Scaling GNN training to billion-edge graphs on a single machine

BOSTON UNIVERSITY

CASP Systems

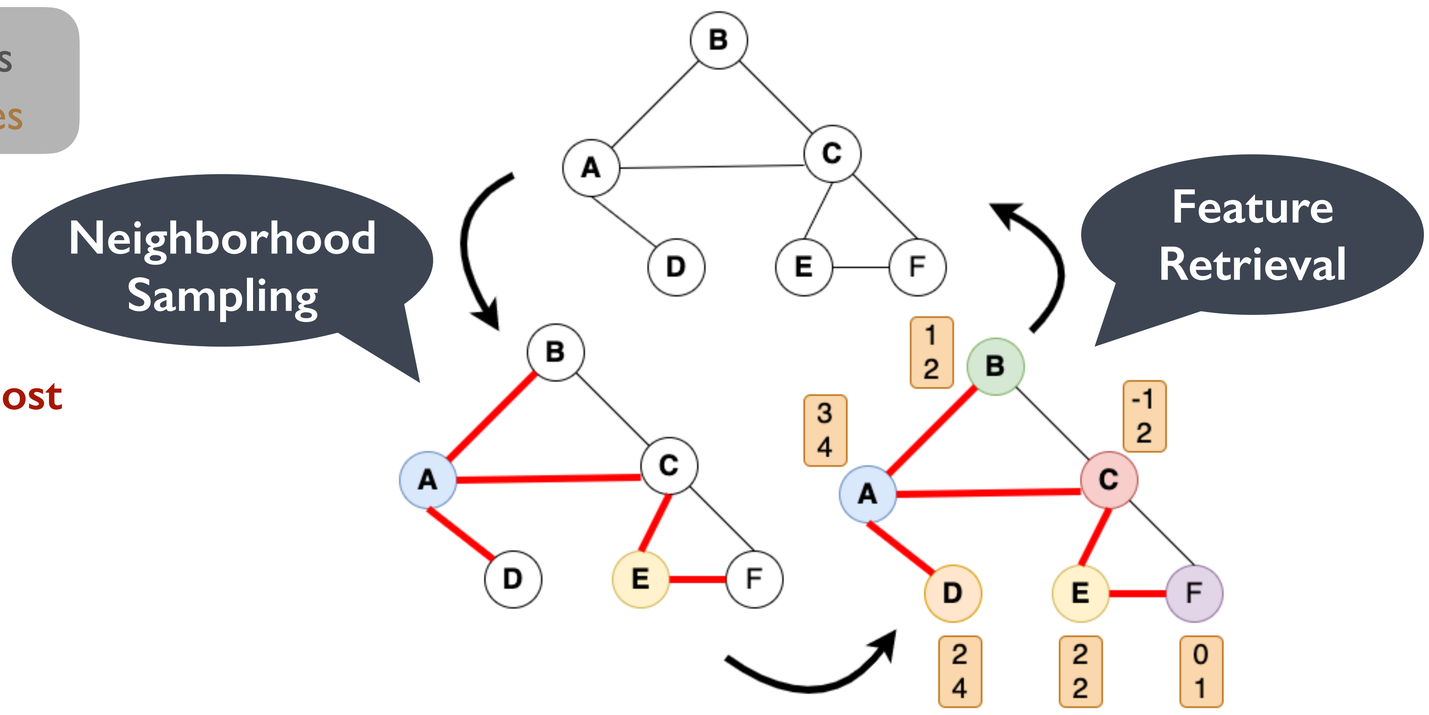
Real-world graphs are massive, comprising billions of nodes and edges

Web graphs	Social networks	Brain graphs
$4*10^9$ pages	$4*10^{11}$ friendships	10^{14} synopses

Existing solutions to larger-than-memory GNN training require excessive hardware

- Distributed training → High communication cost, low efficiency
- Multi-GPU training → High monetary cost
- Out-of-core training with modified algorithm → Degraded accuracy

Sampling-based GNN training has two major performance bottlenecks



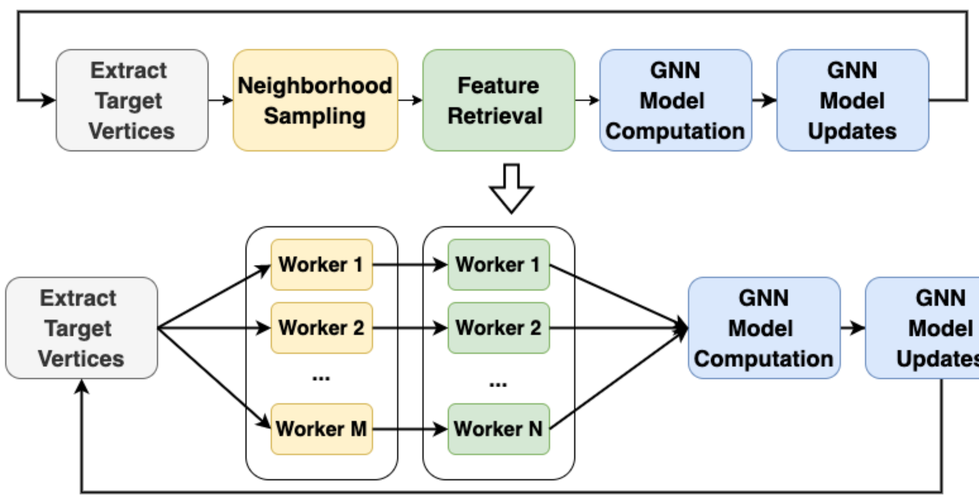
GraphosML
Efficient, large-scale GNN training on modern storage

