Muhammad Adnan

6335 Thunderbird Crescent, V6T 2G9, Vancouver, BC, Canada
☎ +1(236) 978-1115
✉ adnan@ece.ubc.ca
☞http://people.ece.ubc.ca/adnan/

EDUCATION

Ph.D., Electrical and Computer Engineering Sep 2026 (Expected)
University of British Columbia, Vancouver, BC, Canada
Research Area: Cross Stack Optimizations for large Machine Learning Models
Advisor: Prashant J.Nair

M.A.Sc., Electrical and Computer Engineering Nov 2021
University of British Columbia, Vancouver, BC, Canada
Thesis: Accelerating input dispatching for deep learning recommendation models training

B.E., Electrical Engineering July 2013
National University of Sciences and Technology, Islamabad, Pakistan

RESEARCH

Artificial intelligence and machine learning is an essential part of any technology - from self driving cars to E-commerce - any technology is using AI and ML in some way. Currently most of the machine learning jobs are executed in data centers using state of art CPU's and GPU's in a distributed fashion with software frameworks not optimized for such workloads.

Vision

The landscape of deep learning models has undergone significant changes in the last few years, characterized by a steady increase in the size of these models. Leading the way in this trend are recommendation and large language models, which have become increasingly prevalent and influential in the field of artificial intelligence. The emergence of such large models has brought with it a host of system-level challenges, particularly in relation to their training and inference. Training these models across distributed computational resources remains an open problem, while the deployment of large models for inference poses its own set of challenges. My research interest centers at the intersection of architecture and systems, with a particular focus on addressing the challenges posed by large deep-learning models.

Interests

Systems for ML, High Performance Computing, Large Language Models (LLMs), Recommender Systems, Ranking, Hardware Accelerators, Machine Learning.

CONFERENCE Publications

1. “Foresight: Adaptive Layer Reuse for Accelerated and High-Quality Text-to-Video Generation”, *39th Annual Conference on Neural Information Processing Systems* **(NeurIPS’25)**
   Muhammad Adnan, Nithesh Kurella, Akhil Arunkumar, Prashant J.Nair

2. “Heterogeneous Acceleration Pipeline for Recommendation System Training”, *51st International Symposium on Computer Architecture* **(ISCA’24)**
   Muhammad Adnan, Yassaman E.Maboud, Divya Mahajan, Prashant J.Nair

3. “Keyformer: KV Cache reduction through key tokens selection for Efficient Generative Inference”, *7th Annual Conference on Machine Learning and Systems* **(MLSys’24)**
   Muhammad Adnan, Akhil Arunkumar, Gaurav Jain, Prashant J.Nair, Ilya Soloveychik, Purushotham Kamath

4. “Accelerating Recommendation System Training by Leveraging Popular Choices”, *48th International Conference on Very Large Data Bases* **(VLDB’22)**
   Muhammad Adnan, Yassaman E.Maboud, Divya Mahajan, Prashant J.Nair

5. “Slipstream: Semantic-Based Training Acceleration for Recommendation Models”, *27th Design, Automation and Test in Europe Conference* **(DATE’25)**
   Yassaman E.Maboud, Muhammad Adnan, Divya Mahajan, Prashant J.Nair

In Submission

1. HiDE: Hierarchical Pruning and Design Space Exploration for Accelerating Distributed Training”, *In submission at* **(HPCA’26)**
   Muhammad Adnan, Irene Wang, Prashant J.Nair, Divya Mahajan

Workshops

1. “Ad-Rec: Advanced Feature Interactions to Address Covariate-Shifts in Recommendation Networks”, *presented at ML for Systems Workshop at Thirty-Sixth Conference on Neural Information Processing Systems* **(NeurIPS’23)**
   Muhammad Adnan, Yassaman E.Maboud, Divya Mahajan, Prashant J.Nair
2. “Accelerating Recommendation System Training by Leveraging Popular Choices”, *presented at Personalized Recommendation Systems and Algorithms Workshops at Fourth Conference on Machine Learning and Systems* **(MLSys’21)**
   Muhammad Adnan, Yassaman E.Maboud, Divya Mahajan, Prashant J.Nair

PATENTS AND THESIS

1. “Workload-Aware Hardware Architecture Recommendations”
   USSN # 17/965681
   Inventors: Amar Phanishayee, Divya Mahajan, Janardhan Kulkarni, Miguel Castro, Muhammad Adnan,

2. “Accelerating input dispatching for deep learning recommendation models training”
   M.A.Sc. Thesis at University of British Columbia
   Advisor: Prashant J.Nair

