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

Prelim Summary

Alish Kannai
University of Wisconsin Madison
Advisor: Umit Y. Ogras
Email: ahkanani@wisc.edu

Track: System-level Design, Synthesis and Optimization

Dissertation Abstract

Motivation for Chiplet-Based AI Systems: AI inference is increasingly dominated by (i) model parameters and activation footprints that strain memory capacity and bandwidth, and (ii) deployment demands for both high throughput and low latency. Building ever-larger monolithic accelerators to sustain this demand is becoming impractical due to reticle limits and yield degradation, motivating chiplet-based integration as the main path to *scale-out* compute and memory within a single system. Chiplets enable modular composition of compute, memory, I/O, and accelerators, while allowing large systems by integrating smaller dies with higher yield and lower cost. As a result, multi-chiplet 2.5D/3D packages are rapidly emerging as a key substrate for next-generation AI platforms.

Realizing this opportunity, however, requires addressing several fundamental research challenges. As the package size increases, temperature becomes a system-level constraint. At the same time, heterogeneous chiplet-based systems introduce competing trade-offs in performance, energy, and thermal behavior. Moreover, modern AI workloads are increasingly non-uniform across phases and kernels, demanding architectures specialized for distinct execution bottlenecks. Together, these challenges require a cross-layer approach that connects modeling, runtime management, and architecture design.

The first challenge is **thermal** bottleneck. 2.5D/3D integration increases total active silicon per package, raising system power; if power grows faster than area, package-level power density can exceed that of monolithic designs [1]. As shown in Figure 1(a), heat-flow paths in chiplet systems differ qualitatively from monolithic chips: heat spreads laterally through the interposer and heat spreader, while dense die-to-die links introduce additional Joule heating and thermal crosstalk. In 3D stacks, vertical coupling further amplifies hotspots. These effects make *temperature a first-class constraint that can directly limit performance and reliability*, forcing throttling and reducing achievable throughput if not handled proactively.

The second challenge is **heterogeneity-aware optimization**. The full benefit of chiplet integration emerges when the package combines specialized chiplets [2]. Figure 1(b) illustrates a heterogeneous processing-in-memory (PIM) system that combines ReRAM- and SRAM-based chiplets, whose trade-offs differ in execution time, energy, density, and temperature sensitivity [3]. Yet heterogeneity complicates runtime management: workloads have different compute and communication characteristics, chiplets have different thermal limits and performance-energy profiles, and system objectives are inherently multi-dimensional. Static mapping and single-objective optimization are therefore insufficient; maximizing system utility requires **multi-objective scheduling** that adapts to runtime preferences while enforcing thermal constraints.

The third challenge is **workload and phase specialization**. LLM inference is fundamentally asymmetric: the prefill phase is highly parallel and compute-bound, while the decode phase is sequential and memory-bandwidth-bound [4]. Hybrid Mamba–Transformer models preserve this prefill/decode asymmetry while adding state-space-model (SSM) recurrences and element-wise operations that map poorly to matmul-centric accelerators. System-level disaggregation is common in software frameworks, but homogeneous hardware still underutilizes bandwidth during prefill and compute during decode. This motivates a hardware-level shift toward **multi-package, phase-specialized systems** in which each phase runs on a package optimized for its dominant bottleneck, while remaining flexible enough to support mixed kernels.

Dissertation thesis. This dissertation argues that efficient AI inference on chiplet-based systems requires a *cross-layer* methodology that addresses: (i) scalable thermal modeling to make temperature *visible* across design and runtime; (ii) multi-objective scheduling to *exploit heterogeneity* under competing objectives; and (iii) workload- and phase-specialized chiplet architectures for emerging AI models.

Contribution 1 – MFIT. The first contribution establishes the modeling foundation through *MFIT, a multi-fidelity thermal modeling framework for 2.5D and 3D architectures* [1]. No single thermal technique simultaneously provides sign-off accuracy, architectural exploration speed, and runtime suitability. MFIT addresses this gap by systematically abstracting fine-grained FEM models into a family of faster models, including thermal RC and discrete-time state-space models, while preserving physical consistency across fidelities [1]. On systems with 16, 36, and 64 chiplets in 2.5D and 16×3 stacked chiplets in 3D, MFIT reduces evaluation time from days to seconds or milliseconds with negligible loss in accuracy [1].

Contribution 2 – THERMOS. Building on MFIT's thermal modeling, the second contribution turns modeling into control via *THERMOS, a thermally-aware, multi-objective scheduling framework for AI workloads on heterogeneous multi-*

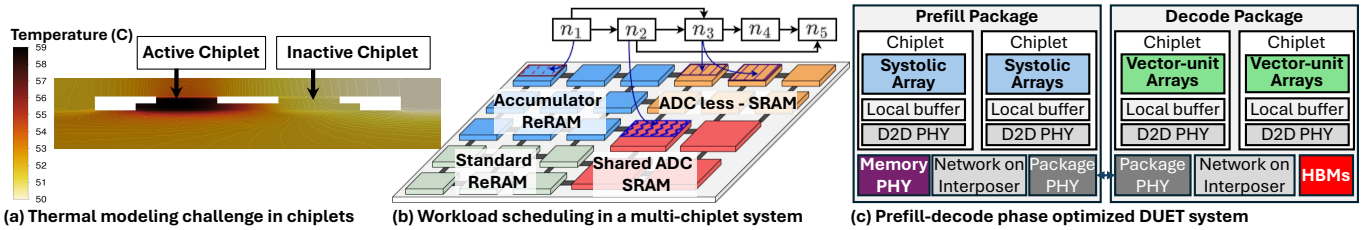

Fig. 1. Dissertation overview: (a) MFIT for thermal modeling, (b) THERMOS for thermally-aware scheduling on heterogeneous multi-chiplet PIM systems, and (c) DUET for disaggregated prefill/decode acceleration of hybrid LLMs.

chiplet PIM architectures [3]. Heterogeneous PIM chiplets offer complementary strengths but also introduce competing objectives. THERMOS formulates scheduling as a constrained multi-objective optimization, learning a *single multi-objective reinforcement learning* policy that accepts a preference vector at runtime and produces a Pareto-optimal policy for latency, energy, or a balanced trade-off, while ensuring chiplet temperatures remain below specified limits [3]. Comprehensive evaluations show that THERMOS achieves faster average execution time and lower average energy consumption than baseline schedulers with negligible overhead [3].

