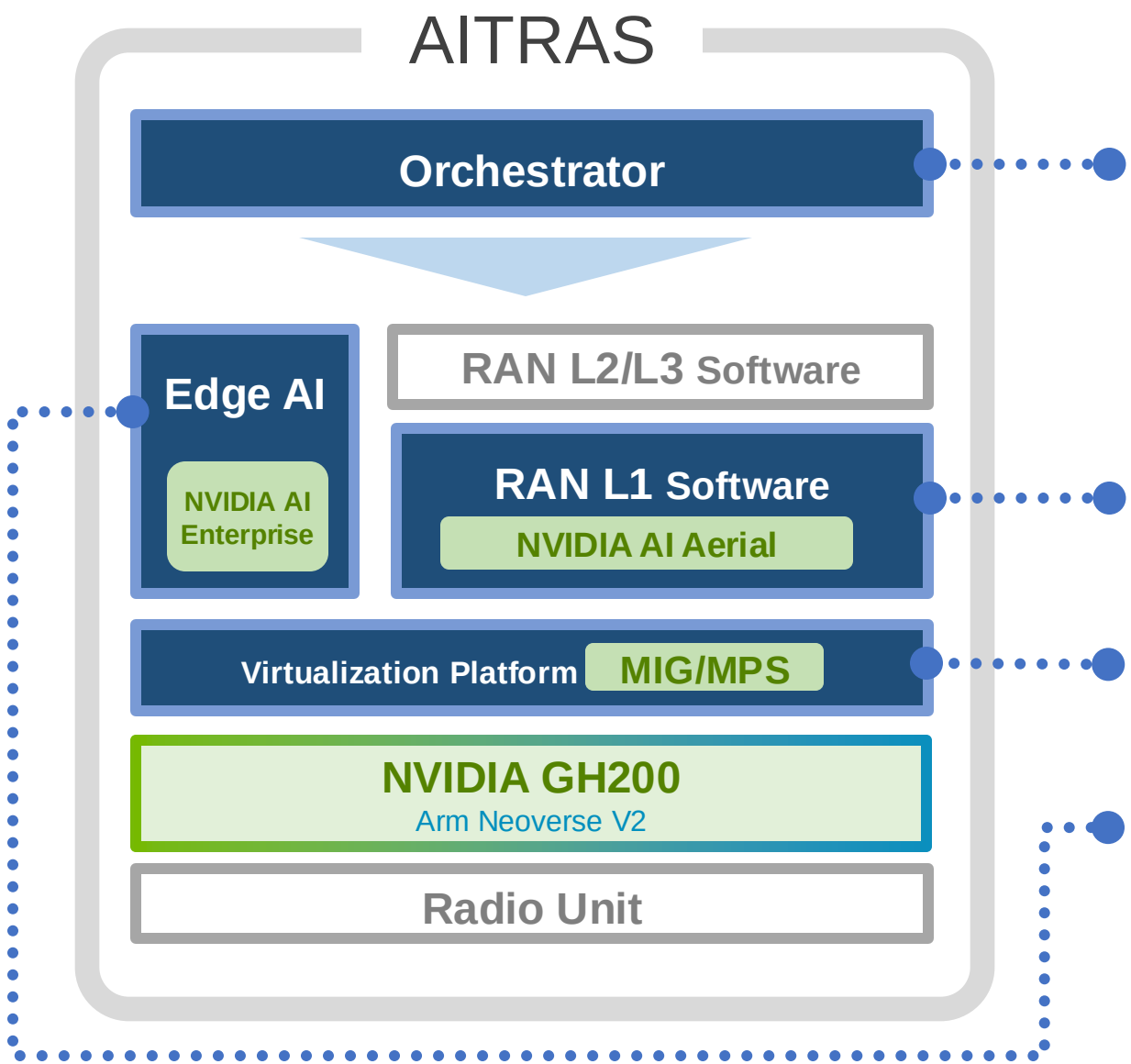
SoftBank

Cloud（LLM）Robot

AITRAS System Architecture



SoftBank-Developed Core Technologies

Management and control software

