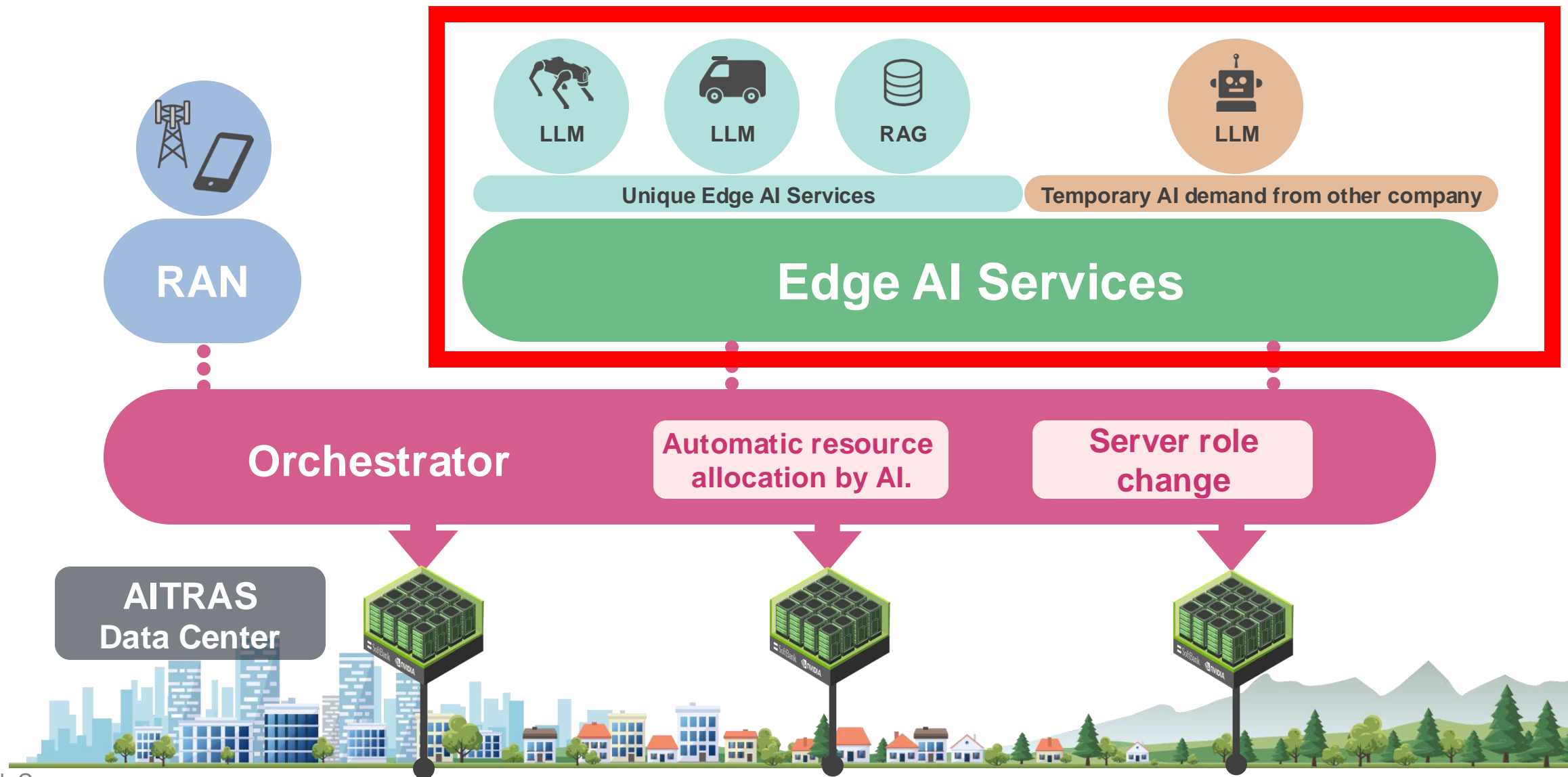
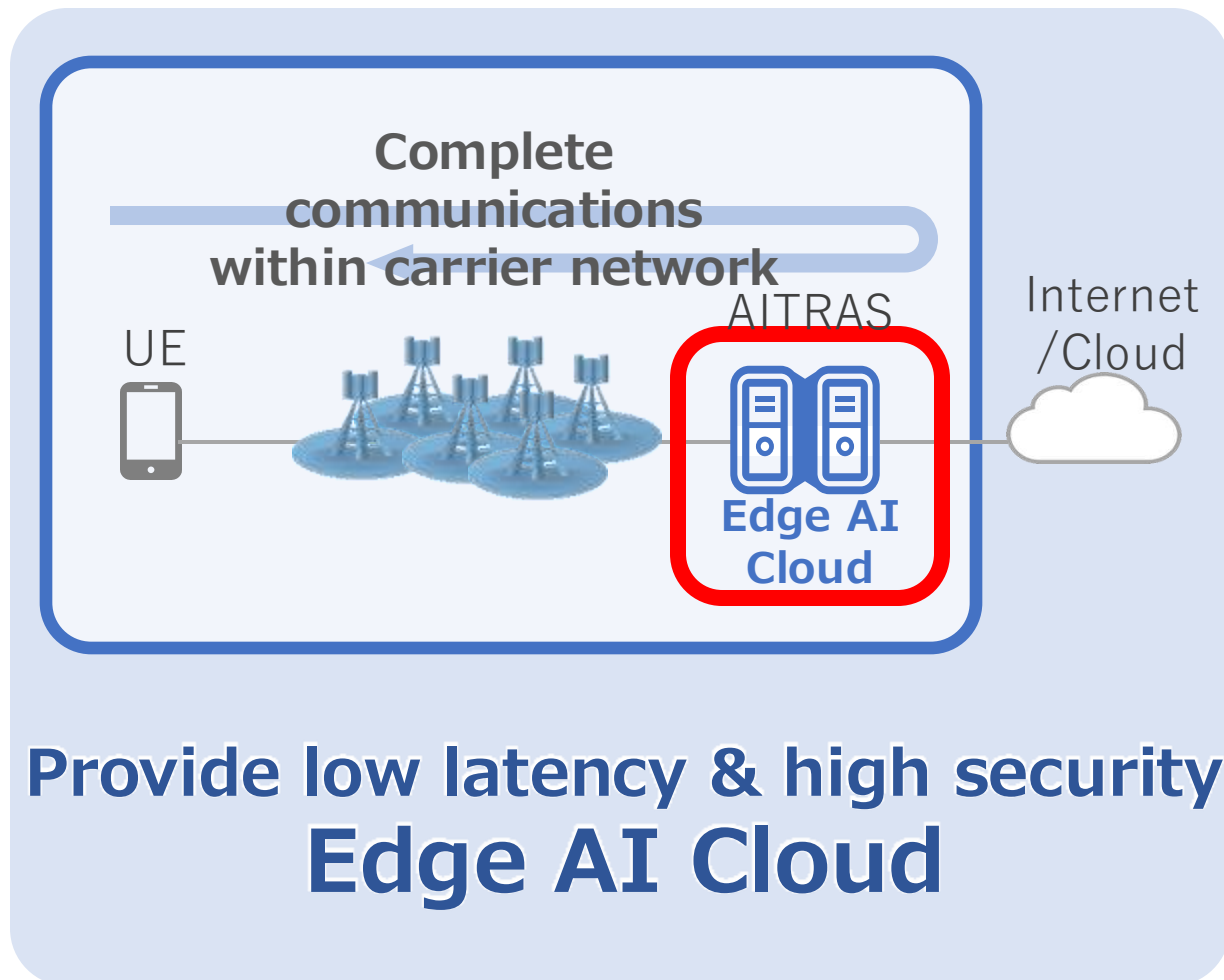


