Suchita Pati *Curriculum Vitae*

CONTACT INFORMATION

AMD Research
2485 Augustine Drive,
Santa Clara, CA-95054, USA.

Mobile: +1-608-572-9690
E-mail: suchita26pati@gmail.com
LinkedIn: suchitapati
Web: http://pages.cs.wisc.edu/∼spati/

RESEARCH INTERESTS

ML Systems, Distributed LLM Acceleration, Communication Collectives, GPU Architecture.

RESEARCH SUMMARY

Accelerating Large Language Models (LLMs) on GPUs. I currently focus on efficient DMA offload and overlap of communication collectives in distributed GPU settings. My past research has involved across-the-stack GPU optimizations for LLMs, including optimizing GEMMs for concurrent executions on GPUs, and offloading memory-intensive LLM phases to in-memory compute units.

EDUCATION BACKGROUND

University of Wisconsin-Madison May'19 - Feb'24
- PhD in Computer Sciences. GPA: 3.94/4.0
- Adviser: Prof. Matthew D. Sinclair
- Thesis: *Cross-Stack Optimizations for Sequence-Based Models on GPUs*

University of Wisconsin-Madison Aug'17 - May'19
- M.S. in Computer Sciences. GPA: 3.86/4.0
- Adviser: Prof. Matthew D. Sinclair

Birla Institute of Technology and Science, Pilani Aug'11 - May'15
- B.E. (Hons.) Electrical and Electronics Engineering. GPA: 9.18/10

CONFERENCE PUBLICATIONS

[1] Suchita Pati, Shaizeen Aga, Mahzabeen Islam, Ryan Quach, Saleel Kudchadker, Mohamed Ibrahim, "DMA-Latte: Expanding the Reach of DMA Offloads to Latency-bound ML Communication", *to appear*, **MICRO 2026**
Optimized DMA-offloaded ML collectives and KV cache transfers leveraging modern DMA architectural features to substantially close the performance gap with GPU core-initiated communication and improve LLM inference performance.

[2] Shagnik Pal, Shaizeen Aga, Suchita Pati, Mahzabeen Islam, Lizy K John, "Design Space Exploration of DMA-based Finer-Grain Compute Communication Overlap", *to appear*, **IISWC 2026**
Characterization & optimization of fine-grained compute–communication overlap for diverse network topologies and improved dataflow.

[3] Suchita Pati, Shaizeen Aga, Nuwan Jayasena, and Matt Sinclair. "GOLDYLOC: Global Optimizations & Lightweight Dynamic Logic for Concurrency", (**ACM TACO 2025**)
Concurrency-aware library tuning and runtime system for efficient concurrent GEMM execution on GPUs.

[4] Anirudha Agrawal, Shaizeen Aga, Suchita Pati, Mahzabeen Islam. "ConCCL: Optimizing ML Concurrent Computation and Communication with GPU DMA Engines", (**ISPASS 2025**)
Optimizing concurrent computation and communication on GPUs via scheduling and resource partitioning optimizations, as well as offloading communication to DMAs.

[5] Suchita Pati, Shaizeen Aga, Mahzabeen Islam, Nuwan Jayasena, and Matt Sinclair. "T3: Transparent Tracking & Triggering for fine-grained overlap of compute & communication", (**ASPLOS 2024**)
A transparent and efficient mechanism to initiate and orchestrate communication as data is generated by producer operations in a fine-grained manner.

[6] Mohamed Ibrahim, Shaizeen Aga, Ada Li, Suchita Pati, Mahzabeen Islam. "JIT-Q: Just-in-time Quantization with Processing-In-Memory for Efficient ML Training", (**MLSys 2024**)

A mechanism to lower memory capacity needs of training by efficiently quantizing high-precision weights when needed using processing-in-memory (PIM) technology.

[7] Suchita Pati, Shaizeen Aga, Mahzabeen Islam, Nuwan Jayasena, and Matt Sinclair. "Tale of Two Cs: Computation vs. Communication Scaling for Future Transformers on Future Hardware", (**IISWC 2023**), **preprint on ArXiV, Jan 2023**
A multi-axial (algorithmic, empirical, hardware evolution) analysis of compute vs. communication scaling for future Transformer models on future hardware.

[8] Suchita Pati, Shaizeen Aga, Nuwan Jayasena, Matt Sinclair, "Demystifying BERT: System Design Implications", to appear in Proc. Int. Symposium on Workload Characterization (**IISWC 2022**), **preprint on ArXiV, April 2021**.
Detailed characterization of Transformer networks, with focus on BERT, and acceleration opportunities with processing-near-memory.