Single-pass reservoir sampling
- SSD-friendly scans

Mini-batch training as a dataflow computation
- Pipeline and data parallelism
- Improved resource utilization

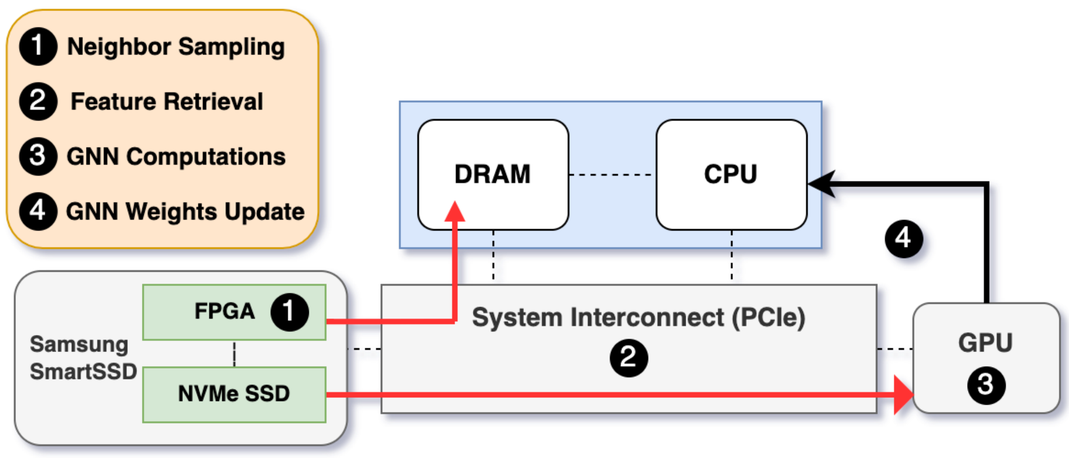
Sampling stage offloading to computational storage

Feature retrieval with Nvidia GPU-Direct technology
- Bypass CPU overhead



On Going Work

- Add sampling stage offloading to computational storage & GPU Direct Storage proposals with baseline solution.
- Measure end to end training time improvement over competitor publications.



Preliminary Experimental Results

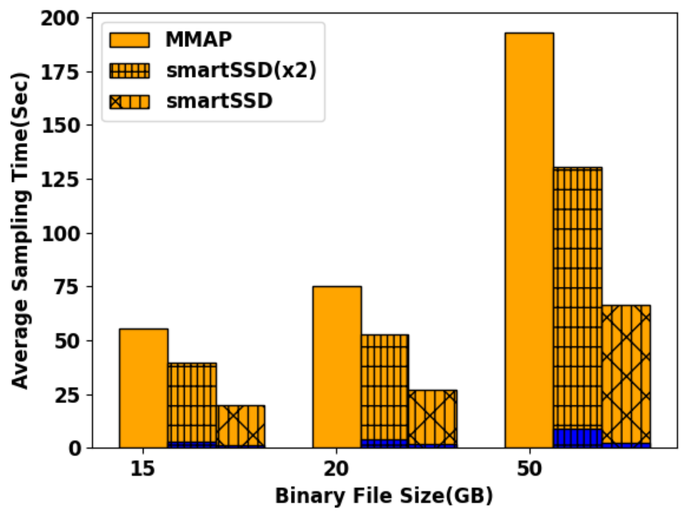
Experimental Setup	Ex-Large (300GB)			Ex-Small (60GB)	
Datasets	GNNLab	DGL	Marius-Mem	Marius-Disk	GraphosML
Products	3.286 s	22.823 s	6.99 s	5.611 s	19.36 s
Papers100M	11.617 s	58.255 s	58.142 s	41.55 s	42.78 s

Datasets	Nodes	Edges	Graph Data Size
Products	$2.5 * 10^6$	$62 * 10^6$	1.5 GB
Papers100M	$111 * 10^6$	$1.6 * 10^9$	65 GB

Node Classification ML Task on GraphSage GNN (5 epochs, 3 layers (20,15,10), mini batch size of 1000).
Both Ex-Large and Ex-Small setups consist of 48 vCPUs, V100 GPU (16GB). The DRAM is 300GB and 60GB respectively.
OGB-Paper100M dataset exceed Ex-Small DRAM memory specifications (60GB).

Our mmap-based baseline has competitive performance to state-of-the-art solutions

Smart SSD enables up to 3x faster neighborhood sampling compared to a CPU mmap-based baseline approach



Average time comparison between reservoir sampling on one or more smartSSD devices via CPU with MMAP over various size binary edger files.

Emmanouil Kritharakis
ekrithar@bu.edu

Po Hao Chen
bupochen@bu.edu

Yuchen Lu
luyc@bu.edu

Vasiliki Kalavri
vkalavri@bu.edu