Technologies Elements of AITRAS

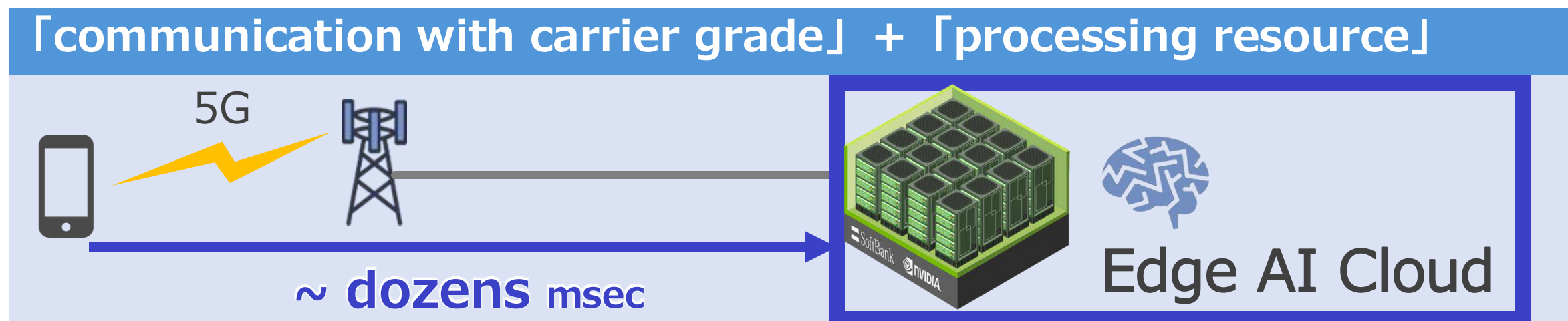
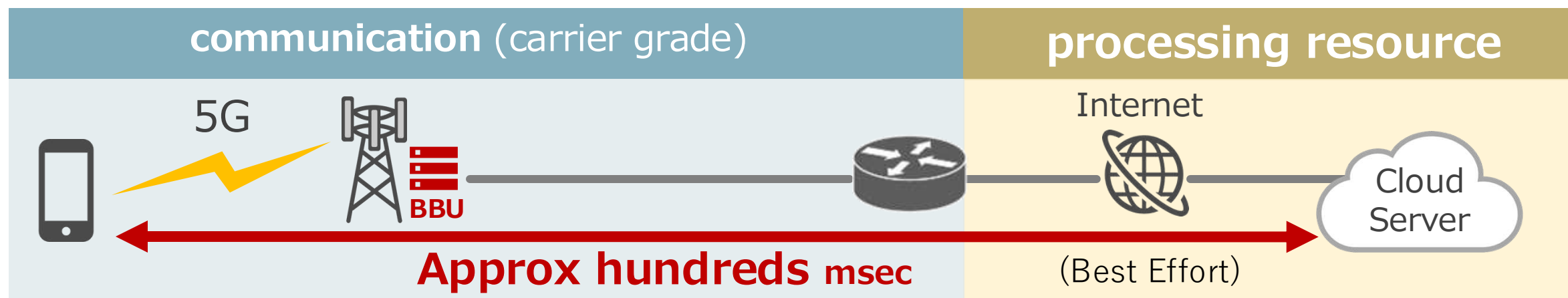


