SAND2021-11383C

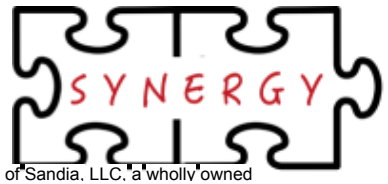
PACT 2021

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



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

People

Geonhwa Jeong
PhD Student
Georgia Tech

Gokcen Kestor
Research Staff
Pacific Northwest National Laboratory

Prasanth Chatarasi
Research Staff Member
IBM Research

Angshuman Parashar
Research Staff
NVIDIA

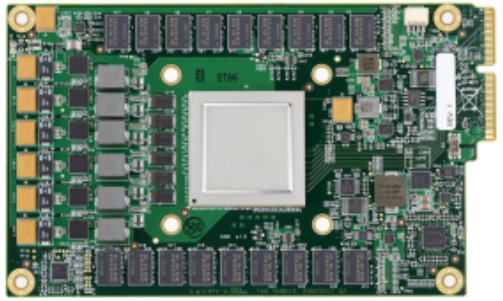
Po-An Tsai
Research Staff
NVIDIA

Sivasankaran Rajamanickam
Research Scientist
Sandia National Laboratories

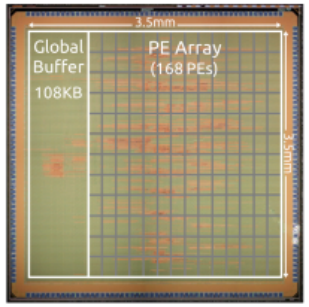
Roberto Gioiosa
Research Staff
Pacific Northwest National Laboratory

Tushar Krishna
Associate Professor
Georgia Tech

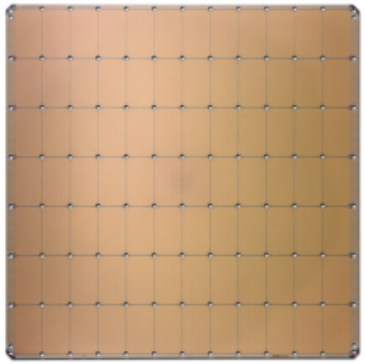
The Era of Domain-Specific Accelerators



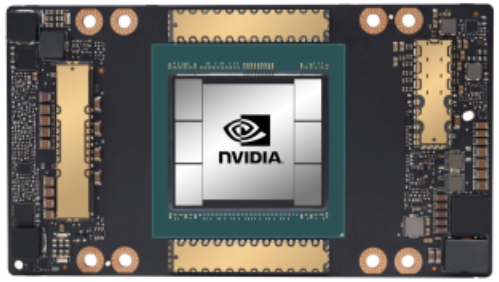
Google TPU [1]



Eyeriss [2]



Cerebras WSE-2 [3]



NVIDIA Ampere GPU [4]

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

[1] In-Datacenter Performance Analysis of a Tensor Processing Unit, Norman P. Jouppi et al., ISCA 2017
[2] Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks, Yu-Hsin Chen et al., JSSC 2017
[3] Cerebras white paper, Cerebras Systems: Achieving Industry Best AI Performance Through A Systems Approach, 2021
[4] NVIDIA A100 Tensor Core GPU Architecture white paper V1.0

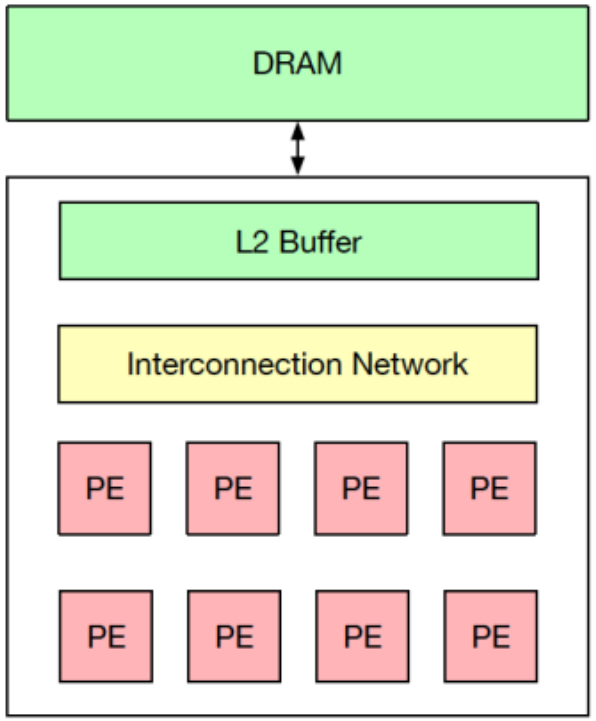
The Era of Domain-Specific Accelerators

Innovation
Novel memory hierarchy
Efficient interconnection network
Custom processing elements

Fragmentation
Custom compiler toolchain
Duplicate engineering overhead
Error-prone frameworks

