EPI Forum

Paris, 6-7 October, 2025

European Processor Initiative

IT TAKES A VILLAGE (AND A CONTINENT) TO BUILD A PROCESSOR: CO-DESIGN ADVENTURES IN EPI

Estela Suarez, Manolis Marazakis, Lilia Zaourar

EPI FORUM

Estela Suarez

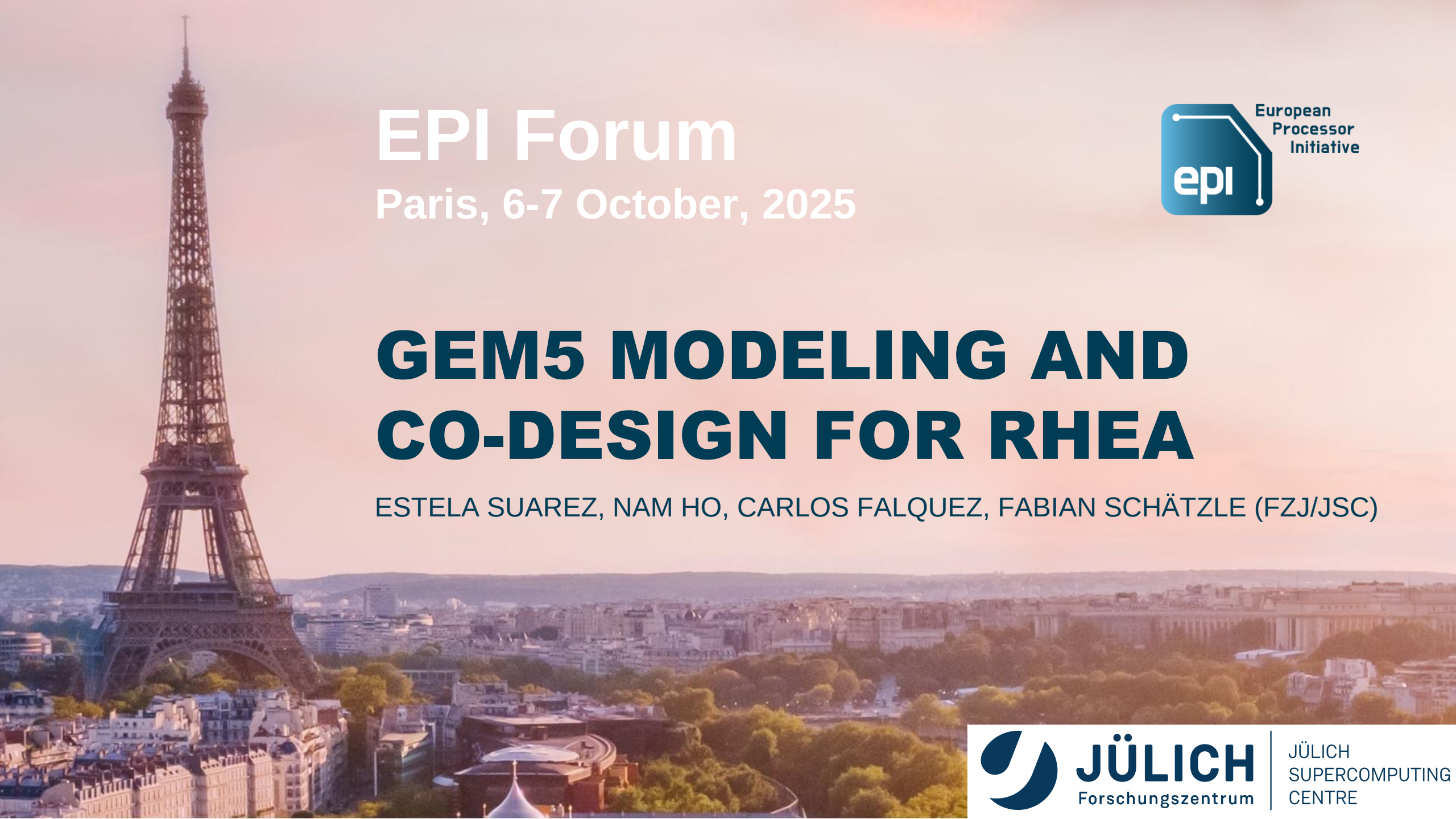
Join Lead of the department Novel System Architecture Design

Manolis Marazakis

Principal Staff Research Scientist

Lilia Zaourar

Expert Co-design



EPI Forum

Paris, 6-7 October, 2025

European Processor Initiative

epi

GEM5 MODELING AND CO-DESIGN FOR RHEA

ESTELA SUAREZ, NAM HO, CARLOS FALQUEZ, FABIAN SCHÄTZLE (FZJ/JSC)

JÜLICH Forschungszentrum

JÜLICH SUPERCOMPUTING CENTRE

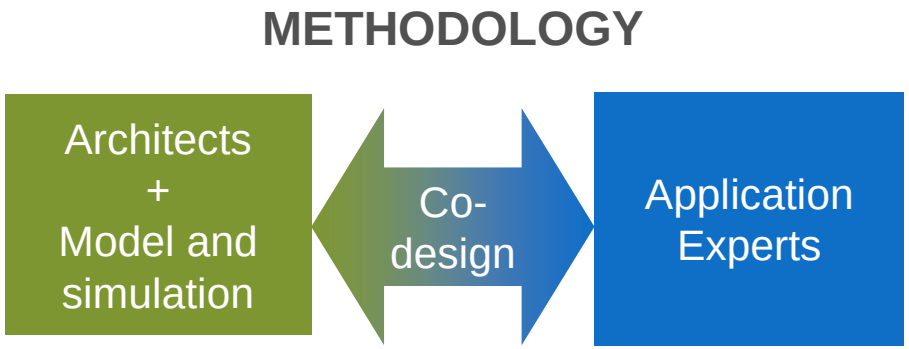
HW ↔ SW co-design

Why?

- Facilitate trade-off decisions to maximize performance and minimize costs under given technology boundary constraints

What?

- Bi-directional and iterative interaction process
- Application experts ↔ Hardware / Software developers



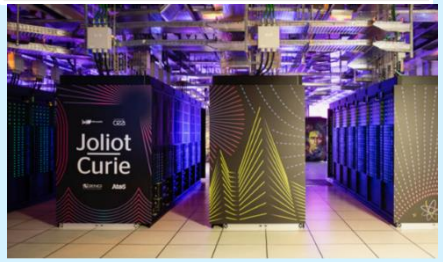
Need for co-design

- Validate your design choices (HW & SW)
- Cores, memory, accelerators, chiplets, interconnect,…
- Metrics evaluation: performance, power, thermal, …
- HW and SW debugging: easier and faster to debug in software mode

EPI: Co-design and Validation

Hardware platforms

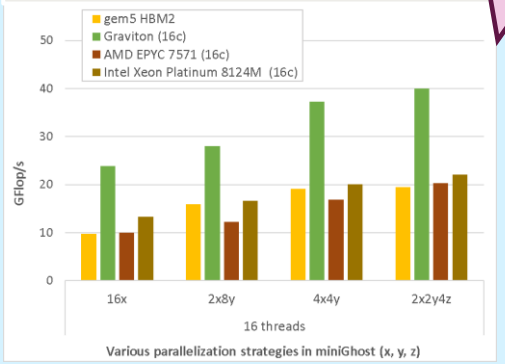
Arm Reference Platforms



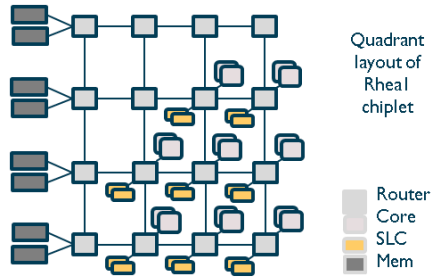
RISC-V SDVs

Validation and Co-design

Validation



Co-design

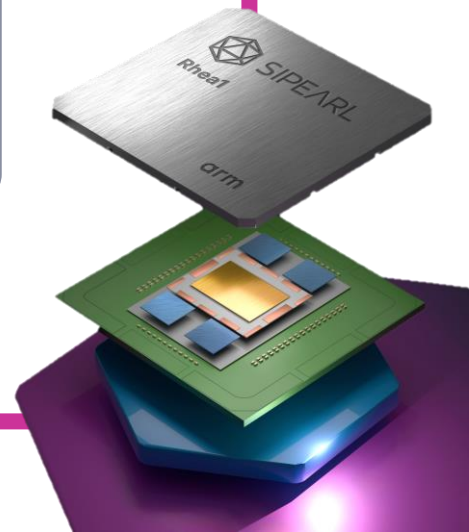


Multi-level:

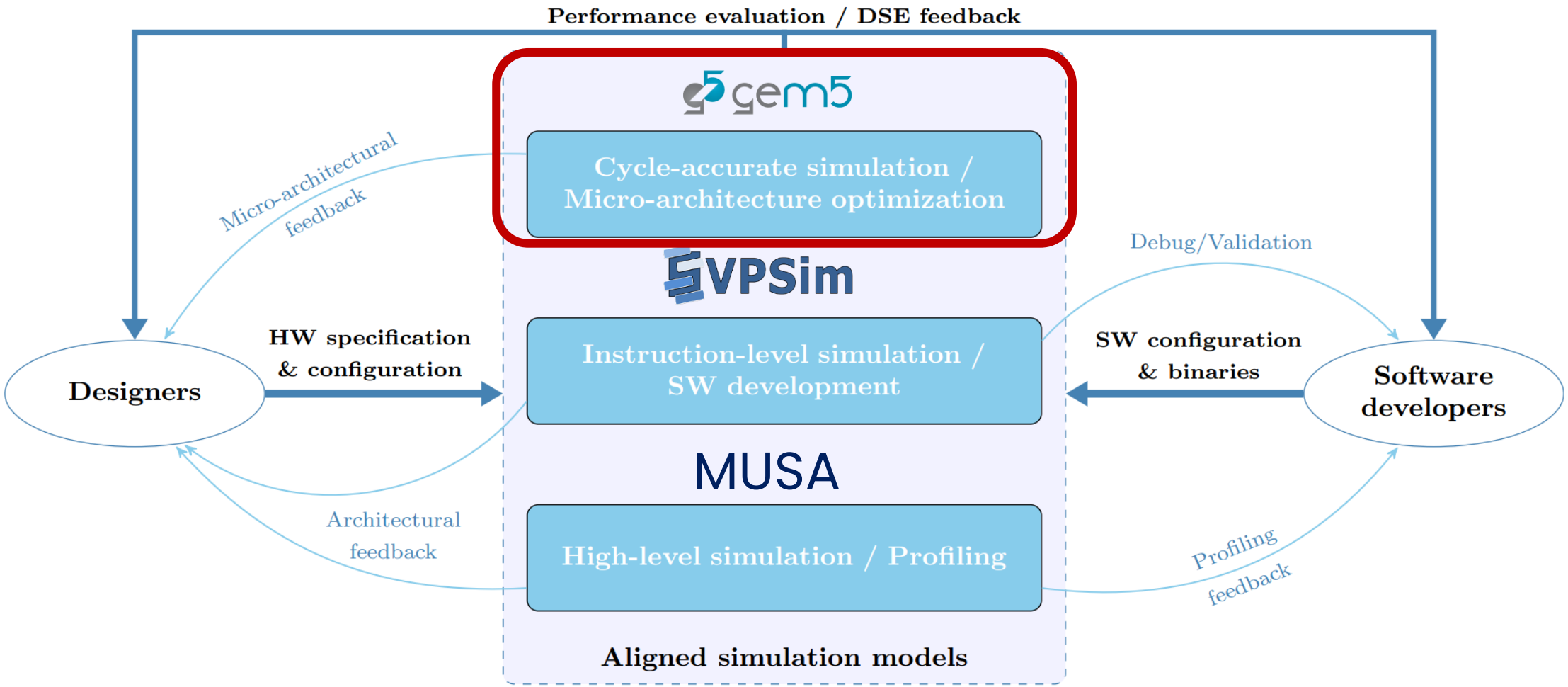
- benchmark suite
- models & simulators

Chip architects

Future Chips



Full HW/SW co-design approach



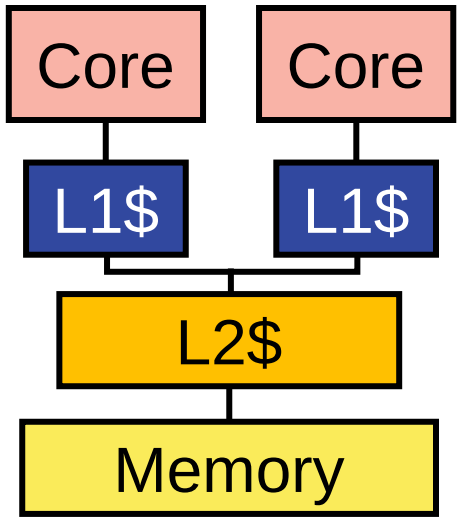
L. Zaourar et al., Multilevel simulation-based co-design of next generation HPC microprocessors. 2021 International Workshop PMPBS (PMBS), SC21

Gem5 simulator

Lowe-Power et al. 2020, *The gem5 Simulator: Version 20.0+.*
https://doi.org/10.48550/arXiv.2007.03152

European Processor Initiative

hardware design



simulation configuration script

```
system = System()
system.cpu = OOO_CPU()
system.cpu.width = 8
system.l1 = Cache()
…
system.l1.mem_side = \
  system.l2.cpu_side
…
system.workload = \
  ‘hello.exe’
simulate()
```

running gem5

```
> ./gem5 script.py
```

application output

```
Hello world!
```

statistics

```
l1.misses 2836
l1.hits  10374
cpu.ipc    1.3
```

Cycle-level computer architecture simulator

- **Event-driven** simulation engine
- **Object-Oriented** modular design

Large number of models

- CPU cores
- DRAM devices
- On-chip interconnect
- Cache coherency

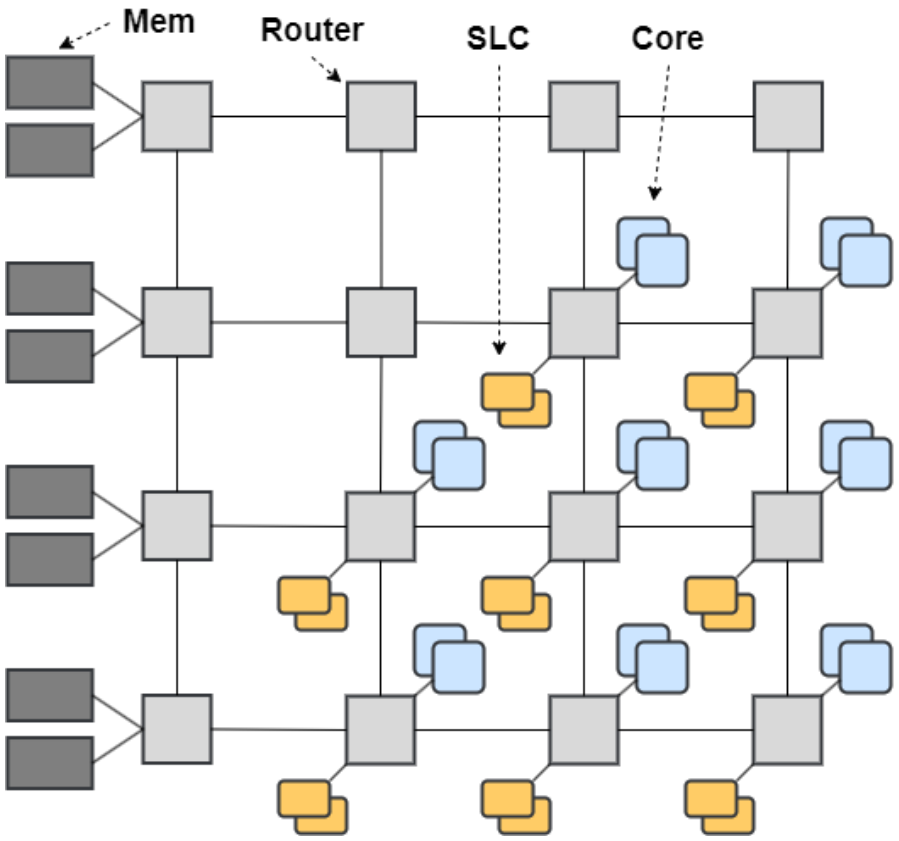
Comprehensive run statistics

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

EPI gem5 benchmarking environment

Set of python scripts for:

