ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

SoftHier Progress Update

Chi Zhang chizhang@iis.ee.ethz.ch

PULP Platform

Open Source Hardware, the way it should be!

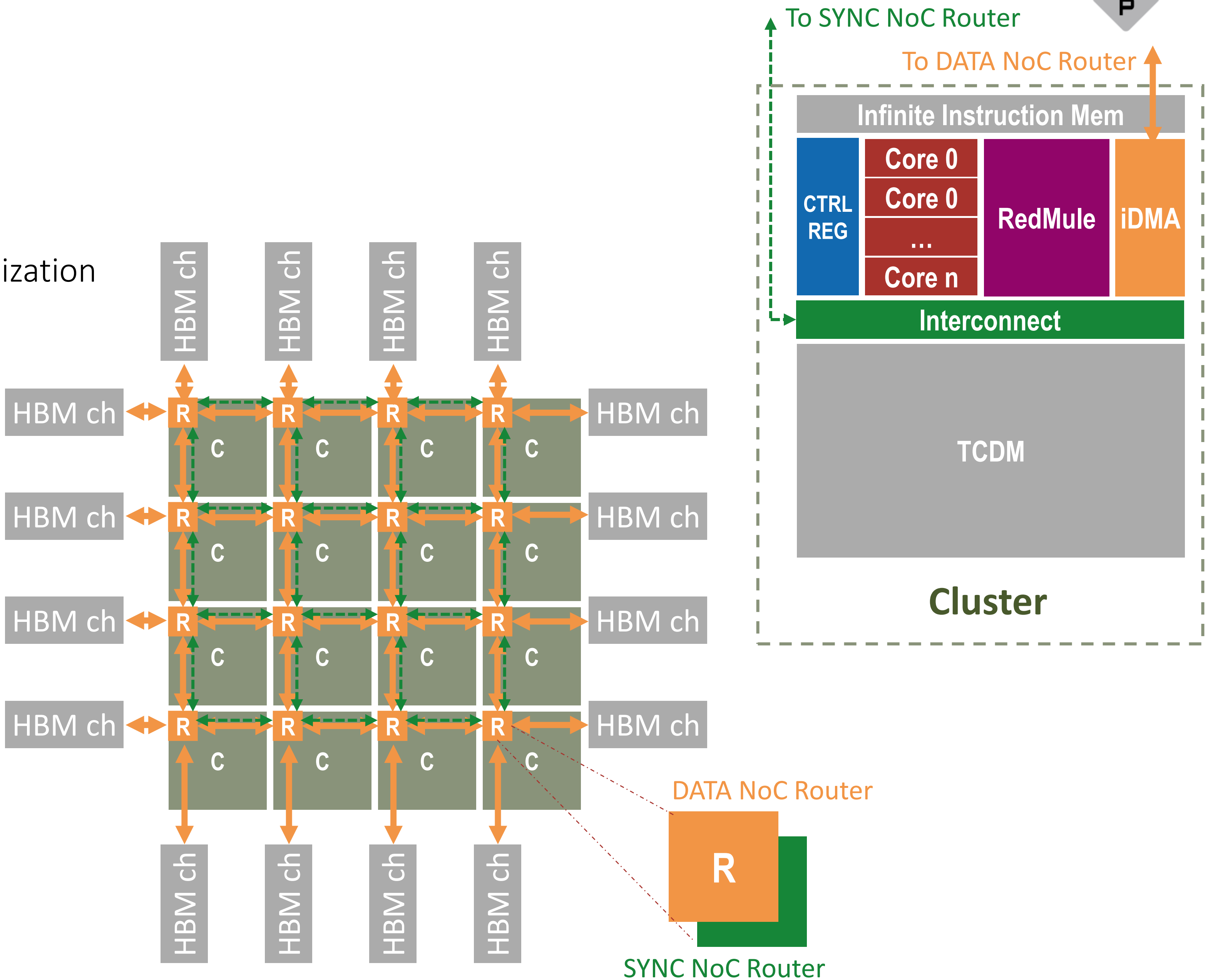
@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

SoftHier: Parameterizable NoC-Based Scalable System

- Features
  - Two separate NoC bus
    - DATA NoC: wide link, transfer bulk data
    - SYNC NoC: narrow link, cluster synchronization
  - Infinite Instruction memory
    - Ignore I$ fetch overhead
- Fully Parameterizable
  - Configure file and push button
    - #Cores, RedMule config
    - L1 (TCDM) size & BW
    - #Clusters (#row, #col)
    - #HBM channels and placement
    - NoC link BW
- SW stack ready

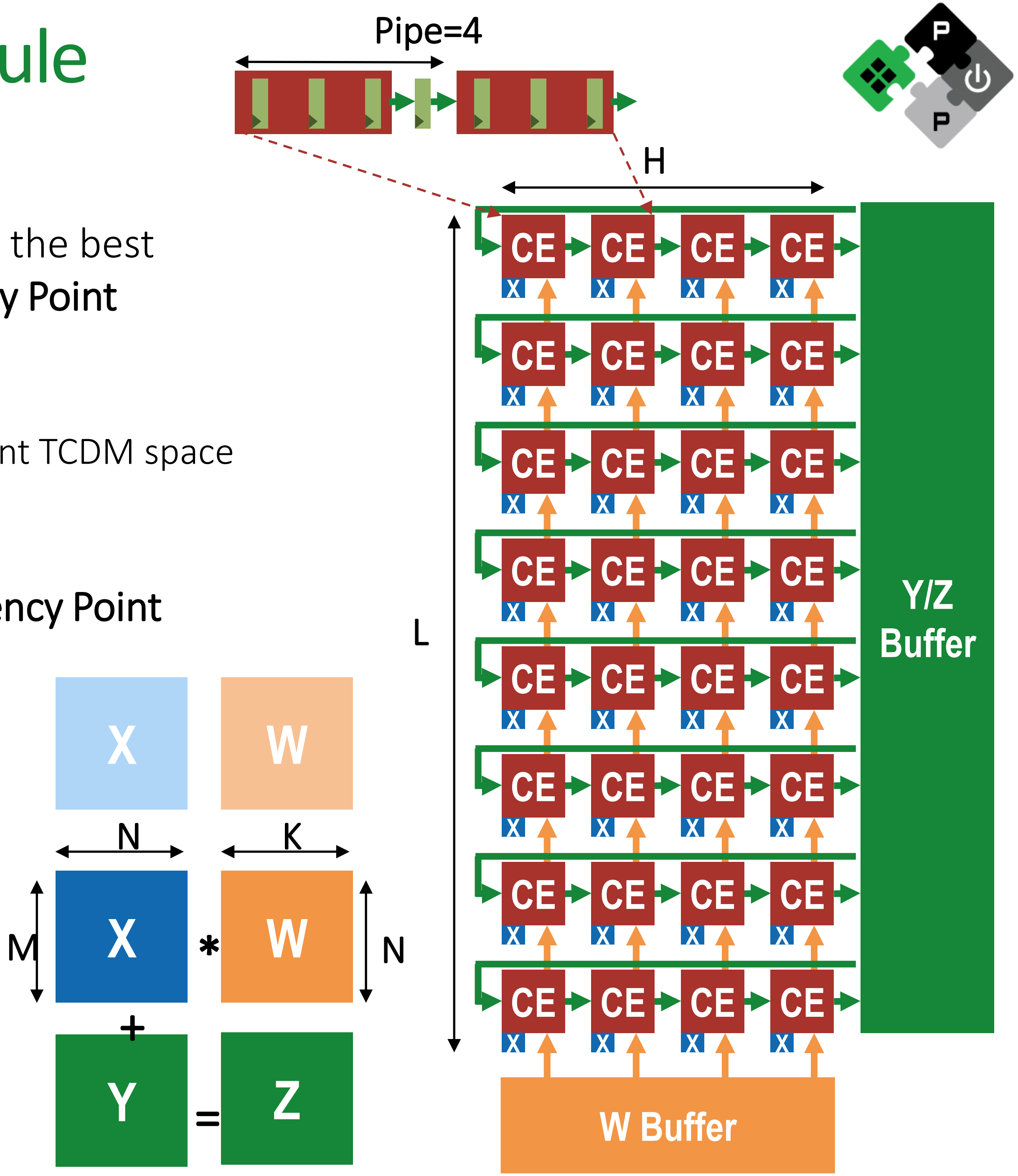


ETH zürich

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

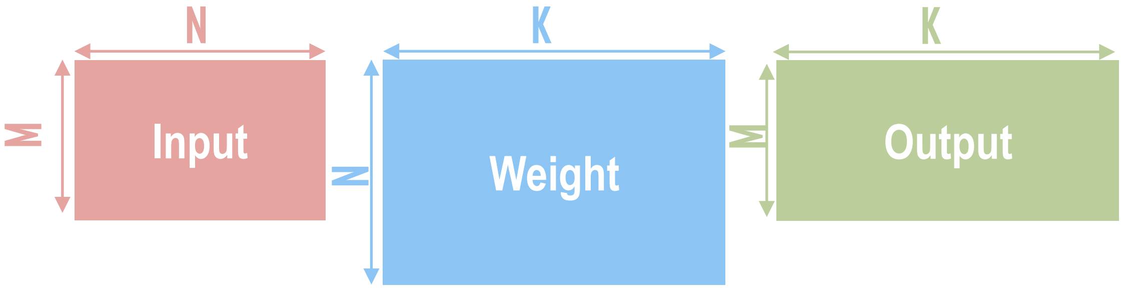
Design Space Exploration of RedMule

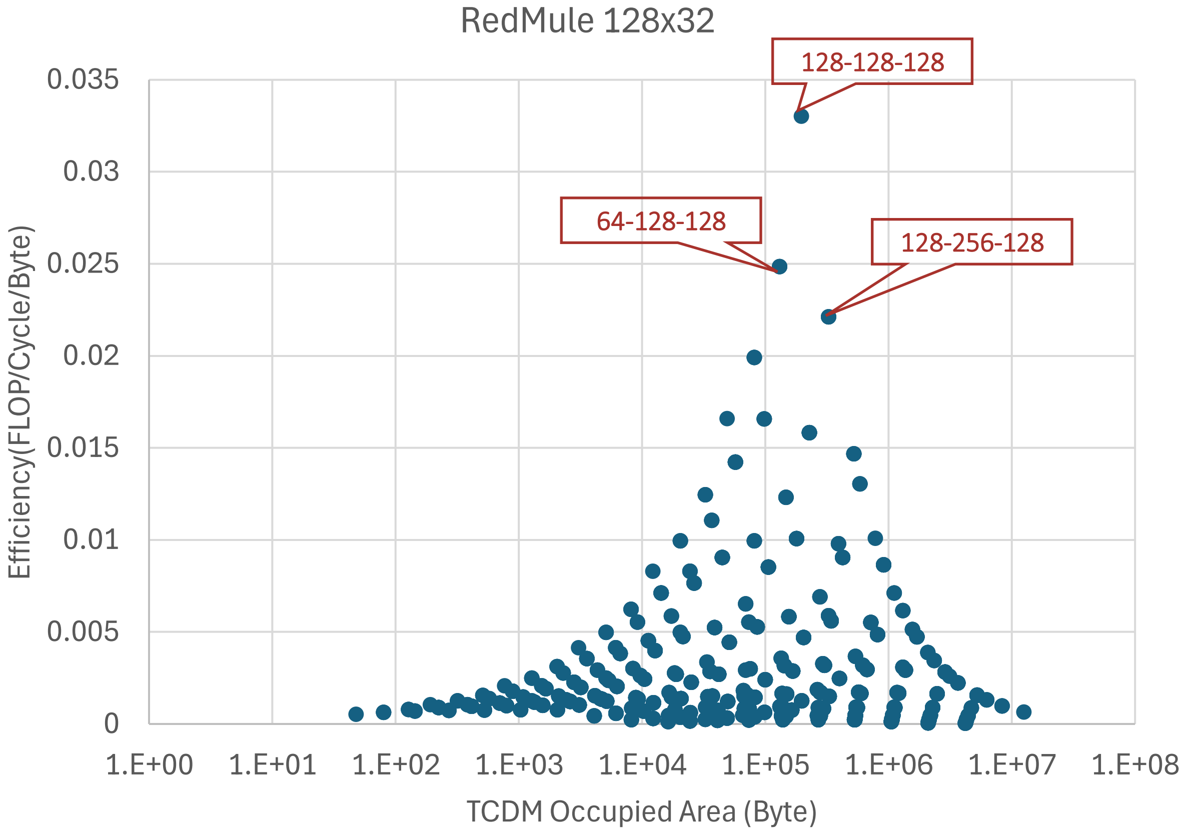
- Goal: We Want to Know
  - What is **optimal MatMul problem size** (M-N-K)to reach the best efficiency of RedMule **in a Cluster:** the **Optimal Efficiency Point**
    - When Matrix too small -> low RedMule utilization
    - When Matrix too large -> we introduce large and redundant TCDM space
    - We are seeking for **best MACs/SRAM in a CLuster**
  - What is the BW needed for RedMule at **Optimal Efficiency Point**
- Design Space Exploration Constraints
  - RedMule CE array constraint: L = H * Pipe, CE Pipe=4
- Key Metric
  - Efficiency Metric:
    - $\frac{Effective\ FLOP/Cycle}{TCDM\ Occupied\ Area} = \frac{RedMule\ Utilization*2*\#CEs}{Elem\ Size*2*(MN+NK+MK)}$



