ESHA CHOUKSE

esha.choukse@microsoft.com ◇ https://aka.ms/eshachoukse

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, Azure Research - Systems *Redmond, WA*
Principal Researcher *Jul 2024 - **Present***
Senior Researcher *Jul 2022 - Jun 2024*
Researcher *May 2021 - Jun 2022*

· Leading the **Efficient AI research** umbrella at Microsoft Azure Research - Systems
· Cross-stack innovations across AI platform, GPU architecture, and datacenter infrastructure.
· Focus on emerging AI workflows, power, energy, and sustainability.

Microsoft *August 2019 - Apr 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.
· Analyzed the hardware characteristics of Functions as a service model in cloud.

NVIDIA - Research Intern *May 2018 - Aug 2018*

Intel - Memory Architecture Intern *May 2017 - Aug 2017*

ARM - Research Intern *May 2016 - Aug 2016*

ARM - Verification Intern *May 2015 - Aug 2015*

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

· Worked on BootROM for ARM cortex M3/A7/A15/A53, Qualcomm Krait, Hexagon processors.

Mentor Graphics - Intern *May 2011 - July 2011*

RESEARCH COMMUNITY SERVICE

- Serving as **Associate Editor** for IEEE Computer Architecture Letters (CAL) since January, 2025.
- Served as **PC Co-chair** for HotCarbon 2024.
- Served as **Publication Chair** for MICRO 2023

- Served on Technical **Program Committee** for SIGCOMM 2025, ISCA 2024, 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.
- Served as **PC Co-chair** for the IEEE ICCD 2021 Special Sessions
- Mentored new PhD students through WiCArch

CONFERENCE/JOURNAL PUBLICATIONS (COMPLETE LIST ONLINE)

TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms

· ASPLOS 2025

· Authors: Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, ***Esha Choukse***, Haoran Qiu, Rodrigo Fonseca, Josep Torrellas, Ricardo Bianchini

DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency

· ***Best Paper Award at HPCA 2025***

· Authors: Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Josep Torrellas, ***Esha Choukse***

Memory Allocation under Hardware Compression

· MICRO 2024

· Authors: Muhammad Laghari, Yuqing Liu, Gagandeep Panwar, David Bears , Chandler Jearls , Raghavendra Srinivas , ***Esha Choukse***, Kirk Cameron, Ali R. Butt, Xun Jian

Mosaic: Harnessing Micro-architectural Resources of Servers in Serverless Environments

· MICRO 2024

· Authors: Jovan Stojkovic, ***Esha Choukse***, Enrique Saurez, Íñigo Goiri, Josep Torrellas

Optimizing GPU Data Center Power

· APCCAS 2024

· Authors: Tawfik Rahal-Arabi, Paul Van der Arend, Ashish Jain, Mehdi Saidi, Rashad Oreifej, Sriram Sundaram, Srilatha Manne, Indrani Paul, Rajit Seahra, Frank Helms, ***Esha Choukse***, Nithish Mahalingam, Brijesh Warrier, Ricardo Bianchini

Splitwise: Efficient generative LLM inference using phase splitting

· ***ISCA 2024 Best paper nominee***

· ***IEEE Micro Top Picks Award, 2024***

· Authors: Pratyush Patel, ***Esha Choukse***, Chaojie Zhang, Íñigo Goiri, Aashaka Shah, Saeed Maleki, Ricardo Bianchini (Microsoft)

· Has been adopted by several companies and open-source software within a year, and is now the new industry standard.

