Suchita Pati *Curriculum Vitae*

Contact Information

University Of Wisconsin-Madison
Department of Computer Sciences
1210 W Dayton St
Madison, WI 53706

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

Research Interests

GPU Architecture, Deep Learning Acceleration, Near-Memory Computing, Architectural Simulation and Profiling

Research Summary

Accelerating training of Natural Language Processing (NLP) models on GPUs. My research involves across-the-stack GPU optimizations for RNN and attention-based models. In the process, I have also characterized state-of-the-art NLP applications, built profiling tools to faithfully characterize their training and developed simulation infrastructure to simulate GPUs executing these applications.

Education Background

University of Wisconsin-Madison May'19 - Present

- PhD Candidate in Dept. of Computer Sciences. GPA: 3.9/4.0
- Adviser: Prof. Matthew D. Sinclair

University of Wisconsin-Madison Aug'17 - May'19

- Master's in Dept. of Computer Sciences. GPA: 3.9/4.0
- Adviser: Prof. Matthew D. Sinclair

Birla Institute of Technology and Science, Pilani Aug'11 - May'15

- B.E. (Hons.) Electrical and Electronics Engineering. GPA: 9.2/10

Conference Publications

[1] Suchita Pati, Shaizeen Aga, Nuwan Jayasena, and Matt Sinclair. "Intelligent Concurrent GEMM Execution", (In Submission)
Concurrency-aware library tuning and runtime system for efficient concurrent GEMM execution.

[2] 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.

[3] 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.

[4] 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.

[5] 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

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

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

[8] Rajesh Kumar, Suchita Pati, Kanishka Lahiri, "Speeding up instruction tracing by hardware profiling AMD SimNow", in AMD Asia Technical Conference (**AATC 2017**)

[9] Suchita Pati, Kanishka Lahiri, "Characterizing SPECjbb2015 – A Server side Java Performance Benchmark", in AMD Asia Technical Conference (**AATC 2016**)

Selected Recognition

1. Qualcomm Innovation Fellowship Finalist, 2020.
2. UW-Madison CS Summer Research Award, 2020.
3. UW-Madison CS Golden Brick Award for service towards WACM, 2019.
4. CRA-W Scholarship to attend Grad Cohort Workshop, 2019.
5. Hiran Mayukh Award, UW Computer Architecture 2018.
6. Grace Hopper Scholar 2018.
7. AMD Spotlight Award 2016.

Research Experience

Graduate Research Assistant (UW-Madison) Jan 2020 - Present

- Advisor: Prof. Matt Sinclair
- Accelerate training of RNN (GNMT, DeepSpeech2) and attention-based (BERT, Transformer) Natural Language Processing (NLP) applications on GPUs, and characterizing state-of-the-art NLP applications on GPUs.

Architecture Research Intern (AMD Research) May 2019 - Present

- Mentors: Nuwan Jayasena and Shaizeen Aga
- Study end-to-end DNN training to extract operational parallelism within the networks and identify opportunities to offload operations for Processing-In-Memory, and building efficient sampling tools to faithfully characterize their behavior.

Graduate Student Researcher (UW-Madison) Aug 2017 - May 2019

- Advisor: Prof. Matt Sinclair
- Enable simulation of GPU architectures with contemporary deep learning applications by extending GPGPU-Sim to support deep learning CUDA libraries like cuDNN and cuBLAS.

Architecture Research Intern (AMD Research) May 2018 - Aug 2018

- 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.

Industry Experience

Design Engineer 2, AMD (Bengaluru, India) Jan. 2017 - Jul. 2017

- 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. 2015 - Dec. 2016

- 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 has been widely used across performance teams at AMD
- Studied *SPECjbb2015* and *NoSQL Database Cassandra* with *Yahoo Cloud Serving Benchmark(YCSB)* by (a) identifying bottlenecks in existing AMD servers, (b) generating instructions and memory access traces for core and SOC simulations and, (c) studying impact of architectural features and projecting 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. 2015 - Jun. 2015

- 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.

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.

Service

- **Artifact Evaluation Committee** for MICRO 2021
- **President**, W-ACM, UW Madison chapter of ACM's Women in Computing 2020-21.
- **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'22, ISPASS'20, HPCA'19

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

- *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