NextRAN-AI – 09/05/2025

Integrated Systems Laboratory (ETH Zürich)

Marco Bertuletti mbertuletti@iis.ee.ethz.ch
Yichao Zhang yiczhang@iis.ee.ethz.ch
Mahdi Abdollahpour mahdi.abdollahpout@unibo.it
Alessandro Vanelli-Coralli avanelli@iis.ee.ethz.ch
Luca Benini lbenini@iis.ee.ethz.ch

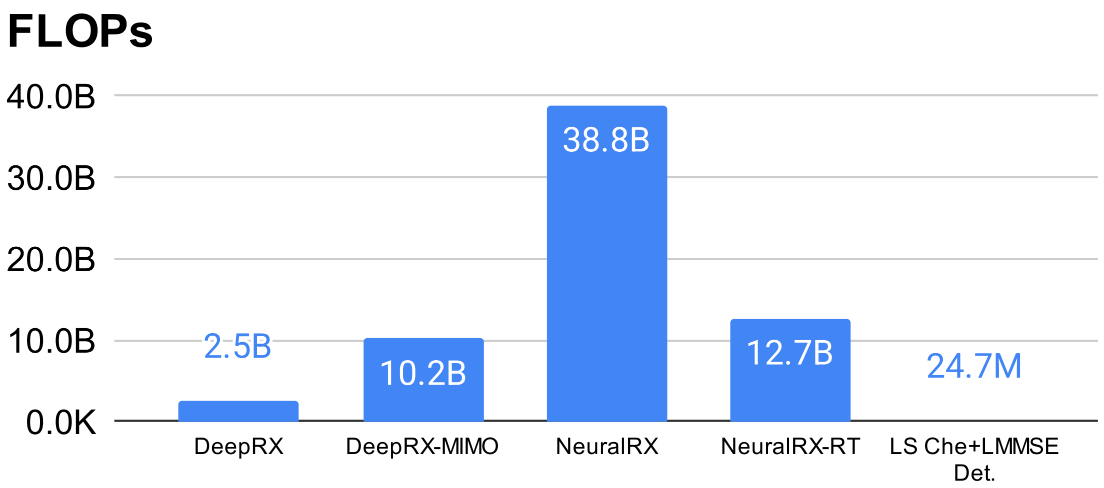
PULP Platform
Open Source Hardware, the way it should be!

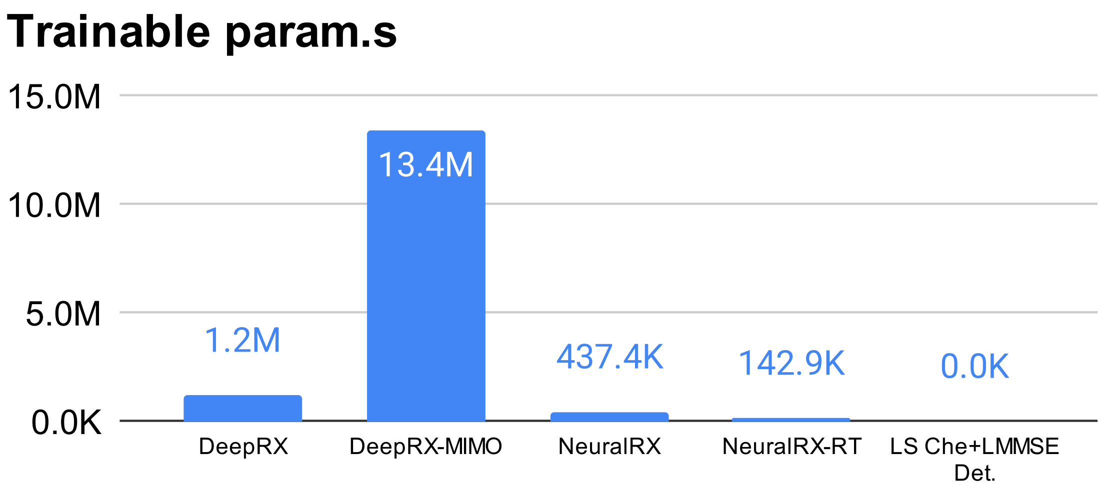
pulp-platform.org
@pulp_platform
company/pulp-platform
youtube.com/pulp_platform

DNNs for combined CHE&MMSE

Name	Processing	NSC	NRXxNTX	Modulation	Model	SNR@BER10^{-3}*
[1] Deep-RX	Ch.Est. + Det.	312	2x1	16QAM	ResNet	6dB vs 8.5dB
[2] Deep-RX MIMO	Ch.Est. + Det.	312	16x4	16QAM	ResNet	18dB vs 22dB
[3] Neural-RX	Ch.Est. + Det.	2604	16x4	16QAM	ResNet	2.6dB vs 3.1dB
[4] Neural-RX RT	Ch.Est. + Det.	2604	16x4	16QAM	ResNet	3.2dB vs 3.1dB

* vs XXdB required for conventional LMMSE Ch.Est. & MMSE MIMO detection





[5] Dianxin'22

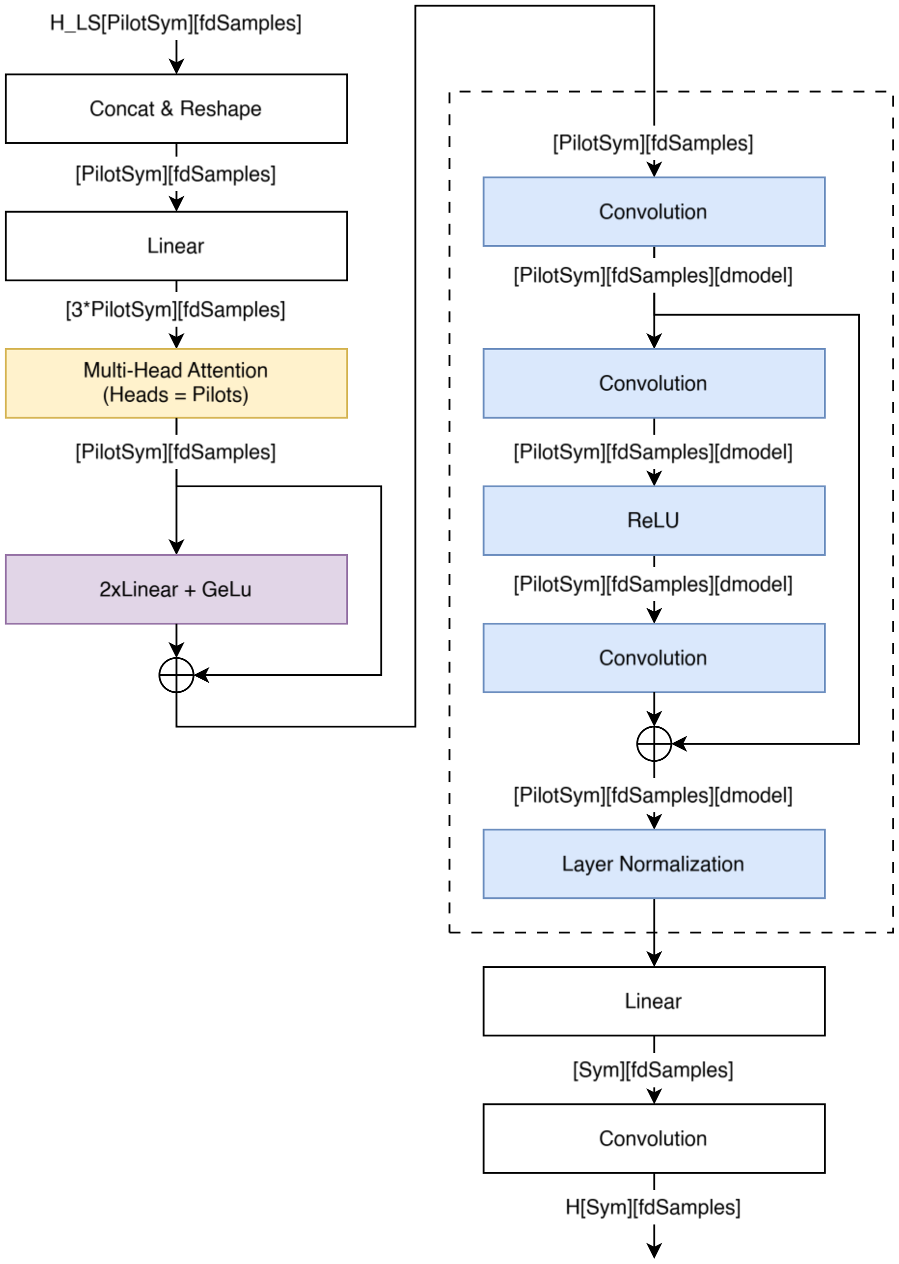
Self-Attention mechanism is deployed to achieve improved channel estimation