- **Need abstractions to unify various accelerator flows.**

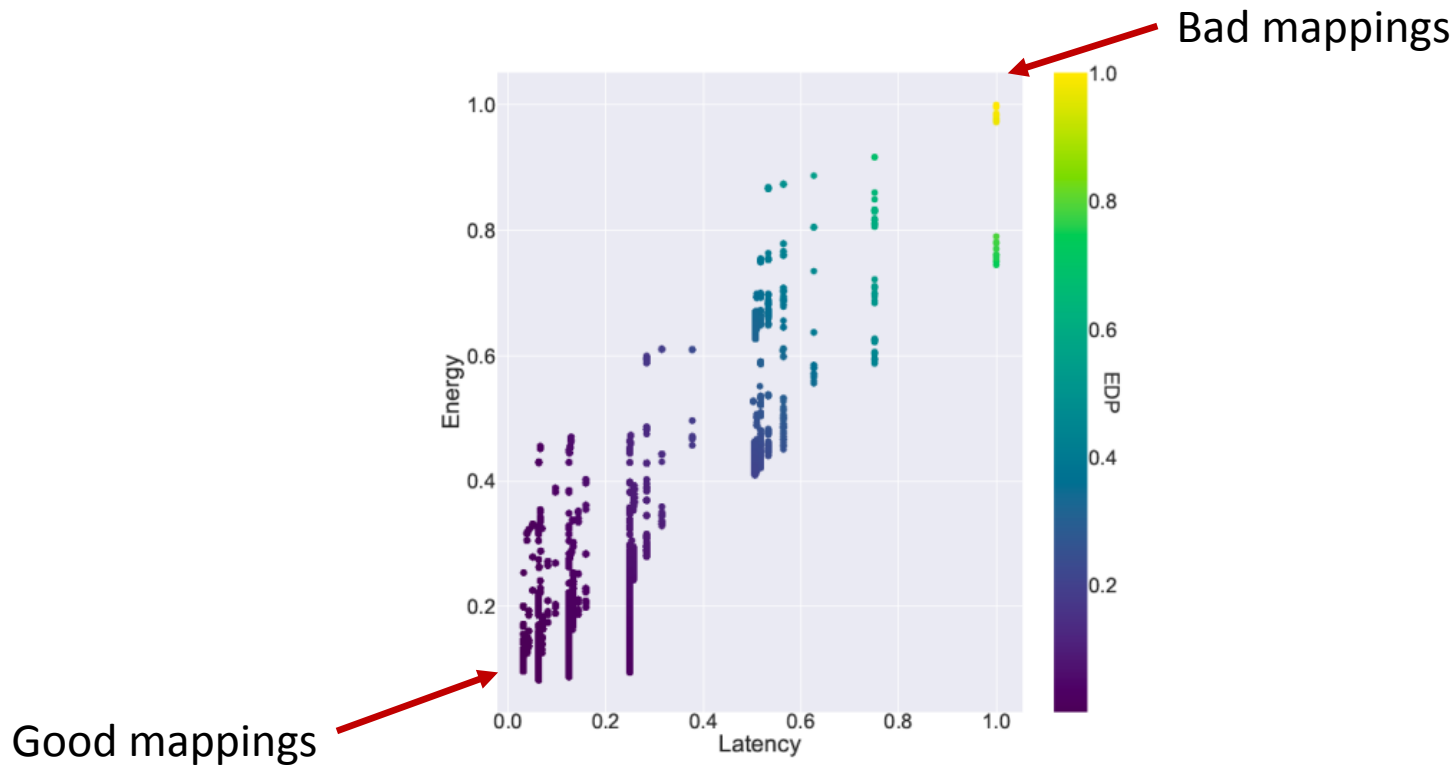
How Do Spatial Accelerators Look Like?



- **Accelerator designs**
  - Programmable scratchpads
  - A huge number of processing elements (PEs)
  - Distribution/reduction network

Motivation

- **How can we solve the given problem using a target accelerator efficiently?**
  - Mapping (tiling, ordering, parallelizing) matters!

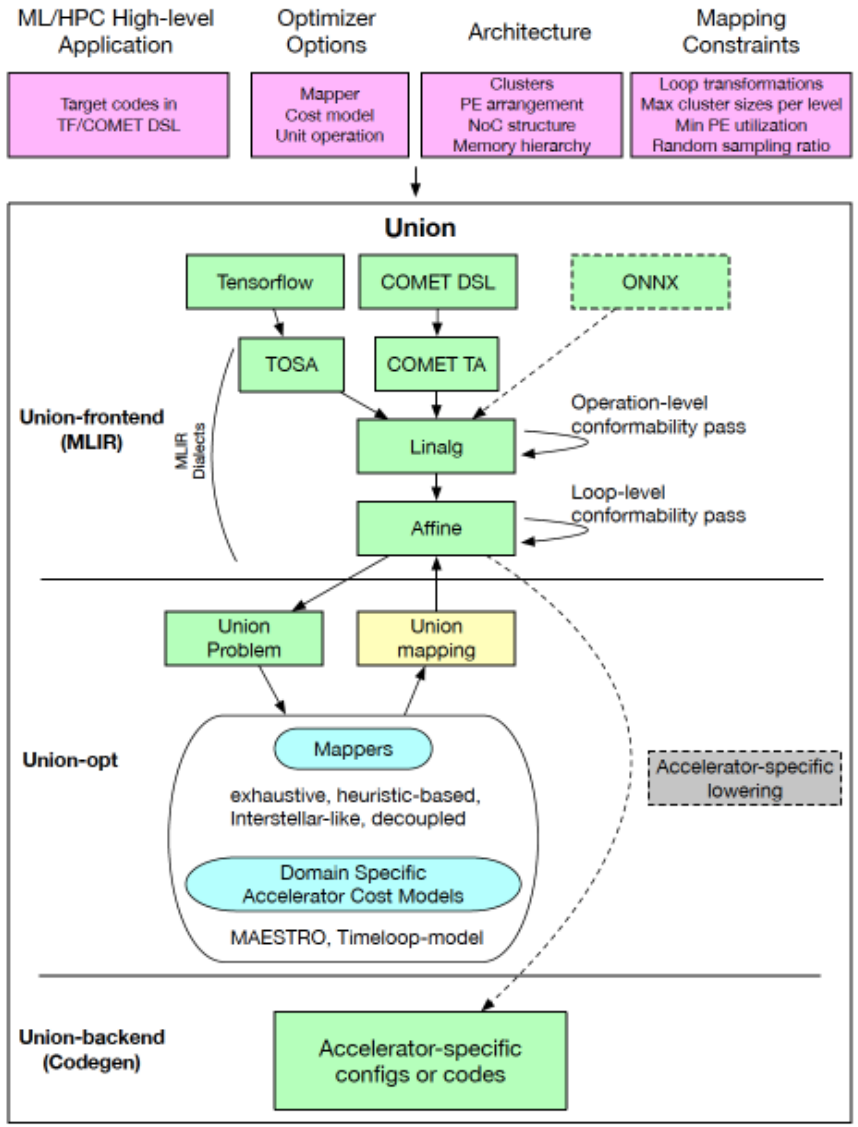


Motivation

- **How can we solve the given problem using a target accelerator efficiently?**
- **Modularity**
  - Accelerators
  - Cost models
  - Mappers
  - High level languages
  - Frameworks
- **Unified abstractions to cover various designs**

