

Your Name

Current department ♦ current organization

✉ youremail@gmail.com  <https://webpage.io/>

in <https://www.linkedin.com/in/your profile/>  (xxx) xxx-xxxx

EDUCATION

202x	PhD	University name, city, country Computer Science and Engineering	GPA
201x	MS	University name, city, country Computer Science and Engineering	GPA
201x	BS	University name, city, country department	GPA

PROFESSIONAL EXPERIENCE

2022-Present **Applied Scientist II**

AWS Generative AI Innovation Center | NY, NY

- As the Technical Lead for developing Generative AI-based solutions for AWS customers, I was responsible for overseeing a diverse portfolio of projects. My key duties included identifying the project scope, defining the critical business metrics, and outlining the deliverables for each engagement. I also served as the primary communicator, liaising with key stakeholders such as Amazon colleagues and external clients to ensure alignment and successful project completion.
- Leading a team of highly skilled scientists, I successfully guided the development of cost-effective and resource-efficient Generative AI solutions. Over the course of 2023-2024, I completed more than 20 such projects, leveraging cutting-edge technologies like Anthropic Cohere models (v3.1), Amazon Titan (v2.0), and LangChain (v0.12). These engagements included a Retrieval Augmented Generation (RAG) model for a proprietary customer data-based Q&A virtual assistant for a sports and entertainment client, a Q&A pipeline using Text-to-SQL for data analysis and personalized memory-enabled chatbots for a financial use case, as well as a Q&A virtual assistant for a manufacturing client.
- Most notably, my recent project resulted in a \$1.8 million statement of work, which is estimated to generate over \$1.2 million in recurring annual revenue for the company. This achievement underscores my ability to deliver high-impact, cost-effective Generative AI solutions that drive tangible business results for AWS customers.
- I have also been actively engaged in knowledge-sharing and thought leadership initiatives. I have authored 4 technical blog posts and 6 research papers, which have been well-received within the industry. I have also delivered more than 10 speaking engagements at both internal Amazon events and external conferences, further cementing my reputation as a subject matter expert in the field of Generative AI.

2022

Applied Scientist I

AWS Machine Learning and Solutions Lab | NY, NY

- As an Applied Scientist, I designed and developed customized machine learning solutions for AWS customers across diverse industries, including Energy, Semiconductor Manufacturing, and Pharmaceuticals. A key aspect of my role was

to deeply understand the most critical and high-impact business problems faced by these customers, and then provide tailored solutions and comprehensive roadmaps to address them.

- I leveraged my expertise in machine learning and data science to implement end-to-end ML pipelines on the AWS ecosystem. This involved data cleaning, feature engineering, model development, training, and deployment for model inference. Through this systematic approach, I was able to deliver robust and scalable ML solutions that drove tangible business outcomes for my clients.

Some of the key projects I spearheaded include:

- Sales Forecast Model: Developed a highly accurate sales forecasting model for one of the largest semiconductor companies, enabling them to make more informed strategic decisions and optimize their supply chain.
- Computer Vision for Animal Behavior Analysis: Designed a computer vision-based solution to analyze animal behavior patterns, which helped a leading pharmaceutical company accelerate their drug discovery process.
- Anomaly Detection for Power Utilities: Implemented a computer vision-based anomaly detection system for a major power utility company, allowing them to proactively identify and address issues in their infrastructure, leading to improved operational efficiency and reliability.
- Time Series Anomaly Detection: Delivered a robust time series anomaly detection solution for an online energy trading company, empowering them to swiftly identify and mitigate market fluctuations, ultimately optimizing their trading strategies and profitability.

2021-2022

Postdoctoral Appointee

Data Science and Learning Division | Argonne National Laboratory | Lemont, Illinois

- Lead researcher for developing rapid ML training systems for high data generating sources, including APS-U at Argonne National Lab and LCLS-II at Stanford Linear Accelerator Center (SLAC).
- Responsible for collaborating with the domain scientist to gather information regarding the ML training pipeline requirements for robust model training.
- Used PyTorch and TensorFlow to train ML models on state-of-the-art HPC systems, including the Argonne Leadership Computing Facility (ALCF).

2020

Research Scientist Intern

Nokia Bell Labs | Stuttgart, Germany

- Lead developer of scalable Machine Learning framework using serverless computing.
- Responsible for designing, analyzing, and debugging key modules of the systems for optimal performance in terms of monetary cost and execution time.
- As a lead developer, I also performed unit testing, integration testing, and function testing with the implementation of new modules in the system.

