

Hierarchical Computation and Communication for Distributed Sparse Operations on Supercomputers with Multi-GPU Nodes

Mert Hidayetoğlu

**Coordinated Science Laboratory
University of Illinois at Urbana-Champaign**

February 26, 2021

Preliminary Exam Committee:

Wen-mei Hwu (Chair)
Weng Cho Chew
Michale Oelze
Andreas Kloeckner
Bill Gropp

Large-Scale Sparse Applications

Conference

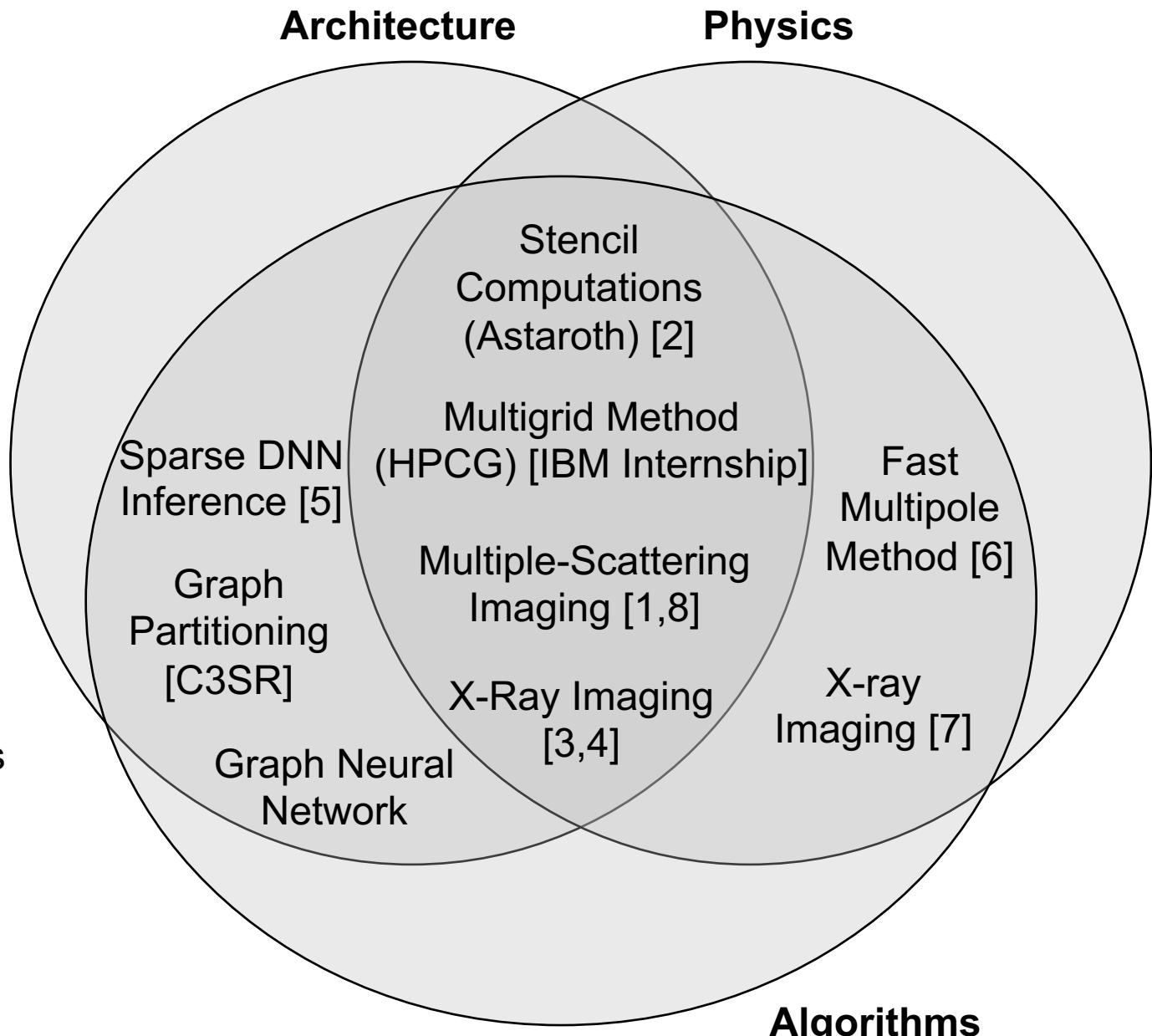
- [1] IPDPS'18 (Outstanding Presentation)
- [2] HPCC'19 (Best Paper)
- [3] SC'19 (SCC Benchmark)
- [4] SC'20 (Best Paper)
- [5] HPEC'20 (Graph Challenge Champion)

Journal

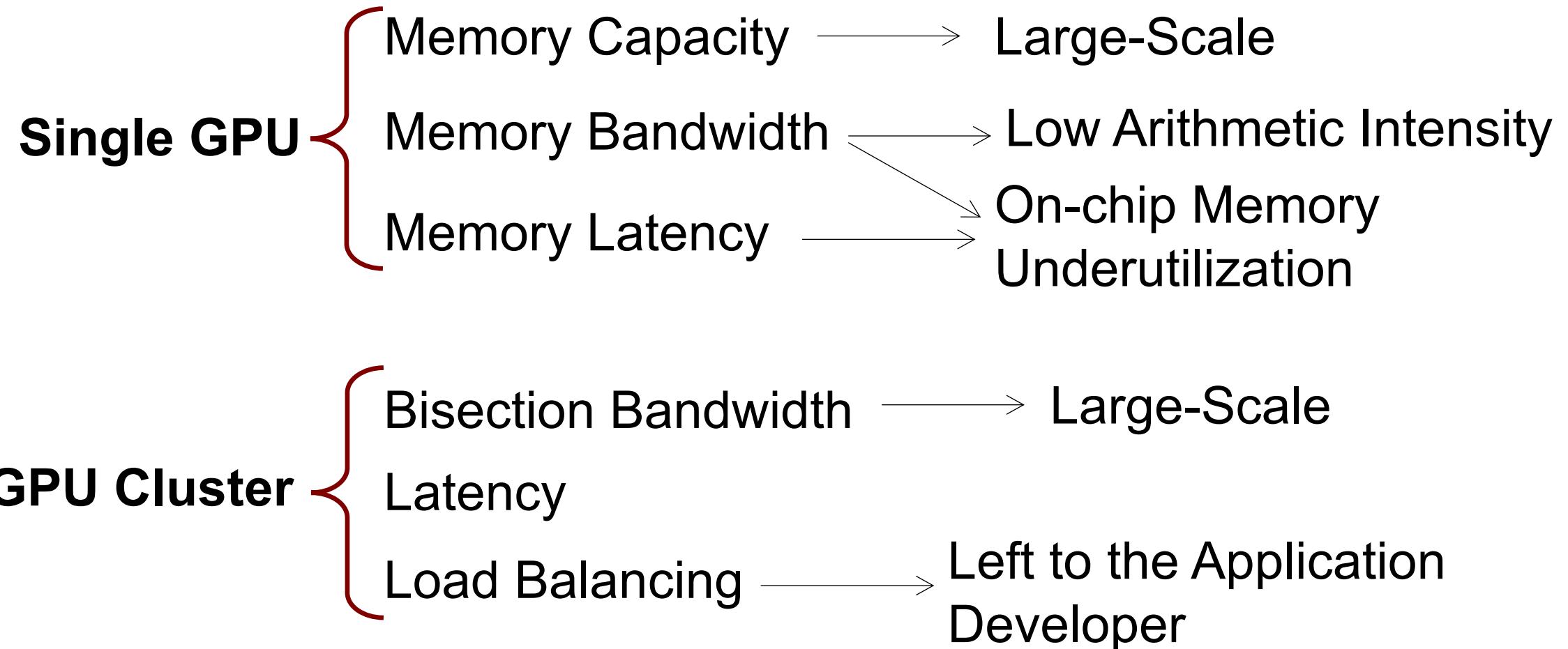
- [6] IEEE TAP (2018)
- [7] OSA Applied Optics (2018)

Book Chapter

- [8] New Trends in Computational Electromagnetics



Performance Bottlenecks



Performance Bottlenecks

Single GPU

GPU Cluster

Summit (2018) 6 GPUs per Node



200 PFLOPS 4,608 Nodes

Aurora (2021) 6 GPUs per Node



1 EFLOPS

Frontier (2021) 4 GPUs per Node



1.5 EFLOPS

El Capitan (2024) 4 GPUs per Node



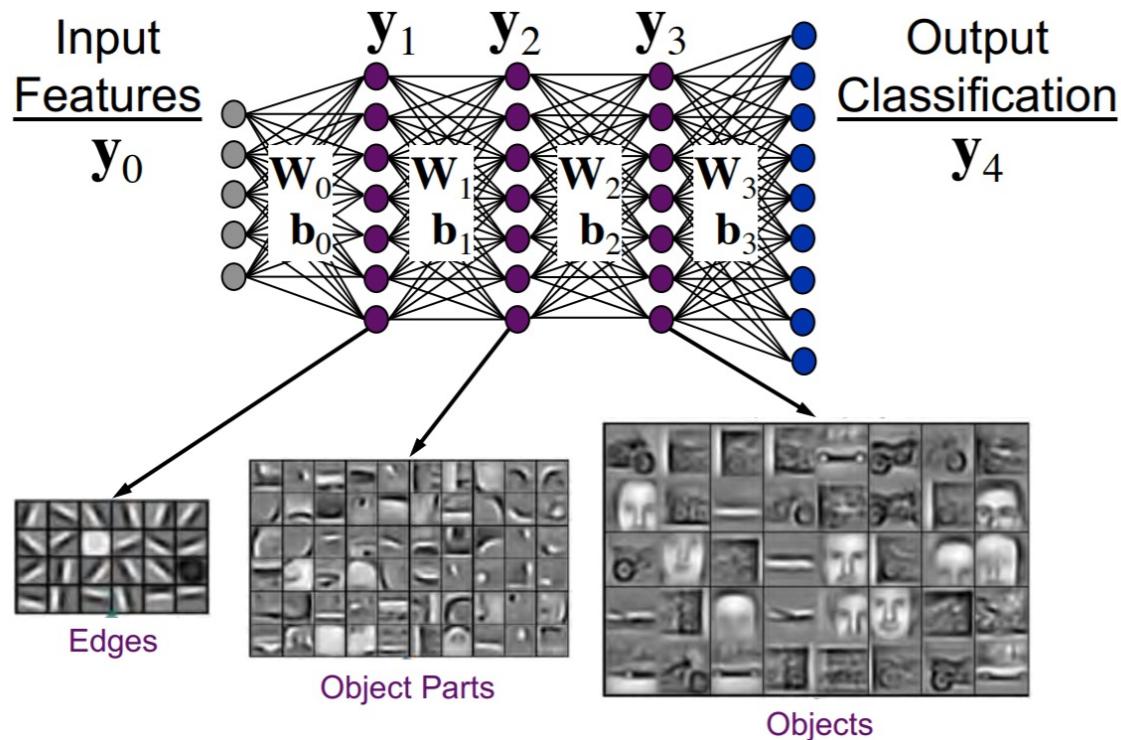
2 EFLOPS

Performance Bottlenecks
e
etic Intensity
mory
ation
ale
lication

Application: Sparse Deep Neural Network



GraphChallenge



Input Data (Y_0)

MNIST corpus of 60,000 handwritten numbers.

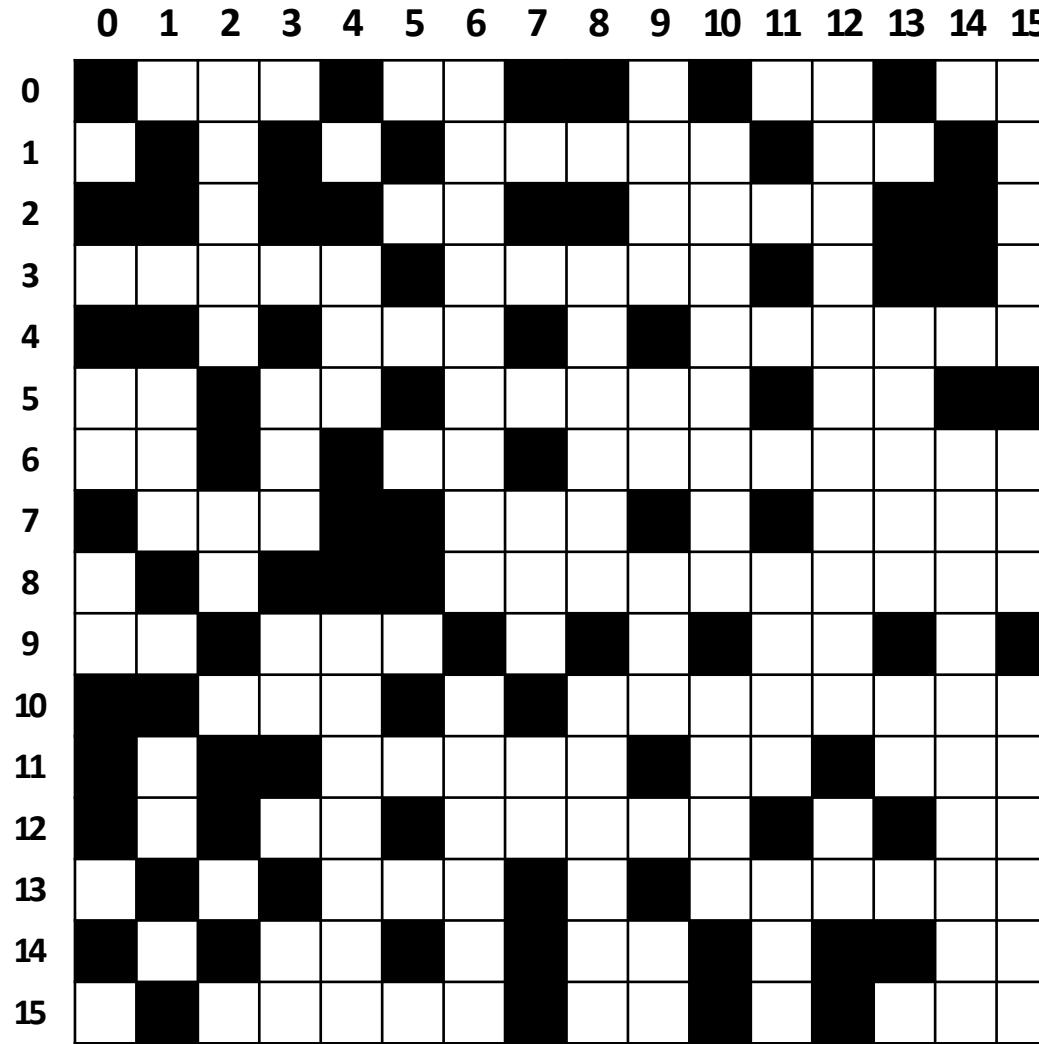
- $32 \times 32 \rightarrow 1,024$ neurons
- $64 \times 64 \rightarrow 4,096$ neurons
- $128 \times 128 \rightarrow 16,384$ neurons
- $256 \times 256 \rightarrow 65,536$ neurons

Neurons/Layer

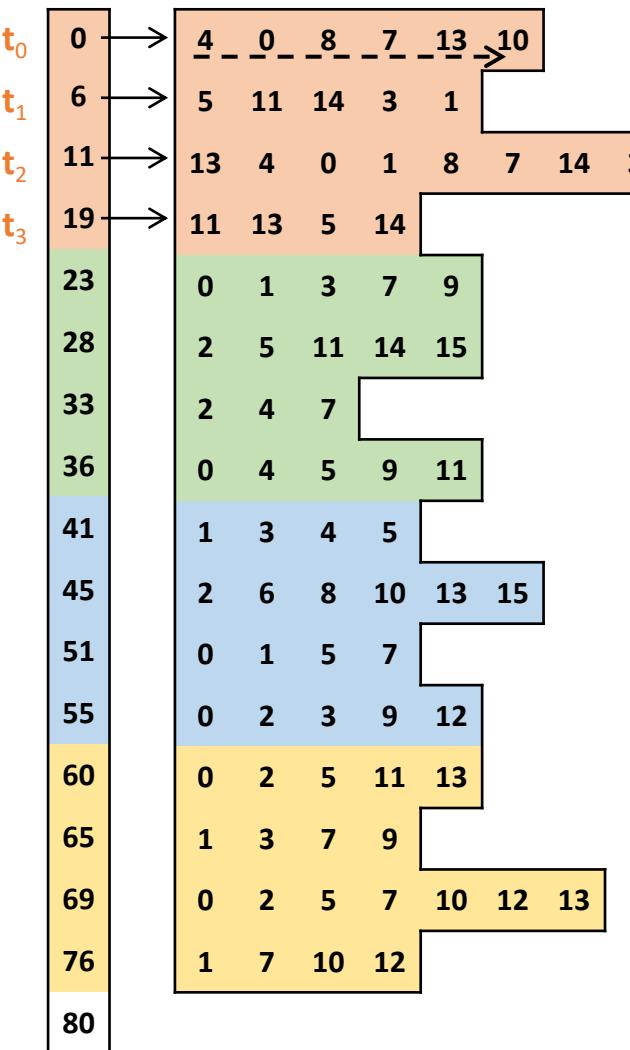
Layers	1024	4096	16384	65536
120	3,932,160	15,728,640	62,914,560	251,658,240
480	15,728,640	62,914,560	251,658,240	1,006,632,960
1920	62,914,560	251,658,240	1,006,632,960	4,026,531,840

