The Architectural Implications of Autonomous Driving

Slides by:
Sukrit Kalra
sukrit.kalra@berkeley.edu

Challenges

“Correct” Decisions

“Real-Time” Decisions

Power Budgets

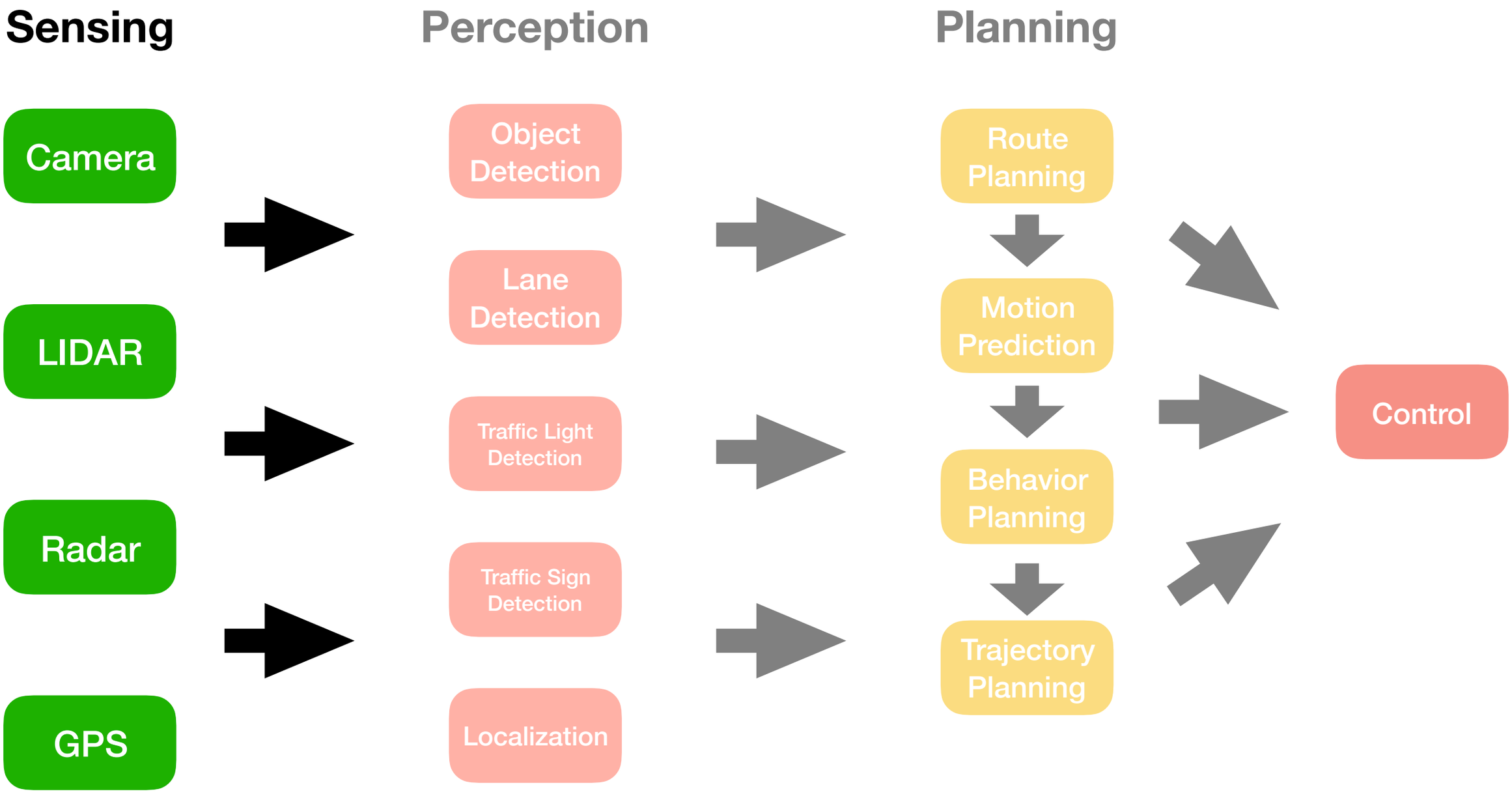
