Towards Cloud-Driven Autonomous Vehicles

Peter Schafhalter
pschafhalter@berkeley.edu

NVIDIA H100 GPU

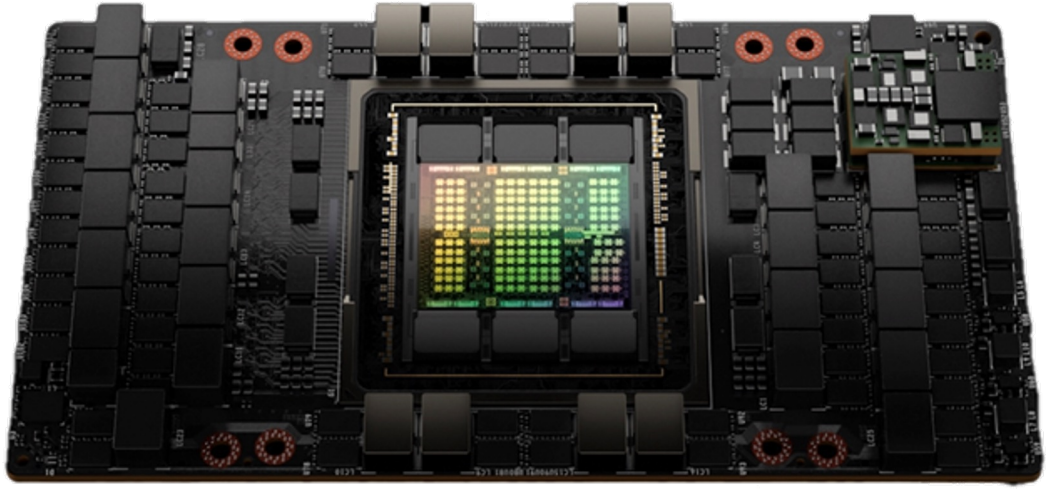


Image courtesy of NVIDIA

- **Cutting-edge GPU**
- **Trains powerful ML models**
- **Costs $30,000**

Tesla Model 3

Image courtesy of Tesla

- **EV with limited self-driving capabilities**
- **Software updates**
- **Costs $30,000**

