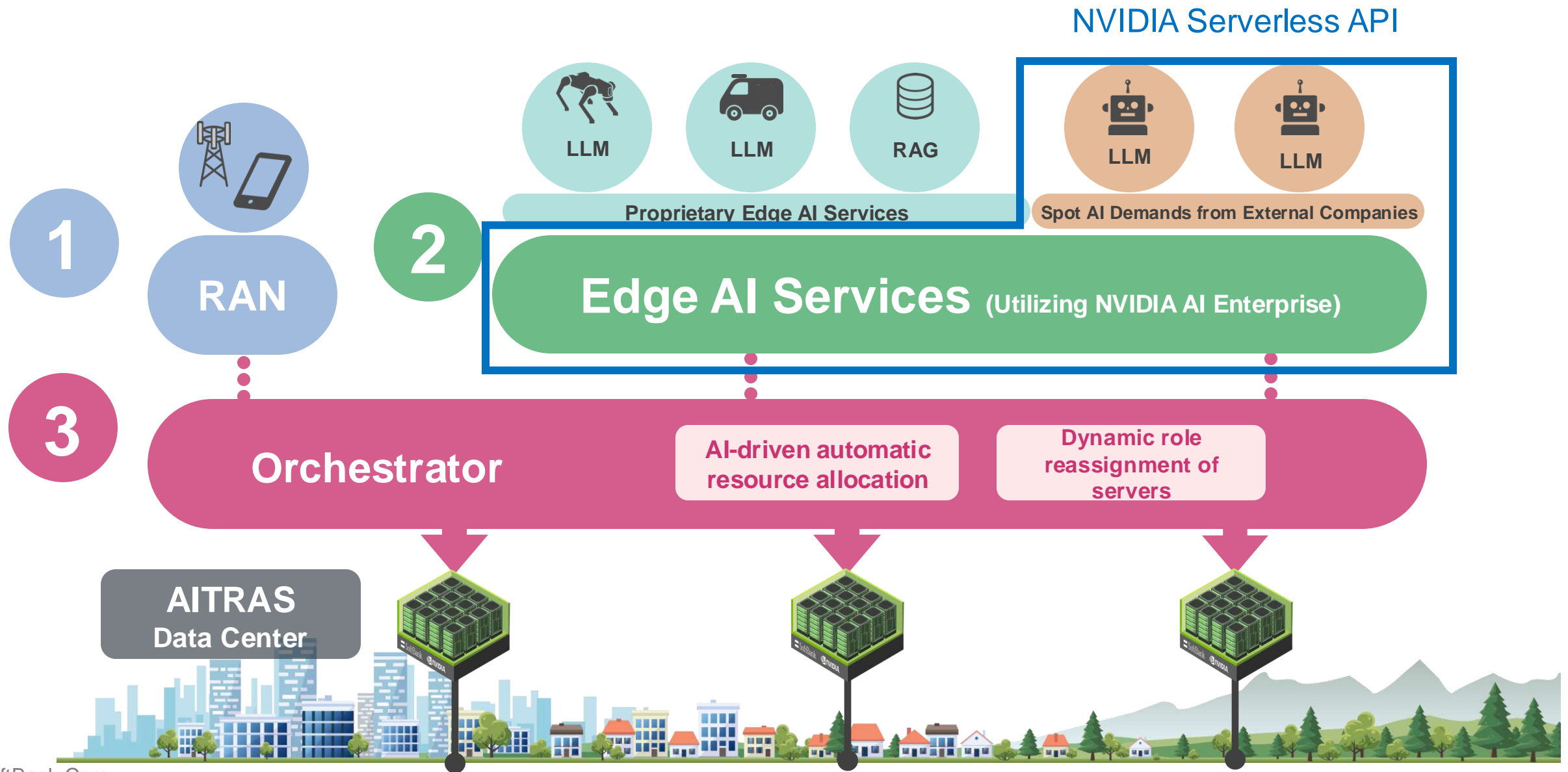


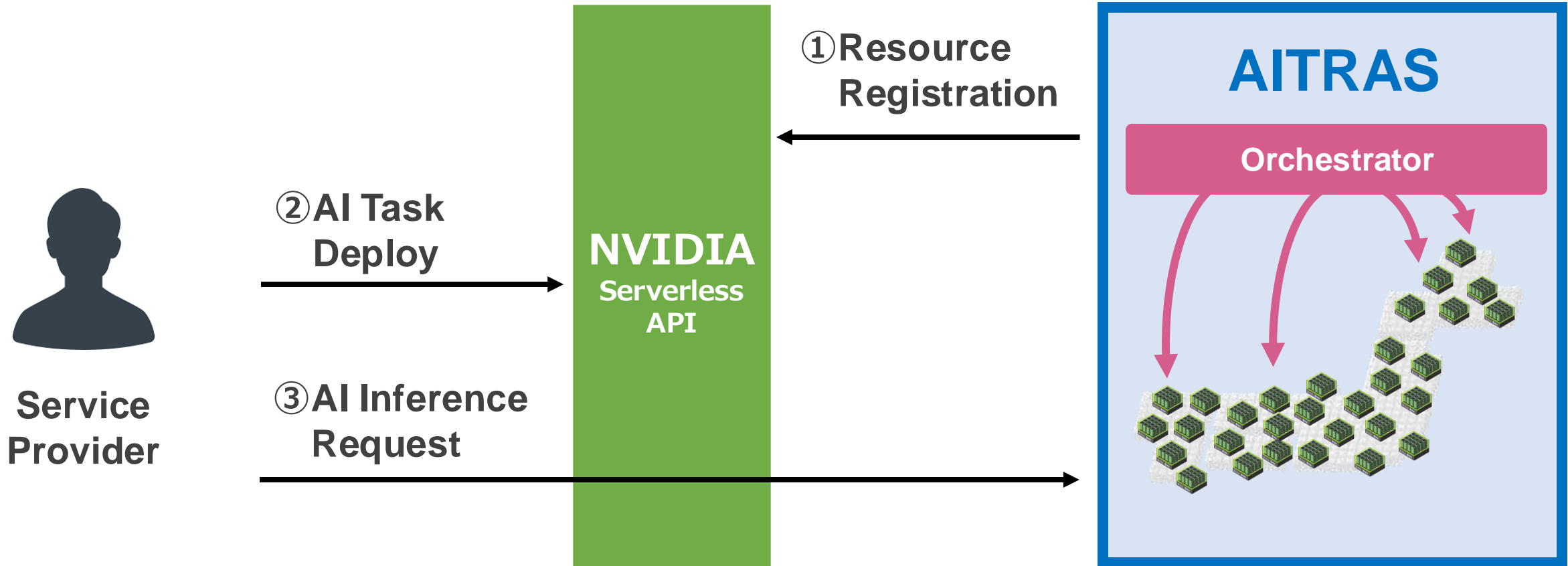
Utilization of Idle Resources by AITRAS

~Leveraging Serverless API powered by NVIDIA AI Enterprise~

