Ishwar Singh Bhati

ishwar.bhati02@gmail.com | +1 503-810-0035 | 1550 Broadstone Pkwy, Apt 1407, Folsom, CA, USA, 95630

Summary

Ph.D. in Computer Engineering with over 15 years of experience in hardware and software co-optimization, high-performance computing, and memory systems architecture. A prolific innovator with 10+ granted/filed US patents and a strong publication record in top-tier conferences (ISCA, VLDB, TMLR). Specialized expertise in optimizing AI/ML applications by leveraging advanced hardware features like AVX/SIMD and AMX on Intel® Xeon to develop innovative solutions for scalable vector search and deep learning architectures. Proven background in ASIC design and verification, including performance modeling and FPGA prototyping for complex memory controllers and processors (DDR2/3, SPARC). Proficient in **C/C++**, with a working knowledge of **Python** and **Bash**.

Research and Professional Experience

Intel Corporation, AI Research Engineer/Scientist | December *2015 – Present*

- Codeveloped a scalable, high-performance vector search library for Intel hardware (https://github.com/IntelLabs/ScalableVectorSearch). **My primary contributions include:**
  - Implemented a highly optimized Inverted File (IVF) index on Intel Xeon, leveraging **AMX** instructions for matrix multiplication.
  - This IVF index build performs on par with a ~5-10x larger GPUs and delivers a solution orders of magnitude faster than existing implementations like **FAISS**.
  - Designed smart software prefetching and an efficient dimensionality algorithm to reduce memory bandwidth requirements and minimize memory latency dependencies.
- Authored and published papers in top-tier conferences, including **VLDB'23** and **TMLR'24**, with the latter receiving a featured certification and selected for a poster presentation at **ICLR'25**.
- Awarded three Divisional Recognition Awards (**DRA**) in 2025 for significant contributions to the scalable vector search library.
- Designed and optimized efficient compute and memory architectures for emerging Deep Learning applications, developing highly optimized GPU kernels for workloads such as **DLRM**, Graph Neural Networks, and MLP layers. This work included projecting performance for future products and proposing new core, cache, and memory features, for which I received a Divisional Recognition Award (**DRA**) in 2018.
- Pioneered research on architectural techniques for **STTRAM-based Last-Level Cache** to enhance memory system performance, with a published paper at **ISCA 2018**.
- Developed a novel, adaptively changing out-of-order (OoO) core width/port mechanism based on simple heuristics to improve efficiency.
- Designed a memory-aware reordering technique to achieve high bandwidth efficiency in GPU architectures.

Oracle (formerly SUN Microsystem), Senior Hardware Engineer | *June 2014 – December 2015*

- Worked on performance modeling and design space exploration for SPARC processors, with a focus on memory controller and database accelerator modules.

Intel Corporation, Graduate Intern| Hillsboro (June 2013 – August 2013)

- Developed a memory model that was 10x faster than a cycle-accurate simulator while maintaining high accuracy, enabling early-stage system design and power analysis.

University of Maryland, Research Assistant, Dept. of ECE | September 2010 – May 2014

- As part of my Ph.D. research, I conducted foundational research on scalable and energy-efficient DRAM refresh mechanisms, resulting in **two US patents** and highly cited publications in **ISCA-2015**, *Transactions on Computers, 2015*, and **ISLPED-2013**. This work has over **450+ independent citations** and is used as a foundation for other industry research.
- Co-designed a parameterized simulation infrastructure to study emerging **Non-volatile Memory (NVM)** technologies, with findings published in the **Intel Technology Journal (ITJ'13)**.
- Implemented reliable full-system simulation infrastructure using **MARSSx86** and **DRAMSim2** to reduce variability and ensure accurate results.

LSI Corporation, Senior ASIC Engineer | *January 2009 – July 2010*

- Co-implemented DDR2/3 memory controller and its PHY layer using 65nm process technology.
- Developed the optimized DDR3 training sequence and write leveling algorithm.
- Created SystemVerilog and VMM-based automated test benches.

