

A 3D-Stacked 8.4TFLOPS/4.3W Many-Core Domain-Specific Processor for AI-Native Radio Access Networks

Marco Bertuletti

mbertuletti@iis.ee.ethz.ch

Yichao Zhang

yiczhang@iis.ee.ethz.ch

Diyoun Shen

dishen@iis.ee.ethz.ch

Alessandro Vanelli-Coralli

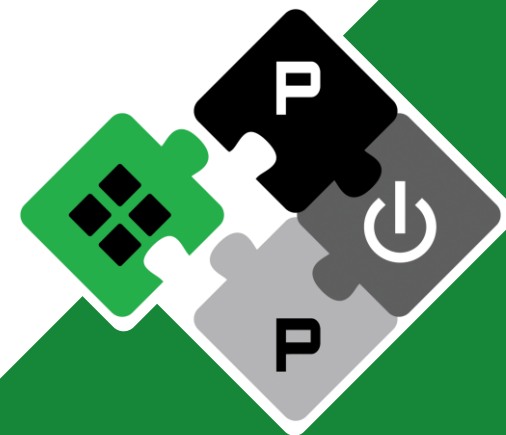
avanelli@iis.ee.ethz.ch

Frank Kagan Gurkaynak

kgf@iis.ee.ethz.ch

Luca Benini

lbenini@iis.ee.ethz.ch



PULP Platform

Open Source Hardware, the way it should be!

pulp-platform.org

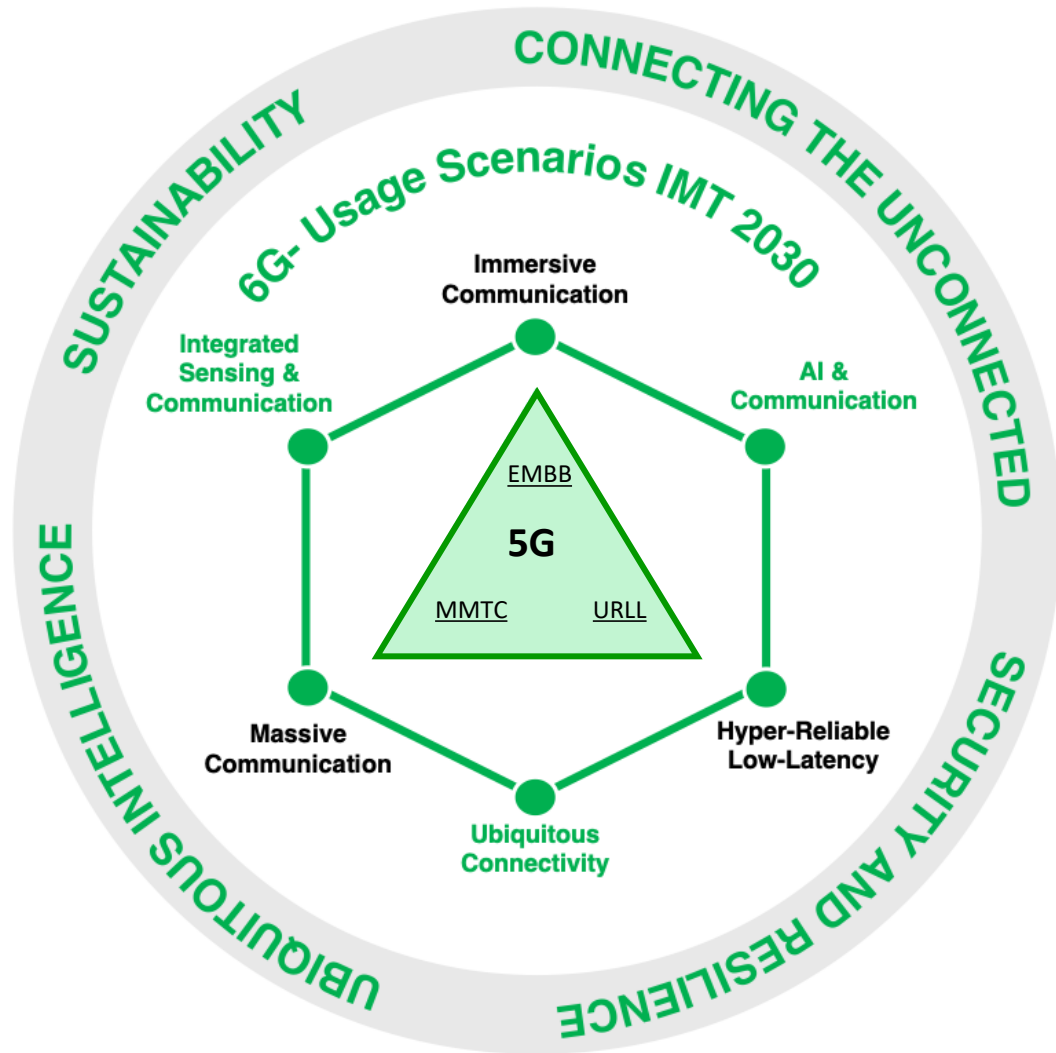
@pulp_platform

company/pulp-platform

youtube.com/pulp_platform

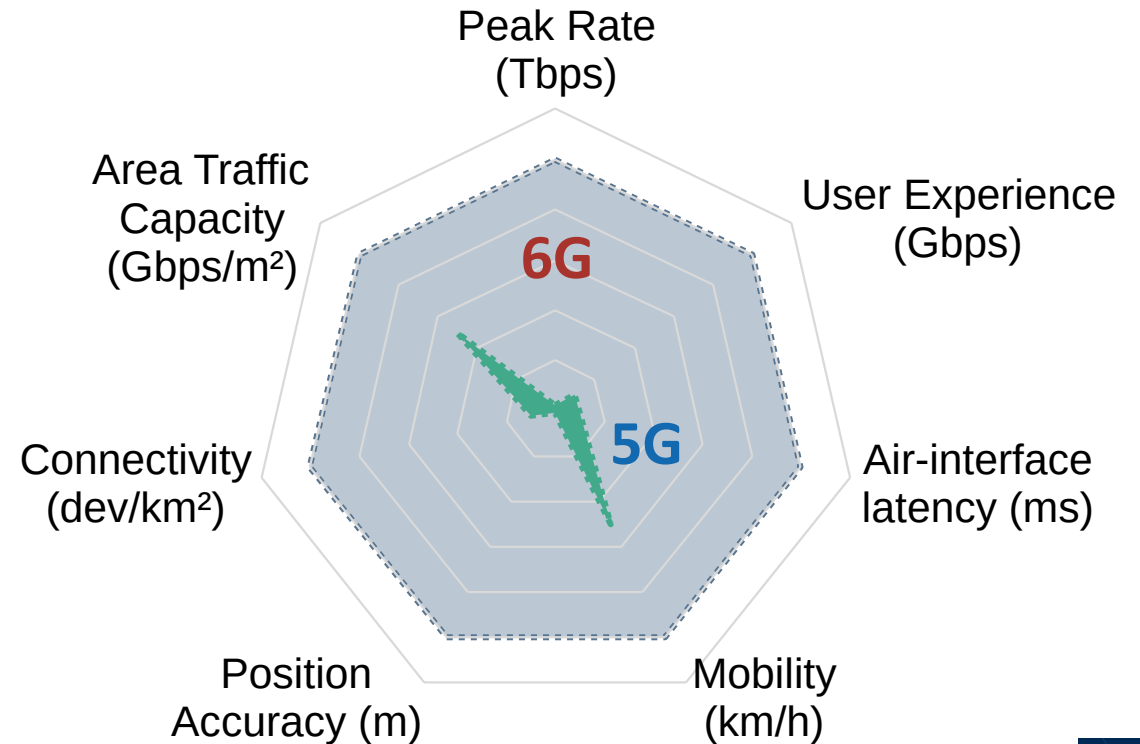


6G is coming fast!



