Tokey Tahmid

Address: Tennessee 37777, USA
Contact: +1(865)438-2188
Email: ttahmid@icl.utk.edu
Website: tokey-tahmid.github.io

SUMMARY

A highly motivated HPC-AI Performance Engineer with a Master's (Level 9) degree in Computer Science, and 4 years of experience in Artificial Intelligence (AI), High-performance Computing (HPC), and Performance Engineering. As a research associate at the Innovative Computing Laboratory (ICL) (University of Tennessee Knoxville), I develop performance API (PAPI) support for specialised AI hardware and software, extend performance optimisation for HPC and AI applications, and disseminate research outcomes through reports, publications, and presentations.

SKILLS

Programming

- Experienced in software development and advanced programming in C, C++, Python, CUDA, and Fortran – in Linux distributions, macOS, and Windows operating systems
- Experienced in parallel programming models: MPI, OpenMP, OpenACC
- Proficient in other programming languages: C#, R, Assembly

HPC-AI Research Expertise

- Robust knowledge of HPC – processor and computer architectures, including CPU, GPU, shared/distributed memory systems, specialised AI architectures (Intel Gaudi, Cerebras WSE, Sambanova RDU), and neuromorphic processors
- Experienced in developing and efficiently executing HPC and AI applications and workflows on HPC clusters and cloud infrastructures (Chameleon, AWS, Azure)
- Experienced with methods and tools for HPC and AI application performance analysis, debugging, optimisation – PAPI, perf, TAU, Score-P, HPCToolkit, VampirTrace, NVIDIA Nsight, Intel Gaudi Profiler, CodeCarbon, PyTorch Profiler, and TensorBoard
- Experienced with scientific software engineering and development practices, including build systems (CMake, GNU Make / Autotools, Meson, Spack); revision control (Git, Git LFS, GitHub, GitLab, Bitbucket); CI/CD pipelines (GitHub Actions, GitLab CI/CD)
- Knowledgeable and experienced in machine learning, deep learning algorithms, reinforcement learning, large language models, generative AI, distributed computing, neuromorphic computing, and spiking neural networks; methods such as support vector machines, decision trees, artificial neural networks, mixed-precision training
- Knowledgeable and experienced in using frameworks such as TensorFlow, PyTorch, Horovod, OpenMPI, MPICH, netCDF, PnetCDF, yaml-cpp, Pandas, NumPy, scikit-learn, and SciPy
- Hands-on experience in using Virtual Machines (VM), containers, workload managers, and workflow orchestration tools – Docker, Singularity, Kubernetes, Slurm, Dask, Apache Spark

- Experienced in benchmarking and doing performance analysis on scientific AI/ML algorithms using HPC techniques and distributed computing on GPU

Technical Research Skills

- Effective working in Agile project management frameworks, generating original research agendas, writing research grant proposals, and solving research problems by implementing creative solutions with analytical thinking
- Experienced in manipulating and analysing complex large datasets for scientific computing and AI research
- Demonstrated experience in writing technical reports, technical documentation, research papers, posters, and presentations
- Excellent communication skills, teamwork, and collaboration skills, with experience working in multidisciplinary research teams

WORK EXPERIENCE

Innovative Computing Laboratory (ICL) | Research Associate | Feb 2025 – Present

- Working on the MINCER project in the Performance Measurement & Modelling group at ICL with Dr. Heike Jagode as the PI
- Developing Performance Application Programming Interface (PAPI) software support for ever-growing AI chips and accelerators such as the Intel Habana Gaudi, Cerebras WSE, and SambaNova RDU
- Research responsibilities include research and maintenance work on PAPI, writing research papers, posters, and proposals
- Work-in-progress paper, "PAPI Support for Specialised AI Architectures", accepted for presentation at **SC25 (PDSW'25 Workshop)**

National Renewable Energy Laboratory (NREL) | Graduate (Summer) Intern | May 2024 – Aug 2024

- Worked as an intern for the "Low Precision and Efficient Programming Languages for Sustainable AI" role with Dr. Weslley Da Silva Pereira
- Demonstrated excellent results (average **speed-up of 2.05×** and **80.75% better energy efficiency**) with mixed-precision training on multiple AI applications at NREL (GitHub)
- Published a position paper titled, "Low Precision for Lower Energy Consumption" at the **2024 ASCR Workshop** on Energy-Efficient Computing for Science (Publication)
- Published the final report titled, "Low Precision and Efficient Programming Languages for Sustainable AI: Final Report for the Summer Project of 2024" at NREL (Publication)

Innovative Computing Laboratory (ICL) | Graduate Research Assistant | Jan 2023 – Jan 2025

- Worked with the Linear Algebra group at ICL and closely with Dr. Piotr Luszczek on multiple HPC-AI projects
- Developed a Benchmarking Infrastructure for FAIR (Findable, Accessible, Interoperable, and Reusable) asset tracking and dataset management for deploying and evaluating scientific ML/AI surrogate models (Currently **benchmarked 5 models with GPU support** and their datasets totaling up to approximately **7.1 TB**) (GitHub)