Contribution 3 – DUET. The third contribution extends the cross-layer methodology to emerging LLM inference via *DUET, a disaggregated accelerator for hybrid Mamba–Transformer LLMs* with prefill- and decode-specific packages [4]. DUET leverages the prefill/compute vs. decode/bandwidth asymmetry to design two heterogeneous packages: a compute-oriented prefill package using systolic-array chiplets with off-package memory for GEMM-heavy execution, and a bandwidth-optimized decode package using vector-unit chiplets with in-package HBMs for token-by-token execution. Since hybrid models interleave attention and SSM blocks, DUET introduces runtime-configurable microarchitectural features to enable each package to support both kernel families. Evaluations across multiple models and workloads show that DUET reduces time-to-first-token (TTFT) and improves decode throughput while reducing time-between-tokens.

Overall impact. Together, MFIT, THERMOS, and DUET establish a coherent dissertation in which chiplet-based AI systems are designed and managed with explicit awareness of thermal behavior, heterogeneity, and workload structure. Supporting co-authored works strengthen this agenda by improving robustness of ML-based schedulers [5], enabling energy-efficient heterogeneous chiplet architectures [2], providing accurate co-simulation and transient thermal evaluation [6], optimizing Mamba execution on edge devices [7], and reducing inter-chiplet communication overhead in hybrid LLMs [8].

Related Publications

- **MFIT** [1]: multi-fidelity thermal models for 2.5D/3D multi-chiplet systems spanning FEM abstractions, thermal RC, and discrete state-space models.
- **THERMOS** [3]: thermally-aware multi-objective-reinforcement-learning scheduler for AI workloads on heterogeneous multi-chiplet PIM architectures with runtime preference control.
- **DUET** [4]: disaggregated hybrid Mamba–Transformer accelerator with phase-specialized packages and runtime-configurable microarchitectures.
- **Runtime Monitoring of ML-Based Scheduling Algorithms Toward Robust Domain-Specific SoCs** [5]: runtime monitoring and adaptation to preserve robustness of ML-based scheduling under unseen workloads.
- **eMamba** [7]: end-to-end hardware acceleration framework for efficient deployment of Mamba models in edge computing.
- **HeMu** [2]: design-space exploration for energy-efficient heterogeneous multi-chiplet DNN inference architectures.
- **CHIPSIM** [6]: co-simulation framework that jointly models compute, communication, and transient thermal behavior in chiplet-based DNN systems.
- **LEXI** [8]: lossless exponent coding that reduces inter-chiplet communication cost in hybrid LLM inference.

References

[1] L. Pfromm, A. Kanani, H. Sharma, P. Solanki, E. Tervo, J. Park, J. R. Doppa, P. P. Pande, and U. Y. Ogras, "MFIT: Multi-Fidelity Thermal Modeling for 2.5d and 3d Multi-Chiplet Architectures," *ACM Transactions on Design Automation of Electronic Systems*, 2025.

[2] H. Sharma, A. Kanani, J. R. Doppa, U. Y. Ogras, and P. P. Pande, "HeMu: Energy-Efficient DNN Inferencing via Heterogeneous-Multi-Chiplet Architectures," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2025.

[3] A. Kanani, L. Pfromm, H. Sharma, J. R. Doppa, P. P. Pande, and U. Y. Ogras, "THERMOS: Thermally-aware Multi-Objective Scheduling of AI Workloads on Heterogeneous Multi-Chiplet PIM Architectures," *ACM Transactions on Embedded Computing Systems*, 2025.

[4] A. Kanani, S. Lee, H. Lyu, J. Lin, J. Park, and U. Y. Ogras, "DUET: Disaggregated Hybrid Mamba-Transformer LLMs with Prefill and Decode-Specific Packages," at DAC 2026.

[5] A. A. Goksoy, A. Kanani, S. Chatterjee, and U. Y. Ogras, "Runtime Monitoring of ML-Based Scheduling Algorithms Toward Robust Domain-Specific SoCs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 11, pp. 4202–4213, 2024.

[6] L. Pfromm, A. Kanani, H. Sharma, J. R. Doppa, P. P. Pande, and U. Y. Ogras, "CHIPSIM: A Co-Simulation Framework for Deep Learning on Chiplet-Based Systems," *IEEE Open Journal of the Solid-State Circuits Society*, 2025.

[7] J. Kim, J. Lee, J. Lin, A. Kanani, M. Sun, U. Y. Ogras, and J. Park, "eMamba: Efficient Acceleration Framework for Mamba Models in Edge Computing," *ACM Transactions on Embedded Computing Systems*, 2025.

[8] M. Sun, A. Kanani, K. Shroff, and U. Ogras, "LEXI: Lossless Exponent Coding for Efficient Inter-Chiplet Communication in Hybrid LLMs," at DAC 2026.