Accelerating GNNs in modern spatial accelerators - challenges and opportunities

Partha Maji

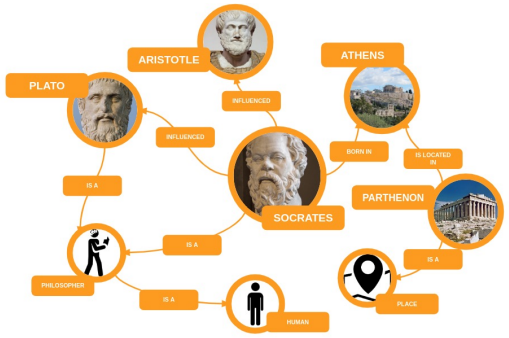
Outline

- Computational View of Graph Neural Network
- Spatial Architectures – Simplified View
- Mapping GNN onto Spatial Architectures

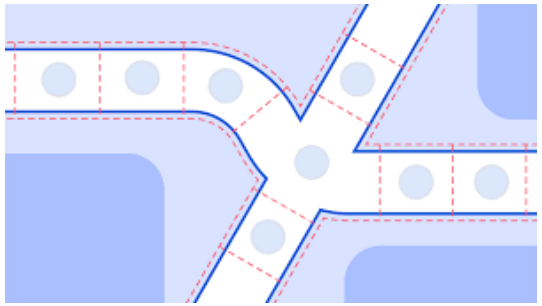
Outline

- Computational View of Graph Neural Network
- Spatial Architectures – Simplified View
- Mapping GNN onto Spatial Architectures

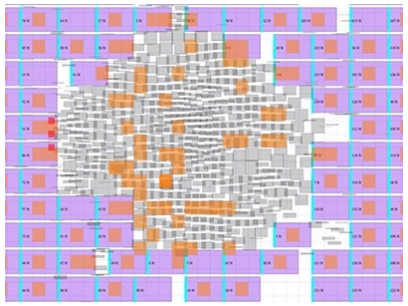
Graph NN in the industry



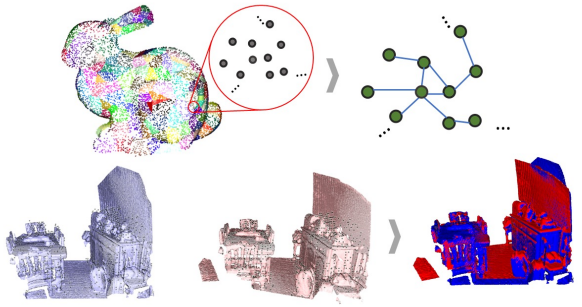
Knowledge Graphs



Transportation Networks

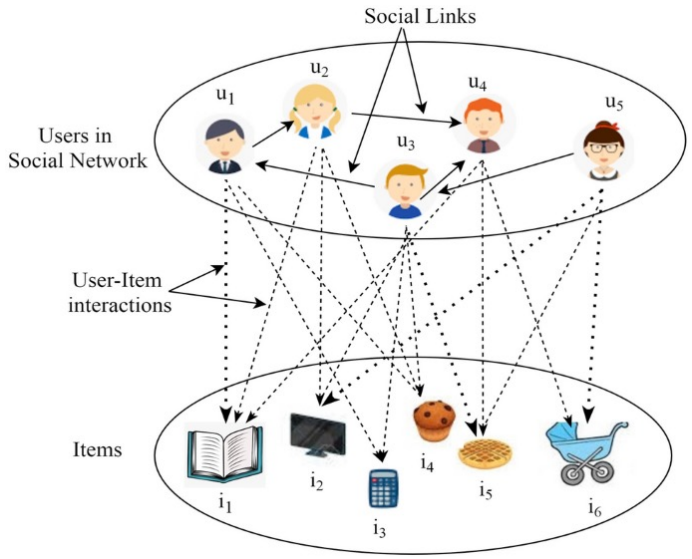


Chip Place & Route

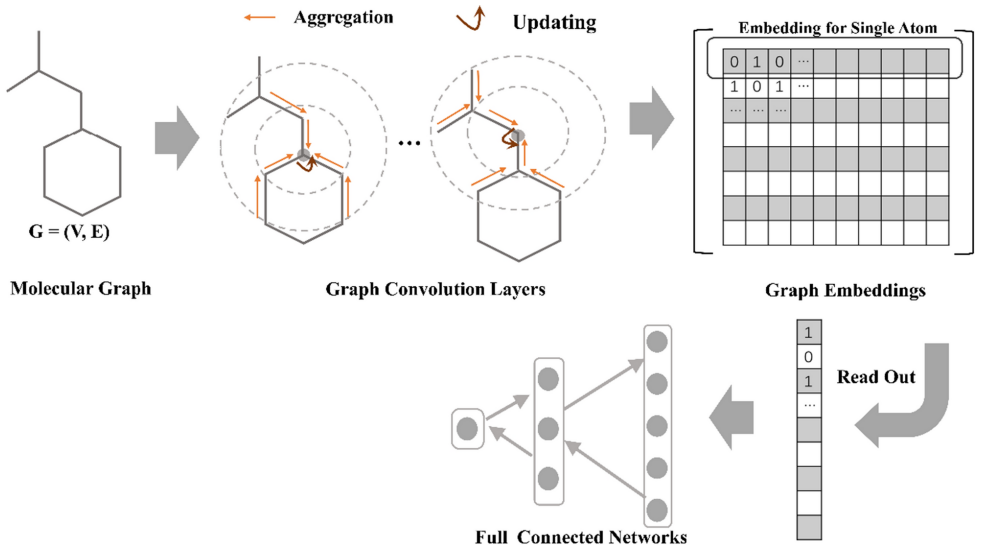


Geometric Deep Learning

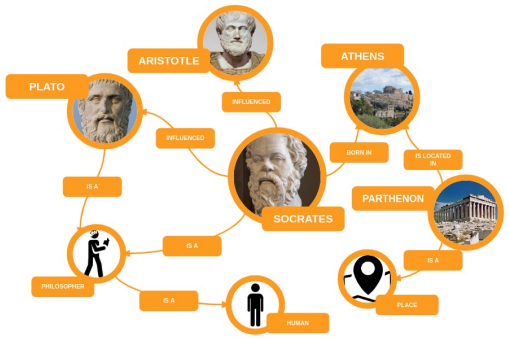
Graph NN in the industry



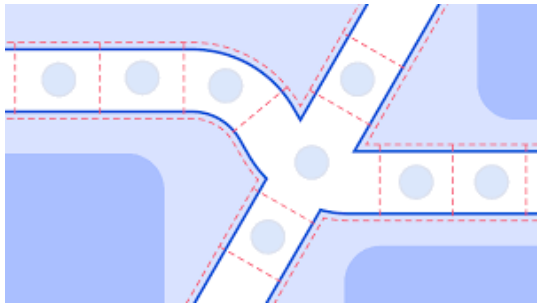
Recommender Systems



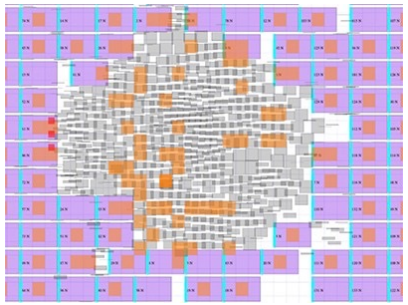
Drug Discovery



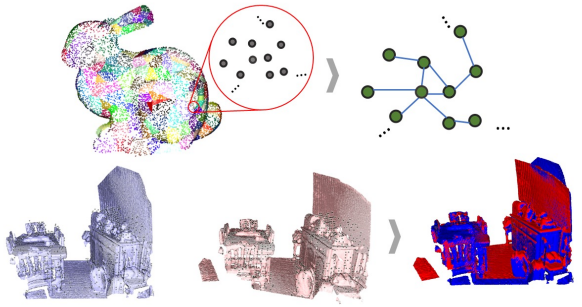
Knowledge Graphs



Transportation Networks

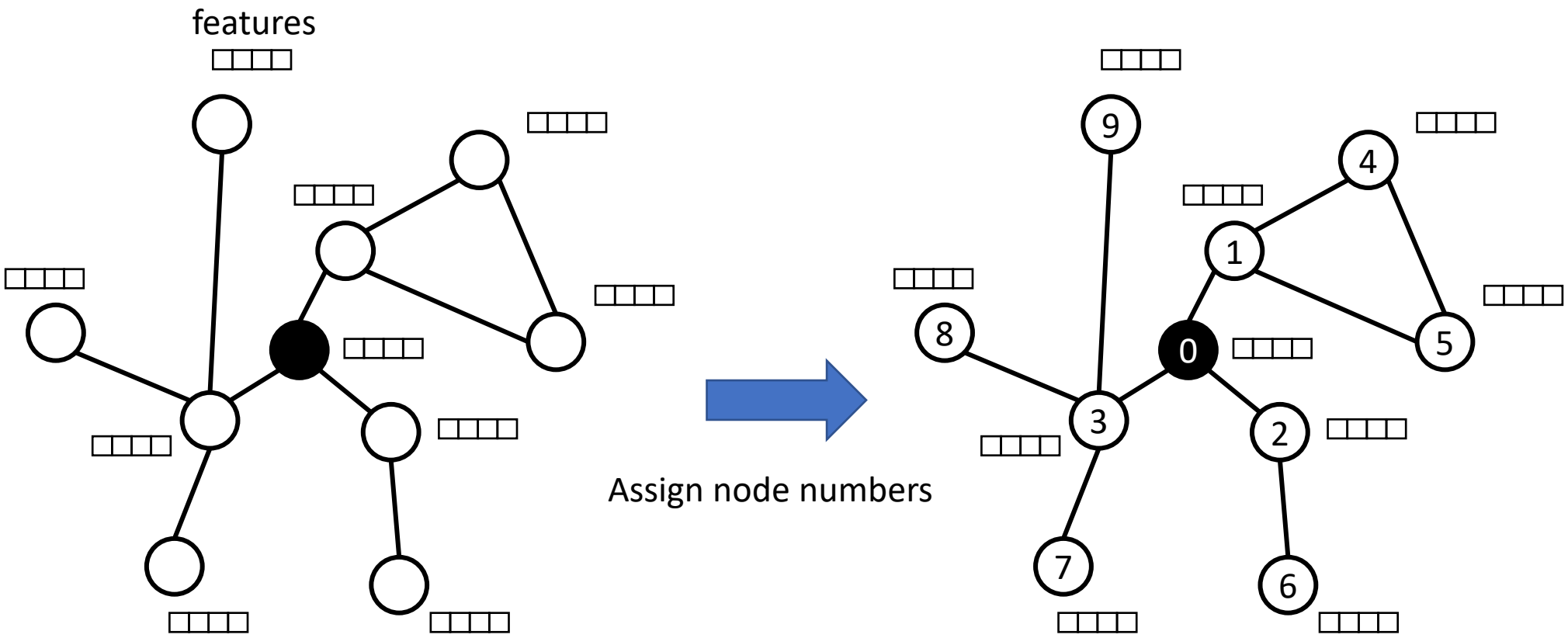


Chip Place & Route



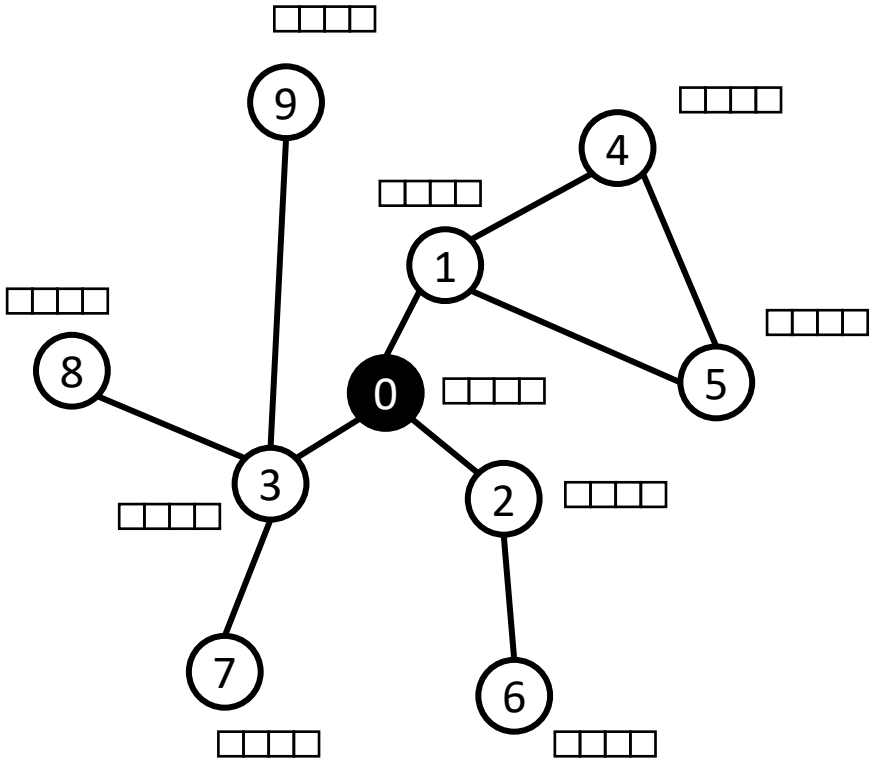
Geometric Deep Learning

How to define a graph?

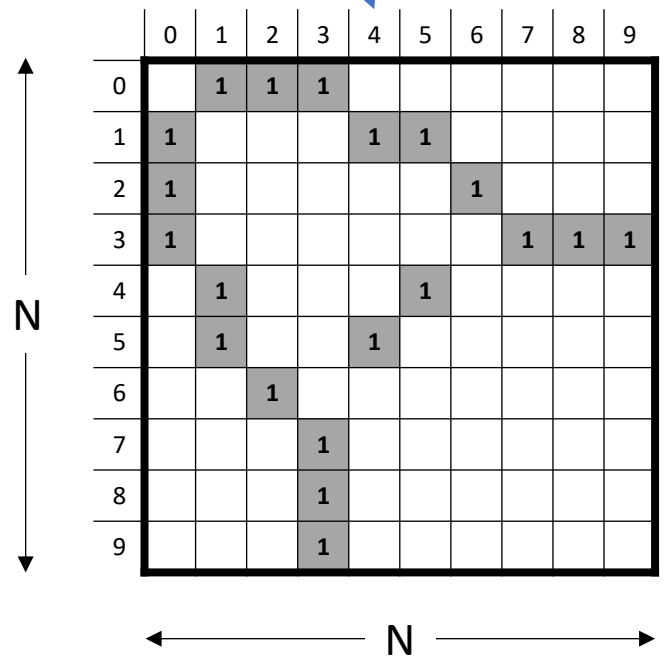


Permutation Invariant!

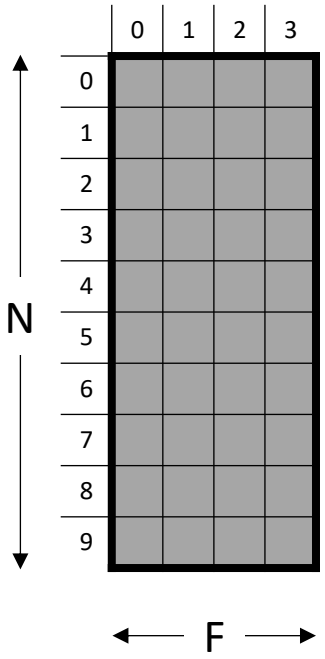
How to define a graph?



Highly sparse! More on this later..

