Thermally-Aware Multi-Chiplet Systems & Emerging Paths for Post-Transformer LLM Acceleration

Alish Kanani

Preliminary Examination
16 Oct2025

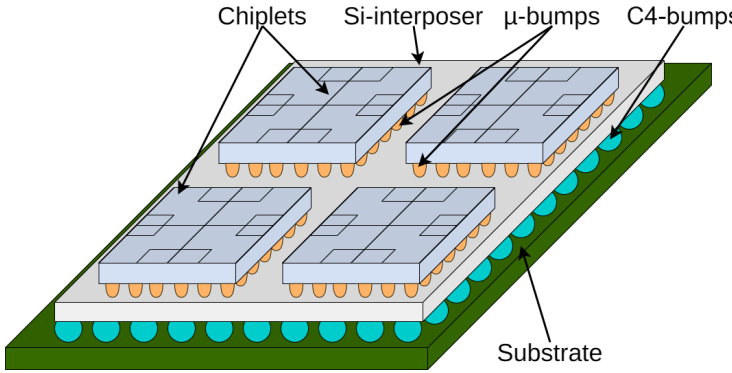
Committee Members: Akhilesh Jaiswal, Jaehyun Park, Joshua San Miguel, Janardhan Rao Doppa & Umit Y. Ogras (Ph.D. Advisor)

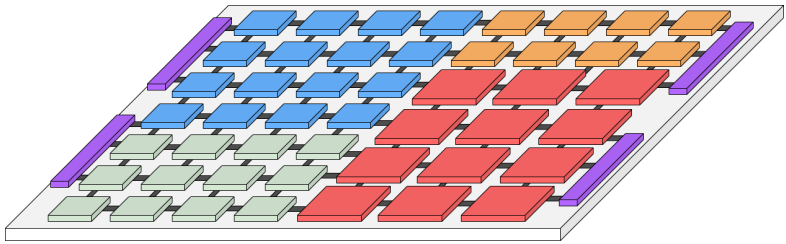
e lab

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Outline

- **Motivation: Chiplet-based platforms**
- **Preliminary Work-1:**
  - MFIT : Multi-Fidelity Thermal Modeling for 2.5D and 3D Multi-Chiplet Architectures
- **Preliminary Work-2:**
  - THERMOS: Thermally-Aware Multi-Objective Scheduling for Heterogeneous Multi-Chiplet PIM Architectures
- **Ongoing and Proposed Work:**
  - Disaggregated Acceleration of Hybrid Mamba–Transformer LLMs via Systolic Prefill and Vector Decode
  - Breaking the Memory Wall in MoE LLMs with Expert Prefetching
- **Timeline**
- **Conclusions**

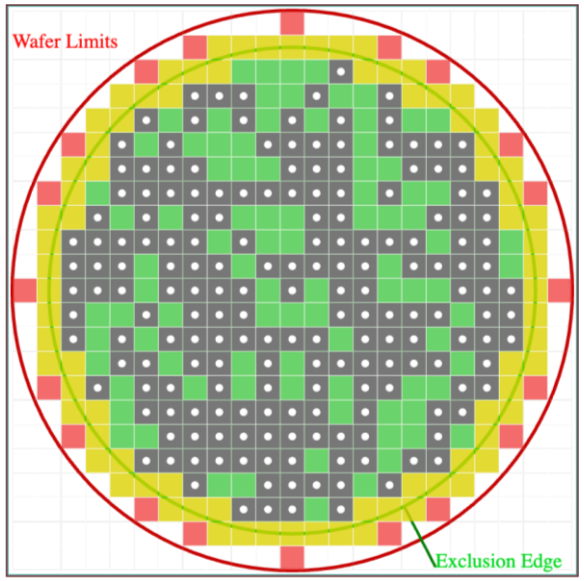


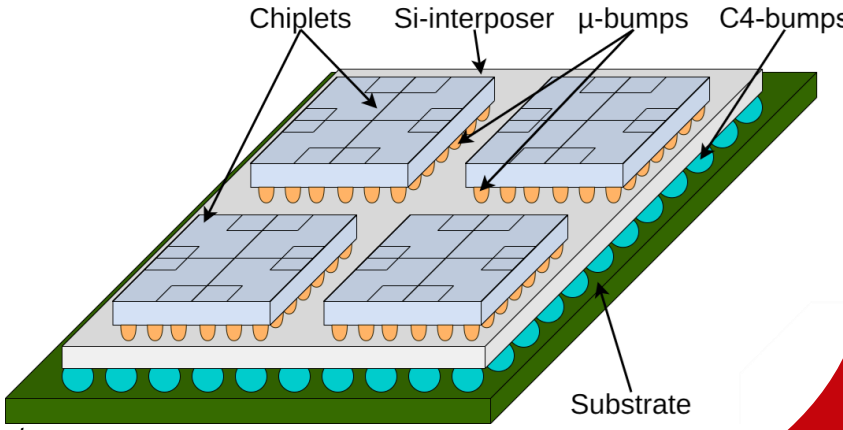


Heterogeneous chiplet system

How did Chiplets Become Crucial?

- **Monolithic 2D fabrication technologies reach physical limits**
  - Manufacturing of a large die is not economical anymore
- **However, the need for more compute power keeps growing**
  - Ever-increasing processing and memory requirements of AI workloads
- **A promising avenue: heterogeneous integration and chiplet-based manycore systems**
  - Integrating many small dies on interposer to sustain performance growth
  - Higher yield and lower fabrication cost





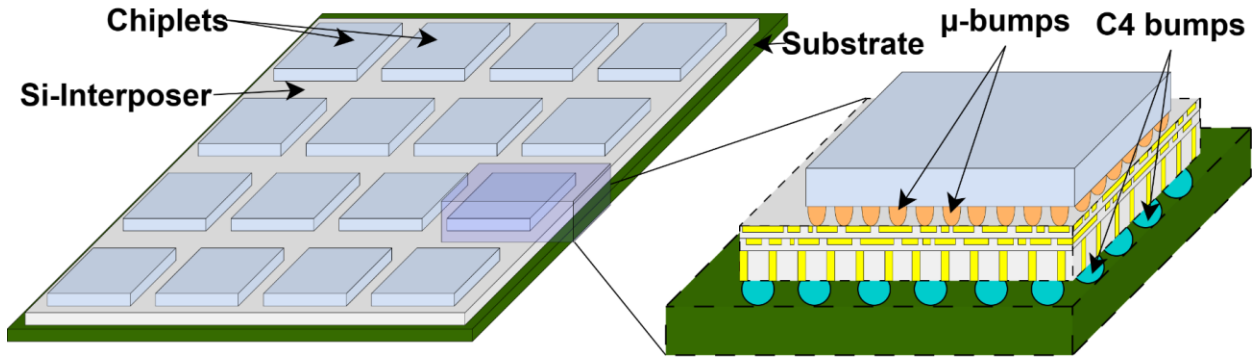
J. A. Cunningham et al., "ITRS2.0:Toward a reframing of the semiconductor technology roadmap.," in *IEEE 32nd Intl. Conf. Computer Design*, Oct 2014

Chiplet-Based Platforms

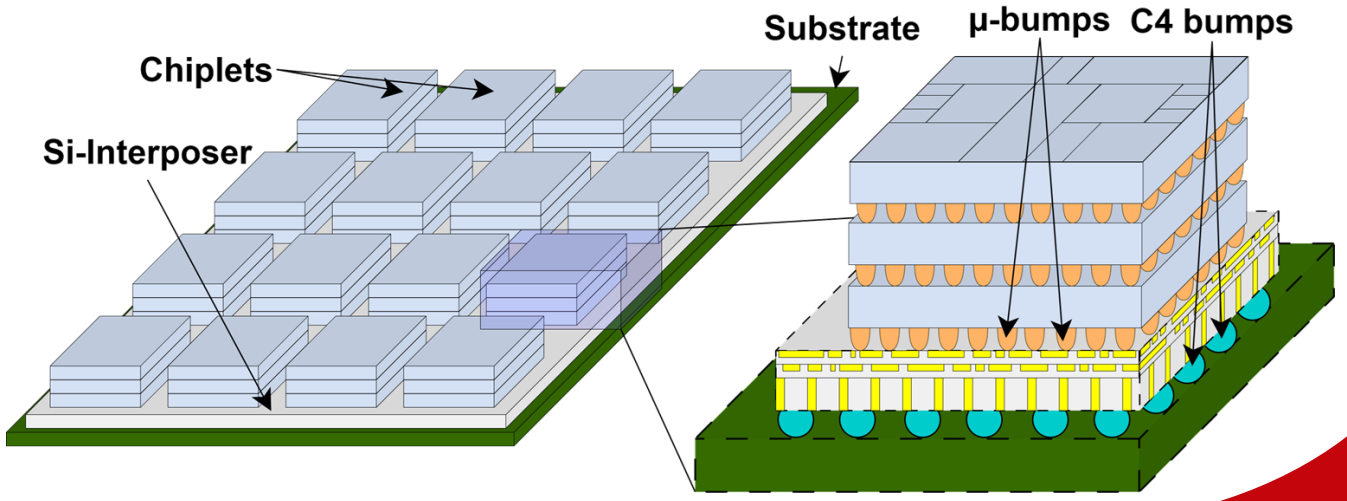
- Small dies fabricated on a single wafer
- Connected through Network-on-Interposer (NoI)
- Compact scale-out implementations
- Economical manufacturing

- Unsolved challenges we address:
  - *Thermal issues*
  - *Workload scheduling*

2.5D Geometry

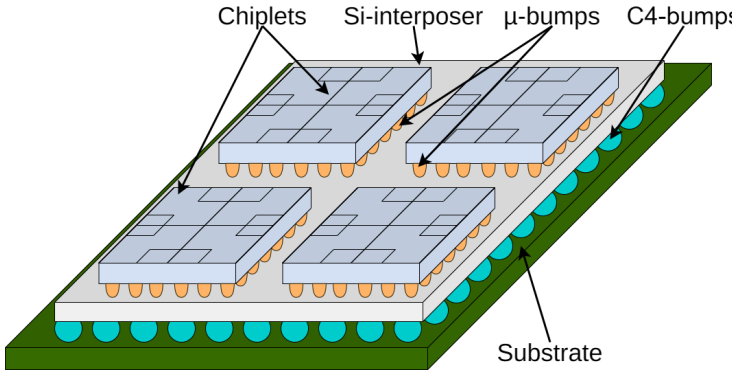


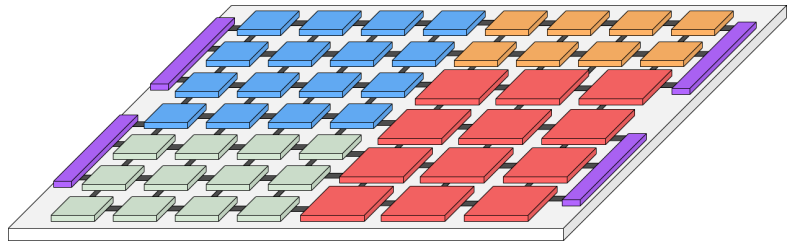
3D Geometry



Outline

- **Motivation: Chiplet-based platforms**
- **Preliminary Work-1:**
  - MFIT : Multi-Fidelity Thermal Modeling for 2.5D and 3D Multi-Chiplet Architectures
- **Preliminary Work-2:**
  - THERMOS: Thermally-Aware Multi-Objective Scheduling for Heterogeneous Multi-Chiplet PIM Architectures
- **Ongoing and Proposed Work:**
  - Disaggregated Acceleration of Hybrid Mamba–Transformer LLMs via Systolic Prefill and Vector Decode
  - Breaking the Memory Wall in MoE LLMs with Expert Prefetching
- **Timeline**
- **Conclusions**





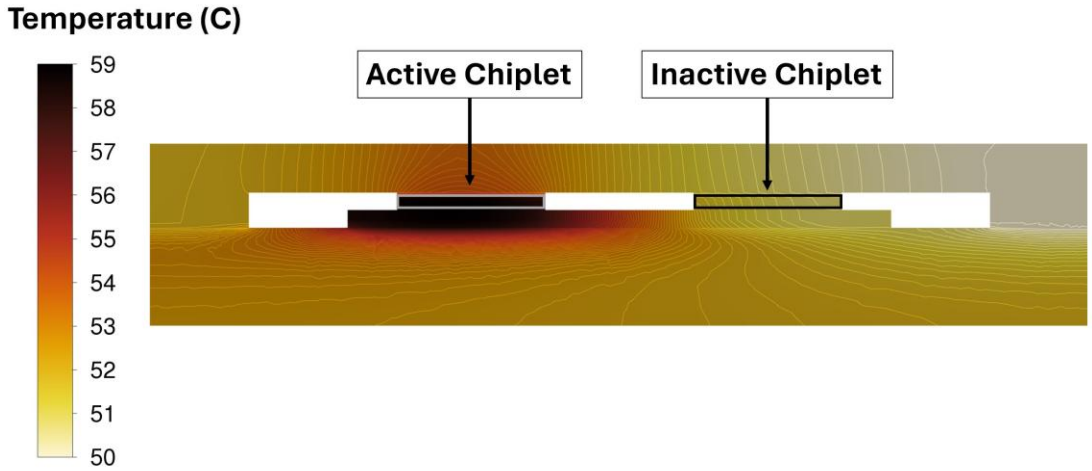
Heterogeneous chiplet system

Pfromm, Lukas, Alish Kanani, et al. "MFIT: Multi-fidelity thermal modeling for 2.5 D and 3D multi-chiplet architectures." *ACM Transactions on Design Automation of Electronic Systems* (2025).

Thermal tool in GitHub: https://github.com/AlishKanani/MFIT

Unique Thermal Paths in 2.5D Platforms

- **Thermal bottlenecks have long been a significant barrier to performance**
- **Chiplet-based systems aggravate this barrier due to their dense integration and add unique challenges**
  - **Monolithic chip:** Heat is spread directly across the die
  - **Chiplet-based platforms:** Heat also flows through the interposer and heat spreader
- **Dense large-scale integration and new thermal conductance paths**
  - Lead to thermal hotpot even at moderate power consumption

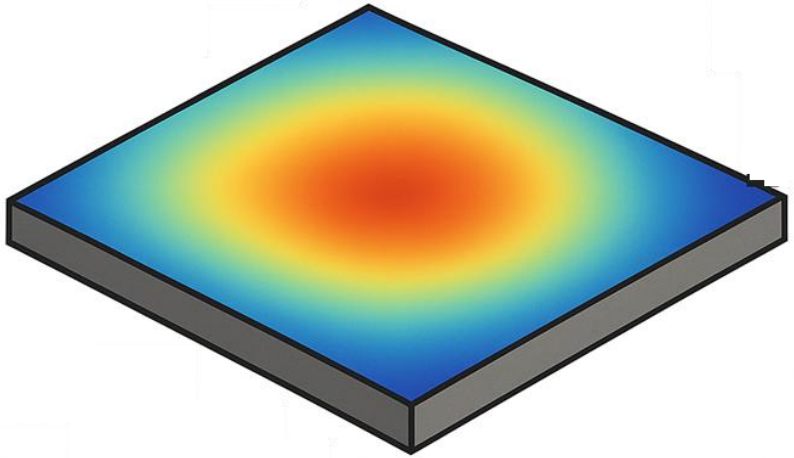
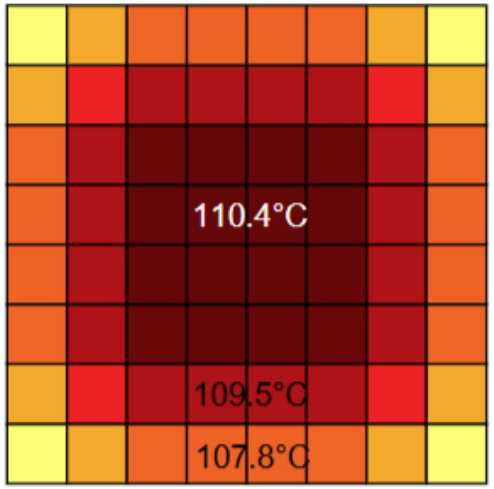
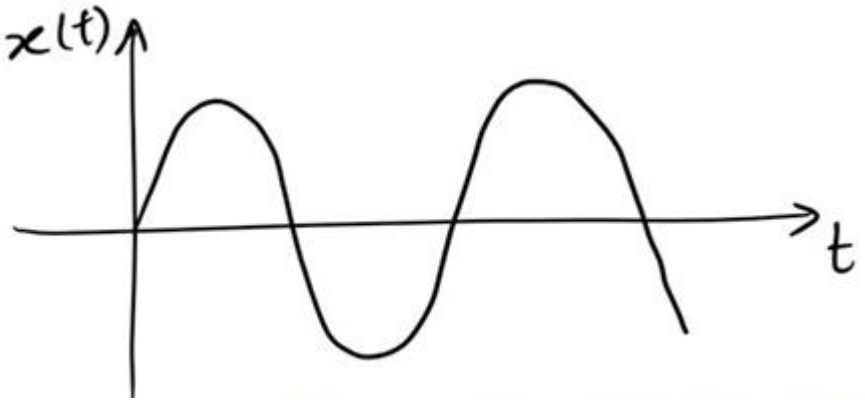
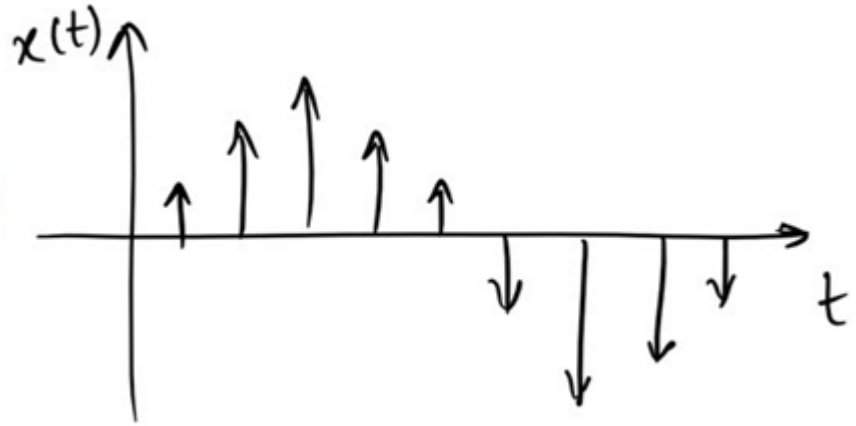


Need new approaches for this unique structure!

