

Frank Cancedda

Tech Lead · AI Infrastructure & Platform

New York, USA · English / Italian

fmcancedda@gmail.com

linkedin.com/in/fcancedda · github.com/fcancedda

repo1.dso.mil/frank.cancedda · cancedda.dev

SUMMARY

Tech Lead at Technergetics, working on AI infrastructure and platform for programs in the DoD Platform One ecosystem — production Kubernetes, AI inference (Triton, vLLM, TensorRT-LLM), and CI/CD for regulated workloads. Led a 3-person platform team through 2024. Built the local dev platform that brings an inference stack of 40+ services and jobs up with one **tilt up**. M.S. Data Science; trains small models for retrieval and reasoning in PyTorch. Runs a self-hosted Kubernetes homelab.

EXPERIENCE

Senior Software Engineer → Tech Lead · Technergetics, LLC

Sep 2022 – Present · Remote (NY)

- Tech lead for the platform domain — own architecture decisions and the technical roadmap for the AI inference stack, serving programs in the DoD Platform One ecosystem. Led a 3-person platform team through 2024: sprint planning, hiring input, and the final call on technical trade-offs.
- Built a Tilt + k3d local orchestration platform mirroring production, bringing a full AI inference stack (vLLM, Triton, Keycloak, Istio) to a single **tilt up** — team onboarding went from half a day to one command.
- Architected and shipped AWS/EKS microservices with Helm; automated infrastructure with Terraform and Ansible.
- Rebuilt release pipelines in GitLab CI (build, test, scan, deploy) — every release reproducible and auditable, removing single-person bus-factor from hand-run deploy scripts.
- Deployed and hardened NVIDIA Triton Inference Server for secure AI inference in regulated environments.
- Designed and implemented RAG pipelines integrating PGVector, TensorRT-LLM, and JWT-secured Java Spring Boot services.

Research Support Specialist · University at Albany, SUNY

Jun 2019 – Jan 2020 · Albany, NY

- Built Python data pipelines integrating FEMA, NOAA, NY511, and social-media APIs, with storage across Azure and Google Cloud (Oracle SQL).
- Implemented an NLP categorization pipeline (scikit-learn) for incoming emergency-alert data.

SELECTED PROJECTS

Cost-Aware Retrieval Policy — 1.6M-parameter policy network that learns per-query search strategies under a token budget; supervised warm-start + REINFORCE; +4.9 NDCG@10 over a tuned static baseline across four BEIR benchmarks. Built on a from-scratch PyTorch transformer lab. [PyTorch](#) · [REINFORCE](#) · [BEIR](#)

Homelab — self-hosted Kubernetes cluster in a 10-inch rack: GitOps (Flux), Authelia SSO, Traefik, Grafana, GitLab CI/CD; hosts live applications, battery-backed at ~50 W. [Kubernetes](#) · [Flux](#) · [Traefik](#) · [Grafana](#)

Vivazo (vivazo.cc) — Rust game engine with first-class WASM support and a formal API-stability system (Stable/Beta/Experimental tiers). [Rust](#) · [WASM](#) · [ECS](#)

Recursive Reasoning for Pathfinding — reimplementation of recursive reasoning architectures (HRM/TRM) for multi-agent pathfinding, extended with per-agent metrics and visualizations; evaluated on MovingAI. [PyTorch](#) · [MAPF](#)

EDUCATION

M.S., Data Science — University at Albany, SUNY

2020 – 2022

B.S., Computer Science & Applied Mathematics — University at Albany, SUNY

2015 – 2019

SKILLS

Platforms	Kubernetes (EKS, k3d, OpenShift), AWS, Linux · Tilt, Flux, Helm, Terraform, Ansible, GitLab CI/CD, Docker
AI / MLOps	Triton Inference Server, vLLM, TensorRT-LLM, PyTorch, PGVector, RAG pipelines
Languages	Python, Go, Rust, TypeScript, Java · Next.js/React · PostgreSQL, Redis · Authelia/OIDC, JWT, Istio, Grafana, Prometheus