ESHA CHOUKSE

(512) · 545 · 7885 ◇ esha.choukse@microsoft.com ◇ eshachoukse.weebly.com

EDUCATION

University of Texas at Austin *Aug 2014 - May 2019*
PhD, Computer Architecture and Embedded Systems, Electrical Engineering

Indian Institute of Technology, Kharagpur *July 2008- Apr 2012*
B.Tech(Hons.) in Electronics and Electrical Communication and Minor in Computer Science

EXPERIENCE

Microsoft *Jul 2022 - Present*
Senior Researcher, Azure Systems Research *Redmond, WA*

- Efficient GPU deployment
- Large language model efficiency through computer architecture research

Microsoft Research *May 2021 - Jul 2022*
Researcher *Redmond, WA*

- Cloud efficiency and sustainability in datacenters
- Novel power management and computer architecture solutions in CPUs and GPUs for cloud

Microsoft *August 2019 - May 2021*
Hardware Engineer 2 *Redmond, WA*

- Architect in the Next Cloud Systems Architecture group in Azure.
- Researching the QoS, live migration and virtualization bottlenecks in Azure stack.
- Working on research for short and long term sustainability solutions from computer architecture perspective.
- Analyzed the hardware characteristics of Functions as a service model in cloud.

NVIDIA Research *May 2018 - Aug 2018*
Research Intern *Austin, TX*

- Analyzed the data of various GPU workloads, and their entropy patterns.
- Designed a compressed memory system designed for the GPU model, which works well with a variety of workload-types of interest.

Intel *May 2017 - Aug 2017*
Memory Architecture Intern *Hillsboro, OR*

- Explored several opportunities to improve memory latency for server chips, resulting in several avenues for memory system optimization.

ARM *May 2016 - Aug 2016*
Research Intern *Austin, TX*

- Evaluated various metrics for using on-chip DRAM to alleviate the challenges with a non-volatile main memory system

ARM *May 2015 - Aug 2015*
Verification Intern *Austin, TX*

- Helped develop a new tool for ISA level testing in interesting multiprocessor scenarios
- Implemented various features like exception level switching and exception handling in the tool

Qualcomm *July 2012 - Aug 2014*
Embedded Systems Engineer *Hyderabad, India and San Diego, CA*

- Implemented Boot flow/ microcode for various Snapdragon chipsets (8974Pro, 8x26, 8x10, 9x35, 8916, 9x45, 8936).
- Worked closely with ARM cortex M3/A7/A15/A53, Qualcomm Krait, Hexagon processors.
- Was involved in Boot code deliverables, pre-silicon emulation/simulation and System-on-Dock testing.
- **Filed a patent** on NAND Flash Failsafe over-the-air upgrades. *US20140173187*
- Was awarded 8 Qualstars (Certificates of merit) within 2 years.

Mentor Graphics *May 2011 - July 2011*
Summer Intern *Noida, India*

- Worked on building the emulation platforms: Testbench Xpress and Veloce.
- Developed various libraries, most importantly for checkpoint restore functionality.

RESEARCH COMMUNITY INVOLVEMENT

Served as publication chair for MICRO 2023, on Technical Program Committee for ASPLOS 2023 (Summer, Fall), MICRO 2022, ISCA 2022, HPCA 2022, ISCA 2021, HPCA 2021, Industry track, MICRO 2020, ICCD 2020.

Served as external reviewer for ASPLOS 2020, ISCA 2020 and HPCA 2020.

Co-chair for the IEEE ICCD 2021 Special Sessions

Organized an MSR workshop during the ASPLOS 2020 PC meeting for better collaboration between Microsoft and the academic community in systems.

Contributions to the principles and best practices being put together at the Green Software Foundation (2020-present)

Helped raise awareness about industry research job interview process through MICRO 2020 Jobs panel and WiCArch talks.