Challenges

“Correct” Decisions

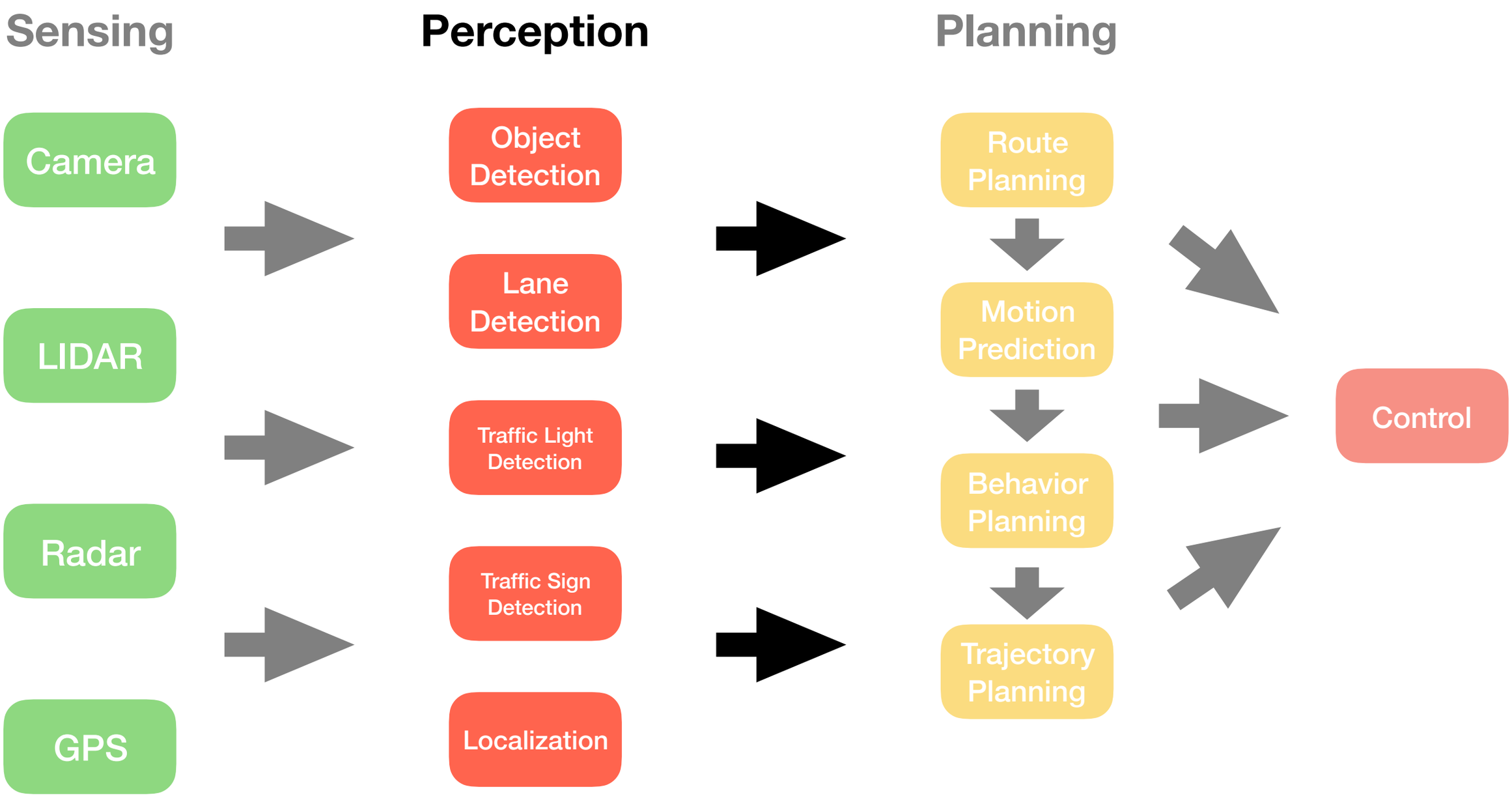
“Real-Time” Decisions

Power Budgets

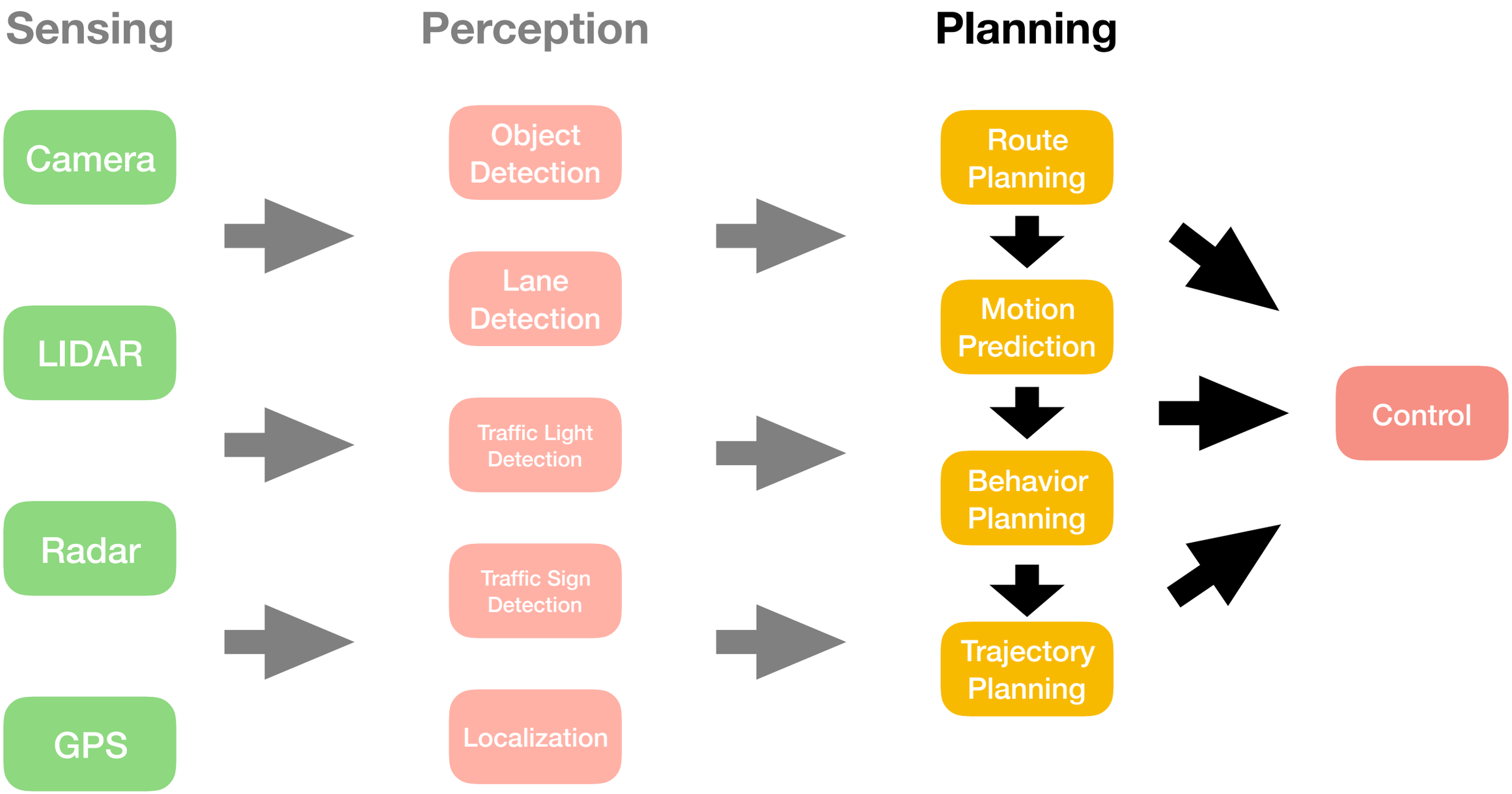
Computational Pipeline