New stringent KPIs

- From 1ms to 0.1ms latency
- From 20Gbps to 1000Gbps throughput



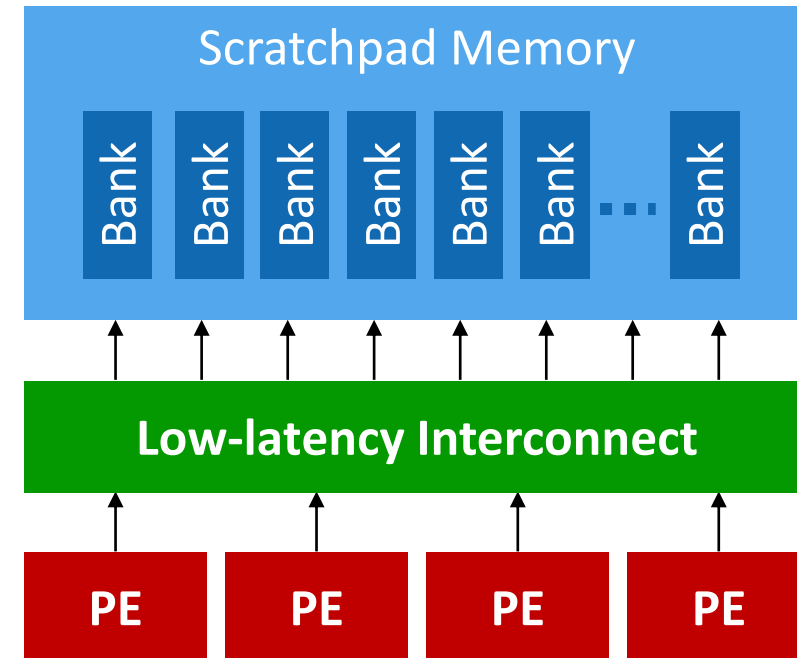
gNB-RU Processing

Must be Software-Defined for flexibility after deployment

Is largely data-parallel

Need to scale-up the HW-architecture and the programming model

Shared Memory Cluster architecture + SPMD execution model



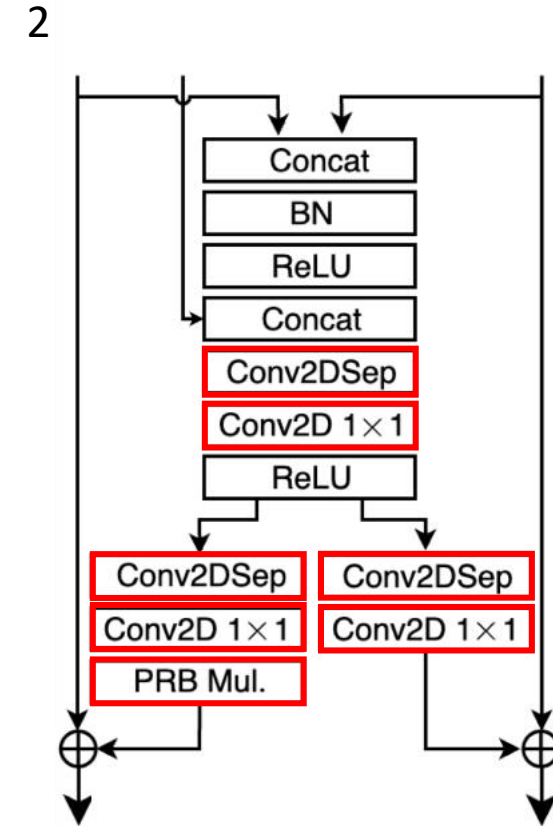
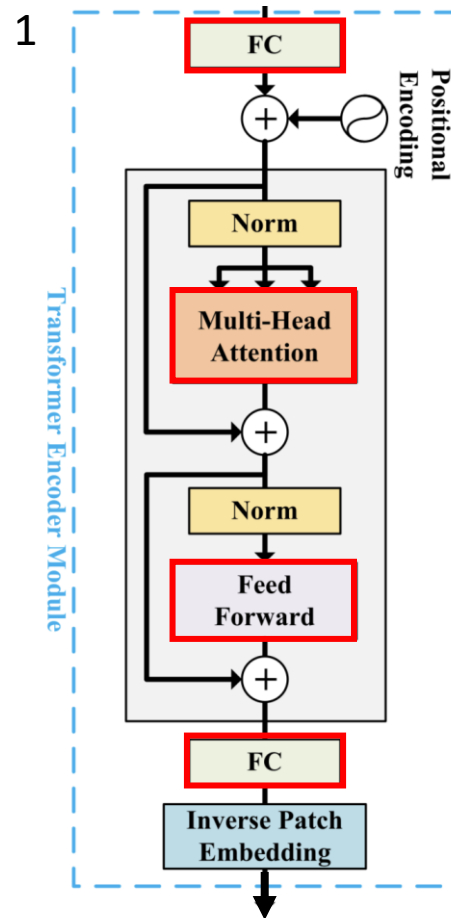
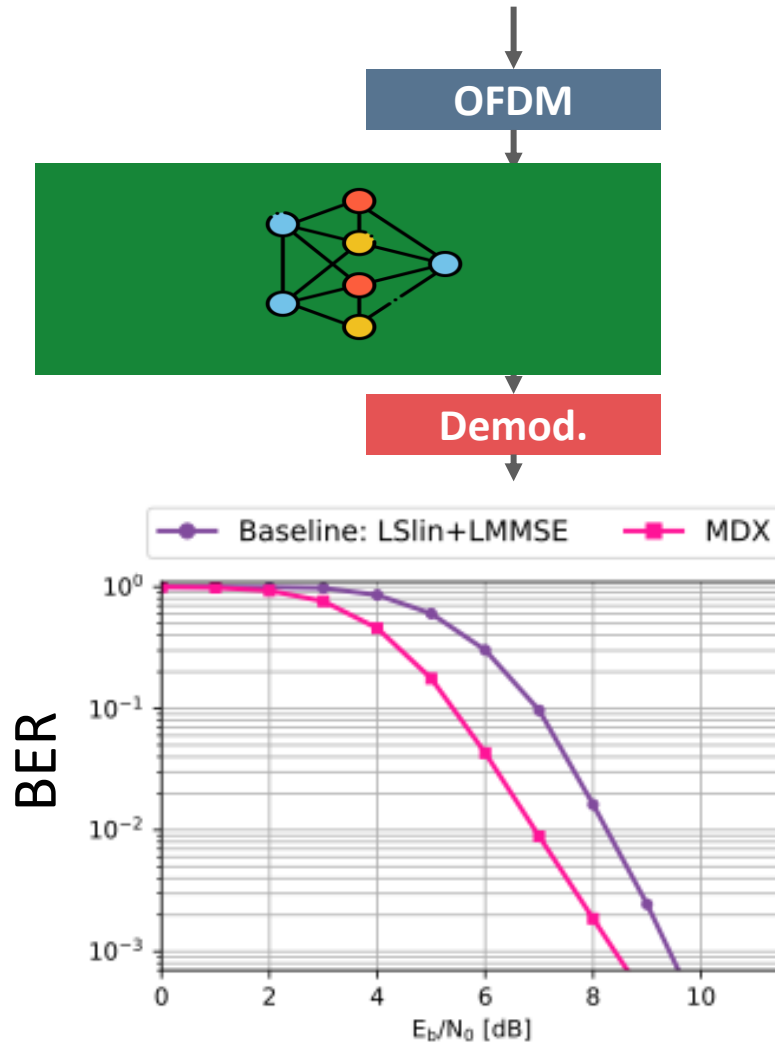
TeraPool is a Physically Feasible Shared-Memory 1024-PE cluster

- 920MHz@TC, 81 mm² in GlobalFoundries' 12P+ FinFET,
- 1.89TOPS peak-performance
- Interconnect-scaling does not impact energy efficiency 50-100GOPS/W

github.com/pulp-platform/mempool



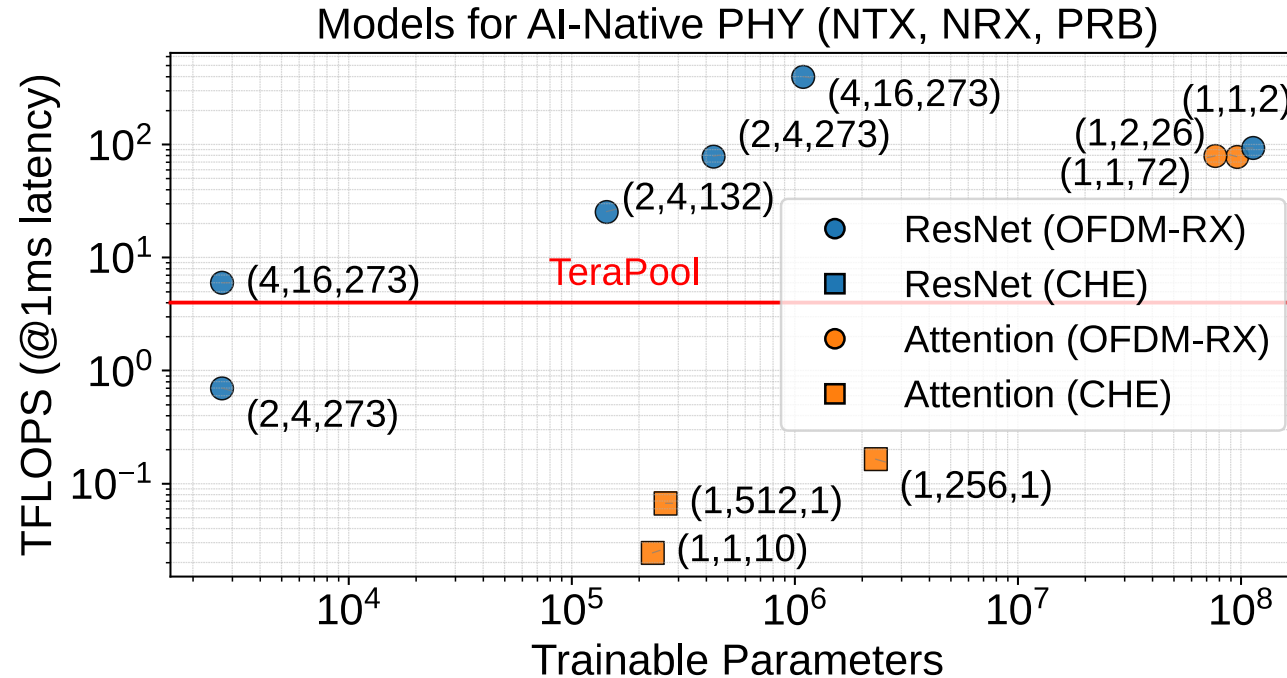
AI-Native RAN Improves the receiver performance!



