Nguyen Hong Quang

DevOps Engineer | AI Platform Architecture
0964848403 | nhq0810@gmail.com | nhq.id.vn | linkedin.com/in/nhqpin | Ha Noi, Vietnam

EDUCATION

University of Information Technology (UIT), VNU-HCM | Ho Chi Minh City, Vietnam | Sep 2018 – Sep 2023
Bachelor of Science in Information Security

EXPERIENCE

C-OpenAI (CMC Corporation), Vietnam | DevOps/AIOps Engineer | Jun 2026 – Present

- Directed GitOps delivery (GitLab CI, ArgoCD, Helm) for **~9 engineering teams**, keeping **100 ArgoCD applications** Synced/Healthy and taking the Console platform to production with **6 services**.
- Architected and operated a **national-scale AI platform** built for **5,000+ concurrent users**, rolled out to central and provincial government departments ahead of commercial customers, at **99.9% uptime**.
- Built a **digital-workforce system** of AI agents, coordinated by a manager agent under human approval, to run diagnostics, GitOps changes and alerting; cut operator workload and troubleshooting effort by **~30%**.
- Scaled the production LLM cluster (vLLM, LiteLLM, HAProxy): GPU node from bare metal to live traffic in **53 minutes**, **zero 5xx**; eliminated a recurring GPU crash (**91+ hours** crash-free) after benchmarking **+25–50% throughput**.
- Cut OCR job failures from **28% to 0.18%** by fixing CPU starvation, and built a two-node OCR cluster with two-tier load balancing serving ~3,700 requests with **zero 5xx errors**.
- Took the initiative to upgrade the platform's security baseline: Vault (OpenBao) secrets, hardened non-root images and infrastructure SOPs, driving Trivy HIGH/CRITICAL findings from 43 to **0**.
- Designed new systems end to end, including a multi-tenant Data Platform with Git-managed tenants, quotas and RBAC, and advised engineering teams on service architecture.
- Led GPU hardware and SaaS PoCs with go/no-go reports (FP8 checkpoints, RTX 5090/A6000 feasibility); shared knowledge via internal seminars and a 6-week LLM-serving research roadmap.

KienlongBank, Vietnam | Tech Lead AI & DevOps/AIOps Engineer | Dec 2024 – May 2026

- Spearheaded the Enterprise AI Platform (KILOBA), leading the team to deploy **10+** AI chatbots from research to production with multi-model LLM integration (OpenAI, Gemini, Bedrock).
- Architected and managed multi-cloud infrastructure (AWS EC2, EKS, VPC, RDS, Bedrock, SageMaker / GCP Vertex AI) serving **15+** production microservices for thousands of banking users.
- Implemented FinOps and a unified Observability stack (Prometheus, Loki, Tempo), reducing cloud compute costs by **~30%** and accelerating incident resolution from hours to minutes.
- Orchestrated a **zero-downtime** AWS Organizational Unit (OU) migration for the core production environment, drastically optimizing IAM posture and availability.
- Engineered GitOps CI/CD pipelines (ArgoCD, GitLab CI) and n8n workflows, reducing deployment cycles from days to minutes and saving **10+ hours/week**.
- Pioneered AIOps (Claude AI + MCP) and LightRAG knowledge graph architectures (Neo4j, Pinecone) to facilitate autonomous monitoring and high-accuracy retrieval.
- Secured platform architecture by enforcing Zero-Trust networking, Kubernetes RBAC, and strict IAM governance across hybrid cloud environments.
- Optimized microservice communication by deploying high-throughput API Gateways (APISIX, Traefik, Nginx) and Redis caching, ensuring low-latency banking operations.

CMC Telecom, Vietnam | Cloud Engineer Intern | Jun 2024 – Nov 2024

- Deployed site-to-site VPNs between customer premises and AWS VPCs. Troubleshot EKS, EC2 and RDS incidents on CMC Cloud and AWS within enterprise SLAs.

VNPro, Vietnam | Cloud Intern | Feb 2024 – Jun 2024

- Administered and hardened AWS lab environments for hundreds of CCNA, CCNP and AWS SAA trainees.

TECHNICAL SKILLS

Cloud & Containers: AWS (EC2, EKS, VPC, RDS, S3, IAM, Organizations), GCP (Vertex AI), Kubernetes, Helm, Docker, Docker Swarm
CI/CD & IaC: GitLab CI/CD, ArgoCD, AWS CodePipeline, Jenkins, Terraform, Ansible
Security: Vault / OpenBao, Kubernetes RBAC, IAM, OIDC SSO, container hardening
Observability: Prometheus, Grafana, Loki, Tempo, OpenTelemetry, Mimir, Alloy, EFK
AI Infrastructure: vLLM, LiteLLM, NVIDIA Triton, GPU clusters, RAG and LightRAG
Tools & Networking: Python, Bash, n8n, k6, guidellm, NVIDIA AIPerf, HAProxy, Nginx, Traefik, APISIX
Language: Vietnamese (native); English (TOEIC LR 600, SW 230, professional working)

SOFT SKILLS

Leadership & Ownership, Problem-Solving under Pressure, Critical Thinking, Cross-team Collaboration, Mentoring & Knowledge Sharing, Fast Technology Adoption, AI-Assisted Engineering, Self-motivated & Proactive