JAY KUBO

Tokyo, Japan · linkedin.com/in/jaykubo · github.com/jkubo · Japan and U.S. dual citizen

I build and run security and AI systems on infrastructure I operate myself — a 20-node Kubernetes cluster across four sites — and for a decade have been the technical-and-business bridge between Japanese enterprises and the cloud platforms they run on, inside Google Cloud, CrowdStrike, and Rakuten. Native Japanese and English, technical register.

EXPERIENCE

Founder, kub0 LLC

Tokyo, Japan | *Jul 2020 - Present*

INFRASTRUCTURE & SECURITY

- 20-node Kubernetes cluster across Austin, Los Angeles, Torrance, and Tokyo — ~1 TB GPU memory, 47 TB DRBD-replicated block storage, WireGuard mesh, Ansible-driven lifecycle behind per-node drain gates. Migrated the whole overlay network live, with zero downtime on active workloads.
- Rebuilt ingress for site-level failure: one Traefik replica per site, required anti-affinity by zone. A full-site outage now shifts traffic to the other two automatically, no manual re-pin.
- Tetragon eBPF as a DaemonSet: 97 TracingPolicies covering ~131 MITRE ATT&CK and 22 ATLAS techniques. vigiles-probe, a Rust eBPF sensor with a BPF_LSM enforcement hook, does live deny-list enforcement across all 20 nodes. Filed the upstream Ubuntu kernel bug (LP#2150798) that unblocks BPF_LSM on arm64 for Tetragon, Tracee, Falco, and KubeArmor.
- CI/CD gate that blocks a merge when a protected host has no OAuth2 companion route, failing closed on any manifest it can't parse. Every control is pinned by a case that must fail and a near-miss that must pass.
- Found a class of alert that passes review and can never fire — one compared a value to zero where the value is always zero; another ran as a job with no database password, so it never evaluated. Neither failed a test. Coverage counts only after a proven red run, not a green pipeline.
- Packaged the platform for GCP Marketplace: Helm chart, RS256 JWT licensing, free tier to 10 nodes and a per-node Pro tier above it. Ran the evaluation end to end on GKE — 47 policies loaded, sub-second enforcement — then priced the tiers against it. Google Cloud Partner Advantage member; approved in Anthropic's Cyber Verification Program.

AI SYSTEMS

- MALINT: an end-to-end malware pipeline — intake against a benign baseline, YARA static analysis, sandboxed detonation under gVisor with Tetragon capture, then a confidence-gated model verdict served on a vLLM inference fabric (AMD ROCm and NVIDIA CUDA) with LoRA fine-tuning on a DGX Spark. ~50,000 samples, ~86,000 behavioral signatures.
- LLM-as-judge evaluation with a 2×2 factorial design (base vs. fine-tuned, zero-shot vs. RAG) to separate the retrieval lift from the fine-tuning lift instead of claiming both. Fine-tune scored 2.79 vs RAG 3.92, so I shipped RAG.
- Project Gaius, open source on PyPI: hybrid semantic and keyword retrieval over 10,000+ facts behind a 7-tool MCP server. 98.6% retrieval hit@10 on LongMemEval-S (ICLR 2025), matching the published same-model baseline offline, no GPU.

Technical Advisor, Japan Deluxe Tours, Inc.

Remote | *May 2015 - Present*

Ten years on one enterprise account — first cloud, migration, production operations, and the business case renewed at every step.

- Recommended and stood up their first AWS cloud in 2015, then took over the full stack when the original developer left. Migrated the customer site to Cloudflare Workers and on-premises Kubernetes; MySQL 8.0 on NVMe with GTID replication, streaming replicas across three sites at 0 s lag.
- Translate goals into a roadmap and present the tradeoffs and the cost case to owners who are not technical, then the same case in engineering terms to whoever has to build it. Every expansion of scope over ten years — cloud, Kubernetes, payments, endpoint security — was argued and approved that way.
- Claude in production for their staff: a Google Chat bot that answers operations questions and ships approved changes, gated by directory roles and logged end to end. Stripe payments (deposits and self-service balance, no stored card); staff Macs on Tailscale with Santa binary authorization and FleetDM for inventory, compliance, and CVE response.

Associate Analyst, CrowdStrike (Falcon Complete Japan)

Austin, TX | *Jan 2025 - May 2025*

- Japanese-language translator on high-priority Falcon Complete escalations — analysts and senior analysts pulled me in, and what I wrote is what went out to the customer.

