SRIKANT BHARADWAJ Ph.D.

+1-470-775-6738 | srikantvv@gmail.com | srikantbharadwaj.github.io/
srikantvv | Google Scholar: Srikant | Justia Patents: Srikant
Redmond, WA, 98005, USA

Systems researcher with 9 years of industrial experience in the field of GPU architecture, high-performance computing, machine learning, and quantum computing.

Professional Experience

- **Microsoft** — Redmond, WA, USA
  Principal Researcher — *Dec 2024 - Present*
  - Leading a team on innovation in hardware-software codesign for efficient artificial intelligence.
  - Managing and mentoring research fellows and interns, fostering their professional development.

  Senior Researcher — *Apr 2022 - Nov 2024*
  - Developed **novel attention execution technique** which improves hardware utilization in large language model inference resulting in 15% reduction in yearly costs.
  - Collaborated with multiple teams to enable a holistic benchmarking guideline for external silicon vendors.
  - Published **several research papers and patents** in key novel techniques to improve efficiency of machine learning workloads.
  - Managed and mentored interns, cultivating essential research methodologies and publishing skills.

- **Advanced Micro Devices (AMD)** — Bellevue, WA, USA
  Senior Silicon Design Researcher — *Feb 2018 - Apr 2022*
  - Conducted research in architecting GPUs, CPUs, and accelerators for high-performance computing used for scientific computing and machine learning workloads.
  - Transferred research technology to the development of **fastest supercomputers**, Frontier (ORNL) and El Capitan (LLNL).
  - Developed quantum and cryogenic software/firmware support infrastructure for AMD CPUs and GPUs.
  - Designed and developed detailed interconnect models for accurate and high-fidelity simulations (Open sourced as *gem5*).
  - Earned **Spotlight Awards** for contributions to internal GPU power modelling infrastructure.

- **Apple** — Cupertino, CA, USA
  Intern - iOS GPU Architecture — *May 2017 - Aug 2017*
  - Designed a tracking scheme for GPU kernel crashes in iPhones and iPads during heavy graphics usages.
  - Architected software-based mechanism which improves productivity of engineers when fixing remote problems occurring during the usage of commercial devices.

- **Nvidia** — Bangalore, India
  ASIC Engineer II — *Jan 2016 - Jun 2016*
  - Designed and implemented the client-side support for VCS Save restore mechanism for integrated GPU in Tegra which saved 78% of test run times.
  - Contributed with a software team to upgrade a software based headless (no-CPU) full chip testing system for Tegra chips which simulate CPU+GPU full chip tests.

- **Oracle** — Bangalore, India
  Associate Software Engineer — *Jul 2014 - Jan 2016*
  - Designed and implemented a way to import virtualized operating system instances into a cluster.
  - Researched timings of server failure by creating an algorithm for convergence of server down times.

- **ST Microelectronics** — Delhi, India
  Intern — *Jan 2014 - Jul 2014*
  - Developed a simulation environment for the subsystem using Specman e and Verilog.
  - Analyzed the RTL regression results and used HDL code coverage and functional coverage to measure the exhaustiveness.

Education

- **Georgia Institute of Technology** — Atlanta, GA, USA
  Ph.D. — *Aug 2020 - Dec 2023*
  - GPA: 4.00/4.00
  - Thesis: Heterogeneous Network-on-Chip Architectures for GPUs

- **Georgia Institute of Technology** — Atlanta, GA, USA
  M.S. — *Aug 2016 - Dec 2017*
  - GPA: 4.00/4.00
  - Thesis: Scalable Translation Lookup Buffer Architectures

- **Birla Institute of Technology and Sciences** — Hyderabad, India
  B.E. — *Aug 2010 - Jul 2014*
  - GPA: 8.32/10.00
  - Thesis: Organic Photovoltaics

Books

Interconnect Modeling for Homogeneous and Heterogeneous Multiprocessors 2022
Srikant Bharadwaj, Tushar Krishna
Springer

Selected Publications (Citations = 540)

C=Conference, J=Journal, T=Thesis, S=In Submission

[C.1] Bharadwaj, S., Cox, G., Krishna, T., Bhattacharjee, A.. **Scalable Distributed Shared Last-Level TLBs Using Low-Latency Interconnects**. In *Proceedings of the International Symposium on Microarchitecture (MICRO), 2018*.