Channel gains not equally important (correlation matrix not diagonal)

- **Encoder** = transformer, focuses on LS-estimates strongly correlated with channel predictions
- **Decoder =** ResNet, receives selected features and produces the channel estimate

https://arxiv.org/pdf/2204.13465

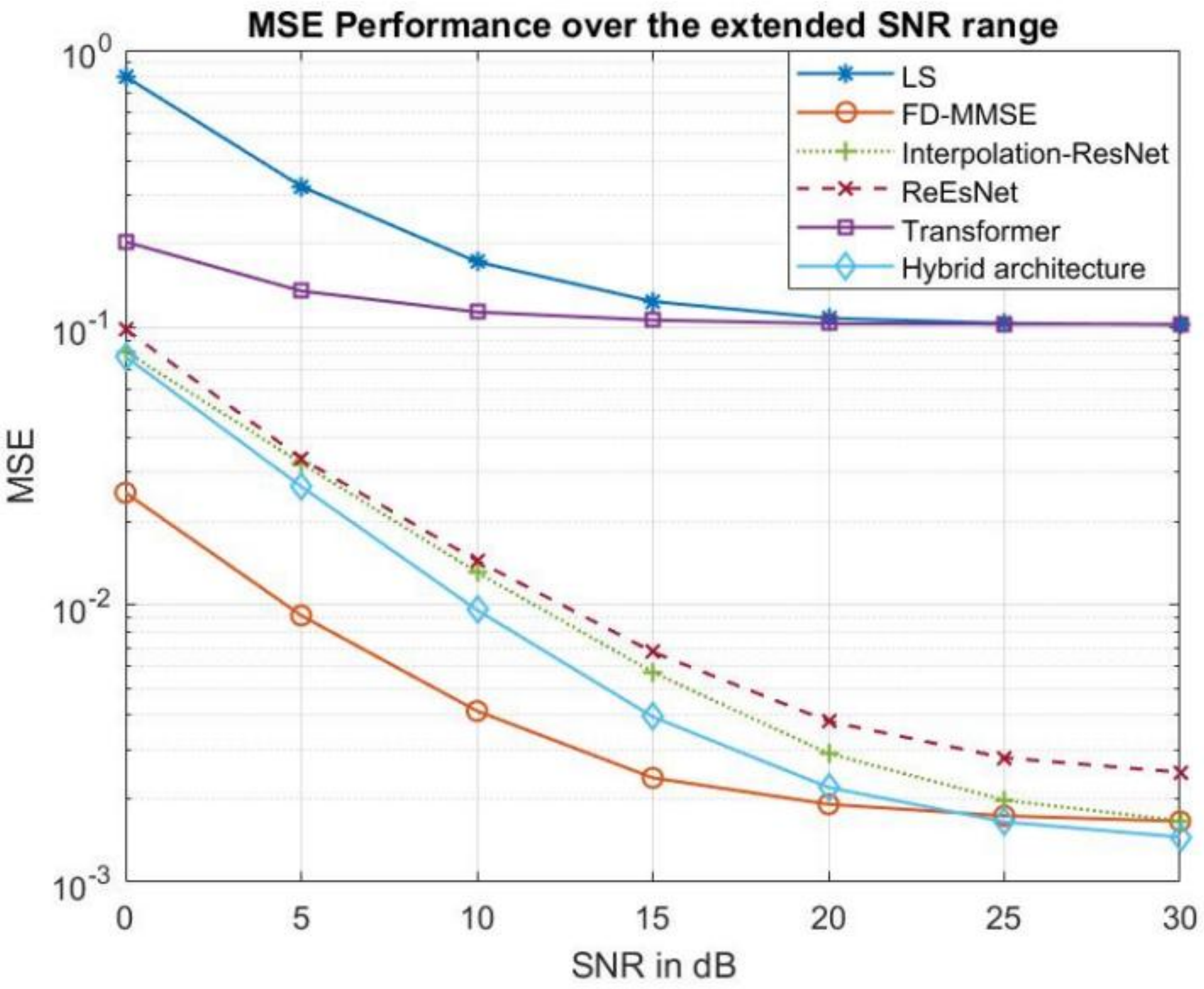
https://github.com/dianixn/Attention_Based_Neural_Networks_for_Wireless_Channel_Estimation



[5] Dianxin'22

Channel estimation transformer-based methods achieve superior performance wrt **ResNet-based** methods

(FD-MMSE assumes **prior statistical knowledge** of the channel and non-causal statistical information)



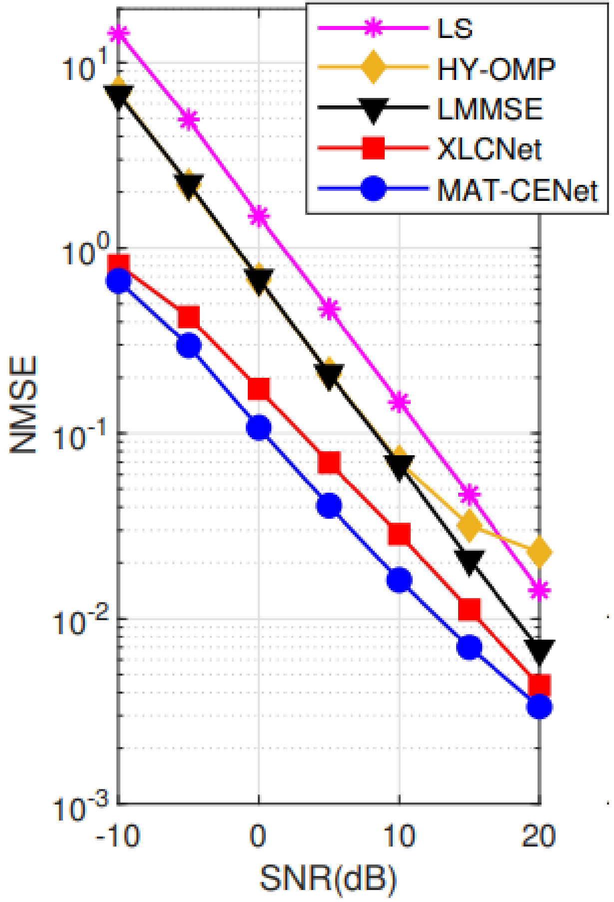
[6] Shuangshuang’24

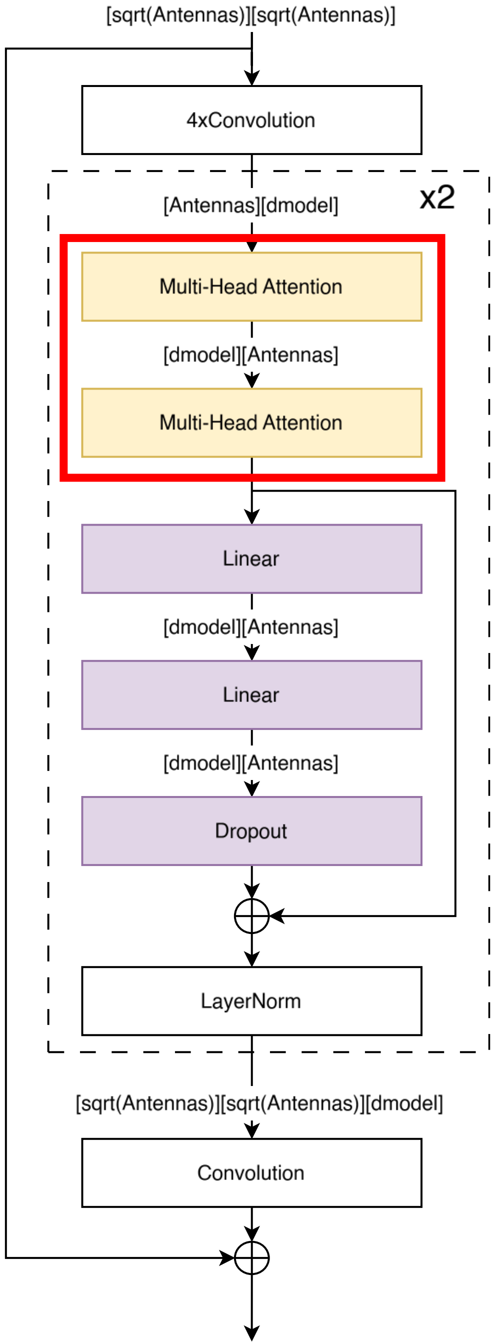
To counteract near/far field effects,

→ Mixed attention

- **Feature-Attention** enhances the channel, suppressing noise and secondary features,
- **Spatial-Attention** focuses on the relation between different spatial positions. The model can recognize spatial correlations.