Nevis Networks, ASIC Design and Verification Engineer | *May 2007 – January 2009*

- Led the SystemC modeling and Full-Chip Verification environment integration.
- Performed entire FPGA prototyping of memory controller using Xilinx's Vertex-4 based board and created synthesizable verification code.

STMicroelectronics, Design Engineer | *June 2006 – May 2007*

- Modeled, performed RTL, and verified multiple modules in a Wireless USB Medium Access Control (MAC) Chip design.

Nevis Networks, ASIC Design and Verification Engineer | *July 2005 – June 2006*

- Involved in the design and verification of the DRAM Control module in a 96-core Network Processor Chip.

Education

PhD in Computer Engineering — Spring, 2014
University of Maryland, College Park, MD — GPA: 4.0/4.0
Advisor: Prof. Bruce Jacob (http://www.ece.umd.edu/~blj/)

B.Tech in Electronics and Communication Engineering — Spring, 2005
Indian Institute of Technology (IIT), Guwahati, India — GPA: 8.6/10

Selected Publications (Google Scholar: https://scholar.google.com/citations?user=DppJvVgAAAAJ&hl=en)

Mariano Tepper, Ishwar Singh Bhati, Cecilia Aguerrebere, Mark Hildebrand, Ted Willke, "LeanVec: Search your vectors faster by making them fit", *Transactions on Machine Learning Research (TMLR), 2024.* **(selected as featured certification and presented as poster in ICLR 2025)**

Mariano Tepper, Ishwar Singh Bhati, Cecilia Aguerrebere, Ted Willke *"GleanVec: Accelerating vector search with minimalist nonlinear dimensionality reduction", arXiv, 2024.*

Cecilia Aguerrebere, Mark Hildebrand, Ishwar Singh Bhati, Theodore Willke, Mariano Tepper, *"Locally-Adaptive Quantization for Streaming Vector Search", arXiv, 2024.*

Cecilia Aguerrebere, Ishwar Bhati, Mark Hildebrand, Mariano Tepper, Ted Willke, *"Similarity search in the blink of an eye with compressed indices", Proceedings of the VLDB Endowment, 2023.*

Ishwar Bhati, Udit Dhawan, Jayesh Gaur, Sreenivas Subramoney, and Hong Wang, *"MARS: Memory Aware Reordered Source", arXiv:1808.03518, August 2018.*

Kunal Korgaonkar, Ishwar Bhati, Huichu Liu, Jayesh Gaur, Sasikanth Manipatruni, Sreenivas Subramoney, Tanay Karnik, Steven Swanson, Ian A. Young, and Hong Wang, "Density Tradeoffs of Non-Volatile Memory as a Replacement for SRAM based Last Level Cache," *Proc. 45th International Symposium on Computer Architecture (ISCA 2018). Los Angeles, CA, June 2018.*

Kaushik Vaidyanathan, Daniel H Morris, Uygar E Avci, Ishwar S. Bhati, Lavanya Subramanian, Jayesh Gaur, Huichu Liu, Sreenivas Subramoney, Tanay Karnik, Hong Wang, and Ian A Young. "Overcoming interconnect scaling challenges using novel process and design solutions to improve both high-speed and low-power computing modes," *Electron Devices Meeting (IEDM), 2017 IEEE International*

Ishwar Bhati, Zeshan Chishti, Shih-Lien Lu, and Bruce Jacob, "Flexible auto-refresh: Enabling scalable and energy-efficient DRAM refresh reductions," *Proc. 42nd International Symposium on Computer Architecture (ISCA 2015)*. Portland, OR, June 2015.

Ishwar Bhati, Mu-Tien Chang, Zeshan Chishti, Shih-Lien Lu, and Bruce Jacob, "DRAM Refresh Mechanisms, Penalties, and Trade-Offs," *IEEE Transactions on Computers*, vol. 64, 2015.

Ishwar Bhati, Zeshan Chishti, and Bruce Jacob, "Coordinated refresh: Energy efficient techniques for DRAM refresh scheduling," *Proc. 2013 International Symposium on Low Power Electronics and Design (ISLPED 2013)*. Beijing China, September 2013.

