Union: A Unified HW-SW Co-Design Ecosystem in MLIR for Evaluating Tensor Operations on Spatial Accelerators

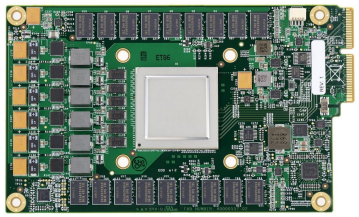
Geonhwa Jeong[1], Gokcen Kestor[2], Prasanth Chatarasi[3], Angshuman Parashar[4], Po-An Tsai[4], Sivasankaran Rajamanickam[5], Roberto Gioiosa[2] and Tushar Krishna[1]

[1]Georgia Tech [2]Pacific Northwest National Laboratory [3]IBM Research [4]NVIDIA [5]Sandia National Laboratories

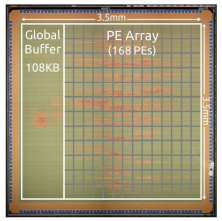
Email: geonhwa.jeong@gatech.edu

Georgia Tech

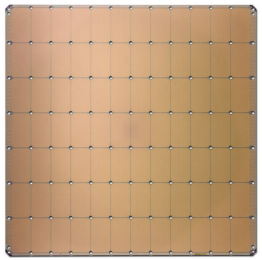