Designing Cloud Servers for Lower Carbon

· ISCA 2024, ***IEEE MICRO Top Picks Award, 2024***

· Authors: Jaylen Wang, Daniel S. Berger, Fiodar Kazhamiaka, Celine Irvene, Chaojie Zhang, ***Esha Choukse***, Kali Frost, Rodrigo Fonseca, Brijesh Warrier, Chetan Bansal, Jonathan Stern, Ricardo Bianchini, Akshitha Sriraman

- Has led to an industry-wide impact on carbon accounting and importance of performance vs carbon/reuse

DyLeCT: Achieving Huge-page-like Translation Performance For Hardware-compressed Memory

- ISCA 2024
- Authors: Gagan Panwar, Muhammad Laghari, ***Esha Choukse***, Xun Jian

SmartOClock: Workload- and Risk-Aware Overclocking in the Cloud

- ISCA 2024
- Authors: Jovan Stojkovic, Pulkit Misra, Íñigo Goiri, Sam Whitlock, ***Esha Choukse***, Mayukh Das, Chetan Bansal, Jason Lee, Zoey Sun, Haoran Qiu, Reed Zimmermann, Savyasachi Samal, Brijesh Warrier, Ashish Raniwala, Ricardo Bianchini
- This work has led to new boosting features in Intel and AMD CPUs.

Characterizing Power Management Opportunities for LLMs in the Cloud

- ASPLOS 2024
- Authors: Pratyush Patel, ***Esha Choukse***, Chaojie Zhang, Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam, Ricardo Bianchini
- Has been deployed at Microsoft, leading to up to 15% lower provisioned power in GPU data centers!

Making Kernel Bypass Practical for the Cloud with Junction

- NSDI 2024
- Authors: Joshua Fried, Gohar Irfan Chaudhry, Enrique Saurez, ***Esha Choukse***, Íñigo Goiri, Sameh Elnikety, Rodrigo Fonseca, Adam Belay

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
- Has led to better power control and telemetry features in NVIDIA H100+

Translation-optimized Memory Compression for Capacity

- MICRO 2022
- Authors: Gagandeep Panwar, Muhammad Laghari, David Bears, Yuqing Liu, Chandler Jearls, ***Esha Choukse***, Kirk W Cameron, Ali R Butt, Xun Jian

Overclocking in Immersion-Cooled Datacenters

- ***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

Buddy Compression: Compressing Device Memory in GPUs

- ISCA 2020.
- Authors: ***Esha Choukse***, Michael Sullivan, Mike O'Connor, Mattan Erez, Jeff Pool, David Nellans, Steve Keckler

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

· SC 2019 (SuperComputing)
· Authors: Sangkug Lym, ***Esha Choukse***, Siavash Zangeneh, Wei Wen, Sujay Sanghavi, Mattan Erez

Compresso: Efficient OS-transparent Main Memory Compression

· MICRO 2018
· Authors: ***Esha Choukse***, Mattan Erez, Alaa Alameldeen

CompressPoints: An Evaluation Methodology for Compressed Memory Systems

· Computer Architecture Letters (CAL) 2018.
· Authors: ***Esha Choukse***, Mattan Erez, Alaa Alameldeen

Bit-Plane Compression: Transforming Data for Better Compression in Many-Core Architectures

· ISCA 2016
· Authors: Jungrae Kim, Michael Sullivan, ***Esha Choukse***, Mattan Erez

WORKSHOP/PREPRINT PUBLICATIONS (COMPLETE LIST ONLINE)

Intelligent Router for LLM Workloads: Improving Performance Through Workload-Aware Scheduling

· EuroMLSys at EuroSys 2025
· Authors: Kunal Jain, A. Parayil, Ankur Mallick, ***Esha Choukse***, Xiaoting Qin, Jue Zhang, Íñigo Goiri, Rujia Wang, Chetan Bansal, Victor Ruehle, Saravan Rajmohan, Kunal Jain, Anoop Kulkarni, Steve Kofsky

EcoServe: Designing Carbon-Aware AI Inference Systems

· arXiv Preprint 2025
· Authors: Yueying Li, Zhanqiu Hu, ***Esha Choukse***, Rodrigo Fonseca, G. Edward Suh, Udit Gupta

Towards Resource-Efficient Compound AI Systems

· arXiv Preprint 2025
· Authors: Gohar Irfan Chaudhry, ***Esha Choukse***, Íñigo Goiri, Rodrigo Fonseca, Adam Belay, Ricardo Bianchini

