

Neil McGlohon, Ph.D.

CONTACT	neil@mcglo.dev	nmcglo.com github.com/nmcglo
EDUCATION	Ph.D., Computer Science Rensselaer Polytechnic Institute, Troy, New York USA Research Focus: High Performance/Parallel Computing	July 2021
	M.S., Computer Science Rensselaer Polytechnic Institute, Troy, New York USA Research Focus: Distributed Computing, Machine Learning	May 2016
	Bachelor of Science in Physics University of Oklahoma, Norman, Oklahoma USA Minor: Computer Science	May 2014
RELEVANT SKILLS	• Self-motivated, fast-learner, quick to ramp-up on new topics • Strong and flexible critical-thinking, problem solving, and analysis skills • Expertise in parallel, distributed, and clustered computing • C/C++, Python, Javascript/NodeJS, MPI, Git, SLURM, Cloud-native architectures, AI inference, HTTP API design, AMQP, vLLM, Tensorflow, PyTorch, Kubernetes, OpenShift, Docker and Podman containerization, AWS, IBM Cloud, L^AT_EX	
FEATURED PROFESSIONAL EXPERIENCE	• Led the design and development of a purpose-built LLM inference engine to provide generative AI inference workloads powered by IBM AIU NorthPole hardware to define the state-of-the-art. • Demonstrated rack-scale distributed AI system deployments of 288 NorthPole cards, capable of collectively producing over 500,000 generated tokens-per-second whilst simultaneously maintaining model-optimal individual per-token latencies as low as 1 millisecond. • Designed, developed, and deployed a cloud-native stack on OpenShift/Kubernetes to expose and control access to IBM AIU NorthPole powered LLM AI inference appliances. Worked with engineers to integrate the services into the IBM watsonx product to enable IBM AIU NorthPole-accelerated LLM inference workloads across the platform. • Acted as primary maintainer of open-source C/C++ project for massively parallel simulation of high performance computing (HPC) communication networks known as CODES. Collaborated with researchers from institutions including Argonne and Sandia National Laboratories. • Participated as a steering-committee member on an NSF funded project-of-projects and acted as a subject-matter-expert and mentor in the areas of high performance computing and AI systems for participating graduate students. • Author on 18 papers published in various conferences and journals in areas of AI/ML, neuromorphic computing, high performance computing, and parallel discrete event simulation.	
EMPLOYMENT EXPERIENCE	– IBM, IBM Research Almaden Research Center, San Jose, California USA <i>Staff Research Scientist</i> October 2022 – Present - Team lead role driving and delegating tasks for the the development of the IBM AIU NorthPole accelerator card’s runtime software stack and SDK packaging/release process. - Subject matter expert in parallel and distributed computing systems, containerization and orchestration, and the serving of LLM inference systems. - Technical representative for IBM AIU NorthPole during business development events and client outreach meetings as well as presentation showcases. - Facilitated knowledge transfer, collaboration, and integration between teams across domains in both scientific and business-focused areas.	

- **Rensselaer Polytechnic Institute**, Center for Computational Innovations
Troy, New York USA

Research Scientist**July 2021 – September 2022**

- Contributor to complex scientific efforts at the Center for Computational Innovations, focusing on software development and benchmark optimization in supercomputing environments.
- Identified and pursued funding opportunities to support innovative research projects in collaboration with the Center Director.
- External collaborator on research projects for the IBM AI Hardware Center.
- Provided technical expertise to enhance design, implementation, and analysis of research.
- Triaged and prioritized center needs for research facilitation of thousands of users across multiple supercomputer systems.

Software Engineer**February 2020 – July 2021**

- Development of research software for simulating high performance computing networks.
- Collaborated with external research partners to create purpose-built software solutions.

- **Cisco Meraki**

San Francisco, California USA

Software Engineering Intern**May – August 2017**

- Research and development of firmware for Meraki's MR Platform access points.
- Designed and developed a lockless thread-safe hashtable utilizing RCU techniques. Contributions submitted to open-source software-defined router project: Click.

