



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



**UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH**

Memory Sandbox 2.0:

A Framework for Enabling HBM2e vs HBM2 Performance and Telemetry Analysis on Xilinx FPGAs

Elias Perdomo^{1,2}, **Joan Teruel**^{1,2},

Sajjad Ahmed¹, Behzad Salami¹, Teresa Cervero¹ and Xavier Martorell^{1,2}

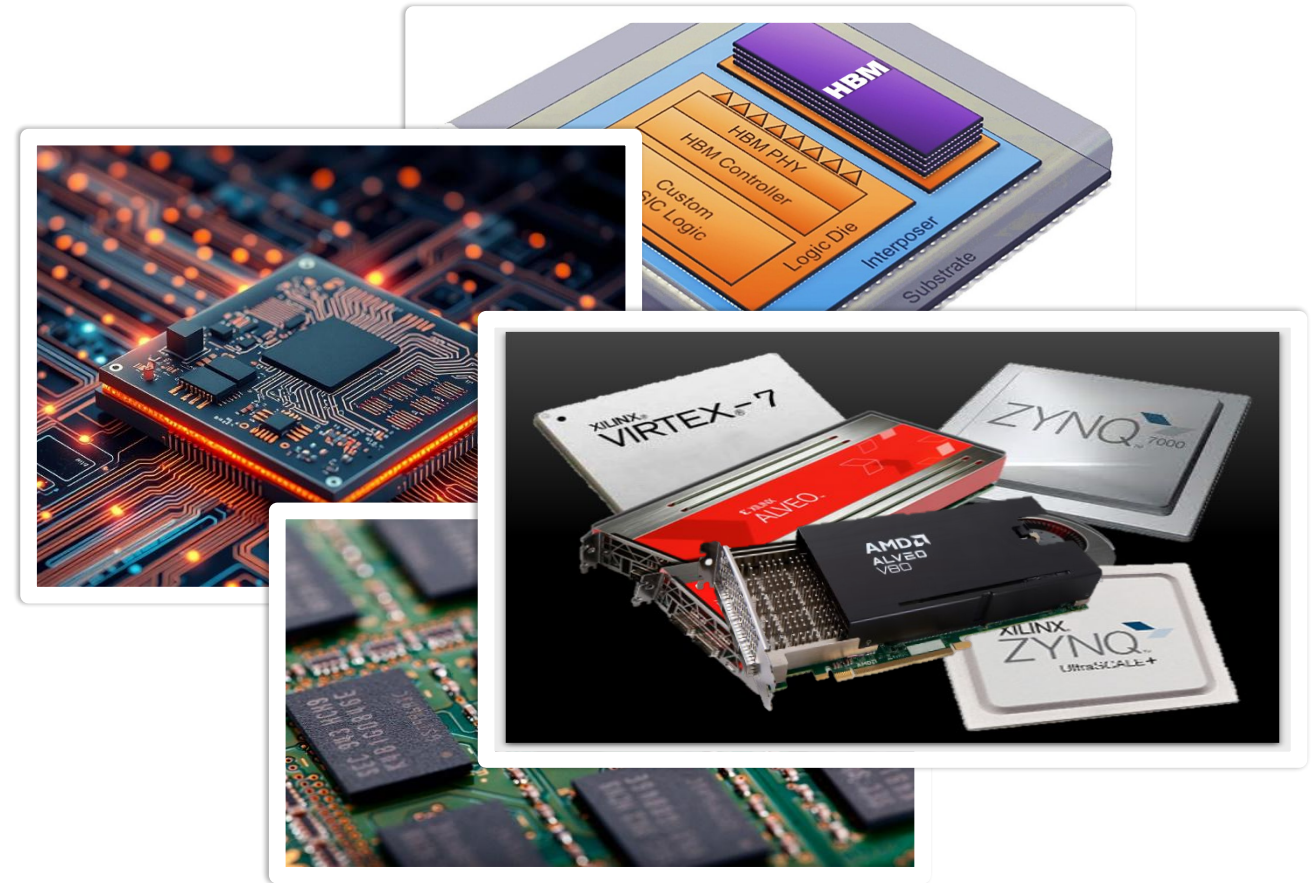
08/10/2025

¹Barcelona Supercomputing Center

²Universitat Politècnica de Catalunya

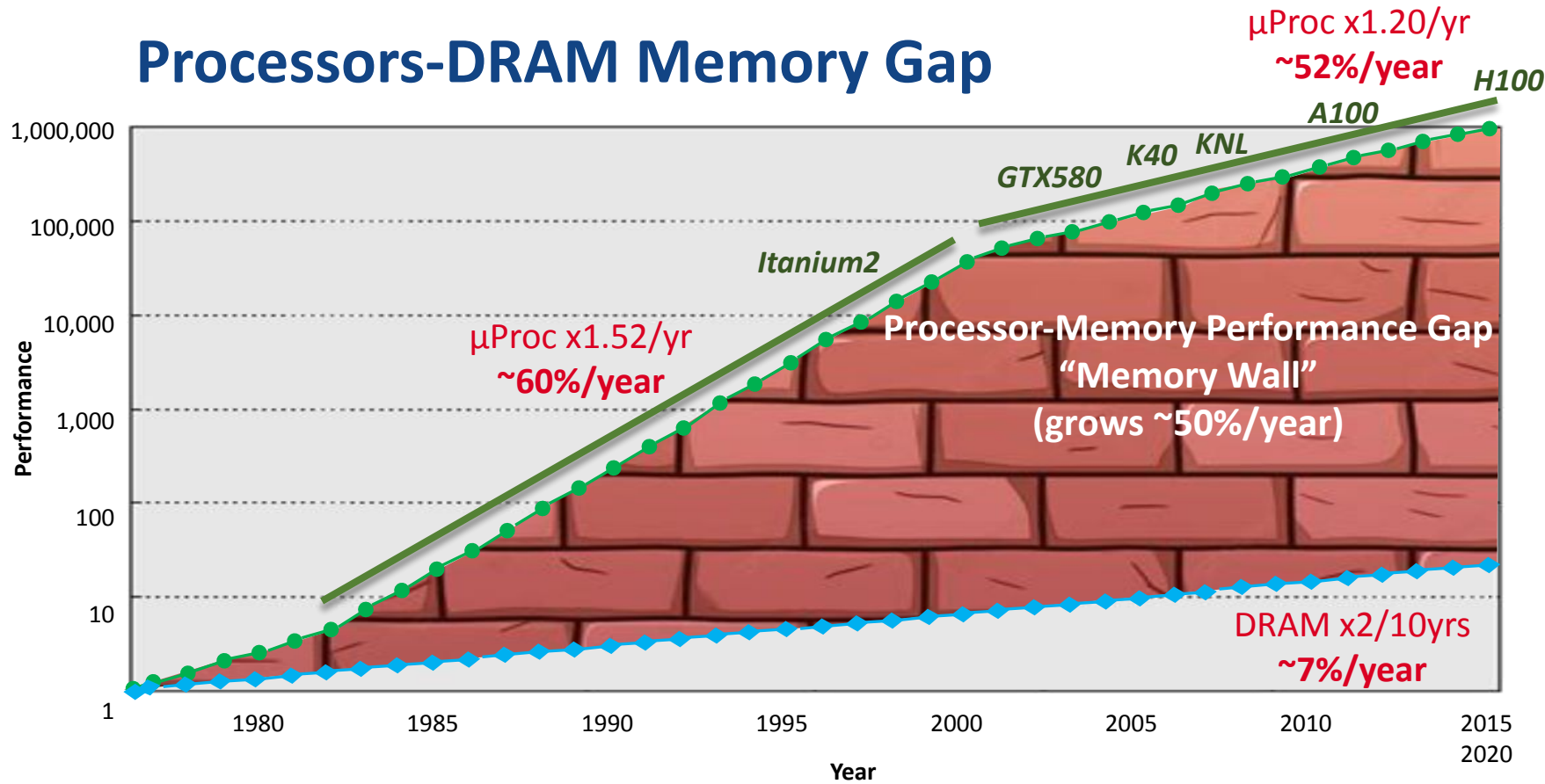
Outline

- Introduction & Problem Definition
- HBM as a solution
- Memory Sandbox 2.0 Architecture
 - ✓ Throughput results
 - ✓ Telemetry results
- Conclusions



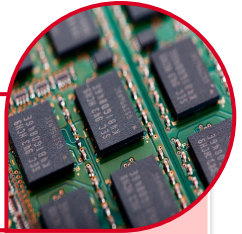
CPU vs Memory speed increase

Processors-DRAM Memory Gap



Place more memory on-chip

- von Neumann architectures
- More's Law
- Dennard scaling



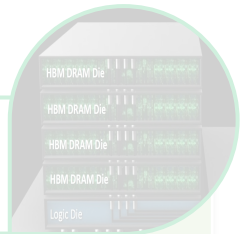
Brand new architectures

- At a rate never seen



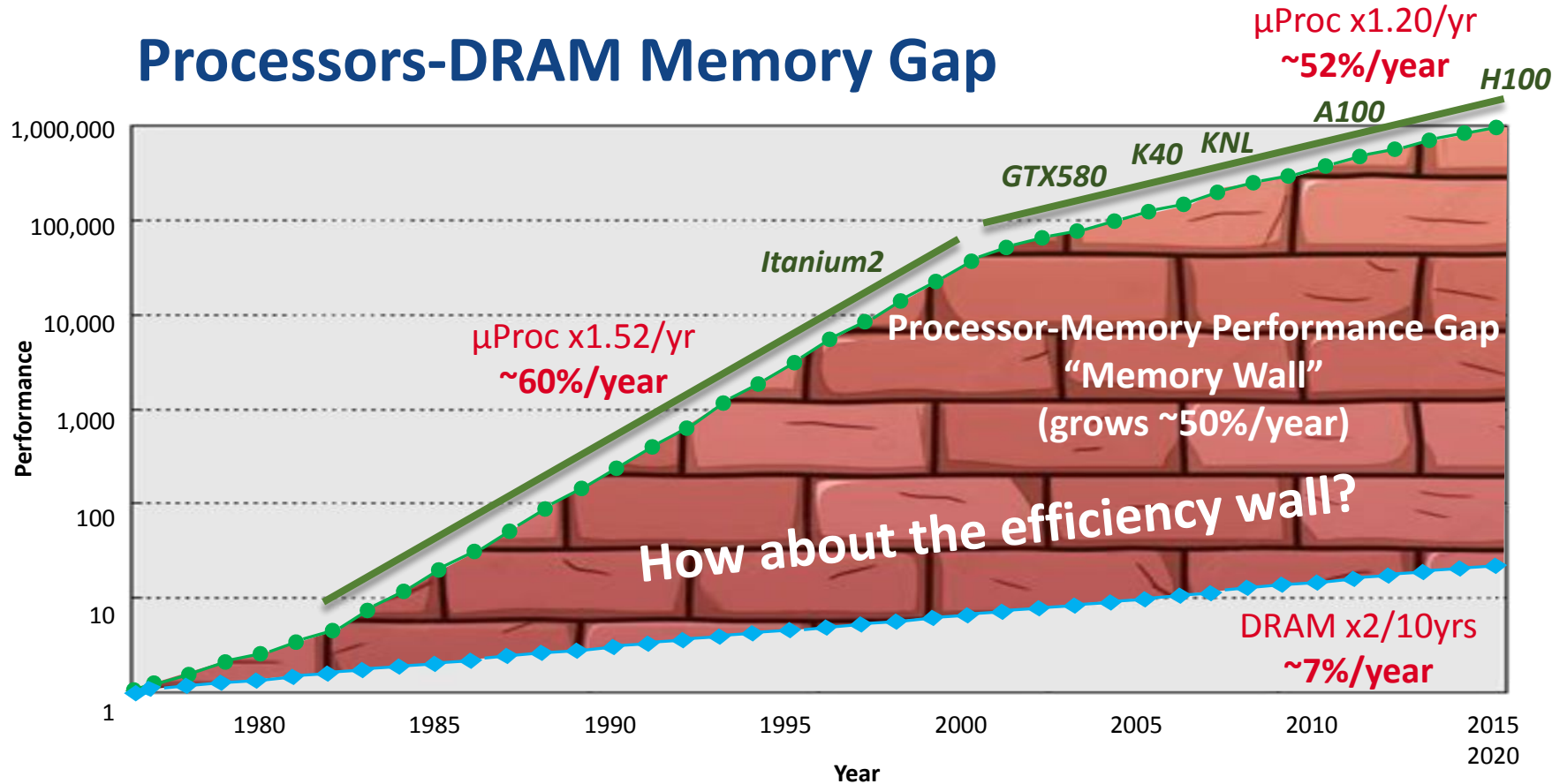
Move processing elements closer/into the memory

