Nitin **Kedia**

Research Fellow, Microsoft Research

kedianitin.com | nitinkedia7@gmail.com | github.com/nitinkedia7 | Google Scholar

Experience

Present **Jul 2023** | **Microsoft Research** — **Bangalore, India**
Pre-Doctoral Research Fellow [🌐]
Mentors: Dr. Ramachandran Ramjee, Dr. Jayashree Mohan, and Dr. Nipun Kwatra
Designing efficient inference systems for Large Language Models. Published in **OSDI'24** and **MLSys'24**.

Jun 2023 **Jan 2023** | **Zeta [🌐]** — **Bangalore, India**
Senior Software Development Engineer | Mentor: Dharmendra Patel
Drove the stability and scalability of Zeta's web platform, successfully delivering it to the company's biggest clients both in India (HDFC Bank – India's largest Private Bank) and the US (FIS).
Designed and implemented a Kubernetes Operator to automate the deployment of web applications, piloting an organization-wide initiative that reduced customer onboarding time from months to days.

Dec 2022 **Jul 2021**
Software Development Engineer II | Mentor: Apurva Jaiswal
Led the design and development of Zeta's API Playground, adopted by 116 internal services.
This became a primary resource for all API documentation and is used to showcase the company's API-enabled stack at industry festivals and client demos.

Jun 2021 **Jul 2020**
Software Development Engineer I | Mentor: Apurva Jaiswal
Developed a frontend application for the internal Pub-Sub service, gaining valuable experience with Zeta's web platform. Subsequently, contributed to the platform by building new features and resolving bugs.

May 2020 **Aug 2019** | **IIT Guwahati** — **Guwahati, India**
Undergraduate Researcher | Mentor: Prof. Moumita Patra
Designed a leader election and task allocation scheme for resource-sharing in vehicular clouds, an emerging paradigm. Developed a simulator to validate the technique.

Jul 2019 **May 2019** | **Flipkart** — **Bangalore, India**
Software Development Intern | Mentor: Sachin Arya
Developed the Reports module for Flipkart's Ads web app using React and GraphQL.

Education

Jun 2020 **Jul 2016** | **Indian Institute of Technology Guwahati** — **Guwahati, India**
Bachelor of Technology, Computer Science and Engineering (GPA 9.32/10.0)

Publications

Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve [pdf] [**OSDI'24**]
Amey Agrawal, **Nitin Kedia**, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee
Published in the 18th USENIX Symposium on Operating Systems Design and Implementation, 2024

Vidur: A Large Scale Simulation Framework For LLM Inference [pdf] [**MLSys'24**]
Amey Agrawal, **Nitin Kedia**, Jayashree Mohan, Ashish Panwar, Nipun Kwatra, Bhargav S. Gulavani, Ramachandran Ramjee, and Alexey Tumanov
Published in the 7th Annual Conference on Machine Learning and Systems, 2024

On Evaluating Performance of LLM Inference Serving Systems [**In Review**]
Amey Agrawal, **Nitin Kedia**, Anmol Agarwal, Jayashree Mohan, Nipun Kwatra, Souvik Kundu, Ramachandran Ramjee, and Alexey Tumanov

Open Source Contributions

Vidur [🐙] We open-sourced the first LLM Inference System Simulator. Has earned 284 GitHub stars till date.

Sarathi-Serve [🐙] Co-authored and open-sourced the artifact reproduced code for Sarathi-Serve, which has garnered 255 GitHub stars. This technique was upstreamed into vLLM by AnyScale, a prominent LLM inference framework. [RFC]

Talks And Posters

Taming Throughput-Latency Trade-off in LLM Inference with Sarathi-Serve

- Talk and poster at OSDI [📹] July 2024 (Santa Clara, California, USA)
- Poster at Microsoft Research Academic Summit [🌐] Jun 2024 (Bangalore, India)

Vidur: A Large-Scale Simulator For LLM Inference Systems

- Poster at MLSys [🌐] May 2024 (Santa Clara, California, USA)
- Talk at Microsoft Research Academic Summit [🌐] Jun 2024 (Bangalore, India)

Awards and Achievements

IIT Joint Entrance Examination (Advanced), 2016 Achieved **All India Rank 601** among 1.2 million candidates.

USENIX OSDI Diversity Grant, 2024 To co-present our talk and poster at the conference.

Awards at Zeta

- Recognised as an **Outstanding Performer twice (2022 and 2021)** in annual performance reviews for consistently exceeding expectations in technical contributions and project delivery.
- Received the **The Mountain Mover award in Q3 2022** for proactively stabilising the static assets serving infrastructure, which had previously caused high-severity incidents, significantly improving system reliability.
- **Ultimate Team award recipient in Q2 2022** for successfully delivering the entire web platform in a standalone installation for its biggest client, meeting strict security requirements within a short timeframe.
- **Ultimate Team award recipient in Q4 2021** for rapidly delivering significant architectural improvements to the API Playground, culminating in a successful demo at a key industry festival.

Institute Merit-cum-Means Scholarship, IIT Guwahati, 2016-20 Recipient of university's scholarship, waiving off 100% of the tuition fee

Top 7.4% Globally in Competitive Programming

- **Codeforces: Achieved a rating of 1852 (92.58th Percentile)**, competing in 49 programming contests which routinely break 10k participants [🌐]
- **Google Kickstart 2019: Ranked 122 globally in Round H** of this contest series among 2100 candidates.

Professional Service

Team Development at Zeta

- **Took over 100 problem solving interviews** (Data Structures and Algorithms).
- **Mentored and onboarded 5 new joiners** including an intern who secured a full-time offer.

Site Reliability Engineering

- **Served as on-call for 18 weeks** at Zeta, solving incidents, issues and queries from both internal and external customers.