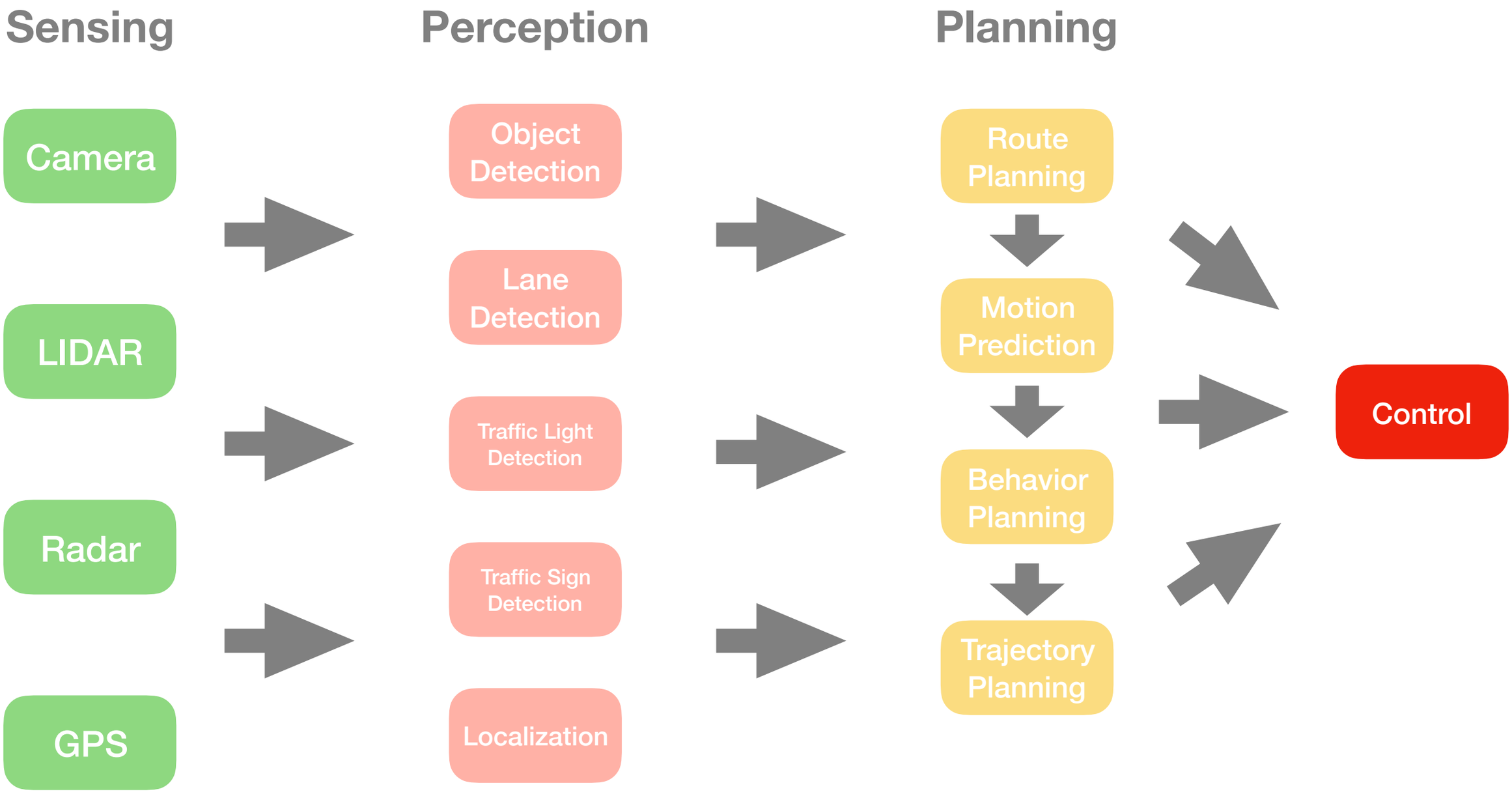
Computational Pipeline



Computational Pipeline



Computational Pipeline



Design Constraint: Performance

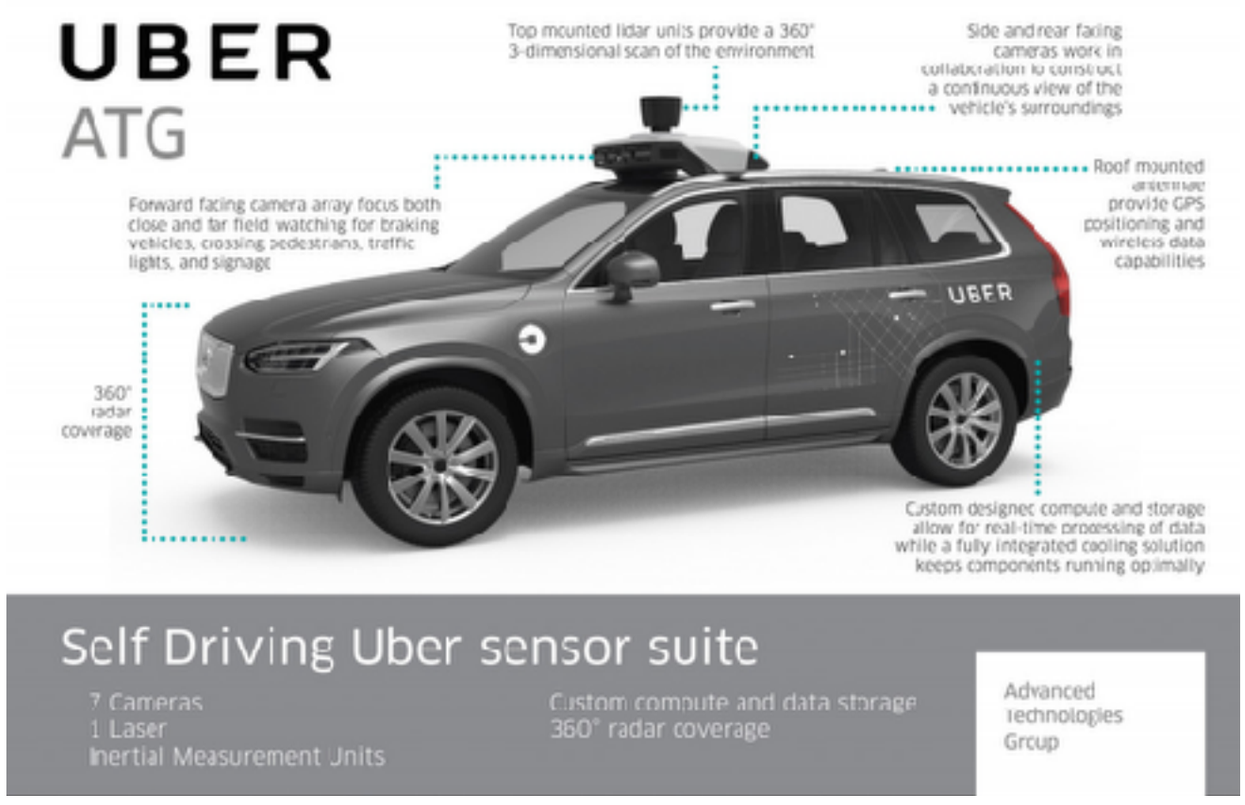
Sensing

