Shubham **Tiwari**

PhD student, University of Washington

shubhamtiwari.net @ tshubham@cs.washington.edu github.com/sbhtwr Google Scholar

+1 (206) 853-0284

Research Interests

Broadly interested in the intersection of distributed systems and machine learning. My current focus is on efficient LLM serving via workload-aware, tiered KVCache management. Previously, I have worked on a broad spectrum of problems – memory management optimizations in hypervisors, network measurement, and congestion control (LEOScope, iBox).

Education

University of Washington, *Seattle* Sept 2023 - Present
Ph.D. in Computer Science (*ongoing*)
Advisors: Simon Peter, Ratul Mahajan

Birla Institute of Technology and Science (BITS), Pilani Aug 2016 - July 2021
B.E. Computer Science and M.Sc. Mathematics
Thesis: *Data-Driven Network Simulation with iBox*

Experience

Microsoft Research, *Redmond* June 2024 - Sept 2024
Research Intern with *Ishai Menache*
Project: Improvements to Azure Compute's VM allocation service.

Microsoft Research, *Bangalore* Aug 2021 - Aug 2023
Research Fellow with *Debopam Bhattacherjee*, *Venkat Padmanabhan*
Projects: LEO Satellite Networks (LEOScope)

Microsoft Research, *Bangalore* Jan 2021 - July 2021
Research Intern with *Venkat Padmanabhan*, *Nagarajan Natarajan*
Project: Data-Driven Network Simulation (iBox)

VMware, *Bangalore* Aug 2020 - Dec 2020
Intern, xLabs with *Jayneel Gandhi*
Project: Page-table Replication (Mitosis)

Samsung Research, *Bangalore* May 2020 - July 2020
Research Intern
Project: Cellular Network Planning

Software-Defined Networking Lab, *BITS Pilani* Jan 2019 - Dec 2019
Research Assistant with *K. Hari Babu*
Project: Passive Estimation of Link Latency (qMon)

Publications

C=Conference, J=Journal, P=Preprint, A=Article

A.1 LEOScope: Building a Global Testbed for Low-Earth Orbit Satellite Networks
Saeed Fadaei, Shubham Tiwari, Aryan Taneja, Saksham Bhushan, Mohamed Kassem, Aravindh Raman, Debopam Bhattacherjee, Lili Qiu, Alan Woodward, Nishanth Sastry
SIGCOMM Computer Communication Review `[nominated for Best of CCR]` **SIGCOMM CCR'25**

C.1 **Boosting Application Performance using Heterogeneous Virtual Channels: Challenges and Opportunities**
Talal Touseef, William Sentosa, Milind Kumar Vaddiraju, Debopam Bhattacherjee, Balakrishnan Chandrasekaran, Brighten Godfrey, Shubham Tiwari
22^{nd} *ACM Workshop on Hot Topics in Networks* **HotNets'23**

P.1 **T3P: Demystifying Low-Earth Orbit Satellite Broadband**
Shubham Tiwari, Saksham Bhushan, Aryan Taneja, Mohamed Kassem, Cheng Luo, Cong Zhou, Zhiyuan He, Aravindh Raman, Nishanth Sastry, Lili Qiu, Debopam Bhattacherjee
Preprint

C.2 **Simulating Network Paths with Recurrent Buffering Units**
Divyam Anshumaan, Sriram Balasubramanian, Shubham Tiwari, Nagarajan Natarajan, Sundararajan Sellamanickam, and Venkata N. Padmanabhan
37^{th} *AAAI Conference on Artificial Intelligence* **AAAI'23**

C.3 **Data-Driven Network Path Simulation with iBox**
Sachin Ashok, Shubham Tiwari, Nagarajan Natarajan, Venkata N. Padmanabhan, and Sundararajan Sellamanickam
ACM SIGMETRICS / IFIP PERFORMANCE 2022 **SIGMETRICS'22**

J.1 **qMon: A method to monitor queueing delay in OpenFlow networks**
Sandhya Rathee, Shubham Tiwari, K Haribabu, and Ashutosh Bhatia
Journal of Communications and Networks **JCN'22**

Projects

ElasticCache: Efficient LLM Serving via Workload-aware, Tiered KVCache Management April 2024 - Present
Advisors: Simon Peter, Ratul Mahajan

- Developing a workload-aware serving system that profiles LLM workflows to determine cache access patterns for efficient utilization of tiered, disaggregated KVCache.
- Working on building a prototype (scheduler + tiered KVCache) on top of vLLM to evaluate our techniques on various LLM workflows and traffic patterns.

Optimizations to Azure Compute's VM Allocation Service June 2024 - Sept 2024
Advisors: Ishai Menache

- Evaluated the impact of enabling fine-grained VM placement strategies at scale.
- Proposed changes to resource allocation service with the potential of saving ($> \$1$M) in operational costs.

LEOScope: Enabling Experimentation Across Low-Earth Orbit (LEO) Satellite Networks July 2022 - Present
Advisors: Debopam Bhattacherjee, Venkat Padmanabhan [code]

- Lead an effort with **Azure Space**, **MSRA**, and academic collaborators to build a platform of a global scale for experimentation across Low-Earth Orbit Satellite networks.
- Drove the effort through engineering challenges such as platform architecture, implementation of experiment scheduler, executor, and the central orchestrator.
- Initiated large-scale measurements based on ping and iperf to characterize satellite network paths.

iBox: Internet in a Box Jan 2021 - June 2022
Advisors: Venkat Padmanabhan, Nagarajan Natarajan [website]

- Built a data-driven network simulator that uses data to recreate end-to-end behavior of a network path.
- Leveraged a combination of internet measurement data and ML models to capture the impact of complex network phenomena such as cross-traffic and packet reordering.
- Integrated iBox with ns-2, ns-3, netem and **Microsoft Teams's** in-house network simulator.
- Resulting papers published at SIGMETRICS'22 and AAAI'23.

qMon: Passive Delay Monitoring in SDNs Jan 2019 - Dec 2019

Advisor: K. Hari Babu [paper] [code]

- Devised **qMon**, a scalable latency monitoring technique with zero data plane footprint.
- Developed an Open vSwitch based prototype to fetch queue length information using OpenFlow and passively estimate link latency.
- Evaluated qMon on a physical testbed under various traffic scenarios.
- Resulting paper published at JCN 2022.

Mitosis: Enabling Page-Table Replication in ESXi Aug 2020 - Dec 2020

Advisor: Jayneel Gandhi

- Implemented page-table replication (Mitosis) in VMware's core virtualization product – ESX.
- Developed prototypes to disambiguate the design and code needed to support page table replication in the ESX kernel.
- Conducted workload profiling to estimate the performance benefits of page-table replication; realized gains of upto 17% in workload execution time.

Miscellaneous

- Demonstrated iBox at TAB – MSR India's annual technical event.
- Presented iBox at SIGMETRICS'22. [video]
- Awarded a grant of 50,000 INR by AUGSD, BITS Pilani for developing a miniature autonomous driving vehicle.