Towards Efficient Large Multimodal Model Serving

· arXiv Preprint 2025
· Authors: Haoran Qiu, Anish Biswas, Zihan Zhao, Jayashree Mohan, Alind Khare, ***Esha Choukse***, Íñigo Goiri, Zeyu Zhang, Haiying Shen, Chetan Bansal, Ramachandran Ramjee, Rodrigo Fonseca

DroidSpeak: KV Cache Sharing for Efficient Multi-LLM Serving

· arXiv Preprint 2024
· Authors: Yuhan Liu, Yuyang Huang, Jiayi Yao, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, Madan Musuvathi, ***Esha Choukse***

Mnemosyne: Parallelization Strategies for Efficiently Serving Multi-Million Context Length LLM Inference Requests Without Approximations

· arXiv Preprint 2024
· Authors: Amey Agrawal, Junda Chen, Íñigo Goiri, Ramachandran Ramjee, Chaojie Zhang, Alexey Tumanov, ***Esha Choukse***

Input-Dependent Power Usage in GPUs

- Sustainable Supercomputing at SC 2024
- Authors: Theo Gregersen, Pratyush Patel, ***Esha Choukse***

Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference

- EMC2 at ASPLOS 2024
- Authors: Jovan Stojkovic, ***Esha Choukse***, Chaojie Zhang, Íñigo Goiri, Josep Torrellas

Myths and Misconceptions Around Reducing Carbon Embedded in Cloud Platforms

- HotCarbon 2023
- Authors: Jialun Lyu, Jaylen Wang, Kali Frost, Chaojie Zhang, Celine Irvene, ***Esha Choukse,*** Rodrigo Fonseca, Ricardo Bianchini, Fiodar Kazhamiaka, Daniel S Berger

PATENTS

Efficient LLM Agent communication with KV cache sharing

- Filed with ***Microsoft***. Based on the paper DroidSpeak.
- Filing date 2025/03/06

Designing LLM Inference Clusters for Performance and Energy Efficiency

- Filed with ***Microsoft***. Based on the DynamoLLM paper.
- Filing date 2025/02/22

Thermal- and Power-Aware Scheduling for LLM Inference in the Cloud

- Filed with ***Microsoft***. Based on the paper from ASPLOS 2025, TAPAS.
- Filing date 2025/02/13

Power Oversubscription in LLM Cloud Providers

- Filed with ***Microsoft***. Based on the paper from ASPLOS 2024.
- Filing date 2024/10/23

Heterogenous accelerators for efficient generative LLM inference using phase splitting

- Filed with ***Microsoft***. Based on the paper Splitwise.
- Filing date 2024/10/23

Systems and methods for autoscaling in datacenters

- Filed with ***Microsoft***. Techniques for autoscaling virtual machines in a datacenter using overclocking.
- Publication date 2023/12/21

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

- Filed with ***Intel***. Increases page hit ratio in DRAM.
- Publication date 2023/12/21

System, apparatus and method for application specific address mapping

- Filed with ***Intel***. Increasing effective DRAM bandwidth per application by spreading the access pattern across banks using speacialized address interleaving.
- Publication date 2021/3

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

· Filed with ***Intel***. Decreasing DRAM access latency upon page miss by using an adaptive auto-close timer that is aware of special access patterns.
· Publication date 2020/1/16

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

· Filed with ***Qualcomm***. Identifying the start of data block in initial boot logic in NAND devices using a magic number.
· Publication date 2014/6/19