Jim Stevens, Paul Tschirhart, Mu-Tien Chang, Ishwar Bhati, Peter Enns, James Greensky, Zeshan Chishti, Shih-Lien Lu, and Bruce Jacob, "An Integrated Simulation Infrastructure for the Entire Memory Hierarchy: Cache, DRAM, Nonvolatile Memory, and Disk," *Intel Technology Journal (ITJ)*, vol. 17, no. 1, 2013.

Patents

Maria Cecilia Aguerrebere Otegui, Ishwar Bhati, Mark Hildebrand, Mariano Tepper, Theodore Willke, "Locally-adaptive vector quantization for similarity search", US Patent application filed, 2024

Mark Hildebrand, Mariano Tepper, Maria Cecilia Aguerrebere Otegui, Ishwar Singh Bhati, Theodore Willke, "TURBO LOCALLY-ADAPTIVE VECTOR QUANTIZATION FOR HIGH-PERFORMANCE DISTANCE COMPUTATIONS", US Patent application filed, 2024

Maria Cecilia Aguerrebere Otegui, Ishwar Singh Bhati, Mark Hildebrand, Mariano Tepper, Theodore Willke, "MULTI-MEANS LOCALLY-ADAPTIVE VECTOR QUANTIZATION FOR MEMORY EFFICIENT AND HIGH-PERFORMANCE STREAMING SIMILARITY SEARCH", US Patent application filed, 2024

Mariano Tepper, Ishwar Singh Bhati, Maria Cecilia Aguerrebere Otegui, Mark Hildebrand, Theodore Willke, "Dimensionality reduction technology to accelerate high-dimensional vector searches and index construction", US Patent application filed, 2024

Supratim Pal, Sasikanth Avancha, Ishwar Bhati, Wei-Yu Chen, Dipankar Das, Ashutosh Garg, Chandra S Gurram, Junjie Gu, Guei-Yuan Lueh, Subramaniam Maiyuran, Jorge E Parra, Sudarshan Srinivasan, Varghese George, "Instructions and logic for vector multiply add with zero skipping", US Patent granted, 2023

Ishwar Bhati, Udit Dhawan, Jayesh Gaur and Sreenivas Subramoney, "Memory aware reordered source", US patent granted, 2020

Ishwar Bhati, Huichu Liu, Jayesh Gaur et al., "Write congestion aware bypass for non-volatile memory, last level cache", US patent granted, 2019

Kunal Korgaonkar, Ishwar Bhati, Huichu Liu et al., "Method and apparatus for reducing write congestion in non-volatile memory based last level caches", US patent granted, 2018

Ishwar Bhati, Zeshan Chishti, and Shih-Lien L. Lu, "Techniques to Reduce Memory Cell Refreshes for a Memory Device", US patent granted, 2016

Ishwar Bhati and Zeshan Chishti, "Coordinating Power Mode Switching and Refresh Operations in a Memory Device," US patent granted, 2015

Ph.D. Thesis

Ishwar Bhati, "Scalable and Energy-Efficient DRAM Refresh Techniques,"Ph.D. thesis, May 2014.

Technical Reports

B. Jacob, Ishwar Bhati, M.-T. Chang, P. Rosenfeld, J. Stevens, P. Tschirhart, Z. Chishti, S.-L. Lu, J. Ang, D. Resnick, and A. Rodrigues, "A Journaled, NAND-flash main-memory system," University of Maryland Systems and Computer Architecture Group Technical Report, 2014.

Mu-Tien Chang, Ishwar Bhati, Jim Stevens, Paul Tschirhart, Peter Enns, Daniel Gerzhoy, Zeshan Chishti, James Greensky, Shih-Lien Lu, and Bruce Jacob, "Producing Reliable Full-System Simulation Results: A Case Study of CMP with Very Large Caches," Institute for Systems Research (ISR) Technical Report UMD-ISR-TR-2012-07, 2012.

References available upon request.