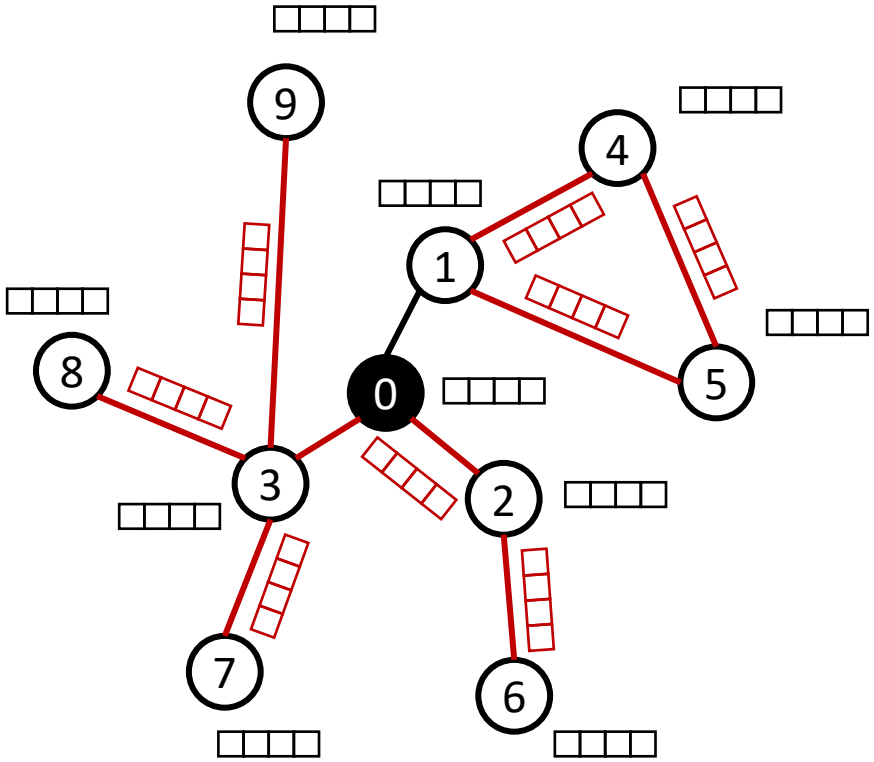


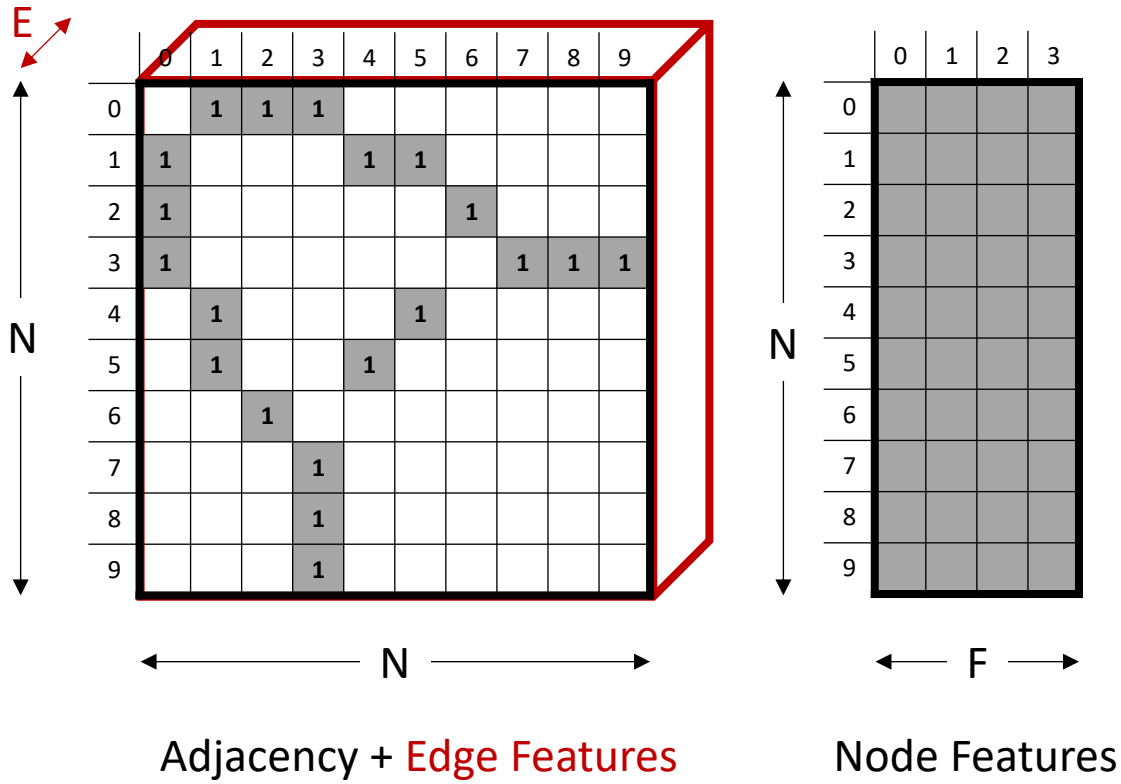
Adjacency



Node Features

How to define a graph?

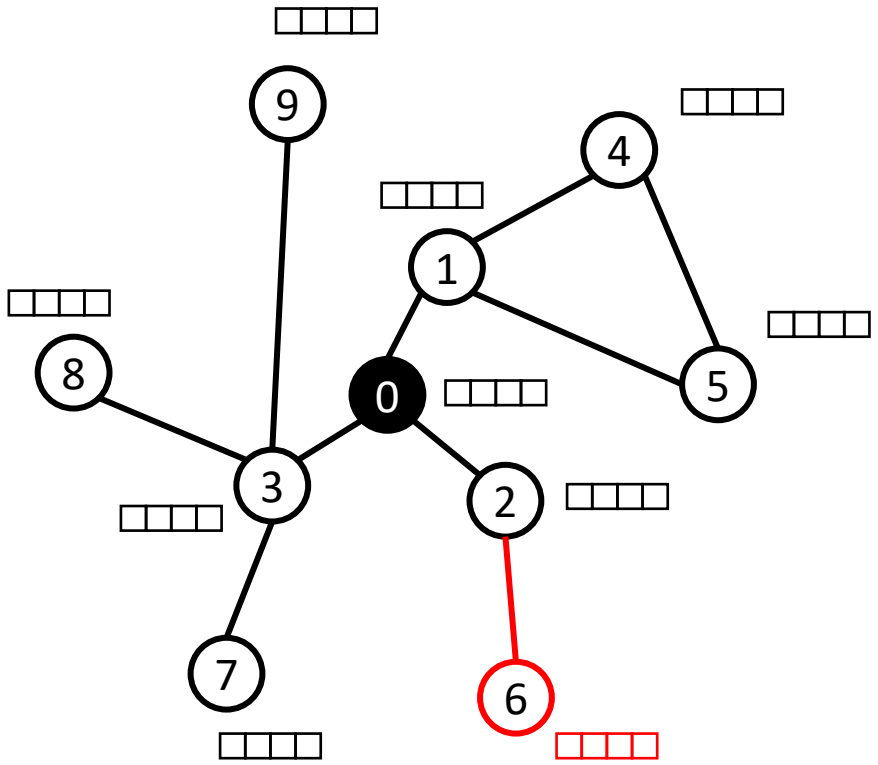




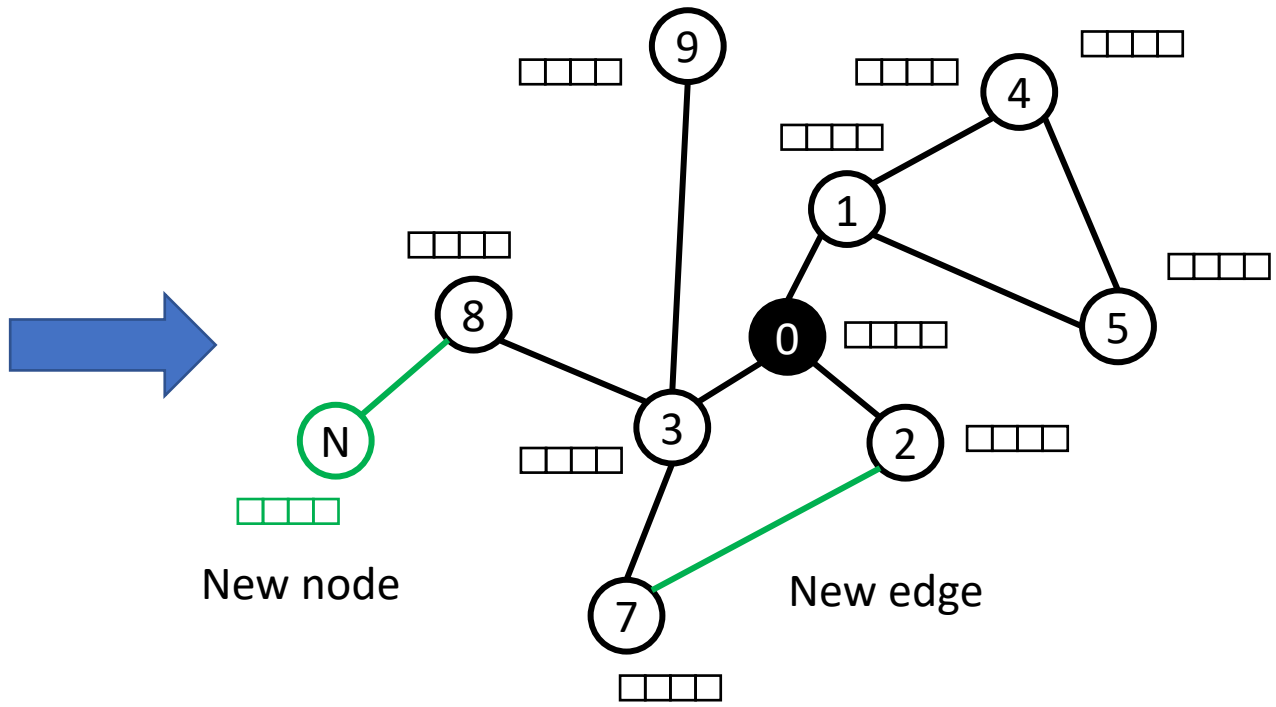
Adjacency + Edge Features

Node Features

Is a Graph always static? No!



Node dropped

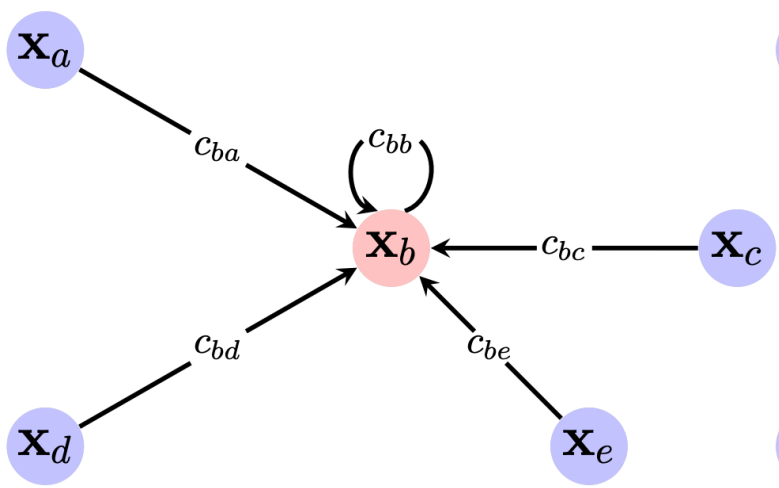


New node

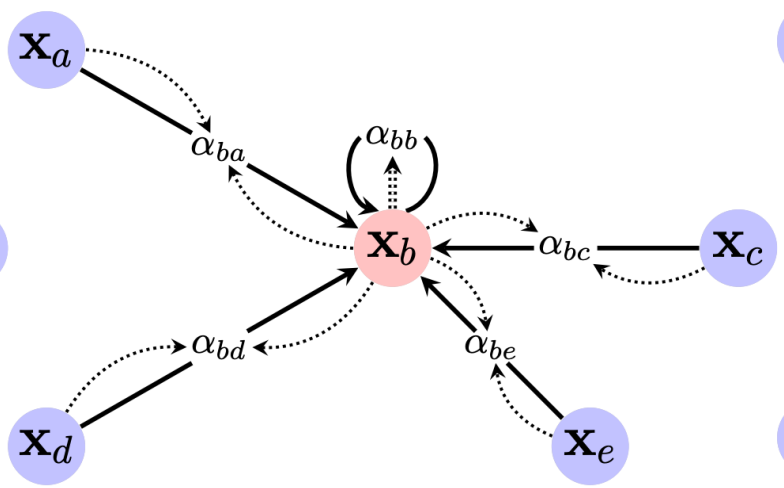
New edge

e.g., User X bought a travel bag, flight tickets for the family and trekking boots. What's the next thing s/he would likely to buy?

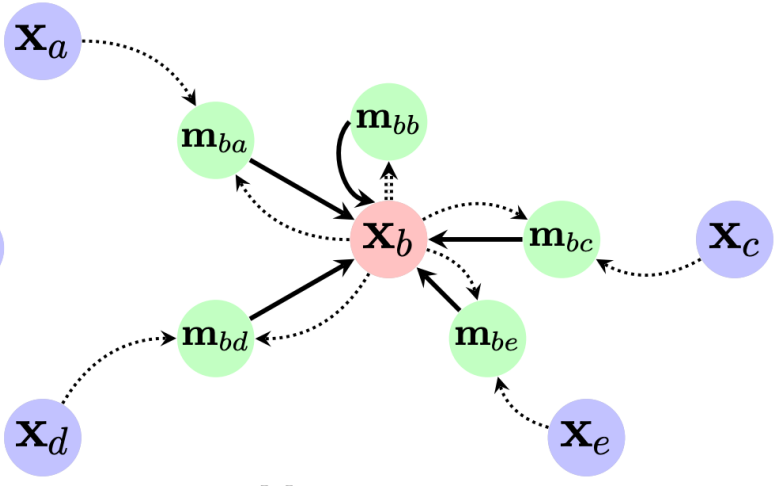
Main GNN Variants



Graph Convolutional Network (GCN)



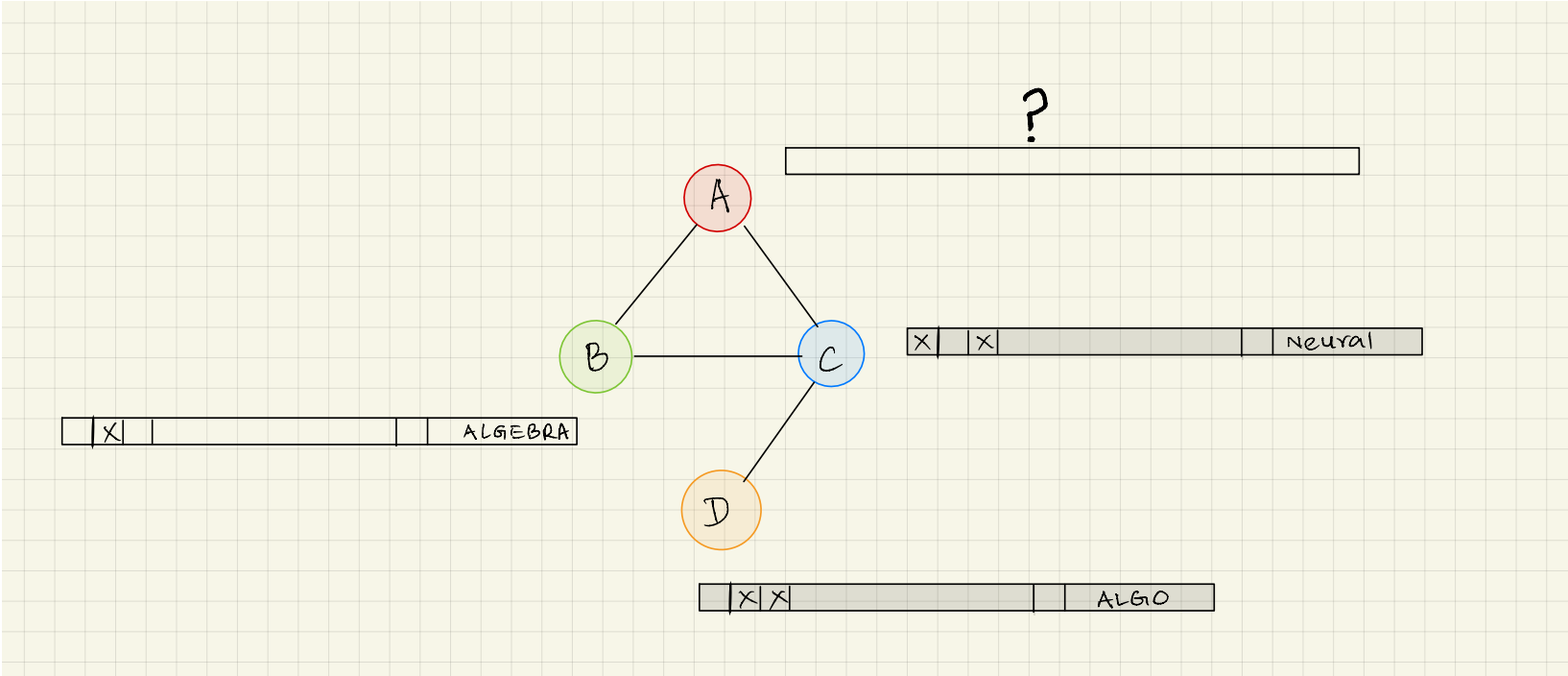
Graph Attention Network (GAT)



