

Education

2021 **Ph.D. Computer Science**



University of Wisconsin–Madison

GPA – **3.82/4.0**

Present

- Ongoing Research: Power impact on concurrent compute and communication in scale-up workloads
- Advisors: Prof. Matt Sinclair and Prof. Shivaram Venkataraman
- Coursework: Advanced Computer Architecture, HPC and Applications in Engineering, Computer Graphics, Operating Systems, Distributed Systems, HW/SW for Next Gen AI Systems, Fault-Tolerant Computing, Real-Time Computing Systems, Optimization, Data Visualization

2015 **B.E.(Hons.) Electrical and Electronics Engineering**



Birla Institute of Technology and Science, Pilani

GPA – **9.52/10.0**

2019

- Class Rank: 3/93
- Coursework: Computer Architecture, Embedded Systems Design, Digital Design, Digital Signal Processing, Microprocessors and Interfacing, Analog and Digital VLSI Design

Publications

In Conference Proceedings

- 2026 **Rutwik Jain**, Yiwei Jiang, Matthew D. Sinclair, Shivaram Venkataraman. Minos: Systematically Classifying Performance and Power Characteristics of GPU Workloads on HPC Clusters. In *Proceedings of the 52nd ACM Conference on Measurement and Analysis of Computing Systems*, SIGMETRICS, June 2026.
- 2024 **Rutwik Jain**, Brandon Tran, Keting Chen, Matthew D. Sinclair, and Shivaram Venkataraman. PAL: A Variability-Aware Policy for Scheduling ML Workloads in GPU Clusters. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, SC '24. IEEE Press, 2024.
- 2022 Prasoon Sinha, Akhil Guliani, **Rutwik Jain**, Brandon Tran, Matthew D. Sinclair, and Shivaram Venkataraman. Not All GPUs Are Created Equal: Characterizing Variability in Large-Scale, Accelerator-Rich Systems. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '22. IEEE Press, 2022.
- 2019 **Rutwik Jain** and Umang Garg. Modelling and Simulation of Onboard Wire Antennas for a 3U CubeSat (undergraduate). In *2019 Photonics and Electromagnetics Research Symposium - Fall (PIERS - Fall 2019)*, pages 871–879, Dec 2019.
- 2018 **Rutwik Jain**, Shubham Sharma, Kushagra Aggarwal, Tushar Goyal, Abhinav Sundhar, Ujjwal Anand, Nishad Sahu, and Joy Parikh. Modes of Operation for a 3U CubeSat with Hyperspectral Imaging Payload (undergraduate). In *Proceedings of the 69th International Astronautical Congress (IAC) 2018, Bremen, Germany*, Oct 2018.

arXiv Preprints

- 2025 Yiwei Jiang, Sangeeta Chowdhary, Nathaniel Morris, **Rutwik Jain**, Srilatha Manne, and Sam Bayliss. Power Aware Dynamic Reallocation For Inference, 2025.

- 2025 Gregor von Laszewski, Wesley Brewer, Jeyan Thiyagalingam, Juri Papay, Armstrong Foundjem, Piotr Luszczek, Murali Emani, Shirley V. Moore, Vijay Janapa Reddi, Matthew D. Sinclair, Sebastian Lobentanzer, Sujata Goswami, Benjamin Hawks, Marco Colombo, Nhan Tran, Christine R. Kirkpatrick, Abdulkareem Alsudais, Gregg Barrett, Tianhao Li, Kirsten Morehouse, Shivaram Venkataraman, **Rutwik Jain**, Kartik Mathur, Victor Lu, Tejinder Singh, Khojasteh Z. Mirza, Kongtao Chen, Sasidhar Kunapuli, Gavin Farrell, Renato Umeton, and Geoffrey C. Fox. AI Benchmark Democratization and Carpentry, 2025.
- 2025 Ben Hawks, Gregor von Laszewski, Matthew D. Sinclair, Marco Colombo, Shivaram Venkataraman, **Rutwik Jain**, Yiwei Jiang, Nhan Tran, and Geoffrey Fox. An MLCommons Scientific Benchmarks Ontology, 2025.