Mentored new PhD students through WiCArch

RESEARCH AND PUBLICATIONS

Splitwise: Efficient generative LLM inference using phase splitting

- arXiv Preprint (Nov 2023)
- Authors: Pratyush Patel (University of Washington), Esha Choukse, Chaojie Zhang, Íñigo Goiri, Aashaka Shah, Saeed Maleki, Ricardo Bianchini (Microsoft)

POLCA: Power Oversubscription in LLM Cloud Providers

- arXiv Preprint (Aug 2023)
- Authors: Pratyush Patel (University of Washington), Esha Choukse, Chaojie Zhang, Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam, Ricardo Bianchini (Microsoft Azure)

Myths and Misconceptions Around Reducing Carbon Embedded in Cloud Platforms

- Workshop on Sustainable Computer Systems (2023)
- Authors: Jialun Lyu, Jaylen Wang, Kali Frost, Chaojie Zhang, Celine Irvene, Esha Choukse, Rodrigo Fonseca, Ricardo Bianchini, Fiodar Kazhamiaka, Daniel S Berger

Towards Improved Power Management in Cloud GPUs

- IEEE Computer Architecture Letters (2023)
- Authors: Pratyush Patel, Zibo Gong, Syeda Rizvi, Esha Choukse, Pulkit Misra, Tom Anderson, Akshitha Sriraman

Translation-optimized Memory Compression for Capacity

- Published at **IEEE Micro 2022**
- Authors: Gagandeep Panwar, Muhammad Laghari, David Bears, Yuqing Liu, Chandler Jearls (Virginia Tech), Esha Choukse (Microsoft Research), Kirk W Cameron, Ali R Butt, Xun Jian (Virginia Tech)

Overclocking in Immersion-Cooled Datacenters

- We show how liquid cooling enables cloud providers to overclock server components, and tradeoff the potential increase in performance with higher power draw and reliability implications with various use-cases.
- Published at **IEEE Micro Top Picks 2022**
- Authors: Pulkit A Misra, Ioannis Manousakis, Esha Choukse, Majid Jalili, Inigo Goiri, Ashish Raniwala, Brijesh Warrier, Husam Alissa, Bharath Ramakrishnan, Phillip Tuma, Christian Belady, Marcus Fontoura, Ricardo Bianchini (Microsoft)

Buddy Compression: Compressing Device Memory in GPUs

- Proposed a compressed data management well-suited to GPU applications, both, HPC and DL.
- Published at **International Symposium on Computer Architecture (ISCA) 2020**.
- Authors: Esha Choukse (UT Austin), Michael Sullivan (NVIDIA), Mike O'Connor (NVIDIA), Mattan Erez (UT Austin), Jeff Pool (NVIDIA), David Nellans (NVIDIA), Steve Keckler (NVIDIA)

Prunetrain: Gradual structured pruning from scratch for faster neural network training

- Pruned the networks during training, resulting in 3x faster training and much lower resource utilization.
- Published at **The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) 2019**.
- Authors: Sangkug Lym (UT Austin), Esha Choukse (UT Austin), Siavash Zangeneh (UT Austin), Wei Wen (Duke University), Sujay Sanghavi (UT Austin), Mattan Erez (UT Austin)

Compresso: Efficient OS-transparent Main Memory Compression

- An OS-transparent, hardware-only mechanism for main memory compression for storage benefits.
- Made optimizations to decrease the compression-related data movement.
- Used novel methodology for accurate and holistic evaluation of a compressed memory system.
- Published in **International Symposium on Microarchitecture (MICRO) 2018**.
- Authors: Esha Choukse (UT Austin), Mattan Erez (UT Austin), Alaa Alameldeen (Intel Labs)

CompressPoints: An Evaluation Methodology for Compressed Memory Systems

- Methodology for choosing better representative regions of workloads for studies that require data-representativeness.
- Published in **IEEE Computer Architecture Letters (CAL)**, July-Dec 2018.