¹ Liu GLOBECOM'23,
² Abdollahpour, GLOBECOM'25

Common Ground? GEMM is the dominant operation

AI-Native RAN Increases Compute in the gNB-RU!



Requires Domain-Specialization to keep high energy-efficiency and flexibility on heterogeneous workloads (GEMM + “old” Wireless)

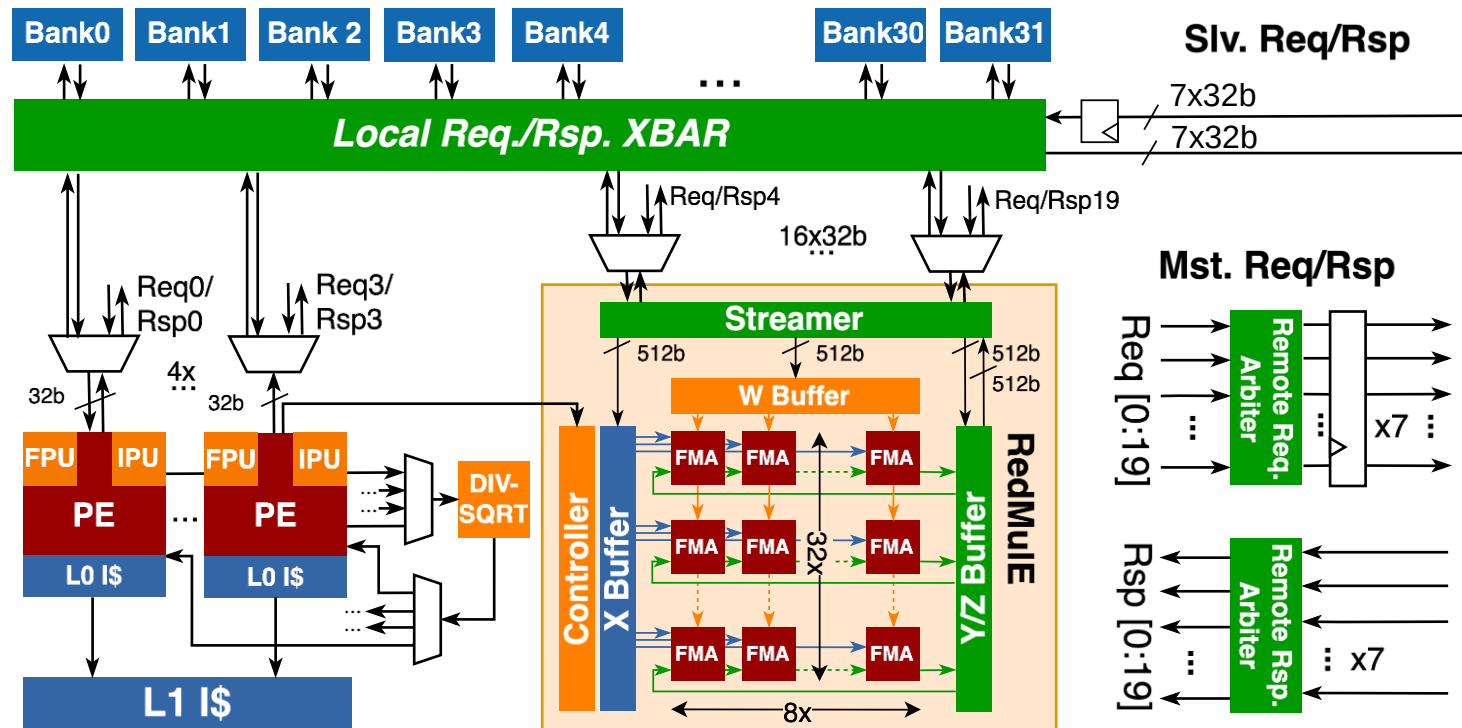
Contributions: scalable, efficient AI-native PHY processing



- **TensorPool: A Domain-specialized, Many-Core for AI-Native RAN**
 - Parallel throughput of 16 TEs for GEMM computation
 - Flexibility of 256 PEs for activations & legacy wireless computations
- **Combined execution of AI-Native RAN models on TEs and PEs**
- **Physical design and 3D integration**

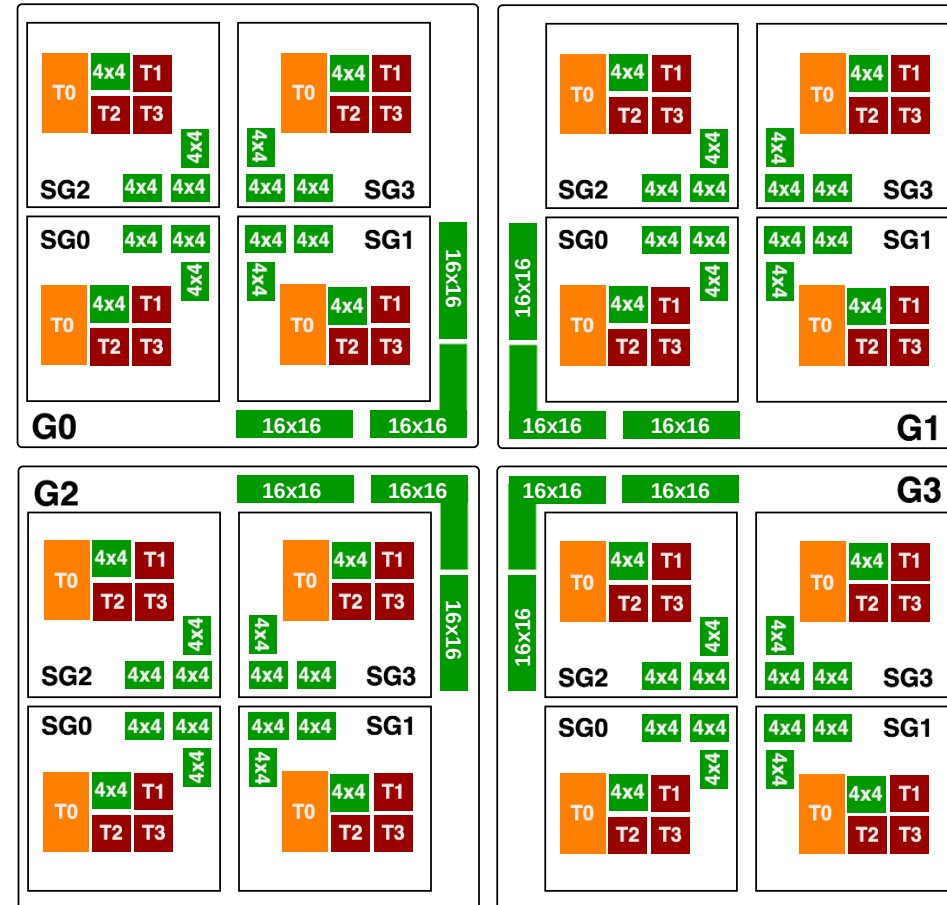
A domain-specialized flavour of the Pool

- 8x32 RedMulE engine provides 256 MAC/cycle, 16 RedMulE Tiles + 48 Snitch Tiles (256 PEs)