AITRAS Edge AI Features

~Supporting the realization of innovative AI services~

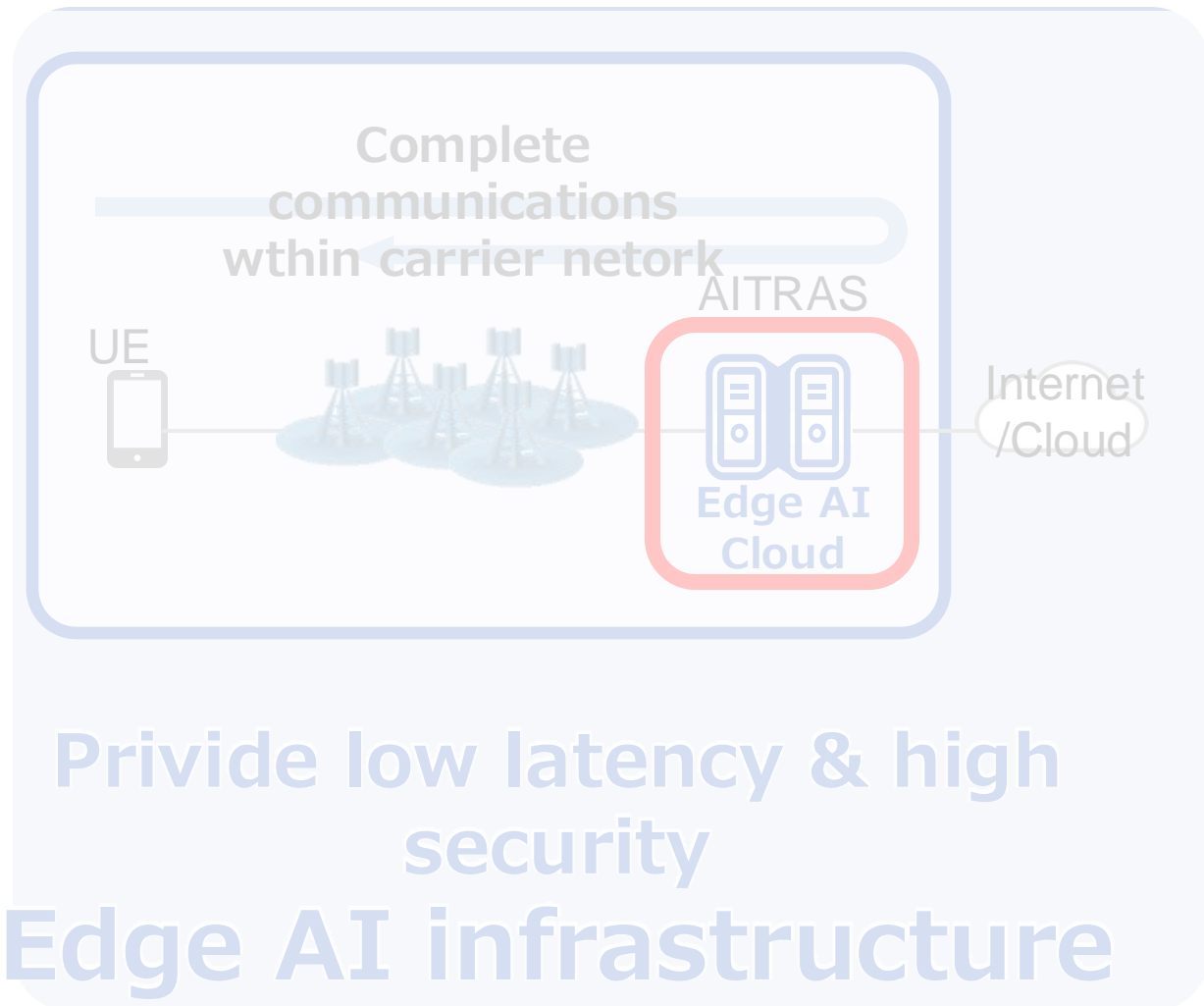


Low Latency and Highly Secure Edge AI Cloud made possible by a telecom operator



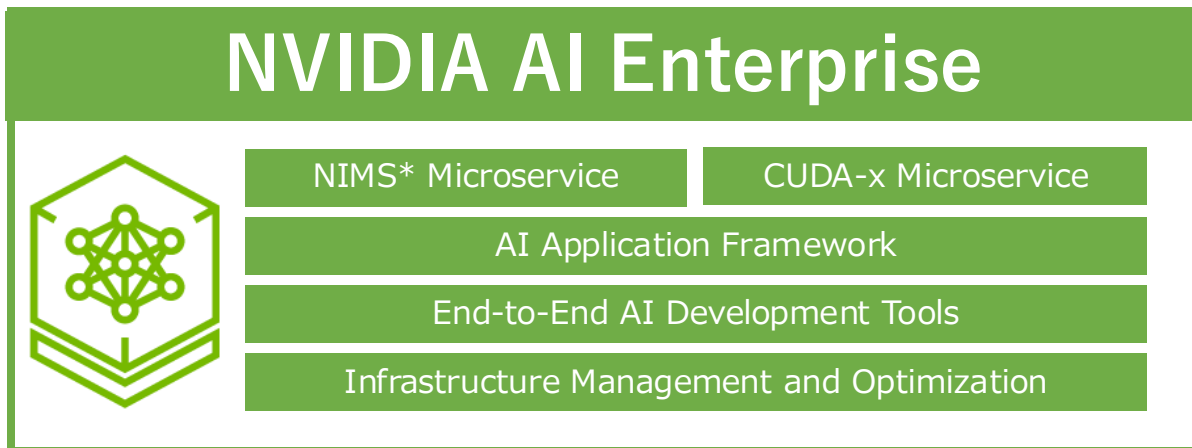
AITRAS's Edge AI Features

-Enabling Innovative AI Services-



Enabling Rapid Development with NVIDIA AI Enterprise