Overview of SOTA

- **The most accurate and direct thermal evaluation: hardware measurements**
  - Thermal imaging or temperature sensors
  - But requires hardware availability (infeasible for future systems)
- **This limitation motivated FEM-based modeling for temperature and analyze the heat flow**
  - Tools, such as ANSYS Fluent and COMSOL, commonly used for FEM simulations
  - Due to computational cost, they are suitable only for small designs and validation
- **Computational overhead and impractical execution time of FEM solvers motivate analytical models that enable rapid thermal evaluation in early design phases**
  - Most common method is constructing thermal RC networks
  - Recently proposed PACT framework employs a similar methodology by utilizing SPICE tools as solvers, focusing on standard-cell-level thermal analysis for 2.5D systems
- **However, these tools are not fast enough for large-scale, thermal-aware DSE and dynamic resource management of multi-chiplet systems**

Space & Time : Continuous vs Discrete

	Continuous	Discrete
Space		
Time		

Multi-Fidelity Thermal Modeling

Increasing speed →

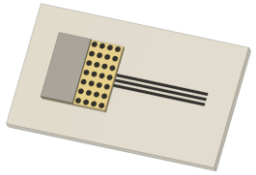
← Increasing accuracy

	Finite Element Method (FEM)		Analytical Models	
	1. Fine-grained	**2. Abstracted**	**3. Thermal RC Model**	**4. Discrete State-Space (DSS)**
Features	Accurate (e.g., real µbumps, links)	Replace micro-structures with equivalent material blocks	Independent of specific geometry, continuous time	Derived a specific architecture, discrete time
Error	Golden reference	< 0.5 °C	< 1.7 °C	Same as thermal RC*
Runtime	Not possible to model entire package	Days	Seconds	Milliseconds
Use case	Validate the abstracted FEM models	Ground truth to tune C values in Thermal RC model	Thermally-aware design space exploration, reference for DSS model	Large-scale optimization, thermal management

$$\boldsymbol{T[k+1] = A \times T[k] + B \times P[k]}$$

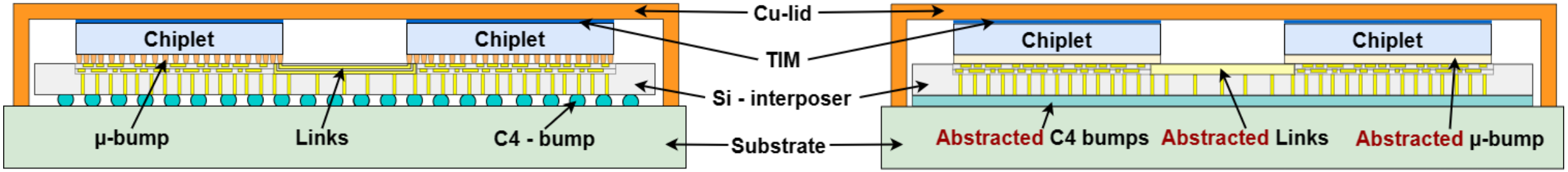
*Loses the generality.
Since the model is derived for a specific configuration, it needs to be regenerated if the architecture changes.

A Novel Multi-Fidelity Thermal Modeling Approach



1. Develop fine-grained FEM Models

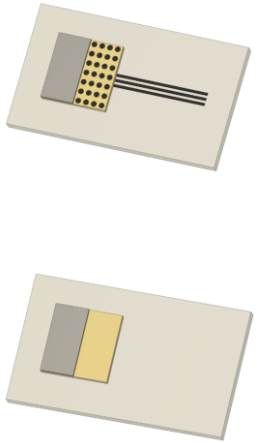
- **Model all the architecture, geometry, and parameters in as much detail possible**
- **Examples:**
  - All links
  - Microbumps
  - C4 (controlled-collapse chip connection) bumps



(a) Detailed FEM Model

(b) Abstracted FEM Model

A Novel Multi-Fidelity Thermal Modeling Approach

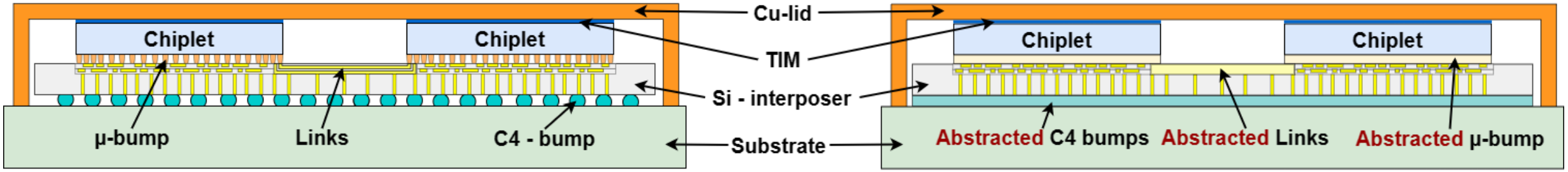


1. **Develop fine-grained FEM Models**
2. **Abstract complex and detailed components**

Iterate until desired fidelity achieved

Objective: < 0.5° temperature difference

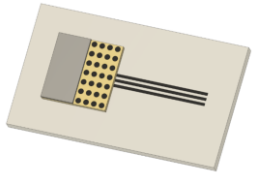
- **Abstract micro-structures that increase the FEM simulation time without significant accuracy impact**



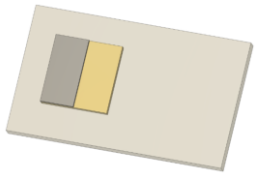
(a) Detailed FEM Model

(b) Abstracted FEM Model

A Novel Multi-Fidelity Thermal Modeling Approach

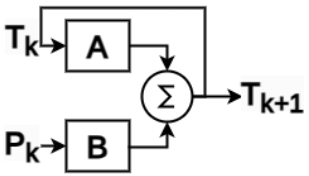


1. Develop fine-grained FEM Models



2. Abstract complex and detailed components

3. Develop thermal RC models



4. Derive design-specific DSS model

Iterate until desired fidelity achieved

Objective: < 0.5° temperature difference

Construct a thermal RC network.

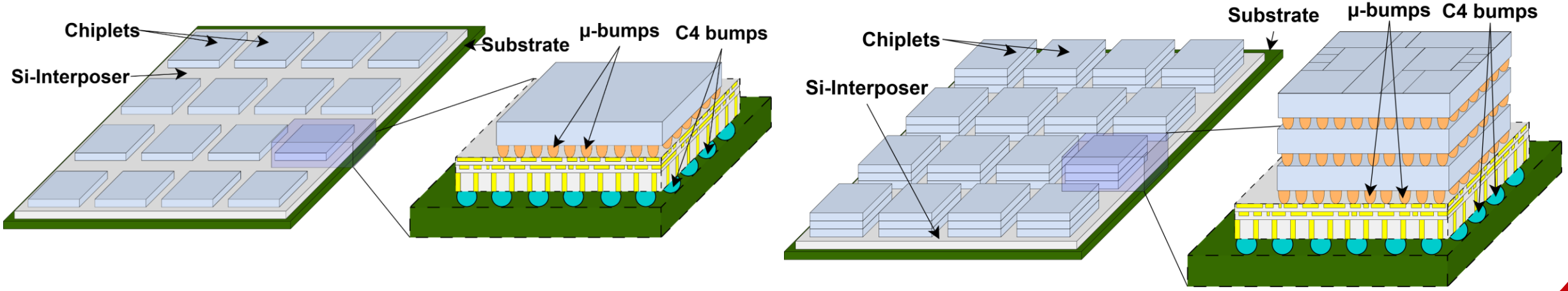
Objective: Minimize the temperature difference with the FEM result

Derive linear discrete state-space models

Objective: Approximate the thermal RC network for the desired fidelity

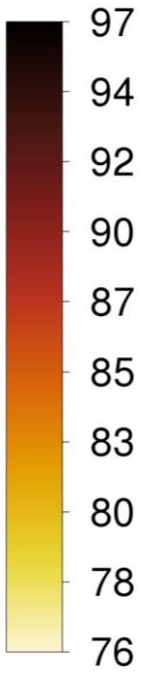
Experimental Setup

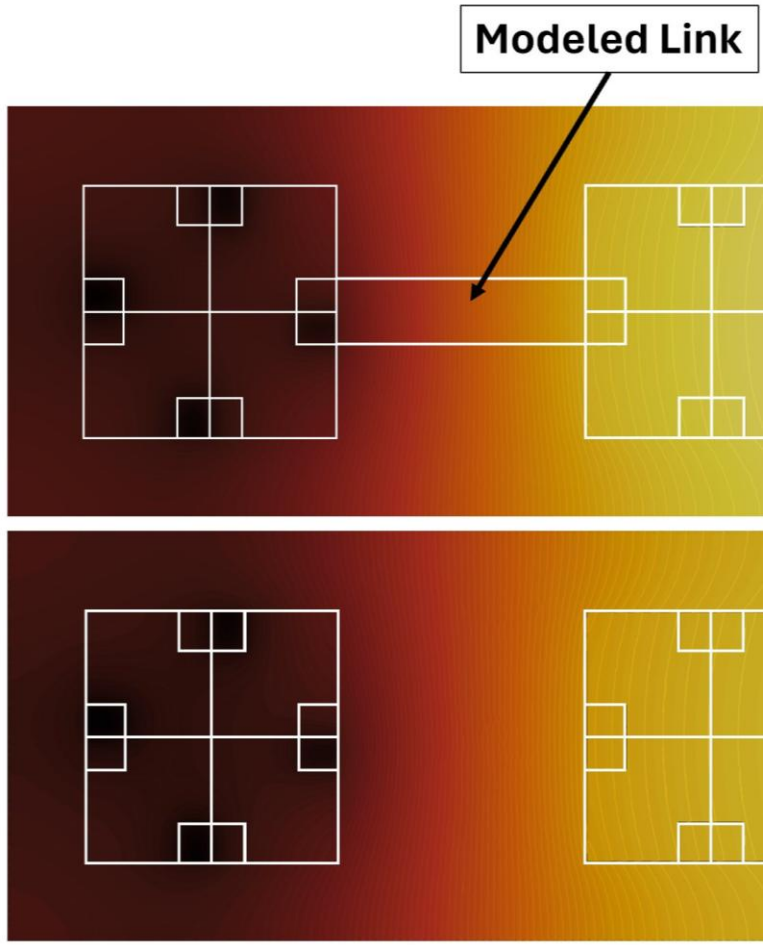
- **FEM simulation is assumed as the golden reference**
  - Evaluate abstracted FEM
  - Thermal RC network
  - Discrete state-space models
- **4x4, 6x6, 8x8 and 4x4x3 chiplet-based systems**
- **Tested on five different neural network workloads**
  - Workloads are a mix of deep neural networks



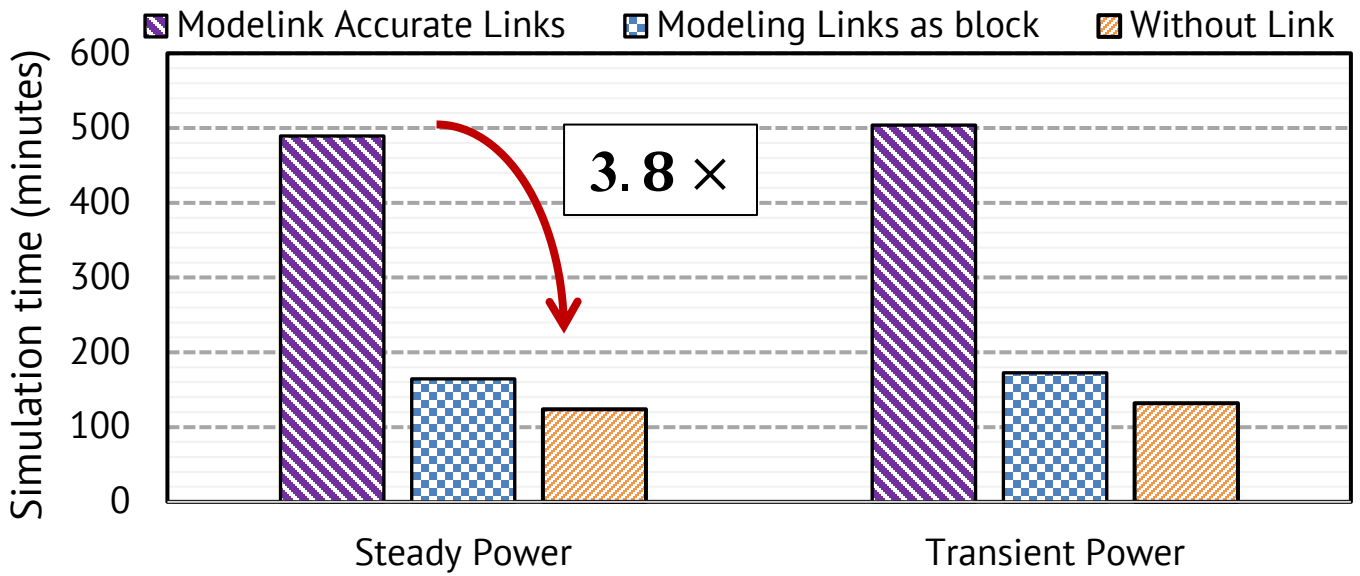
FEM: Accurate vs Abstracted Links

- **One chiplet is active – generating heat**
- **Inactive chiplet heated by *Thermal Crosstalk***
  - ➢ Interposer, heat spreader
- **Three alternative interposer models:**
  1. All links modeled accurately
  2. Links are abstracted as a block
     Material is assumed a mix of Cu and Si
  3. Links are removed



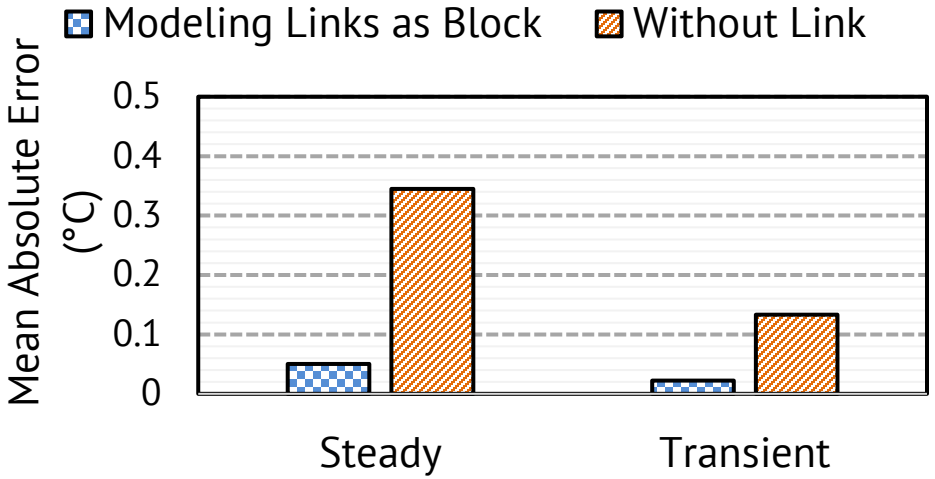


FEM: Accurate vs Abstracted Links



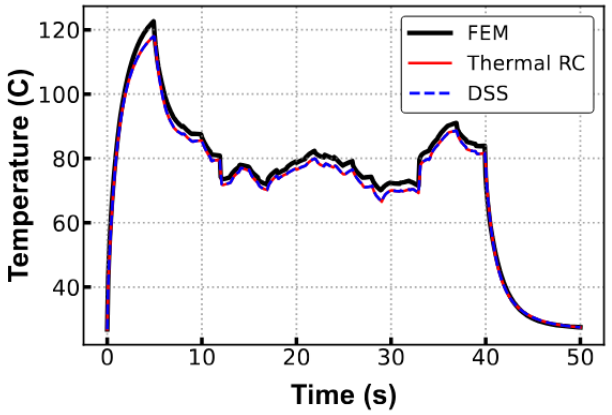
- **Abstracting the links as a single block**
  - Has negligible (<0.05°C) impact
- **Removing links degrades accuracy**

- **Abstracting the links as a single block**
  - Speeds up from 503 min to 132 min
- **Removing links has diminishing returns**

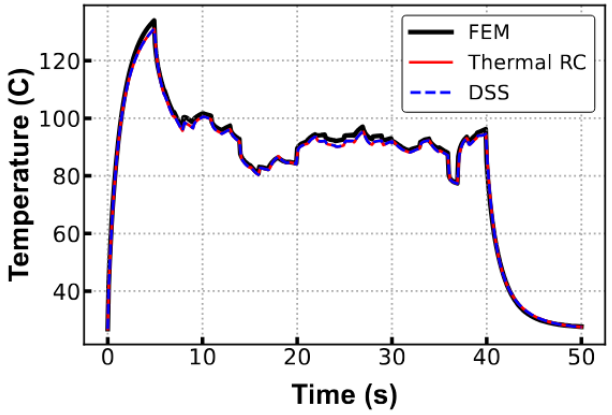


Experimental Results: Accuracy

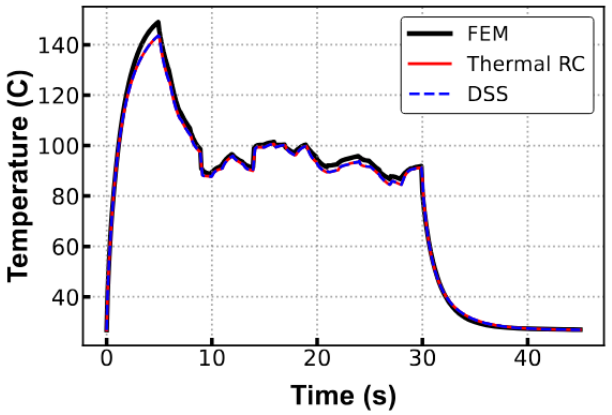
- **Temperature as a function of time**
  - Constant and pseudo-random power inputs to capture transient and steady-state behavior