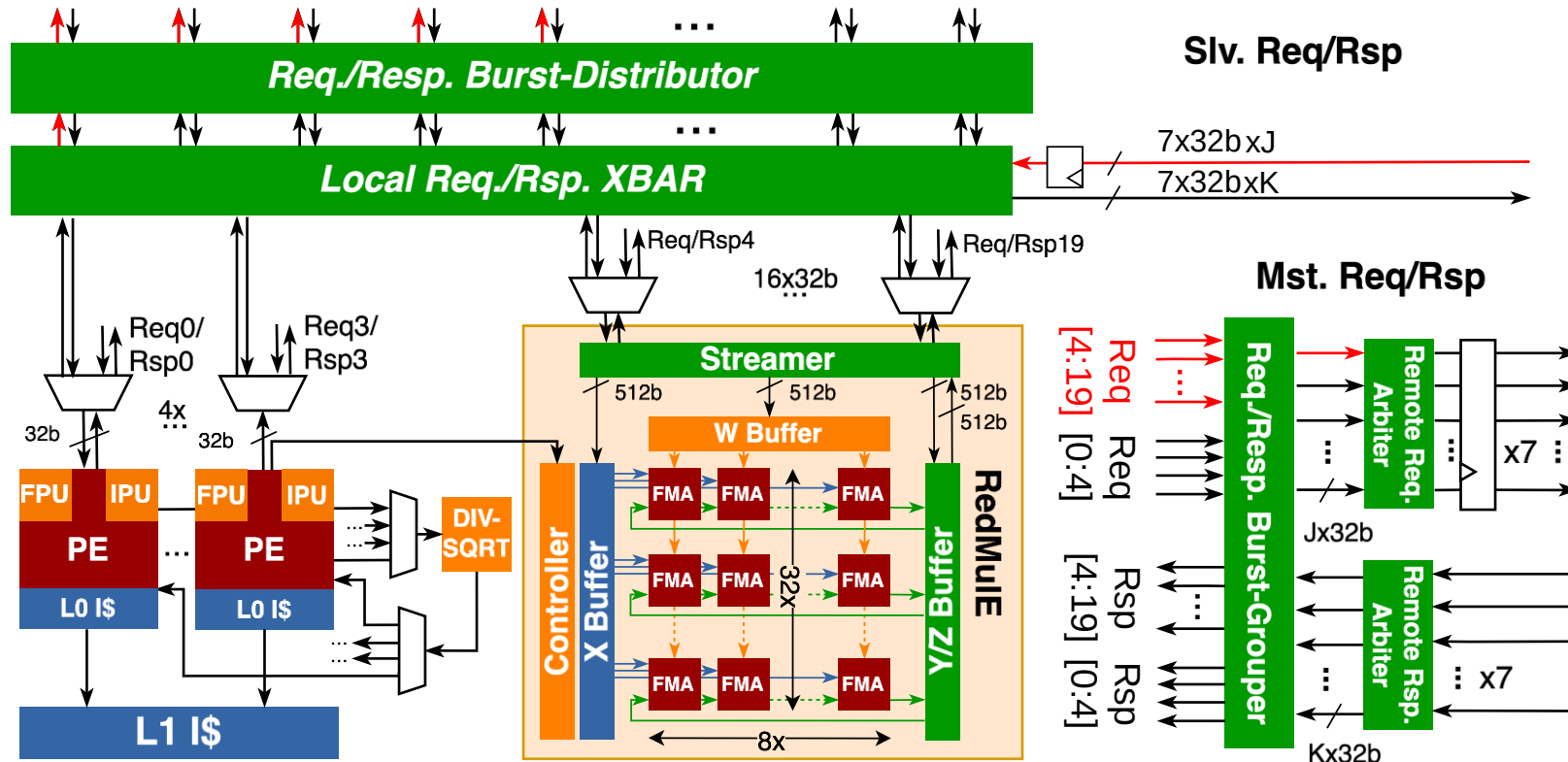


TensorPool (TOP view)

- Hierarchical Implementation
 - 4 PEs (+ 1 Tensor Engine) / Tile
 - 4 Tiles / SubGroup
 - 4 SubGroups/Group
 - 4 Groups / Cluster
 - Peak Performance:
 - (1 TE + 4x4 PEs) / SG
 - $(1 \times 256_{TE} + 2_{f16} \times 16_{PE}) \times 16 \text{ SGs}$
- = **4608** MAC_{f16b} /Cycle/Cluster



Scaling Memory Bandwidth to TEs requirements

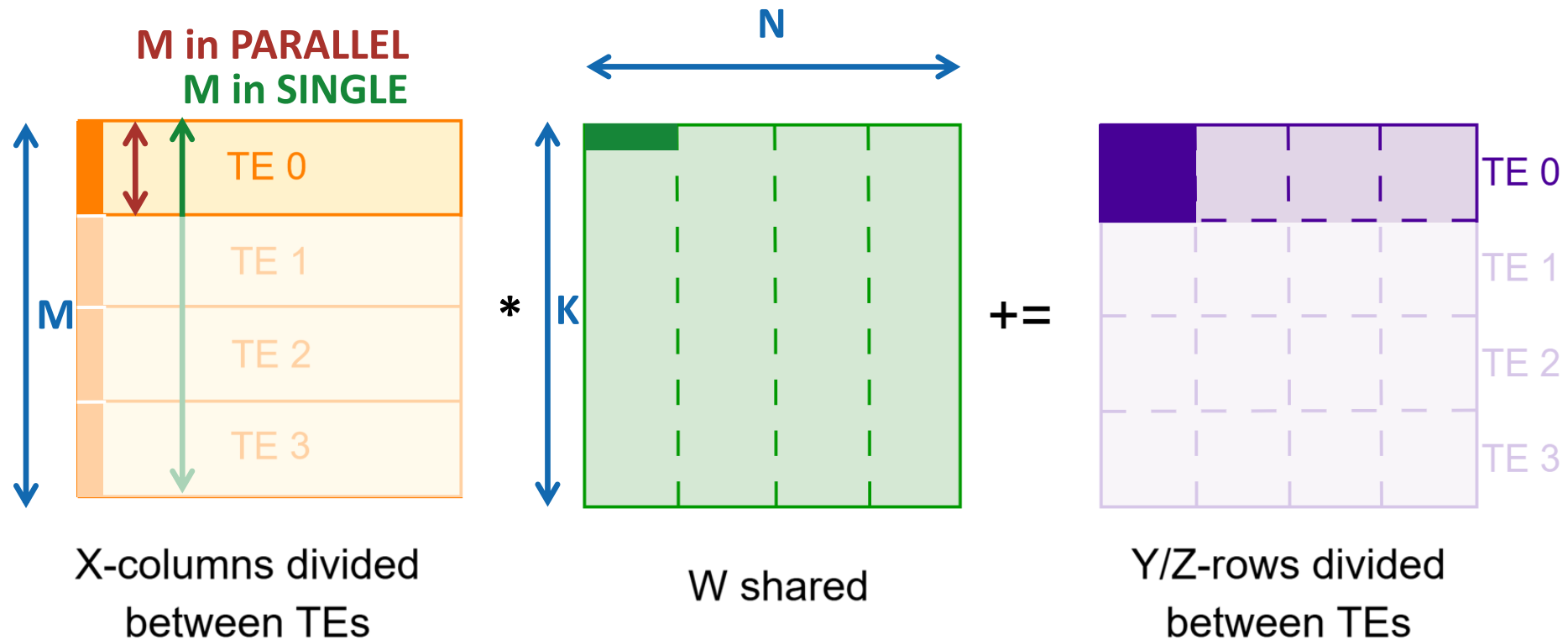


- Streamer with ROBs
- Reads can be bursted = send just the first address
- Write request and read response data-field can be widened

Efficient GEMM Parallelization Scheme



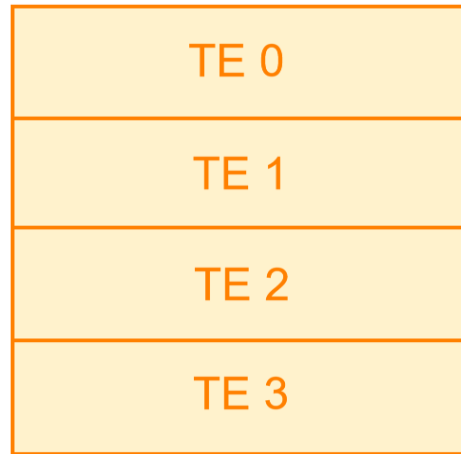
- The problem is split over M dimension
- TEs all access the W matrix in parallel



