Ishwar Singh Bhati

Research Scientist, Intel Labs, Hillsboro, Oregon

E-mail: ishwar.bhati02@gmail.com* | *Mobile: +1 503-810-0035

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

Areas of Interest

Computer Architecture, Memory Systems, Energy Efficient Architecture, Non-volatile Memory, Deep Learning Architecture, Hardware/Software co-optimization, High-performance computing, Parallel programming

Research Summary

Emerging Architecture Research:

- Current work focuses on co-optimizing hardware and software for AI/ML applications. Part of the team developing scalable, high-performance similarity search library for Intel hardware (https://github.com/IntelLabs/ScalableVectorSearch). Published papers in VLDB’23 and TMLR’24 (**selected as featured certification**)
- Designed efficient compute and memory architecture for emerging applications like Deep Learning. Received a Divisional Recognition Award (DRA) in 2018 for this work.
- Developed optimized GPU kernels for ML/DL workloads like Deep Learning Recommendation Model (DLRM), Graph Neural Networks, MLP layers used in other Vison and NLP applications. Projected performance based on detailed as well as high level simulators for future products and recommended core/cache/memory features.
- Optimized DLRM embedding kernels on current Xeon machines and proposed new features for future products.
- I was rewarded and recognized internally multiple times for the above work.

Micro-architecture Research:

- NVM-based LLC: Proposed novel techniques to mitigate long NVM write latency (published in ISCA-2018)
- Memory aware reordered source (MARS): to reshape the memory traffic for efficient memory bandwidth.
- Adaptive Width Aware Core (AWAC): we used simple heuristics to intelligently provision resources in the core dynamically (part of work featured in IEDM'2017)

Scalable DRAM Refresh:

- Comprehensive evaluation and survey of DRAM refresh mechanisms, trade-offs, and penalties. We also clarify prevalent confusion with refresh options and timings available in JEDEC-specified DDR devices. This study was published in Transactions on Computers, 2015 (Weblink).
- Proposed simple modification in DRAM device to enable *refreshes reduction* with optimized auto-refresh commands rather than in-efficient row-level refresh commands. This work was published in ISCA-2015 (Weblink).

Energy Efficient Memory:

- Proposed novel techniques to simultaneously minimize two important types of DRAM energy components: *background and refresh*. Our novel schemes called “*Coordinated Refresh*” schedule refresh operations and power down modes in such a way that energy consumption is reduced while improving performance. This work was accepted in ISLPED-2013 for presentation (Weblink).

High Capacity Memory:

- Co-designed *parameterized simulation* infrastructure to study various emerging Non-volatile Memory (NVM) technologies, organization, and latencies. We simulated a range of workloads to understand performance tradeoffs when NVM is used as part of the memory hierarchy. This study was published in the Intel Technology Journal (ITJ'13) (Weblink) as well as in a Tech report (Weblink).

Accurate Memory Simulations:

- Designed a set of techniques when applied in a full-system simulator to give *reliable, accurate and fewer variable* results. Our techniques implemented on MARSSx86 integrated with DRAMSim2 for case study to show reduced variability in simulations. This work was published in a Tech report (Weblink).

Work Experience

Research Scientist, Intel Labs, Intel, Bangalore (December 2015 – April 2022), Hillsboro (May 202 – present)

- Focusing on architectures for newer applications like Deep Learning.
- Research on design and architectural techniques for STTRAM-based LLC.
- Developed adaptively changing OoO core width/ports based on simple heuristics.
- Designed memory-aware reordering technique for achieving high bandwidth efficiency in GPUs.

Senior Hardware Engineer, Oracle (formerly SUN Microsystem), Santa Clara (June 2014 – December 2015)

- Worked on performance modelling, projection, and design space exploration of SPARC processors.
- Responsible for modeling and maintaining memory-controller and database-accelerator modules.

Research Assistant, Memory Systems Research Lab, Dept. of ECE, in University of Maryland (Sep 2010 – May 2014)

- Proposed novel DRAM refresh and energy-efficient mechanisms.
- Research on applications of persistent memory.
- Implemented reliable full-system simulation infrastructure.

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

- Quantified speed versus accuracy tradeoffs in memory modelling at several levels of abstraction (constant, analytical, queue-based, detailed, etc.)
- Implemented and integrated a memory model, which is 10x faster than the cycle-accurate DRAMSim2 and is within 10% of accuracy.
- These models are targeted to obtain the approximate timing and power behavior of a system early in its design phase
- Technical Mentor: Emily Shriver, Strategic CAD Labs (SCL)

Senior ASIC Engineer, LSI Corporation, India (Jan 2009 - July 2010)

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

ASIC Design and Verification Engineer, Nevis Networks, India (July 2005 - June 2006, May 2007-Jan 2009)

- Involved in the design and verification of the DRAM Control module in a 96-core Network Processor Chip.
- Performed entire *FPGA prototyping* of memory controller using Xilinx's Vertex-4 based board and created synthesizable verification code.
- Led the SystemC modeling and Full-Chip Verification environment integration.

Design Engineer, STMicroelectronics, India (June 2006 - May 2007)

- Responsible for modelling, *RTL, and verification* of a couple of modules in Wireless USB Medium Access Control (MAC) Chip design.

Summer Intern, Kyungpook National University (KNU), Daegu, South Korea (May 2004 - July 2004)

Selected Publications

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)**

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

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.

Personal Information

Date of Birth: 2nd February 1983 Contact Address: 15317 NW Twoponds Dr, Portland, Oregon, USA, 97229
Contact Number: +1 503-810-0035, email: ishwar.bhati02@gmail.com;