Message Passing Network (MPN)

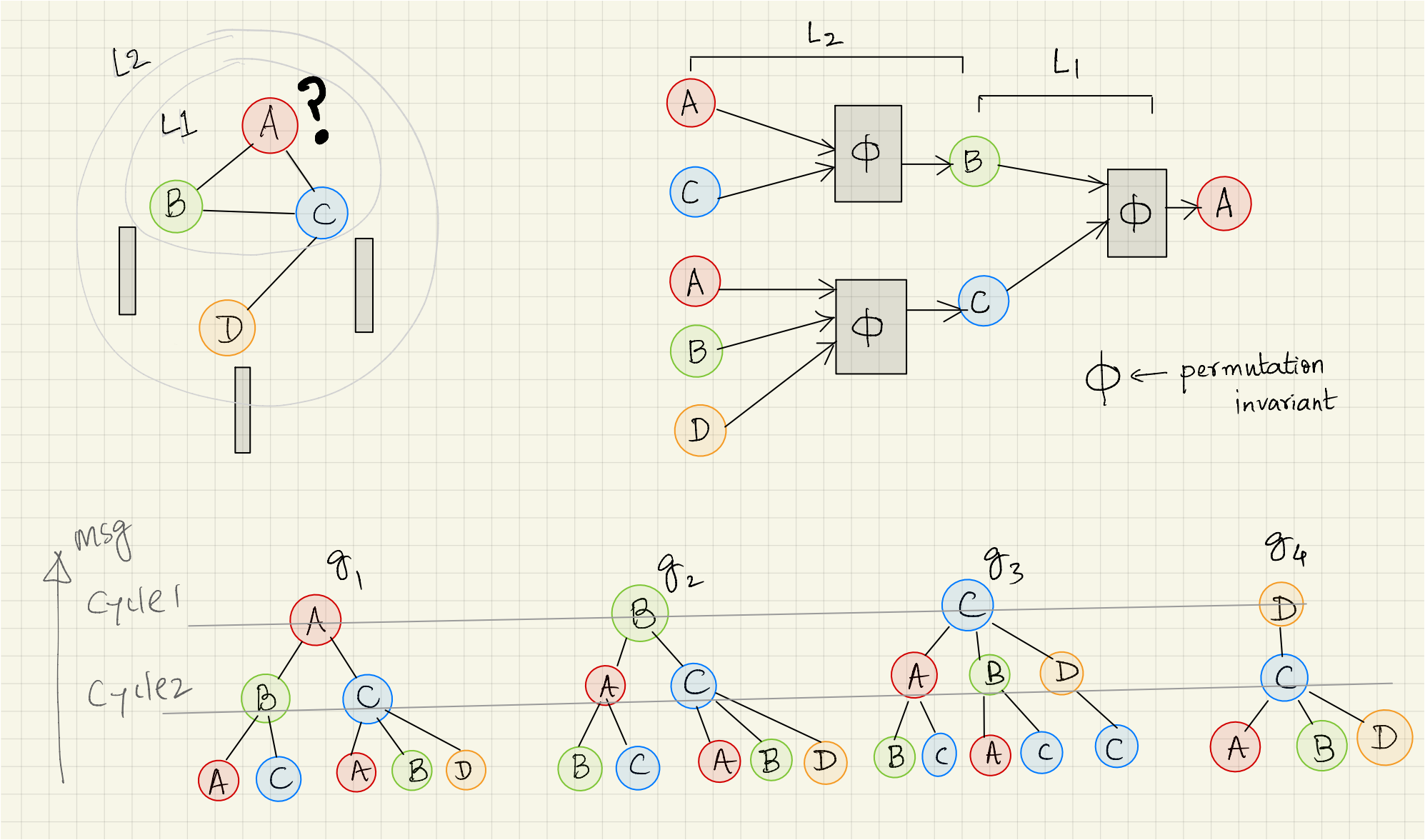
Prediction on a Graph

- Node prediction
- Edge prediction
- Graph prediction

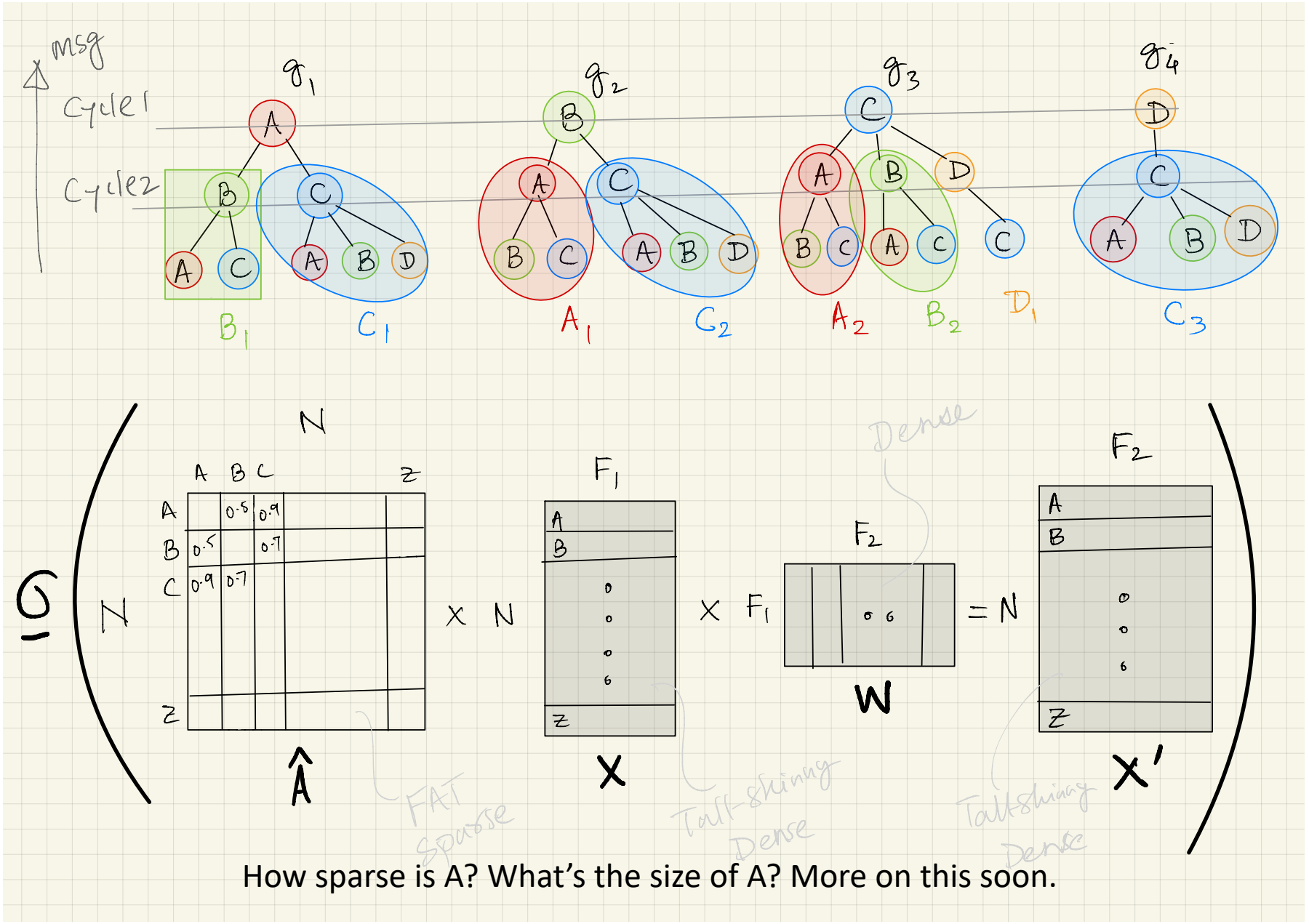


Node Prediction

GNN as message passing (MPNN)

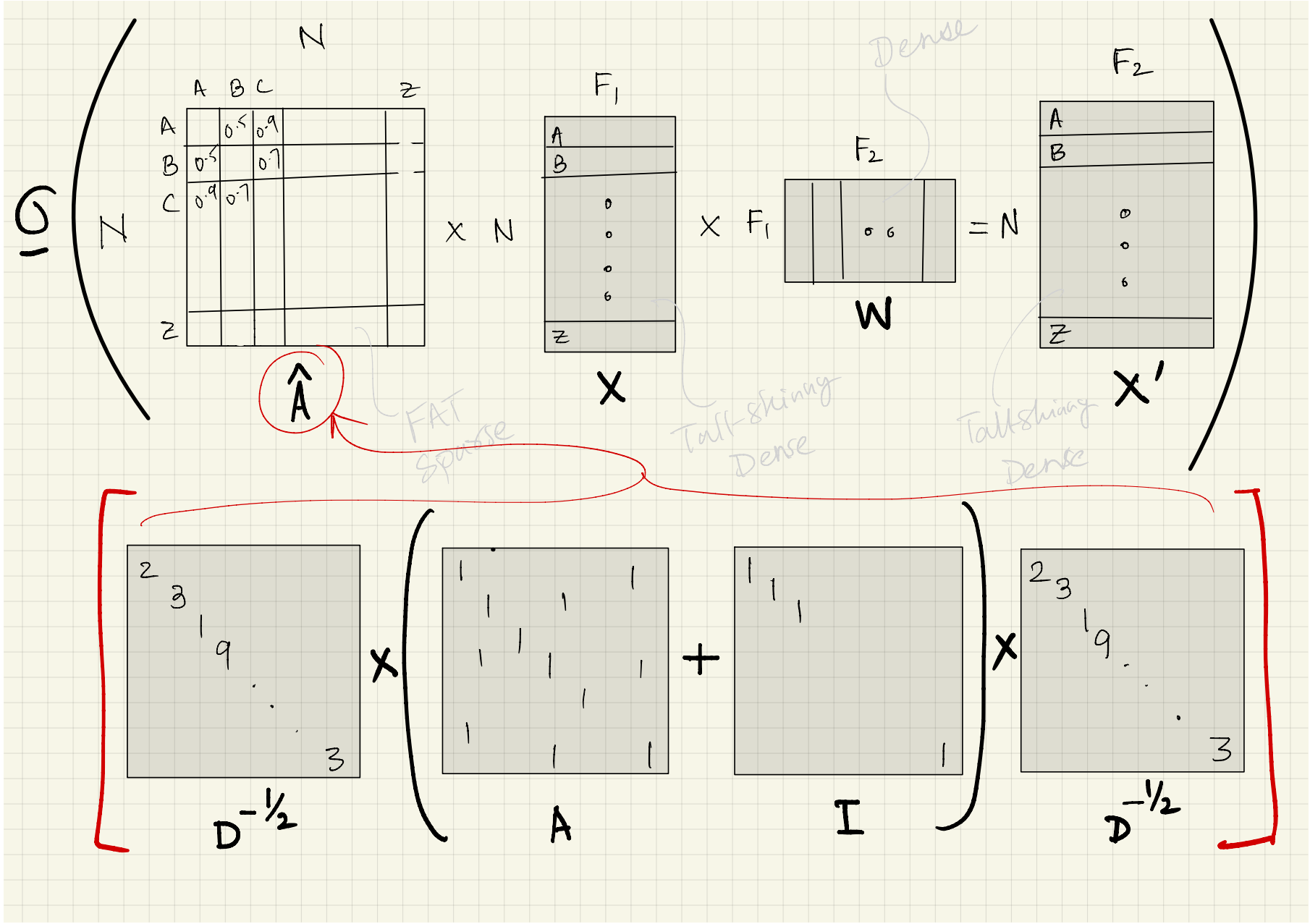


GCN



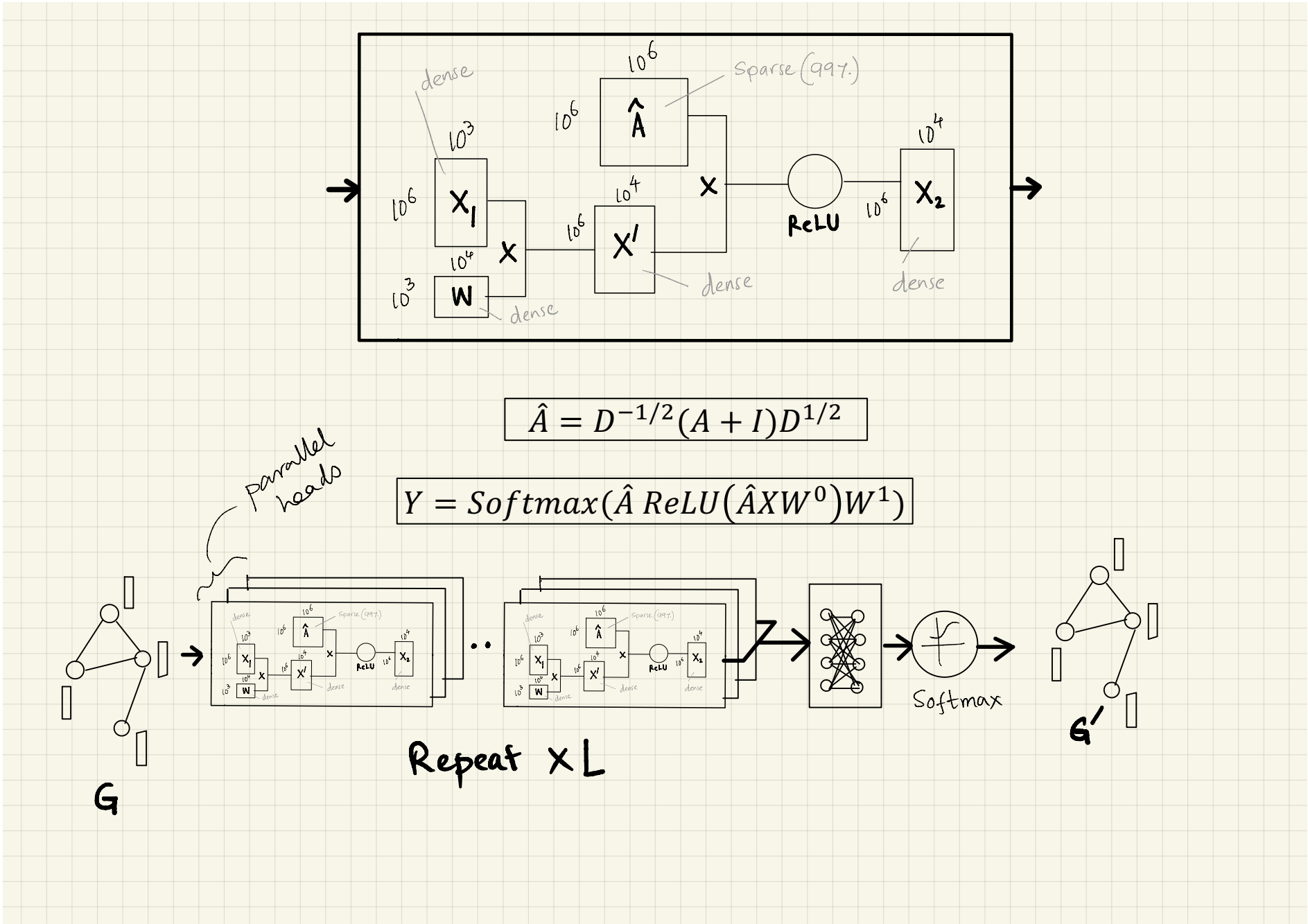
How sparse is A? What's the size of A? More on this soon.