Enhancing the software for wireless signal processing

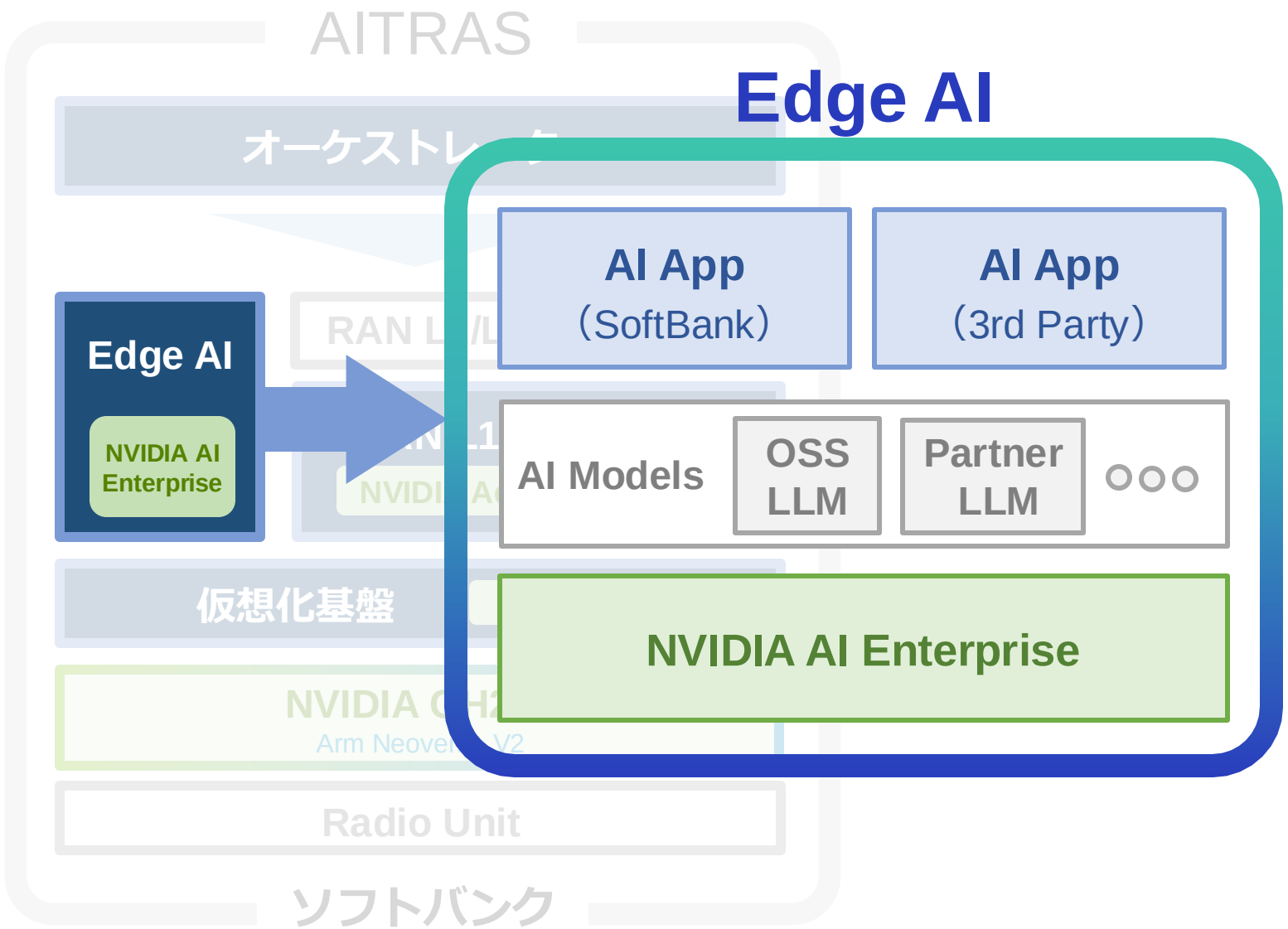
Building and implementing virtualization foundations