- Avoids processing elements replication
- Avoids going outside the die
- Utilizing wide short buses (HBM example)



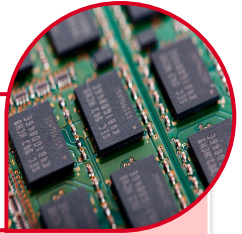
CPU vs Memory speed increase

Processors-DRAM Memory Gap



Place more memory on-chip

- von Neumann architectures
- More's Law
- Dennard scaling



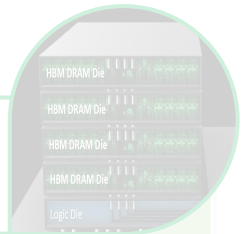
Brand new architectures

- At a rate never seen



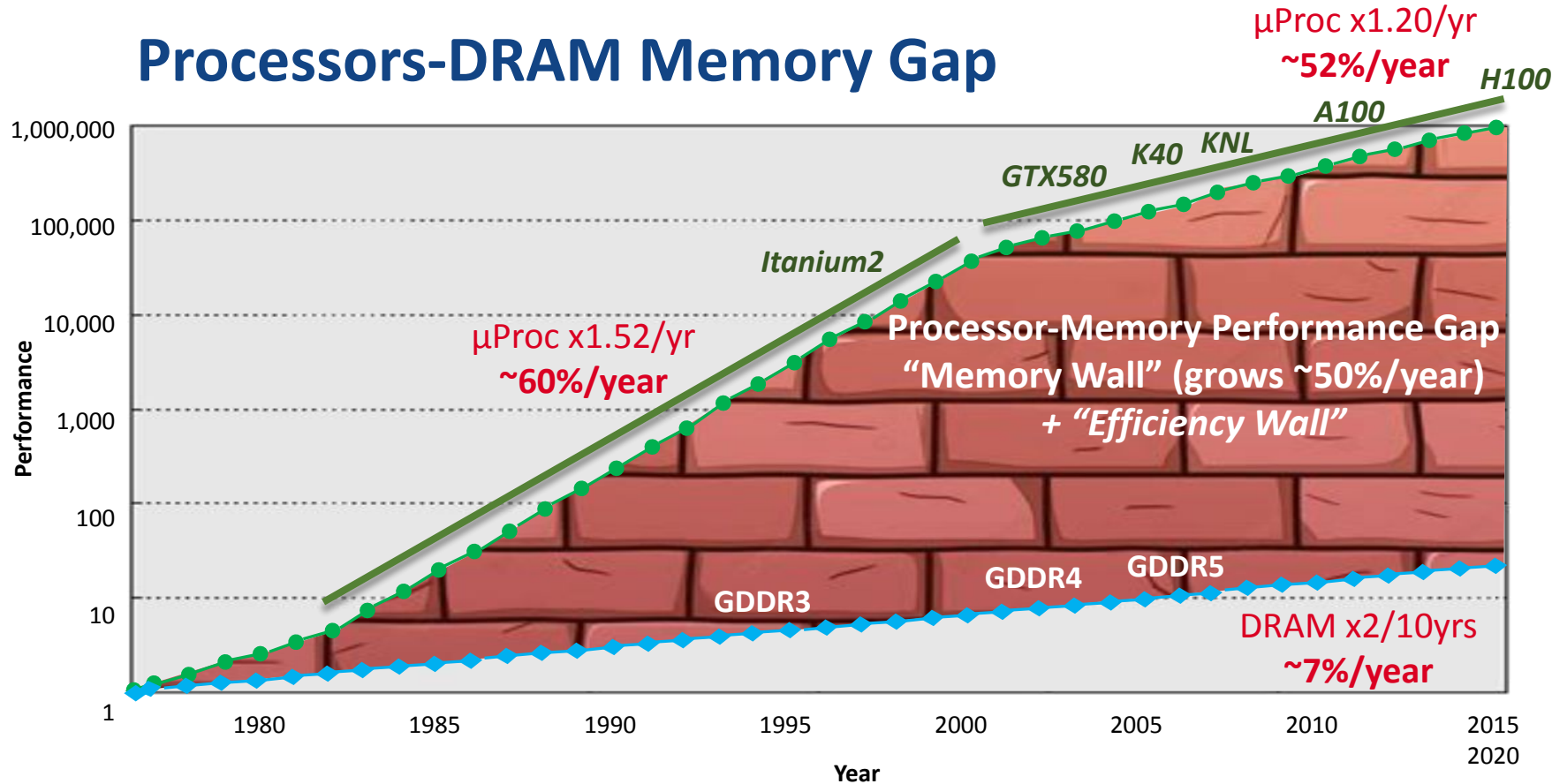
Move processing elements closer/into the memory

- Avoids processing elements replication
- Avoids going outside the die
- Utilizing wide short buses (HBM example)



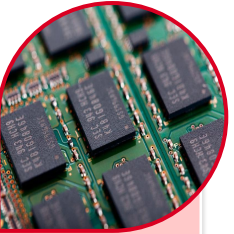
CPU vs Memory speed increase

Processors-DRAM Memory Gap



Place more memory on-chip

- von Neumann architectures
- More’s Law
- Dennard scaling



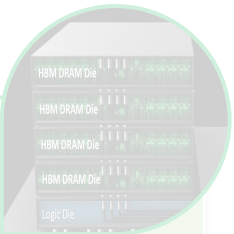
Brand new architectures

- At a rate never seen



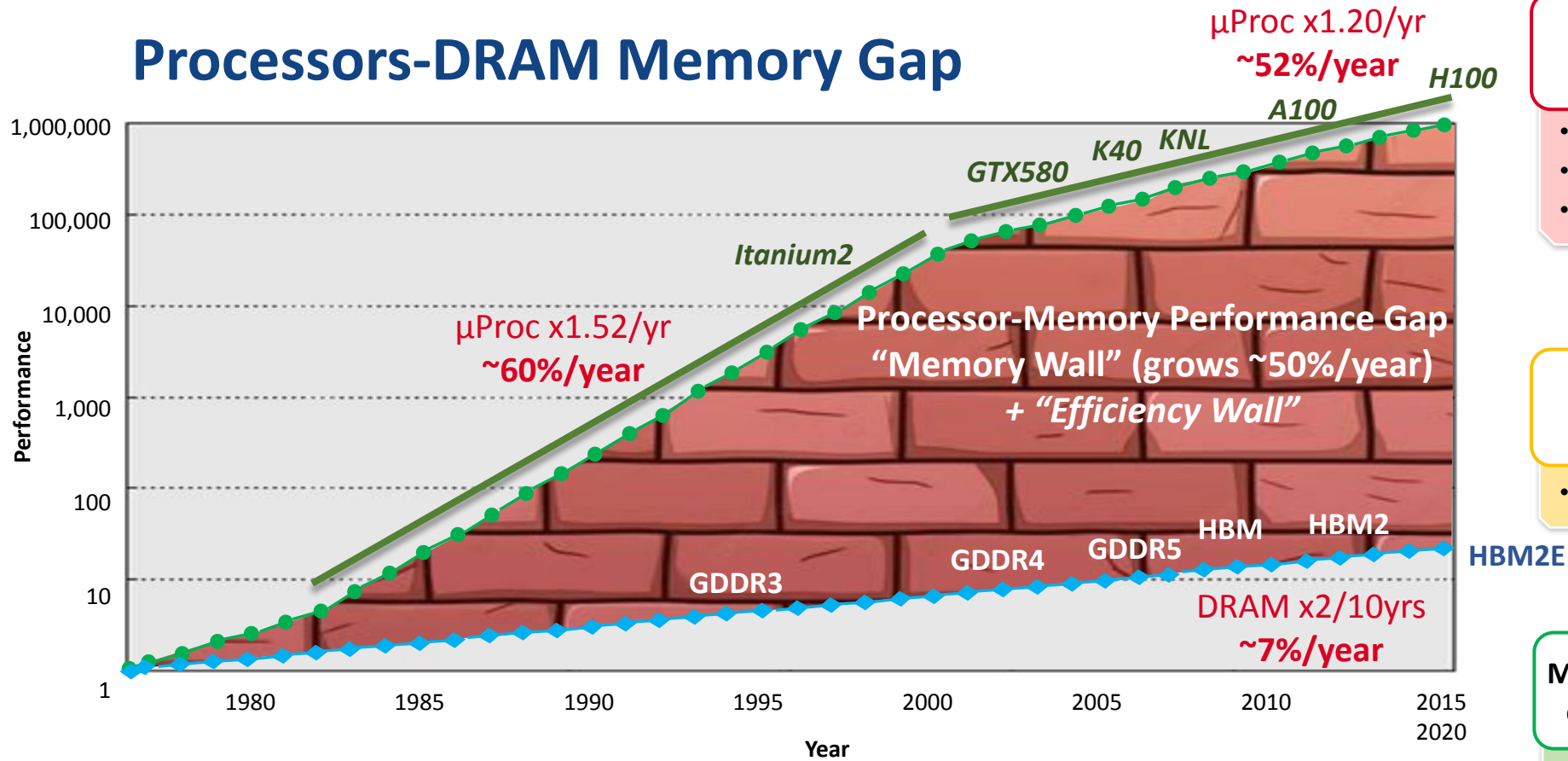
Move processing elements closer/into the memory

- Avoids processing elements replication
- Avoids going outside the die
- Utilizing wide short buses (HBM example)