- **Better reproducibility**
  - Automatic generation of input files for full system simulation
  - Generate benchmark binaries, simulation images and bootloaders in a reproducible way
- **Better model configuration**
  - Replace default procedural configuration (hand-written python scripts) with config-file configuration (python files automatically generated by the framework)



Basic simulation, NoC upper quadrant

Falquez (FZJ-JSC, 2025) https://github.com/FZJ-JSC/gem5-dbc

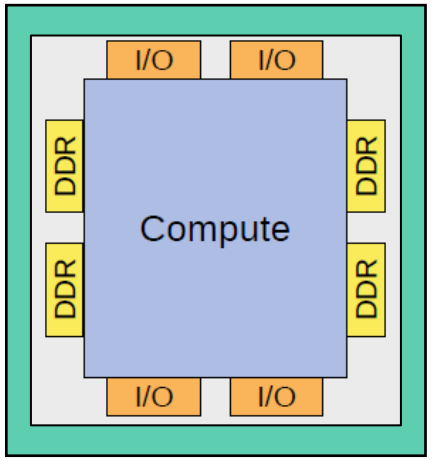
Features added to gem5 in EPI

EPI Requirements	New feature
Accurate **model of cache coherency** and **network traffic**	**Garnet support for AMBA CHI protocol** (detailed interconnection network)
Achieve **bandwidth performance** of **modern Arm mesh networks**	Garnet support for **multiple links** carrying **same VNETs**
Model effect of **NUMA topologies** and latencies, including hybrid memory	Support **NUMA node identifiers** for CPU and memory, including hybrid memory
Modern **Arm CPU caches support** write streaming mode as a cache bypassing technique	Dynamic Write Streaming Mode
Model **modern Arm CPU performance** and instruction throughput	**Improved gem5 CPU model frontend** to achieve instruction throughput closer to current Arm CPUs (V1)

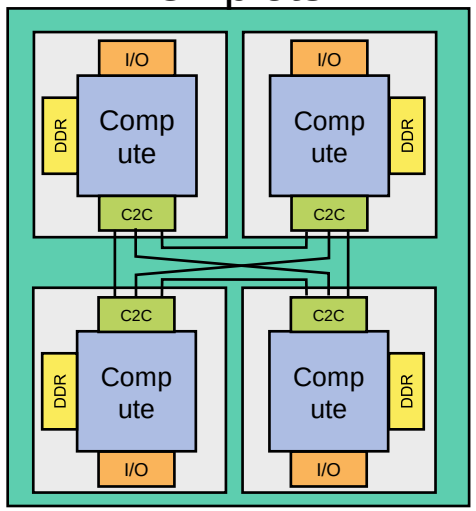
Falquez (FZJ-JSC, 2025) https://github.com/FZJ-JSC/gem5-dbc

Example: Chiplet-to-Chiplet (C2C) communication

Monolithic

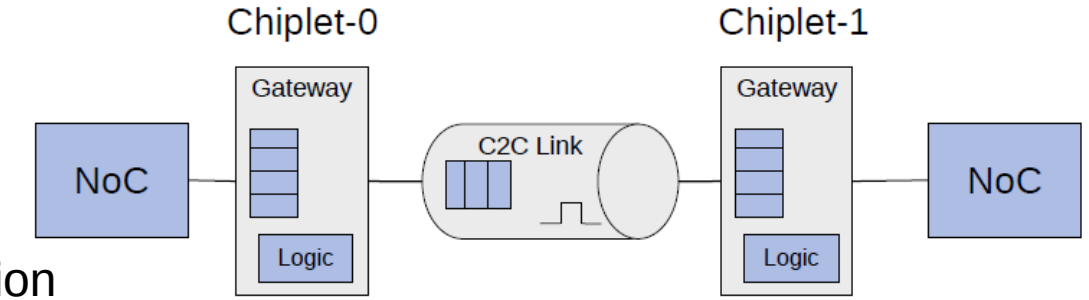


Chiplets

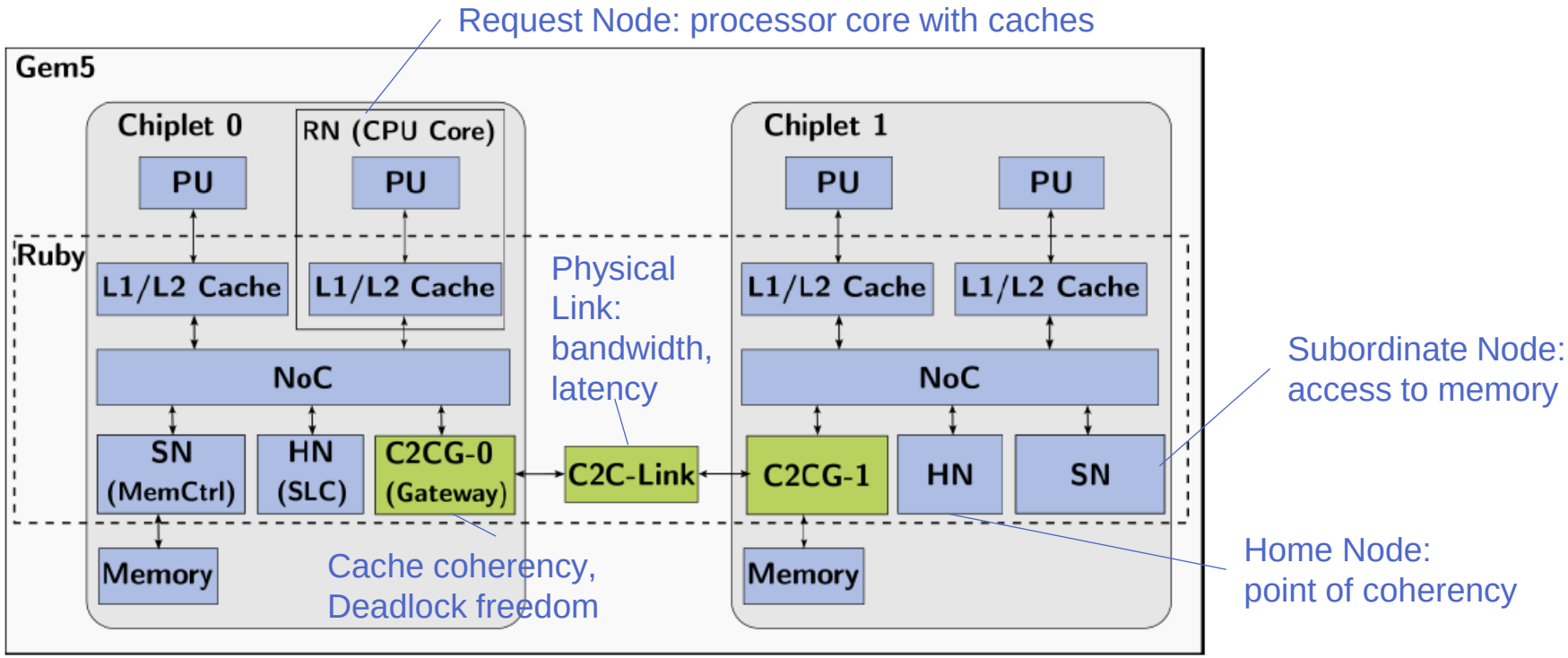


Chiplet processor:

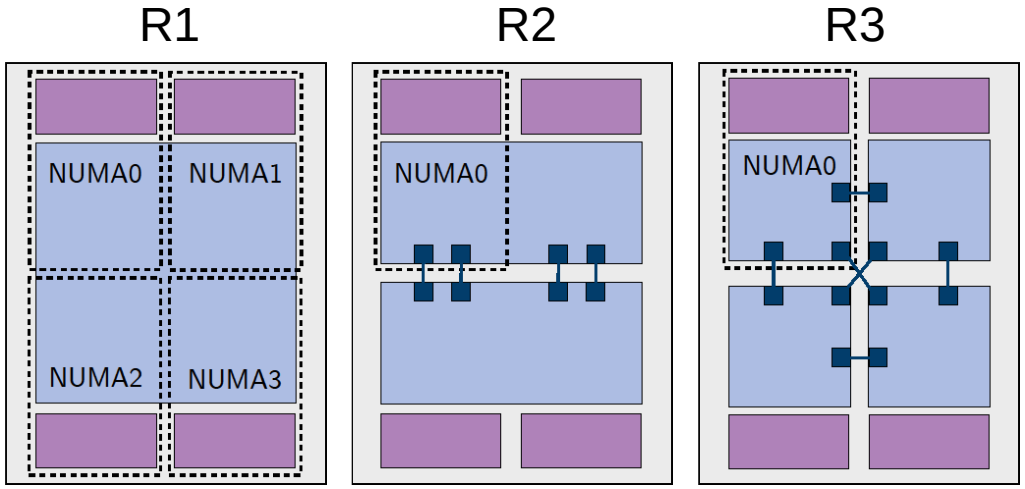
- Each has it own NoC
- Gateways have added logic
- C2C link buffers flits for transmission

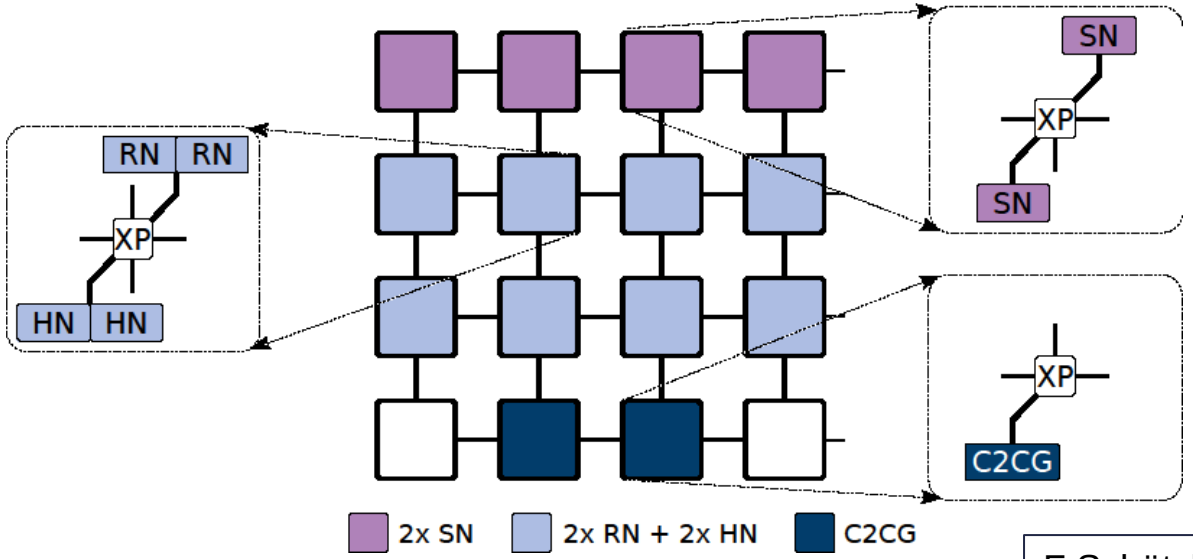


Gem5 simulation of C2C communication:



Example: Chiplet-to-Chiplet (C2C) communication





Evaluation Scenarios:

Design	R1	R2	R3
Core (RN)	Armv8 O3 CPUs; SVE 2x256 bit; L1I: 64KiB, L1D: 64KiB, 4-way, L2: 512KiB		
#Cores per die	64	32	16
NoC	Routing: XY;		
	Mesh 8x8	Mesh 4x8	Mesh 4x4
#L3 (HN)	64 x 1MB	32 x 1MB	16 x 1MB
Memory	HBM2; Bandwidth: 38.4GB/s per channel		
#Channels	32	16	8
C2C	N/A	**#C2CG: {2,4}**	**#C2CG: {3,6}**
		Transaction Table size (#TXN): **{128,256,512}** BW (GB/s): **{64,128,256,512}** TxQueue (TXQ): **{128B,256B}**	

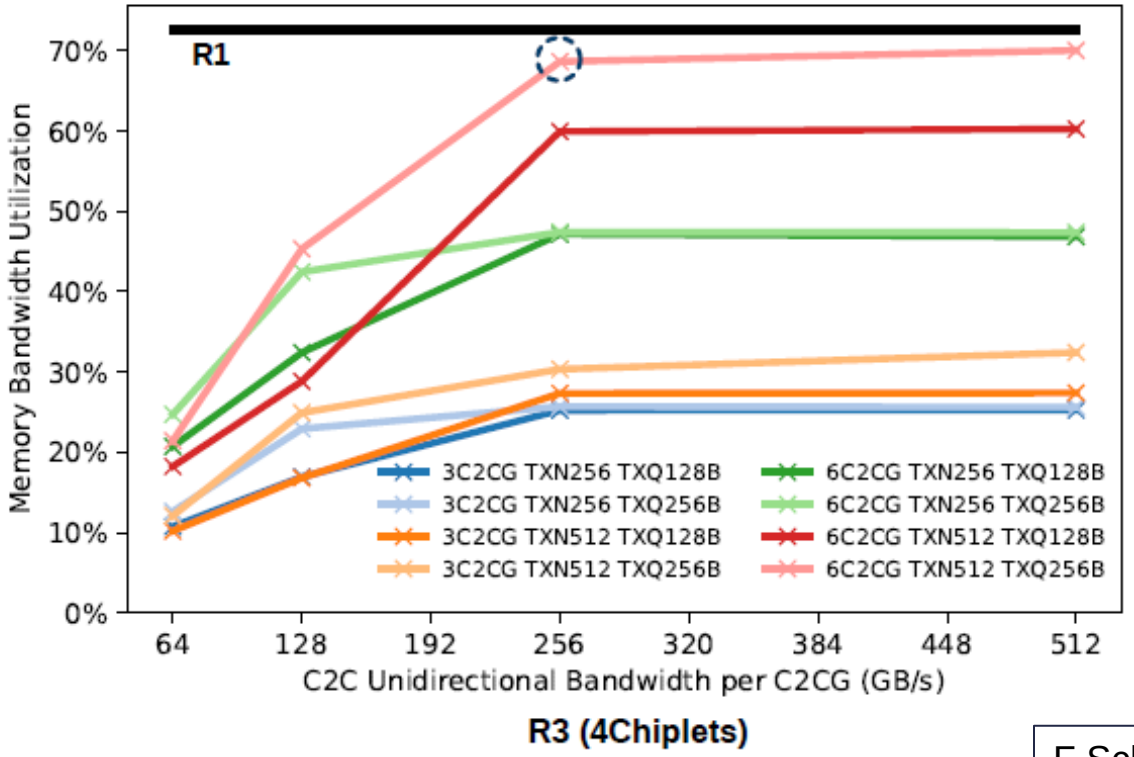
F.Schätzle (FZJ-PGI), ISC 2025, https://ieeexplore.ieee.org/document/11018303

Example: C2C communication

Memory Bandwidth (STREAM Triad)

