14

PhD FORUM: Cross-Layer Design of Heterogeneous Multi-Chiplet Systems for Efficient and Scalable AI Inference

Alish Kanani • ahkanani@wisc.edu

Advisor: Umit Y. Ogras | Expected graduation: July 2027 | University of Wisconsin-Madison

Efficient chiplet-based AI systems requires a cross-layer lifecycle:
MODEL physical limits → SPECIALIZE by workload phase → SCHEDULE to heterogeneous resources

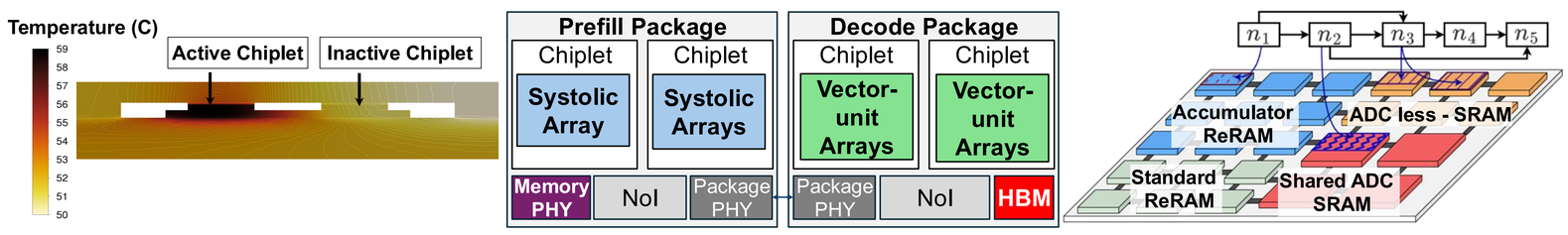
01 One Substrate, Three Bottlenecks

Chiplets scale AI compute and memory but move the design problem across layers

1. Thermal density and 2.5D/3D heat-flow paths limit sustained performance
2. Modern AI phases and kernels demand different compute and memory balance
3. Heterogeneous chiplets create latency-energy-temperature trade-offs

Research question
How should chiplet systems be modeled, specialized, and managed so AI performance keep scaling?

02 Dissertation Roadmap



Thermal modeling > phase-specialized acceleration > thermally-aware scheduling

03 MODEL | MFIT

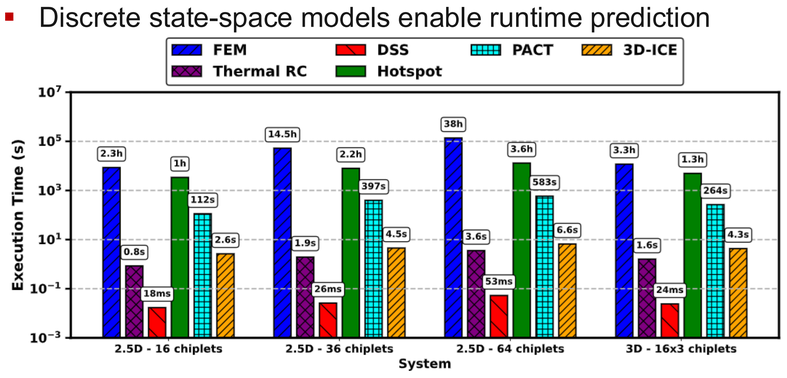
BEST PAPER AWARD

Make temperature visible at the fidelity each design stage can afford

Increasing speed → Increasing accuracy

	Finite Element Method (FEM) 1. Fine-grained	2. Abstracted	Analytical Models 3. Thermal RC Model	4. Discrete State-Space (DSS)
Features	Accurate (e.g., real μbumps)	micro-structures → material blocks	Independent of specific geometry	Specific to architecture
Error	Golden reference	< 0.5 °C	< 1.7 °C	Same as thermal RC*
Runtime	Not possible to model	Days	Seconds	Milliseconds
Use case	Validate the abstracted FEM models	Ground truth to tune Thermal RC model	Thermal aware design space exploration	Thermal management

- Fine-grained FEM anchors physical accuracy
- Abstract FEM preserves package-scale fidelity
- Thermal RC enables fast design-space exploration
- Discrete state-space models enable runtime prediction

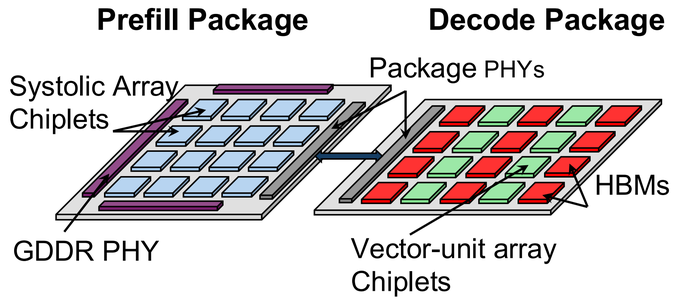


<1.7 C THERMAL RC ERROR | **10k-40k x** VS. FEM RUNTIME | **18-54** MS DSS

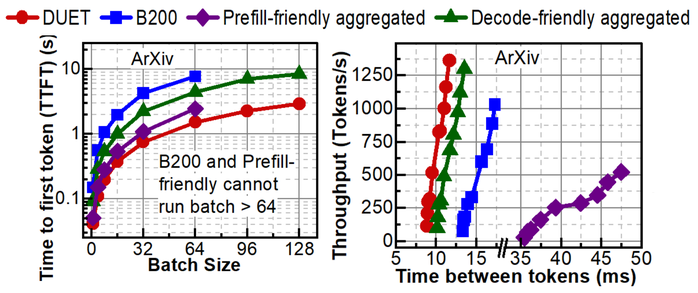
ROLE: Provides the thermal state used by design tools and runtime policies.