Camera

LIDAR

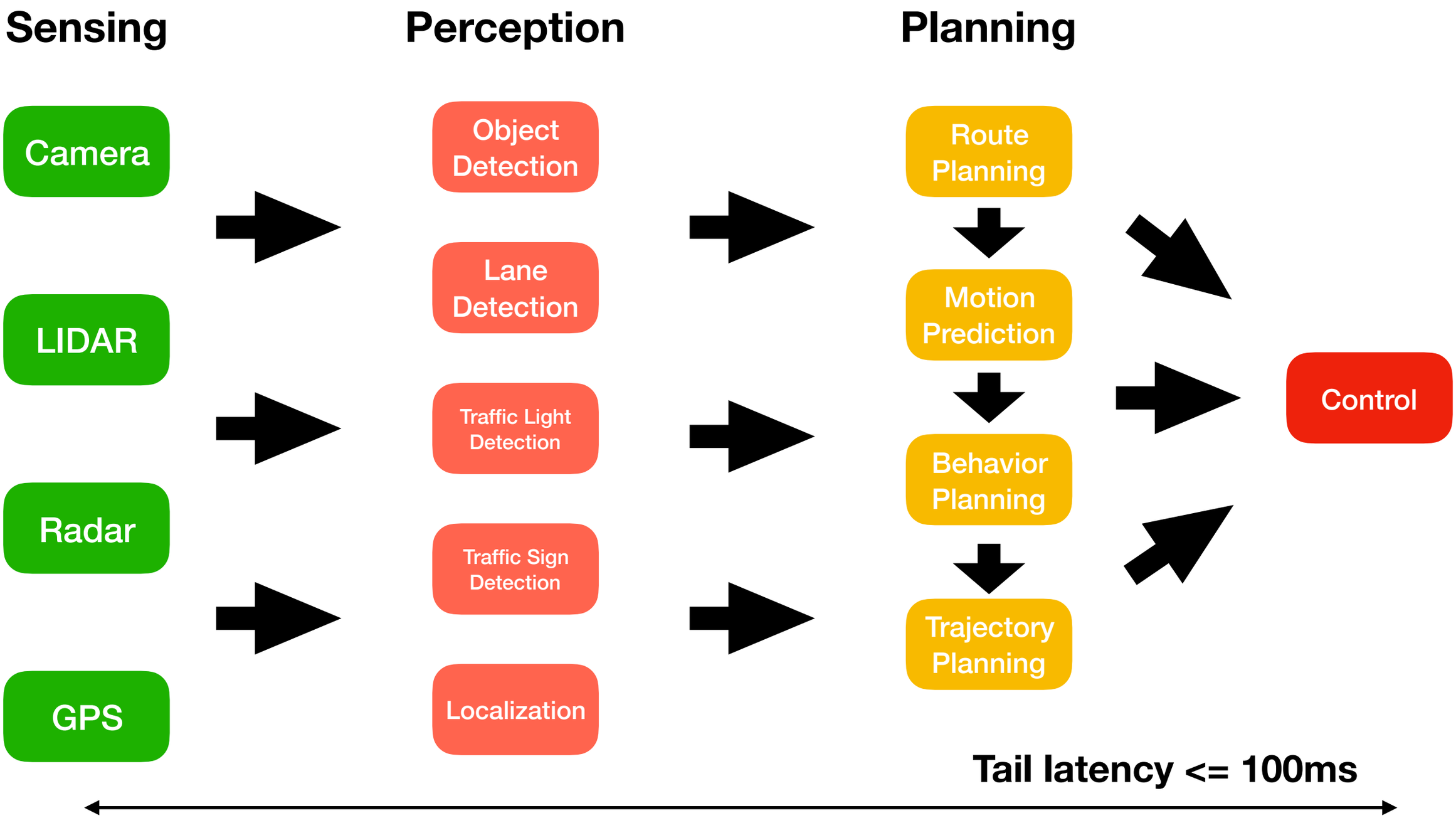
Radar

GPS



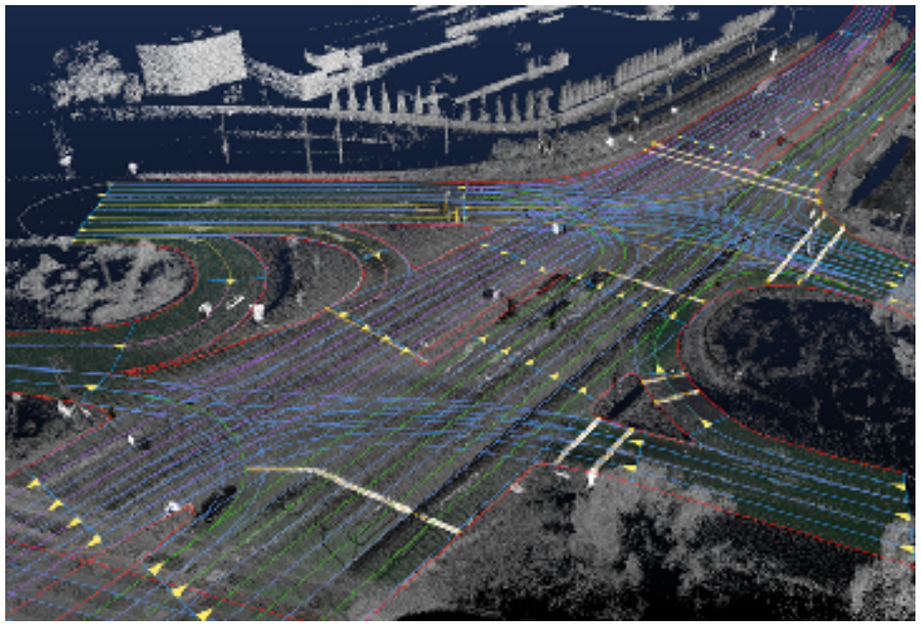
- **Sensors generate 1-2GB of data per second.**
- **Autonomous vehicle system should provide:**
  - **High Throughput**
  - **Low Latency**

Design Constraint: Predictability



Design Constraint: Storage

Localization



41 TB!

4TB of logging data generated by a car each day.