Business logic, company data



Edge AI Cloud



**AITRAS enables rapid
AI service development**

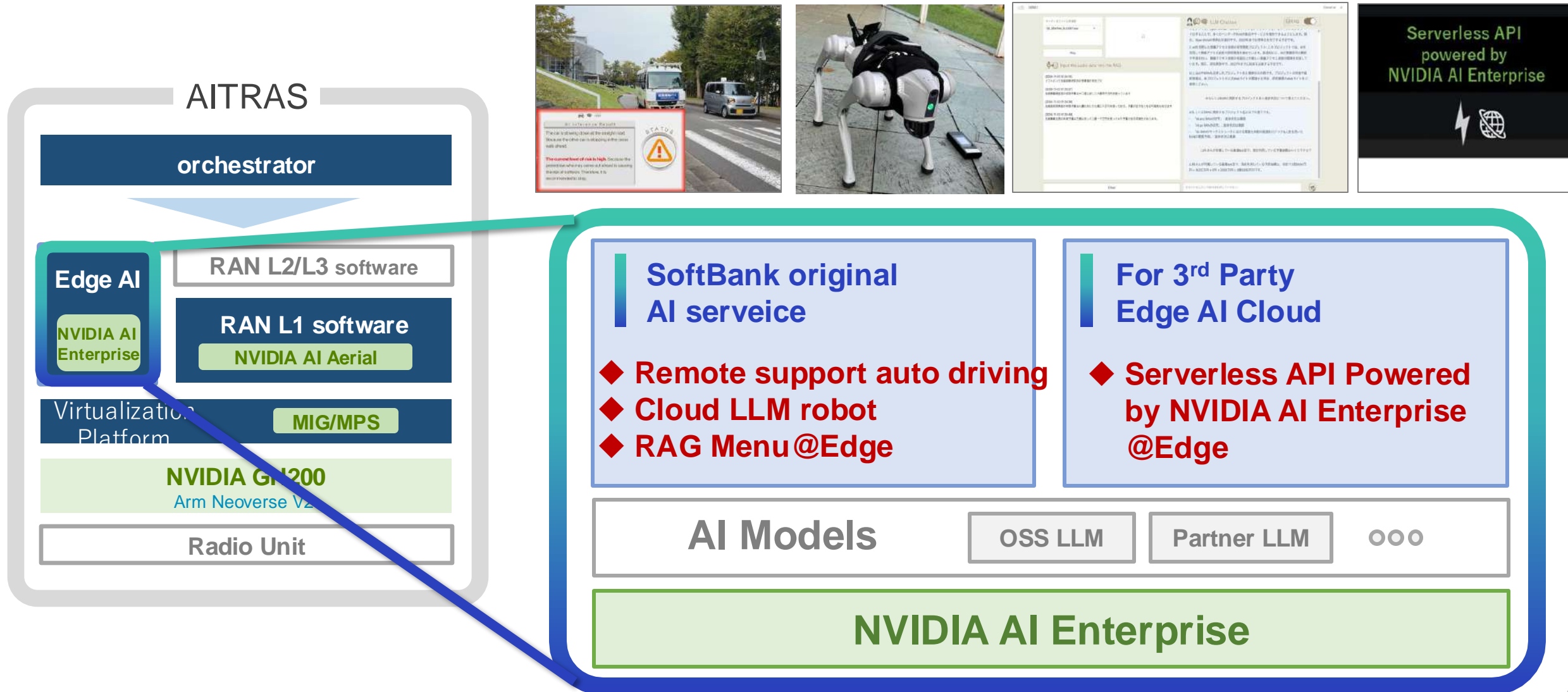


**Collaboration with NVIDIA AI
Enterprise
Functions that powerfully support AI
development**

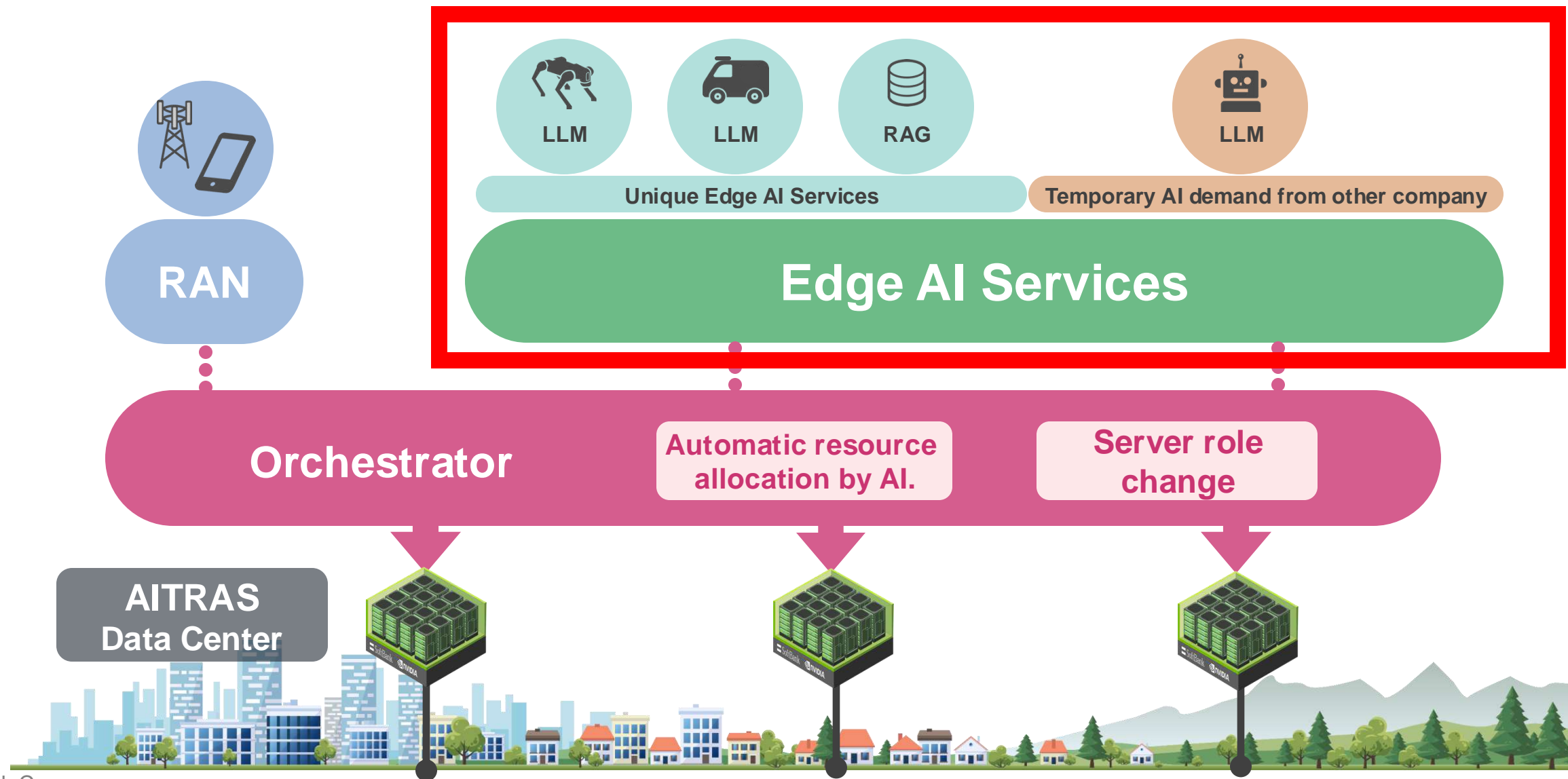


**Low latency & high
security
Edge AI Cloud**