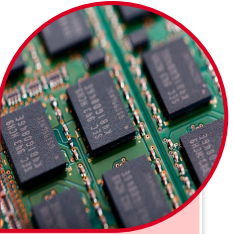
CPU vs Memory speed increase

Processors-DRAM Memory Gap



Place more memory on-chip

- von Neumann architectures
- More's Law
- Dennard scaling



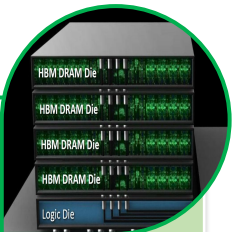
Brand new architectures

- At a rate never seen

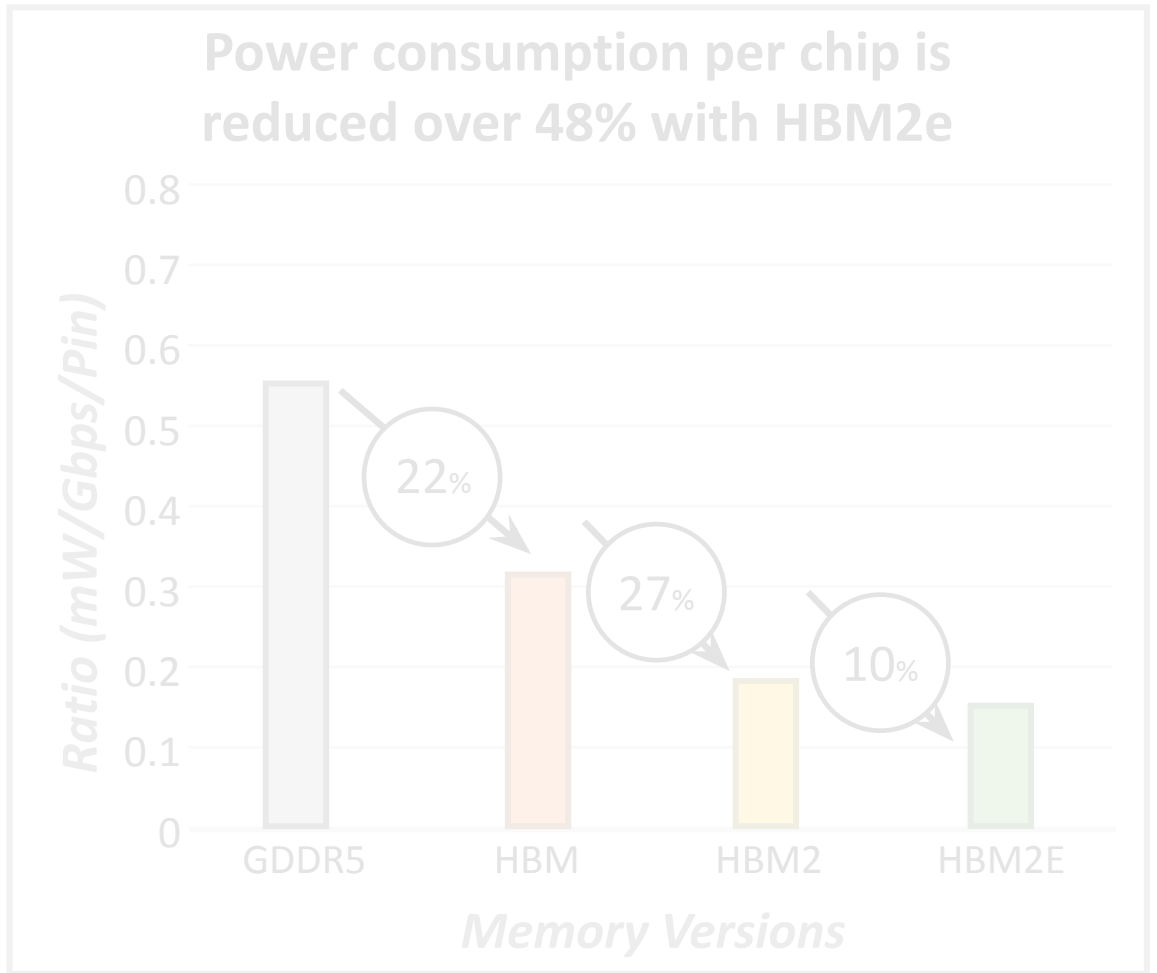
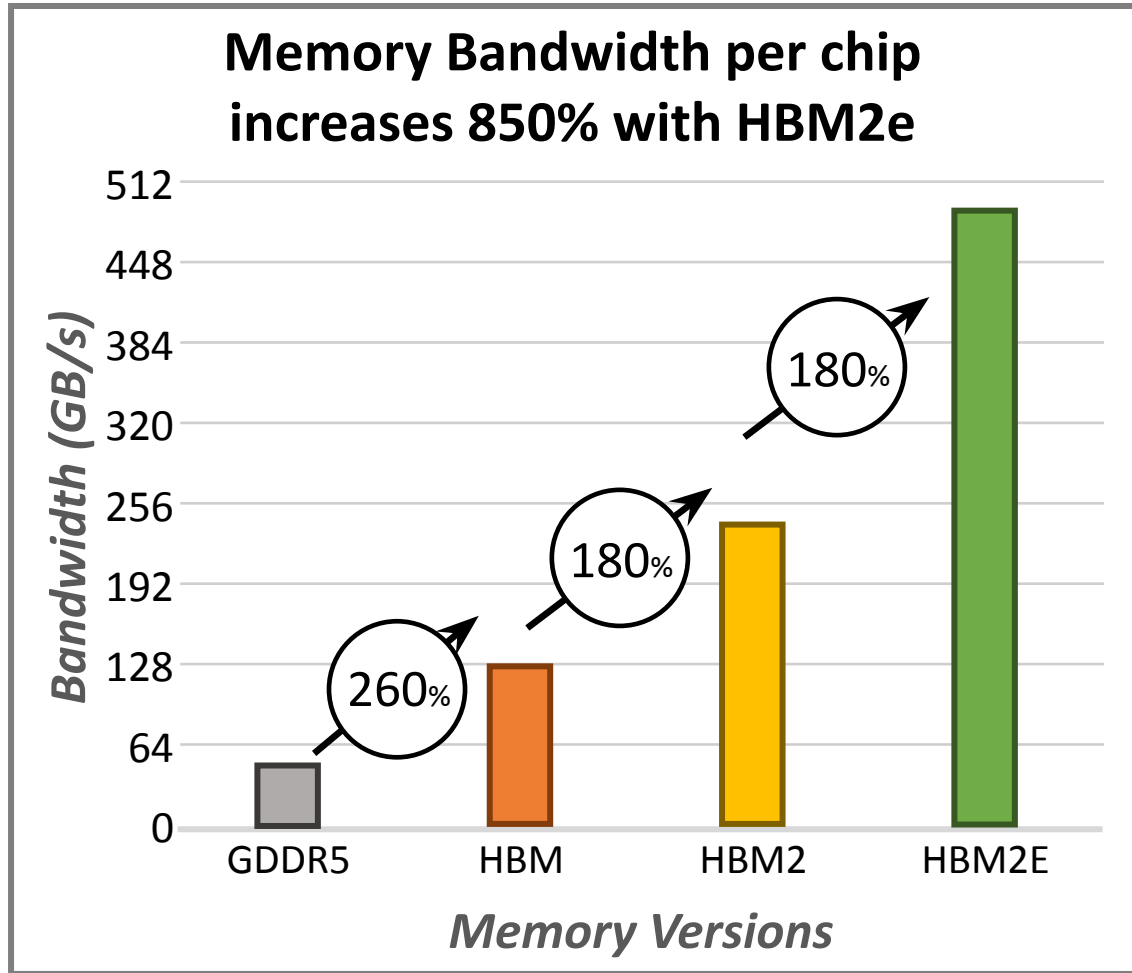


Move processing elements closer/into the memory

- Avoids processing elements replication
- Avoids going outside the die
- Utilizing wide short buses (HBM example)