Total Edges = $32 \times \text{layers} \times \text{neurons/layer}$

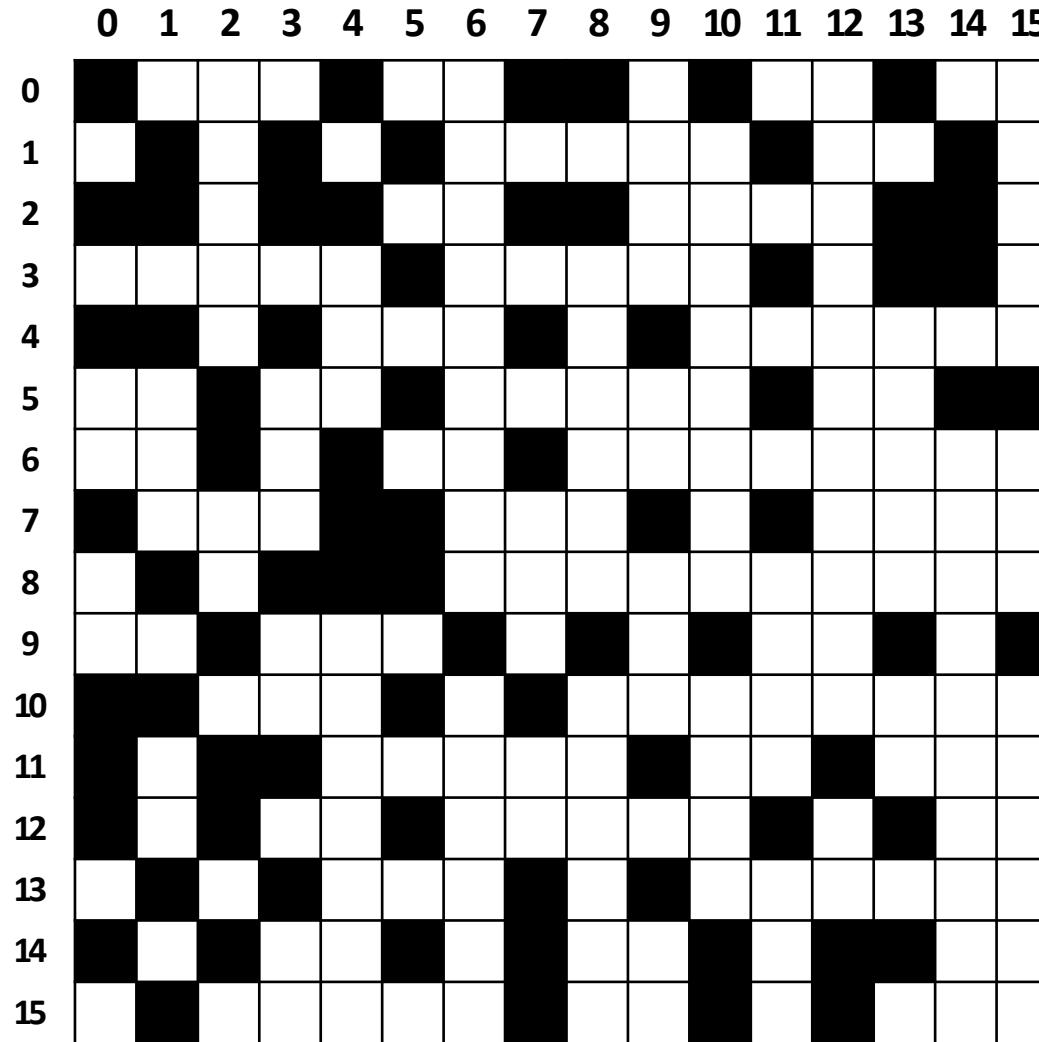
Baseline CSR SpMM



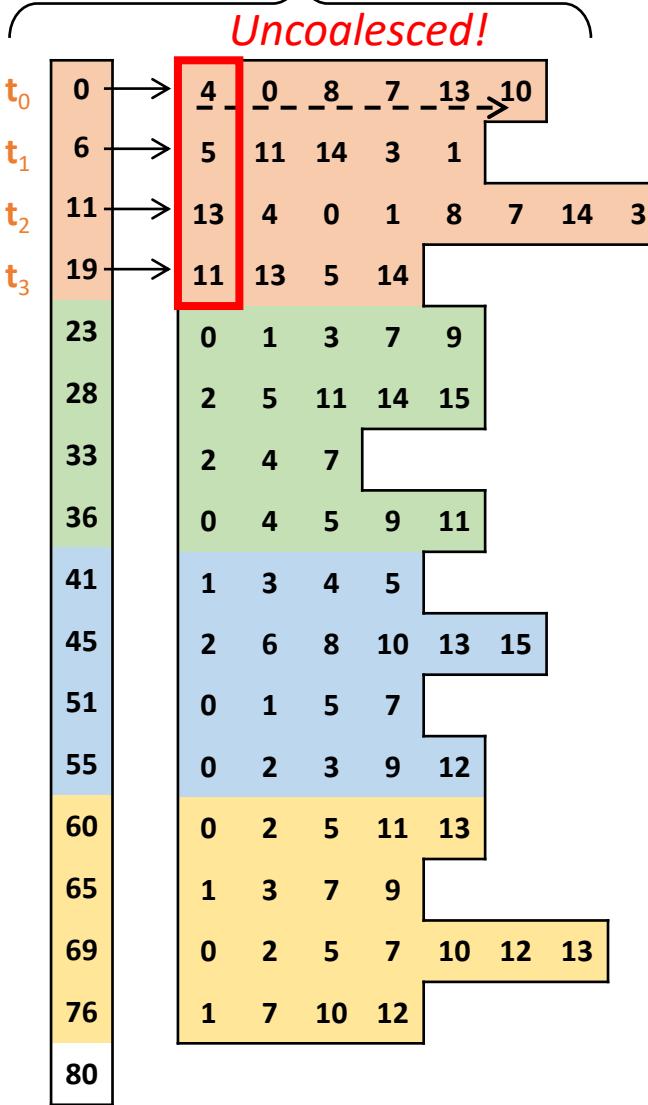
Sparse Weight Matrix



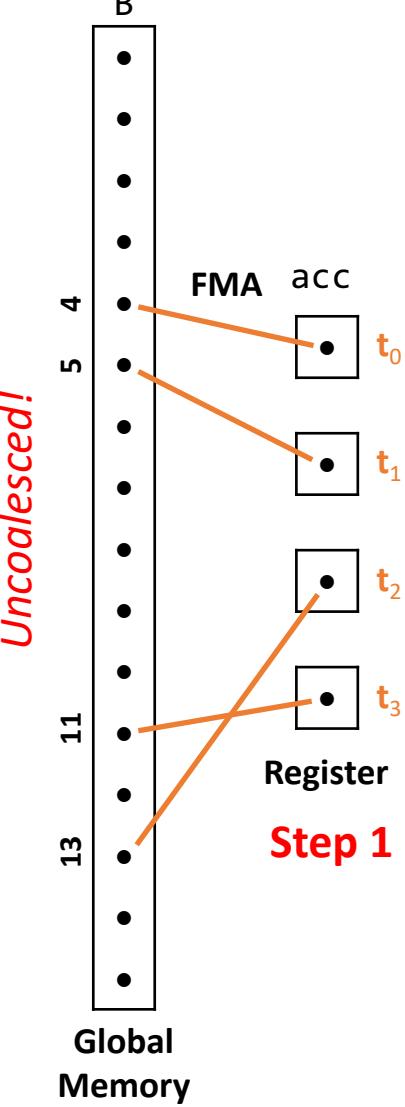
Baseline CSR SpMM



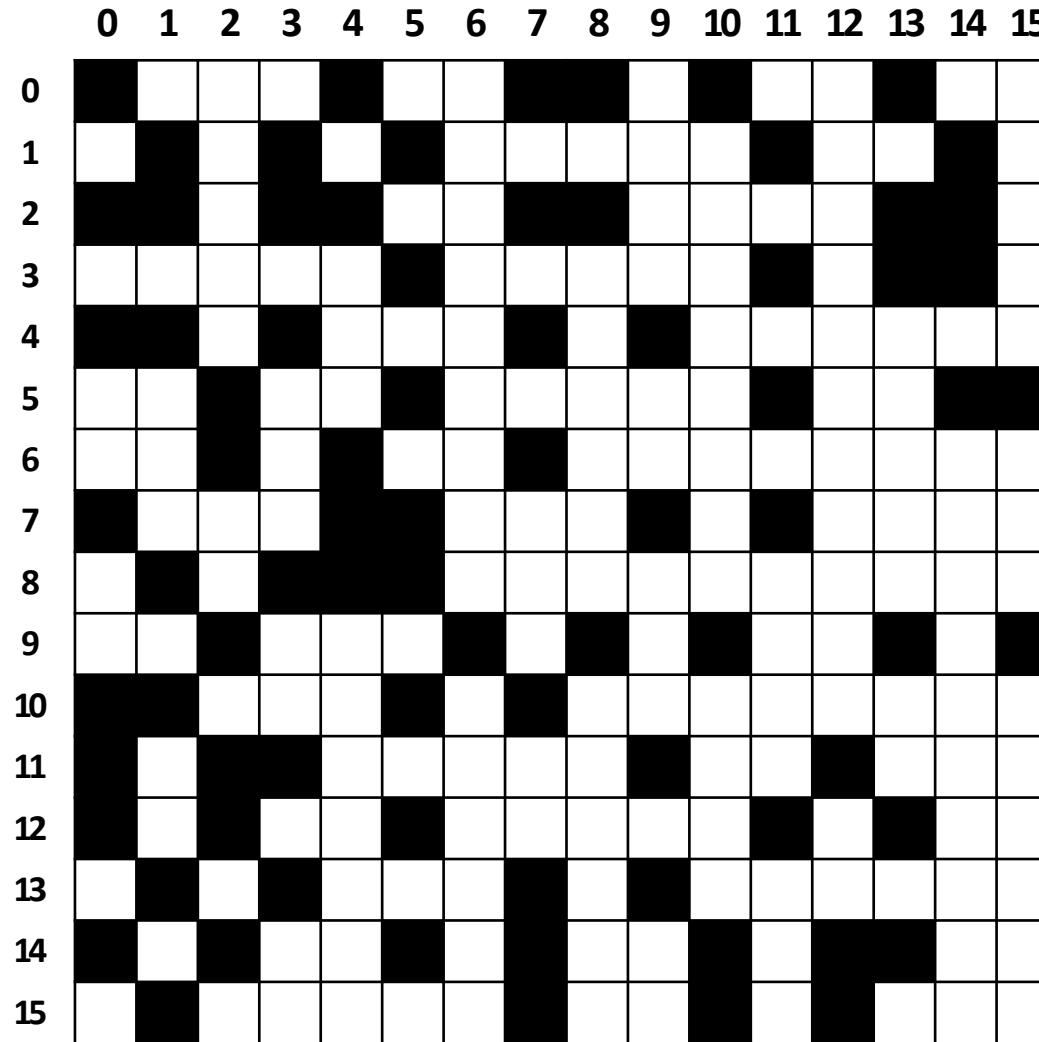
Sparse Weight Matrix



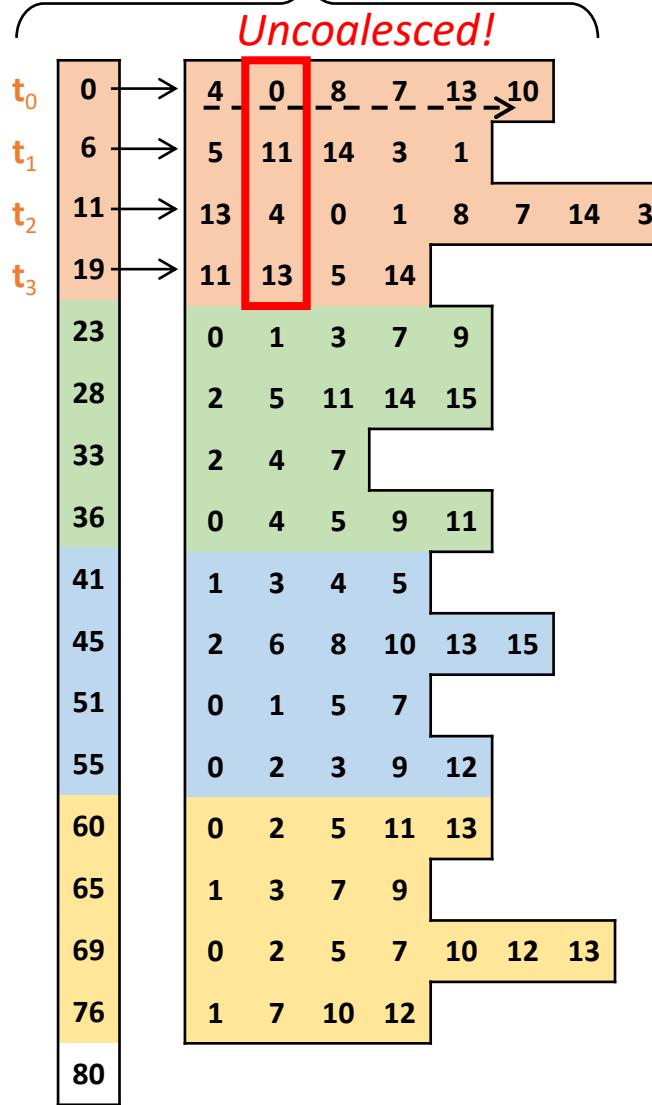
Input Features



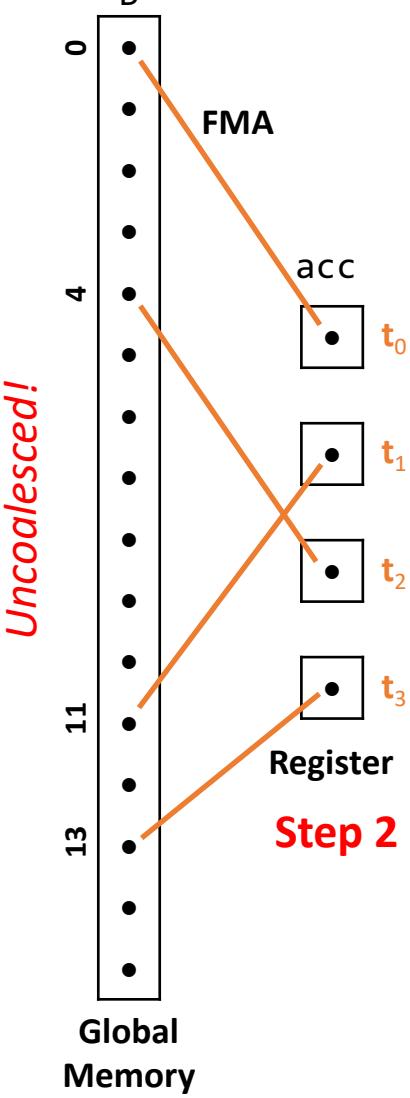
Baseline CSR SpMM



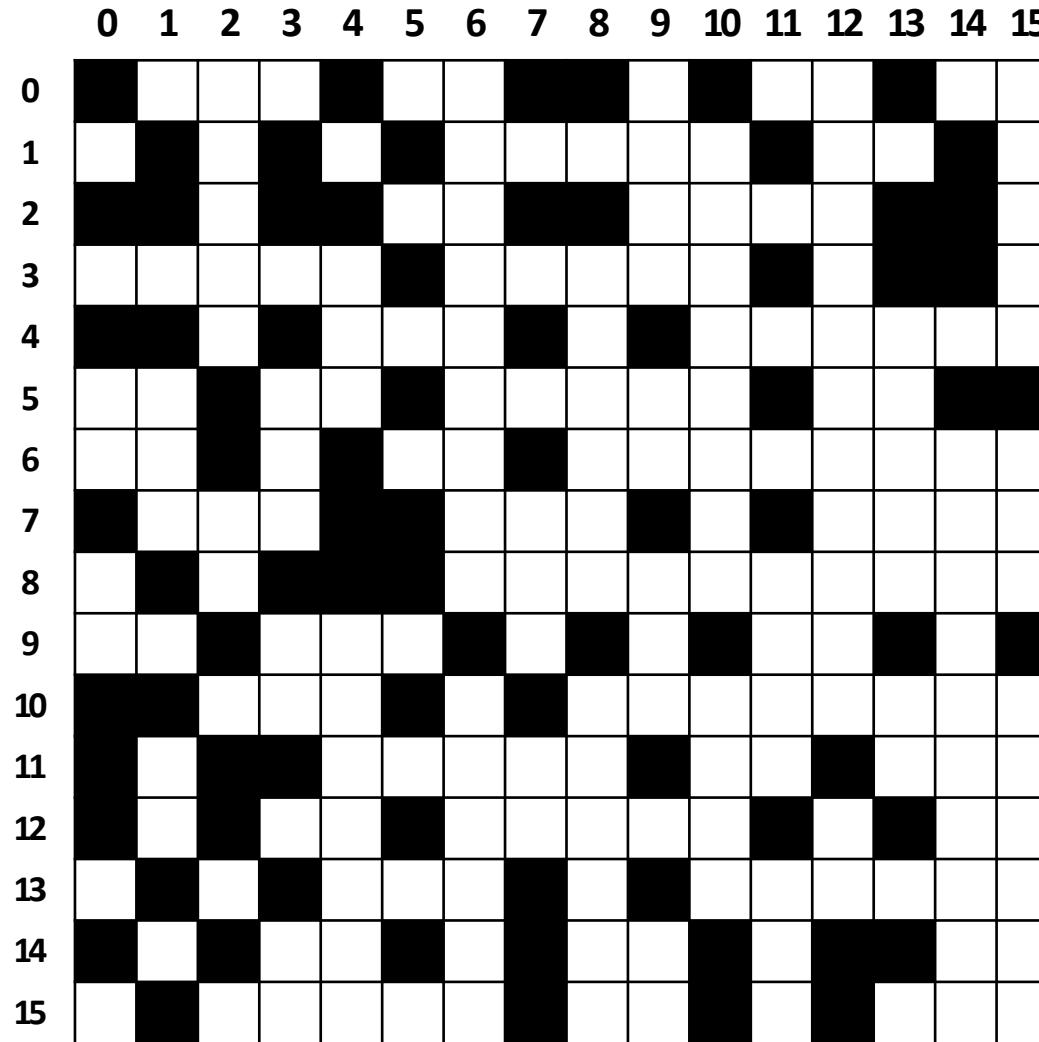
Sparse Weight Matrix



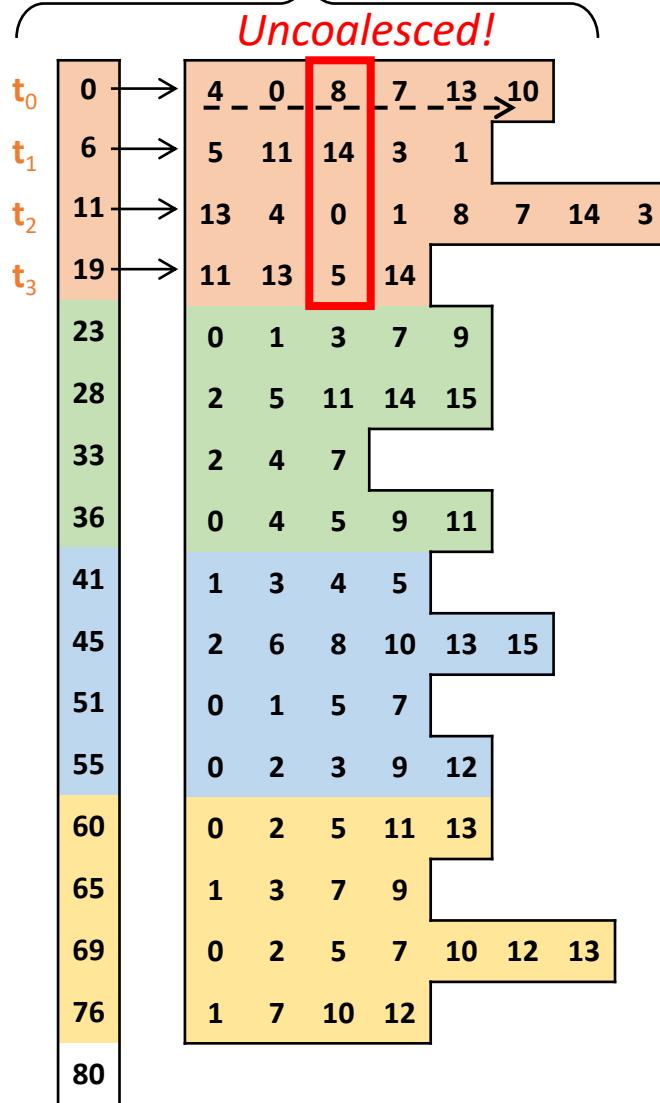
Input Features



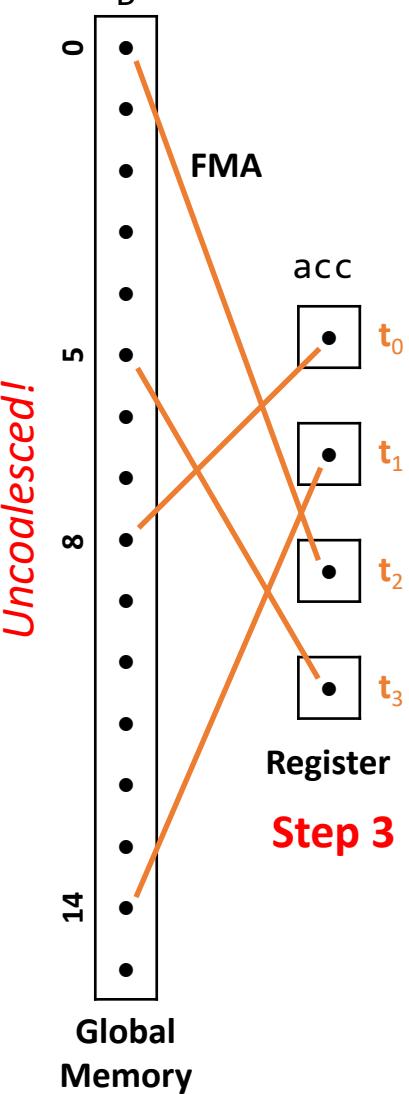
Baseline CSR SpMM



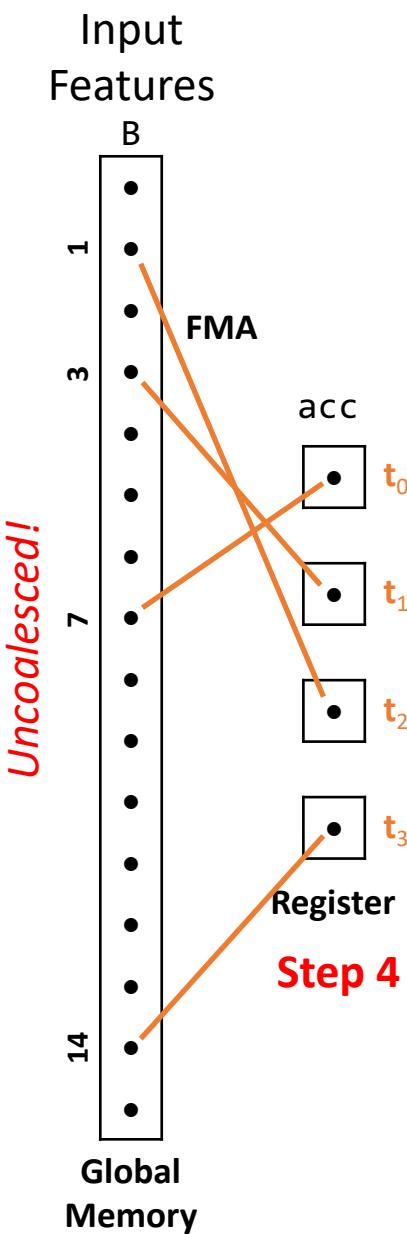
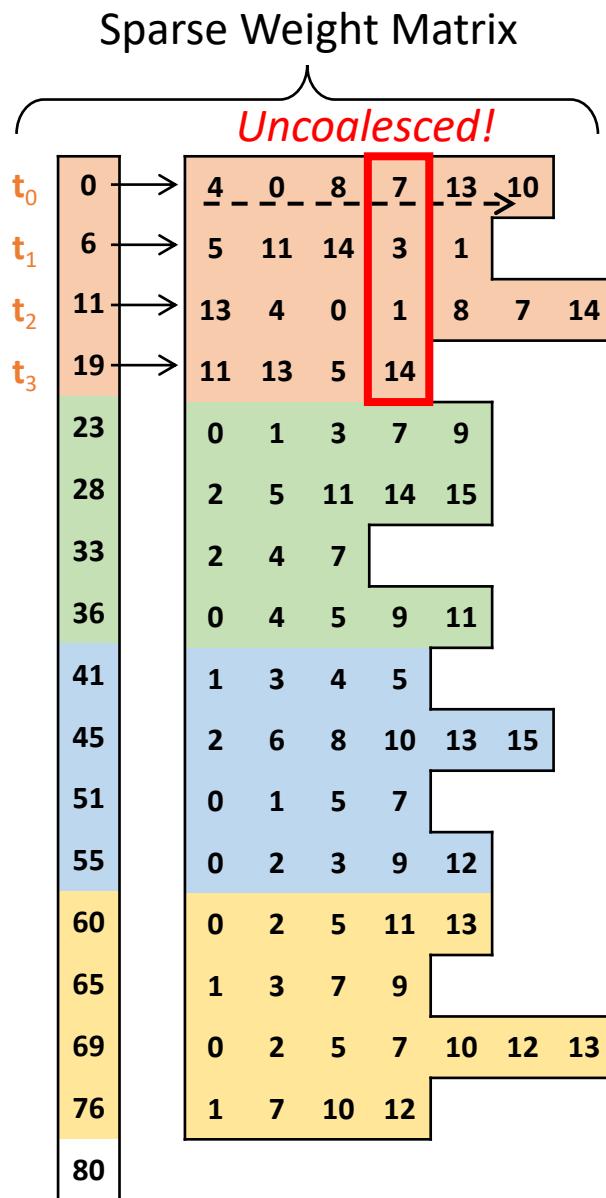
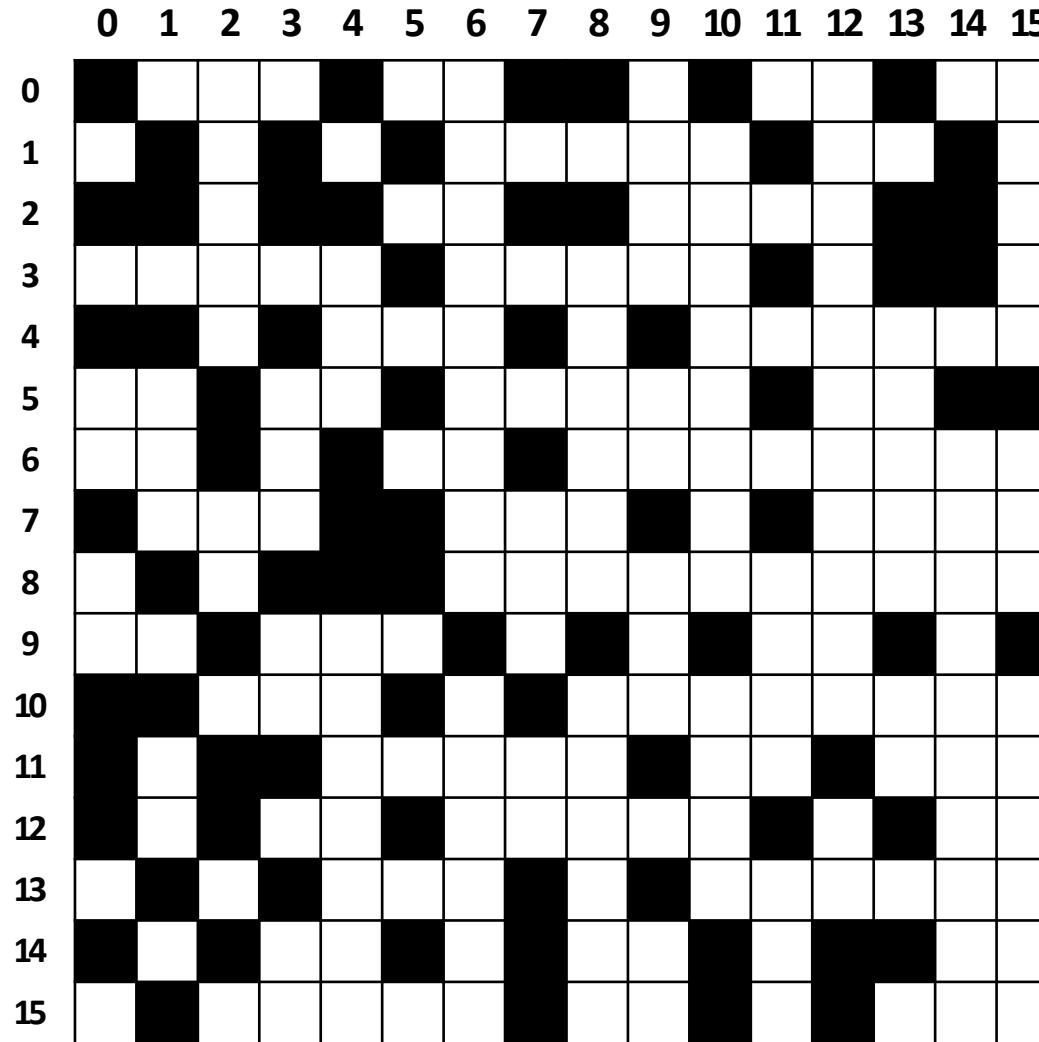
Sparse Weight Matrix



