

Jayjeet Chakraborty

jayjeetc.github.io | jayjeetc@ucsc.edu | github.com/JayjeetAtGithub

OBJECTIVE

To explore the field of Storage Systems, Memory and Interconnects, Linux Kernel Development, Data Processing, Databases, and Distributed Systems; Currently working on applying Computational Storage to Data Management using embeddable Query engines such as Apache Arrow Acero and DuckDB and designing Fast hardware-accelerated data transport protocols for moving Columnar data.

EDUCATION

- **University of California, Santa Cruz** Santa Cruz, CA
Doctor of Philosophy, Computer Science And Engineering, CGPA: 3.56/4 *Expected: June, 2026*
- **National Institute Of Technology, Durgapur** West Bengal, India
B.Tech, Computer Science And Engineering, CGPA: 7.65/10 *Graduated: June, 2021*

EXPERIENCE

- **Software Engineering Intern** Remote
InfluxData Inc. *Summer 2023*
 - Profiled Jaegar queries to InfluxDB IOx (using heaptrack and Flamegraphs) to track down unbounded memory growth when executing sort preserving merge during grouping operations on high-cardinality dictionary-encoded fields.
 - Created reproducer for the above issue and implemented fix in DataFusion (query engine of InfluxDB IOx) after analyzing several alternative solutions for memory efficient Sort's/Group By's/Merge's in query execution engines. [[code](#)]
 - Wrote scripts for benchmarking DataFusion against DuckDB using TPC-H, ClickBench, and H2O.ai, on single and multicore platforms. [[code](#)]
- **Graduate Student Researcher** Santa Cruz, CA
UC Santa Cruz *Fall 2021 - Present*
 - Working on building a high-performance computational storage system in collaboration with Argonne National Labs using the mochi-thallium framework for RDMA-based data transport, mochi-bake for accessing raw storage regions using PMDK, mochi-yokan for storing metadata as K/V pairs, and Apache Arrow Acero and DuckDB for the query execution engine. [[code](#)]
- **IRIS-HEP Research Fellow** Remote
Princeton University and CROSS, UC Santa Cruz *Fall 2020, Winter 2021, Spring 2021*
 - Built scalable and reproducible Popper workflows for running experiments on large datasets stored in a [SkyhookDM](#) cluster. Also, performed experiments on a SkyhookDM deployment and studied the performance gains due to push-down. [[report](#)]
 - Redesigned the SkyhookDM computational storage system to use Apache Arrow as it's in-memory format and extended the Arrow library with a SkyhookFileFormat API to be able to natively connect to a Ceph/RADOS cluster with SkyhookDM plugins and push-down filter and projection queries into the Ceph OSDs. Also, contributed the [Skyhook project](#) to Apache Arrow. [[blog](#)]
- **Winter Research Intern** Varanasi, India
Indian Institute Of Technology, BHU *Winter 2020*
 - Studied and analyzed several terrain rendering techniques and implemented the ROAM(Real-time Optimally Adapting meshes) and Incremental Delaunay Triangulation algorithms for Level Of Detail based rendering of large terrain datasets. [[report](#)]
 - Developed a visualization tool in C++ using OpenGL to render terrains from Li-DAR datasets at 60 fps and benchmarked both the algorithm implementations on GPU.

Google Summer Of Code Student

Remote

Centre for Research in Open Source Software, UC Santa Cruz

Summer 2019

- Extended the Popper workflow engine by adding support for additional container runtimes like Singularity, added more sub-commands, implemented concurrent execution capabilities, and other CI/CD features. Also, wrote unit tests to achieve an 87% test coverage and added documentation for the newly added features. [\[report\]](#)
- Contributed plugins to facilitate the execution of Popper workflows on Virtual machines, Kubernetes clusters, and Slurm based HPC clusters.
- Extended the Popper ecosystem by building Popper workflows for automating the MLPerf benchmark suite and for end-to-end benchmarking of bare-metal machines. Also, developed a Python library to compute the confidence intervals for measuring the variability in CPU, Memory, Disk, and Network performance of CloudLab machines.

Software Engineering Intern

Mumbai, India

LogN Software

Summer 2018

- Worked on developing the backend APIs for a client-facing Ionic application using Django Rest Framework and MySQL. Also, built and integrated parts of UI of the application. Also, built the payment gateway of the application using the Stripe Payments SDK.

SKILLS

Programming Languages: C, C++, Rust, Python, Go, Java, JavaScript, Bash, MATLAB.

Tools: Git/GitHub, GitHub Actions, Docker, Podman, Kubernetes, Prometheus/Grafana, Ansible, Azure/GCP/AWS.

Profilers: Valgrind, Perf, FlameGraphs, Samplify, Bytehound, Heaptrack, Intel VTune, Instruments, Open Telemetry/Jaeger.

Frameworks: Flask, React, Boost, Qt, NumPy, Pandas, Matplotlib, Seaborn, Android Framework.

Operating Systems: Ubuntu/Debian, CentOS, MacOS.

PUBLICATIONS

- Jayjeet Chakraborty, Ivo Jimenez, Sebastiaan Alvarez Rodriguez, Alexandru Uta, Jeff LeFevre, and Carlos Maltzahn. Skyhook: Towards an Arrow-Native Storage System. CCGrid, 2022. [\[paper\]](#)
- Sebastiaan Alvarez Rodriguez, Jayjeet Chakraborty, Aaron Chu, Ivo Jimenez, Jeff LeFevre, Carlos Maltzahn, Alexandru Uta. Zero-Cost, Arrow-Enabled Data Interface for Apache Spark. SCDM, 2021. [\[paper\]](#)
- Jayjeet Chakraborty, Carlos Maltzahn, Ivo Jimenez. Enabling Seamless Execution of Computational and Data Science Workflows on HPC and Cloud with the Popper Container-Native Automation Engine. Paper at CANOPIE-HPC Workshop 2020, 12 November, 2020. [\[paper\]](#)
- Jayjeet Chakraborty, Ivo Jimenez, Carlos Maltzahn, Arshul Mansoori, Quincy Wofford. Popper 2.0: A Container-native Workflow Execution Engine For Testing Complex Applications and Validating Scientific Claims. Poster at 2020 Exascale Computing Project Annual Meeting, Houston, TX, February 3-7, 2020. [\[poster\]](#)

TALKS

- Presented a talk on "Embedding Apache Arrow inside Storage Systems" at The Data Thread conference held on June 23, 2022. [\[video\]](#) [\[slides\]](#)
- Presented SkyhookDM and its recent developments at the SNIA Storage Developers Conference 2021, September 28-29, 2021. [\[slides\]](#)
- Talked about "Reproducible and Automated Storage systems experimentation with Popper" at the Second K8S-HEP Meetup, December 1-2, 2020. [\[slides\]](#)
- Presented the paper entitled "Enabling Seamless Execution of Computational and Data Science Workflows on HPC and Cloud with the Popper Container-Native Automation Engine." at the CANOPIE-HPC Workshop, November 12, 2020. [\[slides\]](#)