[9] Suchita Pati, Shaizeen Aga, Matt Sinclair, Nuwan Jayasena. "SeqPoint: Identifying Representative Iterations of Sequence-Based Neural Networks", in Proc. Int. Symposium on Performance Analysis of Systems and Software (**ISPASS 2020**)
Tool for efficient sampling and characterization of SQNN training on GPUs.

[10] Jonathan Lew, Deval Shah, Suchita Pati, Shaylin Cattell, Mengchi Zhang, Amruth Sandhupatla, Christopher Ng, Negar Goli, Matt Sinclair, Tim Rogers, Tor Aamodt, "Analyzing Machine Learning Workloads Using a Detailed GPU Simulator", extended abstract in Proc. Int. Symposium on Performance Analysis of Systems and Software (**ISPASS 2019**), **extended version on ArXiV, Nov. 2018.**
Open-sourced infrastructure to simulate GPUs with state-of-the-art machine learning algorithms.

[11] Rajesh Kumar, Suchita Pati and Kanishka Lahiri, "DARTS: Performance Counter Driven Sampling Using Binary Translators", extended abstract in Proc. Int. Symposium on Performance Analysis of Systems and Software (**ISPASS 2017**)
Fast and efficient tool to identify representative workload phases and collect their instructions traces using performance counters, binary translation and machine learning.

Other Publications

[12] Suchita Pati, "Cross-Stack Optimizations for Sequence-Based Models on GPUs", PhD Thesis, UW-Madison, 2024

[13] Reese Kuper, Suchita Pati, Matt Sinclair, "Improving GPU Utilization in ML Workloads Through Finer-Grained Synchronization", in The 3rd Young Architect Workshop co-located w/ ASPLOS (**YArch 2021**)

[14] Suchita Pati, "Exploring GPU Architectural Optimizations for RNNs", in The 1st Young Architect Workshop co-located w/ HPCA (**YArch 2019**)

Internal Publications

[15] Suchita Pati, Mahzabeen Islam, Shaizeen Aga, "Efficient Communication Offload via MI300X DMA Collectives Optimizations", AMD Global Technical Authors Conference (**GTAC'25**)

[16] Mahzabeen Islam, Shaizeen Aga, Suchita Pati, et al., "Specializing DMAs for ML", AMD Global Technical Authors Conference (**GTAC'25**)

[17] Mahzabeen Islam, Suchita Pati, Mohamed Ibrahim, Shaizeen Aga, Gilbert Lee, "A Case for sDMA-based ML Collectives", AMD Global Technical Authors Conference (**GTAC'24**)

[18] Suchita Pati, Rithik Jain, Shaizeen Aga, "Genie - Cross-platform, Cross-framework GenAI GPU Database and Learnings", AMD Global Technical Authors Conference (**GTAC'24**)

[19] Shaizeen Aga, Suchita Pati, Rithik Jain, "Lessons Learned from LLaMA/Falcon Inference Characterization on state of the art GPUs", AMD Global Technical Authors Conference (**GTAC'24**)

[20] Mohamed Ibrahim, Shaizeen Aga, Ada Li, Mahzabeen Islam, Suchita Pati, Anish Saxena, "Just-in-time (JIT) Quantization for Lowering LLM Training Capacity Needs", AMD Global Technical Authors Conference (**GTAC'23**)

[21] Mahzabeen Islam, John Alsop, Shaizeen Aga, Suchita Pati, Vamsi Alla, "Interference Characterization of Concurrent Computations in ML", AMD Global Technical Authors Conference (**GTAC'23**)

[22] Rajesh Kumar, Suchita Pati, Kanishka Lahiri, "Speeding up instruction tracing by hardware profiling AMD SimNow", AMD Asia Technical Conference (**AATC'17**)

[23] Suchita Pati, Kanishka Lahiri, "Characterizing SPECjbb2015 – A Server side Java Performance Benchmark", AMD Asia Technical Conference (**AATC'16**)

