

## Sonny McManigle

### HPC Supercomputing GPU Cluster Engineer

 sonny.mcmanigle@gmail.com |  724-518-7869

 Pittsburgh, PA |  LinkedIn: [www.linkedin.com/in/sonny-mcmanigle](https://www.linkedin.com/in/sonny-mcmanigle)

## Professional Summary

Results-driven Senior HPC Datacenter Architect and Engineer with over 25 years of experience. Expert in designing, deploying, and automating high-performance computing infrastructures, specializing in bare metal GPU deployments and high-speed HPC architectures. Proven leader in datacenter modernization, AI/ML infrastructure optimization, and end-to-end automation, achieving measurable performance and efficiency gains in petabyte-scale HPC environments. Core expertise includes large-scale GPU platforms (air and liquid cooled), high speed Ethernet and InfiniBand networking, distributed storage, and GPU Kubernetes on bare metal.

## Core Technical Competencies

### HPC Infrastructure & Architecture

- **Bare Metal GPU:** MaaS, Tinkerbell, ClusterAPI, EKS Anywhere; automated provisioning and lifecycle management.
- **Supercomputing:** Liquid only build outs of GB200 and GB300 NVL72 servers
- **High-Speed Networking:** Mellanox InfiniBand (800G), Mellanox Ethernet (800G), Blue-Field DPU, NVIDIA UFM, NVIDIA NetQ, NVIDIA AIR, Arcamesh, Mellanox Spectrum and Quantum, Carrier BGP
- **HPC Storage:** Quobyte, GPFS (IBM Spectrum Scale), ArcaStream, NVMe direct-attach, Oracle Exabyte, distributed filesystems, VAST, DDN.
- **GPU Platforms:** GB200 NVL72, GB300 NVL72, NVIDIA DGX-1 HGX, Blackwell B300s, RTX8000s, A40s, NVIDIA AI Factory and GPUOperator deployment; multi-GPU k8 cluster architecture.

### Containerization & Orchestration

- **Kubernetes on Bare Metal:** EKS Anywhere, ClusterAPI, Helm, custom operators and CRDs.
- **Platforms:** KVM, Qemu, Proxmox, Docker, Podman, Portainer, Amazon EKS, Google GKE, GitLab CI/CD runners.
- **Orchestration:** Tinkerbell for bare metal provisioning, ClusterAPI for cluster management, Helmsman.

### Automation & Infrastructure as Code (IaC)

- **IaC Tools:** Terraform, Ansible, Python, Terragrunt, CloudFormation, Troposphere.

- **Scripting:** Python, Go, Bash, PowerShell, custom MCP support.
- **CI/CD:** GitLab CI/CD, Jenkins, automated deployment pipelines.

### **Datacenter & Management (DCIM)**

- NetBox; automated DCIM as IaC, power estimation, capacity planning, cross-regional migration planning, liquid cooling environmental preparation and visual elevation design.

### **Monitoring & Security**

- **Monitoring:** CheckMK, Datadog, Prometheus, Grafana, ELK Stack, Splunk
- **Security:** HashiCorp Vault, AWS IAM, Okta, Juniper/Palo firewalls, PCI compliance.

## Professional Experience

HPC Supercomputing GPU Cluster Engineer - GlobalAI

NVIDIA Premium NeoCloud Partner

April 2026 - *Present*

- **Architected** a multi-region firmware platform orchestration and automation for compliance within sovereign and air-gapped supercomputing deployments
- **Architected** a DCIM automation platform for hardware reporting and compliance automation defining a source of truth hardware neocloud deployment.
- **Administered** bare metal HPC platform deployments of compute, storage, spectrum ethernet, quantum infiniband, Blackwell and Rubin nvl trays and liquid CDU build, firmware and operational management of SuperMicro hardware
- **Collaborated** with a team of highly technical GPU Supercomputing HPC specialists to manage a multi-region Supercompute deployment across customers and hardware in a managed and orchestrated manner.
- **Developed** automated auditing and compliance mechanisms for sovereign, air-gapped and specific customer vertical needs
- **Administered** SLURM for Fortune 100 clients to a 99% uptime. Hardware and platform procurement and orchestration through operational support

### **Staff Site Reliability Engineer (SRE) – PathAI**

January 2022 – *Present* | Boston, MA / Remote

- **Led** the architecture and operations of a petabyte-scale hybrid cloud HPC datacenter, specializing in bare metal GPU clusters integrated with AWS for AI-driven pathology research.
- **Architected** a vendor compute shift and liquid B300 GPU Architecture design and deployment.
- **Led** and managed a regional full datacenter move selecting and scoring regions and

vendors and designing the move event including move partner and carrier relationships.

- **Designed and deployed** automated bare metal provisioning using Tinkerbell and ClusterAPI for 100+ specialized GPU nodes.
- **Architected** large-scale GPU clusters, achieving a **60% reduction in ML model training time** through optimized GPU scheduling and storage tuning.
- **Engineered** multi-tier HPC storage (Quobyte, GPFS) and InfiniBand network (100G/200G/800G roadmap) utilizing NVIDIA UFM Subnet Manager.
- **Executed** a cross-regional datacenter move with zero-downtime, resulting in **millions in cost savings** through optimized refresh cycles.
- **Developed** a comprehensive DCIM automation platform (IaC), improving datacenter accuracy to **99.5%** and reducing provisioning time by **70%**.
- **Managed** 15+ Kubernetes clusters across multiple AWS regions (500+ worker nodes) supporting GPU-intensive ML workloads, implementing Karpenter for dynamic provisioning.
- **Achieved 99.9% uptime SLA** for critical ML infrastructure by implementing Datadog, Prometheus, Grafana, and ELK stack.

### Senior Enterprise Architect – Argo AI

*October 2021 – November 2022 | Pittsburgh, PA*

- **Designed and deployed** an OpenLDAP enterprise directory service with automated CI/CD and integration with HRIS/iDP systems.
- **Provided** architecture review for a VDI environment, including GPU offloading using NVIDIA DGX-1.
- **Reviewed and optimized** AWS Direct Connect and GCP Cloud Interconnect designs for routing and security.

### Lead Systems Engineer II – American Eagle Outfitters

*May 2007 – October 2021 | Warrendale, PA*

- **Architected** a hybrid Office365/Exchange 2016 environment and managed Kubernetes Cluster Resources on GCP.
- **Led** DNS architecture (Infoblox, MS DNS, BIND) and enterprise architecture of Active Directory Forest Schemas.
- **Achieved annual PCI compliance** by leveraging scripting and orchestration tools for store-based deployments.

### Key Achievements

- **60% Reduction** in ML model training time through GPU scheduling and storage tuning.
- **99.9% Uptime** achieved for critical ML infrastructure supporting 24/7 workloads.
- **70% Reduction** in provisioning time and **99.5%** datacenter accuracy via full DCIM Automation (IaC).
- **Millions Saved** through strategic datacenter regional migration and hardware refresh optimization.

- **300% Infrastructure Scale** growth to meet increasing model demand.
- **Petabyte-Scale** HPC storage architecture designed and implemented.

## Certifications

- Microsoft Certified Professional
- VMware Certified Professional
- Red Hat Certified Engineer
- Cisco Certified Engineer
- ProofPoint Certified
- GCP Certified
- Novell Certified Engineer

*References available upon request.*