Imperial College London

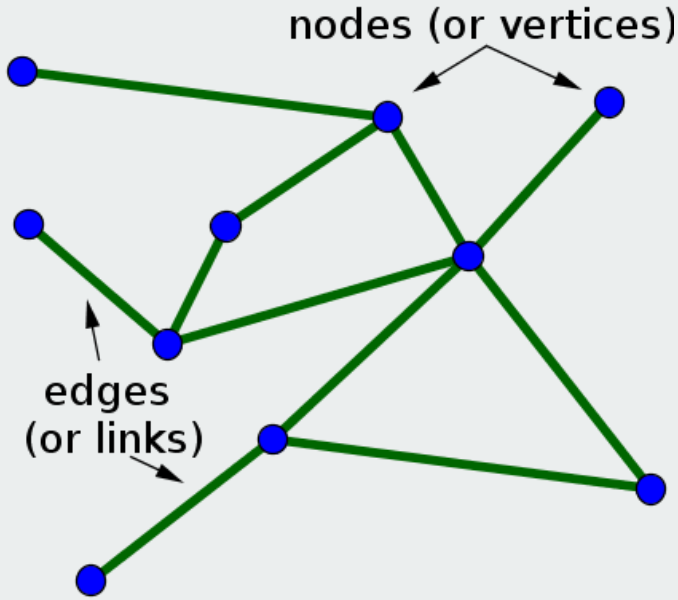
AMPLE: Event-Driven Accelerator for Mixed-Precision Inference of Graph Neural Networks

Pedro Gimenes, Imperial College London

EuroMLSys '25, Rotterdam, Netherlands

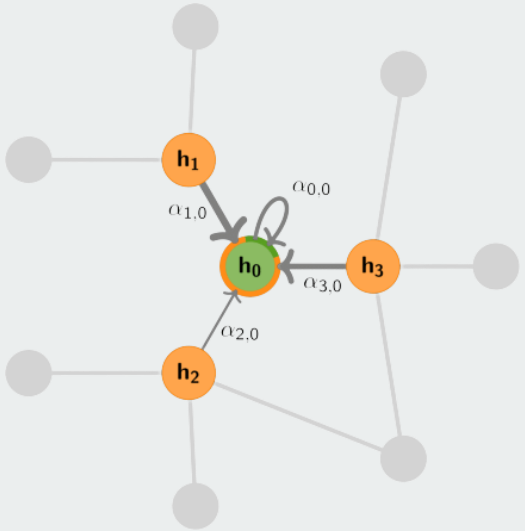
Background

- Graphs capture relationships between entities, which are represented as nodes, interconnected by edges.
- This structure enables modeling social networks, biological interactions, recommendation systems, etc.



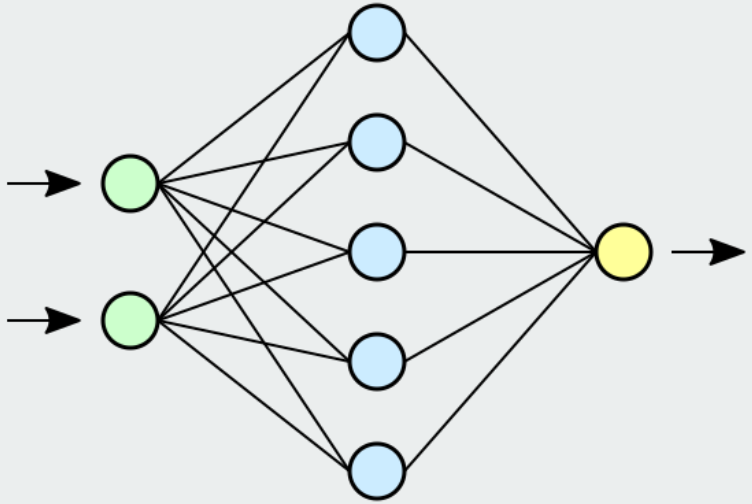
Graph Neural Networks

- **Aggregation**
  - Permutation invariant function over neighboring node embeddings
  - **Irregular** memory access



Graph Neural Networks

- **Aggregation**
  - Permutation invariant function over neighboring node embeddings
  - **Irregular** memory access
- **Transformation**
  - Fully-connected layer over aggregation results
  - **Regular** memory access



Motivation

- Computation **complexity** linked to graph size
- **CPU**:
  - Low parallelization
  - Inefficient memory management
- **GPU**:
  - No inter-phase pipelining
  - Little support for low-precision computation

Motivation

- Computation **complexity** linked to graph size
- **CPU**:
  - Low parallelization
  - Inefficient memory management
- **GPU**:
  - No inter-phase pipelining
  - Little support for low-precision computation

Current **FGPA/ASIC** accelerators present…

- Double-buffering
- No support for **mixed-precision**
- Requirement for **on-chip buffering** of features and weights

AMPLE: Event-Driven Accelerator for Mixed-Precision Inference of Graph Neural Networks

