Design Space Exploration and Runtime Optimization of Reconfigurable Systems for Dynamic Workloads

DAC PhD Forum 2026

Jiahao Lin
Electrical and Computer Engineering Department
University of Wisconsin-Madison, USA
jlin445@wisc.edu
Est. Graduation: Summer 2027
Prev. ASP-DAC/DATE PhD Forum: No

Advisor: Umit Y. Ogras
Electrical and Computer Engineering Department
University of Wisconsin-Madison, USA
uogras@wisc.edu

I. Abstract

Motivation. Emerging computing workloads are increasingly data-intensive and exhibit dynamic, irregular behavior at run-time, challenging the efficiency of conventional accelerator designs and scheduling techniques. For example, modern wireless communication applications, such as dynamic spectrum access (DSA), must cope with unknown spectrum environments, varying sample rates, and diverse transmitter algorithms, which together cause rapid waveform-level signal variations at run-time [1]. Large language model (LLM) services exhibit highly variable workloads due to differing sequence lengths, request arrival rates, and the distinct computational characteristics between prefill and decode phases. Emerging agentic scenarios further amplify this unpredictability through function calls and external tool usage.

Executing such applications efficiently demands hardware that reconfigures at run-time to match shifting compute, memory, and bandwidth demands. Reconfigurable and adaptive systems, such as Coarse-Grained Reconfigurable Arrays (CGRAs) and heterogeneous chiplet-based systems, have emerged as promising candidates, offering ASIC-like efficiency while enabling run-time adaptation. However, ad-hoc hardware configurations and traditional static scheduling techniques struggle to cope with run-time irregularity and dynamic control flow. This motivates a holistic approach that combines *design-time* hardware design space exploration (DSE) with *run-time* dynamic scheduling and resource management to fully exploit the potential of reconfigurable systems.

Contributions. This dissertation develops design space exploration and runtime optimization methodologies for reconfigurable systems targeting two representative dynamic workloads: DSA and hybrid LLM serving. The central methodology combines design-time architectural optimization with run-time adaptive scheduling to maximize throughput and minimize latency. Specifically, the contributions are:

- A unified methodology that couples design-time DSE with run-time dynamic scheduling, enabling reconfigurable systems to adapt to workload irregularity without sacrificing efficiency.
- For *DSA*, we design optimized CGRA architectures and develop communication-aware runtime schedulers that handle irregular, high-throughput spectrum sensing workloads.
- For *hybrid LLM serving*, we explore heterogeneous chiplet-based accelerators with runtime-adaptive scheduling to efficiently serve Transformer-Mamba models under dynamic request patterns.
- A evaluation framework that quantifies the combined benefit of architectural optimization and runtime scheduling on realistic, time-varying workloads.

II. Supporting Paper

The primary supporting paper, *"CADAS: Communication-Aware Dynamic Scheduler on CGRAs for Large-Volume and Real-Time Processing"* [2], presents a holistic methodology for designing optimized domain-specific CGRA architectures and efficiently utilizing them at run-time for data-intensive, dynamic applications. As illustrated in Fig. 1, CADAS identifies four primary performance bottlenecks on a CGRA-based system: reconfiguration/switch time, processing power, data communication bandwidth, and input data bandwidth. They guide both design-time DSE and run-time decisions.

Design-Time Optimization. CADAS first conducts a systematic DSE on hardware configurations tailored to application-specific constraints, including area budget, throughput requirement, and throughput efficiency. The DSE explores key architectural parameters: system scaling and tiling (monolithic vs. modular with multiple arrays), PE array geometry (aspect ratio), and distributed memory sizing. We identify four primary performance bottlenecks that guide the exploration. The optimized configurations achieve a $2.8\times$ performance gain over the default monolithic design, establishing a fair baseline for evaluating run-time scheduling.

Run-Time Optimization. Building on the optimized architecture, CADAS proposes a communication-aware dynamic scheduling approach. Two key observations motivate the scheduler design: (1) the arrangement of on-chip memory critically determines data flow and bisection bandwidth, and

