

# Siddhant Ray

Chicago, IL, USA  
+1-(773)-457-4156  
siddhant.r98@gmail.com  
<https://siddhant-ray.github.io/>  
<https://github.com/Siddhant-Ray> (GitHub)

## Education

- 2023 – 2028 **The University of Chicago**, *PhD in Computer Science*  
Advisor - Junchen Jiang and Nick Feamster
- 2020 – 2022 **ETH Zürich**, *MSc in Electrical Engineering and Information Technology*  
Advisor - Laurent Vanbever
- 2016 – 2020 **VIT Vellore**, *B.Tech in Electronics and Communication Engineering*

## Experience

- Sep 2023 – **Graduate Research Assistant**, *Computer Science Department, The University of Chicago*  
Present
  - *Project 1*: Developing Transformer based models for sharp latency change prediction to enable packet-level queue management for tail-latency reduction.
  - *Project 2*: Joint quality-latency optimization for Retrieval Augmented Generation(RAG) LLM systems by combining query-level configuration selection and resource scheduling.
- Sep 2022 – **Cloud Networks Researcher**, *Advanced Network Architectures Lab, UPC Barcelona*  
Mar 2023
  - Reinforcement learning based resource sharing, offloading and allocation for cloud-edge systems.
  - Developed an approximation for a Mixed-Integer Optimal Matching Algorithm for edge resource allocation to reduce execution time by 2.5-3X.
- Oct 2021 – **Graduate Research Assistant**, *Law, Economics, and Data Science Group, ETH Zurich*  
Sep 2022
  - *Project 1*: Improving semantic labelling for text corpora using new NLP models, sentence simplification, paraphrase mining and clustering for topic modelling.
  - *Project 2*: Creating NLP (RoBERTa based) models to analyse political discourse in meat policy documents to enable actor-narrative clustering.
- May 2019 – **Software Development Intern**, *Capgemini Engineering*  
July 2019
  - Developed a K-Shortest Path Searching algorithm in an ONOS based Software Defined Layer 2 VPN.
  - Algorithm applied dynamic constraints of network resources (e.g.required edges) to be used for path calculation.
- May 2018 – **Software Development Intern**, *BlueStacks*  
July 2018
  - Worked on a machine learning algorithm to predict the App Engine's appropriate display screen based on the customer's past experiences.
  - Developed an automation script for generating SVG cards for the App Engine's game front end and an address verification tool using the EasyPost API.

## Publications

- 2024 Jiayi Yao, Hanchen Li, Yuhua Liu, **Siddhant Ray**, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. *CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion*, 2024. To appear at EuroSys 2025.
- 2024 **Siddhant Ray**, Xi Jiang, Jack Luo, Nick Feamster, and Junchen Jiang. *SwiftQueue: Optimizing Low-Latency Applications with Swift Packet Queuing*, 2024. In Submission.
- 2024 Yuhua Liu, Hanchen Li, Yihua Cheng, **Siddhant Ray**, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, and Junchen Jiang. *CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving*. In *Proceedings of the ACM SIGCOMM 2024 Conference*, 2024.

- 2024 Hanchen Li, Yuhua Liu, Yihua Cheng, **Siddhant Ray**, Kuntai Du, and Junchen Jiang. *Eloquent: A More Robust Transmission Scheme for LLM Token Streaming*. In *Proceedings of the 2024 SIGCOMM Workshop on Networks for AI Computing*, 2024.
- 2022 Alexander Dietmüller, **Siddhant Ray**, Romain Jacob, and Laurent Vanbever. *A New Hope for Network Model Generalization*. In *Proceedings of the 21st ACM Workshop on Hot Topics in Networks*, 2022.
- 2020 **Siddhant Ray** and Budhaditya Bhattacharyya. *Machine Learning based Cell Association for mMTC 5G Communication Networks*. *International Journal of Mobile Network Design and Innovation*, 10(1):10–16, 2020.

## Selected Projects

- 2022 Advancing Packet-Level Traffic Predictions with Transformers (Master Thesis) - [code, thesis]
- 2021 Towards a New Framework for Integration of Network Planes (Research Project) - [code]
- 2021 Attentive Neural Networks for News Classification (Research Project) - [code]
- 2021 Investigating Possible Inductive Biases in Local Sparse Attention ViT Architectures Against Traditional CNNs (Course Project) - [code, paper]
- 2020 Maximizing Cross Traffic Flows in a L2/L3 Network with Programmable Switches (Course Project) - [code, poster]
- 2020 Machine Learning based Cell Association for 5G Communication Networks (Bachelor Thesis) - [code]

## Honors and Awards

- 2023 – 2028 **Liew Family Graduate Fellowship**, University of Chicago
- 2022 **Winner at Datathon**, *Microsoft Challenge*, ETH Zurich
- 2020 **Best Outgoing Student**, *SENSE department*, VIT Vellore
- 2019 **Runner-Up at VIT Hack**, *Education Track*, VIT Vellore
- 2016 – 2019 **Merit Scholarship for Academic Excellence**, VIT Vellore

## Service

- CoNEXT'24 **Artifacts Evaluation Committee**

## Relevant Courses

- Graduate Approximation Algorithms, Algorithms, Advanced Computer Networks, System Security, Network Security, Distributed Computing, Discrete Event Systems, Networks Seminar, Deep Learning, Learning Theory, Mathematics of Data Science, Neural Network Theory
- Undergraduate Computer Networks, Operating Systems, Wireless Communication, Linear Algebra

## Technical Skills

- Programming Python, C++, C, Bash, Rust, SQL, Java, T<sub>E</sub>X
- Software Linux, Git, Docker, P4 switches, ONOS, Google Cloud, AWS, Maven, MATLAB, NetSim, Cadence
- Frameworks Mininet, FRRouting, PyTorch, TensorFlow, Sklearn, NLTK, Flask, SciPy, Scapy, BS4, NS-3, Langchain, vLLM

## Leadership and Volunteering

- 2019 – 2020 **Technical Advisor**, IETE VIT
- 2018 – 2019 **Organizer**, TEDx VIT Vellore
- 2017 – 2020 **President** (2018 – 2019) & **Outreach Worker**, Anokha NGO