Main contributions:

- **Event-driven** node programming flow
- **Mixed-precision** node dataflow
- On-chip **Streaming Pre-fetcher** to enable computation on large graphs

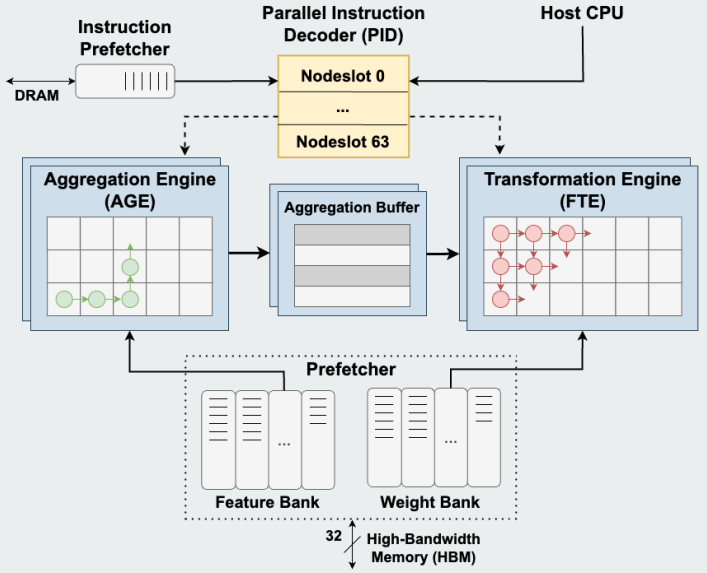
Imperial College London

Architecture

AMPLE: Event-Driven Accelerator for Mixed-Precision Inference of Graph Neural Networks

High-Level Architecture

- **PID:** Parallel Instruction Decoder
  - Communication with host device
- **AGE:** Aggregation Engine
  - Node aggregation using Network-on-Chip
- **FTE:** Transformation Engine
  - Node transformation with systolic array
- **Prefetcher**
  - Fetch node embeddings from HBM



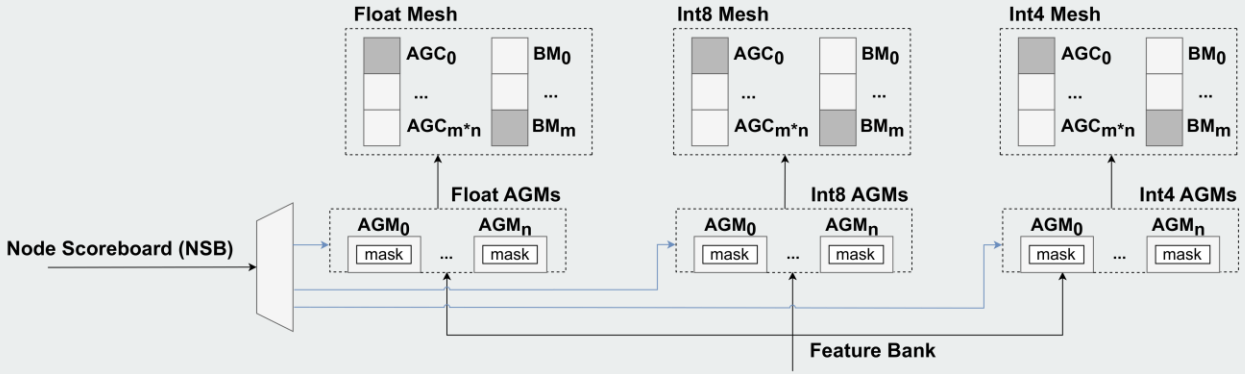
Event-Driven Programming

- The Host device schedules workload into the accelerator by asynchronously programming a memory-mapped register bank in the Node Instruction Decoder (NID).
- Subsequently, the NID drives other functional units including the Aggregation Engine (AGE) and Transformation Engine (FTE) to perform the inference computation.

Slot	Node ID	Precision	State	Neighbors	Adjacency List Pointer	Updated Feature Pointer
0	267	float	Transformation	32	0x3BC90188	0x4FE8B774
1	268	float	Aggregation	8	0xCAF5C03F	0xE672109F
…	…	…	…	…	…	…
63	330	int4	Prefetch	1	0x78E26A27	0xA4D89ED9