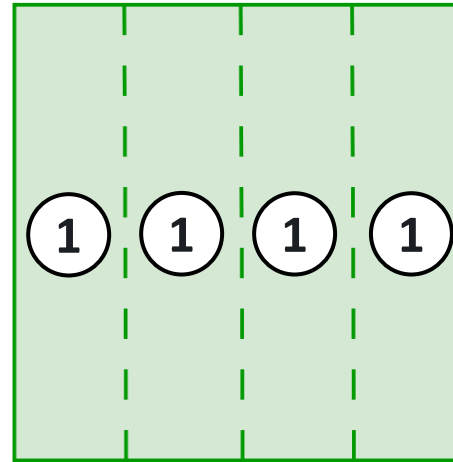
GEMM Parallelization: Offset on W columns



- We have to avoid conflicts on the shared W matrix
- Implemented offset on W columns

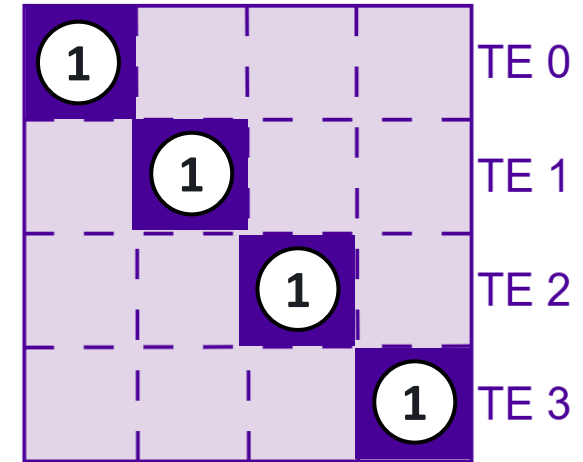


X-columns divided
between TEs



W shared

$+=$

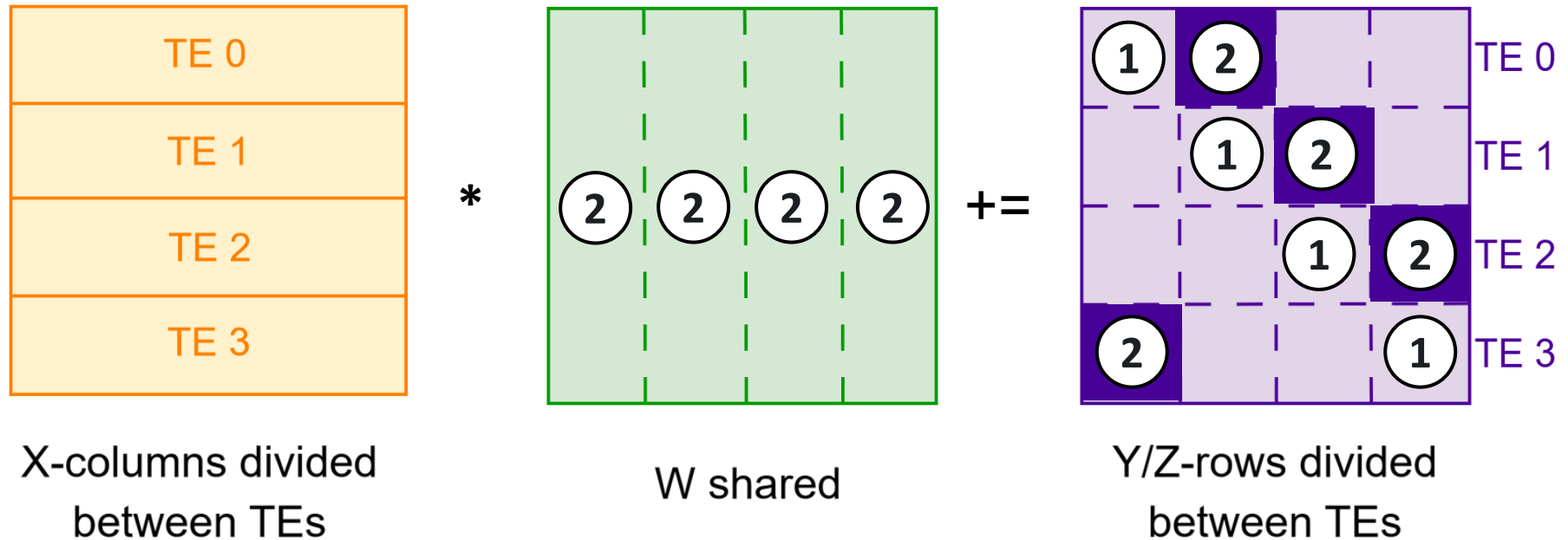


Y/Z-rows divided
between TEs

GEMM Parallelization: Offset on W columns



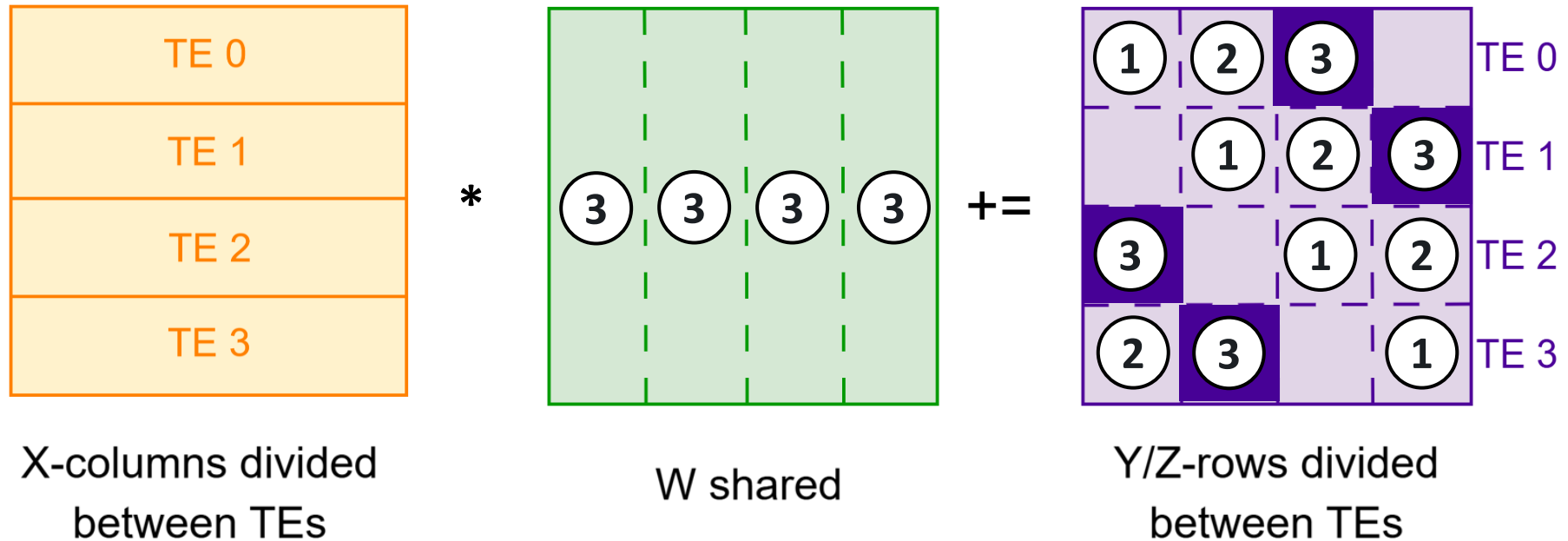
- We have to avoid conflicts on the shared W matrix
- Implemented offset on W columns