Datacenter of GPUs

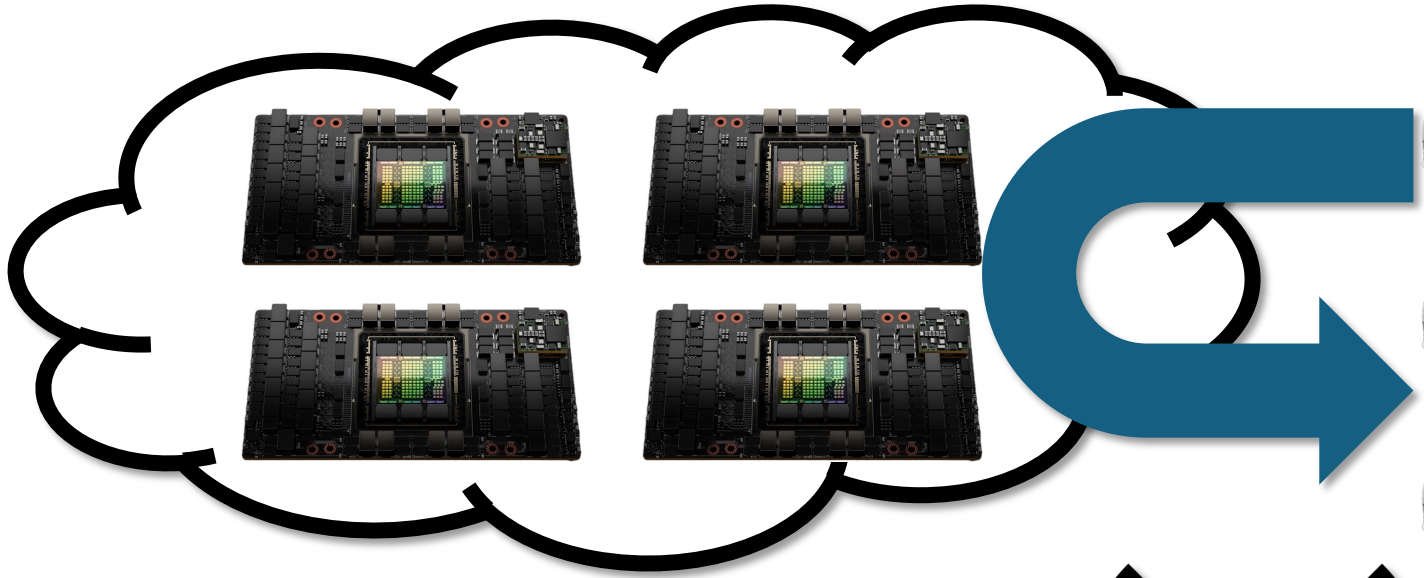


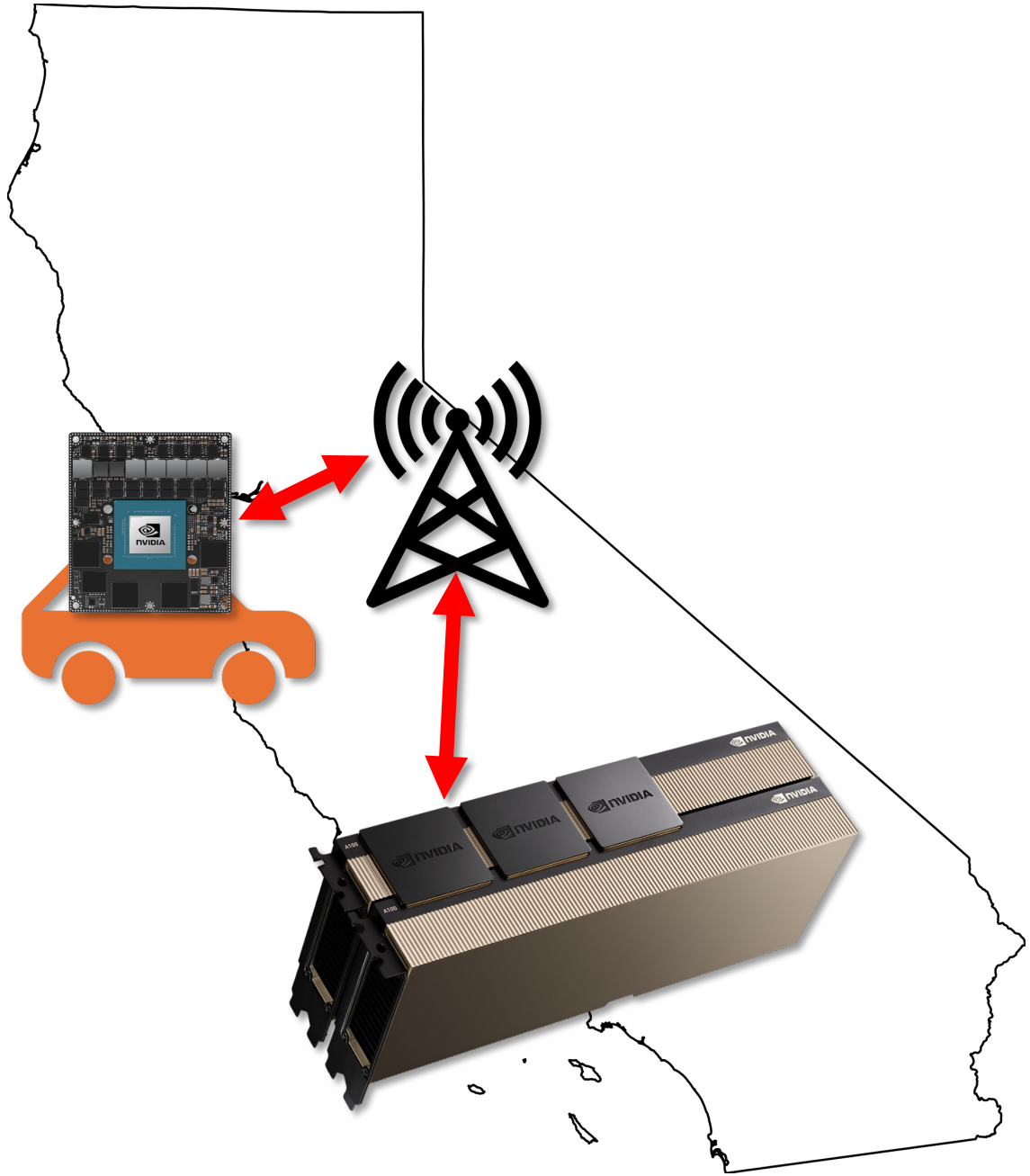
Image courtesy of NVIDIA

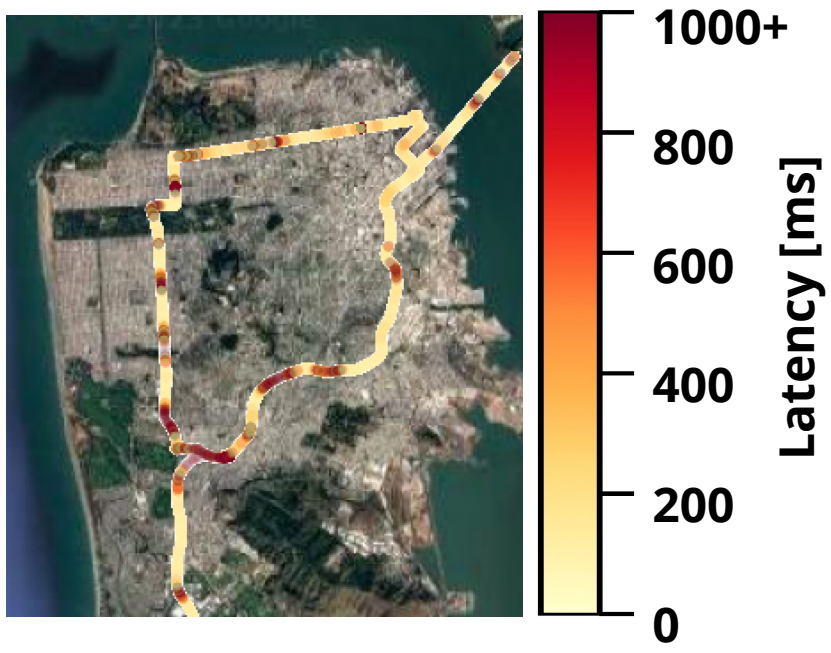
Fleet of Vehicles

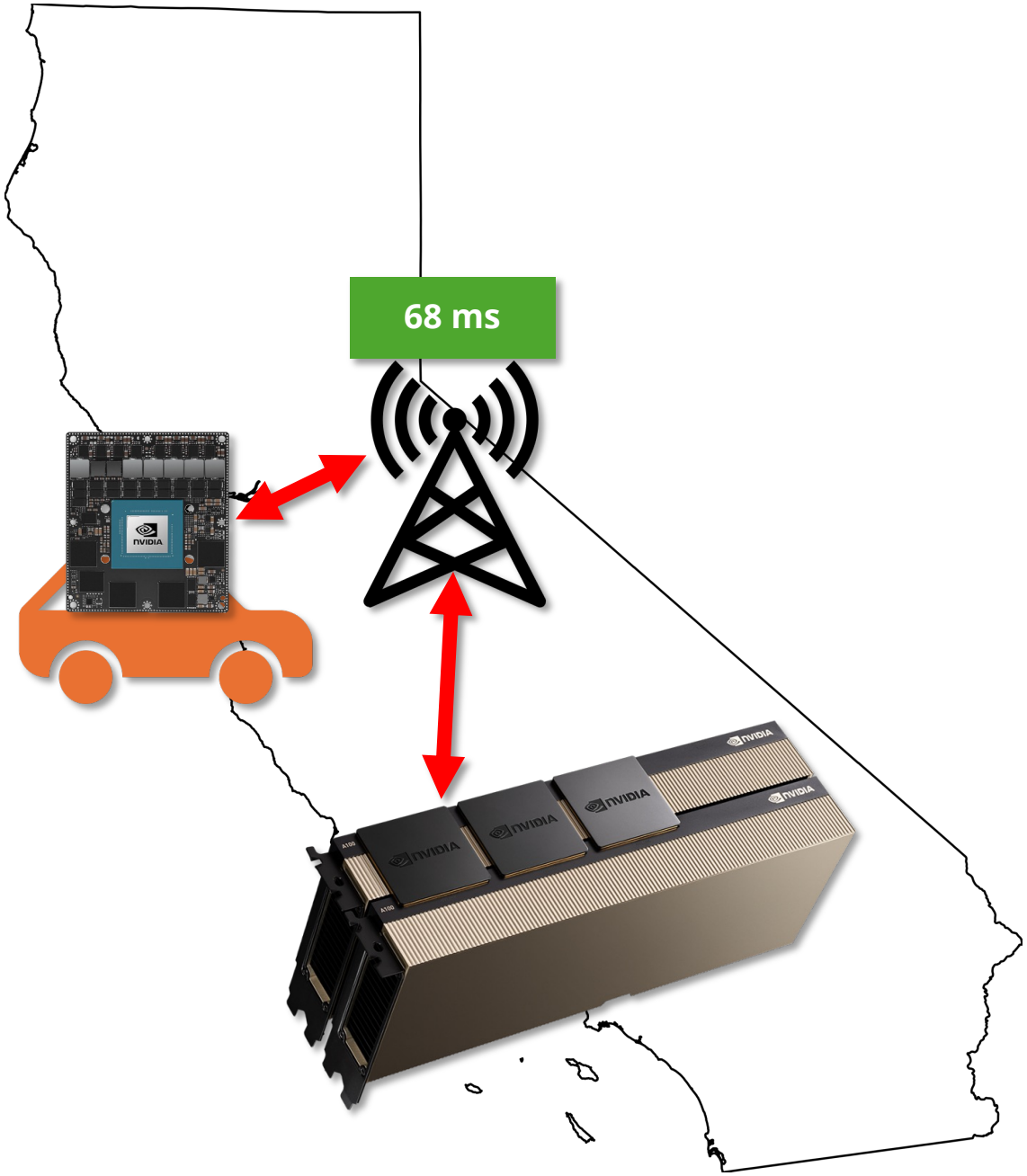
Image courtesy of Tesla

Challenge #1: Latency

Challenge #2: Reliability

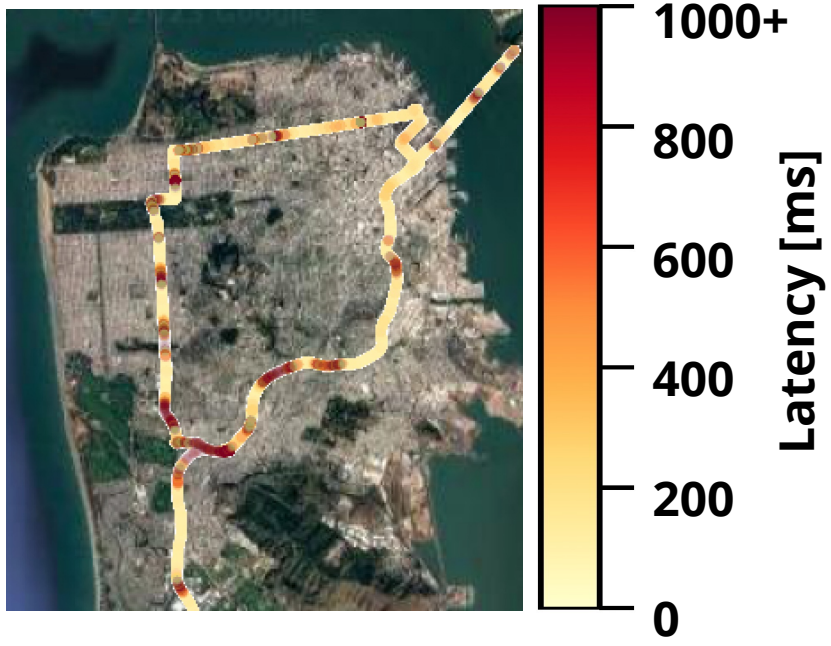

NVIDIA
NVIDIA
NVIDIA
NVIDIA
NVIDIA
NVIDIA

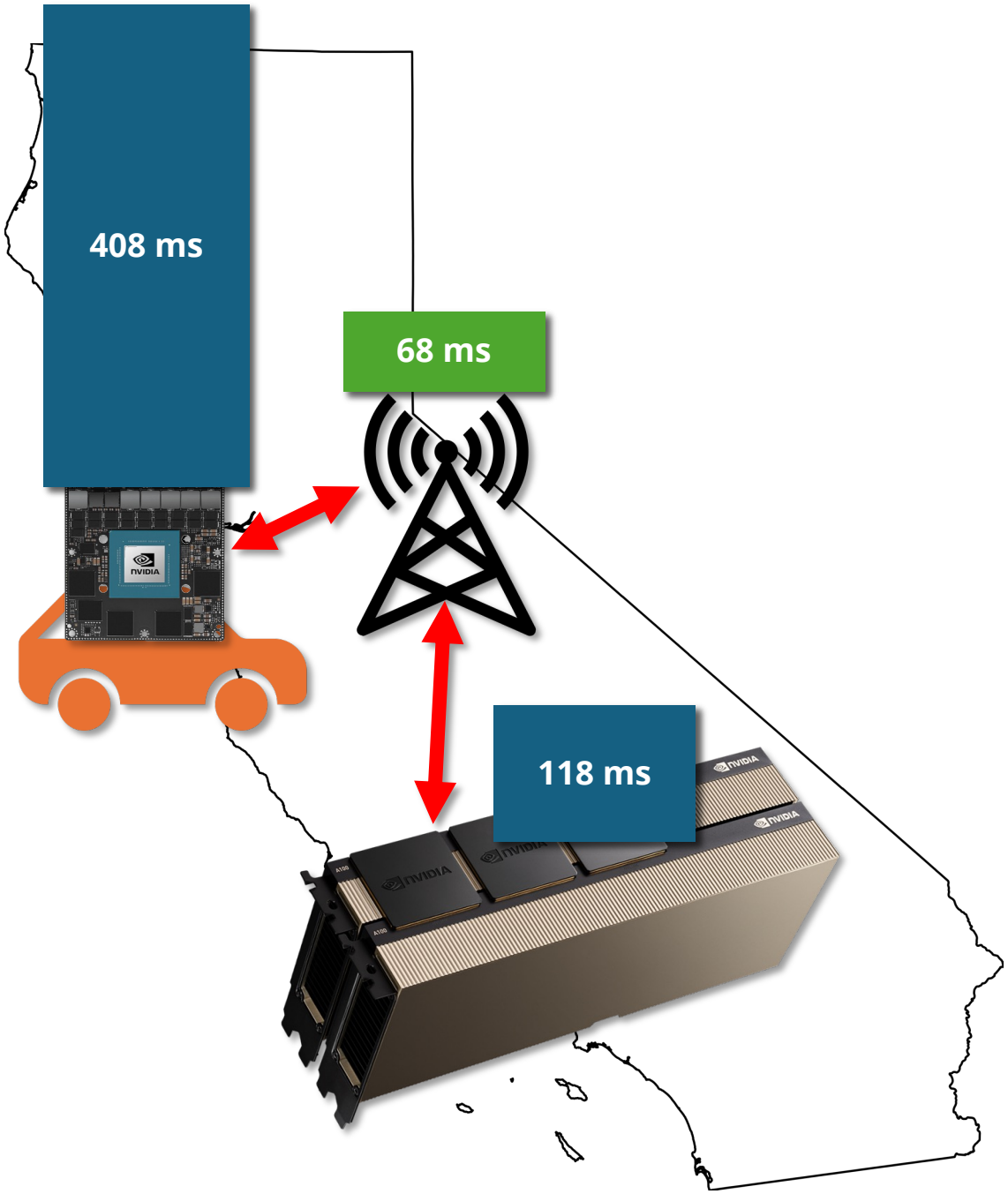




Round-trip latency:

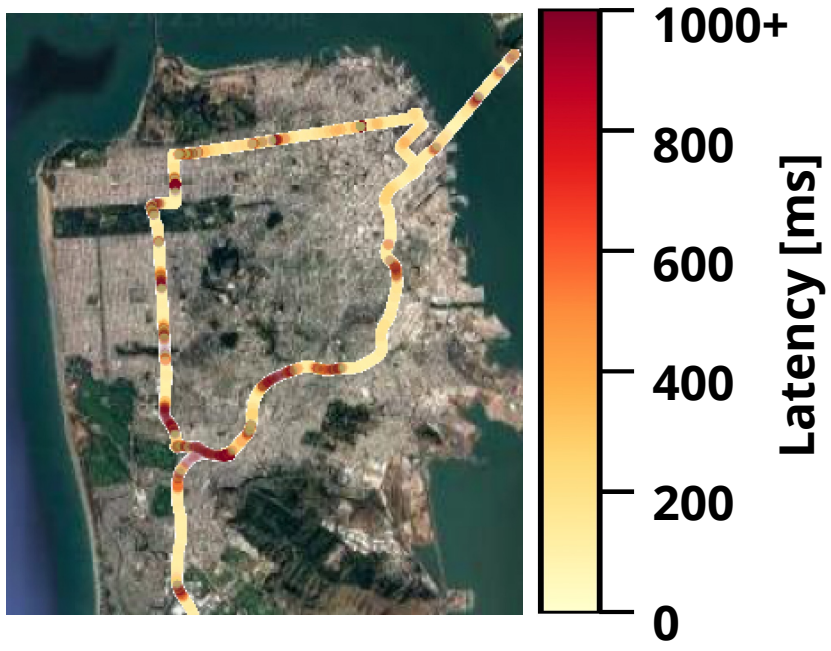
- Median: 68 ms

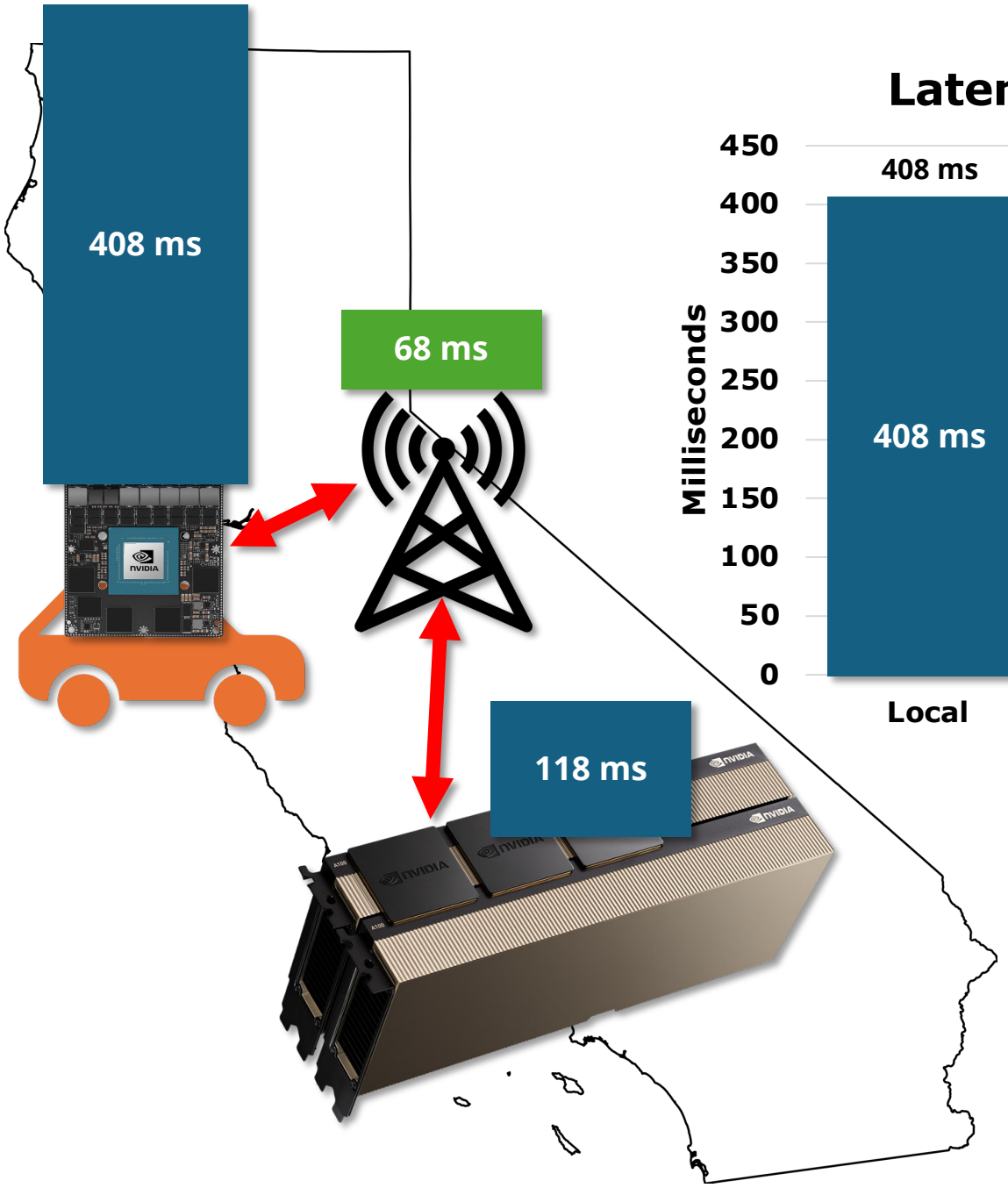




Round-trip latency:

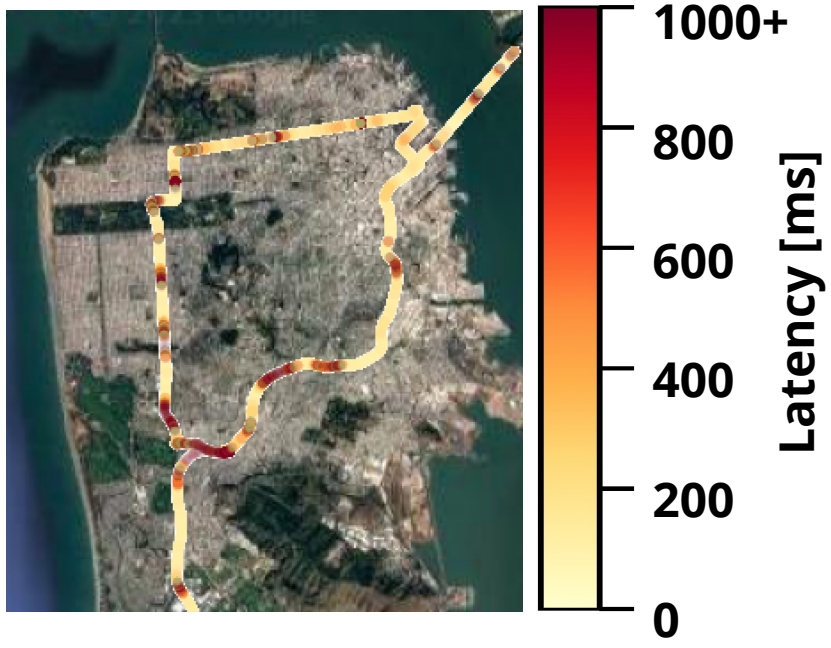
- Median: 68 ms

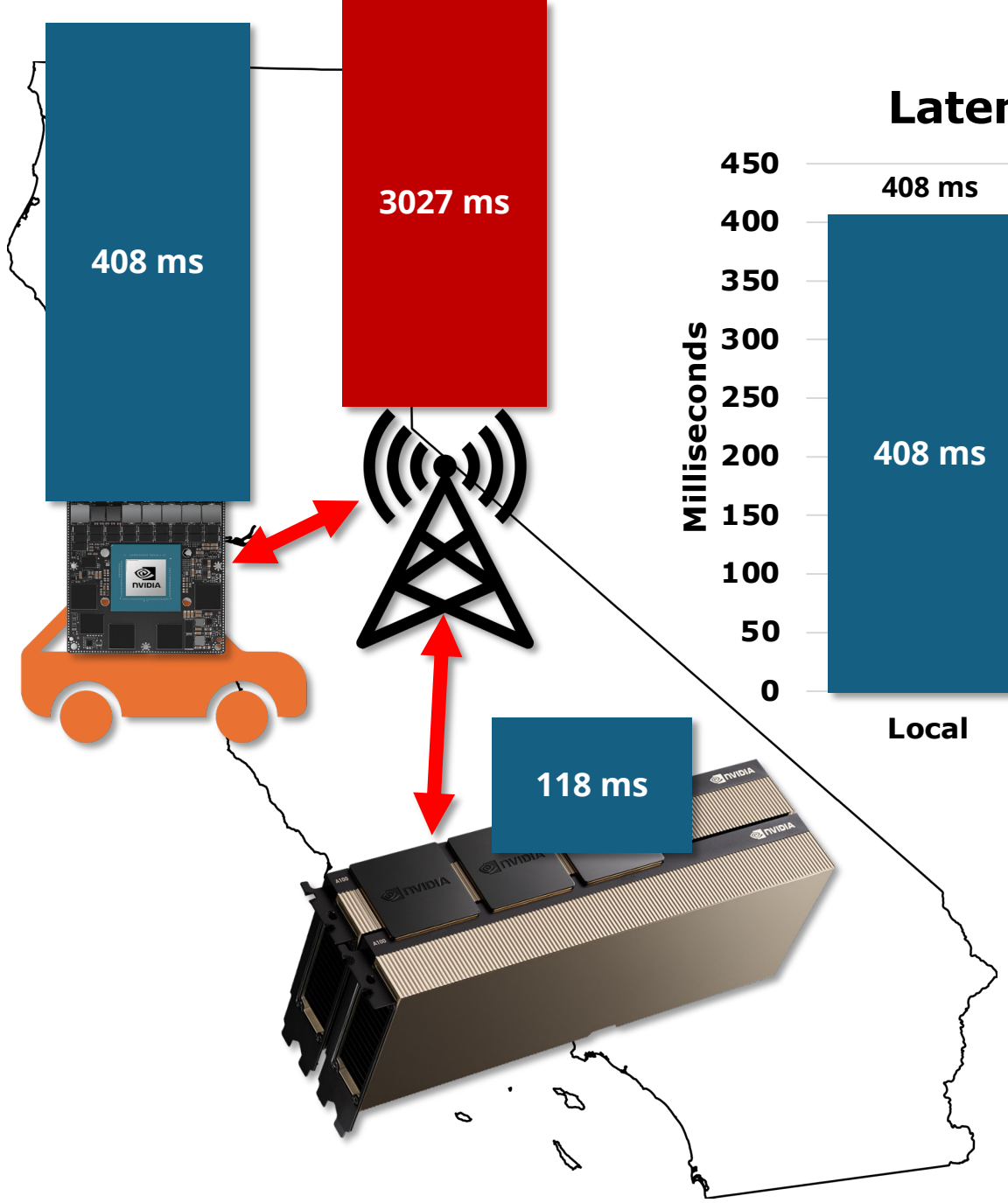




Round-trip latency:

- Median: 68 ms





Round-trip latency:

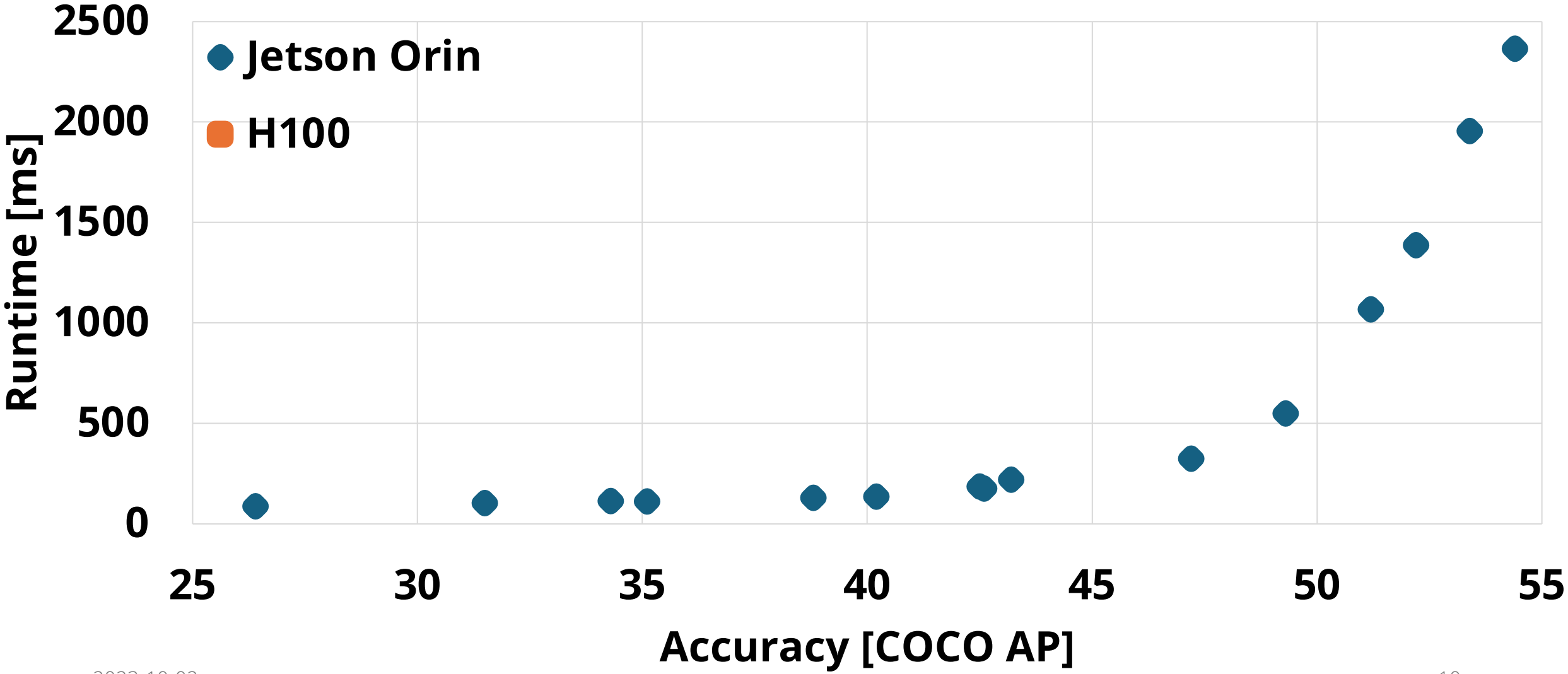
- Median: 68 ms
- 99th percentile: 3027 ms