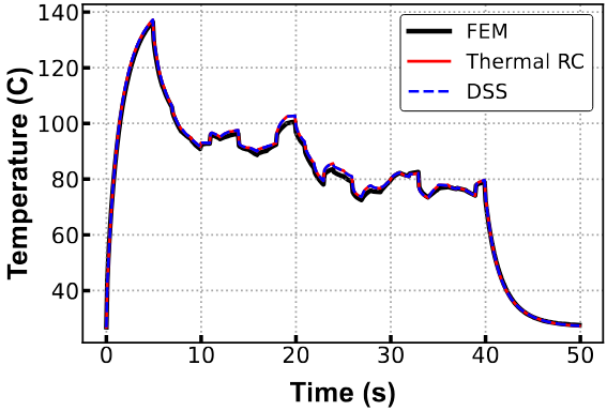
(a) 2.5D - 16 chiplet system



(b) 2.5D - 36 chiplet system



(c) 2.5D - 64 chiplet system

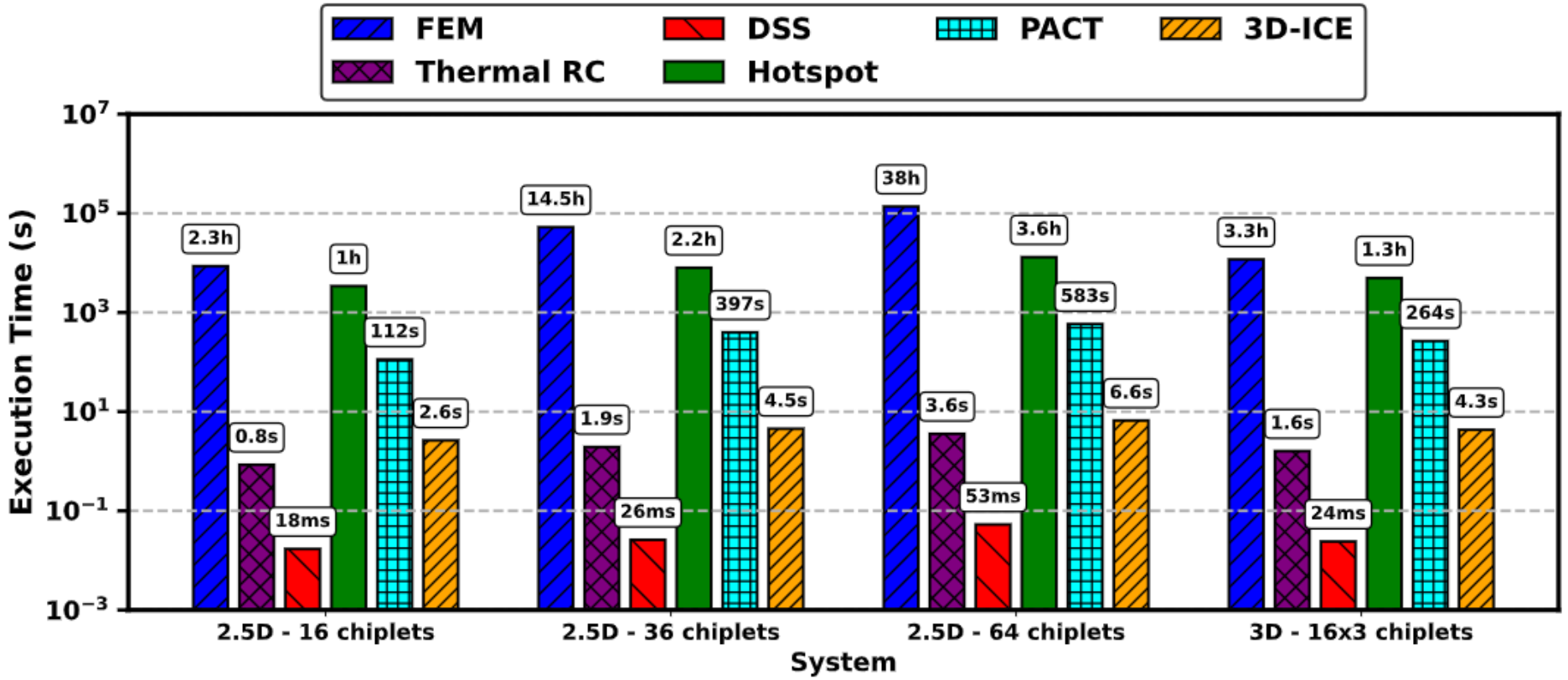


(d) 3D - 16x3 chiplet system

Negligible difference in transient temperature (even smaller difference in steady-state)

Experimental Results: Speedup

- **Thermal RC model is ~10,000x faster than FEM**
- **DSS model is >50x faster compared to our thermal RC**



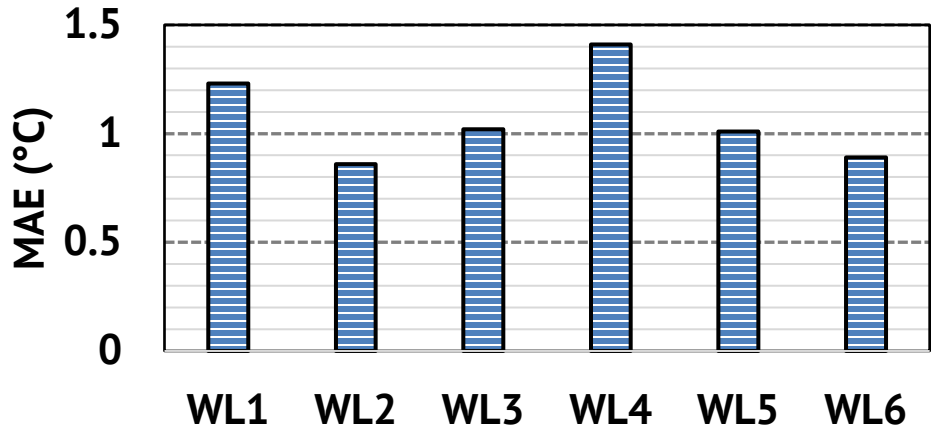
[1] Z. Yuan, et al, "PACT: An Extensible Parallel Thermal Simulator for Emerging Integration and Cooling Technologies," in IEEE TCAD, 2021

[2] A. Sridhar, et al, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling" ICCAD 2010

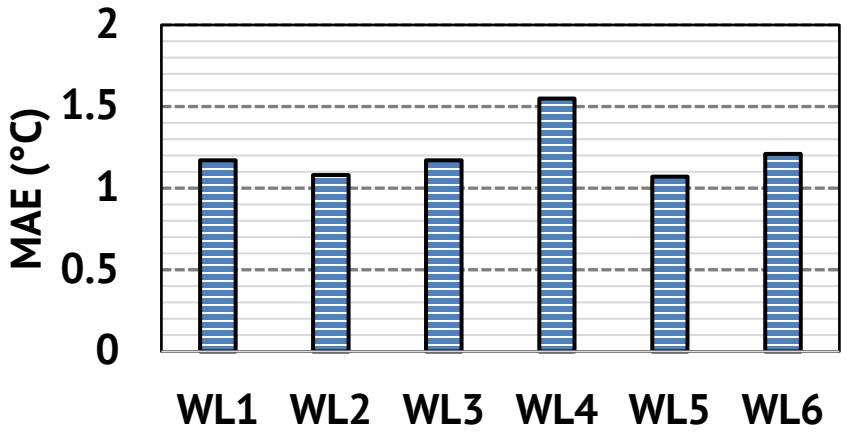
[3] Wei Huang, et al, "HotSpot: a compact thermal modeling methodology for early-stage VLSI design," in *IEEE Transactions on VLSI* 2006

Experimental Results: Error

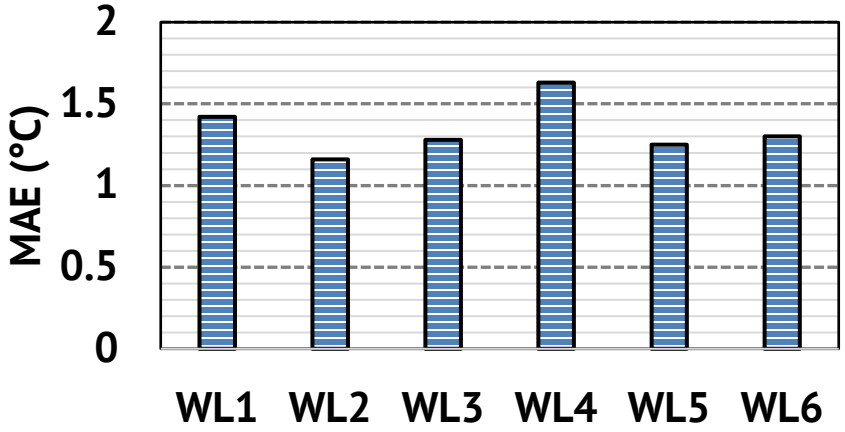
- **<1.7°C Mean Absolute Error wrt FEM**



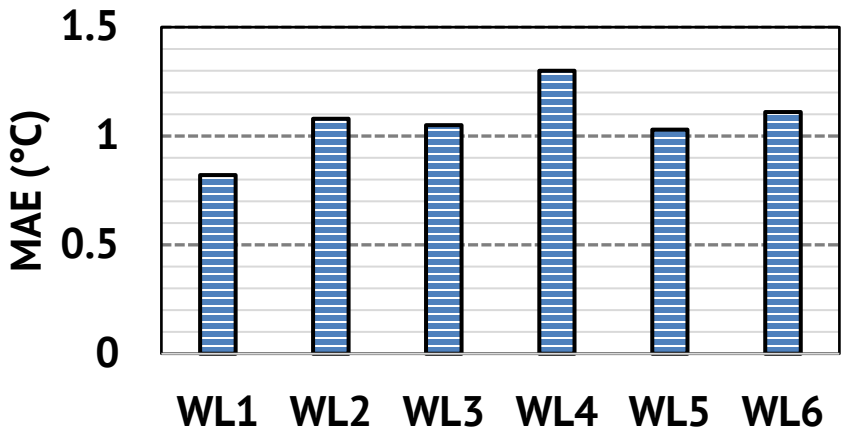
4x4 Chiplet system. Errors wrt FEM model



8x8 Chiplet system. Errors wrt FEM model



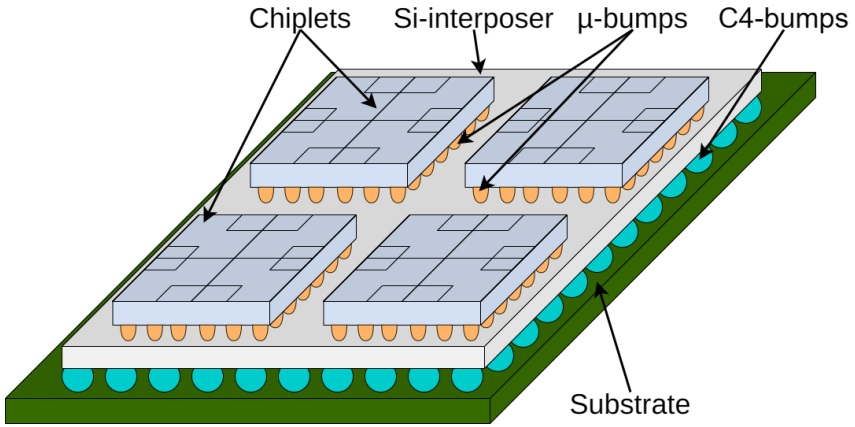
6x6 Chiplet system. Errors wrt FEM model



3D-16x3 Chiplet system. Errors wrt FEM model

Qualitative analysis

	Non-uniform grid	Anisotropic Materials	Non-homogeneous layers	Heat dissipation from both boundaries	Flexible with Architecture changes
Hotspot	✗	✗	✓	✓	✓
PACT	✗	✗	✓	✗	✓
3D-ICE	✓	✗	✗	✗	✓
Thermal RC(ours)	✓	✓	✓	✓	✓
DSS (ours)	✓	✓	✓	✓	✗

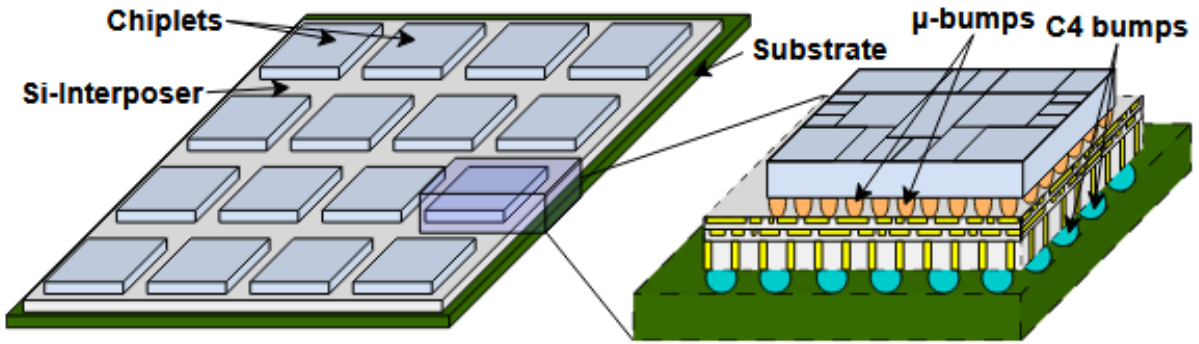


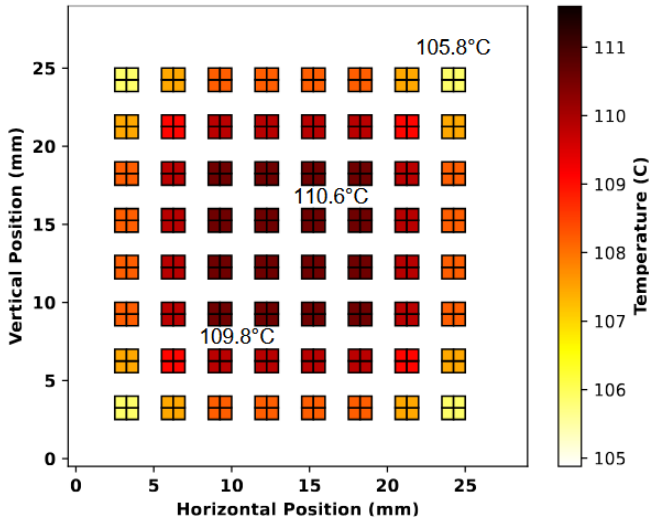
[1] Z. Yuan, et al, "PACT: An Extensible Parallel Thermal Simulator for Emerging Integration and Cooling Technologies," in IEEE TCAD, 2021
[2] A. Sridhar, et al, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling" ICCAD 2010
[3] Wei Huang, et al, "HotSpot: a compact thermal modeling methodology for early-stage VLSI design," in *IEEE Transactions on VLSI* 2006

MFIT Conclusions

- We proposed MFIT, a multi-fidelity thermal modeling tool for every stage of the chiplet system design process
- MFIT balances speed and accuracy, matching the requirements of the current design stage, while supporting a broad range of chiplet system configurations

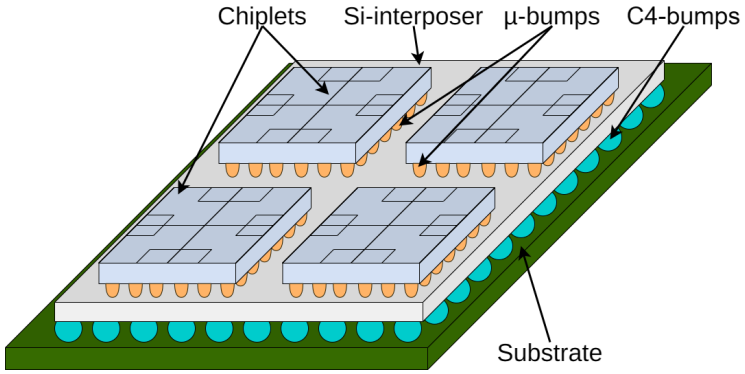
- Published at ACM Transactions on Design Automation of Electronic Systems, 2025
- GitHub repo: https://github.com/AlishKanani/MFIT

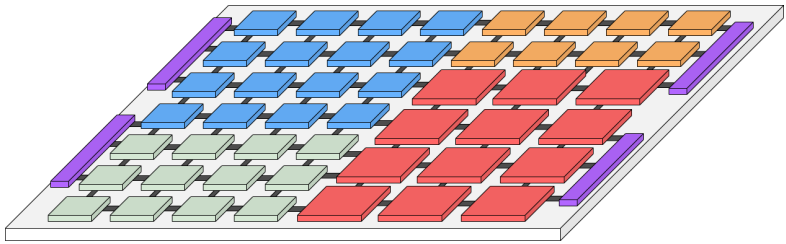




Outline

- **Motivation: Chiplet-based platforms**
- **Preliminary Work-1:**
  - MFIT : Multi-Fidelity Thermal Modeling for 2.5D and 3D Multi-Chiplet Architectures
- **Preliminary Work-2:**
  - THERMOS: Thermally-Aware Multi-Objective Scheduling for Heterogeneous Multi-Chiplet PIM Architectures
- **Ongoing and Proposed Work:**
  - Disaggregated Acceleration of Hybrid Mamba–Transformer LLMs via Systolic Prefill and Vector Decode
  - Breaking the Memory Wall in MoE LLMs with Expert Prefetching
- **Timeline**
- **Conclusions**