GEMM Parallelization: Offset on W columns



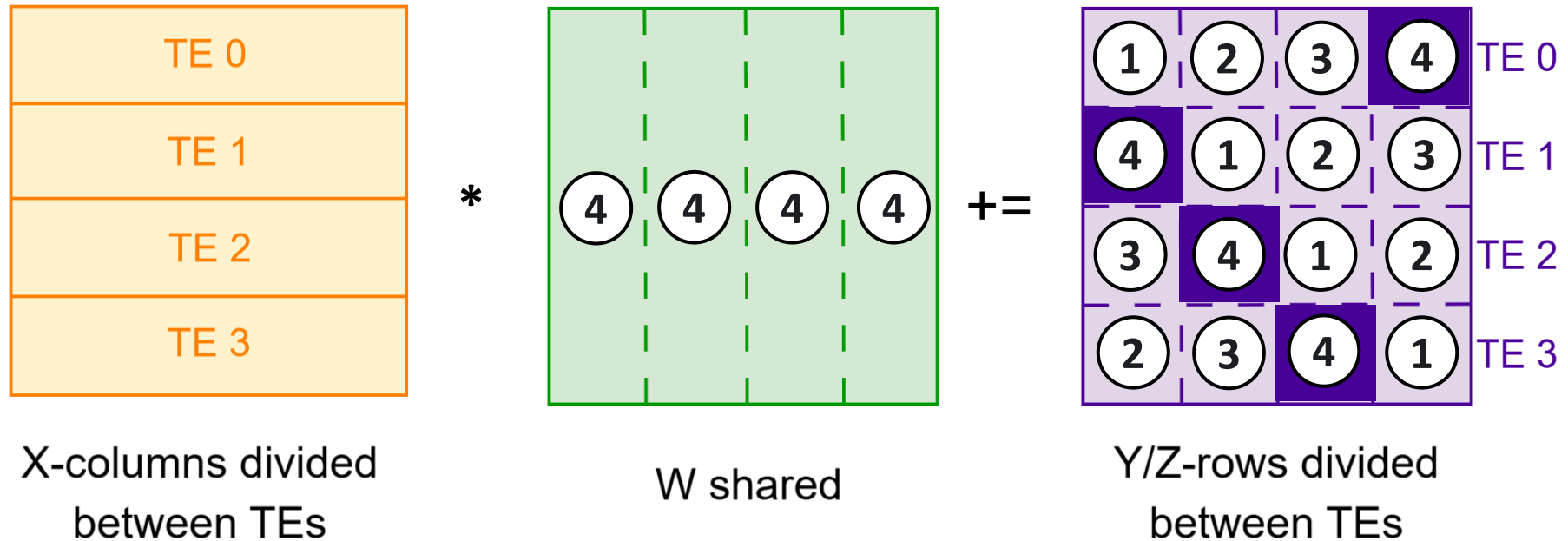
- We have to avoid conflicts on the shared W matrix
- Implemented offset on W columns



GEMM Parallelization: Offset on W columns



- We have to avoid conflicts on the shared W matrix
- Implemented offset on W columns