The Network is Key to Self-Driving

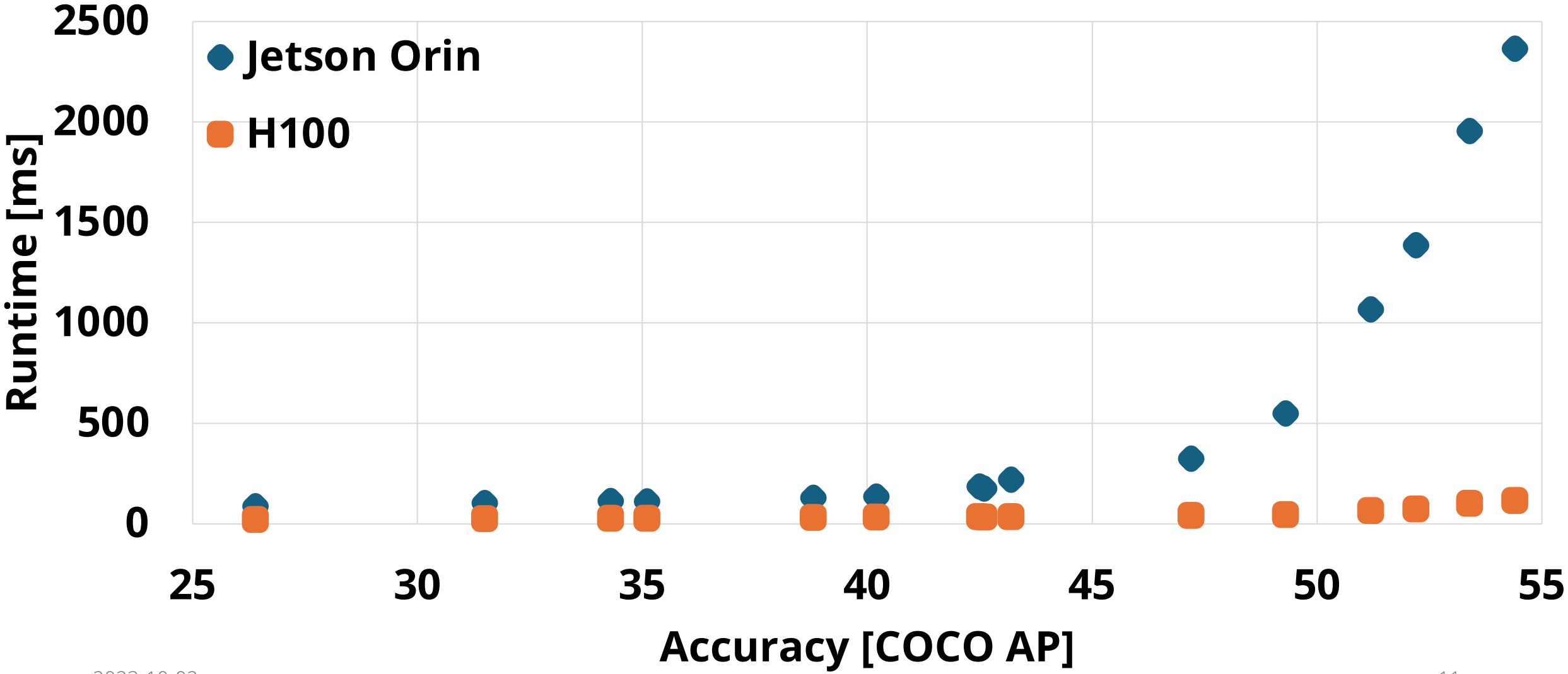
Opportunity: Cloud resources are faster and more plentiful than processors designed for the car

Challenge: How to manage the network?