Selected Recognition

1. AMD Spotlight Award (Q3'2025), for developing DMA prototypes that deliver performance, energy, and programmability wins for ML collectives.
2. AMD Spotlight Award (Q1'2025) for architectural innovations that improve ML communication performance and power.
3. AMD Spotlight Award (Q3'2024) for characterizing state-of-the-art generative AI models that identified opportunities to improve hardware and software performance.
4. AMD Spotlight Award (Q1'2024) for performance projections of LLM inference executions on PIM.
5. UW-Madison Student Research Grants Competition - Conference Presentation Funds
6. Qualcomm Innovation Fellowship Finalist, 2020.
7. UW-Madison CS Summer Research Award, 2020.
8. UW-Madison CS Golden Brick Award for service towards WACM, 2019.
9. CRA-W Scholarship to attend Grad Cohort Workshop, 2019.
10. Hiran Mayukh Award, UW Computer Architecture 2018, given to a computer architecture student who has shown great potential during their first year of graduate studies.
11. Grace Hopper Scholar 2018.
12. AMD Spotlight Award 2016.

Patents

1. Suchita Pati, Mahzabeen Islam, Shaizeen Aga, "Data exchange scheduling within a processing group.", *Pending*
2. Shaizeen Aga, Suchita Pati, Mahzabeen Islam, Brad Beckmann, Mario Baldi, "Systems and methods to offload all-to-all collective communications", *Pending*
3. Just-in-time KV (key-value) cache generation for generating AI inference.
   Shaizeen Aga, Akila Subramaniam, Suchita Pati, US Patent App. 19/000,019 *Pending*
4. Suchita Pati, Shaizeen Aga, Nuwan Jayasena, Matt Sinclair, "Dynamic control of work scheduling.", US Patent App. 18/091,443, *Pending*
5. Johnathan Alsop, Suchita Pati, Shaizeen Aga, and Sergey Blagodurov, "Systems and methods for scheduling workloads.", US Patent App. 18/088,884, *Pending*
6. Shaizeen Aga, Suchita Pati, Nuwan Jayasena, "Fused data generation and associated communication", U.S. Patent No. 12474926, *Granted*

Academic and Professional Experience

Member of Technical Staff (AMD Research & Advanced Development) Jun'23 - Present
- Mentors: Shaizeen Aga
- Optimizing ML communication in distributed multi-GPU environments to enhance the scalability and efficiency of Large Language Models.

Graduate Research Assistant (UW-Madison) Jan'20 - Jun'23
- Advisor: Prof. Matt Sinclair
- Optimizing GPU hardware-software stack for Natural Language Processing (NLP) training via efficient use of operational parallelism, characterizing state-of-the-art models (Transformers, RNNs), and their distributed setups with focus on communication collectives.

Architecture Research Intern (AMD Research) May'19 - Jun'23
- Mentors: Nuwan Jayasena and Shaizeen Aga
- Accelerating NLP operations using near-memory-processing, proposing efficient multi-GPU scheduling policy, and developing a profiling tool for sequence-based models.

Graduate Student Researcher (UW-Madison) Aug'17 - May'19
- Advisor: Prof. Matt Sinclair
- Enabled simulation of GPU architectures with contemporary machine learning applications by extending GPGPU-Sim to support deep learning CUDA libraries like cuDNN and cuBLAS.

Architecture Research Intern (AMD Research) May'18 - Aug'18
- Mentor: John Kalamatianos
- Improve performance and energy efficiency of next generation AMD CPUs for DoE's exascale applications through intelligent control of type and aggressiveness of data prefetchers.

Design Engineer 2, AMD (Bengaluru, India) Jan.'17 - Jul.'17
- Mentor: Kanishka Lahiri
- Identified performance bottlenecks in the latest AMD server, EPYC, with focus on cache-to-cache transfer latency, NUMA latency, data prefetching and prefetch throttling.
- Worked with software team to tune SPECJbb2015 and the JVM on EPYC for publishing best possible benchmark scores.
- Mentored intern on characterization of the in-memory NoSQL database Redis with YCSB.

Design Engineer 1, AMD (Bengaluru, India) Jul.'15 - Dec.'16
- Mentor: Kanishka Lahiri
- Devised DARTS, an efficient workload trace sampling methodology using Dynamic Binary Translators (AMD SimNow), performance counter data and machine learning, which significantly reduced tracing effort and is widely used across performance teams at AMD
- Studied *SPECjbb2015* and *NoSQL Database Cassandra* with *Yahoo Cloud Serving Benchmark(YCSB)* to (a) identify hardware bottlenecks in AMD servers, (b) generate instructions and memory traces for simulations and, (c) study impact of architectural features and project their performance on future server architectures.
- Mentored two interns on setting up a distributed cluster with Cassandra database server and YCSB clients.