04 SPECIALIZE | DUET

Match each LLM phase with a package built around its dominant bottleneck



- Prefill: systolic-array chiplets + off-package GDDR
- Decode: vector-unit chiplets + in-package HBM
- Runtime-configurable attention and SSM kernels
- Chiplet counts and array sizes scale independently

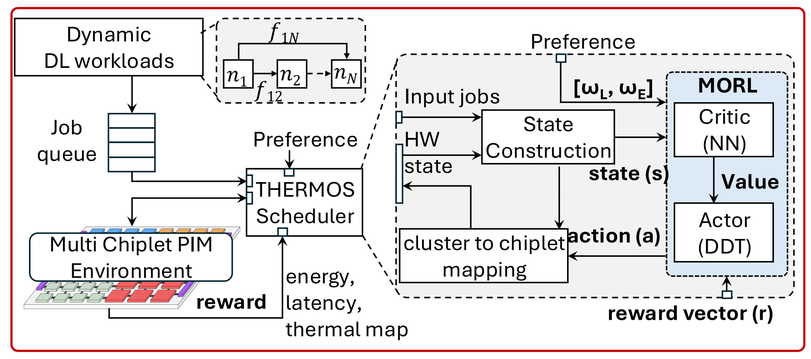


4 x FASTER TTFT | **1.4 x** HIGHER THROUGHPUT | **1.5 x** LOWER TBT

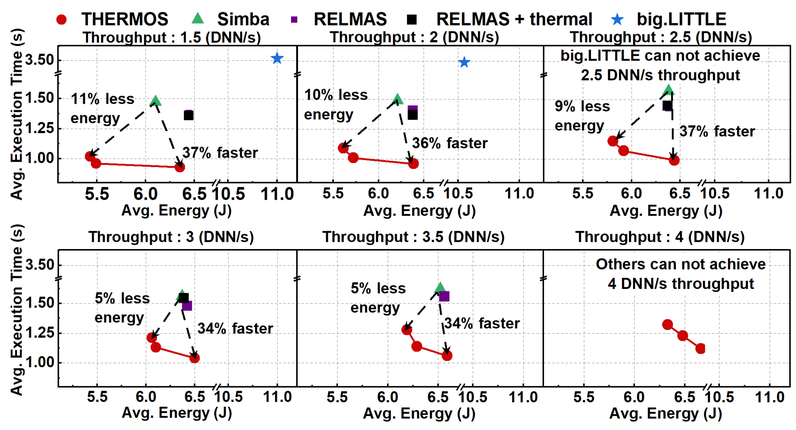
ROLE: converts workload structure into package-level compute and memory specialization.

05 SCHEDULE | THERMOS

Turn thermal visibility into adaptive decisions across heterogeneous resources



- Multi-Objective RL maps each layer to a PIM cluster
- Runtime preference vector selects latency, energy, or balance
- Proximity-driven stage minimizes inter-chiplet communication
- Thermal constraints prevent unsafe chiplet assignments



89% FASTER EXECUTION | **57%** LOWER ENERGY | **0.14%** RUNTIME OVERHEAD

ROLE: exploits heterogeneity while enforcing temperature and runtime preferences

06 One Cross-Layer Design Loop

MODEL → **SPECIALIZE** → **SCHEDULE**

Thermal models expose limits • specialized packages match phase bottlenecks • runtime scheduling adapts heterogeneous resources

What changes at each step ?

Thermal Visibility
Temperature becomes a quantitative design signal instead of a late-stage constraint

Architecture Specialization
Chiplet packages scale compute, memory, and interconnect around workload-specific demand

Runtime Adoption
One policy reacts to workload mix, thermal headroom, and changing user objectives

DISSERTATION IMPACT

A reusable methodology for efficient AI inference across the chiplet lifecycle - from thermal modeling, through workload-specialized hardware, to adaptive runtime scheduling

- Scales across 2.5D and 3D integration
- Supports heterogeneous technologies and objectives
- Connects physical constraints to system performance
- Targets both conventional DNNs and emerging hybrid LLMs

07 Broader Research

THERMAL + SIMULATION
MFIT (TODAES'25) • CHIPSIM (OJSSC'25) • HYDRA (TCAD'26) • Thermoelectric Cooler (Underreview)

AI ARCHITECTURES
DUET (DAC'26) • HeMu (TCAD'25) • eMamba (TEC'25) • LEXI (DAC'26)

RUNTIME CONTROL
THERMOS (TECS'25) • robust ML-based scheduling (TCAD'24) • ReVolt (TCAD'26) • APEX (TCAD'26)

BOTTOM LINE
Chiplets are not only a packaging solution. They are a cross-layer opportunity to co-design thermal behavior, AI acceleration, and runtime control.

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON | e lab | 63 THE CHIPS TO SYSTEMS CONFERENCE SHAPING THE NEXT GENERATION OF ELECTRONICS