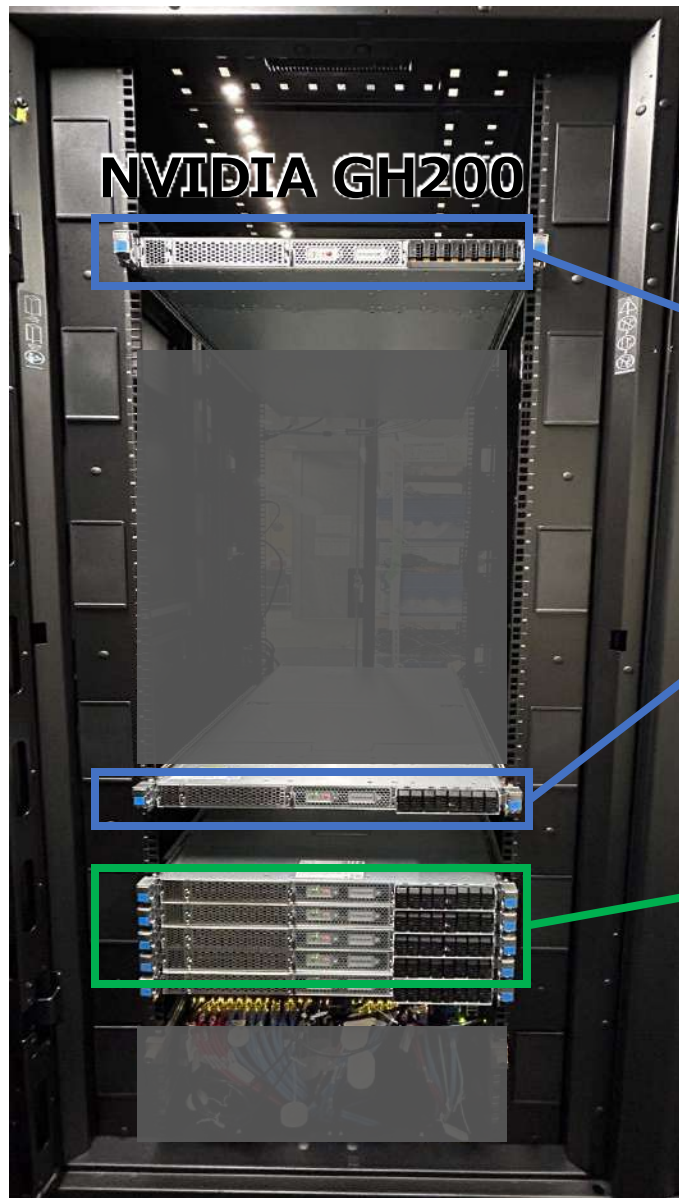
AITRAS Edge AI Use Cases



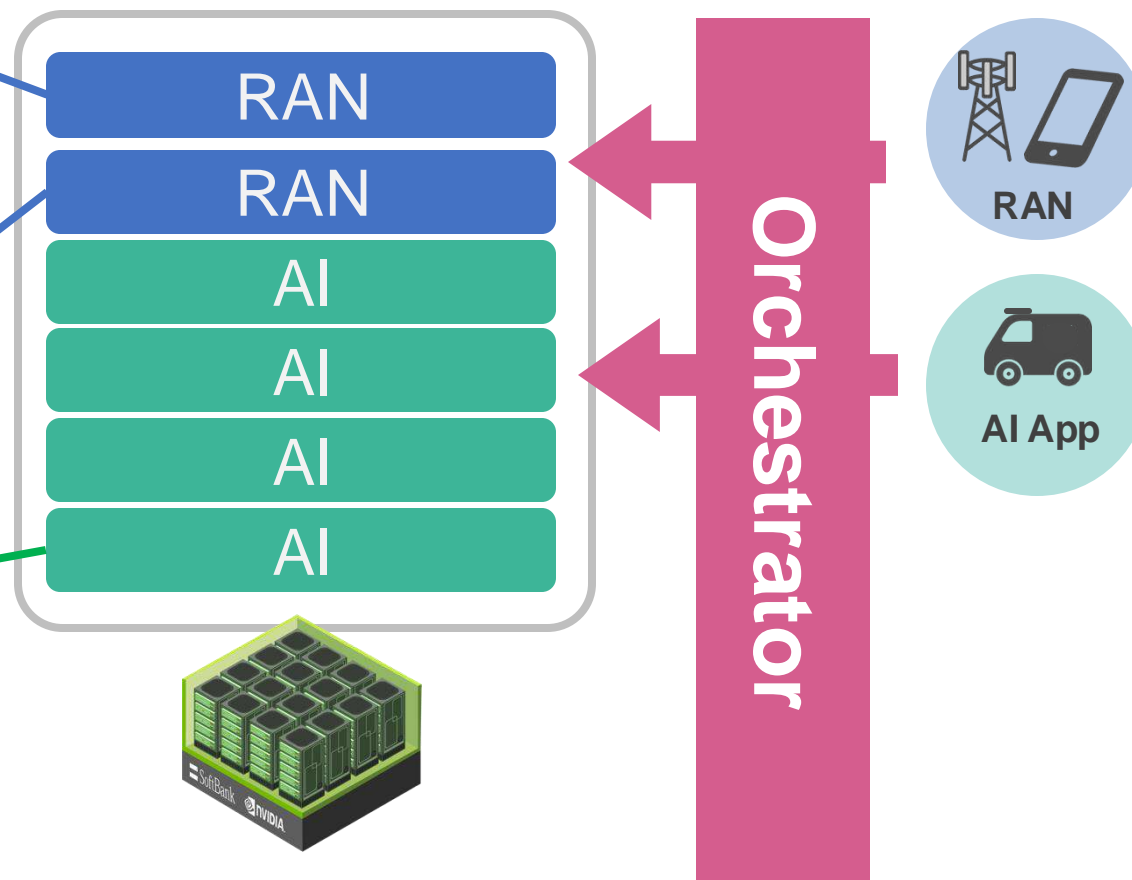
Technology Elements of AITRAS



AITRAS Orchestrator Basic Functions

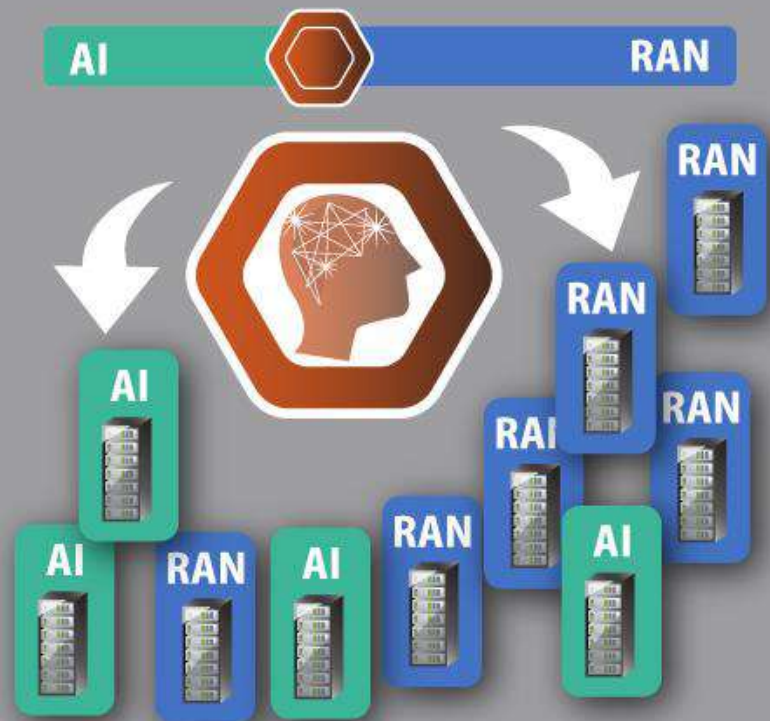


- ✓ Switching between RAN and AI roles for each server
- ✓ Deploy applications according to the assigned role.