TALKS

- “Recommendation Models training challenges across Distributed Systems” lecture at **GaTech 2024**.
- “Training Big Sparse Recommendation Models on Commodity Servers” tutorial at **IISWC 2024** and **HPCA 2023**.
- “Life in Grad school panelist” at Undergrad Architecture Mentoring (uArch) Workshop in conjunction with **ISCA 2024** and **ISCA 2021**.
- “Real Time Systems Workshop: Using Real Time Operating System”, at Bahrain Polytechnic in Jan. 2016

SERVICE

- PC Member, 32nd IEEE International Symposium on High-Performance Computer Architecture (HPCA'26).
- PC Member, 2025 IEEE International Symposium on Workload Characterization (IISWC'25).
- PC Member, Machine Learning for Systems Workshop at NeurIPS 2024 & 2025.
- PC Member, Energy Efficient Machine Learning and Cognitive Computing (EMC^2) workshop colocated with ASPLOS 2024.
- Reviewer, SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'24).
- PhD Student Reviewer, 55th IEEE/ACM International Symposium on Microarchitecture (MICRO'22).
- Reviewer, IEEE Computer Architecture Letters (IEEE CAL).
- Reviewer, IEEE Transactions on Computers (IEEE TC).

ACADEMIC EXPERIENCE

- **Graduate Research Assistant** 2019-Present
  Systems and Architectures (STAR) Lab.
- **Graduate Teaching Assistant** Fall-2019, Fall-2020
  Introduction to Computer Architecture: Undergraduate-Level
- **Graduate Teaching Assistant** Fall-2020, Summer-2020, Fall 2021
  Digital Systems Design: Undergraduate-Level
- **Graduate Teaching Assistant** Fall-2022
  Introduction to Microcomputers: Undergraduate-Level
- **Graduate Teaching Assistant** Winter-2023
  Digital Systems and Microcomputers : Undergraduate-Level

INDUSTRIAL EXPERIENCE

- **Microsoft** June 2025 - Sep 2025
  Research Intern (Azure Research - Systems)
  Inverstigating Efficent LLM Serving from KV Cachce Management.
  Mentors: Pantea Zardoshti, Esha Chouksi, Rodrigo Fonseca
- **d-Matrix** May 2023 - Dec 2023, Sep 2024 - Dec 2024
  ML Performance Intern
  Investigating performance of generative language models and video generation models for inference.
  Mentors: Akhil Arunkumar, Nithesh Kurella
- **Microsoft Research** June 2021 - Sep 2021
  Research Intern (ML Systems)
  Investigated design space exploration of domain specific training accelerators for DNN workloads.
  Mentor: Amar Phanishayee, Divya Mahajan
- **National Instruments (NI)** 2013 - 2019
  Account Manager
  Responsible for providing technical support to the customers in Middle East, preparing demo's for Sales & Marketing and managing the team of applications engineers in Pakistan.

PROJECTS
Graduate

- **Optimizing memory hierarchy for Deep Neural Networks**
  A Real Time Machine Learning project (RTML).
- **Optimizing the search algorithm for finding efficient mappings for DNN accelerators.**
- **Avoiding Cache Pollution from Mis-speculated Loads for efficient Cache Management.**
  Course project for Advanced Computer Architecture.
- **MAC-ECC: An Approach for an Optimized Memory Reliability**
  Course project for Security and Reliability.

HONORS AND AWARDS

- Recipient of prestigious **NSERC Canada Graduate Scholarship - Doctoral (CGS D)** award.
- Selected as **Machine Learning and Systems Rising Star** in the 2023 cohort by **MLCommons**.
- Recipient of **Graduate Support Initiative (GSI) - 2023-2024** award.
- Recipient of **ISCA Travel Grant** for attending ISCA 2024 conference.
- Recipient of **VLDB Endowment Travel SPEND Award** for attending VLDB 2022 conference.
- Nominated for **Rookie of the year at EMEIA level** at NI Week 2017.
- **Certified LabVIEW Developer (CLD)**.
- Nominated for **President's gold medal** for best senior year project.
- Received **NUST merit based scholarship** for 4 years for academic achievements during undergraduate.
- Received **Commandant's plaque of Excellence Award** being high achiever.

LEADERSHIP

- President of **Pakistani Students Association in Canada** for helping present and future Pakistani students studying in Canada.
- Lead for **Planet NI STEM Program** for promoting STEM Education in Pakistan
- Event lead for **EME Olympiad 2011 & 2012**

SKILLS

C, C++, Python, Pytorch, Bash scripting, Cadence Virtuoso, Architecture Simulators, Deep Neural Network Performance Simulators

PROFESSIONAL MEMBERSHIP

ACM

REFEREES

Prof. Prashant J.Nair
University of British Columbia

KAIS 4014
2332 Main Mall
Vancouver, BC, Canada
✉ prashantnair@ece.ubc.ca

Prof. Divya Mahajan
Georgia Insitute of Technology

North Avenue
Atlanta,
GA 30332, USA
✉ divya.mahajan@gatech.edu