approach

- Core Kernels

Kernel Name	Computational Model	Short Form	Description
indexSelect	MP	is	Indexes the input along specified dimension by using index entries.
scatter	MP	sc	Reduces given input based-on index vector using entries.
sgemm	SpMM /GEMM	sg	Generalized matrix multiplication of two given matrices.
SpGEMM	SpMM /GEMM	sp	Matrix multiplication of two sparse matrices.

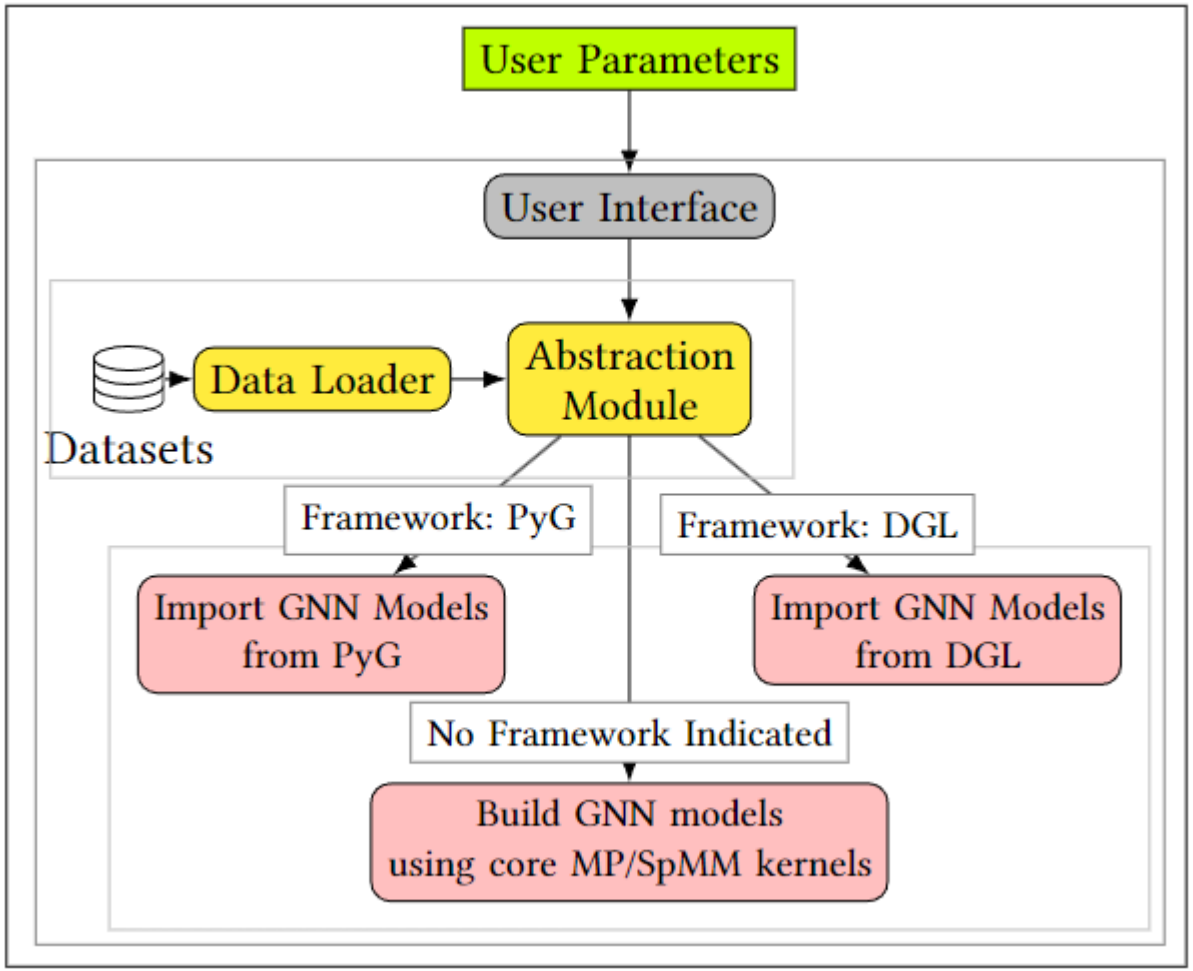
Our Design Approach

- Computational Schema
  - **Orange** -> Core Kernel
  - **Yellow** -> Data



Our Design Approach

- Software Architecture



Evaluation

- GNN Models
  - GCN (Graph Convolutional Network)
  - GIN (Graph Isomorphism Network)
  - SAG (Graph SAGE)