GCN Data Flow



- On a static graph $\hat{A}$ can be pre-computed once.
- On a dynamic graph $\hat{A}$ needs to be computed on-the-fly.

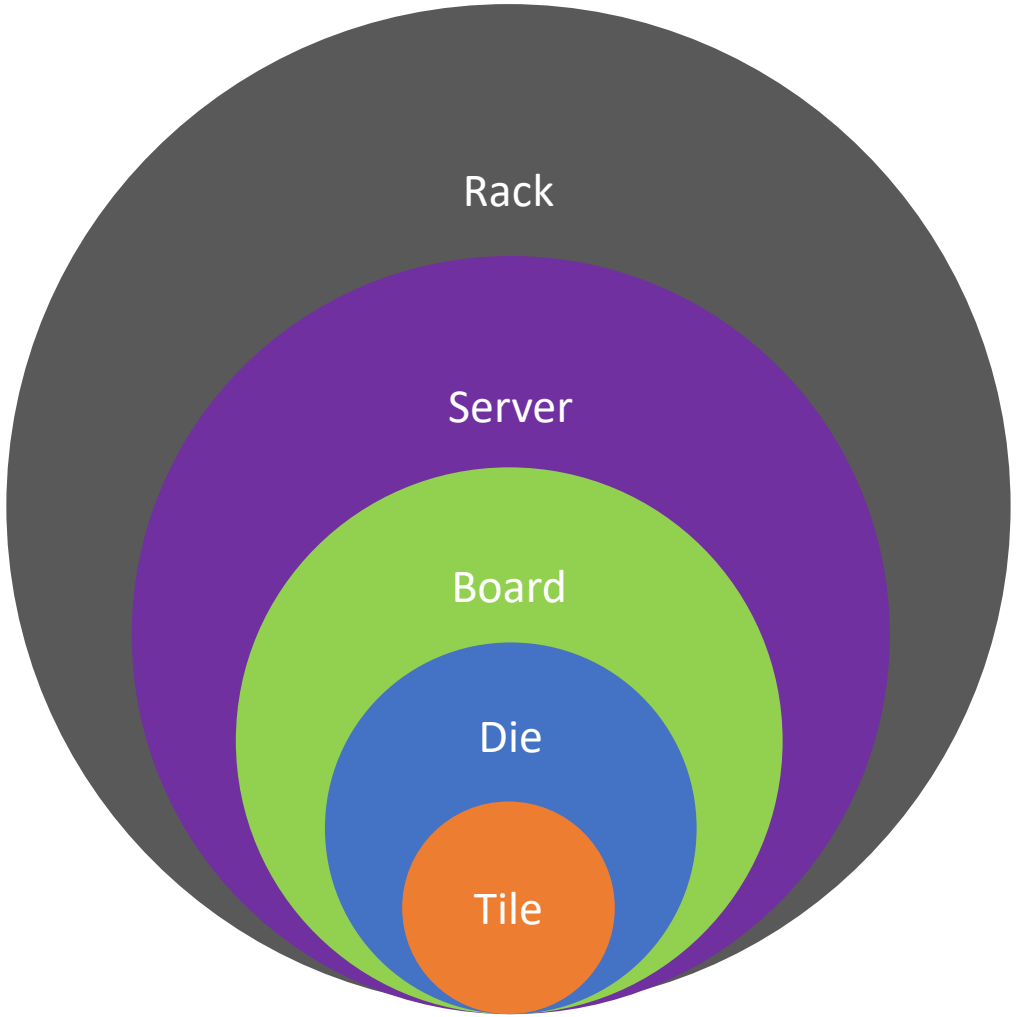
GCN
Data Flow



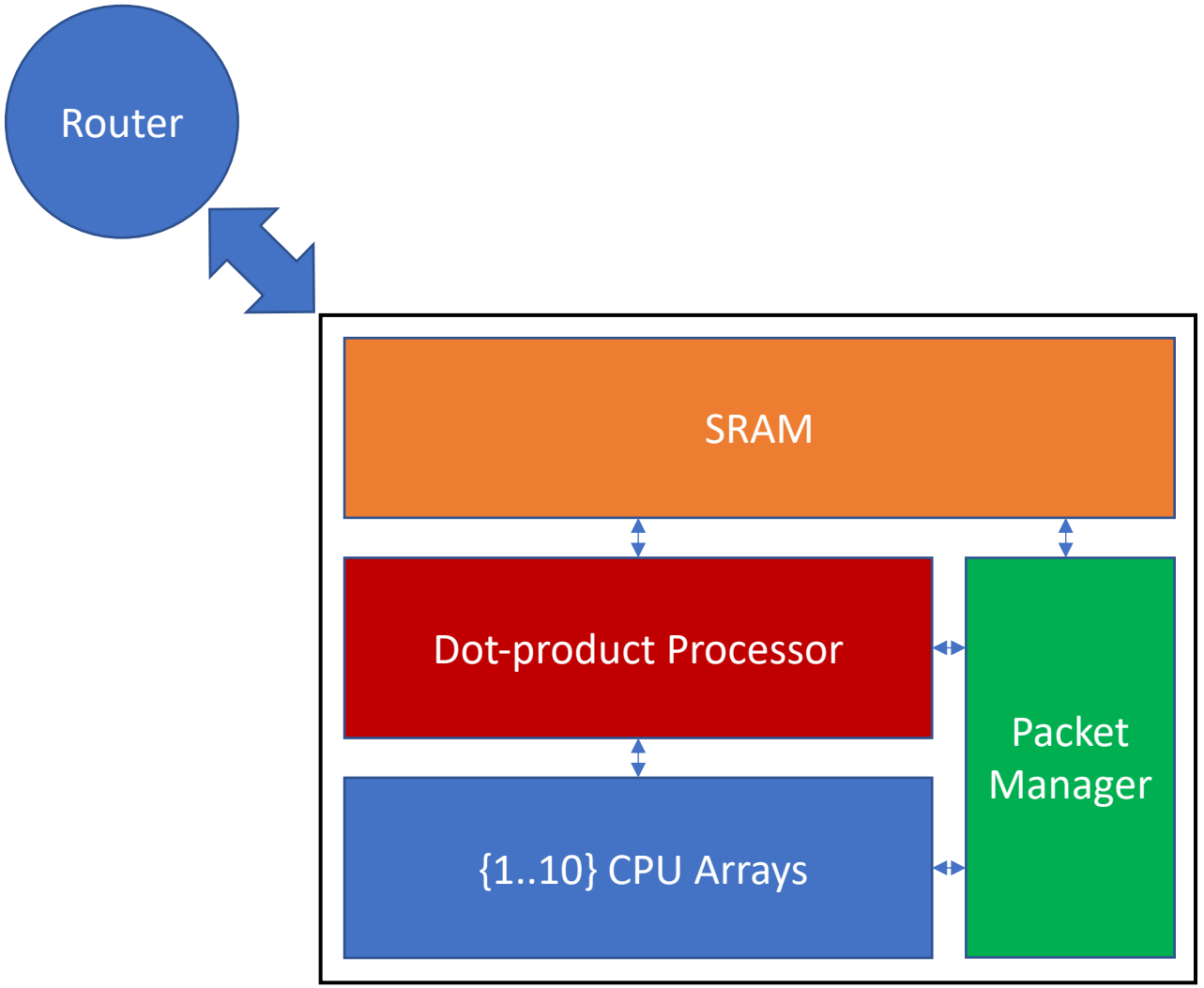
Outline

- Computational View of Graph Neural Network
- Spatial Architectures – Simplified View
- Mapping GNN onto Spatial Architectures

Building Blocks Spatial Accelerators

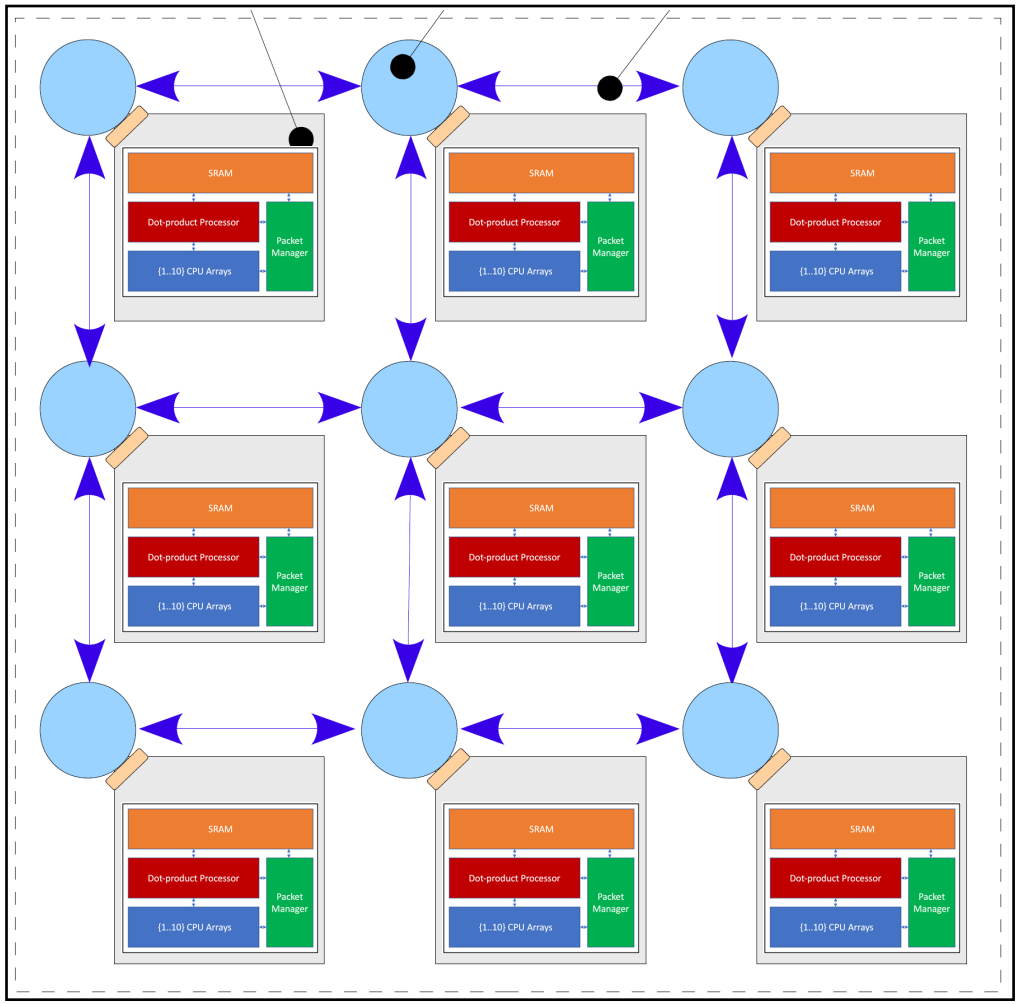


Building Blocks Spatial Accelerators – a Tile



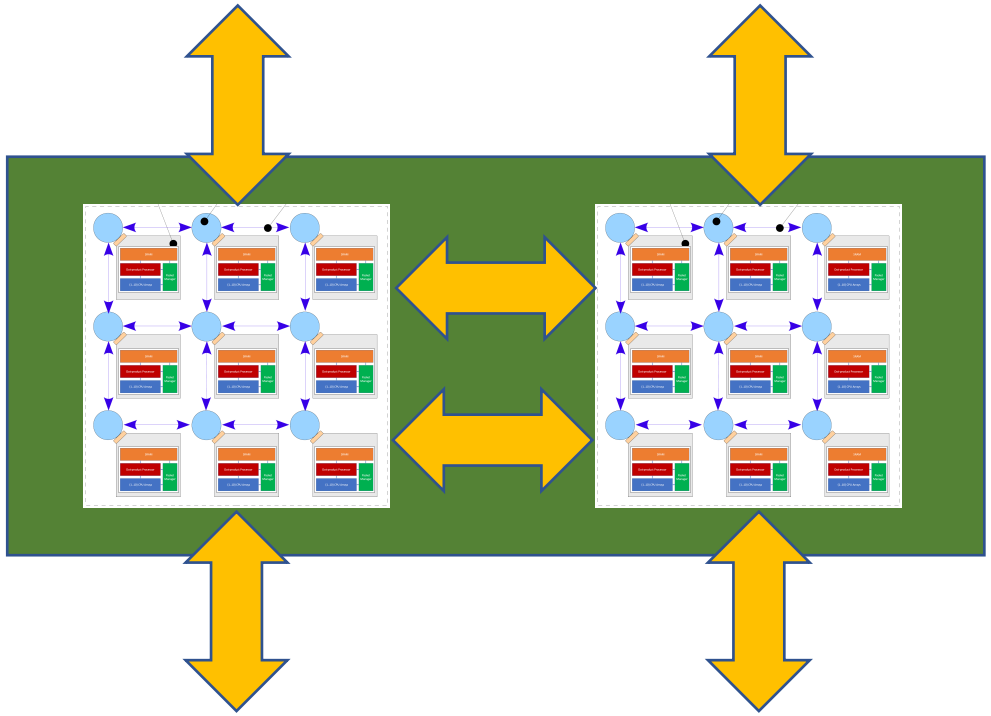
Building Blocks Spatial Accelerators – a Die

- Most tiled based architecture will have an array of tiles connected over a NOC
- e.g., 120 tiles in Tenstorrent
- 1216 tiles in GraphCore IPU
- Most common NOC topologies are 2-D Mesh and 2-D Torus
- Transactions are packetized



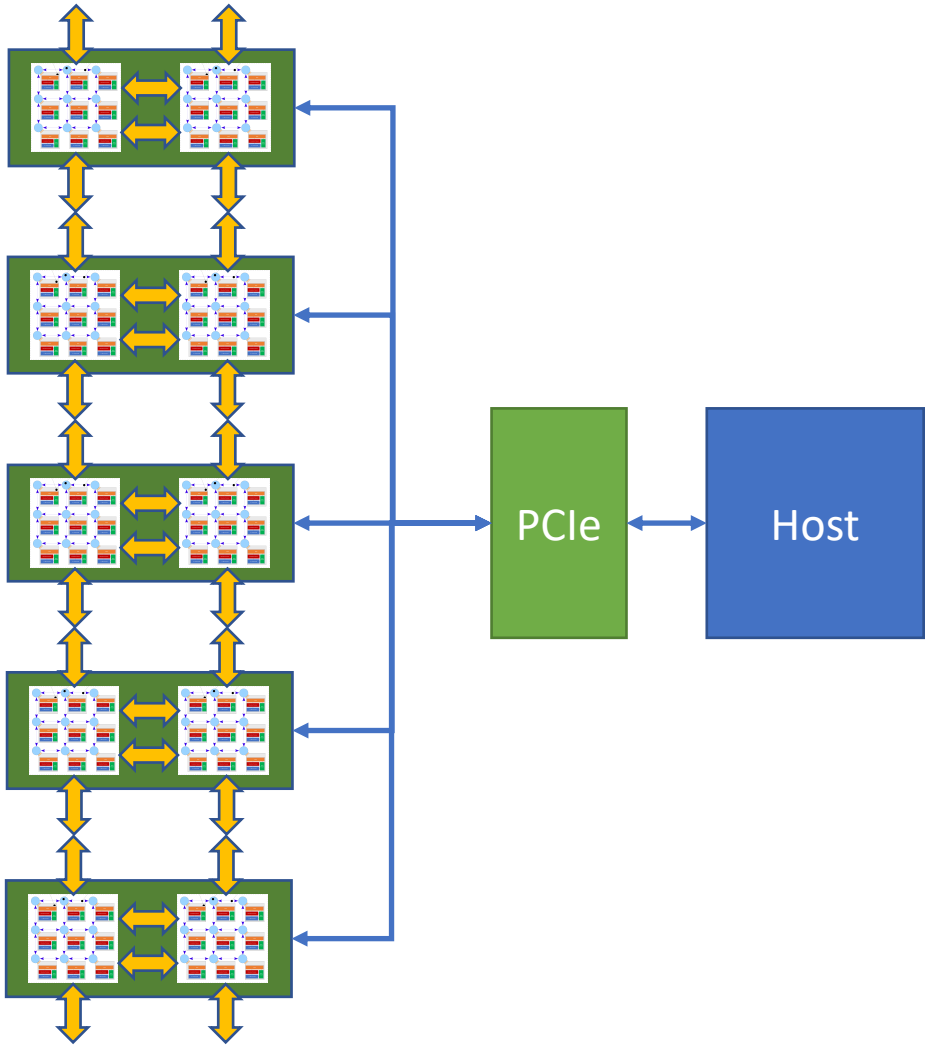
Building Blocks Spatial Accelerators – a Board

