Nirav Atre

21 Unger Ln, Pittsburgh, PA 15217
+1 (669) 377-4414 • natre@cs.cmu.edu • https://www.cs.cmu.edu/~natre

Education

Carnegie Mellon University • Pittsburgh, PA, USA — Aug, 2018 – Present
Ph.D. in Computer Science (Computer Networking)
Advisor: Prof. Justine Sherry

University of Toronto • Toronto, Canada — Sep, 2013 – May, 2018
B.A.Sc. in Computer Engineering
Advisors: Prof. Jonathan Rose, Prof. Ravi Adve
- CGPA: 3.9/4.0 • Graduated with highest honors in engineering

Research Interests

Computer systems, computer networking, performance modeling. Building high-performance networked systems (on x86, FPGAs) and proving theoretical properties about their performance and stability.

Honors and Awards

- CNIB Hochhausen Prize for Excellence in Accessible Design in Engineering. *CNIB, University of Toronto.* 2018
- Certificate of Recognition for Outstanding Senior Design (Capstone) Project. *University of Toronto.* 2018
- NSERC University of Toronto Excellence Award (UTEA-NSE). *NSERC, University of Toronto.* 2015
- Applied Science and Engineering (APSC) Undergraduate Research Fellowship. *University of Toronto.* 2014
- Dean's Honor List (8 of 8 semesters). *University of Toronto.* 2013 – 2018

Publications

- Zhipeng Zhao, Hugo Sadok, **Nirav Atre**, James C. Hoe, Vyas Sekar, and Justine Sherry. Achieving 100Gbps Intrusion Prevention on a Single Server. In Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation (**OSDI**), November 2020.
- **Nirav Atre**, Justine Sherry, Weina Wang, and Daniel S. Berger. Caching with Delayed Hits. In Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (**SIGCOMM**), August 2020.

Industry Experience

Software Engineering Intern, Microsoft — May, 2020 – Present
Azure Physical Networking (PhyNet) team — *Virtual*
- Helped develop a performance model for the next generation of Azure's SmartNIC
- Designed and implemented a simulator to help uncover NIC performance bottlenecks

Software Engineering Intern, Google — May, 2018 – Aug, 2018
Android Project Treble team — *Mountain View, CA, USA*
- Contributed to a source-to-source compiler (*hidl-gen*) for Android HIDL (HAL Interface Definition Language)
- Helped implement several framework-level improvements for Android Q, including Safe Unions and Java Native Handles (the corresponding AOSP changelists are public, and can be found here)

Software Engineering Intern, Intel — Sep, 2016 – Aug, 2017
Deep Learning Accelerator (DLA) team for Intel FPGAs — *Toronto, Canada*
- Helped develop a C++ software stack for the DLA from the ground-up (pre-Alpha contributor)
- Implemented key features, including a cache-aware memory management framework for the accelerator overlay
- Contributed to Intel's LLVM-based OpenCL compiler (*i++*), a high-level synthesis (HLS) tool for Intel FPGAs

Software Engineering Intern, Google — May, 2016 – Aug, 2016
Android Nearby team — *Mountain View, CA, USA*
- Designed a fast, audio-based indoor localization protocol, realizing a 3X speedup over the existing design
- Implemented a prototype audio codec in C++ (native code), including several optimized DSP components

Research Projects

Adversarial Job Scheduling in Network Functions, SNAP Lab, CMU 2020 – Present

Denial-of-Service (DoS) attacks are the bane of enterprise-scale NF deployments. An important consideration in the design of public-facing NFs is the worst-case 'Amplification Factor' (AF) of a DoS attack. In this work, we design a practical scheduling policy with theoretical bounds on the worst-case AF. However, unlike prior work, we make no assumptions about the system's ability to distinguish between attack/innocent traffic, and we impose no hard limits on the resources consumed by each packet or flow, significantly increasing the applicability of our approach.

Caching with Delayed Hits, SNAP Lab, CMU 2018 – 2020

Traditional caching models assume that cache requests result in one of two outcomes: hits and misses. In reality, high-throughput caches experience a phenomenon known as 'delayed hits', which subvert expectations of caching models and simulators. In this work, we quantify the effect of delayed hits on practical caching systems. We then develop BELATEDLY, a latency-optimal, offline caching strategy for the delayed hits problem. Using insights from BELATEDLY, we present a practical caching strategy that outperforms existing algorithms by 10-40%.

An FPGA-based Flow Table with Constant Tail Latency, SNAP Lab, CMU 2018 – 2019

Traditional FPGA implementations of Flow Tables rely on caching, relegating uncached flows to DRAM. However, for latency-sensitive datacenter applications, DRAM accesses impose a significant performance overhead. In this work, we design a high-density, concurrent Flow Table that serves *all* steady-state connections with a deterministic, constant-time worst-case latency. The resulting data-structure is implemented as part of an FPGA-based Intrusion Detection System capable of achieving 100Gbps.

Continuous, Real-Time Face Recognition on Mobile Devices, University of Toronto 2017 – 2018
Senior Design (Capstone) Project, Advisor: *Prof. Ravi Adve*

To bridge the gap between (a) poor accuracy and generality of existing mobile-based face recognition algorithms, and (b) low throughput of state-of-the-art Deep Neural Networks on embedded devices, we used recent advancements in statistical learning theory to develop an accurate, yet real-time, *offline* facial recognition system.

Implementing Real-Time Eye-Tracking in a Mobile Context, University of Toronto Summer, 2015
Summer Research Intern, Advisor: *Prof. Jonathan Rose*

- Helped improve the accuracy of an infrared-based gaze-point estimator by 15%
- Developed a regression test suite and fully-automated build/test pipeline for the eye-tracking software in Python

Funded by the *University of Toronto Excellence Award* (UTEA) in Natural Sciences and Engineering (NSE).

Smartphone-Based Automation of Ankle Physiotherapy, University of Toronto Summer, 2014
Summer Research Fellow, Advisor: *Prof. Jonathan Rose*

- Contributed to myAnkle, an Android application to quantitatively measure balance
- Presented the research at an undergraduate research conference (UnERD, 2014) at the University of Toronto

Funded by an *APSC Undergraduate Research Fellowship*.

Research Talks

Caching with Delayed Hits

- at ACM SIGCOMM August, 2020
- at the University of Cambridge (Systems Research Group Seminar) July, 2020
- at Microsoft Research (MSR Systems Research Group) July, 2020

Teaching Experience

Undergraduate Teaching Assistant, University of Toronto Aug, 2015 – Apr, 2016

TA for first-year Calculus and Programming (C++) courses. Prepared practice exams, discussion outlines, and organized bi-weekly review sessions for 20+ students.