SCIENTIFIC
SERVICE

- NSF CAREERS Cyberteam Program
Steering Committee Member (2020–2022)
- ACM SIGSIM Principles of Advanced Discrete Simulation (PADS)
Organizing Committee, Ph.D. Colloquium Chair (2022)
Organizing Committee, Publicity Chair (2023)
Program Committee Member (2022–2025)
- ACM SIGPLAN Benchmarking in the Data Center: Expanding to the Cloud (BID)
Program Committee Member (2022)
- IEEE IPDPS Scalable Deep Learning over Parallel and Distributed Infrastructure (ScaDL)
Program Committee Member (2022)
- Peer-reviewer for 6 ACM/IEEE computing conferences and journals. (2019–Present)

ACADEMIC
EXPERIENCE

- **Rensselaer Polytechnic Institute**, Troy, New York USA

Graduate Research Assistant**August 2014 – July 2021**

M.S. and Ph.D. research, graduate level coursework and projects. Areas of research: high performance and parallel computing, distributed systems, cloud computing, machine learning, neuro-morphic computing and simulation.

Teaching Assistant: Computer Science I**August – December 2014**

Support to professor during course teaching fundamentals of computer science using Python. Facilitated laboratory classes and discussions, hosted office hours, graded homework and exams. Worked in group of eight teaching assistants.

- **University of Oklahoma**, Norman, Oklahoma USA

Undergraduate Research Assistant**May 2012 – May 2014**

Explored electrical transport properties of antimony measured at cryogenic temperatures. Created and refined laboratory procedures for measuring differential conductance of a material – allowing for a deeper understanding of the interface between a topological insulator and a superconductor. Advisor: Dr. Sheena Murphy.

National Science Foundation REU**May – August 2012, 2013**

Participated in condensed matter research and analysis under an NSF Materials Research Science and Engineering Center (MRSEC) grant as part of an NSF Research Experience for Undergraduates program.

PUBLICATIONS

- M. DeBole, R. Appuswamy, **N. McGlohon**, B. Taba, S. Esser, F. Akopyan, J. Arthur, A. Amir, A. Andreopoulos, P. Carlson, A. Cassidy, P. Datta, M. Flickner, R. Gandhasri, G. Garreau, M. Ito, J. Klamro, J. Kusnitz, N. McClatchey, J. McKinstry, T. Nayak, C. Tadeo Ortega Otero, H. Penner, W. Risk, J. Sawada, J. Sivagnaname, D. Smith, R. Sousa, I. Terrizzano, T. Ueda, T. Gray-Donald, D. Cox, and D. Modha. *A Scalable NorthPole System with End-to-End Vertical Integration for Low-Latency and Energy-Efficient LLM Inference*. ArXiv 2511.15950v1 Pre-print article. November 2025
- R. Appuswamy, M. DeBole, B. Taba, S. Esser, A. Cassidy, A. Amir, A. Andreopoulos, D. Bablani, P. Datta, J. Kusnitz, N. McClatchey, **N. McGlohon**, J. McKinstry, T. Nayak, D. Smith, R. Sousa, I. Terrizzano, F. Akopyan, P. Carlson, R. Gandhasri, G. Garreau, N. Mimura Gonzalez, M. Ito, J. Klamro, Y. Nakamura, C. Tadeo Ortega Otero, W. Risk, J. Sawada, K. Schleupen, J. Sivagnaname, M. Stallone, T. Ueda, M. Flickner, J. Arthur, R. Panda, D. Cox, and D. Modha. *Breakthrough low-latency, high-energy-efficiency LLM inference performance using NorthPole*. IEEE High Performance Extreme Computing (HPEC), Boston, USA. September 2024.
- F. Akopyan, W. Risk, J. Arthur, A. Cassidy, M. DeBole, C. Tadeo Ortega Otero, J. Sawada, E. Colgan, M. Criscuolo, P. Mann, H. Baier, K. Schleupen, A. Amir, A. Andreopoulos, R. Appuswamy, D. Bablani, P. Carlson, P. Datta, S. Esser, M. Flickner, R. Gandhasri, G. Garreau, M. Ito, J. Klamro, J. Kusnitz, N. McClatchey, **N. McGlohon**, J. McKinstry, Y. Nakamura, T. Nayak, J. Sivagnaname, D. Smith, R. Sousa, B. Taba, I. Terrizzano, T. Ueda, and D. Modha. *Breakthrough edge AI inference performance using NorthPole in 3U VPX form factor*. IEEE High Performance Extreme Computing (HPEC), Boston, USA. September 2024.
- E. Cruz, S. Qian, A. Shulka, **N. McGlohon**, S. Rakheja, C. D. Carothers. *Performance Evaluation of Spintronic-Based Spiking Neural Networks Using Parallel Discrete-Event Simulation*. ACM Transactions on Modeling and Computer Simulation (TOMACS), March 2024.
- E. Cruz, S. Qian, A. Shulka, **N. McGlohon**, S. Rakheja, C. D. Carothers. *Evaluating Performance of Spintronics-Based Spiking Neural Network Chips using Parallel Discrete Event Simulation*. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS), Virtual. June 2022.
- N. McGlohon**, K. S. Hemmert, K. A. Brown, M. Levenhagen, S. Chunduri, R. B. Ross, C. D. Carothers. *Exploration of Congestion Control Techniques on Dragonfly-class HPC Networks Through Simulation*. IEEE SC'21 Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computing Systems (PMBS), Virtual. November 2021.
- J. Ma, J. Goodhue, K. Nelson, A. Sherman, E. Brown, C. D. Carothers, G. Collier, A. Del Maestro, A. Elledge, W. Figurelle, J. Huffman, G. Khanna, **N. McGlohon**, S. Najafi, J. Nucciarone, A. Schwartz, B. Segee, S. Valcourt, R. Zottola. *Leveraging Northeast Cyberteam Successes to Build the CAREERS Cyberteam Program: Initial Lessons Learned*. PEARC Workshop on Strategies for Enhancing HPC Education and Training (SEHET), Virtual. July 2021.
- N. McGlohon**, C. D. Carothers. *Toward Unbiased Deterministic Total Ordering of Parallel Simulations with Simultaneous Events*. ACM SIGSIM Winter Simulation Conference (WSC), Virtual. December 2021.
- K. A. Brown, **N. McGlohon**, S. Chunduri, R. B. Ross, E. Borch, C. D. Carothers, K. Harms. *A Tunable Implementation of Quality-of-Service Classes for HPC Networks*. ISC High Performance (ISC), Virtual. June 2021.
- N. McGlohon**, C. D. Carothers. *Unbiased Deterministic Total Ordering of Parallel Simulations with Simultaneous Events*. ArXiv 2105.00069v1 Pre-print article. May 2021.
- N. McGlohon**, R. B. Ross, C. D. Carothers. *Evaluation of Link Failure Resilience in Multi-Rail Dragonfly-Class Networks*. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS), Miami, USA. June 2020.
- N. McGlohon**, N. Wolfe, M. Mubarak, C. D. Carothers. *Fit Fly: A Case Study on Interconnect Innovation Through Parallel Simulation*. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS), Chicago, USA. June 2019.