Evaluation

- Datasets

Dataset	Nodes	Feature Length	Edges	Short Form
Cora [47]	2,708	1,433	5,429	CR
CiteSeer [47]	3,327	3,703	4,732	CS
PubMed [53]	19,717	500	44,438	PB
Reddit [28]	232,965	602	11,606,919	RD
LiveJournal [2]	4,847,571	1	68,993,773	LJ

Evaluation

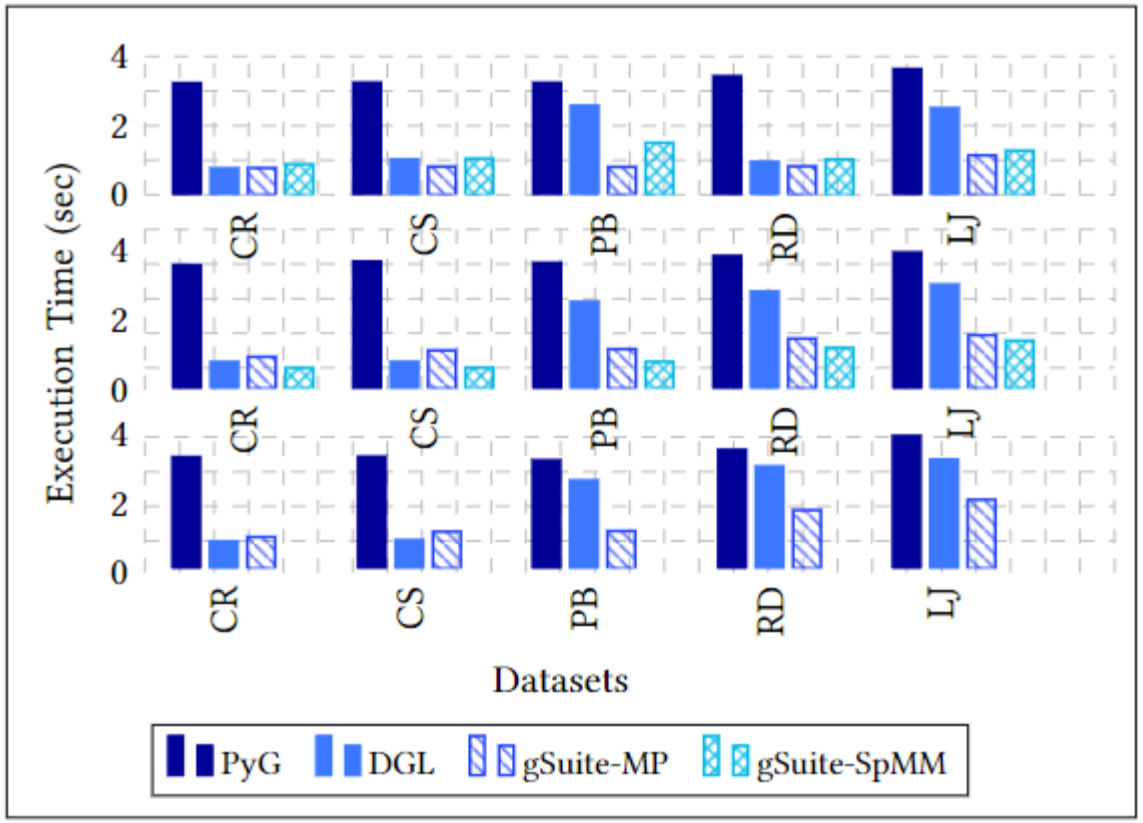
- Experimental Setup
  - NVIDIA V100 GPU 32GB
  - Profiling:
    - NVIDIA nvprof (CUDA Toolkit 10.2)
    - GPGPU-Sim (v4.0)

