Divya Kiran Kadiyala

Website: dkadiyala3.github.io | Email: dkadiyala3@gatech.edu

Summary

PhD candidate with a strong research background in computer architecture and memory system design. My research focuses on developing and evaluating architecture-level and memory-system optimizations, including CXL-enabled hierarchies, to improve the performance and scalability of AI/ML, HPC, and cloud workloads. Experienced in performance modeling, distributed AI/ML modeling, and processor architecture analysis.

Education

Georgia Institute of Technology (Georgia Tech) ..... 2019 – 2025
Doctor of Philosophy (PhD), Electrical and Computer Engineering
Thesis: Memory system optimizations for parallel and bandwidth-intensive applications
Advisor: Prof. Alexandros Daglis

Arizona State University ..... 2015 – 2017
Master of Science in Engineering (MSE), Electrical Engineering
Specialization: VLSI and Mixed Signal Circuits

KL University, Guntur, A.P., India ..... 2009 – 2013
Bachelor of Technology (BTech), Electronics & Communications Engineering

Research & Work Experience

School of Computer Science, *Graduate Research Assistant, Georgia Tech* ..... May 2020 – Dec 2025
Thesis: Exploring memory system optimizations to accelerate parallel and bandwidth-intensive workloads

1. Acceleration of HPC workloads on Hardware Transactional Memory (HTM) [*HPCA'23*]
   - **Problem**: Capacity aborts leads to severe performance loss reducing the utility of commercial HTMs.
   - **Insight**: Track only the critical memory accesses to increase the effective capacity of on-chip HTM buffers.
   - **Contribution**: Developed HinTM, a novel hardware–software co-design technique that leverages software hints to track only critical memory accesses within a transaction.
   - **Result**: Achieved up to 8.7× speedup over baseline HTM by eliminating 64% of transactional capacity aborts.
2. Acceleration of memory-bound server workloads using I/O bandwidth harvesting [*under peer review*]
   - **Problem**: Reduced per-core bandwidth in server CPUs leads to high memory latency and queuing delays.
   - **Insight**: Augment the memory bandwidth by dynamically harvesting idle I/O bandwidth in many-core CPUs.
   - **Contribution**: Developed SURGE, a software-assisted architectural mechanism that opportunistically harvests unused I/O bandwidth to access additional memory via high-speed serial links, such as CXL.
   - **Result**: Reduced memory queuing delay by 33% and achieved up to 1.5× speedup over a DDR-only baseline.
3. Acceleration of distributed AI training workloads on large-scale clusters [*arXiv*]
   - **Problem**: Increased model sizes force cluster scale-out and cause slowdowns from communication bottlenecks.
   - **Insight**: Expanding memory capacity in training nodes limits scale-out requirements improving throughput.
   - **Contribution**: Developed COMET, a holistic cluster design methodology for rapid design-space co-exploration to evaluate the impact of memory expansion on distributed deep learning training performance.
   - **Result**: Identified viable designs for memory expansion to improve training throughput for LLMs and DLRMs.

Hewlett Packard Labs, Milpitas, CA, *Research Associate Intern* ..... Summer 2024 – Present
Supervisors: Dr. Puneet Sharma & Dr. Lianjie Cao
Project: Improving memory efficiency and scalability of AI/ML Training using CXL

- Designed composable AI/ML training architectures leveraging disaggregated memory expansion techniques.
- Boosted collective communication via CUDA & ROCm optimizations tailored to algorithm–topology co-design.

Samsung MSL Lab, San Jose, CA, *Systems Technology Research Intern* ..... Summer 2022
Supervisor: Michael Choi
Project: CXL enabled Memory Pooling solutions for HPC and cloud infrastructure

- Designed and evaluated novel memory fabric topologies based on the CXL 3.0 specification
- Built performance models to analyze the impact of memory expansion on HPC, AI/ML, and Cloud workloads.

Luminous Computing, Mountain View, CA, *CPU Architecture Intern* ..... Summer 2021
Supervisor: Dr. Muhammad Tauseef Rab
Project: Developed performance models for RISC-V based microarchitectures

- Built SystemC transactional models to evaluate microarchitecture enhancements in novel RISC-V processors.

Cadence Design Systems, San Jose, CA, *Sr. Applications Engineer* ..... Jul 2017 – Jul 2019
Supervisor: Kumar Subramani
Worked on development of novel EM/IR & Power Sign-off solutions

- Developed and automated Power & EM/IR flows for Power-Grid View (PGV) validation and design sign-off.
- Collaborated with R&D and customers to validate new Voltus IR/EM features and resolved critical design issues.

Selected Publications

1. **Harvesting idle I/O resources for boosting memory bandwidth** [*under peer review*]
   D. K. Kadiyala, and A. Daglis
2. **Enabling Flexible and Composable AI Systems via Memory Disaggregation** [*under peer review*]
   D. K. Kadiyala, L. Cao, P. Sharma, S. Sury, and A. Daglis
3. **Geode: A Zero-shot Geospatial Question-Answering Agent with Explicit Reasoning and Precise Spatio-Temporal Retrieval**
   D. Gupta, A. Ishaqui, and D. K. Kadiyala
   ISCA Workshop Emerging Vision and Graphics System and Architectures (EVGA), June 2024
4. **COMET: A Comprehensive Cluster Design Methodology for Distributed Deep Learning Training**
   D. K. Kadiyala, S. Rashidi, T. Heo, A. R. Bambhaniya, T. Krishna, and A. Daglis
   preprint arXiv, 2022
5. **Safety Hints for HTM Capacity Abort Mitigation**
   A. Jain*, D. K. Kadiyala*, and A. Daglis
   High-Performance Computer Architecture (HPCA), 2023. Acceptance rate: 25.0%
   * Equal Contribution
6. **Exploring Memory Expansion Designs for Training Mixture-of-Experts Models**
   T. Heo, S. Rashidi, C. Man, D. K. Kadiyala, W. Won, S. Srinivasan, M. Elavazhagan, M. Kumar, A. Daglis, and T. Krishna
   Workshop on Hot Topics in System Infrastructure, (HotInfra), June 2023
7. **Physically Unclonable Functions Using Foundry SRAM Cells**
   L. T. Clark, S. B. Medapuram, D. K. Kadiyala, and J. Brunhaver
   IEEE Transactions on Circuits and Systems I (TCAS), 2019. Acceptance rate: 30.0%
8. **SRAM Circuits for True Random Number Generation Using Intrinsic Bit Instability**
   L. T. Clark, S. B. Medapuram, and D. K. Kadiyala
   IEEE Transactions on Very Large Scale Integration Systems, (TVLSI), 2018. Acceptance rate: 37.3%

Technical Skills

Programming Languages	:	C, C++, CUDA, Perl, Python, System Verilog, Bash Scripting
Performance modeling	:	ZSim, ASTRA-Sim, DRAMSim, gem5, SESC, Garnet2.0