Backend Engineer (Routing & Token) @ AI startup

募集職種

採用企業名
株式会社Unsung Fields

求人ID
1575704

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

最寄駅
みなとみらい線、　みなとみらい駅

給与
800万円 ~ 1400万円

勤務時間
09:00 - 18:00（60-minute break）

休日・休暇
Two-day weekends,holidays,special leaves,120+ days off annually

更新日
2026年02月04日 14:07

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

Backend Engineer (Routing & Token)

Role overview

As a Backend Engineer, you will be responsible for the control-plane and gateway layer that connects the customer to our compute serving infrastructure. You will build and power the customer request lifecycle end to end, including handling

requests via Cloudflare worker nodes, managing authentication and tenancy, validating requests, routing to the correct model endpoints, enforcing quotas and rate limits, and implementing reliability mechanisms to ensure platform stability under load. This layer is also key to achieving significant service performance latency through optimization. The goal is to build a backend that is fast, secure by default, abuse-resistant, and highly operable at scale.

You will collaborate with the inference and platform teams on developing the backend architecture. We expect you to be a central, hands-on contributor to the code stack, driving both the building and technical decision-making as an expert from the ground up. You will work closely with the President and engineering leadership on backend routing decisions that accurately reflect real capacity and failure domains, and ensure the system provides the necessary telemetry for rapid debugging.

Responsibilities

- **Architect Intelligent Routing Logic:** Design and implement a dynamic "Intelligent Router" that uses real-time metrics and ML-based scoring to select the optimal GPU Pool for every request. You will ensure efficient GPU utilization and prevent SLA violations by routing traffic based on node health and congestion.
- **Implement Model-Based Parsing:** Build logic within the Edge Gateway to parse request bodies, identify model parameters, and execute "Model-Based Routing" to direct specific workloads to the appropriate specialized GPU clusters.
- **Build Distributed Caching Systems:** Engineer a "Distributed KV Cache" strategy that allows GPU pools to share KV caches, significantly reducing duplicate calculations and improving inference speed. You will also manage local edge caches for rate limiting and quota enforcement.
- **Optimize Edge Gateway & Security:** Leverage Cloudflare Anycast to accept user requests at the edge location nearest to them, solving distance-based latency issues. You will implement security layers to filter malicious requests at the gateway, ensuring they never reach the origin servers.
- **High-Throughput Stream Management:** Optimize the Token Flow using Cloudflare Network Interconnect, ensuring that inference streams are delivered with minimal jitter and latency (TTFT).
- **Protocol & Traffic Engineering:** Fine-tune the communication protocols between the Edge and the DC (GPU Pool), handling connection pooling and keep-alives to sustain high throughput.

[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)
Social insurance (health insurance, employee pension insurance, employment insurance, workers' compensation insurance)
Overtime pay: Standard overtime pay

スキル・資格

You may be a fit if you have the following skills:

- **Advanced Edge Engineering:** Expert-level TypeScript/JavaScript or Rust/WASM experience specifically within Cloudflare Workers environments. You understand V8 isolates and edge runtime limitation
- **Complex Routing Algorithms:** Proven experience building custom load balancers or routing logic You can translate "ML-based scoring" into performant edge code.
- **Network Protocol Technical Expertise:** Strong grasp of HTTP/2, HTTP/3, WebSockets, and Anycast networking principles to minimize latency.
- **Distributed Systems Caching:** Experience designing distributed caching mechanisms (Redis, KV stores) where consistency and hit rates are critical for performance.
- **ML/Inference Knowledge:** Understanding of how LLM KV caches work and how model parameters impact compute requirements.
- **Security Engineering:** Experience implementing WAF rules or DDoS mitigation logic at the edge

会社説明