- 2 or more Die share a board
- The Die/Chip connected via dedicated bus (which is part of a system-wide interconnect)
- e.g., 3 x 64 GB/s inter-die connections
- e.g., 4 x 64 GB/s inter-board connections
- North/South connections are also part of system-wide interconnect

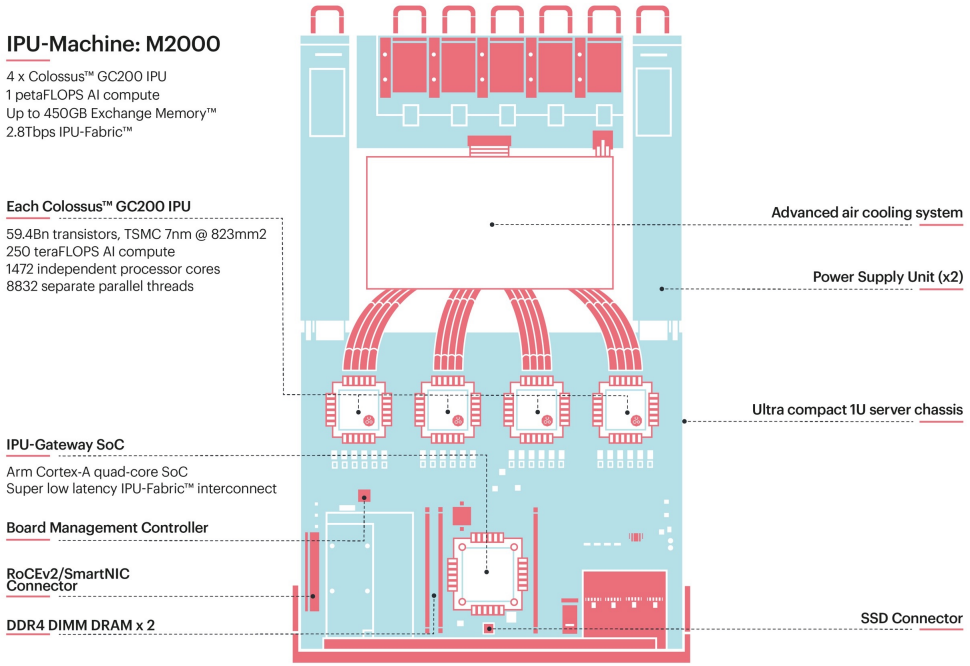


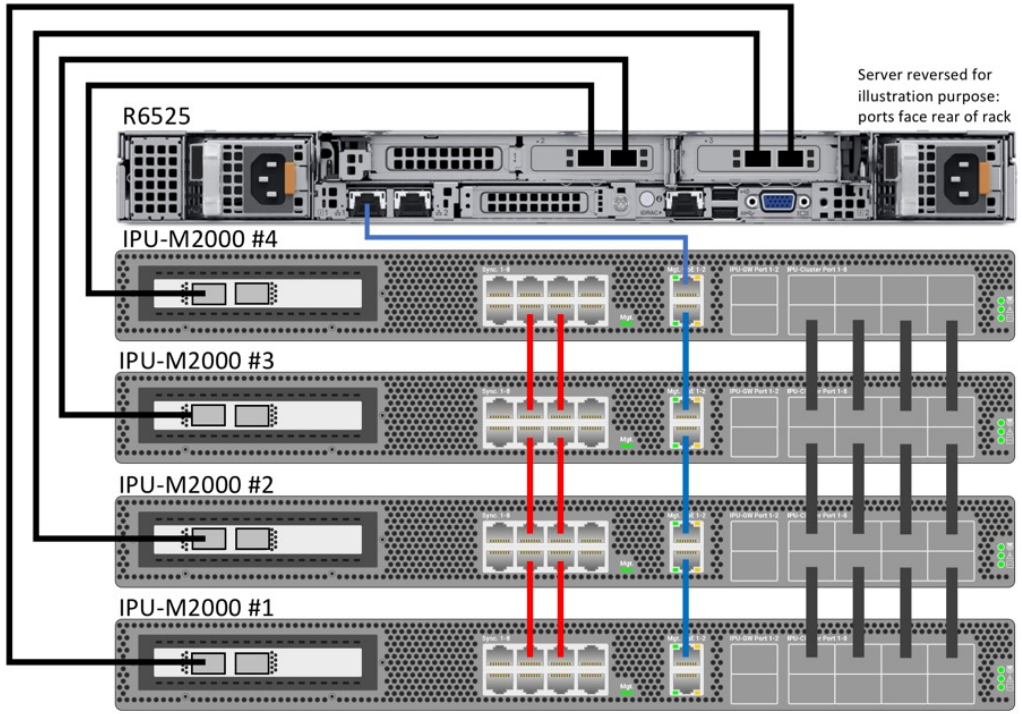
Building Blocks Spatial Accelerators – a Server

- 16/32/64 or more boards are connected over a system-wide interconnect
- North/South/East/West connections are also part of system-wide interconnect
- Interconnect – proprietary or ethernet protocol



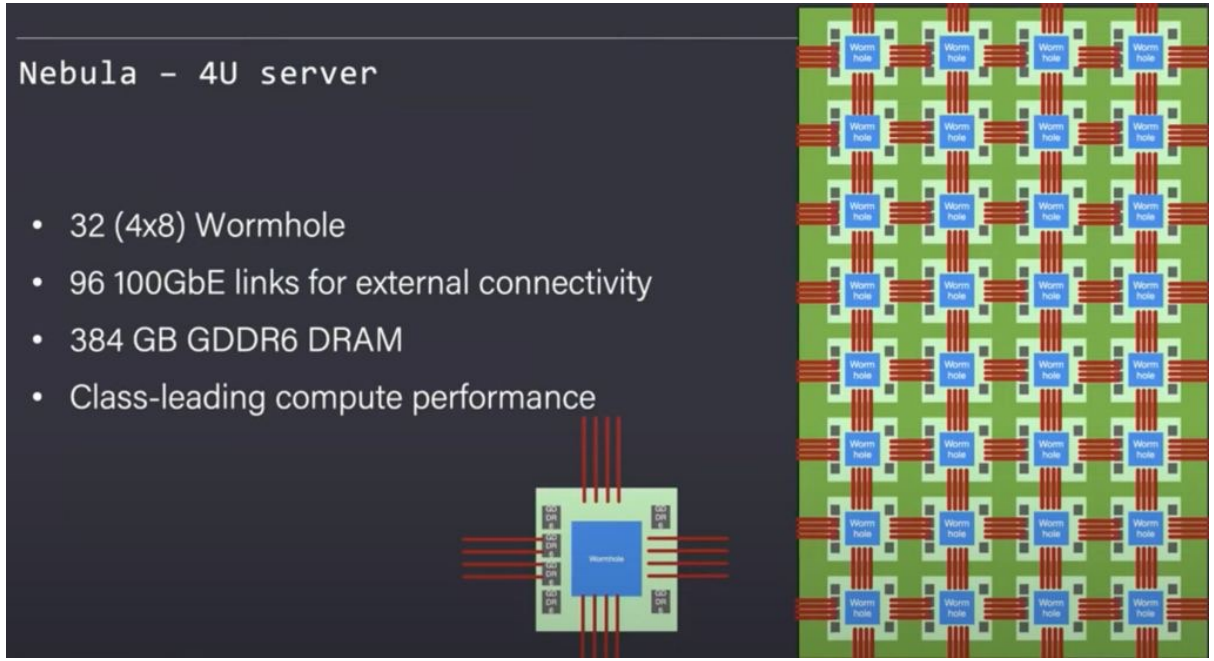
Building Blocks Spatial Accelerators – a Rack

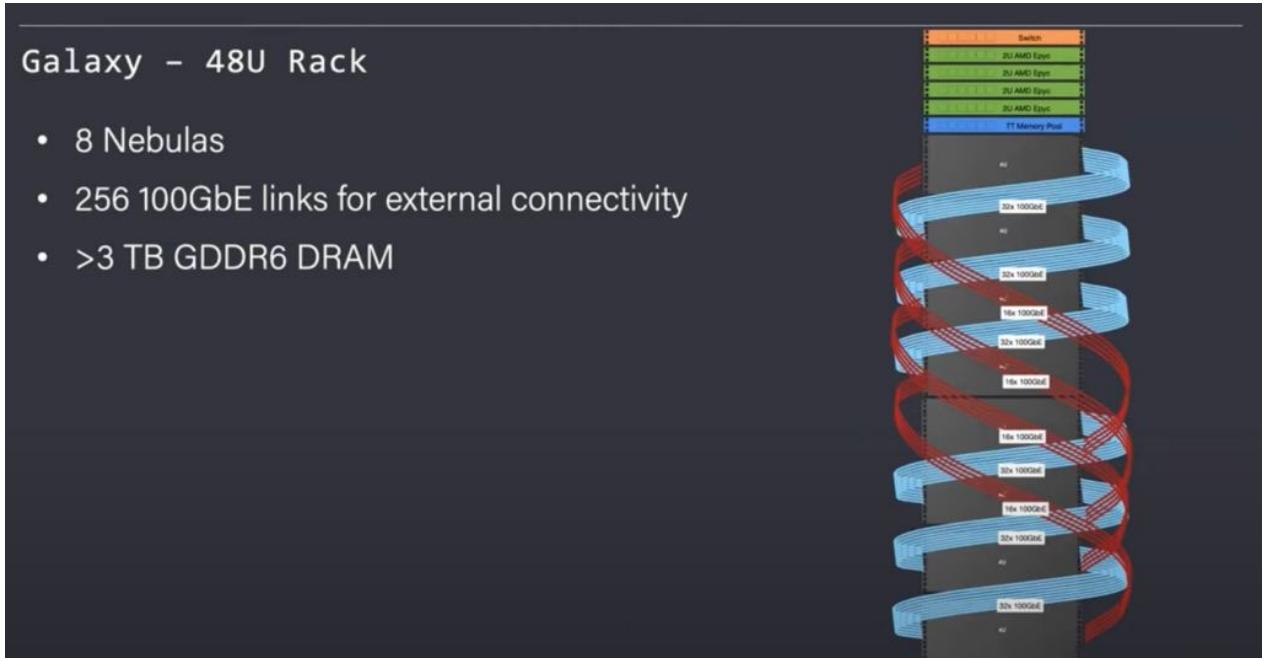




Source: https://www.nhr.kit.edu/userdocs/ftp/graphcore/

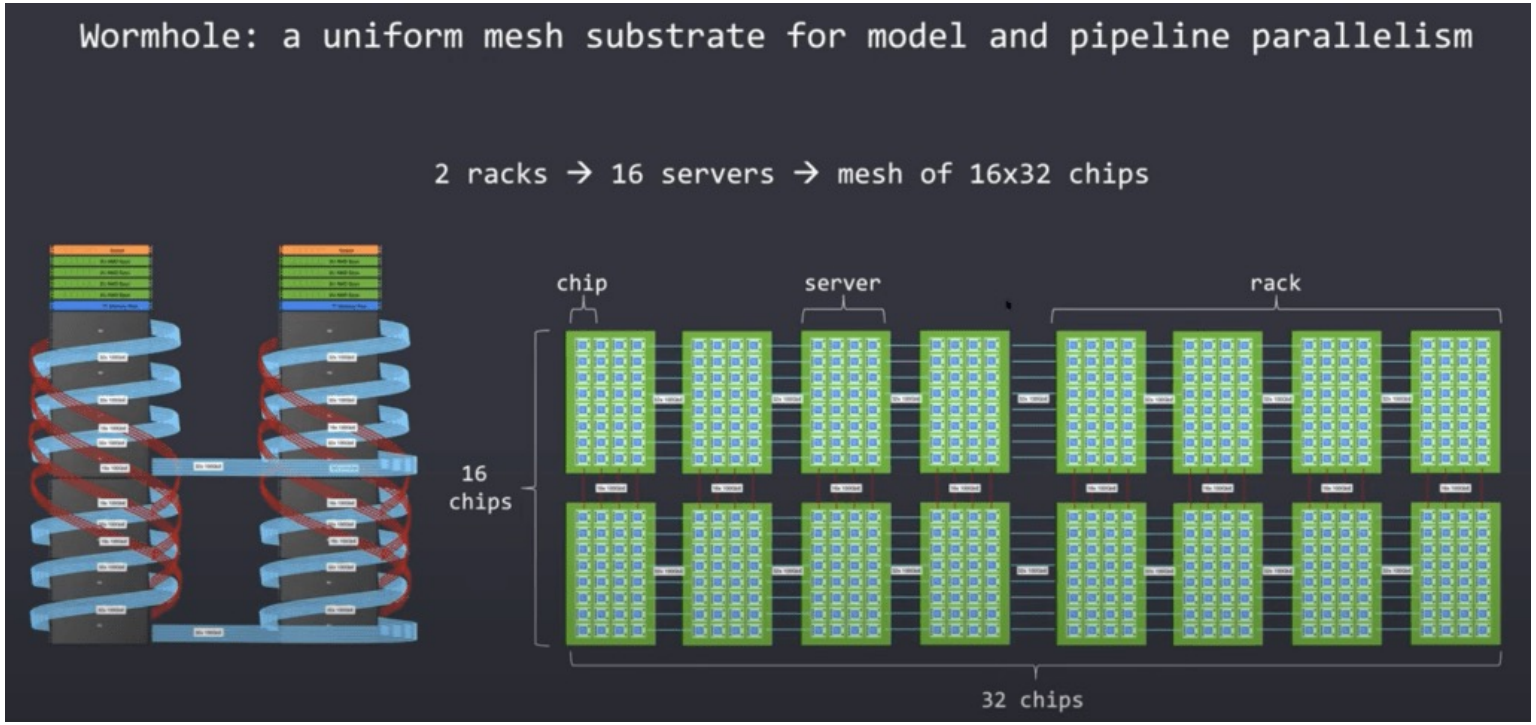
Building Blocks Spatial Accelerators – a Rack





