

Neil McGlohon, Ph.D.

ABOUT

nmcglohon@gmail.com
nmcglo.com

github.com/nmcglo
linkedin.com/in/nmcglo

Systems engineer and applied scientist with 6+ years building performance-critical distributed and parallel software, recently leading software development and rack-scale deployment of an LLM inference engine and runtime stack for IBM's NorthPole AI accelerator card. Finely sharpened specialties, with the generality to take on any new stack in stride.

SKILLS AND EXPERTISE

AI/ML: PyTorch, TensorFlow, vLLM, Ray, Langflow, agentic systems, neuromorphic hardware
Parallel & Distributed Computing: HPC, MPI, NCCL, SLURM, cloud-native architectures
Cloud & Infrastructure: Kubernetes, OpenShift, Helm, Docker, Podman, Grafana, Prometheus, Infiniband & RoCE, NFS, Longhorn, MetalLB, AWS, IBM Cloud
Software Engineering: C/C++, Python, JavaScript/Node.js/Vue.js, Java, GitOps (Git, Jenkins, ArgoCD), CMake, AMQP, S3, REST, SQL, MongoDB, asynchronous event-driven systems

PROFESSIONAL EXPERIENCE

- **IBM, IBM Research**
Almaden Research Center, San Jose, California USA
Staff Research Scientist **October 2022 – July 2026**
 - Led design and development of a purpose-built LLM inference engine for the IBM AIU NorthPole AI inference accelerator card, delivering state-of-the-art low latency and energy efficiency.
 - Led team of 5 engineers building NorthPole's performance-critical runtime stack.
 - Deployed rack-scale distributed AI systems of 288 NorthPole cards, generating over 500,000 tokens per second in aggregate while holding per-token latency as low as 1 millisecond.
 - Collaborated with model team to evaluate our quantized and retrained models using tools and benchmarks including EleutherAI's lm-eval and Hugging Face's OpenLLM leaderboards.
 - Designed, prototyped, and shipped a cloud-native OpenShift stack exposing NorthPole LLM inference through a controlled API and chat front-end; integrated back-end into IBM watsonx, bringing NorthPole-accelerated inference to workloads across the platform.
 - Created NorthPole runtime client/server applications leveraging the YOLOv9 model to perform object detection, identification and tracking on real-time video streams, providing over 300 megapixels per-second of image processing consuming less than 60W of card power.
 - Guided broader team as subject-matter expert in parallel and distributed systems, containerization, orchestration, and LLM inference serving including vLLM plugin development and integration.
 - Represented NorthPole at business development meetings, client outreach and showcase events, presenting the platform to prospective industry and public-sector partners. Bridged scientific and business-focused teams, driving knowledge transfer and cross-team collaboration.
- **Rensselaer Polytechnic Institute, Center for Computational Innovations**
Troy, New York USA
Research Scientist **July 2021 – September 2022**
 - Development and optimization of workloads on the HPC systems for internal and partner research.
 - Served on steering committee of an NSF-funded project-of-projects, mentoring participating graduate students as a subject-matter expert in HPC and AI systems.
 - External collaborator with IBM AI Hardware Center on research into distributed AI training benchmarking and simulation of composable systems.
- Software Engineer** **February 2020 – July 2021**
 - Maintained the open-source Rensselaer Optimistic Simulation System (ROSS) project, a parallel-discrete-event-simulation (PDES) engine capable of providing strong-scaling linear speedup even when distributed across 2-million supercomputer CPU cores.
 - Maintained CODES, an open-source framework for massively parallel simulation of HPC communication networks using ROSS, collaborating with researchers at Argonne and Sandia National Laboratories.
 - Collaborated with external research partners, writing specifications and architecting bespoke software solutions to capture bleeding-edge experimental results.

	– Cisco Meraki San Francisco, California USA <i>Software Engineering Intern</i> May – August 2017 - Researched and developed firmware for Meraki’s MR platform access points. - Designed and built a lockless, thread-safe hashtable using RCU techniques, and upstreamed the contribution to Click, an open-source software-defined router project.
SELECTED PROJECTS	– Homelab Infrastructure Platform 2022 – Present nmcglo.com/homelab github.com/nmcglo/home-gitops - Built and operate a 4-node ARM Kubernetes cluster running 30+ self-hosted services, each packaged as a Helm chart and continuously deployed from Git by ArgoCD. All secrets injected from 1Password operators to avoid any secret data being committed to Git. - Designed a walled-garden network for autonomous AI agents: a deny-by-default restricted VLAN, per-agent scoped Git identities, and a dedicated Jenkins instance running on-demand build pods in an isolated namespace, reducing the potential consequences and risk from errant agents. - Serve local LLM inference on RTX 5070 Ti, RTX Pro 2000, and Rockchip RK3588 NPU hosts, powering interactive and unattended coding agents entirely on-premises.
EDUCATION	Ph.D., Computer Science July 2021 Rensselaer Polytechnic Institute, Troy, New York USA Research Focus: High Performance/Parallel Computing M.S., Computer Science May 2016 Rensselaer Polytechnic Institute, Troy, New York USA Research Focus: Distributed Computing, Machine Learning B.S., Physics May 2014 University of Oklahoma, Norman, Oklahoma USA Minor: Computer Science
ACADEMIC EXPERIENCE	– Rensselaer Polytechnic Institute, Troy, New York USA <i>Graduate Research Assistant</i> August 2014 – July 2021 Ph.D. research, coursework and projects. Areas of focus: high performance and parallel computing, distributed systems, cloud computing, machine learning, neuromorphic computing and simulation. <i>Teaching Assistant: Computer Science I</i> August – December 2014 Supported the instructor of an introductory Python course: led lab sessions and discussions, held office hours, and proctored and graded exams/assignments. – University of Oklahoma, Norman, Oklahoma USA <i>Undergraduate Research Assistant</i> May 2012 – May 2014 Created process to measure differential conductance of cryogenic quantum-confined antimony films. <i>National Science Foundation REU</i> May – August 2012, 2013 Participated in condensed-matter NSF Research Experience for Undergraduates program including first observations of weak anti-localization in cryogenic antimony films.
SCIENTIFIC SERVICE	– NSF CAREERS Cyberteam Program Steering Committee Member – 2020–2022 – ACM SIGSIM Principles of Advanced Discrete Simulation Organizing Committee, Publicity Chair – 2023 Organizing Committee, Ph.D. Colloquium Chair – 2022 – Program Committee Memberships ACM SIGSIM Principles of Advanced Discrete Simulation – 2022–2025 ACM SIGPLAN Benchmarking in the Data Center: Expanding to the Cloud – 2022 IEEE IPDPS Scalable Deep Learning over Parallel and Distributed Infrastructure – 2022
PUBLICATIONS	Author on 18 papers published in conferences and journals spanning AI/ML, neuromorphic computing, high performance computing, and parallel discrete event simulation. Full list at nmcglo.com/cv.