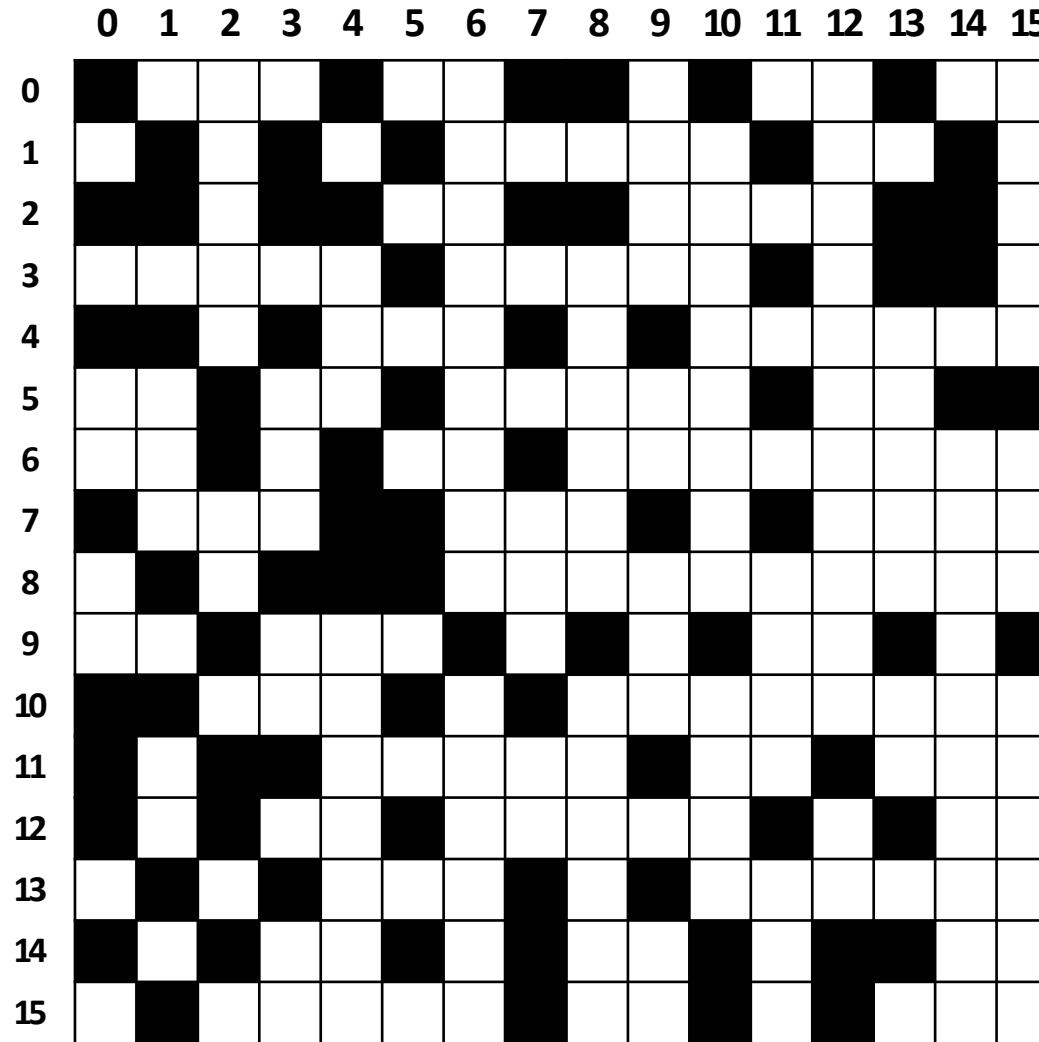
Input Features



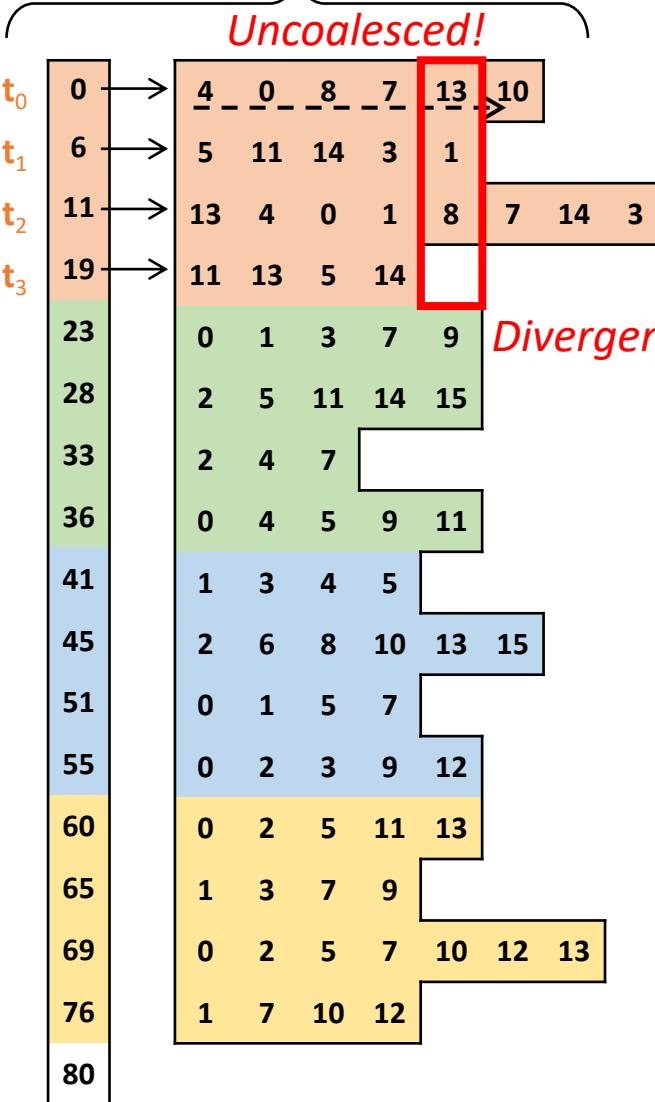
Baseline CSR SpMM



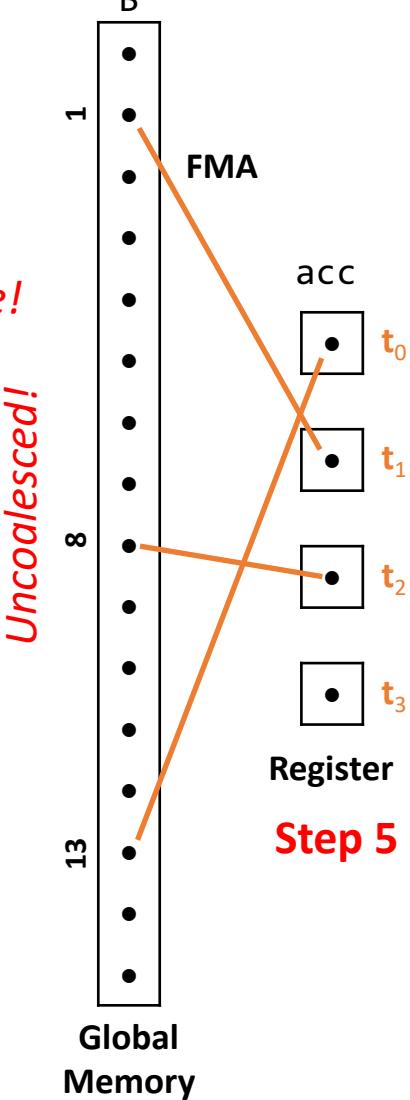
Baseline CSR SpMM



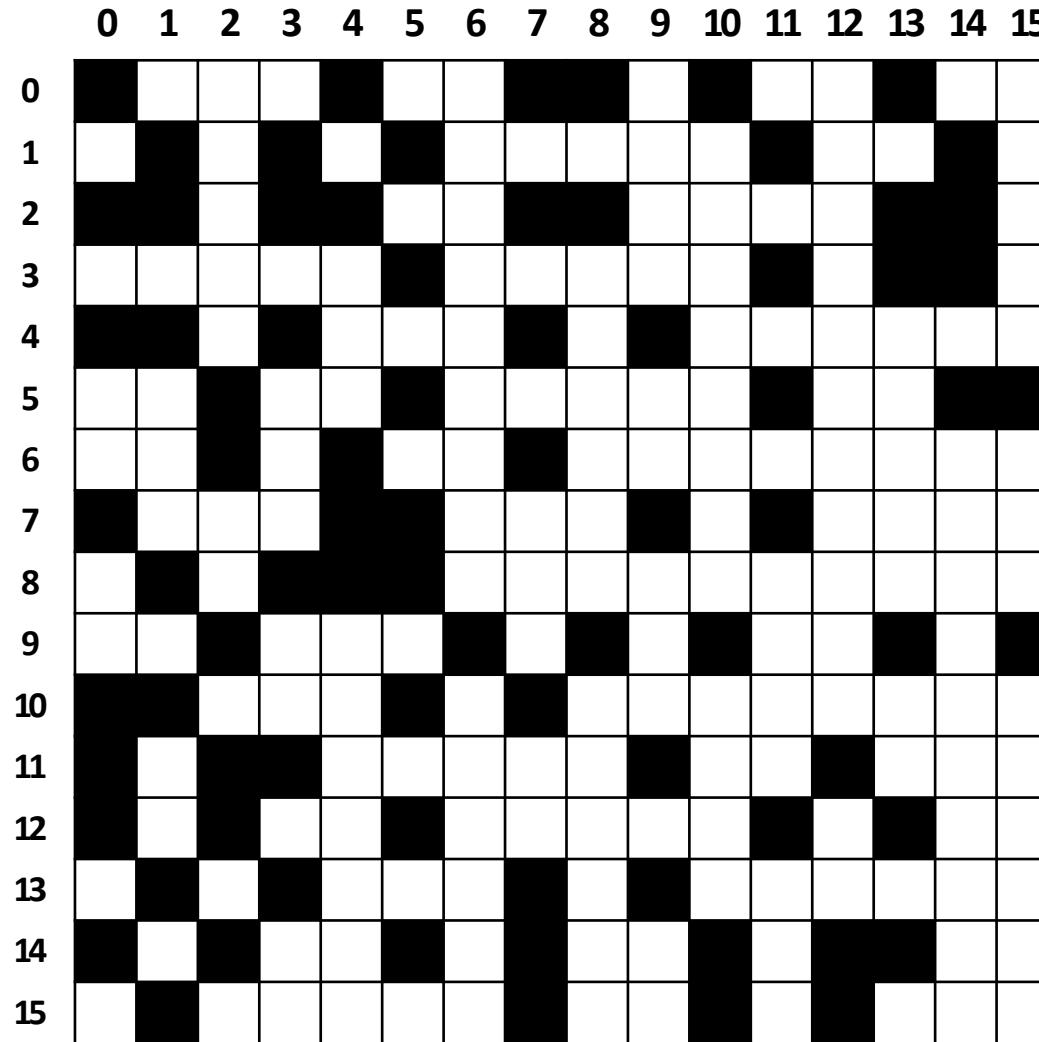
Sparse Weight Matrix



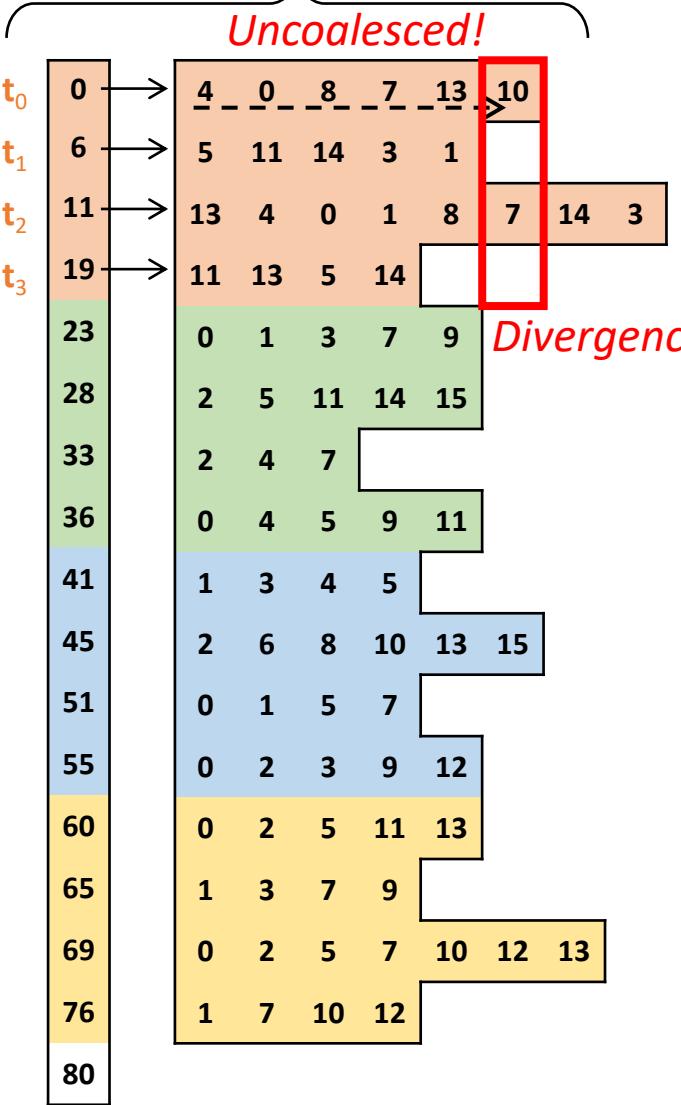
Input Features



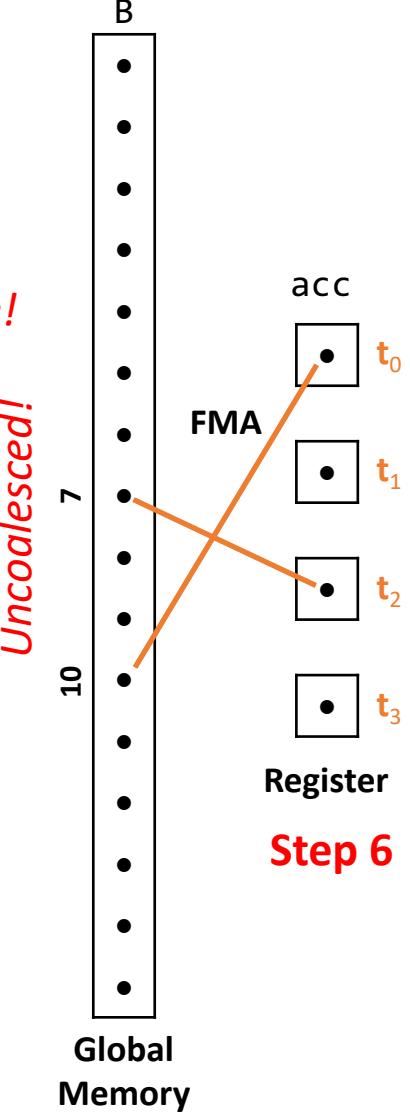
Baseline CSR SpMM



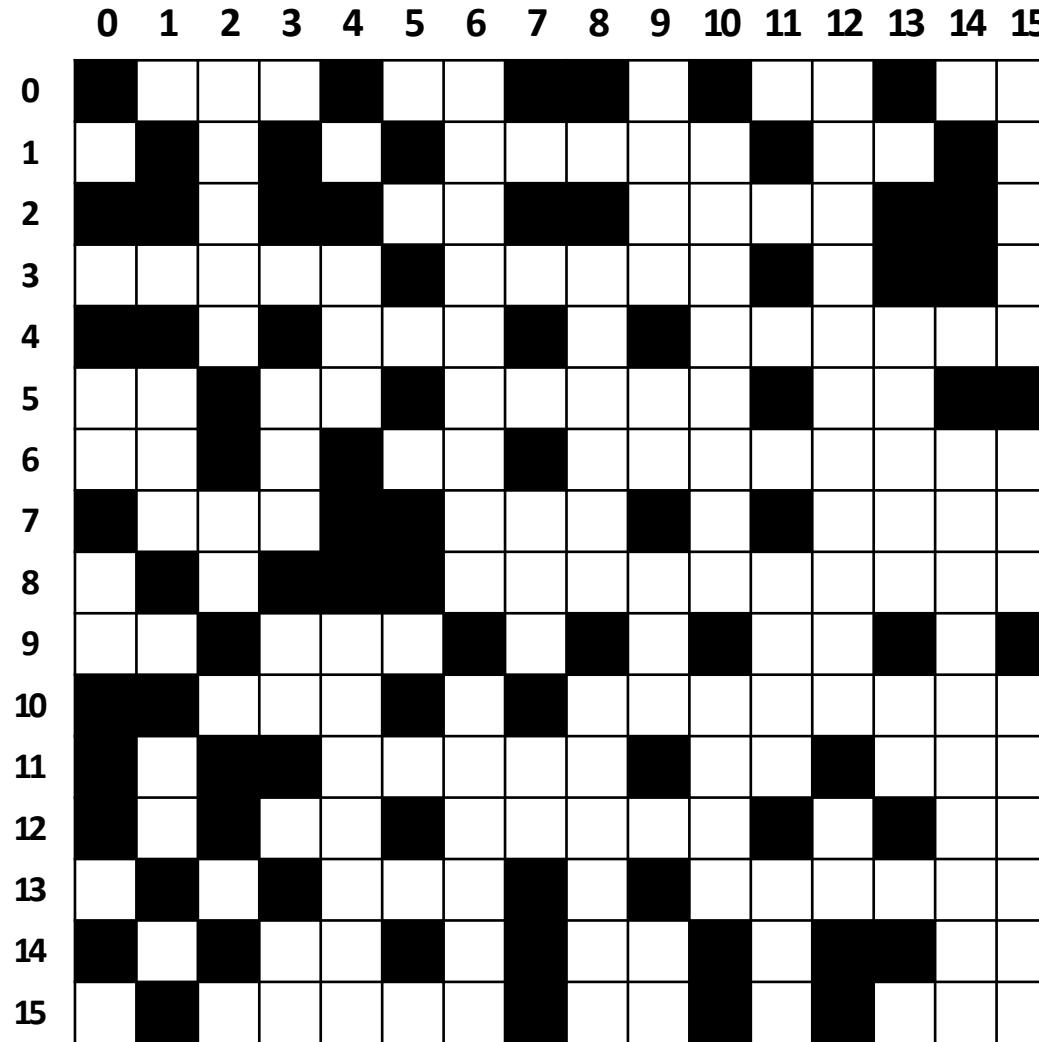
Sparse Weight Matrix



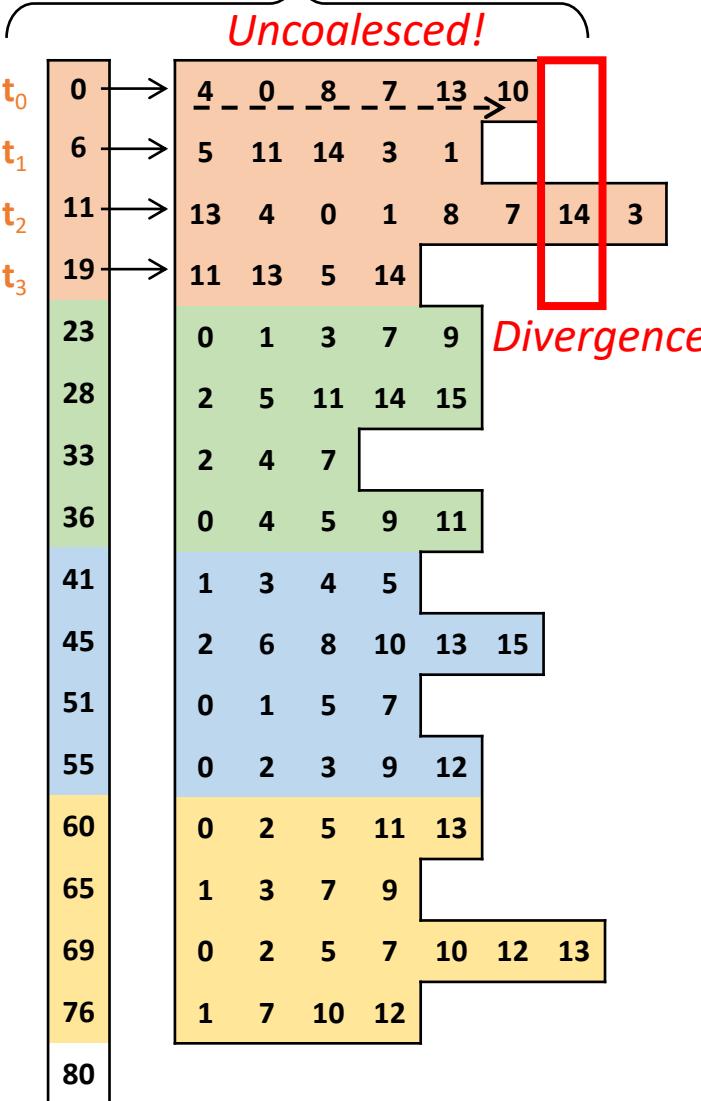
Input Features



Baseline CSR SpMM



Sparse Weight Matrix

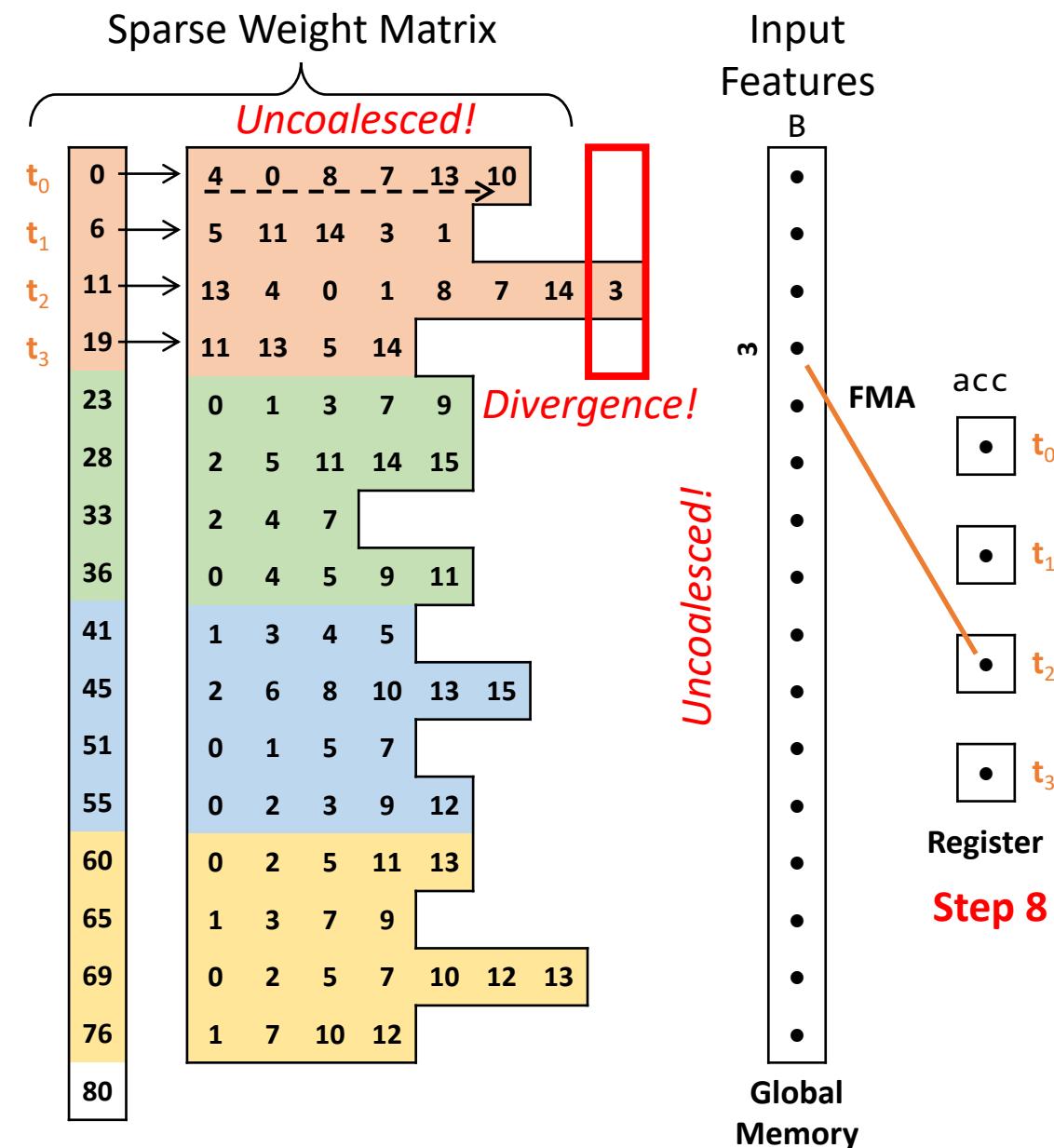
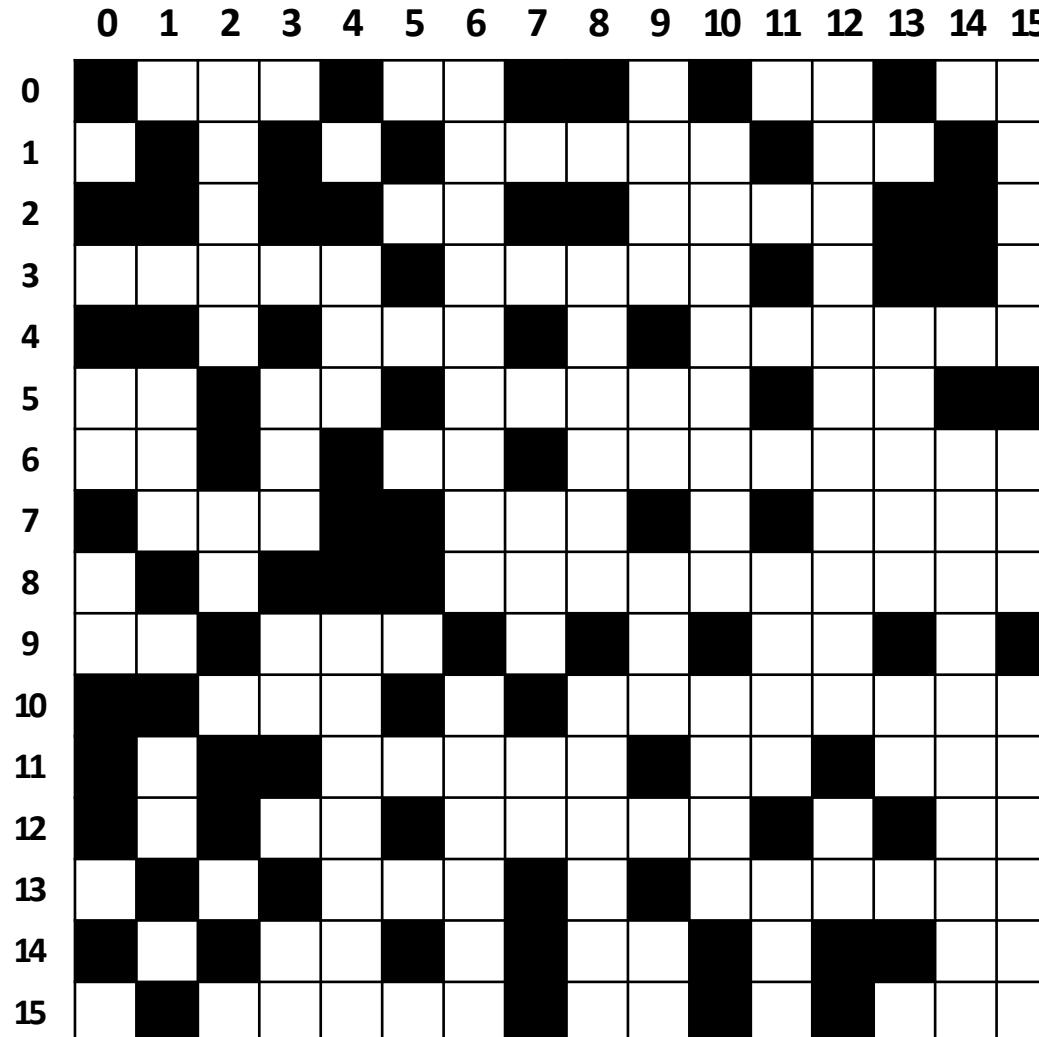


Input Features

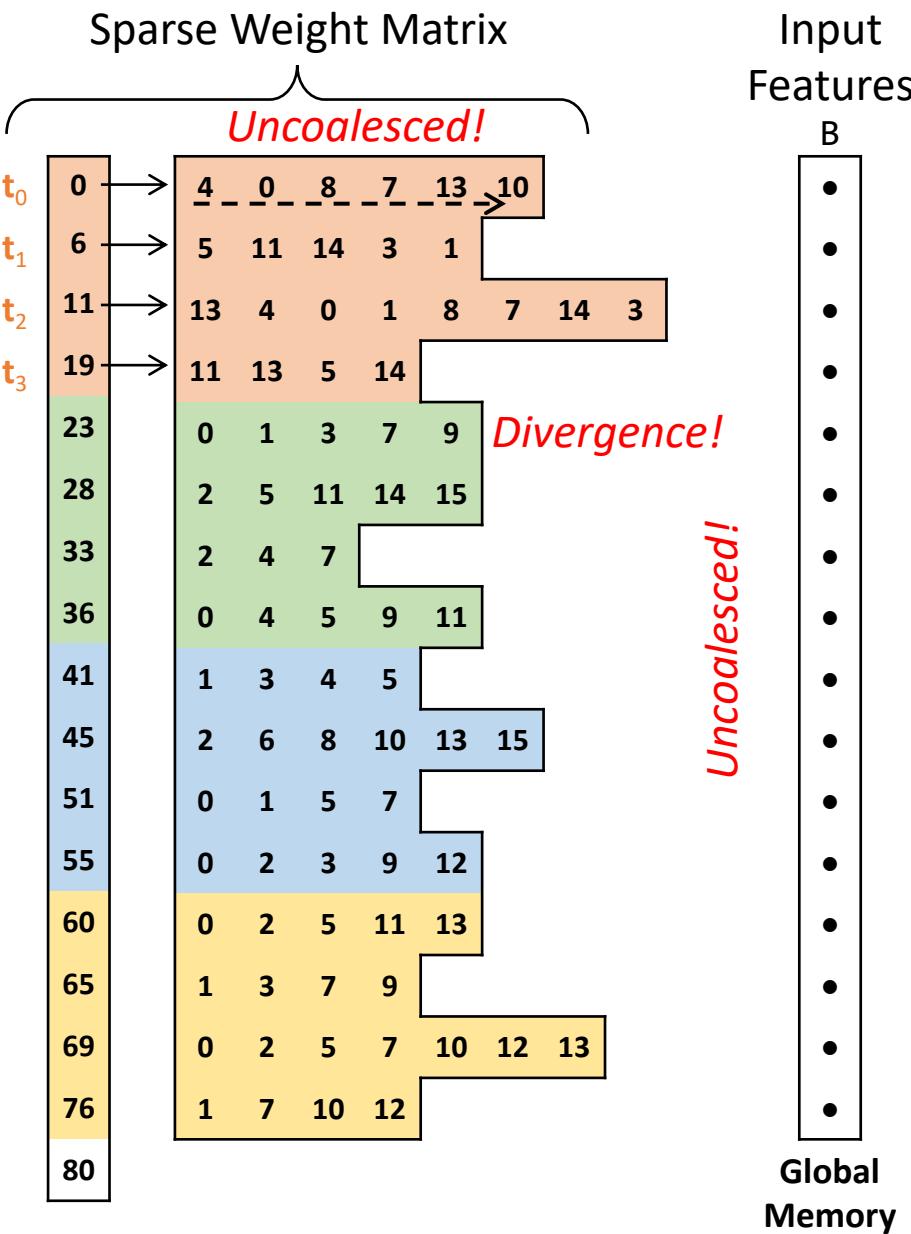
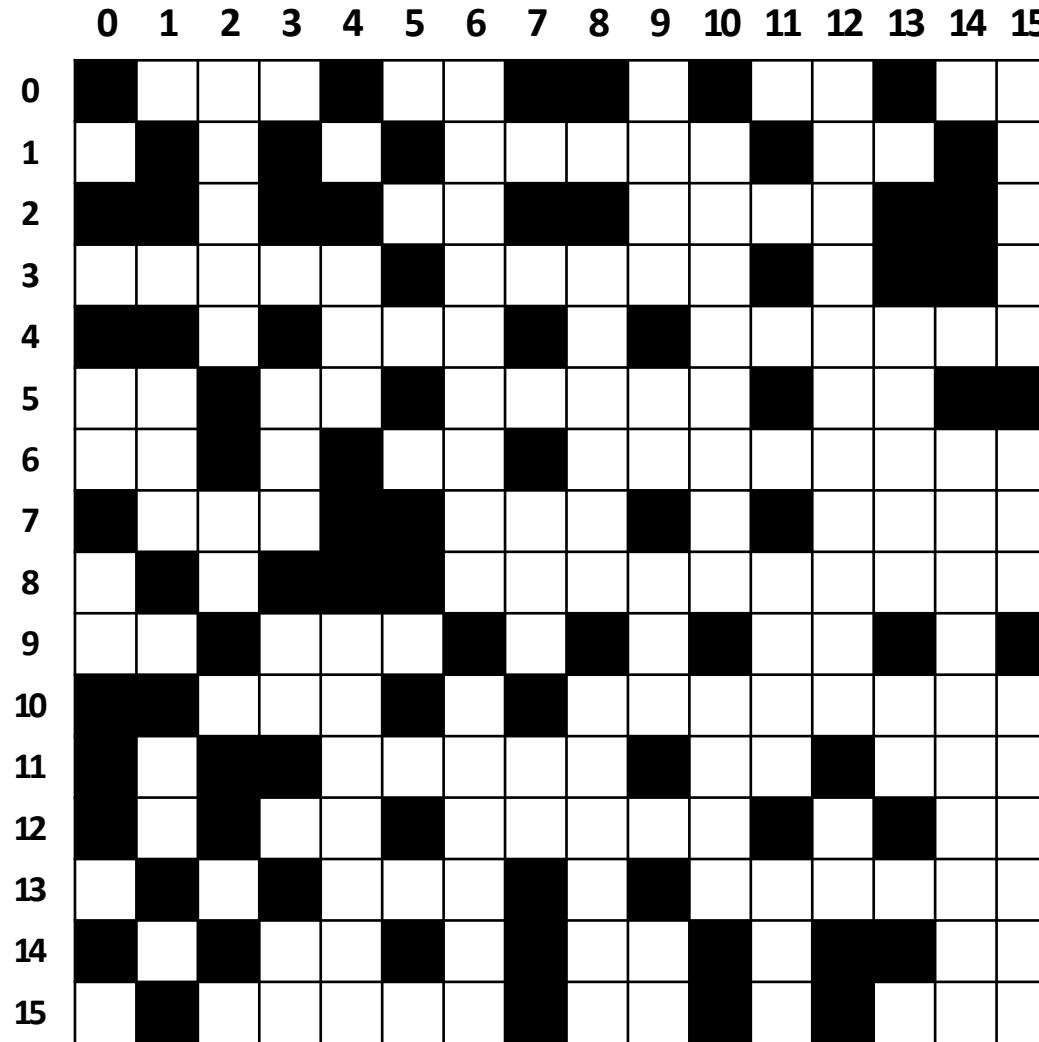


acc t_0
 t_1
 t_2
 t_3
FMA
Register
Step 7

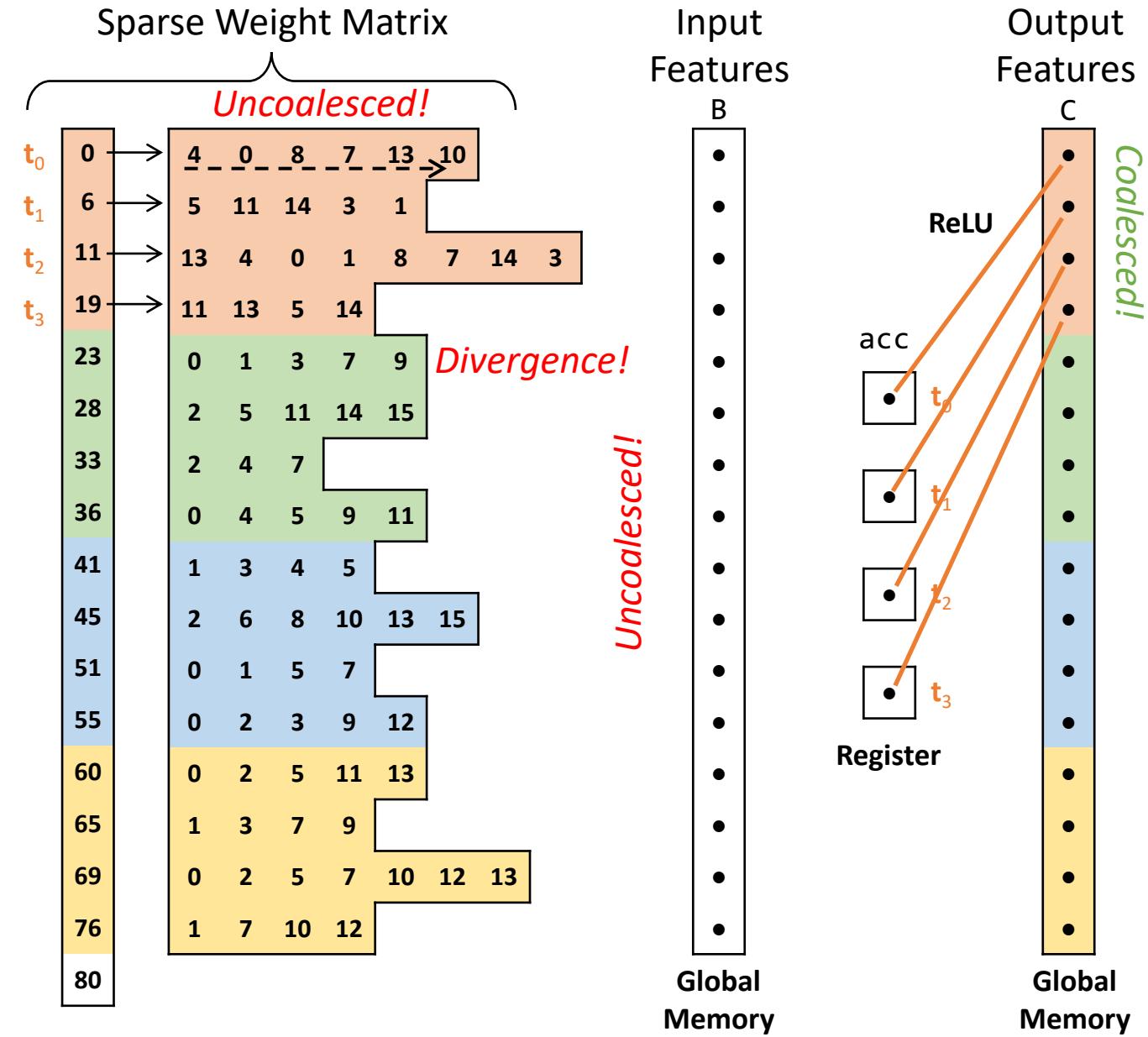
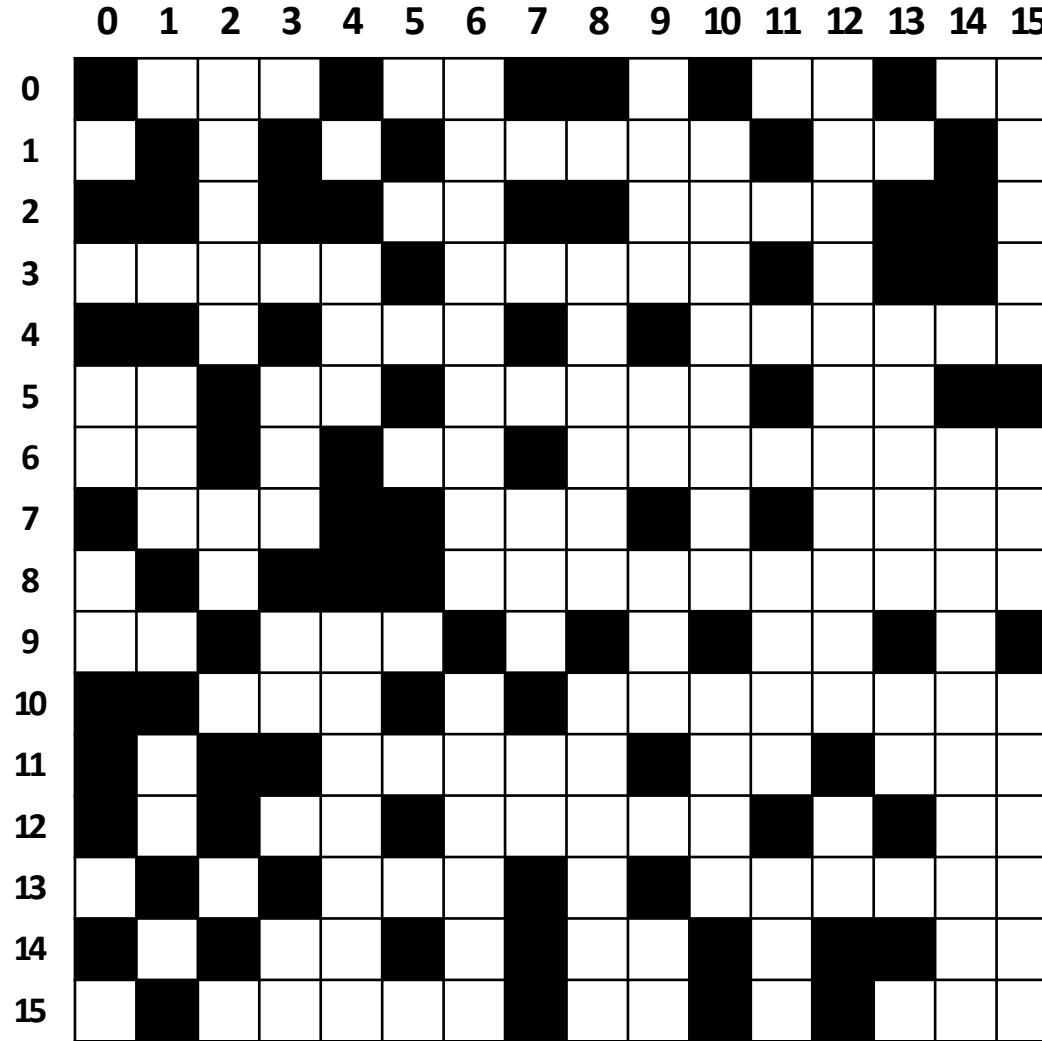
Baseline CSR SpMM



Baseline CSR SpMM



Baseline CSR SpMM



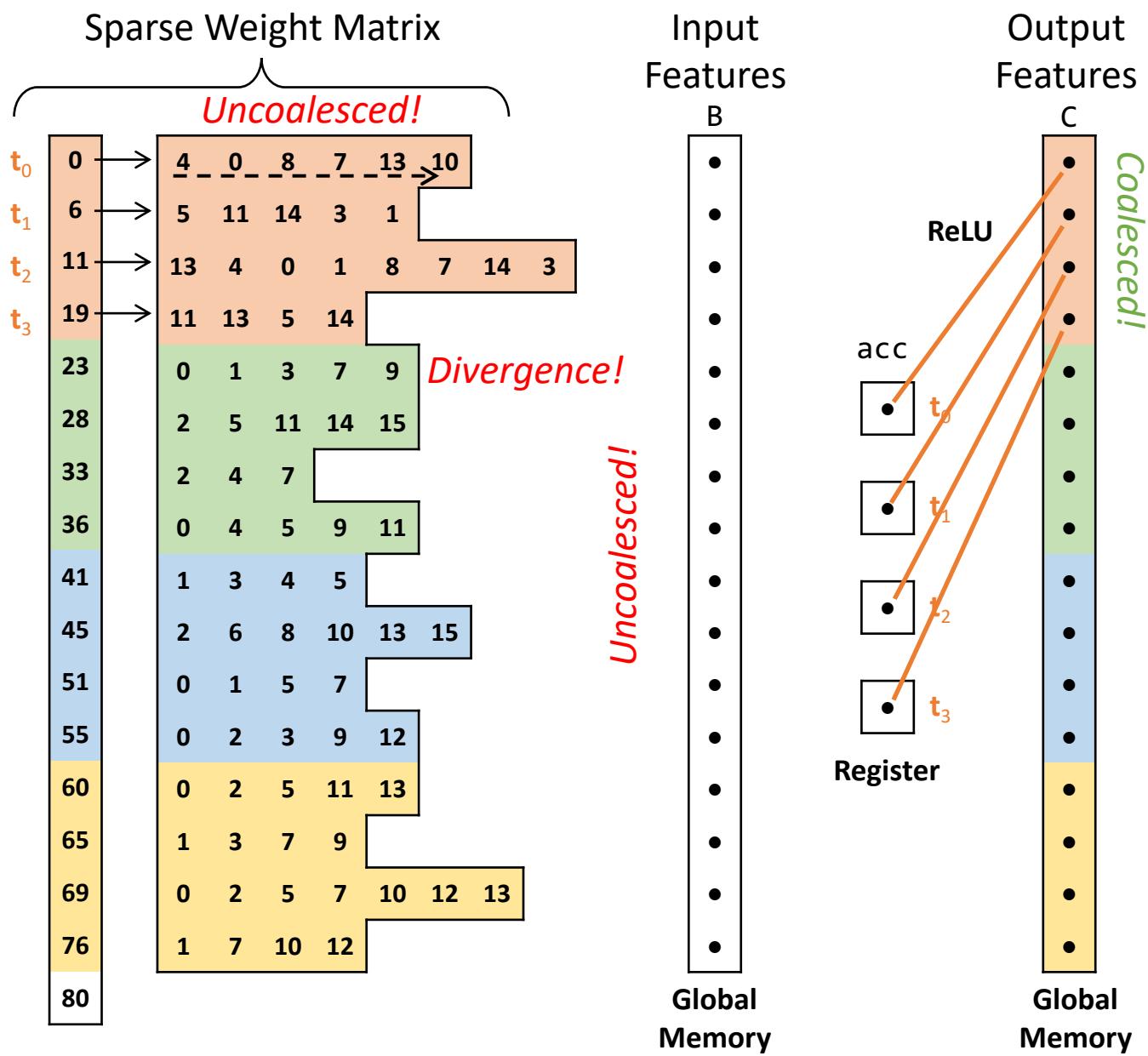
Baseline CSR SpMM

Data Access Redundancies

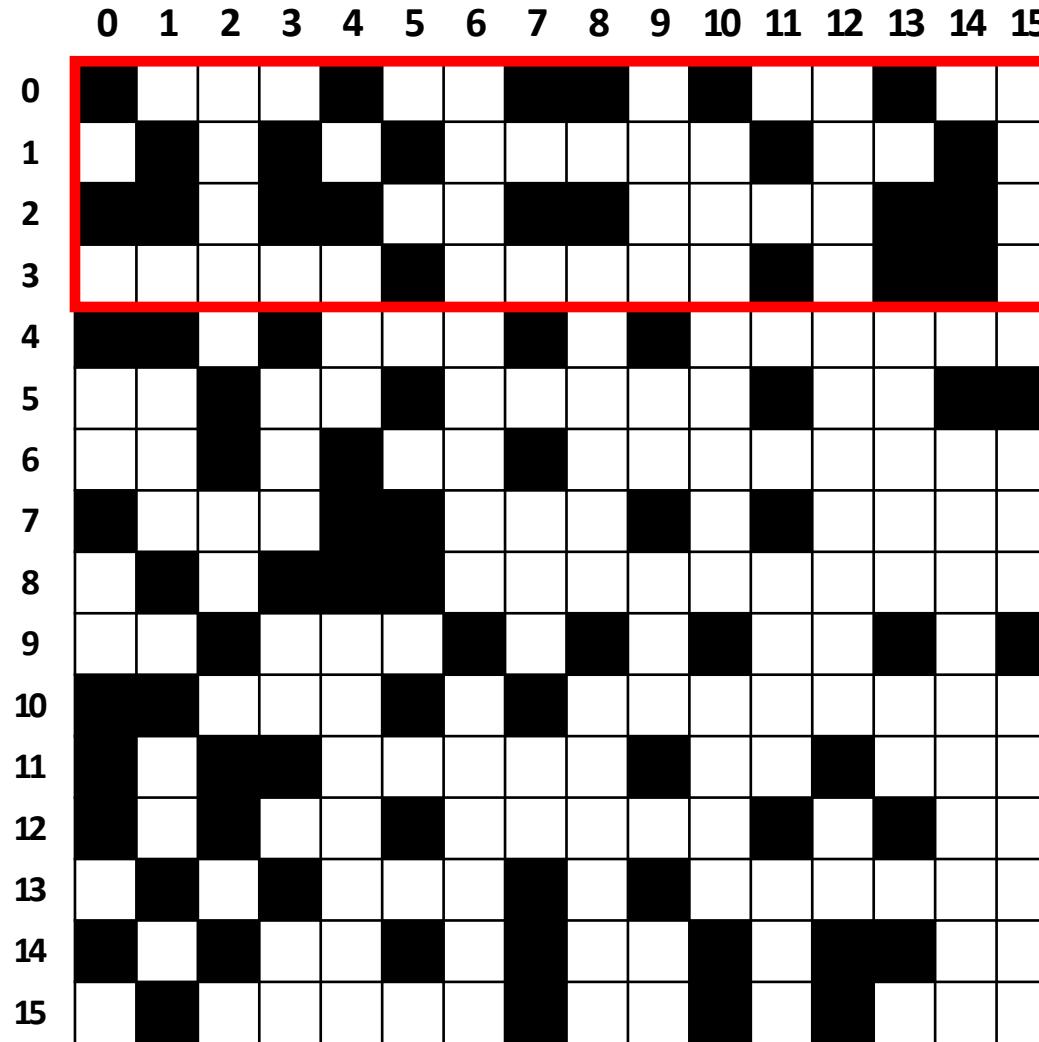
- Input features by different threads
- Weight matrix by all output features