Providing Edge AI Service Menu

- LLM Robot
- Multimodal LLM
- High-performance RAG
- NVIDIA Serverless API*

*Serverless API powered by NVIDIA AI Enterprise

Positioning of this Demo



Original AI Service

- Autonomous Driving Monitoring App
- **Cloud Robot**
- RAG Menu @edge

NVIDIA Serverless API

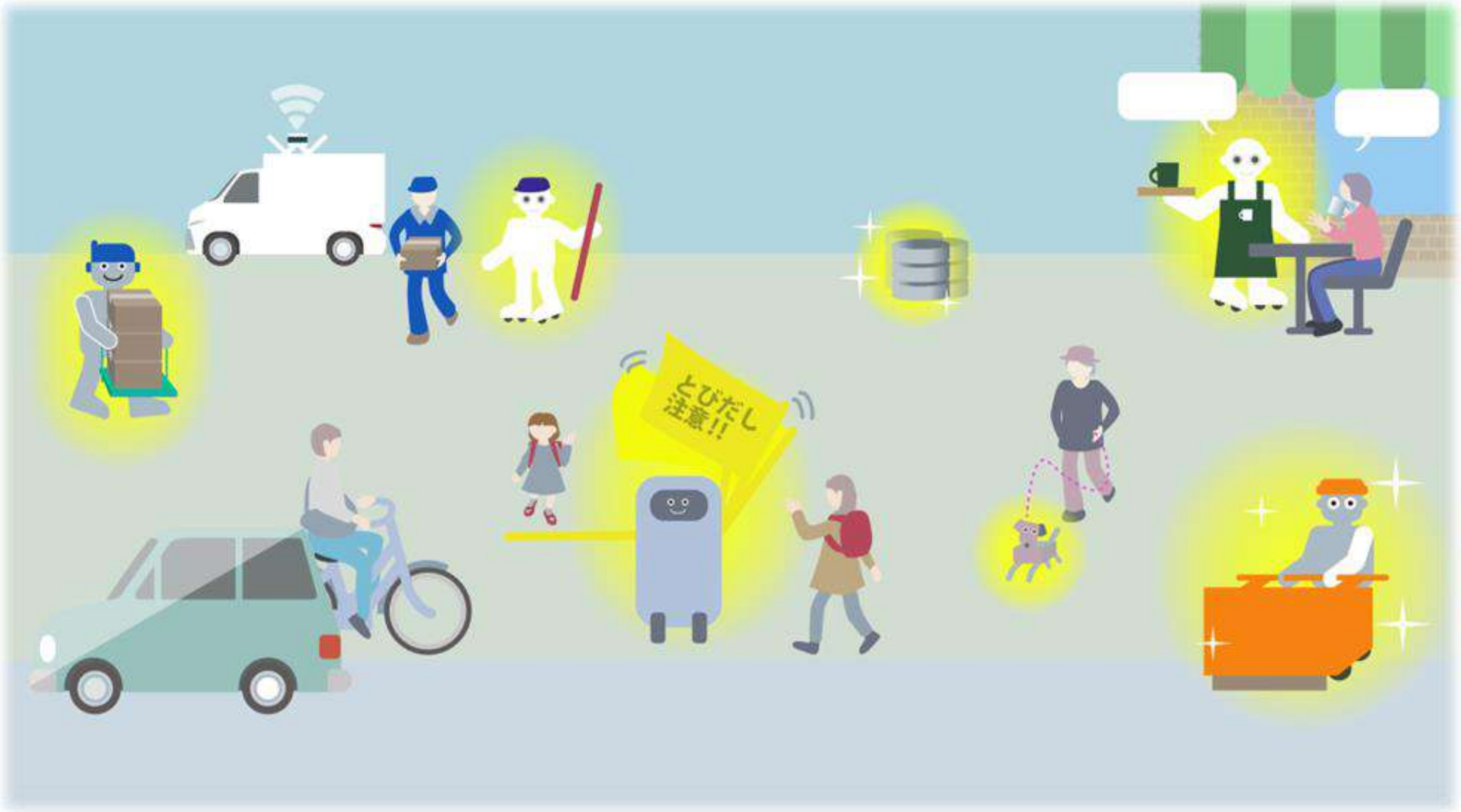
Allocates unused resources to meet customer demands on demand