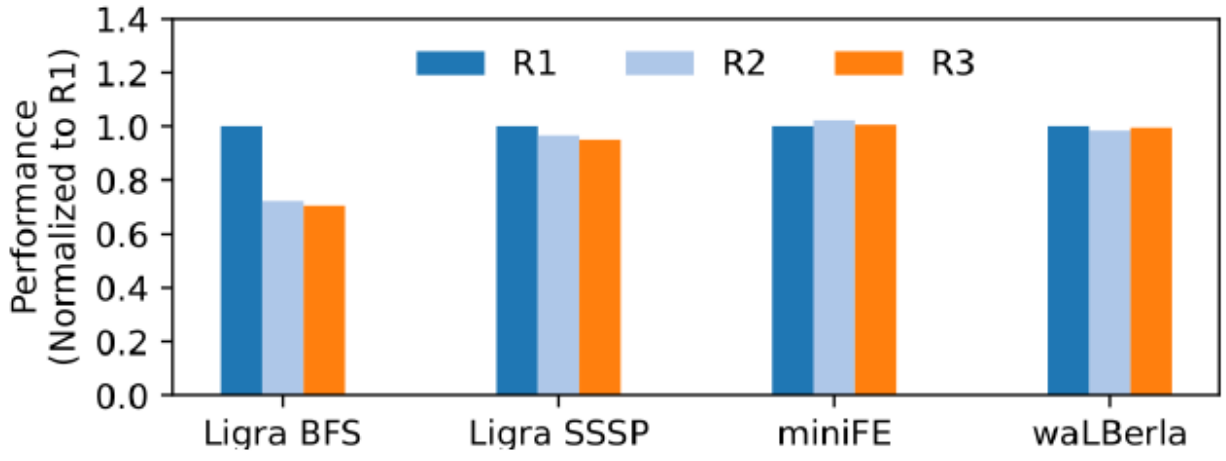
- Allocate vectors in one chiplet and activate CPUs on remote chiplet

Tune C2C parameters to make design trade-offs



Performance on HPC benchmarks

- Ligra shows high C2C traffic, decreasing performance, due to higher latency
- miniFE and waLBerla show low data sharing and hence low C2C traffic



F.Schätzle (FZJ-PGI), ISC 2025, https://ieeexplore.ieee.org/document/11018303

EPI Gem5 model:

Falquez (FZJ-JSC, 2025)
https://github.com/FZJ-JSC/gem5-dbc

AN OPEN AND POWERFUL TOOL FOR PROCESSOR CO-DESIGN

- Describe detailed microarchitecture details
- Run applications on *not-yet-existing* hardware
- Identify sources of performance losses and gains
- Provide quantitative insight
- Applicable to address different case studies
- Guide choices of processor architecture in a quantitative manner

See Next

EPI Forum

Paris, 6-7 October, 2025

European Processor Initiative

THANK YOU!

ESTELA SUAREZ, NAM HO, CARLOS FALQUEZ, FABIAN SCHÄTZLE (FZJ/JSC)

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

EPI Forum

Paris, 6-7 October, 2025

European Processor Initiative

HIGHLIGHTS OF WHAT-IF ANALYSIS USING GEM5: NUMA, TOPOLOGY, NOC, HYBRID MEMORY

POLYDOROS PETRAKIS, VASSILIS PAPAEFSTATHIOU, MANOLIS MARAZAKIS (FORTH)

FORTH
FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

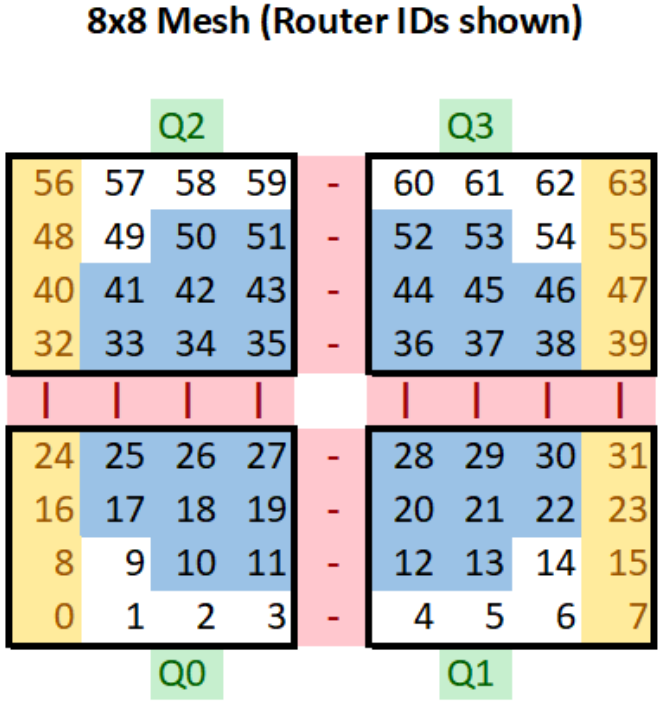
“ELEMENTARY, MY DEAR WATSON”

- What-if analysis of alternative designs and the impact of their parameters
- 4 “teasers” of case studies based on simulation of Rhea1-based Linux system, in gem5:
  - NUMA impact (NoC-level, Application-level)
  - Topology alternatives: relative placement of cores and SLC slices
  - NoC-level effects: impact of multiple VNETs
  - Impact of page-level migration across hybrid memory tiers

Holmes-style deductive reasoning to investigate elusive behaviors and trade-offs in modern HPC processors

“Come, Watson! The game is afoot”... *and so are our simulations*

4 NUMA NODES IN AN 8X8 NOC

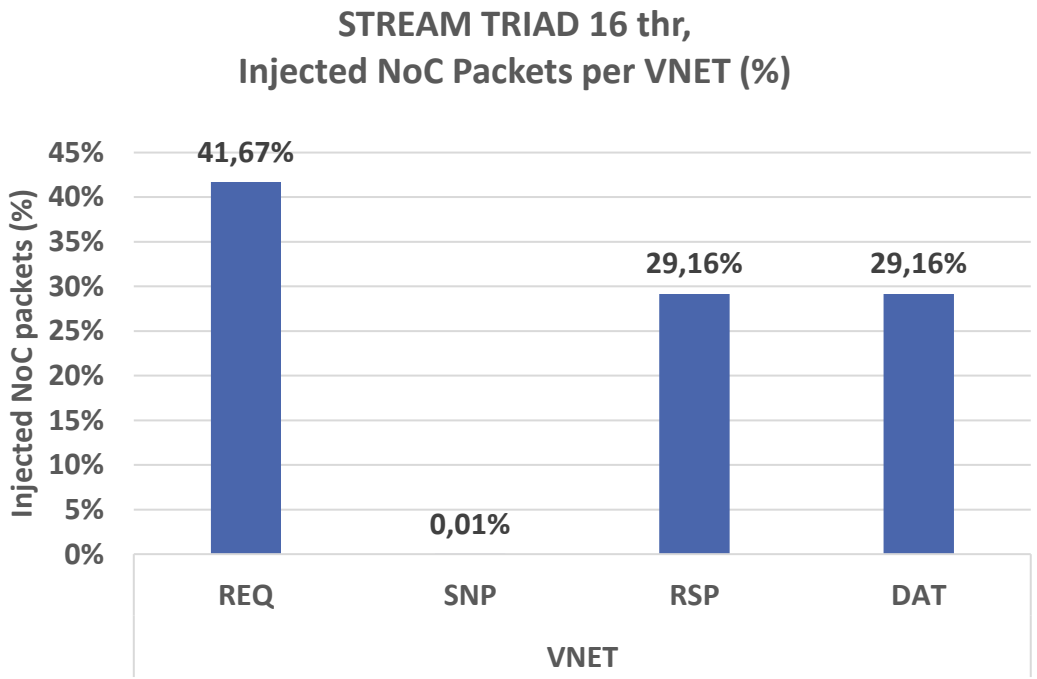


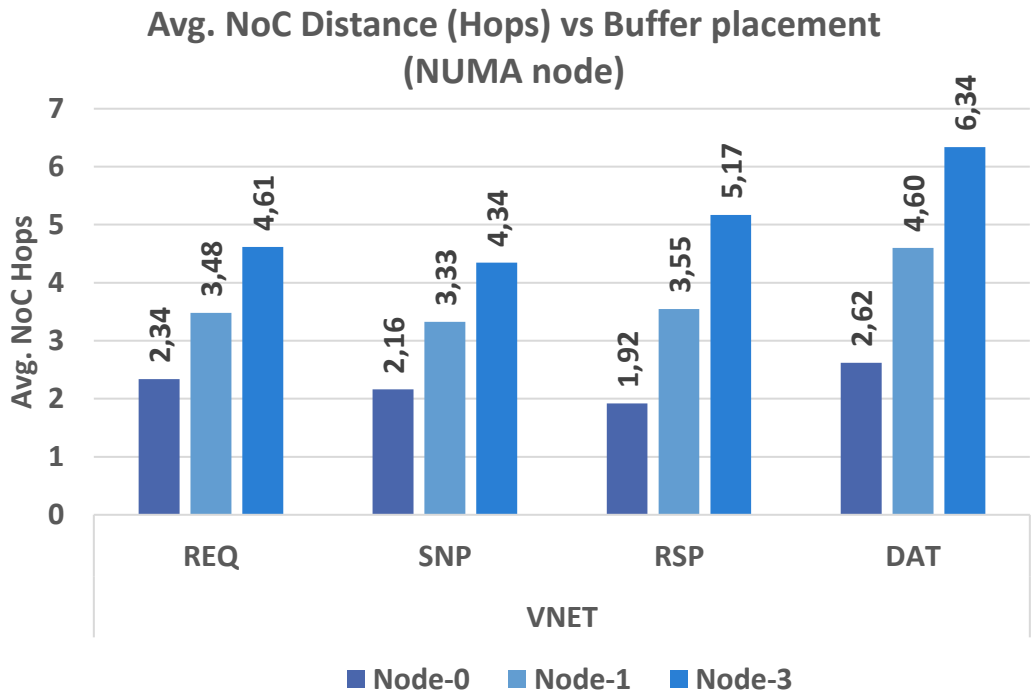
- 64 cores, with 64 SLC slices (1MB/ Slice)
- We run STREAM TRIAD with 16 threads bound on NUMA-node0 (first 16 cores)
- We use numactl and membind to direct memory allocation placement
- The allocated buffer is placed on a different NUMA-NODE each time
  - We tested with NUMA-NODES 0, 1 and 3
- We experiment with different “inter-NUMA” link-latencies
- Example command:
  - numactl --membind=1 ./**sve_triad.exe** -s 10000000
- Problem size: 10M elem. array size - Total: 228MiB

MEASURING NUMA IMPACT (1/2)

AMBA-CHI VNETs:
REQ = Request, SNP = Snoop
RSP = Response, DAT = Data

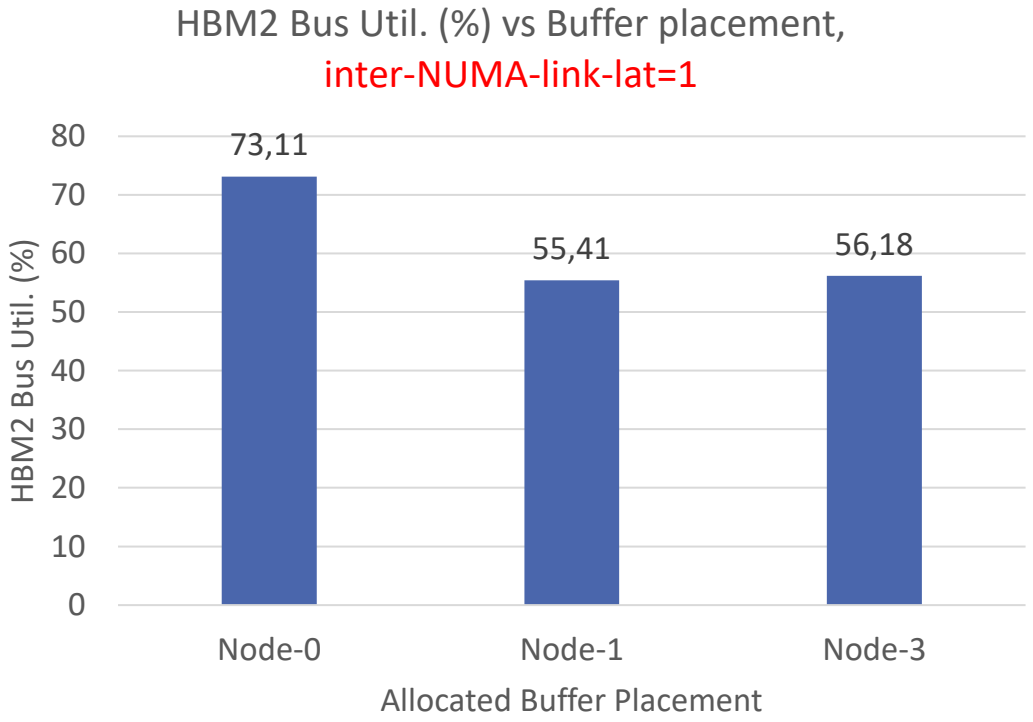
Avg. NoC hops increase as we move the allocated buffer further away from NODE-0.

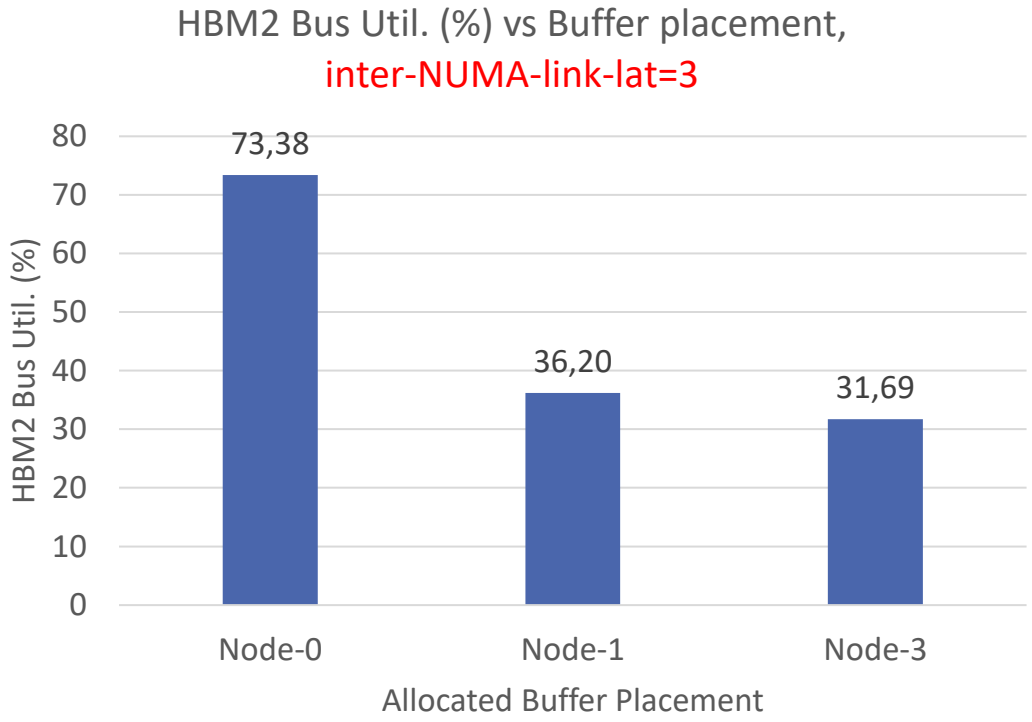




MEASURING NUMA IMPACT (2/2)

Note: Reported Bus Util. (%) considers only the HBM2 controllers of the "target" NUMA node (where the buffer is allocated).