Data Access Latencies

- Irregular access to input features
- Uncoalesced access to weight matrix



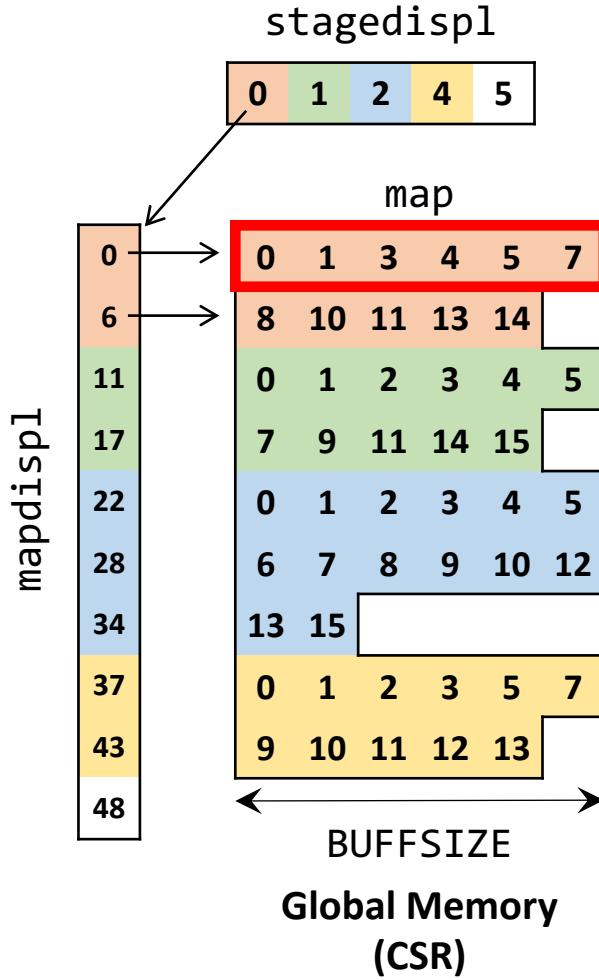
Tiled SpMM



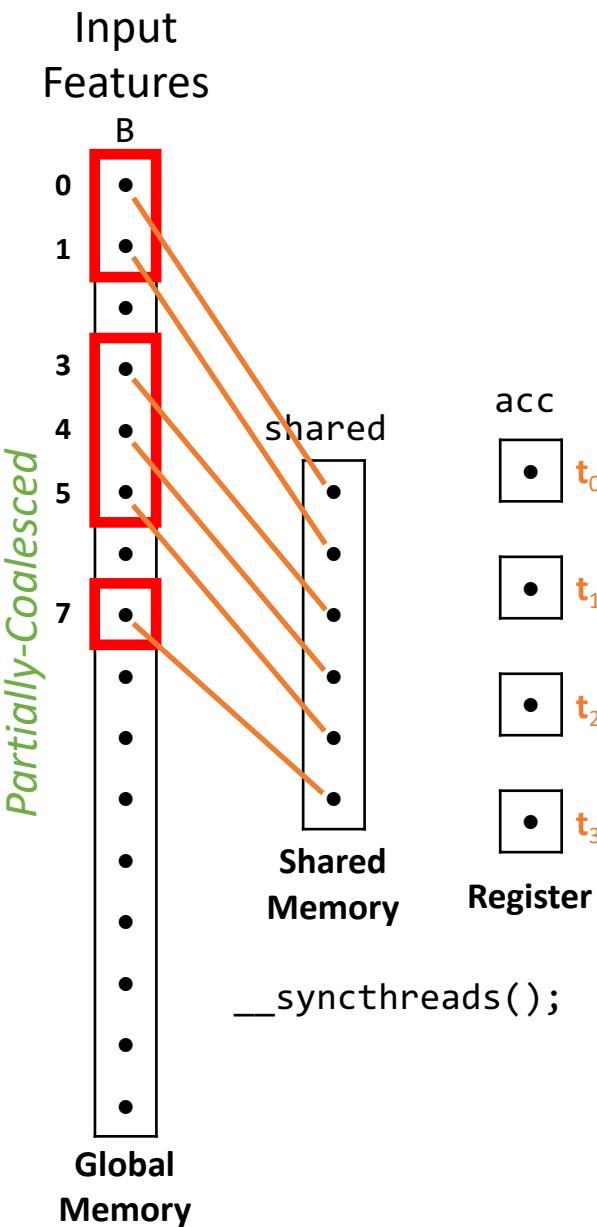
Input
Features

B	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	•															
1	•															
2		•														
3			•													
4				•												
5					•											
6						•										
7							•									
8								•								
9									•							
10										•						
11											•					
12												•				
13													•			
14														•		
15																•

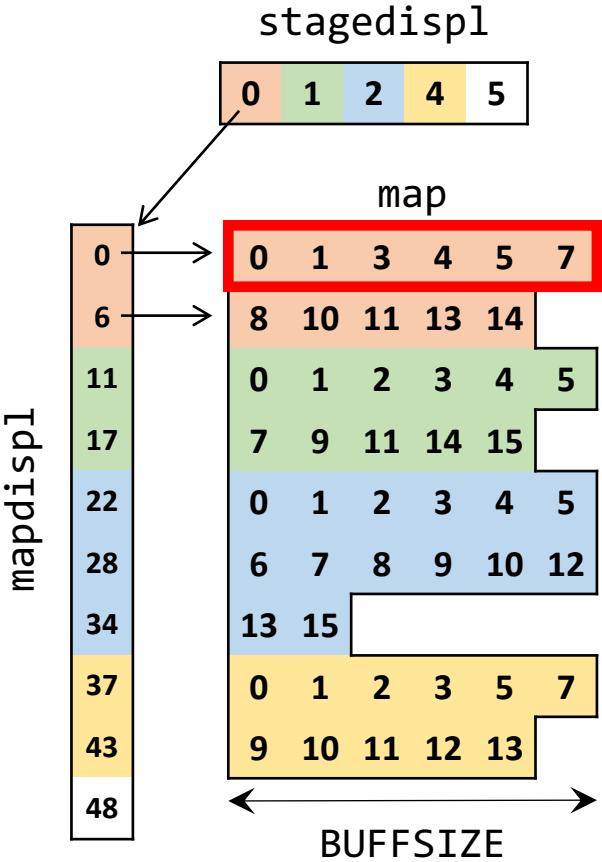
Tiling Data Structures



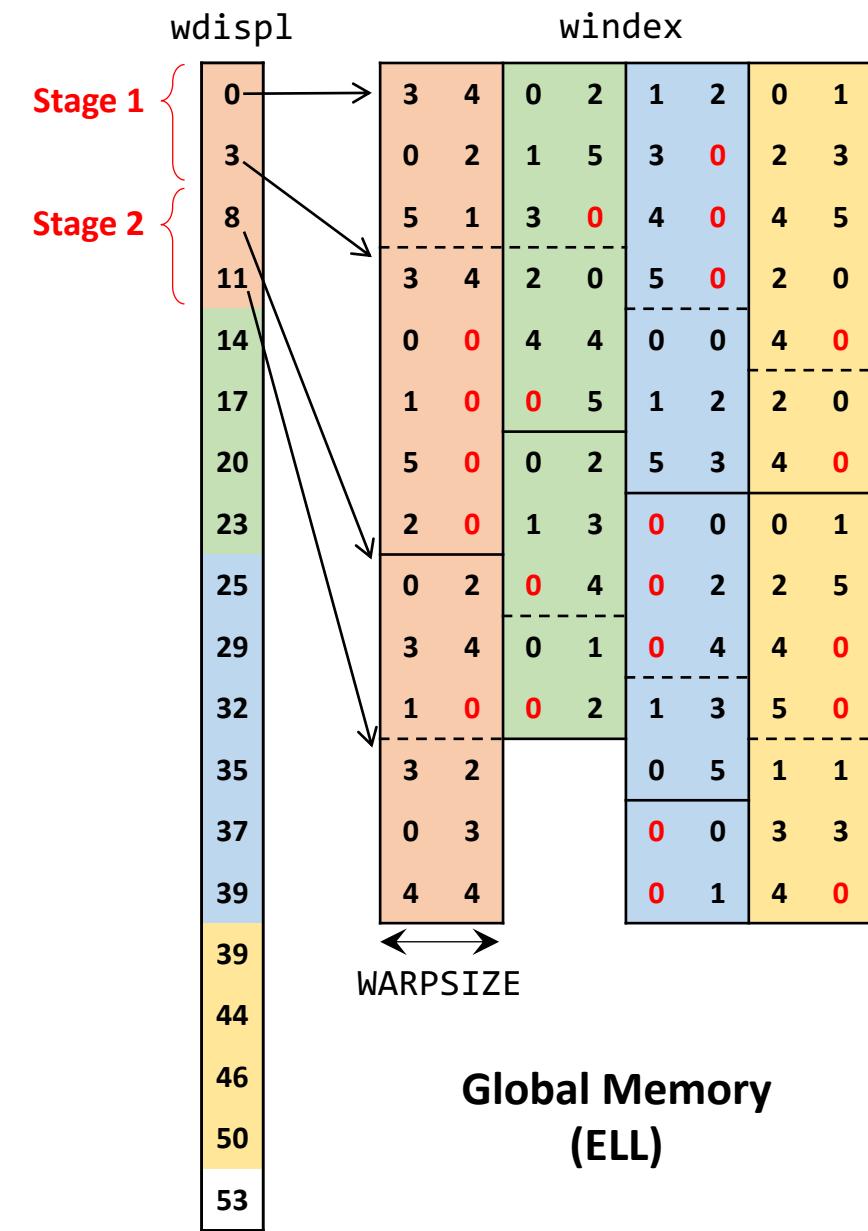
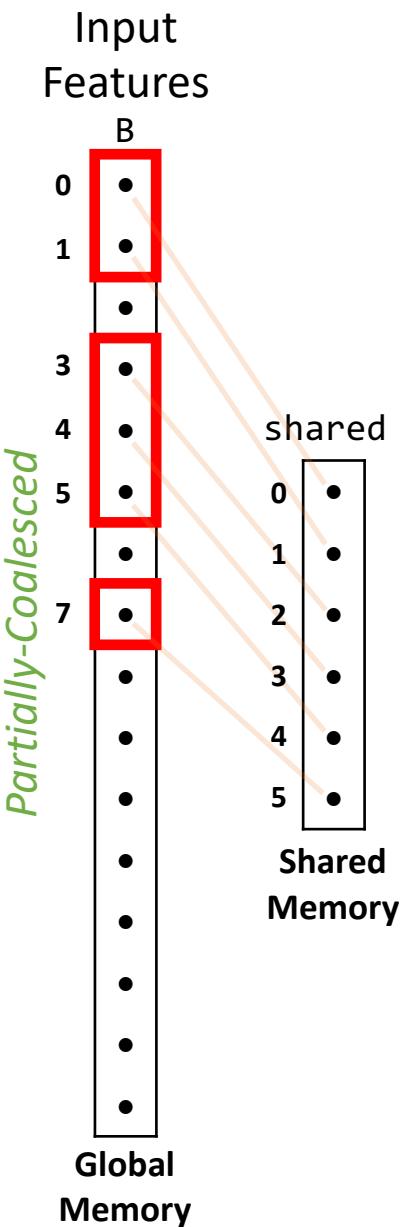
First Staging



Tiling Data Structures

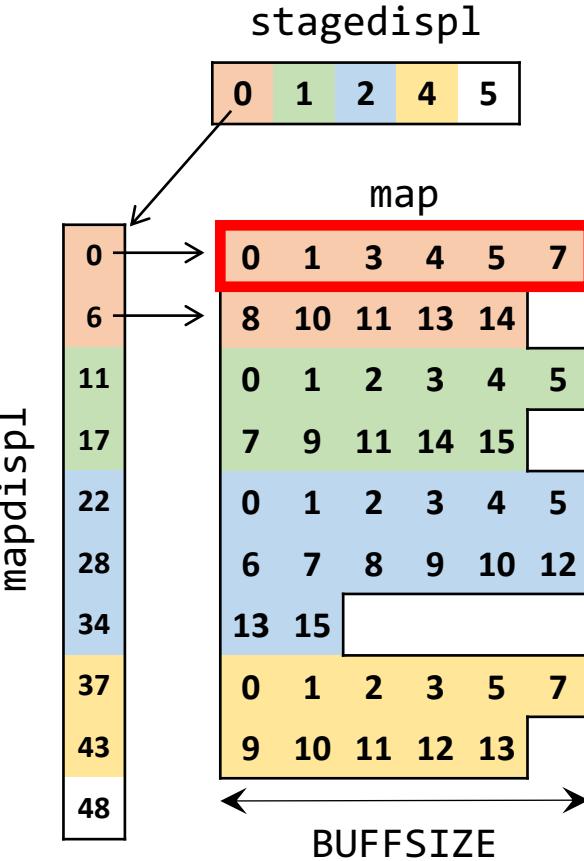


First Staging

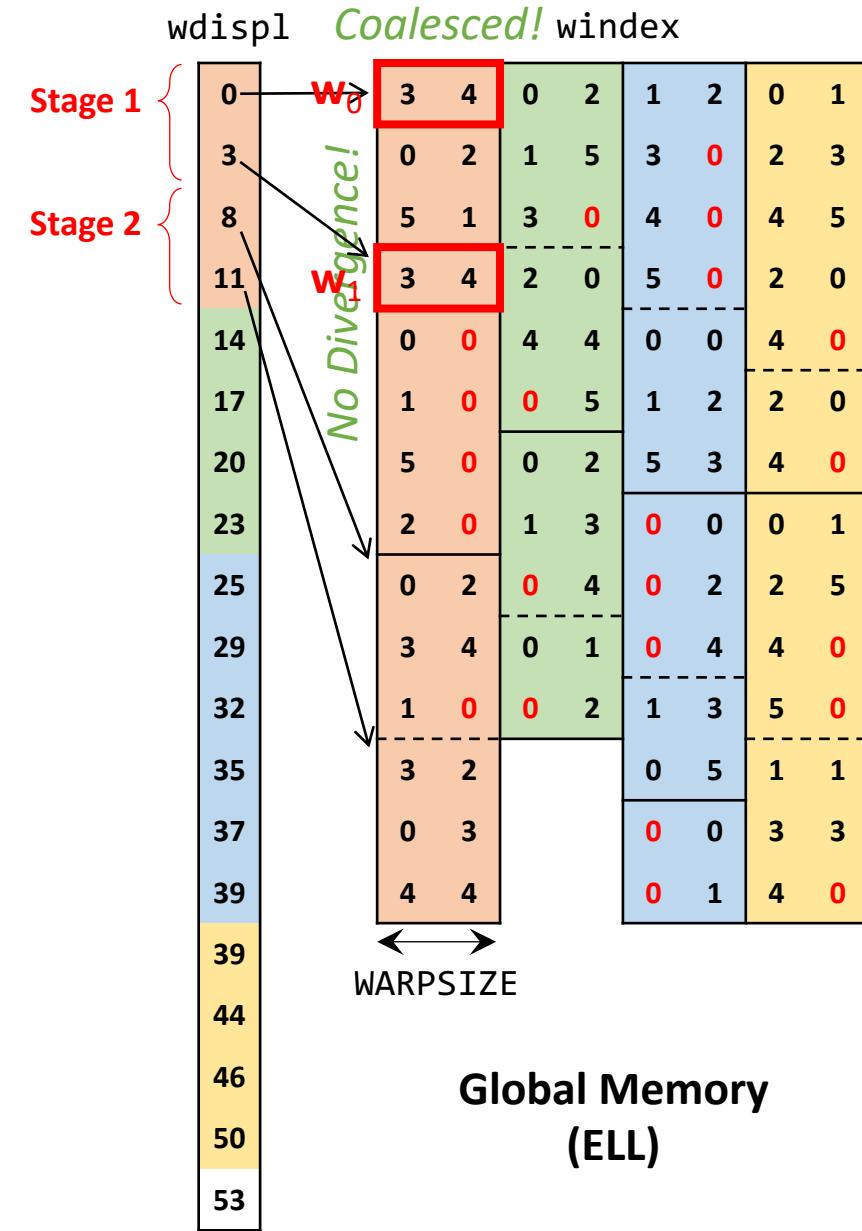
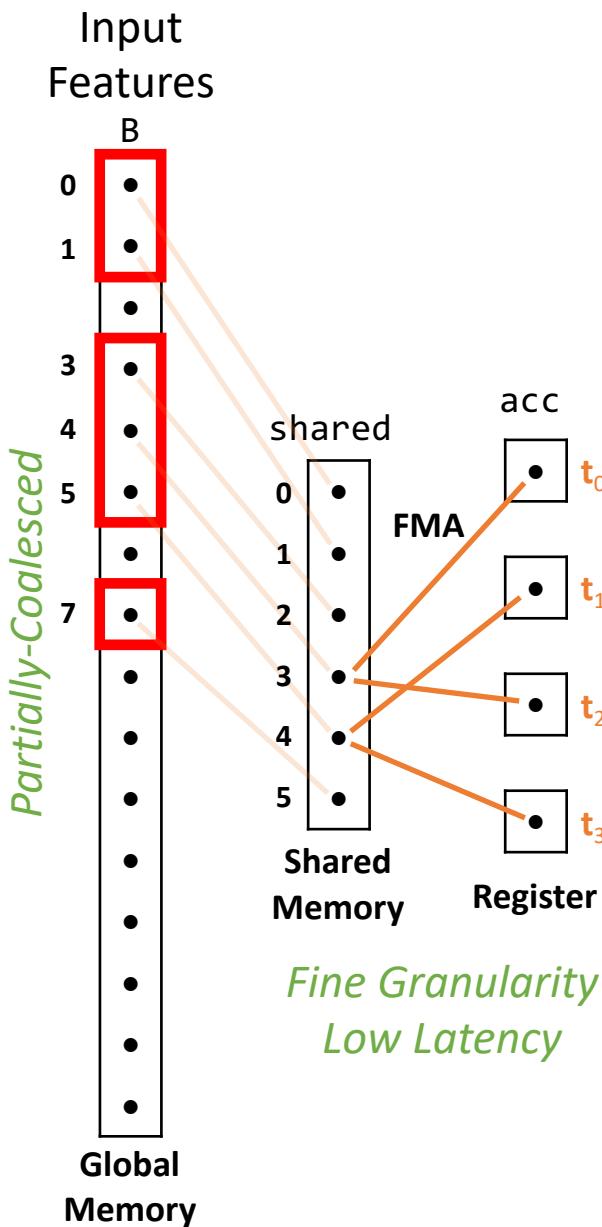


Global Memory (ELL)

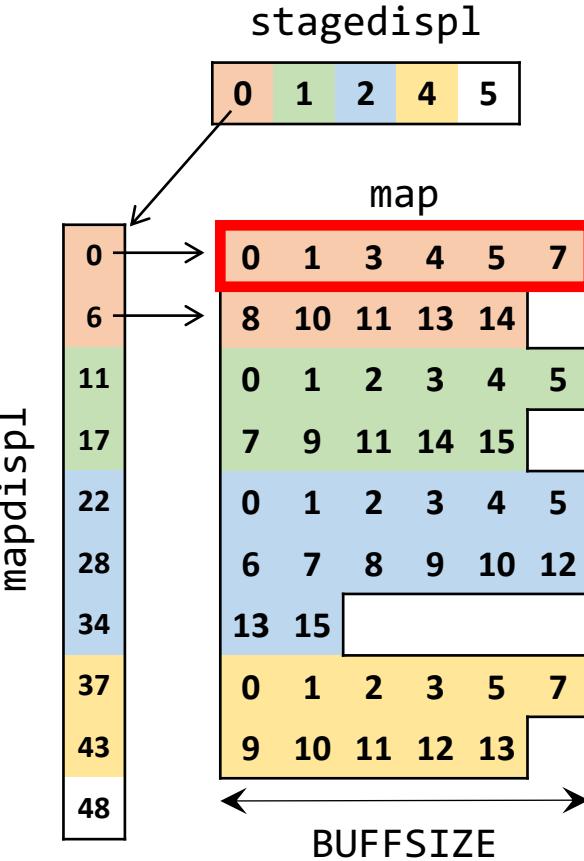
Tiling Data Structures



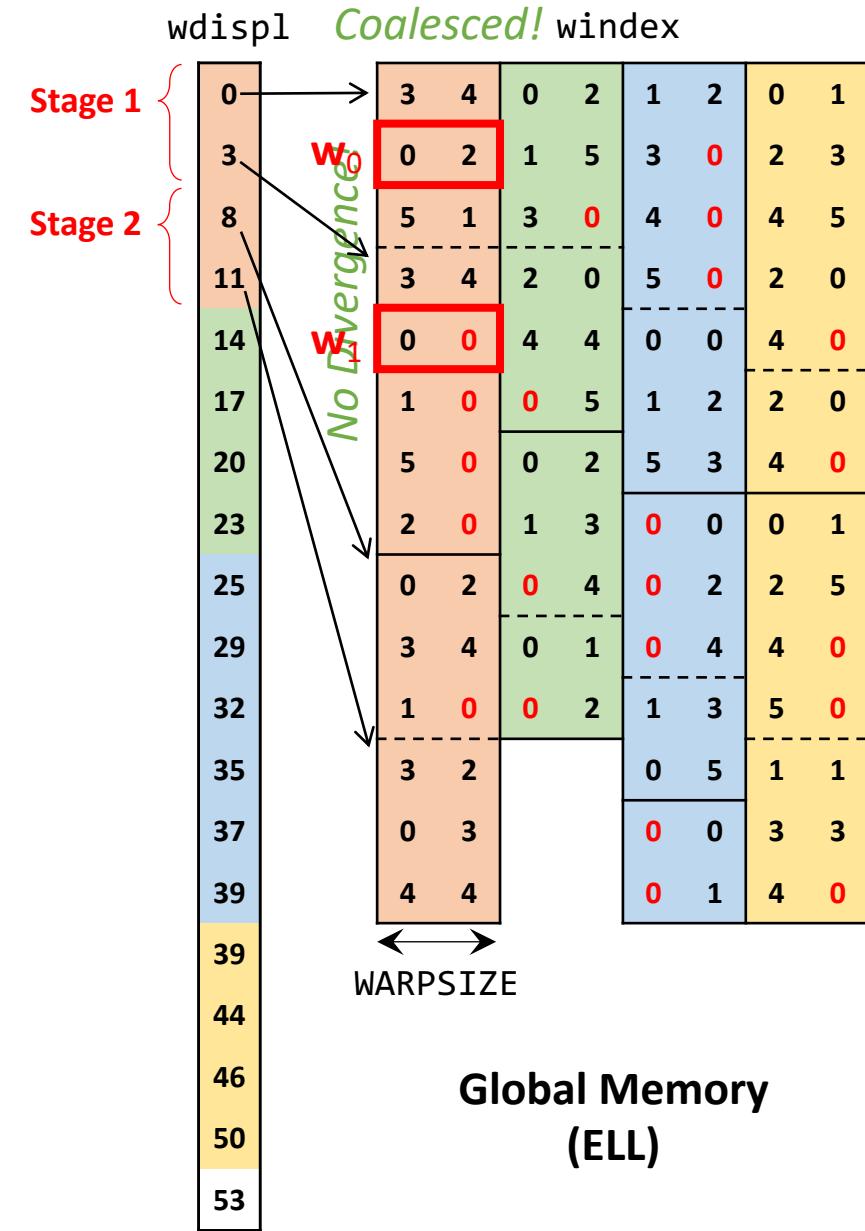
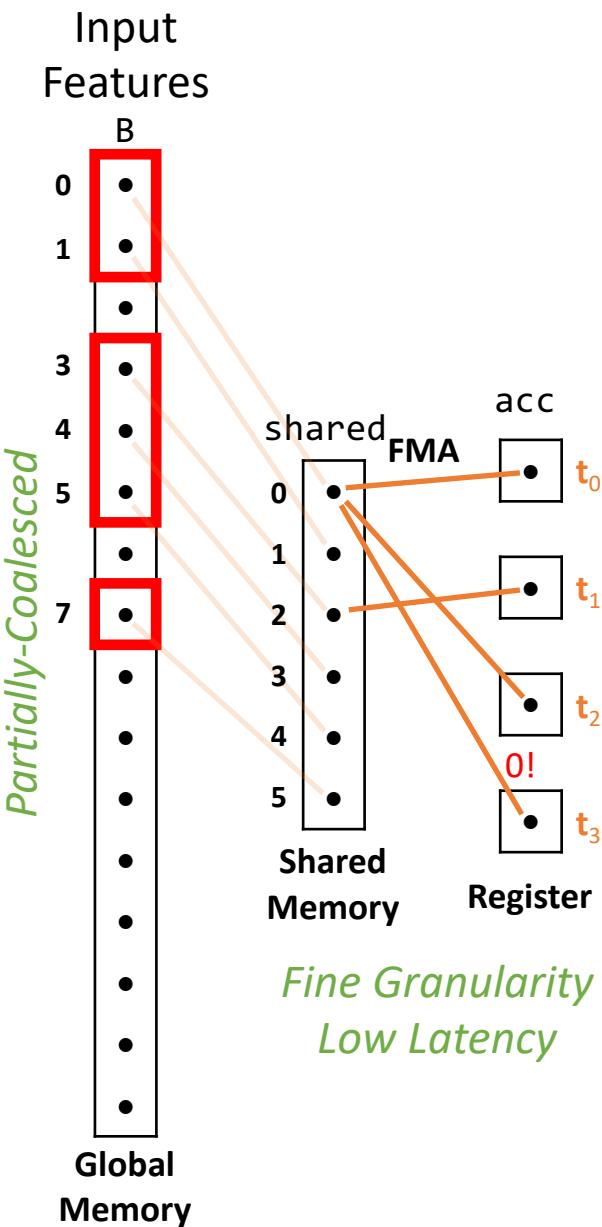
First Staging



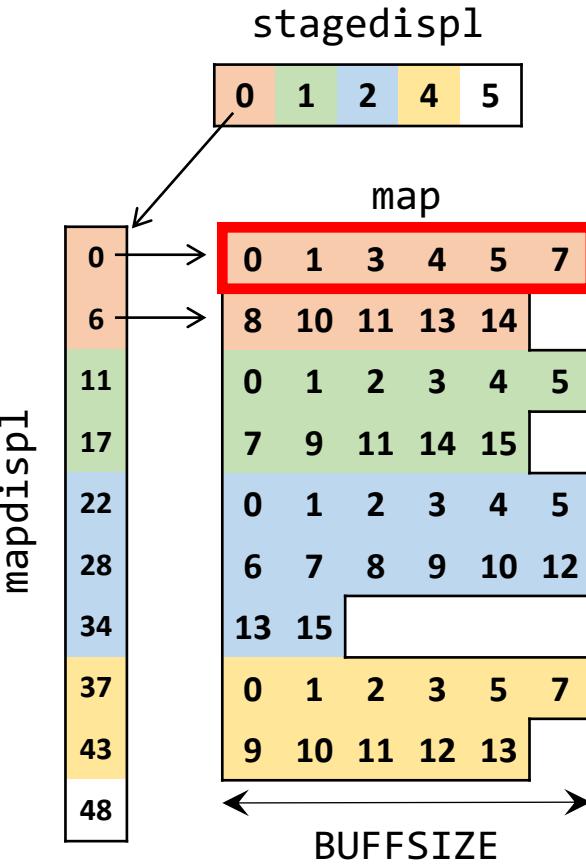
Tiling Data Structures



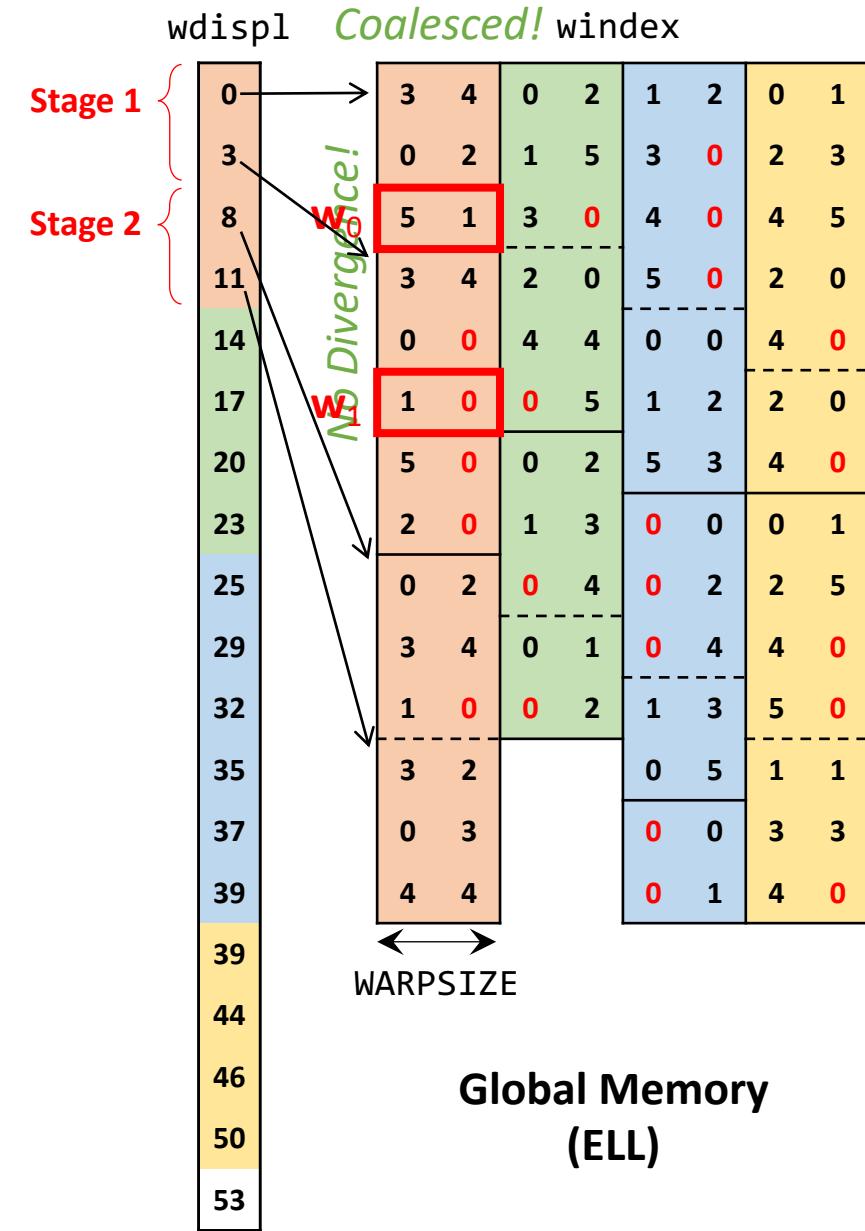
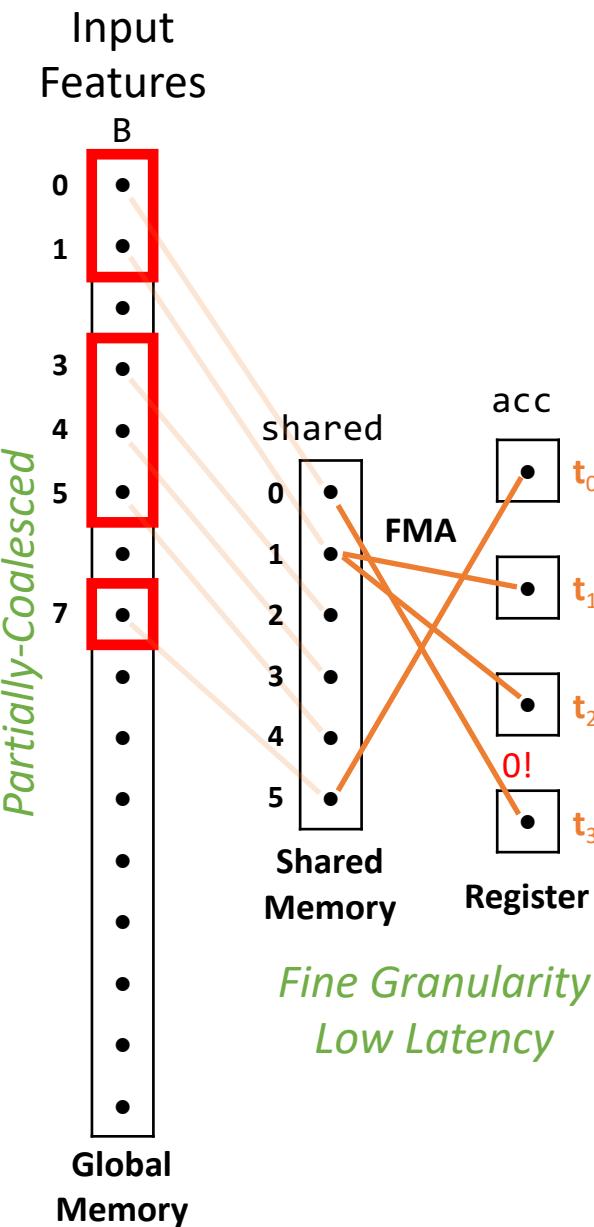
First Staging