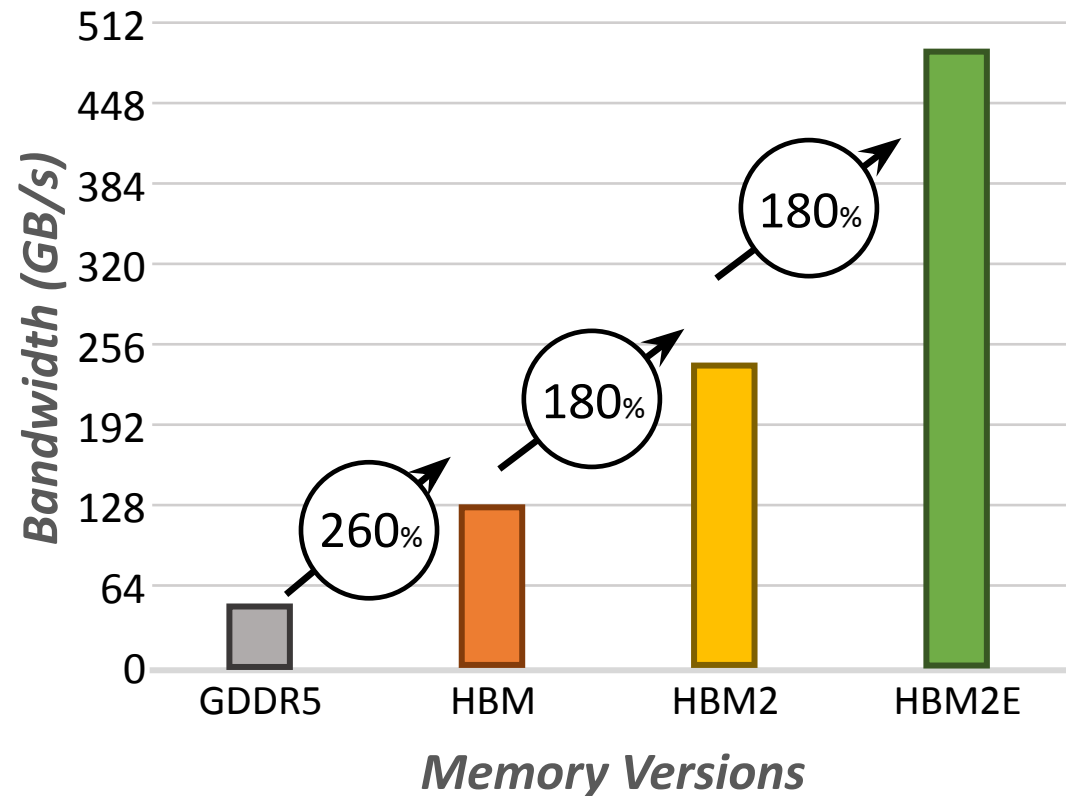


Bandwidth and power consumption comparison

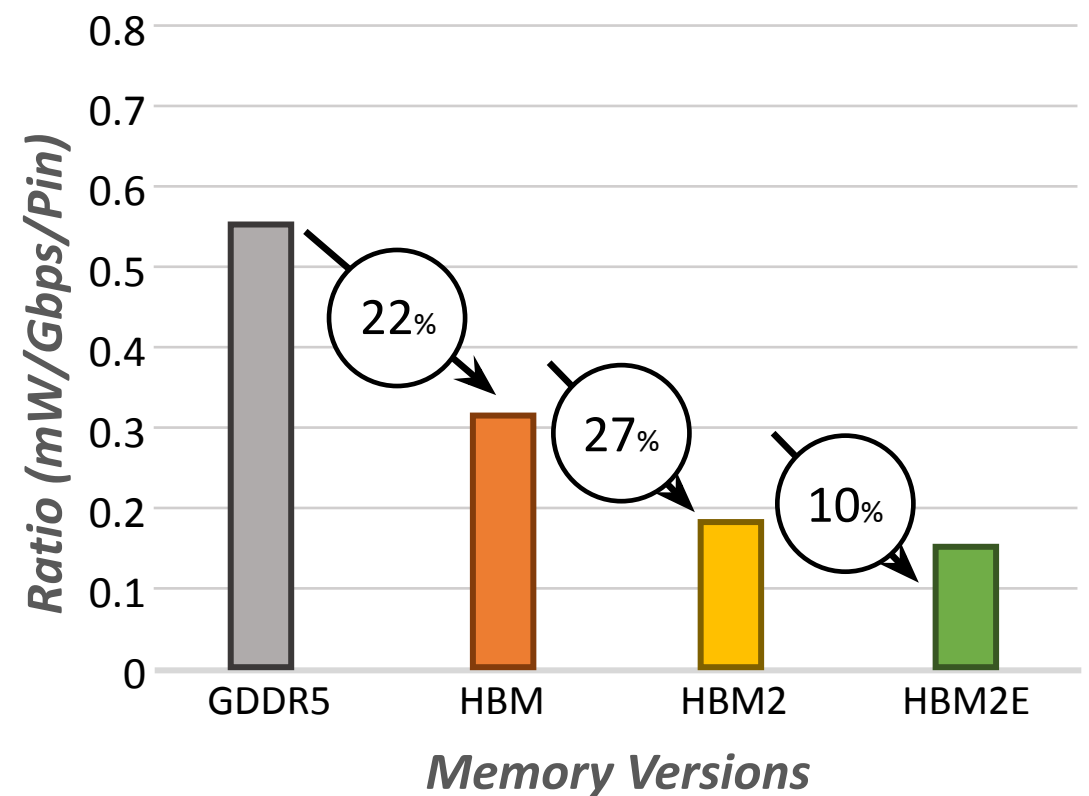


Bandwidth and power consumption comparison

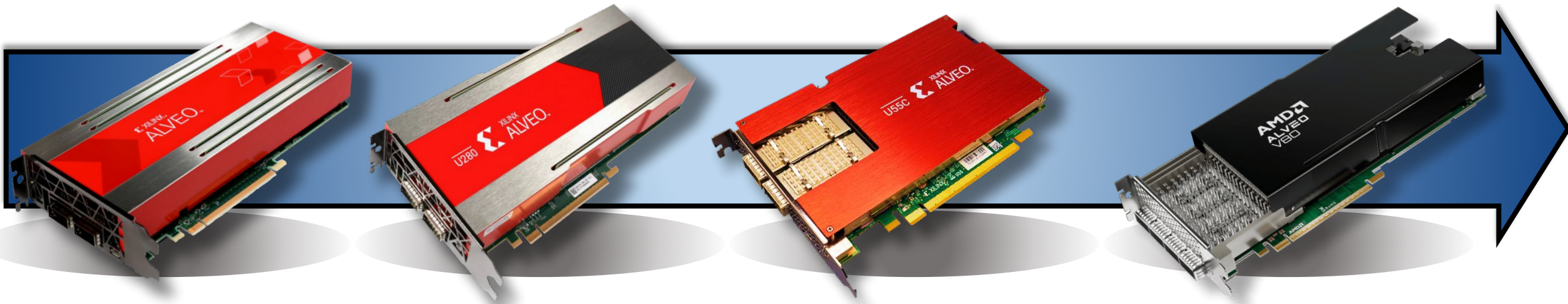
Memory Bandwidth per chip increases 850% with HBM2e



Power consumption per chip is reduced over 48% with HBM2e



AMD committed to a transition to HBM



**Alveo u200
(Oct/2018)**

DDR Total Capacity
64GB

HBM Total Capacity
0

HBM version
HBM2

**Alveo u280
(Nov/2018)**

DDR Total Capacity
32GB

HBM Total Capacity
8GB

HBM version
HBM2

**Alveo u55c
(Nov/2021)**

DDR Total Capacity
0GB

HBM Total Capacity
16GB

HBM version
HBM2

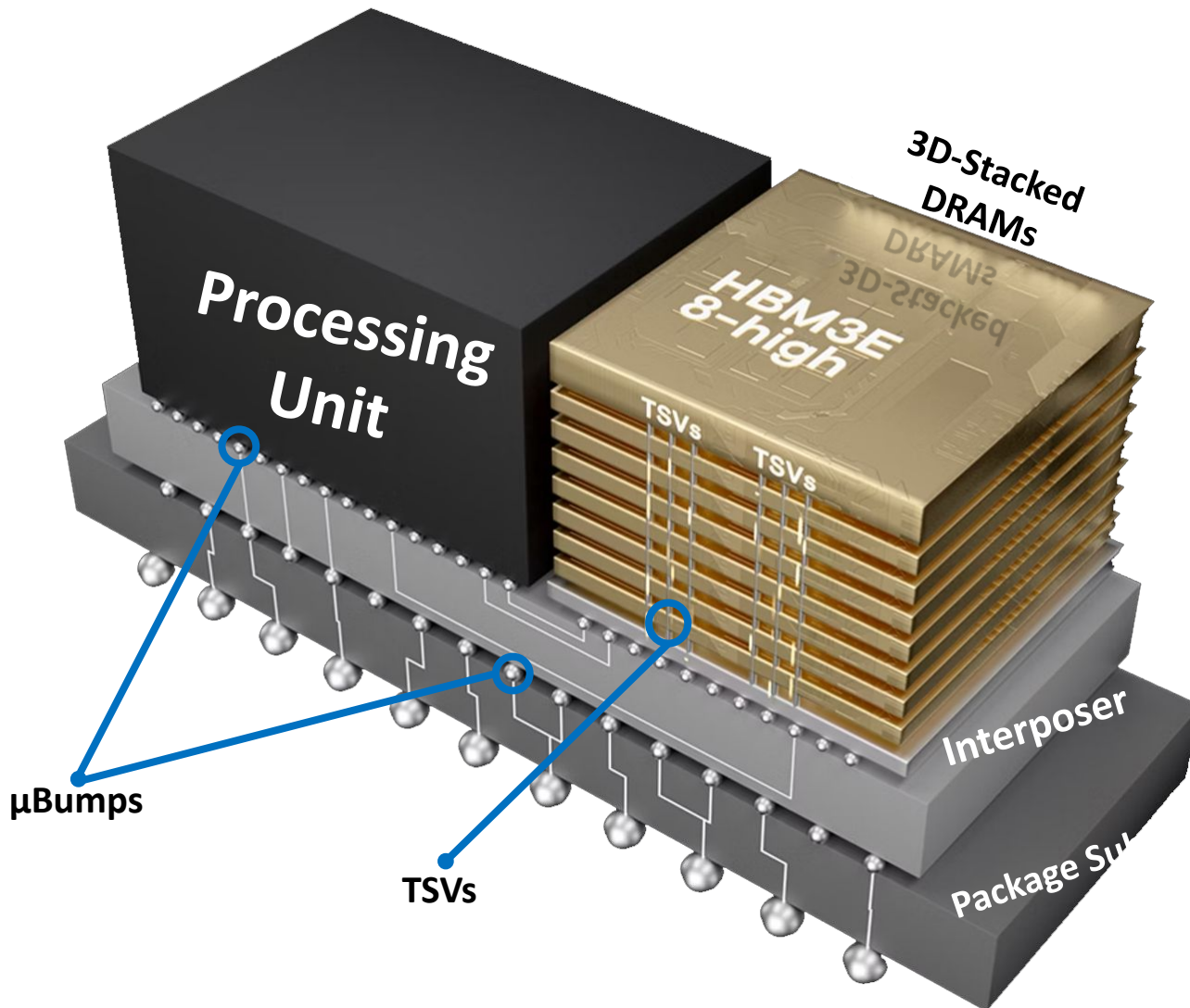
**Versal v80
(May/2024)**

DDR Total Capacity
4GB*(ARM)

HBM Total Capacity
32GB

HBM version
HBM2e

HBM technology and Standards Evolution



HBM STANDARDS EVOLUTION

Max Pin Transfer Rate	Max Capacity per stack	Max Bandwidth
HBM1		
1 Gb/s	1 GB	128 GB/s
HBM2		
2.4 Gb/s	4/8 GB	460 GB/s
HBM2E (Current FPGA)		
3.6 Gb/s	8/16/24 GB	820 GB/s
HBM3		
6.4 Gbps	16/24 GB	1.2 TB/s
HBM3e		
9.8 Gbps	36 GB	1.6 TB/s
HBM4*		
8 Gbps	64 GB	2 TB/s

Roadmap

March 2024

Makinote
'BSC FPGA-Shell'



November 2024

Memory
Sandbox 1.0



October 2025

Memory
Sandbox 2.0

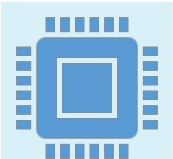
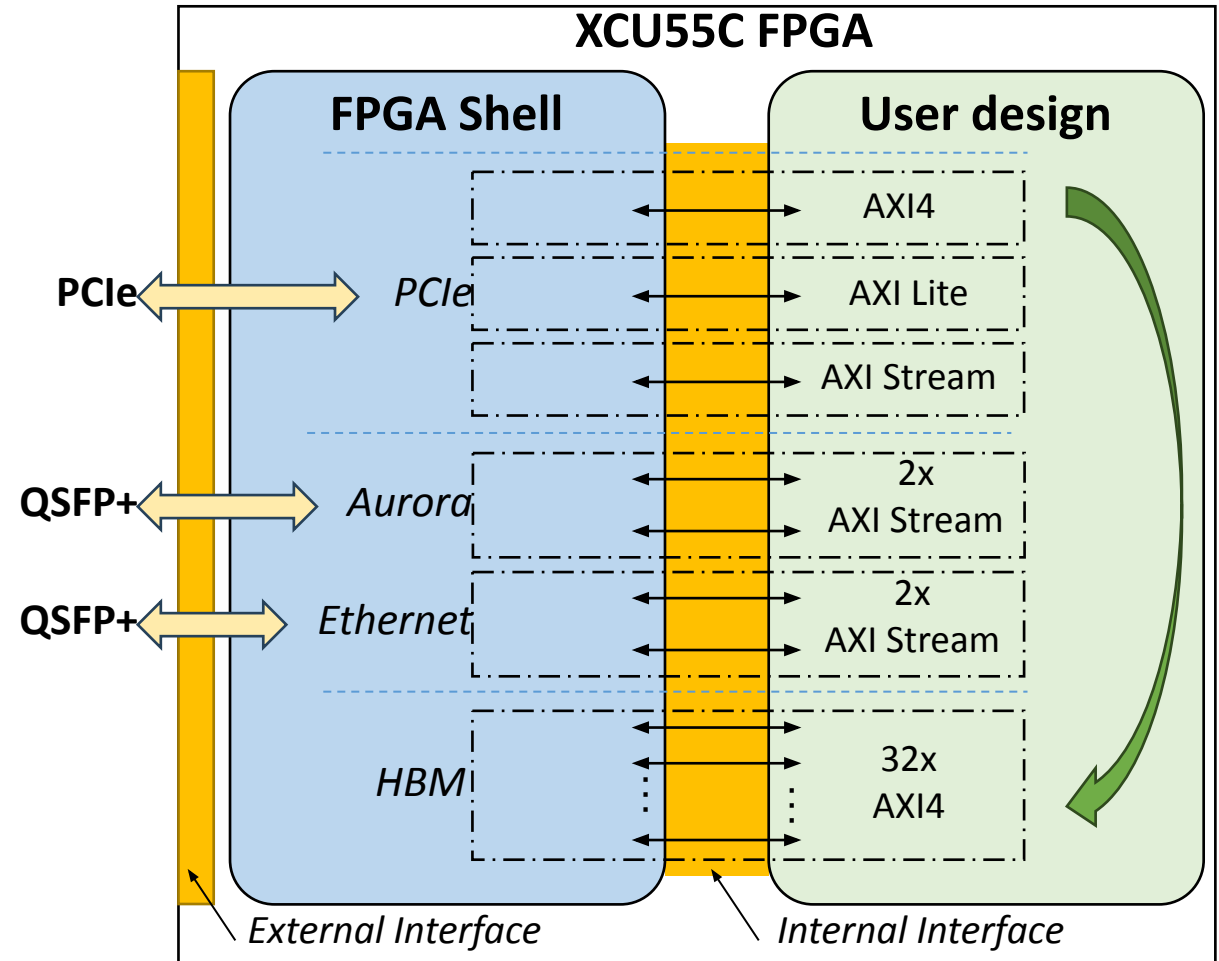


Makinote



Makinote 'BSC FPGA-Shell'

- Basic FPGA-SHELL for automating I/O interfacing between FPGA and custom designs.