NVIDIA AI Enterprise

Industry-standard AI framework

A Society Coexisting with Robots

SoftBank



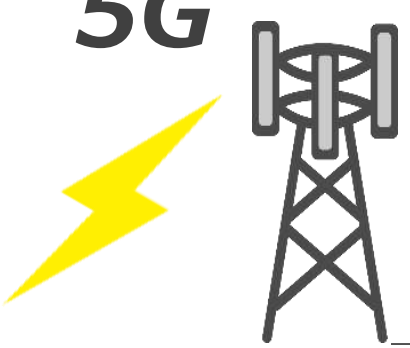
Robots with advanced decision-making abilities are actively supporting daily life

Offloading Processing

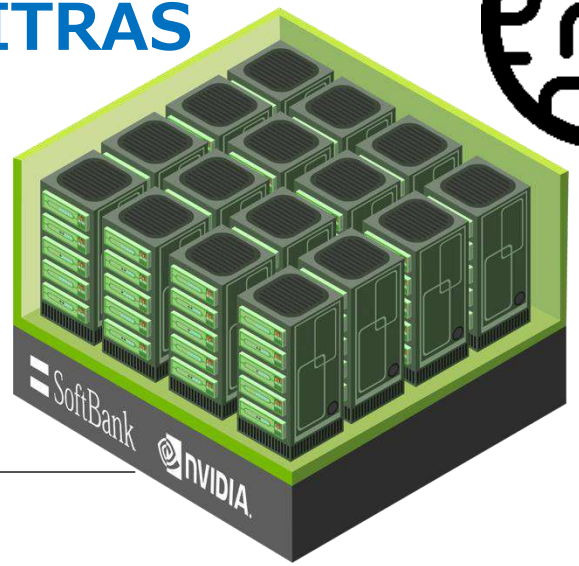
Advanced decision-making requires high-performance computing.