Intern, Analog Devices Inc. (Bengaluru, India) Jan.'15 - Jun.'15
- Mentor: Anand Venkitasubramani
- Implemented Universal Verification Methodology in SystemC for Verification of SoCs.
- Enabled efficient development and reuse of verification environments for verification of SOCs, obviating the need for different design and verification environments.

Service

- **Program Committee** for MICRO'26, HPCA'26, ICS'26, ISPASS'26, IISWC'25, ICS'25, ISPASS'25, IISWC'24, EMC2'24, GTAC'24 (AMD-internal)
- **Reviewer** ACM TACO, FGCS journal.
- **President**, W-ACM, UW Madison chapter of ACM's Women in Computing 2020-21.
- **Artifact Evaluation Committee** for MICRO 2021
- **Secretary, Activity Chair, Mentor**, W-ACM 2017-20.
- **Member**, Corporate Social Responsibility Committee, AMD 2015-17
- **Member**, BITS-Embryo, BITS-Pilani Alumni Assoc. 2013-16

Grants

Travel grants for IISWC'23, IISWC'22, ISPASS'20, HPCA'19

Research Projects

Classical Simulation of Quantum Circuits: Stabilizer Formalism & Beyond Spring'20
- Mentor: Prof. Dieter van Melkebeek, Course: CS880
- Developed a quantum circuit simulator to simulate Clifford and non-Clifford gates leveraging stabilizer frame representation.
- Simulated the Quantum Fourier Transform circuit using our simulator and compared the result and performance with IBM's open source simulator

Tail Latency and Predictable Local Storage Systems Fall'18

- Mentor: Prof. Remzi Arpaci-Dusseau, Course: CS739
- Added support to measure tail latency in Ceph distributed storage system and identified cluster configs for optimal performance.
- Studied Ceph behavior under long latency. Modeled fail fast and redirected requests to study performance benefits.

Effective Prefetching for Multi-core/Multiprocessor Systems Spring'18

- Mentor: Prof. Joshua San Miguel, Course: CS757
- Proposed techniques to reduce cache interference and coherence downgrades/invalidations caused by local prefetchers in multi-core systems employing directory-based protocols.
- Employed techniques to improve global prefetch effectiveness and thus, performance, by tuning local prefetcher aggressiveness.

Transparent File Compression Spring'18

- Mentor: Andrea C. Arpaci-Dusseau , Course: CS736
- Integrated bzip2 kernel compression with ext2 file-system and implemented a smart user-level program to perform on-demand decompression and delayed compression using heuristics derived from file characteristics.
- Resulted in 50% disk space saving with only 10% increase in file access time.

Remembering Prediction between Context Switches Fall'17

- Mentor: Prof. Mikko Lipasti , Course: CS752
- Analyzed the impact of context switches on the TAGE branch predictor accuracy and identified hot-spots of destructive interference caused by intermediate processes
- Reduced its impact by storing/restoring TAGE structures between context switches and making the predictor aware of the process identity.

Skills

Languages: C, C++, Python, Shell, Verilog, Matlab, SystemC.
Simulators and Tools: GPGPU-Sim, ZSIM, gem5, xv6 OS, Intel Pin, AMD SimNow, Linux Perf, Intel PCM, CodexL, rocprof, AMD Propriety perf. tools.

Talks/Posters

- *Accelerating DNN Training By Better Utilizing Hardware Resources*, Preliminary Examination, UW-Madison, Dec'22
- *Demystifying BERT: System Design Implications*, IISWC, Nov'22
- *System Design Implications for Transformers*, UW Architecture Affiliates, Oct'22
- *Improving GPU Utilization in ML Workloads Through Finer-Grained Synchronization*, UW Architecture Affiliates, Oct'21
- *Demystifying BERT: Implications for Accelerator Design.* AMD Research, Dec'20
- *SeqPoint: Identifying Representative Iterations of Sequence-based Neural Networks.* UW Architecture Affiliates, Oct'20
- *SeqPoint: Identifying Representative Iterations of Sequence-based Neural Networks*, ISPASS, Aug'20
- *SeqPoint: Identifying Representative Iterations of Sequence-based Neural Networks*, AMD Research, Dec'19
- *Analyzing Machine Learning Workloads Using a Detailed GPU Simulator*, Poster at ISPASS, April'19
- *Exploring GPU Architectural Optimizations for RNNs*, YArch/HPCA Feb'19