Bakshree Mishra

⌂ *bakshree.cs.illinois.edu*   ✉ *bmishra3@illinois.edu*   in *linkedin.com/in/bakshree*

RESEARCH STATEMENT

My research focuses on efficient orchestration of LLM inference across multiple heterogeneous edge and cloud devices, addressing performance, scalability, and data-movement bottlenecks at system and architectural levels.

EDUCATION

PhD in Computer Science   2021-Present
University of Illinois Urbana-Champaign   GPA: **3.92/4**
ADVISOR: Prof. Sarita Adve (*sadve@illinois.edu*)
RESEARCH INTERESTS: Computer Architecture, Hardware-Software Codesign, Machine Learning & Systems

M.Tech in Computer Science   2015-2017
National Institute of Technology, Rourkela   GPA: **9.69/10**
ADVISORS: Prof. Bansidhar Majhi (NIT Rourkela), Mr. Tarjinder Singh (Intel)

B.Tech in Computer Science and Engineering   2010-2014
College of Engineering and Technology, Bhubaneswar   GPA: **8.85/10**

PATENTS

• Boschi, G. and Makkadayil, S. and Manjunath, R. and **Mishra, B.** and Campinoti, A. "Register Fault Detector." U.S. Patent US12,373,295 B2, issued 2025.

• Singh, T. and SR, S. and Sumiran, R. and **Mishra, B.** and Makkadayil, S. and Thyagarajan, V. and Baireddy, V. "Graph Reordering and Tiling Techniques." U.S. Patent Application 17/533,976, filed 2021.

• Makkadayil, S. and Paul, S. and Saifee, S. and **Mishra, B.** and Thyagarajan, V. and Velayudha, M. and Khellah, M. and Udofia, A. "Parallel Pruning and Batch Sorting for Similarity Search Accelerators." U.S. Patent Application 17/358,495, filed 2021.

PUBLICATIONS

• **Mishra, B.** et al., *Efficient Orchestration of Language Model Inference on Edge Devices,* **submitted to OSDI/MobiSys**

• **Mishra, B.** et al., *Investigating NoC Bottlenecks on 3D Stacked GPU Architectures,* **submitted to ISCA/HPCA/ASPLOS**

• Suresh, V. and **Mishra, B.** and Jing, Y. and Zhu, Z. and Jin, N. and Block, C. and Mantovani, P. and Giri, D. and Zuckerman, J. and Carloni, L. and Adve, S. *Mozart: Taming Taxes and Composing Accelerators with Shared-Memory* In Proceedings of the 2024 International Conference on Parallel Architectures and Compilation Techniques (PACT '24). [Paper]

• **Mishra, B.** and Chakraborty, D. and Makkadayil, S. and Patil, S. D. and Nallani, B. *Hardware Acceleration of Computer Vision and Deep Learning Algorithms on the Edge using OpenCL.* EAI Endorsed Transactions on Cloud Systems, 2019 [Paper]

WORK EXPERIENCE

Graduate Research Assistant   *University of Illinois Urbana-Champaign*   May 2022 – Present
Research on heterogeneous and disaggregated accelerator systems for efficient LLM inference; developed analytical and empirical performance, communication, and energy models.

Research Associate   *AMD, Bellevue*   May 2025 – Nov 2025
Graduate (PhD) Intern   *AMD, Bellevue*   May 2024 – Aug 2024
Analyzed NoC and data-movement bottlenecks in 3D-stacked GPU architectures across a variety of workloads. Developed analytical models for LLM inference on these architectures under different orchestration strategies.

ML and IP Design Engineer   *Intel Corporation, Bangalore*   Jun 2017 – Aug 2021
Design and FPGA validation for ML accelerators targeted for production Intel platforms.
Previously: Graduate Technical Intern, Intel (May 2016 – May 2017).

TEACHING EXPERIENCE

Graduate Teaching Assistant   *University of Illinois Urbana-Champaign*   Aug 2021 – May 2022
CS 233 (Computer Architecture) and CS 225 (Data Structures); recognized among Teachers Ranked as Excellent for CS 225 (Spring 2022).

Earlier Industry Experience

Assistant System Engineer	*Tata Consultancy Services, Bhubaneswar*	Jun 2014 – Jul 2015
Summer Intern	*Tata Consultancy Services, Bhubaneswar*	Jun 2013 – Aug 2013

Worked on development of e-governance and enterprise web applications.

Projects

Efficient Orchestration of Language Model Inference on Edge Devices *2024–Present*
(Associated manuscript under double-blind review)
- Developed an analytical–empirical model for predicting LLM inference latency in a multi-device edge setup.
- Analyzed inference partitioning, KV-cache reuse, and migration trade-offs in multi-device edge deployments.
- Early results published as a workshop paper; full manuscript under peer review.

Investigating NoC Bottlenecks on 3D Stacked GPU Architectures *2025*
(Associated manuscript under double-blind review)
- Identified NoC power-density and scalability bottlenecks limiting bandwidth utilization in 3D-stacked GPU.
- Developed an analytical model capturing LLM inference latency and communication behavior under alternative interconnect organizations.
- Proposed communication algorithms inspired by spatial accelerators and performed trade-off analyses.

Mozart: Composable Acceleration with Disaggregated Accelerator Systems *2021–2024*
(PACT 2024)
- Designed a lightweight accelerator synchronization interface (ASI) reducing invocation overhead in disaggregated accelerator pipelines.
- Evaluated chained and pipelined accelerator compositions across diverse workloads on FPGA.
- Demonstrated performance trade-offs with monolithic and scalability of disaggregated acceleration.

Selected Research Projects at Intel *2016–2021*
- Designed and deployed OpenCL-based FPGA accelerators for real-time computer vision workloads, achieving up to 5× throughput improvement; work was presented at IEEE WinTechCon and published in *EAI Endorsed Transactions on Cloud Systems*, 2019.
- Developed functional safety IP including RTL design, waveform analysis, and verification; resulted in filed and granted U.S. patents.

Select Awards and Honors

• Among Teachers Ranked as Excellent for CS225	Spring 2022
• Co-authored 3 accepted internal papers, presented a demo at Intel DTTC, Portland, OR	2019-2021
• Multiple Intel Divisional and Departmental Recognition Awards	2017-2020
• Ranked 2nd among all Masters (~110) students in CS Department at NIT Rourkela	2017
• CET Merit Scholarship (Undergrad scholarship 2010-2014)	2010
• National Talent Search Examination Scholar and CBSE Xth board rank-holder in State	2008

Technical Skills

- Analytical and empirical performance modeling for ML and accelerator workloads
- Power profiling, DVFS analysis, and application benchmarking on heterogeneous systems
- RTL and SystemC-based accelerator design, waveform analysis, and FPGA deployment
- OpenCL/HLS-based prototyping and end-to-end system evaluation

Service, Mentoring, & Professional Development

• Planning committee member of GradWCS, hosting monthly Faculty/Student luncheons	2023 - present
• Co-started and run weekly coffee meet-ups for women in comp. arch in CS and ECE	2022 - present
• Completed formal mentor training through UR2PhD (Undergraduate Research to PhD)	2025
• Co-chair of the Systems and Architecture session for the 19th CSL Student Conference	2024
• Named one of Top 50 Volunteers in Intel India for community service initiatives	2020
• Recipient of Intel Seed Grant for renovation of nurses' dining hall at Karunashraya Hospice	2019