Sarabjeet Singh, Ph.D.

sarab@cs.utah.edu | San Jose, CA | cs.utah.edu/~sarab | LinkedIn | Google Scholar

CORE RESEARCH AREAS

Computer Architecture	Memory Systems Design	Privacy-Preserving Computing
Memory & Dataflow Optimization	Hardware/Software Co-Design	Trusted Execution / Confidential Computing
FHE Acceleration	Multi-GPU / Distributed Encrypted Computation	AI/LLM Inference Hardware

PROFESSIONAL EXPERIENCE

Samsung Semiconductor – Sr. Engineer, Performance Architecture
Sep 2026 – Present

San Jose, CA | Manager: Hyojong Kim

- Work on the **hardware performance simulator** for next-generation AI inference hardware.
- Study how trends in **large language models (LLMs)** change hardware metrics, to drive architectural decisions.

Duality Technologies – Scientist
Jul 2024 – May 2026

Manager: Yuriy Polyakov

DARPA DPRIVE – TREBUCHET: 12 nm FHE Accelerator ASIC

$15M DARPA effort (Apr 2021 – Nov 2025, completed): custom 12 nm ASIC for homomorphic machine and deep learning, carried to fully verified RTL and a top-level floorplan (176 mm², 1 GHz timing closure). Custom ISA, microcode scheduler, and node-array architecture; validated on encrypted CNN training and inference and AES transciphering.

Co-author, "Accelerating FHE to Enable Privacy-Preserving Deep Learning," GOMACTech 2025, the U.S. Government Microcircuit Applications & Critical Technology Conference (see Publications).

- Designed and implemented components of TREBUCHET, a custom 12 nm FHE accelerator ASIC for encrypted machine and deep learning.
- Developed **microcode generators** for FHE kernels, including kernel-to-hardware mapping, dataflow optimization, and functional-unit design evaluation.
- Balanced kernels on a **compute-memory roofline** through functional-unit design choices and careful microcode generation; investigated NTT variants with loop optimizations and on-the-fly twiddle generation.
- Developed a novel **KeySwitching dataflow** to improve data reuse and managed technical collaboration with the NYU subcontractor team led by Prof. Brandon Reagen for simulator development.

ARPA-H SHARE – Sharing Data for Rapid Collaboration on Encrypted Data

ARPA-H award to Duality Technologies (Scalable Solutions Office; up to $6.3M; started September 2024).

- Technical Lead for building a **deployment-grade platform** for privacy-preserving healthcare analytics across distributed patient-data owners.
- Designed secure algorithms for statistical and ML workloads while balancing **security, latency, and usability trade-offs.**
- Integrated FHE, federated learning, differential privacy, and post-quantum cryptography to support private model training and evaluation.
- Led development of an open-source-ready platform intended for real-world healthcare collaboration, coordinating technical work with ICF and Dana-Farber Cancer Institute.

NVIDIA Research – Research Co-op
May 2022 – Aug 2022

Project: Homomorphic Encryption at Modern GPU Datacenter Scale | Mentor: Michael Sullivan

- Explored mappings of FHE kernels on **multi-GPU architectures** to maximize shared-memory reuse and reduce communication overheads.
- Evaluated acceleration opportunities using **in-switch compute, NVSwitch, AllReduce/Broadcast/Transpose, and kernel fusion.**

AMD Research – Research Co-op
May 2020 – Aug 2020

Project: Processing-in-Memory | Mentor: John Kalamatianos

- Profiled **BERT/NLP workloads** to identify code segments suitable for processing-in-memory acceleration.
- Automated a PIM code-generation flow translating **x86 instructions to PIM instructions, memory requests, and PIM/DRAM commands.**