Design Constraint: Thermal and Power

JACK STEWART TRANSPORTATION 02.06.16 08:00 AM

SELF-DRIVING CARS USE CRAZY AMOUNTS OF POWER, AND IT'S BECOMING A PROBLEM

Shelley, a self-driving Audi TT developed by Stanford University, uses the brains in the trunk to speed around a racetrack autonomously

NIKKI KAHN/THE WASHINGTON POST/GETTY IMAGES

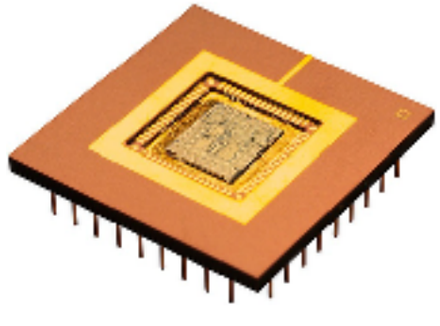
“Datacenter on wheels.”

Power requirements of up to 3kW.

Key Idea

GPUs

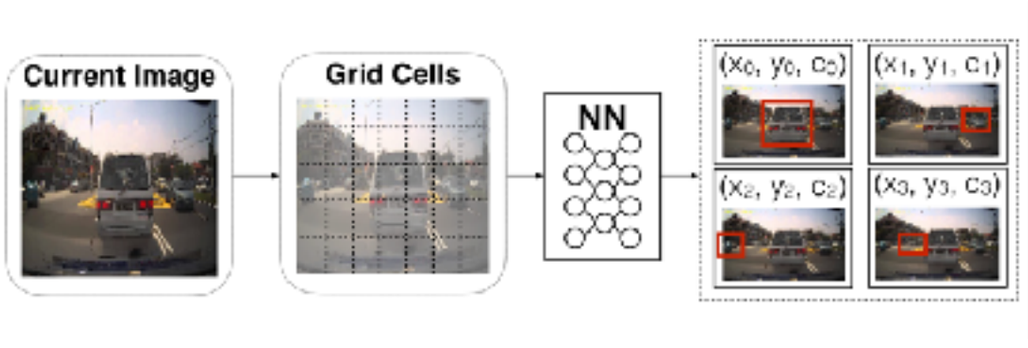
FPGAs



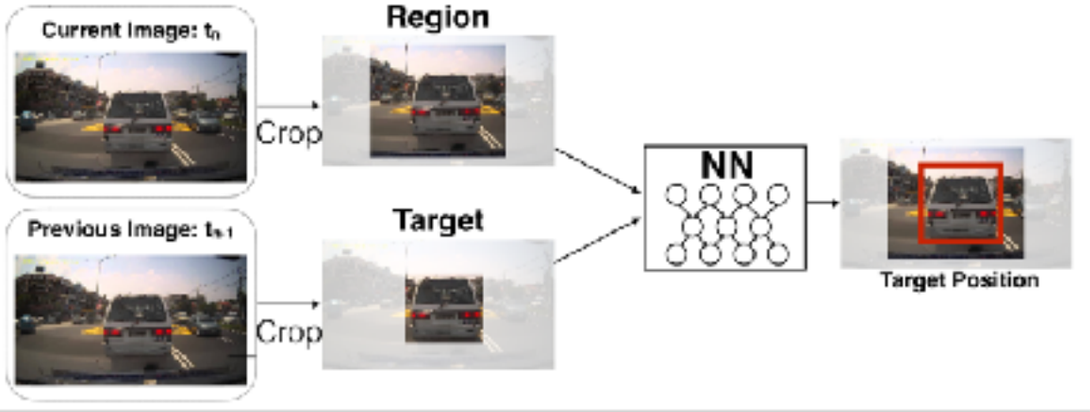
ASICs

Exploit different accelerator platforms to achieve predictability and performance while reducing the power requirements.