**XLCNet = ResNet-based network, HY-OMP = Hybrid field orthogonal matching pursuit*

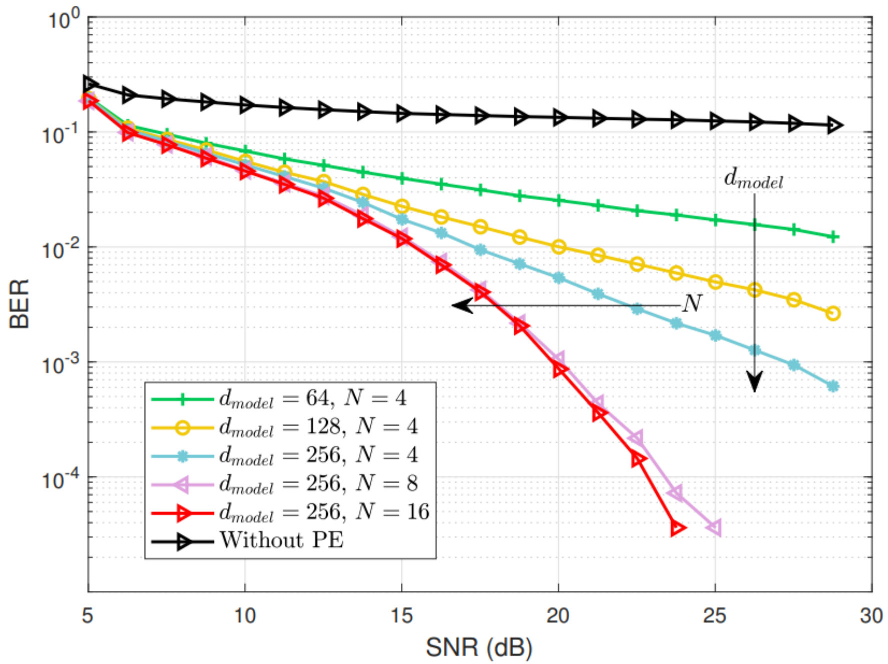


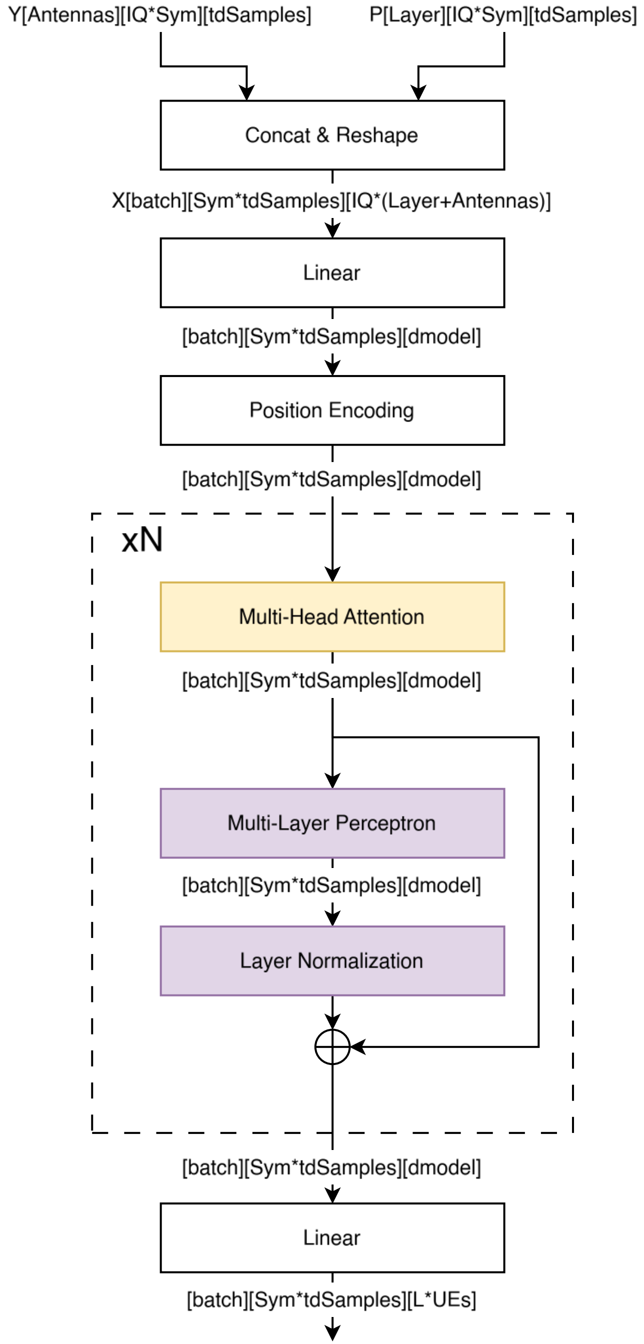


[7] Zou'25

Multi-layer data transmission with superimposed pilots (different power levels for pilots and data)

→ Making the network deeper with multiple cascaded transformer blocks increases performance





ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

AI-models for Channel Estimation

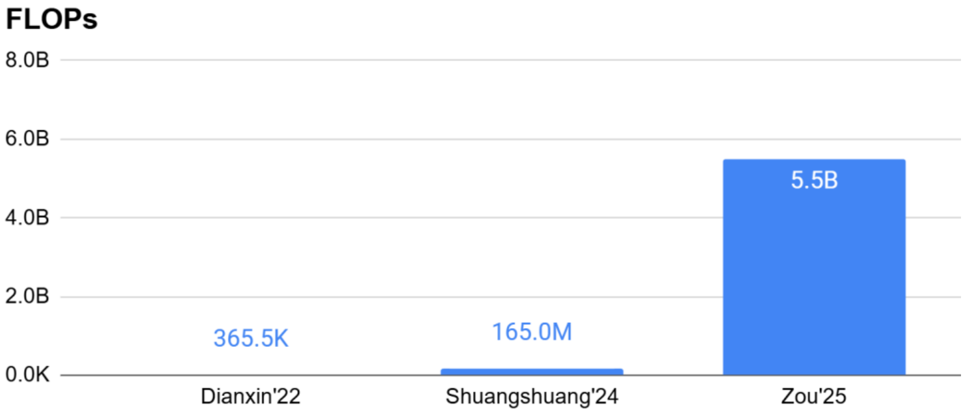
Name	Processing	fd/tdSamples	NRXxNTX	Modulation	Model	SNR@BER10^{-3}
[1] Deep-RX	Ch.Est. + Det.	312	2x1	16QAM	ResNet	6dB vs 8.5dB*
[2] Deep-RX MIMO	Ch.Est. + Det.	312	16x4	16QAM	ResNet	18dB vs 22dB*
[3] Neural-RX	Ch.Est. + Det.	2604	16x4	16QAM	ResNet	2.6dB vs 3.1dB*
[4] Neural-RX RT	Ch.Est. + Det.	2604	16x4	16QAM	ResNet	3.2dB vs 3.1dB*
[5] Dianxin’22	Ch.Est.	72	1x(-)	QPSK	Attention	-
[6] Shuangshuang’24	Ch.Est.	1	256x1	-	Attention	-
[7] Zou’25	Ch.Est.	96	4x32	-	Attention	-