- ~30 JavaScript automations for Falcon and internal SOC tooling: bulk hash extraction, one-click VirusTotal submission, polling where a customer package had no native alert. One captured a file from a host online for under 60 seconds on a ticket open for days — evidence manual work would not have gotten. The team adopted the library immediately and kept using it after the layoff.
- Curated daily threat-intelligence feeds and trained a GovCloud senior analyst to run Japan commercial-cloud operations alone.

Application Support Analyst, EPAM Systems (Google Cloud)

Austin, TX | *Nov 2019 - Mar 2021*

- Frontline support for Japanese enterprise customers on Google Cloud, customer-facing under Google: BigQuery, Dataflow, Cloud Composer, Cloud Spanner, Cloud SQL, App Engine, Dialogflow. Gaming, e-commerce, rail, trading houses, telecom, banking.
- Bridged U.S. and Japan on high-priority escalations. A Chrome extension I built cut Platinum P0/P1 SLA misses to near zero. Carried Won't Fix rulings back to enterprise customers; moved stalled private cases onto the public tracker.
- Filed P2s from those accounts into Google's issue tracker: Cloud SQL export blocked for viewer IAM, a Dataproc fluentd leak that ran the master to OOM, billing accounts missing from Console.

Security Engineer & Data Analyst, Rakuten

San Mateo, CA | *Sep 2015 - Jul 2018*

- On-call DBA on a 24×7 rotation for 100+ Japan-region services, roughly 5,000 VMs. Authored the company MySQL and Oracle security standards from CIS benchmarks across thousands of instances, with continuous compliance-audit tooling behind them.
- Built an internal vulnerability-intelligence platform to collect, search, alert on, and visualize CVEs in real time. Sold it across subsidiaries — Ebates, Slice, Buy.com, Linkshare — alongside whitebox and blackbox penetration tests for the same brands.
- Led Rakuten's first PII/SPII sweep for GDPR across the full database landscape, coordinating with thousands of internal teams, much of it in Japanese.

EDUCATION

B.A., Mathematics and Computer Science (Courant Institute) · New York University · May 2015

CERTIFICATIONS

Certified Falcon Administrator (CCFA) · CrowdStrike · Apr 2025 - Apr 2028

Remote Pilot Certification (Part 107) · FAA · Issued Oct 2025

Advanced Open Water Diver · PADI · Issued Aug 2025

Private Pilot License · FAA · Issued Dec 2024

Continuous Monitoring Certification (GMON) · GIAC · Expired May 2026

TRAINING

Assessing and Exploiting Control Systems and IIoT · Black Hat · Apr 2024

Abusing and Protecting Kubernetes, Linux and Containers · Black Hat · Mar 2024

Reverse Engineering Firmware with Ghidra · Black Hat · Dec 2023

Machine Learning (inaugural) · Black Hat · Aug 2023

Reversing Signal with Software-Defined Radio · Black Hat · Aug 2023

Advanced Infrastructure Hacking · Black Hat · Mar 2023

ICS/SCADA Security Essentials · SANS · Dec 2022

A Crash Course in Practical Fast Forensics with a Red Teaming Perspective (Japanese) · Black Hat · Nov 2022

OPEN SOURCE & WRITING

PyPI `gaius-memory` · Ubuntu kernel bug LP#2150798 · cluster engineering writing at jkubo.com

SKILLS

Spoken languages: Japanese (native; business and technical) · English (native; business and technical)

Security: eBPF/BPF_LSM · Tetragon · detection engineering and validation · YARA · malware detonation and sandboxing (gVisor) · penetration testing · vulnerability management (NVD, CISA KEV, EPSS) · threat intelligence · osquery · Santa · CrowdStrike Falcon

AI systems: agents and tool calling · MCP servers · RAG and hybrid retrieval · LLM evaluation (LLM-as-judge) · LoRA fine-tuning · vLLM (ROCm and CUDA)

Cloud & infrastructure: GCP (BigQuery, Dataflow, Spanner, Cloud SQL, GKE, Marketplace) · AWS · Azure · Cloudflare · Kubernetes/K3s · Docker · Helm · Ansible · MySQL · PostgreSQL · OpenTelemetry/Grafana

Programming: Python · Go · Rust · JavaScript · Bash · SQL

Customer work: discovery with executives and practitioners · POC and evaluation design · demos and workshops · escalation management under SLA