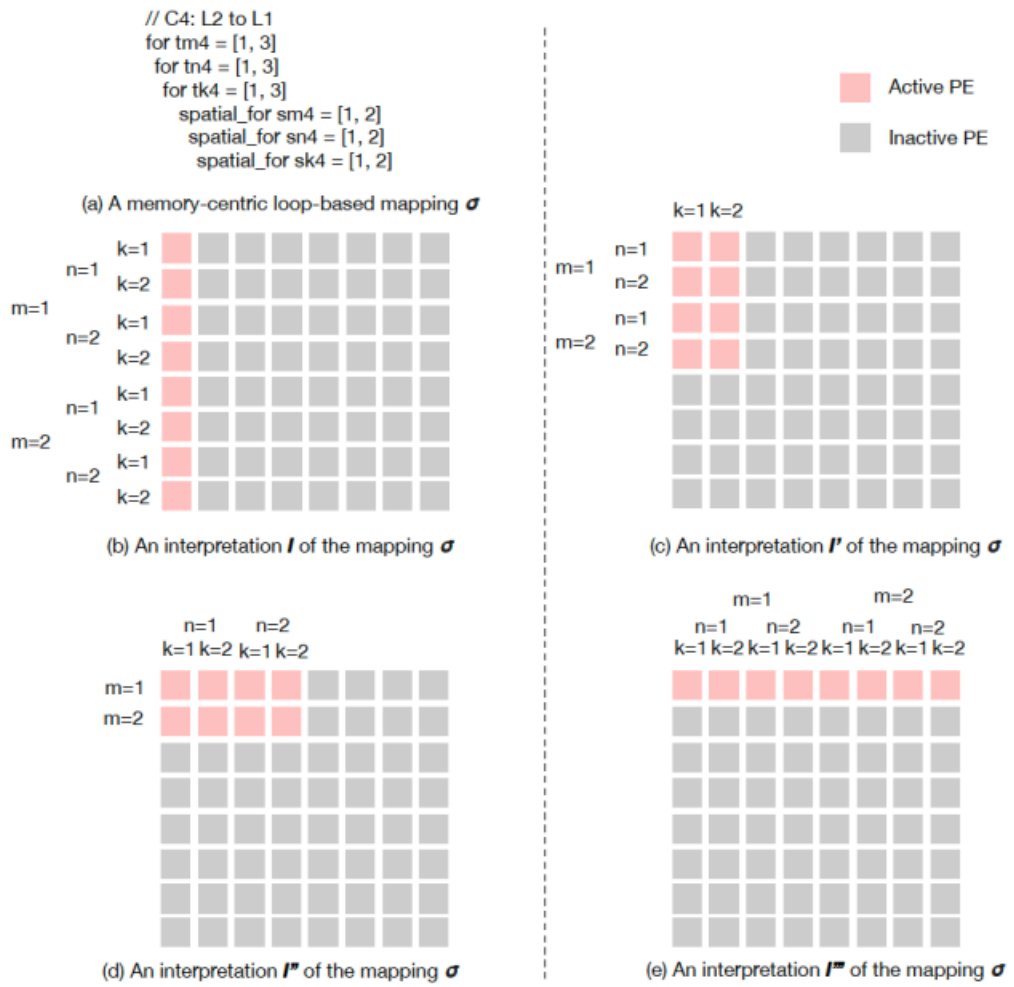
We propose ***Union***, a Unified HW-SW Co-Design Ecosystem in MLIR for Evaluating Tensor Operations on Spatial Accelerators.

