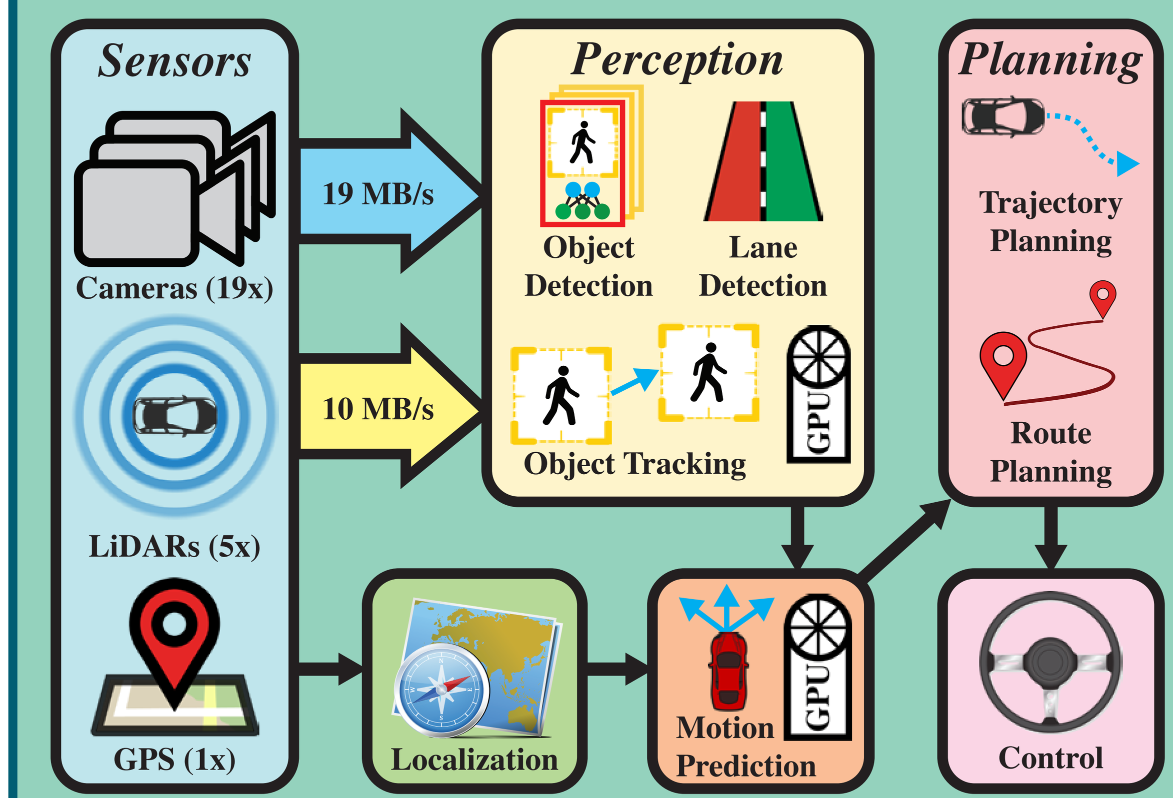
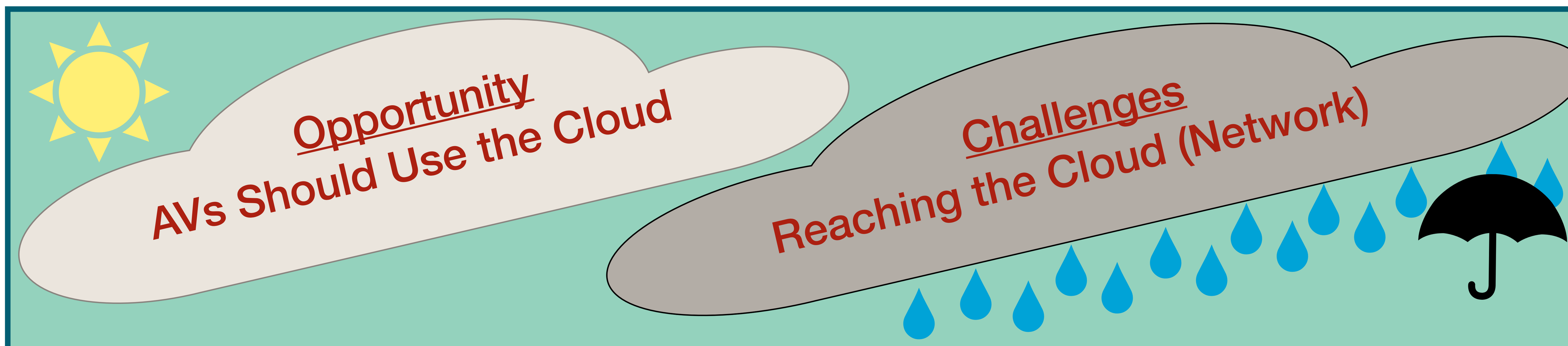
Towards Cloud-Assisted Autonomous Driving