Evaluation

- Results
  - Execution Time
  - Instruction Breakdown
  - Issue Stall Distribution
  - Warp Occupancy Distribution
  - L1/L2 Cache Hit Rate
  - Compute/Memory Utilization

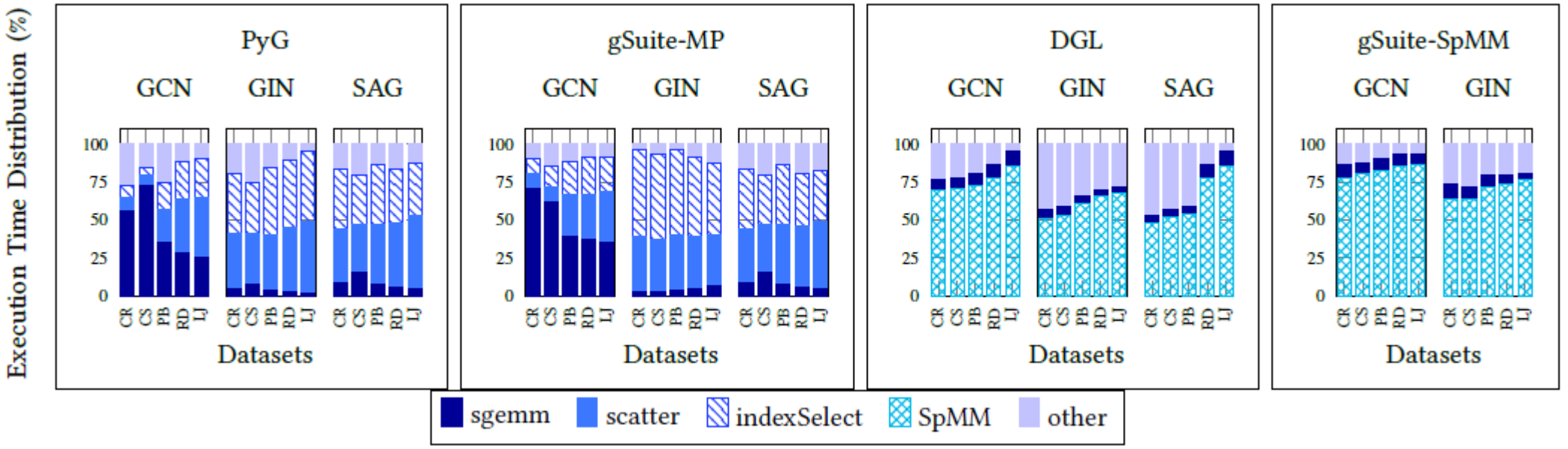
Evaluation

- Results
  - Execution Time (Wall Clock Time)