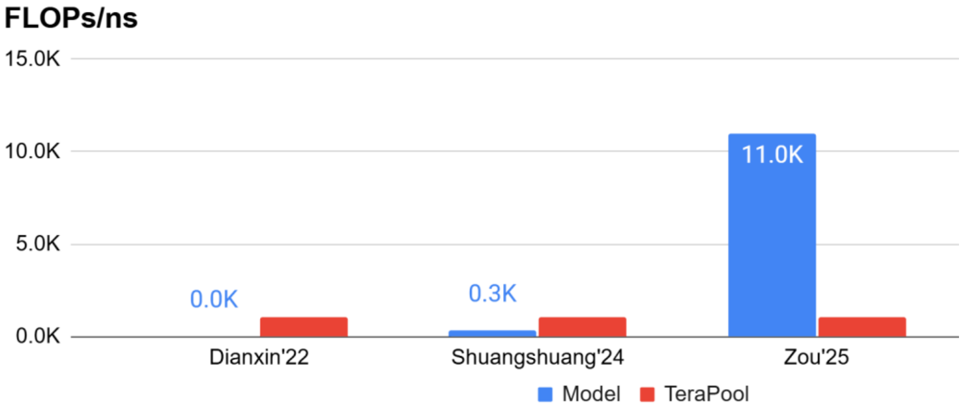
* vs XXdB required for conventional LMMSE Ch.Est. & MMSE MIMO detection

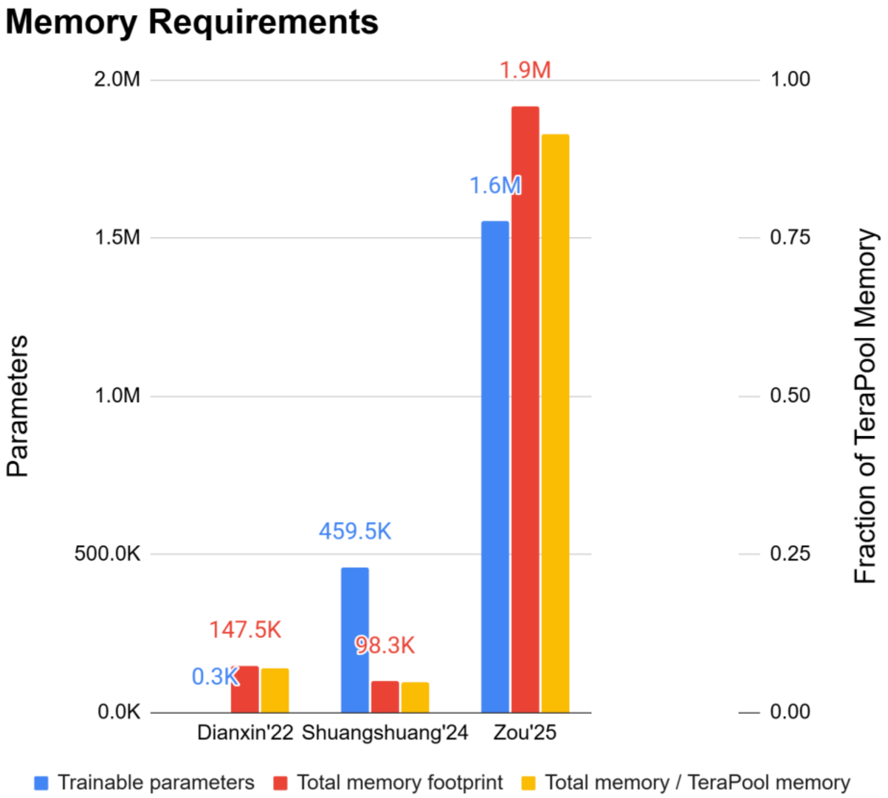
[1] DeepRX, https://arxiv.org/abs/2005.01494
[2] DeepRX-MIMO, https://arxiv.org/abs/2010.16283
[3] NeuralRX, https://arxiv.org/pdf/2312.02601
[4] NeuralRX-RT, https://arxiv.org/abs/2409.02912
[5] Dianxin’22, https://arxiv.org/pdf/2204.13465
[6] Shuangshuang’24, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10827075
[7] Zou’25, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10890516&tag=1

Computational complexity & Memory Footprint (Attention-based)

Total memory footprint depends on tiling. We assume all *loops* stay in L1.