The Era of Domain-Specific Accelerators



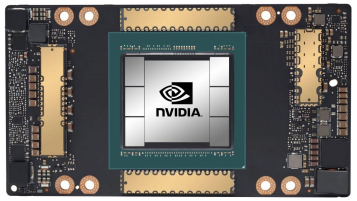
Google TPU



Eyeriss



Cerebras WSE-2



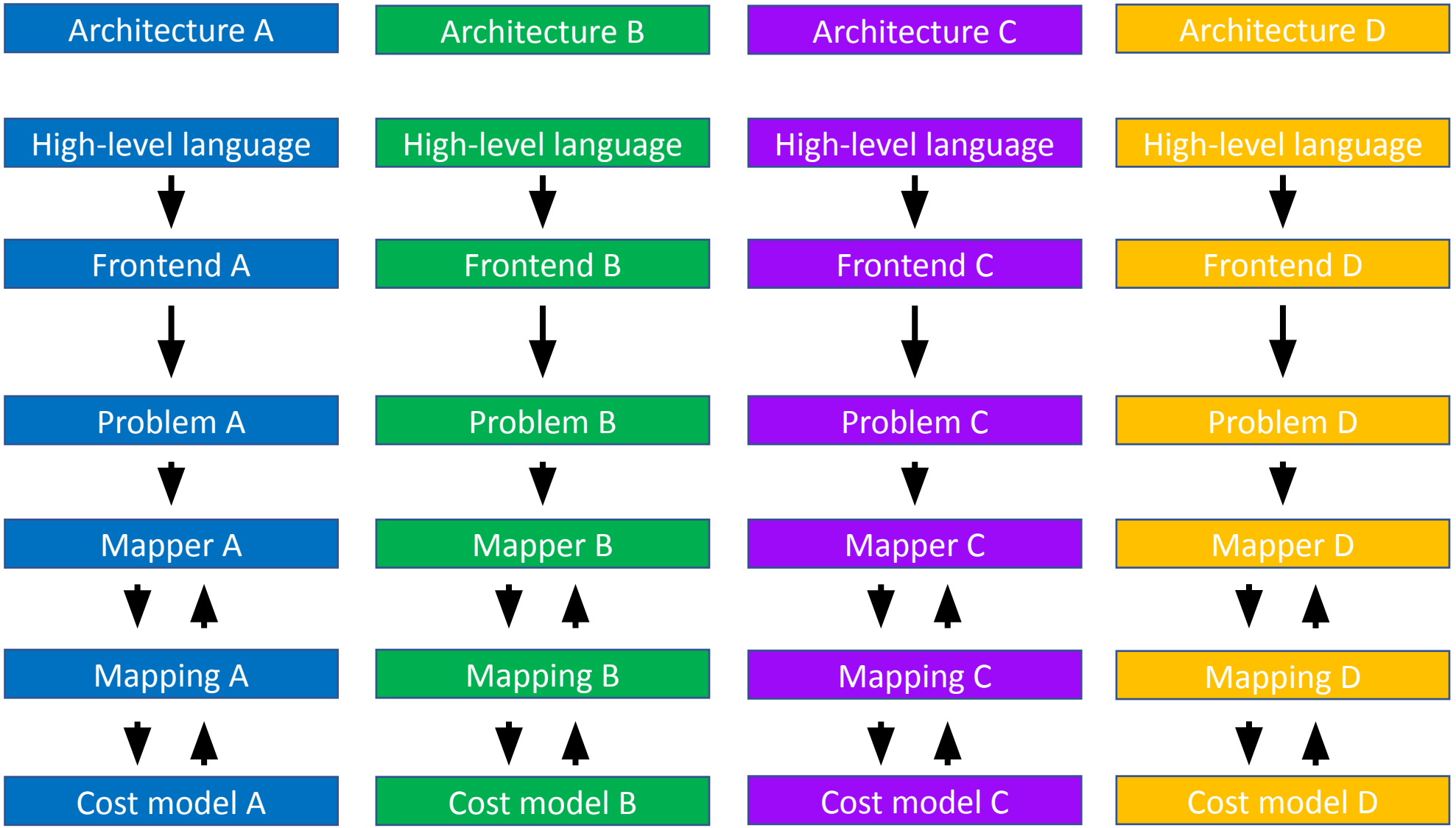
NVIDIA Ampere GPU

- **Moore's law and Dennard's scaling do not work anymore.**
- **Accelerators include large parallel compute units to meet the extreme compute demands.**

Innovation
Novel memory hierarchy
Efficient interconnection network
Custom processing elements

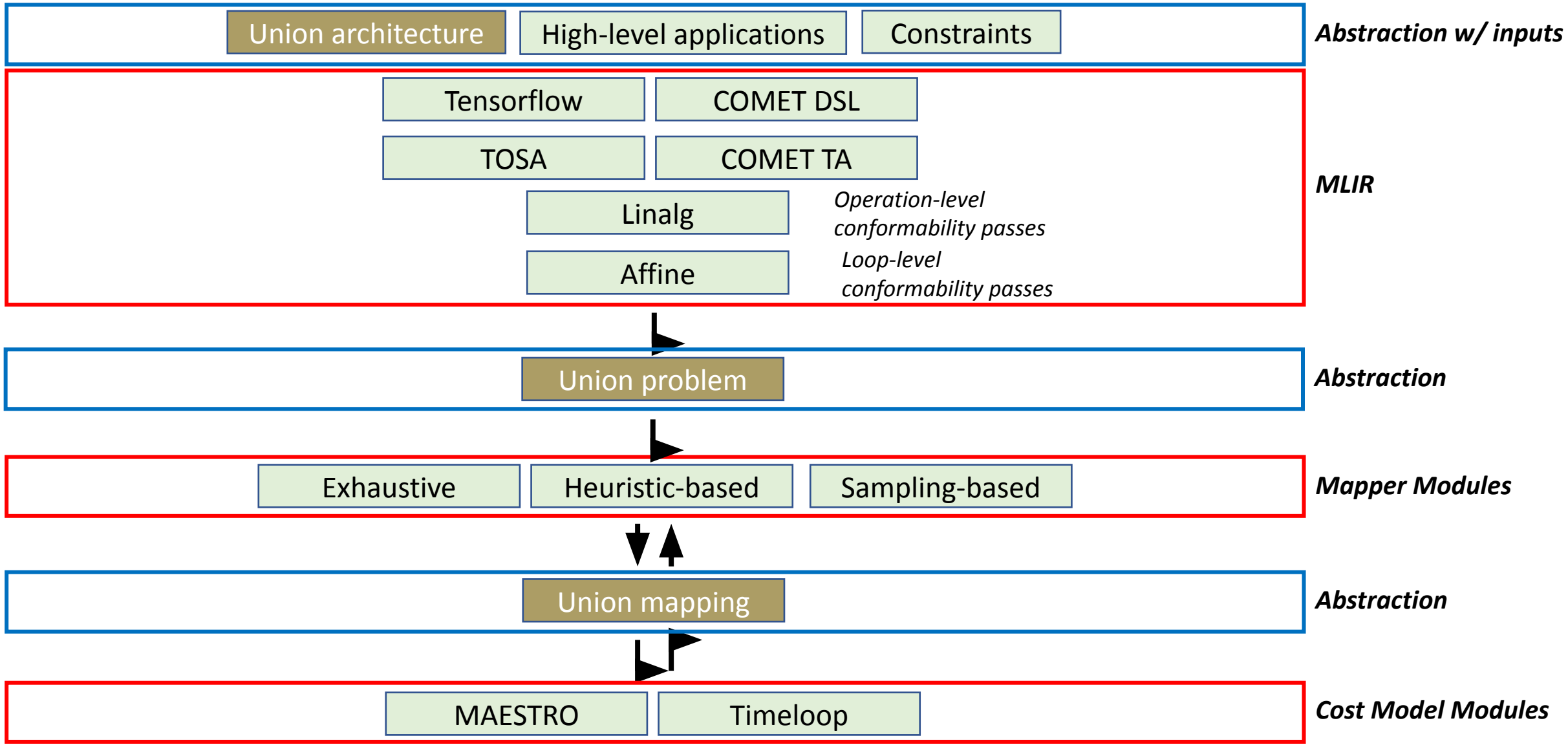
Fragmentation
Custom compiler toolchain
Duplicate engineering overhead
Error-prone frameworks

