Jayjeet Chakraborty

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

Introduction

I am a graduate student working in the field of big data processing, accelerated computing, HW/SW co-design, storage systems, and distributed systems. Currently, I am working on GPU accelerated data processing applications at NVIDIA as an intern in their RAPIDs cuDF team.

Education

- **University of California, Santa Cruz** — Santa Cruz, CA
  Master of Science, Computer Science And Engineering, CGPA: 3.50/4 — *Expected Grad.: January, 2025*
- **National Institute Of Technology, Durgapur** — Durgapur, India
  B.Tech, Computer Science And Engineering, CGPA: 7.65/10 — *Graduated: June, 2021*
- **Hem Sheela Model School, Durgapur** — Durgapur, India
  Senior Secondary Education, CBSE, Percentage: 92.6% — *Graduated: April, 2017*

Experience

- **GPU Software Engineering Intern (RAPIDs)** — Santa Clara, CA
  NVIDIA Corporation — *Summer '24*
  - Implemented TPC-H queries using low-level 'libcudf' APIs, performed benchmarks on platforms such as NVIDIA H100 and Grace Hopper, and profiled benchmark runs using Nsight Systems.
  - Implemented a TPC-H derived data generator using 'libcudf', CUDA, and Thrust following the TPC-H 3.0.1 specifications. Reduced data generation duration for scale factor 100
  - Integrated the 'libcudf' TPC-H query implementations with the data generator to build self-contained benchmarks using NVBench that can be run in automation allowing tracking the performance regression of 'libcudf' over time on TPC-H workloads.
  - Exposed a 'stream' parameter on several 'libcudf' public APIs to enable better stream-ordered execution. Additionally, worked on examples showing interoperation between 'StringView' (Umbra strings) and 'String' data types.
- **Software Engineering Intern** — Remote
  InfluxData Inc. — *Summer '23*
  - 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.
  - Wrote scripts for benchmarking DataFusion against DuckDB using TPC-H, ClickBench, and H2O.ai, on single and multicore platforms.
- **Graduate Student Researcher** — Santa Cruz, CA
  UC Santa Cruz — *Spring '22 - Present*
  - Studying ANN (Approximate Nearest Neighbour) algorithms and their performance characteristics with the goal of designing HW/SW co-designed systems for accelerating vector search and vector databases.
  - 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 Acero (from Apache Arrow) for the embedded compute engine. [code]

- Working on joining NanoEvents data generated from HEP experiments in Apache Arrow intermediate format using the Cylon framework.

- **IRIS-HEP Summer and Winter Fellow** Remote
  Princeton University and CROSS, UC Santa Cruz *Summer 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 SkyhookDM to store and process data in Arrow IPC and Parquet format and also extended the Arrow framework with a Skyhook Dataset API to be able to natively connect to a Ceph/RADOS cluster with SkyhookDM plugins and push-down compute tasks.
  - Completed working on SkyhookDM, an Arrow-Native storage system based on Ceph. Integrated Coffea, a distributed scientific data processing framework from CERN with Skyhook to enable compute offloading in HEP data analytics. Also, contributed the Skyhook project to Apache Arrow.

- **Winter Research Intern** Varanasi, India
  Indian Institute Of Technology, BHU *Winter '20*
  - 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 '19*
  - 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 '18*
  - 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.

Publications

- A. Lamb, Y. Shen, D. Heres, J. Chakraborty, M. O. Kabak, C. Sun, L-C. Hsieh. Apache Arrow DataFusion: A Fast, Embeddable, Modular Analytic Query Engine, SIGMOD 2024, Santiago, Chile. [paper]
- 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]

Presentations

- Presented an invited talk on "Analyzing the Performance of Vector Databases" at Broadcom Inc. on June 6, 2024.

- Presented a talk on "Towards Faster Columnar Data Transport using RDMA" at the Computational I/O Stack Workshop 2023, UC Santa Cruz, August 17, 2023. [slides]

- Presented a talk on "Optimizing Data Access with Compute Offloading, Fast Hardware-Accelerated Data Transport, and Modern Query Languages" at the PyHEP.dev Workshop 2023, Princeton University, New Jersey, July 25-28, 2023. [slides]

- Presented a talk on "Embedding Apache Arrow inside Storage Systems" at The Data Thread conference held on June 23, 2022. [video]

- Presented a demo + talk on "Data Management with Skyhook" at the AGC Tools Workshop, UW Madison, April 2022. [slides]

- Presented SkyhookDM and it's 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]

- Presented the IRIS-HEP Summer Fellowship project on making SkyhookDM and Ceph Experiments Reproducible at the CROSS Research Symposium, October 6-9, 2020 and at IRIS-HEP Topical Meeting, October 5, 2020. [slides]

Skills

Languages: C, C++, CUDA, Rust, Python, Go, Java, JavaScript, Bash, MATLAB.
Tools: Git, Docker/Kubernetes, Ansible, Azure/GCP/AWS, Prometheus/Grafana, Jaegar, Intel VTune/NVIDIA NSight Systems, Perf, GDB
Frameworks: Flask, React, Boost, Qt, NumPy, Pandas, Matplotlib, Seaborn, Android Framework.
Operating Systems: Debian/Ubuntu, CentOS, MacOS, WSL

Teaching/Mentoring Experience

- Served as a mentor in Google Summer Of Code 2021 and OSRE (Open Source Research Experience) 2021 as a part of the organization "Centre for Research in Open Source Software" at UC Santa Cruz.

- Took workshops on Open Source Software Development topics like Linux, Git, GitHub, and Python programming during college fests every year.

- Mentored 2 freshman year students from IIT Roorkee during Kharagpur Winter of Code (KWoC) 2018 on Back-End Web Development projects. Also, guiding new contributors in the getpopper.io community.