Key Technologies of AITRAS



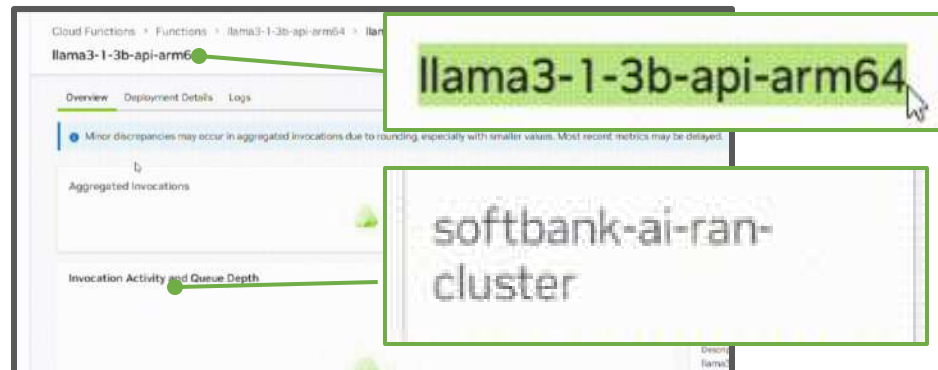
AITRAS × NVIDIA Serverless API



Providing Idle Resources Managed by AITRAS Across Japan to Customers via NVIDIA Serverless API