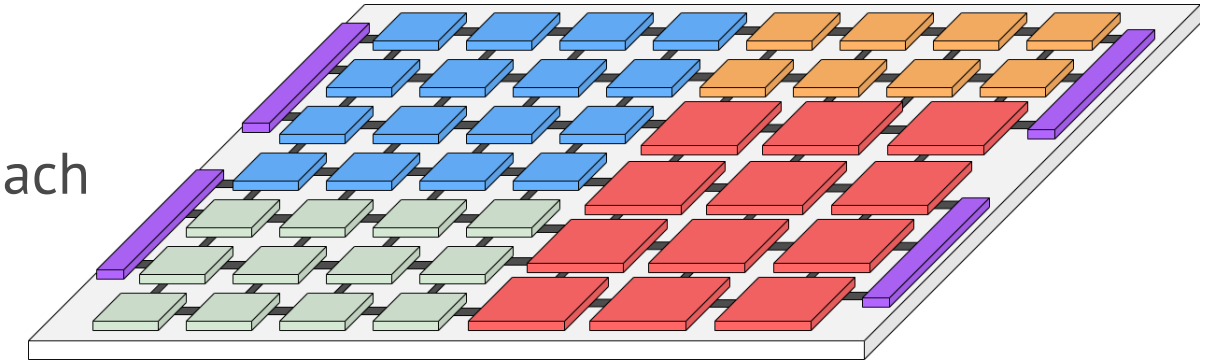


Heterogeneous chiplet system

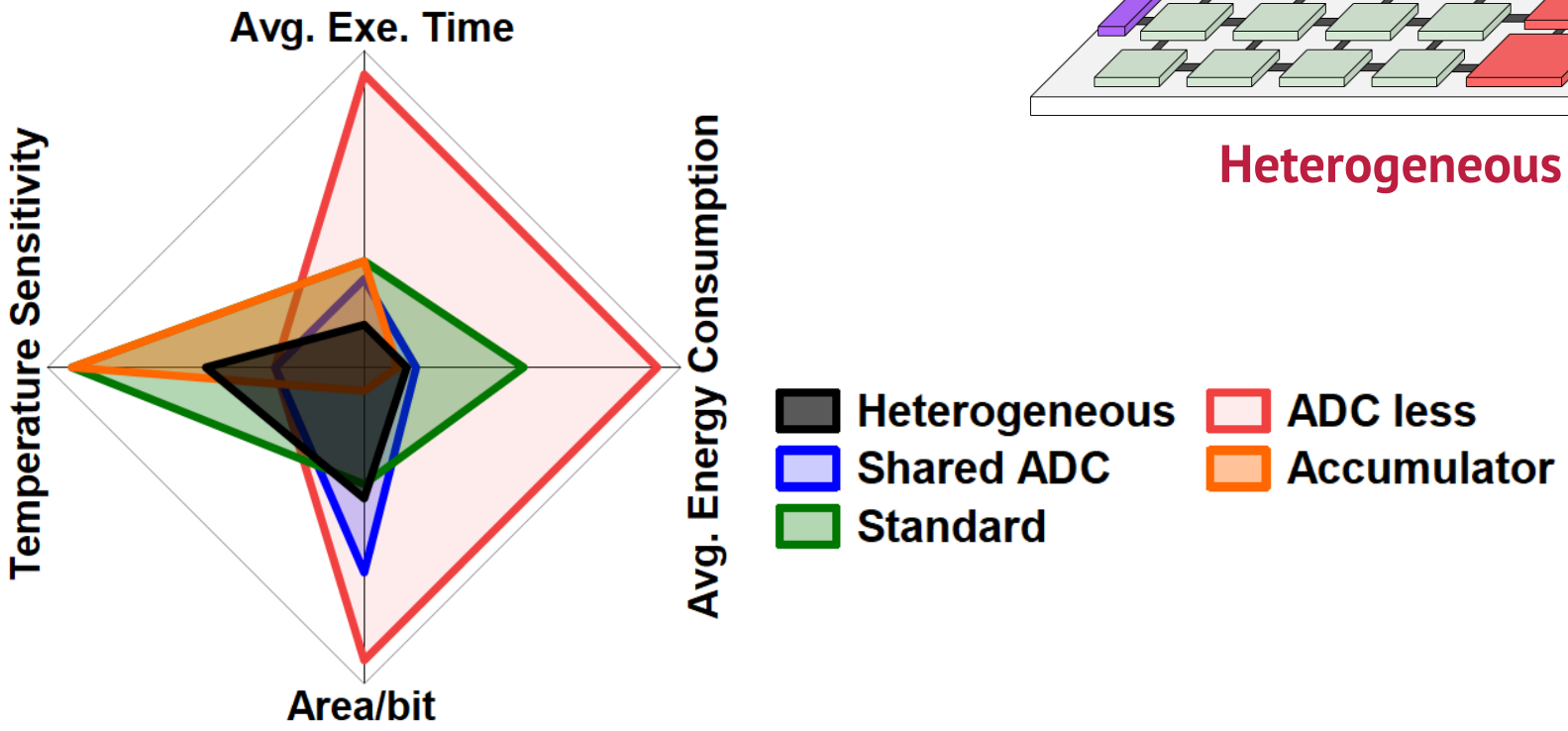
Alish Kanani, et al. "THERMOS: Thermally-Aware Multi-Objective Scheduling of AI Workloads on Heterogeneous Multi-Chiplet PIM Architectures." presented in ESWEEK in Oct. 2025 and published in *ACM Transactions on Embedded Computing Systems*

Why Heterogeneous Architectures?

- **Chiplets can have different characteristics**
  - High performance, more energy efficient, etc.
  - Heterogeneity can leverage the strengths of each



Heterogeneous chiplet system



Related Work

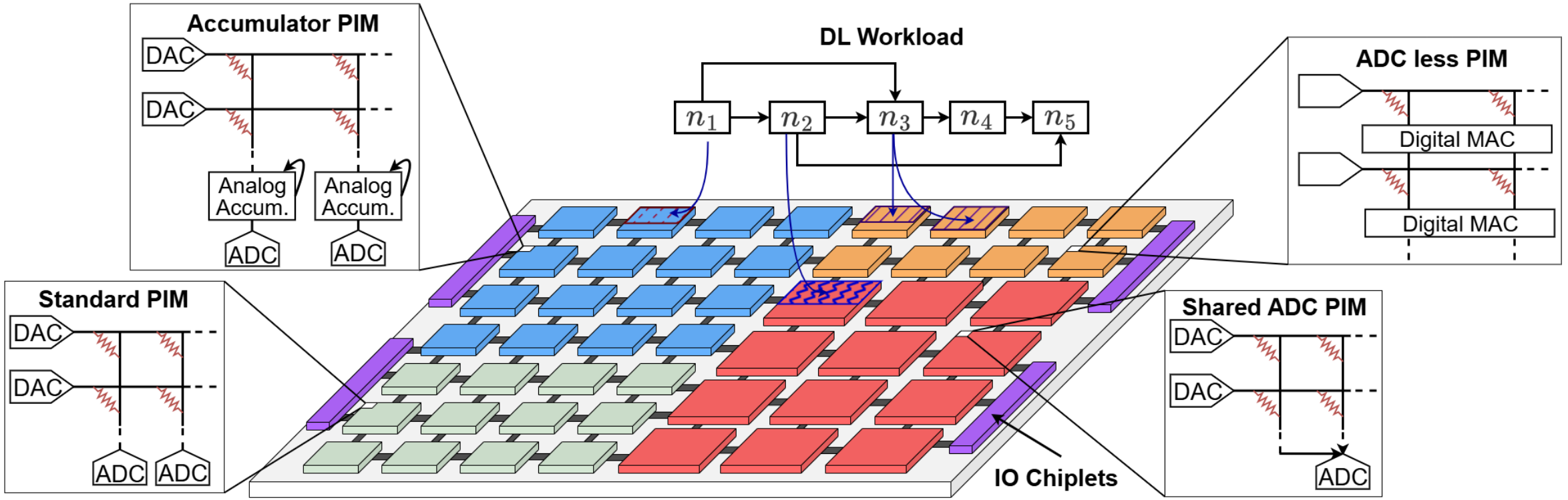
- **Big-Little chiplets:** heterogeneous chiplet sizes; heuristic scheduling; not thermally-aware
- **SIAM / Simba:** homogeneous chiplets; heuristic schedulers; single-objective
- **RELMAS:** RL-based scheduling, but still not thermally aware or multi-objective

Work	Heterogeneous	Scheduling Method	Multi-Objective	Thermally-aware
Big-Little				×
SIAM (ES				×
Simba (N				×
RELMAS (DAC'24)	Simba + Eyeriss	Single objective RL	×	×
THERMOS	**Different PIM chiplets**	**Multi objective RL + Heuristic**	✓	✓

None of the prior work consider multi-objective optimization with thermal awareness

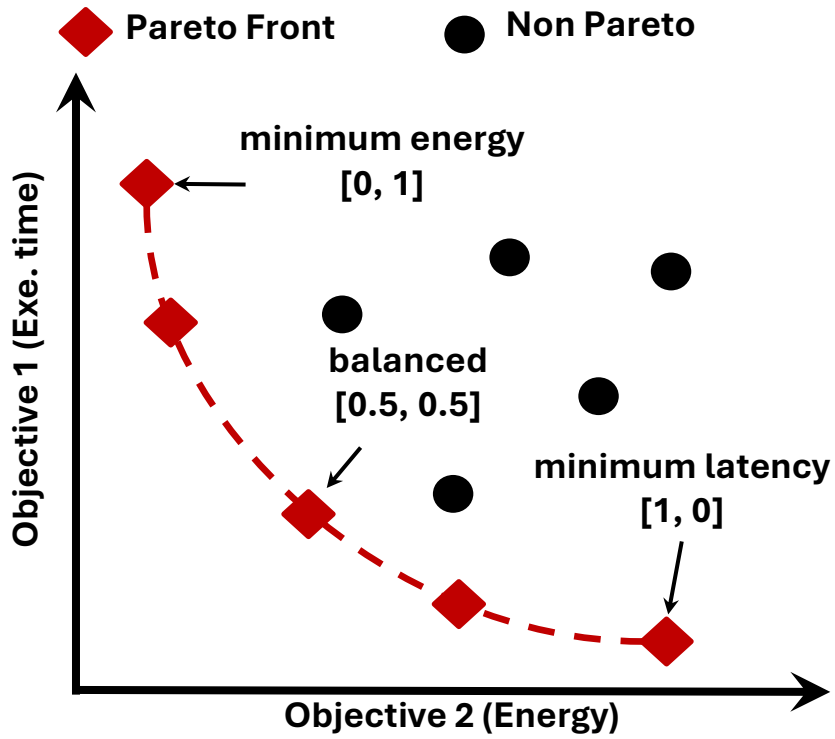
Target Architecture Overview

- **Four processing-in-memory (PIM) chiplet clusters**
  - ReRAM: Standard, Accumulator
  - SRAM: ADC-less, Shared ADC
- **Interconnected via a network on interposer (NoI)**



DL Workload Scheduling to Chiplets

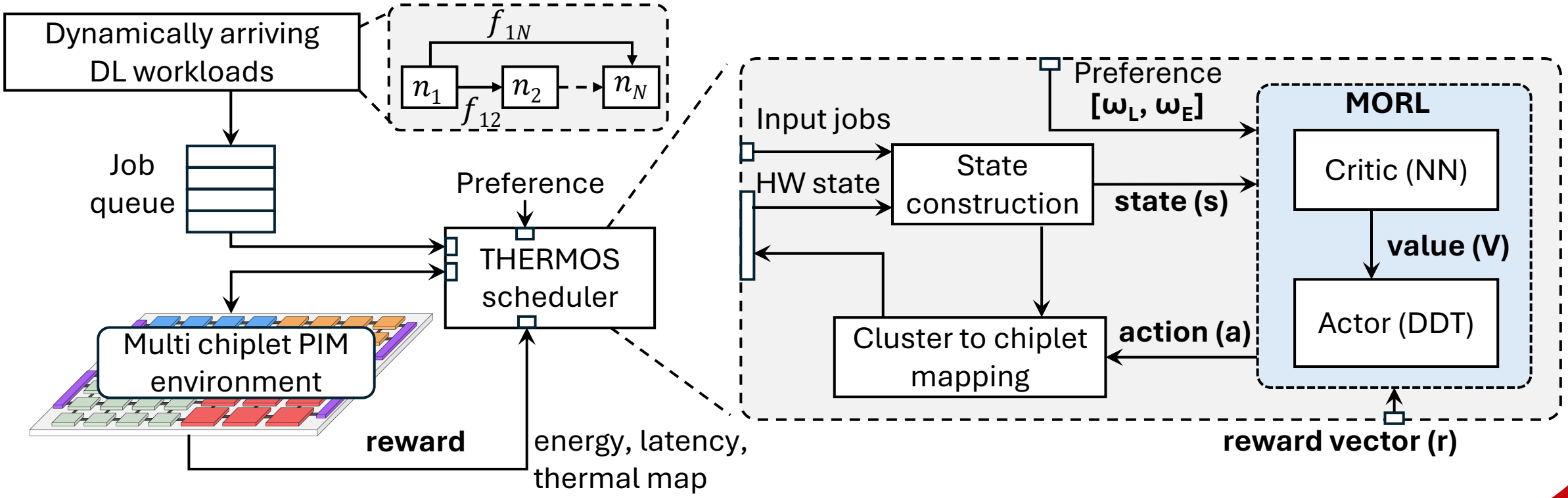
- **Multi-objective: co-optimize execution time & energy**
  - Runtime adaptability via **preference vectors**: minimum latency, minimum energy, balanced
  - Examples:
    - [0.5, 0.5]: Balanced
    - [1, 0]: Performance only
- **Must respect thermal thresholds**
  - Lower threshold for ReRAM, higher for SRAM
- **Decision space**
  - Given layer → which chiplet cluster → which chiplet



THERMOS Framework: Overview

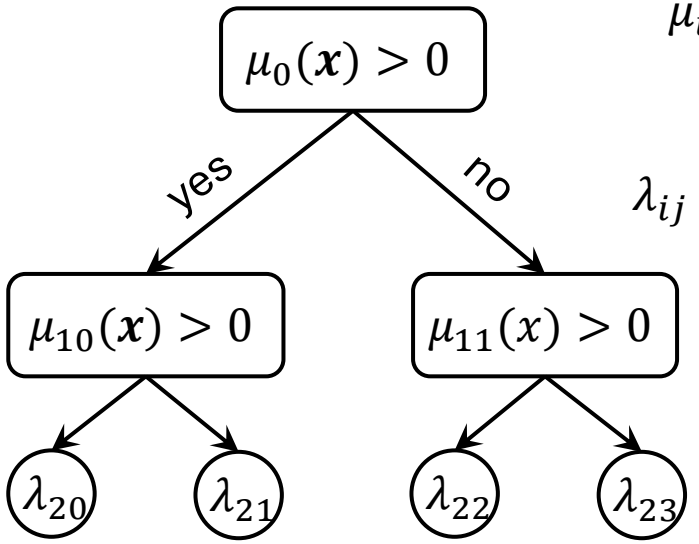
Hierarchical scheduler with two levels

1. **Multi-Objective RL (MORL) agent**: selects chiplet cluster (preference-aware)
2. **Proximity-driven algorithm**: maps layers to chiplets within cluster



Lightweight RL Agent: DDT

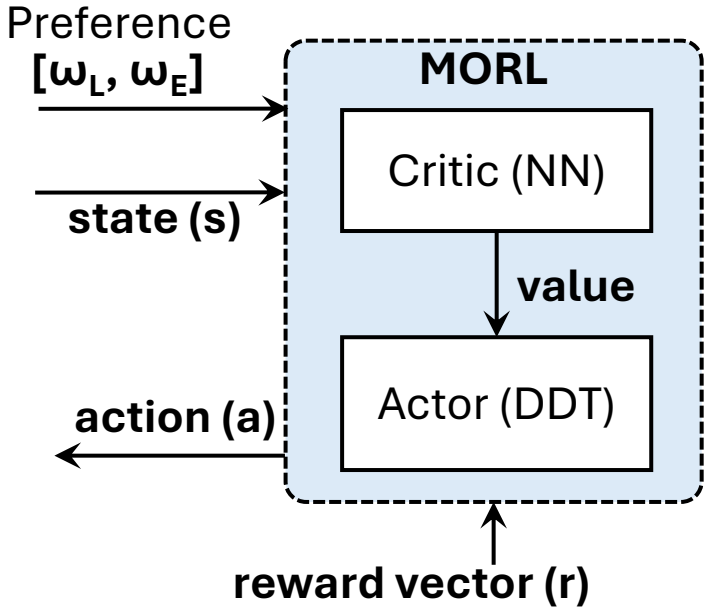
- **Differential decision tree (DDT)**
  - Interpretable and lightweight
- **Efficient compared to neural network policies**