Find Optimal GEMM Size(M-N-K) for RedMule (@ saturated BW)

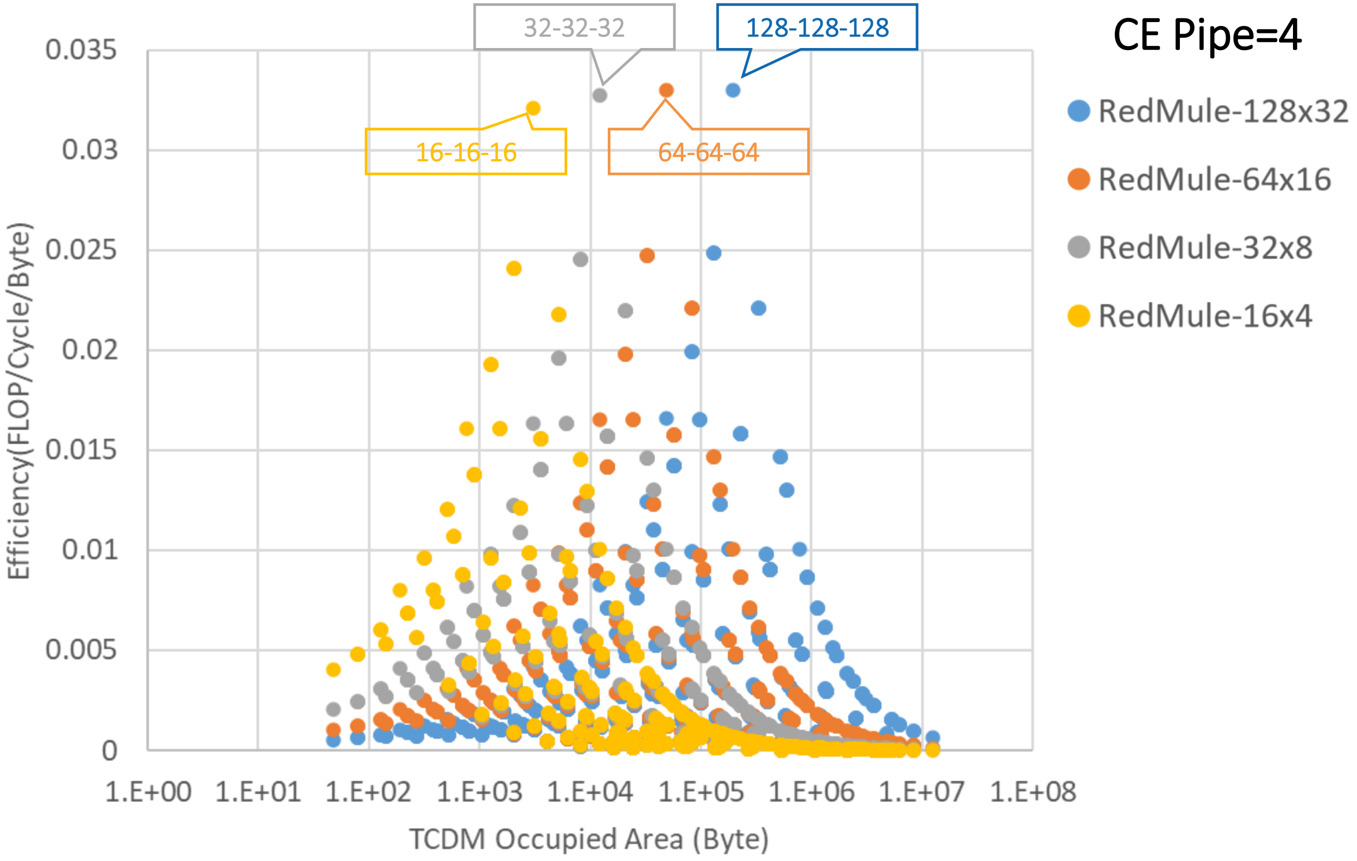
- GEMM Dimension
  - $M \in [8, 16, 32, 64, 128, 256, 512]$
  - $N \in [8, 16, 32, 64, 128, 256, 512]$
  - $K \in [8, 16, 32, 64, 128, 256, 512]$
- RedMule Config
  - CE array = 128x32
  - TCDM BW = 1024 Elem/Cycle
  - Element = FP16
- Run GEMM on One Cluster with One RedMule
- Collect Data
  - Efficiency Metric:
    - $\frac{Effective\ FLOP/Cycle}{TCDM\ Occupied\ Area} = \frac{RedMule\ Utilization * 2 * \#CEs}{Elem\ Size * 2 * (MN+NK+MK)}$





Find Optimal GEMM Size (M-N-K) for RedMule (@ saturated BW)

- At RedMule Constraints: CE array constraint: L = H * Pipe



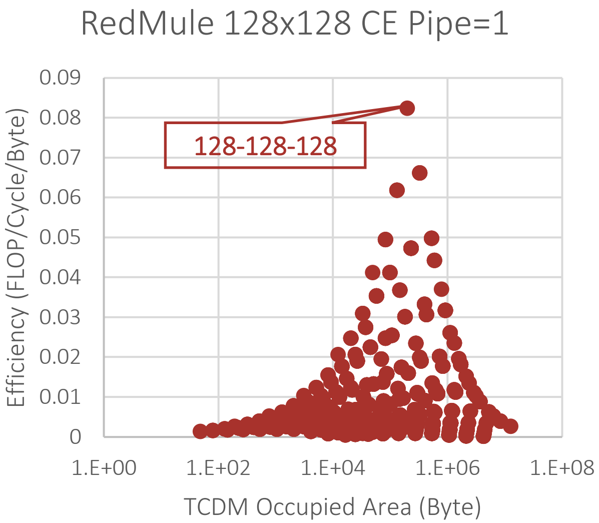
To use RedMule at best efficiency (Optimal Efficiency Point)

$$M = N = K = \sqrt{CE\ array\ * CE\ Pipeline} = L$$

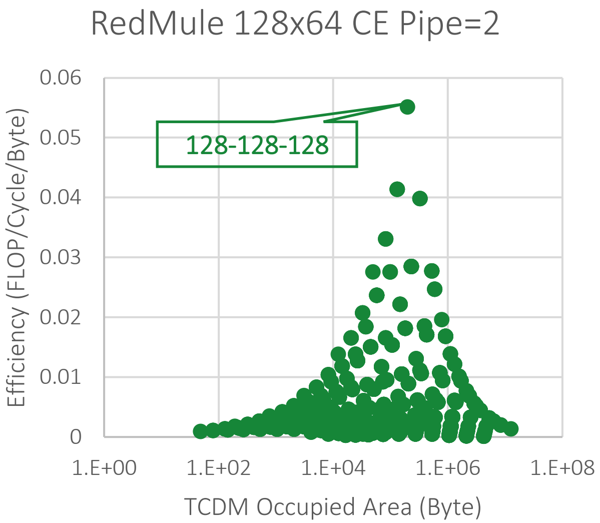
ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Find Optimal GEMM Size (M-N-K) for RedMule (@ saturated BW)

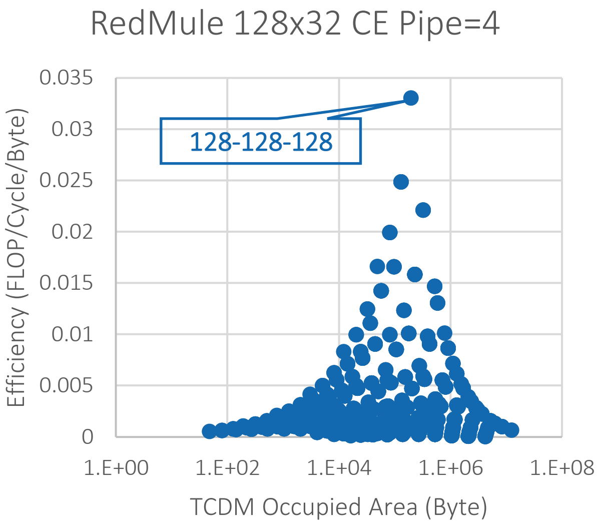
- At RedMule Constraints: CE array constraint: L = H * Pipe



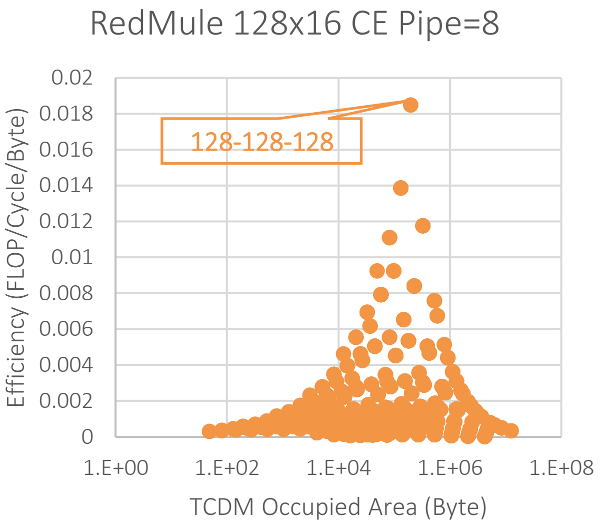
RedMule 128x128 CE Pipe=1



RedMule 128x64 CE Pipe=2



RedMule 128x32 CE Pipe=4



RedMule 128x16 CE Pipe=8

To use RedMule at best efficiency (Optimal Efficiency Point)

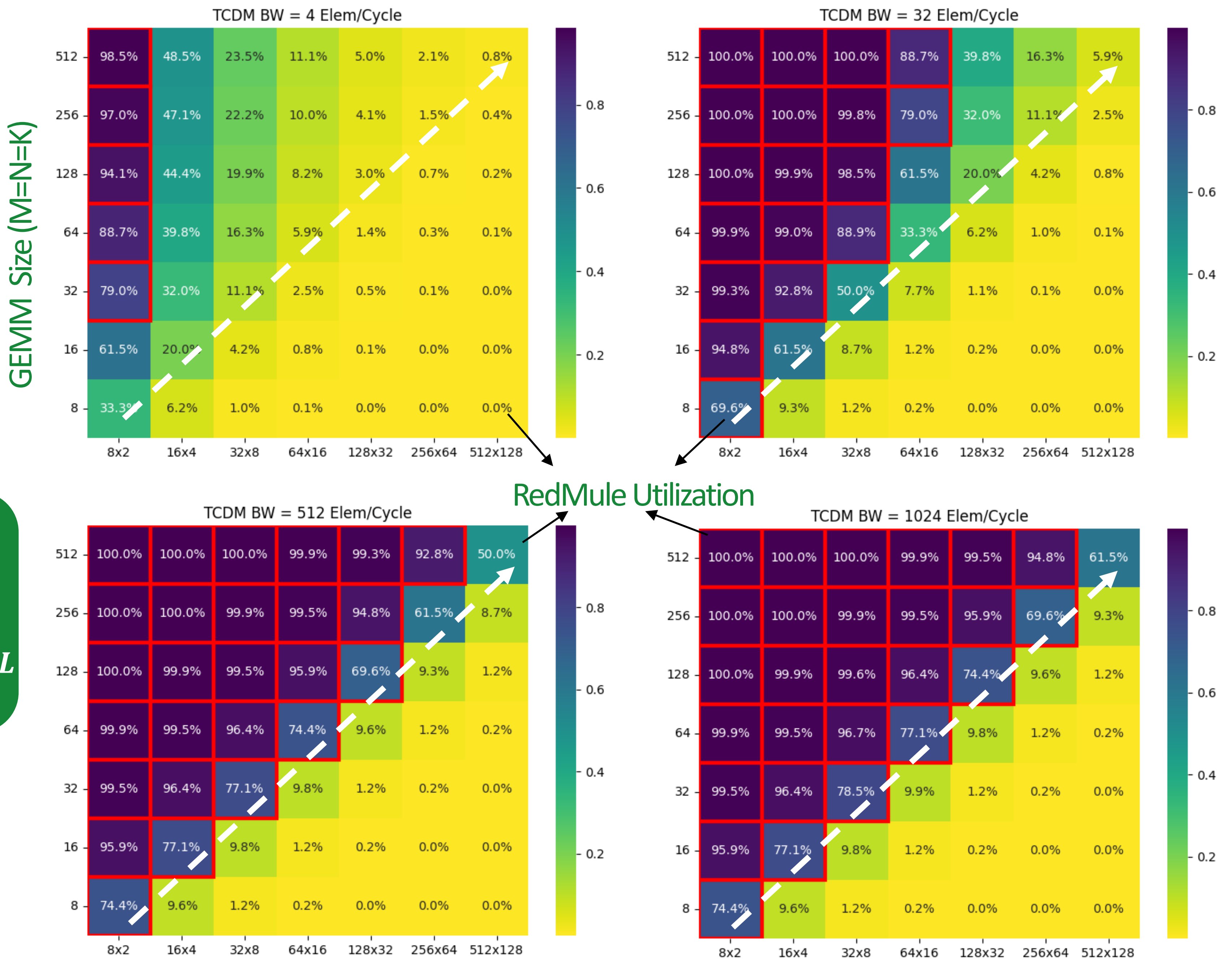
$$M = N = K = \sqrt{CE\ array\ * CE\ Pipeline} = L$$