Fragmentation Example

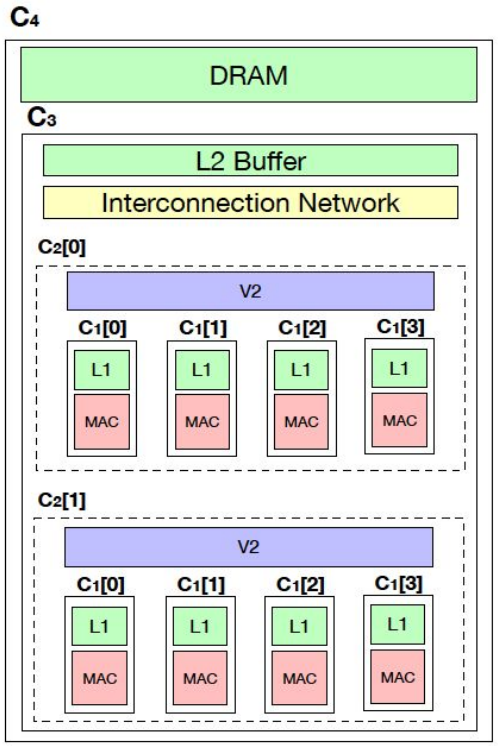


and many more!

Union Overview



Union Abstractions




```
Problem:
  Operation: GEMM
  Shape:
    Name: Example
    Dimensions: [M, N, K]
    Data-space:
      - Name: Input
        Projection:
        - [ [M], [K] ]

      - Name: Weight
        Projection:
        - [ [K], [N] ]

      - Name: Output
        Projection:
        - [ [M], [N] ]
        Read-write: true

  Instance:
    M: 16
    N: 64
    K: 32
```

Union Problem

```
Name: C4
  Virtual: False
  Dimension: X
  Local:
    Memory: DRAM
  Sub-tree:
    Name: C3
      Virtual: False
      Dimension: Y
      Local:
        Memory: L2 Buffer

      Sub-tree:
        Name: C2[1...2]
          Virtual: True
          Dimension: X

          Sub-tree:
            Name: C1[1...4]
              Virtual: False
              Local:
                Memory: L1 Buffer
                Compute: MAC Unit
```

Union Architecture

```
// C4: DRAM to L2
target_cluster: C4
temporal_order: MNK
temporal_tile_sizes: 16, 32, 16
spatial_tile_sizes: 16, 32, 16

// C3: L2 to V2
target_cluster: C3
temporal_order: MNK
temporal_tile_sizes: 8, 16, 8
spatial_tile_sizes: 8, 8, 8

// C2: V2 to L1
target_cluster: C2
temporal_order: MNK
temporal_tile_sizes: 8, 8, 8
spatial_tile_sizes: 8, 8, 2

// C1: L1 to MAC
target_cluster: C1
temporal_order: MNK
temporal_tile_sizes: 1, 1, 1
spatial_tile_sizes: 1, 1, 1
```

Union Mapping

Conclusion

- We propose **Union**, a unified framework for evaluating tensor operations on spatial accelerators with unified abstractions.
- Our MLIR based framework allows to map **both HPC and ML tensor operations using multiple mappers to multiple cost models** for spatial accelerators.
- **We present a few case studies** to demonstrate the flexibility of the framework by evaluating different operations, mappings, and hardware features with a single framework.

Question? Please send me an email:
geonhwa.jeong@gatech.edu

Thank you for listening! This work is accepted to PACT'21.
Code available at https://github.com/union-codesign/union