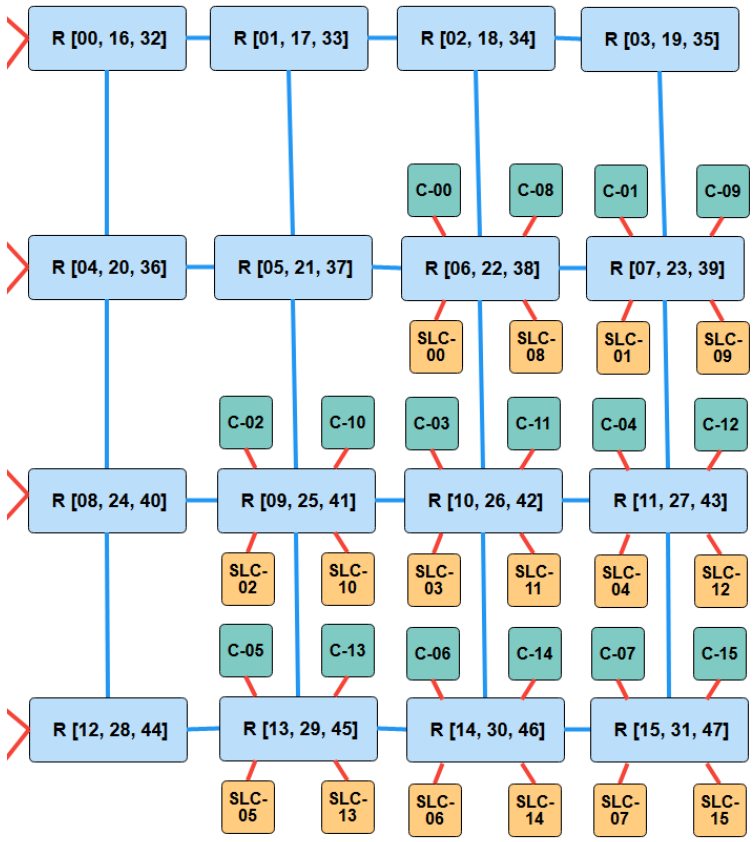


TOPOLOGIES EXPLORATION

Case Study #2: Topology

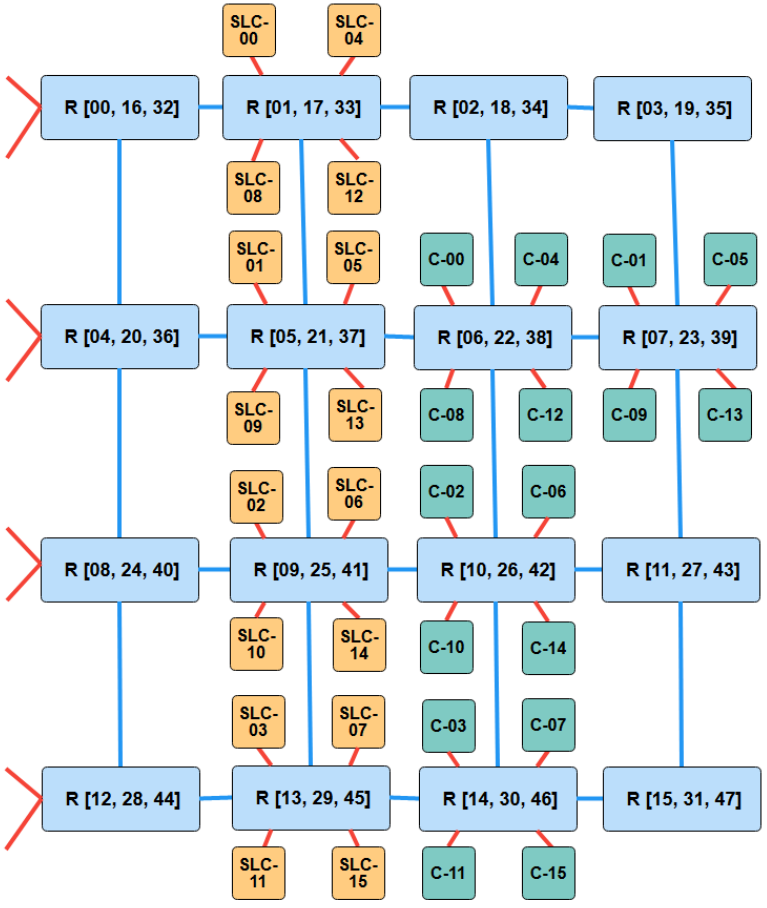
HBM2 (omitted here) is attached at the left side of each Topology.

[cores & SLC slices intermingled]



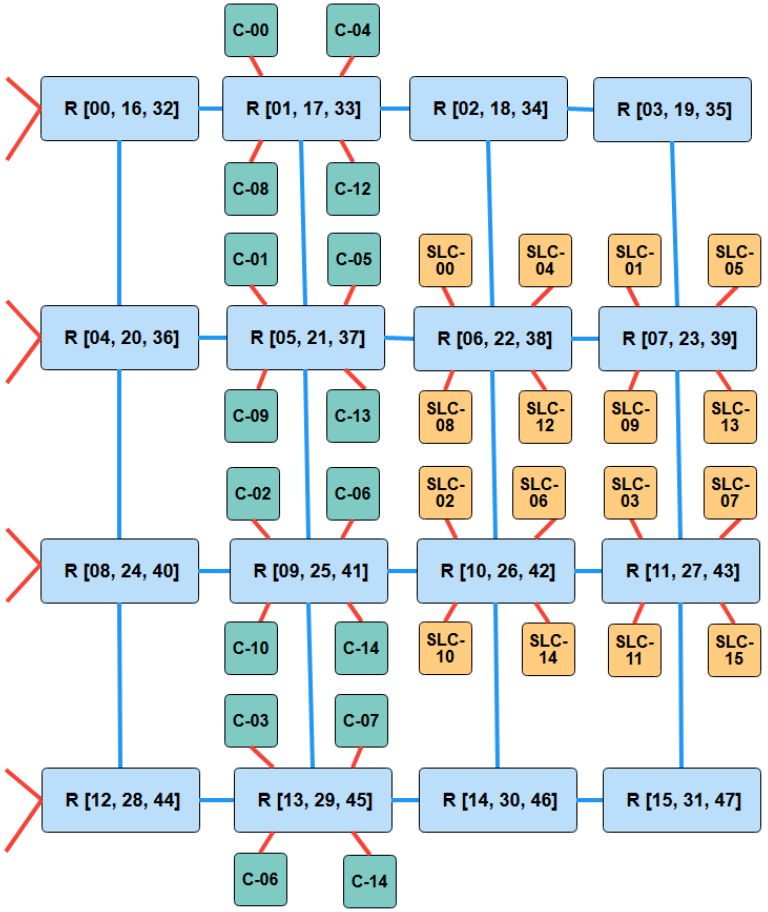
Topology-1,
Default

[SLC slices placed closer to memory]



Topology-2
SLC close to HBM2

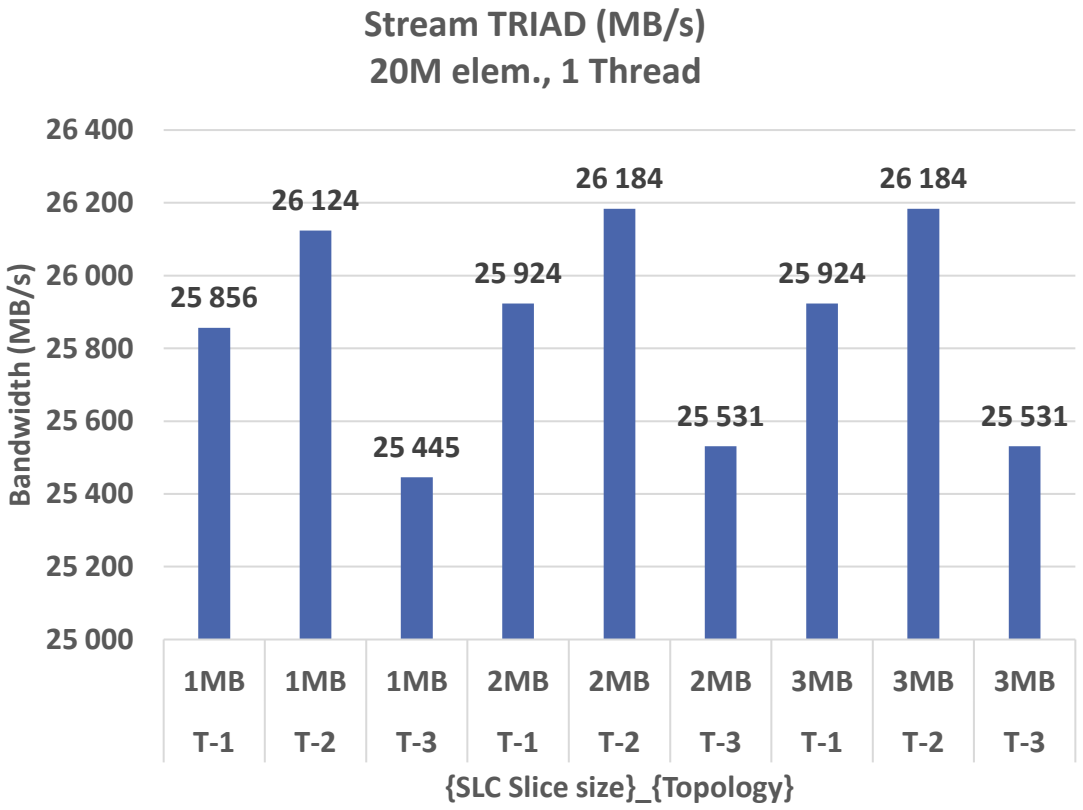
[cores placed closer to memory]

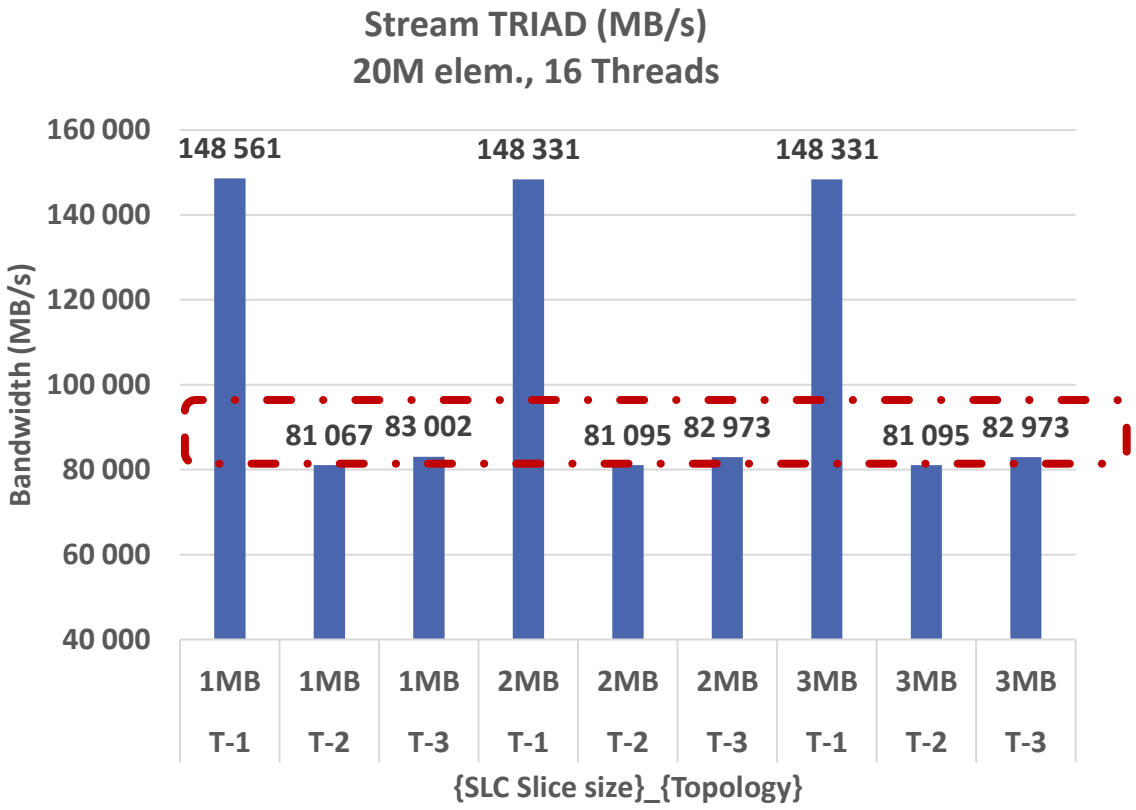


Topology-3
Cores close to HBM2

STREAM TRIAD – SLICE SIZE VS NOC TOPOLOGY

Significant performance drop for T-2 & T-3, with 16 threads.





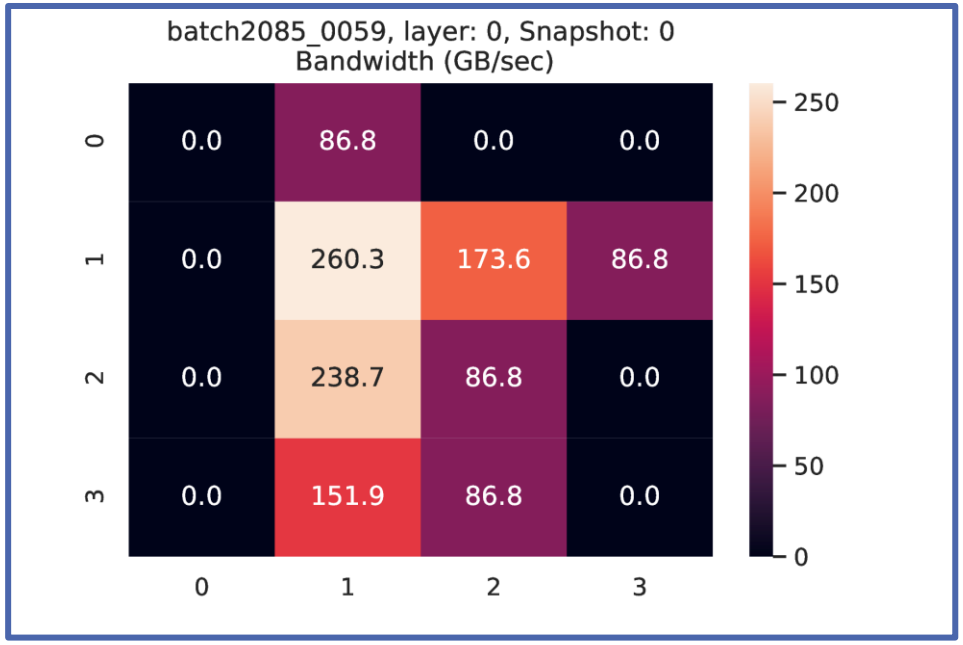
Case Study #3: VNETs

NOC HEATMAPS, STREAM TRIAD, 16THR. @ T-2

European Processor Initiative

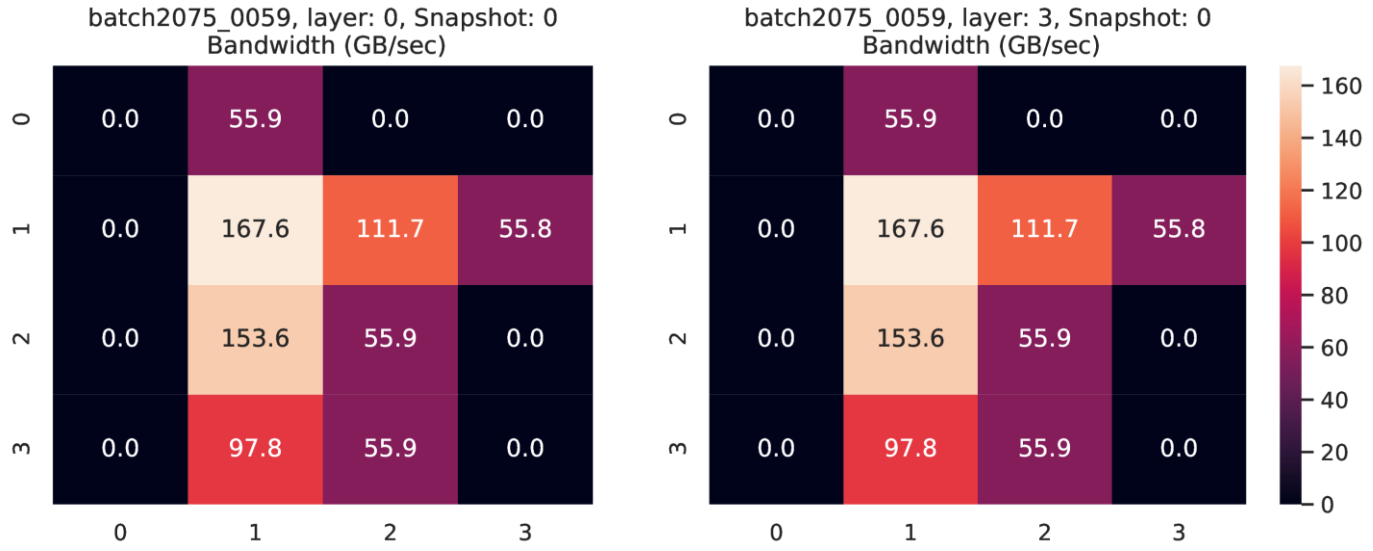
Multi-VNETs OFF

CPU Cores Requests via VNET (0)