TCDM Bandwidth Requirement

- RedMule Utilization Heatmap
  - GEMM Size vs RedMule CE array
  - Vary TCDM BW
  - We’re satisfied at > 69% uti

To feed RedMule efficiently TCDM Bandwidth (elem per cycle):

$$BW \geq 4\sqrt{CE\ array\ * CE\ Pipeline} = 4L$$



TCDM BW = 4 Elem/Cycle

	8x2	16x4	32x8	64x16	128x32	256x64	512x128
512	98.5%	48.5%	23.5%	11.1%	5.0%	2.1%	0.8%
256	97.0%	47.1%	22.2%	10.0%	4.1%	1.5%	0.4%
128	94.1%	44.4%	19.9%	8.2%	3.0%	0.7%	0.2%
64	88.7%	39.8%	16.3%	5.9%	1.4%	0.3%	0.1%
32	79.0%	32.0%	11.1%	2.5%	0.5%	0.1%	0.0%
16	61.5%	20.0%	4.2%	0.8%	0.1%	0.0%	0.0%
8	33.3%	6.2%	1.0%	0.1%	0.0%	0.0%	0.0%

TCDM BW = 32 Elem/Cycle

	8x2	16x4	32x8	64x16	128x32	256x64	512x128
512	100.0%	100.0%	100.0%	88.7%	39.8%	16.3%	5.9%
256	100.0%	100.0%	99.8%	79.0%	32.0%	11.1%	2.5%
128	100.0%	99.9%	98.5%	61.5%	20.0%	4.2%	0.8%
64	99.9%	99.0%	88.9%	33.3%	6.2%	1.0%	0.1%
32	99.3%	92.8%	50.0%	7.7%	1.1%	0.1%	0.0%
16	94.8%	61.5%	8.7%	1.2%	0.2%	0.0%	0.0%
8	69.6%	9.3%	1.2%	0.2%	0.0%	0.0%	0.0%

TCDM BW = 512 Elem/Cycle

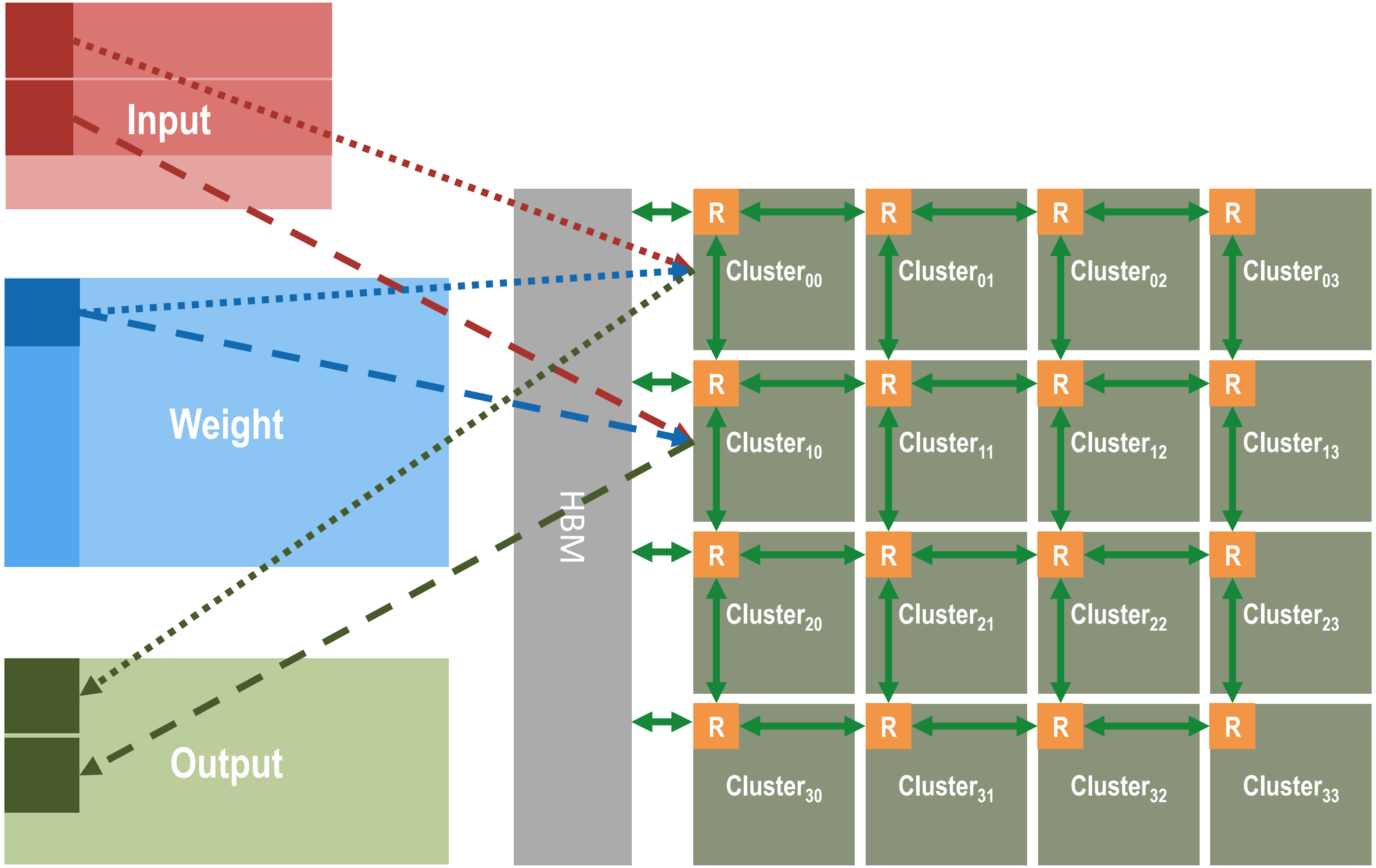
	8x2	16x4	32x8	64x16	128x32	256x64	512x128
512	100.0%	100.0%	100.0%	99.9%	99.3%	92.8%	50.0%
256	100.0%	100.0%	99.9%	99.5%	94.8%	61.5%	8.7%
128	100.0%	99.9%	99.5%	95.9%	69.6%	9.3%	1.2%
64	99.9%	99.5%	96.4%	74.4%	9.6%	1.2%	0.2%
32	99.5%	96.4%	77.1%	9.8%	1.2%	0.2%	0.0%
16	95.9%	77.1%	9.8%	1.2%	0.2%	0.0%	0.0%
8	74.4%	9.6%	1.2%	0.2%	0.0%	0.0%	0.0%

TCDM BW = 1024 Elem/Cycle

	8x2	16x4	32x8	64x16	128x32	256x64	512x128
512	100.0%	100.0%	100.0%	99.9%	99.5%	94.8%	61.5%
256	100.0%	100.0%	99.9%	99.5%	95.9%	69.6%	9.3%
128	100.0%	99.9%	99.6%	96.4%	74.4%	9.6%	1.2%
64	99.9%	99.5%	96.7%	77.1%	9.8%	1.2%	0.2%
32	99.5%	96.4%	78.5%	9.9%	1.2%	0.2%	0.0%
16	95.9%	77.1%	9.8%	1.2%	0.2%	0.0%	0.0%
8	74.4%	9.6%	1.2%	0.2%	0.0%	0.0%	0.0%

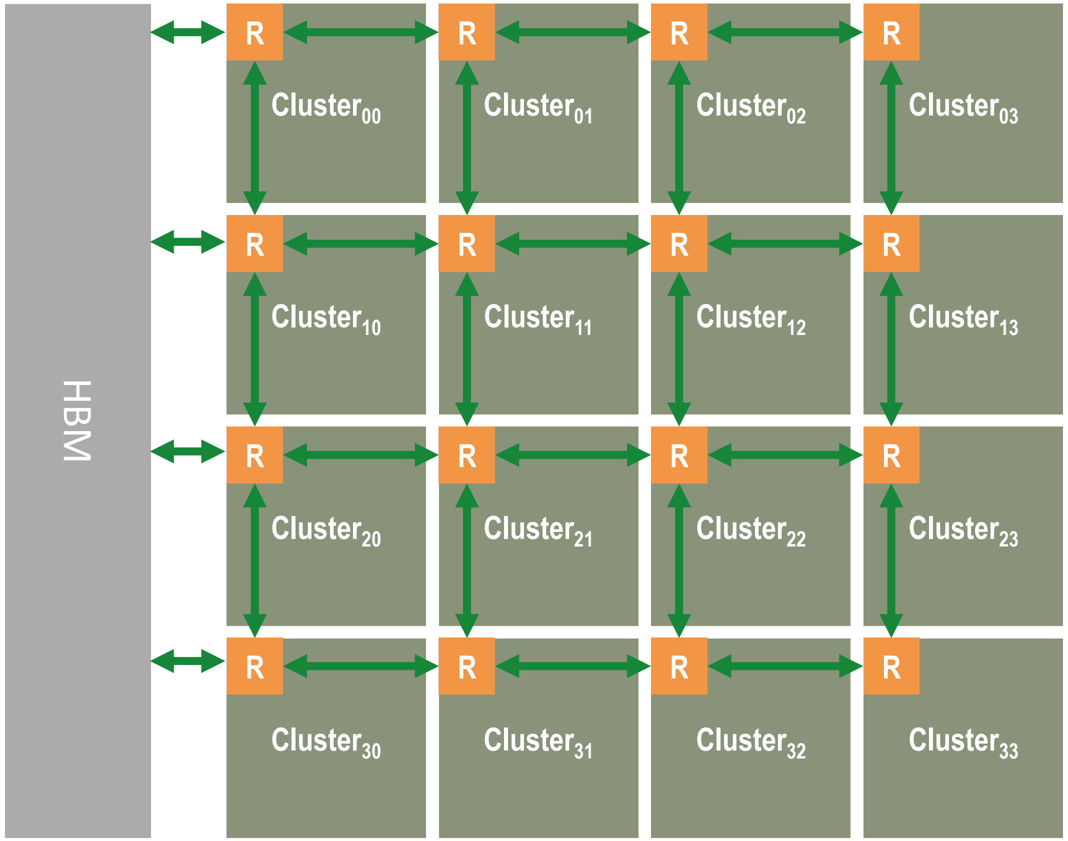
For Large GEMM: Enable Cluster-to-Cluster Comm or Not?