Union Overview



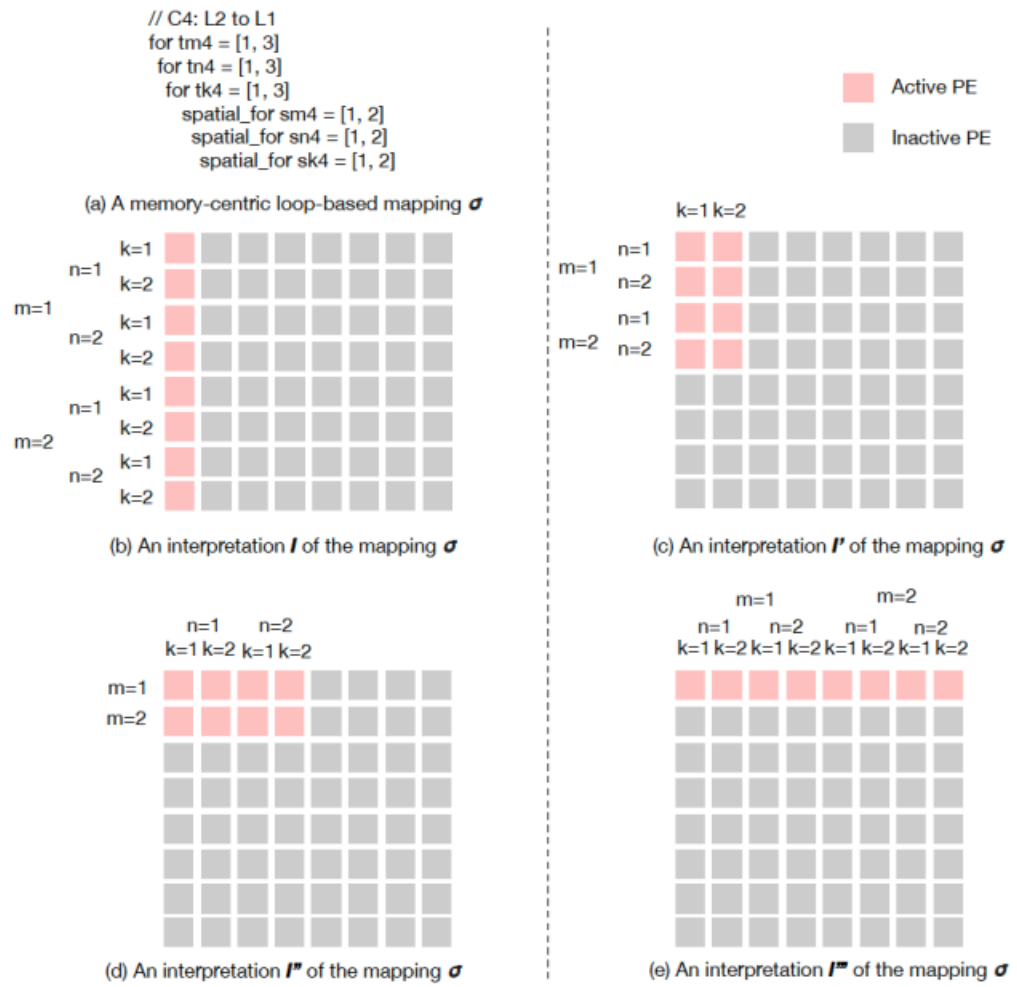
Limitations of Current Abstractions

- **Loop-centric** approach is intuitive and easy to understand.
- **Memory-target loop-centric** approach used in Timeloop/Interstellar can be enhanced by changing into cluster-target to support a broader set of mappings.



Limitations of Current Abstractions

- **Cluster-target** notion has been used in MAESTRO data-centric notation while Marvel and Timeloop are using memory-centric notion.



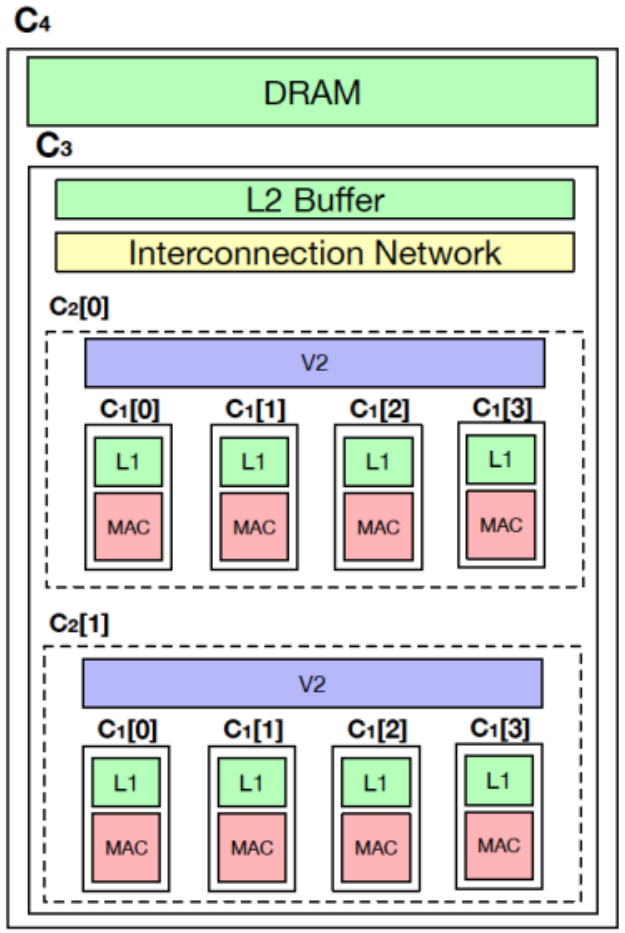
Union Abstractions

```
Problem:
  Operator: 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
```

(a) 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
```

(b) Union architecture



(c) Target accelerator architecture

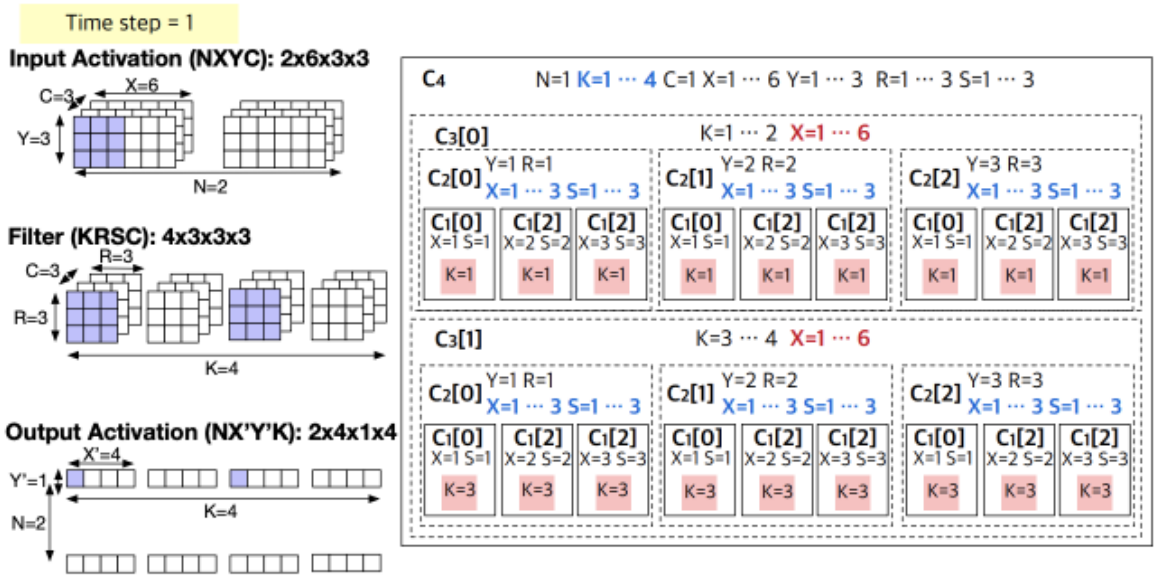
```
// C4: DRAM to L2
target_cluster: C4
temporal_order: MNK
temporal_tile_sizes: 16, 32, 16
spatial_tile_sizes: 16, 32, 16

