Paul Richter

Miami, FL · paul@boronmail.com · github.com/Paulr13 · deepscaleanalytics.com

SUMMARY

AI infrastructure engineer running self-hosted systems in daily production since 2019: a multi-host Linux fleet (Guix System, NixOS) that grew into a multi-GPU LLM inference stack with always-on voice AI and image/video generation pipelines. Python/Bash automation and full-stack troubleshooting: CUDA driver issues, storage, constrained-network links, service recovery.

EXPERIENCE

AI Infrastructure Engineer (Independent) — DeepScale Analytics · Miami, FL 2019 – Present

- Operate a multi-host Linux fleet across two geographic sites (FL/WI): declarative configs on Guix System and NixOS, Caddy reverse proxy with TLS, SSH-based access, cron automation, LTE failover.
- Built a multi-GPU inference fleet (10× RTX 3090, production since 2023): Llama-class GGUF models, tensor-parallel llama.cpp/vLLM behind OpenAI-compatible endpoints; benchmarked throughput before/after tuning.
- Run an always-on voice AI service (STT → LLM → TTS) behind a public HTTPS endpoint; sub-second warm response, VAD-gated capture, in production daily.
- Operate ComfyUI/FLUX image- and video-generation pipelines on dedicated GPU hosts; batch workflows tuned to ~90s end-to-end.
- Run LLM-driven automation agents in production: scheduled monitoring and job pipelines, plus a vector memory layer (Qdrant) serving agent recall.
- Troubleshooting depth: CUDA/LD_LIBRARY_PATH environment failures, disk and backup strategy (additive rsync), service restart and recovery procedures.

Software Engineer — Deepscale Intelligence (NMU spinout) · Marquette, MI 2019 – 2024

- Early member of an NMU spinout, involved from the company's start (2019) through completion of my M.S.
- Helped secure ~$113k in MTRAC state translational-research funding and a $25k NSF I-Corps award (regional and national cohorts), the feeder into a planned SBIR/STTR application.
- Presented the platform to manufacturing audiences at national industry conferences (FabTech and others).
- Developed ModTalk, a communications project in the spinout's early portfolio.

Warehouse Operations & IT Support — UPS 2018 – 2020

- Internal help desk / tech support for hub operations: troubleshooting scanners, label printers, and terminal issues on the floor while working operations.
- Hub operations: unloaded and stacked packages, moved and parked trailers, coordinated drivers — building toward a full-time tech direction.

IT Support (part-time, concurrent roles) — Northern Michigan University · Marquette, MI

- NMU LTE Network Call Center: phone support for the university's LTE subscriber network — connectivity troubleshooting, device configuration, and account issues for students and employees. 2020 – 2024
- ThinkPad Help Desk: walk-up support for students, faculty, and staff — laptop hardware repair, OS and software troubleshooting, ticket resolution.
- Both roles ran concurrently with each other and the DeepScale startup work above.

AI Trainer (Contract) — Outlier AI (Scale AI) · Remote Jun 2024 – Mar 2025

- Evaluated, ranked, and rewrote LLM outputs on coding and reasoning tasks; wrote comparisons and corrections used for model training.

SKILLS

AI serving: llama.cpp, vLLM, GGUF quantization, OpenAI-compatible APIs, Whisper (STT), TTS pipelines, ComfyUI, FLUX, CUDA · **Systems:** Linux (Guix System, NixOS), Bash, SSH fleet ops, cron, systemd/shepherd services, remote administration · **Languages:** Python, Bash; C

Data/storage: Qdrant vector DB, rsync backup pipelines · **Web/API:** Caddy, TLS/HTTPS endpoints, FastAPI

EARLY EXPERIENCE (HIGH SCHOOL)

- FIRST Robotics · three seasons with Enstrom Helicopter Co. (Marinette, WI): robot build, wiring, programming.
- Business Professionals of America: national competitor (Web Design); state competitor (Animation, team).

EDUCATION

M.S. & B.S., Computer Science — Northern Michigan University