Kaustubh Shivdikar

shivdikar.k [at] northeastern.edu | wiki.kaustubh.us | kaustubhcs | kaustubhcs

Availability: May 2024 onwards

SUMMARY

Computer Architecture Research Engineer enhancing GPU and CGRA microarchitecture, targeting graph computations and homomorphic encryption. Diverse experience ranging from GPU cluster computing to designing architecture simulators.

CORE COMPETENCIES

- **Machine Learning:** Neural Networks Acceleration (DNN on GPU), Secured Large Language Models (FHE on LLM)
- **Architecture Simulator Design:** Sniper (x86), NaviSim (AMD GPU), NeuraSim (CGRA), SST (Accelerators), FHESim (FPGA)
- **Homomorphic Encryption:** CKKS Scheme, Bootstrapping, Number Theoretic Transforms (NTT), Modular Reduction
- **Programming Languages & Frameworks:** C, C++, CUDA, ROCm, OpenCL, PyTorch, TensorFlow, Microsoft SEAL, OpenFHE

EDUCATION

Northeastern University, Boston, MA **Expected Graduation April 2024**
MS and Ph.D. *in Electrical and Computer Engineering (GPA: 3.95/4.0)*
Major: Computer Engineering, **Ph.D. Dissertation**: Enabling Accelerators for Graph Computing

Veermata Jijabai Technological Institute (VJTI), University of Mumbai **May 2016**
BTech *in Electrical Engineering with distinction*

INDUSTRY EXPERIENCE

Intel, Parallel Computing Lab, Santa Clara, CA **May 2020 - Dec 2020**
Computer Architecture Research Intern (Mentor: Fabio Checconi)

- Identified key memory bottlenecks leading to compute underutilization using the Sniper simulator
- Integrated vectorized memory instructions to accelerate sparse graph workloads
- Amortized memory access latency of graph workloads by pipelining computations

Intel, Parallel Computing Lab, Santa Clara, CA **May 2019 - Dec 2019**
Computer Architecture Research Intern (Mentor: Pradeep Dubey)

- Designed a hash-based Sparse GEMM kernel for Intel's new microarchitecture
- Developed a novel compressed sparse matrix data layout format that allows parallel writes
- Integrated near-memory processing with custom atomic memory instructions

Omron Adept Technologies, Amherst, NH **May 2018 - Dec 2018**
GPU Research Intern (Mentor: George Paul)

- Designed a massively parallel graph traversal algorithm using Breadth-First-Search for robot path planning
- Implemented real-time path planning on the Intel HD graphics platform using OpenCL

MIT Media Lab, Cambridge, MA **March 2015 - Dec 2015**
India Initiative Research Intern, Department of Urban Studies and Planning (Mentor: Kira Intrator)

- Developed LSTM models to analyze housing crisis using LIDAR and satellite data *[Published in Cityscape Journal]*

RESEARCH EXPERIENCE

NUCAR, Northeastern University Computer Architecture Research Lab (Advisor: David Kaeli)

NeuraChip (Decoupled CGRA GNN Accelerator)

- Designed a cycle-accurate multi-threaded CGRA simulator to benchmark on-chip network topologies
- Accelerated GNNs by mapping to CGRA compute units **(16x speedup over A100 & MI100 GPU)** *[Submitted ISCA 2024]*

OpenFHE (Fully Homomorphic Encryption)

- Extended AMD GPU's CDNA microarchitecture to accelerate FHE **(14x speedup over V100)** *[Published MICRO 2023]*
- Accelerated FHE-based large language models (LLM) on AMD MI100 GPUs

Correlation of Luggage and Specific Passengers (CLASP)

- Developed a CNN-based model for real-time, multi-camera, people and object tracking
- Improved model inference performance on NVIDIA GPU cluster exploiting model-level parallelism and load balancing
- Deployed the system at a simulated airport checkpoint with real-time system performance tracking

PUBLICATIONS

[1] H. Peng, X. Xie, **K. Shivdikar**, A. Hasan, S. Huang, O. Khan, C. Ding, and D. Kaeli, "**MaxK-GNN: Towards Theoretical Speed Limits for Accelerating Graph Neural Networks Training**," Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2024) *[Link]*

[2] **K. Shivdikar**, Y. Bao, R. Agrawal, M. Shen, G. Jonatan, E. Mora, A. Ingare, N. Livesay, J.L. Abellán, J. Kim, A. Joshi, and D. Kaeli, "**GME: GPU-based Microarchitectural Extensions to Accelerate Homomorphic Encryption**," 56th IEEE/ACM International Symposium on Microarchitecture (MICRO 2023) *[Link]*

[3] N. Livesay, G. Jonatan, E. Mora, **K. Shivdikar**, R. Agrawal, A. Joshi, J.L. Abellán, J. Kim, and D. Kaeli, "**Accelerating Finite Field Arithmetic for Homomorphic Encryption on GPUs**," (IEEE Micro 2023) *[Link]*