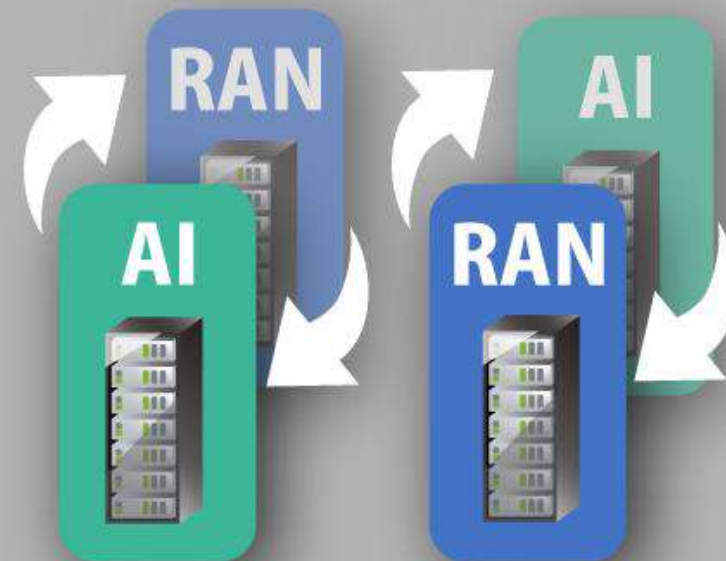


Two Important Functions of the AITRAS Orchestrator

Allocate the demands of AI and RAN to the optimal resources

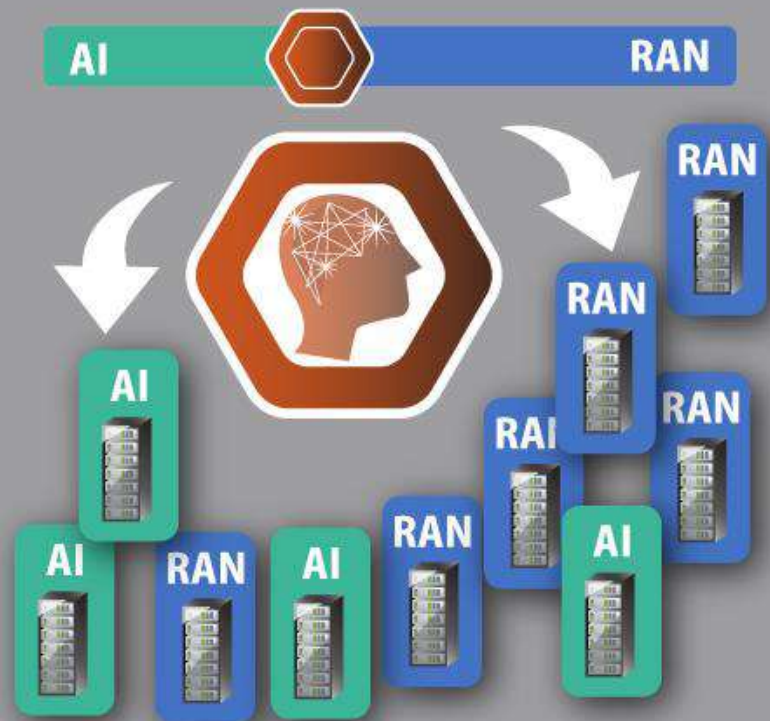


Dynamic change of server roles

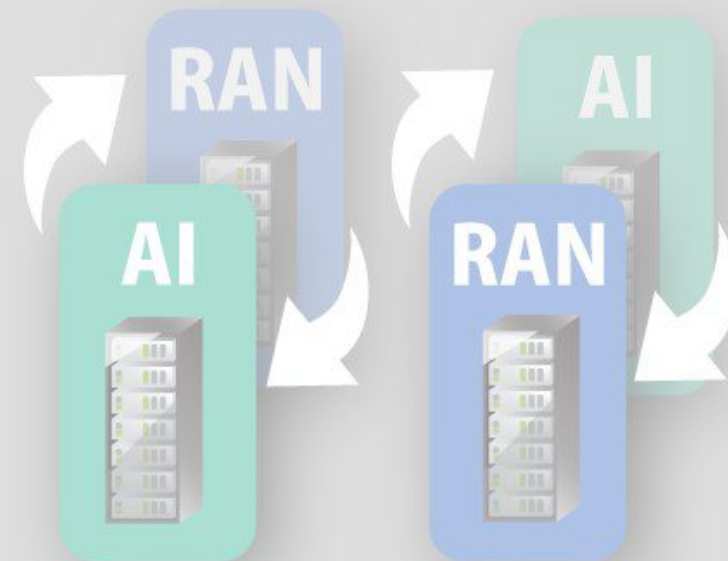


Two Important Functions of the AITRAS Orchestrator

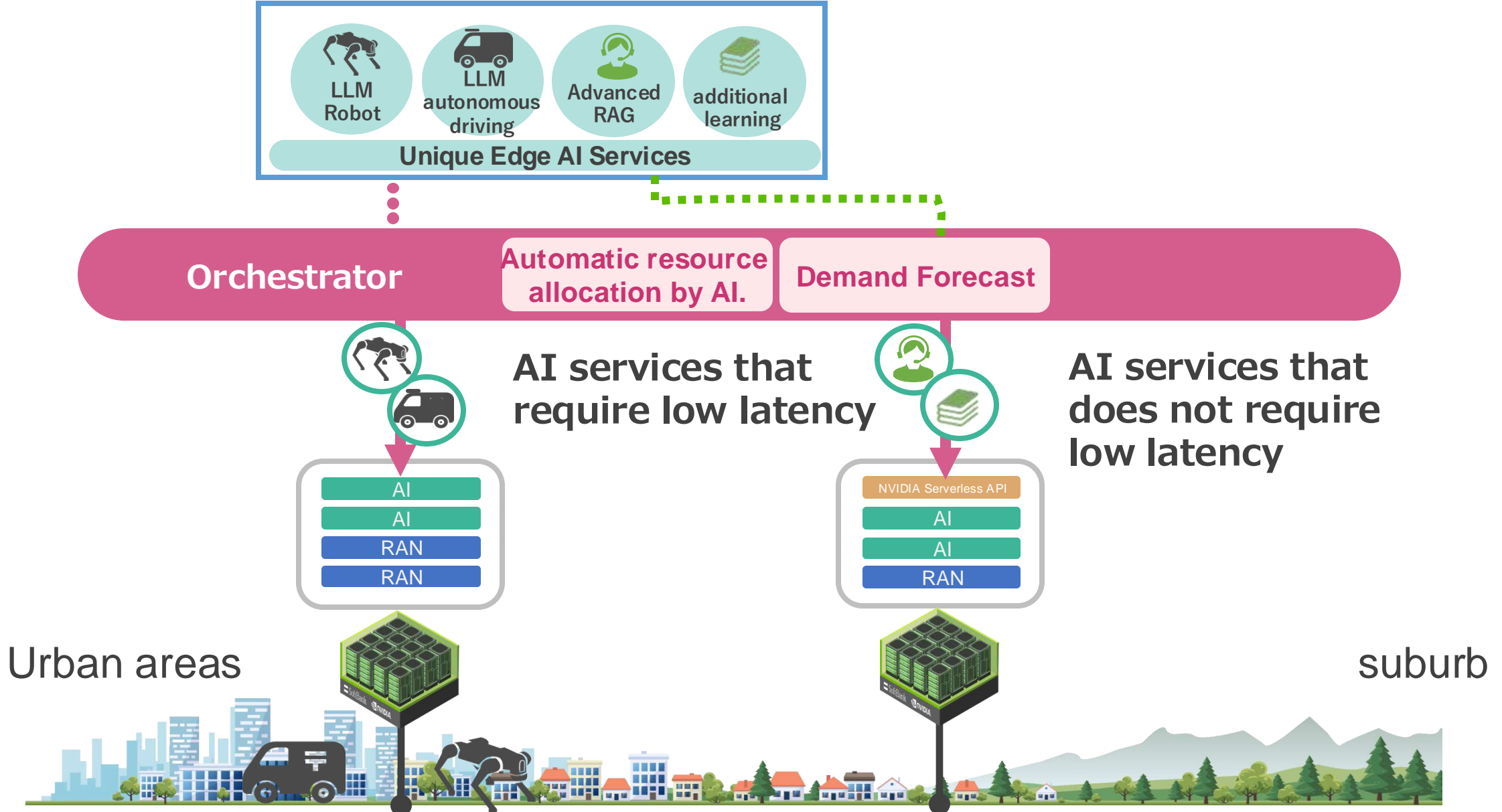
Allocate the demands of AI and RAN to the optimal resources



