Alan Zamani

Perth, Australia | alanzamani.dev

Summary

Nine years building production systems in Australia, eight at the country's largest supercomputing centre. Designs and runs Python services that schedule, allocate and monitor GPU workloads across multi-tenant clusters. Built vScheduler, an open-source HPC scheduler for interactive GPU sessions, in production since 2022, with the accompanying paper published at ISC High Performance 2026.

Experience

HPC Software Engineer — Visualisation Specialist, Pawsey Supercomputing Research Centre (CSIRO) – Perth, Australia — Mar 2018 – present

Designs and operates the centre's interactive GPU platform: scheduling, remote access and the infrastructure underneath it.

- Designed, built and shipped vScheduler, a Python scheduling and orchestration service allocating GPU-backed interactive sessions across two production clusters — a Linux HPC partition and a Windows GPU cluster — unifying heterogeneous hardware under one scheduling model. In production since 2022 across 51 GPU nodes.
- Implemented two allocation models in one system: on-demand scheduling via a custom load balancer that ranks nodes on live utilisation and health, plus reservation-based allocation validated against a booking service. No open-source alternative combines both; the capability previously required commercial products such as AWS NICE DCV and NICE EnginFrame.
- Built the control plane end to end: socket-based telemetry agents streaming session and health events, a MySQL-backed state store as the single source of truth, systemd-managed services, and policy enforcement for quotas, walltime and fair-share.
- Shipped a Slurm-style operator CLI (vmanage, valloc, vkill, vquota, vinfo, vreport) plus an HTML reporting and observability layer.
- Architected for failure isolation and portability: separated control and data planes, per-site deployable modules and an optional booking backend. Open-sourced at github.com/PawseySC/vScheduler.
- Operates the clientless remote-access platform — an Apache Guacamole HTML5 gateway brokering RDP, VNC and SSH behind a reverse proxy and institutional SSO.
- Provisions the estate with Ansible across bare metal, KVM/Proxmox virtual machines and an in-house OpenStack private cloud.
- Operates the centre's render farm and NVIDIA Omniverse environment.

Software Engineer, Stake – Sydney, Australia — Mar 2017 – Feb 2018

Built the customer-facing trading platform in a regulated financial-services environment.

- Built account opening, order placement and portfolio monitoring in Angular, integrated with backend REST APIs and market-data feeds.

Publications

A Scheduler System for Interactive Visualization on HPC Resources — June 2026

First author. System design, implementation and evaluation.
Ali Zamani, Yathu Sivarajah, Maciej Cytowski, Ugo Varetto
10.23919/ISC.2026.11520479 (ISC High Performance 2026, Hamburg)

Projects

vScheduler Jan 2022

An open-source HPC scheduler for interactive GPU sessions, in production at Pawsey since 2022 across 51 GPU nodes. Unifies a Linux HPC partition and a Windows GPU cluster under one scheduling model, with both on-demand and reservation-based allocation.

- Stack: Python, Slurm, MySQL, Linux, Ansible
- https://github.com/PawseySC/vScheduler

vScheduler MCP Gateway In development

A production-grade Model Context Protocol server exposing GPU scheduling and cluster operations as tools callable by LLM agents. OAuth 2.1, per-tool rate limiting, structured audit logging, OpenTelemetry tracing and Prometheus metrics, packaged as a Helm chart.

- Stack: MCP, OAuth 2.1, OpenTelemetry, Prometheus, Helm

Kubernetes ML Training Platform In development

Gang-scheduled GPU training job queueing with Kueue, experiment tracking with MLflow, model serving with KServe and full-stack observability. Deliberately mirrors the vScheduler architecture on cloud-native primitives, with the MCP gateway driving live training queues.

- Stack: Kubernetes, Kueue, MLflow, KServe, Prometheus/Grafana

Skills

Python, Kubernetes, GPU scheduling, Slurm, Linux, Docker/Helm, Ansible, OpenStack, AWS, Prometheus/Grafana

Education

UNSW Sydney, MSc in Petroleum Engineering Jan 2015 – Jan 2017

- Computational research. Thesis: pore network modelling for coal seams — 3D numerical flow simulation over networks extracted from micro-CT imaging, benchmarked against Lattice Boltzmann.

Certifications

Certified Kubernetes Administrator: Cloud Native Computing Foundation — in progress