University of Utah – Graduate Research Assistant — Aug 2019 – Jun 2024

Advisor: Prof. Rajeev Balasubramonian | Utah Arch Research Group

- Conducted Ph.D. research on accelerating modern cryptography with a focus on **data movement, memory systems, and hardware/software co-design.**
- Designed and evaluated accelerator and memory-system techniques for FHE, lattice-based cryptography, sparse neural networks, and cache/memory hierarchy optimization.
- Published research across **IEEE S&P, HPCA, IEEE Micro, ACM TACO, IISWC, ICPE, and ACM JETC.**

Ashoka University – Junior Research Fellow — Sep 2018 – Jun 2019

Project: Design of Main Memory Architectures for Next Generation Datacenter Servers | Supervisor: Prof. Manu Awasthi

Indian Institute of Technology (IIT) Gandhinagar – Junior Research Fellow — Jan 2018 – Aug 2018

Project: Design of Main Memory Architectures for Next Generation Datacenter Servers | Supervisor: Prof. Manu Awasthi

Hexagon Capability Center India (HCCI), Hyderabad – Software Analyst — Aug 2017 – Dec 2017

Project: Oracle Application Express (APEX) platform

EDUCATION

Ph.D., Computer Science, University of Utah — Aug 2019 – Jun 2024

Thesis: Accelerating modern cryptography with a focus on data movement. Advisor: Prof. Rajeev Balasubramonian.

B.Tech., Indian Institute of Technology (IIT) Gandhinagar — 2013 – 2017

Major: Mechanical Engineering. Minor: Computer Science and Engineering.

SERVICE, RECOGNITION & LEADERSHIP

- **Technical Program Committee:** IEEE/ACM ISCA 2027, IEEE/ACM MICRO 2026, IEEE/ACM ISCA 2026, IEEE HPCA 2026, IEEE ISPASS 2026, IEEE Micro 2022.
- Finalist, Samsung Open Innovation Contest for AxDIMM Technology, December 2021.
- Founding Member and Steering Committee Member, Computer Architecture Student Association (CASA), 2020–2022.
- Teaching Mentor, University of Utah: CS/ECE 6810 Computer Architecture (Fall 2020); CS 3810 Computer Organization (Spring 2022).

PUBLICATIONS

7 peer-reviewed conference and journal papers (first author on 5), plus 1 paper at a U.S. government conference. Full list with citations on Google Scholar.

Conference and journal papers

1. **Sarabjeet Singh**, Shreyas Singh, Sumanth Gudaparthi, Xiong Fan, Rajeev Balasubramonian. **"Hyena: Balancing Packing, Reuse, and Rotations for Encrypted Inference."** *IEEE Symposium on Security and Privacy (S&P)*, 2024, pp. 3091–3108. [PDF] [DOI]
2. **Sarabjeet Singh**, Xiong Fan, Ananth Krishna Prasad, Lin Jia, Anirban Nag, Rajeev Balasubramonian, Mahdi Nazm Bojnordi, Elaine Shi. **"XCRYPT: Accelerating Lattice-Based Cryptography with Memristor Crossbar Arrays."** *IEEE Micro*, vol. 43, no. 5, pp. 45–54, 2023. [PDF] [DOI]
3. **Sarabjeet Singh**, Neelam Surana, Kailash Prasad, Pranjali Jain, Joycee Mekie, Manu Awasthi. **"HyGain: High-Performance, Energy-Efficient Hybrid Gain Cell-Based Cache Hierarchy."** *ACM Transactions on Architecture and Code Optimization (TACO)*, vol. 20, no. 2, 2023. [PDF] [DOI]
4. Sumanth Gudaparthi, **Sarabjeet Singh**, Surya Narayanan, Rajeev Balasubramonian, Visvesh Sathe. **"CANDLES: Channel-Aware Novel Dataflow-Microarchitecture Co-Design for Low Energy Sparse Neural Network Acceleration."** *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2022, pp. 876–891. [PDF] [DOI]
5. **Sarabjeet Singh**, Manu Awasthi. **"Efficacy of Statistical Sampling on Contemporary Workloads: The Case of SPEC CPU2017."** *IEEE International Symposium on Workload Characterization (IISWC)*, 2019, pp. 70–80. [PDF] [DOI]
6. **Sarabjeet Singh**, Manu Awasthi. **"Memory Centric Characterization and Analysis of SPEC CPU2017 Suite."** *ACM/SPEC International Conference on Performance Engineering (ICPE)*, 2019, pp. 285–292. [PDF] [DOI]
7. Sneha N. Ved, **Sarabjeet Singh**, Joycee Mekie. **"PANE: Pluggable Asynchronous Network-on-Chip Simulator."** *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 15, no. 1, 2019. [PDF] [DOI] [Code]