Computing Resources
Small

Computing Resources
Large

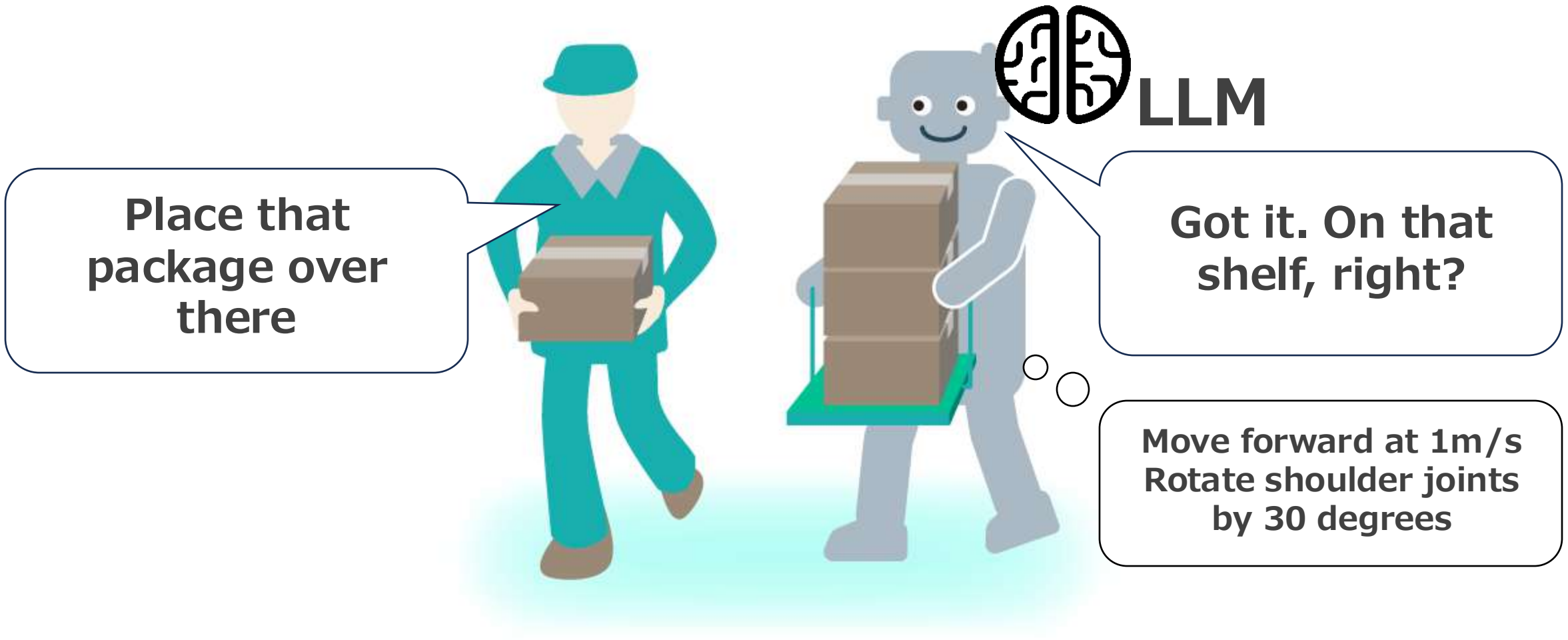


AITRAS



AITRAS can leverage more advanced computing power.

What is an LLM Robot?

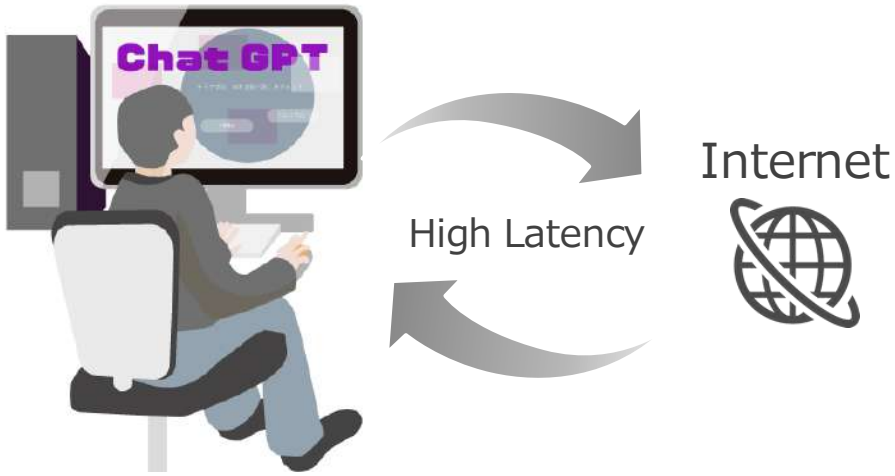


Capable of responding to new situations with common sense.