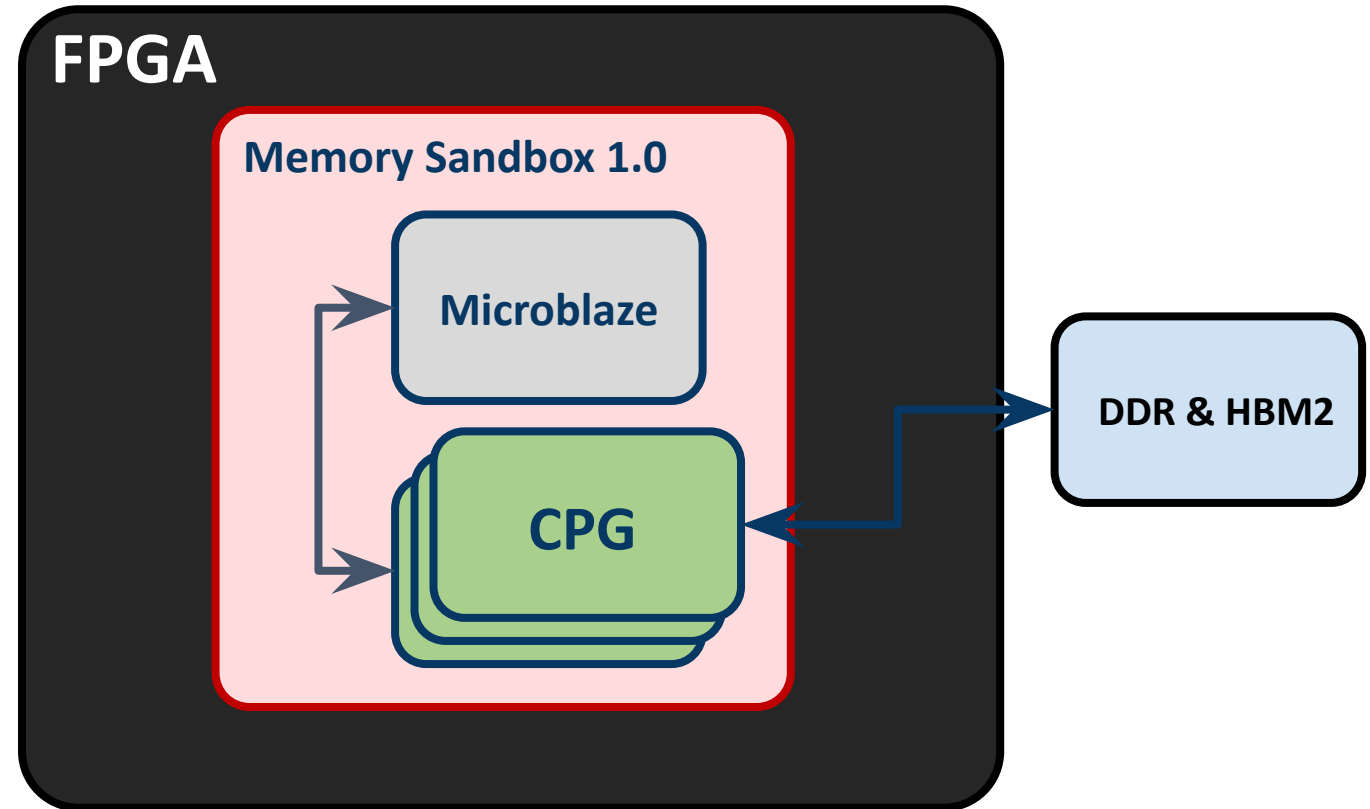


Memory Sandbox 1.0

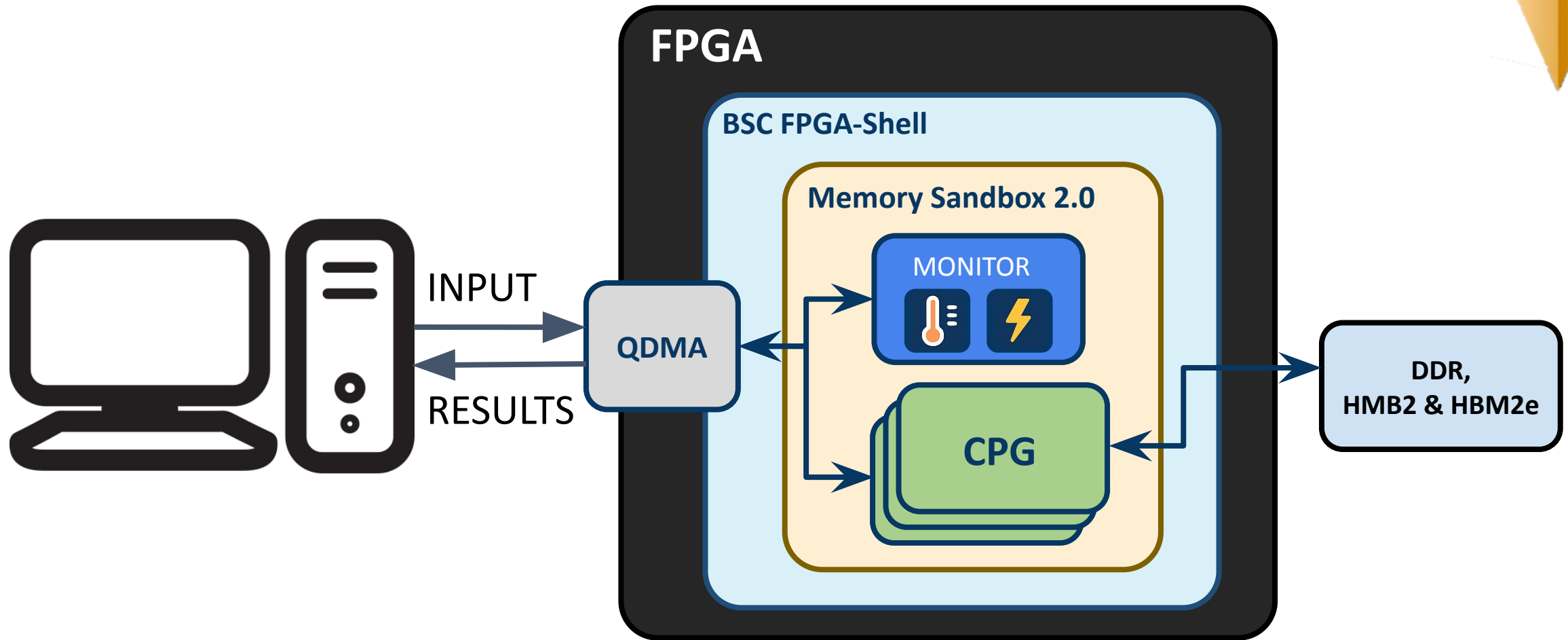


Memory Sandbox 1.0

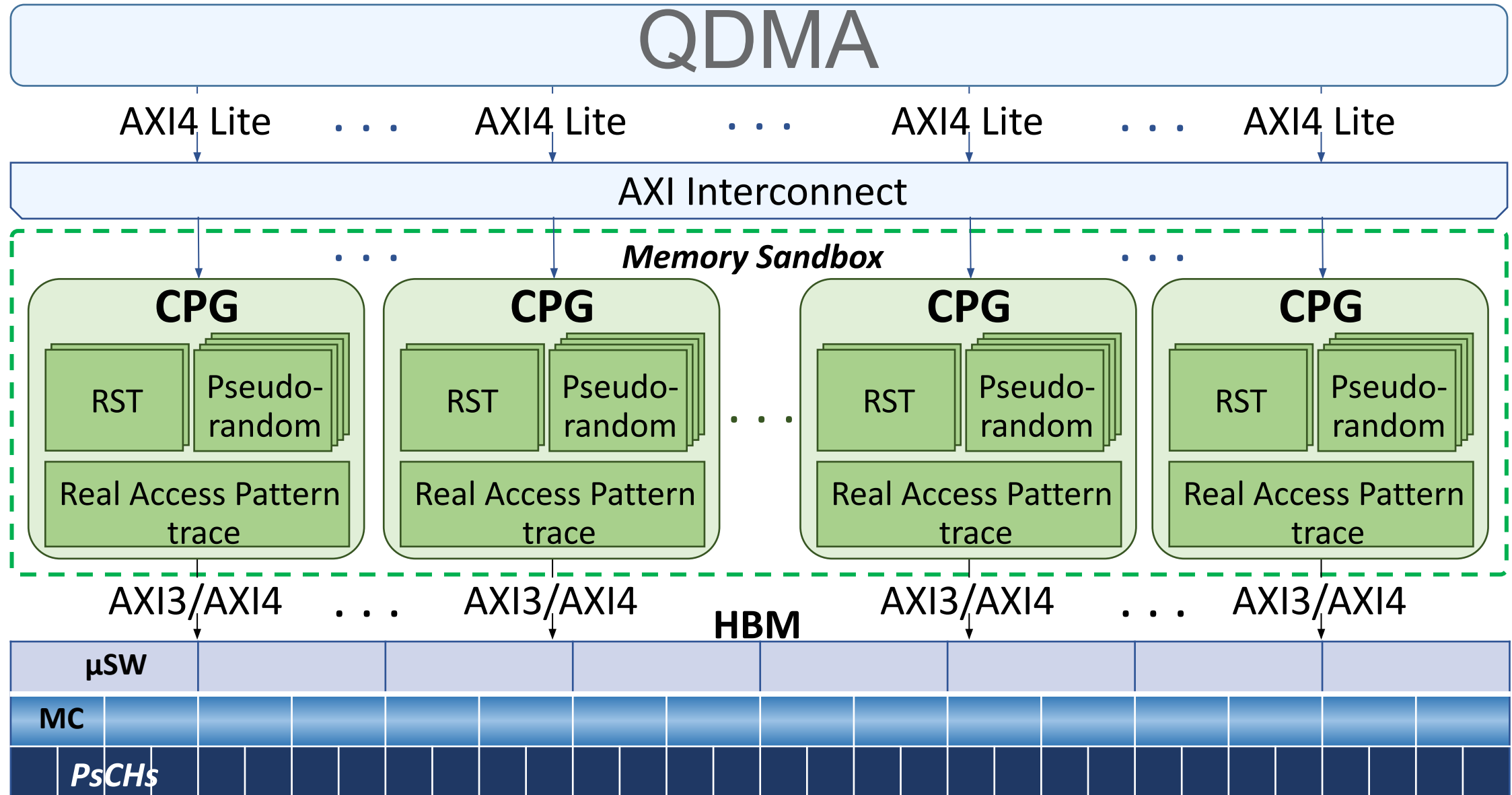
- Profiler without FPGA-Shell Automation or FPGA to PC communication.
- DDR & HBM
 - Throughput
 - Latency