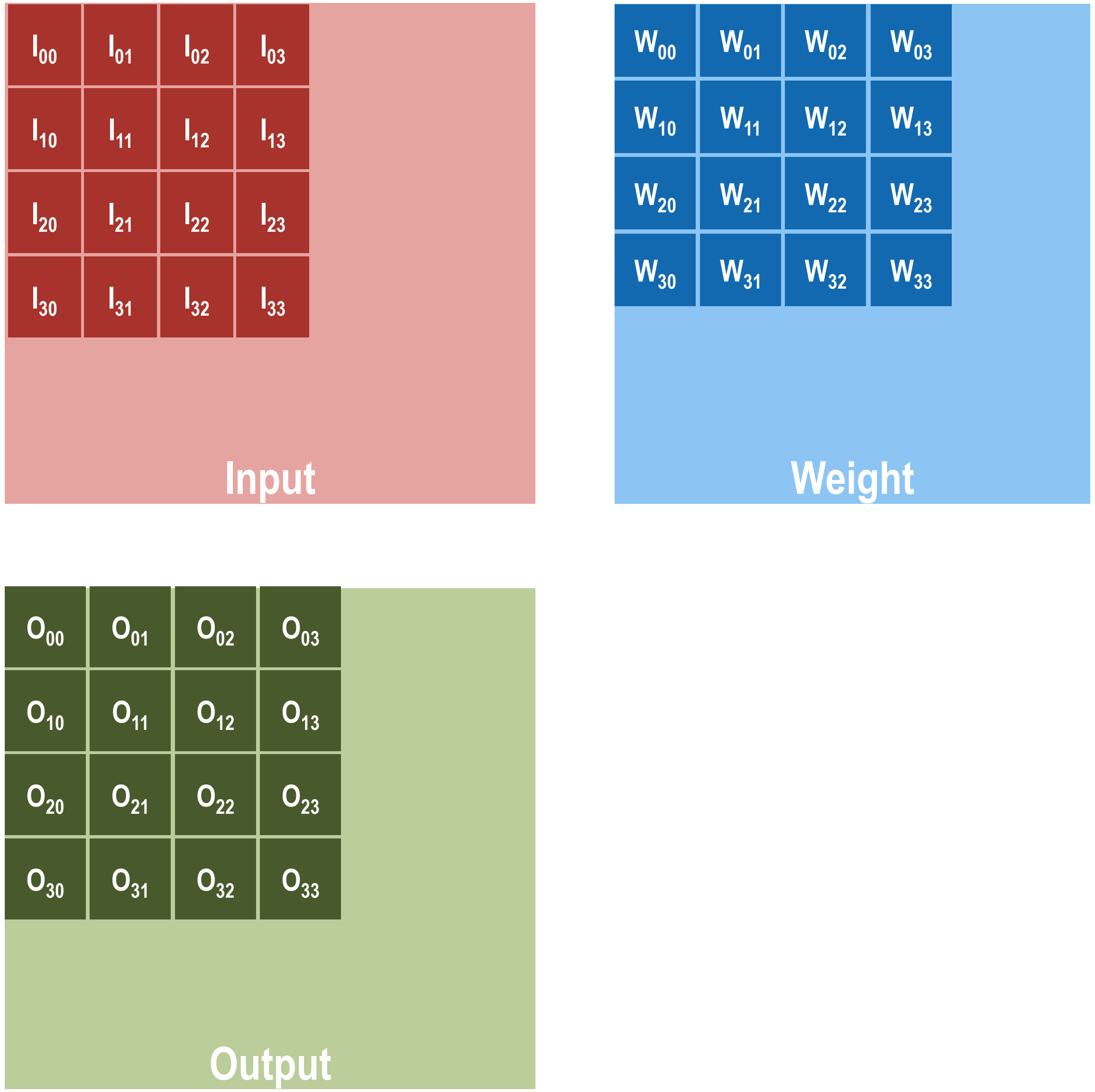
- No Cluster-to-Cluster Comm.
  - Each clusters take care of different output tiles (384x384).
  - iDMA transfers input matrix tile + weight matrix tile from HBM
  - No inter-cluster tile reuses



For Large GEMM: Enable Cluster-to-Cluster Comm or Not?

- Leverage Cluster-to-Cluster Comm.
  - Reuse tiles, reduce HBM accesses, save BW limitation
  - Need “smart” tile mapping and scheduling





Input

I_{00}	I_{01}	I_{02}	I_{03}
I_{10}	I_{11}	I_{12}	I_{13}
I_{20}	I_{21}	I_{22}	I_{23}
I_{30}	I_{31}	I_{32}	I_{33}

Weight

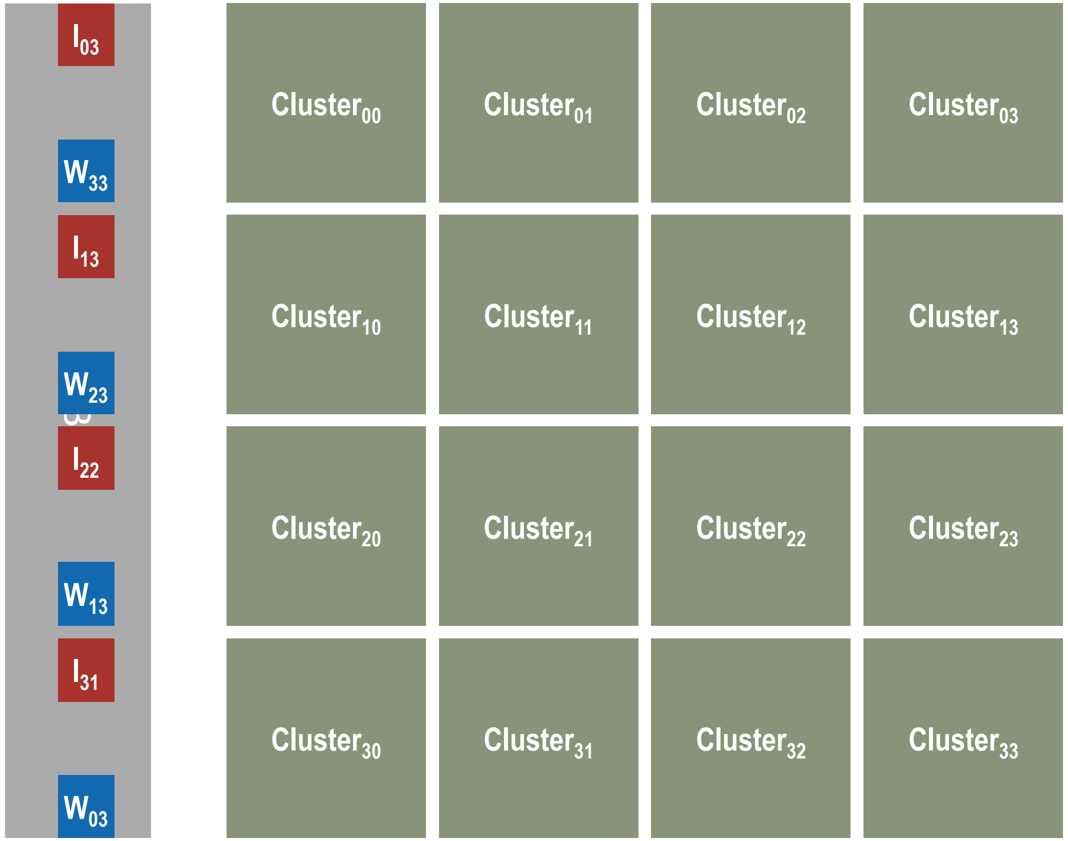
W_{00}	W_{01}	W_{02}	W_{03}
W_{10}	W_{11}	W_{12}	W_{13}
W_{20}	W_{21}	W_{22}	W_{23}
W_{30}	W_{31}	W_{32}	W_{33}

Output

O_{00}	O_{01}	O_{02}	O_{03}
O_{10}	O_{11}	O_{12}	O_{13}
O_{20}	O_{21}	O_{22}	O_{23}
O_{30}	O_{31}	O_{32}	O_{33}

Example Solution for GEMM: Step1 Distribute Tiles to Clusters

- Load All (16 Input tiles, 16 weight tiles) from HBM
- Cluster contain different tiles from each other



$$O_{00} = \boxed{I_{00} \cdot W_{00}} + I_{01} \cdot W_{10} + I_{02} \cdot W_{20} + I_{03} \cdot W_{30}$$

$$O_{01} = I_{00} \cdot W_{01} + \boxed{I_{01} \cdot W_{11}} + I_{02} \cdot W_{21} + I_{03} \cdot W_{31}$$

$$O_{02} = I_{00} \cdot W_{02} + I_{01} \cdot W_{12} + \boxed{I_{02} \cdot W_{22}} + I_{03} \cdot W_{32}$$

$$O_{03} = I_{00} \cdot W_{03} + I_{01} \cdot W_{13} + I_{02} \cdot W_{23} + \boxed{I_{03} \cdot W_{33}}$$

$$O_{10} = I_{10} \cdot W_{00} + I_{11} \cdot W_{10} + I_{12} \cdot W_{20} + \boxed{I_{13} \cdot W_{30}}$$

$$O_{11} = \boxed{I_{10} \cdot W_{01}} + I_{11} \cdot W_{11} + I_{12} \cdot W_{21} + I_{13} \cdot W_{31}$$

$$O_{12} = I_{10} \cdot W_{02} + \boxed{I_{11} \cdot W_{12}} + I_{12} \cdot W_{22} + I_{13} \cdot W_{32}$$

$$O_{13} = I_{10} \cdot W_{03} + I_{11} \cdot W_{13} + \boxed{I_{12} \cdot W_{23}} + I_{13} \cdot W_{33}$$

$$O_{20} = I_{20} \cdot W_{00} + I_{21} \cdot W_{10} + \boxed{I_{22} \cdot W_{20}} + I_{23} \cdot W_{30}$$

$$O_{21} = I_{20} \cdot W_{01} + I_{21} \cdot W_{11} + I_{22} \cdot W_{21} + \boxed{I_{23} \cdot W_{31}}$$

$$O_{22} = \boxed{I_{20} \cdot W_{02}} + I_{21} \cdot W_{12} + I_{22} \cdot W_{22} + I_{23} \cdot W_{32}$$

$$O_{23} = I_{20} \cdot W_{03} + \boxed{I_{21} \cdot W_{13}} + I_{22} \cdot W_{23} + I_{23} \cdot W_{33}$$

$$O_{30} = I_{30} \cdot W_{00} + \boxed{I_{31} \cdot W_{10}} + I_{32} \cdot W_{20} + I_{33} \cdot W_{30}$$

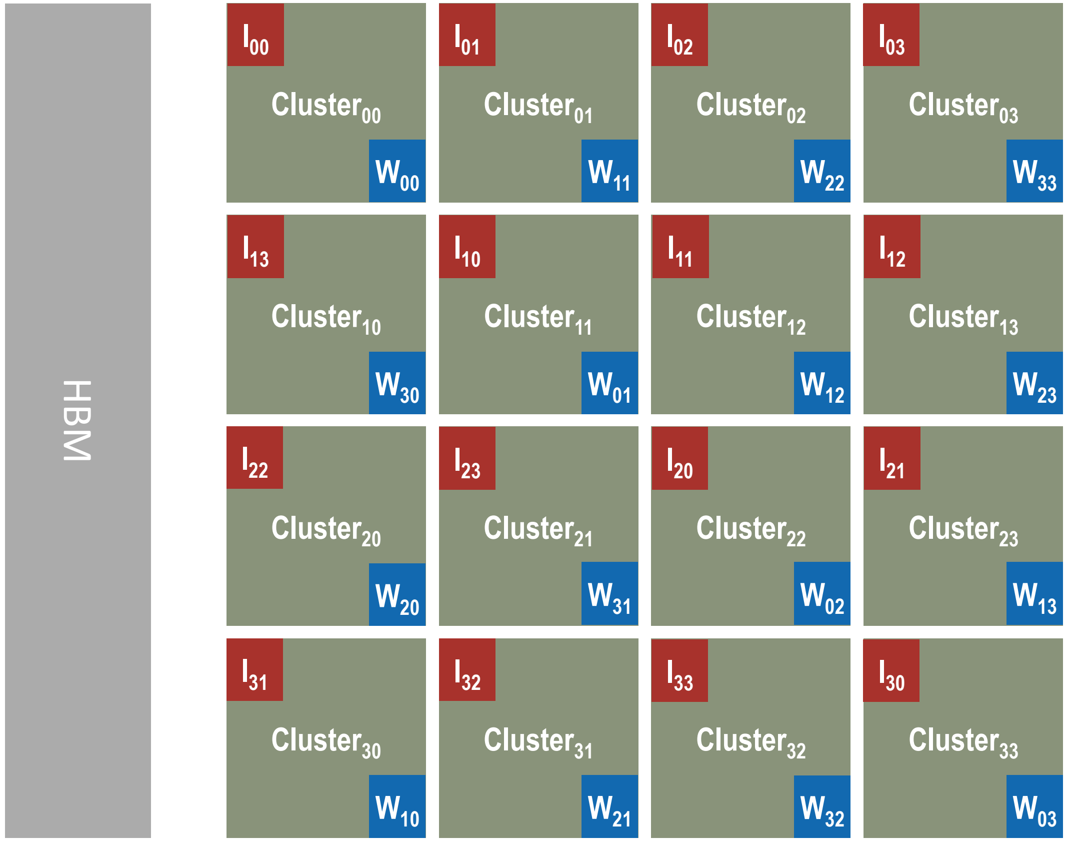
$$O_{31} = I_{30} \cdot W_{01} + I_{31} \cdot W_{11} + \boxed{I_{32} \cdot W_{21}} + I_{33} \cdot W_{31}$$

$$O_{32} = I_{30} \cdot W_{02} + I_{31} \cdot W_{12} + I_{32} \cdot W_{22} + \boxed{I_{33} \cdot W_{32}}$$

$$O_{33} = \boxed{I_{30} \cdot W_{03}} + I_{31} \cdot W_{13} + I_{32} \cdot W_{23} + I_{33} \cdot W_{33}$$

