SARTHAK CHAKRABORTY

+1 (217)-841-2791 | sarthak.chakraborty@gmail.com | Github | Website | Google Scholar

RESEARCH INTERESTS

My research interests lie in developing efficient algorithms and system abstractions for ML and cloud, which encompass aspects like reliability, correctness, and performance. My research experiences include ML for systems, distributed systems, cloud resource management and reliability, data-driven systems, and federated learning.

EDUCATION

- **PhD in Computer Science** August 2023 - May 2028 (expected)
  University of Illinois Urbana-Champaign
  Advisor: Dr. Indranil Gupta (indy@illinois.edu)
  Cumulative GPA: 4.0/4.0
  Funded by ***Illinois Distinguished Fellowship***

- **Dual Degree (B. Tech + M. Tech) in Computer Science and Engineering** July 2016 - April 2021
  Indian Institute of Technology (IIT) Kharagpur, India
  Cumulative GPA: 9.74/10.00
  Institute Rank: 2 | Department Rank: 2

PUBLICATIONS

[1] (NeurIPS '25) **Sarthak Chakraborty**, Suman Nath, Xuchao Zhang, Chetan Bansal, Indranil Gupta. **Generative Caching for Structurally Similar Prompts and Responses.** In *Thirty-Ninth Annual Conference on Neural Information Processing System*, 2025. [LINK] (Accpt. Rate: 24.52%)

[2] (SoCC' 25) Chirag Shetty, **Sarthak Chakraborty**, Hubertus Franke, Larisa Shwartz, Chandra Narayanaswami, Indranil Gupta, Saurabh Jha. **CPU-Limits kill Performance: Time to rethink Resource Control.** In *Proceedings of the 2025 ACM Symposium on Cloud Computing*, 2025. [LINK]

[3] (NSDI '24) Shubham Agarwal, Subrata Mitra, **Sarthak Chakraborty**, Srikrishna Karanam, Koyel Mukherjee, Shiv Saini. **Approximate Caching for Efficiently Serving Diffusion Models.** In *21st USENIX Symposium on Networked Systems Design and Implementation*, 2024. [LINK] (Accpt. Rate: 18.6%)

[4] (ASE '23) **Sarthak Chakraborty**, Shubham Agarwal, Shaddy Garg, Abhimanyu Sethia, Udit Narayan Pandey, Videh Aggarwal, Shiv Saini. **ESRO: Experience Assisted Service Reliability against Outages.** In *The 38th IEEE/ACM International Conference on Automated Software Engineering*, 2023. [LINK] (Accpt. Rate: 21.3%)

[5] (ESEC/FSE '23) Shubham Agarwal, **Sarthak Chakraborty**, Shaddy Garg, Sumit Bisht, Chahat Jain, Ashritha Gonuguntla, Shiv Saini. **Outage-Watch: Early Prediction of Outages using Extreme Event Regularizer.** In *The 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 682-694, 2023. [LINK] (Accpt. Rate: 25.6%)

[6] (WWW '23) **Sarthak Chakraborty**, Shaddy Garg, Shubham Agarwal, Ayush Chauhan, Shiv Saini. **CausIL: Causal Graph for Instance Level Microservice Data.** In *Proceedings of The Web Conference 2023*, pp. 2905-2915, 2023. [LINK] (Accpt. Rate: 19.2%)

[7] (NeurIPS '22) Azam Ikram, **Sarthak Chakraborty**, Subrata Mitra, Shiv Saini, Saurabh Bagchi, Murat Kocaoglu. **Root Cause Analysis of Failures in Microservices through Causal Discovery.** *Advances in Neural Information Processing Systems 35*, pp. 31158-31170, 2022. [LINK] (Accpt. Rate: 25.6%)

[8] (ICBC '22) **Sarthak Chakraborty**, Sandip Chakraborty. **Proof of Federated Training: Accountable Cross-Network Model Training and Inference.** In *2022 IEEE International Conference on Blockchain and Cryptocurrency*, pp. 1-9, 2022. [LINK] (Accpt. Rate: 18.6%)

[9] (WWW '21) Lovish Chopra*, **Sarthak Chakraborty***, Abhijit Mondal, and Sandip Chakraborty. **PARIMA: Viewport Adaptive 360-degree Video Streaming.** In *Proceedings of The Web Conference 2021*, pp. 2379–-2391, 2021. [LINK] (Accpt. Rate: 20.6%)

INDUSTRIAL RESEARCH

- **Research Intern - Microsoft Research** May 2024 - Aug 2024
  Group: M365 Research + Cloud Systems Reliability Team *Redmond, USA*
  - Designed a novel program caching strategy to enable faster execution workflow of agentic systems by reducing LLM calls
  - Identified repeated and systematic workflows in which syntactic and semantic prompt caching fails, and enabled program caching by analyzing structural similarities across LLM prompts and their responses.
  - Clustered structurally similar LLM prompts, and synthesized a program that identifies the variable component of the prompt and replaced it with placeholders to generate corresponding responses
  - Improved cache hit rate by 68% and reduced workflow execution time by 33% on a proprietary agentic system