// 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
```

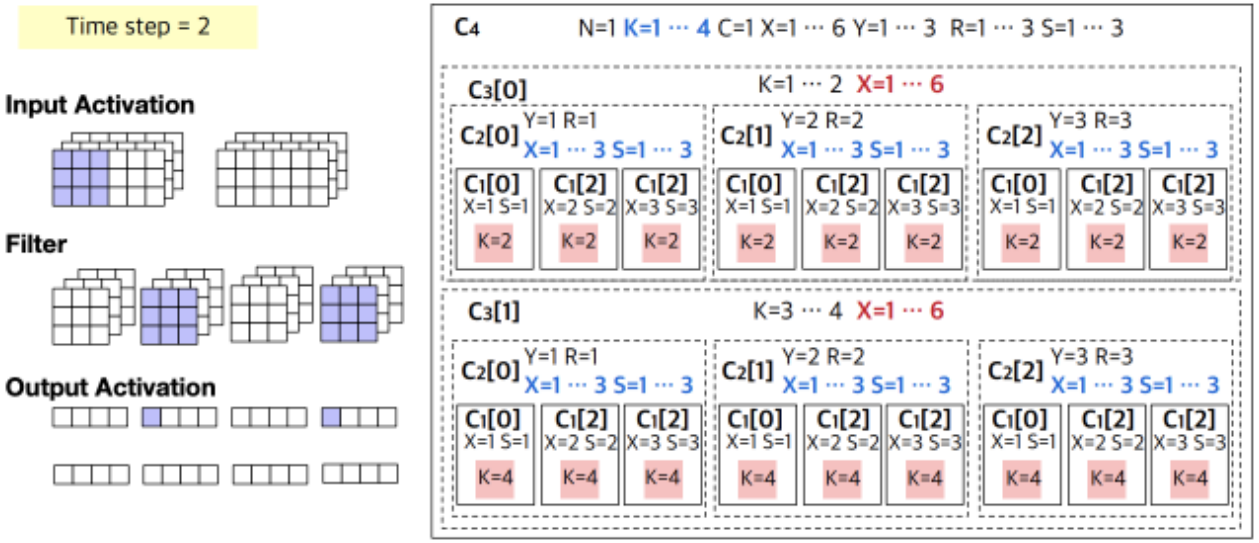
(d) Union mapping