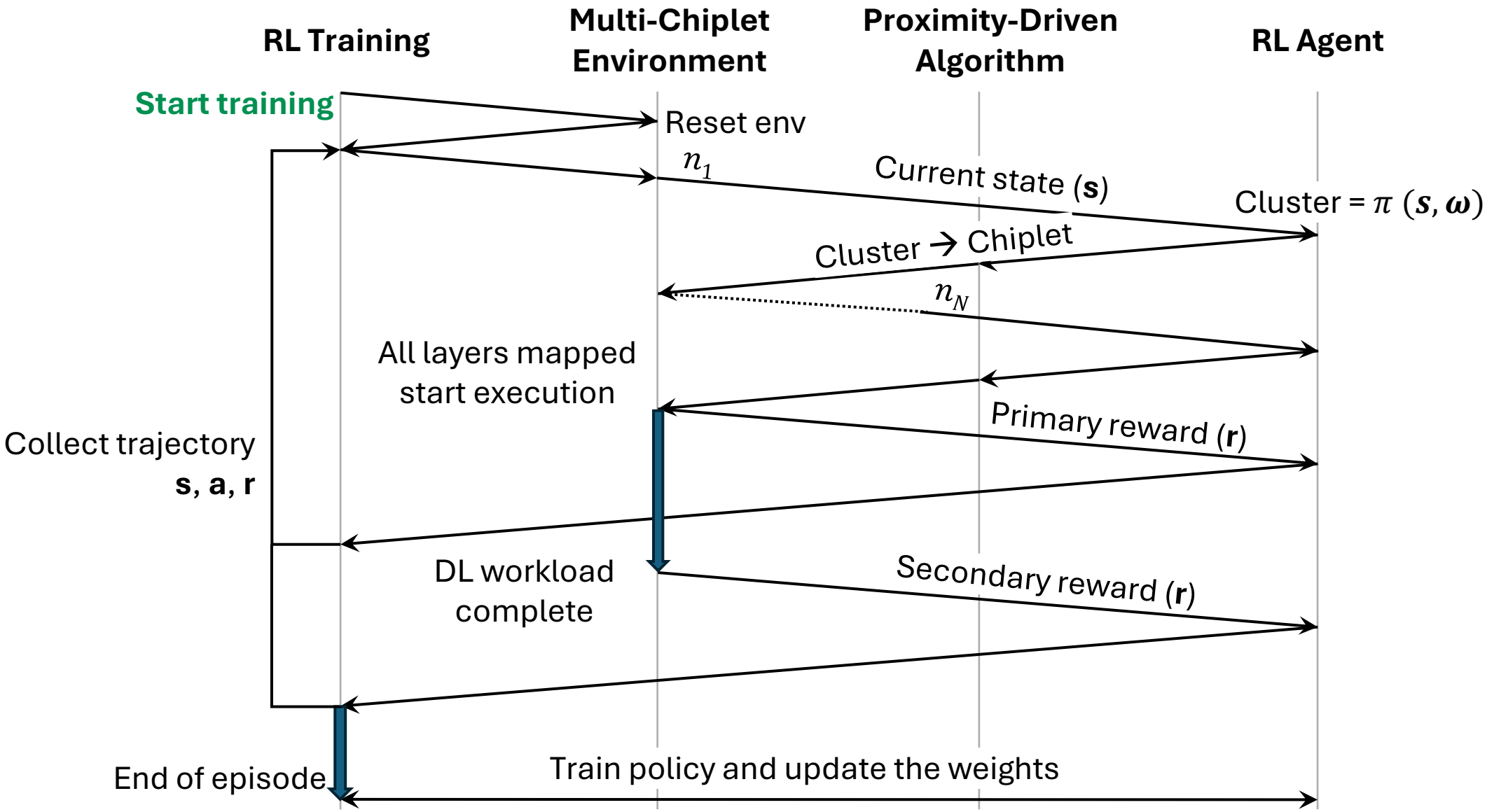
$$\mu_{ij}(\boldsymbol{x}) \coloneqq \frac{1}{1 + e^{-\alpha(\boldsymbol{\beta}_{ij}^T \boldsymbol{x} - \phi_{ij})}}$$

where, $\boldsymbol{x} = [\boldsymbol{s}, \boldsymbol{\omega}]^T$

$$\lambda_{ij} \coloneqq [p(v_1), p(v_2), p(v_3), p(v_4)]$$

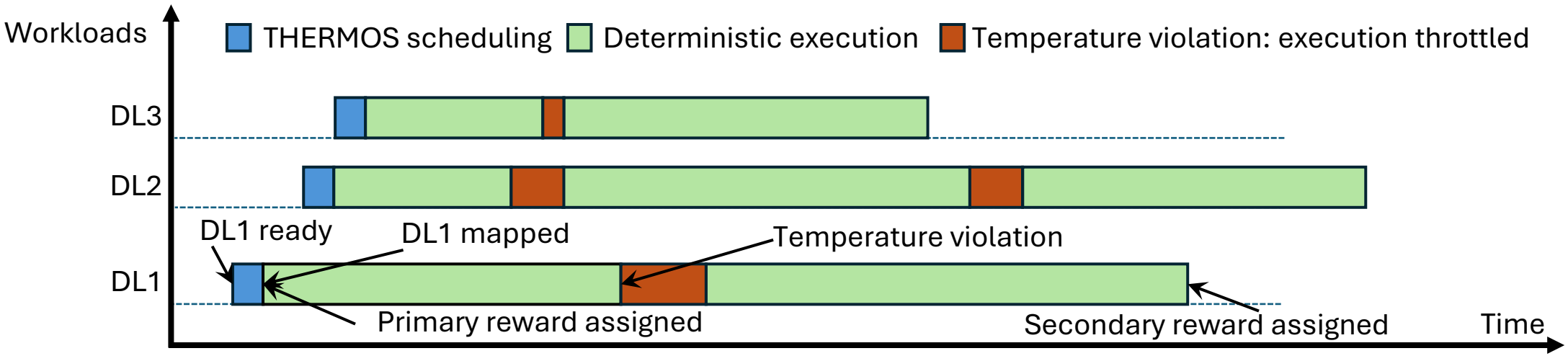


Multi-Objective RL (MORL): Training



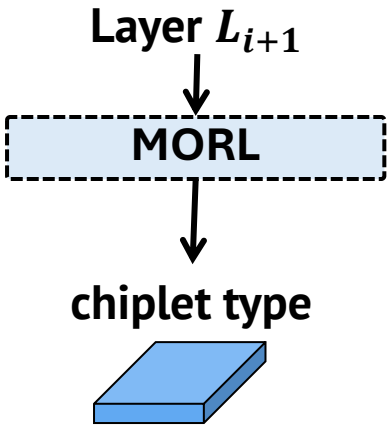
Handling Thermal Constraints: Delayed Reward

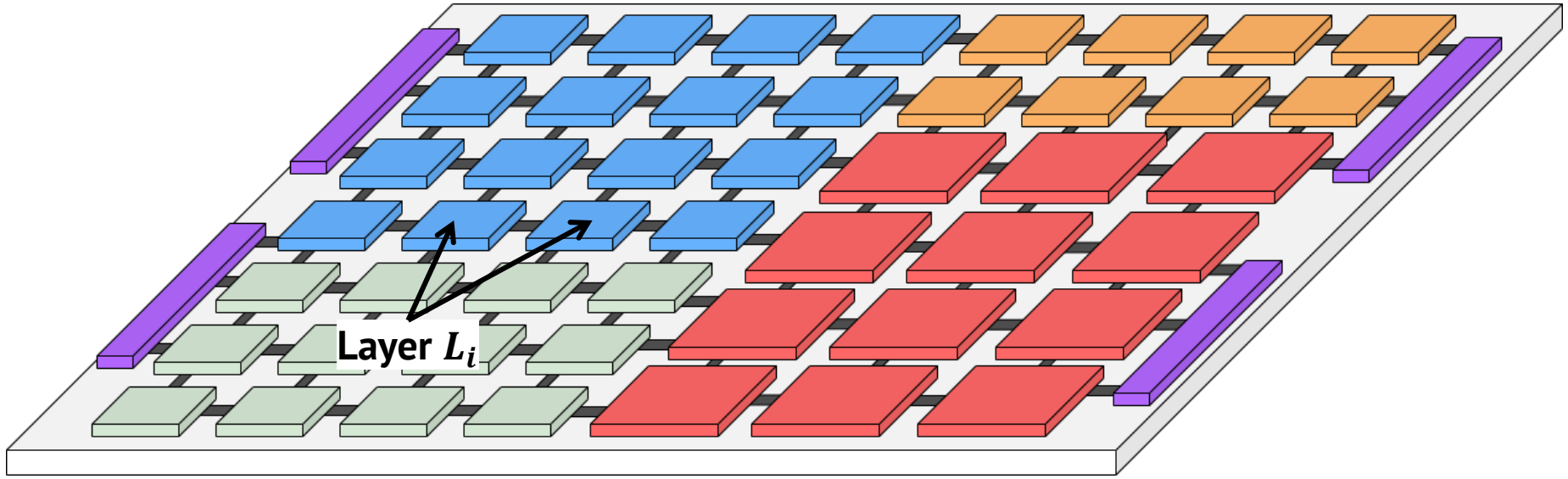
- **Temperature-sensitive PIM requires throttling**
  - Throttling prevents accuracy degradation
- **THERMOS splits rewards: primary (deterministic) + secondary (thermal effects).**



Proximity-driven algorithm

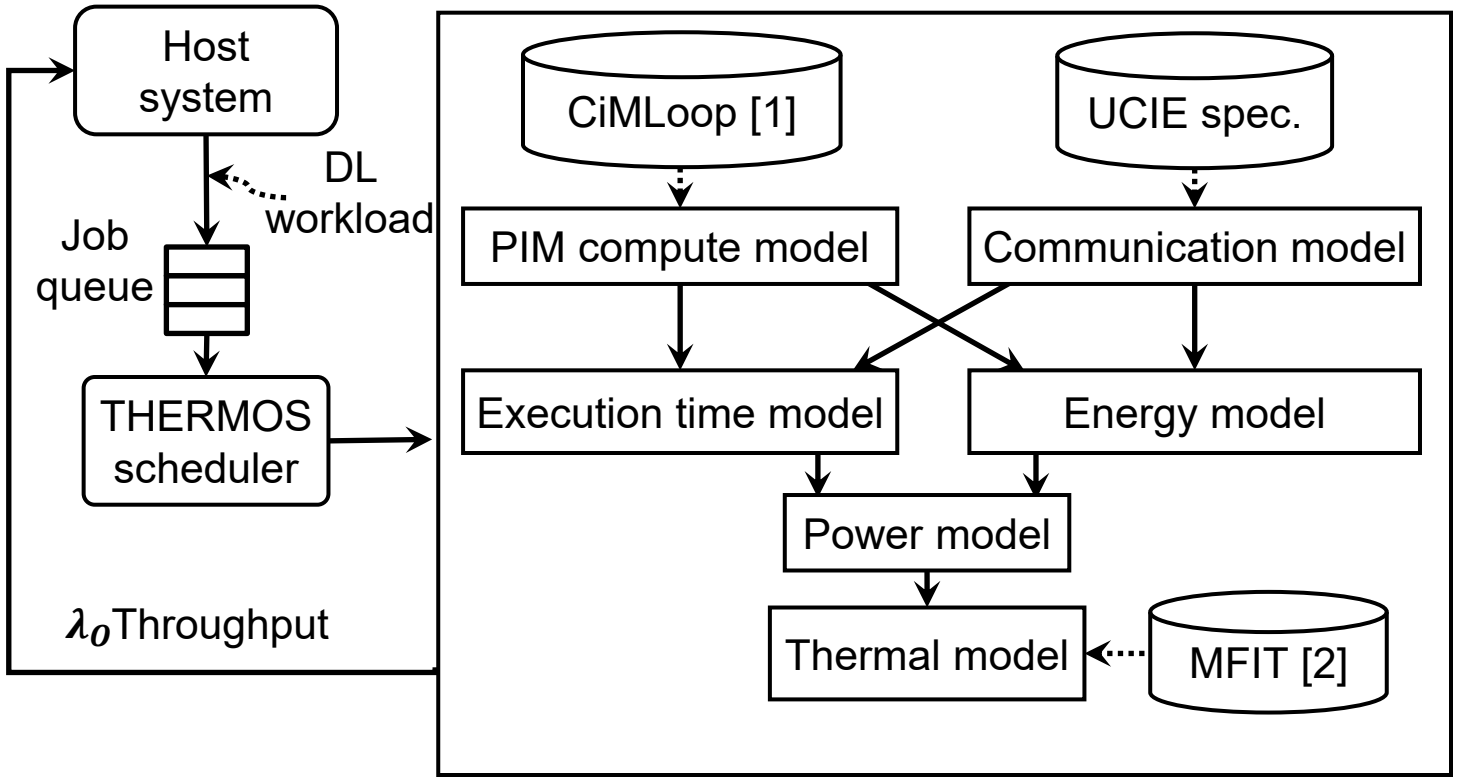
- **MORL selects a PIM cluster, layers must be mapped to specific chiplets**
- **Proximity-driven algorithm minimizes inter-chiplet communication**
  - Prioritize chiplets used in the **previous layer**
  - Compute weighted distance from those chiplets to candidates
  - **Fill nearest chiplet(s) to capacity** before moving farther





Experimental Setup

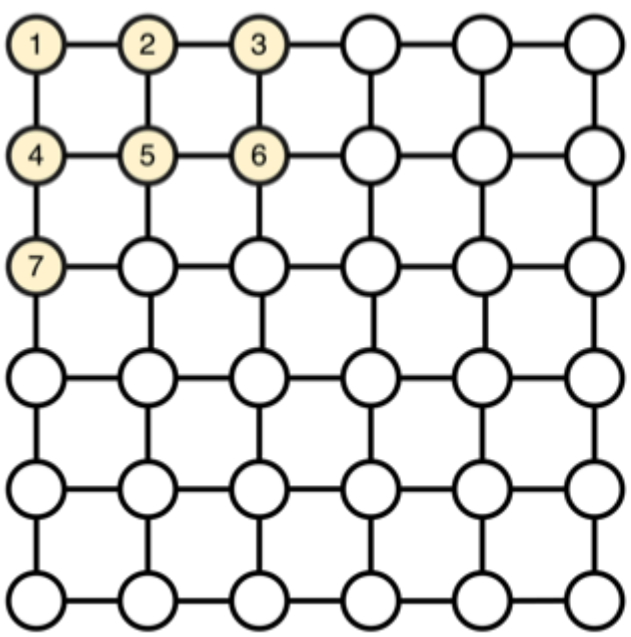
- **Chiplet system**
  - 25 Standard
  - 28 Shared ADC
  - 10 Accumulator
  - 15 ADC-less
- **Simulators**
  - CiMLoop (Compute)
  - UCIe (Communication)
  - MFIT (Thermal)
- **Ensemble of Models:**
  - AlexNet, ResNet, MobileNet
    …



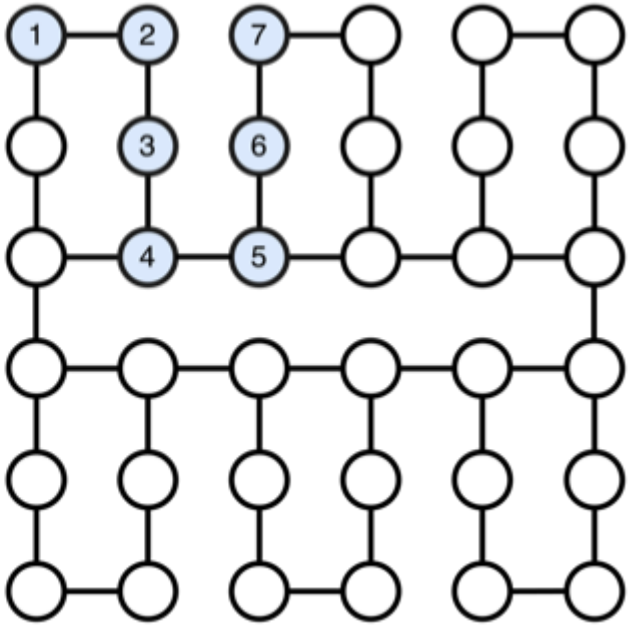
Simulation framework for THERMOS

[1] Andrulis, T., Emer, J.S. and Sze, V. "CiMLoop: A flexible, accurate, and fast compute-in-memory modeling tool." *ISPASS*. IEEE 2024.
[2] Pfromm, L., Kanani, A., Sharma, H., Solanki, P., Tervo, E., Park, J., Doppa, J., Pande, P.P. and Ogras, U. "MFIT: Multi-fidelity thermal modeling for 2.5 D and 3D multi-chiplet architectures." *TODAES, ACM* 2025.

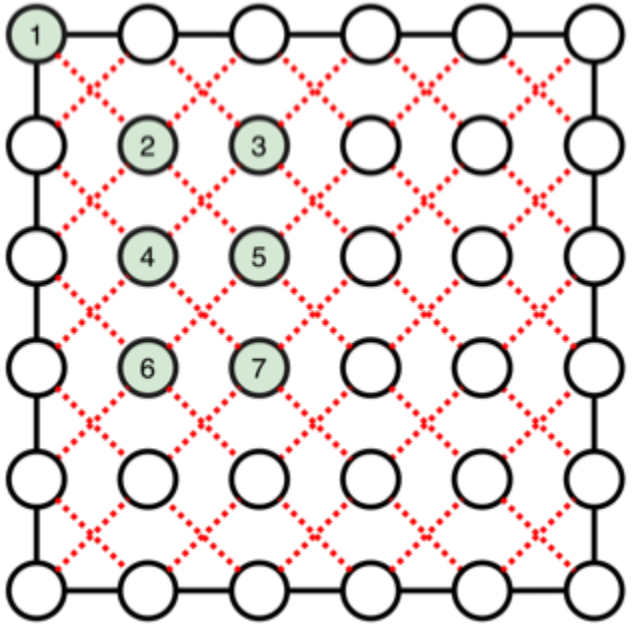
Different Network on Interposer Topologies



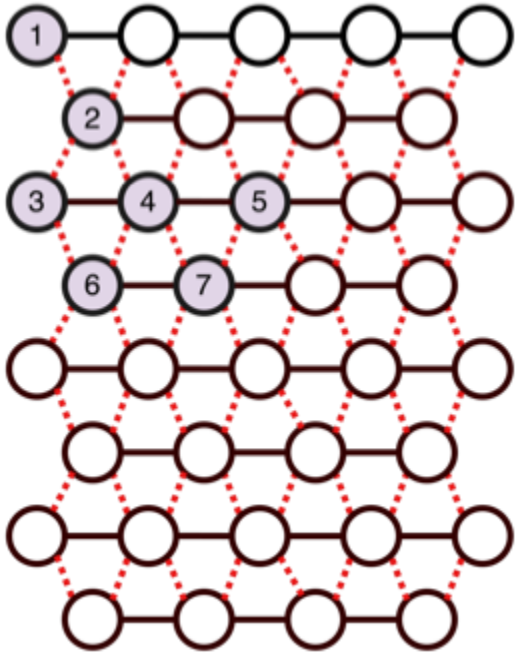
Mesh



Floret [1]



Kite [2]