[C.2] Alsop, J., Sinclair, M., Bharadwaj, S., et al. **Optimizing GPU Cache Policies for Machine Learning Workloads**. In *IEEE International Symposium on Workload Characterization (IISWC), 2019*.

[C.3] Bharadwaj, S., Yin, J., Beckmann, B., Krishna, T. **Kite: A Family of Heterogeneous Interposer Topologies Enabled via Accurate Interconnect Modeling**. In *Proceedings of the Design Automation Conference (DAC), 2020*.

[J.1] Lowe-Power, J., Bharadwaj, S., et al. **The gem5 simulator: Version 20.0+**. *arXiv, 2020*.

[J.2] Resch, S., Gutierrez, A., Bharadwaj, S., Eckert, Y., Oskin, M., Loh, G. **Accelerating Variational Quantum Algorithms Using Circuit Concurrency**. In *arxiv, 2021*.

[C.4] Bharadwaj, S., Das, S., Eckert, Y., Oskin, M., Krishna, T. **DUB: Dynamic Underclocking and Bypassing in Network-on-Chip for Heterogeneous GPU Workloads**. In *IEEE/ACM International Symposium on Networks-on-Chip (NOCS), 2021*.

[C.5] Bharadwaj, S., Das, S., Beckmann, B., Mazumdar, K., Kosonocky, S. **Predict; Don't React for Enabling Efficient Fine-Grain DVFS in GPUs**. In *rchitectural Support for Programming Languages and Operating Systems (ASPLOS), 2024*.

[C.6] Sanovar, R., Bharadwaj, S., St. Amant, R., Rühle, V., Rajmohan, S. **Lean Attention: Hardware-Aware Scalable Attention Mechanism for the Decode-Phase of Transformers**. In *Proceedings of Machine Learning and Systems (MLSys), 2025*.

[J.3] Bharadwaj, S., BR, S., Kumar, M. (2016). **Importing a zone into zone-cluster configuration**. In *Oracle Yearly Proceedings*.

[J.4] Kang, H., Bharadwaj, S., Hensman, J., Krishna, T., Ruhle, V., Rajmohan, S. **TurboAttention: Efficient Attention Approximation For High Throughputs LLMs**. *arXiv, 2024*.

[C.7] Bharadwaj, S., Krishna, T. **InC2: Design of Interconnection Systems for Composable Chiplet Architectures**. In *IEEE/ACM International Workshop on Network on Chip Architectures (NoCArc), 2024*.

[S.1] Chen, Y., Xia, M., Gong, W., Mallick, A., Bharadwaj, S., Jazbec, M., Siddiqui, S. A., Weller, A., Rühle, V. **Semi-autoregressive Decoding for Efficient LLM Inference**. In *International Conference on Learning Representations (ICLR), 2025*.

[S.2] Ruttenberg, M., Bharadwaj, S., Gutierrez, A., Eckert, Y., Oskin, M. **Application-Aware Reconfiguration of High Bandwidth Memory in GPUs**. In *Proceedings of the International Symposium on Computer Architecture (ISCA), 2025*.

[T.1] Bharadwaj, S. **Scaling Address Translation in multi-core architectures**. Georgia Institute of Technology Thesis, 2017.

[T.2] Bharadwaj, S. **Design of High-Performance and Energy-Efficient Interconnection Systems for Heterogeneous Multi-Chiplet Graphics Processing Units**. Georgia Institute of Technology Thesis, 2023.

SELECTED PATENTS (17)

G=GRANTED, F=FILED

[G.1] Bharadwaj, S., Das, S. **Routing Flits in a Network-on-Chip based on Operating States of Routers**. Patent Application No.: 16/188900.

[G.2] Bharadwaj, S. **Credit Based Flow Control Mechanism for Use in Multiple Link Width Interconnect Systems**. Patent Application No.: 16/271371.

[G.3] Bharadwaj, S. **Packet router with virtual channel hop buffer control**.

[G.4] Gutierrez, A. T., Resch, S., Eckert, Y., Loh, G. H., Oskin, M. H., Bharadwaj, S. **Quantum circuit mapping for multi-programmed quantum computers**.

[G.5] Bharadwaj, S. **Dynamic Voltage Frequency Scaling Based on Active Memory Barriers**. Patent Application No.: 16/425414.

[G.6] Gutierrez, A. T., Sangaiah, K. R., Bharadwaj, S. **VLIW power management**.

[F.1] Bharadwaj, S., et al. **Compiler Directed Fine-Grained Power Management**. Patent Application No.: 17/033000.