Example Solution for GEMM: Step1 Distribute Tiles to Clusters

- Load All (16 Input tiles, 16 weight tiles) from HBM
- Cluster contain different tiles from each other



HBM

I_{00} Cluster$_{00}$ W_{00}	I_{01} Cluster$_{01}$ W_{11}	I_{02} Cluster$_{02}$ W_{22}	I_{03} Cluster$_{03}$ W_{33}
I_{13} Cluster$_{10}$ W_{30}	I_{10} Cluster$_{11}$ W_{01}	I_{11} Cluster$_{12}$ W_{12}	I_{12} Cluster$_{13}$ W_{23}
I_{22} Cluster$_{20}$ W_{20}	I_{23} Cluster$_{21}$ W_{31}	I_{20} Cluster$_{22}$ W_{02}	I_{21} Cluster$_{23}$ W_{13}
I_{31} Cluster$_{30}$ W_{10}	I_{32} Cluster$_{31}$ W_{21}	I_{33} Cluster$_{32}$ W_{32}	I_{30} Cluster$_{33}$ W_{03}

$$O_{00} = \boxed{I_{00} \cdot W_{00}} + I_{01} \cdot W_{10} + I_{02} \cdot W_{20} + I_{03} \cdot W_{30}$$

$$O_{01} = I_{00} \cdot W_{01} + \boxed{I_{01} \cdot W_{11}} + I_{02} \cdot W_{21} + I_{03} \cdot W_{31}$$

$$O_{02} = I_{00} \cdot W_{02} + I_{01} \cdot W_{12} + \boxed{I_{02} \cdot W_{22}} + I_{03} \cdot W_{32}$$

$$O_{03} = I_{00} \cdot W_{03} + I_{01} \cdot W_{13} + I_{02} \cdot W_{23} + \boxed{I_{03} \cdot W_{33}}$$

$$O_{10} = I_{10} \cdot W_{00} + I_{11} \cdot W_{10} + I_{12} \cdot W_{20} + \boxed{I_{13} \cdot W_{30}}$$

$$O_{11} = \boxed{I_{10} \cdot W_{01}} + I_{11} \cdot W_{11} + I_{12} \cdot W_{21} + I_{13} \cdot W_{31}$$

$$O_{12} = I_{10} \cdot W_{02} + \boxed{I_{11} \cdot W_{12}} + I_{12} \cdot W_{22} + I_{13} \cdot W_{32}$$

$$O_{13} = I_{10} \cdot W_{03} + I_{11} \cdot W_{13} + \boxed{I_{12} \cdot W_{23}} + I_{13} \cdot W_{33}$$

$$O_{20} = I_{20} \cdot W_{00} + I_{21} \cdot W_{10} + \boxed{I_{22} \cdot W_{20}} + I_{23} \cdot W_{30}$$

$$O_{21} = I_{20} \cdot W_{01} + I_{21} \cdot W_{11} + I_{22} \cdot W_{21} + \boxed{I_{23} \cdot W_{31}}$$

$$O_{22} = \boxed{I_{20} \cdot W_{02}} + I_{21} \cdot W_{12} + I_{22} \cdot W_{22} + I_{23} \cdot W_{32}$$

$$O_{23} = I_{20} \cdot W_{03} + \boxed{I_{21} \cdot W_{13}} + I_{22} \cdot W_{23} + I_{23} \cdot W_{33}$$

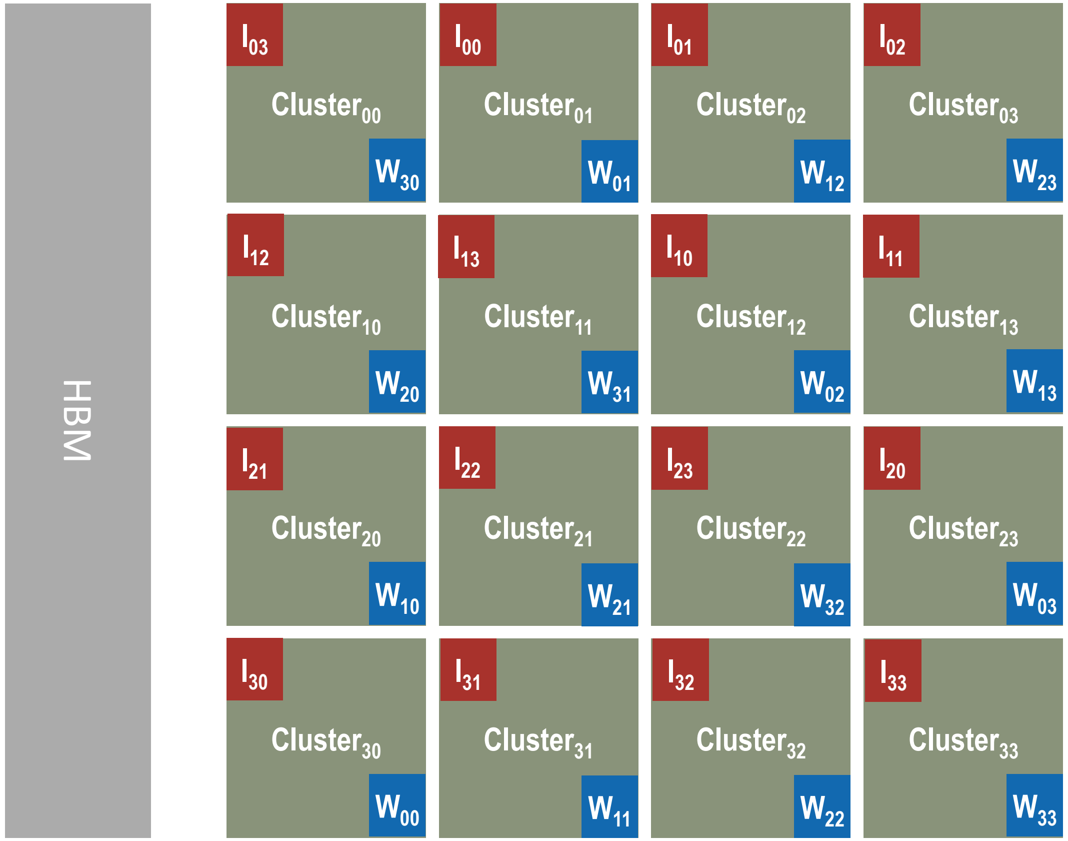
$$O_{30} = I_{30} \cdot W_{00} + \boxed{I_{31} \cdot W_{10}} + I_{32} \cdot W_{20} + I_{33} \cdot W_{30}$$

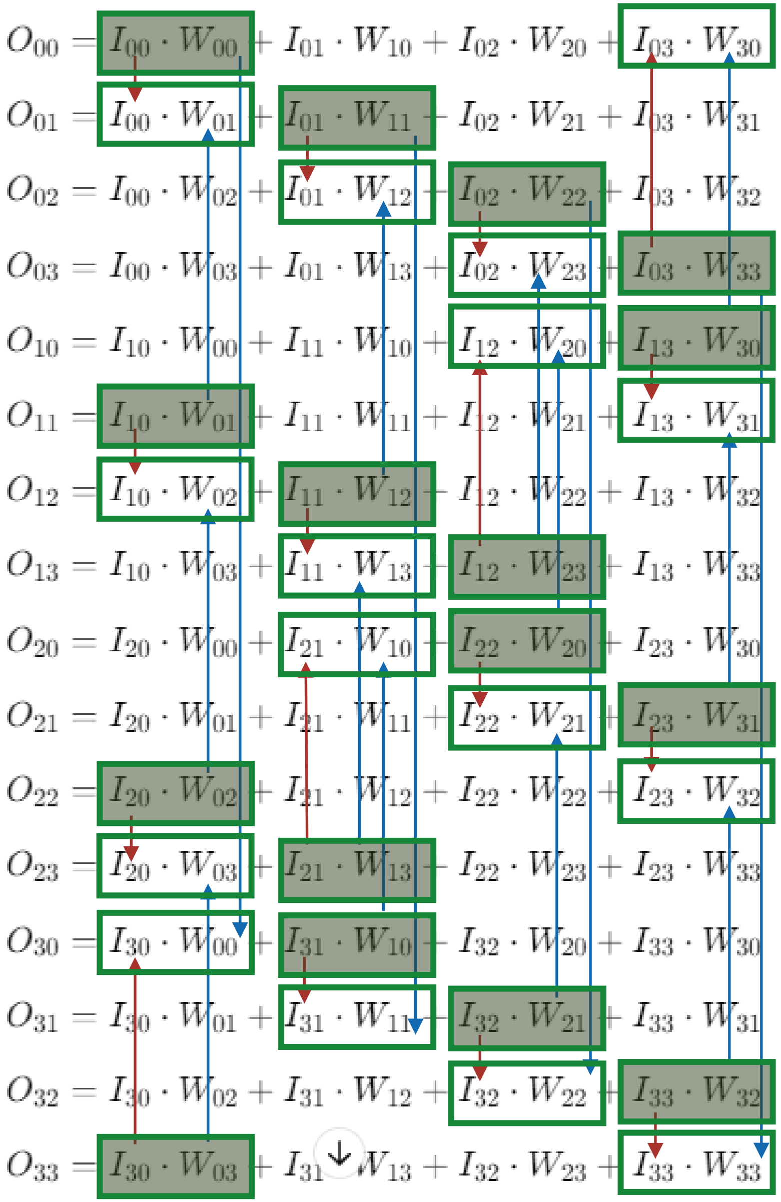
$$O_{31} = I_{30} \cdot W_{01} + I_{31} \cdot W_{11} + \boxed{I_{32} \cdot W_{21}} + I_{33} \cdot W_{31}$$

$$O_{32} = I_{30} \cdot W_{02} + I_{31} \cdot W_{12} + I_{32} \cdot W_{22} + \boxed{I_{33} \cdot W_{32}}$$

$$O_{33} = \boxed{I_{30} \cdot W_{03}} + I_{31} \cdot W_{13} + I_{32} \cdot W_{23} + I_{33} \cdot W_{33}$$

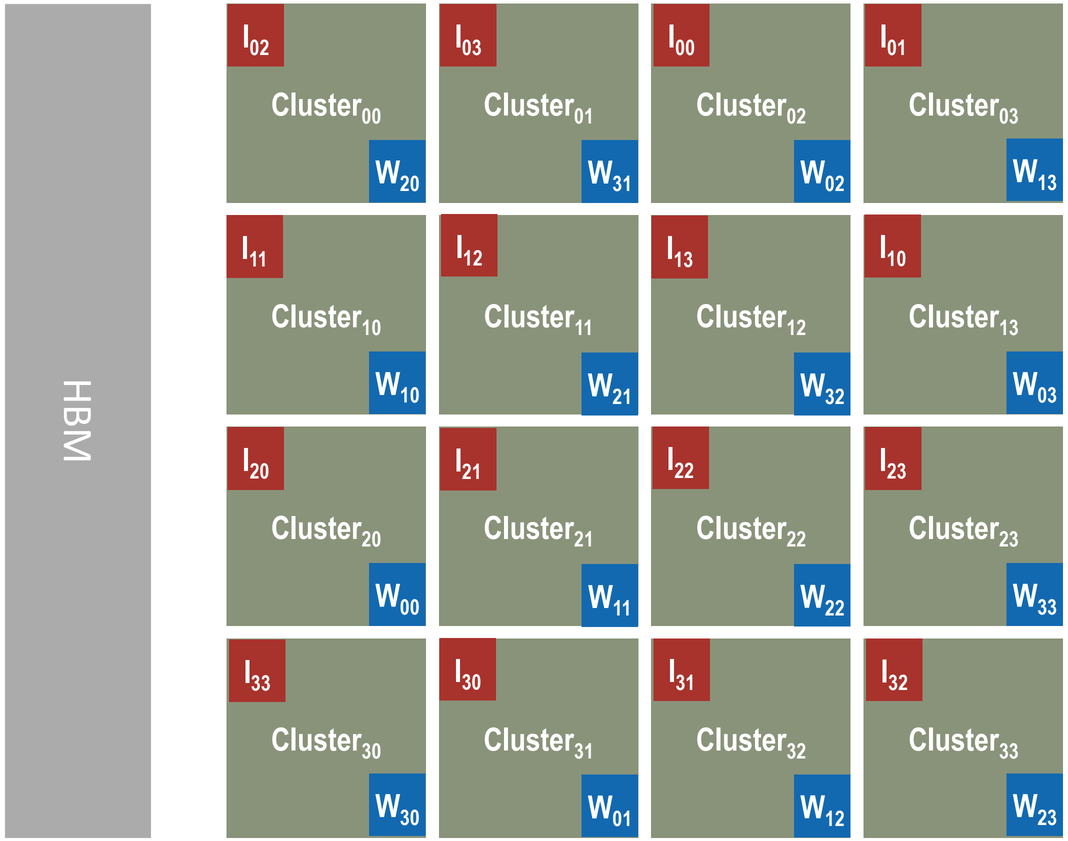
Example Solution for GEMM: Step2 Inter-Cluster Tile Exchanging