```
// C4: DRAM to L2
for tm3 = 0
 for tn3 = 0…1
  for tk3 = 0…1
   spatial_for sm3 = 0
   spatial_for sn3 = 0
   spatial_for sk3 = 0
// C3: L2 to V2
for tm2 = 0…1
 for tn2 = 0…1
  for tk2 = 0…1
   spatial_for sm2 = 0
   spatial_for sn2 = 0…1
   spatial_for sk2 = 0
// C2: V2 to L1
for tm1 = 0
 for tn1 = 0
  for tk1 = 0
   spatial_for sm1 = 0
   spatial_for sn1 = 0
   spatial_for sk1 = 0 … 3
// C1: L1 to MAC
for tm0 = 0…7
 for tn0 = 0…7
  for tk0 = 0…1
   spatial_for sm0 = 0
   spatial_for sn0 = 0
   spatial_for sk0 = 0
```

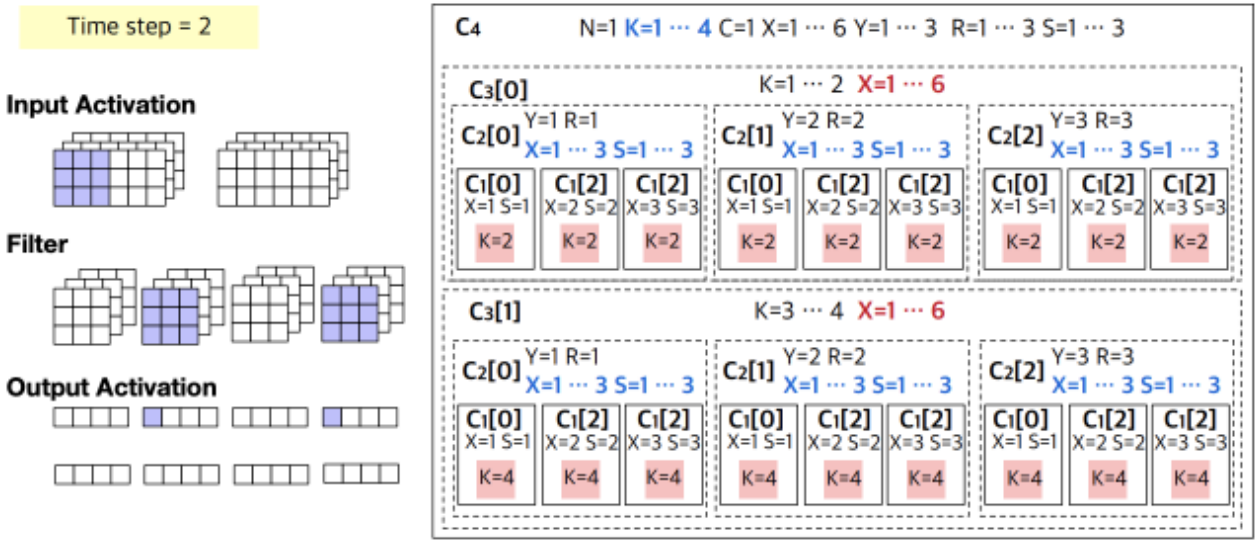
(e) Loop nest representation

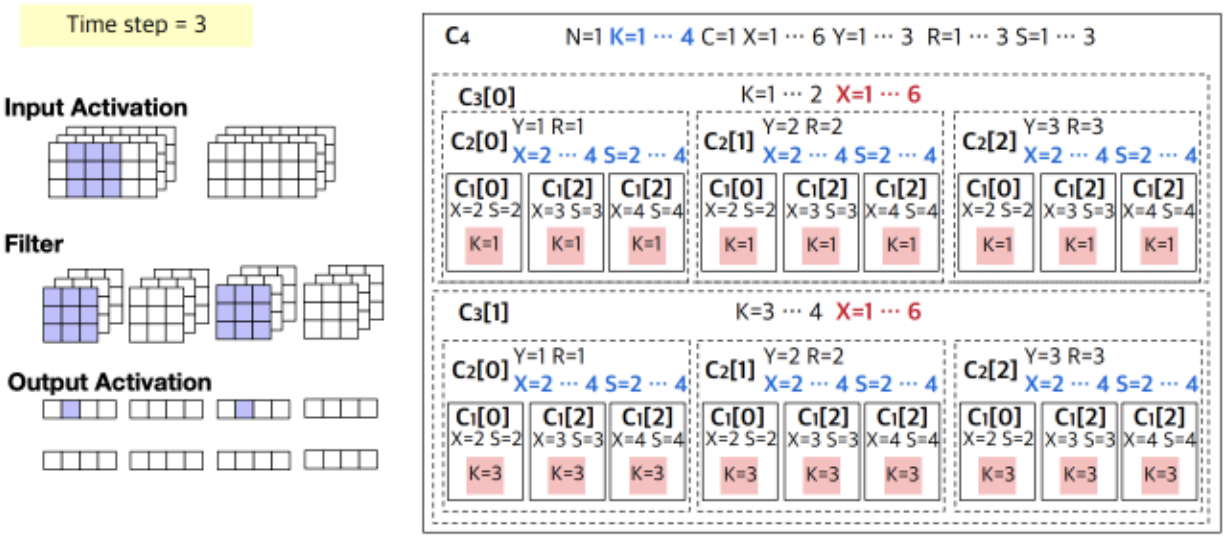
Mapping Example



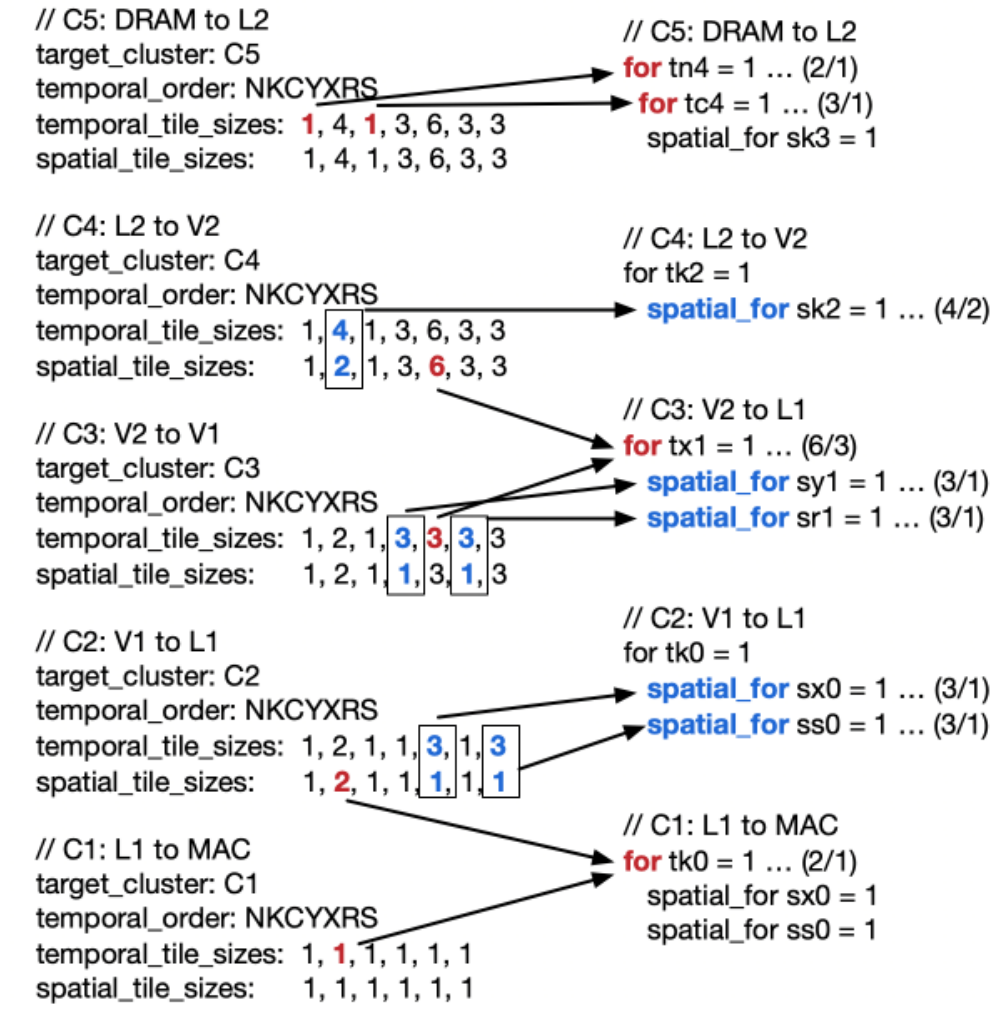


Mapping Example





Mapping Example



(a) Union mapping for a K_YR_XS mapping (b) Loop nest representation

Case Studies Using Union

TABLE III
TENSOR CONTRACTION PROBLEMS AND THE CORRESPONDING GEMM DIMENSION SIZES FOR TTGT

Name	Equation	Tensor Dimension Sizes	GEMM Dimension Sizes