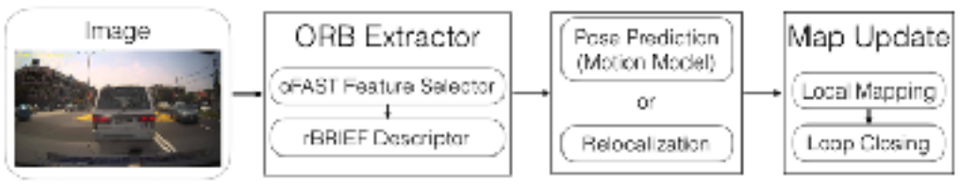
Implementation



Detector

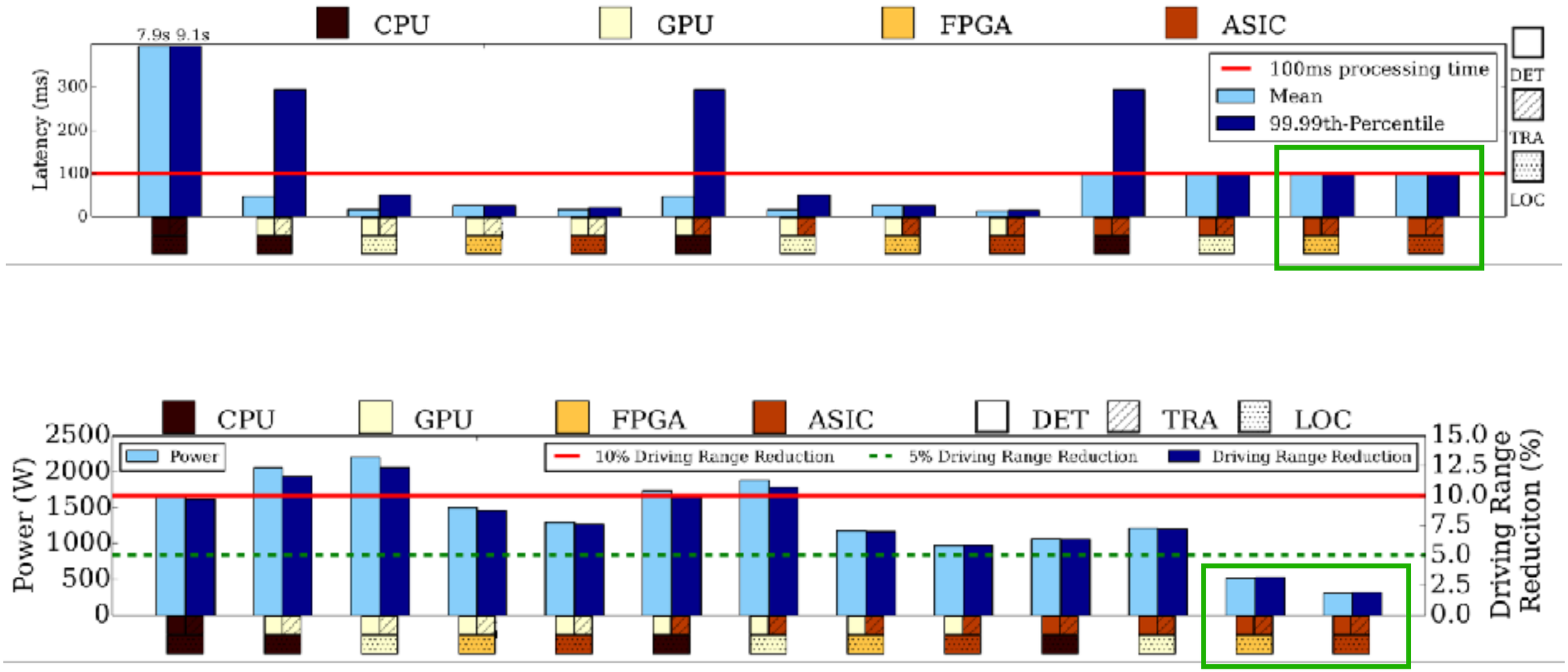


Tracker



Localization

Metrics of Success



Discussion

Are DNNs the major computational part of the perception pipeline?

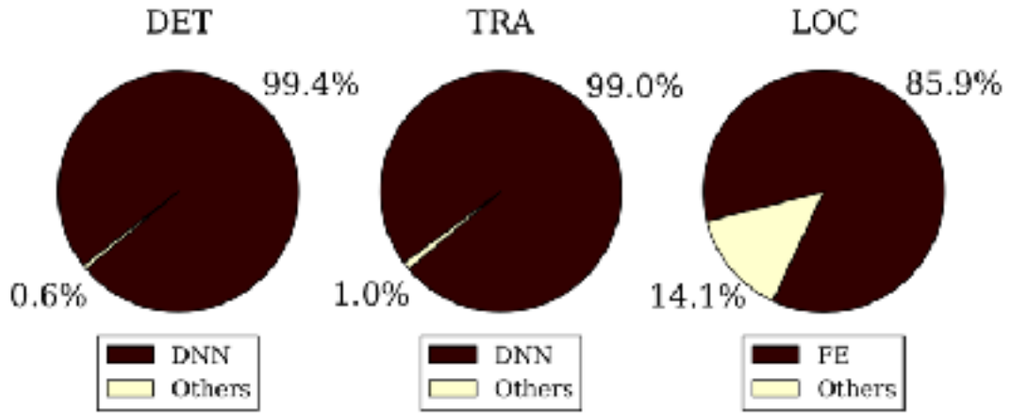
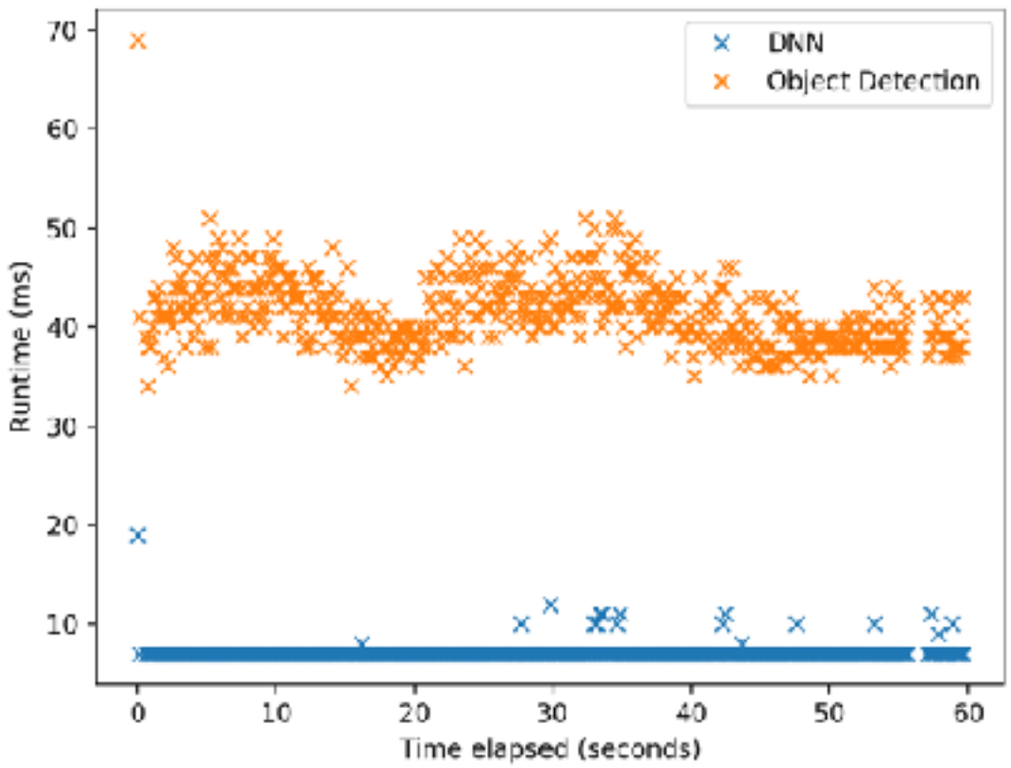