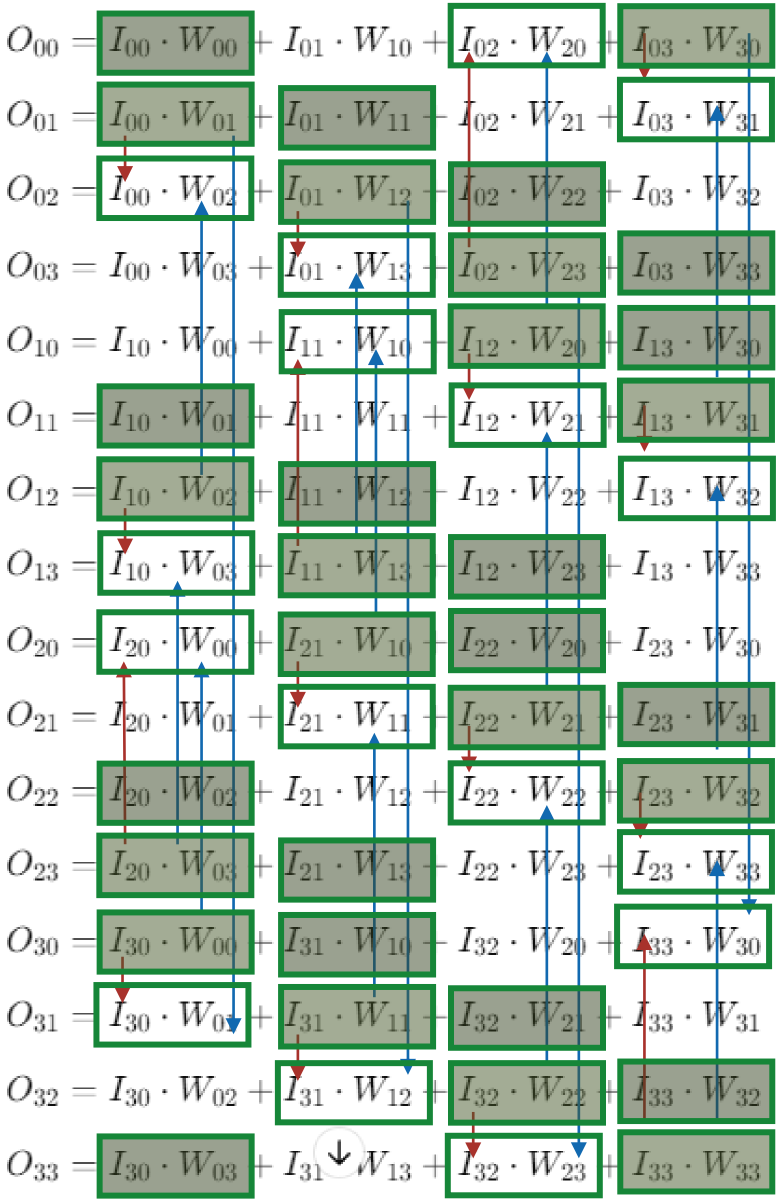




$$O_{00} = \boxed{I_{00} \cdot W_{00}} + I_{01} \cdot W_{10} + I_{02} \cdot W_{20} + \boxed{I_{03} \cdot W_{30}}$$
$$O_{01} = \boxed{I_{00} \cdot W_{01}} + \boxed{I_{01} \cdot W_{11}} + I_{02} \cdot W_{21} + I_{03} \cdot W_{31}$$
$$O_{02} = I_{00} \cdot W_{02} + \boxed{I_{01} \cdot W_{12}} + \boxed{I_{02} \cdot W_{22}} + I_{03} \cdot W_{32}$$
$$O_{03} = I_{00} \cdot W_{03} + I_{01} \cdot W_{13} + \boxed{I_{02} \cdot W_{23}} + \boxed{I_{03} \cdot W_{33}}$$
$$O_{10} = I_{10} \cdot W_{00} + I_{11} \cdot W_{10} + \boxed{I_{12} \cdot W_{20}} + \boxed{I_{13} \cdot W_{30}}$$
$$O_{11} = \boxed{I_{10} \cdot W_{01}} + I_{11} \cdot W_{11} + I_{12} \cdot W_{21} + \boxed{I_{13} \cdot W_{31}}$$
$$O_{12} = \boxed{I_{10} \cdot W_{02}} + \boxed{I_{11} \cdot W_{12}} + I_{12} \cdot W_{22} + I_{13} \cdot W_{32}$$
$$O_{13} = I_{10} \cdot W_{03} + \boxed{I_{11} \cdot W_{13}} + \boxed{I_{12} \cdot W_{23}} + I_{13} \cdot W_{33}$$
$$O_{20} = I_{20} \cdot W_{00} + \boxed{I_{21} \cdot W_{10}} + \boxed{I_{22} \cdot W_{20}} + I_{23} \cdot W_{30}$$
$$O_{21} = I_{20} \cdot W_{01} + I_{21} \cdot W_{11} + \boxed{I_{22} \cdot W_{21}} + \boxed{I_{23} \cdot W_{31}}$$
$$O_{22} = \boxed{I_{20} \cdot W_{02}} + I_{21} \cdot W_{12} + I_{22} \cdot W_{22} + \boxed{I_{23} \cdot W_{32}}$$
$$O_{23} = \boxed{I_{20} \cdot W_{03}} + \boxed{I_{21} \cdot W_{13}} + I_{22} \cdot W_{23} + I_{23} \cdot W_{33}$$
$$O_{30} = \boxed{I_{30} \cdot W_{00}} + \boxed{I_{31} \cdot W_{10}} + I_{32} \cdot W_{20} + I_{33} \cdot W_{30}$$
$$O_{31} = I_{30} \cdot W_{01} + \boxed{I_{31} \cdot W_{11}} + \boxed{I_{32} \cdot W_{21}} + I_{33} \cdot W_{31}$$
$$O_{32} = I_{30} \cdot W_{02} + I_{31} \cdot W_{12} + \boxed{I_{32} \cdot W_{22}} + \boxed{I_{33} \cdot W_{32}}$$
$$O_{33} = \boxed{I_{30} \cdot W_{03}} + I_{31} \cdot W_{13} + I_{32} \cdot W_{23} + \boxed{I_{33} \cdot W_{33}}$$

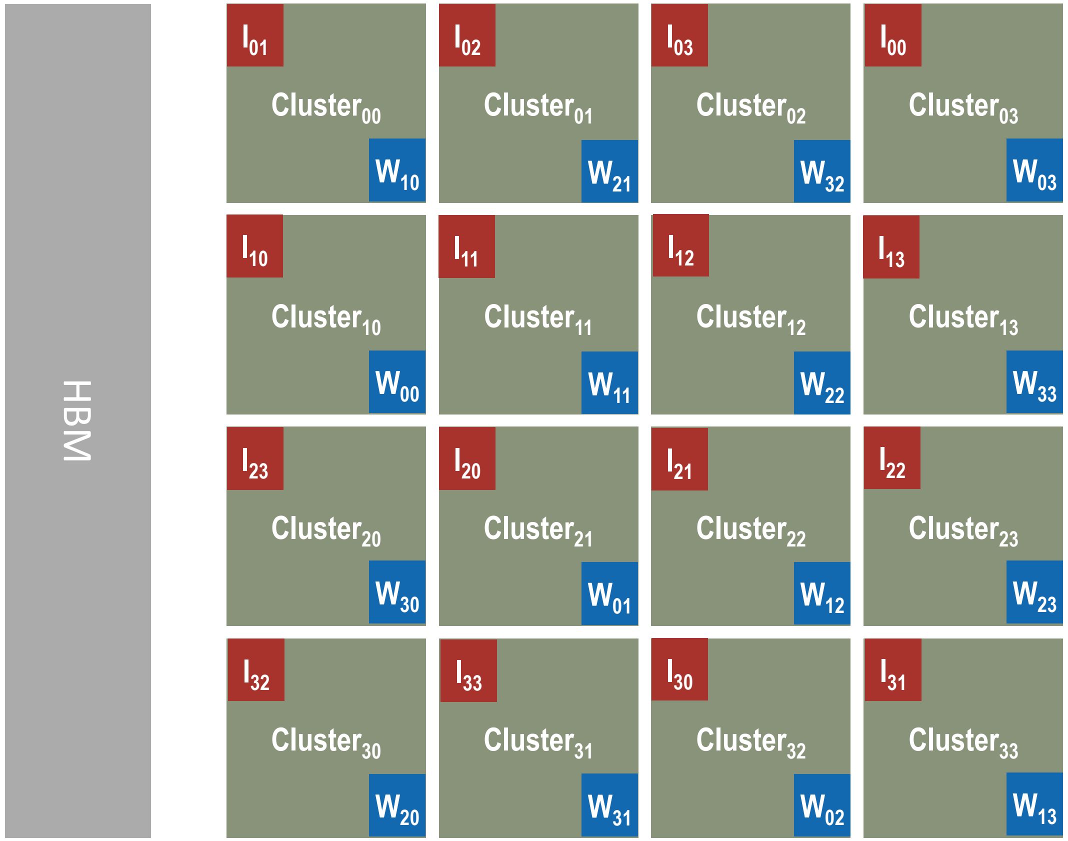
Example Solution for GEMM: Step2 Inter-Cluster Tile Exchanging