High-speed processing is essential for controlling robots

SoftBank

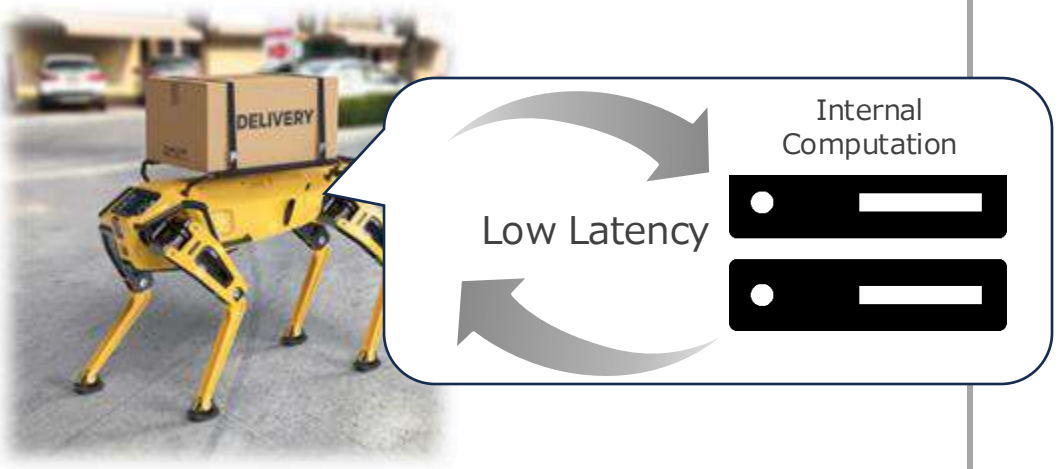
Latency of General LLMs
Time from Question to Answer