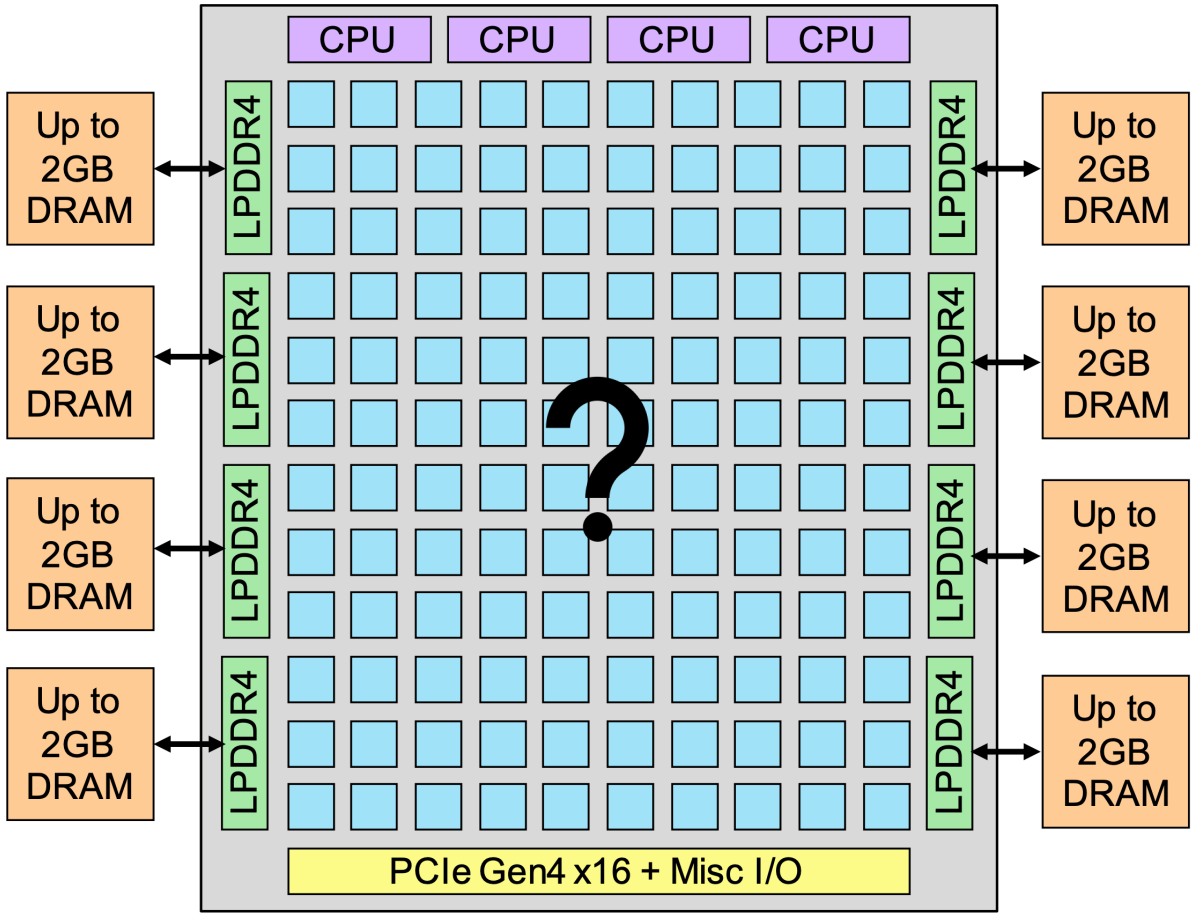
Source: https://semianalysis.com/tenstorrent-wormhole-analysis-a-scale-out-architecture-for-machine-learning-that-could-put-nvidia-on-their-back-foot/

Building Blocks Spatial Accelerators – a scale out rack



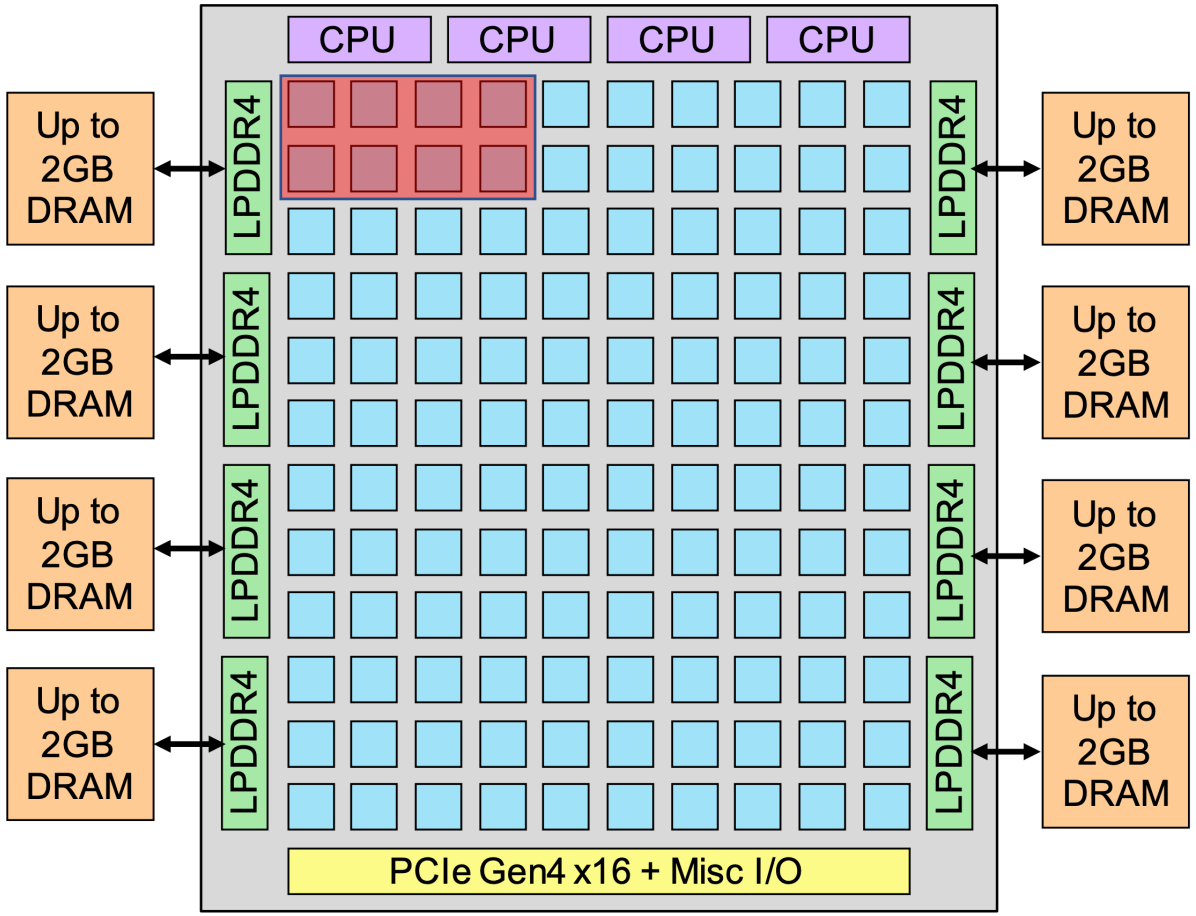
Source: https://semianalysis.com/tenstorrent-wormhole-analysis-a-scale-out-architecture-for-machine-learning-that-could-put-nvidia-on-their-back-foot/

120 core spatial accelerator



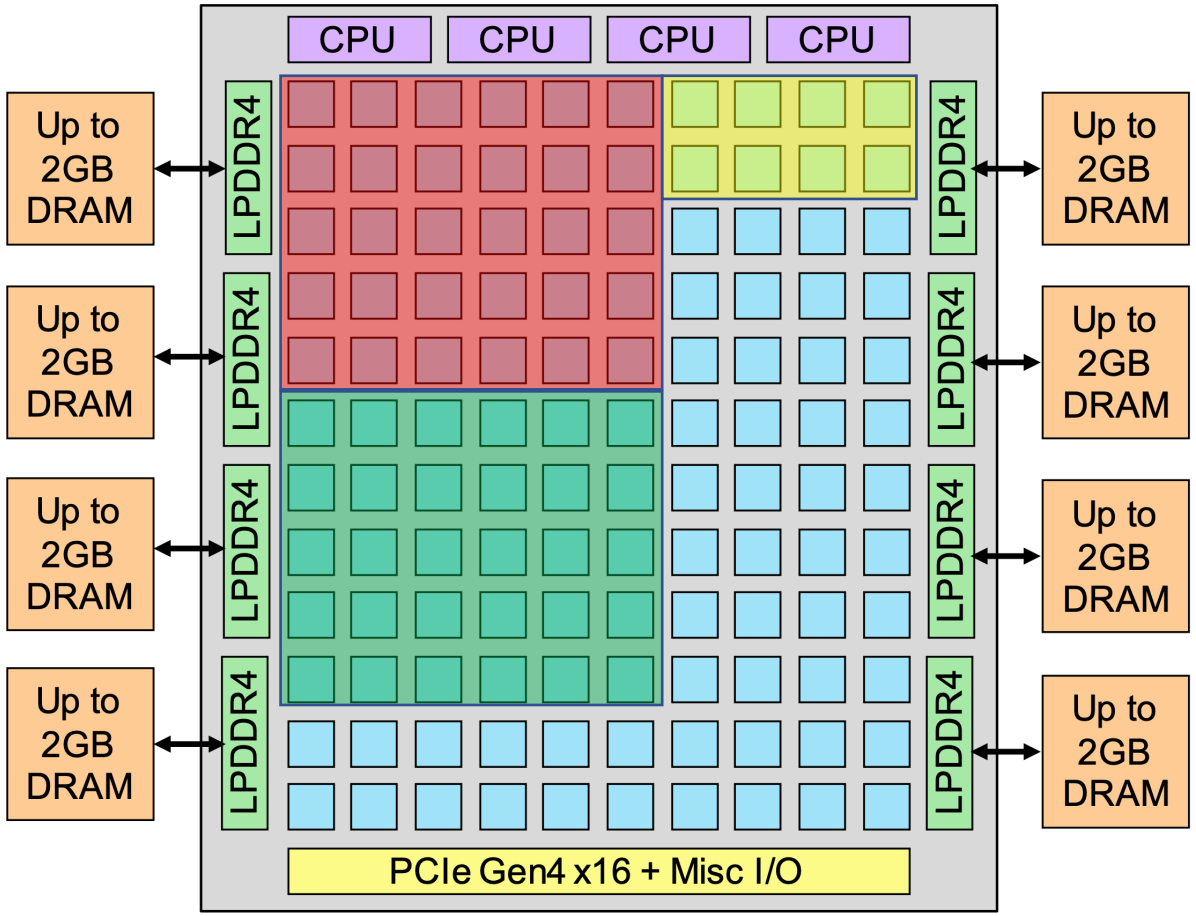
Source: https://www.linleygroup.com/mpr/article.php?id=12287