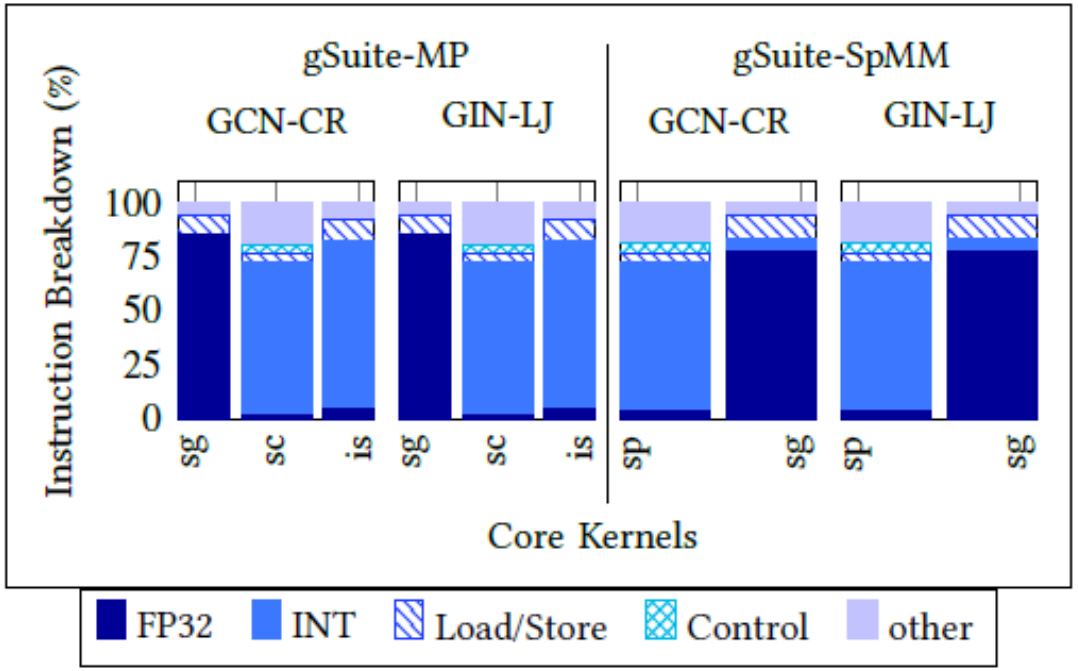
Evaluation

- Results
  - Execution Time (Distribution of Kernels)