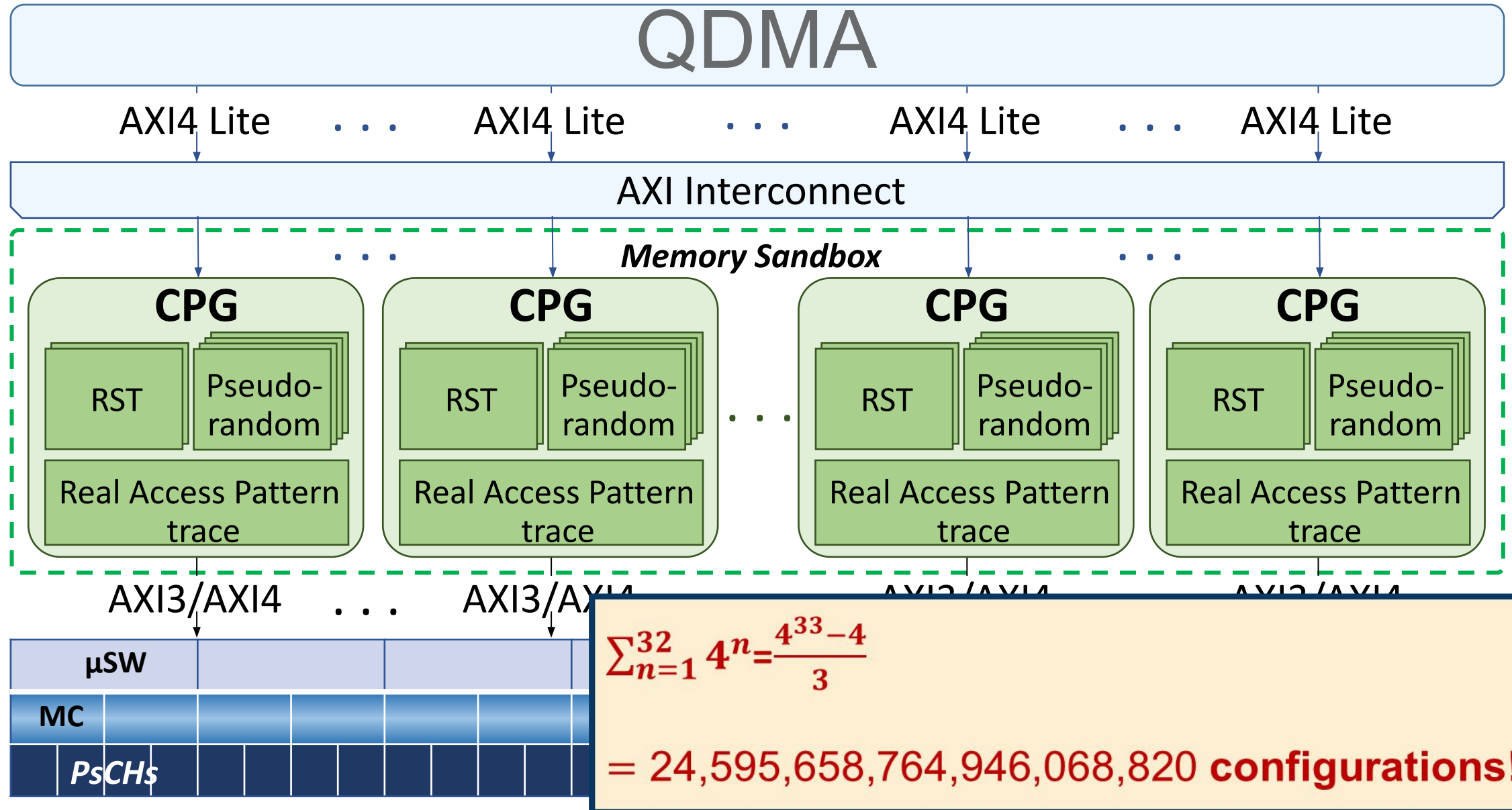
Memory Sandbox 2.0



Memory Sandbox 2.0: Throughput and Latency



Memory Sandbox 2.0: Throughput and Latency



Memory Sandbox 2.0: Throughput and Latency

- Higher configurability
- More control over measurements
- Further insights
 - *Memory transaction clock cycles*
 - *Precise bandwidth*

1) front-end

Run-time Parameters

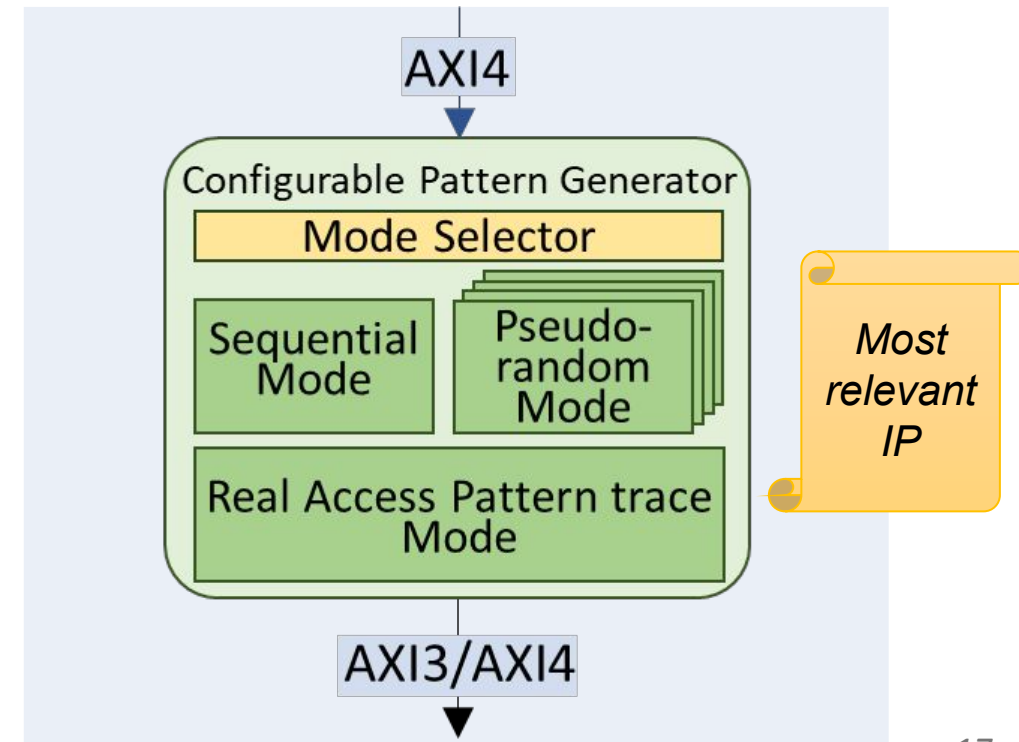
- *Burst Size*
- *Number of transactions*
- *PSCH Address Init*
- *PSCH Address End*
- *Rand Options*

Design Parameters

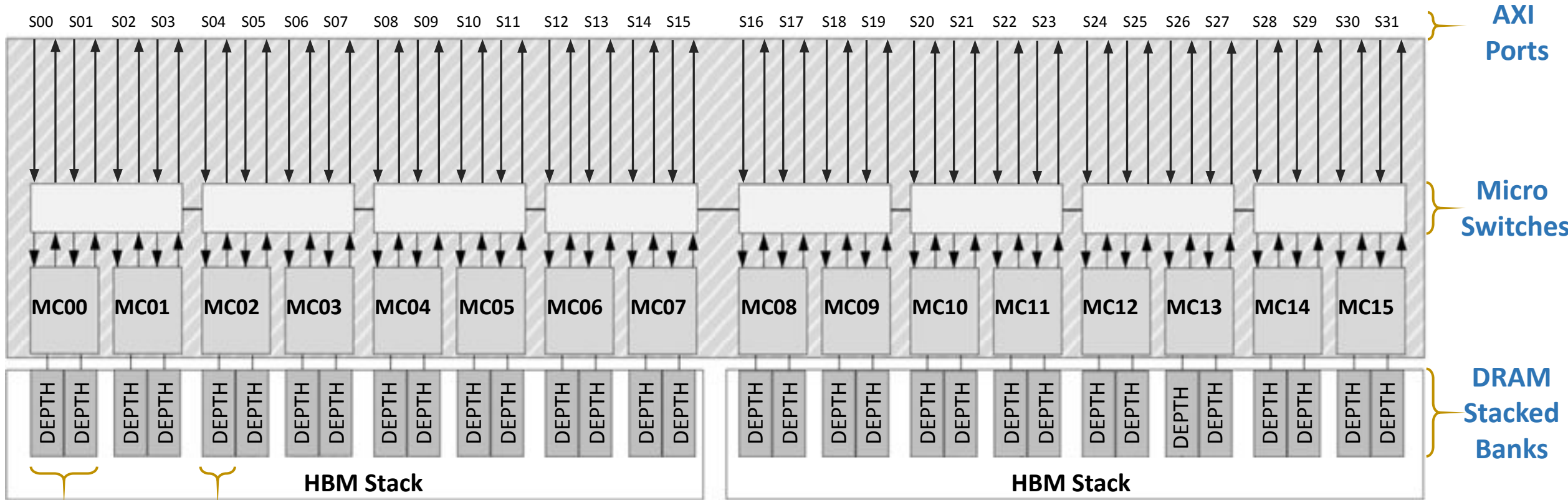
- *Memory: DDR/HBM*
- *Pattern mode*
- *Address mapping*