[4] **K. Shivdikar**, G. Jonatan, E. Mora, N. Livesay, R. Agrawal, A. Joshi, J.L. Abellán, J. Kim, and D. Kaeli, "**Accelerating Polynomial Multiplication for Homomorphic Encryption on GPUs**," IEEE International Symposium on Secure and Private Execution Environment Design (SEED 2022) *[Link]*

[5] M. Jayaweera, **K. Shivdikar**, Y. Wang, and D. Kaeli, "**JAXED: Reverse Engineering DNN Architectures Leveraging JIT GEMM Libraries**," International Symposium on Secure and Private Execution Environment Design (SEED 2021) *[Link]*

[6] **K. Shivdikar**, "**SMASH: Sparse Matrix Atomic Scratchpad Hashing**," Master's Thesis, Northeastern University, 2021 *[Link]*

[7] T. Baruah, **K. Shivdikar**, S. Dong, Y. Sun, S.A. Mojumder, K. Jung, J.L. Abellán, Y. Ukidave, A. Joshi, J. Kim, and D. Kaeli, "**GNNMark: A Benchmark Suite to Characterize Graph Neural Network Training on GPUs**," IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2021) *[Link]*

[8] C. Bunn, H. Barclay, A. Lazarev, F. Yusuf, J. Fitch, J. Booth, **K. Shivdikar**, and D. Kaeli, "**Reproducing performance of a multi-physics simulations of the Tsunamigenic 2004 Sumatra Megathrust earthquake on the AMD EPYC 7551 architecture**," International Conference for High Performance Computing, Networking, Storage, and Analysis (SC 2019) *[Link]*

[9] **K. Shivdikar**, K. Paneri, and D. Kaeli, "**Speeding up DNNs using HPL based Fine-grained Tiling for Distributed Multi-GPU Training**," Boston Area Architecture Workshop (BARC 2018) *[Link]*

[10] K. Intrator and **K. Shivdikar**, "**Missing 'Middle Scenarios': Uncovering Nuanced Conditions in Latin America's Housing Crisis**," Habitat III Symposium (Cityscape 2017) *[Link]*

[11] S. Thakkar, **K. Shivdikar**, C. Warty, "**Video Steganography Using Encrypted Payload for Satellite Communication**," IEEE Aerospace Conference, (AeroConf 2017) *[Link]*

[12] M. Srinivasan, A. Kak, **K. Shivdikar**, and C. Warty, "**Dynamic Power Allocation using Stackelberg Game in a Wireless Sensor Network**," IEEE Aerospace Conference (AeroConf 2016) *[Link]*

[13] **K. Shivdikar** and A. Kak, "**Automatic Image Annotation using a Hybrid Engine**" 12th IEEE Annual India Council International Conference (INDICON 2015) *[Link]*

AWARDS & TALKS

Travel Grant Award **MICRO 2023**
Awarded a Student Travel Grant to present our work, GME, at MICRO 2023 in Toronto

RISE: Research, Innovation, Scholarship, and Entrepreneurial Expo, 2018 **Graduate Innovator Award**
Awarded for Pi-Tiles: A Distributed Matrix-Multiplication Approach over Tiled Data using 64 Raspberry Pis

RISE: Research, Innovation, Scholarship, and Entrepreneurial Expo, 2019 **Best Poster Award**
Awarded for The Prime Hexagon

TEDx Talks

- Magic Pillow: Enhancing bed-time stories with augmented reality *[Link]* **TEDx BITSHyderabad**
- GoodBye: An afterlife digital vault service *[Link]* **TEDx SIESGST**

PROFESSIONAL SERVICE

IEEE International Symposium on High-Performance Computer Architecture (HPCA - 2023) **Submission Chair**
Developed an algorithm for matching submitted papers to the reviewer's expertise

REU Pathways Mentorship Program **Mentor**
Mentored three high school students through research projects on parallel processing applications in graph computing

Basics of GPU Programming using CUDA **Course Instructor**
Instructed a parallel programming class for undergraduate and graduate students for the Spring 2020 semester

Student Cluster Contest (Super Computing Conference 2017 & 2018) **Mentor**
Mentored the Northeastern team for Student Cluster Contest at Super Computing Conference 2017 and 2018

GRANTS AWARDED

Collaborative Research: CSR: Medium: Architecting GPUs for Practical Homomorphic Encryption-based Computing **July 2023**
NSF CISE CNS Medium Core Programs, Grant number: 2312275 *Awarded: $600,000*

Integrating GPUs and FPGAs with Processing-in-Memory to Accelerate Fully Homomorphic Encryption **September 2023**
NSF CHEST IUCRC, Project: P18-22, WP30_23 *Awarded: $98,270*