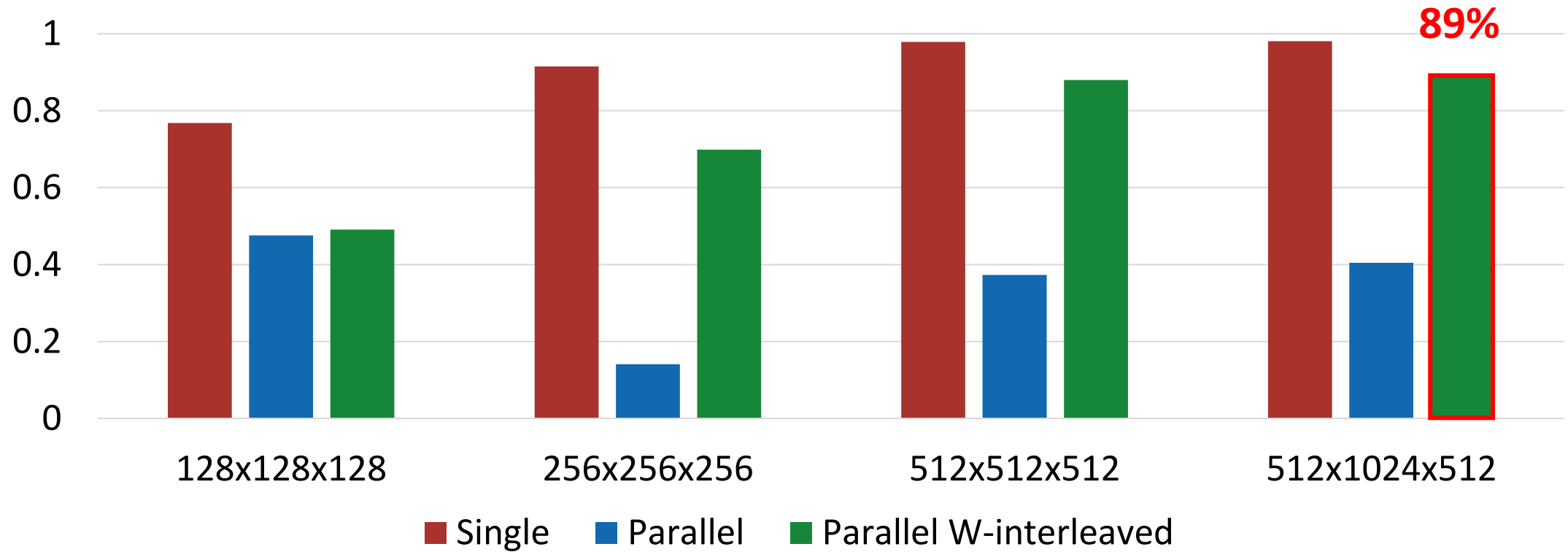


High Parallel GEMM Utilization



94% Single-TE utilization, **89%** Parallel-TE utilization

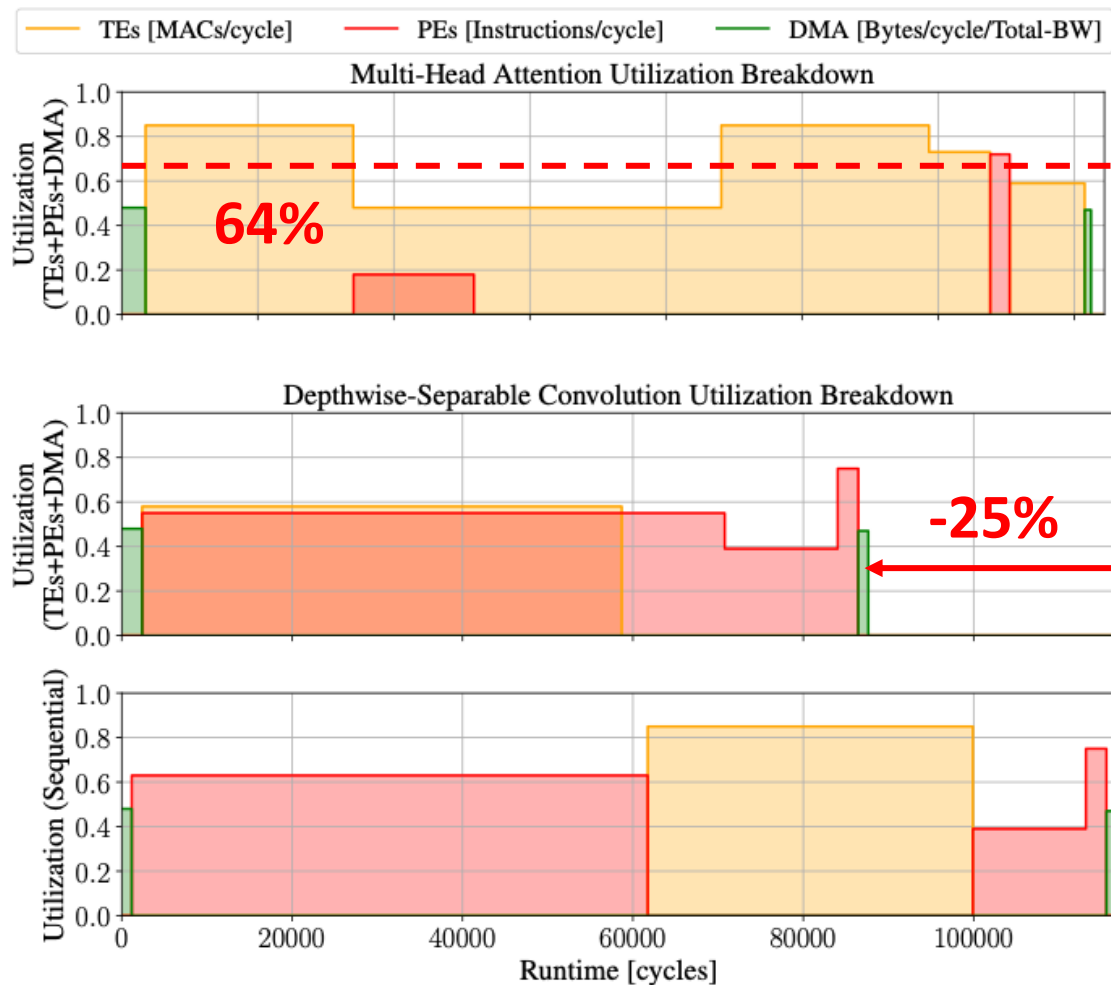
GEMM Utilization



Parallelization of uplink models



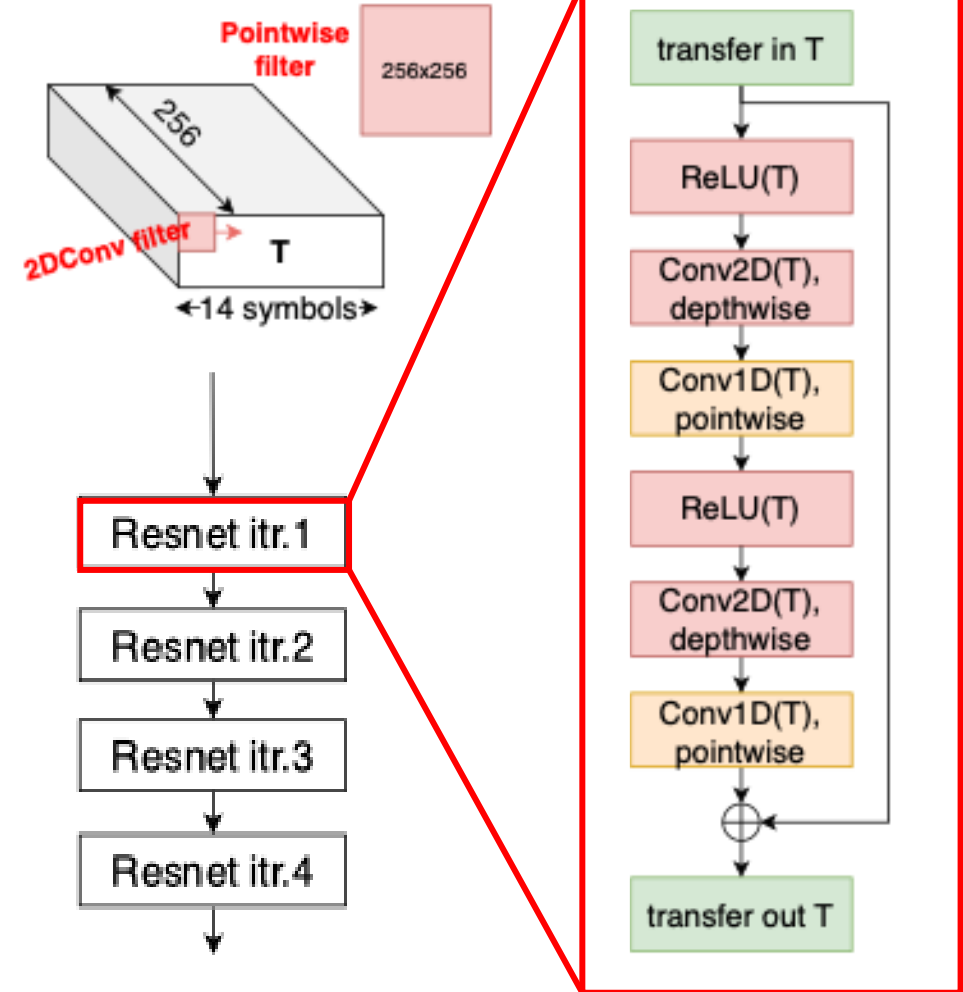
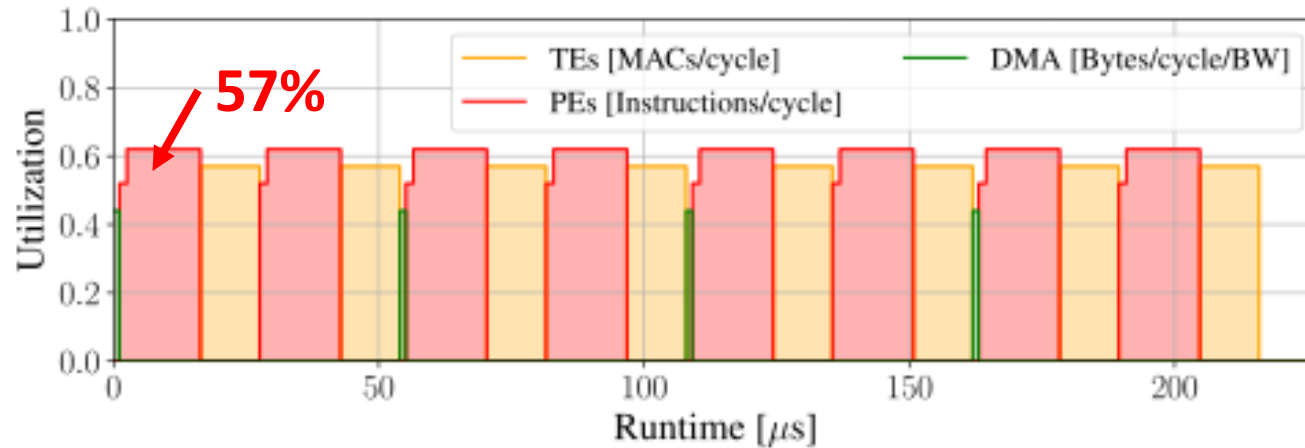
- **64%** Average TE utilization on Multi-Head Attention
- Up to **25%** runtime reduction on Concurrent operation of DMA, TEs, PEs in Depthwise Convolution



End2End Inference on a ResNet based model < 1ms



0.23ms per transceiver, 57% GEMM utilization



PnR in TSMC N7

- **6x** more throughput on GEMM
(2x FMAs and 3x utilization 89% vs 30%)
- **9.1x** Area&Energy Efficiency

	TensorPool	TeraPool
Node	7nm	12nm
Area [mm2]	26.6	81.7
T [ns]	1.1	1.1
Peak FLOPs/cycle	9,216	4,096

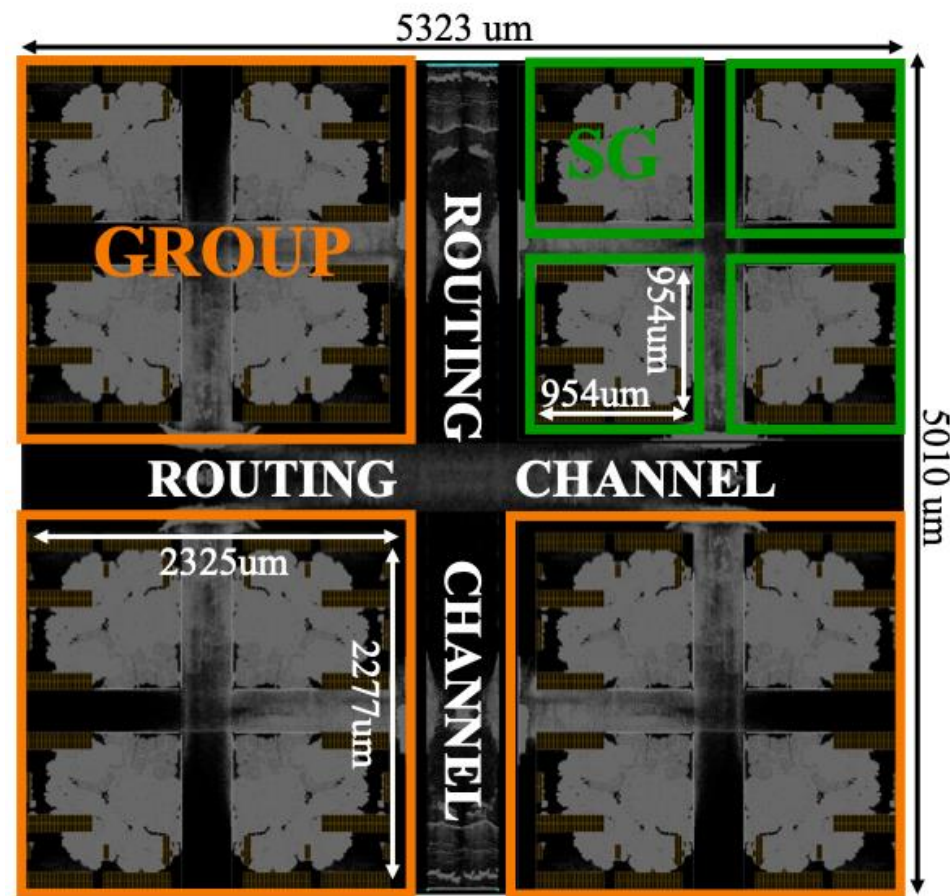
GEMM Utilization	88.94%	29.74%
GEMM FLOPs/cycle	7286	1218
GEMM Power	4.3	6.3
GEMM GFLOPs/s/W/mm2	57.53	6.24