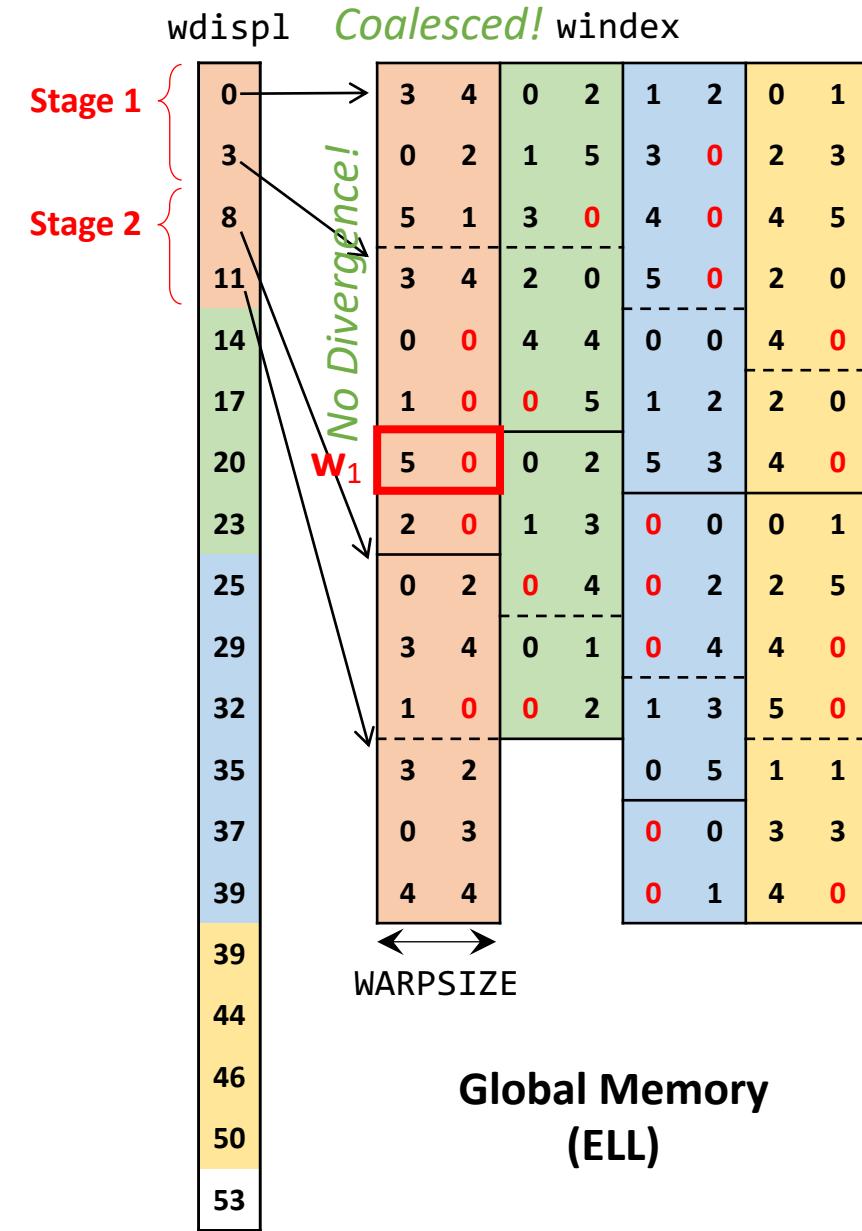
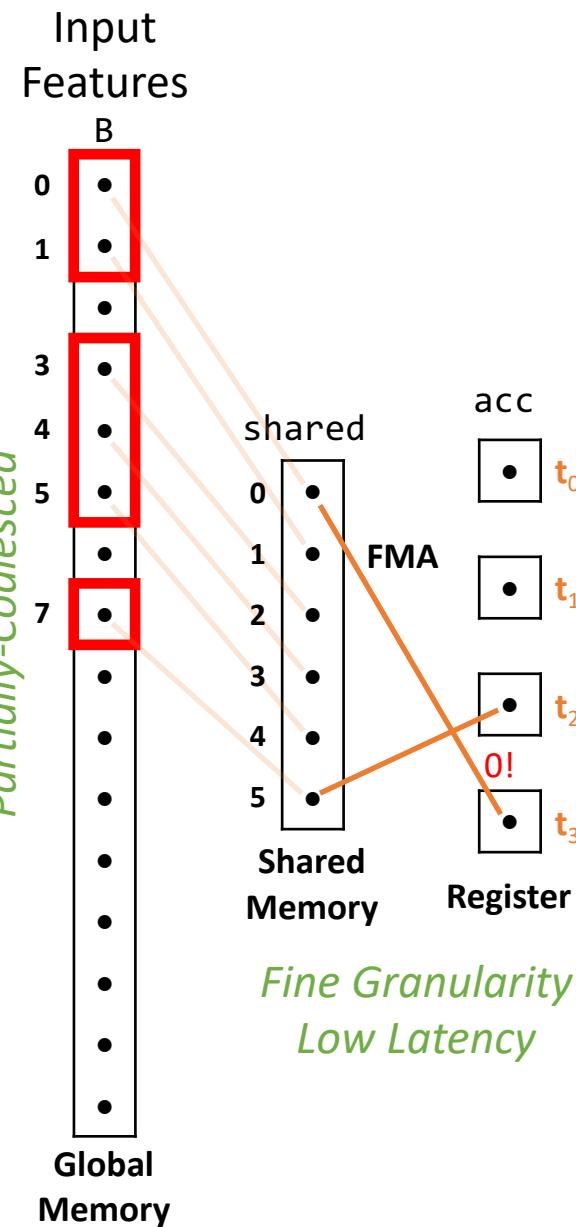
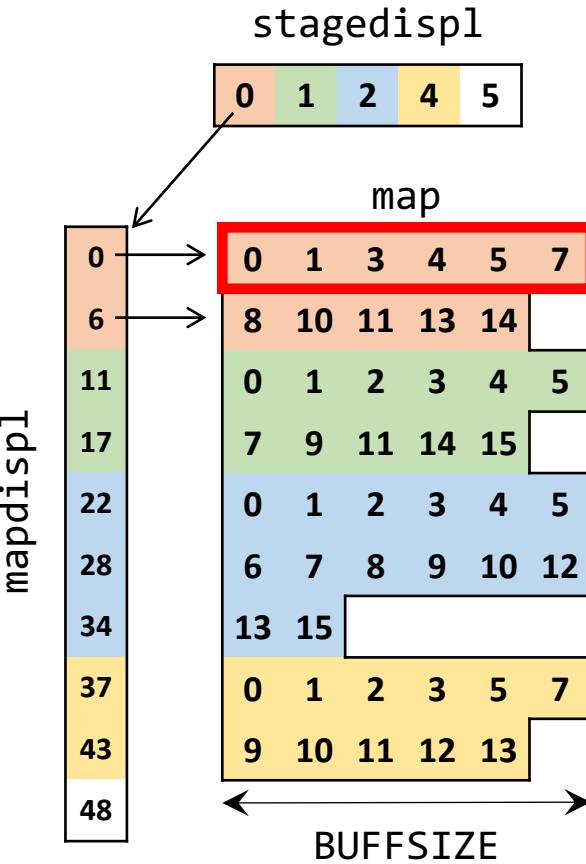
Tiling Data Structures



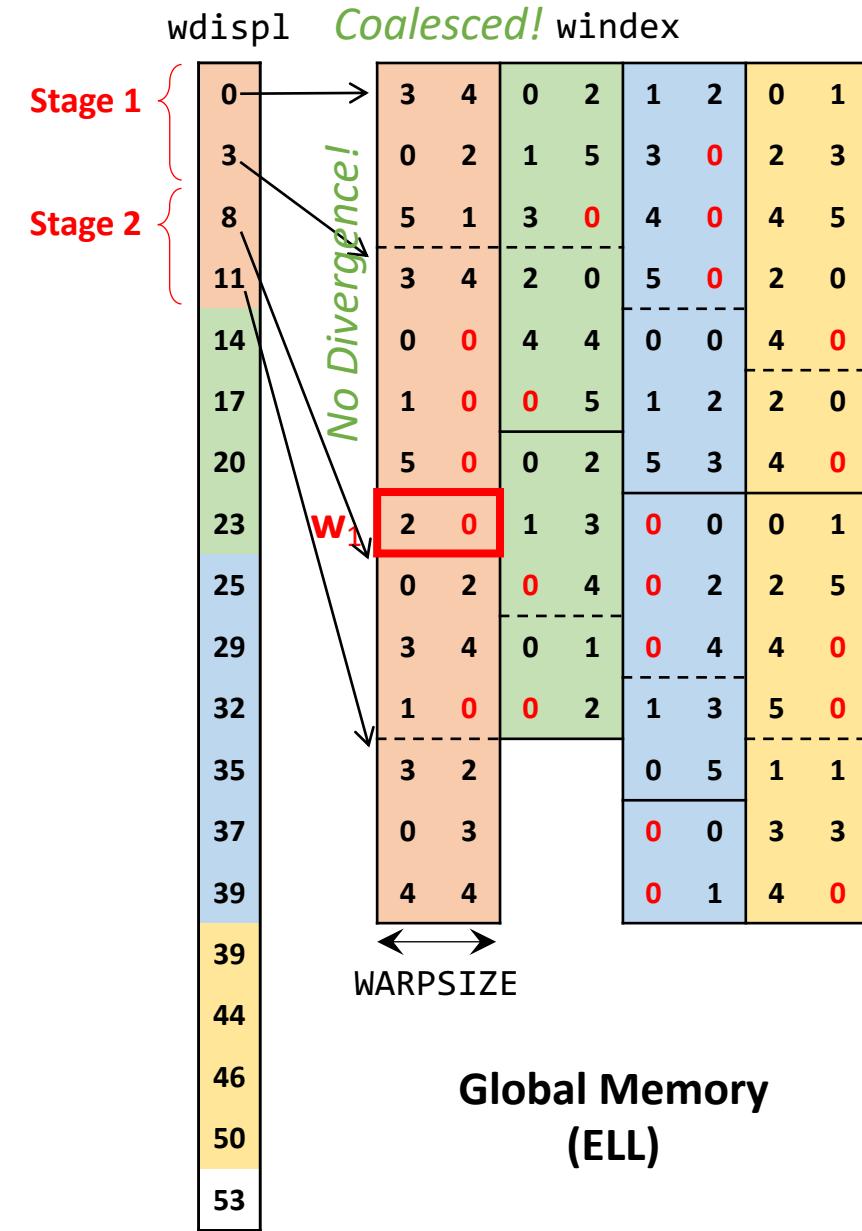
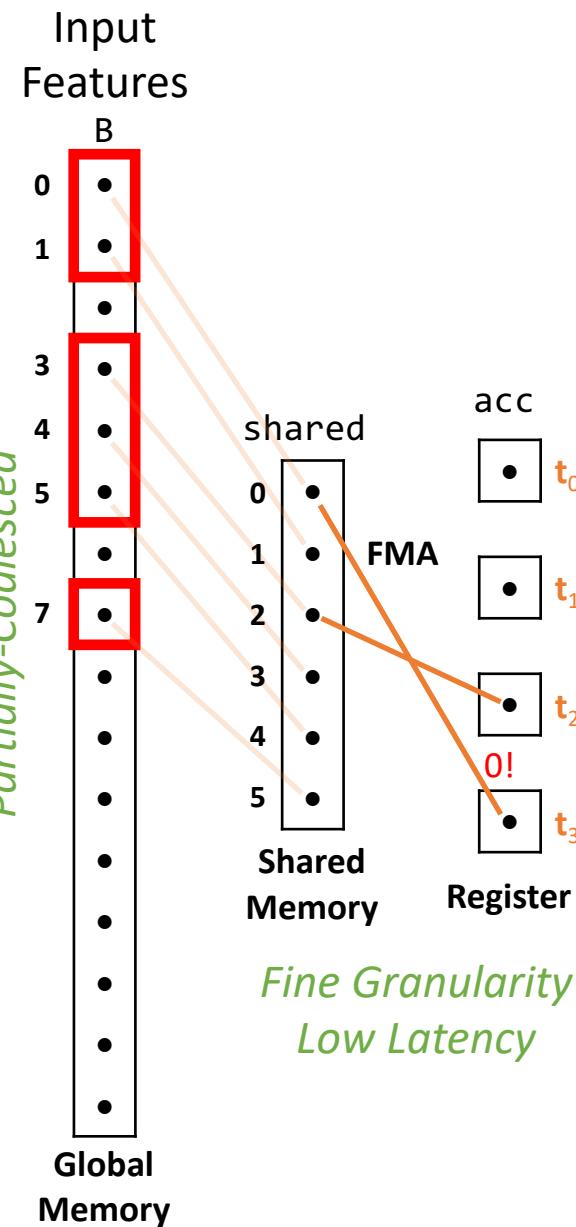
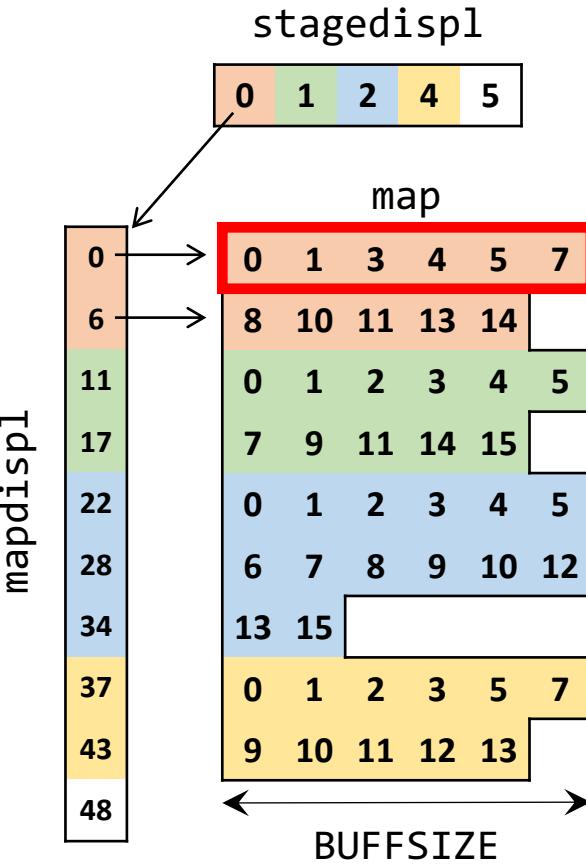
First Staging



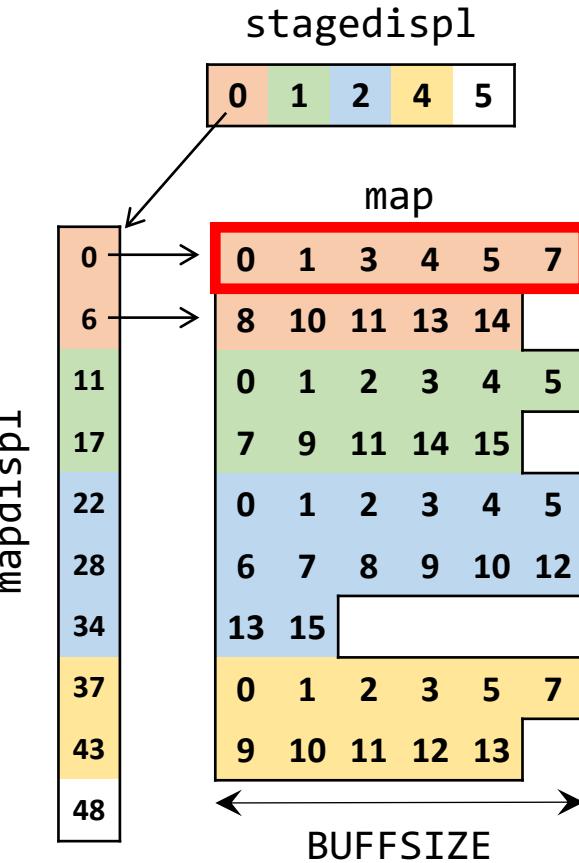
Tiling Data Structures



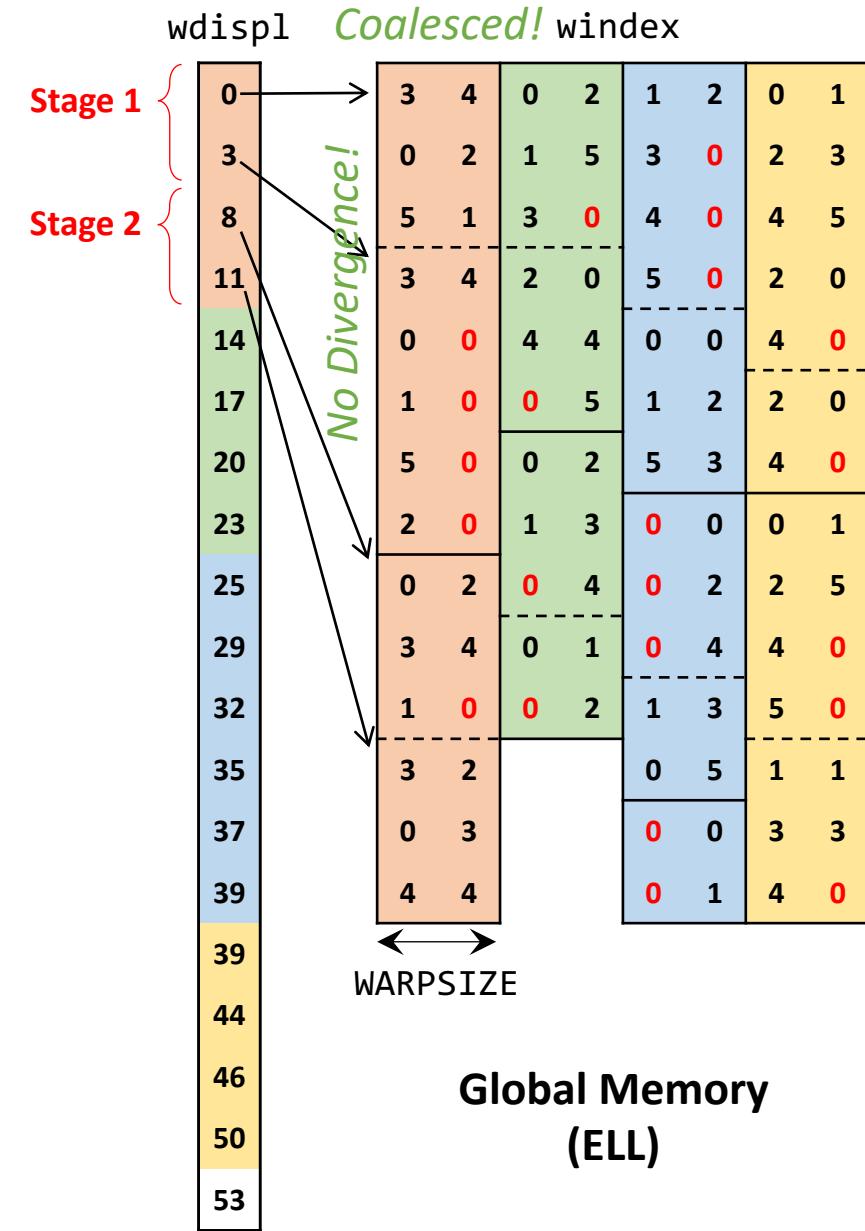
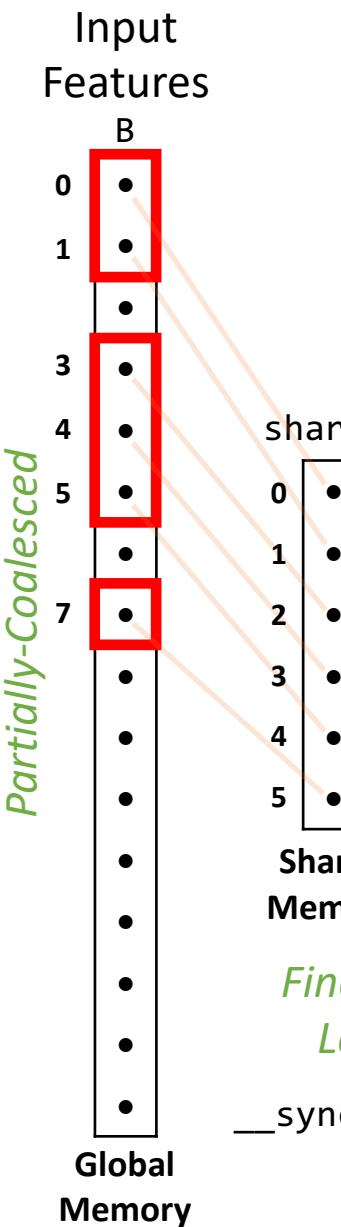
Tiling Data Structures



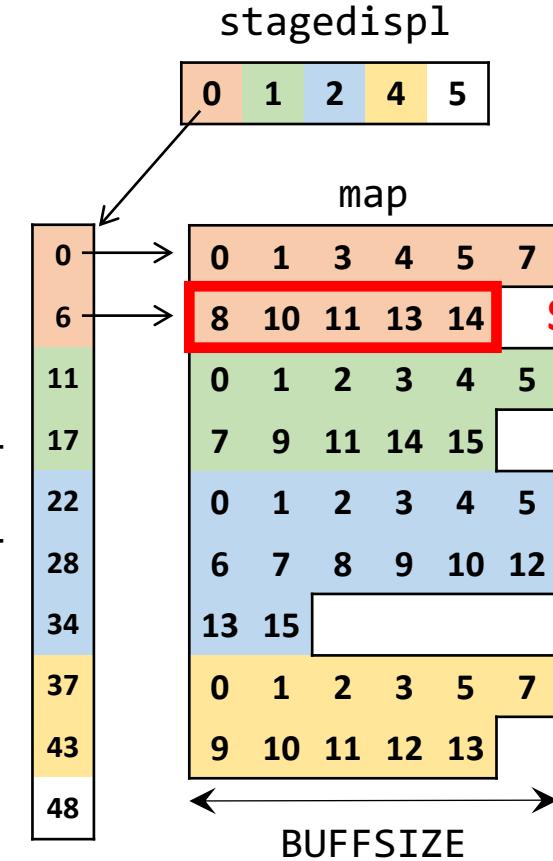
Tiling Data Structures



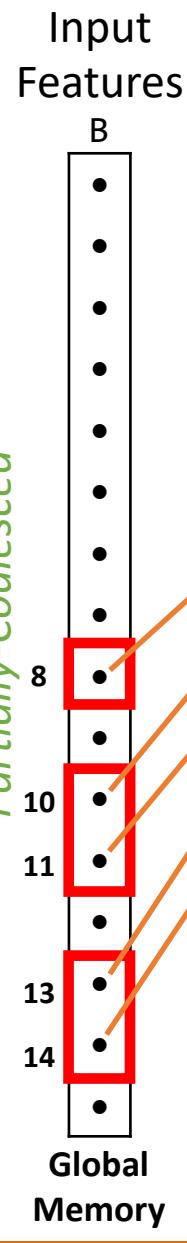
First Staging



Tiling Data Structures



Global Memory
(CSR)



Partially-Coalesced

8 10 11 13 14

0 1 2 3 4 5

Shared
Memory

Global
Memory

acc

t_0

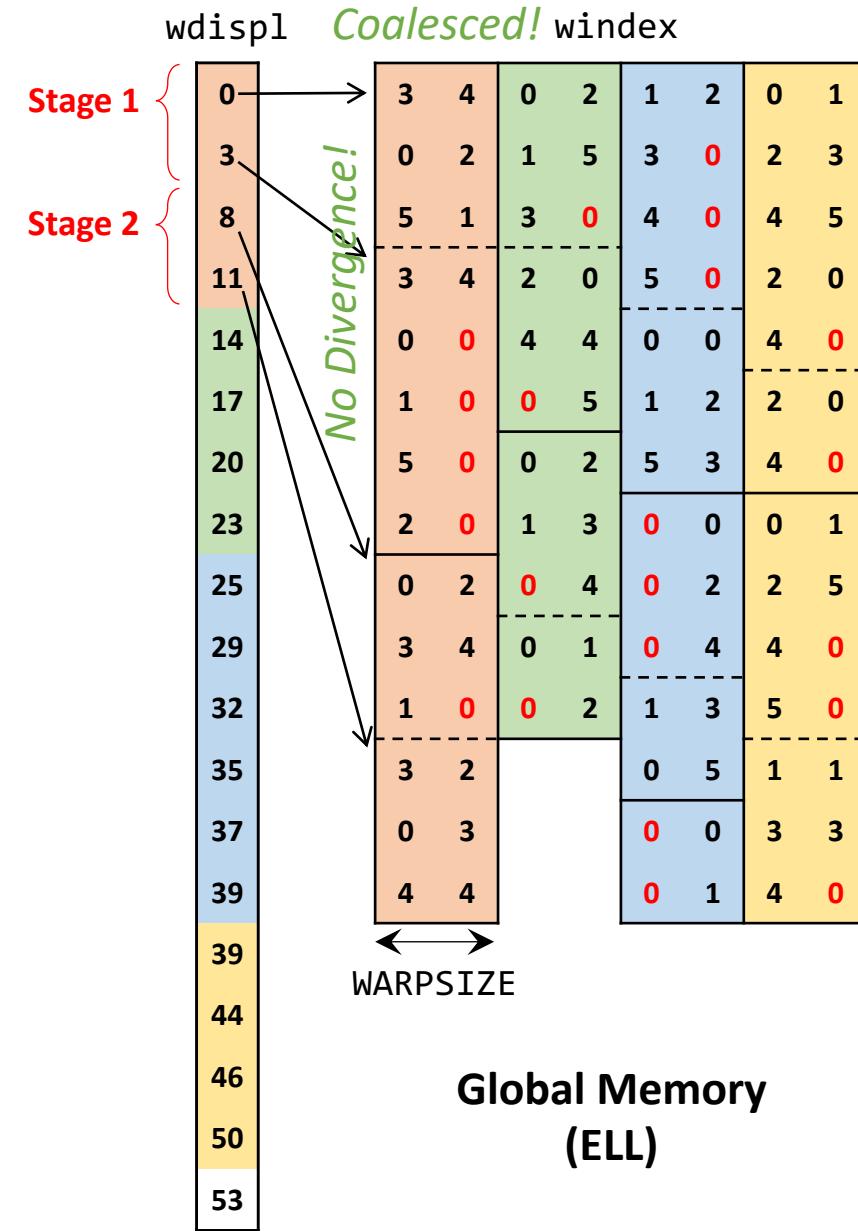
t_1

t_2

Register

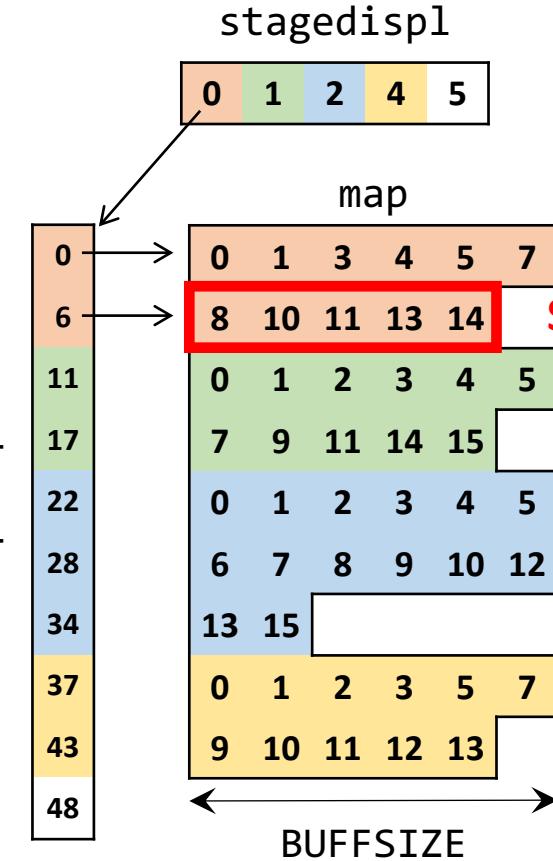
Fine Granularity
Low Latency

`__syncthreads();`



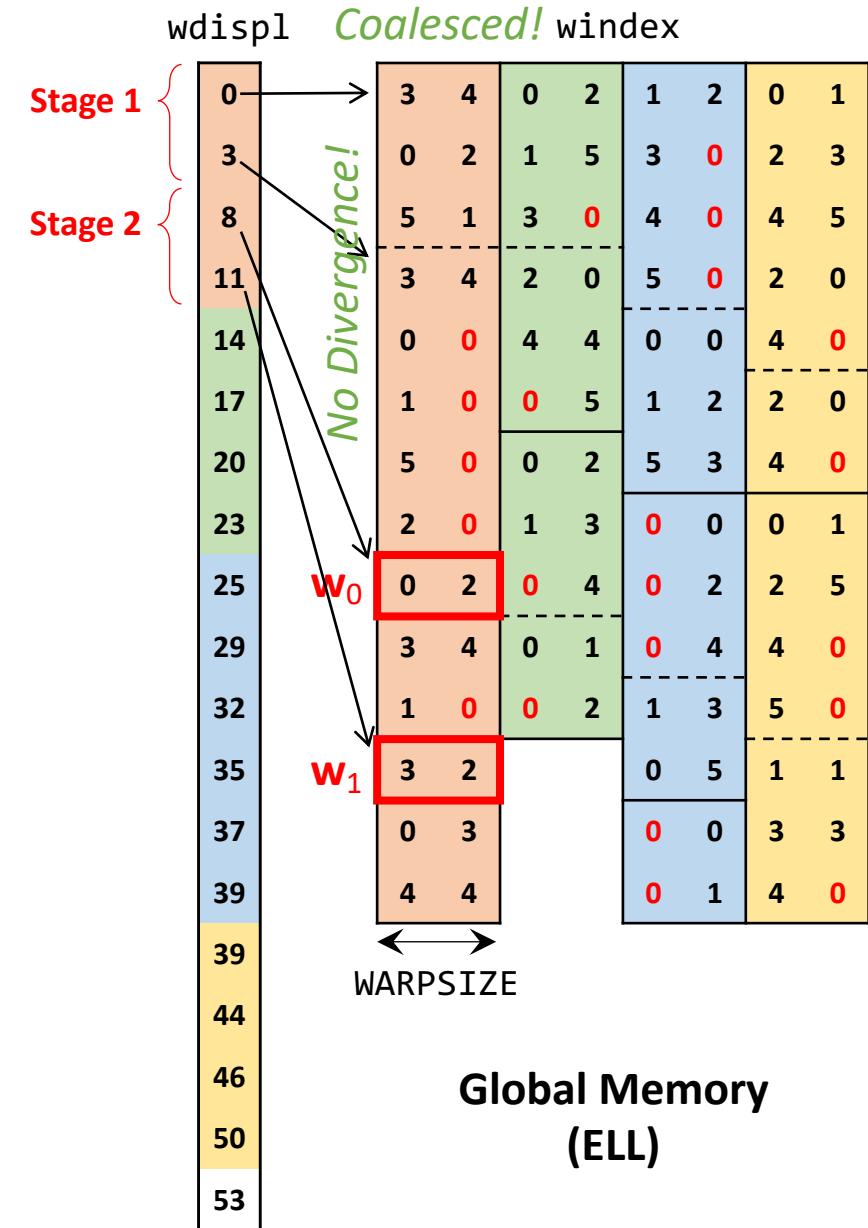
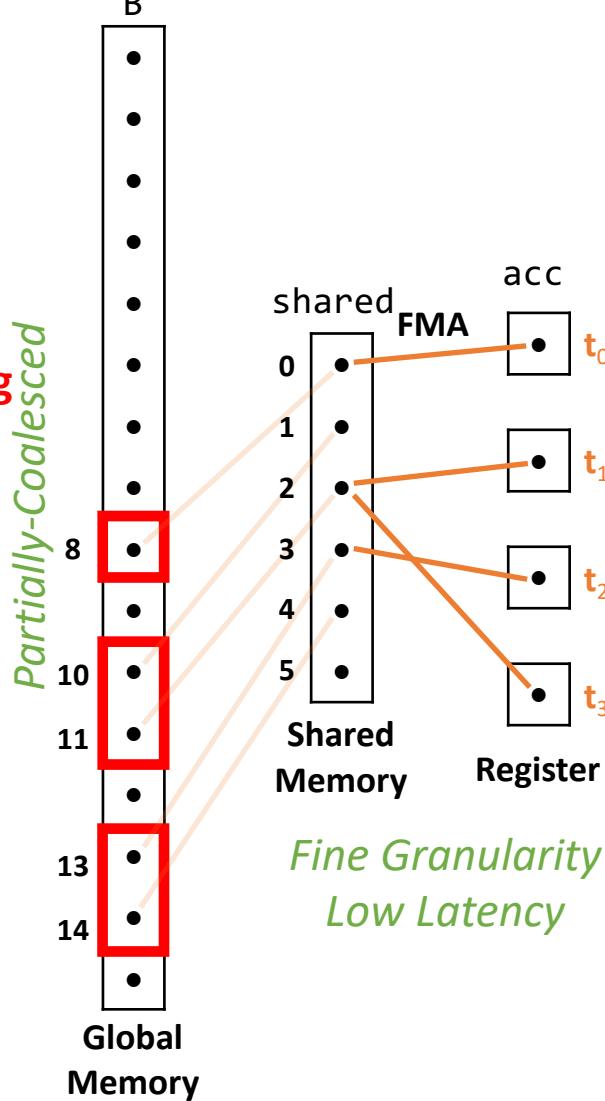
Global Memory
(ELL)

