Alireza Farshin

Networked System Researcher

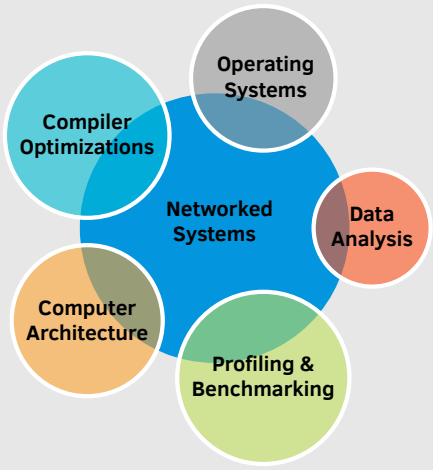
aliireza.github.io

alireza-farshin

Google Scholar

aliireza

Technical Skills Overview



Programming & Tools

Linux • Shell Scripting • C • C++

DPDK • FastClick • Perf • LLVM

Git • Gnuplot • LaTeX

Python • MATLAB • R

TensorFlow • Spark • Pandas

Education

PhD., Information Communication Technology
Specialization: Communication Systems
School of EECS
KTH Royal Institute of Technology
2017 - 2023 | Stockholm, Sweden

MSc., Electrical Engineering
Specialization: Digital Electronic Circuits
Amirkabir University of Technology
2015 - 2017 | Tehran, Iran

BSc., Electrical Engineering
Specialization: Electronics
Sharif University of Technology
2010 - 2015 | Tehran, Iran

Research Summary

I have improved the performance of the Network Functions Virtualization (NFV) service chains running at 100/200-Gbps commodity hardware by using **multi-disciplinary low-level optimization** techniques. I have been actively doing (networked) system programming in C/C++ and running automated experiments on the UNIX environment (e.g., Linux), using bash & Python scripts. My research has resulted in:

- **Google PhD fellowship 2021** award in systems and networking, which recognizes outstanding graduate students doing exceptional and innovative research in areas relevant to computer science.
- Top systems conference papers (**EuroSys'19**, **ATC'20**, **ASPLOS'21**, and **NSDI'22**).
- 20 patent applications filed in collaboration with Ericsson research (see **here**).
- News articles at Ericsson Blog (about **memory management**, **packet processing**, and **packet reordering**), KTH news (about **CPU cache** and **traffic order**), **Tech Xplore**, and **Framtidens Forskning**.
- Community award at NSDI'22 for "**Packet Order Matters**".
- Open-source contributions (**CacheDirector**, **DDIO-Bench**, **PacketMill**, **DDC-RA**, and **IOMMU-Bench**).

Watch this **video** and read my full **CV** & **dissertation** for more information.

Experience

Jan 2024 - Now — **Distributed Systems Researcher at Networking SW Arch.** — **NVIDIA**
Stockholm, Sweden

Aug 2023 - Jan 2024 — **Senior Researcher at Connected Intelligence Unit** — **RISE**
Stockholm, Sweden

- Improving packet processing at multi-100-Gbps rates.
- Using large language models (LLMs) to build and configure networked systems (see **NetBuddy**).
- Developing pruning techniques and improving inference of LLMs.

Aug 2017 - Aug 2023 — **Postdoctoral and Doctoral Researcher** — **KTH NSlab**
Stockholm, Sweden
Advisors: Professor **Dejan Kostić** & Professor **Gerald Q. Maguire Jr.**

- Proposed a slice-aware memory management technique to exploit the non-uniform cache architecture (NUCA) in Intel processors and implemented **CacheDirector** to send packets to the right slice of the Last-Level Cache (LLC).
- Implemented a set of benchmarks (**DDIO-Bench**) to study the effectiveness of Data Direct I/O Technology (DDIO) at 100 Gbps.
- Implemented **PacketMill** to grind the whole packet processing stack and produce a customized binary for a given network function.
- Analyzed a KTH campus trace (via Spark & Pandas) to extract flow-related characteristics & predict packet interarrival time with LSTM (via Keras/TensorFlow) to reorder packets using **Reframer**.
- Implemented a set of benchmarks (**IOMMU-Bench**) to study the impact of IOTLB misses on throughput and extended **Page Pool API** & mlx5 Linux driver to use 2-MiB hugepages for packet buffers to mitigate the IOTLB wall.
- Designed & Implemented a constraint-based C++ framework (**DDC-RA**) to allocate resources in a disaggregated data center.

Tools: DPDK, FastClick, Linux kernel, iPerf, Perf, LLVM, Intel PCM, Intel PMU Profiling Tools, Cache Allocation Technology (CAT), Spark, Pandas, Tensorflow, Gecode.
Hardware: Intel Xeon & AMD EPYC processors, NVIDIA/Mellanox & Intel NICs.