Suchita **Pati**

(+1) 608-572-9690 | spati@cs.wisc.edu | pages.cs.wisc.edu/ spati/ | suchitapati

Education

University of Wisconsin - Madison
CGPA - 3.94/4

Ph.D. in Computer Science, Adviser - Prof. Matthew D. Sinclair
May. 2019 - present

University of Wisconsin - Madison
CGPA - 3.94/4

M.S. in Computer Science, Adviser - Prof. Matthew D. Sinclair
Sep. 2017 - May 2019

Birla Institute of Technology & Science, Pilani
CGPA - 9.18/10

B.E.(Hons) in Electrical & Electronics Engineering
Aug. 2011 - Jun. 2015

Research Experience

UW-Madison
Madison, USA

Graduate Research Assistant (Mentor: Prof. Matthew D. Sinclair)
Jan.'20 - Present

- Characterizing RNN and Attention-based network training and identifying performance bottlenecks on GPUs
- Enabled simulation of GPU architectures with contemporary Deep Learning applications by extending GPGPU-Sim to support deep learning CUDA libraries like cuDNN and cuBLAS.

AMD Research
Santa Clara, USA

Architecture Research Intern (Mentor: Nuwan Jayasena and Shaizeen Aga)
May'19-Dec.'19, Jan'20-Present (remote)

- Studied end-to-end DNN training to extract operational parallelism within the networks and identified opportunities to offload operations for Processing-In-Memory.

AMD Research
Boston, USA

Architecture Research Intern (Mentor: Dr. John Kalamatianos)
May. 2018 - Aug. 2018

- Focused on improving performance and energy efficiency of the next generation AMD CPUs for DoE's Exascale Applications.
- Proposed an efficient data prefetch control mechanism which reduced overall data movement in the cache hierarchy and thus, power while also improving performance across HPC, Server, Cloud, Desktop and SPEC workloads.

Industrial Experience

AMD
Bengaluru, India

Design Engineer 2, *Server Performance Group*
Jan. 2017 - Jul. 2017

- Identified performance bottlenecks in the latest AMD server, EPYC, with focus on cache-to-cache transfer latencies, 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 setup, tuning and characterization of the in-memory NoSQL database Redis with YCSB

AMD
Bengaluru, India

Design Engineer 1, *Server Performance Group*
Jul. 2015 - Dec. 2016

- Devised DARTS, a novel and efficient workload trace sampling methodology using Dynamic Binary Translators (AMD SimNow), performance counter data and machine learning techniques which significantly reduced the tracing effort and has been widely used across performance teams at AMD
- Workload characterization for *SPECjbb2015* and *NoSQL Database Cassandra* with *Yahoo Cloud Serving Benchmark(YCSB)* through:
  - Analysis of micro-architectural statistics from hardware to identify bottlenecks in current AMD and Intel servers.
  - Generation of instruction traces and highly-threaded memory access traces from hardware and creating suites for each workload using statistical and machine learning techniques to drive future core and SoC simulators.
  - Cycle-accurate simulations of future architectures to study the impact of architectural features on these workloads.
  - Projecting performance of workloads on future architectures using current hardware and cycle-accurate simulations.

Analog Devices Inc.
Bengaluru, India

Intern
Jan. 2015 - Jun. 2015

- Implemented Universal Verification Methodology in SystemC for Verification of SoCs.
  - Implemented UVM-SystemC using Cadence UVM-SC Library, SystemC Verification (SCV) Library and Transactional level Modeling (TLM).
  - Enabled efficient development and reuse of verification environments for verification of SOCs, obviating the need for different design and verification environments.

Honors & Awards

2020	**Summer RA Award**, CS department funding to pursue research over summer.	*UW-Madison*
2019	**YArch/HPCA'19 Student Travel Award**, to attend and present at YArch'19	*HPCA'19*
2018	**Hiran Mayukh Award**, Given to a computer architecture student who has shown great potential during his or her first year of graduate studies at Wisconsin.	*UW-M Computer Architecture*
2018	**CRA-W Grad Cohort**, Invited to attend the Grad Cohort Workshop 2019	*CRA-W*
2018	**Scholar, Grace Hopper Conference (GHC) 2018** , Accepted to attend GHC'18	*AnitaB.org*
2016	**Spotlight Award**, for successfully characterizing AMD's first server workload in several years.	*AMD*

