Alireza Farshin | Curriculum Vitae

aliireza.github.io • Alireza Farshin • aliireza
alireza-farshin • alirezafarshin

I am a distributed systems researcher at NVIDIA. My research interests include computer networks and networked systems. During my doctoral studies, I improved the performance of Network Functions Virtualization (NFV) service chains by using **low-level optimization** techniques. Watch

Work Experience

- **Networking Software & Systems Research Team at NVIDIA** — **Stockholm, Sweden**
  Distributed Systems Researcher/Software Architect — *January 2024–now*

- **Connected Intelligence Unit at RISE Research Institutes of Sweden** — **Stockholm, Sweden**
  Senior Researcher in AI / Machine Learning and Networking — *August 2023–January 2024*
  - Improving packet processing at multi-100-Gbps rates (see FAJITA).
  - Using large language models (LLMs) to build and configure networked systems (see NetBuddy, FlowMage, and NetConfEval).
  - Developing pruning techniques and improving inference of LLMs (see [W2]).

- **Network Systems Laboratory (NSLab) at KTH** — **Stockholm, Sweden**
  Postdoctoral Researcher — *March 2023–August 2023*

- **Network Systems Laboratory (NSLab) at KTH** — **Stockholm, Sweden**
  Doctoral Researcher — *August 2017–March 2023*

- **ICT Doctoral Programme Council at KTH** — **Stockholm, Sweden**
  Student Representative of the Division of Communication Systems (CoS) — *May 2018–December 2020*

- **Mobile Telecommunication Company of Iran (MCCI)** — **Tehran, Iran**
  Portal Specialist — *December 2015–June 2016*
  Vendor Manager & Portal/Application Supervisor:
  - eCare Application: My MCI Application for iOS and Android
  - eSales Website: eVoucher

- **CafeYab** — **Tehran, Iran**
  Co-founder and CEO — *Fall-2013*
  An application for iOS and Android for finding nearby Coffee Shops

- **Informatics Services Corporation (ISC)** — **Tehran, Iran**
  Internship — *June 2013–September 2013*
  - Ported an RF unit controller from PIC-16F877A to AtMega64A and tested the new module.
  - Designed a remote-control system with HM-T and HM-R FSK modules.

Publications

Conference Publications

[C1] Hamid Ghasemirahni, **Alireza Farshin**, Mariano Scazzariello, Gerald Q. Maguire Jr., Dejan Kostić, Marco Chiesa. FAJITA: Stateful Packet Processing at 100 Million pps, *The 20th International Conference on emerging Networking EXperiments and Technologies (**CoNEXT**)*. 2024. Download

[C2] Changjie Wang, Mariano Scazzariello, **Alireza Farshin**, Simone Ferlin, Dejan Kostić, Marco Chiesa. NetConfEval: Can LLMs Facilitate Network Configuration?, *The 20th International Conference on emerging Networking EXperiments and Technologies (**CoNEXT**)*. 2024. [**IRTF/IETF ANRP Winner!**] Download

[C3] Hamid Ghasemirahni, Tom Barbette, Georgios Katsikas, **Alireza Farshin**, Massimo Girondi, Amir Roozbeh, Marco Chiesa, Gerald Q. Maguire Jr., Dejan Kostić. Packet Order Matters! Improving Application Performance by Deliberately Delaying Packets In *19th USENIX Symposium on Networked Systems Design and Implementation (**NSDI**)*. 2022. Acceptance rate (Spring): 28/104 ≈ 26.9%. [**Community Award Winner!**] Download

[C4] **Alireza Farshin**, Tom Barbette, Amir Roozbeh, Gerald Q. Maguire Jr., Dejan Kostić. PacketMill: Toward per-core 100-Gbps Networking In *International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*. 2021. Acceptance rate: 75/398 ≈ 18.8%. Download

[C5] **Alireza Farshin**, Amir Roozbeh, Gerald Q. Maguire Jr., Dejan Kostić. Reexamining Direct Cache Access to Optimize I/O Intensive Applications for Multi-hundred-gigabit Networks In *USENIX Annual Technical Conference (**ATC**)*. 2020. Acceptance rate: 65/348 ≈ 18.6%. Download

[C6] **Alireza Farshin**, Amir Roozbeh, Gerald Q. Maguire Jr., Dejan Kostić. Make the Most out of Last Level Cache in Intel Processors In *The European Conference on Computer Systems (**EuroSys**)*. 2019. Acceptance rate: 45/207 ≈ 21.7%. Download

Journal Publications

[J1] **Alireza Farshin**, Luigi Rizzo, Khaled Elmeleegy, Dejan Kostić. Overcoming the IOTLB wall for multi-100-Gbps Linux-based networking In *PeerJ Computer Science (**PeerJ CS**)*. 2023. **Impact factor: 2.41**. Download