Layer mapping



Source: https://www.linleygroup.com/mpr/article.php?id=12287

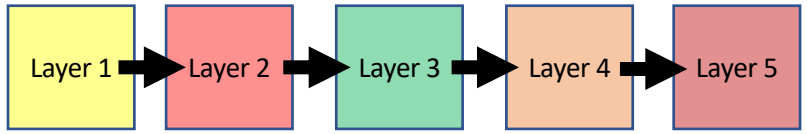
Layer mapping

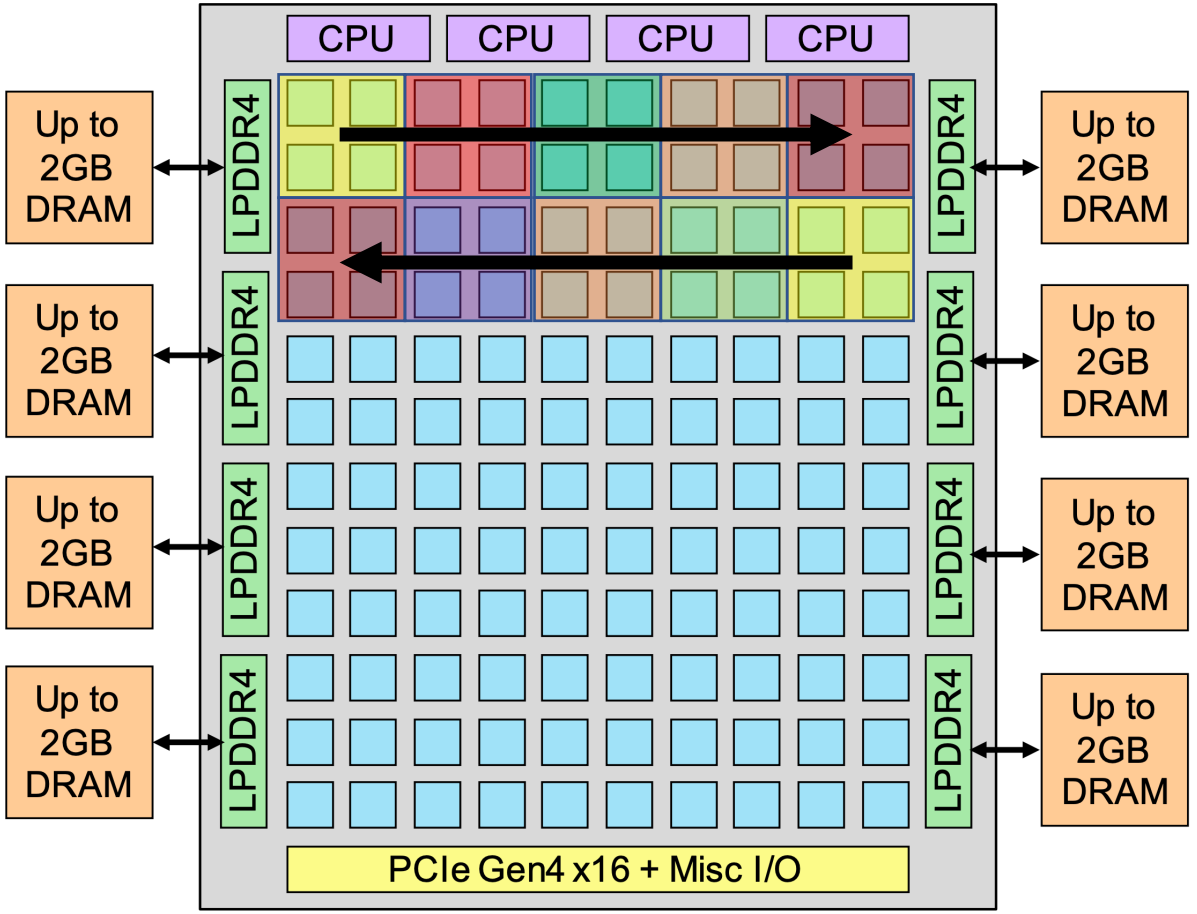


Source: https://www.linleygroup.com/mpr/article.php?id=12287

Layer mapping – general practice

- Spatial mapping of layers
- Load balancing



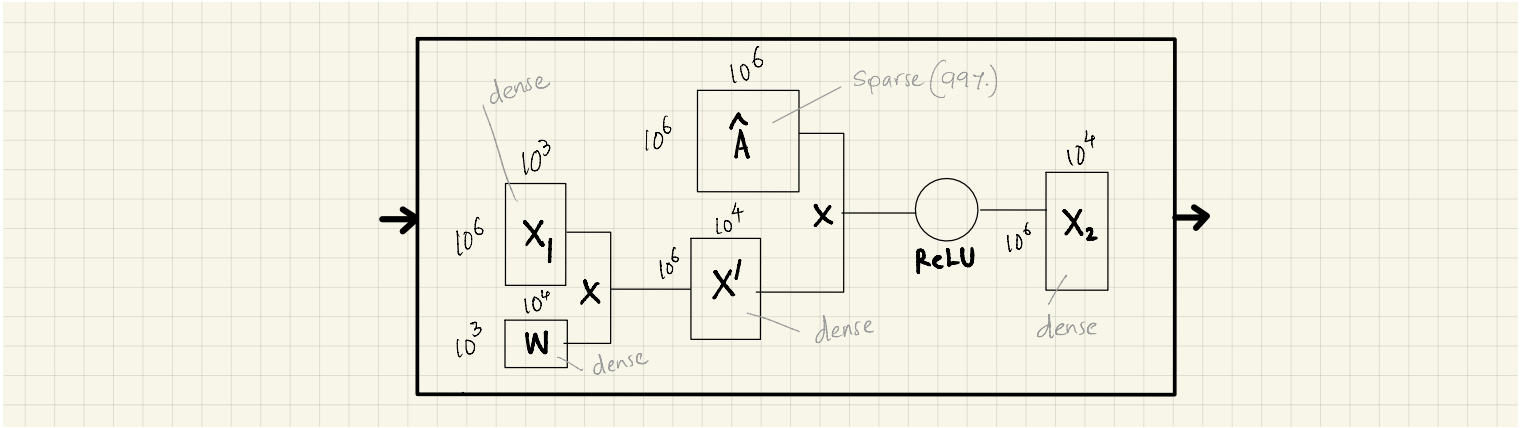


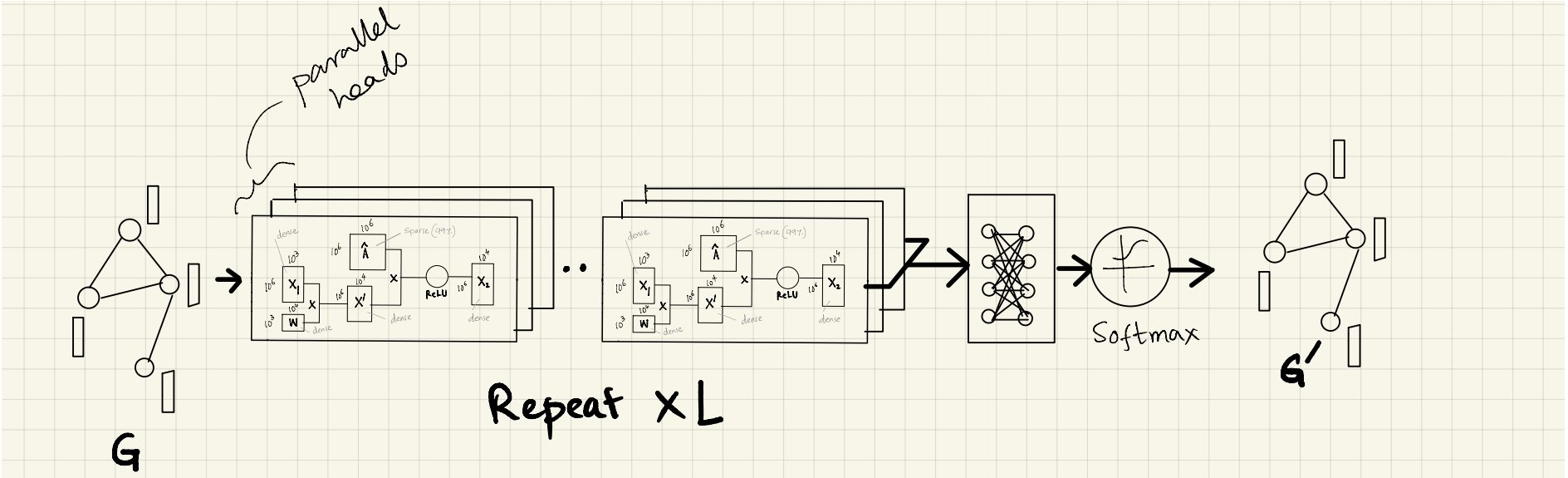
Source: https://www.linleygroup.com/mpr/article.php?id=12287

Outline

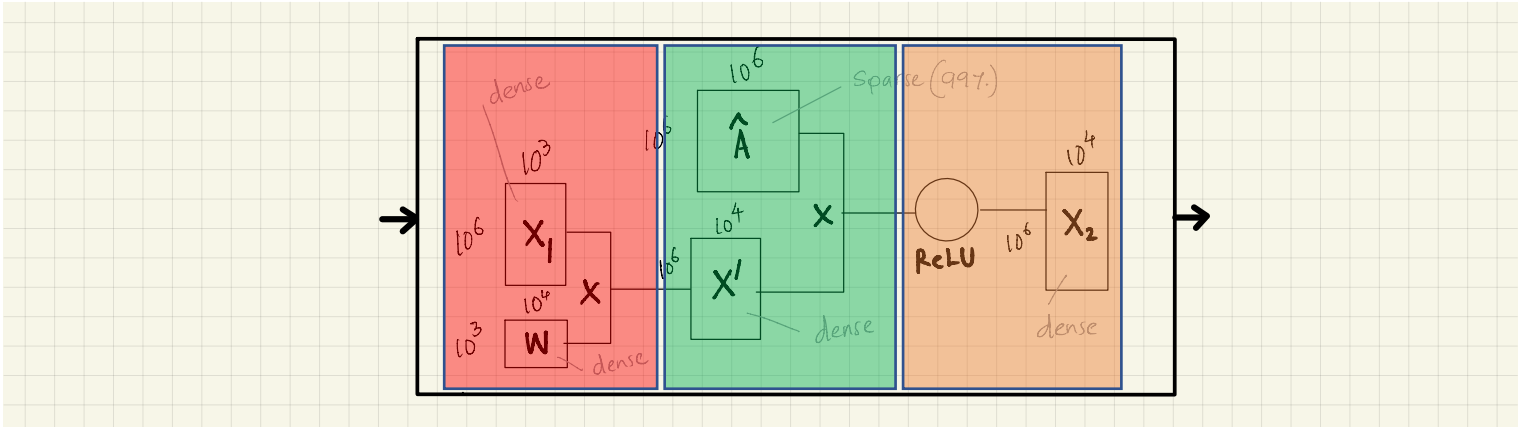
- Computational View of Graph Neural Network
- Spatial Architectures – Simplified View
- Mapping GNN onto Spatial Architectures

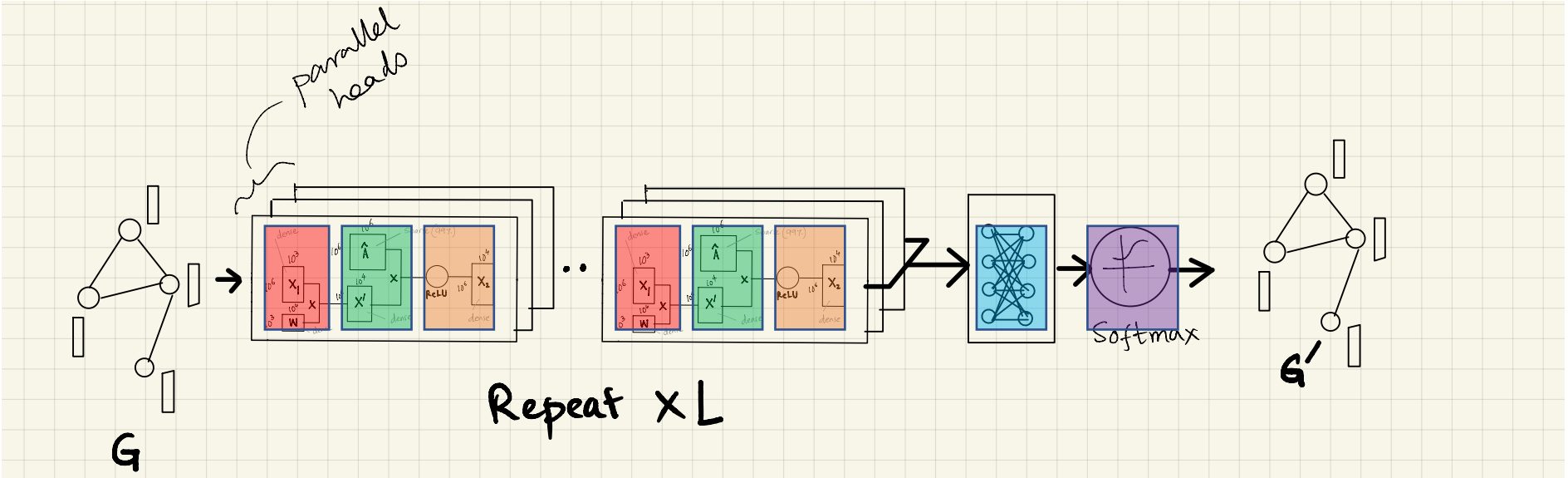
Back to GNN – How do we map GNN on to a spatial architecture?



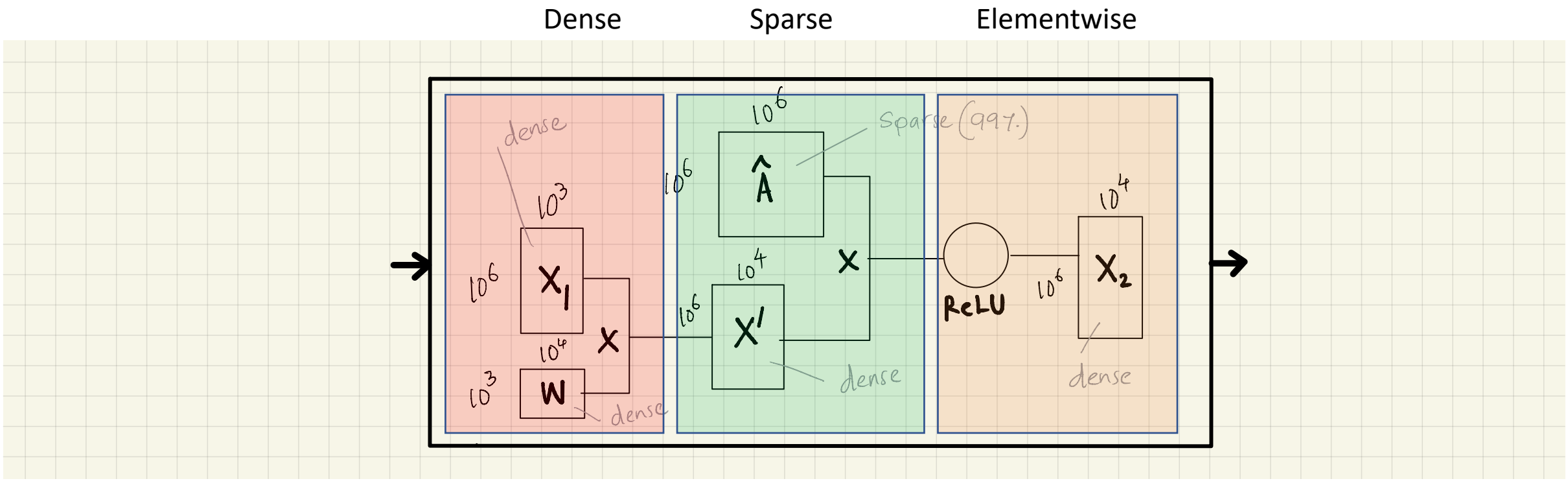


Back to GNN – How do we map GNN on to a spatial architecture?





Relative Arithmetic Intensity per layer



Arithmetic Intensity Scale: 10000 100 1

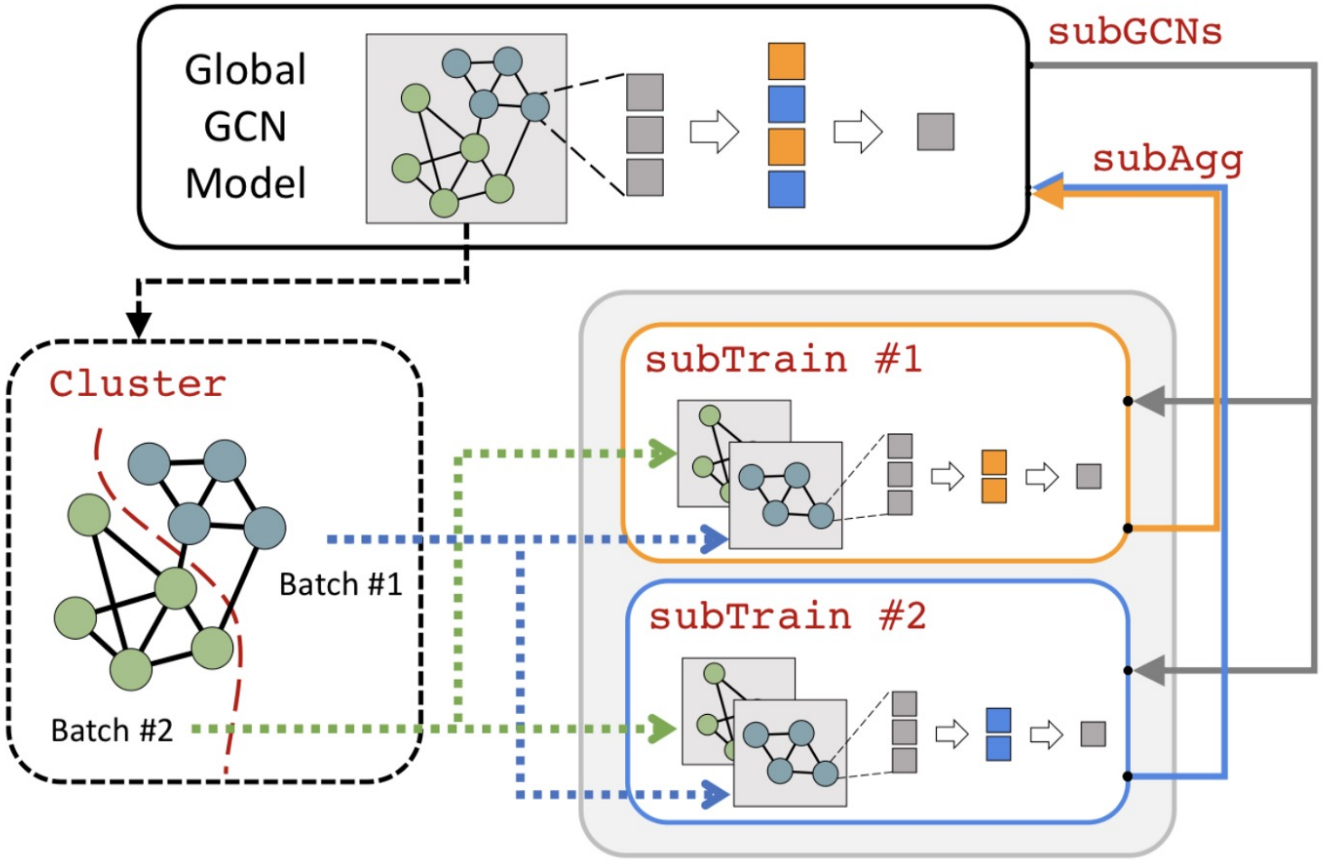
Industrial Graphs >99.9% Sparsity

Challenges:

- **Algorithmic:** Long-distance nodes loses expressivity (averages out)
- **Computational:** High memory footprint for full graph
  - Tensors are huge
    - Features: 100M nodes, 256 features/Node, 4 Layer GNN, FP32 → 400GB
    - Adjacency matrix: 100M x 100M
  - Gather-Scatter traffic
- Nobody in the industry operate on full graph (except in molecular AI)

Dataset	Nodes	Edges
Yelp	716,847	6,977,410
Amazon	1,598,960	132,169,734
OAG-Paper	15,257,994	220,126,508
Twitter-Full	~15,500,000	?
OGBN-products	2,449,029	123,718,280
OGBN-Papers100M	111,059,956	1,615,685,872
WebCrawl2012	~3,500,000,000	~128,000,000,000