Alexander Krentsel, Peter Schafhalter, Joseph E. Gonzalez, Sylvia Ratnasamy, Scott Shenker, Ion Stoica
UC Berkeley, akrentsel@berkeley.edu

Autonomous Vehicle Pipeline

- AV control system comprises Compound AI system of distinct ML-based modules executing tasks.





Plentiful Powerful Resources
Accelerate inference to run more accurate models, or the same models in less time

5G Networks Are Fast
Can transmit sensor data in real time

Manage network reliability
Handle connectivity, bandwidth, and latency fluctuations

Maximize cloud benefit
Prioritize critical services that gain the most from the cloud's capabilities

AV Requirements

- AVs target maximizing safety, which translates to maximally accurate individual components, which requires:

High Fidelity Data
Better information for decision-making

State-of-the-Art Computation
To enable running the most accurate models

Timely Results
Run in real-time to beat human reaction times

Human reaction time SLO: 390ms [0] to 1.2s [1]

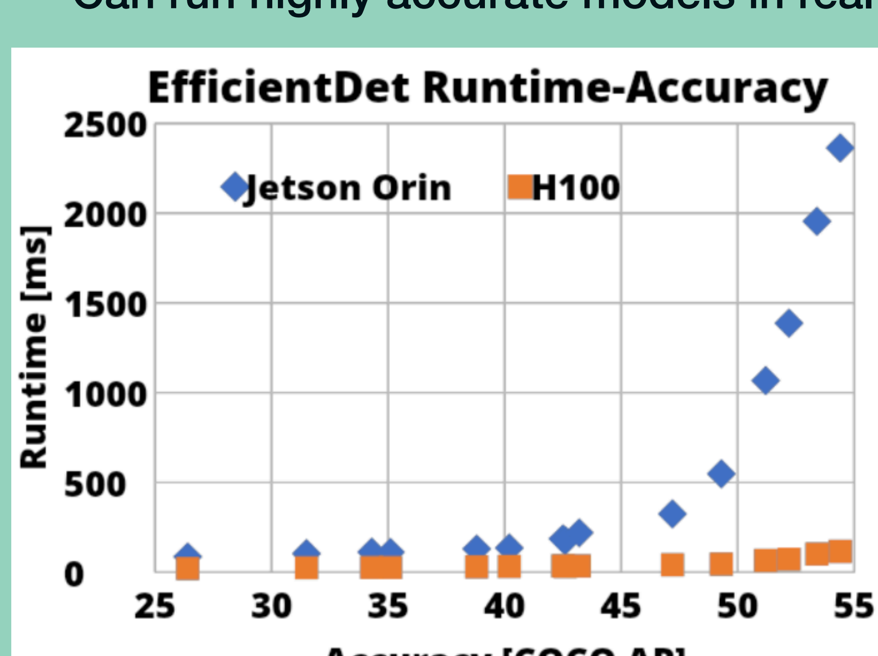
Compute Hardware

- Compute and data determine runtime and accuracy.
- State-of-the-art compute on-car is NVIDIA's Jetson Orin [2].

Limited in AVs
physical constraints include power, cooling, stability

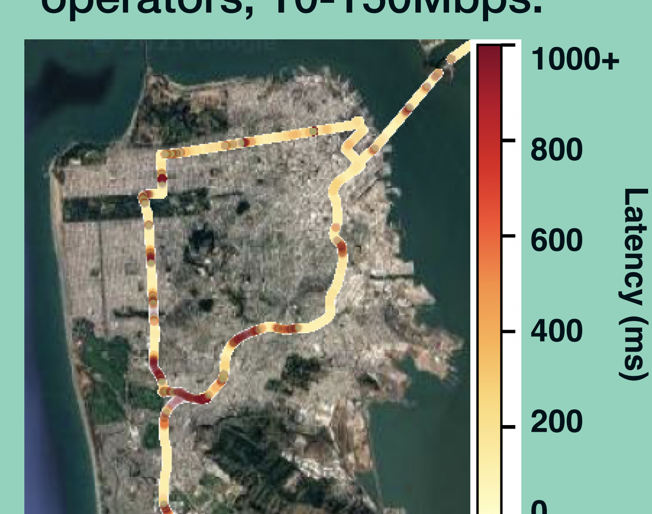
Plentiful in Cloud
Specialized data centers, frequent hardware upgrades