More than 1 second

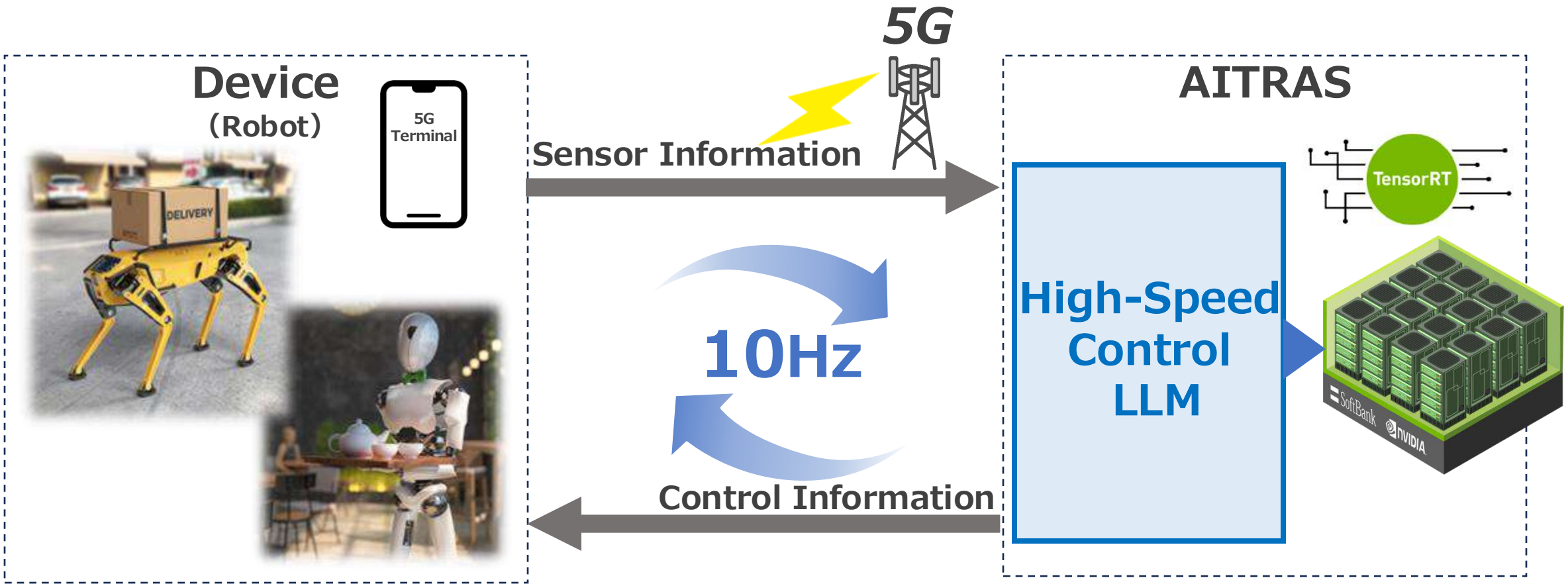
Gap

Robot Control Latency
Output of Control Information



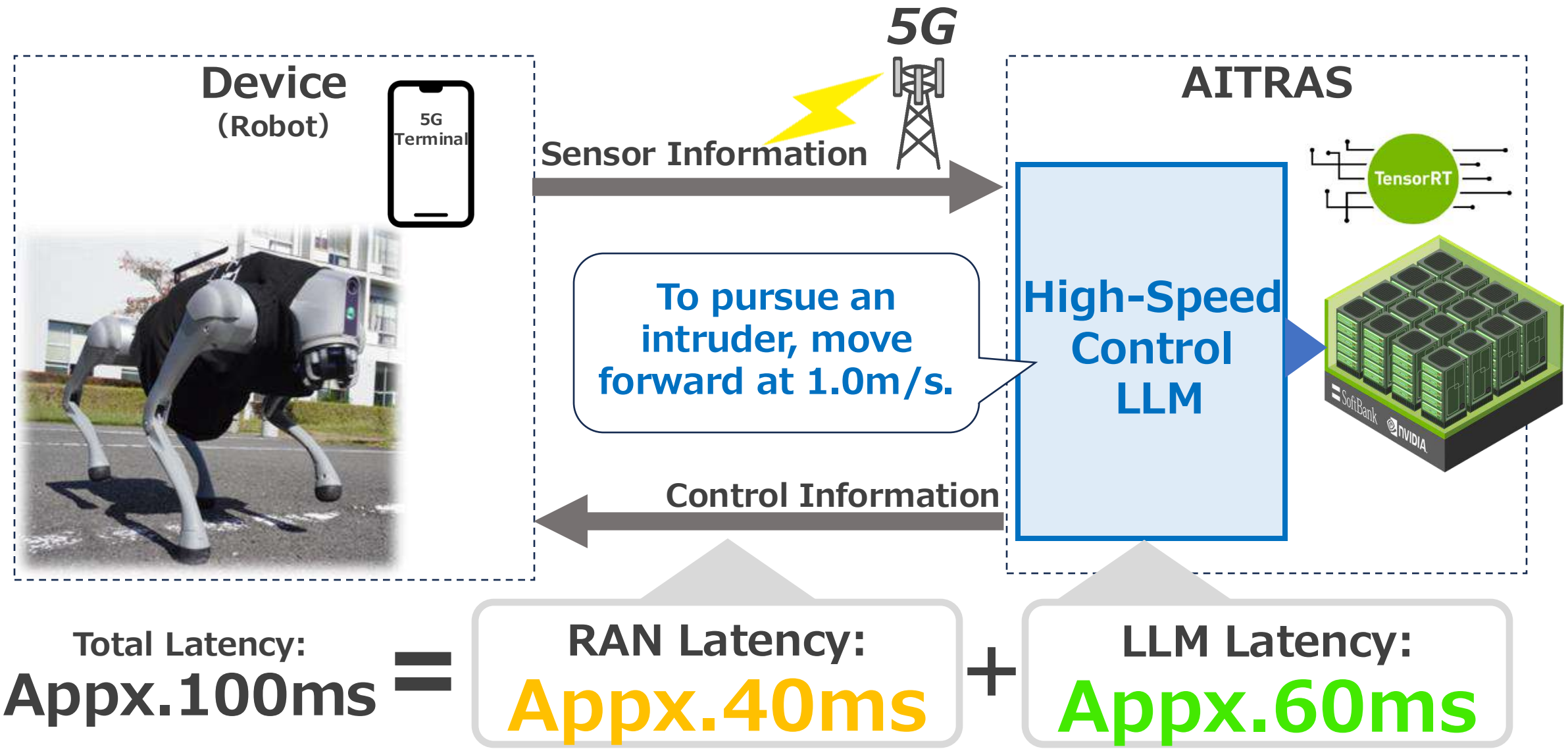
Around 0.1 seconds

