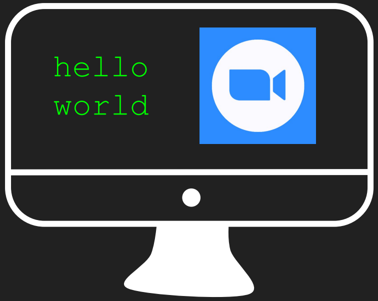
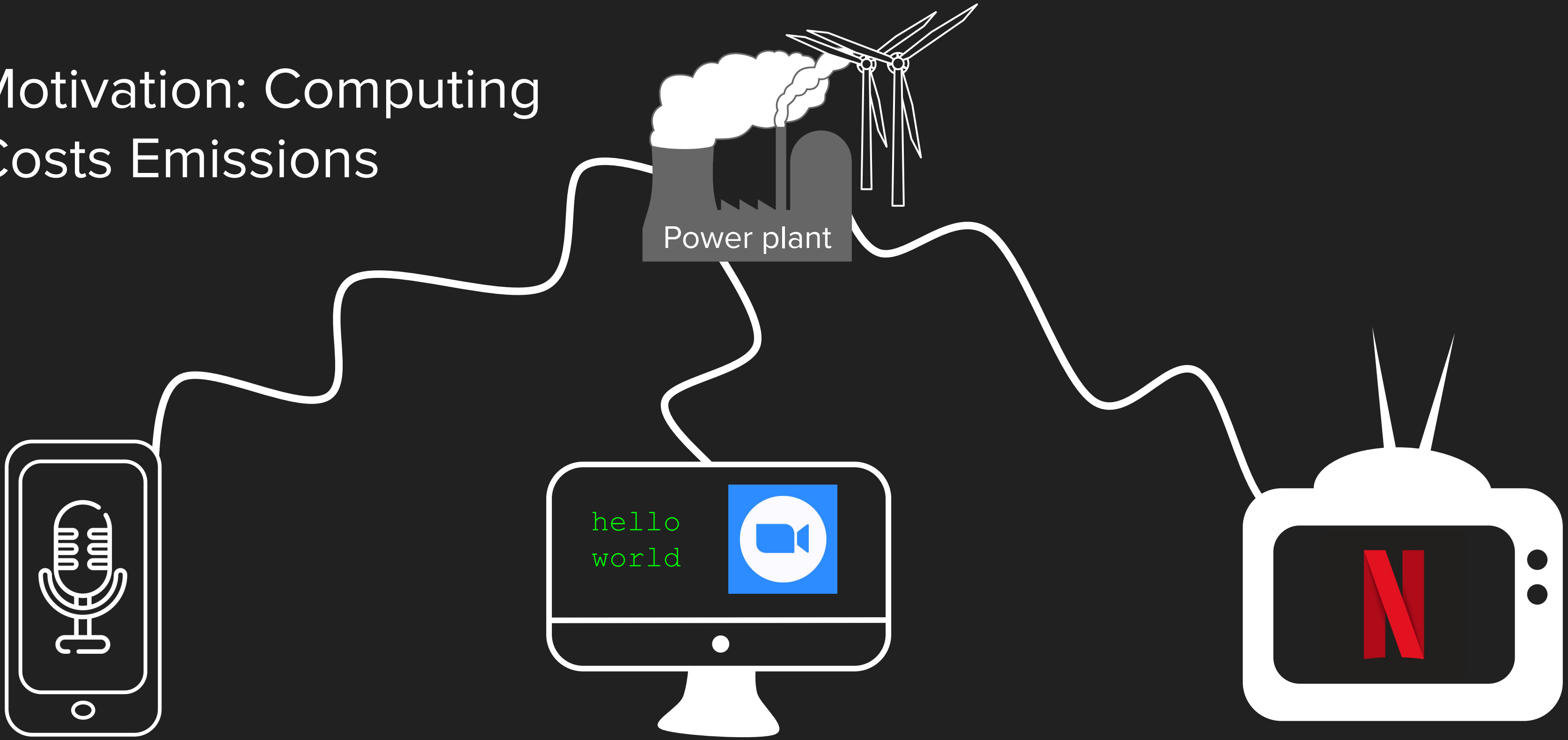
Data Centers on Wheels: Emissions from Computing Onboard Autonomous Vehicles