Dynamic change of server roles

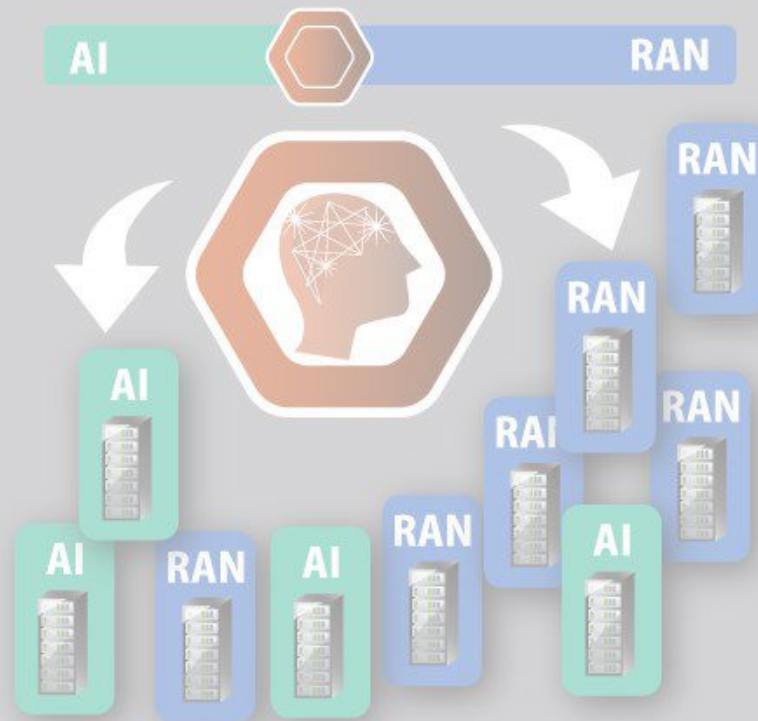


Assigning Services According to Use

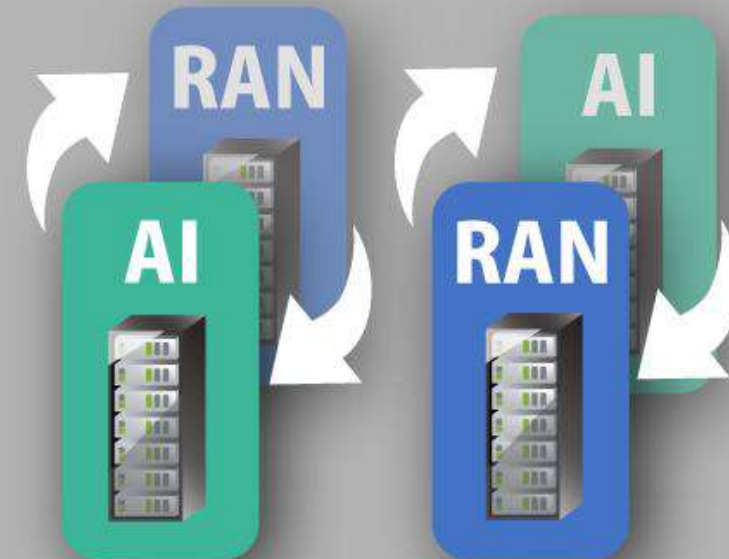


Two Important Functions of the AITRAS Orchestrator

Allocate the demands of AI and RAN to the optimal resources



Dynamic change of server roles

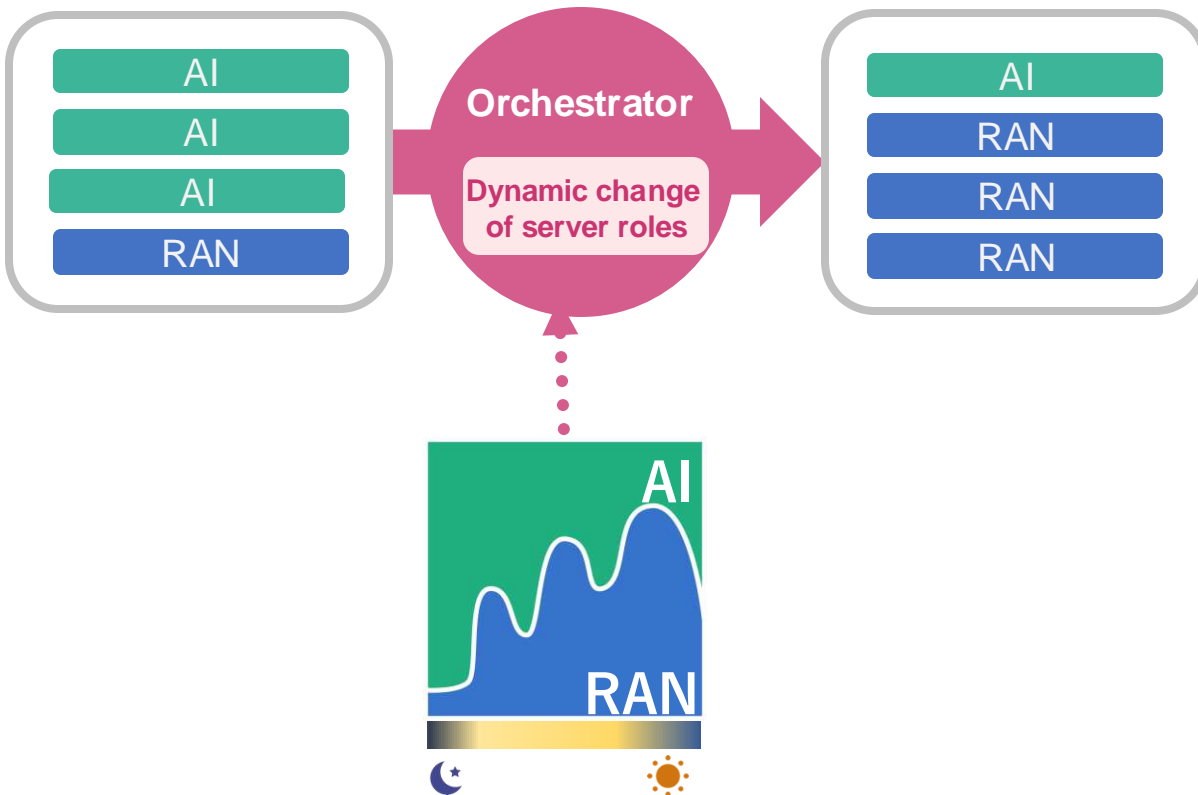


Optimally change the role of the server

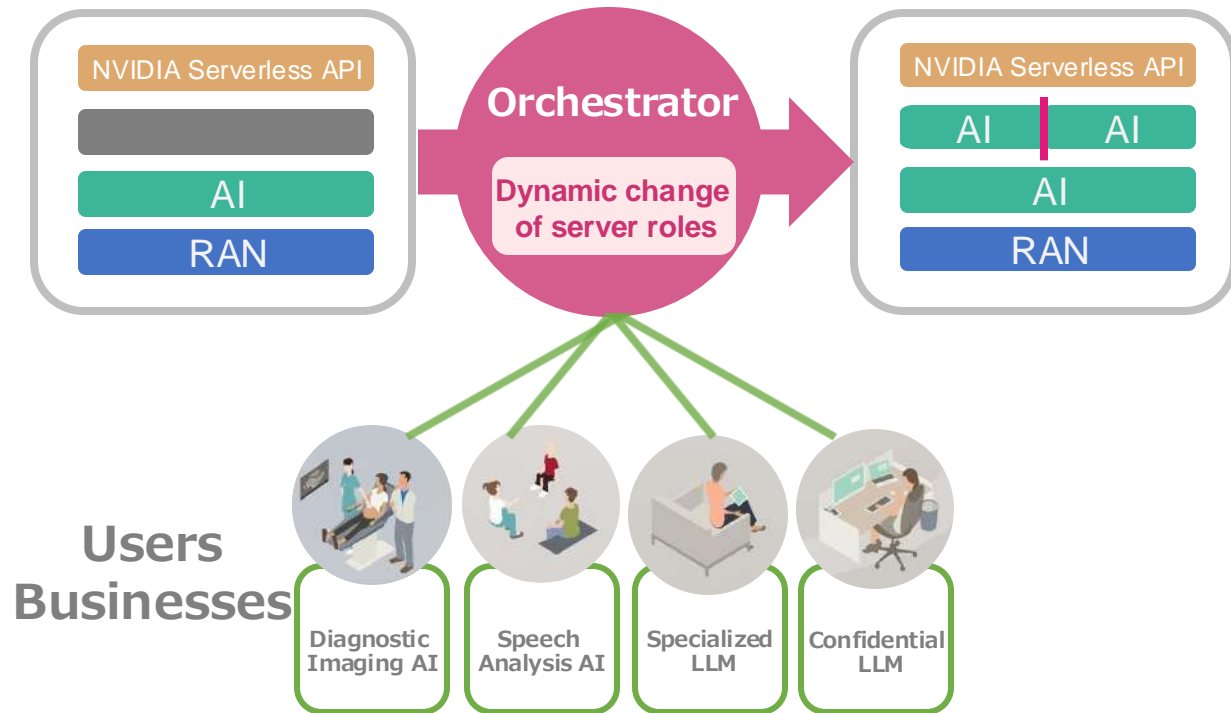
Change the server role according to
RAN's demand
(AI \leftrightarrow RAN)

4 : 00

12 : 00



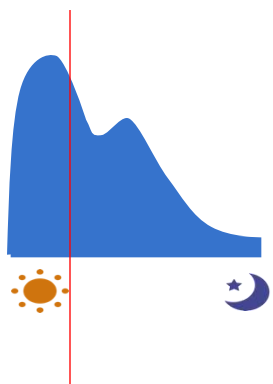
Split GPUs on demand
(Multi-Instance GPU:MIG)



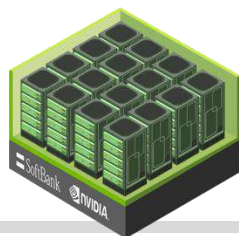
Demo

Demo Scenario

RAN has higher demand during the daytime.