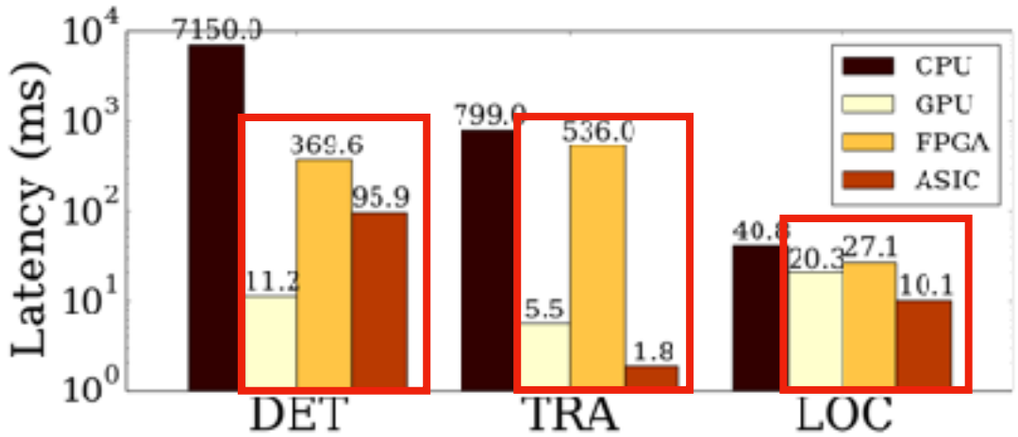
Figure 7. Cycle breakdown of the object detection (DET), object tracking (TRA) and localization (LOC) engines. The Deep Neural Networks (DNNs) portion in DET and TRA, and the Feature Extraction (FE) portion in LOC account for more than 94% of the execution in aggregation, which makes them ideal candidates for acceleration.



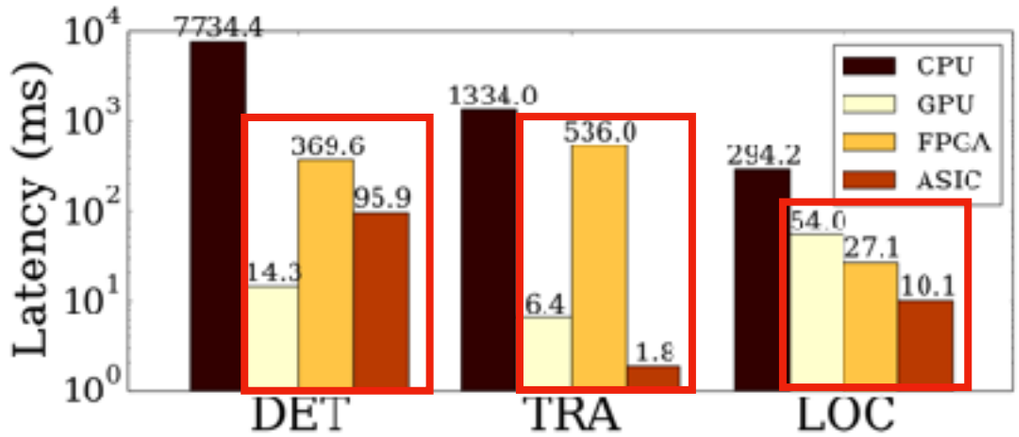
Discussion

What is the reason for zero variability in both ASICs and FPGAs?

What is the source of runtime variability for the GPU?



Mean Latency



99.99th Percentile Latency

Discussion

What about the cost of inter-component communication across the devices?