- Authors: Esha Choukse (UT Austin), Mattan Erez (UT Austin), Alaa Alameldeen (Intel Labs)

Bit-Plane Compression

- Bandwidth savings in GPGPUs using Bit-plane compression over transformed data
- Published in **International Symposium on Computer Architecture (ISCA) 2016**.
- Authors: Jungrae Kim (UT Austin), Michael Sullivan (NVIDIA), Esha Choukse (UT Austin), Mattan Erez (UT Austin)

PATENTS

Systems and methods for autoscaling in datacenters

- Techniques for autoscaling virtual machines in a datacenter using overclocking.
- US Patent App. 17/752,689, 2023
- Publication date 2023/12/21

System, apparatus and method for application specific address mapping

- Increasing effective DRAM bandwidth per application by spreading the access pattern across banks using speacialized address interleaving.
- US Patent App. US10936507B2
- Publication date 2021/3

Techniques for setting a 2-level auto-close timer to access a memory device

- Decreasing DRAM access latency upon page miss by using an adaptive auto-close timer that is aware of special access patterns.
- US Patent App. 20200019513
- Publication date 2020/1/16

Virtual boundary codes in a data image of a read-write memory device

- Identifying the start of data block in initial boot logic in NAND devices using a magic number.
- US Patent App. 14060736
- Publication date 2014/6/19

TERM PROJECTS

Reinforcement Learning based driving simulation

Machine Learning Term Project, UT Austin *Jan 2017 - Apr 2017*

- Used Q-Learning to train a simulated car in an environment with other cars, pedestrians, and traffic lights.

Health Radar Android App

Mobile Computing Term Project, UT Austin *Sep 2016 - Dec 2016*

- Bluetooth Beacons based Android app for tracking and reporting exposure to diseases.

Microarchitecture: Cache and Memory design

Microarchitecture Term Project, UT Austin *Jan 2016 - Apr 2016*

- Designed and implemented the memory subsystem for an x86 processor in structural verilog.
- Added features like MSHRs, Write-buffer, Victim-buffer, Critical Word First and Row Buffers.

Krispy Kache: Remote cache injection in multicore systems

Parallel Computer Architecture Term Project, UT Austin *Sep 2015 - Dec 2015*

- Evaluated Software-based and Hardware-prediction based approaches to cache-block injection into remote-core's caches.

Parity Based Erasure Correction for Fused Data Structures

Distributed Systems Term Project, UT Austin *Sep 2015 - Dec 2015*

- Implemented and evaluated EvenOdd, Replication and RS-encoding mechanisms in fused backups for servers.

Memory Aware Warp Scheduling in GPGPUs

Parallelism and Locality Term Project, UT Austin *Feb 2015 - Apr 2015*

- Extended the idea of Large Warps to make scheduling decisions based on locality of thread accesses

Hourglass- Solving Distributed Priority Inversion

Real-time OS Term Project, UT Austin *Feb 2015 - Apr 2015*

- Defined the problem of distributed priority inversion in a multi-process embedded system
- Solved the problem by implementing locks using a special, new OS construct
- Implemented in uCOS-III

Shared stack using Queue delegation and Elimination in a Multicore system

Multicore Computing Term Project, UT Austin *Sep 2014 - Dec 2014*

- Implemented a fast shared stack using queue delegation and elimination

Architecture Design and FPGA implementation of Efficient Face Recognition

Senior Design Project, IIT Kharagpur *Sep 2011-Apr 2012*

- Designed and implemented a time and computation efficient eigenface algorithm on Xilinx FPGA.

RELEVANT COURSES

- Parallelism and Locality
- Computer Architecture
- Distributed Systems
- Machine Learning
- Real time Operating Systems
- Dynamic Compilation
- Microarchitecture
- Compilers
- Parallel Computer Architecture
- Multicore Computing
- Mobile Computing

TECHNICAL STRENGTHS

Programming/Scripting	C, C++, Java, MPI, CUDA, Cilk, pThreads, python
Testing Platforms and simulators	Pin, Trace32, GPGPUSim, zsim
Processor Architectures	ARM Cortex A/M for v8, Hexagon, Intel x86, NVIDIA Tesla