[J2] **Alireza Farshin**, Saeed Sharifian. A modified knowledge-based ant colony algorithm for virtual machine placement and simultaneous routing of NFV in distributed cloud architecture In *The Journal of Supercomputing (**SUPE**)*. 2019. **Impact factor: 2.469**. Download

[J3] **Alireza Farshin**, Saeed Sharifian. A chaotic grey wolf controller allocator for Software Defined Mobile Network (SDMN) for 5th generation of cloud-based cellular systems (5G) In *The Journal of Computer Communications (**COMCOM**)*. 2017. **Impact factor: 2.816**. Download

[J4] **Alireza Farshin**, Saeed Sharifian. MAP-SDN: a metaheuristic assignment and provisioning SDN framework for cloud datacenters In *The Journal of Supercomputing (**SUPE**)*. 2017. **Impact factor: 2.469**. Download

Patent Applications

[P1] **Alireza Farshin**, Omri Kahalon, Liran Liss, Aviad Shaul Yehezkel, Vishwanath Venkatesan. Distributed Memory with Pre-Computed Key and Value (KV) Entries for Inference Operations. US Patent Application 19/173,029. Filed in April 2025.

[P2] **Alireza Farshin**, Omri Kahalon, Liran Liss, Aviad Shaul Yehezkel, Vishwanath Venkatesan. Distributed Memory with Subsets of Model Weights or Parameters based on Inference Task Types. US Patent Application 19/173,027. Filed in April 2025.

[P3] **Alireza Farshin**, Omri Kahalon, Liran Liss, Aviad Shaul Yehezkel, Vishwanath Venkatesan. Distributed Memory of Inference Operations. US Patent Application 19/173,023. Filed in April 2025.

[P4] **Alireza Farshin**, Omri Kahalon, Vishwanath Venkatesan, Timothy Stamler. Orchestration of Distributed Inference Operations. US Patent Application 18/926,233. Filed in October 2024.

[P5] Amir Roozbeh, **Alireza Farshin**, Marco Chiesa, Dejan Kostić. Network Entity and Method Performed Therein for Handling one or more Packets in a Computer Environment (*Other name:* System and Methods for Minimizing Branch Mispredictions and Executing Highly Optimized Code for Networking Applications). PCT Application PCT/SE2023/050880. Download

[P6] Amir Roozbeh, **Alireza Farshin**, Marco Chiesa. Network Entity and Method Performed Therein for Handling one or more Packets in a Computer Environment (*Other name:* System and Methods for

Programmatically Storing Packet Payloads in the Per-Port Queue Memory). US Provisional Patent Application 63/511,198. Filed in June 2023.

[P7] Amir Roozbeh, **Alireza Farshin**, Marco Chiesa. System and Method Performed Therein for Handling one or more Packets in a Computer Environment (*Other name:* System and Methods for Disaggregated Packet Construction). PCT Application PCT/SE2023/050538. Download

[P8] Amir Roozbeh, **Alireza Farshin**, Marco Chiesa. Entity and Method Performed Therein for Handling Packets in a Computer Environment (*Other name:* System and Methods for Performing Millions Low-Latency Key-Value Insertion on Switches). PCT Application PCT/SE2023/051174. Download

[P9] Amir Roozbeh, **Alireza Farshin**, Marco Chiesa, Dejan Kostić, Hamid Ghasemirahni. Hint Entity, Receiver Node, System and Methods Performed Therein for Handling Data in a Computer Environment (*Other name:* System and Methods for Network-Accelerated State Prefetching). PCT Application PCT/SE2022/051036. Download

[P10] Amir Roozbeh, Chakri Padala, **Alireza Farshin**, Dejan Kostić, Gerald Q. Maguire Jr. Processing Unit, Packet Handling Unit, Arrangement and Methods for Handling Packets (*Other name:* System and Methods for Probing and Polling Multiple I/O Operations on I/O Devcies for Priority-Based Packet Processing). PCT Application PCT/SE2022/050710. Download

[P11] Amir Roozbeh, **Alireza Farshin**, Marco Chiesa, Tom Barbette, Dejan Kostić. Packet Processing Including an Ingress Packet Part Distributor. PCT Application PCT/EP2023/063619. Download

[P12] Amir Roozbeh, **Alireza Farshin**, Dejan Kostić. System and Method for Organizing Physical Queues into Virtual Queues. PCT Application PCT/EP2022/051103. Download

[P13] Amir Roozbeh, **Alireza Farshin**, Marco Chiesa, Fabio Luciano Verdi. System and Method for Accurate Traffic Monitoring on Multi-Pipeline Switches. PCT Application PCT/EP2021/084572. Download