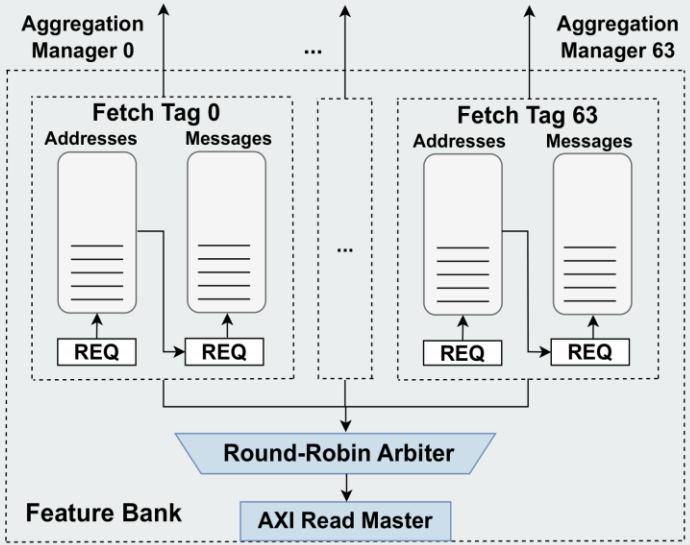
Mixed-Precision Arithmetic

- Computing units arranged in a Network-on-Chip (NoC) where the precision ratio of processing elements (PEs) can be tuned at compile time.
- PEs are dynamically allocated according to precision-wise resource availability, bypassing static pipeline gaps due to node degree variability.



Large Graph Processing

- Storage elements for node embeddings access HBM banks concurrently, alleviating memory boundedness of sparse graph data.
- Two-stage fetching mechanism
  - Store neighbors in Adjacency Queue (AQ)
  - Use pointers from AQ to store neighbor embeddings in Message Queue (MQ)
- Computing over large graphs supported through **partial response mechanism**, whereby the Prefetcher continues fetching data in the background while the Aggregation Engine is working.



Imperial College London

Experiments

AMPLE: Event-Driven Accelerator for Mixed-Precision Inference of Graph Neural Networks

Experiments

- Evaluated using three foundational Graph Neural Network models: GCN, GIN, GraphSAGE
- Six datasets spanning a range of graph sizes
  - **Citation:** Cora, Citeseer, Pubmed
  - **Social media:** Flickr, Reddit, Yelp
- Baselines included computation on an Intel Xeon CPU and RTX A6000 GPU with Pytorch Geometric kernels.

Model	Aggregation	Residual	Normalization
GCN	sum	✗	aggregation
GIN	sum	aggregation	✗
GraphSAGE	mean	transformation	transformation

	Name	Nodes	Mean Degree	Features	DQ Ratio
CR	Cora	2,708	3.9	1,433	2.1 %
CS	CiteSeer	3,327	2.7	3,703	2.7 %
PB	PubMed	19,717	4.5	500	2.9 %
FL	Flickr	89,250	10.0	500	0.2 %
RD	Reddit	232,965	99.6	602	2.7 %
YL	Yelp	716,847	19.5	300	0.4 %

Results

- AMPLE results were obtained from Modelsim 19.2 simulation at 200MHz frequency, obtained from Vivado 23.1 synthesis on Alveo U280 FPGA card.
- AMPLE led to an improvement in mean inference time compared to CPU/GPU baselines across all models.

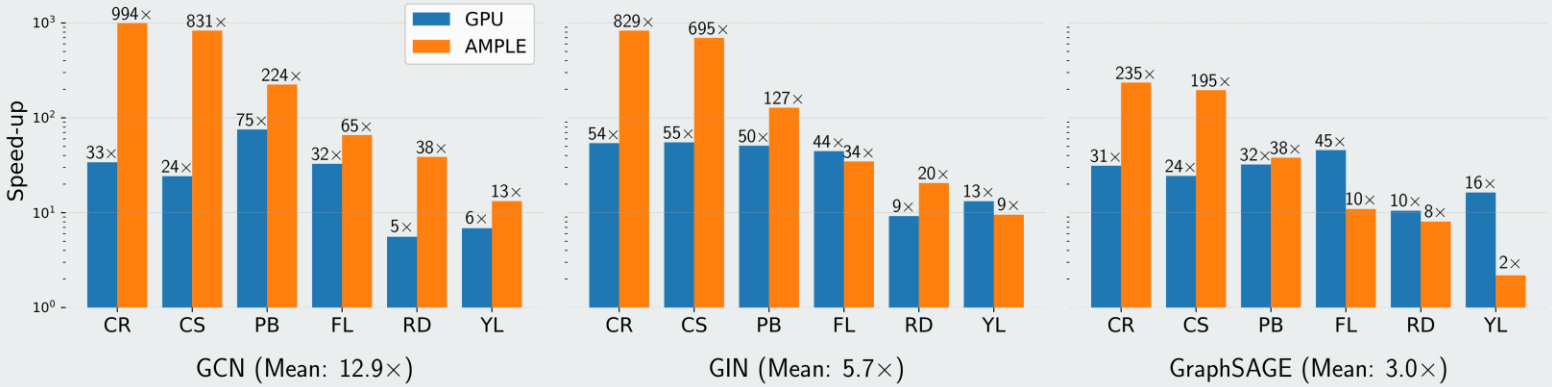


Figure: GPU and AMPLE speedup for each model and dataset relative to Intel Xeon CPU baseline

Imperial College London

Thanks!