Rui Shao

+86-15623713170 / +81-07090122378 sr1054461216@gmail.com github.com/ryan42xyz

About

Infrastructure engineer working on large-scale production systems on AWS (10+ regions) and Kubernetes (25+ clusters with 600+ nodes), with a focus on safe deployment, traffic control, observability, and cost-aware operations, increasingly leveraging AI-assisted agents to support decision-making and operational workflows.
Built platforms at the boundary of infrastructure and backend systems, combining automation and AI-driven analysis to make complex environments easier to operate, safer to change, and faster to recover under real production pressure.

Experience

Datavisor *Infrastructure Engineer* 2025/03 – Present (China, Japan)

Platform Control & Production Infrastructure

- Built backend control-plane services to manage **deployments, releases, traffic switching, and access governance** across multi-cloud, multi-cluster environments.
- Designed end-to-end change execution workflows with rollback and approval gates, reducing deployment time from **~30 minutes to ~5 minutes** while improving release reliability.
- Implemented cross-cluster traffic switching and failover supporting canary and blue–green releases, reducing incident recovery time by **~80%**.
- Operated and evolved production Kubernetes platforms on AWS, including cluster upgrades, node lifecycle management, and multi-AZ / multi-region architectures with defined RTO/RPO targets.

Observability, On-call & Reliability

Service Health, Incident Response

- Built and operated observability systems spanning infrastructure, middleware, and service-level signals across multiple Kubernetes clusters.
- Shifted monitoring from resource-centric dashboards to **service impact and SLO-driven views**, improving signal quality and reducing alert fatigue.
- Developed an **AI-assisted on-call investigation workflow** that aggregates anomalies, affected services/customers, dependencies, recent changes, and logs into a unified context, enabling faster human-led triage and informed incident response, capacity planning, and performance tuning.

Cost & Operational Efficiency

- Optimized cloud costs via selective use of Spot and Reserved Instances, achieving **20–40% compute cost reduction** without violating service SLOs.
- Reduced storage and data transfer costs through lifecycle policies, tiered storage, and topology-aware optimizations.

Intel *Cloud Software Development Engineer* 2022/06 – 2025/02 (China)

- Built core **distributed task execution and data-processing systems** on Kubernetes, forming a strong foundation in **concurrency control, state consistency, and failure handling** under parallel workloads.
- Engineered a **high-throughput gRPC streaming service**, using profiling-driven optimization (**pprof**), memory reuse, and async pipelines to cut latency from **200ms → 50ms** and significantly reduce GC and memory pressure.
- Designed **correctness-first scheduling and version update mechanisms** using transactional boundaries and distributed coordination, preventing race conditions and inconsistent intermediate states.

Skills

Programming Languages: Python, Golang, Java, Shell, SQL
Tools: AWS, GCP, Kubernetes, Docker, MySQL, YugaByte, Clickhouse, Redis, Kafka, Elasticsearch, Prometheus, InfluxDB, Grafana, Loki, gRPC, Helm, Git
Languages: Chinese, English

Education

Master, University of Science & Technology Beijing (Beijing, China) 2019/09 – 2022/06
Computer Science and Technology

Bachelor, Wenhua College (Wuhan, China) 2014/09 – 2018/06
Electronic Information Engineering