[F.2] Bharadwaj, S., et al. **Dynamically configurable overprovisioned microprocessor**. Patent Application No.: 17/037727.

[F.3] Shridhar, U., Ruehle, V. J., Bharadwaj, S. **Fine-Grained Selective Quantization to Maximize Hardware Resource Utilization**.

[F.4] Bharadwaj, S., et al. **Application aware dynamic tuning of DRAM parameters by leveraging thermal headroom**.

[F.5] Ganapathy, S., Eckert, Y., Gutierrez, A., Sangaiah, K. R., Bharadwaj, S. **Using chiplet-level performance information for configuring chiplets in a processor**.

[F.6] Ruttenberg, M., Bharadwaj, S., Eckert, Y., Oskin, M. H., Gutierrez, A. **Workload based tuning of memory timing parameters**.

[F.7] Resch, S., Gutierrez, A., Eckert, Y., Bharadwaj, S., Oskin, M. H. **Running Instances of a Quantum Program Concurrently on a Quantum Processor**.

[F.12] Srikanth, S., Sangaiah, K. R., Gutierrez, A. T., Bharadwaj, S., Kalamatianos, J. **VLIW Dynamic Communication**.

[F.8] Silva Tavares, J. A., Das, M., Ruehle, V. J., Bharadwaj, S. **Adaptation of task performable by pre-trained model into parallel hardware**.

[F.9] Ruttenberg, M., Bharadwaj, S., Eckert, Y., Gutierrez, A., Oskin, M. H. **Distribution of data and memory timing parameters across memory modules based on memory access patterns**.

[F.10] Sanovar, R., Ruehle, V., Bharadwaj, S. **Hardware-aware attention mechanism with dynamic workload distribution for transformer models**.

[F.11] Ehrett, W. P., Gutierrez, A., Bharadwaj, S., Sangaiah, K. R., Shukla, P., Srikanth, S., Dasika, G., Kalamatianos, J. **Semiconductor device for performing data reduction for processing arrays**.

PROFESSIONAL SERVICE

- **Reviewer**
  Conferences and Journals
  - **Submission Chair**: ISCA 2023 Industry Track
  - **Conference Reviewer**: HPCA 2025, MASCOTS 2019, ICCD 2023, MICRO 2024, ISCA 2024, ASSYST 2023, SCOPE 2025
  - **Journal Reviewer**: Transactions on Computers
  - Artifact Evaluator: ISCA 2023
  - Reviewed and evaluated more than 40 research papers.
- **Open source Contributions**
  - Maintainer of gem5, the most popular computer architecture simulator.
  - Open sourced HeteroGarnet, an interconnection system simulator, used by several industry and academic research laboratories.
  - Active contributor of Microsoft's ONNX runtime.

Selected Academic Projects

- **Intel GPU Architecture Analytical Modelling** *2016*
  Tools: OpenCL, CPU+GPU Systems Advisor: Dr. Hyesoon Kim, Georgia Institute of Technology
  - Investigating architecture for analyzing performance bottlenecks of parallel applications in CPU + GPU systems
  - Developing an analytical model for estimating the execution time of massive parallel programs
  - Developing micro-benchmarks in OpenCL to be used for ensuring accuracy in the analytical model
- **Digital String Tuner** *2014*
  Tools: Piccolo Microcontroller Funded by Texas Instruments for Analog Design Competition
  - Designed and implemented a standalone guitar tuner using Piccolo Microcontroller
  - Reached the pre-final round of the contest and was shortlisted in the top innovative designs

Honors and Awards

- **Academic Performance Scholarship** *2011*
  Angiras Foundation, USA
  - Received scholarship in recognition of outstanding academic achievements
- **Rookie of the Year Nomination** *2014*
  Oracle
  - Nominated for co-implementing a method to import instances of running Solaris instances to another cluster
- **Spotlight Award** *2019*
  AMD
  - Recognized for significant contributions to internal GPU power modelling infrastructure.
- **Spotlight Award** *2021*
  AMD
  - Recognized for developing a novel firmware technique to reduce power utilization in processing-in-memory architecture.

References

1. **Dr. Tushar Krishna**
   Associate Professor
   Georgia Institute of Technology
   Email: tushar@ece.gatech.edu
   Relationship: PhD Advisor
2. **Dr. Matthew Sinclair**
   Assistant Professor
   University of Wisconsin-Madison
   Email: sinclair@cs.wisc.edu
   Relationship: Research Collaborator