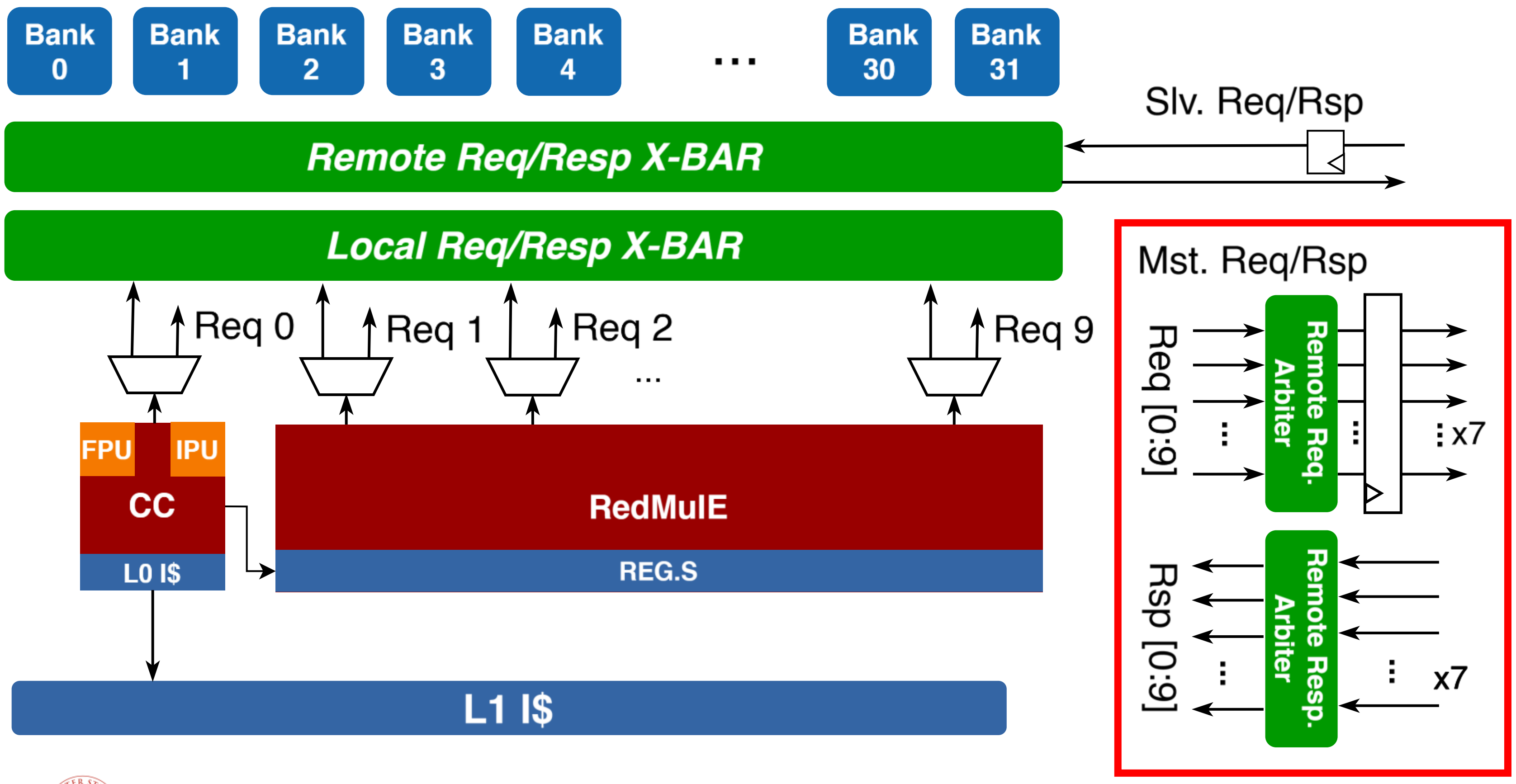




WIP – Burst RedMulE transactions

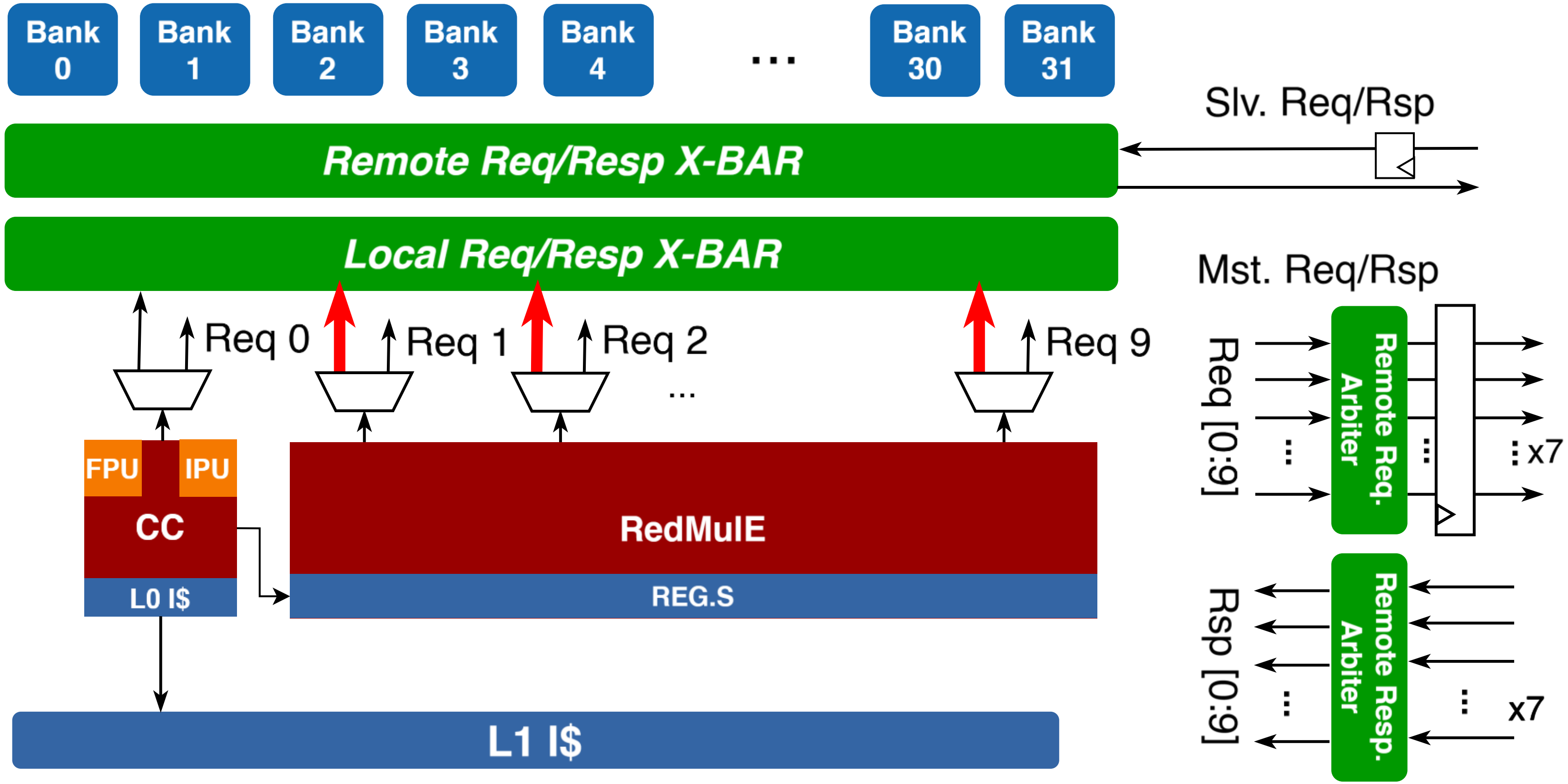
RedMulE wide access can cause conflicts at the Tile interface

(shared interconnect resources to access out-of-Tile memory locations)