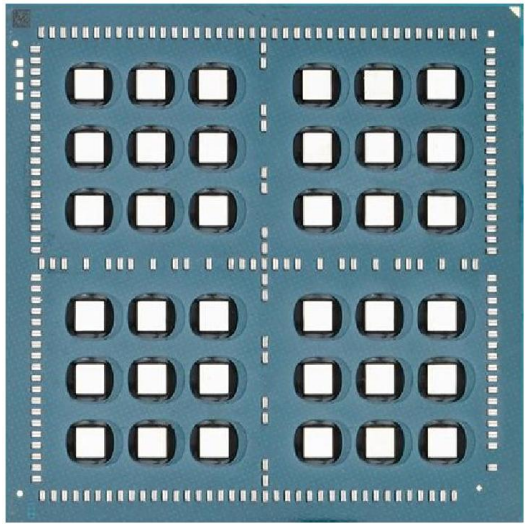
Hexamesh [3]

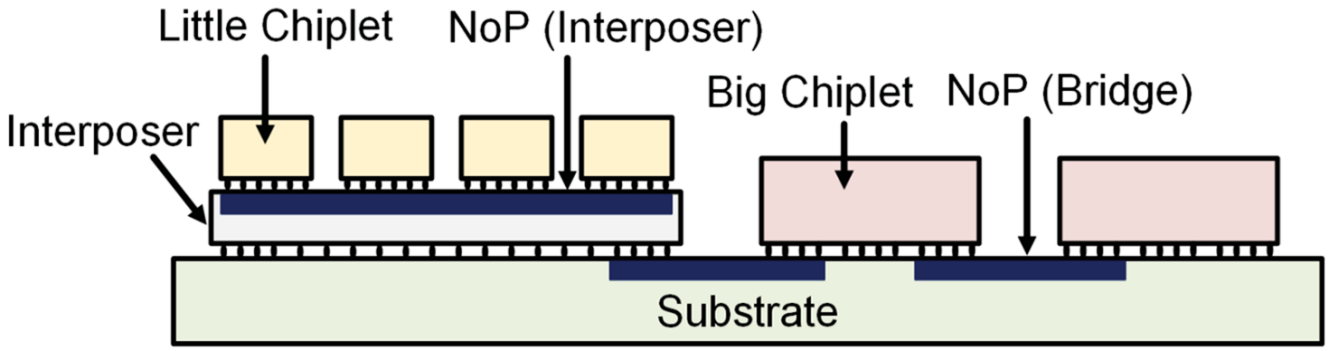
DL1 (ResNet18)→ DL2 (ResNet50) → DL3 (VGG19)→…

[1] Sharma, Harsh, et al. "Florets for chiplets: Data flow-aware high-performance and energy-efficient network-on-interposer for CNN inference tasks." *ACM TECS,* 2023
[2] Bharadwaj, Srikant, et al. "Kite: A family of heterogeneous interposer topologies enabled via accurate interconnect modeling." *2020 IEEE DAC*, 2020
[3] Iff, Patrick, et al. "Hexamesh: Scaling to hundreds of chiplets with an optimized chiplet arrangement." 2023 *IEEE DAC*, 2023

Baseline

- **Simba [1]**
  - Nearest neighbor scheduling
  - Minimize communication
- **Big-little chiplets [2]**
  - Based on crossbar utilization
  - Maximize compute utilization
- **RELMAS [3]**
  - Single objective RL
  - Flat approach to select individual chiplets

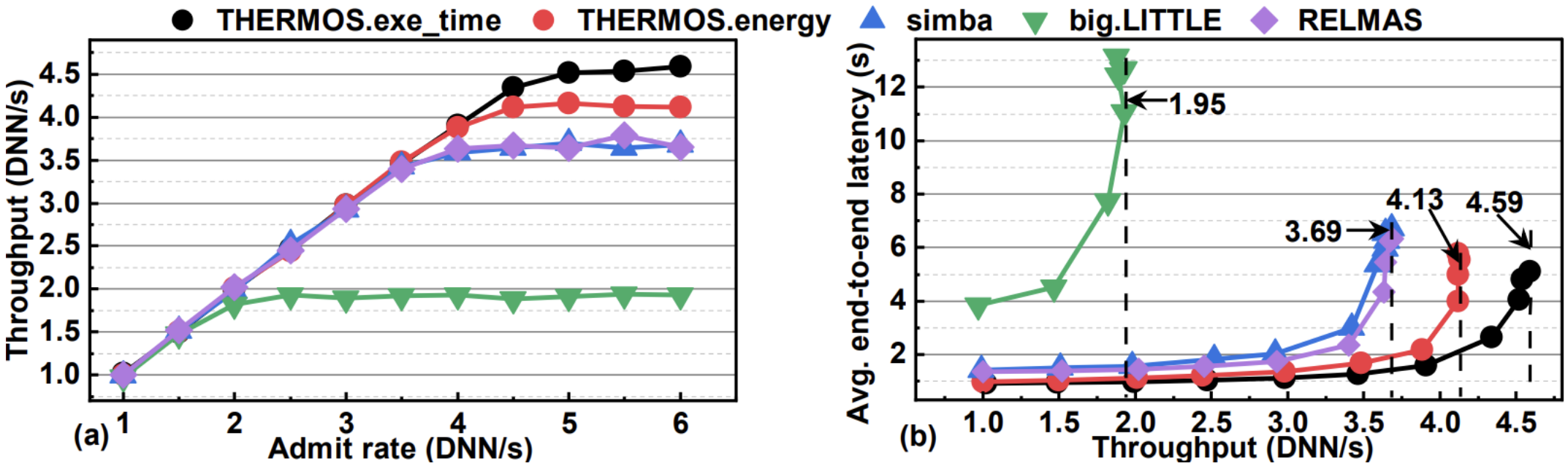




[1] Shao, Yakun Sophia, et al. "Simba: Scaling deep-learning inference with multi-chip-module-based architecture." *MICRO,* 2019.
[2] Krishnan, Gokul, et al. "Big-little chiplets for in-memory acceleration of dnns: A scalable heterogeneous architecture." *ICCAD,* 2022.
[3] Blanco, Francesco Giulio, et al. "A deep reinforcement learning based online scheduling policy for deep neural network multi-tenant multi-accelerator systems." *DAC,* 2024.

Results: Throughput & End-to-end latency

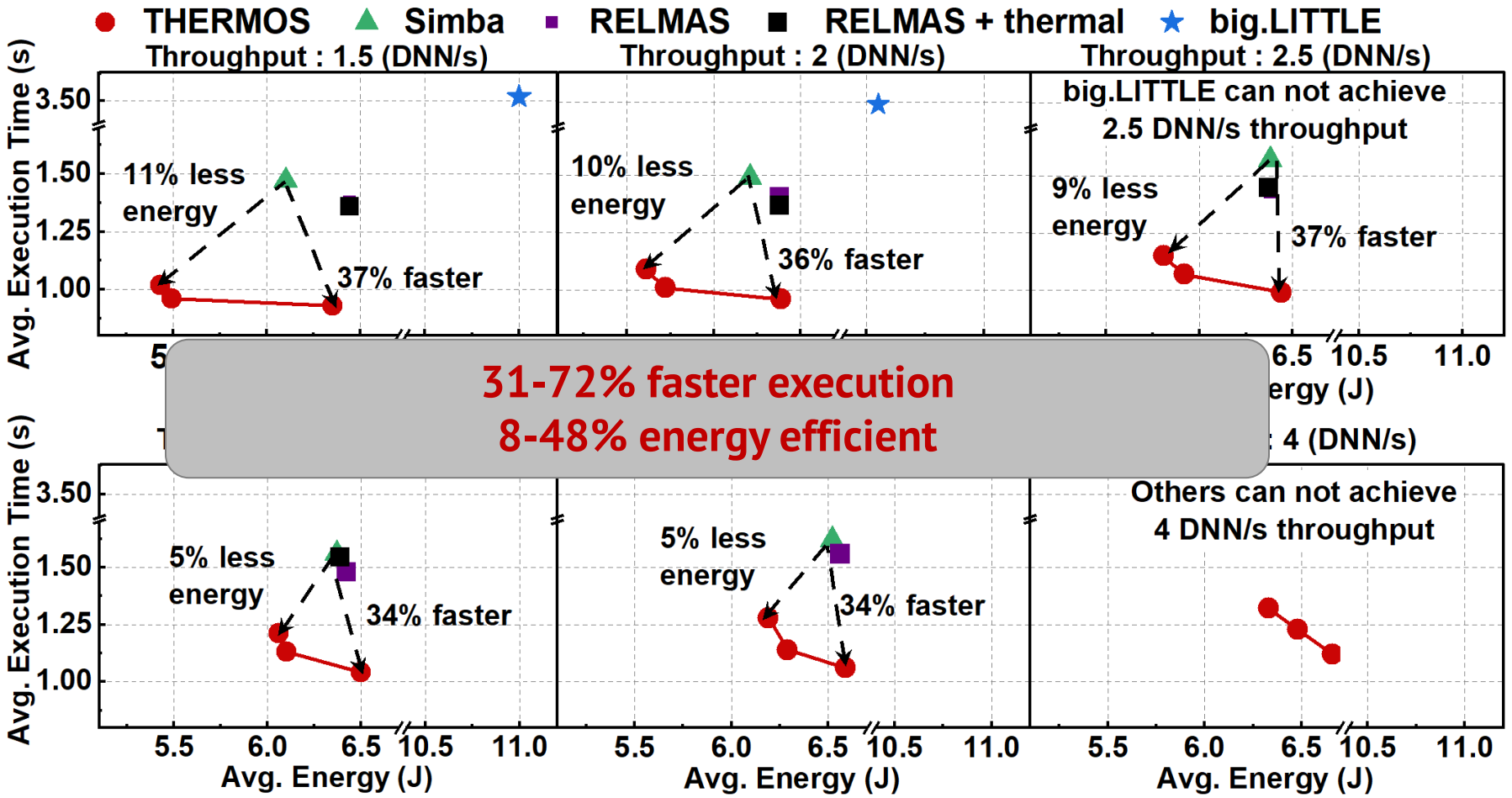
- THERMOS achieves higher throughput across admit rates
- Baselines saturate early → queueing delays



Up to **4.59 DNN/s** throughput vs **<3.7** for baselines

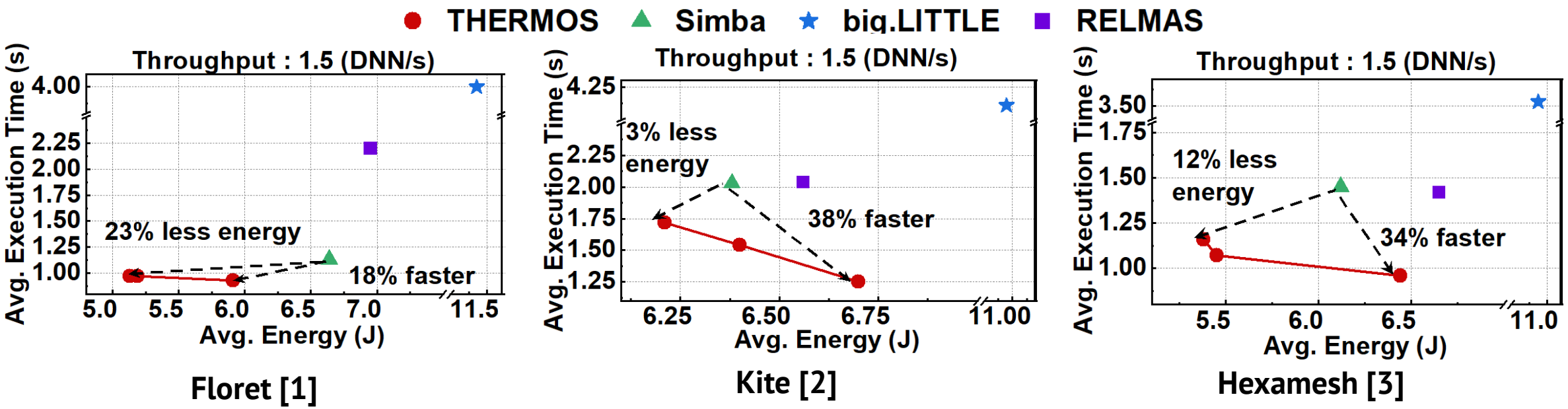
Results: Pareto Fronts (Mesh NoI)

- Single policy achieves Pareto-optimal trade-offs
- Outperforms baselines in both execution time and energy



Results: Across NoI Topologies

- Robust performance across Mesh, Floret, Hexamesh, Kite
- THERMOS adapts to system-level differences

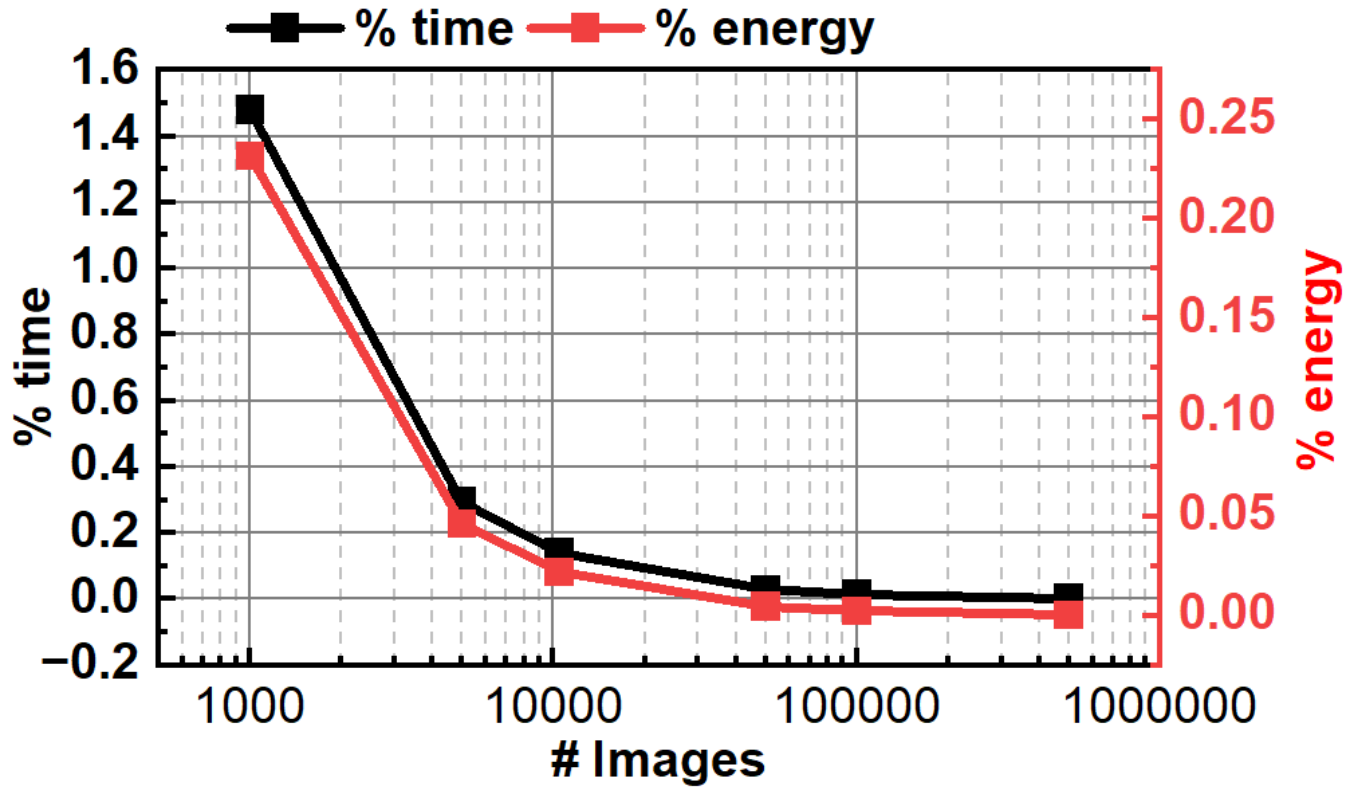


Up to 88% faster execution
57% energy efficient

[1] Sharma, Harsh, et al. "Florets for chiplets: Data flow-aware high-performance and energy-efficient network-on-interposer for CNN inference tasks." *ACM TECS*, 2023
[2] Bharadwaj, Srikant, et al. "Kite: A family of heterogeneous interposer topologies enabled via accurate interconnect modeling." *2020 IEEE DAC*, 2020
[3] Iff, Patrick, et al. "Hexamesh: Scaling to hundreds of chiplets with an optimized chiplet arrangement." 2023 *IEEE DAC*, 2023

Results : Overhead Analysis

- Implemented on Nvidia Jetson Xavier NX
- Runtime overhead is negligible

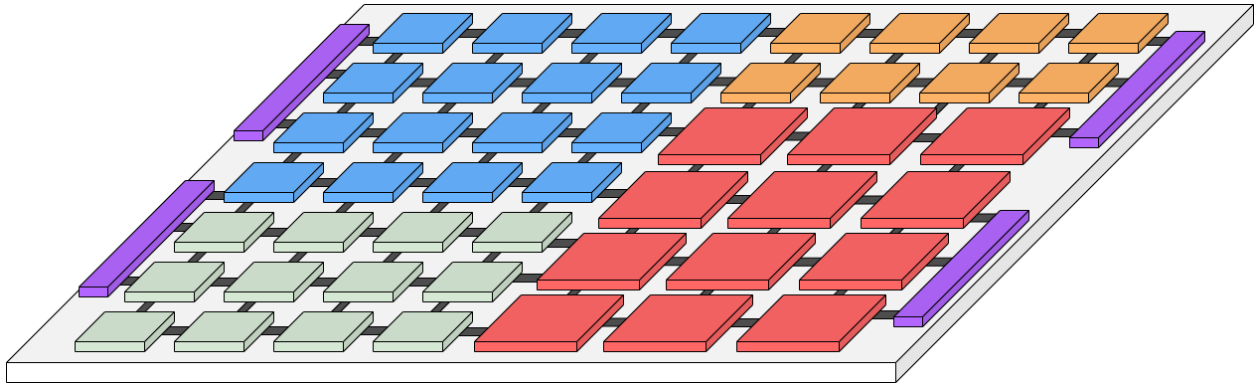


0.14% of DL execution time, 0.022% of DL energy with 10,000 images

THERMOS Conclusions

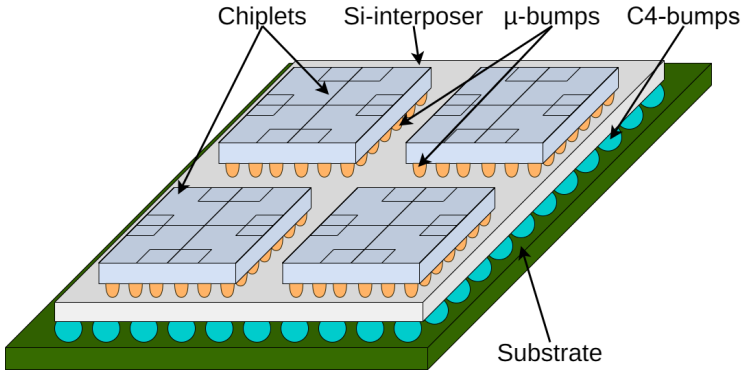
- Heterogeneous chiplets are essential for future AI workloads
- THERMOS: adaptable, thermally-safe, high-throughput scheduling
- Significant performance + energy benefits with negligible overhead
- Broader vision: scaling chiplet architectures for reliable AI deployment