intensli2	C[a, b, c, d] = A[d, b, e, a] * B[e, c]	a = b = c = d = e = 64 a = b = c = d = e = 16	M = 262144, N = 64, K = 64 M = 4096, N = 16, K = 16
ccsd7	C[a, b, c] = A[a, d, e, c] * B[e, b, d]	a = b = c = d = e = 64 a = b = c = d = e = 16	M = 4096, N = 64, K = 4096 M = 256, N = 16, K = 256
ccsd-t4	C[a, b, c, d, e, f] = A[d, f, g, b] * B[g, e, a, c]	a = b = c = d = e = f = g = 32 a = b = c = d = e = f = g = 16	M = 32768, N = 32768, K = 32 M = 4096, N = 4096, K = 16

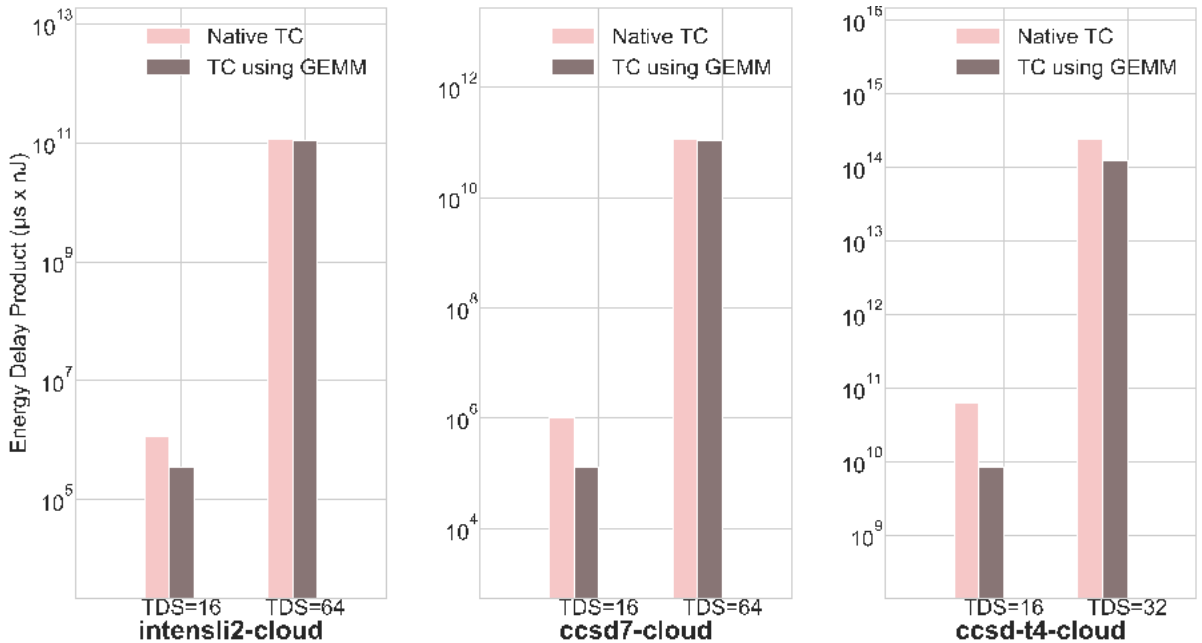
TABLE IV
DNN LAYER DIMENSIONS USED IN EVALUATION

Layer	Dimensions
ResNet50-1	N=32 K=C=64 X=Y=56 R=S=1
ResNet50-2	N=32 K=C=64 X=Y=56 R=S=3
ResNet50-3	N=32 K=512 C=1024 X=Y=14 R=S=1
DLRM-1	N=512 NIN=1024 NON=1024
DLRM-2	N=512 NIN=1024 NON=64
DLRM-3	N=512 NIN=2048 NON=2048
BERT-1	N=256 NIN=768 NON=768
BERT-2	N=256 NIN=3072 NON=768
BERT-3	N=256 NIN=768 NON=3072

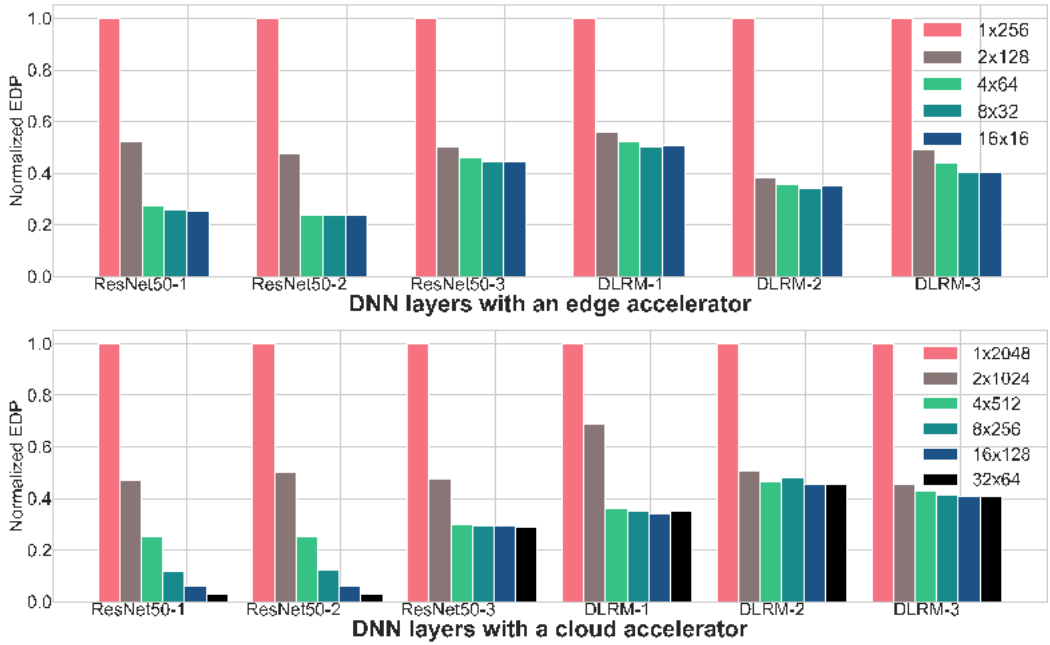
TABLE V
ACCELERATOR CONFIGURATIONS

Type	# of PEs	L1 Buffer Size	L2 Buffer Size	NoC Bandwidth