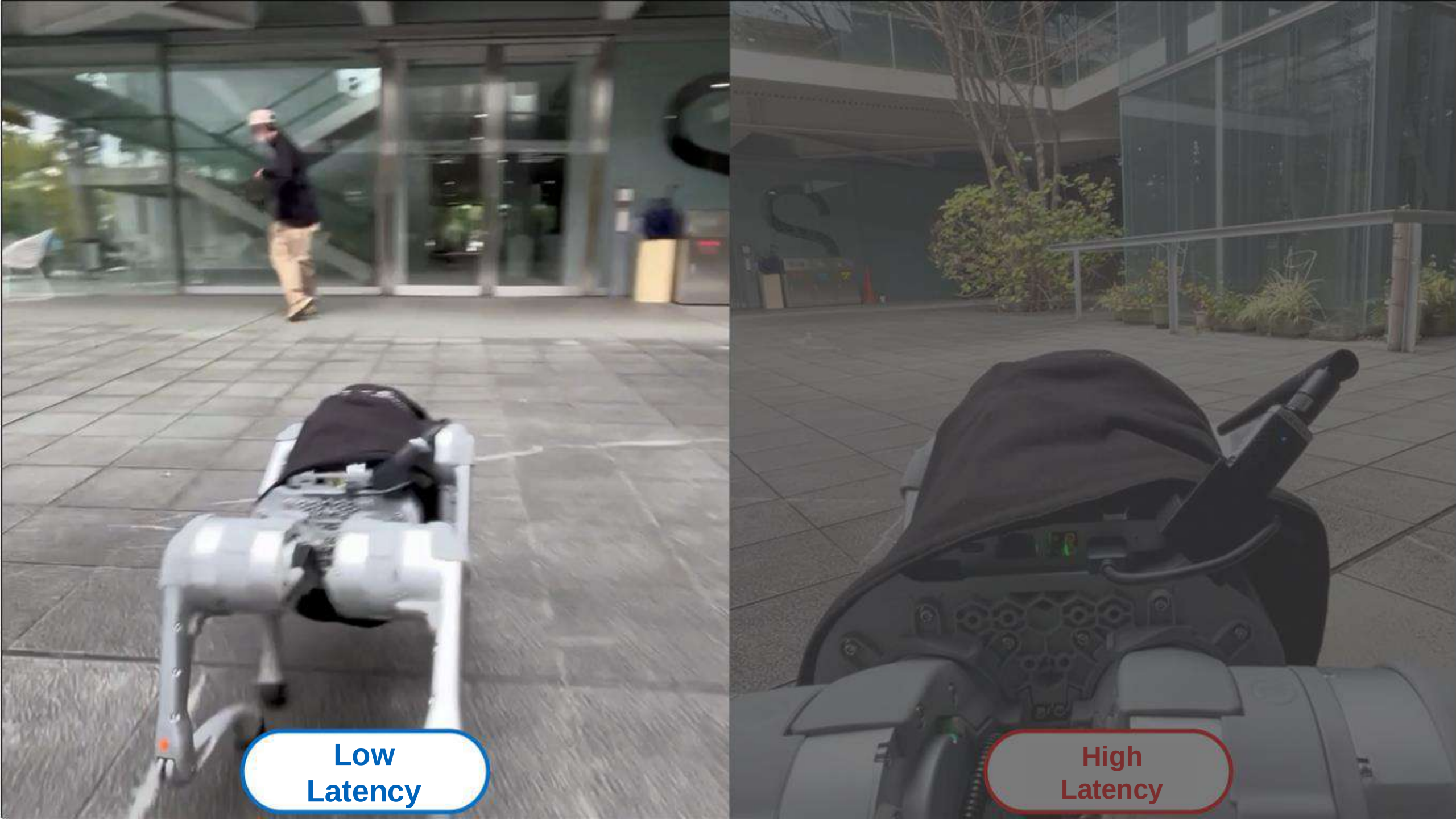
Low-Latency LLM Robots in Action



Robots Controlled 10 Times Per Second with LLM

Low-Latency LLM Robots in Action




Low
Latency
High
Latency