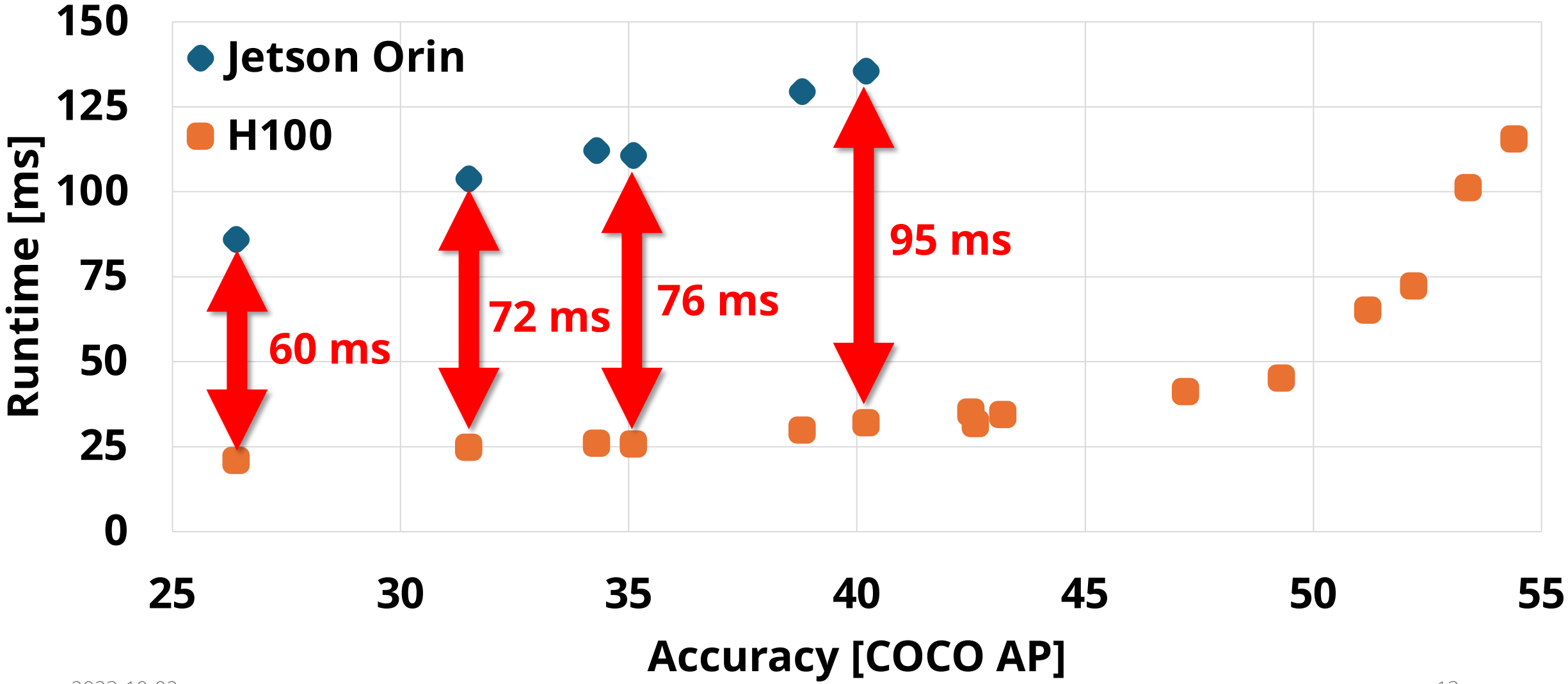
Vehicle vs. Cloud: Object Detection



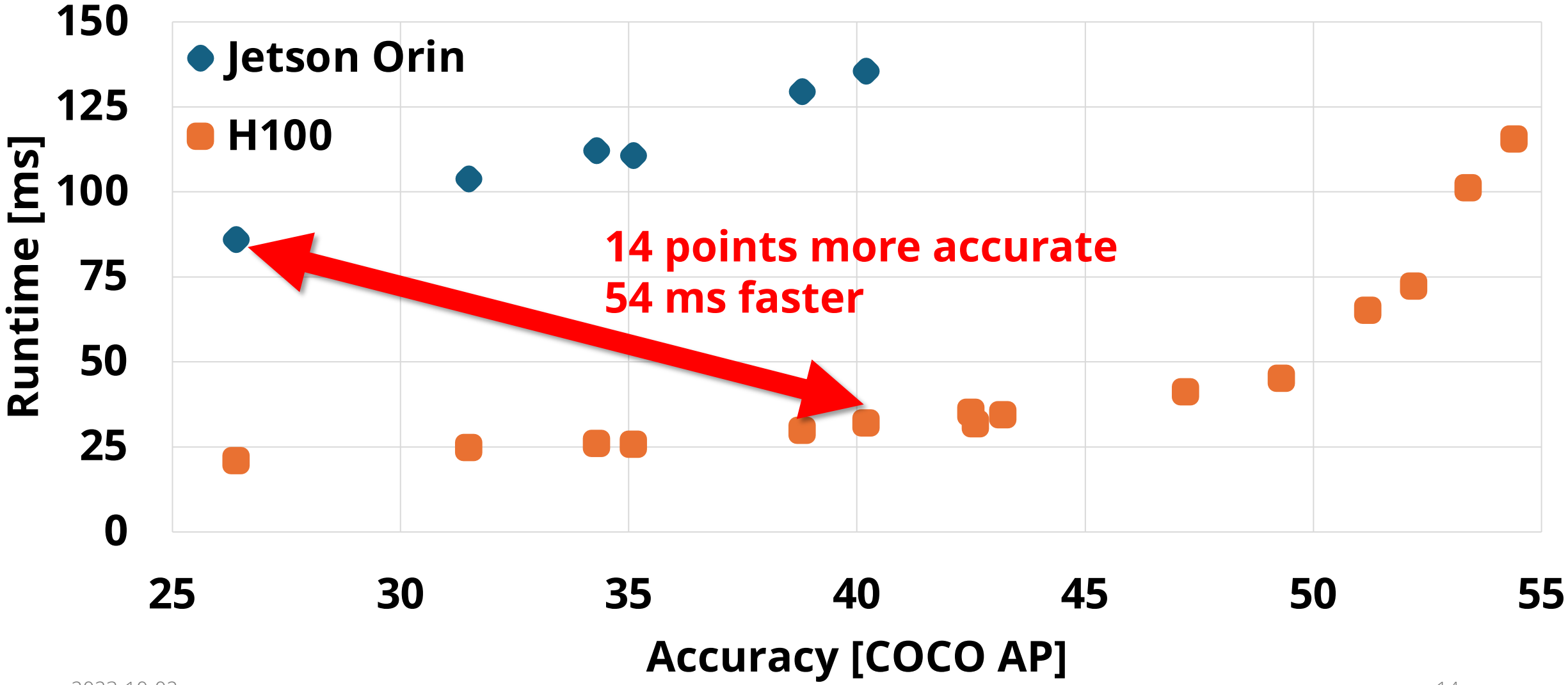
Vehicle vs. Cloud: Object Detection



Vehicle vs. Cloud: Object Detection



Vehicle vs. Cloud: Object Detection



Managing Network Reliability

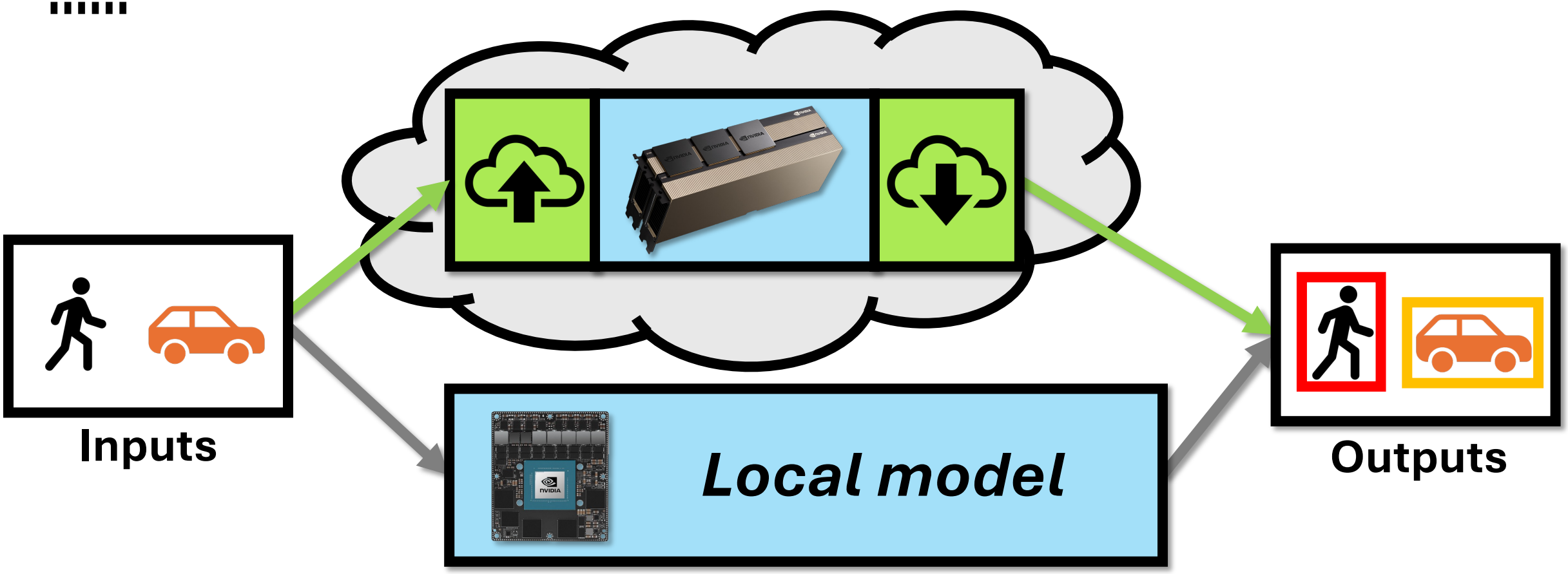
Availability

Managing Network Reliability

Availability

Connection Quality
latency, bandwidth

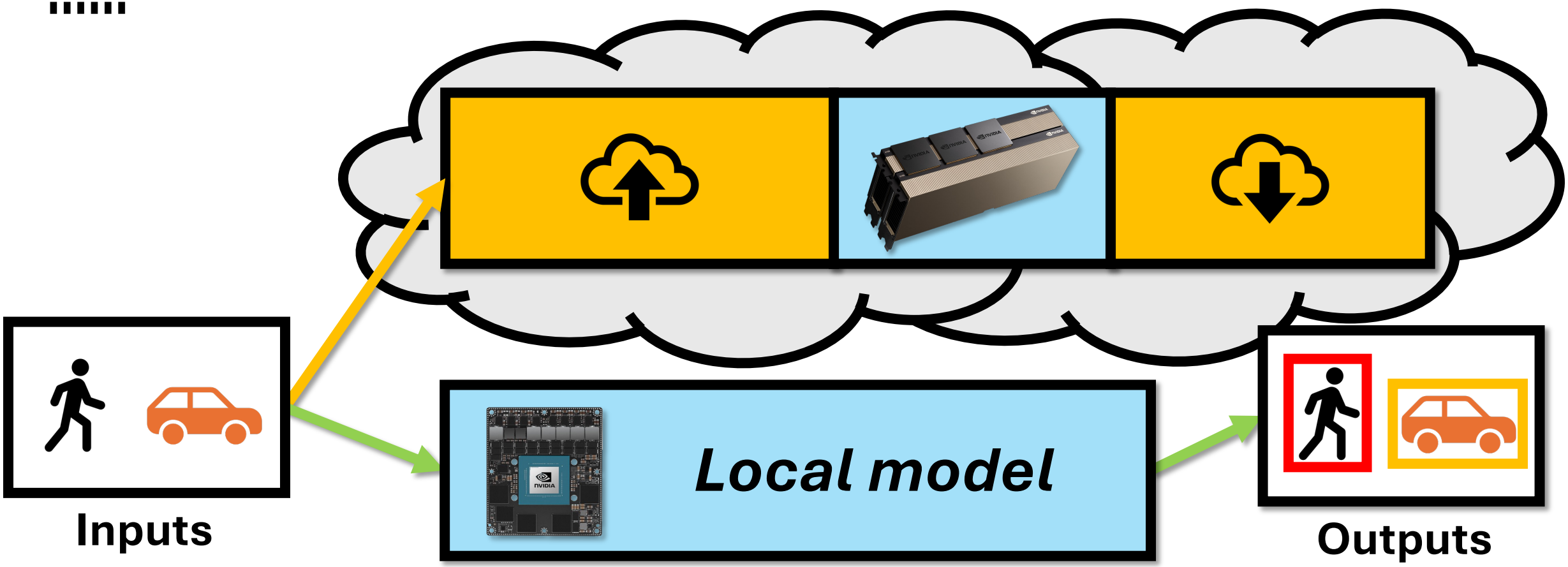
Speculative Execution



Availability

Connection Quality

Speculative Execution



Availability

Connection Quality

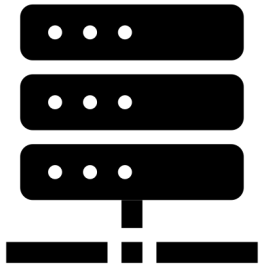
Impact: Avoid Collisions with Cloud

Detecting a traffic jam with DETR-ResNet-101

Jetson Orin – collision

A100 in Cloud – no collision

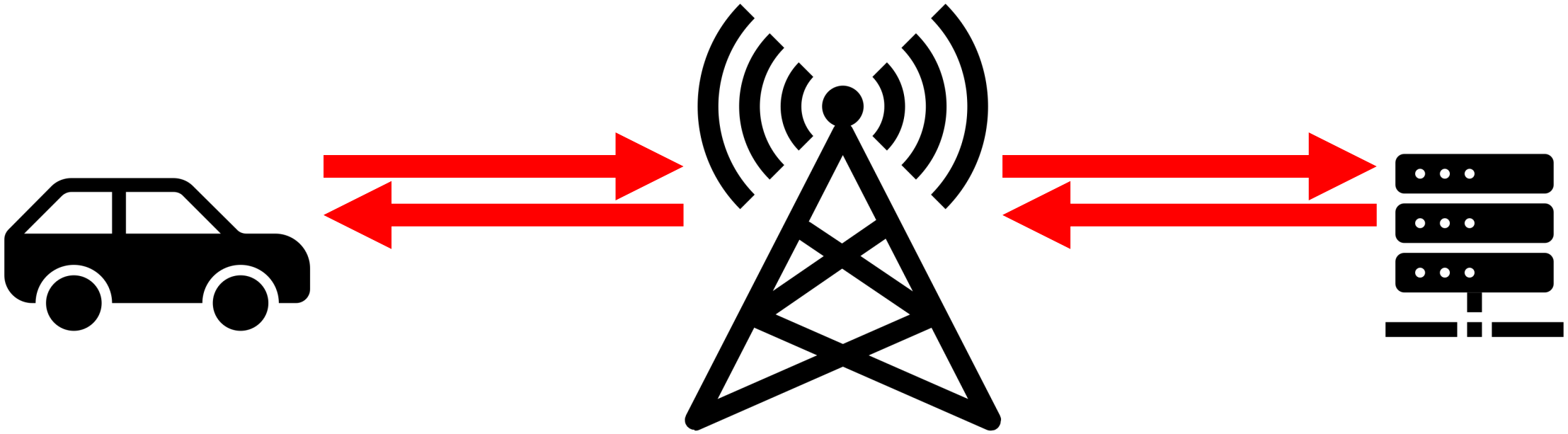
The Network is a Scarce Resource



The Network is a Scarce Resource

Network latency factors:

Round trip latency

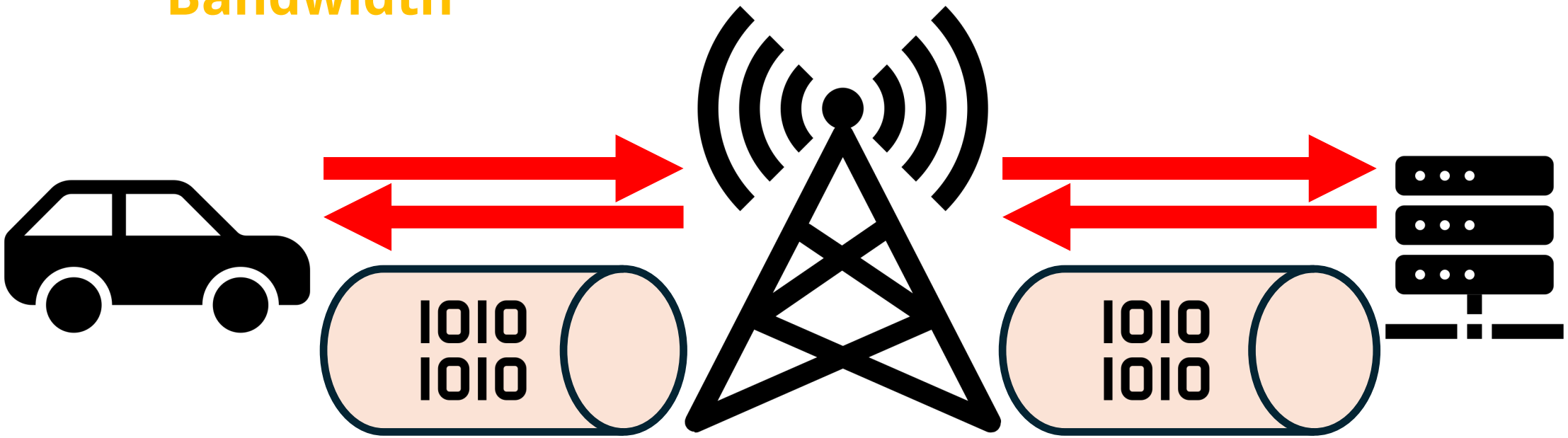


The Network is a Scarce Resource

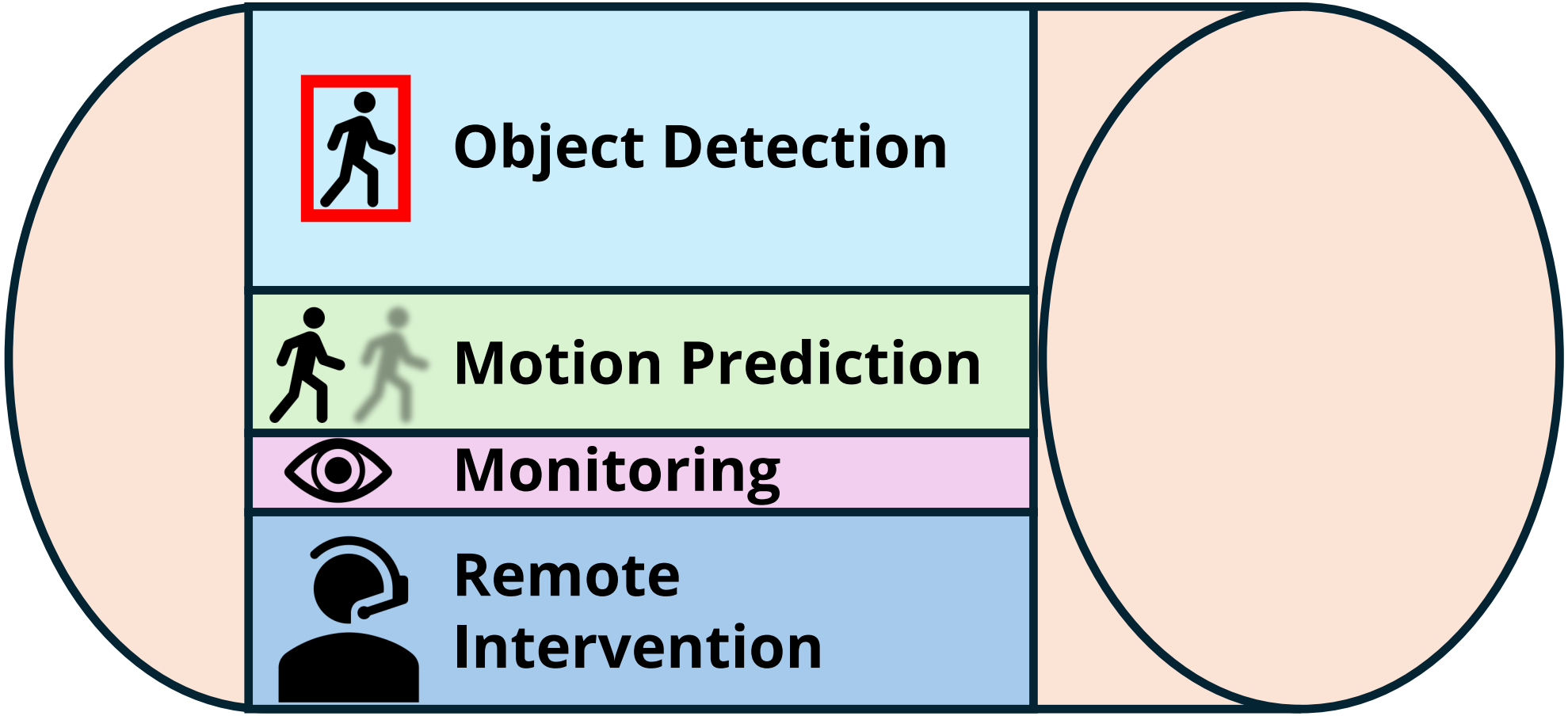
Network latency factors:

Round trip latency

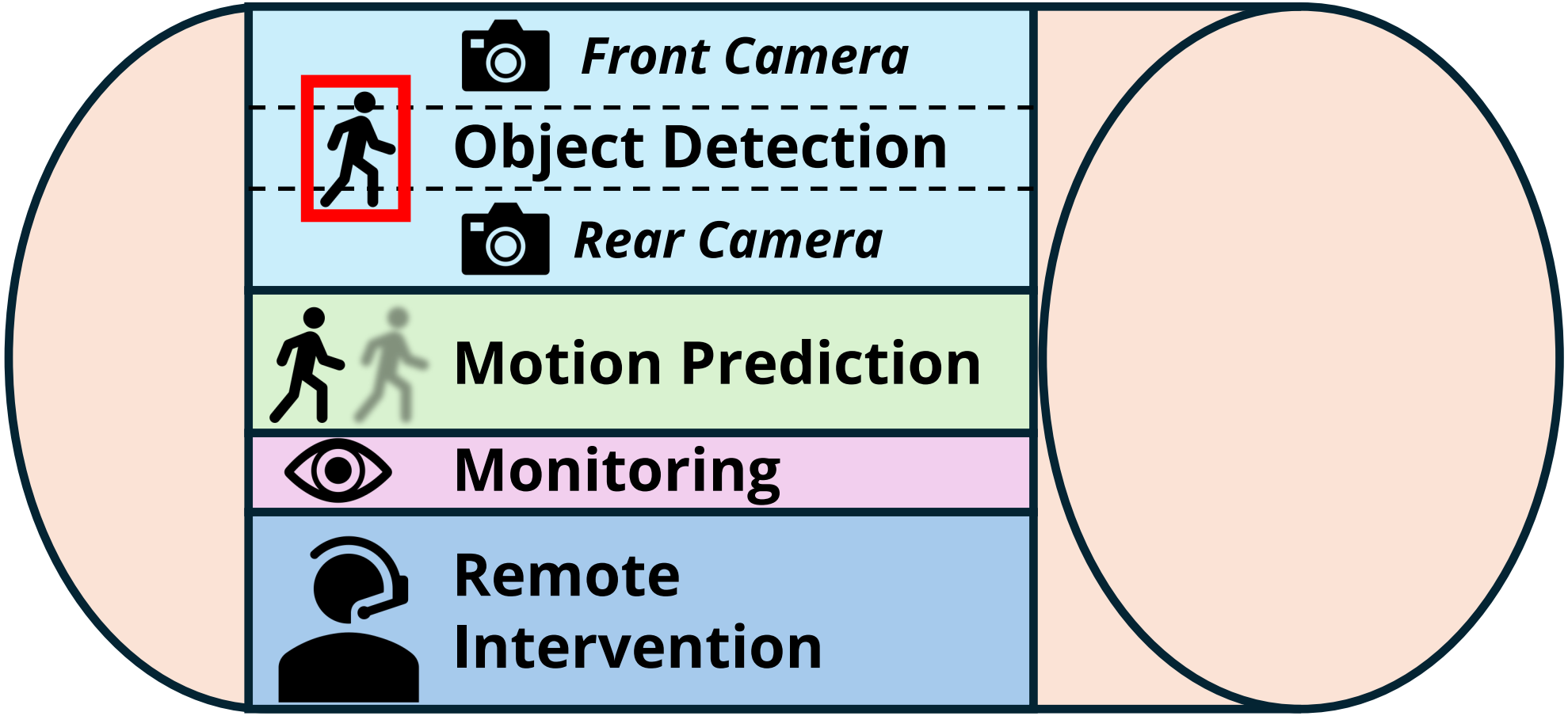
Bandwidth



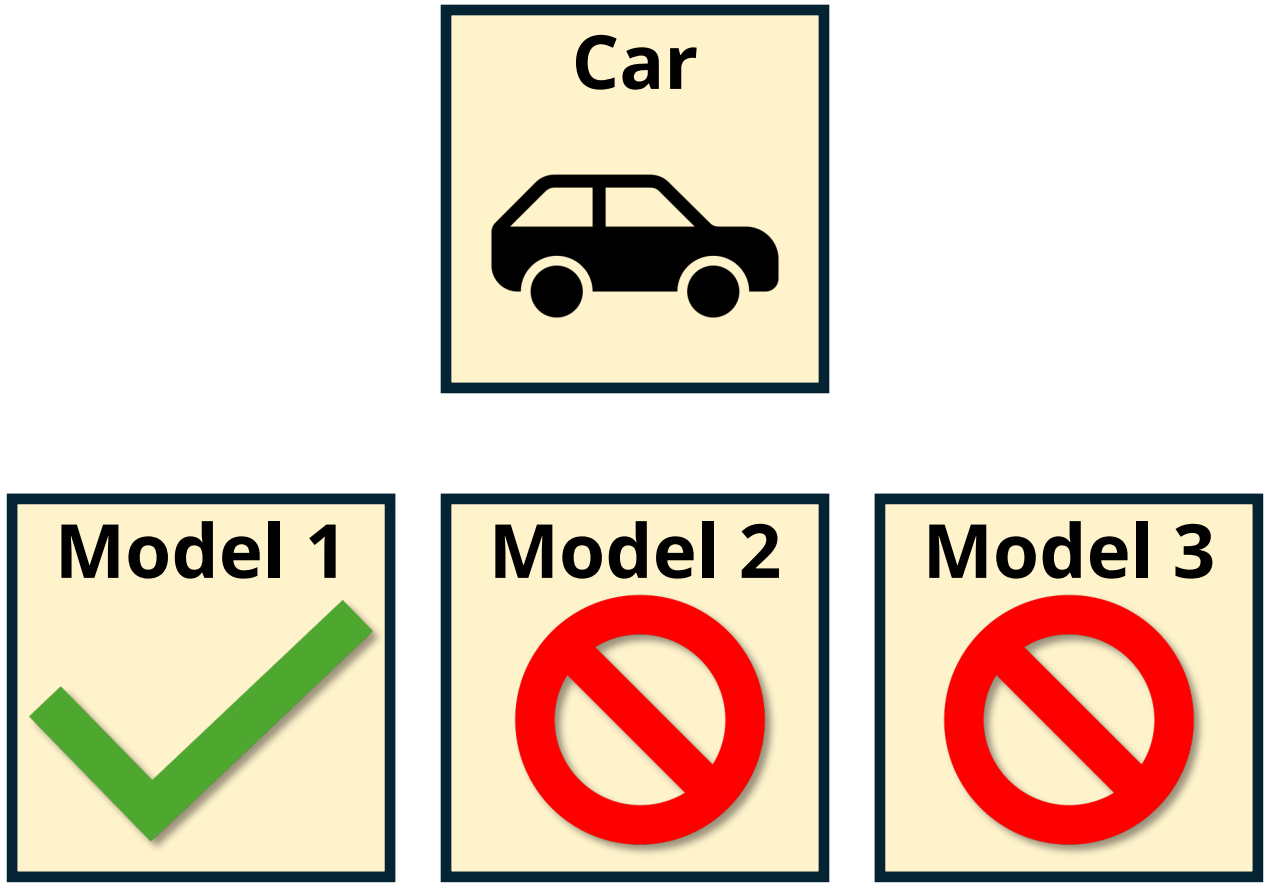
Allocating Bandwidth



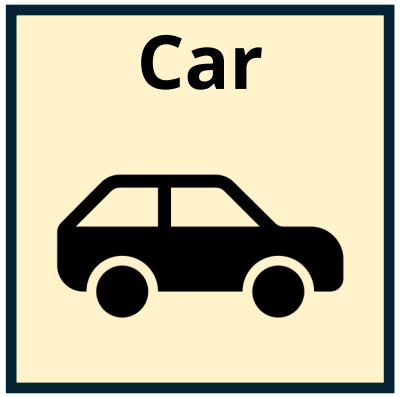
Allocating Bandwidth

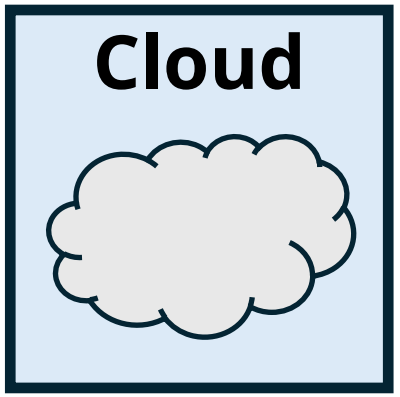


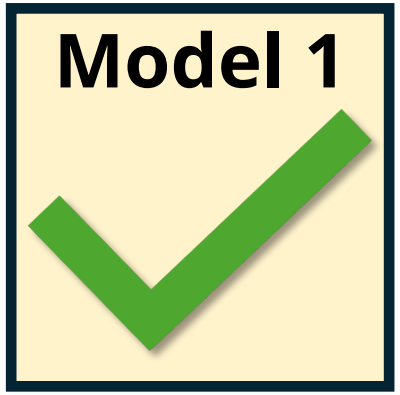
Prioritizing Services

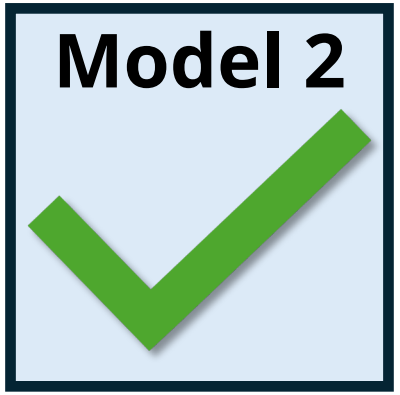


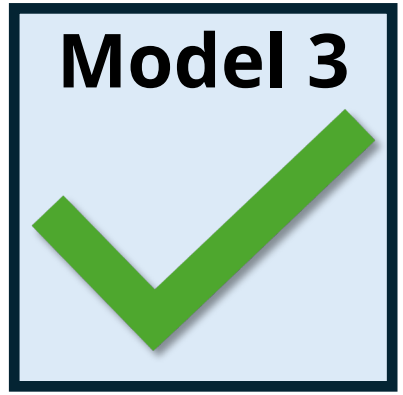
Prioritizing Services



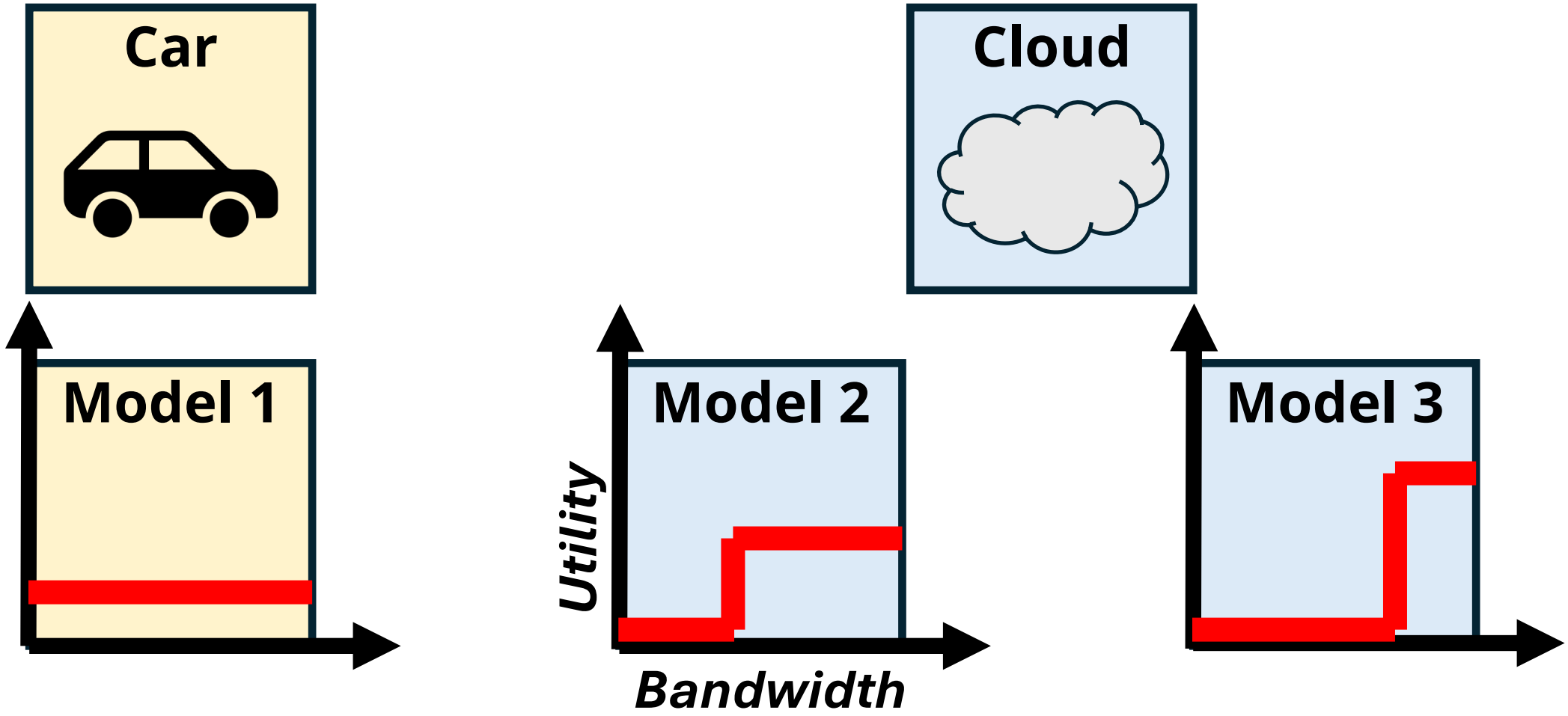




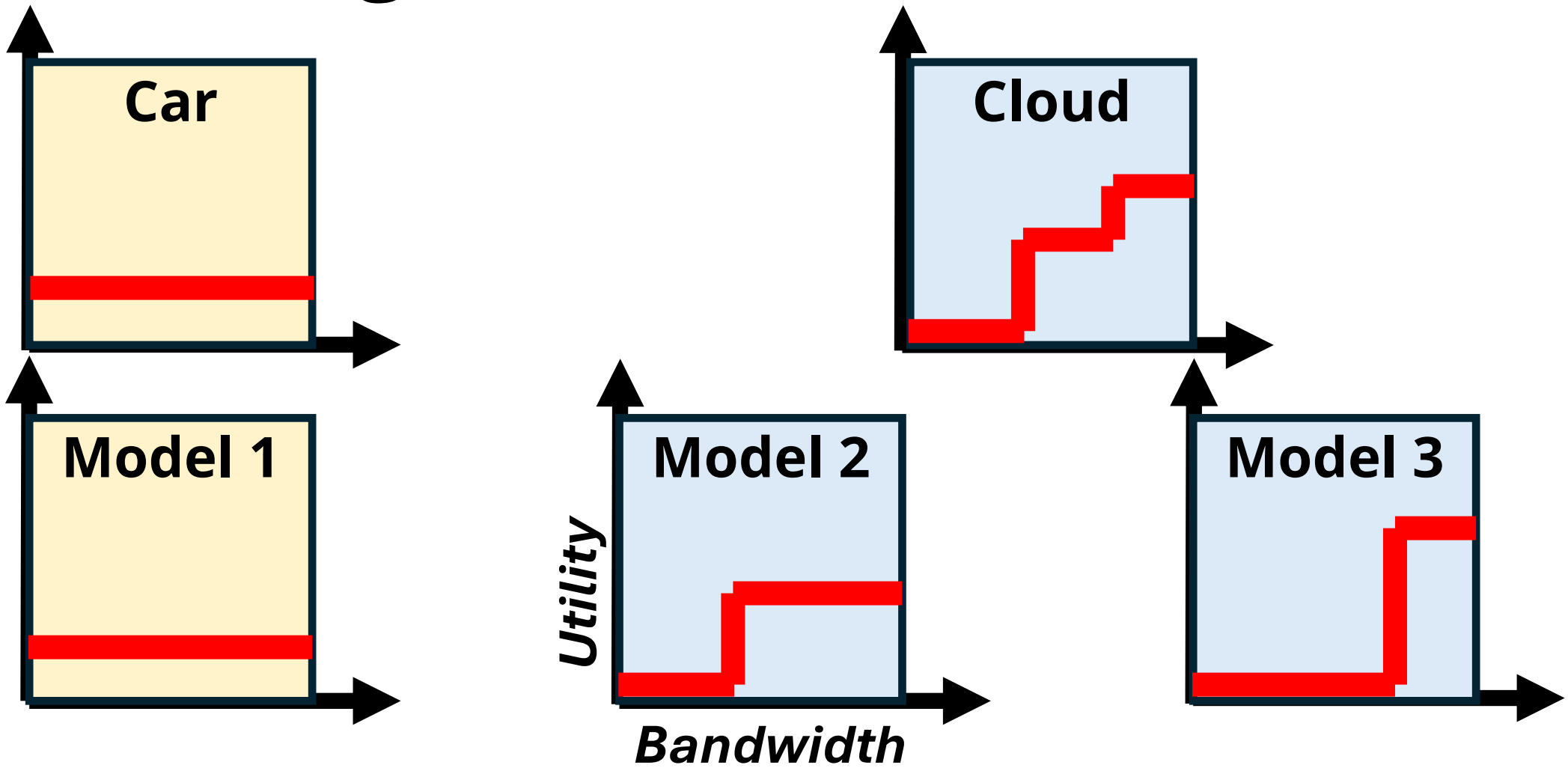




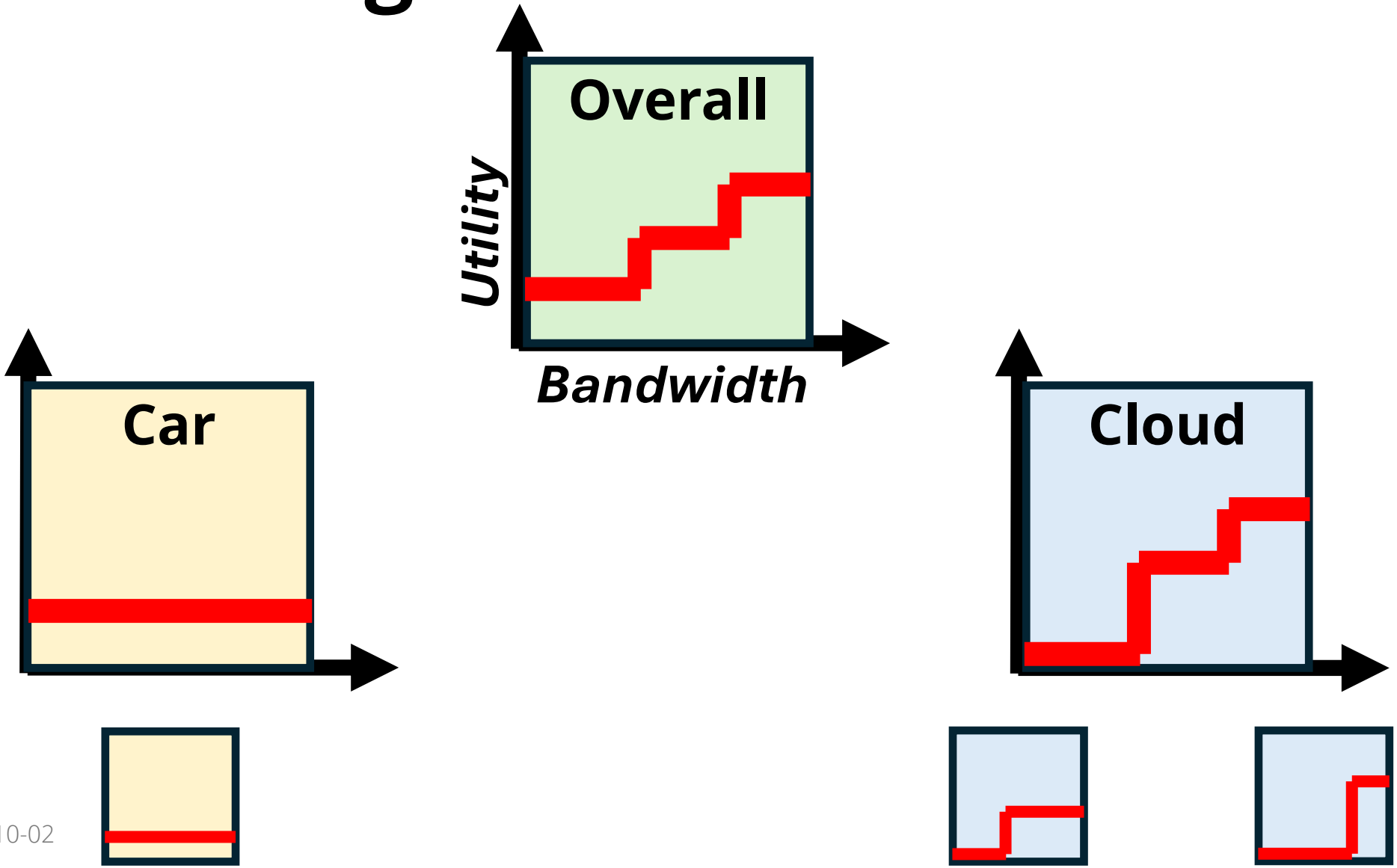
Prioritizing Services



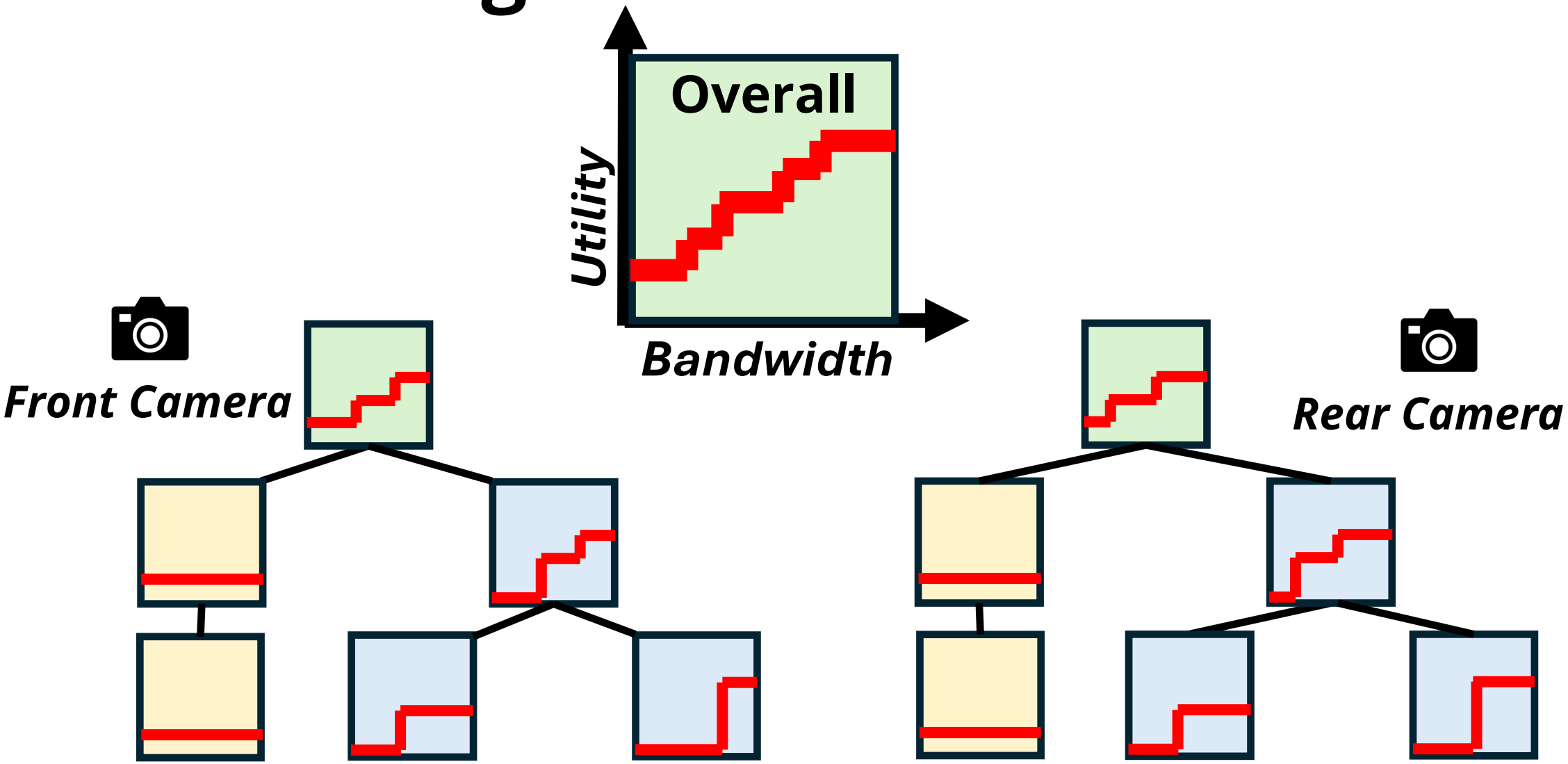
Prioritizing Services



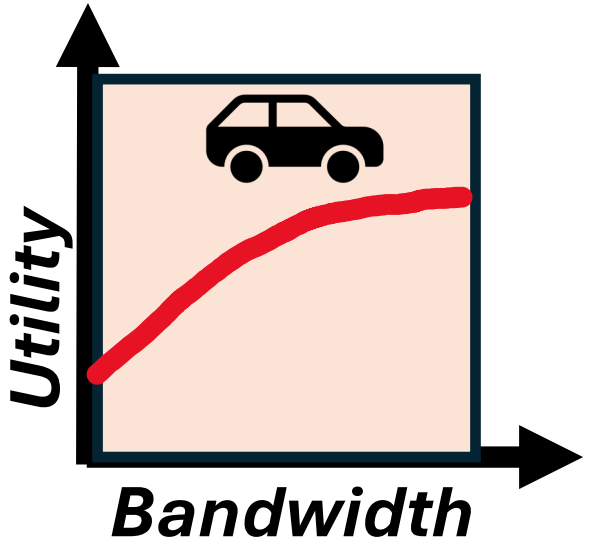
Prioritizing Services

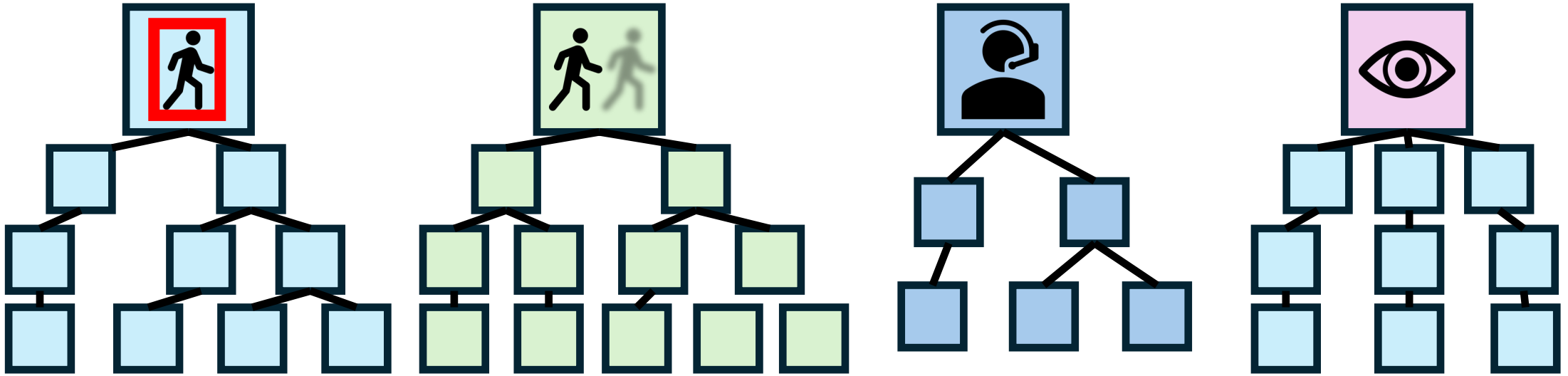


Prioritizing Services

