Minatomirai | Infrastructure Platform Engineer @ AI startup

募集職種

採用企業名
株式会社Unsung Fields

求人ID
1573183

業種
ソフトウエア

会社の種類
中小企業 (従業員300名以下)

雇用形態
正社員

勤務地
神奈川県, 横浜市西区

給与
800万円 ~ 1400万円

更新日
2026年03月19日 13:00

応募必要条件

職務経験
6年以上

キャリアレベル
中途経験者レベル

英語レベル
ビジネス会話レベル

日本語レベル
無し

最終学歴
大学卒： 学士号

現在のビザ
日本での就労許可は必要ありません

募集要項

Infrastructure Platform Engineer

Role overview

As an Infrastructure Engineer, you will own the **GPU platform** that runs production inference: cluster architecture, deployment reliability, observability, capacity management, and incident response mechanisms. Your job is to make the platform **predictable and reliable**—even as we scale hardware, models, tenants, and traffic patterns.

You'll work closely with serving/runtime and gateway teams to ensure the platform enforces the right isolation, exposes the right telemetry, and supports safe changes without downtime. This role blends strong systems intuition with real production discipline: reliable rollouts, clean operational tooling, and fast incident response.

Responsibilities

- **Own GPU cluster architecture and operations:** provisioning, node images, driver/runtime lifecycle, GPU plugin/operator lifecycle, and standardized deployment patterns for serving pools and system services.
- **Define and maintain the production baseline:** golden node configurations, cluster hardening, upgrade paths, and “known good” compatibility matrices (drivers ↔ CUDA ↔ runtime ↔ kernel).
- **Build reliability into the platform:** SLOs/SLIs, alerting quality, runbooks, incident tooling, and postmortems with real follow-through (automation, guardrails, and elimination of repeat incidents).
- **Enable safe delivery:** canary deploys, progressive rollouts, rollback paths, and configuration safety (validation, guardrails, change controls, and safe defaults).
- **Own fleet health and maintenance workflows:** node draining, GPU quarantining, automated remediation, scheduled maintenance, and safe “break-glass” procedures with auditability.
- **Capacity and utilization:** scheduling constraints, binpacking/fragmentation management, warm pools, autoscaling primitives, and quota enforcement hooks that align with product tiers and fairness goals.
- **Observability:** metrics/logs/tracing across gateway → serving → GPU; latency breakdowns, saturation signals, queue depth, GPU memory/compute metrics, and fleet health dashboards that help correlate customer symptoms to root causes.
- **Production readiness for heterogeneous environments:** manage differences across hardware generations and evolving server platforms, minimizing reliability risk while improving utilization.
- **Security baseline:** secrets management, least-privilege access, audit trails for operator actions, and secure operational workflows.
- **Partner with networking:** topology, failure domains, load balancing, and performance-sensitive traffic paths that impact tail latency and availability.
- **Build operational tooling:** fleet management, debugging workflows, safe admin actions, capacity tooling, and maintenance automation that reduces MTTR and improves operator efficiency.
- **Collaborate across teams:** align rollout plans, health semantics, capacity signals, and failure handling so the entire platform behaves predictably under load.

[Employment Type]
Full-time employee
*Probationary period: 3 months

[Salary]
Annual Salary: ¥8,000,000 - ¥14,000,000
Monthly Salary: ¥666,667 - ¥1,166,667 (Monthly Base Salary: ¥666,667 - ¥1,166,667)
■Salary Increases: Available

[Working Hours]
9:00 AM - 6:00 PM (60-minute break)

[Work Location]
Queen's Tower A, 10th Floor, 2-3-1 Minatomirai, Nishi-ku, Yokohama, Kanagawa Prefecture, 220-6010
■Access: 7-minute walk from Sakuragicho Station (all lines), direct access from Minatomirai Station (Toyoko Line, Minatomirai Line)
■Non-smoking workplace
■Changes to work location: Company-designated offices
■Transfers/Secondments: None

[Holidays and Leave]

120 days off per year Days
Full two-day weekend
Annual paid vacation (minimum 10 days after the seventh month of employment)

[Benefits]
Partial transportation allowance (up to ¥15,000 per month)https://www.careercross.com/login
Social insurance (health insurance, employee pension insurance, employment insurance, workers' compensation insurance)
Overtime pay: Standard overtime pay

スキル・資格

Requirements

- 5+ years in infrastructure/SRE/platform engineering for production distributed systems.
- Strong Kubernetes experience in production (or equivalent orchestration), with real ops ownership.
- Experience operating GPU clusters or other high-performance compute fleets (or similarly performance-sensitive infrastructure).
- Strong debugging skills across Linux, networking, and distributed systems failure modes.
- Strong operational discipline: automation-first mindset, measurable reliability, careful change management, clear communication during incidents.
- Willing to participate in an on-call rotation for owned systems.

Nice to have

- Experience with high-throughput gateways/service meshes (e.g., Envoy), **OpenTelemetry**, and multi-region architectures.
- Experience with Slurm/HPC-style scheduling, **RDMA**/**IB**, or performance-sensitive networking.
- Experience building internal developer platforms and “golden paths” for consistent deploy/rollback workflows.
- Experience managing GPU driver/runtime upgrades safely across a fleet (compatibility testing + staged rollouts).

- Familiarity with observability patterns for latency-sensitive systems (request correlation, sampling strategy, high-cardinality metrics control).

会社説明