2.25x

3x

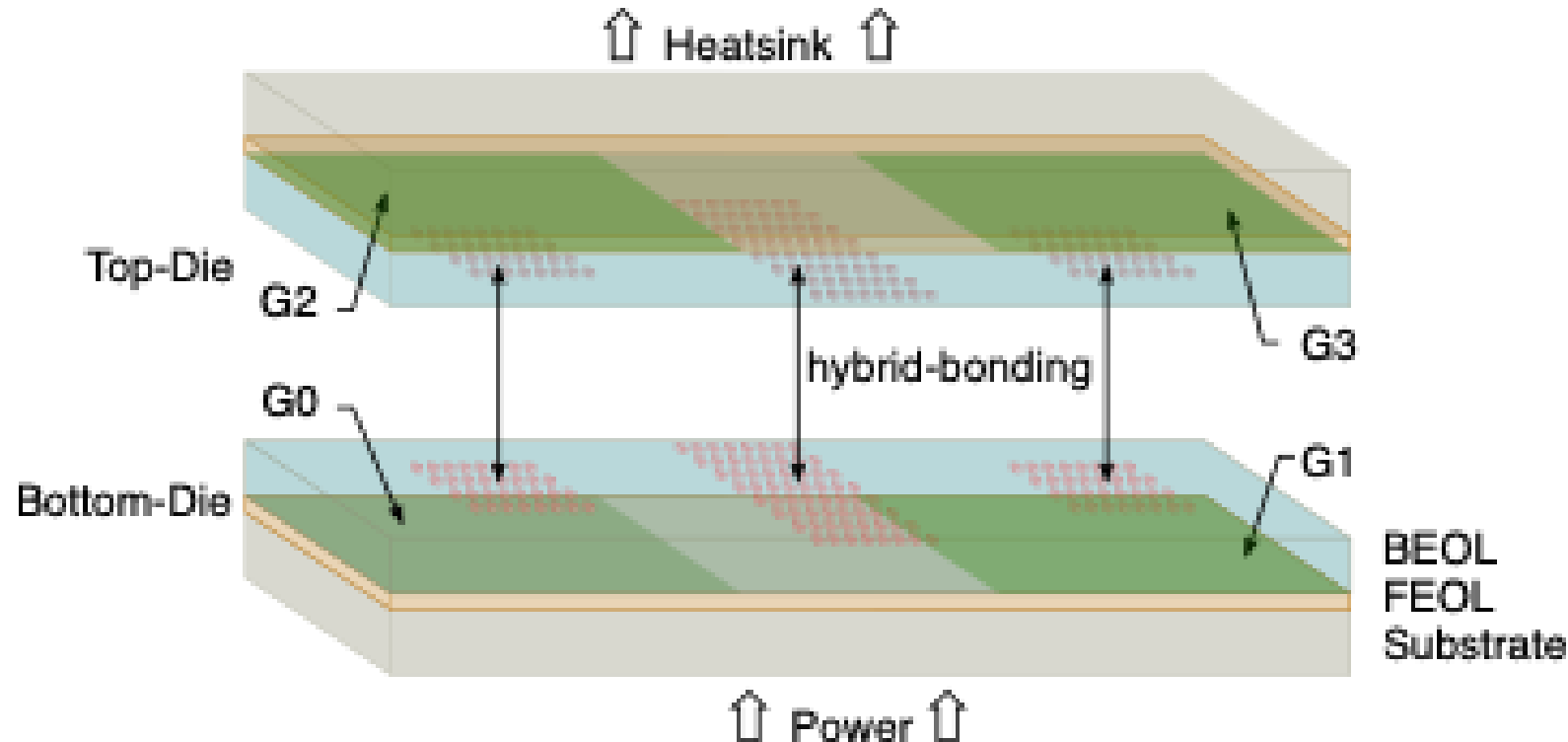
6x

9.1x



3D is an area efficiency booster

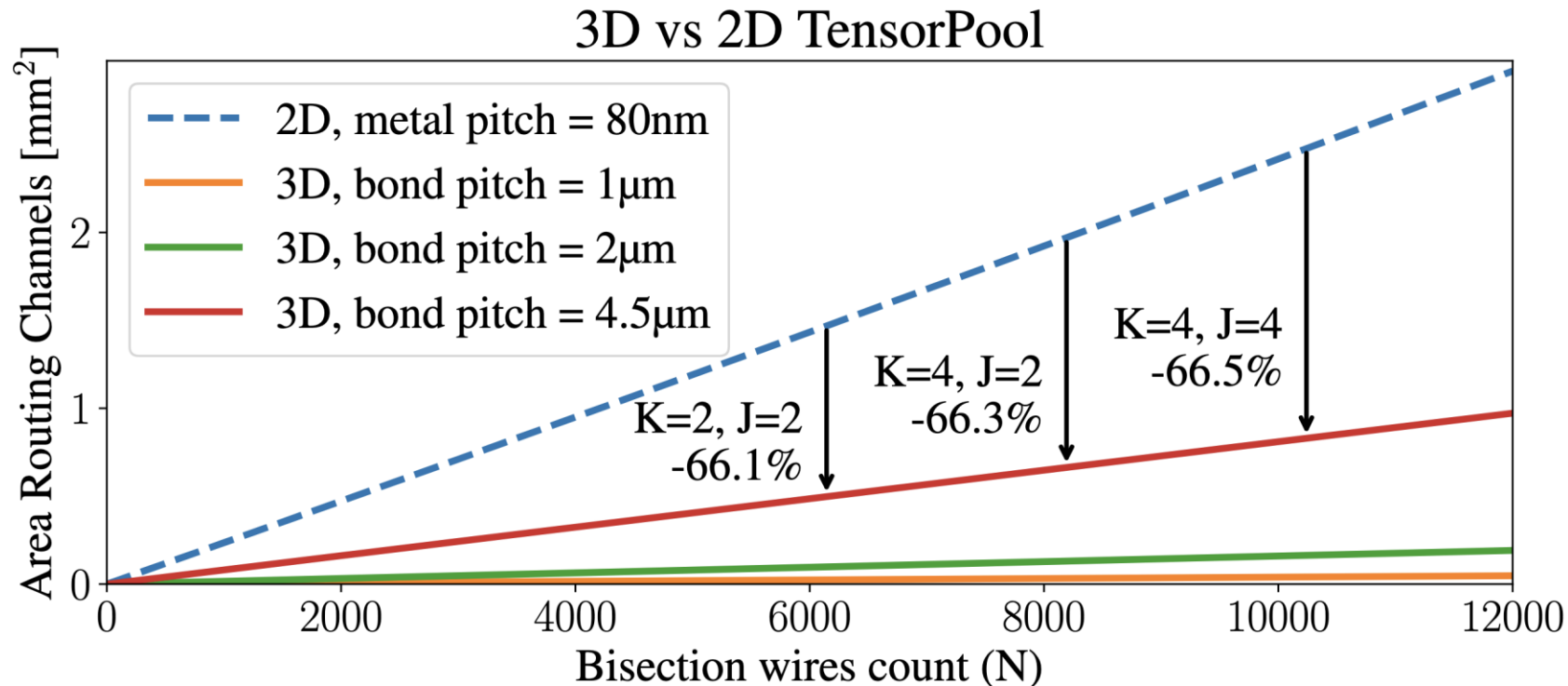
- 3D eliminates the horizontal channel → **2.3x** less footprint vs 2D
- Routing the Group to Group connections shows no timing-degradation w.r.t 2D



Quantifying the Area-Efficiency Gains



- The area of the channels is there to route the point-to-point connections
- The wider the connection the larger the channel and the gain when you go 3D



SoA Highlight on Perf. & Area for gNB-RU



Similar Many-Core architectures have been proposed for AI-Native RAN

	NVIDIA L4 (Ada-Lovelace)	TensorPool-3D
Tech. Node	4nm	7nm
f [MHz]	2040	900
Num. L1-Clusters	60	1
L1-Size / Num. TEs / Num. PEs	128 KiB / 240 / 7424	4 MB / 16 / 256
Area L1-Cluster [mm ²]	1.7▼	22.94
GOPS / L1-Cluster	2017 / 1390*	6623
GOPS / Area L1-Cluster	1680 / 267**	288
Power	72	4.32

**Larger L1 =
more data-reuse**

**more FLOPS per
cluster on equal Area
Efficiency**

▼ Based on die shot from <https://locuza.substack.com/p/nvidias-ad102-officially-revealed>

<https://www.tomshardware.com/pc-components/gpus/gb202-die-shot-beautifully-showcases-blackwell-in-all-its-glory-gb202-is-24-percent-larger-than-ad102>

* Normalized to the frequency of A100 (implemented in 7nm), ** Normalized by (7/4)² to account for technology scaling

Summary

- Low-Latency interconnect + Outstanding read/writes + Bursts + Grouped Req/Resp =
- **94%** FMA utilization single-TE
- **89%** FMA utilization 16-TEs
- Improvement on TeraPool thanks to domain specialization:
 - **6x** more throughput on GEMM
 - **9.1x** Area&Energy Efficiency

TensorPool