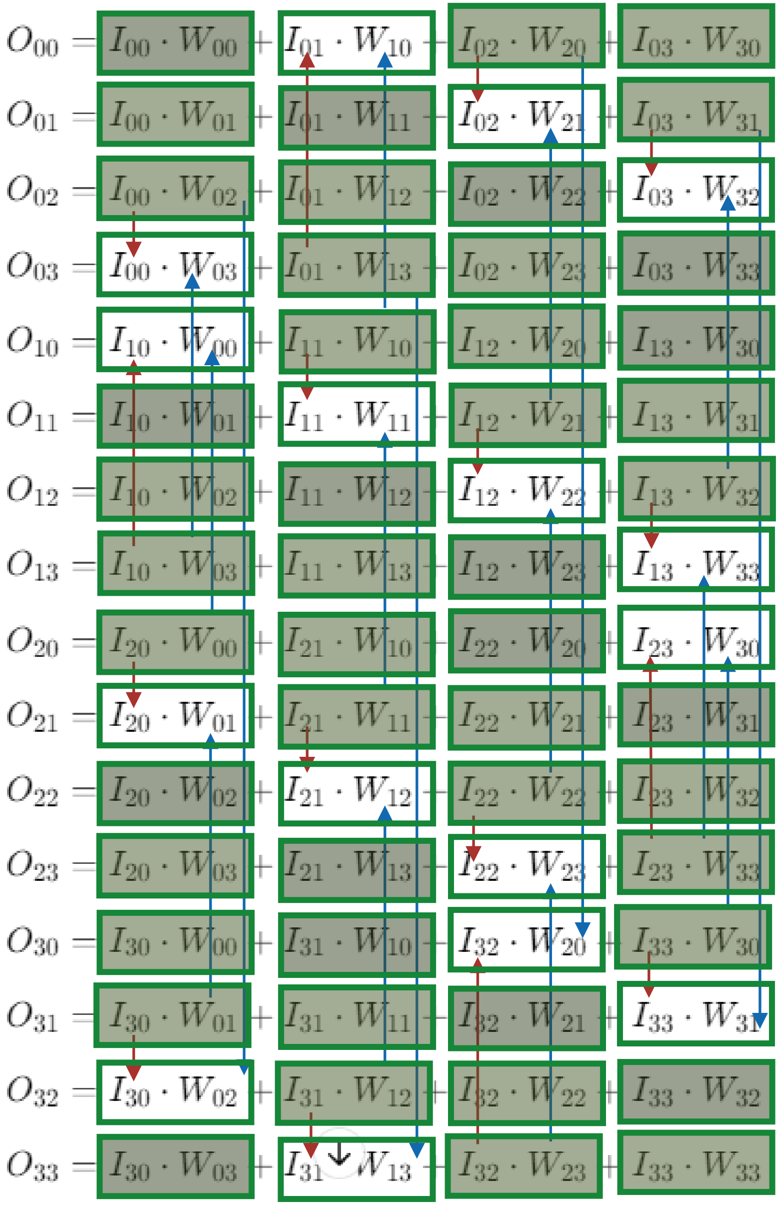


$$
\begin{aligned}
O_{00} &= I_{00} \cdot W_{00} + I_{01} \cdot W_{10} + I_{02} \cdot W_{20} + I_{03} \cdot W_{30} \\
O_{01} &= I_{00} \cdot W_{01} + I_{01} \cdot W_{11} + I_{02} \cdot W_{21} + I_{03} \cdot W_{31} \\
O_{02} &= I_{00} \cdot W_{02} + I_{01} \cdot W_{12} + I_{02} \cdot W_{22} + I_{03} \cdot W_{32} \\
O_{03} &= I_{00} \cdot W_{03} + I_{01} \cdot W_{13} + I_{02} \cdot W_{23} + I_{03} \cdot W_{33} \\
O_{10} &= I_{10} \cdot W_{00} + I_{11} \cdot W_{10} + I_{12} \cdot W_{20} + I_{13} \cdot W_{30} \\
O_{11} &= I_{10} \cdot W_{01} + I_{11} \cdot W_{11} + I_{12} \cdot W_{21} + I_{13} \cdot W_{31} \\
O_{12} &= I_{10} \cdot W_{02} + I_{11} \cdot W_{12} + I_{12} \cdot W_{22} + I_{13} \cdot W_{32} \\
O_{13} &= I_{10} \cdot W_{03} + I_{11} \cdot W_{13} + I_{12} \cdot W_{23} + I_{13} \cdot W_{33} \\
O_{20} &= I_{20} \cdot W_{00} + I_{21} \cdot W_{10} + I_{22} \cdot W_{20} + I_{23} \cdot W_{30} \\
O_{21} &= I_{20} \cdot W_{01} + I_{21} \cdot W_{11} + I_{22} \cdot W_{21} + I_{23} \cdot W_{31} \\
O_{22} &= I_{20} \cdot W_{02} + I_{21} \cdot W_{12} + I_{22} \cdot W_{22} + I_{23} \cdot W_{32} \\
O_{23} &= I_{20} \cdot W_{03} + I_{21} \cdot W_{13} + I_{22} \cdot W_{23} + I_{23} \cdot W_{33} \\
O_{30} &= I_{30} \cdot W_{00} + I_{31} \cdot W_{10} + I_{32} \cdot W_{20} + I_{33} \cdot W_{30} \\
O_{31} &= I_{30} \cdot W_{01} + I_{31} \cdot W_{11} + I_{32} \cdot W_{21} + I_{33} \cdot W_{31} \\
O_{32} &= I_{30} \cdot W_{02} + I_{31} \cdot W_{12} + I_{32} \cdot W_{22} + I_{33} \cdot W_{32} \\
O_{33} &= I_{30} \cdot W_{03} + I_{31} \cdot W_{13} + I_{32} \cdot W_{23} + I_{33} \cdot W_{33}
\end{aligned}
$$

Example Solution for GEMM: Step2 Inter-Cluster Tile Exchanging



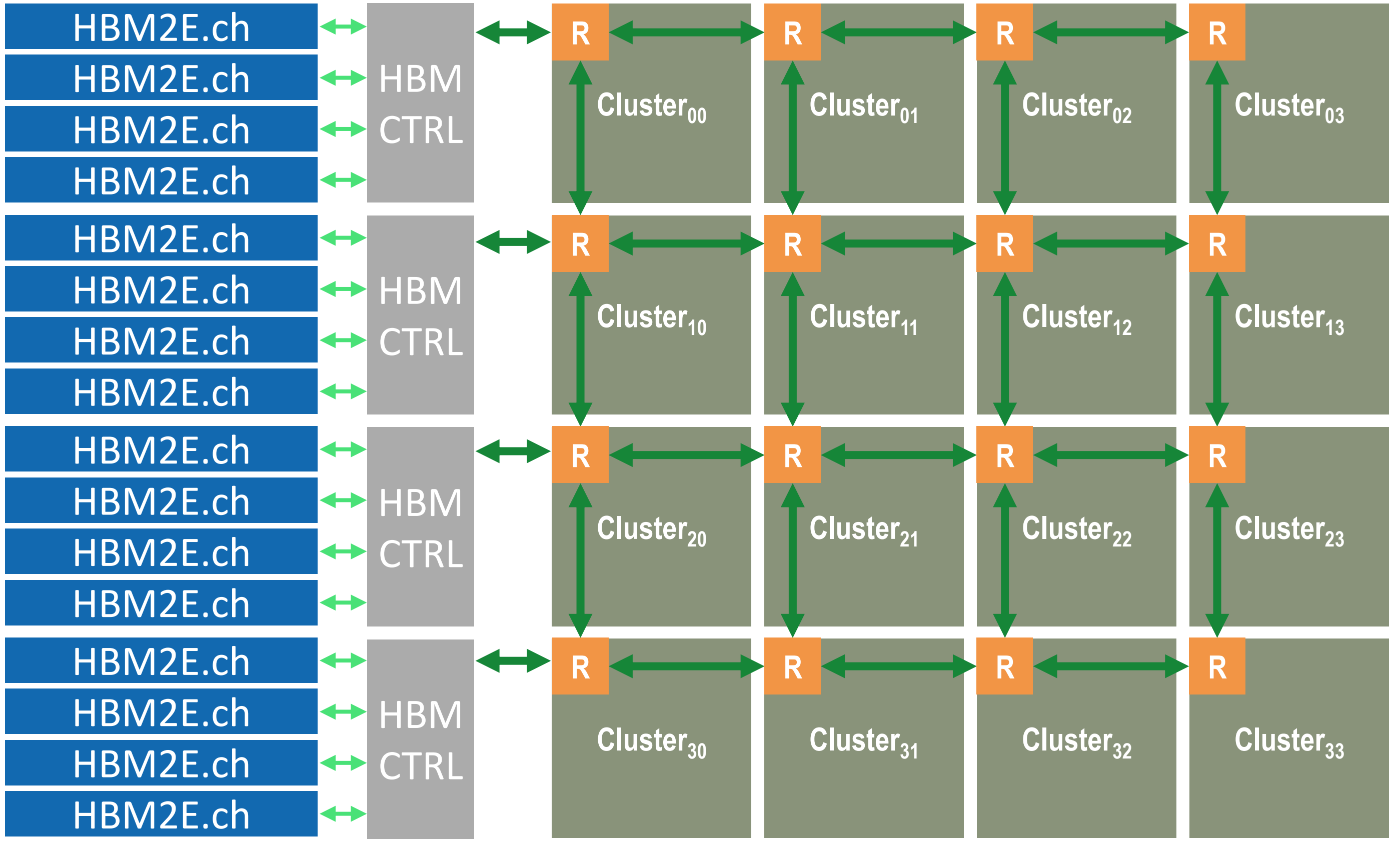
I_{01} Cluster_{00} W_{10}	I_{02} Cluster_{01} W_{21}	I_{03} Cluster_{02} W_{32}	I_{00} Cluster_{03} W_{03}
I_{10} Cluster_{10} W_{00}	I_{11} Cluster_{11} W_{11}	I_{12} Cluster_{12} W_{22}	I_{13} Cluster_{13} W_{33}
I_{23} Cluster_{20} W_{30}	I_{20} Cluster_{21} W_{01}	I_{21} Cluster_{22} W_{12}	I_{22} Cluster_{23} W_{23}
I_{32} Cluster_{30} W_{20}	I_{33} Cluster_{31} W_{31}	I_{30} Cluster_{32} W_{02}	I_{31} Cluster_{33} W_{13}



$$
\begin{aligned}
O_{00} &= I_{00} \cdot W_{00} + I_{01} \cdot W_{10} + I_{02} \cdot W_{20} + I_{03} \cdot W_{30} \\
O_{01} &= I_{00} \cdot W_{01} + I_{01} \cdot W_{11} + I_{02} \cdot W_{21} + I_{03} \cdot W_{31} \\
O_{02} &= I_{00} \cdot W_{02} + I_{01} \cdot W_{12} + I_{02} \cdot W_{22} + I_{03} \cdot W_{32} \\
O_{03} &= I_{00} \cdot W_{03} + I_{01} \cdot W_{13} + I_{02} \cdot W_{23} + I_{03} \cdot W_{33} \\
O_{10} &= I_{10} \cdot W_{00} + I_{11} \cdot W_{10} + I_{12} \cdot W_{20} + I_{13} \cdot W_{30} \\
O_{11} &= I_{10} \cdot W_{01} + I_{11} \cdot W_{11} + I_{12} \cdot W_{21} + I_{13} \cdot W_{31} \\
O_{12} &= I_{10} \cdot W_{02} + I_{11} \cdot W_{12} + I_{12} \cdot W_{22} + I_{13} \cdot W_{32} \\
O_{13} &= I_{10} \cdot W_{03} + I_{11} \cdot W_{13} + I_{12} \cdot W_{23} + I_{13} \cdot W_{33} \\
O_{20} &= I_{20} \cdot W_{00} + I_{21} \cdot W_{10} + I_{22} \cdot W_{20} + I_{23} \cdot W_{30} \\
O_{21} &= I_{20} \cdot W_{01} + I_{21} \cdot W_{11} + I_{22} \cdot W_{21} + I_{23} \cdot W_{31} \\
O_{22} &= I_{20} \cdot W_{02} + I_{21} \cdot W_{12} + I_{22} \cdot W_{22} + I_{23} \cdot W_{32} \\
O_{23} &= I_{20} \cdot W_{03} + I_{21} \cdot W_{13} + I_{22} \cdot W_{23} + I_{23} \cdot W_{33} \\
O_{30} &= I_{30} \cdot W_{00} + I_{31} \cdot W_{10} + I_{32} \cdot W_{20} + I_{33} \cdot W_{30} \\
O_{31} &= I_{30} \cdot W_{01} + I_{31} \cdot W_{11} + I_{32} \cdot W_{21} + I_{33} \cdot W_{31} \\
O_{32} &= I_{30} \cdot W_{02} + I_{31} \cdot W_{12} + I_{32} \cdot W_{22} + I_{33} \cdot W_{32} \\
O_{33} &= I_{30} \cdot W_{03} + I_{31} \cdot W_{13} + I_{32} \cdot W_{23} + I_{33} \cdot W_{33}
\end{aligned}
$$