- Published a paper titled, “Towards the FAIR Asset Tracking across Models, Datasets, and Performance Evaluation Scenarios“ at the 2023 IEEE High Performance Extreme Computing Conference (**HPEC23**) (Publication)

TENNLab - Neuromorphic Architectures, Learning, Applications | Graduate Research Assistant | May 2023 – Jan 2025

- Worked on Neuromorphic applications and Spiking Neural Networks (SNN) with Dr. Catherine Schuman
- Developed SpikeRL, a scalable and energy-efficient infrastructure (**4.26× faster and 2.25× more energy efficient** than state-of-the-art) for deep reinforcement learning (DRL) based spiking neural networks (SNN) with MPI for distributed training and mixed precision for optimisation (GitHub)
- Published and presented a paper titled, “SpikeRL: A Scalable and Energy-efficient Framework for Deep Spiking Reinforcement Learning” at the 2025 International Conference on Neuromorphic Systems (**ICONS 2025**) (Publication)
- Published and presented a paper titled, “Towards Scalable and Efficient Spiking Reinforcement Learning for Continuous Control Tasks” at the 2024 International Conference on Neuromorphic Systems (**ICONS 2024**) (Publication)

Chowkosh Limited | Research Assistant | Sep 2021 – Mar 2022

- Developed a Deep Learning based Android Malware Detection model in Android applications
- Tested machine learning models: Random Forest (RF), Logistic Regression, and Support Vector Machine (SVM), as well as deep learning models: BERT and MLP for malware detection
- Evaluated the performance difference between ML and DL classifiers in **detecting zero-day attacks** and compared the results with state-of-the-art methods

BRAC UNIVERSITY | Undergraduate Thesis Research | May 2020 – October 2021

- Developed a model that generates real-time character animation for biped locomotion in Unity ML agents using Reinforcement learning (RL) and Imitation learning algorithms
- The novel approach of combining RL and IL provided a better and easy-to-implement solution to character animation than the state-of-the-art methods
- Published and presented a paper titled, “Character Animation Using Reinforcement Learning and Imitation Learning Algorithms” at the 2021 Joint 10th International Conference on Informatics, Electronics & Vision (**ICIEV**) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (**icIVPR**). (Publication)

Cye Retail Tech Ltd | Full Stack Software Developer Intern | May 2022 - Jul 2022

- Acquired experience in developing frontend and backend web and Android applications

EDUCATION

Master of Science in Computer Science | University of Tennessee, Knoxville | Jan 2023 - Dec 2024
CGPA: 3.66 out of 4.0
Thesis: Energy-Efficient Computing for Scalable and Sustainable AI
Bachelor of Science in Computer Science | BRAC University | Jan 2017 - Dec 2021
CGPA: 3.27 out of 4.0

Thesis: Character Animation Using Reinforcement Learning and Imitation Learning Algorithms

PUBLICATIONS

[1] PAPI Support for Specialised AI Architectures. SC25 (PDSW'25 Workshop).
[2] SpikeRL: A Scalable and Energy-efficient Framework for Deep Spiking Reinforcement Learning. ICONS2025.
[3] Energy-Efficient Computing for Scalable and Sustainable AI. University of Tennessee 2024.
[4] Towards Scalable and Efficient Spiking Reinforcement Learning for Continuous Control Tasks. ICONS2024.
[5] Low Precision for Lower Energy Consumption. ASCR Energy Efficient Workshop 2024.
[6] Low Precision and Efficient Programming Languages for Sustainable AI: Final Report for the Summer Project of 2024. National Renewable Energy Laboratory 2024.
[7] Towards the FAIR Asset Tracking Across Models, Datasets, and Performance Evaluation Scenarios. HPEC2023.
[8] Character animation using reinforcement learning and imitation learning algorithms. ICIEV and icIVPR 2021.

PROJECTS

LLM Chatbot | University of Tennessee, Knoxville | Jan 2024 – May 2024

- Developed a Chatbot for International Students using Google's Gemini API (GitHub)

Operating System Development | University of Tennessee, Knoxville | Aug 2023 – Dec 2023

- Built a fully functional Operating System from scratch in the COSC562 - OS Design/Implementation course (GitHub)

Jailbreak GPT Project | University of Tennessee, Knoxville | Aug 2023 – Dec 2023

- Analysed Jailbreak-llms dataset containing adversarial prompts for Large Language Models (LLMs) to identify the unethical use of LLMs as the COSC545 - Fundamentals of Digital Archeology course final project (GitHub)

LIBRA - A Subscription-Based Online Gaming Platform | BRAC University | Sep 2020 – Dec 2020

- Developed a software interface for a subscription-based online gaming platform for the final project of the CS471 - Software Engineering course (GitHub)