[P14] Amir Roozbeh, Chakri Padala, **Alireza Farshin**. System and Method for Cache pooling and Efficient Usage and I/O Transfer in disaggregated and Multi-Processor Architectures via Processor Interconnect. PCT Application PCT/SE2021/051016. Download

[P15] Amir Roozbeh, **Alireza Farshin**, Chakri Padala, Dejan Kostić, Gerald Q. Maguire Jr. System, Method, and Apparatus for Fine-grained Control of I/O Data Placement in Memory Subsystem. PCT Application PCT/SE2021/050803. Download

[P16] Amir Roozbeh, **Alireza Farshin**, Tom Barbette, Dejan Kostić, Gerald Q. Maguire Jr. Methods and Systems for Efficient Metadata and Data Delivery between a Network Interface and Applications. PCT Application PCT/IB2021/052976. Download

[P17] Amir Roozbeh, **Alireza Farshin**, Dejan Kostić, Gerald Q. Maguire Jr. Method and System for Efficient Input/Output Transfer in Network Devices. PCT Application PCT/SE2020/051107 (Download) & PCT/SE2020/051108 (Download).

[P18] Amir Roozbeh, **Alireza Farshin**, Dejan Kostić, Gerald Q. Maguire Jr, Hamid Ghasemirahni, Tom Barbette. Reordering and Reframing Packets. PCT Application PCT/IB2020/054991. Download

[P19] Chakri Padala, Amir Roozbeh, **Alireza Farshin**, Dejan Kostić, Gerald Q. Maguire Jr. Efficient Loading of Code Portions to a Cache. PCT Application PCT/SE2020/050527. Download

[P20] Amir Roozbeh, **Alireza Farshin**, Dejan Kostić, Gerald Q. Maguire Jr. Entities, System and Methods Performed Therein for Handling Memory Operations of an Application in a Computer Environment. PCT Application PCT/SE2019/050948. Download (US12111766B2 Granted)

[P21] Amir Roozbeh, **Alireza Farshin**, Dejan Kostić, Gerald Q. Maguire Jr. Methods and Devices for Controlling Memory Handling. PCT Application PCT/SE2020/050161. Download (US12111768B2 Granted)

[P22] Amir Roozbeh, Dejan Kostić, Gerald Q. Maguire Jr., **Alireza Farshin**. Memory Allocation in a Hierarchical Memory System. PCT Application PCT/SE2019/050596. Download (US12293227B2 Granted)

[P23] Amir Roozbeh, **Alireza Farshin**, Dejan Kostić, Gerald Q. Maguire Jr. Methods and Nodes for Handling Memory. PCT Application PCT/SE2018/051311. Download (US11714753B2 Granted)

Workshop Papers, Extended Abstracts, Preprints, Technical Reports, Demo, and Posters....

[W1] Adel Sefiane, **Alireza Farshin**, Marios Kogias. MLSynth: Towards Synthetic ML Traces In *The 2nd Workshop on Networks for AI Computing (**NAIC**)*. 2025. Download

[W2] Laura Puccioni, **Alireza Farshin**, Mariano Scazzariello, Changjie Wang, Marco Chiesa, Dejan Kostić. Deriving Coding-Specific Sub-Models from LLMs using Resource-Efficient Pruning In *The Second International Workshop on Large Language Models for Code (**LLM4Code**)*. 2025. Download

[W3] Hamid Ghasemirahni, **Alireza Farshin**, Dejan Kostić, Marco Chiesa. Just-in-Time Packet State Prefetching, *ArXiv Preprint*. 2024. Download

[W4] Hamid Ghasemirahni, **Alireza Farshin**, Mariano Scazzariello, Marco Chiesa, Dejan Kostić. Deploying Stateful Network Functions Efficiently using Large Language Models In *The Workshop on Machine Learning and Systems (**EuroMLSys**)*. 2024. Download

[W5] Changjie Wang, Mariano Scazzariello, **Alireza Farshin**, Dejan Kostić, Marco Chiesa. Making Network Configuration Human Friendly, *ArXiv Preprint*. 2023. Download

[W6] **Alireza Farshin**, Amir Roozbeh, Christian Schulte, Gerald Q. Maguire Jr., Dejan Kostić. Scheduling - A Secret Sauce For Resource Disaggregation, *Technical Report*. 2021. Download

[W7] **Alireza Farshin**, Tom Barbette, Amir Roozbeh, Gerald Q. Maguire Jr., Dejan Kostić. PacketMill: Toward per-core 100-Gbps Networking In *International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*. 2021. Download

[W8] **Alireza Farshin**, Amir Roozbeh, Gerald Q. Maguire Jr., Dejan Kostić. Optimizing Intel Data Direct I/O Technology for Multi-hundred-gigabit Networks In *The European Conference on Computer Systems (**EuroSys**)*. 2020. Download

