HARSH SHARMA

harshari.github.io • harsh.sharma@wsu.edu

RESEARCH SUMMARY

My general research interests are at the intersection of artificial intelligence (AI) and computing system with a focus on exploiting their synergistic strengths: AI for design and optimization of computing systems, and the design of optimized computing systems for AI applications. The current focus of my research is on AI-driven design and Optimization of *Chiplet-based Systems* for enabling high-performance and low-energy computing for various applications including training/inference of large AI models. Specific topics include:

- Enabling server-scale system design with low-latency interconnect networks.
- Hardware and software co-design to create chiplet systems for training/inference with large AI models including CNNs, GNNs, and Transformers.
- Design of high-performance and energy-efficent manycore systems to overcome Moore's law.
- Design of defect-aware chiplet-based systems to reduce carbon footprint at scale.
- Accelerating the design of robust, reliable, and environmentally sustainable paradigms.

EDUCATION

Ph.D. Candidate, Computer Engineering, 3.93 GPA **2021–Present**
Advisors: Partha Pratim Pande & Janardhan Rao Doppa
Washington State University *Pullman, Washington*

Bachelor of Engineering, Electronics and Communication Engineering **2017–2021**
NSIT, Delhi University *New Delhi, India*
Department ranker (Top 5%)

INDUSTRIAL EXPERIENCE

Machine Learning Research Intern **June 2020–December 2020**
Lenskart.com *New Delhi, India*
Developed AR tools with vision model to boost online sales by 35% during COVID19 Pandemic.

AWARDS AND HONORS

- Harvard Scholar at HPAIR Conference, Kazakhstan. Technology Track (**top 1%**)
- Best Paper Award at ACM/IEEE Embedded Systems Week Conference, 2023
- Best Paper Award at ACM/IEEE Embedded Systems Week Conference, 2022 [†]
- ACM SIGDA Richard Newton Young Fellowship, 2022

SELECTED PUBLICATIONS

1. **[Best Paper Award] Harsh Sharma**, Lukas Pfromm, Rasit Topaloglu, Janardhan Rao Doppa, Umit Y. Ogras, Ananth Kalyanraman, Partha Pratim Pande. Florets for Chiplets: Data Flow-aware High-Performance and Energy-efficient Network-on-Interposer for CNN Inference Tasks. *ACM Transactions on Embedded Computing Systems, Hamburg*, 2023.
2. **Harsh Sharma**, Lukas Pfromm, Janardhan Rao Doppa, Umit Y. Ogras, Partha Pratim Pande. A Heterogeneous Chiplet Architecture for Accelerating End-to-End Transformer Models. *Design Automation and Test in Europe DATE*, 2024. Under Review.
3. **Harsh Sharma**, Sumit K. Mandal, Janardhan Rao Doppa, Umit Y. Ogras, Partha Pratim Pande. Achieving Datacenter-scale Performance through Chiplet-based Manycore Architectures. *Design Automation and Test in Europe DATE*, 2023.

[†] `https://school.eecs.wsu.edu/2022/10/14/cases-best-paper-award/`

4. [Best Paper Award] **Harsh Sharma**, Sumit K. Mandal, Janardhan Rao Doppa, Umit Y. Ogras, Partha Pratim Pande. SWAP: A Server-Scale Communication-Aware Chiplet-Based Manycore PIM Accelerator. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Phoenix/Shanghai*, 2022.
5. **Harsh Sharma**, Lukas Pfromm, Janardhan Rao Doppa, Umit Y. Ogras, Partha Pratim Pande. Network-on-Interposer Design for CNN Inferencing in Presence of Defective Chiplets. IEEE Transactions on Very Large Scale Integration (VLSI) Systems (TVLSI). Under Review.
6. **Harsh Sharma**, Dhananjay Gadre, Sangeeta Gadre, Smriti Srivastava. Science on a stick: An experimental and demonstration platform for learning several physical principles. *American Journal of Physics*, 2022.

SELECTED PROFESSIONAL AND OUTREACH ACTIVITIES

Conference and Invited Talks

- SWAP: A Server-scale Communication aware Chiplet-based PIM Accelerator at ESWEEK 2022.
- Achieving Datacenter-scale Performance through Chiplet-based Manycore Architectures DATE23.
- Florets for Chiplets: Data Flow-aware High-Performance and Energy-efficient Network-on-Interposer for CNN Inference Tasks at TUHH Hamburg, Germany- 2023.
- Talk on *AI-Driven Design and Optimization of Chiplet-based Manycore Systems for Server-Scale Applications* at WSU Pullman-2023.
- Talk on *AI-Driven Design and Optimization strategies for more Moore* at NSIT Delhi (Virtual)-2023.
- Talk on *Accelerating the Future of Electronics* at Boston University (Virtual)-2023. [‡]

Reviewer

- ESWEEK 2022, ICCAD 2023, DAC 2022, DAC 2023, DATE 2022

SKILLS

- **Programming Languages.** Python, Bash, C/C++, HTML/CSS, LaTeX, Java, MATLAB
- **Tools/Packages.** Git, SQL, PyTorch, TensorFlow, Python data science tools

[‡]Based on `https://medium.com/@harshari/accelerating-the-future-of-electronics-e23cc42d9d39`