Journal Articles

- 2023 Matthew D. Sinclair, Parthasarathy Ranganathan, Gaurang Upasani, Adrian Sampson, David Patterson, **Rutwik Jain**, Nidhi Parthasarathy, and Shaan Shah. Fifty Years of the International Symposium on Computer Architecture: A Data-Driven Retrospective. *IEEE Micro*, volume 43, pages 109–124. IEEE Computer Society, nov 2023.
- 2020 Anuj K. Ojha, **Rutwik Jain**, and Praveen Kumar AV. Magnetic quadrupole mode excitation of a cylindrical dielectric resonator antenna using planar feed. *Microwave and Optical Technology Letters*, volume 62, pages 484–490, 2020.

Work Experience

- May 2024 **Research Associate, AMD Research and Advanced Development (RAD)**
↓ Power Efficiency for AI Compute (PEAC) Team Austin, TX
- Present
 - Worked with Srilatha (Bobbie) Manne’s team to develop an event-based model for accurate on-chip power estimation on AMD Instinct GPUs.
 - Used the power estimation model to improve excursion management, particularly for LLM workloads. Also developed FW and SW level solutions for managing excursions.
 - Currently working on mitigation techniques for at-scale variability and Power Management (PM) problems.
- Jul 2019 **ASIC Design Engineer, NVIDIA**
↓ Memory Subsystem Design Team Bangalore, India
- Jul 2021
 - Microarchitecture design and RTL implementation for the SoC’s memory management IP.
 - This included new feature implementation, modification of existing μ Arch for PPA optimization, logic synthesis for area and timing accuracy, iterative design to incorporate DV and PD feedback.
- Jan 2019 **ASIC Intern, NVIDIA**
↓ Memory Subsystem Design Team Bangalore, India
- Jun 2019
 - Developed a debug tool that uses a generated waveform database to identify, trace and correlate transactions as they flow through the Memory Management Unit (MMU).
 - Presented my work on developing and scaling the tool in a lightning talk at India-level internal NVIDIA conference, NTech 2019.

Research

- Aug 2021 **Characterizing Variability in Large-scale, Accelerator-rich Systems**
↓ Advisors: Prof. Matt Sinclair and Prof. Shivaram Vankataraman
- Aug 2022
 - Characterized GPU variability in GPU supercomputers such as ORNL Summit and analyzed how it is affected by cluster attributes like size, cooling, and GPU vendors.
 - Profiled different applications (image processing, machine translation, molecular dynamics, graph analytics) on the same cluster to evaluate application-dependence.
 - Co-authored and presented our work at Supercomputing (SC), 2022.
- Aug 2022 **PAL: Variability-aware Scheduling for Large-scale, GPU-based Systems**
↓ Advisors: Prof. Matt Sinclair and Prof. Shivaram Vankataraman
- May 2024
 - Designed a variability-aware scheduling policy that makes application-specific GPU allocations when scheduling jobs in large-scale GPU-based systems, based on variability profiles from hardware.
 - Authored and presented our work at Supercomputing (SC), 2024.

May 2024 **Excursion Management in Accelerators**



In collaboration with AMD Research

Present

- Developed an accurate event-based on-chip power estimation methodology to detect power excursions on Instinct GPUs.
- Obtained < 13% prediction errors across state-of-the-art deep learning workloads including LLaMA training, inference and diffusion models.
- Used this digital power meter to demonstrate potential for effectively managing power excursions.

Course Projects

Aug 2023 **Parallelized Implementation of Coordinated Animal Motion – Boid Flocking**



Course: CS 759 High Performance Computing and Applications in Engg, Fall 2023

Dec 2023

- Wrote CPU-based C++ and OpenMP implementations and CUDA-based GPU implementation of flocking behavior to compare frame rate performance and scalability.
- The CUDA kernel with additional shared memory optimizations allowed flocking simulations to scale gracefully as number of particles increases up to 2^{14} .