2) back-end



HBM topology in the Alveo Ultrascale+



Memory Channel
128b

Pseudo Channel
64b

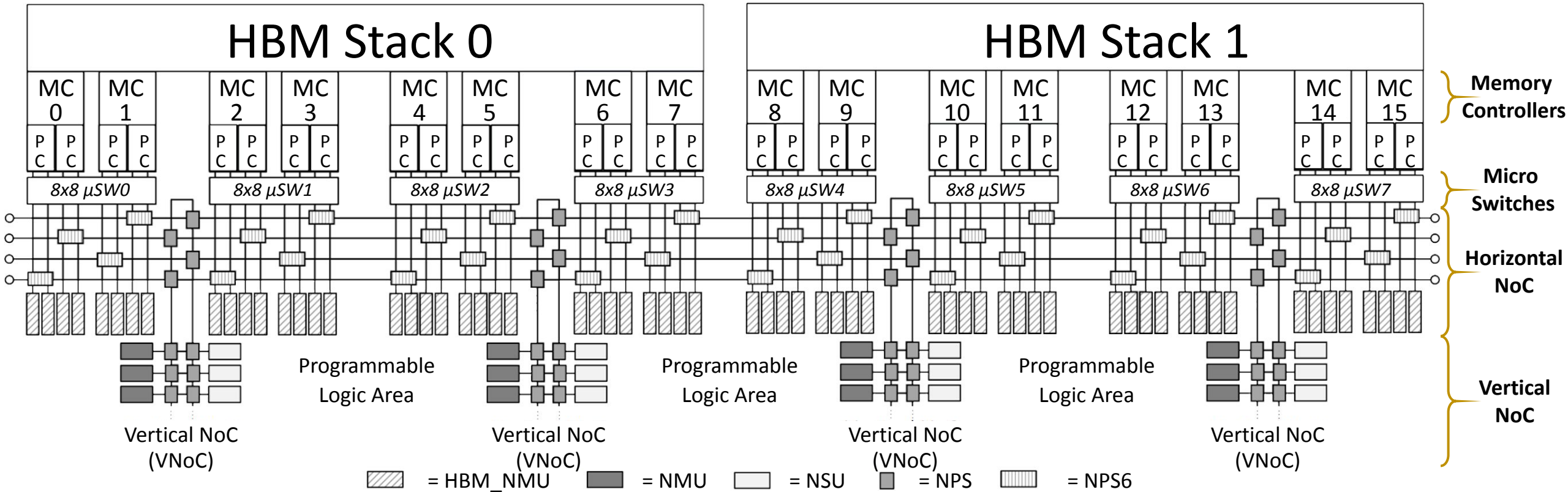


FPGA General Interconnect

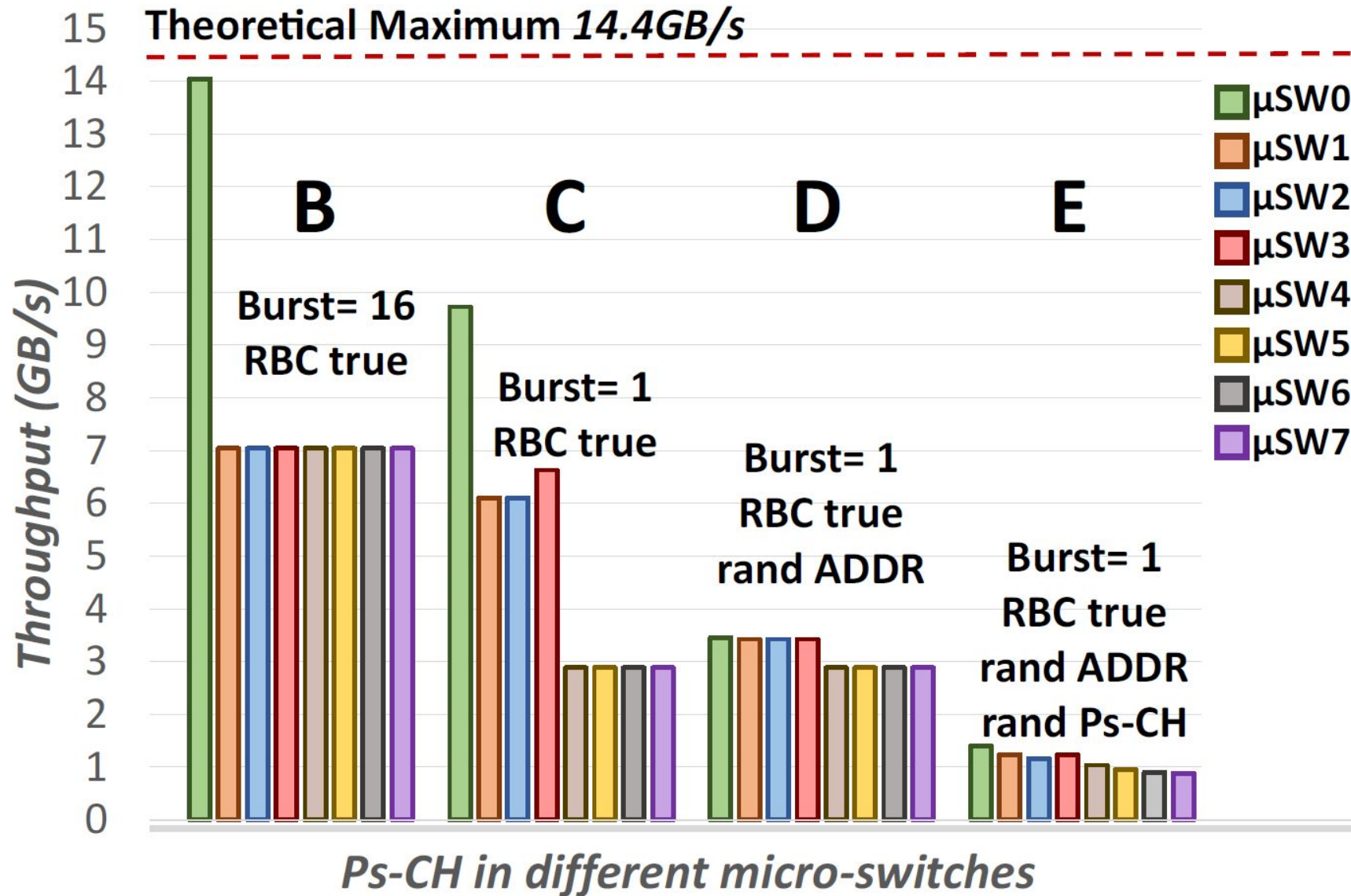


Board	Pseudo-Channel Depth	Stack Size
U55C	512 MB	8GB
U280	256 MB	4GB

HBM2e topology in the Versal V80



Throughput Results for HBM2 accessing different micro-switches: U280

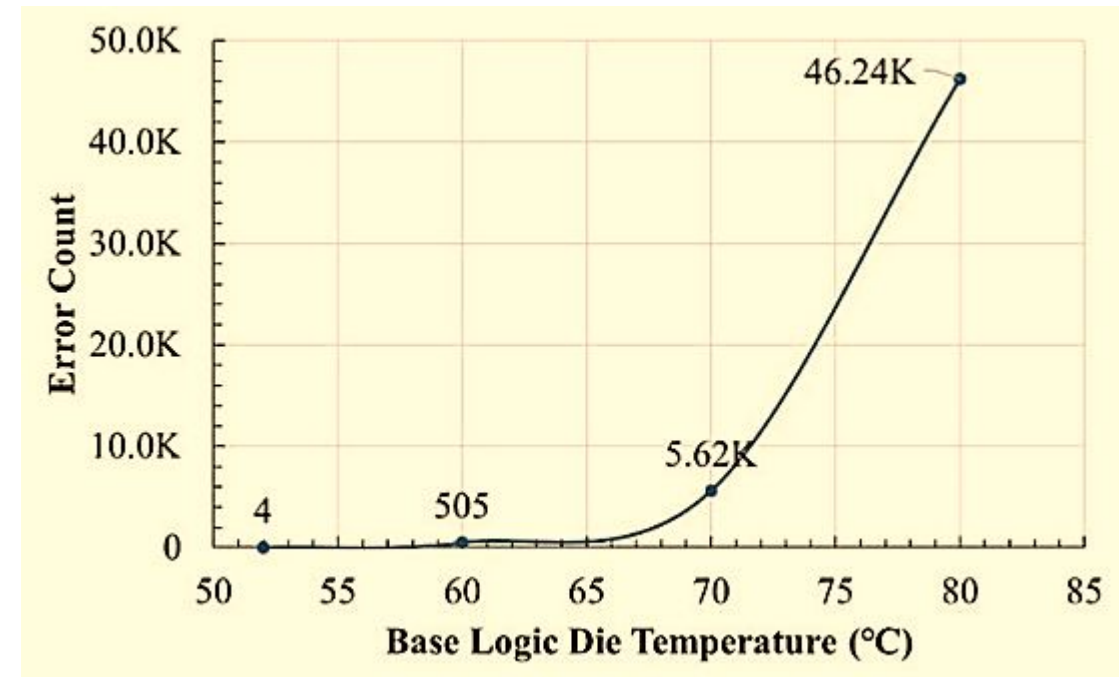


Memory Sandbox 2.0 Architecture: Telemetry

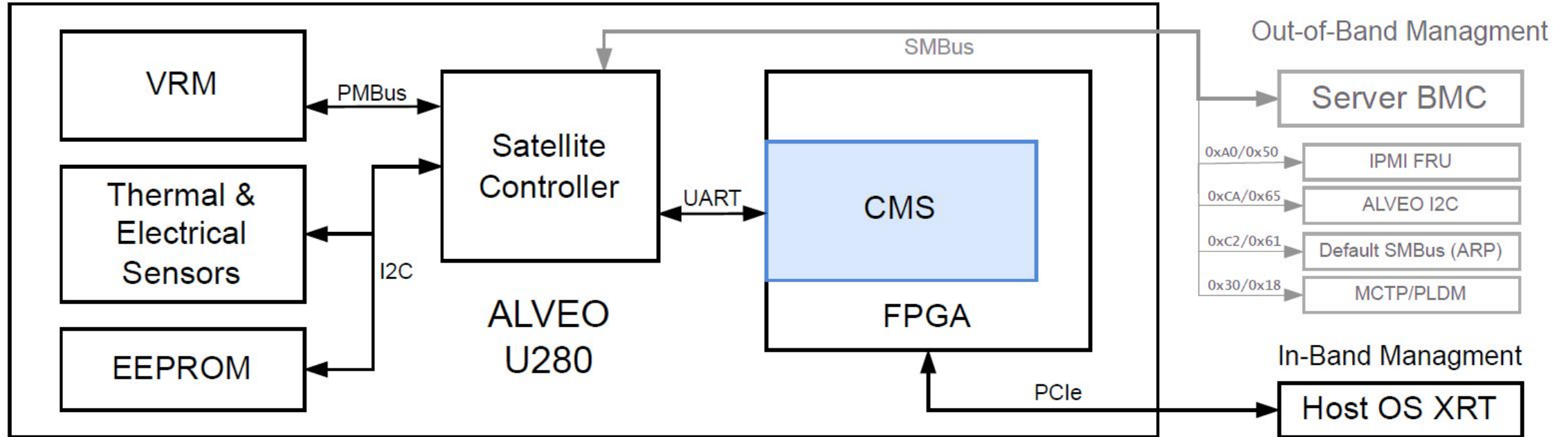
Why Telemetry?

- To better determine the correct state of the FPGA
- To characterize the FPGA when utilizing different resources
- To profile designs on the FPGA

Architectural behavior matters but so does physical!