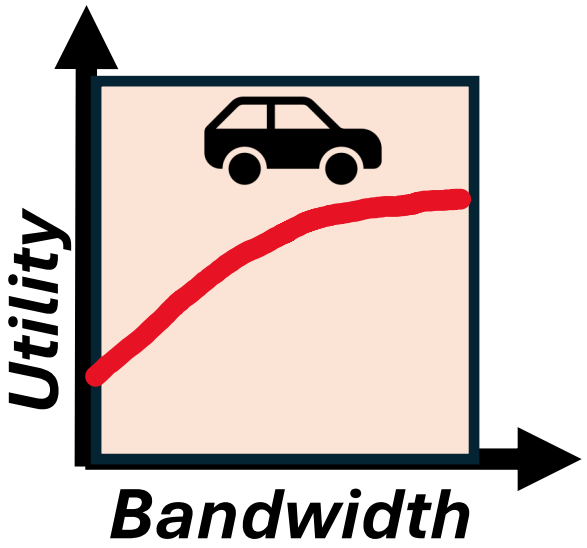


Prioritizing Services



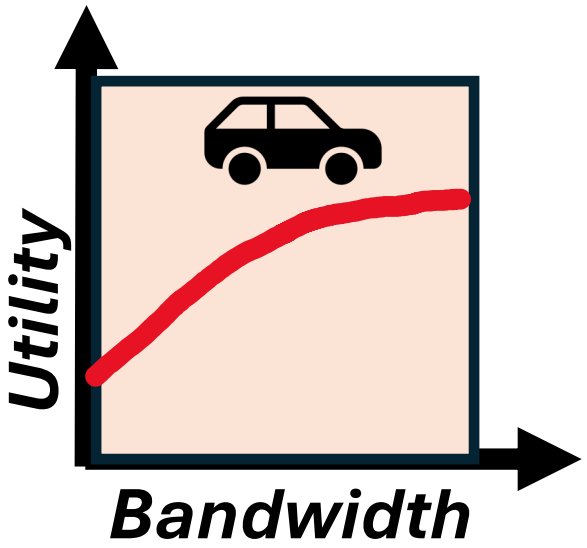


Open Problems



Generating utility curves various services.

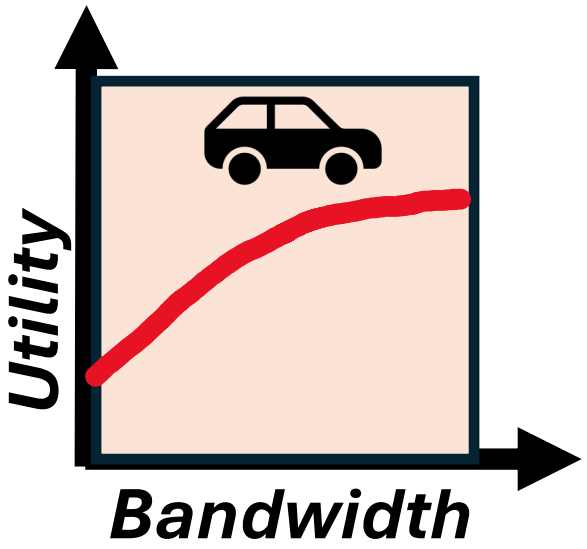
Open Problems



Generating utility curves various services.

Combining utility curves across (sub)services.

Open Problems



Generating utility curves various services.

Combining utility curves across (sub)services.

Dynamic utility curves when the benefit of the cloud changes.

Autonomous Driving ❤️ Cloud

- Network is the bottleneck
- Manage network via systems approaches
  - Speculative execution to address connectivity, connection quality
  - Bandwidth allocation to share network

IROS paper
tinyurl.com/cloud-avs

✉ **pschafhalter@berkeley.edu**