[W9] **Alireza Farshin**, Amir Roozbeh, Gerald Q. Maguire Jr., Dejan Kostić. Make the Most out of Last Level Cache in Intel Processors In *The European Conference on Computer Systems (**EuroSys**)*. 2019. Download

Funded Projects

- **SEMLA - Securing Enterprises via Machine-Learning-based Automation**
  ~9.5 Million SEK *November 2023–October 2025*
  Funded by Vinnova - Cyber Security for Industrial Advanced Digitalization 2023
  Acted as one of the co-PIs from RISE AB
  other partners: KTH (led by Marco Chiesa), Saab AB, and RedHat AB
- **Realizing Low-Latency Internet Services via Low-Level Optimization of NFV Service Chains**
  $140,000 USD *August 2021–August 2023*
  Funded by Google PhD Fellowship in Systems and Networking

Open-Source Contributions

- **iommu-bench**: Understanding the IOTLB Wall for Multi-100-Gbps Linux-based Networking [Link]
- **DDC-RA**: A Constrained-based Scheduler for Disaggregated Data Centers (DDC) [Link]
- **PacketMill**: Toward per-core 100-Gbps Networking [Link]

ddio-bench: Understanding Intel Data Direct I/O Technology [Link]
Slice-aware Memory Management: Exploiting NUCA Characteristic of LLC in Intel Processors [Link]
CacheDirector: Sending Packets to the Right Slice by Exploiting Intel Last-Level Cache Addressing [Link]

Education

- **KTH Royal Institute of Technology** — **Stockholm, Sweden**
  Ph.D. in Information and Communication Technology, School of EECS — *August 2017–March 2023*
  Advisors: Prof. Dejan Kostić and Prof. Gerald Q. Maguire Jr.

  Dissertation Title: Realizing Low-Latency Packet Processing on Multi-Hundred-Gigabit-Per-Second Commodity Hardware (see my Dissertation)

  I also received my licentiate degree (Halfway to Ph.D.) in June 2019, see my Thesis.

- **Amirkabir University of Technology** — **Tehran, Iran**
  M.Sc. Electrical Engineering - Digital Electronic Circuits, Department of EE — *September 2015–July 2017*
  Advisor: Associate Prof. Saeed Sharifian

  Thesis: Resource Allocation in Software-Defined Networks for 5G Applications

  I used bio-inspired metaheuristic algorithms to perform resource allocation.

- **Sharif University of Technology** — **Tehran, Iran**
  B.Sc. Electrical Engineering - Electronics, EE Department — *September 2010–July 2015*
  Advisor: Associate Prof. Mehran Jahed

  Thesis: Design of Exoskeletal System for Wrist and Forearm

Honors, Awards, and Professional Services

2025: PC Member for eBPF'25.
2025: NetConfEval [C2] received IRTF/IETF Applied Networking Research Prize 2025.
2024: Packet Order Matters! [C3] was featured in the WIPO Green Technology Book.
2024: Giving a talk with Luigi Rizzo at Netdev 0x18 about IOTLB Wall.
2024: PC Member for eBPF'24.
2023: Faculty at Digital Futures.
2023: Reviewer for IEEE Computer Architecture Letters.
2022: Packet Order Matters! [C3] was featured in the Ericsson Blog and KTH.
2022: "Framtidens Forskning" has published a Swedish article on my research.
2022: PC Member for SIGCOMM'22 posters and demos program.
2022: Packet Order Matters! [C3] received the "Community Award" at NSDI'22.
2022: Giving a talk, Optimization Techniques for NFV, at Cisco Engineering Switzerland.
2021: Awarded Google PhD Fellowship 2021 in Systems and Networking. [Interview with KTH EECS]
2021: PacketMill [C4] was featured in the Ericsson Blog.
2021: Giving a talk with Tom Barbette at FOSDEM'21. [Watch]
2020: EuroSys'20 Shadow Program Committee.
2019: CacheDirector [C6] was featured in the Ericsson Blog, Tech Xplore, AlphaGalileo, and KTH.
2018: External Reviewer for NSDI'19.
2015: **Ranked 107th** among more than 20,000 participants in Iran's universities entrance exam for M.Sc.
2010: **Ranked 46th** among more than 460,000 participants in Iran's universities entrance exam for B.Sc.

Skills

Languages: English (Fluent), Persian (Native), Swedish (Novice)
Programming Languages: `C/C++`, `Python`, `MATLAB`, `Scala`, `R`, `Assembly-X86`, bash.
Tools & Libraries: DPDK, FastClick, Perf, LLVM, TensorFlow, Pandas, Spark, Gecode, Git, gnuplot, LaTeX.

Hobbies

Playing Piano and Bass Guitar, Jamming with Friends, Reading Books, Watching Movies and TV Series.