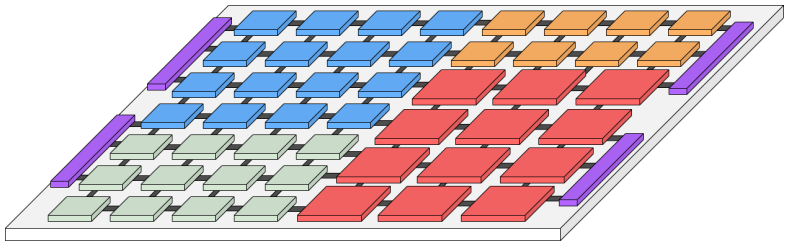
- Published at ACM Transactions on Embedded Computing Systems 2025 (presented at ESWEEK 2025)



Outline

- **Motivation: Chiplet-based platforms**
- **Preliminary Work-1:**
  - MFIT : Multi-Fidelity Thermal Modeling for 2.5D and 3D Multi-Chiplet Architectures
- **Preliminary Work-2:**
  - THERMOS: Thermally-Aware Multi-Objective Scheduling for Heterogeneous Multi-Chiplet PIM Architectures
- **Ongoing and Proposed Work:**
  - Disaggregated Acceleration of Hybrid Mamba–Transformer LLMs via Systolic Prefill and Vector Decode
  - Breaking the Memory Wall in MoE LLMs with Expert Prefetching
- **Timeline**
- **Conclusions**





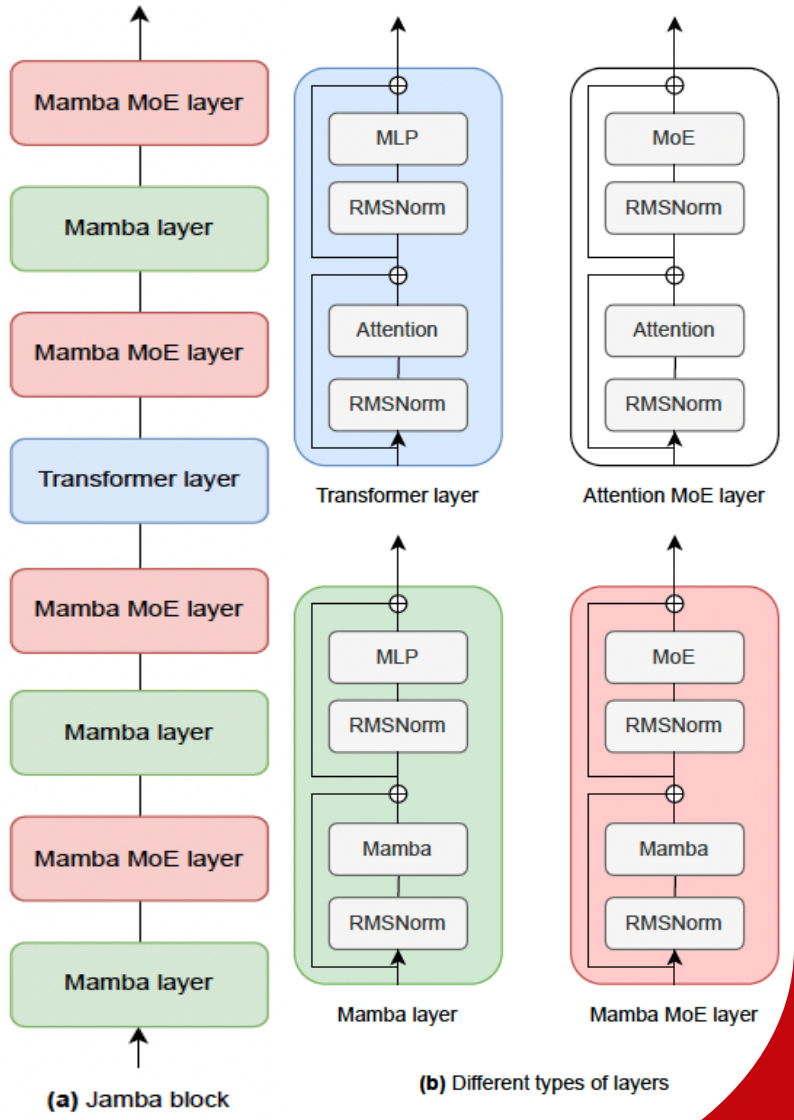
Heterogeneous chiplet system

Background: Hybrid LLMs

- **Transformers :** accurate context retrieval through attention
- **Mamba state-space models (SSM) :** long history in compact dense states

Growing trend of LLMs with a hybrid architecture[1-4] from industry

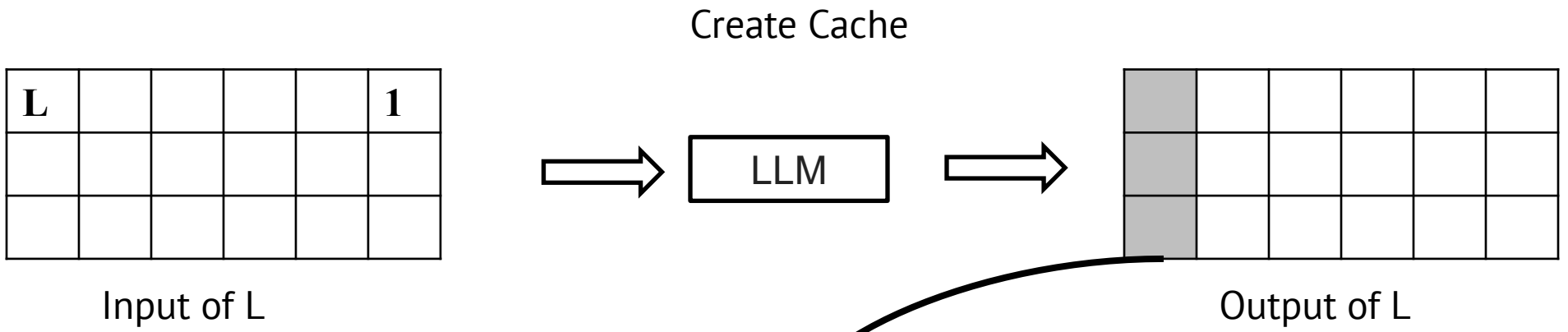
1. **Different compute-memory BW requirements for the prefill and decode stage**
2. **Hybrid LLMs have a different kernel compared to standard attention-based (matrix multiplication)**



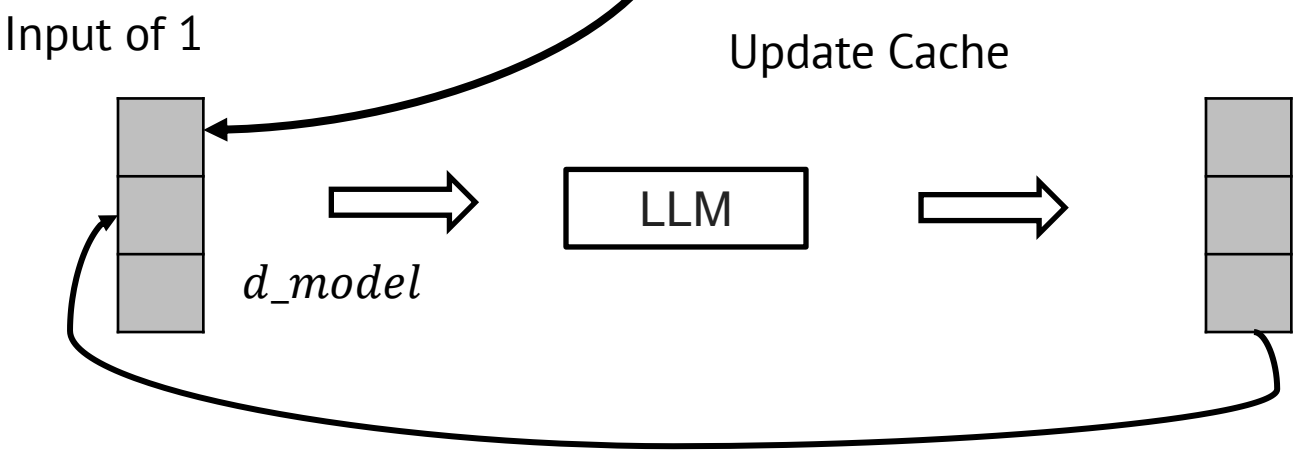
[1] Blakeman, Aaron, et al. "Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models." arXiv preprint arXiv:2504.03624 (2025).
[2] Lieber, Opher, et al. "Jamba: A hybrid transformer-mamba language model." *arXiv preprint arXiv:2403.19887* (2024).
[3] https://github.com/foundation-model-stack/bamba?tab=readme-ov-file
[4] Team, Tencent Hunyuan, et al. "Hunyuan-TurboS: Advancing Large Language Models through Mamba-Transformer Synergy and Adaptive Chain-of-Thought." arXiv preprint arXiv:2505.15431 (2025).

Background: LLM inference

- **Prefill**



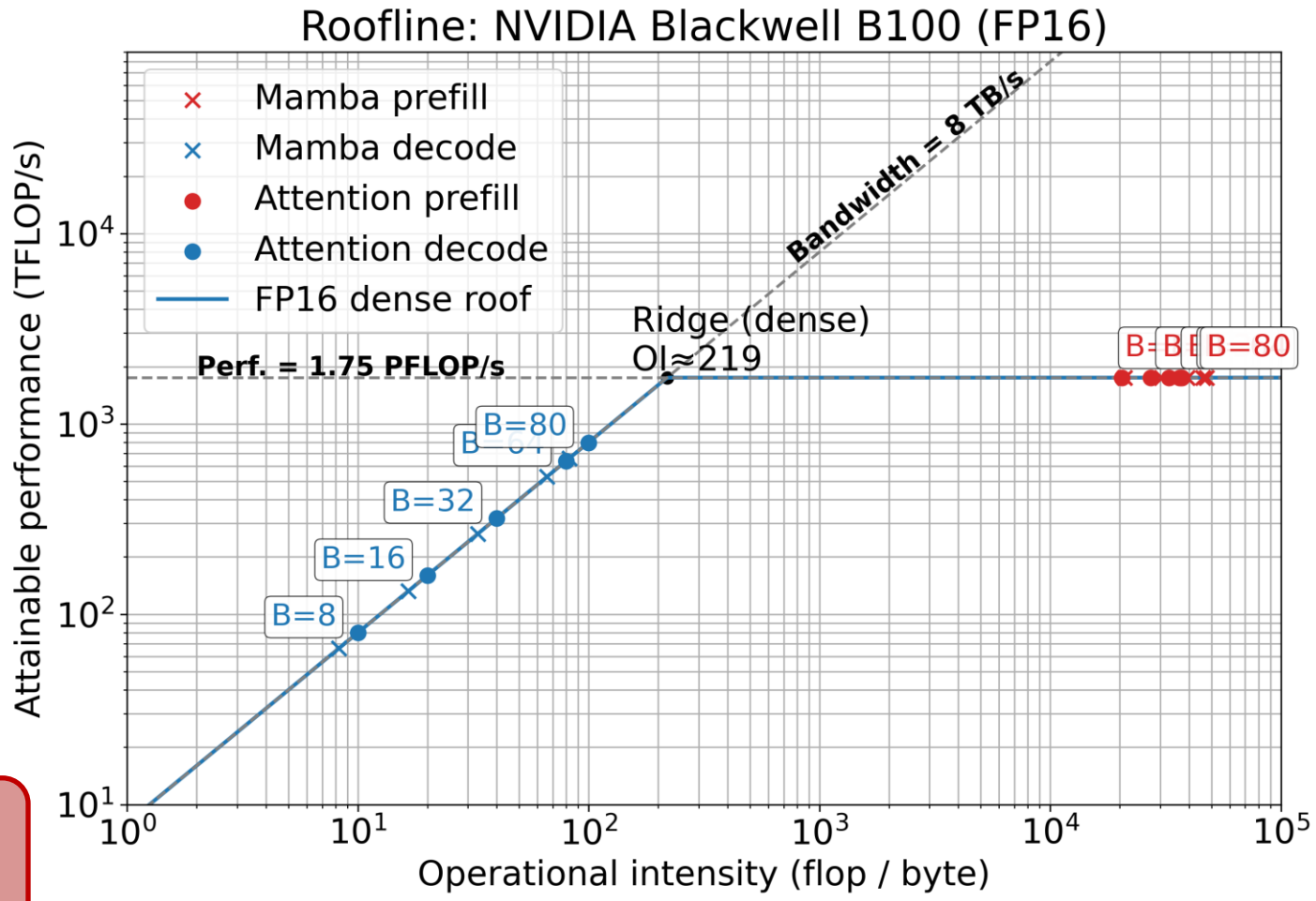
- **Decode**



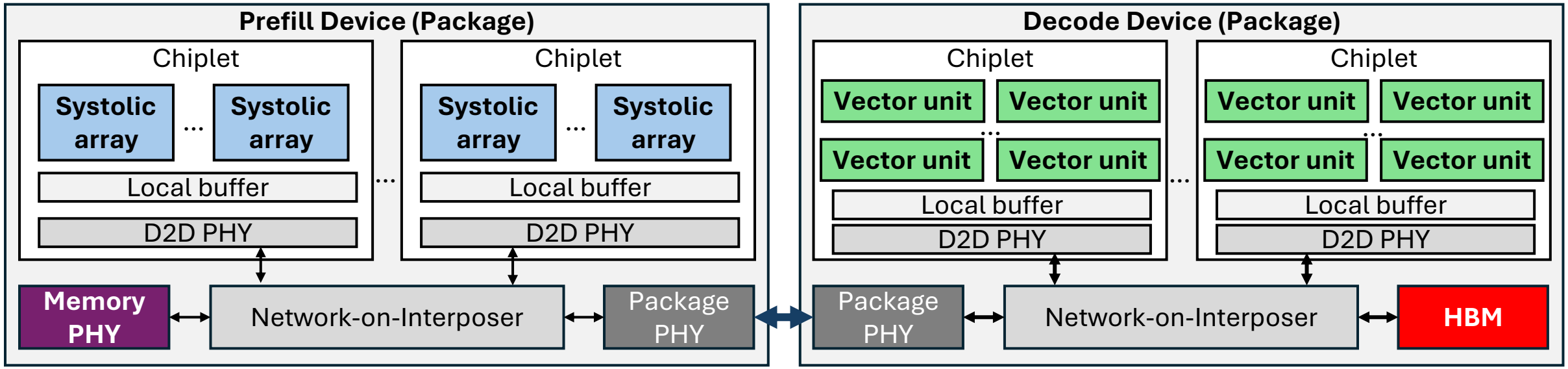
1. Different compute-memory BW requirements

- **Model: Nemotron-H-56B**
- **Seq. length : 4096**
- **Blackwell**
  - FP16: 1.75 PFLOPS
  - HBM bandwidth: 8 TB/s
- **Max batch size is 80 for B100**
  - Requires 184.09 GB
  - Total memory in B100 192GB

Prefill: Increase FLOPS
Decode: Increase BW



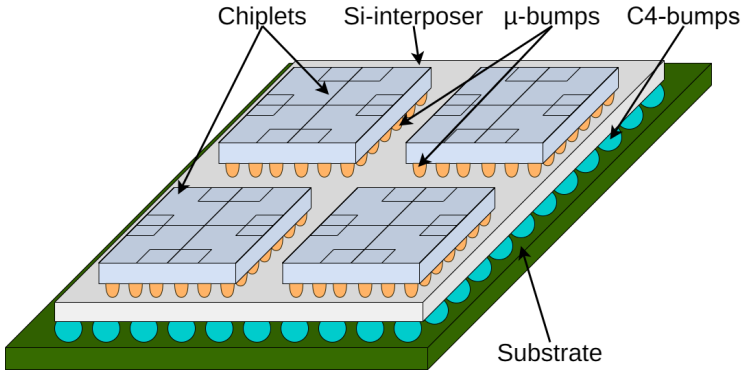
Proposed Solution : Disaggregated Acceleration

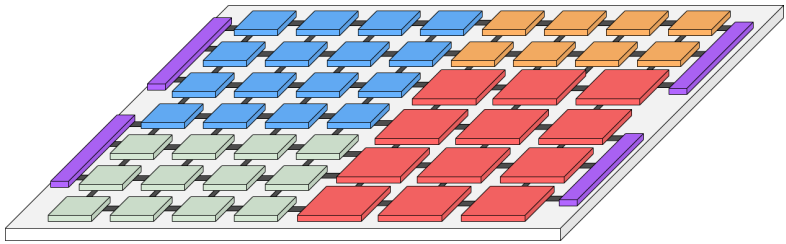


- **Prefill device: Chiplets with modified systolic array, No HBM**
- **Decode device: Chiplets with configurable vector unit + HBM**