Edge	256	0.5 KB	100 KB	32 GB/s
Cloud	2048	0.5 KB	800 KB	256 GB/s

Case Studies Using Union

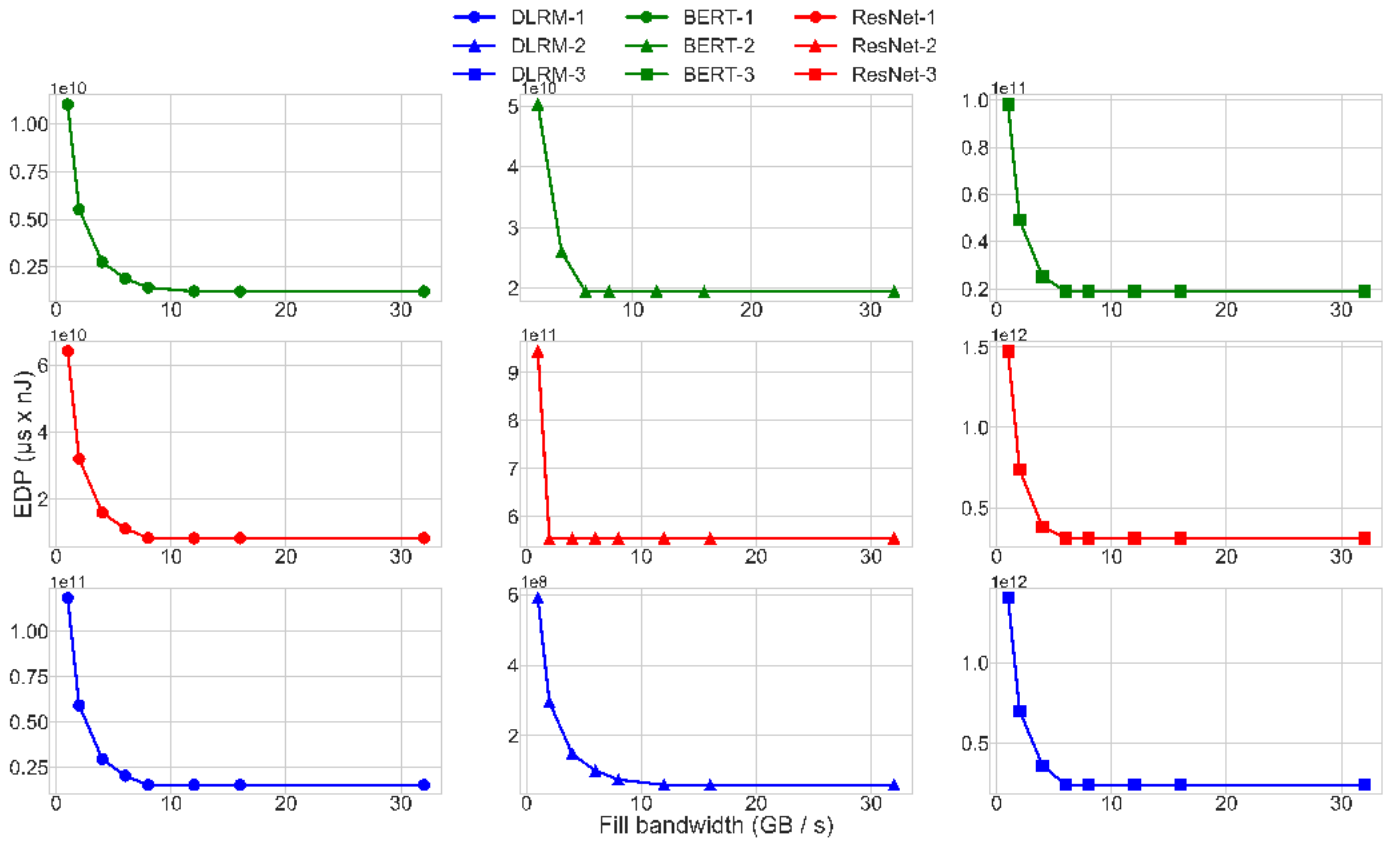


An experiment with different algorithms for the same problem



An experiment with different aspect ratio

Case Studies Using Union



An experiment with different fill bandwidth for multi-chiplet accelerators

Conclusion

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

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

Thank you for listening!
Code available at https://github.com/union-codesign/union

Acknowledgement

- Support for this work was provided through U.S. Department of Energy's (DOE) Office of Advanced Scientific Computing Research as part of the Center for Artificial Intelligence-focused Architectures and Algorithms.
- Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

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

Thank you for listening!

Code available at https://github.com/union-codesign/union