Partitioned GNNs

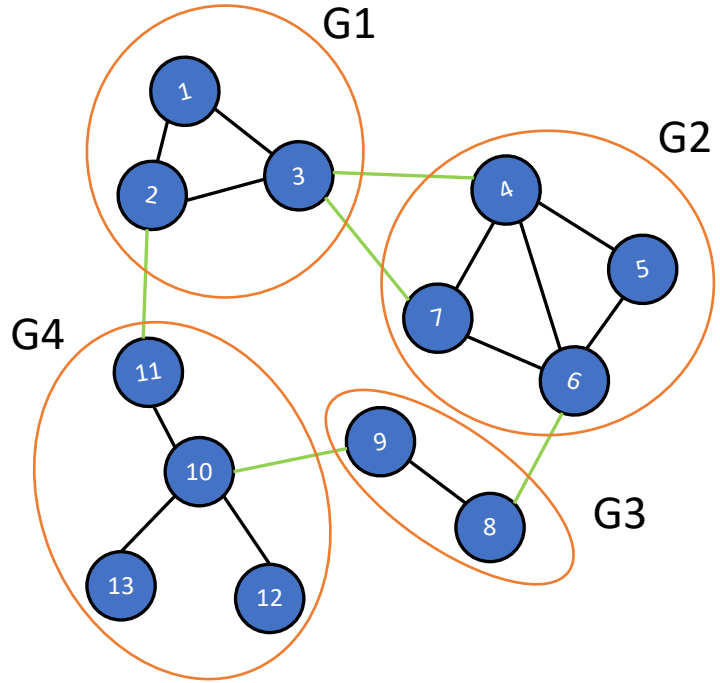


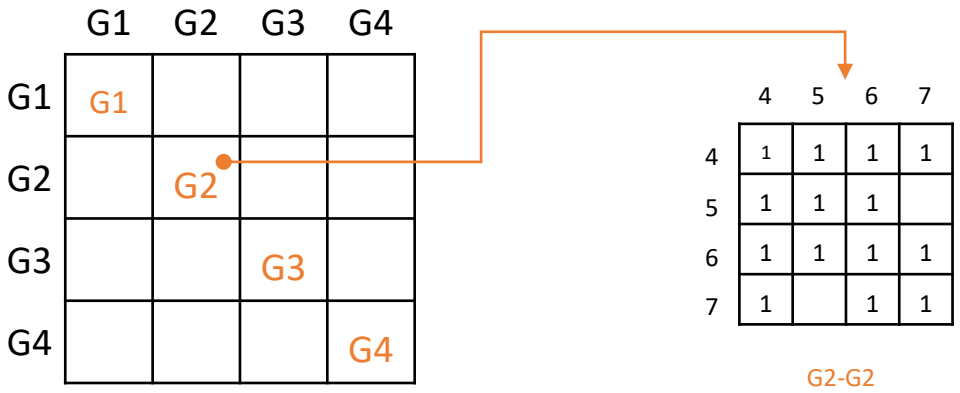
GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron R. Wolfe, Jingkang Yang, Arindam Chowdhury, Chen Dun, Artun Bayer, Santiago Segarra, Anastasios Kyrillidis

Density-based partitioning

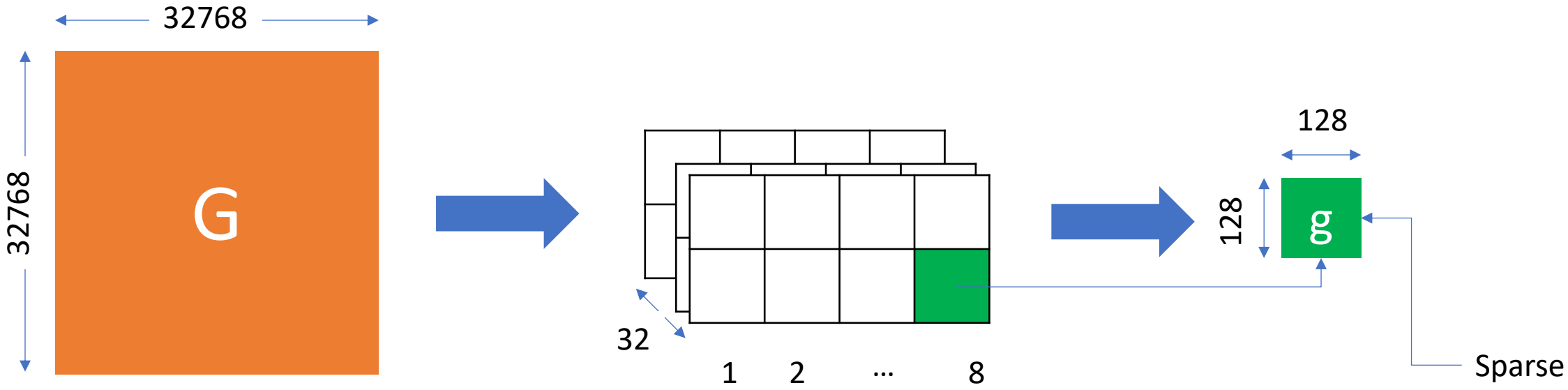
- Find densely connected clusters
- Construct new batch from clusters
- Train on sub-graphs



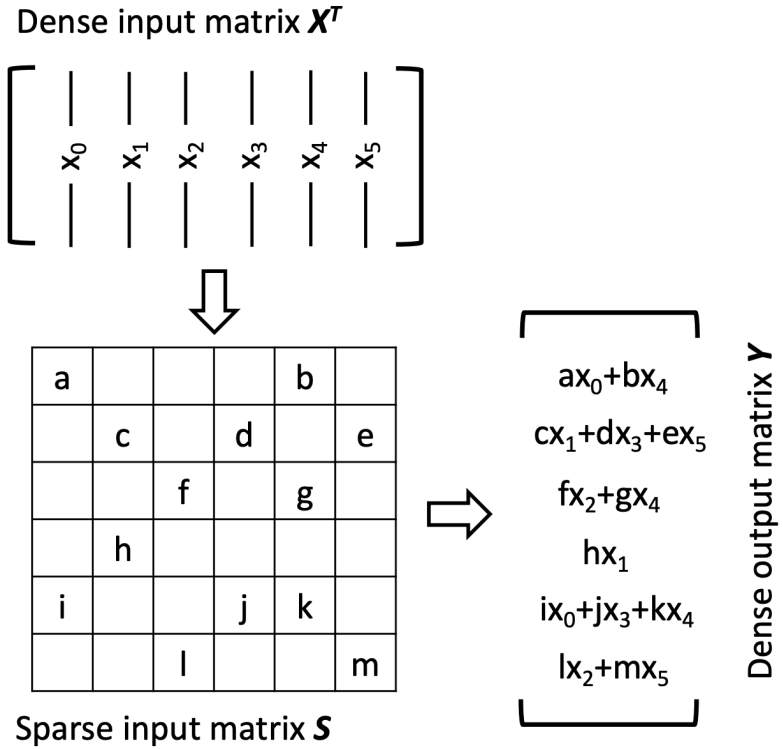


$$\hat{A} = \sum G + \sum S$$

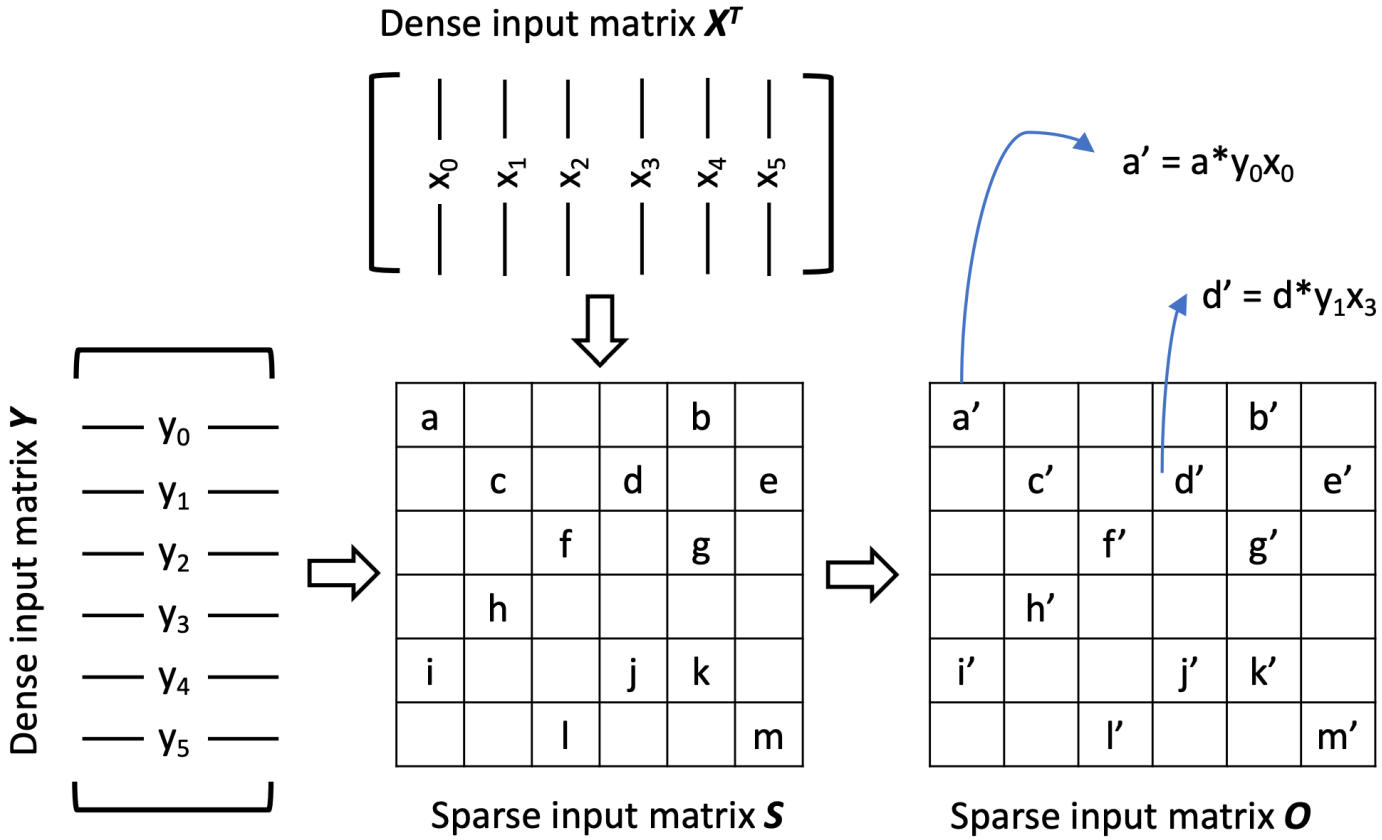
Density-based partitioning



Sparsity does not go away!



SpMM (GCN)



SDDMM (GAT)

Sparse MatMul (SpMM)

matrix

a	b		
	c	d	
e		f	g
			h

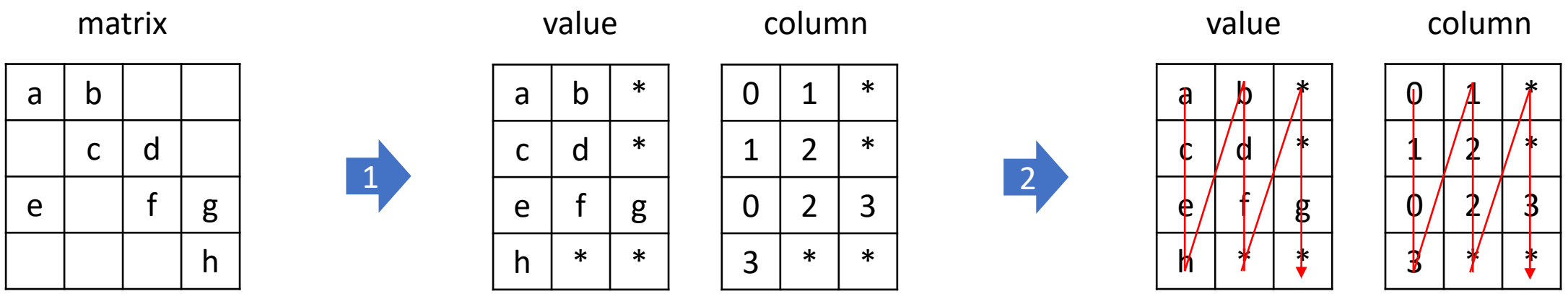
value

a	b	*
c	d	*
e	f	g
h	*	*

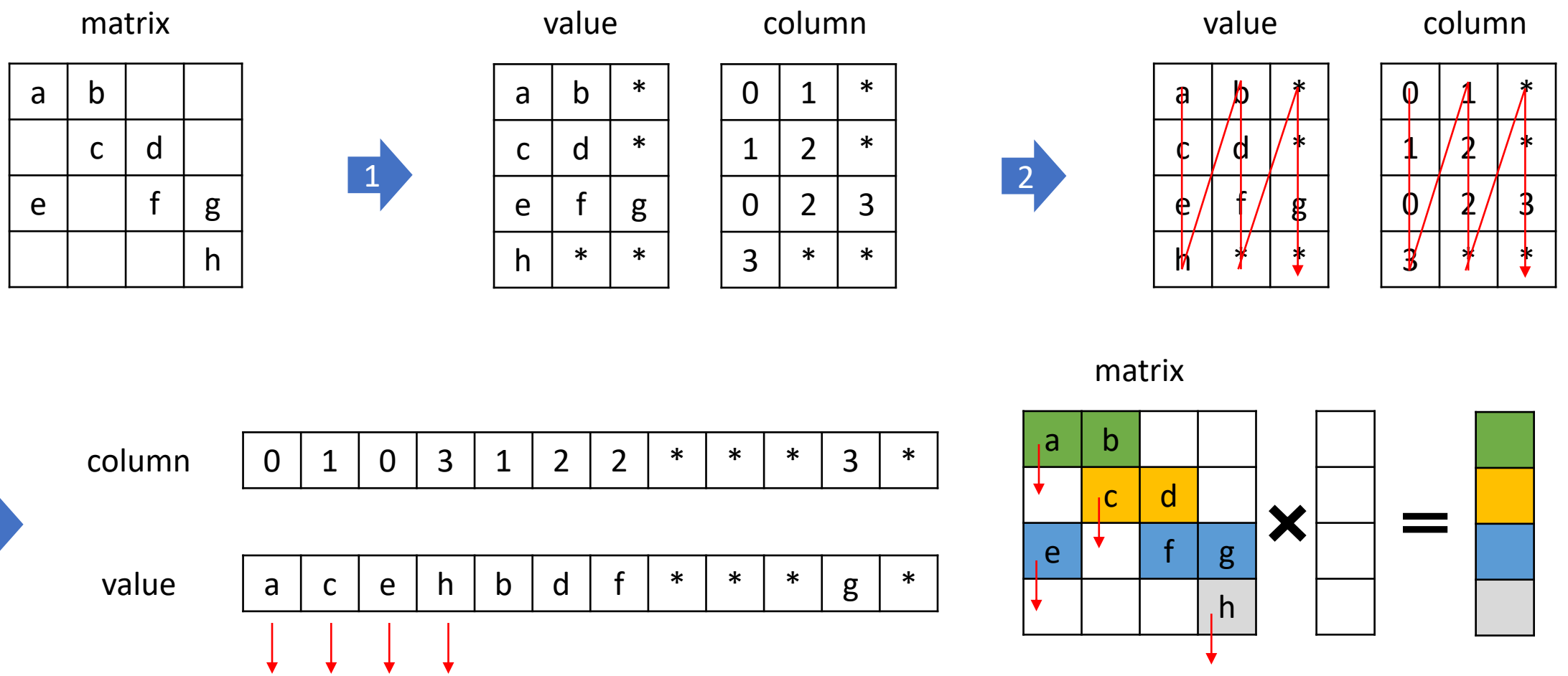
column

0	1	*
1	2	*
0	2	3
3	*	*

Sparse MatMul (SpMM)



Sparse MatMul (SpMM)



End-to-end flow

- Find strong clusters
- Partition into sub-graphs
- Create mini-batch from sub-graphs
- Store sub-graphs into compressed storage format
- SpMV is processed by walking through the column indices
- Partial sums are updated in every cycle
- Place every sub-graphs into fixed number of device tiles
- Sub-graphs are balanced in the spatial architecture

Outline

- Computational View of Graph Neural Network
- Spatial Architectures – Simplified View
- Mapping GNN onto Spatial Architectures

Take-home summary

- Spatial architectures are becoming very common as large models start to dominate (Scale out)
- Graph NN workloads are different from other DNNs
- Most industrial graphs are highly sparse (>99.9%)
- Huge scope for Innovation in sparse workloads like graphs
  - e.g., TACO: http://tensor-compiler.org/ at Stanford
- Like spatial ASICs, FPGAs could be well suited to many emerging models