- Y. Kang, X. Wang, **N. McGlohon**, M. Mubarak, S. Chunduri, Z. Lan. *Modeling and Analysis of Application Interference on Dragonfly+*. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS), Chicago, USA. June 2019.
- M. Mubarak, **N. McGlohon**, M. Musleh, E. Borch, R. B. Ross, R. Huggahalli, S. Chunduri, S. Parker, C. D. Carothers, K. Kumaran. *Evaluating Quality of Service Traffic Classes on the Megafly Network*. ISC High Performance (ISC), Frankfurt, Germany. June 2019.
- M. Plagge, C. D. Carothers, E. Gonsiorowski, **N. McGlohon**. *NeMo: A Massively Parallel Discrete-Event Simulation Model for Neuromorphic Architectures*. ACM Transactions on Modeling and Computer Simulation (TOMACS). September 2018.
- M. Plagge, **N. McGlohon**, C. Ross, C. D. Carothers. *Simulation and Visualization of Custom Neuromorphic Hardware using NeMo*. Neuromorphic Computing Symposium on Architectures, Models, and Applications, Oak Ridge National Laboratory, USA. July 2017.
- N. McGlohon**, S. Patterson. *Distributed Semi-Stochastic Optimization with Quantization Refinement*. American Control Conference (ACC), Boston, USA. July 2016.
- S. Patterson, **N. McGlohon**, K. Dyagilev. *Optimal k -Leader Selection for Coherence and Convergence Rate in One-Dimensional Networks*. IEEE Transactions on Control of Network Systems (TCNS). January 2016.
- S. Patterson, **N. McGlohon**, K. Dyagilev. *Efficient, Optimal k -Leader Selection for Coherent, One-Dimensional Formations*. European Control Conference (ECC), Linz, Austria. July 2015.