Memory Sandbox 2.0 Architecture: Telemetry

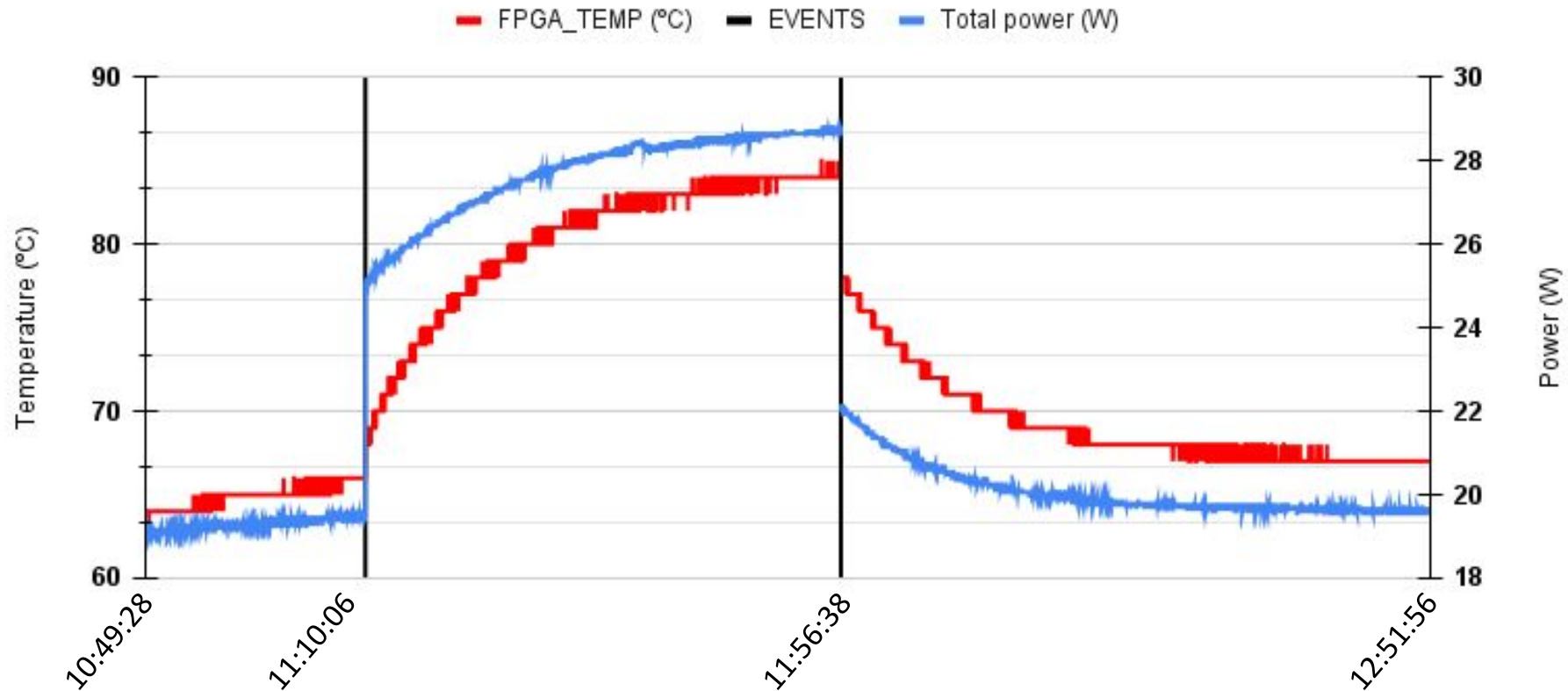


Memory Sandbox 2.0: Telemetry Results

- We achieve systematic correlation between telemetry and memory system behavior

Power and Temperature readings

Alveo U55C



Conclusions

- With our Memory Sandbox 2.0 we provide a modular and extensible framework designed to analyze memory system behavior across performance and telemetry dimensions. By refactoring our original tool, we have made it capable of easily integrating new memory technologies (such as HBM2e) and supporting custom access patterns, enabling more systematic design space exploration.
- With the enhanced parameterization and user interface improvements we allow developers to replicate hardware-relevant memory behavior, such as sequential, pseudo-random, and sparse patterns, without requiring a full accelerator implementation, as demonstrated with our SpMV-inspired access pattern.
- Moreover, moving most of the configuration parameters to runtime allows them to execute multiple experiments on the same bitstream.
- Each Configurable Pattern Generator (CPGs) can emulate diverse access schemes, allowing to study how memory-bound accelerators or CPUs behave in heterogeneous computing environments.
- To validate these capabilities, we are currently conducting a comparative analysis of HBM2 (Alveo U280) and HBM2e (Versal V80) in terms of throughput, latency, and telemetry (temperature, power).

Future Work

- V80 characterization
- Side by side telemetry-throughput experiments
- Trazer fueled pattern generation
- V80 Telemetry improvements

**Unofficially, still under discussion: Inclusion of CPGs into Xilinx shells/IP.*

Acknowledgements

This work has been co-financed by the Barcelona Zettascale Laboratory under project reference REGAGE22e00058408992, with support from the Spanish Ministry for Digital Transformation and Public Services, within the framework of the Recovery and Resilience Facility and the European Union – NextGenerationEU; from the research projects ST4HPC (PID2023-147979NB-C21) and the ACAP project (PID2023-146511NB-I00) supported by the Spanish Ministry of Science and Innovation (MICIU), the State Research Agency (AEI), and the European Regional Development Fund (FEDER). Additional support was provided by the Lenovo–BSC Contract Framework (2022) and the LOCA project (CEX2021-001148-S) funded by MCIN/AEI /10.13039/501100011033 and by “ERDF/EU”.



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



**UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH**

Memory Sandbox 2.0:

A Framework for Enabling HBM2e vs HBM2 Performance and Telemetry Analysis on Xilinx FPGAs

Elias Perdomo^{1,2}, **Joan Teruel**^{1,2},

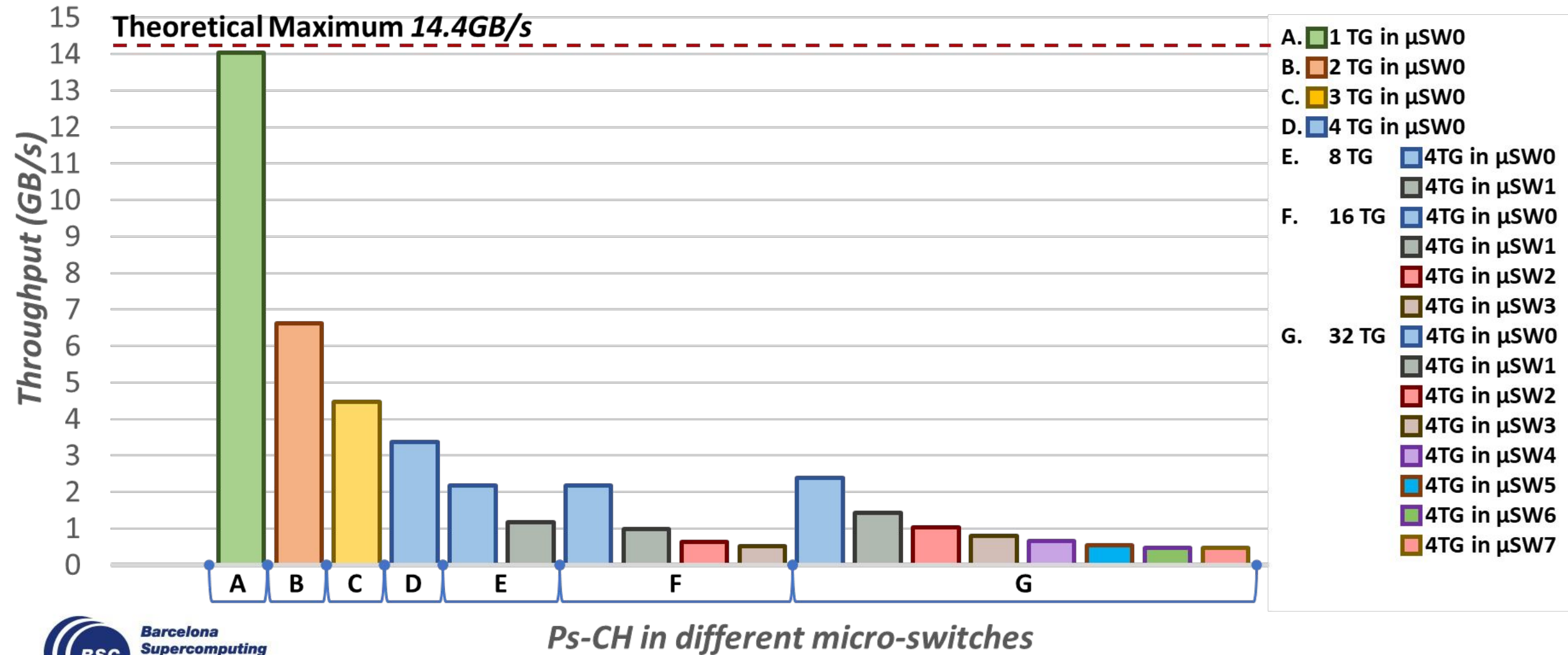
Sajjad Ahmed¹, Behzad Salami¹, Teresa Cervero¹ and Xavier Martorell^{1,2}

08/10/2025

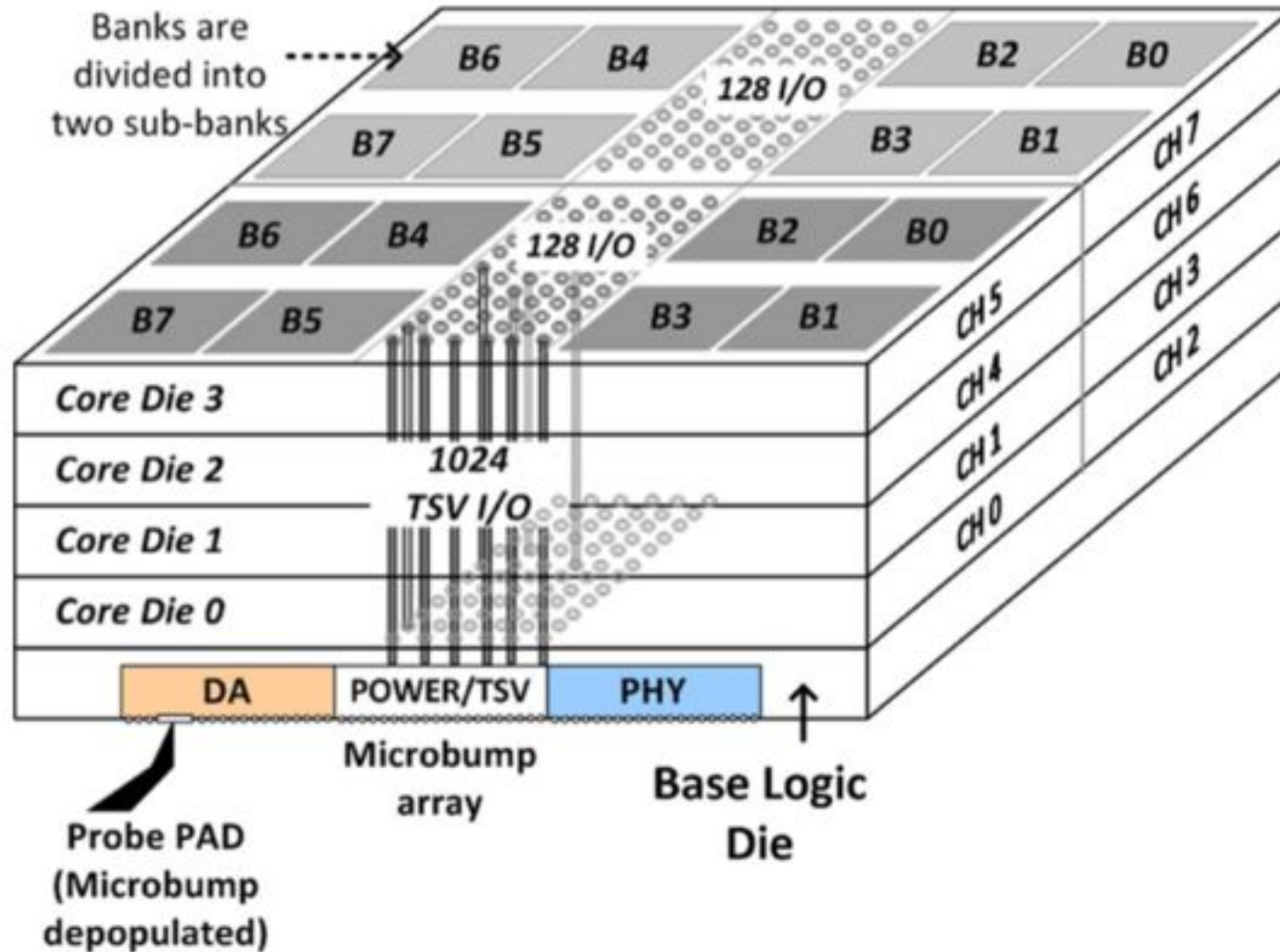
¹Barcelona Supercomputing Center

²Universitat Politècnica de Catalunya

Throughput Results for several AXI Ports accessing PSCH0



HBM Architecture



HBM DRAM DIES