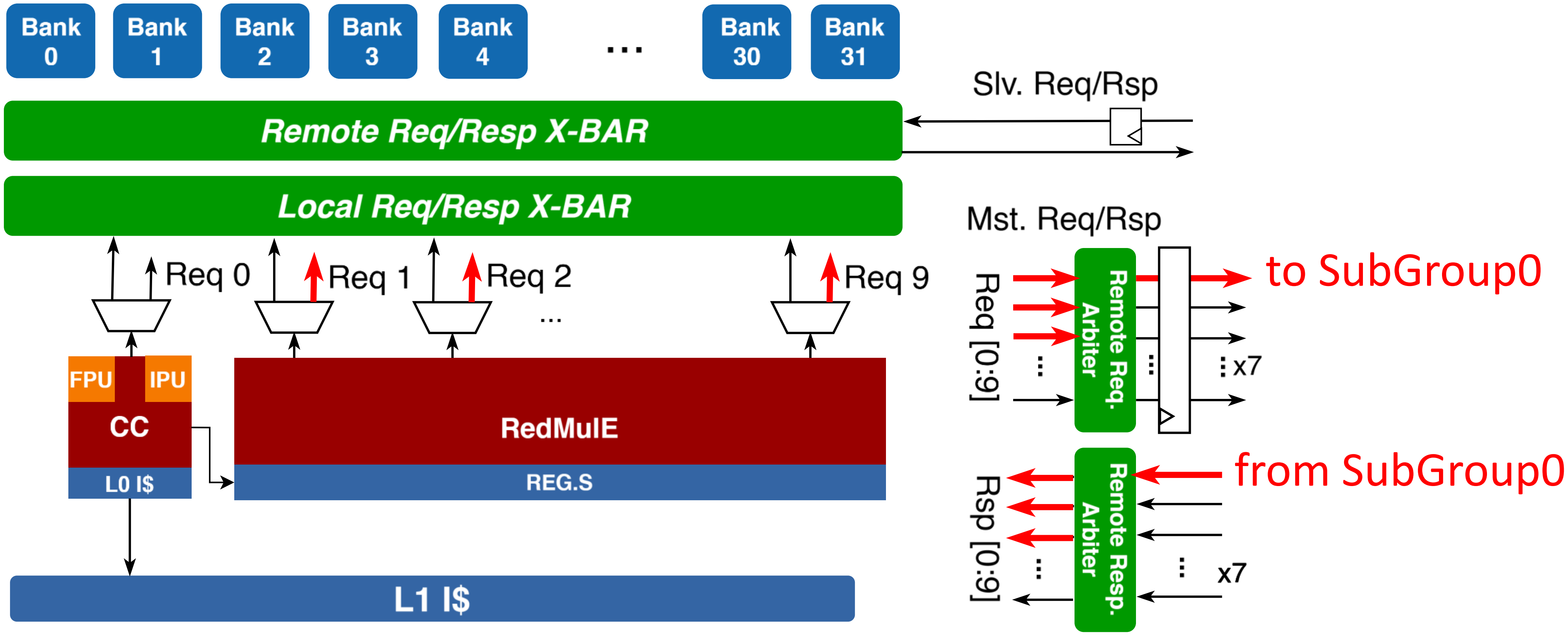
WIP – Burst RedMulE transactions

- **Local parallel access** → No conflict, all requests to different banks

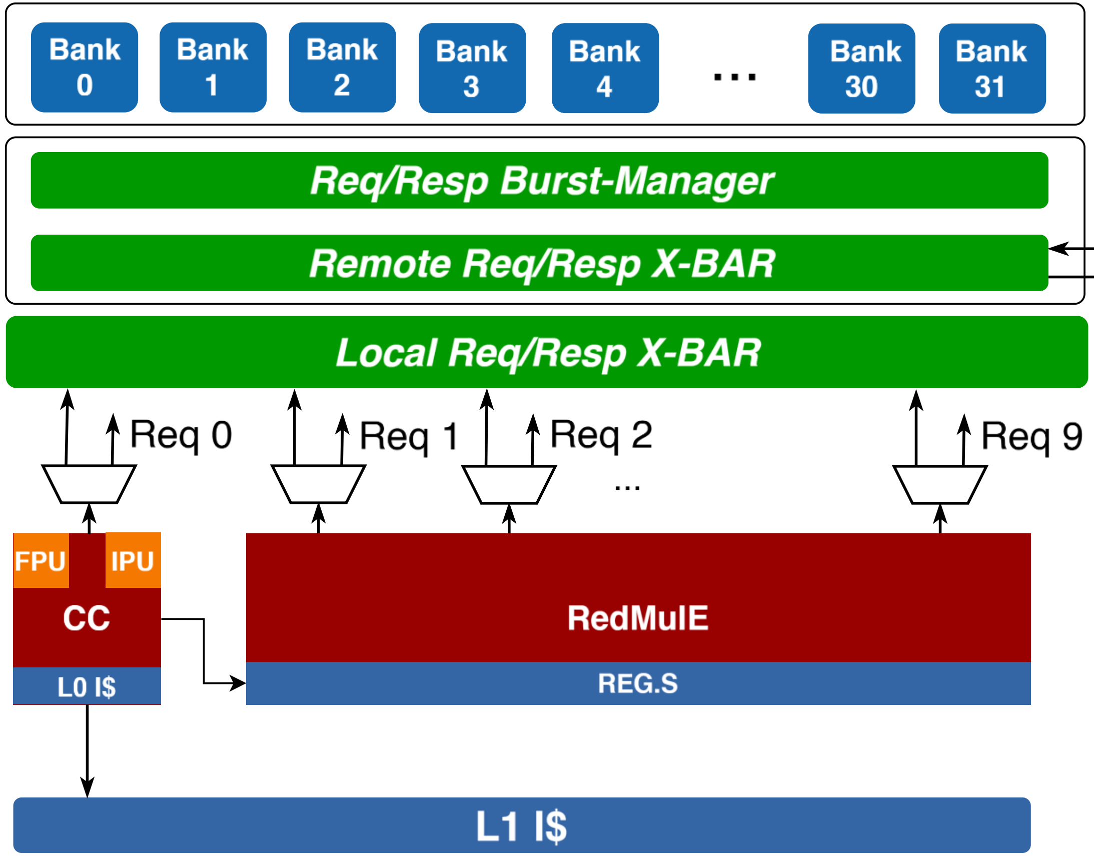


WIP – Burst RedMulE transactions

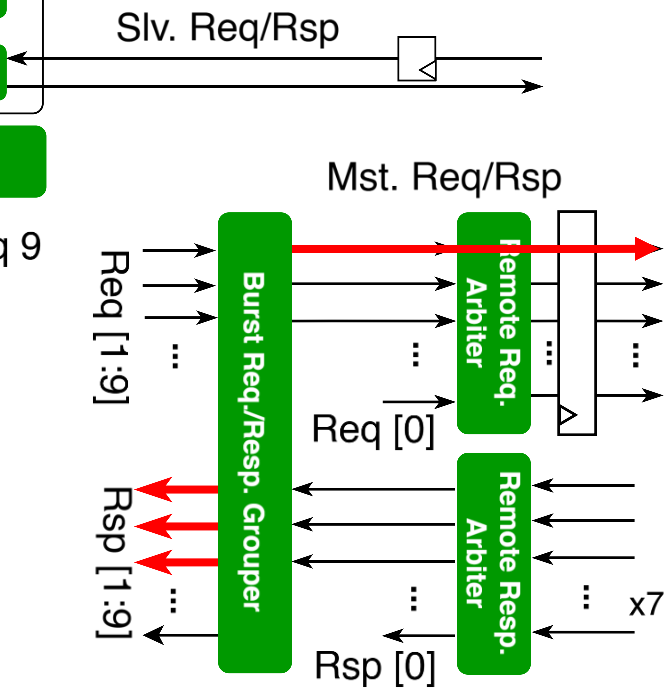
- **Local parallel access** → No conflict, all requests to different banks
- **Remote parallel access** → Conflict, all requests to the same remote port



WIP – Burst RedMulE transactions



2. Distribute the burst to the sequential addresses in the Tile



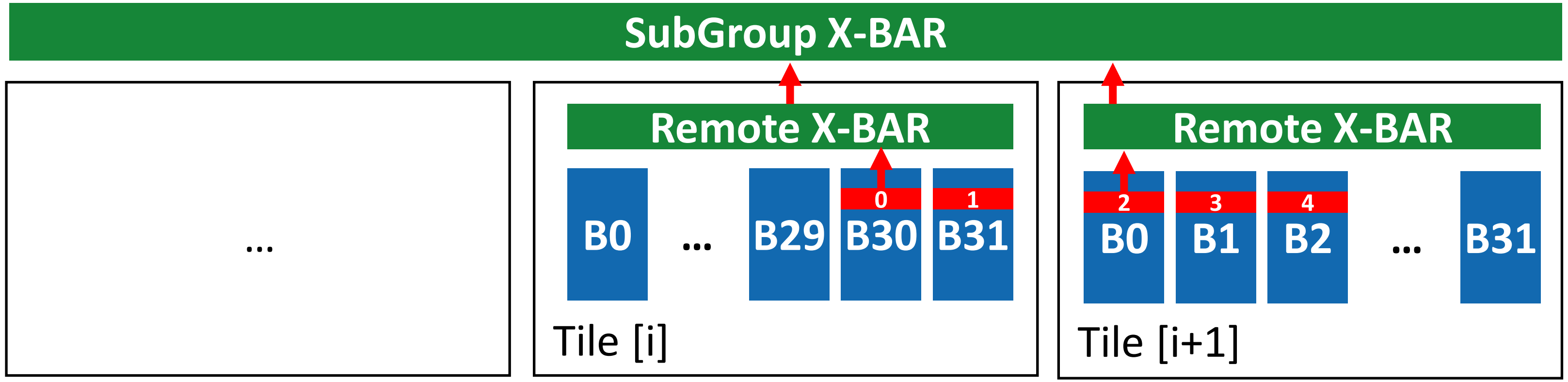
1. Replace the parallel request with a **burst request** → send **a) first address, b) burst length**

3. Retire responses

WIP – Burst RedMulE transactions

When we retire responses

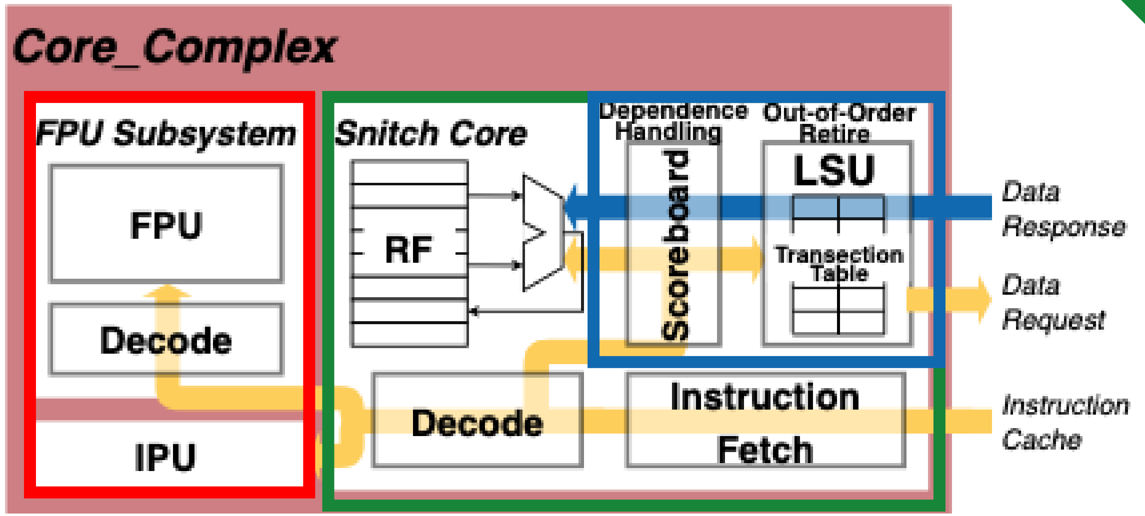
- The **burst manager** ensures that responses to a burst come back in-order
- When the burst is longer than the Tile boundary?
  - We **cut** the burst request in two bursts