Government conference papers

- Ahmad Al Badawi, Yuriy Polyakov, Ronen Cohen, Carlo Pascoe, **Sarabjeet Singh**, David Bruce Cousins, Kurt Rohloff, Matthew French, Benedict Reynwar, Troy Hicks, Kellie Canida, Clynn Mathew, Ajey Jacob, Kenneth M. Zick, Akshath Mahajan, Mukund Ramakrishnan, Rugved Mhatre, Austin Ebel, Brandon Reagen. **"Accelerating FHE to Enable Privacy-Preserving Deep Learning."** *Government Microcircuit Applications & Critical Technology Conference (GOMACTech)*, Pasadena, CA, March 2025, paper 37.2. [Program]

Workshop papers

- **Sarabjeet Singh**, Shreyas Singh, Rajeev Balasubramonian. **"EPIC: Efficient Packing for Inference using Cheetah."** *6th Workshop on Cognitive Architectures (CogArch), held with HPCA*, 2022. [PDF]

Other writing

- **Sarabjeet Singh**. **"Post Quantum Cryptography."** *ACM SIGARCH Computer Architecture Today* (blog), February 2021. [Link]

SOFTWARE & RESEARCH ARTIFACTS

- **PANE**: open-source simulator for synchronous, asynchronous, and mixed NoCs (ACM JETC 2019). [GitHub]
- **SPEC CPU2017 simulation points**: whole and regional Pinballs with reproducible traces of simulation points (IISWC 2019), shared with researchers on request.

PH.D. RESEARCH PROJECTS

- **Efficient Integrity Verification using Custom DIMM**: reducing the cost of integrity verification in trusted environments with near-memory support.
- **ZPU: Zero-Knowledge Proof Processing Units**: high-throughput processing units for MSM/NTT kernels in zero-knowledge proofs.
- **Reducing Expensive Key Movement in FHE using Processing-in-Memory**: partitioning FHE kernels on a CPU-UPMEM model.
- **Secure AI using Samsung's AxDIMM** (Oct 2021 – Jun 2022): finalist, Samsung Open Innovation Contest for AxDIMM Technology.
- **Efficient Metadata for Memory Protection** (Aug 2020 – Mar 2021).
- **AMBOP: Adaptive Multiple Best Offset Prefetcher** (Mar 2019 – Aug 2019).

TECHNICAL SKILLS

Architecture & Systems Computer architecture, ASICs, memory systems, processing-in-memory, hardware/software co-design, ML/AI dataflow, dataflow optimization, performance modeling	**Privacy-Preserving Computing** Trusted Execution/Enclaves, Confidential Computing, fully homomorphic encryption, federated learning, post-quantum cryptography, Privacy-enhancing technologies (PETs), Differential privacy, Secure multi-party computation
Programming & Hardware C, C++, Python, Verilog, shell scripting, SQL	**Simulation & Performance Tools** gem5, Sniper, ZSim, ChampSim, NVMain, DRAMSim2, USIMM, Pin, perf, Intel VTune
Design Tools CACTI, Cadence Spectre, RTL Compiler, Innovus, Virtuoso	**Libraries & Frameworks** OpenFHE, SEAL, NVFlare, AI training & inference