Sacheendra Talluri

Amsterdam | Dutch citizen | Willing to relocate or be remote   sacheendra.t@gmail.com | Website | GitHub | Google Scholar

Distributed systems engineer with 7+ years in performance analysis, large-scale cluster scheduling, and fault-tolerant cloud services. Delivered throughput, latency, and efficiency gains at IBM Research, Databricks, and startups. Skilled in benchmarking, workload modeling, and building schedulers, GPU query engines, and high-throughput pipelines.

Experience

Ph.D. Candidate, Vrije Universiteit Amsterdam - Amsterdam, the Netherlands   July 2019 – May 2025

- Reduced datacenter job runtimes by up to 24% by evaluating API integration between Spark and Kubernetes schedulers [4, 5].
- Led a long-term project to collect crowdsourced data about cloud and LLM service reliability [2, 3].
- Built the world's fastest GPU-based JSON query engine, achieving a 2.9x speedup over state-of-the-art [1].
- Modeled LLM fine-tuning, GPUs, and job schedulers in OpenDC datacenter simulator. Achieved 5.5-3.6% MAPE error [GitHub].
- Published in top-tier venues (VLDB, TPDS) and earned a Best Artifact Award for reproducible research tools.
- Mentored 12 BSc/MSc students, resulting in 4 publications, 1 best thesis award, and 1 runner-up.

Research Visitor, IBM Research - Haifa, Israel   July 2023 – September 2023

- Extended the Go compiler with dynamic dataflow analysis, improving debugging workflows [GitHub].
- Built a Kubernetes error diagnostics system that automatically links configuration issues to root causes, reducing error triage time.

Software Engineer, IPBurger - Remote   October 2018 – May 2019

- Reduced proxy connection time by 90% and improved uptime by deploying caching and replication strategies.
- Led a 3-person team in implementing the control plane for a fault-tolerant proxy infrastructure serving thousands of connections.

Database Internals Intern, Databricks - Amsterdam, the Netherlands   February 2018 – May 2018

- Analyzed 600 TB of Apache Spark logs to produce the first published study of large-scale storage access patterns [Paper].
- Simulated and benchmarked caching algorithms, revealing that simpler policies can match complex ones at large cache sizes [Paper].

Part-time Infrastructure Engineer, ReactiveSearch.io - Remote   January 2014 – April 2016

- Automated Elasticsearch cluster provisioning per customer, scaling to 100k ops/sec without manual intervention.
- Designed a stream processing pipeline leveraging Elasticsearch's query engine for real-time analytics.

Skills

Programming Languages: Python, Go, Java, Kotlin, JavaScript, Node.js, CUDA, C, C++, Scala
Platforms: Linux (Storage and Network APIs), Kubernetes, Apache Spark, AWS, GCP
Data Management: Elasticsearch, Redis, PostgreSQL, MongoDB, Apache Kafka, HDFS, Hive, MinIO, Zookeeper, etcd
Data Analysis: Numpy, Pandas, PyTorch
DevOps: Terraform, Docker, GitHub Actions, Prometheus, InfluxDB

Education

Vrije Universiteit Amsterdam Ph.D. in Computer Science (Thesis Completed)   Amsterdam, The Netherlands
Thesis: Design and Characteristics of Cloud Computing Systems Along the Performance-Failure Continuum   December 2025

Delft University of Technology M.Sc. in Computer Science – GPA: 8/10 (top 20%)   Delft, The Netherlands
Thesis: Workload Characterization, Modeling, and the Design of Cache Policies for Big Data Workloads   December 2018

Dhirubhai Ambani University B.Tech. in ICT – GPA: 8.6/10 (top 15%)   Gandhinagar, India

Other Experience

Secretary, SPEC (of SPEC CPU fame) Cloud research group   November 2019 – present
Lead TA and Lab Coordinator, Distributed Systems at VU Amsterdam   November 2019 – February 2025
Teaching Assistant, Storage Systems at VU Amsterdam   September 2020 – November 2022
Treasurer, TU Delft Debating Club   August 2017 – July 2018
Teaching Assistant, Big Data Processing & Distributed Systems at TU Delft   November 2017 – March 2018
Reviewer, ACM/SPEC ICPE, IEEE/ACM CCGRID, IEEE TPDS

Selected Publications

[1] **Sacheendra Talluri**, Guido Walter Di Donato, Luca Danelutti, Koen R. Vlaswinkel, Marco Arnaboldi, Arnaud Delamare, Marco D. Santambrogio, Daniele Bonetta. "GpJSON: High-performance JSON Data Processing on GPUs." VLDB 2025.

[2] **Sacheendra Talluri**, Dante Niewenhuis, Xiaoyu Chu, Jakob Kyselica, Mehmet Cetin, Alexander Balgavy, and Alexandru Iosup. "Cloud Uptime Archive: Open-Access Availability Data of Web, Cloud, and Gaming Services." Revision TPDS.

- Matthijs Jansen, **Sacheendra Talluri**, Krijn Doekemeijer, Nick Tehrany, Alexandru Iosup, Animesh Trivedi. "Columbo: A Reasoning Framework for Kubernetes' Configuration Space." ICPE 2025.

[3] Xiaoyu Chu, **Sacheendra Talluri**, Qingxian Lu, Alexandru Iosup. "An Empirical Characterization of Outages and Incidents in Public Services for Large Language Models." ICPE 2025.

[4] Aratz Manterola Lasa, **Sacheendra Talluri**, Tiziano De Matteis, Alexandru Iosup. "The Cost of Simplicity: Understanding Datacenter Scheduler Programming Abstractions." ICPE 2024.

[5] **Sacheendra Talluri**, Nikolas Herbst, Cristina L. Abad, Tiziano De Matteis, Alexandru Iosup. "ExDe: Design Space Exploration of Scheduler Architectures and Mechanisms for Serverless Data-processing." FGCS 2024.

Awards

AWS Cloud Research Credit 5k and 10k USD	2020, 2022
Best Artifact Award at ACSOS'21	2021
Graduate travel grant by ACM SIGSOFT and UCC	2019, 2022

References

Prof. dr. Alexandru Iosup, Full Professor, Vrije Universiteit Amsterdam	a.iosup@vu.nl
Dr. Animesh Trivedi, Senior Research Scientist, IBM Research Zurich	animesh.trivedi@ibm.com
Dr. Daniele Bonetta, Assistant Professor, Vrije Universiteit Amsterdam	d.bonetta@vu.nl