TeraPool's PEs-to-L1-banks Interconnect Design

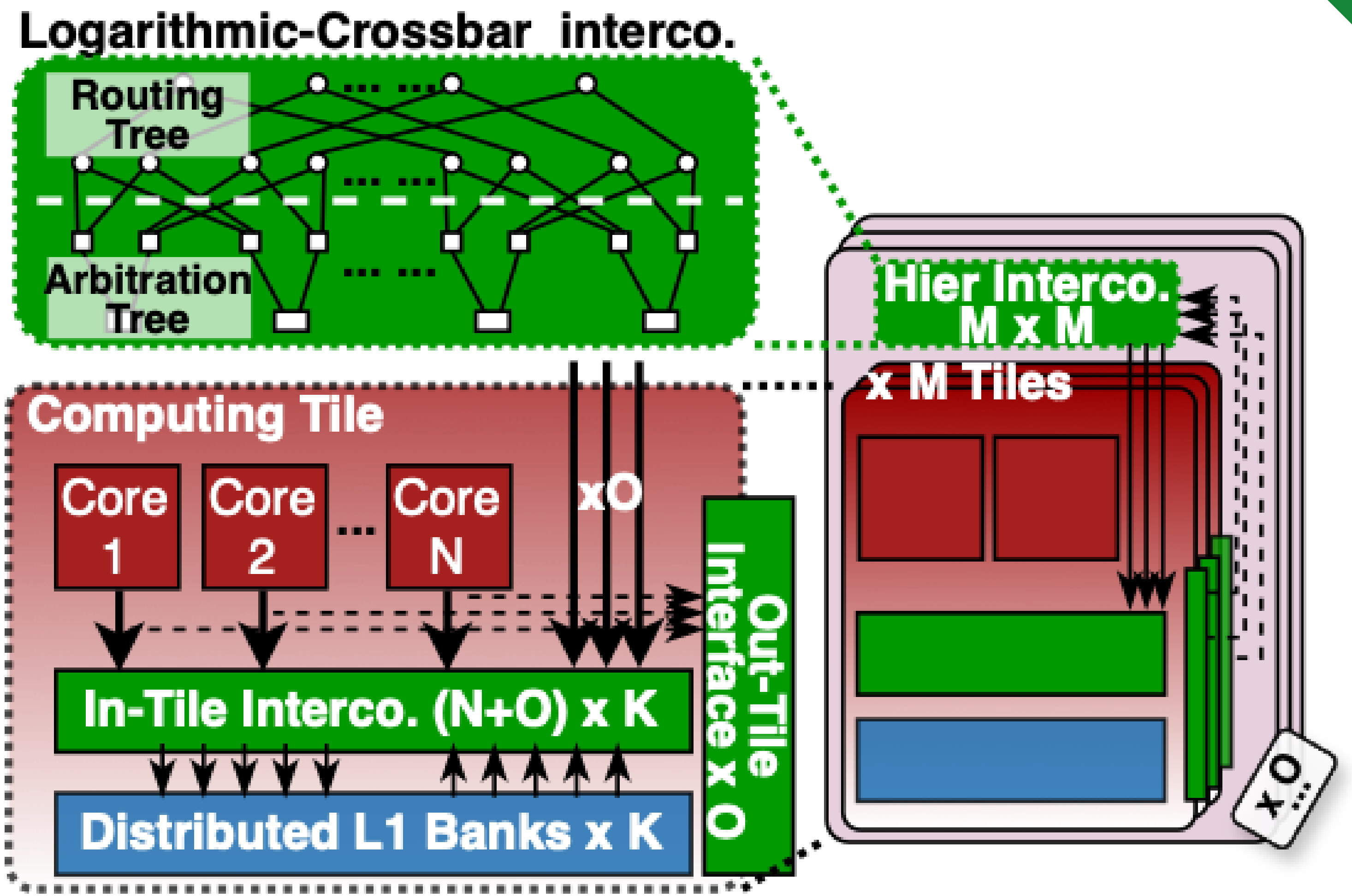
TeraPool is NUMA (Non-Uniform Mem Access) Cluster

- **TeraPool is a large single cluster:**
  - 1024 cores with 4096 SW-managed L1 banks;
  - All cores shared all L1 banks, **uni-address**;
- **Large-scale requires hierarchical design:**
  - Different latency to each Hier -> NUMA;
- **PPA consideration:**
  - 1000+ small cores **-> single-pipeline tiny core**
  - Extendable **-> extension port per core**
  - Core's outstanding support **-> Transaction Table per core**
  - Low-latency interconnect **-> Full-Comb. Log. Crossbar**



Multi-stage Hierarchical Crossbar design

- **Fully combinational Logarithmic Crossbar:**
  - 1 cycle low-latency memory access.
  - Spill reg can be added on hierarchy interface, breaking long distance path.
- **Each out-Tile interface to different Hier-blocks:**
  - **Conflict happens when:** Cores in same Tile, access to same Hier-block, in same cycle.



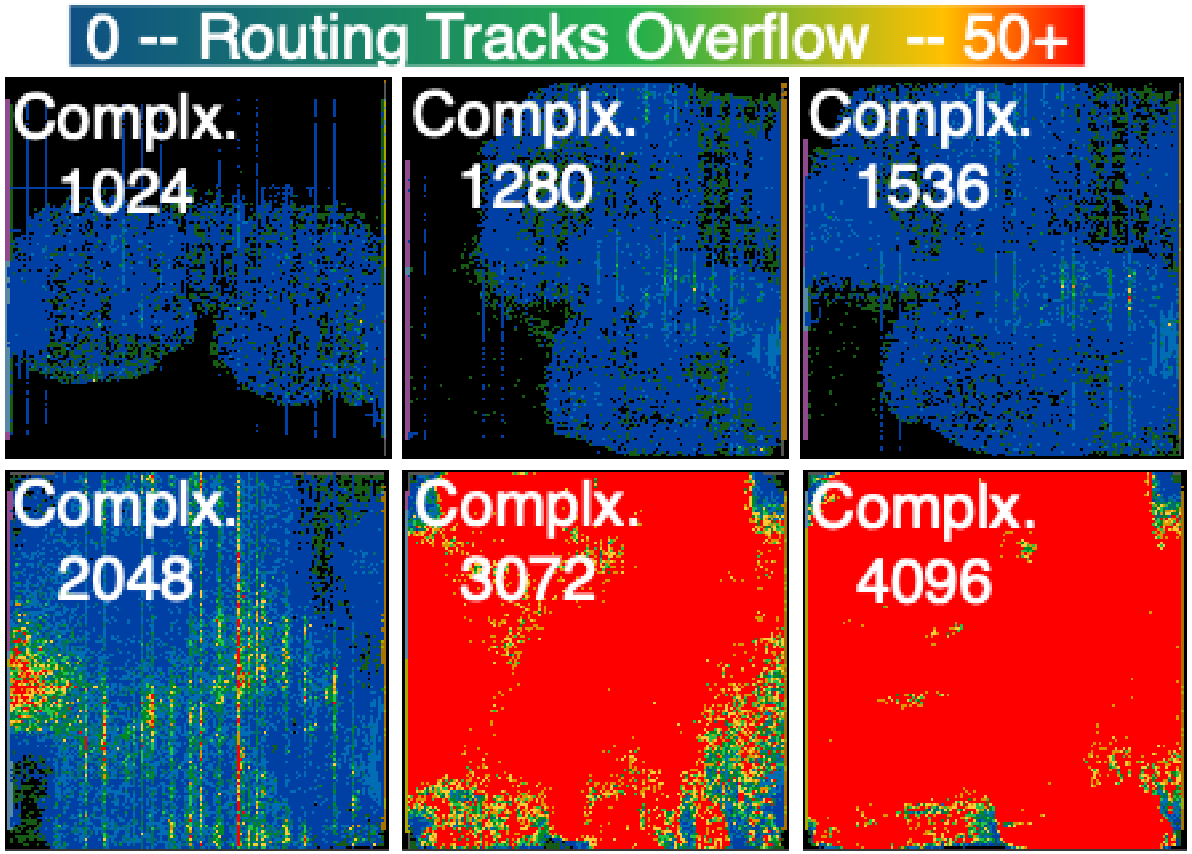
How many implement hierarchy we need?

