Rajesh Shashi Kumar

rajesh.shashikumar@arm.com ◇ linkedin.com/in/rajeshwisc ◇ rajesh-s.github.io ◇ Google Scholar

EDUCATION

M.S. Electrical and Computer Engineering, *09/08/2021 - 05/31/2023*
Specialization: Computer Systems & Architecture Grade: 3.94/4
University of Wisconsin-Madison, Madison, WI, USA

B.E. Electronics & Communication Engineering *08/01/2013 - 05/31/2017*
PES Institute of Technology, Bangalore, Karnataka, India Grade: 9.11/10

WORK EXPERIENCE

ARM *06/20/2023 - Present*
Senior Engineer *Austin, Texas, USA*

- Conducting Architecture research to improve scalability and efficiency of datacenter systems for emerging workloads.
- Specialized in hardware-software co-design for distributed heterogeneous memory systems for AI/ML.
- Designed novel architecture for interconnects and memory coherence in next-generation datacenter systems.
- Developed analytical and simulation models for architecture performance modeling and characterization.
- Investigated workloads: NGINX, Vector Packet Processing, OpenMP, LLM inference serving
- Experience in performance tuning for scalable GenAI inference workflows, with expertise in heterogeneous memory systems and working knowledge of full-stack frameworks such as PyTorch and vLLM.

Advanced Micro Devices Research *05/23/2022 - 08/26/2022*
Research Co-op *Austin, Texas, USA*

Investigated efficient mapping of sparse linear algebra operations in graph workloads at-scale for GPU architectures.

Qualcomm *09/03/2019 - 08/11/2021*
Network on Chip Design Engineer *Bangalore, India*

- Designed multi-core hardware cache coherent interconnects. My role involved close collaboration with performance, processor, power, physical design, & memory subsystem teams.
- Extensive experience optimizing RTL and architecture for advanced nodes, managing system performance trade-offs, performance analyses and debugging for mobile and automotive SoCs.

Analog Devices *07/03/2017 - 08/14/2019*
Design Engineer *Bangalore, India*

- Designed hardware interconnects, offload engines, peripheral IP & memory protection as a part of the Digital Signal Processors team. Proven track record in evaluating design feasibility and closing timing at high operational frequencies.
- Contributed to ECO root cause analysis, silicon bring-up, and formal verification of boot flow security schemes.

Analog Devices *01/01/2017 - 06/20/2017*
Performance Verification Intern *Bangalore, India*

Performance analysis of DSP accelerators and system interconnects. Developed an in-house tool for flow-based stimulus generation and capturing SoC interconnect performance metrics.

ACADEMIC RESEARCH EXPERIENCE

Heterogeneous Architectures Lab *Aug 2022 - May 2023*
Advisor: Prof. Matthew Sinclair *Madison, WI*

Efficient multi-chiplet GPU coherence with intelligent acquires and releases to promote inter-kernel data reuse.

Heterogeneous Architectures Lab *Jan 2022 - May 2022*
Advisors: Prof. Shivaram Venkataraman & Prof. Matthew Sinclair *Madison, WI*

Characterizing spatio-temporal effects of DVFS on performance variability in multi-tenant GPU systems for ML.

National Institute of Technology *May 2015 - July 2015*
Advisor: Prof. Santanu Sarkar *Rourkela, India*

Design of process, voltage, temperature independent CMOS bandgap voltage and current reference circuits.

PUBLICATION

- **CPElide: Efficient Multi-Chiplet GPU Implicit Synchronization**
  Preyesh Dalmia, Rajesh Shashi Kumar, Matthew Sinclair
  57th IEEE/ACM International Symposium on Microarchitecture (MICRO 2024)
- **Contention Aware Machine Learning: Understanding CPU ML Inference Scalability Bottlenecks**
  Josh Minor, Minjun Wu, Rajesh Shashi Kumar, Shashank Hegde, Eric Van Hensbergen
  Internal Research Paper, Arm, Texas, USA, 2024

PATENTS

- **Cache Synchronization for Chiplet Accelerators**. Preyesh Dalmia, Rajesh Shashi Kumar, Matthew Sinclair.
  Granted US Patent No.: US12292836B2.

INVITED TALKS

- **Efficient Multi-Chiplet GPU Implicit Synchronization**
  - Architecture Affiliates, University of Wisconsin–Madison. October 2022
- **Power-Efficient AI Inference at Scale**
  - Arm Global Engineering Conference, Birmingham, UK. July 10, 2025

TEACHING

Department of Electrical & Computer Engineering, Wisconsin-Madison *August 2022 - May 2023*
Teaching Assistant *Madison, WI*

· ECE/CS532: Matrix methods in Machine Learning (Fall 2022)
· ECE210: Introductory Experience in Electrical & Computer Engineering (Spring 2023)
· ECE342: Circuit Analysis (Spring 2023)

Department of Computer Sciences, Wisconsin-Madison *August 2021 - May 2022*
Teaching Assistant *Madison, WI*

· CS200: Object-oriented Programming I (Fall 2021)
· CS300: Object-oriented Programming II (Spring 2022)

TECHNICAL SKILLS

Programming Languages	Python, C++, C, CUDA, HIP, System Verilog, Bash
Runtimes	OpenMP, MPI, PyTorch
Performance profilers	Nsight Systems, Streamline, VTune, Perf
Architectural simulators	gem5, GPGPUsim, Interconnect Workbench
Interconnect protocols	CHI, CHI-C2C ACE, AXI, PCIe
Digital design tools	Xcelium, VCS, ModelSim, Spyglass, Vivado, JasperGold
Implementation tools	[Synthesis:DC,Genus,Yosys], [STA:Tempus], [LEC:Conformal]
Productivity Tools	Docker, SLURM, Platform LSF, Git, LaTeX

ACADEMIC PROJECTS

- Characterizing DVFS induced performance variability in Systems for ML. Link to report.
- Improving TLB performance in GPUs using locality prefetching. Link to source code.
- Performance analysis of bidirectional-exchange and ring primitives for AllReduce. Link to source code.
- Deepspeech training and inference performance profiling on Nvidia P100. Link to source code.
- Fine-grained GPU cache bypassing in hardware for GPGPU workloads. Link to source code.

- Design of an operational trans-conductance amplifier using the gm/ID methodology to meet the EEG biosignal pre-processing feature extraction requirements as a part of Bachelor's thesis under Prof. Partha Ray.

AWARDS

Spot Award: Analog Devices 2018

· Awarded in Dec 2018, for root cause analysis and ECO implementation to fix critical clock infrastructure issues affecting preboot, which was much needed for the tapeout confidence of a DSP SoC.

· Awarded in May 2018, for an efficient Interconnect design in NIC400 from scratch resulting in significant performance improvement within the area and power budget with no impact to the project schedule.

IEEE IBM Watson Student Showcase: Top five winning projects worldwide December 2015

Awarded for design of an NLP application to aid in verbal skill development using IBM Watson
Publication: *IEEE Computer Issue No.01 - Jan. (2016 vol.49)* Link to publication