- **Research Associate 2 - Adobe Inc. (BigData Intelligence Lab)** Jul 2021 - Aug 2023
  Group: Systems and Insights Group *Bangalore, India*
  - Published 5 papers, filed 2 patents and successfully integrated research technologies into 2 products within 2 years
  - Mentored 10 undergraduate interns and 1 PhD intern during summer internships at Adobe over a span of 2 years
  - Selected as a speaker to present at Adobe Tech Summit 2022, a company-wide global internal technical conference
  - ***ML for System Reliability***
    - **Runtime Prediction of Incoming Jobs in Multi-Tenant System (Integrated into Product)**: Designed a pipeline that engineers features based on system state and predicts the latency of incoming Spark jobs with a MAPE value of ~ 0.4. The pipeline is used by the internal developers to decide if preventive measures are necessary
    - **Root Cause and Remediation Consolidation System** [Paper Link]: Constructed a knowledge graph utilizing both structured alerts data and unstructured incident reports data, enabling the consolidation of potential root causes and remediation strategies for a series of triggered alerts. The knowledge graph was employed to predict possible root causes in real-time, resulting in an improvement of $\sim 27\%$ over prior methods
    - **Designing Instance-aware Causal Graph from Telemetry Data** [Paper Link]: Formulated an algorithm that builds the causal structure among performance metrics at the instance level in a microservice-based cloud system, integrating domain knowledge derived from system architecture. This resulted in a 25% enhancement in the accuracy of graph estimation
  - ***Cloud System Reliability***
    - **Outage Prediction in Production System (Integrated into Product)** [Paper Link]: Implemented the inference and training pipeline of an outage forecasting model by inventing a novel distribution learning approach, exhibiting AUC of 0.8. To implement the pipeline into product, we leveraged Shapley value based explainability system to localize faulty system alerts.
    - **Root Cause Analysis via Intervention Modeling of Faults** [Paper Link]: Developed a hierarchical and localized causal discovery algorithm to model microservice metrics and detect the root causes of faults. Significantly reduced computation time against popular baselines and evaluated against real-world production data
  - ***System Optimizations for ML***
    - **Approximate Caching to Efficiently Serve Diffusion Models** [Paper Link]: Built an end-to-end diffusion model serving system and innovated a novel caching technique to reduce the cost and latency of text-to-image generation by intelligently reusing intermediate states. We are able to achieve 21% GPU savings and 19.8% reduction in latency without compromising the accuracy of generation

- **Research Intern - Adobe Inc. (BigData Intelligence Lab)** Apr 2020 - Jul 2020
  Topic: Architecting Large-Scale Asynchronous Federated Learning *Bangalore, India*
  - Designed a scalable and flexible framework for federated learning to support synchronous and asynchronous model training, with on-device learning on heterogeneous target devices such as android mobiles, web browsers, and desktops
  - Devised an algorithmic strategy to aggregate stale gradients effectively and deployed the framework on over 100 clients to perform image classification and boundary prediction tasks with real-world production models
  - Supported on-device federated learning on heterogeneous target devices including android mobiles (tfLite), web browsers (tfjs), IoT (Raspberry Pie) and desktop
  - Among one of the 6 interns in a pool of 64 interns to receive a pre-placement offer for full-time role in the research team

ACADEMIC RESEARCH

- **Reliability for LLM Agents** May 2025 - Ongoing
  University of Illinois Urbana-Champaign | In collaboration with IBM Research *Dr. Indranil Gupta*
  - Working on building a reliability and safety abstraction for the emerging agentic ecosystems in various domains
  - Exploring ways of combining ideas from formal verification, databases, and distributed systems within the agentic pipeline to assure that the state of the environment is always consistent, even when agents execute erroneous actions

- Designing a novel mechanism of incorporating "checkpointing" and "rollbacks" of agentic actions into the agentic world

- **Resource Management of Cloud Applications using Cloud Observability** Aug 2023 - July 2025
  University of Illinois Urbana-Champaign | In collaboration with IBM Research *Dr. Indranil Gupta*
  - Studied how various Kubernetes knobs integrate with the OS and affect resource management for cloud applications
  - Identified inefficiencies in the scaling mechanism and implemented an autoscaler that integrates horizontal, vertical, and node scaling, while avoiding overused Kubernetes configurations and preserving multi-tenancy.
  - Verified cycle-free and monotonicity properties of the autoscaler and shown that it reduces CPU over-allocation by over 38% while consistently maintaining high cluster utilization of over 70%