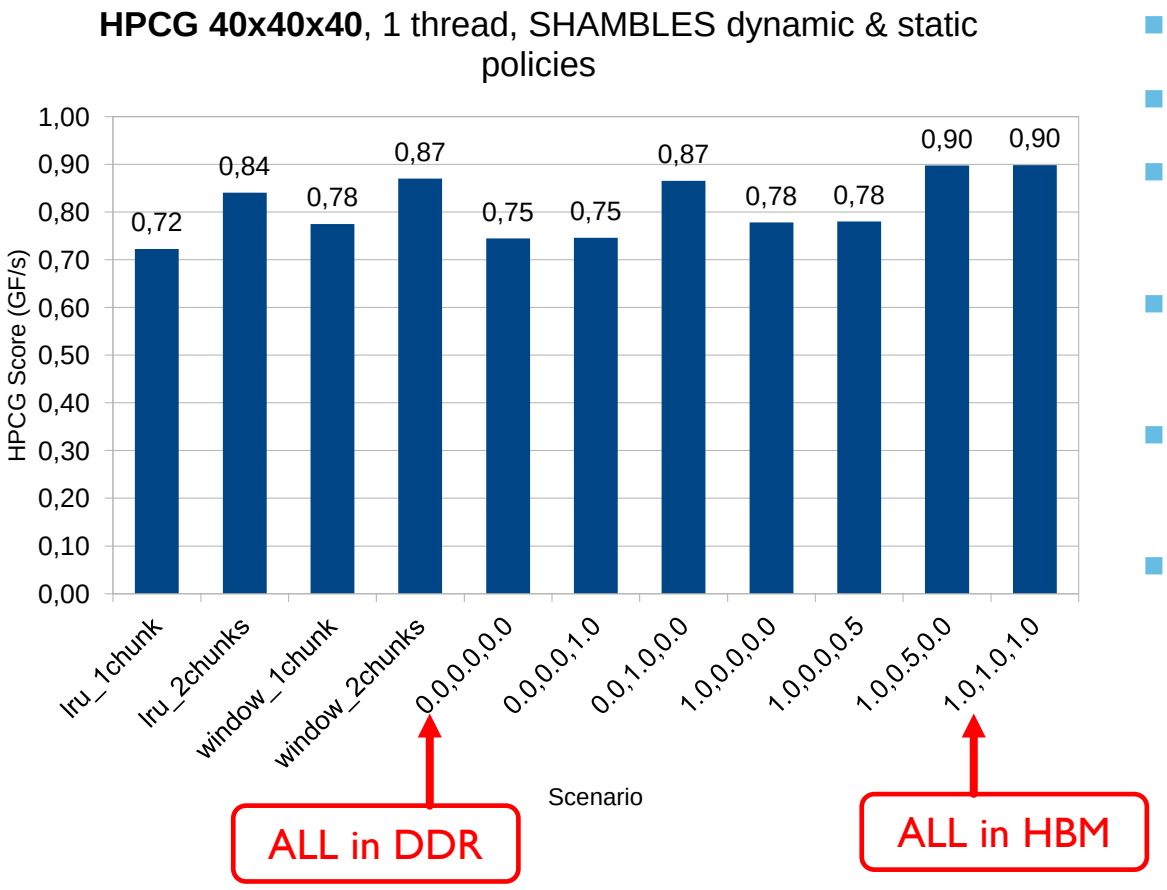
Multi-VNETs ON

CPU Cores Requests via VNETs (0 & 3)



- We examine the most congested VNET for T-2
- Aggregate Peak Bandwidth on Router with coordinates (1,1)
  - 260.3 GB/s vs 2*167.6 = 335.2 GB/s (split in 2 layers)
- STREAM TRIAD Reported BW: 81 GB/s ? 144 GB/s

DYNAMIC MIGRATION BET. MEMORY TIERS (HBM, DDR)



{0.0,1.0,0.0} → 100% of array B in HBM, while A,C are in DDR.

- Best case: 0.90 GF/s (All in HBM)
- Worst case: 0.75 GF/s (All in DDR)
- Both LRU and WINDOW policies are close to worst case when using only 1 chunk.
- Increasing # of chunks (2), results in performance improvement for both policies.
- **Window policy achieves 0.87 GF/s, while only 40% of the problem size is in HBM.**
- Reported memory bandwidth ranges from 4.45 – 5.5 GB/s

HPCG: https://github.com/hpcg-benchmark/hpcg
SHAMBLES: https://github.com/CARV-ICS-FORTH/shambles

EPI Forum

Paris, 6-7 October, 2025

European Processor Initiative

THANK YOU!

POLYDOROS PETRAKIS, VASSILIS PAPAEFSTATHIOU, MANOLIS MARAZAKIS (FORTH)

FORTH
FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

European Processor Initiative

From Concept to Computation: A Comprehensive Co-Design Approach for Future HPC Processors

Mohamed Benazouz, Denis Dutoit,
Ayoub Mouhagir, Mohamed Ouazzi
Anthony Philippe, Jean-Christophe Weil,
Lilia Zaourar

Contexte

Co-design

- Avoid costly late redesigns
- Ensure balanced & scalable system
- Accelerate best-fit exascale architecture

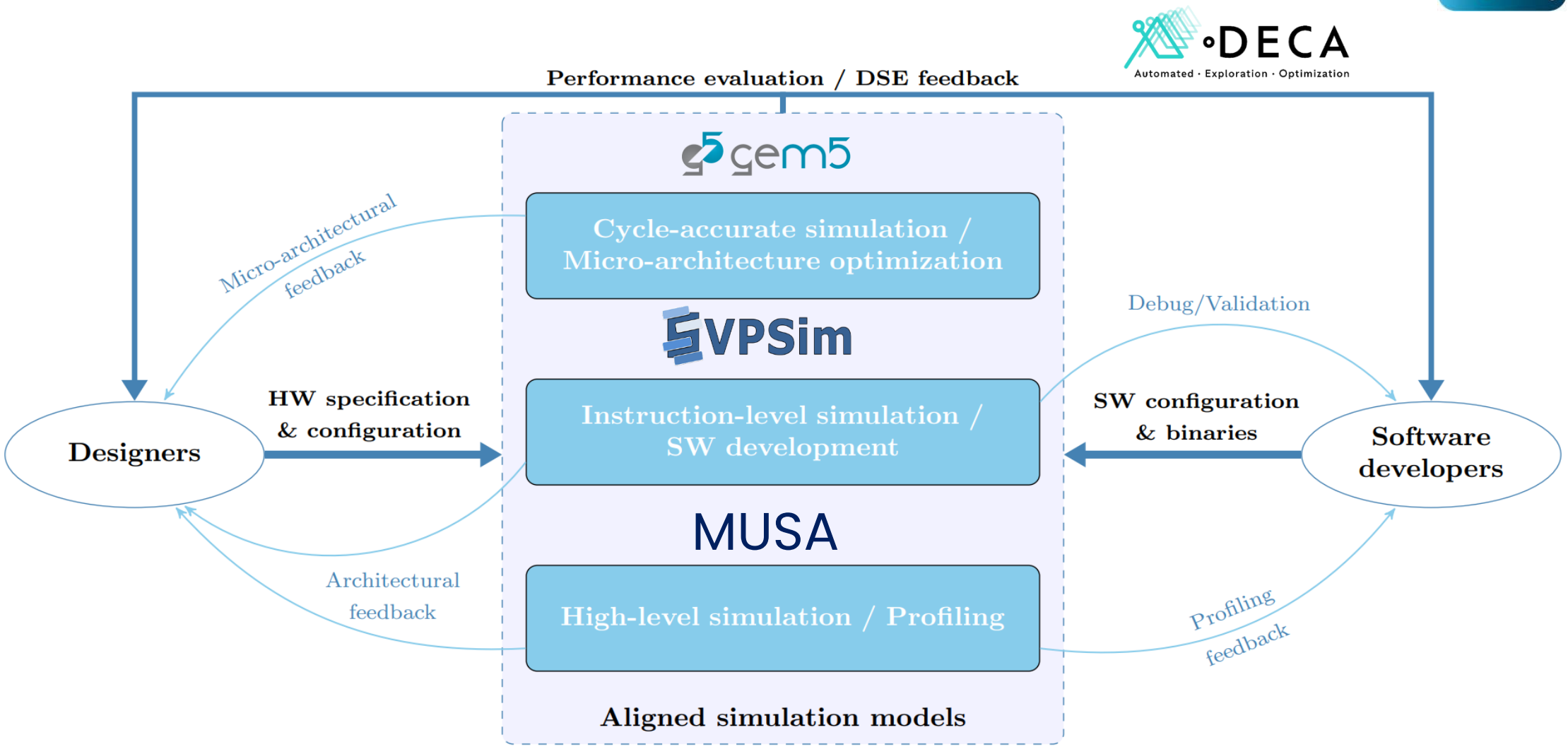
1- Modeling & KPI Evaluation

- Virtual prototypes of system components
- Evaluate KPIs: performance, power, scalability
- Detect bottlenecks before hardware freeze

2- Design space exploration

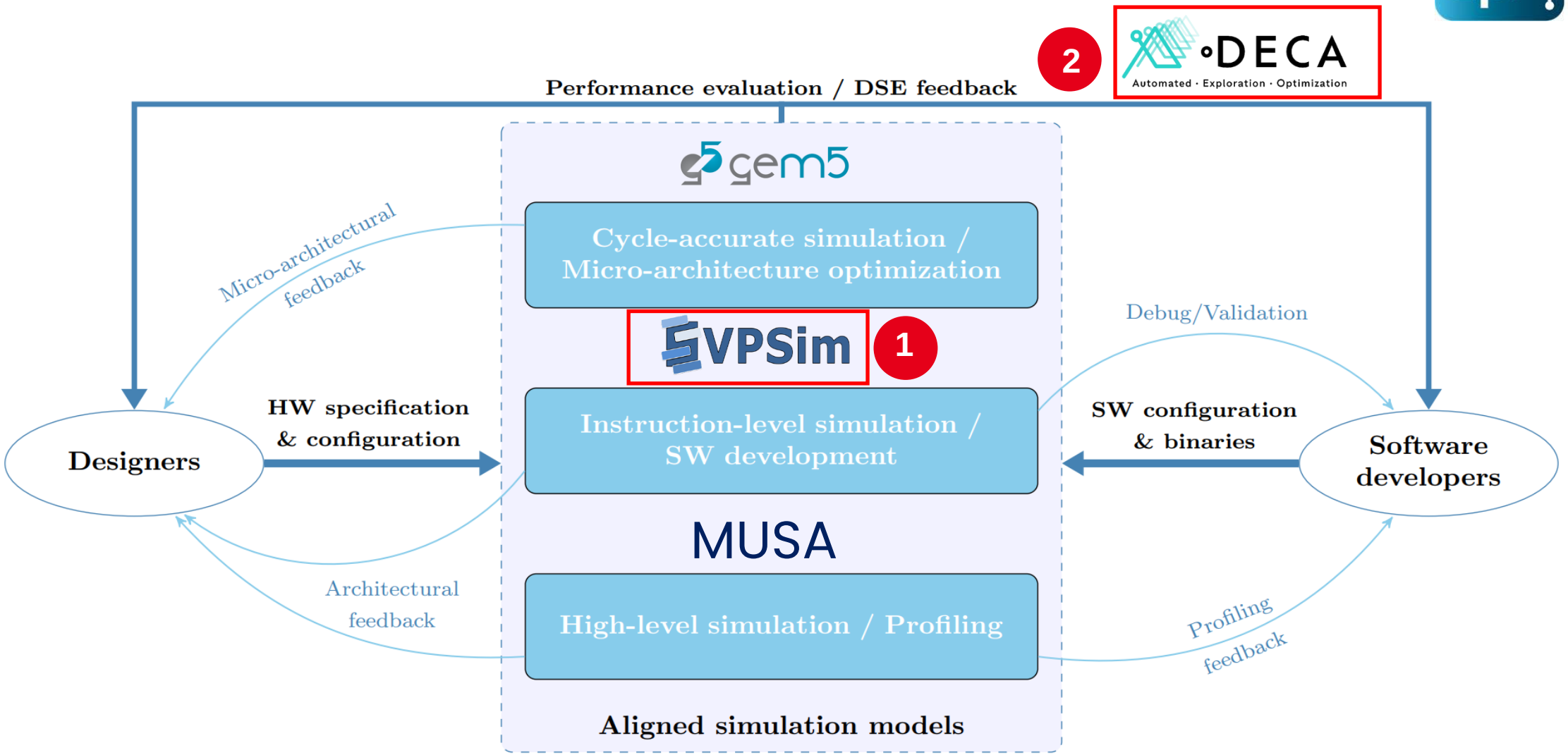
- Explore multiple HW-SW configurations
- Automate trade-off analysis
- Select optimal configurations for workloads

Full HW/SW co-design approach



✓ *L. Zaourar et al., Multilevel simulation-based co-design of next generation HPC microprocessors. 2021 International Workshop PMPBS (PMBS), Super Computiong 2021*

Full HW/SW co-design approach



✓ *L. Zaourar et al., Multilevel simulation-based co-design of next generation HPC microprocessors. 2021 International Workshop PMPBS (PMBS), Super Computing 2021*

Tools for HW/SW Co-design

VPSim

1 Efficient evaluation of Key Performance Indicators
→ **Modeling & Simulation**

- Virtual Prototype help designing complex SoCs
  - Fast design evaluation
  - Allow early performance analysis in the design flow
  - Provide a virtual HW platform to SW designers
  - Functional validation
- Virtual Prototype parallelize design phases and increase design productivity