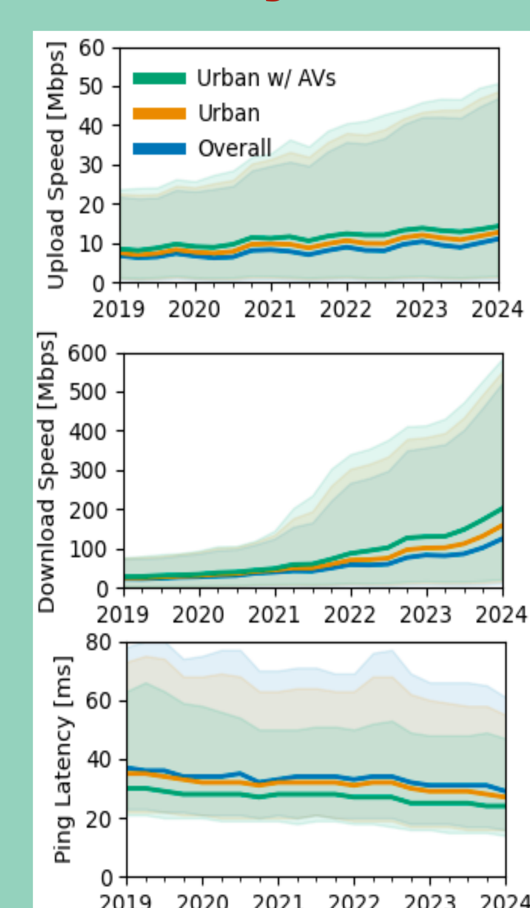
Powerful Cloud HW
Can run highly accurate models in real time



Network Reliability

- Network properties have <u>high variance across location/time</u>
- In San Francisco, latency:
  - Fast median: 68ms
  - Long Tail: 3027ms [99th]
- Uplink BW configured low by operators, 10-150Mbps.





Varying Cloud Benefit

- Each service gets variable benefit from leveraging the cloud
  - Based on model resource requirements and relative accuracies
- Feasibility depends on bandwidth available
  - ↑ bandwidth ⟹ ↓ network time ⟹ ++ cloud compute time
- Not enough bandwidth to use cloud for *all* services
  - must choose best subset to offload to cloud