2017-2021

Graduate Research/Teaching Assistant

Computer Science and Engineering Department | University of Nevada, Reno | Reno, Nevada

Supervisor: Feng Yan, PhD

- Responsible for performing research for machine learning systems, efficient resource

- utilization, cloud computing, and high-performance computing.
- Teaching Assistant for the following courses:
 - CPE 301 Embedded Systems Lab
 - CS 219 Computer Organization and Architecture

2014-2016

Graduate Research/Teaching Assistant

Computer Science and Engineering | Koç University | Istanbul, Turkey

- Researched energy efficient scheduling and workload management in distributed data centers.
- Teaching Assistant for the following courses:
 - COMP 132 Advanced Programming
 - COMP 303 Computer Architecture
 - EQUR 101 Quantitative Reasoning Using Computers
 - COMP 130 Introduction to Programming

2013-2014

IT Consultant

IBM Global Business Services | Islamabad, Pakistan

- Responsible for developing new modules using Java for a web-based application.
- Performed requirement gathering, end-to-end stakeholder management, client communication, writing summary reports, resource management, and task allocation.
- Responsible for unit testing and system testing of the new modules and reports based on the specifications.

THESES

2021

Toward Autonomous and Efficient Machine Learning Systems

Computer Science and Engineering Department | University of Nevada, Reno | Reno, Nevada

Dissertation Advisor: Dr. Feng Yan

2016

Workload Management in Distributed Data Centers: Spatial and Thermal Awareness for Energy Efficiency

Department of Computer Science and Engineering | Koç University | Istanbul, Turkey

Thesis Advisor: Dr. Oznur Ozkasap

2012

IPTV Services on IP based Multimedia Subsystem

Department of Telecommunication Engineering | FAST National University of Computer and Emerging Sciences | Islamabad, Pakistan

Thesis Advisor: Mr. Aamer Hafeez

PUBLICATIONS

2024

[PGA virtual assistant featured in AWS CEO's LinkedIn page](#)

[PGA TOUR virtual assistant featured by AWS LinkedIn page](#)

[PGA TOUR virtual assistant featured on AWS AI/ML vice president's LinkedIn Page](#)

[PGA TOUR virtual assistant featured on AWS Generative AI Innovation Center's Director's LinkedIn Page](#)

2023

[PGA TOUR Virtual Assistant Featured in Sports and Business Journal](#) [Tech Lead]

Ahsan Ali, Oznur Ozkasap. *Spatial and Thermal Aware Methods for Efficient Workload Management in Distributed Data Centers*. Future Generation of Computer Systems (**FGCS**).

Improving Asset Health and Grid Resilience using Machine Learning. **AWS AI/ML Blogs**.

Anomaly Detection in Power Transmission Lines Using Vision Transformer Foundation Models. Amazon Machine Learning Conference (**AMLC**).

Deep Variational Approach for Markov Processes to Sequence Animal Behavior for Drug Discovery. Amazon Machine Learning Conference (**AMLC**).

Predicting New and Existing Product Sales in Semiconductors Using Amazon Forecast. **AWS AI/ML Blogs**.

2022

Ahsan Ali, Riccardo Pincirolì, Feng Yan, Evgenia Smirni. *Optimizing Inference Serving on Serverless Platforms*. Very large Databases (**VLDB**) Conference.

Ahsan Ali, Peter Kenesei, Antonino Miceli, Hemant Sharma, Dennis Trujillo, Ryan Coffee, Jana Thayer, Ian Foster, Zhengchun Liu. *fairDMS: Rapid Model Training by Data and Model Reuse*. High-Performance Parallel and Distributed Computing (**HPDC 2022**).

Ahsan Ali, Syed Zawad, Istemi Ekin Akkus, Paarijaat Aditiya, Akin, Ruichuan Chen, Feng Yan. *Enabling Scalable and Adaptive Machine Learning Workflows via Serverless Computing*. High Performance Parallel and Distributed Computing (**HPDC 2022**).

2021

Zhengchun Liu, **Ahsan Ali**, Peter Kenesei, Antonino Miceli, Hemant Sharma, Nicholas Schwarz, Dennis Trujillo, Hyunseung Yoo, Ryan Coffee, Naoufal Layad, Jana Thayer, Ryan Herbst, Chun Hong Yoon, Ian Foster. *Bridge Data Center AI Systems with Edge Computing for Actionable Information Retrieval*. 3rd Workshop on Extreme-Scale Experiment-in-the-Loop Computing (**XLOOP**). Super Computing (**SC 2021**). Virtual. Best paper award.