➡ **Best trade off between accuracy and simulation time**

DECA
Automated · Exploration · Optimization

2 Efficient tuning of architecture parameters
→ **automatic exploration of the design space**

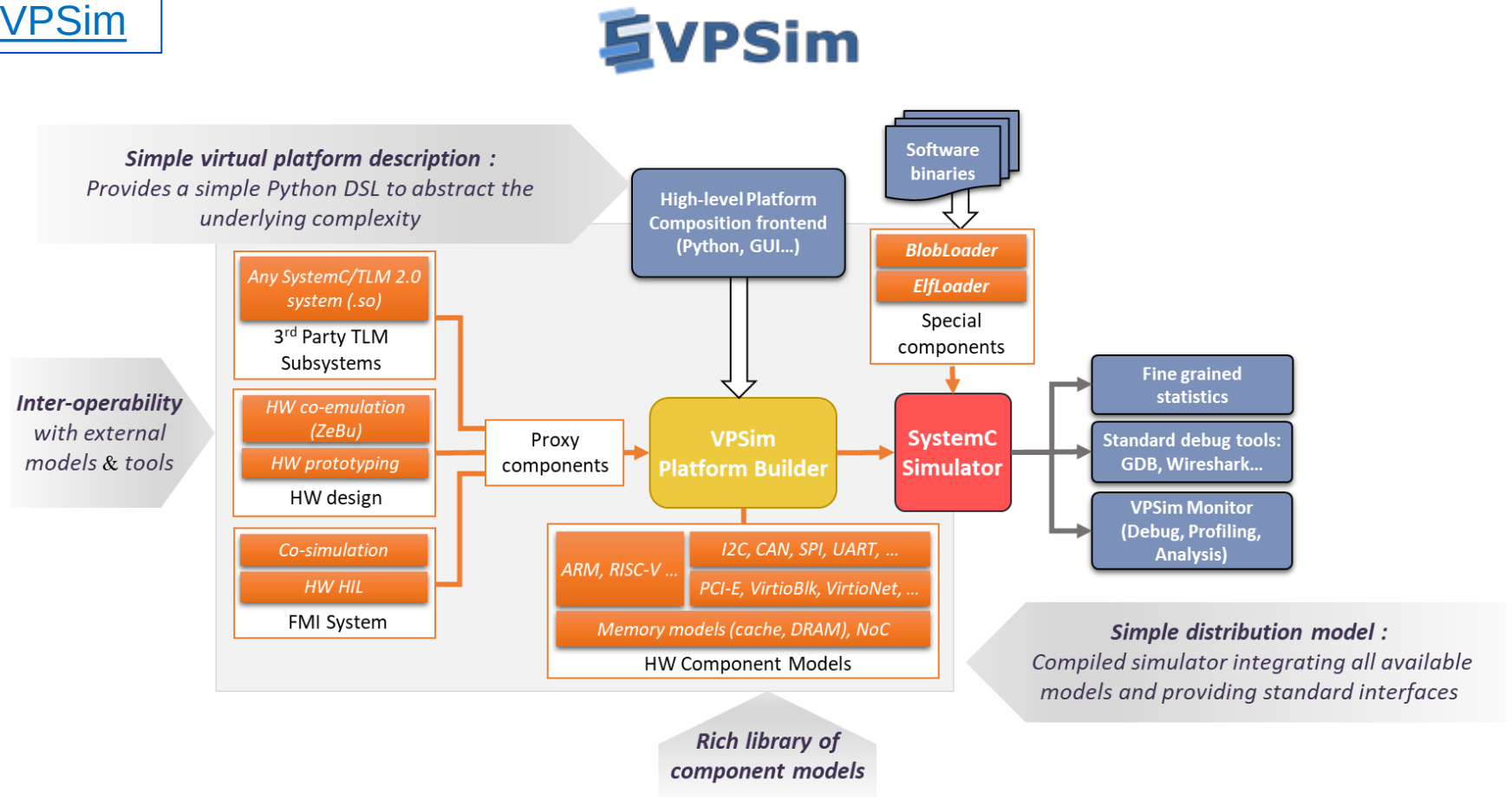
- High level exploration to find quickly best configurations
  - Automatic techniques to handle complexity
  - Leverage architects intuition and benchmarking
- Advanced algorithms for optimizing multi-objective exploration efficiently
  - Tackle various conflicting KPI : performance, power, area, safety, security, sustainability

➡ **Allow quick creation of configurations, evaluation exploration of the different combinations**

1- VPSim Virtual Prototyping overview

- EPI project :

https://github.com/CEA-LIST/VPSim



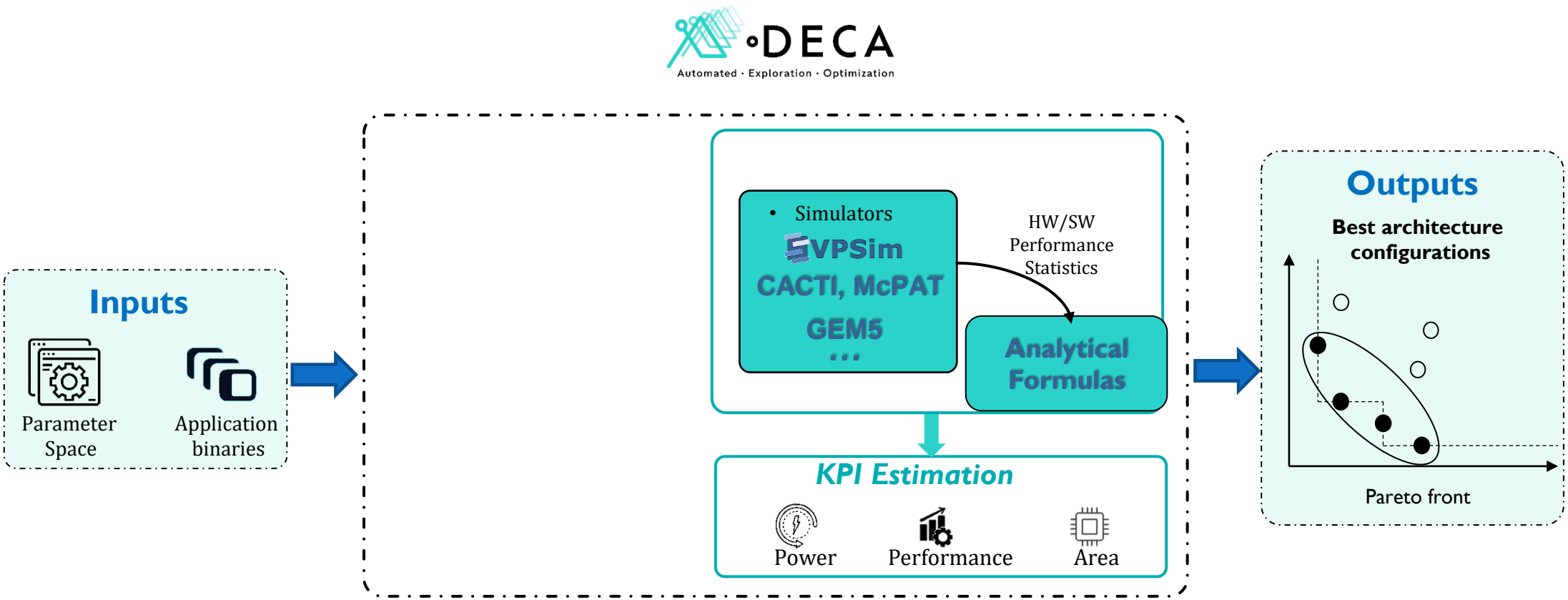
Performance counters (Non-Exhaustive List)

	Performance counter	Description
CPU	Executed instructions	Number of executed instructions per cpu
	Executed FP instructions	Number of executed Floating Point instructions per cpu
	Loads/stores	Number of loads and stores per cpu
Cache	Cache protocol transactions	Stats about the cache protocol transactions per cache
	Hits/Misses	Number of hits/misses per cache
	Writes/Reads	Number of reads and writes per cache
Network on-Chip	Total Packets	Number of packets that traversed the NoC
	Total distance	Total number of hops made by a packet across routers and channels from its source node to its destination node.
	Total latency	Total latency of packets through the NoC
	Average latency	Average network latency
	Packets per router	Number of packets that traversed each router
	Contention delay per router	Contention delay of packets at each router level
	Memory Reads Crossing Numa Nodes	Number of memory reads initiated by a CPU in one NUMA node that access memory located in another NUMA node.
	Memory Writes Crossing Numa Nodes	Number of memory writes initiated by a CPU in one NUMA node that access memory located in another NUMA node.
	Packets Crossing Numa Nodes	Number of NoC packets crossing one NUMA node to another NUMA node.

2- Automated Design Space Exploration

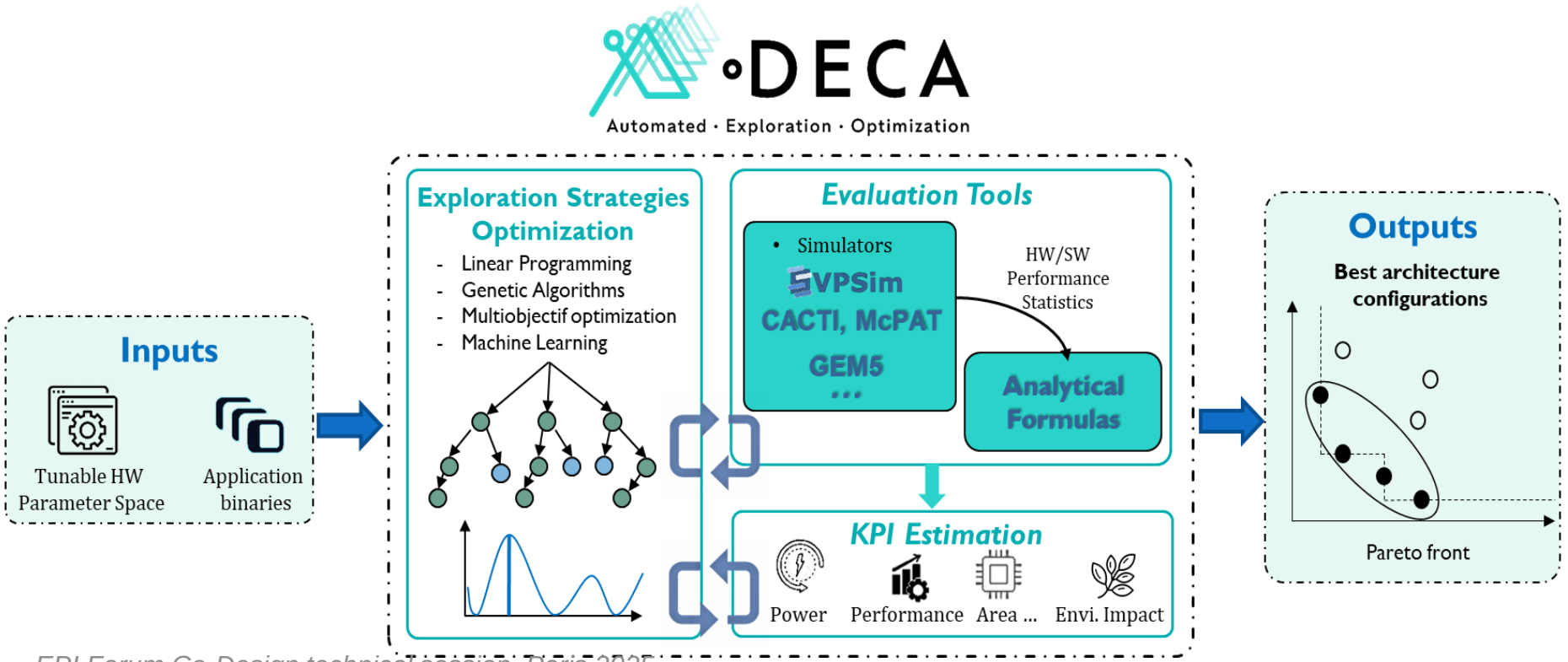
Optimized : PPA Performances, Power, Area

2- Automated Design Space Exploration



2- A-DECA : Design Space Exploration

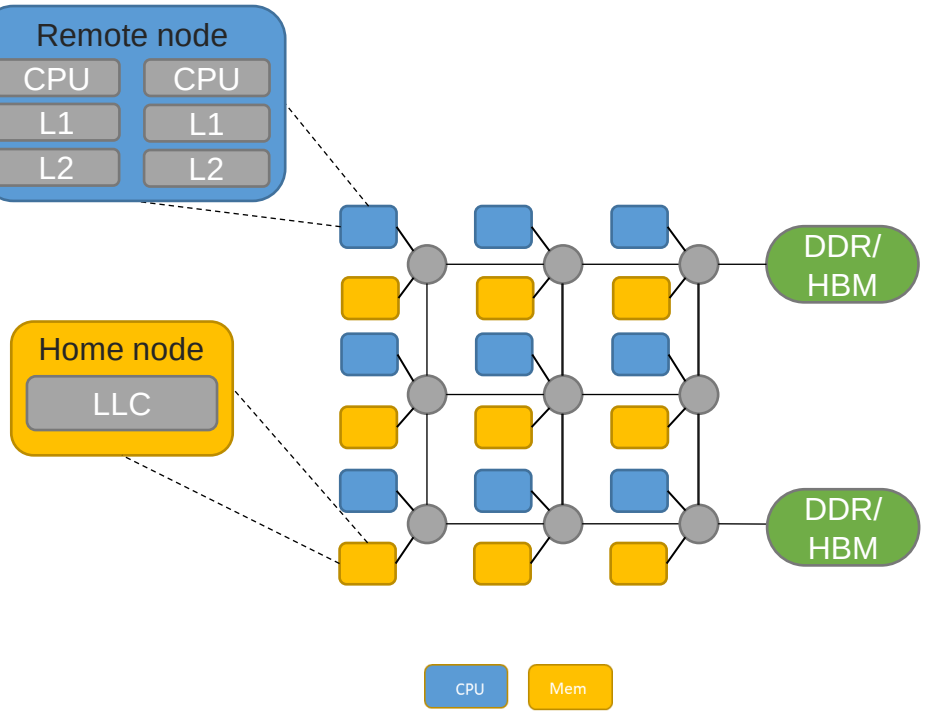
- Allows engineers to quickly create configurations, evaluate them and explore the different combinations
  - Combination of different simulation tools and analytical formulations to fully estimate PPA
  - Modular easy-to-use and updatable approach, various advanced optimization algorithms
  - Genericity : not limited to a restricted set of parameters due to the simulator.