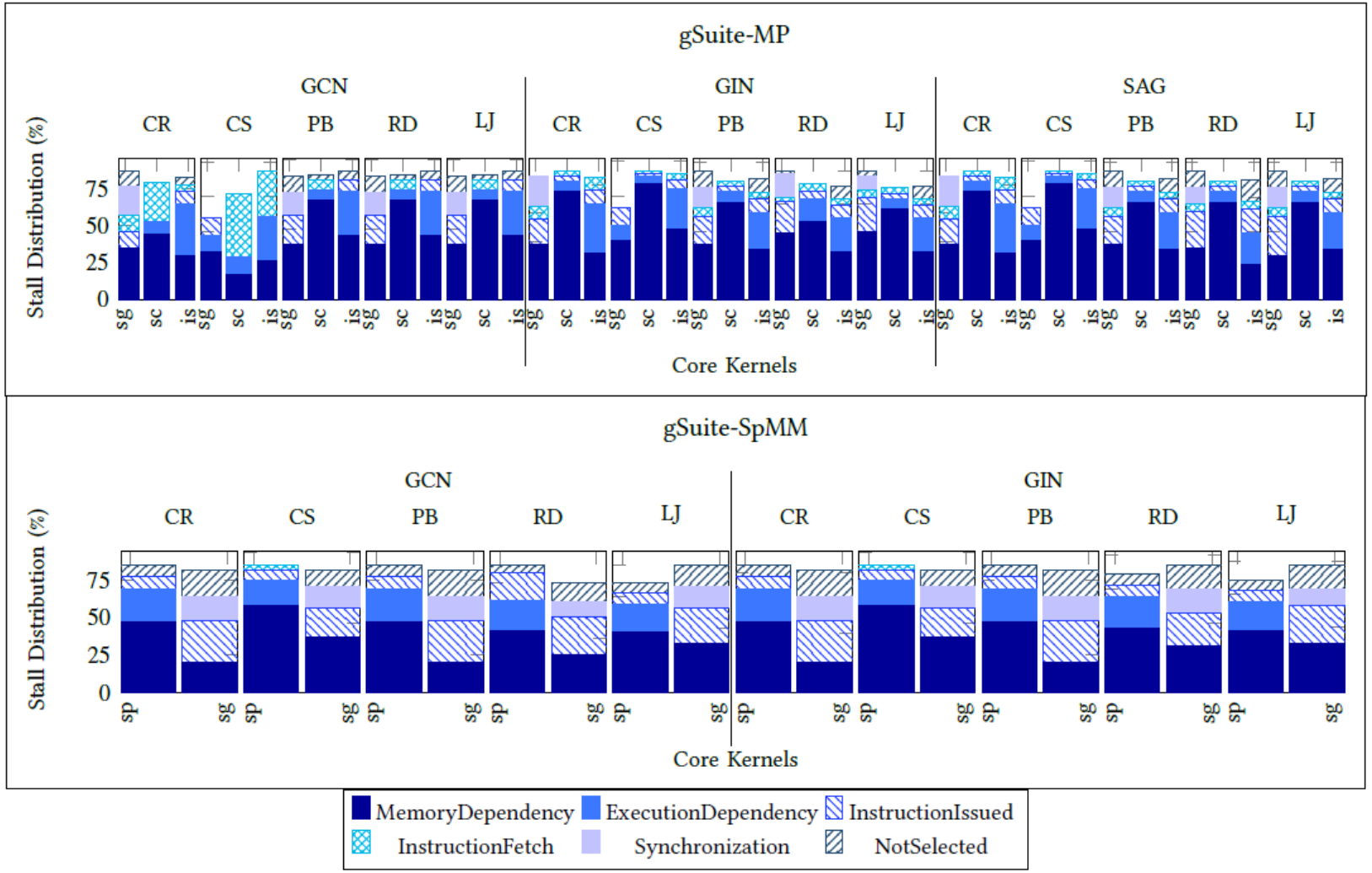
Evaluation

- Results
  - Instruction Breakdown



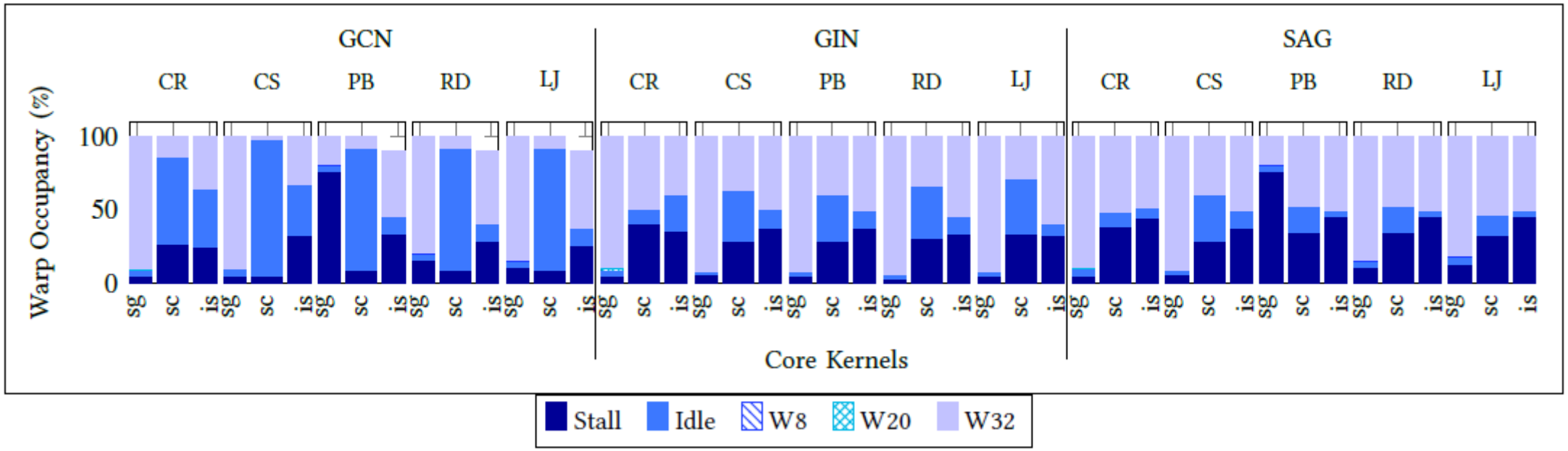
Evaluation

- Results
  - Issue Stall Distribution



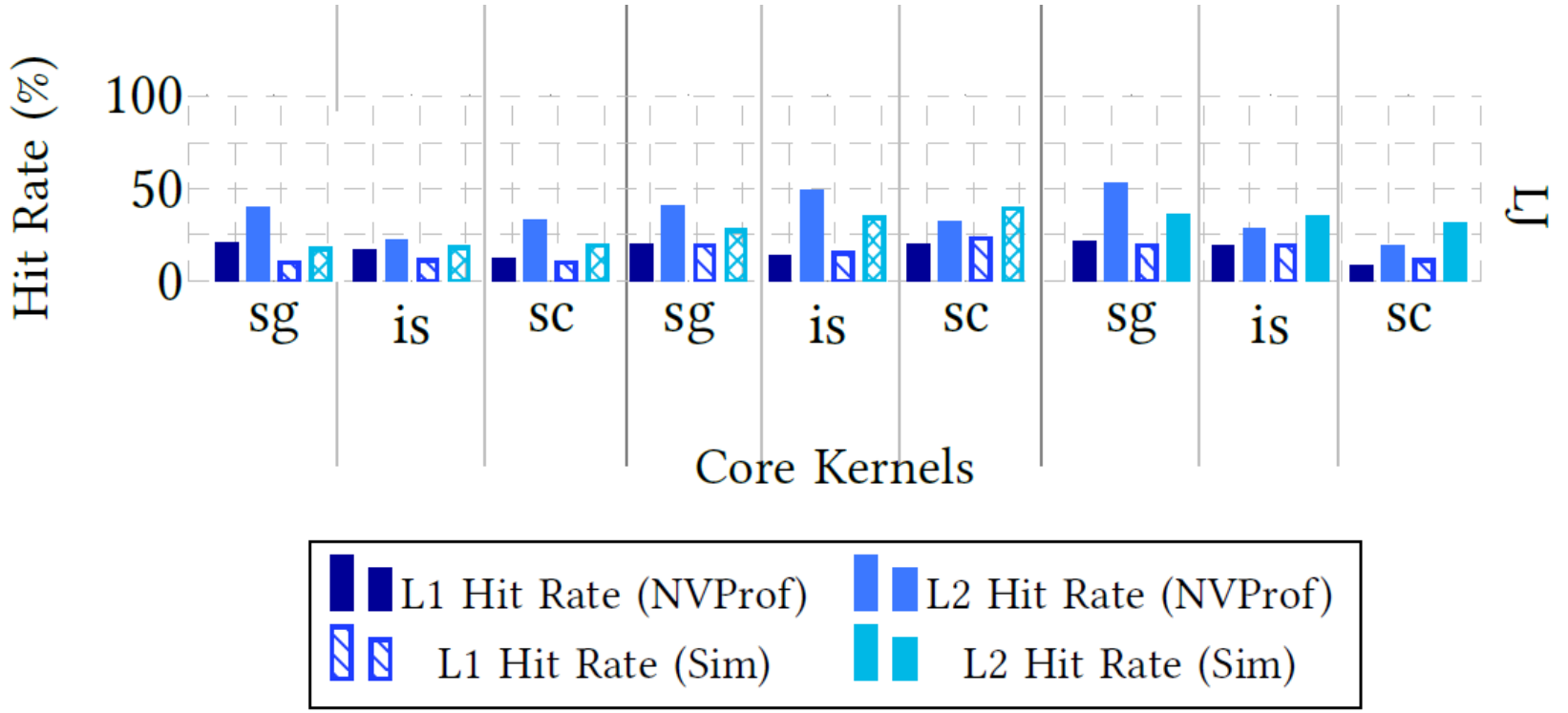
Evaluation

- Results
  - Warp Occupancy Distribution



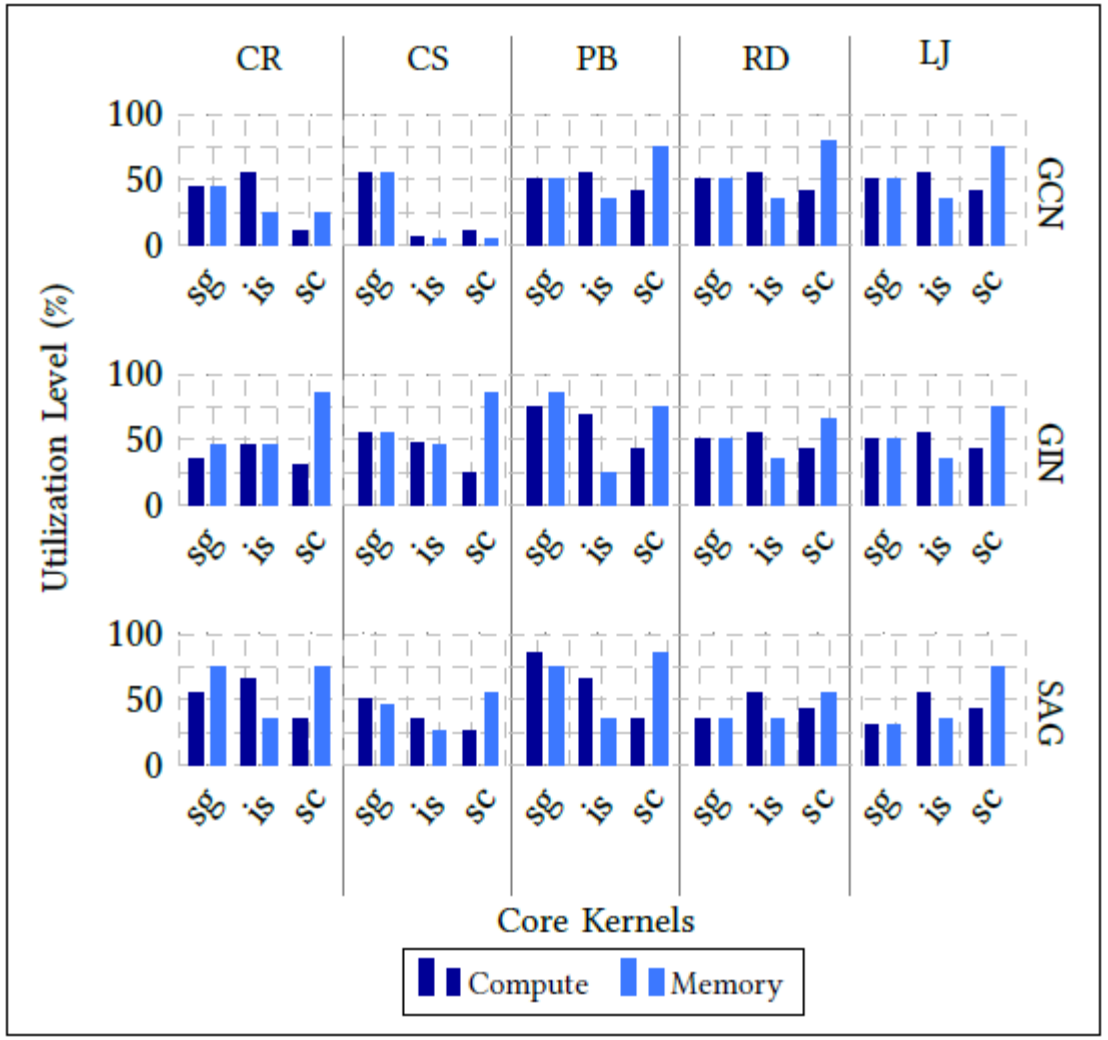
Evaluation

- Results
  - L1/L2 Cache Hit Rate



Evaluation

- Results
  - Compute/Memory Utilization



Conclusion

- Framework-independent and flexible benchmark suite for GNN Inference
- Architectural implications on GNN computation
- Suggestions based on our findings
- Open source, we welcome any suggestion/contribution!

GitHub

github.com/tekdogan/gsuite

zenodo

zenodo.org/record/7071370

Future Work

- Support for GNN Training
  - Implement propagations, weights, etc.
- Support Different Architectures
  - FPGAs, AMD GPUs, etc.
  - Implement core kernels with OpenCL

Acknowledgment

Ayse Yilmazer-Metin, Ph.D.

ASELSAN, Inc.

Thank you!

Webmail

tekdogan20@itu.edu.tr

Linkedin

linkedin.com/in/tekdogan