AI
RAN
RAN
RAN



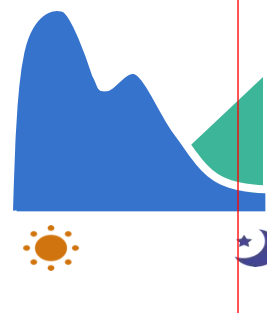
**Dynamic change
of server roles**

**Automatic resource
allocation by AI**

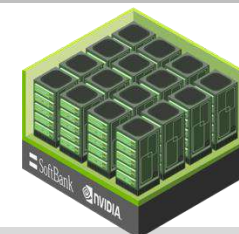


**As RAN demand decreases
during late-night hours, servers
are reassigned to appropriate
roles to run AI apps.**

4 : 00



AI
AI
AI
RAN

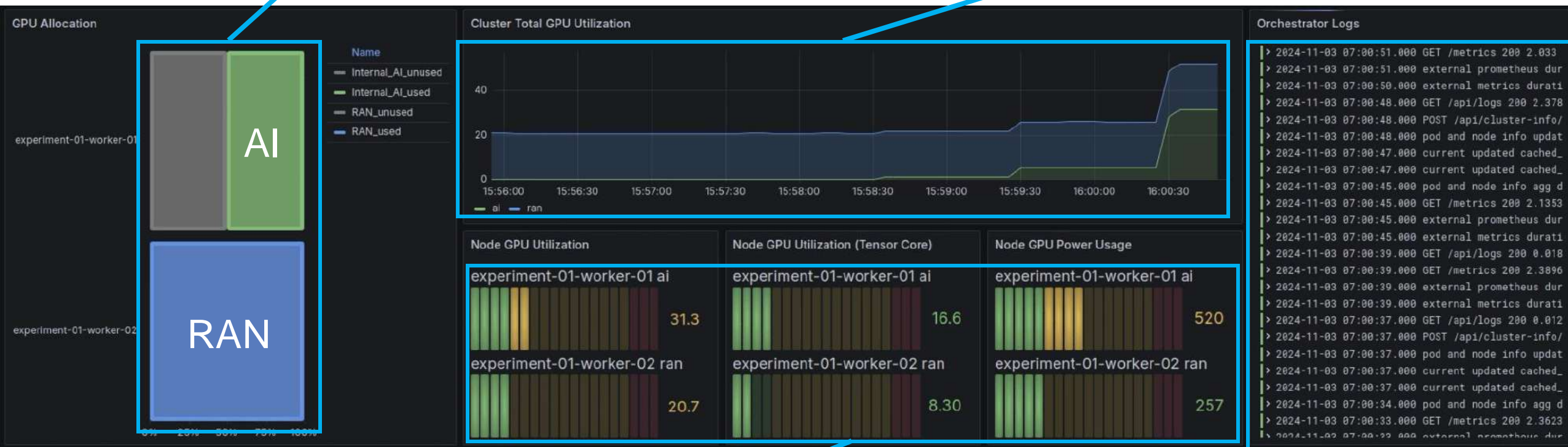


**Allocate
appropriately
sized GPUs
to AI & run
AI apps**

Demo Screens

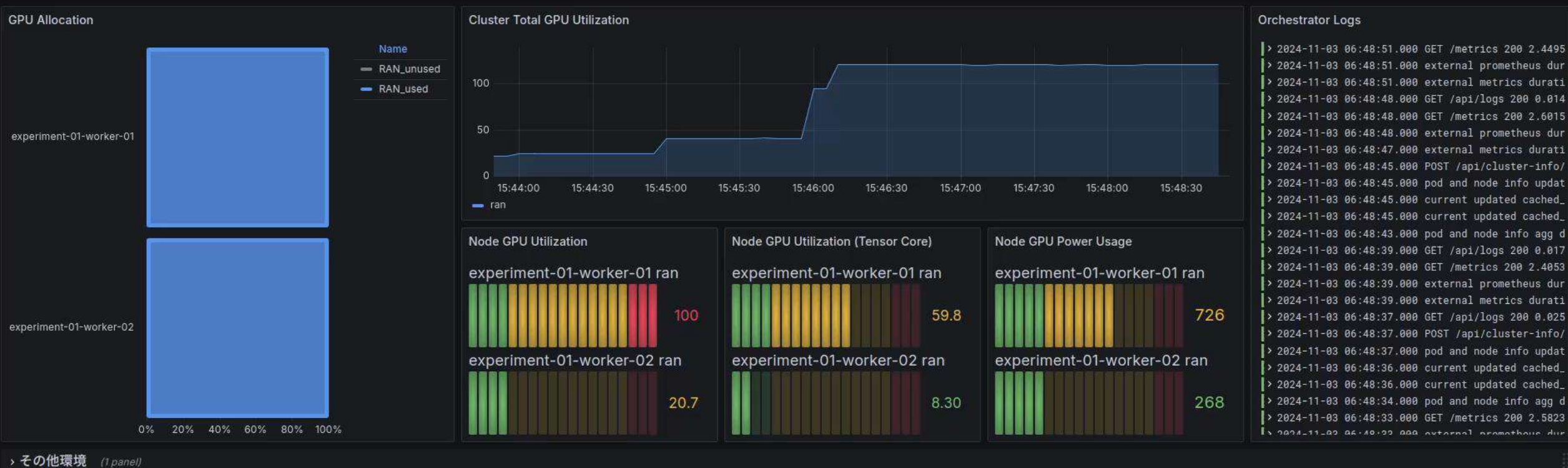
Node type and GPU allocation status

GPU Usage Graph



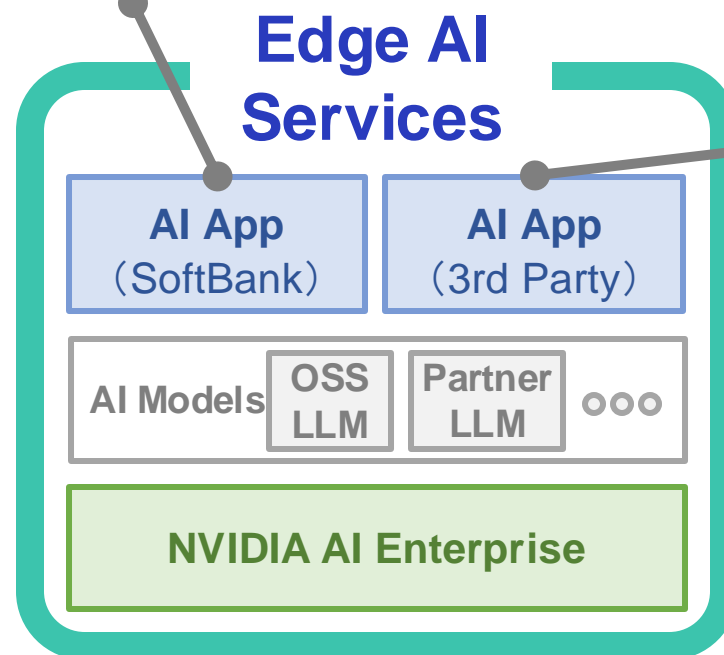
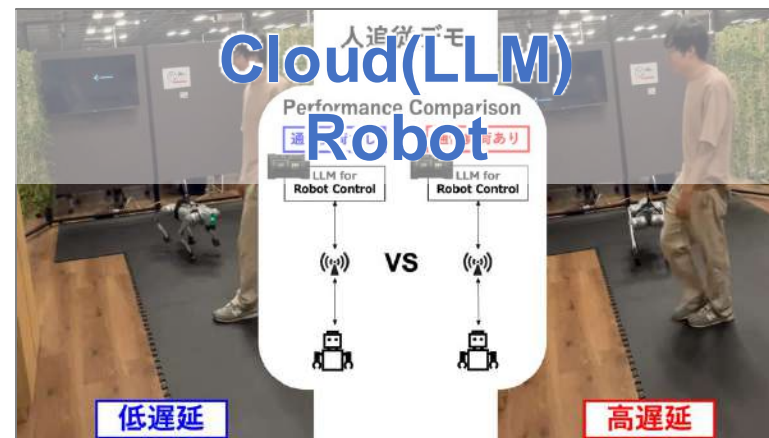
CUDA cores, Tensor cores,
Instantaneous value of node power consumption

Orchestrator
operation log



AI-RANオーケストレーターが管理するクラスターに2つのノードが存在し、RANのアプリケーションがそれぞれデプロイされています。

Use Cases with NVIDIA AI Enterprise@AITRAS



**Serverless API powered
by NVIDIA AI Enterprise
@edge**