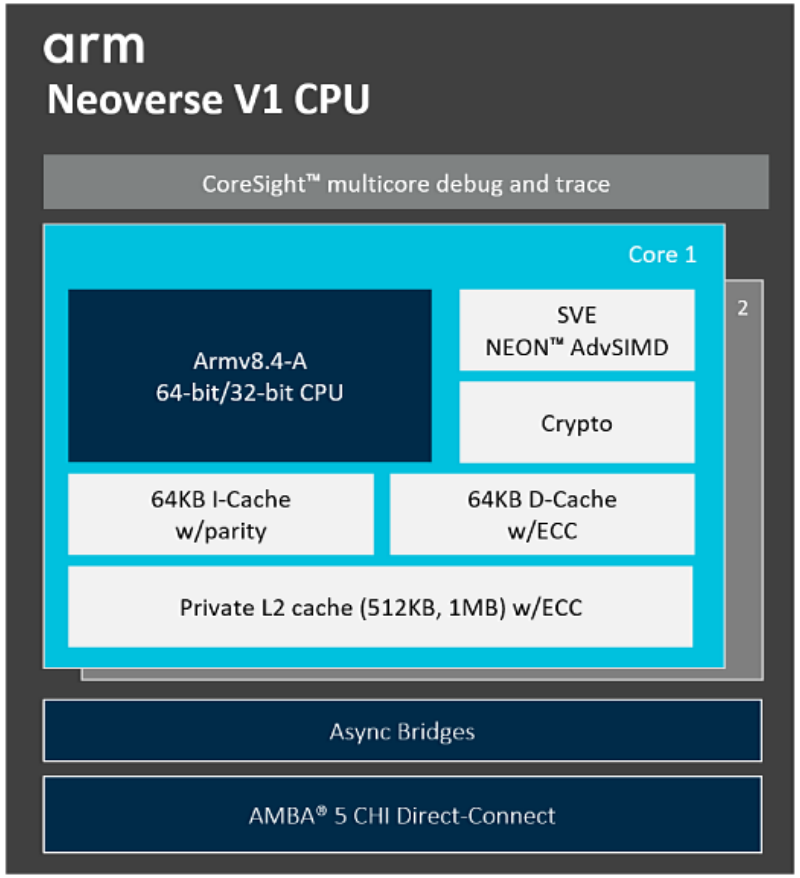


Optimize : PPA Performances, Power, Area

HPC System-level design

- **Rhea like architecture**
  - ARMv8.4 Neoverse V1 cores
  - Private L1/L2 caches, Shared Last Level Cache
  - External DDR, HBM
  - Mesh Network-on-Chip

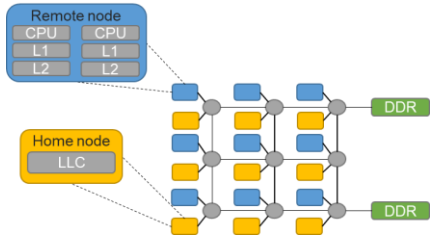




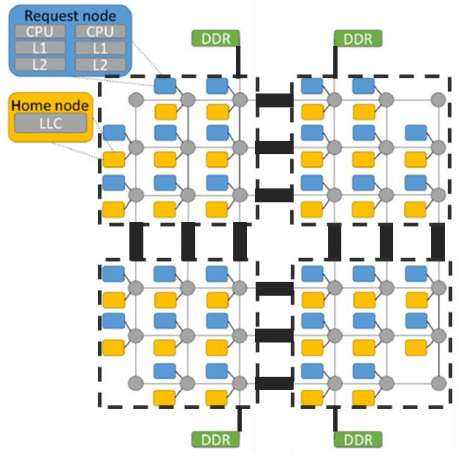
Co-design concerns

- EPI Addresses lot of design concerns :
  - Calibration and Validation Process

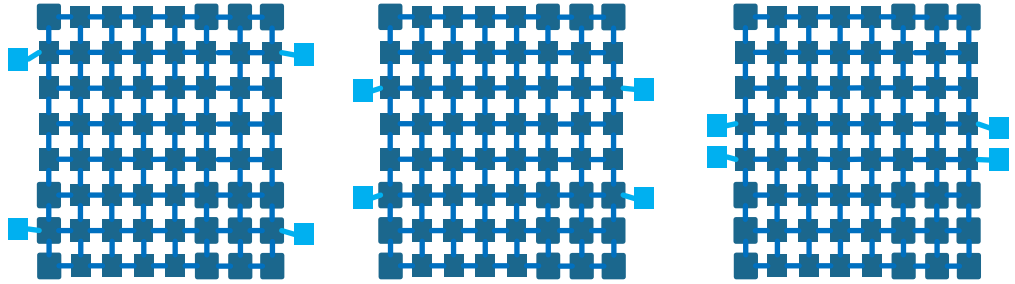
  - Core parameters & memory hierarchy



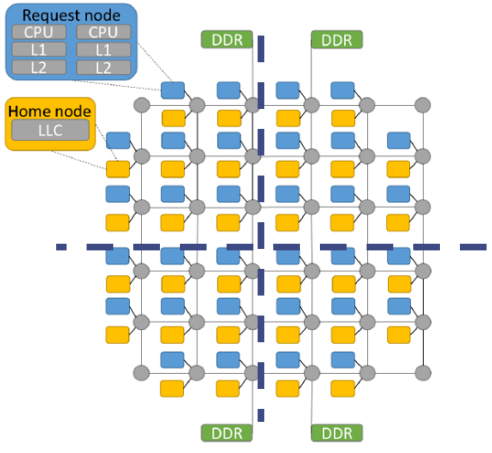
  - Non Uniform Memory Acess



  - Placement external memories HBM/DDR



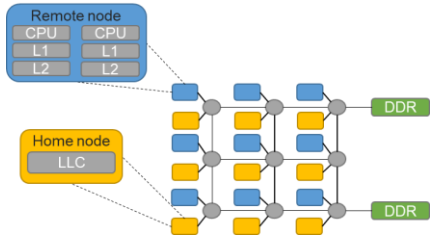
  - .....
  - From Physical to Logical Quadrants reconfiguration



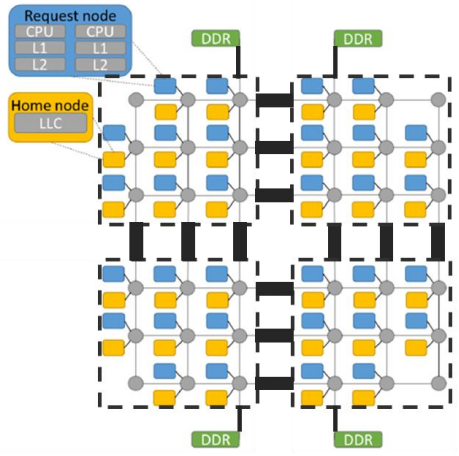
Co-design concerns

- EPI Addresses lot of design concerns :
  - Calibration and Validation Process

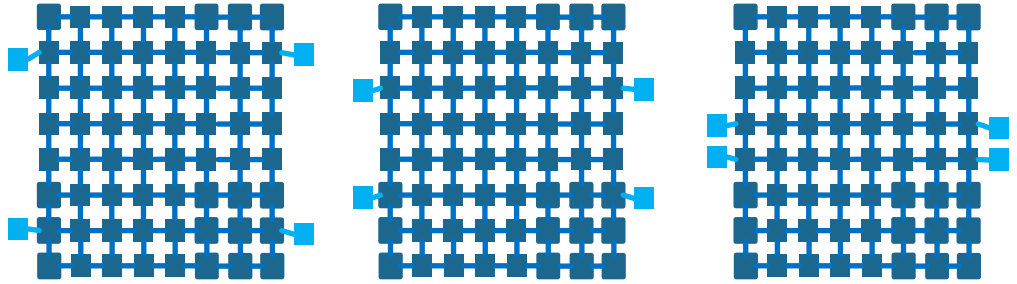
  - Core parameters & memory hierarchy



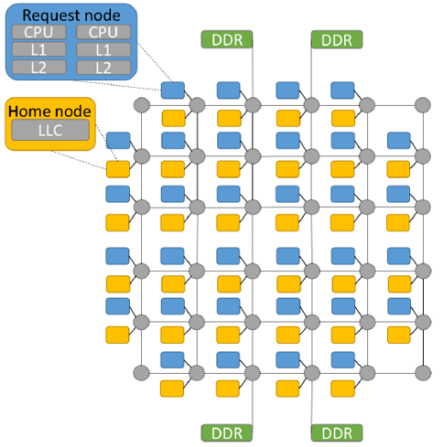
  - Non Uniform Memory Acess



  - Placement external memories HBM/DDR



  - .....
  - From Physical to Logical Quadrants reconfiguration

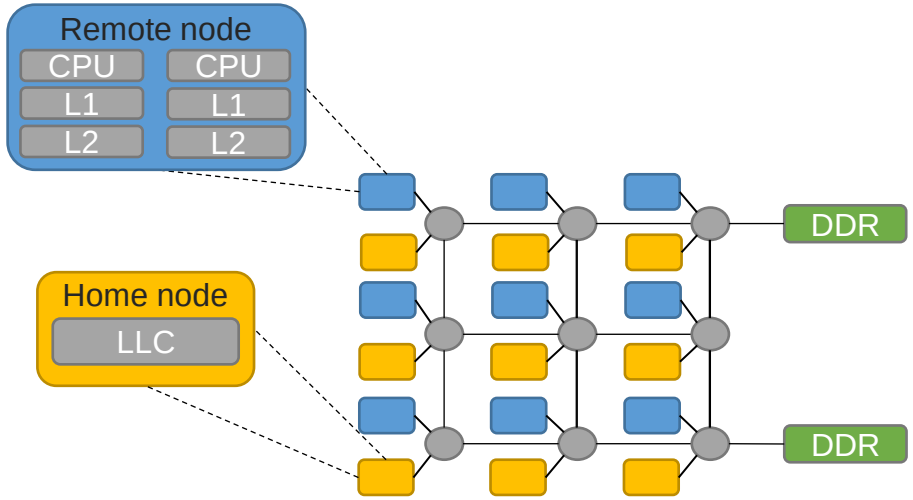


Use case 1 : @core level Rhea

- **Inputs**
  - **The explored design space**

Architectural parameters	Candidate values
Number of cores	1, 2, 4, 8, 16, 32, 64
Number of cores per cluster	1, 2, 4, 8
RAM size (GB)	2, 4, 8
L1 Data/Instruction cache size (KB)	8, 16, 32, 64
L2 cache size (KB)	128, 256, 512, 1024
LLC* size (KB)	512, 1024, 2048
L1/L2/LLC line size (B)	16, 32, 64, 128, 256
L1/L2/LLC associativity	1, 2, 4, 8
NoC X/Y dimension	1, 2, 4, 8, 16
Number of memory channels	1, 2, 4, 8, 16, 32

*Last Level Cache (L3)



  - **HPC benchmark applications**
    - STREAM: memory bandwidth
    - DGEMM: matrix computation
    - WalBerla: numerical simulation

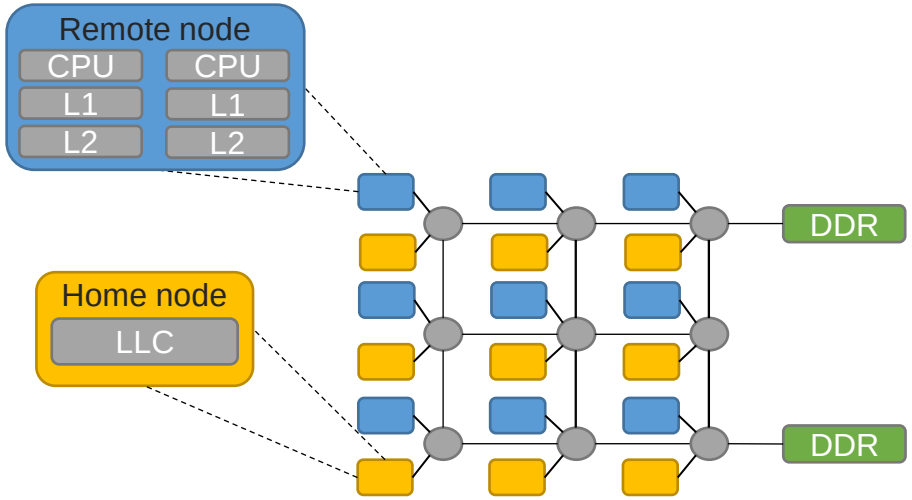
Use case 1 : @core level Rhea

- **Inputs**
  - **The explored design space**

Architectural parameters	Candidate values
Number of cores	1, 2, 4, 8, 16, 32, 64
Number of cores per cluster	1, 2, 4, 8
RAM size (GB)	2, 4, 8
L1 Data/Instruction cache size (KB)	8, 16, 32, 64
L2 cache size (KB)	128, 256, 512, 1024
LLC* size (KB)	512, 1024, 2048
L1/L2/LLC line size (B)	16, 32, 64, 128, 256
L1/L2/LLC associativity	1, 2, 4, 8
NoC X/Y dimension	1, 2, 4, 8, 16
Number of memory channels	1, 2, 4, 8, 16, 32

*Last Level Cache (L3)

$1\ simulation = \sim 3\ min$
$2\ x\ 10^6$ **configurations**
Total = **7 years**



  - **HPC benchmark applications**
    - STREAM: memory bandwidth test
    - DGEMM: matrix computation
    - WalBerla: numerical simulation

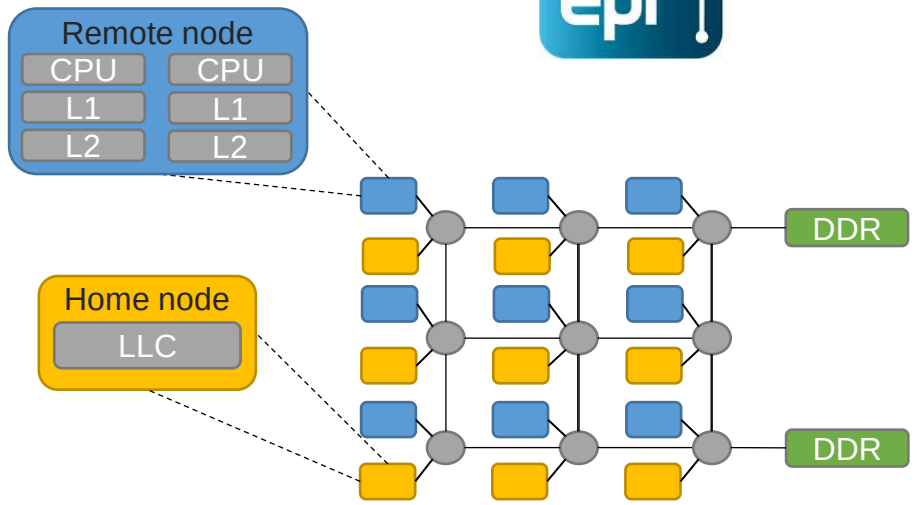
Use case 1 : @core level Rhea

- **Inputs**
  - **The explored design space**

Architectural parameters	Candidate values
Number of cores	1, 2, 4, 8, 16, 32, 64
Number of cores per cluster	1, 2, 4, 8
RAM size (GB)	2, 4, 8
L1 Data/Instruction cache size (KB)	8, 16, 32, 64
L2 cache size (KB)	128, 256, 512, 1024
LLC* size (KB)	512, 1024, 2048
L1/L2/LLC line size (B)	16, 32, 64, 128, 256
L1/L2/LLC associativity	1, 2, 4, 8
NoC X/Y dimension	1, 2, 4, 8, 16
Number of memory channels	1, 2, 4, 8, 16, 32

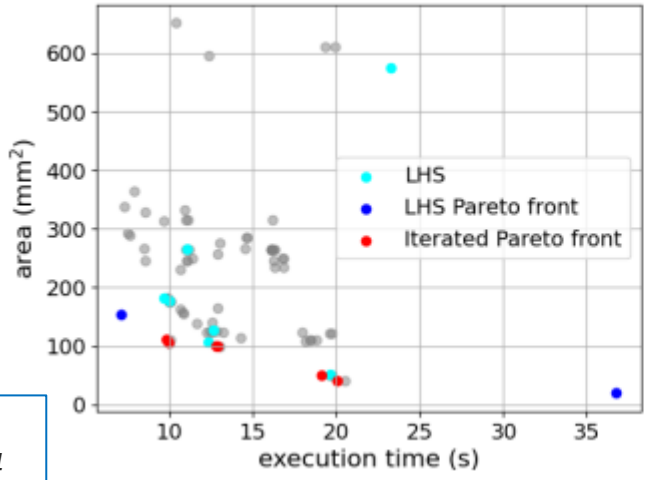
*Last Level Cache (L3)

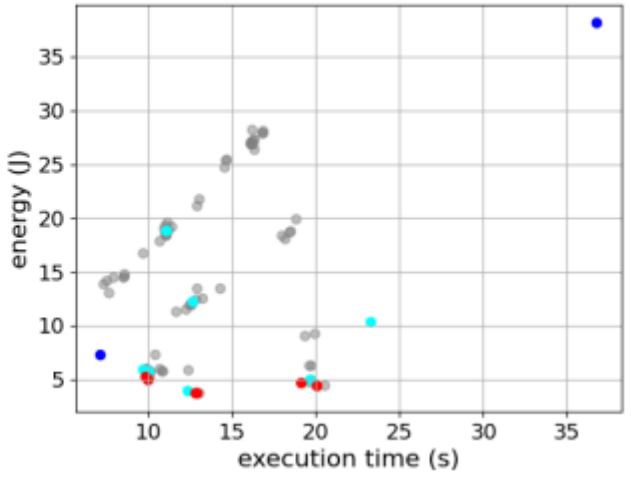
- **Co-design**
  - VPSim Simulation & McPat
  - A-DECA Exploration
    Bayesian optimization