- **More hierarchies -> more arbitration -> more conflict.**
- **Less hierarchies -> more complex crossbar -> physical routing limited.**
- **For one crossbar, *n* input *k* output:**
  - Routing complexity = $n \times k$;
  - Combinational stage delay: $\log_2 n + \log_2 k$;
- **GF12nm, 13 Metal layers:**

Routing Quality of Logarithmic-Staged Crossbar Interconnect at Different Complexities (GF12nm, 13M)

Interco. Complexity	Congestion* H	Congestion* V	Congestion* Overall	Area (kGE)	Critical Path (ns)
1024	0.56%	0.12%	0.34%	361	0.91
1280	1.72%	0.47%	1.09%	503	1.06
1536	3.25%	0.82%	2.04%	669	1.08
2048	34.46%	15.09%	24.77%	923	1.13
3072	172.30%	294.31%	233.31%	1274	1.27
4096	247.10%	368.90%	308.00%	1485	1.47

* The average routing track overflow rate for each horizontal, vertical layer, and overall design. The in/outputs span from the center of the west/east boundary.



How many implement hierarchy we need? -> Three

Hierarchical Interconnect Design Analysis for 1024 PEs Fully Connect 4096 L1-Memory Banks

Hierarchy Interco.[*]	Interconnect Quality						Design Challenge			
	ZeroLd (cyc)	AMAT (cyc)	Throughput (req/pe/cyc)	Total Complex.	Critical Complex.	Critical Comb. Delay	Physical Routing	Path Balance	Base Block Scale	Interco. Perform.
1024C	1.000	1.130	0.885	4194304	4194304	22				
4C-256T	2.992	6.081	0.245	87040	65536	16				
8C-128T	2.984	10.075	0.124	54272	16384	14				
16C-64T	2.969	18.077	0.062	74752	4096	12				
4C-16T-16G	4.867	5.318	0.431	163840	320	8.3				
4C-32T-8G	4.742	5.443	0.409	122880	1024	10				
8C-16T-8G	4.734	5.794	0.358	90112	512	9				
8C-32T-4G	4.484	6.676	0.272	69632	1024	10				
16C-8T-8G	4.719	6.669	0.273	110592	1536	10.6				
16C-16T-4G	4.469	8.612	0.178	90112	1280	10.3				
4C-16T-4SG-4G	6.367	8.457	0.270	121856	4096	12				
8C-8T-4SG-4G	6.359	9.198	0.230	89088	1024	10				
16C-4T-4SG-4G	6.344	11.049	0.159	109568	1536	10.6				

[*] 4096 1 KiB SPM banks are split across each hierarchy with PEs, using a banking factor of 4.

[*] The hierarchy is denoted as αC–βT–γSG–δG, where δ is the number of Groups, each with γ SubGroups, β Tiles per SubGroup, and α PEs per Tile.

TeraPool: 3 Hierarchical, Multi-Stage Xbars Interconnect

- **Only one Spill Reg add at “outgoing req/rsp” ports of each Hier-interfaces, avoiding critical path at interconnects.**

“Tile”:

8 cores -> 32 Banks.

1 cycle accessing.

“SubGroup”

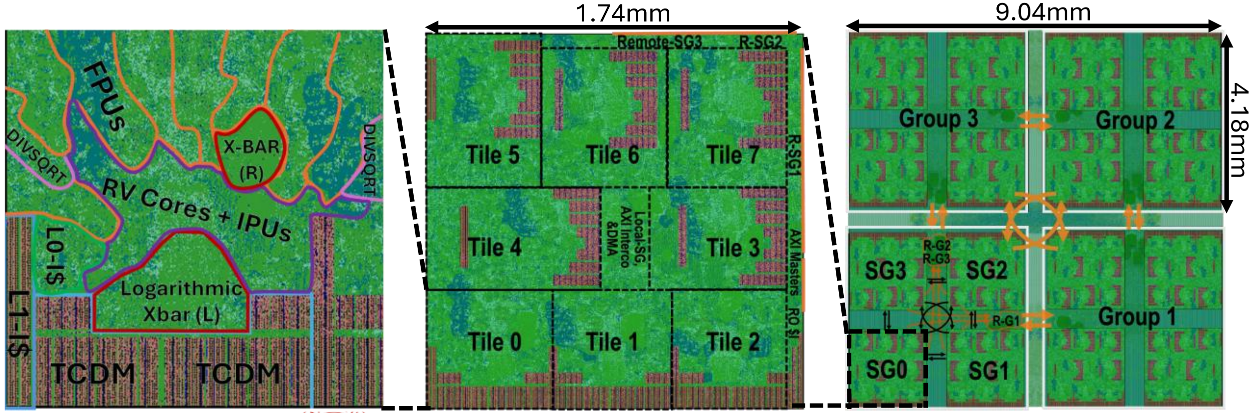
8 Tiles.

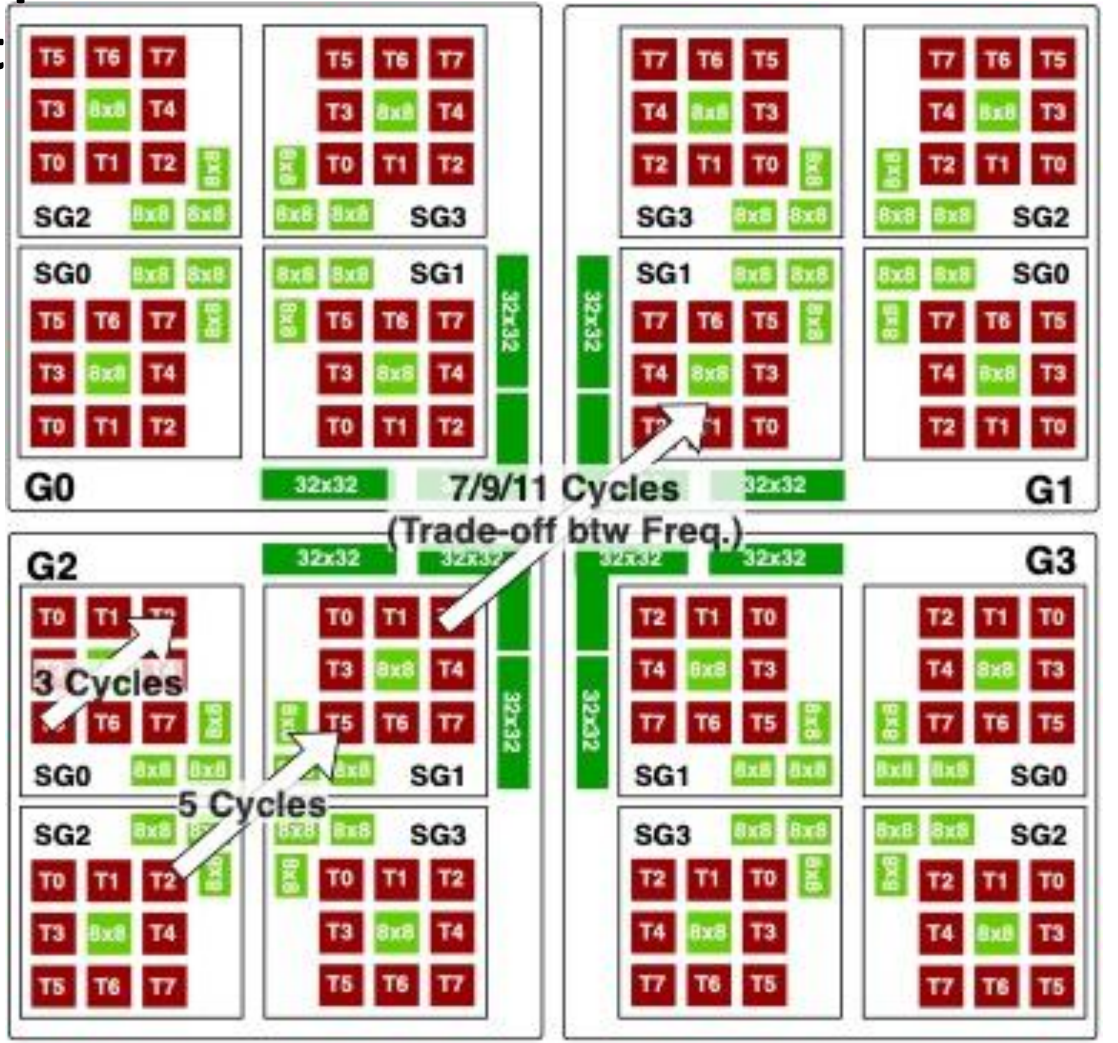
3 cycles accessing.

“Groups”

4 SubGroups each, 5 cycles.

4 Groups total, 9 cycles.





Thank you!

Q&A

ETH zürich
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA