Shreyansh Sancheti

Bengaluru, India | +91 8446267174 | hire.ssancheti@gmail.com
shreyanshjain7174 | shreyansh-sancheti | portfolio-ssancheti.vercel.app

Summary

AI Infrastructure / Platform Engineer at Microsoft Azure Compute (Hyper-V). **4+ years** in **Kubernetes**, **LLM inference**, distributed systems and AI agent tooling; **15+ merged Ceph PRs**, systems serving **10,000+ SMEs** and **1000+ deployments**.

Technical Skills

Languages: **Python**, **Go**, **Rust**, C/C++, Swift, Bash, PowerShell, TypeScript, SQL
AI Infrastructure: **LLM inference**, **vLLM**, **KServe**, **llm-d**, inference autoscaling (HPA), model serving, multi-tenant isolation, **MCP**, RAG, LangGraph, LangChain, Azure OpenAI, agentic workflows
Cloud & Platform: **Kubernetes** (operators, CRDs), **Azure**, AWS, Docker, **Terraform**, Ansible (configuration management), IaC, containerd, Hyper-V, gVisor, Cilium, Ceph, Linux, backend systems, distributed systems, open source
Observability & Delivery: **Prometheus**, Grafana, OpenTelemetry, Kusto, ELK, **CI/CD** (GitHub Actions, GitLab CI), incident response, root cause analysis (RCA), capacity planning, high availability, cost optimization
Customer-Facing / GTM: solution design, enterprise integration, technical advisory, REST/gRPC APIs, workflow automation (n8n), human-in-the-loop AI, demos, conference talks

Projects

Agentic Operator, Clawdlinux (Founder) | *Go, Kubernetes, gVisor, Cilium, MCP* | GitHub

- Built an open-source **Kubernetes operator** that translates natural-language intent into infrastructure action, with governance controls for autonomous AI agents.
- Isolated agent workloads with **gVisor** sandboxing and **Cilium** egress policy for multi-tenant safety; added cost controls and offline-verifiable audit trails.
- Exposed cluster state and actions through an **MCP server**; awarded the **Mercatus Emergent Ventures** research grant and presented at Agentic Summit Bangalore.

ZeroDB | *Rust, Distributed Systems* | GitHub

- Built a durable **agent memory database** in Rust with an fsync-before-ack write-ahead log, group commit and an MCP API.
- Benchmarked on Azure against Redis, Valkey and DiceDB with 16 concurrent writers: **48.6K acks/s** at **502 μs p99**.
- Validated correctness and crash recovery with deterministic simulation testing across **2,000** seeded runs.

Piko | *Swift, iOS, On-device ML* | GitHub

- Building a personal intelligence app for iPhone with **on-device** voice dictation using Apple SpeechAnalyzer and Foundation Models.
- Designed a privacy-first architecture where **zero audio leaves the device**; no cloud inference or server round trip.
- Planned system-wide integration through a keyboard extension and Live Activity for hands-free capture in any app.

Professional Experience

Microsoft Mar 2026 – Present
Software Engineer II, Container Platform & Infrastructure (Azure Compute / Hyper-V)

- Drive migration from legacy **HCSShim/HCS API** integrations to modern container platform abstractions, enabling **Live Migration** across Azure-scale container infrastructure.
- Built automated **parity and regression validation frameworks** across container runtime layers before production rollout.
- Develop **PowerShell/Go** deployment automation and CI/CD workflows covering package rollout, environment setup and validation.
- Added reliability validation and observability to improve production readiness.
- Perform **root cause analysis** of distributed failures across Azure clusters using **Kusto** telemetry.

Arista Networks Feb 2025 – Feb 2026
Software Engineer, Network Infrastructure & Performance (C++)

- Built the **BESS forwarding agent** anomaly detection with predictive task scheduling, **eliminating 100% packet loss** during port updates.
- Worked on platform-independent routing (Caravan) and high-performance networking software in **C++**.
- Automated incident detection on control-plane metrics with root cause correlation, cutting **MTTR by 40%**.
- Built **Grafana** dashboards and alerting for **400G+** throughput systems; CI/CD improvements raised deployment reliability **35%**.

XBattery Mar 2025 – Oct 2025
Senior Infrastructure Consultant, AI & Cloud Platform

- Built an AI-powered **Kubernetes observability** platform (Prometheus, Grafana, anomaly detection) serving **10,000+ SMEs** with a **99.9% uptime** SLA and **60%** less manual intervention.
- Delivered **Terraform, Ansible and GitLab CI** pipelines for configuration management and infrastructure automation, enabling **50+ daily zero-downtime deployments**.
- Correlated metrics, logs and traces across **500+ VMs**, cutting false-positive alerts by **75%**.
- Automated **root cause analysis** workflows integrated with ticketing systems, accelerating incident resolution by **50%**.

IBM Jan 2023 – Feb 2025
Software Engineer II, Cloud Storage & Automation

- Built **Ceph** performance monitoring (Prometheus, Python) that detected anomalies across **1000+ enterprise deployments**.
- Automated troubleshooting with Python/Bash log analysis, reducing incident response time by **40%**.
- Designed **Rook-Ceph Kubernetes operators** with health monitoring, cutting **MTTR by 60%**; **15+ PRs merged** to upstream open-source Ceph.
- Built a Docker/Ansible S3 performance benchmarking framework, a **10x** productivity gain for storage testing.

Red Hat Jan 2022 – Dec 2022
Associate Software Engineer, Distributed Systems

- Improved distributed storage **throughput by 25%** through automated profiling and bottleneck analysis.
- Cut production incidents by **30%** by adding monitoring and alerting to Ceph **distributed systems**.
- Reduced QA cycle time by **40%** with Docker-based performance test automation.
- Completed **100+ code reviews** on open-source storage contributions, improving code quality and team velocity.

Talks & Open Source

- **SREcon26 EMEA** (Dublin, Oct 2026), lightning talk: why the **Kubernetes autoscaler** fails **LLM inference** workloads and what **vLLM, KServe and llm-d** do instead.
- Speaker at **Ceph Days** conferences and **Agentic Summit Bangalore** on distributed storage and AI agent infrastructure.
- **15+ merged PRs to open-source Ceph** (PRs) covering Rook-Ceph operators, monitoring and S3 performance tooling.
- Contributing to **containerd** and **hcsshim**, the Windows container runtime layer used across Azure.

Education

Walchand Institute of Technology Aug 2018 – Jul 2022
B.Tech, Computer Science and Engineering *CGPA: 9.3/10.0*