Syed Zawad, **Ahsan Ali**, Pin-Yu Chen, Ali Anwar, Yi Zhou, Nathalie Baracaldo, Yuan Tian, and Feng Yan. *Curse or Redemption? How Data Heterogeneity Affects the Robustness of Federated Learning*. Proceedings of the AAAI Conference on Artificial Intelligence (**AAAI 2021**). Virtual. (Acceptance rate: 1692/7911 = 21%).

2020

Riccardo Pincirolì, **Ahsan Ali**, Feng Yan, Evgenia Smirni. *CEDULE+: Resource Management*

for Burstable Cloud Instances Using Predictive Analytics. IEEE Transactions on Network and Service Management (**TNSM 2020**). (Impact factor: 4.682).

Ahsan Ali*, Riccardo Pincioli*, Feng Yan, Evgenia Smirni. *BATCH: Machine Learning Inference Serving on Serverless Platform with Adaptive Batching*. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (**SC2020**). Atlanta, GA. (*Equal Contribution and Co-First Authors; Acceptance rate: 85/380 = 22%).

Zheng Chai*, **Ahsan Ali***, Syed Zawad*, Stacey Truex, Ali Anwar, Nathalie Baracaldo, Yi Zhou, Heiko Ludwig, Feng Yan, Yue Cheng. *TiFL: A Tier-based Federated Learning System*. Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing (**HPDC 2020**). Stockholm, Sweden. (*Equal Contribution and Co First Authors; Acceptance rate: 16/71 = 22%).

2019 **Ahsan Ali**, Riccardo Pincioli, Feng Yan, Evgenia Smirni. *It's Not a Sprint; It's a Marathon: Stretching Multi-resource Burstable Performance in Public Clouds*. ACM/IFIP/USENIX Middleware 2019 (**Middleware 2019**). Trento, Italy. (Acceptance rate: 32%).

2018 **Ahsan Ali**, Riccardo Pincioli, Feng Yan, Evgenia Smirni. *CEDULE: A Scheduling Framework for Burstable Performance in Cloud Computing*. IEEE International Conference on Autonomic Computing (**ICAC 2018**). Trento, Italy. (Acceptance rate: 15/49=30%).

2016 **Ahsan Ali** and Oznur Ozkasap. *Workload Management in Distributed Data Centers: Thermal and Spatial Awareness*. IEEE International Conference on Smart Cloud (**Smart Cloud 2016**). NY, NY.

Ahsan Ali and Oznur Ozkasap. *Price/Cooling Aware and Delay Sensitive Scheduling in Geographically Distributed Data Centers*. IEEE/IFIP International Workshop on Analytics for Network and Service Management (**AnNet 2016**). Istanbul, Turkey.

2015 **Ahsan Ali** and Oznur Ozkasap. *Environment Friendly Energy Efficient Distributed Data Centers*. 30th International Symposium on Computer and Information Sciences, Springer International Publishing, 69-79. London, UK.

SELECTED HONORS and AWARDS

2021 **Social Liaison Officer**
Postdoctoral Society of Argonne (PSA) | Argonne National Laboratory

2021 **Technical Program Committee Co-Chair**
Postdoctoral Research and Career Symposium | Argonne National Laboratory

- 2019 **Middleware Conference Travel Grant**
ACM/IFIP Middleware Conference
Presented: It's Not a Sprint; It's a Marathon: Stretching Multi-resource Burstable Performance in Public Clouds (US \$1,500)
- 2018 **Graduate Student Association Travel Award**
University of Nevada, Reno for IEEE International Conference on Autonomic Computing (ICAC 2018)
Presented: CEDULE: A Scheduling Framework for Burstable Performance in Cloud Computing (US \$500)
- 2016 **Koç University Student Travel Award**
Koç University | Istanbul, Turkey for IEEE International Conference on Smart Cloud (Smart Cloud 2016)
Presented: Workload Management in Distributed Data Centers: Thermal and Spatial Awareness (US \$1,500)
- 2014-2016 **Higher Education Commission Pakistan Student Fellowship Government of Pakistan**
Fellowship covered monthly stipend (US \$900), accommodation, and tuition fees
- 2010-2011 **Vice President**
FAST National University of Computer and Emerging Sciences Engineering Society