- **Outputs**
  - Multi-objectif optimisation :
    Execution Time, Energy, Area
  - 800 exploration iterations : **3h** (simulation & exploration)

- **HPC benchmark applications**
  - STREAM: memory bandwidth test
  - DGEMM: matrix computation
  - WalBerla: numerical simulation

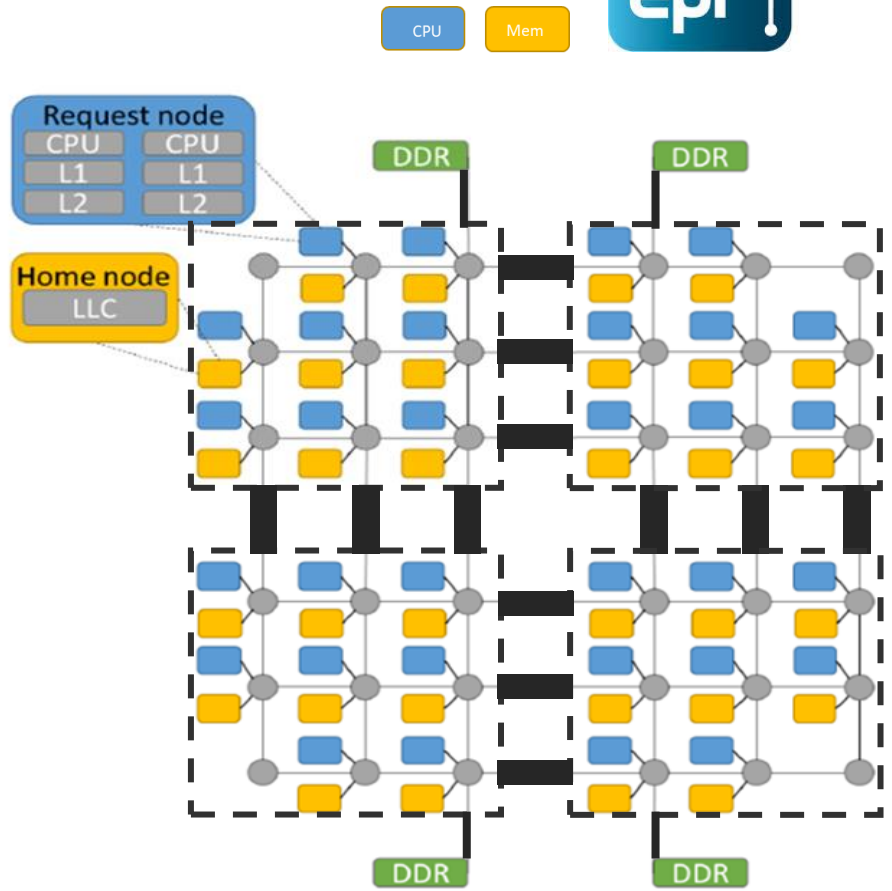




✓ *V. Fu et al., "Design space exploration of HPC systems with Random Forest-based Bayesian Optimization" Best paper Rapio/HIPEAC 2024*

Use case 2 : scaling up with NUMA

- NUMA (Non-Uniform Memory Access)
  - Multiple processors to access shared memory
  - Differences in distances (latency & bandwidth) between processors and memory regions
  - Reduces communication overhead, and boosts performance
- Design concerns
  - Memory placement
  - CPU node partitioning
  - Interconnect latency & bandwidth
  - Technology : various interfaces from various IP providers
- Application & Software
  - Adapt application code
  - System software aware

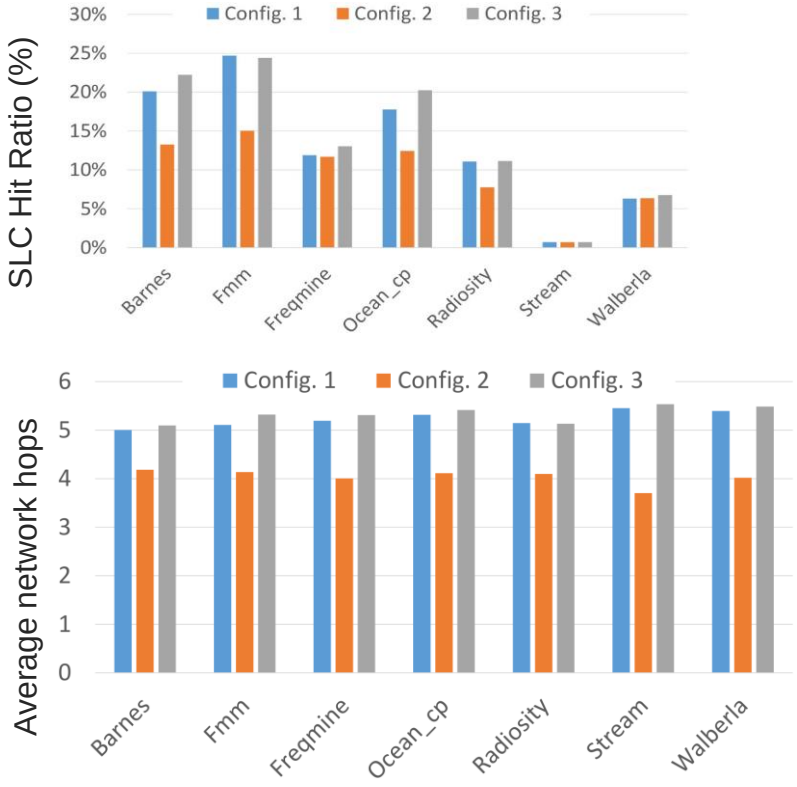


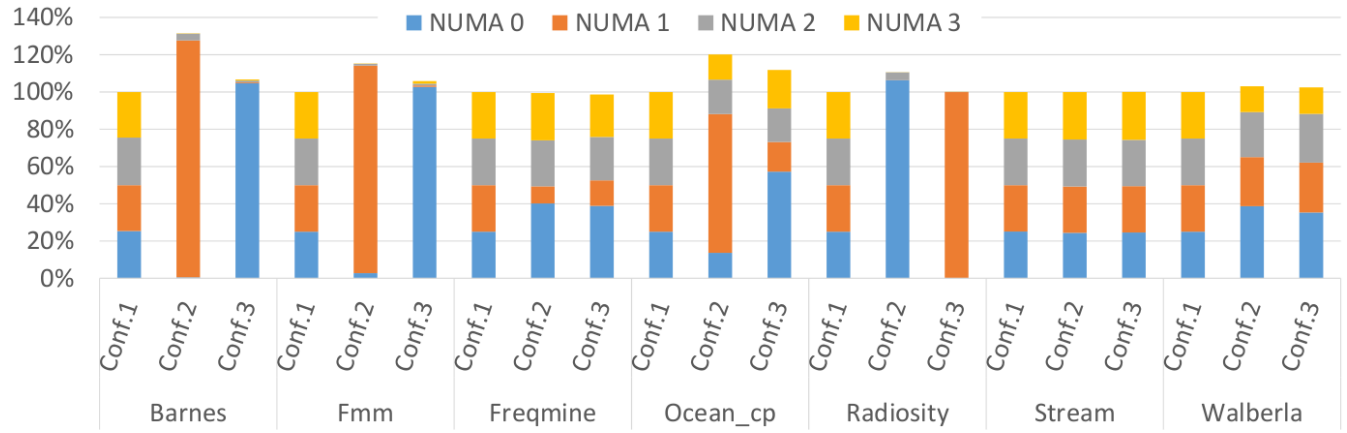
Effects of NUMA access on different workload performances is crucial to analyze and quantify design trade-off

Use case 2 : scaling up with NUMA

- What is the impact of different SLC address range assignments?
- How does application load balancing affect performance?
  - Consider the impact of SLC address range assignment on a variety of HPC applications

#	NUMA Nodes	SLC address range assignment
1	Single node	Interleaved over all memory controllers
2	4 nodes	Interleaved to local memory controller (SLC Cache Group)
3	4 nodes	Interleaved over all memory controllers





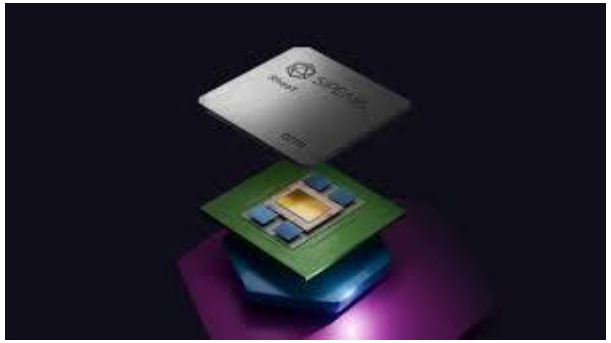
Memory Reads/Writes per NUMA node normalized to Configuration 1

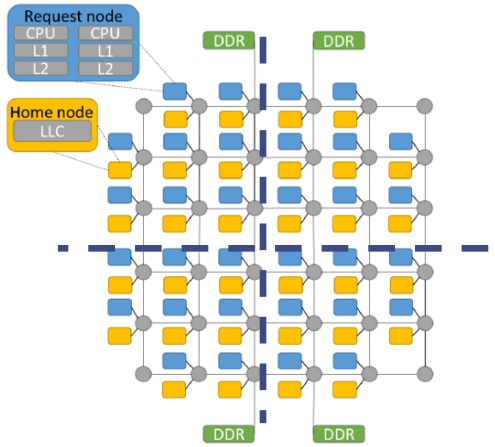
- SLC Cache Group functionality consistently reduces average network packet distance
- Good load balancing is critical for applications to benefit from NUMA functionality

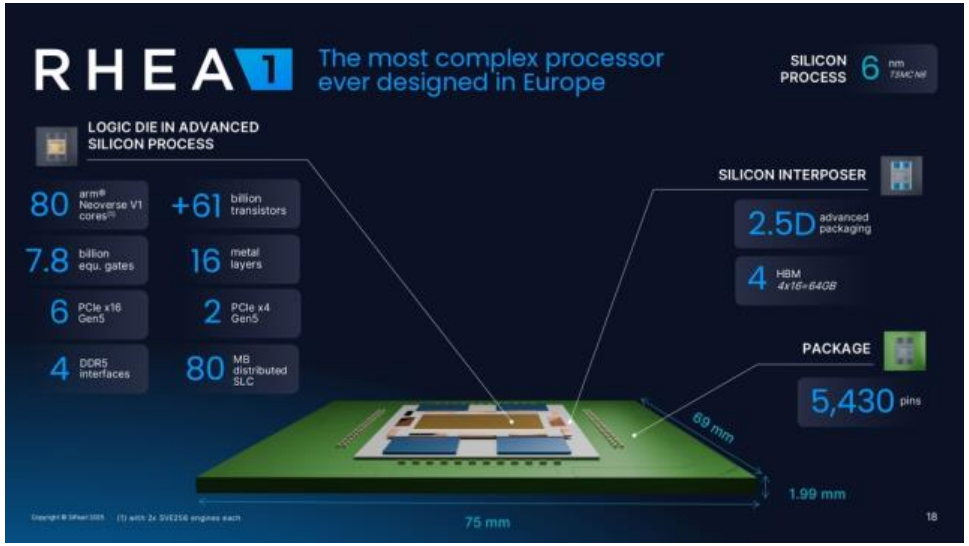
✓ *L. Zaourar et al., "Case Studies on the Impact and Challenges of Heterogeneous NUMA Architectures for HPC. ARCS 2024*

In summary

- EPI : Full co-design approach (VPSim & A-DECA)
- Before HW freeze
  - VPSim: Calibration and Validation Process
  - Core parameters (processing, cluster, caches)
  - Placement external memories HBM/DDR
  - Non-Uniform Memory Access
- After HW freeze
  - Post-tape out optimization







Conclusion

The Village

- Simulation & virtual prototyping

@Core Level

The continent

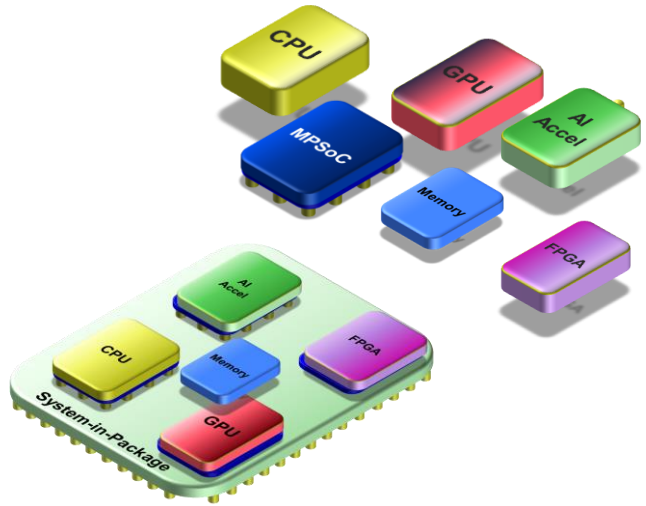
- Advanced architectural models
- Scaling exploration through automation

@System Level

The world

- Tools evolving with AI
- Beyond : CPU, GPUs, Memory, Photonics,
- Chiplets

@ Global ecosystem



EuroHPC Joint Undertaking

Co-design is no longer a luxury — it's the glue that holds together villages, continents, and soon the world of HPC innovation

Conclusion

The Village

- Simulation & virtual prototyping

@Core Level

The continent

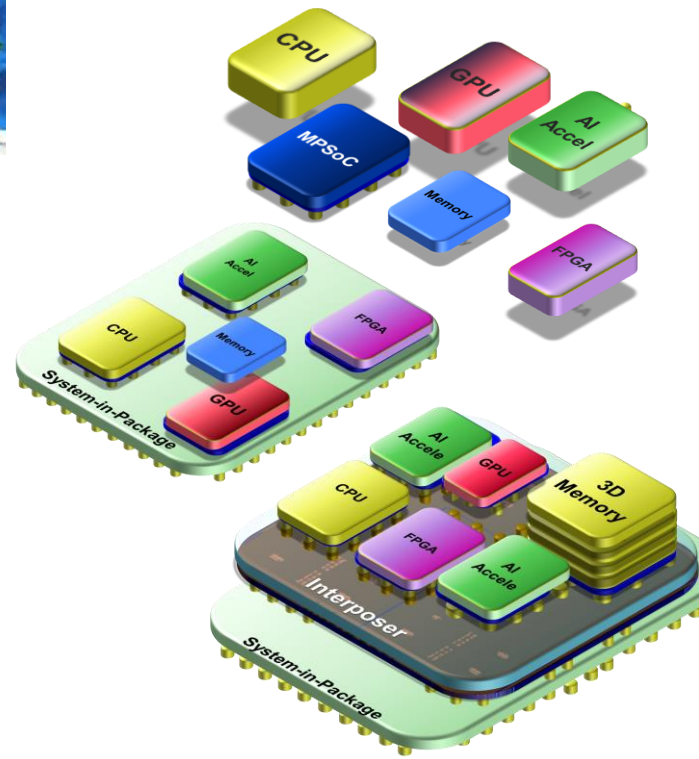
- Advanced architectural models
- Scaling exploration through automation

@System Level

The world

- Tools evolving with AI
- Beyond : CPU, GPUs, Memory, Photonics,
- Chiplets

@ Global ecosystem



Co-design is no longer a luxury — it's the glue that holds together villages, continents, and soon the world of HPC innovation