Outline

- **Motivation: Chiplet-based platforms**
- **Preliminary Work-1:**
  - MFIT : Multi-Fidelity Thermal Modeling for 2.5D and 3D Multi-Chiplet Architectures
- **Preliminary Work-2:**
  - THERMOS: Thermally-Aware Multi-Objective Scheduling for Heterogeneous Multi-Chiplet PIM Architectures
- **Ongoing and Proposed Work:**
  - Disaggregated Acceleration of Hybrid Mamba–Transformer LLMs via Systolic Prefill and Vector Decode
  - Breaking the Memory Wall in MoE LLMs with Expert Prefetching
- **Timeline**
- **Conclusions**





Heterogeneous chiplet system

LLMs on Memory Limited Systems: MoE Prediction

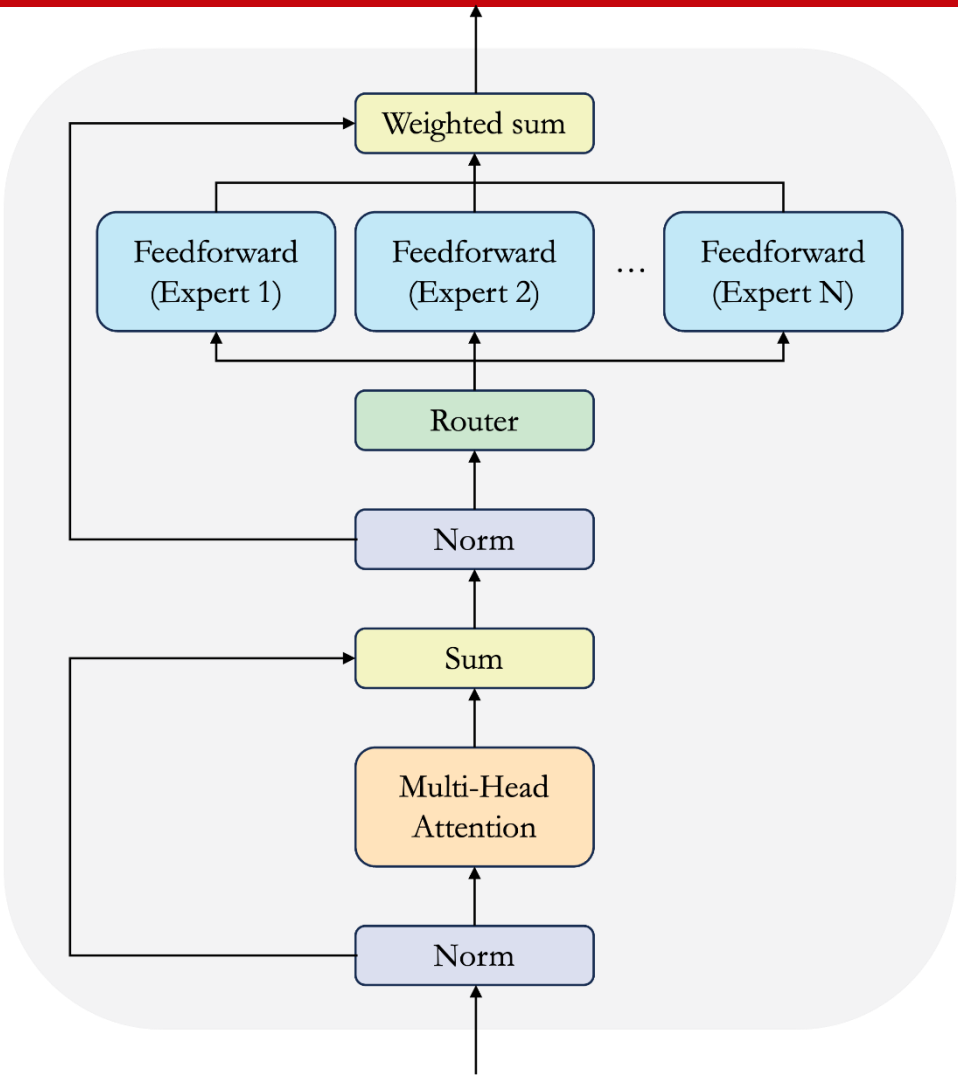
- **MoE provides an opportunity to have fewer active parameters**

Current implementation:

- **Experts are unknown until the router decides**
- **Weight loading from outside memory is an extra penalty**

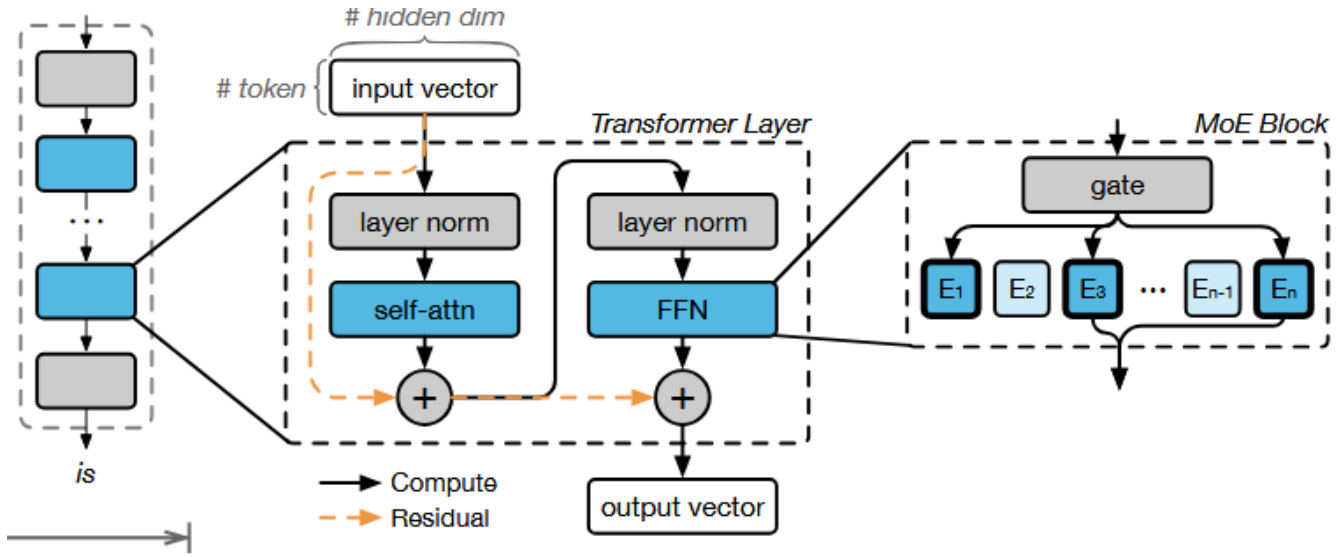
What if:

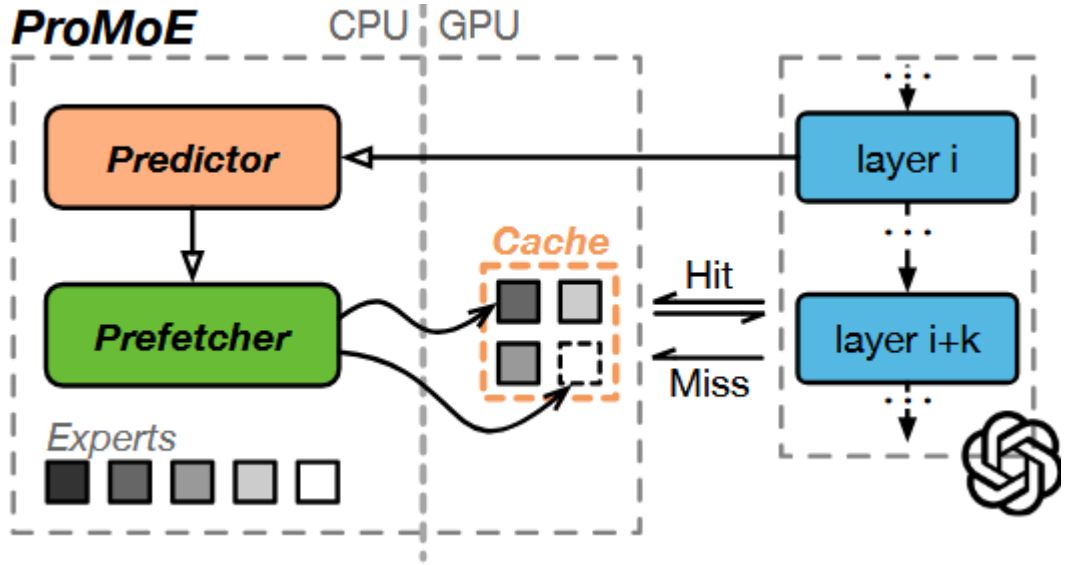
- **Experts for a given token can be known early**



Idea 1: Prefetch Top K + Δ Experts

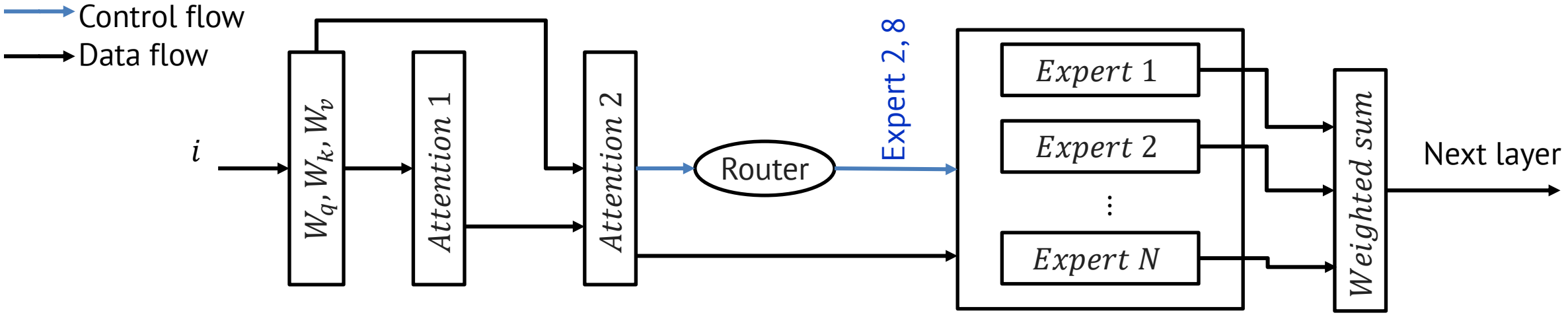
- **Assume that an LLM requires 4 experts per token**
- **Instead of predicting top 4 experts, predict top 6**
  - Higher prediction accuracy





Song, Xiaoniu, et al. "Promoe: Fast moe-based llm serving using proactive caching." arXiv preprint arXiv:2410.22134 (2025).

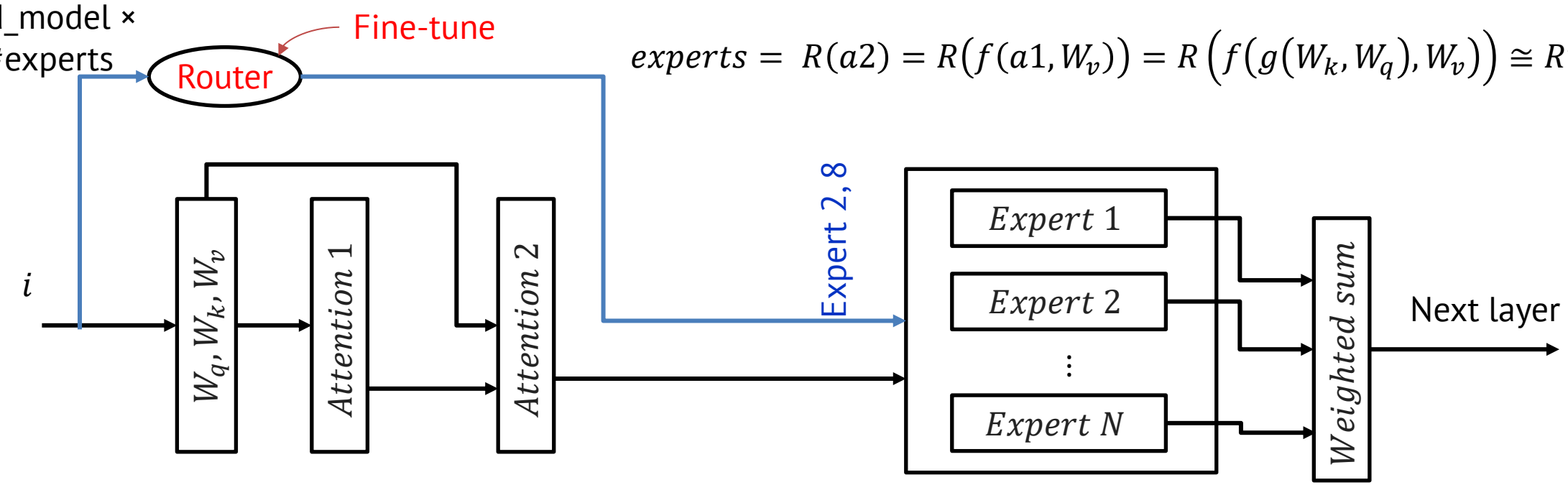
Idea 2: Re-structure the Router



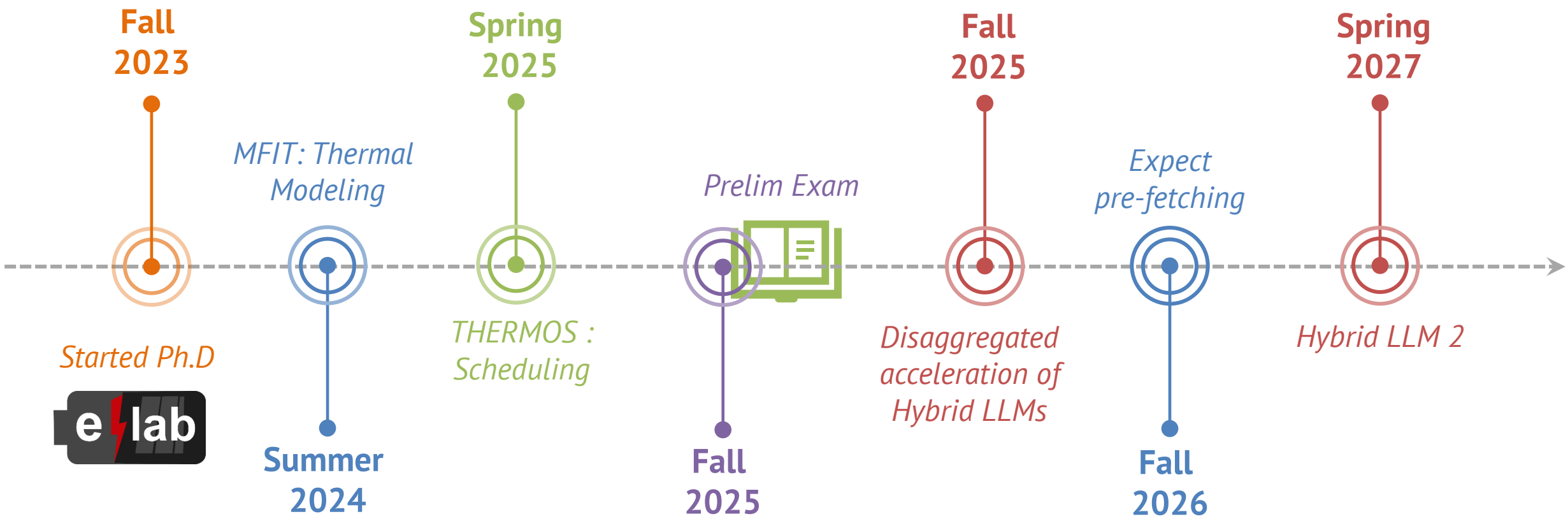
d_model × #experts

Fine-tune

$$experts = R(a2) = R\big(f(a1, W_v)\big) = R\Big(f\big(g(W_k, W_q), W_v\big)\Big) \cong R'(i)$$



Tentative Timeline



List of Publications

[1] Pfromm, L.*, **Kanani, A.***, Sharma, H., Solanki, P., Tervo, E., Park, J., Doppa, J., Pande, P.P. and Ogras, U., 2025. *"MFIT: Multi-fidelity thermal modeling for 2.5 D and 3D multi-chiplet architectures,"* ACM TODAES.

[2] Park, J., **Kanani, A.**, Pfromm, L., Sharma, H., Solanki, P., Tervo, E., Doppa, J.R., Pande, P.P. and Ogras, U.Y., 2024. *"Thermal modeling and management challenges in heterogenous integration: 2.5 D chiplet platforms and beyond,"* IEEE VTS. (INVITED)

[3] Goksoy, A.A., **Kanani, A.**, Chatterjee, S. and Ogras, U., 2024. *"Runtime Monitoring of ML-Based Scheduling Algorithms Toward Robust Domain-Specific SoCs,"* IEEE TCAD.

[4] **Kanani, A.**, Pfromm, L., Sharma, H., Doppa, J., Pande, P. and Ogras, U., 2025. *"THERMOS: Thermally-Aware Multi-Objective Scheduling of AI Workloads on Heterogeneous Multi-Chiplet PIM Architectures,"* ACM TECS.

[5] Kim, J., Lee, J., Lin, J., **Kanani, A.,** Miao, S., Ogras, U. and Park, J., 2025. *"eMamba: Efficient Acceleration Framework for Mamba Models in Edge Computing,"* ACM TECS.

Under review:

[1] Sharma, H., **Kanani, A.,** Doppa, J., Ogras, U. and Pande, P.P., 2025."*HeMu: Energy-Efficient DNN Inferencing via Heterogenous-Multi-Chiplet Architectures,*" IEEE TCAD.

[2] Pfromm, L., **Kanani, A.,** Sharma, H., Doppa, J., Pande, P. and Ogras, U., 2025. *"CHIPSIM: A Co-Simulation Framework for Deep Learning on Chiplet-Based Systems,"* IEEE OJ-SSCS.

[3] Sun, M., **Kanani, A.**, Shroff K. and Ogras U., 2025 *"ECLIPS - Entropy-Aware Exponent Compression for Efficient LLM Inference on Chiplet Systems,"* IEEE DATE.

*Equal contribution

Thank you!

e lab

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON