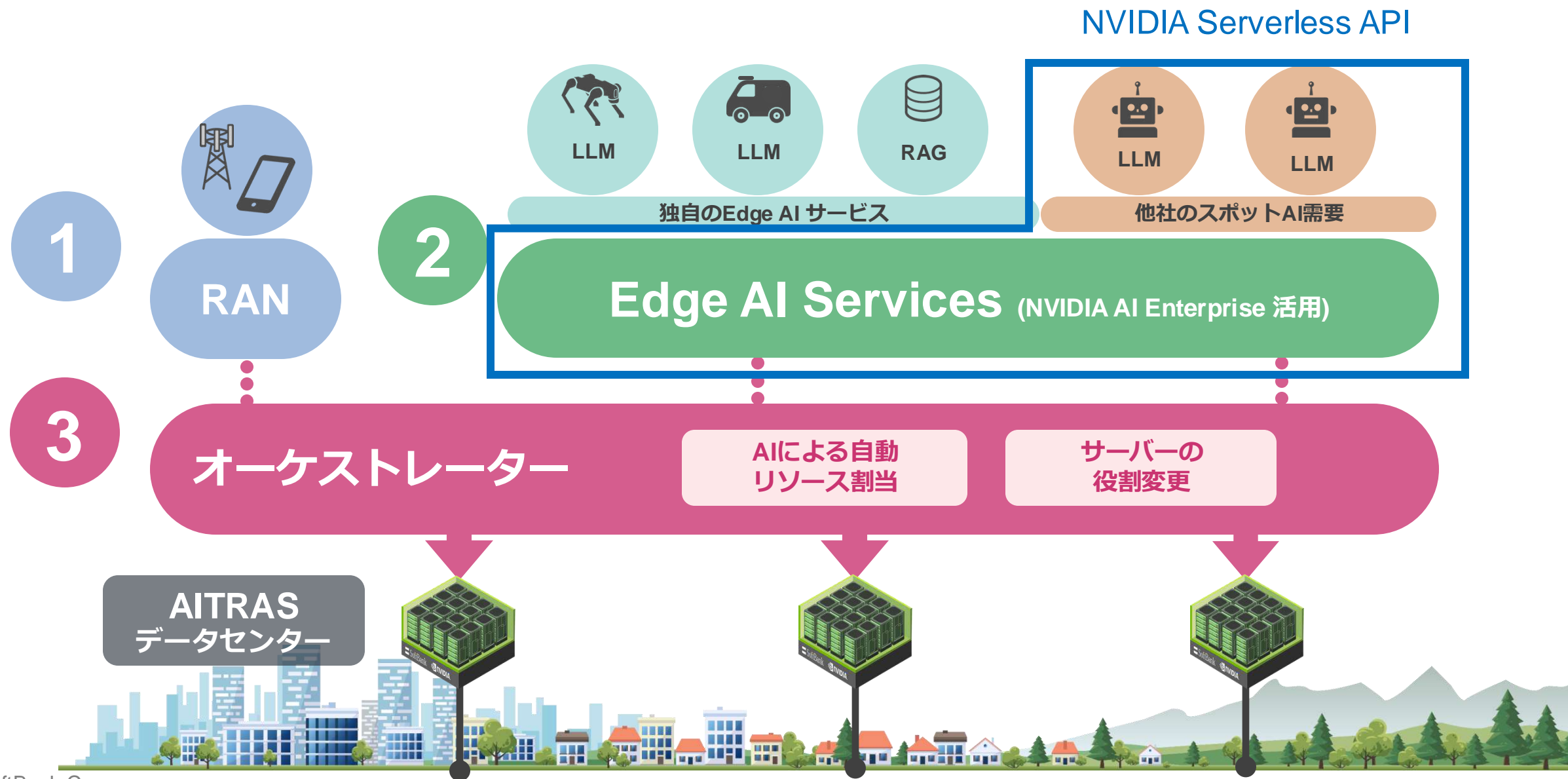


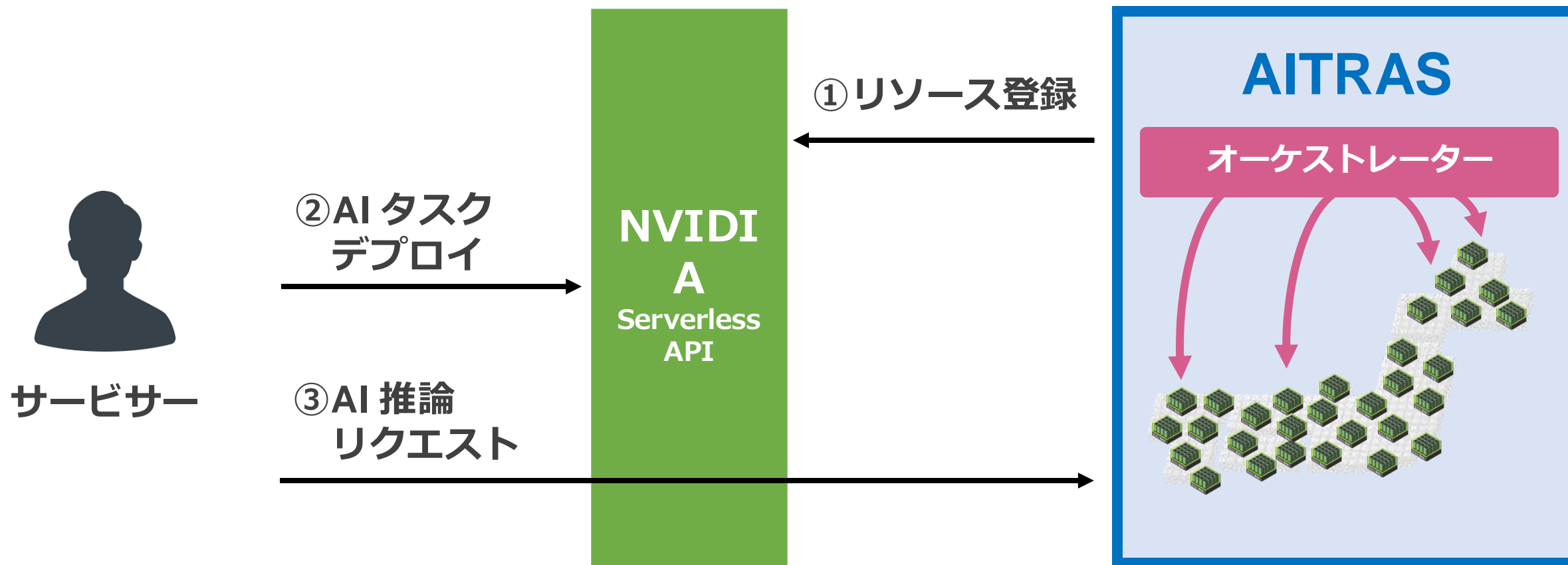
AITRASの余剰リソース活用

～Serverless API powered by NVIDIA AI Enterpriseの活用～

AITRASの要素技術



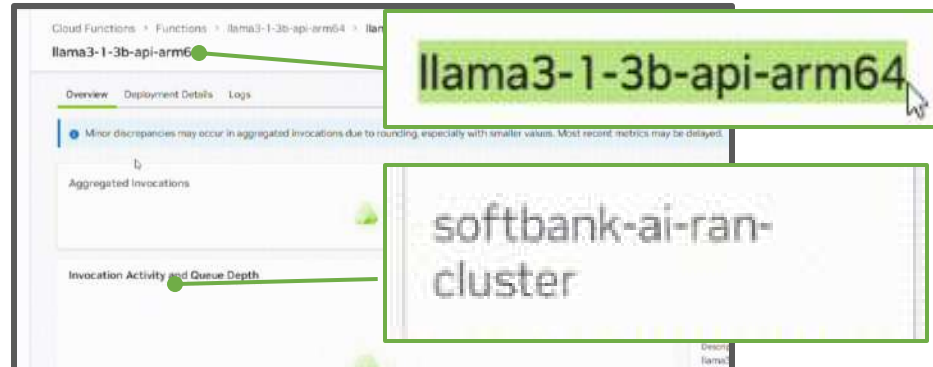
AITRAS × NVIDIA Serverless API



**AITRAS が管理する日本各地の空きリソースを
NVIDIA Serverless API で顧客に提供**

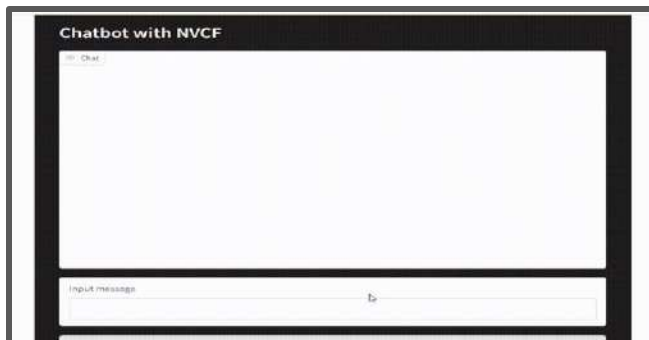
デモンストレーション

② AI タスクデプロイ



コンソール画面

③ AI 推論リクエスト・実行



チャットボット アプリ

llama3を
使った
チャットボット

NVIDIA
Serverless
API

① リソース登録

オーケストレーター

サーバーの
役割変更

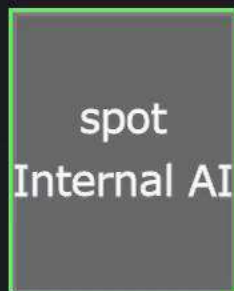
AI
AI
AI
RAN

変更

AI
AI
AI (External)
RAN

GPU Allocation

experiment-01-worker-01



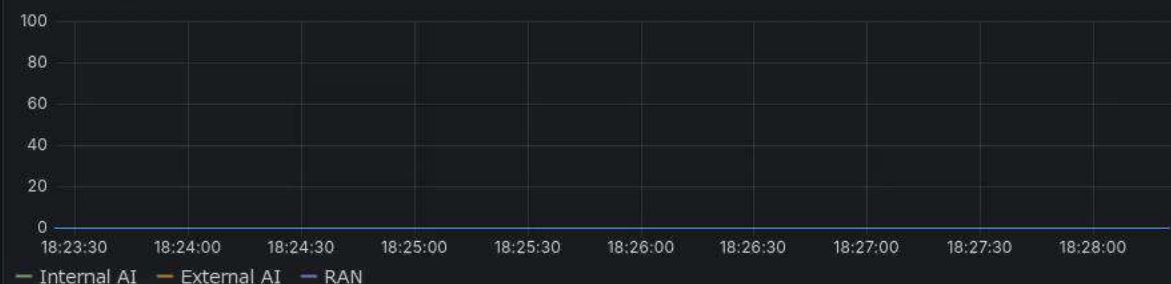
experiment-01-worker-02



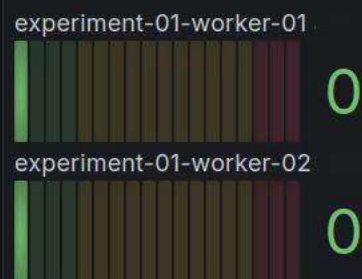
Name

- Internal_AI_unused
- Internal_AI_used
- RAN_unused
- RAN_used

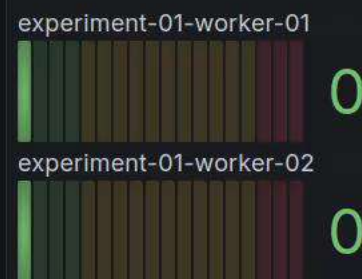
Cluster Total GPU Utilization



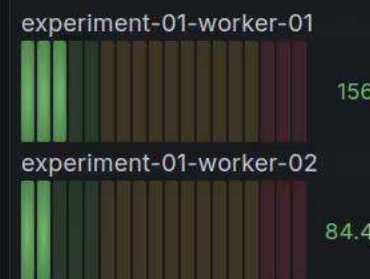
Node GPU Utilization



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)

オーケストレーターはリソースの余剰が存在するタイミングで
外部のAI需要を取り込むことができます。

 SoftBank