INVITED TALKS AND PANELS

- Invited talk at EMC2 workshop at ASPLOS 2024
- Invited seminars at
  - Indian Institute of Technology, Bombay in December 2024
  - University of Minnesota, Twin Cities in December 2024 (online)
  - University of California, San Diego in April 2024
  - Intel architecture talks - talk on Splitwise in March 2024
  - Open Compute Project (OCP) - talk on Splitwise in Composable Memory Systems workstream in February, 2024
  - University of Washington, Seattle in February 2024
  - Indian Institute of Science, Bangalore in December 2019
  - WiCArch in April 2019 (online)
- Guest lectures at
  - Cornell Tech, New York City in March 2025 (online)
  - Georgia Tech, Atlanta in November 2025 (online)
  - CMU, Pittsburgh in December 2024 (online)
- Panelist at Green Software Foundation, global summit, 2023
- Panelist at various Microsoft conferences like MLADS 2023 - Sustainability, MLADS 2024 - AI Efficiency, and Women Innovators 2024.
- Panelist at uArch workshop, ISCA 2021
- Panelist at MICRO JOBS workshop, 2020
- Chaired various paper sessions at MICRO 2019, ISCA 2021, MICRO 2024, ASPLOS 2024

SELECTED MEDIA COVERAGE OF PAPERS

DroidSpeak: KV Cache Sharing for Efficient Multi-LLM Serving

· ***E-Week*** - 'Droidspeak': AI Agents Now Speak Their Own Language Courtesy of Microsoft (`https://www.eweek.com/news/droidspeak-ai-language-microsoft/`)
· ***Tech Xplore*** - Microsoft collaboration develops DroidSpeak for better communication between LLMs (`https://techxplore.com/news/2024-11-microsoft-collaboration-droidspeak-communication-llms.html`)

· ***Reddit*** - Droidspeak: AI models work together faster when they speak their own language(https://www.reddit.com/r/technews/comments/1gypnxv/droidspeak_ai_models_work_together_faster_when/)

Splitwise: Efficient generative LLM inference using phase splitting

· Microsoft Research Blog - Splitwise improves GPU usage by splitting LLM inference phases(https://www.microsoft.com/en-us/research/blog/splitwise-improves-gpu-usage-by-splitting-llm-inference-phases/)
· ***Mark Tech Post*** - Are Your AI Models Hungry for Too Much Power? This Paper from Microsoft Introduces Splitwise to Split the Bill (https://www.marktechpost.com/2024/01/12/are-your-ai-models-hungry-for-too-much-power-this-paper-from-microsoft-introduces-splitwise-to-split-the-bill/)
· ***EDGE*** - Microsoft Azure researchers introduce Splitwise to improve GPU efficiency for LLMs (https://sp-edge.com/updates/25229)
· ***Azure CTO blog*** - https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/09/26/sustainable-by-design-innovating-for-energy-efficiency-in-ai-part-2/?mscm=1)
· ***Azure AI blog*** - AI Frontiers: Human insights on AI training (https://techcommunity.microsoft.com/blog/azure-ai-services-blog/ai-frontiers-human-insights-on-ai-training/4120574)
· ***Tweakers*** - Microsoft Azure: Splitwise tech reduces LLMs' power requirements by 20 percent (https://tweakers.net/nieuws/217228/microsoft-azure-splitwise-tech-verlaagt-stroomvereisten-llms-met-20-procent.html)

Designing Cloud Servers for Lower Carbon

· IEEE Spectrum - Servers Get a Second Life for Sustainability (https://spectrum.ieee.org/amp/server-reuse-2669736974)

Characterizing Power Management Opportunities for LLMs in the Cloud

· Microsoft Ignite Keynote by Azure CTO - Inside Microsoft AI innovations with Mark Russinovich (https://www.youtube.com/watch?v=c4SUhWBybXo&ab_channel=MicrosoftEvents)

PAST STUDENTS I WORKED WITH THAT HAVE GRADUATED

- Gagandeep Panwar was a PhD student at Virginia Tech when I worked with him, and is now at AMD Research.
- Theo Gregersen was an BS/MS student at UW, Seattle when I was working with him, and is now a PhD student at CMU.
- Kunal Jain was an undergraduate student at IIIT Hyderabad, and is now an incoming PhD student at UT Austin.