Tiling Data Structures

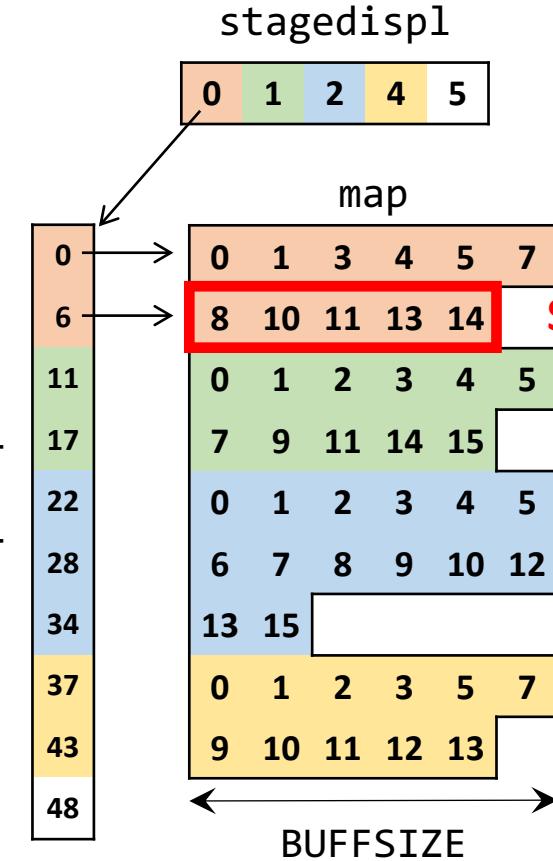


Global Memory
(CSR)

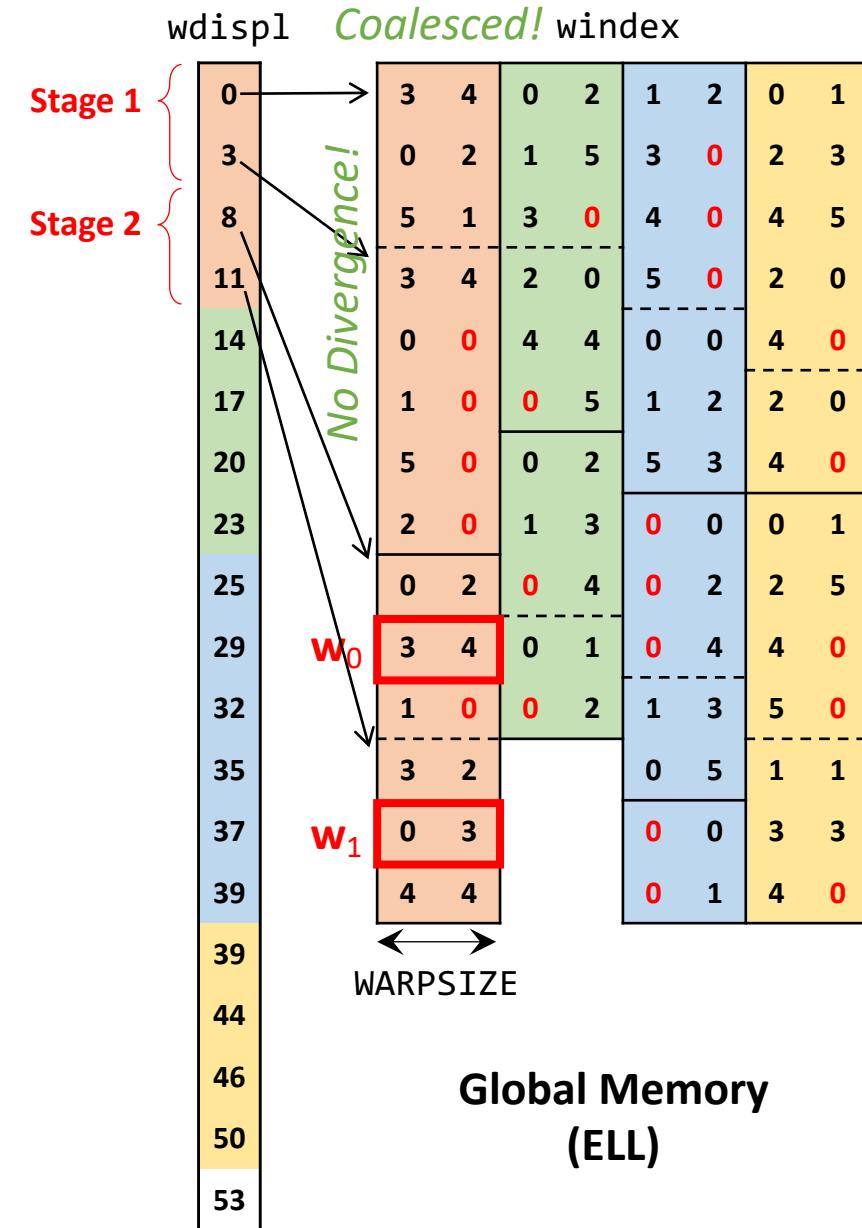
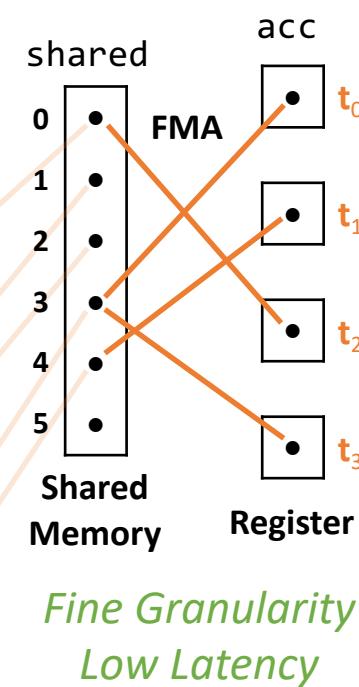
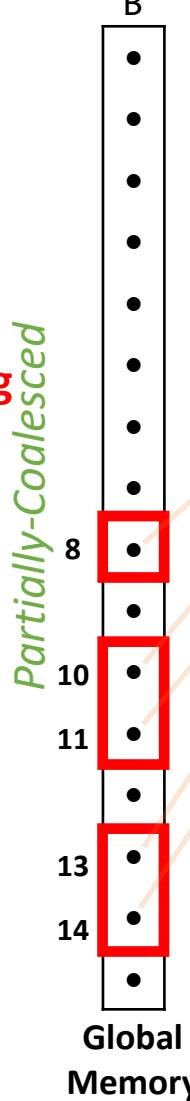
Input
Features



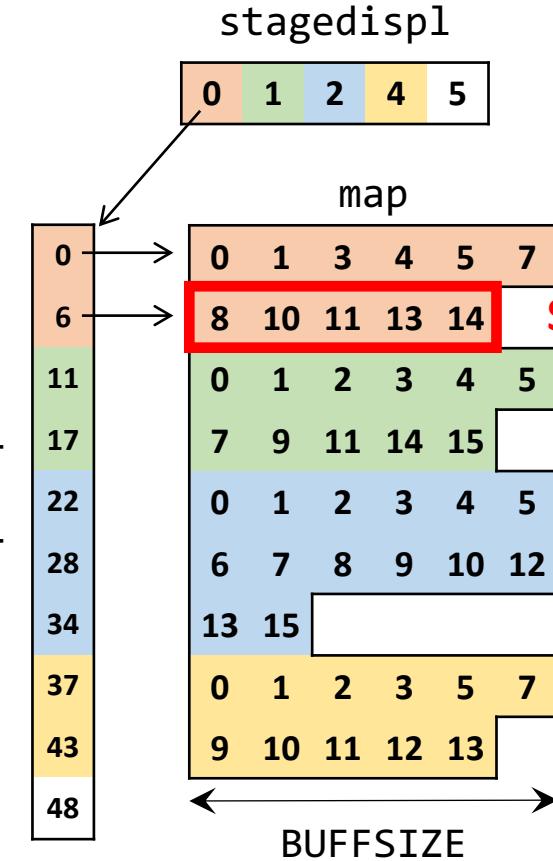
Tiling Data Structures



Input
Features

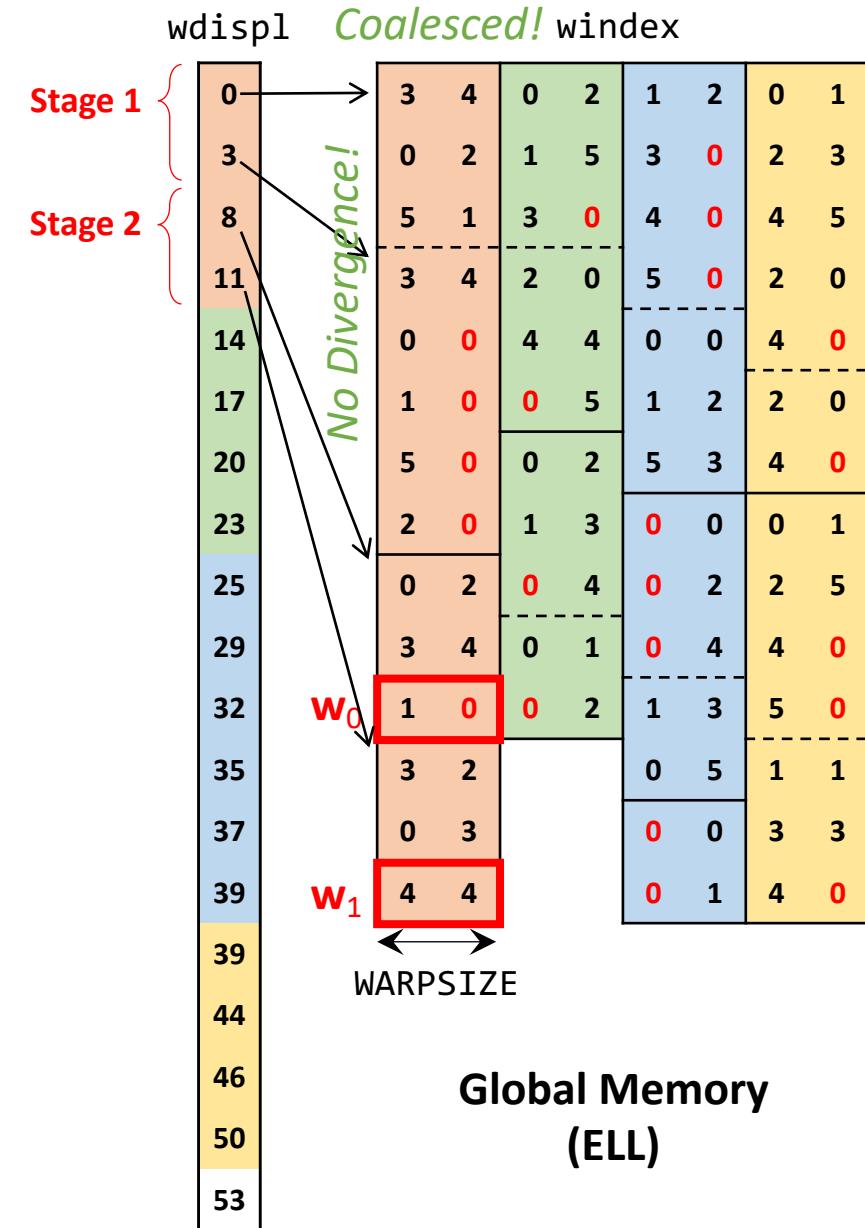
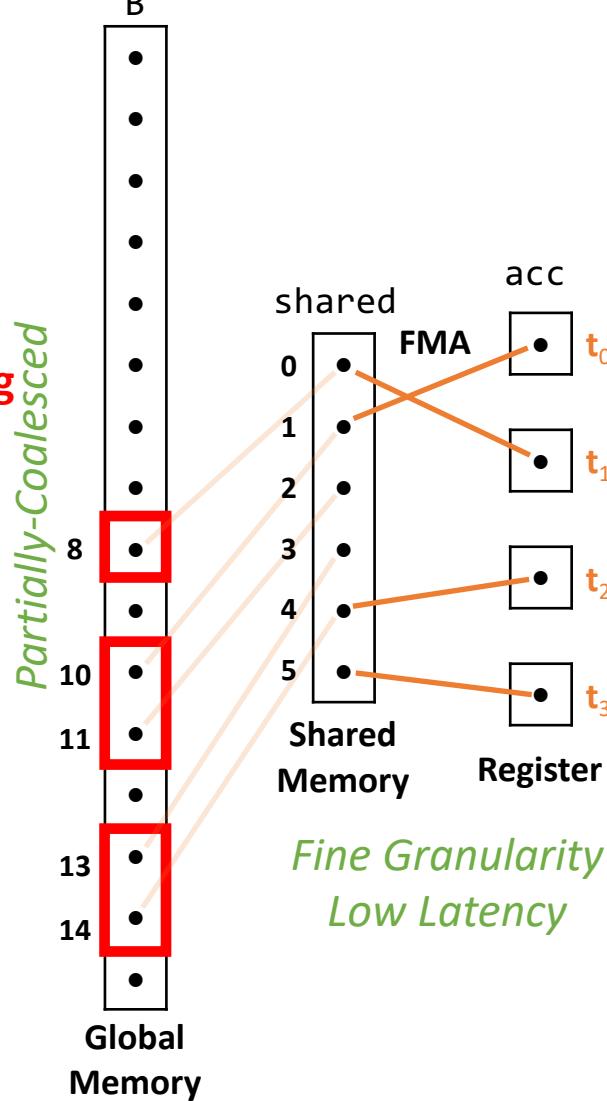


Tiling Data Structures

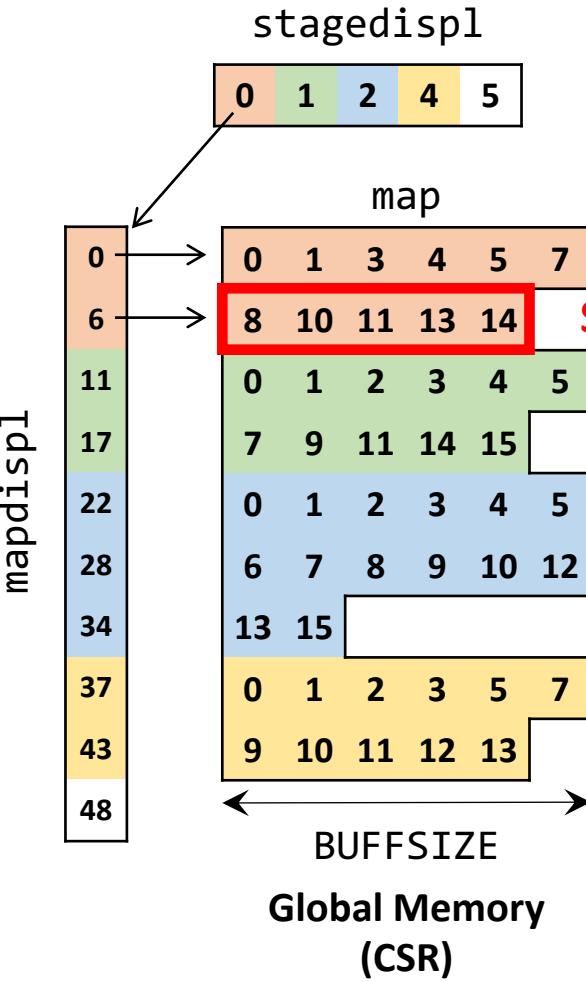


Global Memory
(CSR)

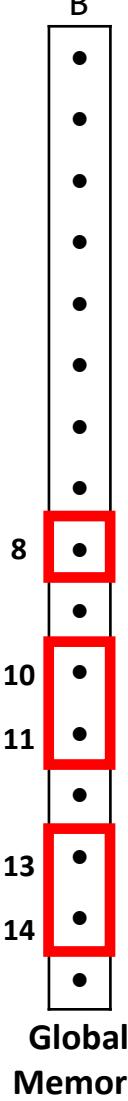
Input
Features



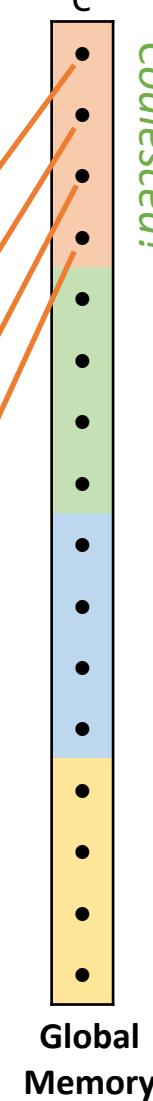
Tiling Data Structures



Input
Features



Output
Features



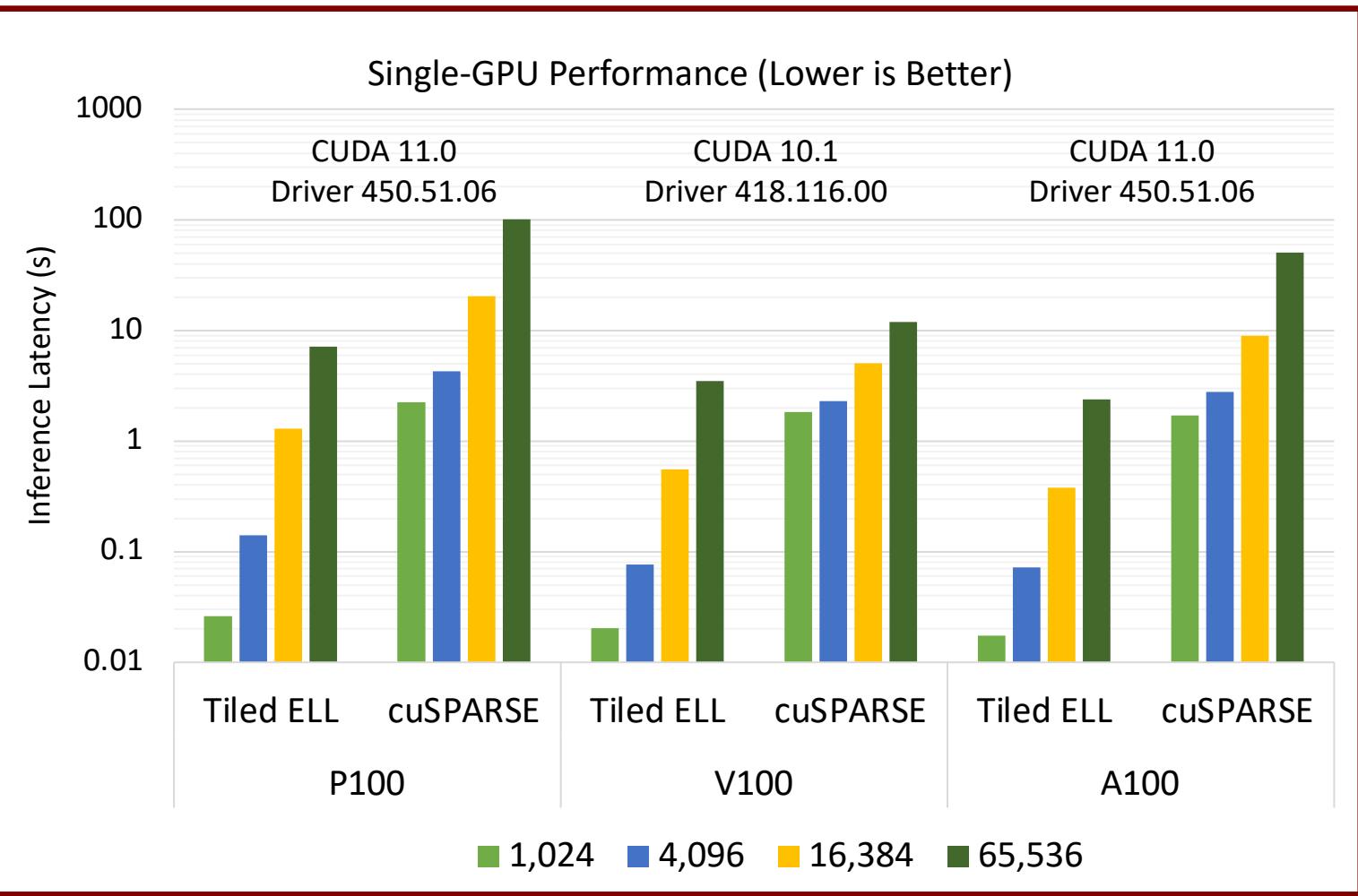
wdispl *Coalesced!* windex

The global memory is shown as a grid of values. The columns are grouped by `WARP SIZE`. The first column is labeled `wdispl`. The second column is labeled *No Divergence!*. The third column is labeled `Coalesced!`. The fourth column is labeled `windex`.

<code>wdispl</code>	3 4	0 2	1 2	0 1
0	0 1	1 5	3 0	2 3
3	5 1	3 0	4 0	4 5
8	3 4	2 0	5 0	2 0
11	0 0	4 4	0 0	4 0
14	1 0	0 5	1 2	2 0
17	5 0	0 2	5 3	4 0
20	2 0	1 3	0 0	0 1
23	0 2	0 4	0 2	2 5
25	3 4	0 1	0 4	4 0
29	1 0	0 2	1 3	5 0
32	3 2	0 1	0 5	1 1
35	0 3	0 2	0 0	3 3
37	4 4	0 1	0 1	4 0
39	3 2	0 0	1 3	5 0
44	0 3	0 2	0 5	1 1
46	4 4	0 1	0 1	4 0
50				
53				

Global Memory (ELL)

Application: Sparse Deep Neural Network



Application: Sparse Deep Neural Network

Inference Throughput (TeraEdges/Second)

Number of V100 GPUs (Six per Node)

Neurons	Layers	Single V100	Single A100	3	6	12	24	48	96	192	384	768
1024	120	10.51 (0.225s)	16.74 (1.59×)	18.92	22.46	25.52	28.52	27.77	29.17	27.89	29.12	29.13
	480	12.87 (0.073s)	20.99 (1.63×)	21.47	24.34	26.92	28.73	28.43	29.30	28.80	29.10	23.06
	1920	14.30 (0.264s)	20.68 (1.45×)	22.26	24.77	27.33	28.70	28.58	28.60	28.73	28.83	28.83
4096	120	9.45 (0.100s)	14.27 (1.51×)	20.69	31.36	47.82	62.03	70.31	75.81	79.11	81.13	82.20
	480	11.74 (0.322s)	18.63 (1.59×)	28.18	40.58	56.54	67.63	73.16	77.27	80.02	79.97	82.22
	1920	13.88 (1.08s)	19.86 (1.43×)	30.53	44.48	62.74	72.57	73.72	76.25	79.99	80.67	82.32
16384	120	6.15 (0.614s)	11.60 (1.89×)	16.31	28.85	50.74	64.33	89.18	111.44	146.88	114.87	111.30
	480	7.45 (2.027s)	14.31 (1.92×)	19.82	32.88	50.83	71.45	95.78	112.61	138.62	138.30	139.44
	1920	7.84 (7.704s)	15.27 (1.95×)	20.86	33.62	57.08	77.73	104.83	120.63	146.11	146.30	146.40
65536	120	3.47 (4.352s)	8.15 (2.35×)	10.90	18.77	34.20	51.14	73.67	100.72	162.19	173.25	179.58
	480	3.83 (15.769s)	9.08 (2.37×)	12.13	20.39	37.63	56.66	75.29	108.06	166.15	170.26	169.30
	1920	3.93 (61.474s)	9.33 (2.37×)	12.47	20.88	38.81	58.08	77.55	112.01	167.43	170.06	171.37

Application: Sparse Deep Neural Network



GraphChallenge



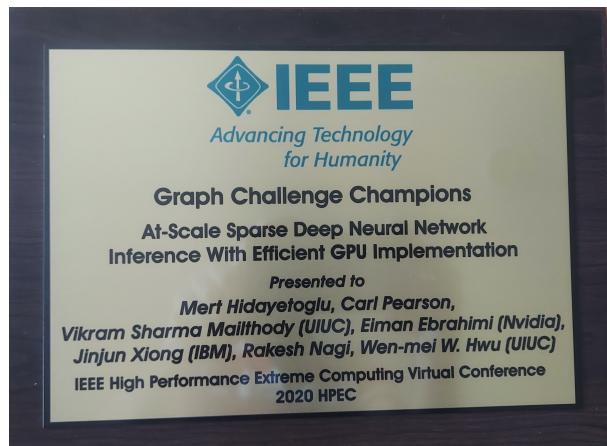
center for
cognitive computing
systems research



Graph Challenge Champions

2020 Champions

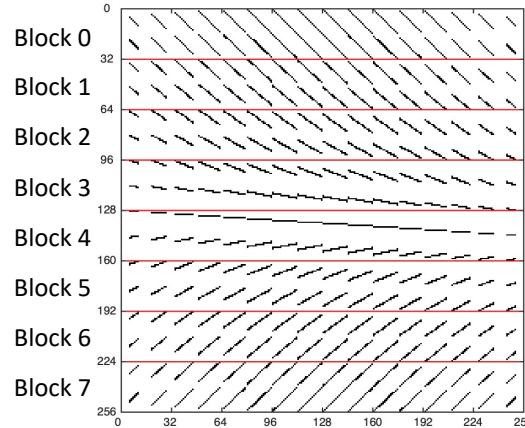
- *Scaling Graph Clustering with Distributed Sketches* - Benjamin Priest (LLNL), Alec Dunton (CU Boulder), Geoffrey Sanders (LLNL)
- *At-Scale Sparse Deep Neural Network Inference With Efficient GPU Implementation* - Mert Hidayetoglu, Carl Pearson, Vikram Sharma Mailthody (UIUC), Eiman Ebrahimi (Nvidia), Jinjun Xiong (IBM), Rakesh Nagi, Wen-mei W. Hwu (UIUC)
- *A Novel Inference Algorithm for Large Sparse Neural Network using Task Graph Parallelism* - Dian-Lun Lin, Tsung-Wei Huang (Univ of Utah)
- *TriC: Distributed-memory Triangle Counting by Exploiting the Graph Structure* - Sayan Ghosh, Mahantesh Halappanavar (PNNL)



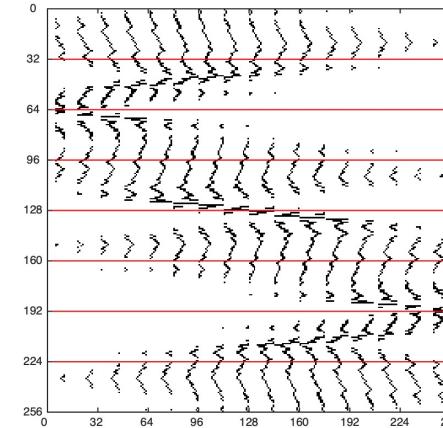
Open Source and Reproducible:
https://github.com/merthidayetoglu/SpDNN_Challenge2020