- **Cross-Chain Training of Learning Models via Blockchain Interoperability** Aug 2020 - Apr 2021
  Master's Thesis Project | IIT Kharagpur | Paper Link *Dr. Sandip Chakraborty*
  - Developed an end-to-end system to train a common machine learning model in a cross-silo setting over multiple smart contract enabled federated networks via the concept of blockchain interoperability
  - Incorporated permissioned blockchain networks to store auditable model states learned by the federated system
  - Constructed a relay-based cross-chain transfer mechanism to transfer the model state from one network to the other via HTTP channel. Signatures ensured that the data transferred was verifiable and authentic

- **PARIMA: Viewport Adaptive 360-degree Video Streaming** Jul 2019 - May 2020
  Bachelor's Thesis Project | IIT Kharagpur | Paper Link *Dr. Sandip Chakraborty*
  - Designed an online viewport-adaptive video streaming algorithm along with a client-server streaming platform
  - Implemented a novel PARIMA algorithm: an augmented Passive-Aggressive(PA) model and time series(ARIMA) model for viewport detection using video content as well as personalized head movement tracking.
  - Employed a pyramidal adaptive bitrate allocation scheme and improved the Quality of Experience by $\sim 30\%$
  - Used HEVC video encoding, GPAC for segmenting video chunks and 'MP4Client' for client streaming of video

- **Advanced Optimization Methods for Machine Learning** May 2019 - Aug 2019
  MITACS Globalink Research Intern | University of Waterloo | Github Link *Dr. Hans de Sterck*
  - Designed a randomized ALS algorithm targeted for CP Decomposition and Completion of Sparse Tensors
  - Computed leverage scores for the rows of factor matrices to sample non-zero data points using weighted reservoir sampling
  - Measured improvement of 25% in RMSE with 30% sampling rate against benchmark algorithms for tensor completion like conventional ALS, SGD, CCD++ and RRALS algorithms

PATENTS

[1] Sunav Choudhary, Atanu R. Sinha, *Sarthak Chakraborty*, Sai Shashank Kalakonda, Liza Dahiya, Purnima Grover, Kartavya Jain. **LiveStream Key Moment Identification.** [Granted] (US Patent No. 12294755)

[2] *Sarthak Chakraborty*, Sunav Choudhary, Atanu R. Sinha, Sapthotharan Nair, Manoj Ghuhan A, Yuvraj Gagneja, Atharva Anand Joshi, Atharv Tyagi, Shivi Gupta. **Generating Concise and Common User Representations for Edge Systems from Event Sequence Data stored on Hub Systems.** [Granted] (US Patent No. 12182829)

TEACHING AND SERVICES

- **Volunteering:** EuroSys 26 Shadow Reviewer, OSDI 25 Artifact Evaluation Committee
- **Reviewer:** ATC 23 (sub-review), IEEE TCSVT 23, EuroSys 23 (sub-review), IEEE TSNM 22, DSN 22 (sub-review)
- **Undergraduate TA:** Database Management Systems (CS43002), Theory of Computation (CS41001)

TERM PROJECTS

- Designed an externally synchronous replicated file system that performs faster than the traditional synchronous file system by implementing fast and slow path for multiple system calls.
- [Link] Developed a reduced version of Hadoop for running Map-Reduce tasks from scratch by implementing a distributed logging service, failure detection, distributed file system and an SQL wrapper parser.
- [Link] Developed a Distributed Collaboration System where multiple users can collaborate on a single document at once that maintained consistency along with a passive replication scheme. It used a master-worker architecture of servers.
- [Link] Designed an in-memory disk emulator with 4kB block size and built an ext2 like simple file system on top of the disk
- [Link] MRP: Implemented a reliable message-oriented communication protocol over an unreliable User Datagram protocol

- [Link] TinyC: Implemented a compiler for a subset of C functionalities to translate the C code to x86 Assembly Language
- [Link] KGP-RISC: Designed a 32-bit single cycle CPU(RISC based architecture) in Verilog VHDL and tested it on FPGA

SKILLS

• **Languages**	Python, C, C++, Java, SQL, Golang, Verilog, MIPS
• **Packages and Frameworks**	Kubernetes, Docker, LangChain, Kafka, MongoDB, scikit-learn, PyTorch, Keras, TensorFlow, DGL, Tensorflow-Federated, Git, Hyperledger Fabric

HONOURS AND AWARDS

- Recipient of ACM SIGSOFT CAPS Student Travel Grant to attend ESEC/FSE 2023 — 2023
- Recipient of the Illinois Distinguished Fellowship for academic achievements among incoming graduate students — 2023
- Recipient of the Goralal Syngal Memorial Scholarship awarded by the Institute for academic excellence — 2020, '19
- Received the prestigious MITACS Globalink Research Fellowship for a research internship in Canada — 2019