Refereed Publications

- **S. Pati**, S. Aga, M. Sinclair, N. Jayasena ,"SeqPoint: Identifying Representative Iterations of Sequence-Based Neural Networks", *to appear in Proc. Int. Symposium on Performance Analysis of Systems and Software (ISPASS), 2020*.
- J. Lew, D. Shah, **S. Pati**, S. Cattell, M. Zhang, A. Sandhupatla, Christopher Ng, N. Goli, M.D. Sinclair, T.G. Rogers, T. Aamodt,"Analyzing Machine Learning Workloads Using a Detailed GPU Simulator", *in Proc. Int. Symposium on Performance Analysis of Systems and Software (ISPASS), 2019*.
- J. Lew, D. Shah, **S. Pati**, S. Cattell, M. Zhang, A. Sandhupatla, Christopher Ng, N. Goli, M.D. Sinclair, T.G. Rogers, T. Aamodt,"Analyzing Machine Learning Workloads Using a Detailed GPU Simulator", *preprint on ArXiV, November 2018*.
- R. Kumar, **S. Pati**, and K. Lahiri, "DARTS: Performance Counter Driven Sampling Using Binary Translators," *in Proc. Int. Symposium on Performance Analysis of Systems and Software (ISPASS), 2017*.

Other Publications

- **S. Pati**, "Exploring GPU Architectural Optimizations for RNNs", *in The 1st Young Architect Workshop co-located w/ HPCA (YArch-2019)*
- R. Kumar, **S. Pati**, and K. Lahiri, "Speeding up instruction tracing by hardware profiling AMD SimNow," *in AMD Asia Technical Conference, 2017*.
- **S. Pati**, and K. Lahiri, "Characterizing SPECjbb2015 – A Server side Java Performance Benchmark," *in AMD Asia Technical Conference, 2016*.

Research Projects

Tail Latency and Predictable Local Storage Systems

UW-Madison

Mentor: Prof. Remzi Arpaci-Dusseau

Nov. 2018 - Dec. 2018

- Added support to measure tail latency in Ceph distributed storage system and identified cluster configs for optimal performance.
- Studied Ceph behavior under long latencies. Modeled fail fast and redirected requests to study performance benefits.

Effective Prefetching for Multicore/Multiprocessor Systems

UW-Madison

Mentor: Prof. Joshua San Miguel

Mar. 2018 - May. 2018

- Proposed techniques to study and reduce cache interference and coherence downgrades/invalidations caused by local prefetchers in multi-core and multi-processor systems employing directory-based protocols.
- Employed techniques to improve global prefetch effectiveness and thus, performance, by tuning local prefetcher aggressiveness.

Transparent File Compression

UW-Madison

Mentor: Prof. Andrea C. Arpaci-Dusseau

Mar. 2018 - May. 2018

- Integrated bzip2 kernel compression with ext2 file-system and implemented a smart user-level program to perform on-demand de-compression and delayed compression using heuristics derived from file characteristics.
- Resulted in 50% disk space saving with only 10% increase in file access time.

Hit on Transition Prediction: Remembering Prediction between Context Switches

UW-Madison

Mentor: Prof. Mikko Lipasti

Oct. 2017 - Present

- Analyzed the impact of context switches and context switch interval on the TAGE branch predictor accuracy with varying number of simultaneously executing processes.
- Identified hot-spots of destructive interference caused by intermediate processes and reduced its impact by storing/restoring TAGE structures between context switches and making the predictor aware of the process identity.

Design and Implementation of Digital Filters for Bio-medical Signal Processing

BITS Pilani

Mentor: Dr. Prabhakar Rao

Aug. 2014 - Dec. 2014

- Designed and simulated an efficient cascade of Digital Finite Impulse Response and Infinite Impulse Response filters using Matlab for removal of high frequency power line interference and low frequency baseline wander from Electrocardiogram for noise free analysis and detection of diseases.

Skill Sets

- **Programming Languages:** C, C++, Python, Shell, Verilog, Matlab, SystemC, 8086 Assembly Language Programming.
- **Tools and Simulators:** gem5, GPGPU-Sim, ZSIM, xv6 OS, Intel Pin, AMD SimNow, Perf, Intel PCM, CodexL, Xilinx ISE, Matlab