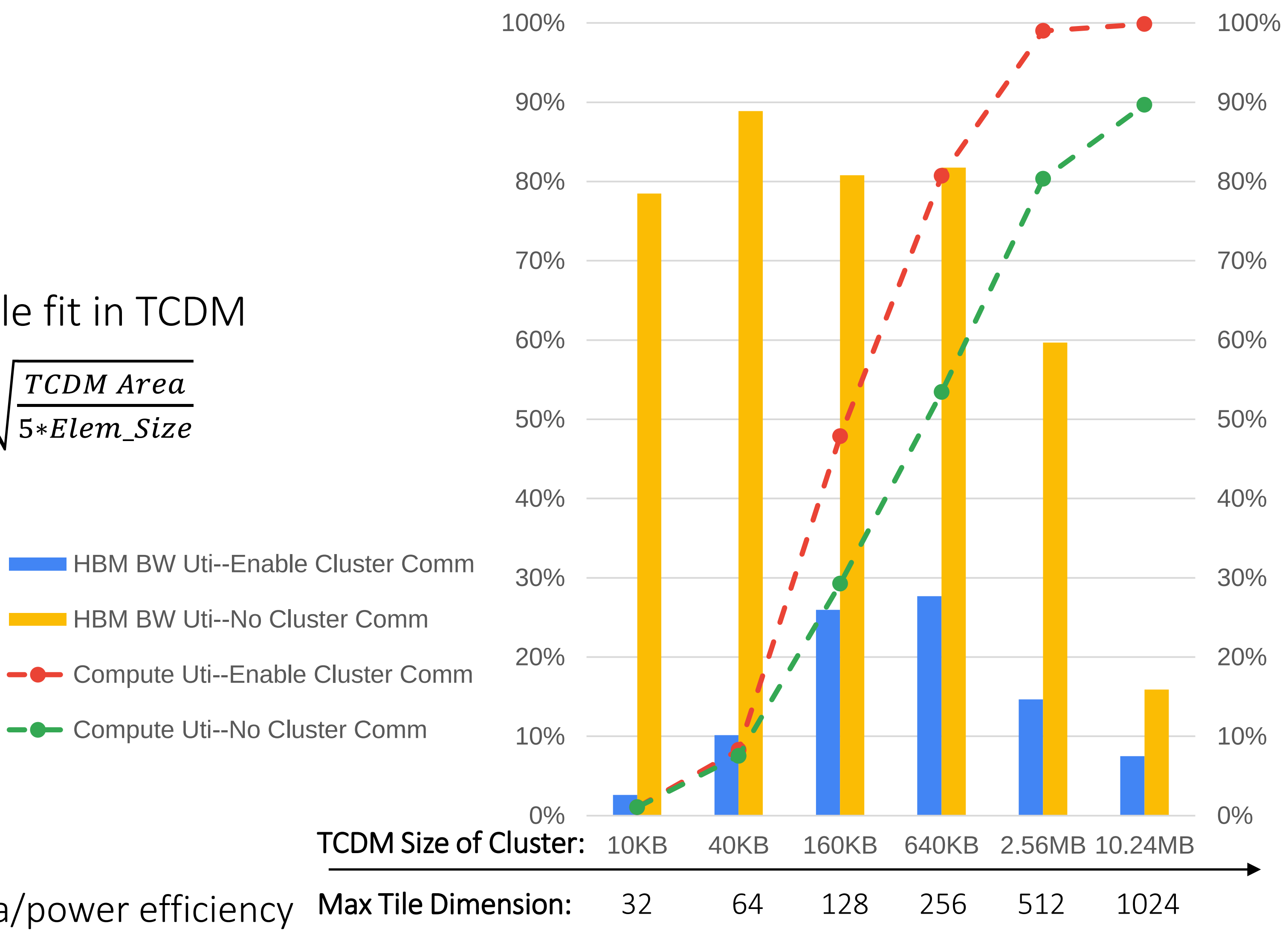
Experiment Setup: Enable Cluster-to-Cluster Comm or Not?

- RedMule
  - Each Cluster has One RedMule
  - RedMule CE array 128x32 (8TFLOPs @1GHz)
  - TCDM BW = 1024 GB/s
- HBM
  - Place HBM on left side
  - Each HBM CTRL mange 4 HBM2E channel
    - 256Byte address interleaving
    - Each HBM CTRL provide Max.BW = 205 GB/s
- NoC
  - Mesh 4x4
  - NoC link width = 2048 bits
    - Link BW = 256GB/s



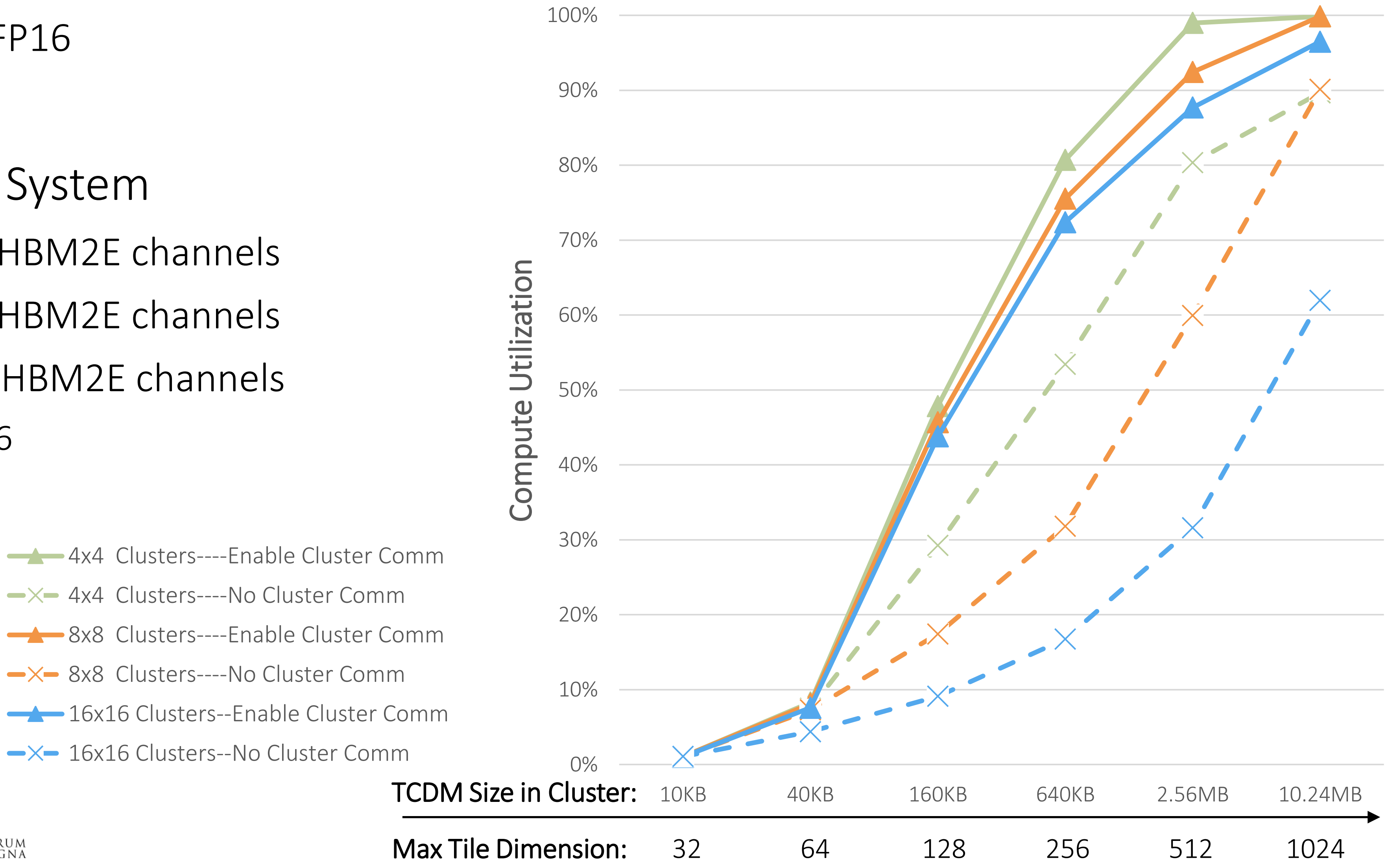
Experiment Result: Enable Cluster-to-Cluster Comm or Not?

- Benchmark GEMM
  - 16384 x16384 x16384
  - GEMM Elem Size = FP16
  - RedMule 128x32
- Varying Cluster TCDM Size
  - Tiling strategy: Max possible tile fit in TCDM
    - Tile dimension$M = N = K = \sqrt{\frac{TCDM\ Area}{5{*}Elem_Size}}$
- Results
  - No inter-cluster comm
    - HBM BW limited
    - Need 4x more TCDM size to saturate compute power
  - Enable inter-cluster comm
    - Reduce HBM traffic, better area/power efficiency



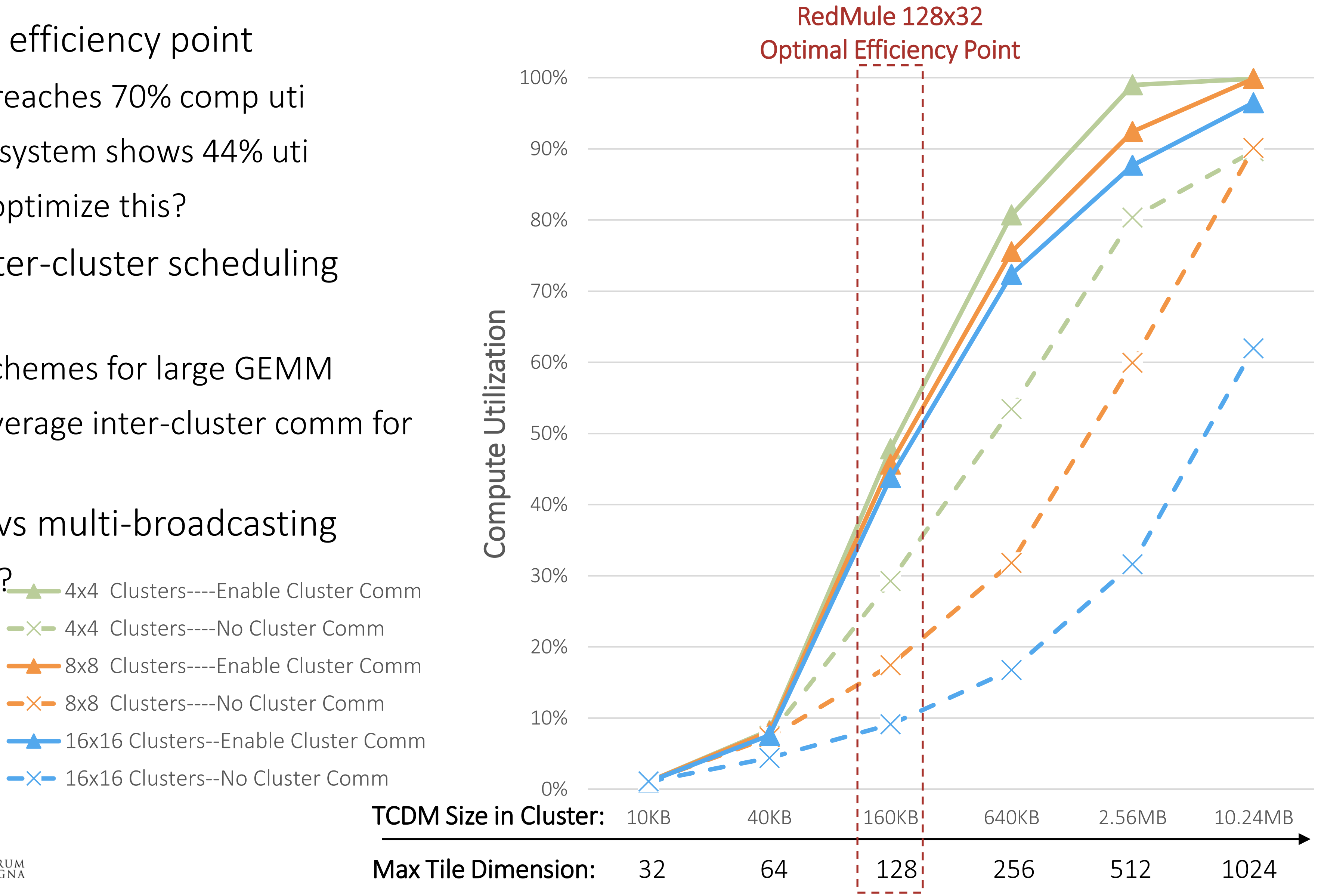
Enable Cluster-to-Cluster Comm is Scalable

- Benchmark GEMM
  - 16384 x16384 x16384
  - GEMM Elem Size = FP16
  - RedMule 128x32
- Scale-out SoftHier System
  - 4x4 Clusters + 16 HBM2E channels
  - 8x8 Clusters + 32 HBM2E channels
  - 16x16 Clusters + 64 HBM2E channels
    - 2048 TFLOPS @FP16
    - 3.2 TB/s HBM BW



Further Discussion …

- New Questions Comes Out
  - @ RedMule optimal efficiency point
    - RedMule in cluster reaches 70% comp uti
    - But end-to-end the system shows 44% uti
    - Why? How can we optimize this?
  - Tile mapping and inter-cluster scheduling scheme
    - Is there any more schemes for large GEMM
    - How can we also leverage inter-cluster comm for MHA?
  - Inter-cluster comm vs multi-broadcasting
    - Which one is better?



SoftHier Progress Update

Thomas Benz tbenz@iis.ee.ethz.ch

PULP Platform

Open Source Hardware, the way it should be!

@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

iDMA

- Various contributions and bugfixes merged in iDMA
  - Detailed tracer will soon be enabled in Snitch (port to GVSoC pending)
  - Release early next week

- Transposition engine implemented by student overhauled
  - Support for packed SIMD types is still ongoing effort
  - Update in Snitch is ongoing

SoftHier Progress Update

Luca Colagrande colluca@iis.ee.ethz.ch

PULP Platform
Open Source Hardware, the way it should be!

@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

Snitch cluster

- A lot of maintenance work:
  - Merged PRs: #165, #163, #161
  - WIP PRs: #115, #71, #158

- Implemented MHA and MLP layers for Snitch/Occamy [Link]
  - Work on streamlining data generation functions and scripts, to reuse base layer functions (e.g. GEMM, Layernorm) in composite layers (e.g. MHA, MLP)
  - Set up a proper Python package infrastructure to cross-reference these functions

- WIP on full encoder block