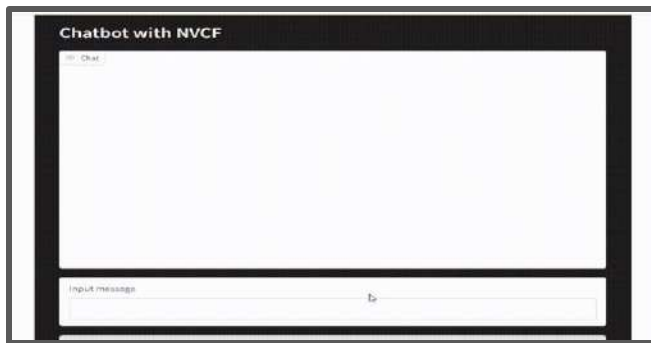
Demonstration

② AI Task Deploy



コンソール画面

③ AI Inference Request and Execution

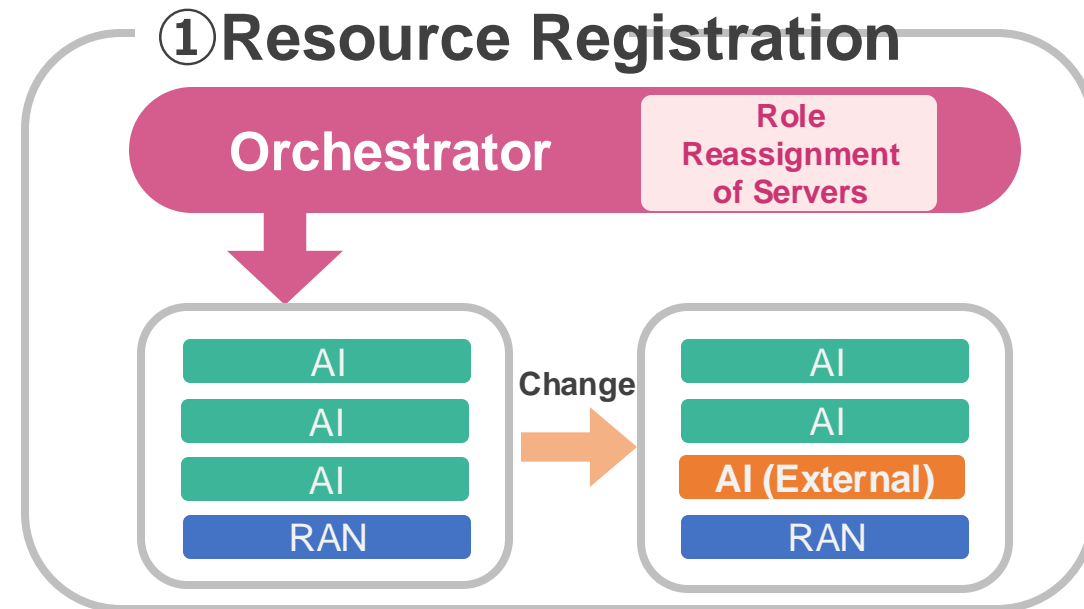


Chatbot Application

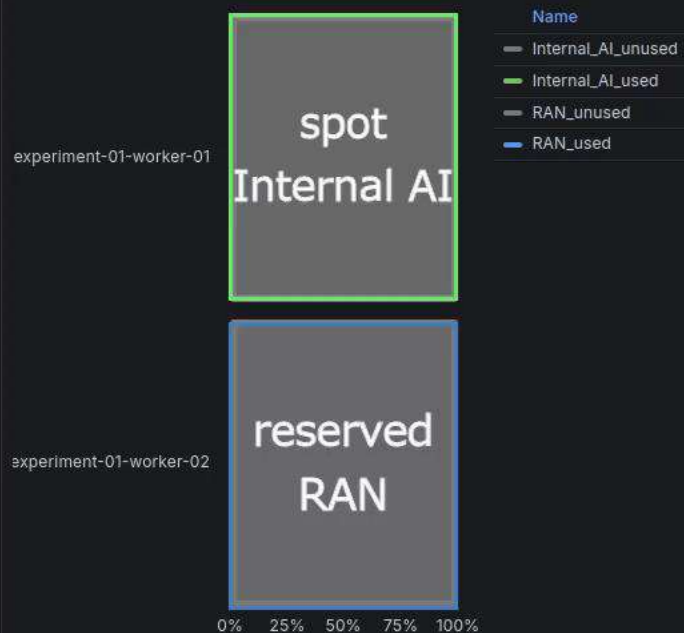
Chatbot
powered by
llama3

NVIDIA
Serverless
API

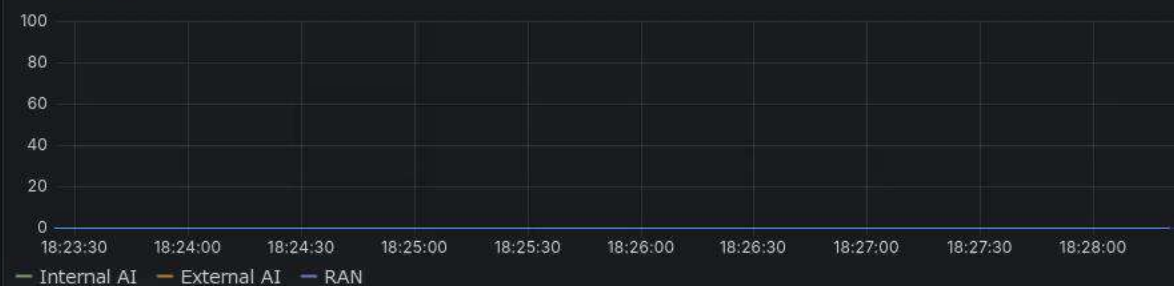
① Resource Registration



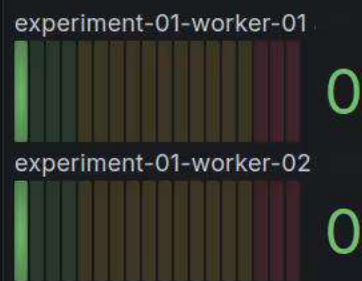
GPU Allocation



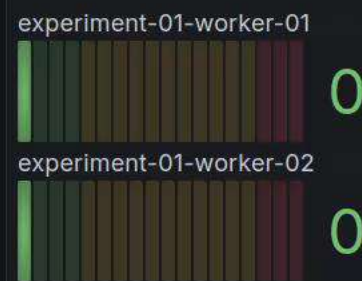
Cluster Total GPU Utilization



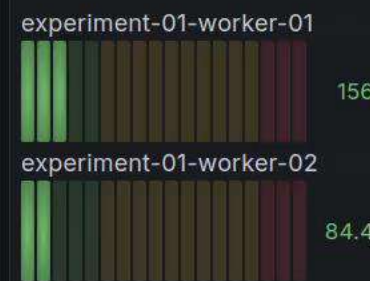
Node GPU Utilization



Node GPU Utilization (Tensor Core)



Node GPU Power Usage



Orchestrator Logs

```
> 2024-11-05 09:28:24.000 GET /api/logs 200 2.3
> 2024-11-05 09:28:24.000 GET /metrics 200 2.31
> 2024-11-05 09:28:24.000 external prometheus d
> 2024-11-05 09:28:23.000 external metrics dura
> 2024-11-05 09:28:21.000 POST /api/cluster-inf
> 2024-11-05 09:28:21.000 pod and node info upd
> 2024-11-05 09:28:21.000 current updated cache
> 2024-11-05 09:28:21.000 current updated cache
> 2024-11-05 09:28:19.000 pod and node info agg
> 2024-11-05 09:28:18.000 GET /metrics 200 2.53
> 2024-11-05 09:28:18.000 external prometheus d
> 2024-11-05 09:28:17.000 external metrics dura
> 2024-11-05 09:28:14.000 GET /api/logs 200 0.0
> 2024-11-05 09:28:12.000 GET /api/logs 200 0.0
> 2024-11-05 09:28:12.000 GET /metrics 200 2.84
> 2024-11-05 09:28:12.000 external prometheus d
> 2024-11-05 09:28:12.000 external metrics dura
> 2024-11-05 09:28:10.000 GET /api/logs 200 3.4
> 2024-11-05 09:28:10.000 POST /api/cluster-inf
> 2024-11-05 09:28:10.000 pod and node info upd
> 2024-11-05 09:28:10.000 current updated cache
> 2024-11-05 09:28:10.000 current updated cache
> 2024-11-05 09:28:07.000 pod and node info agg
> 2024-11-05 09:28:06.000 GET /metrics 200 2.55
```

> その他環境 (1 panel)

Orchestrator can absorb external AI demand when there is a resource surplus.



— SoftBank