Aug 2022 **PICLE: Prefetcher for Inter-CU Locality Exploitation in GPU TLBs**



Course: CS 757 Adv Computer Architecture II, Fall 2022

Dec 2022

- Implemented a GPU L1-TLB prefetcher based on Valkyrie [Baruah et.al, PACT'20] which exploits inter-TLB locality to improve GPU virtual memory performance using gem5.

Jan 2022 **Project Eagle: Open-source HW/SW Stack for AI Applications**



Course: CS 838 Next-Gen HW/SW Design for AI/ML, Spring 2022

May 2022

- Implemented RTL for a dual-ported, lockup-free L1 cache with prefetching for the memory system of UPCYCLE, a tiled accelerator architecture.

Jan 2022 **Replicated Block Store**



Course: CS 739 Distributed Systems, Spring 2022

May 2022

- Implemented a primary-backup replica based distributed block store.
- Analyzed and reported read and write performance, crash recovery and reintegration time, and evidence for correctness and linearizability guarantees.

Skills

Programming: C, C++, OpenMP | Scripting: Python | RTL Design: Verilog
GPU computing: CUDA, ROCm stack | HPC: MPI | Optimization: Julia & JuMP
Cycle-level Simulators: gem5 | RTL Design and Debug: ModelSim, VCS Synopsys
Version Control: Git | Containers: Docker, Apptainer | Visualization: Tableau

Talks & Presentations

- Jun 2026 **CRISPIER: A CRISP-Based DVFS Implementation in gem5 for AMD GPUs**, gem5 Workshop at ISCA 2026. Raleigh, NC.
- Jun 2026 **Minos: Systematically Classifying Performance and Power Characteristics of GPU Workloads on HPC Clusters**, ACM SIGMETRICS 2026, Ann Arbor, MI.
- Sep 2025 **Embracing Variability in Accelerator-Rich Systems: Power Management Approaches Across the Stack**, Preliminary Examination, University of Wisconsin-Madison.
- Jan 2025 **Characterizing GPU variability in Large-Scale, Accelerator-Rich Systems**, MLCommons Science and HPC Working Group, Online.
- Jan 2025 **PAL: A Variability-Aware Policy for Scheduling ML Workloads in GPU Clusters**, HPC, Systems and Applications Sync, AMD, Online.
- Dec 2024 **Excursion management in MI Accelerators**, AMD Research, Online.
- Nov 2024 **PAL: A Variability-Aware Policy for Scheduling ML Workloads in GPU Clusters**, SC 2024. Atlanta, GA.
- Jan 2024 **GPU Variability: Characterization Study and Mitigation Strategies**, AMD Research. Online.
- Nov 2022 **Not All GPUs are Created Equal: Characterizing GPU variability in Large-Scale, Accelerator-Rich Systems**, SC 2022. Dallas, TX.

- Oct 2022 **Variability-Aware Scheduling for Efficient Exascale Computing**, UW-Madison's Architecture Affiliates 2022. Madison, WI.
- Oct 2018 **Modes of Operation of a 3U CubeSat with Hyperspectral Imaging Payload**, 69th International Astronautical Congress (IAC) 2018. Bremen, Germany.

Teaching

- FA 21 **CS/ECE 552: Introduction to Computer Architecture**
- SP 22 Taught Verilog and HDL concepts in discussion section. Also helped students with in-class exercises and debugging projects (pipelined processor and cache implementation)
- FA 22 **CS/ECE 252: Introduction to Computer Engineering**
Assisted students with homework and projects through office hours, graded homework assignments.

Awards & Honors

- AMD spotlight award (Q4 2025) for technical contributions to power estimation, excursion management, and power profiling tools.
- CS Summer Research Assistantship from UW-Madison for the 2022 summer term.
- First-year CS Departmental Scholarship from UW-Madison for academic year 2021-22.
- Merit Scholarship at BITS Pilani: 40% tuition waiver for seven semesters from 2015 to 2019.