

Manu Zhephod Chert

manu@zhephod.xyz | github.com/icehunt | zhephod.xyz | x.com/zhephod

HIGHLIGHTS

- Built and operated Kubernetes platforms running thousands of high-performance code sandboxes for AI agents
- Designed multi-cluster orchestration, placement, and isolation (Karpenter, Istio, GitOps) across AWS fleets
- Cut sandbox connection time from 30+ seconds to under 10 seconds
- Brought up Intel iGPU split virtualization (GVT-g) on KVM/QEMU/VFIO so guests shared one integrated GPU

EXPERIENCE

Contour AI

May 2026 – Present

Lead Infrastructure Engineer

San Francisco, California

- Built the production EKS platform from scratch with OpenTofu, Karpenter, Argo CD, and GitOps so product teams could ship without an ops bottleneck
- Scheduled latency-sensitive workloads onto dedicated Karpenter node pools; isolated cluster traffic with Istio ambient mTLS and Tailscale-gated APIs
- Ran Temporal workers in production for long-running AI-agent jobs, including IAM, secrets, placement, and queue monitoring
- Stood up Grafana, Mimir, and Victoria Logs with Temporal/queue/Karpenter alerts; shipped per-PR preview environments and automated Terragrunt plan/apply

Julius.AI

October 2024 – May 2026

Lead Infrastructure Engineer

San Francisco, California

- Scaled code sandbox infrastructure across multiple Kubernetes clusters to thousands of concurrent users
- Optimized sandbox spin-up, bringing initial connection time from 30+ seconds to under 10 seconds
- Overhauled Terraform so environment deploys went from hours of error-prone applies to minutes
- Cut sandbox persistent-storage cost 90% with a more scalable backend; led the infrastructure side of SOC2

zhephod.xyz

April 2023 – October 2024

Solutions Consultant

Earth (Digital Nomading)

- Designed scalable scRNA inference for Phenomic AI and mentored data scientists building a full-stack platform for their models
- Added cloud autoscaling for build queues at an AI hardware startup; used Nvidia Nsight to improve client inference performance

Dewpoint Therapeutics (Condensates Drug Discovery Startup)

October 2020 – March 2023

Cloud Engineer/DevOps Contractor

Toronto, Ontario (Remote)

- Scaled Phenomic's imaging AI platform to process millions of microscopy images at once
 - Collaborated with software engineering to run a multi-node Dask cluster across hundreds of Kubernetes nodes, cutting processing from 1 week to 4 hours
 - Bypassed S3 limits for 3× faster preprocessing of large image datasets; influenced design for performance, disaster recovery, and TCO
- Designed and led the initial cloud infrastructure to improve access to cloud resources and simplify later development

Phenomic AI (Drug Discovery Startup)

April 2019 – September 2020

Cloud/DevOps Engineer

Toronto, Ontario

- Built cloud infrastructure for an ML drug-discovery pipeline, cutting experiment iteration from 1 month+ to 1 week
- Migrated Omero storage from EBS to EFS IA: \$25,000/year (90%) savings, better scalability, and less maintenance

TECHNICAL SKILLS

Languages:	Python, Terraform, C, Rust, Java, Bash, JavaScript
Orchestration:	Docker, Kubernetes, EKS, Argo CD, Karpenter, Istio, KVM, QEMU, VFIO, Temporal
Observability:	Prometheus, Grafana, Mimir, Loki, Victoria Logs, CloudWatch
Cloud / IaC:	AWS, GCP, OpenTofu, Terragrunt, Terraform, Tailscale
CI/CD:	GitHub Actions, GCP Cloud Build, GitLab CI, Drone CI, TeamCity

EDUCATION

University of Toronto

Toronto, Ontario

Honours Bachelor of Science, Specialist in Computer Science