Open Problems: Enhancing Tiled SpMM by Matrix Reordering

Original Matrix



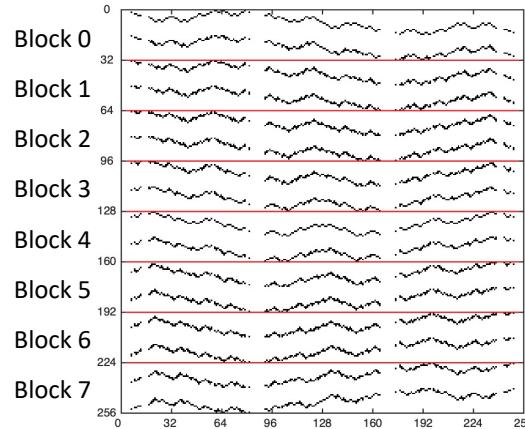
Row Reordering



Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

Data Reuse: 2.52

Column Reordering

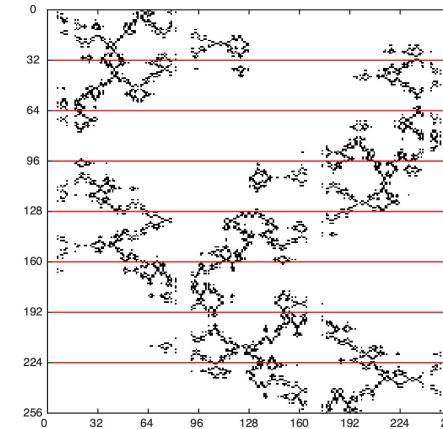


Data Reuse: 2.52

Row Reordering

Data Reuse: 4.44

Column Reordering

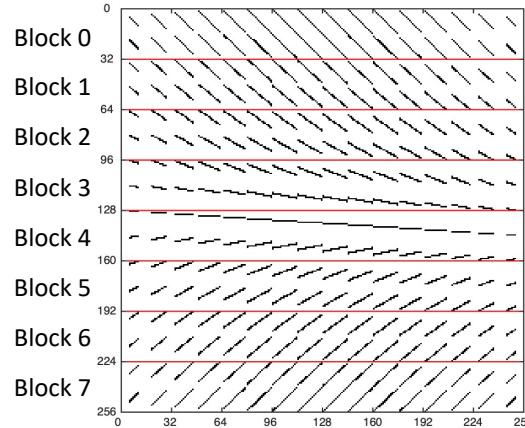


Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

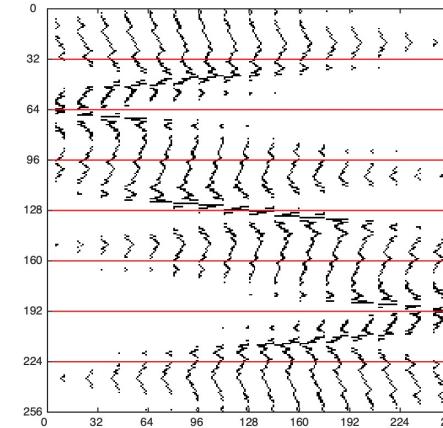
Data Reuse: 4.44

Open Problems: Enhancing Tiled SpMM by Matrix Reordering

Original Matrix



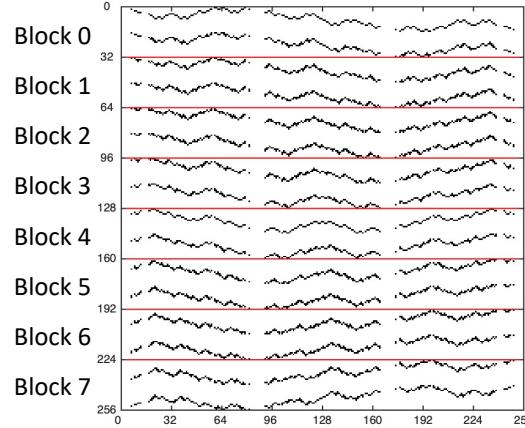
Row Reordering
→



Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

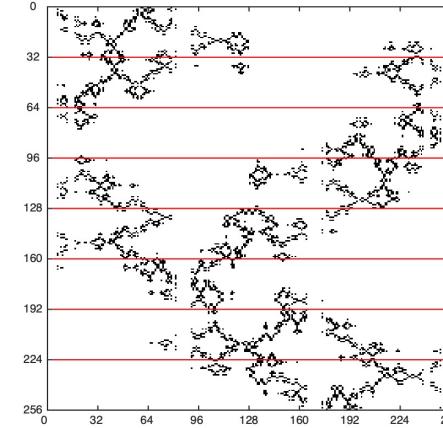
Data Reuse: 4.44

Column Reordering
↓



Data Reuse: 2.52

Row Reordering
→

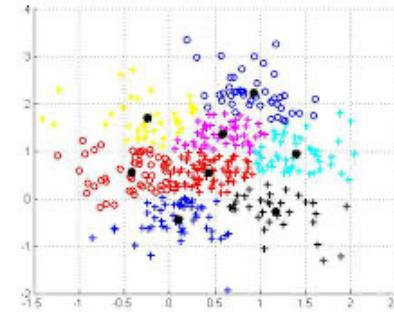


Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

Data Reuse: 4.44

Data Reuse: 2.52

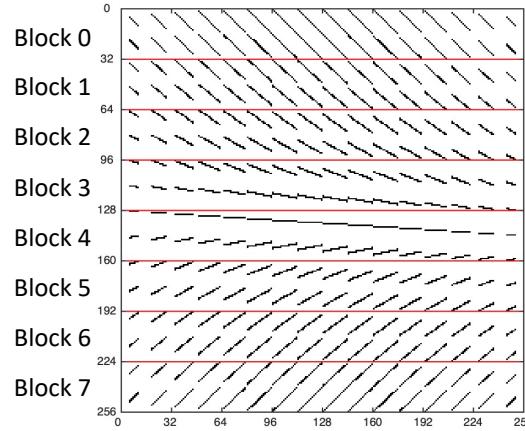
K-Means Clustering



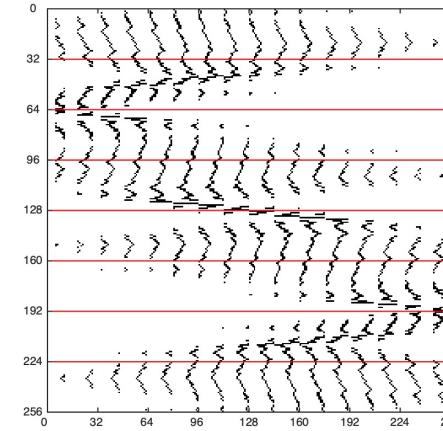
$O(kNM)$

Open Problems: Enhancing Tiled SpMM by Matrix Reordering

Original Matrix



Row Reordering

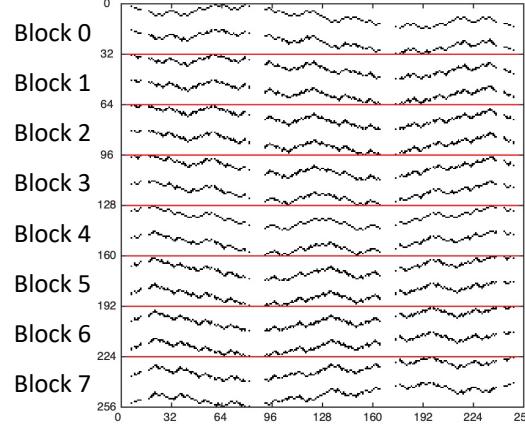


Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

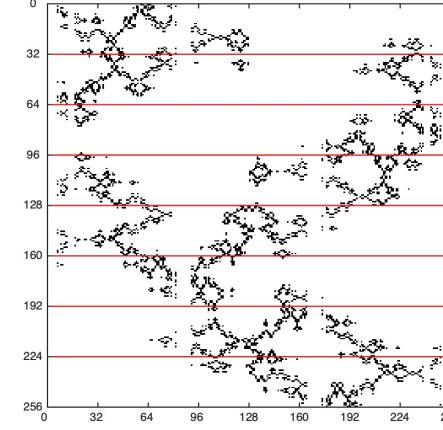
Data Reuse: 2.52

Column Reordering

Column Reordering



Row Reordering

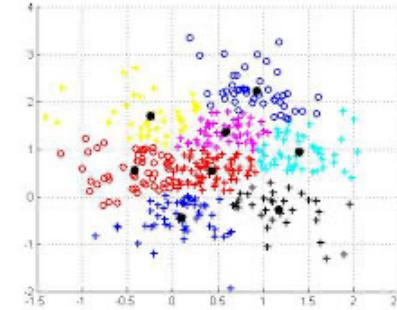


Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

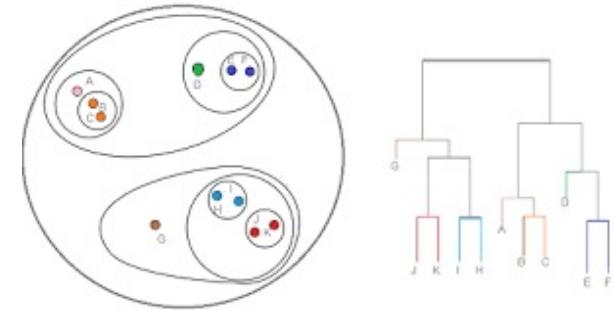
Data Reuse: 2.52

Data Reuse: 4.44

K-Means Clustering

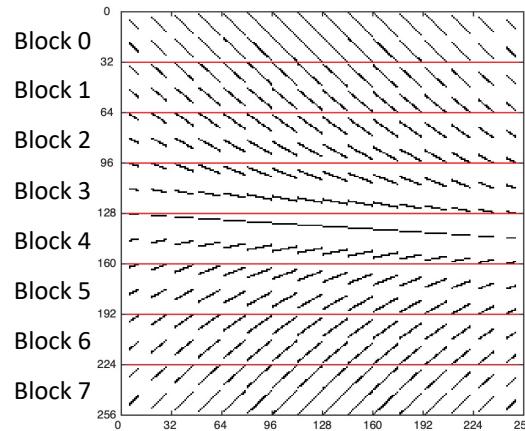


Hierarchical Clustering

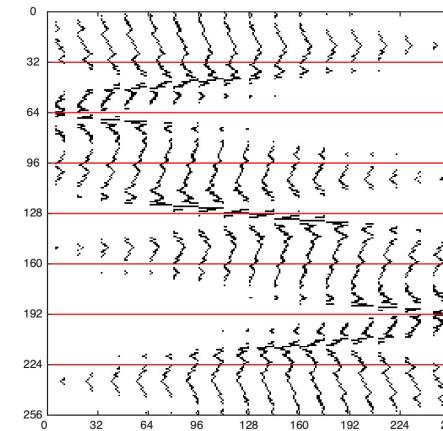


Open Problems: Enhancing Tiled SpMM by Matrix Reordering

Original Matrix



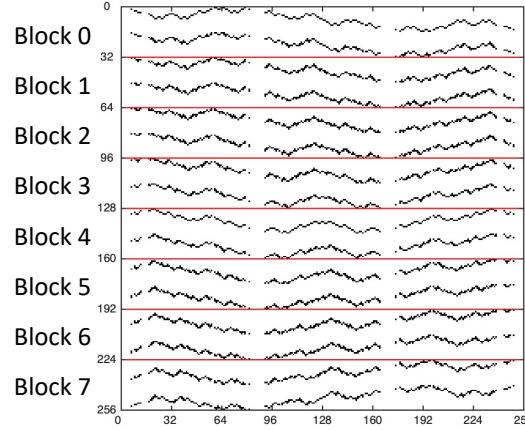
Row Reordering



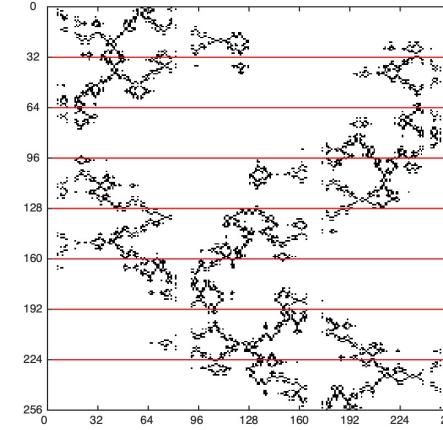
Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

Data Reuse: 4.44

Column Reordering

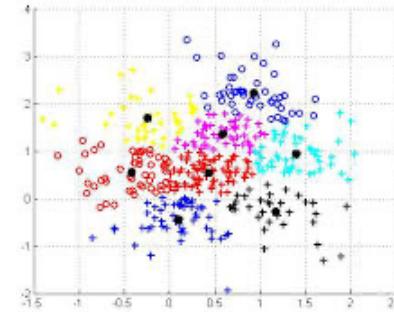


Row Reordering



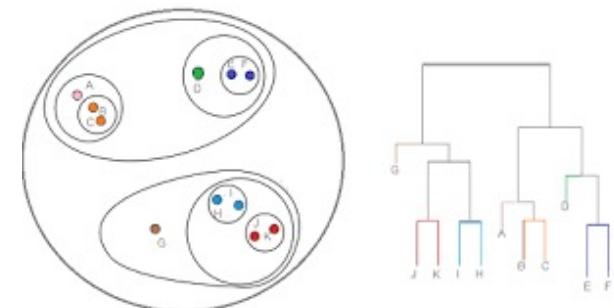
Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

K-Means Clustering

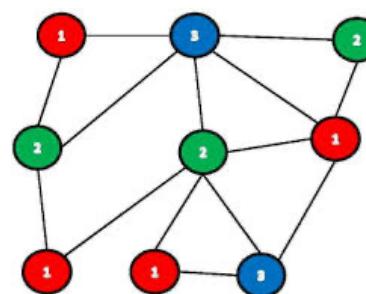


$O(kNM)$

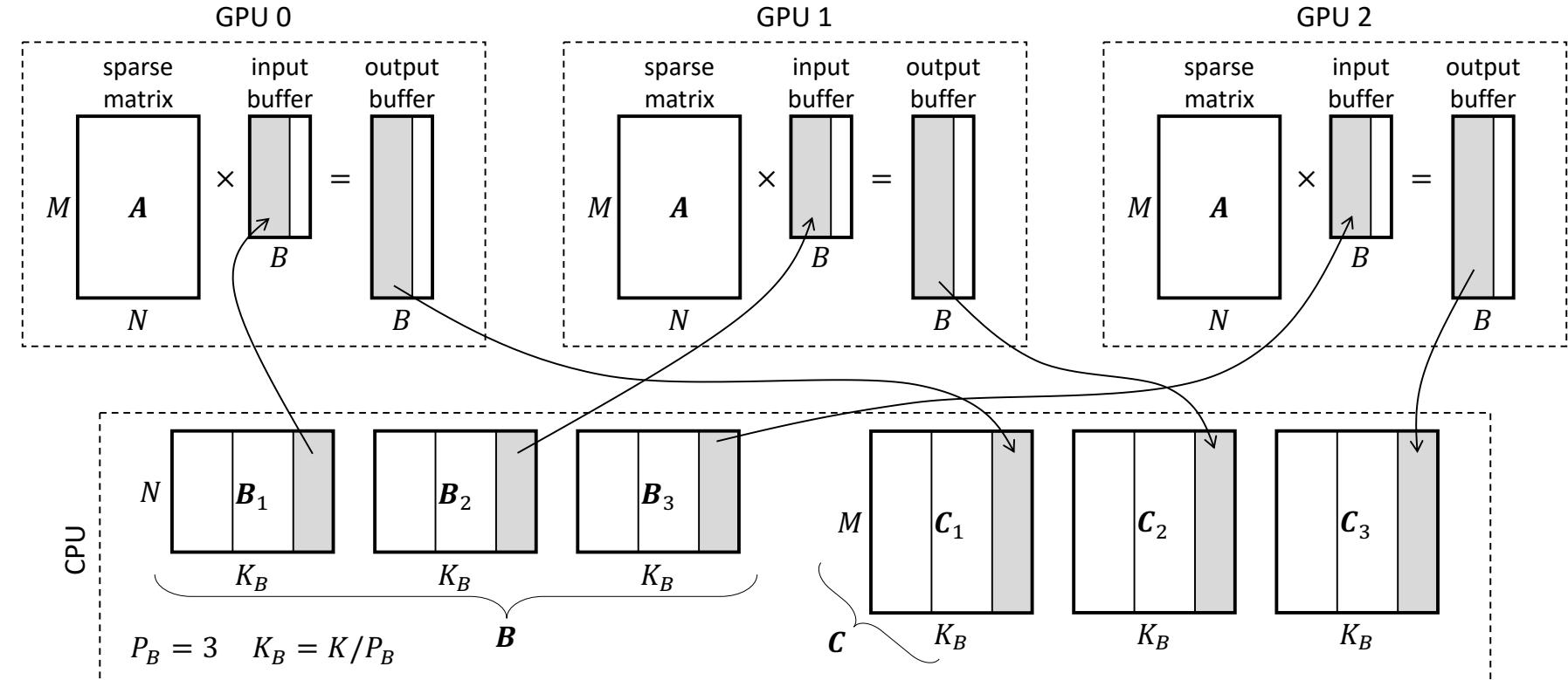
Hierarchical Clustering



Graph Coloring



Partitioning SpMM: Batch Parallelization & Streaming



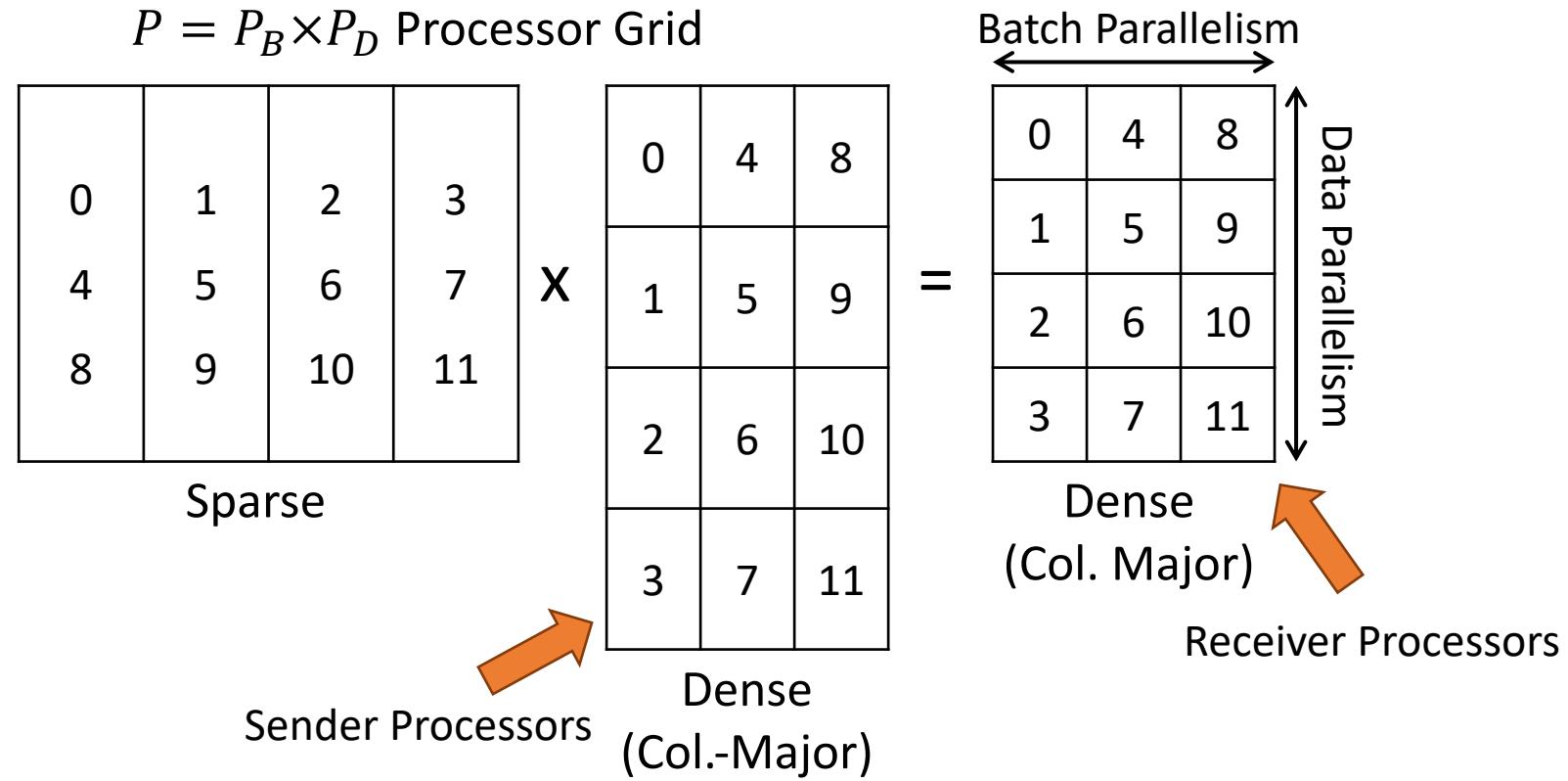
- Batch parallelism duplicates the partitions

Per-process memory

Total Memory ↑

Communications

Partitioning SpMM: Data Parallelization



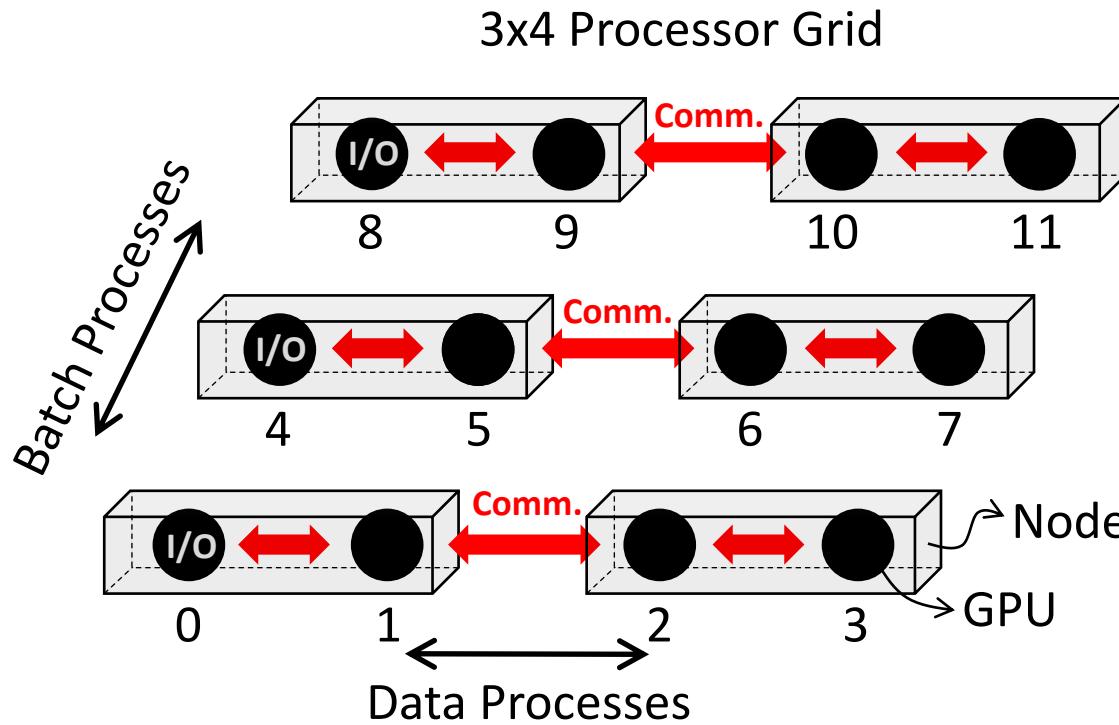
- **Batch parallelism duplicates the partitions**

Per-process memory — Total Memory ↑ Communications —

- **Data parallelism partitions the matrix**

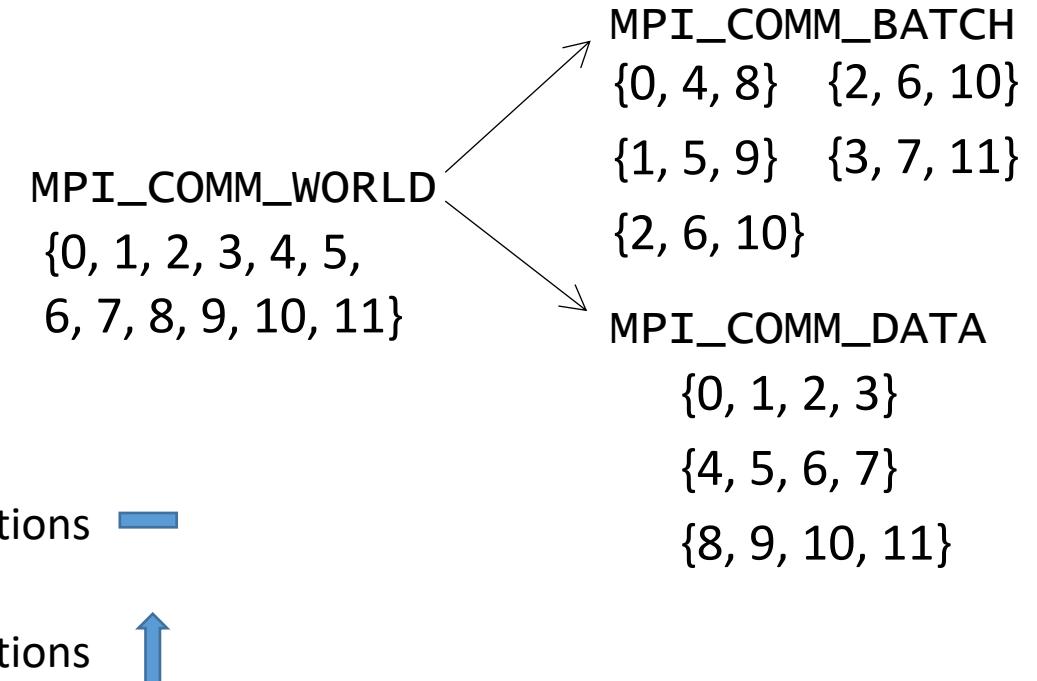
Per-process memory ↓ Total Memory — Communications ↑

Partitioning SpMM: Topology-Aware Rank Placement



Defining Orthogonal Communicators

```
MPI_Comm_split( MPI_Comm comm, int  
color, int key, MPI_Comm* newcomm)
```



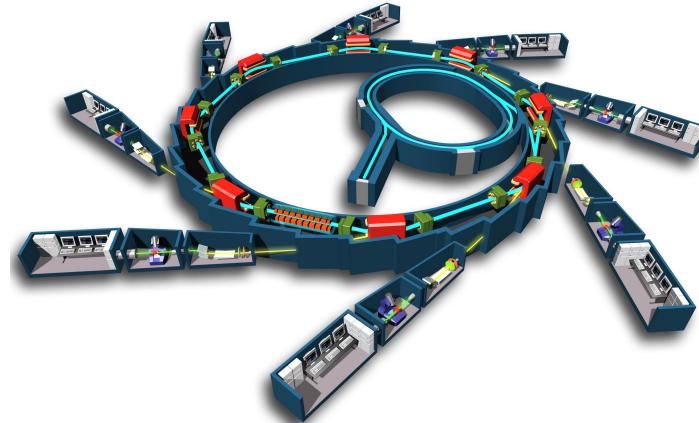
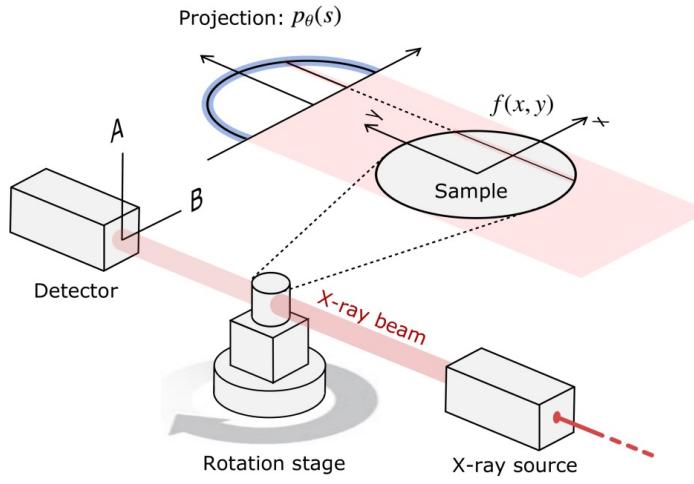
- **Batch parallelism duplicates the partitions**



- **Data parallelism partitions the matrix**

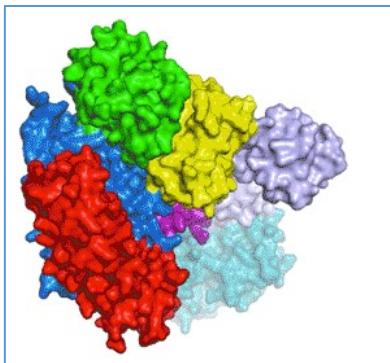


Application: X-Ray Imaging

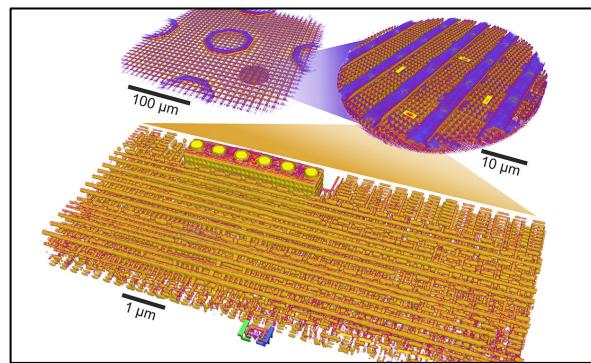


Soleil Light Source Beamlines
Paris, France

Molecular Reactions



Chip Imaging

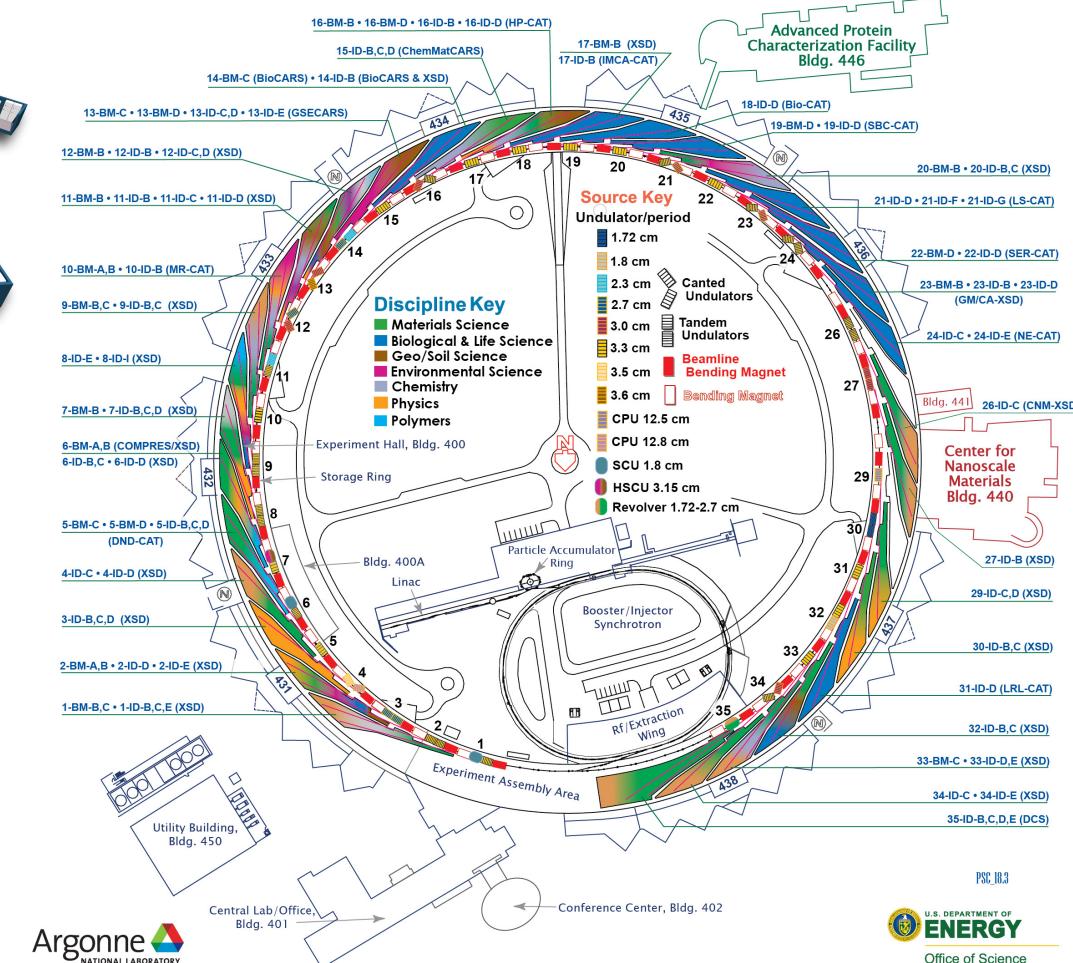


ARGONNE NATIONAL LABORATORY 400-AREA FACILITIES

ADVANCED PHOTON SOURCE

(Beamlines, Disciplines, and Source Configuration)

ADVANCED PROTEIN CHARACTERIZATION FACILITY CENTER FOR NANOSCALE MATERIALS



Application Performance: X-Ray Imaging

DOE Synchrotron Light Sources



Advanced Light Source
at Lawrence Berkeley
National Laboratory
Upgrade Scheduled 2024



Stanford Synchrotron Radiation
Light Source at SLAC National
Accelerator Laboratory
Upgraded 2015



Advanced Photon Source at
Argonne National Laboratory
Upgrade Scheduled 2022



National Synchrotron Light
Source II at Brookhaven
National Laboratory
Upgraded 2015

Diamond in Oxford, UK
Upgrade Scheduled 2024



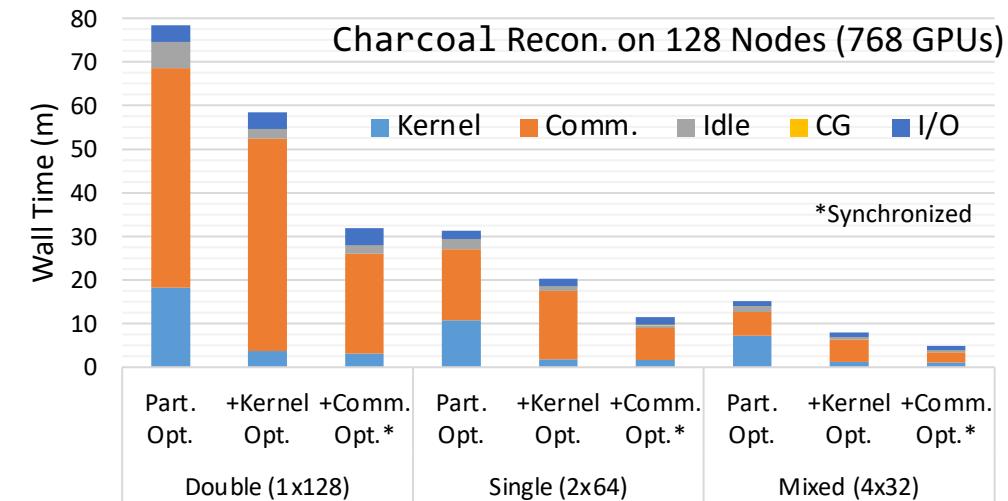
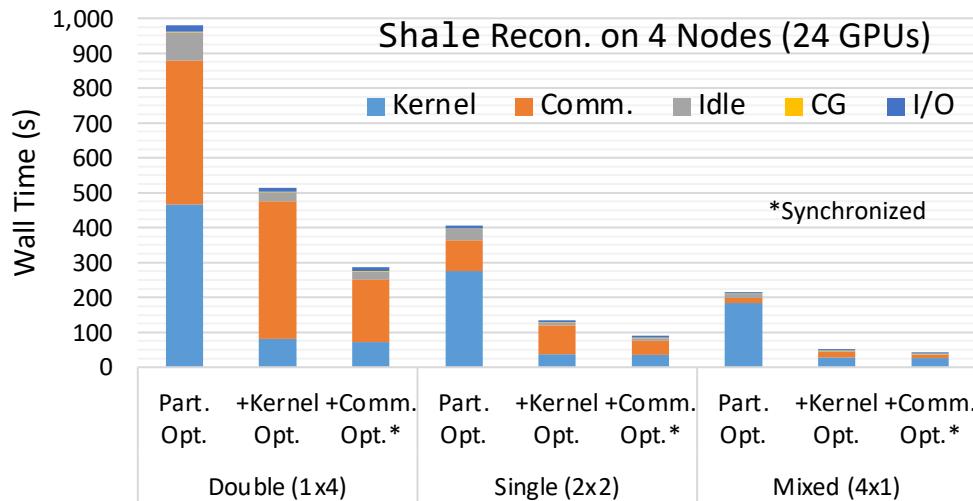
DESY in Hanofer, Germany
Upgrade Scheduled 2025



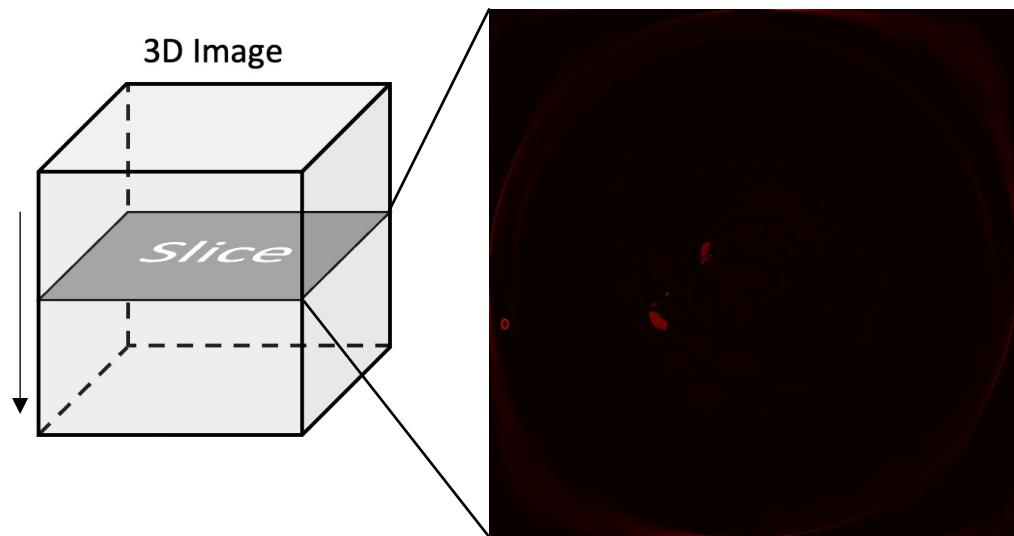
SPRING-8 in Sayo, Japan
Upgrade Scheduled TBD