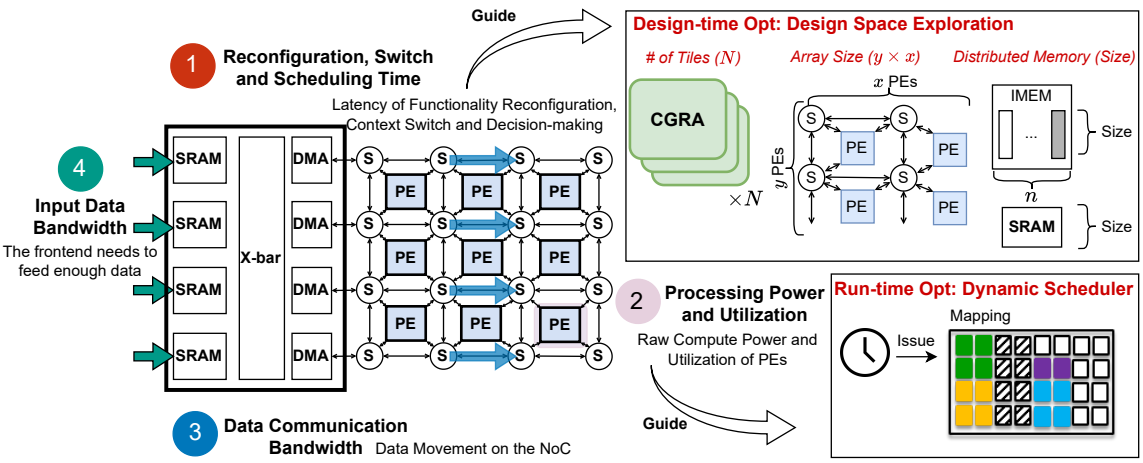

Fig. 1. Overview of CADAS [2]: four performance bottlenecks of CGRA-based systems guide the optimization that couples *design-time* DSE (tile count, array geometry, and distributed memory sizing) with a *run-time* dynamic scheduler to sustain high throughput under dynamic workloads.

(2) interconnect I/O congestion is a dominant bottleneck in data-intensive streaming applications. The scheduler integrates a scoreboard and preloading mechanism within a hardware/software (HW/SW) co-design framework. The scoreboard, implemented as a content-addressable memory, enables fast identification of reusable kernel configurations to minimize hard context switches. When a new placement is needed, a communication-aware dynamic placer minimizes inter-kernel communication latency by optimizing kernel placement along the data-flow direction, maximizing interconnect bandwidth utilization. The preloading mechanism speculatively loads successor kernels based on the application's analysis tree structure, further reducing reconfiguration overhead.

Evaluation. The proposed methods are evaluated on real-time spectrum sensing benchmarks comprising 48 sub-band scenarios with three complexity levels. The scheduling method achieves up to 1.6× throughput improvement over a baseline and 1.3× over an adapted state-of-the-art (SOTA) dynamic scheduling strategy on the DSE-optimized configurations.

III. Other Works

Beyond CADAS, several of my other works contribute to the dissertation theme across both target domains.

Dynamic Spectrum Access.

- We provide a comprehensive overview of the challenges and system requirements for real-time spectrum sensing in modern RF autonomy systems [1], identifying the key demands on reconfigurable hardware: fast switching at waveform timescales, real-time adaptation for resource management, high throughput for dense wideband spectrum, and broad application scope across diverse modulation types. This survey motivates the architectural and scheduling co-design pursued in CADAS.
- K-PACT [3] complements CADAS with a framework for kernel clustering, placement, and prefetching that exploits the hierarchical structure of spectrum sensing analysis trees. By grouping frequently co-occurring kernels and proactively prefetching successor configurations, K-PACT further reduces reconfiguration overhead beyond what the CADAS scoreboard achieves alone.
- We also investigate reconfigurable system design trade studies for dynamic spectral sensing [4], analyzing the interplay between hardware architecture choices and dynamic workload characteristics across different system scales.

Hybrid LLM Serving.

- DUET [5] addresses hybrid Transformer-Mamba LLM acceleration at the micro-architectural level by designing disaggregated accelerator packages with prefill- and decode-optimized chiplets, tailored to the distinct computational and memory access characteristics of each inference phase.
- eMamba [6] presents an efficient acceleration framework for Mamba models in edge computing, contributing to our understanding of state-space model (SSM)-specific hardware optimizations.

References

[1] J. Lin, H. U. Suluhan, H. Chung, A. Dutta, A. Vipperla, G. Gubash, et al. "An Overview of Challenges and Requirements for Real-Time Spectrum Sensing in Modern RF Autonomy Systems". In: *IEEE Design & Test* (2025).

[2] J. Lin, H. U. Suluhan, C. Chakrabarti, A. Akoglu, and U. Ogras. "CADAS: Communication-Aware Dynamic Scheduler on CGRAs for Large-Volume and Real-Time Processing". In: *ACM Transactions on Embedded Computing Systems* 25.2 (2026).

[3] H. U. Suluhan, J. Lin, S. Gener, C. Chakrabarti, U. Ogras, and A. Akoglu. "K-PACT: Kernel Planning for Adaptive Context Switching–A Framework for Clustering, Placement, and Prefetching in Spectrum Sensing". In: *Proceedings of the 44th IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. 2025, pp. 1–9.

[4] A. Dutta, J. Lin, H. U. Suluhan, G. Gubash, A. Akoglu, U. Y. Ogras, et al. "Reconfigurable System Design and Trade Studies for Dynamic Spectral Sensing". In: *2025 59th Asilomar Conference on Signals, Systems, and Computers*. 2025, pp. 1–6.

[5] 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". In: *Proceedings of the 63rd ACM/IEEE Design Automation Conference (DAC)*. 2026.

[6] J. Kim, J. Lee, J. Lin, A. Kanani, M. Sun, U. Y. Ogras, et al. "eMamba: Efficient Acceleration Framework for Mamba Models in Edge Computing". In: *ACM Transactions on Embedded Computing Systems* 24.5s (2025), pp. 1–22.