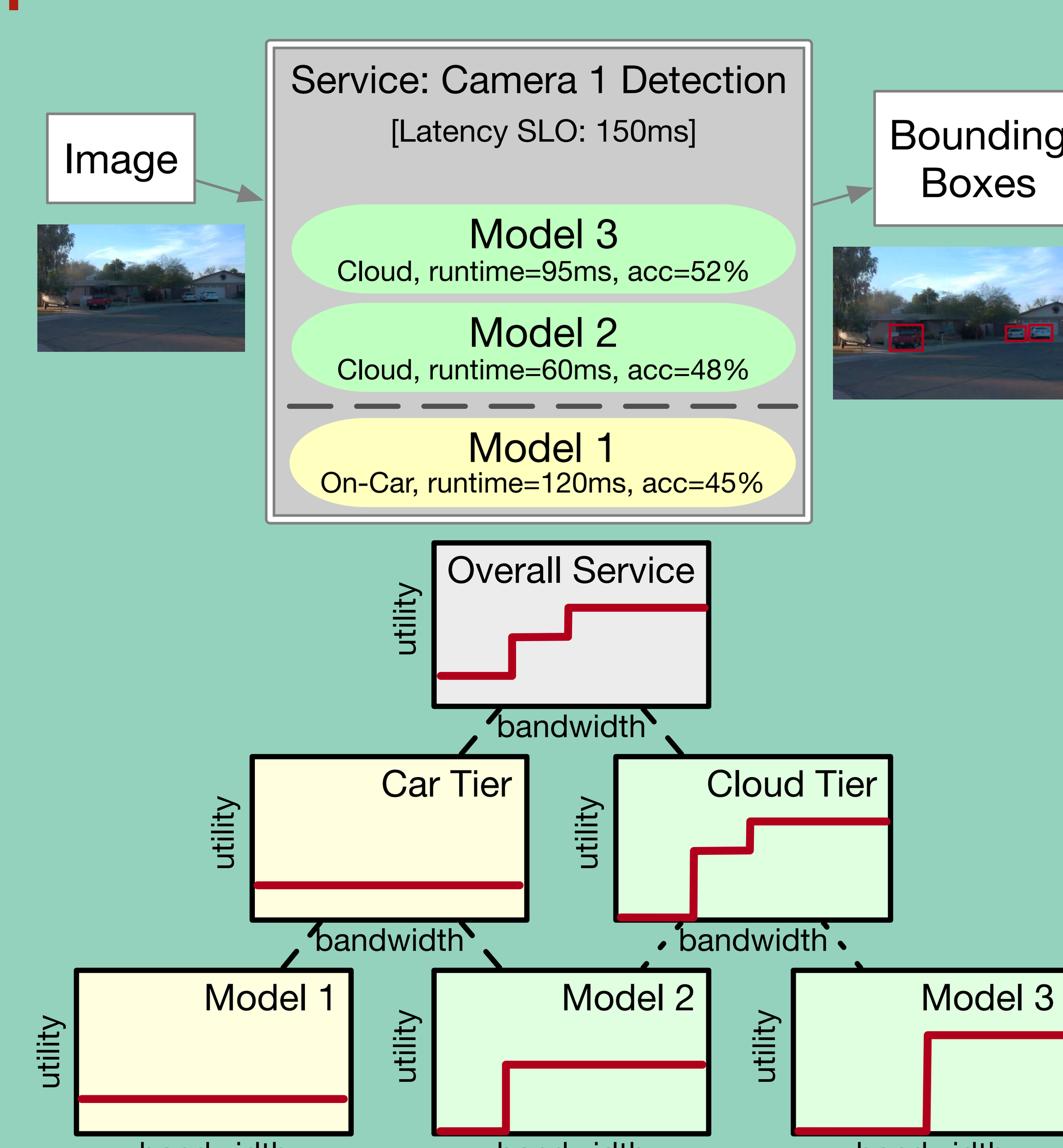
<u>Goal</u>: allocate bandwidth to subset of services that (1) can **feasibly run** in cloud and (2) get **maximal benefit**

EfficientDet Models		Network Latency (ms)	Remaining Computation Time (ms)
Model Name	*Input Size*		
ED0	512x512	44.84	105.16
ED1	640x640	47.56	102.44
ED2	768x768	50.89	99.11
ED3	896x896	54.82	95.18
ED4	1024x1024	59.36	90.64
ED5	1152x1152	64.50	85.50
ED6	1280x1280	70.25	79.75
ED7	1536x1536	83.56	66.44

Assuming 40ms ping and 100 Mbps bandwidth, latency consumed by each of the EfficientDet [5] family of models

A Tiered Approach

- Structure the AV pipeline as having *tiers* of execution for each service
  - Each tier has access to different resources
  - Local tier is 0 cost, cloud cost is bandwidth required to transmit input quickly enough to complete within deadline
- Utility curves to capture bandwidth benefit
  - Extract bandwidth utility curves based on service target latency SLO and accuracy of each model, adapting from Cao & Zegura's Utility curve model [3]
  - Compose utility curves [6] to get overall service utility curve
- Allocate bandwidth to maximize total utility
  - Alternatively can do max-min fair allocations, or other allocation based on policy
- Always run on-car model for reliability



References

[0] B. Wolfe, B. Seppelt, B. Mehler, B. Reimer, and R. Rosenholtz, "Rapid holistic perception and evasion of road hazards." Journal of experimental psychology, 2020.
[1] G. Johansson and K. Rumar, "Drivers' brake reaction times," Human factors, vol. 13, no. 1, pp. 23–27, 1971.
[2] "NVIDIA Introduces DRIVE AGX Orin," https://tinyurl.com/6pjsxzw7.
[3] Zhiruo Cao and E W Zegura. Utility max-min: an application-oriented bandwidth allocation scheme. In IEEE INFOCOM '99. Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies.
[4] P. Schafhalter, S. Kalra, L. Xu, J. E. Gonzalez, and I. Stoica, "Leveraging cloud computing to make autonomous vehicles safer," in 2023 IEEE/RSJ IROS, pp. 5559–5566.
[5] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[6] Alok Kumar *et. al* BwE: Flexible, hierarchical bandwidth allocation for WAN distributed computing. In Proceedings of the 2015 ACM SIGCOMM '15, pages 1–14, New York, NY, USA, August 2015.