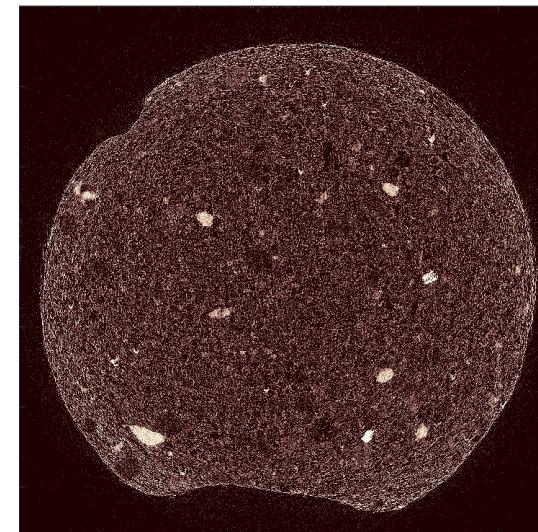
Communication Bottleneck



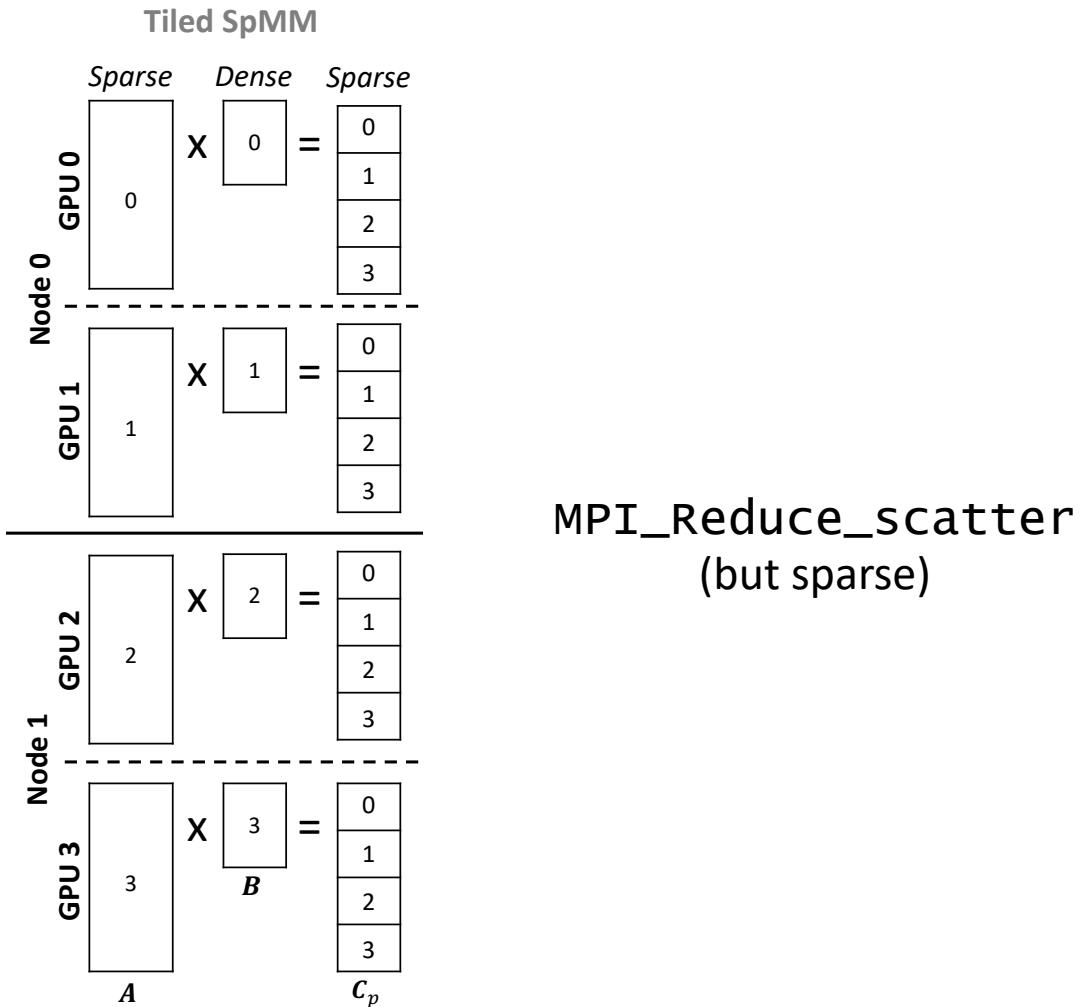
1792 x 2048 x 2048



4198 x 6613 x 6613

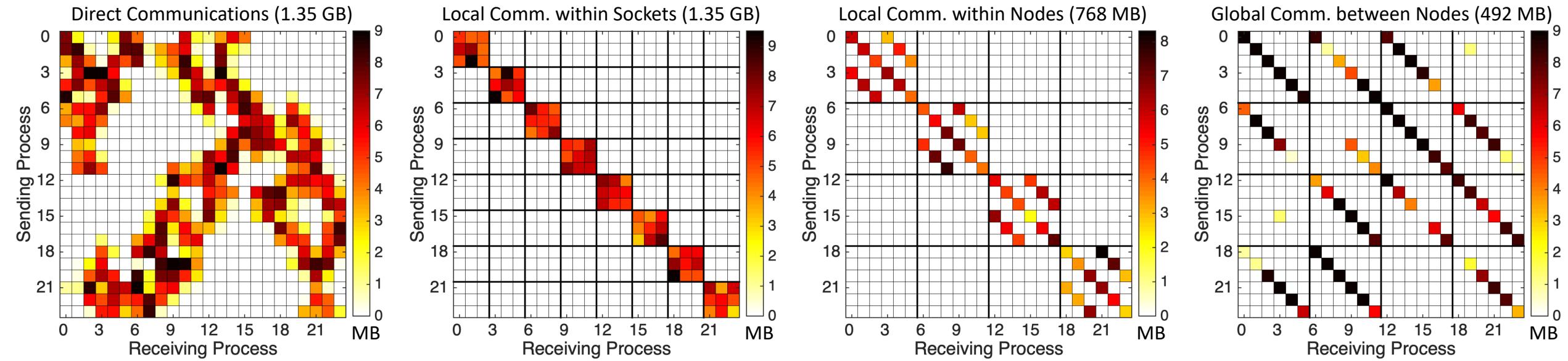
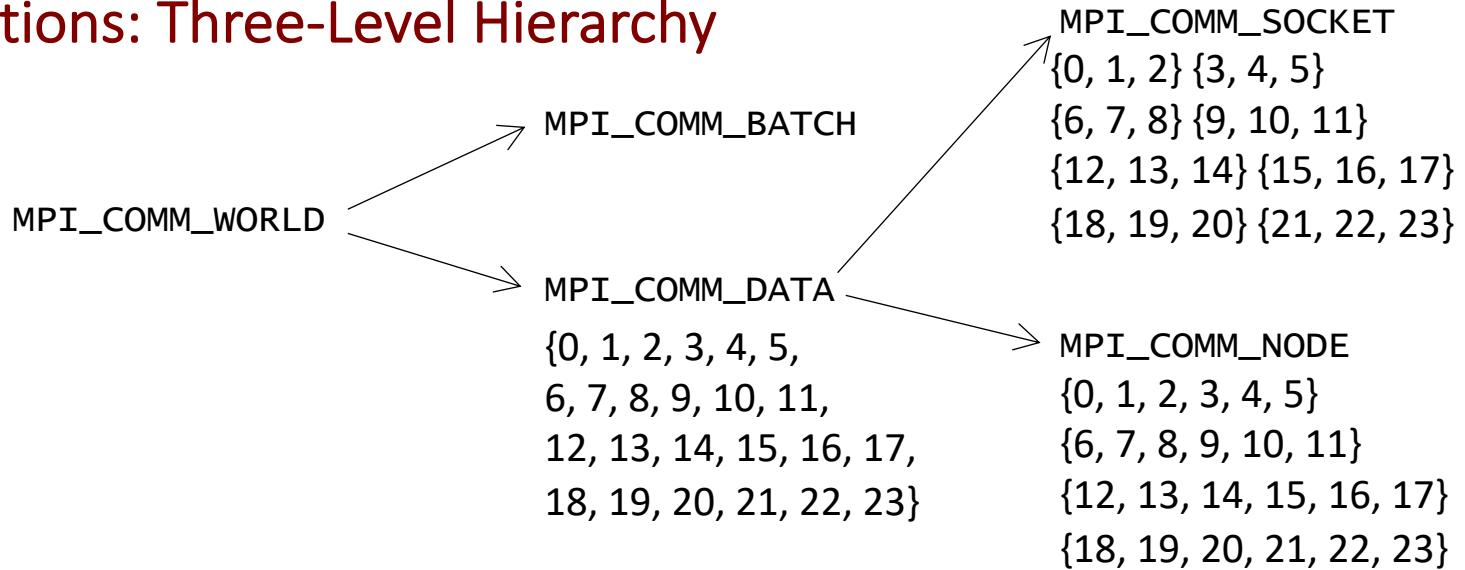


Hierarchical Communications: Overview

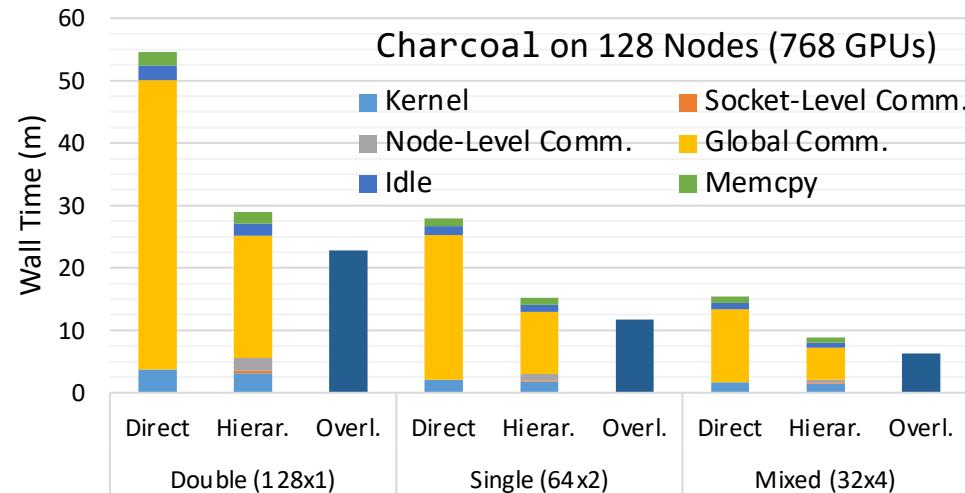


`MPI_Reduce_scatter`
(but sparse)

Hierarchical Communications: Three-Level Hierarchy



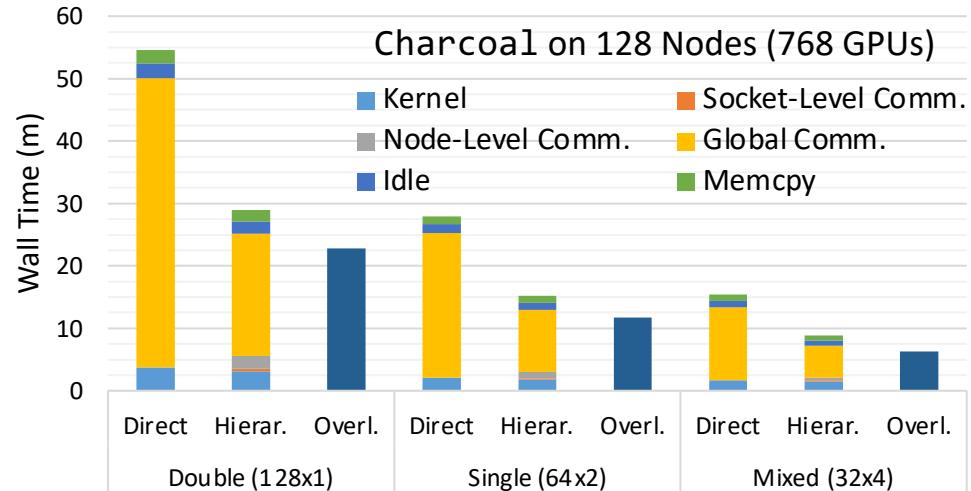
Hierarchical Communications: Benchmarking on Summit



Communication Volume and Effective Bandwidth

	Prec.	Socket-Level Comm.		Node-Level Comm.		Global Comm.		Memcpy	
		Data	B/W	Data	B/W	Data	B/W	Data	B/W
Direct	Double					36.6 TB	1.61 TB/s	35.2 TB/s	
	Single	N/A		N/A		18.3 TB	1.61 TB/s	34.9 TB/s	
	Mixed					9.16 TB	1.59 TB/s	34.6 TB/s	
Hierar.	Double	36.6 TB	174 TB/s	21.4 TB	21.3 TB/s	15.2 TB	1.58 TB/s	34.9 TB/s	
	Single	18.3 TB	170 TB/s	10.7 TB	22.8 TB/s	7.58 TB	1.55 TB/s	34.5 TB/s	
	Mixed	9.16 TB	164 TB/s	5.35 TB	23.5 TB/s	3.79 TB	1.49 TB/s	33.6 TB/s	

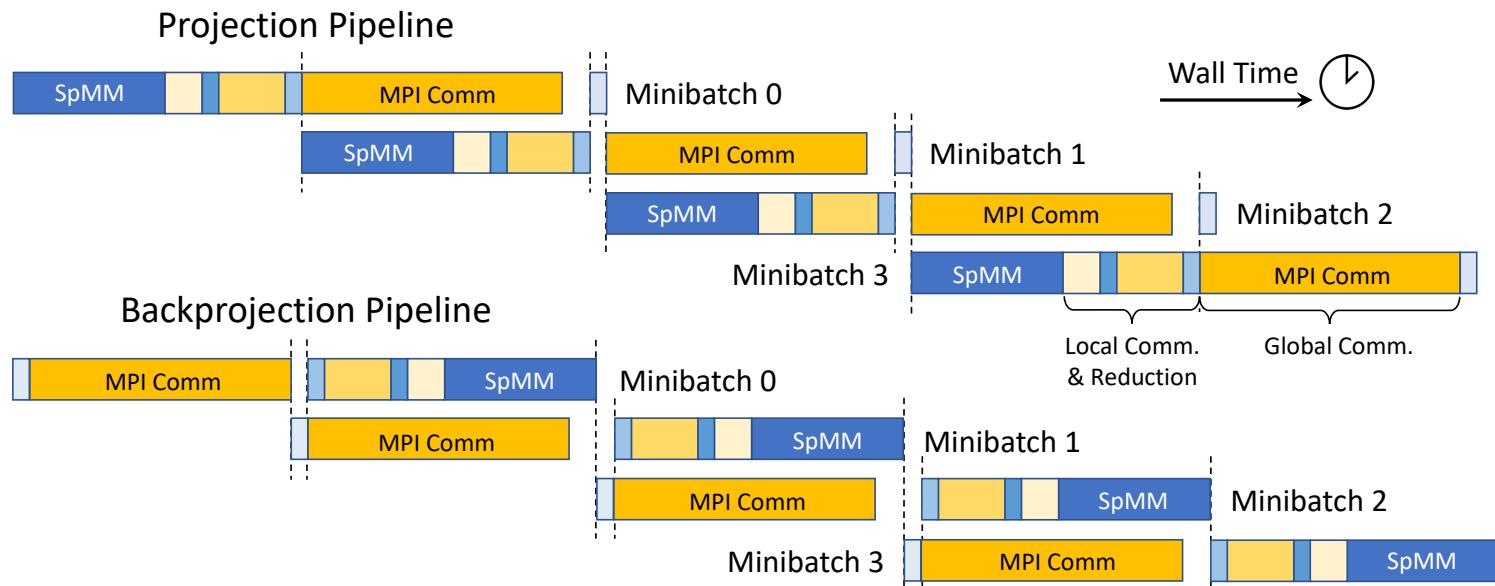
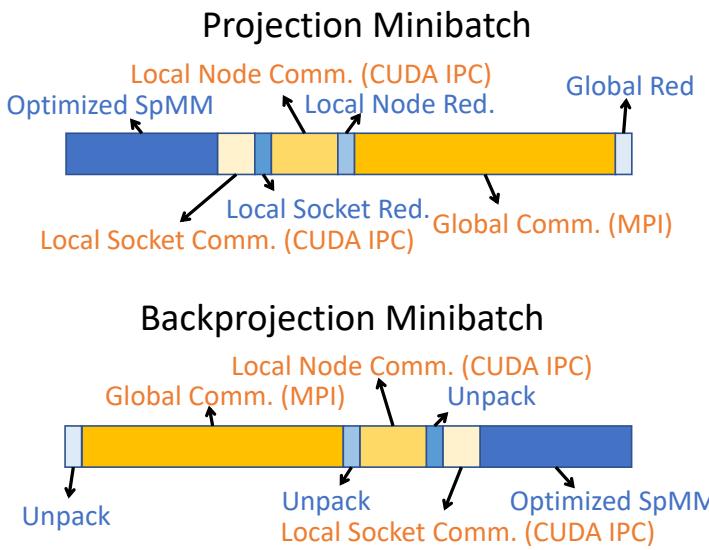
Hierarchical Communications: Benchmarking on Summit



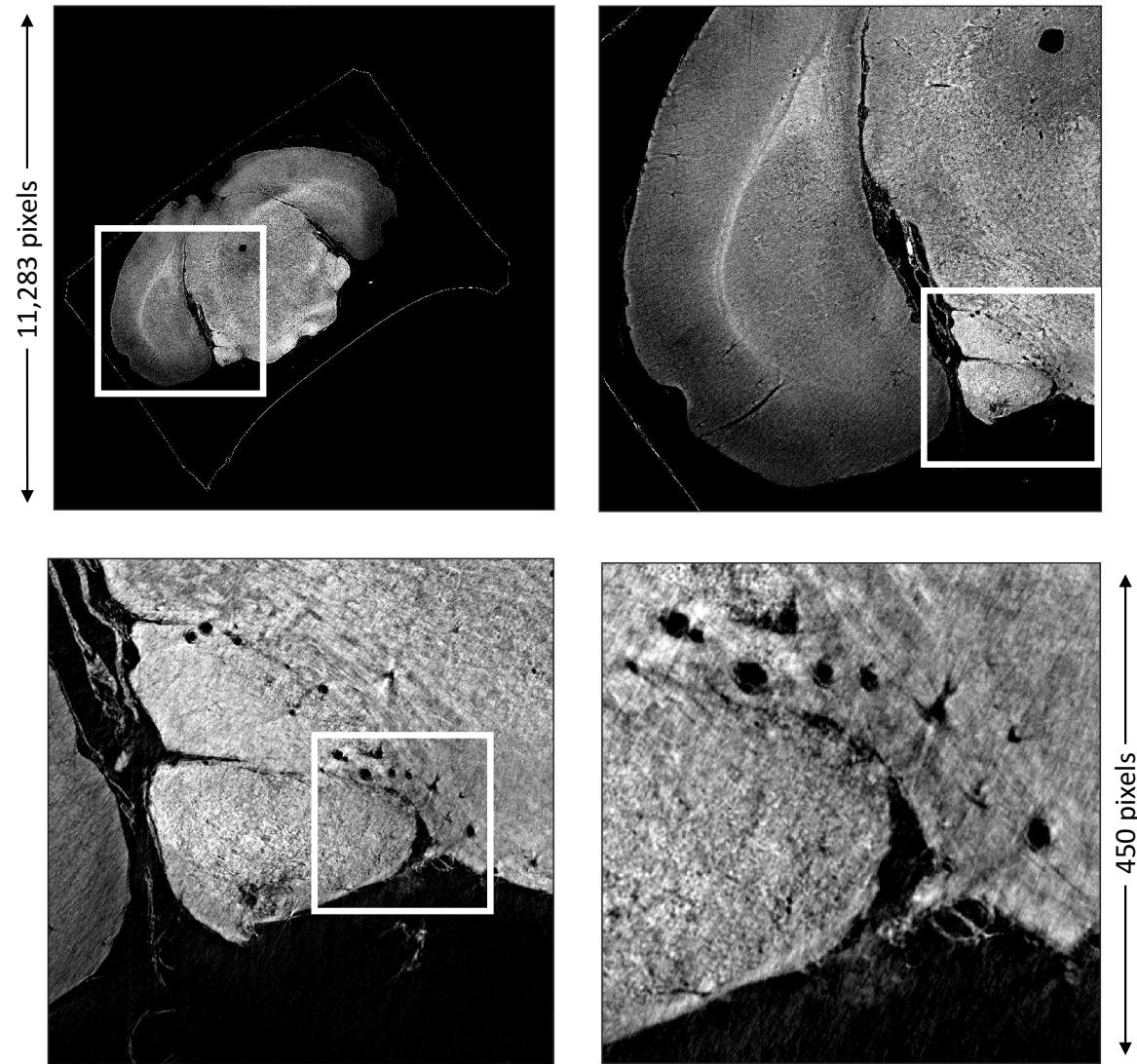
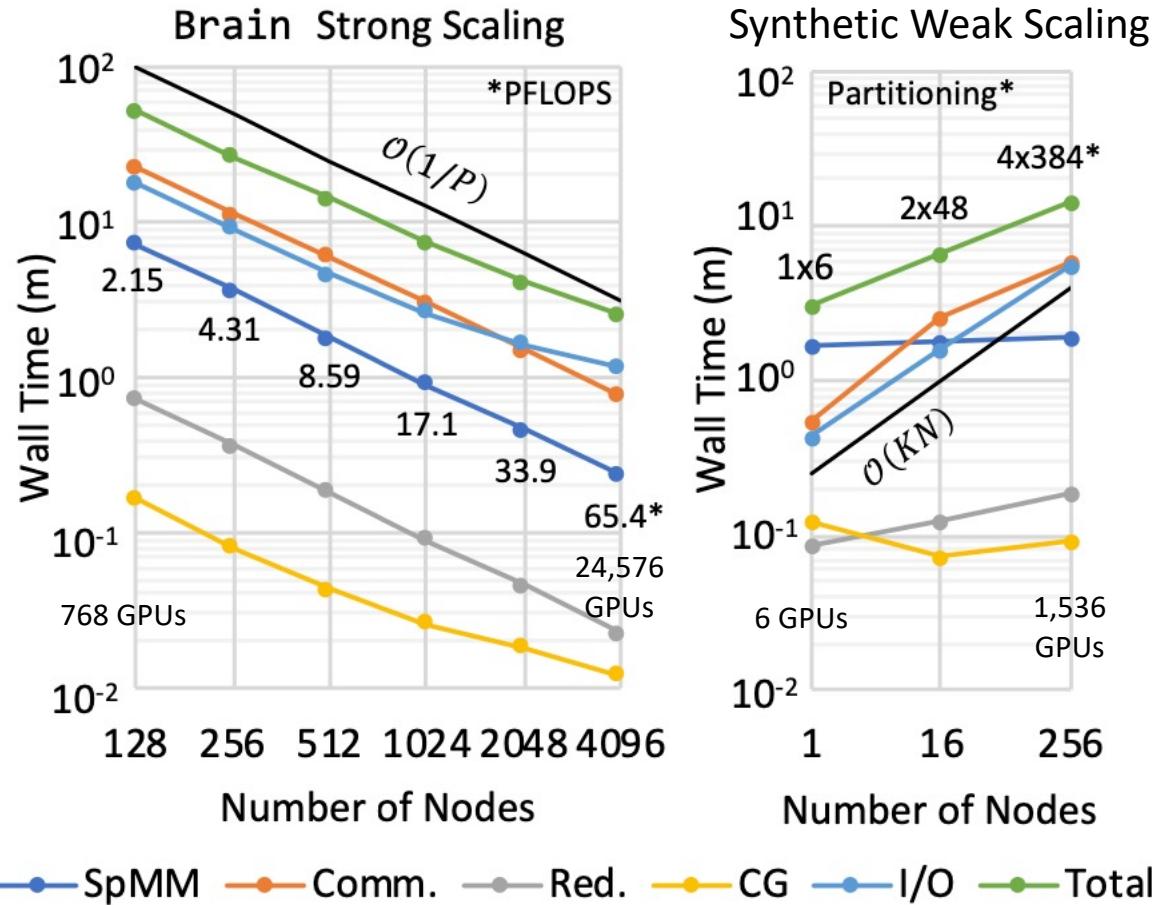
Communication Volume and Effective Bandwidth

	Prec.	Socket-Level Comm.		Node-Level Comm.		Global Comm.		Memcpy
		Data	B/W	Data	B/W	Data	B/W	B/W
Direct	Double					36.6 TB	1.61 TB/s	35.2 TB/s
	Single	N/A		N/A		18.3 TB	1.61 TB/s	34.9 TB/s
	Mixed					9.16 TB	1.59 TB/s	34.6 TB/s
Hierar.	Double	36.6 TB	174 TB/s	21.4 TB	21.3 TB/s	15.2 TB	1.58 TB/s	34.9 TB/s
	Single	18.3 TB	170 TB/s	10.7 TB	22.8 TB/s	7.58 TB	1.55 TB/s	34.5 TB/s
	Mixed	9.16 TB	164 TB/s	5.35 TB	23.5 TB/s	3.79 TB	1.49 TB/s	33.6 TB/s

Pipelining Batch Processing



Hierarchical Communications: Scaling on Summit



We thank Bobby Kasthuri of UChicago/Argonne, Rafael Vescovi and Ming Du of Argonne for sharing the mouse brain dataset.

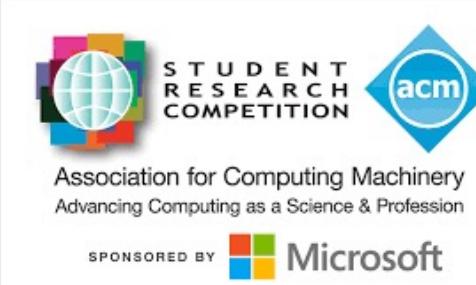
Recognition: Best Paper @ SC20

SC20 Student Cluster Reproducibility Committee Chooses Benchmark Wisely

April 15, 2020

<https://github.com/merthidayetoglu/MemXCT-GPU>

We are excited to announce that the SC20 Reproducibility Committee has selected the SC19 paper "MemXCT: Memory-Centric X-ray CT Reconstruction with Massive Parallelization", by Mert Hidayetoğlu, Tekin Biçer, Simon Garcia de Gonzalo, Bin Ren, Doğa Gürsoy, Rajkumar Kettimuthu, Ian T. Foster, and Wen-mei W. Hwu, to serve as the Student Cluster Competition (SCC) benchmark for the Reproducibility Challenge this year. A team of reviewers selected the paper from 45 finalists, based on the paper's Artifact Descriptor (AD) and its suitability to the SCC. The authors and the Reproducibility Committee have been working to create a reproducible benchmark that builds on the paper's results. At SC20, the sixteen SCC teams will be asked to run the benchmark, replicating the findings from the original paper under different settings and with different datasets.



ACM STUDENT RESEARCH COMPETITION SC20 GRADUATE – 1ST PLACE WINNER



Memory-Centric 3D Image Reconstruction with Hierarchical Communications on Multi-GPU Node Architecture

Mert Hidayetoglu
University of Illinois

BEST PAPER - WINNER

Petascale XCT: 3D Image Reconstruction with Hierarchical Communications on Multi-GPU Nodes

Mert Hidayetoglu, Tekin Bicer, Simon Garcia de Gonzalo, Bin Ren,
Vincent De Andrade, Doga Gursoy, Rajkumar Kettimuthu,
Ian T. Foster, Wen-mei W. Hwu

University of Illinois, Argonne National Laboratory,
Barcelona Supercomputing Center, College of William and Mary



ACM SIGHPC CERTIFICATE OF APPRECIATION

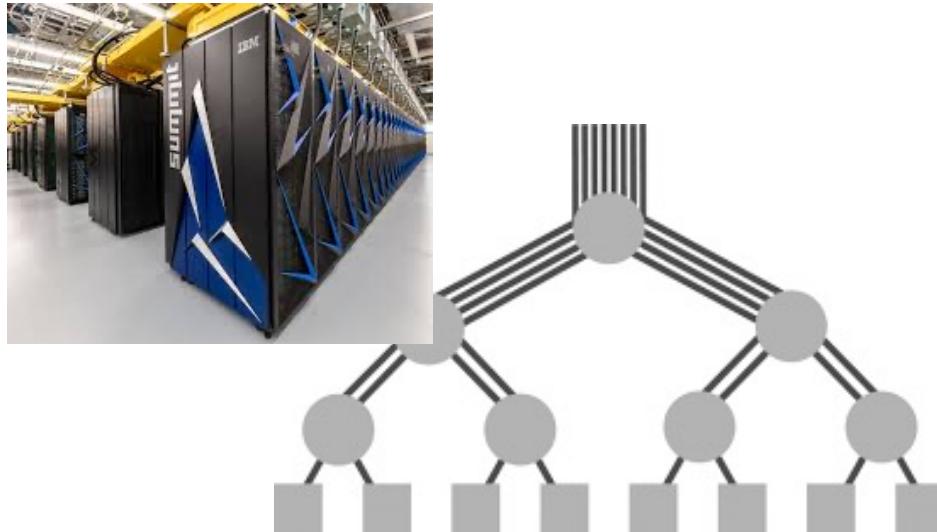
MemXCT: Memory-Centric X-ray CT Reconstruction with Massive Parallelization

Mert Hidayetoglu, Tekin Bicer, Simon Garcia de Gonzalo,
Bin Ren, Doga Gursoy, Rajkumar Kettimuthu,
Ian T. Foster, Wen-mei W. Hwu

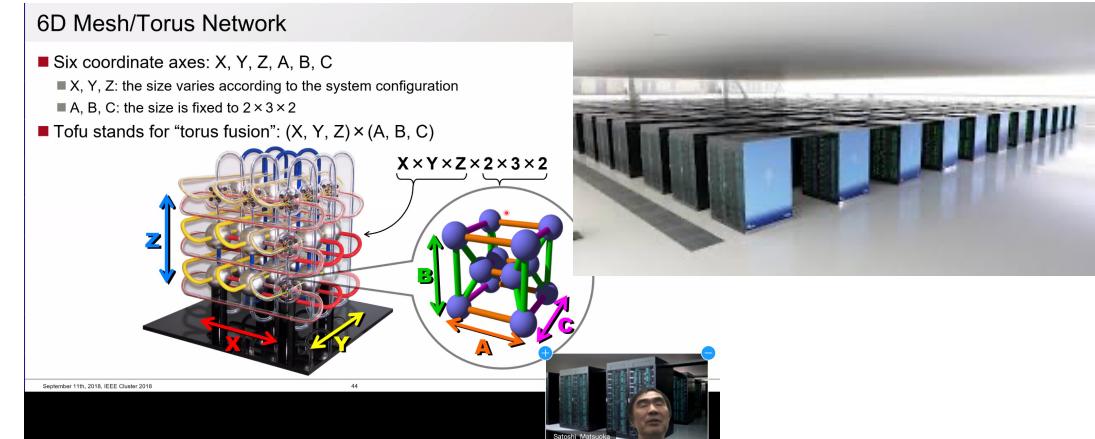


Open Problems: Extending Comm. Hierarchy for Inter-Node Communication

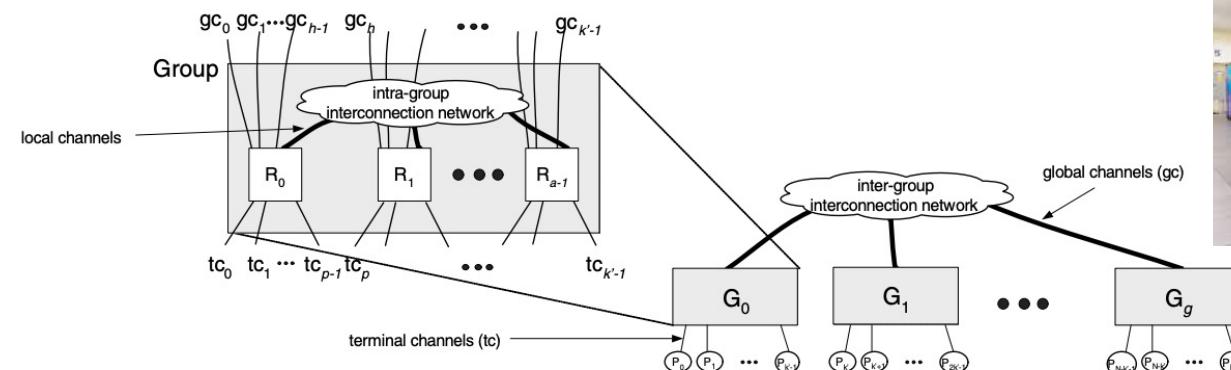
Fat-Tree: OLCF Summit



Tofu: RIKEN Fugaku



Dragonfly: ALCF Theta



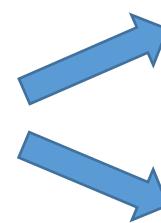
Conclusions

Large-Scale Sparse Applications

Conclusions

Large-Scale Sparse Applications

Multi-GPU Node Architecture



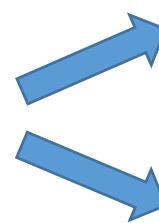
Single GPU: Tiled SpMM

GPU Cluster: Hierarchical Comm.

Conclusions

Large-Scale Sparse Applications

Multi-GPU Node Architecture



Single GPU: Tiled SpMM

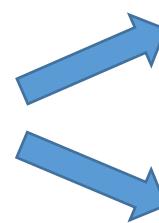
GPU Cluster: Hierarchical Comm.

Possible Contributions: cuSPARSE, MPI Collectives, NCCL

Conclusions

Large-Scale Sparse Applications

Multi-GPU Node Architecture



Single GPU: Tiled SpMM

GPU Cluster: Hierarchical Comm.

Possible Contributions: cuSPARSE, MPI Collectives, NCCL

Further Dissertation Research