Soumya Sudhakar
Low Energy Autonomy and Navigation (LEAN) Group
CICS - November 2, 2022

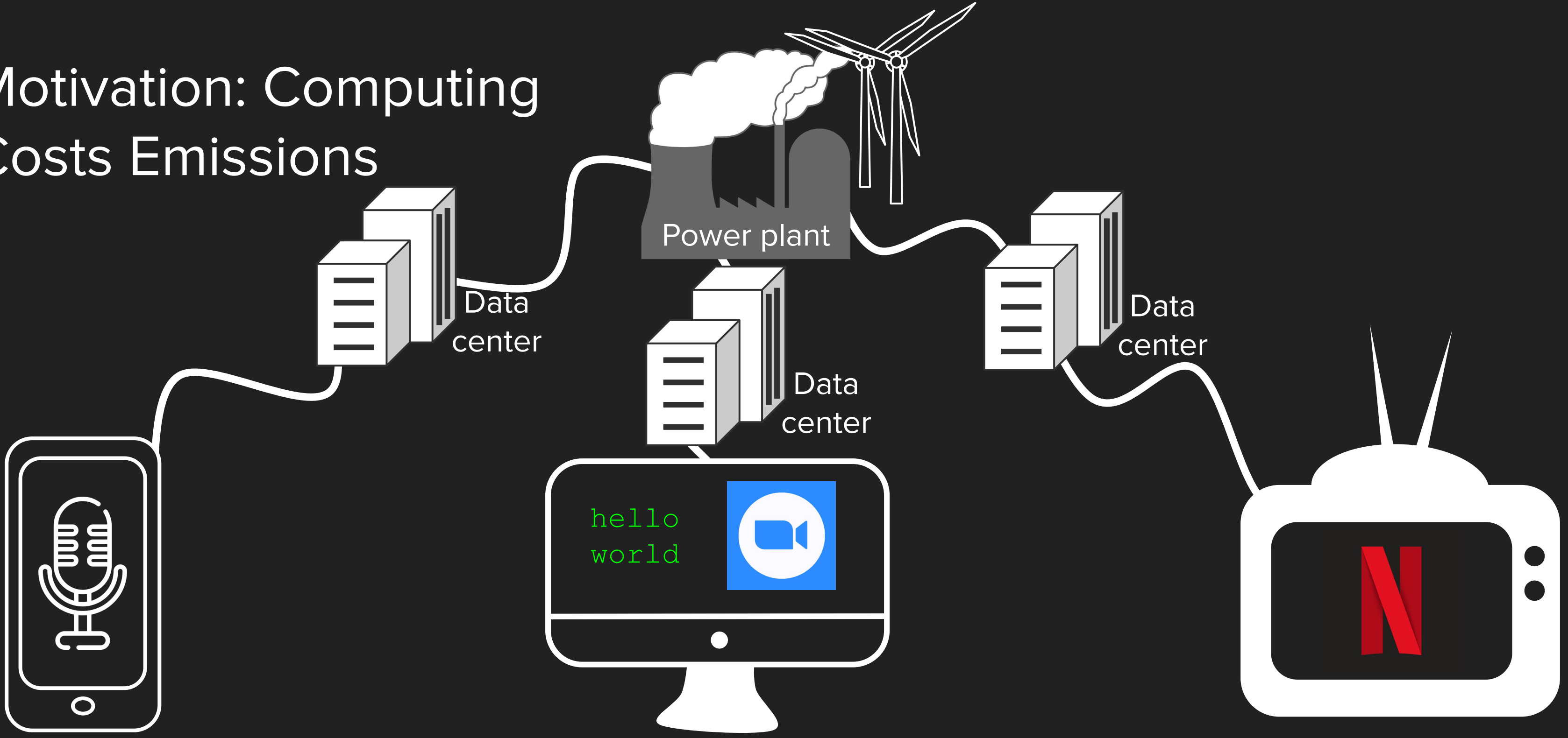
Motivation: Computing Costs Emissions



Motivation: Computing Costs Emissions



Motivation: Computing Costs Emissions



Concern for Emissions from Data Centers and ML Workloads

Data centers could cause serious environmental damage — if we don't regulate them now

NextWeb 2021

How to stop data centres from gobbling up the world's electricity

Nature 2018

ARTIFICIAL INTELLIGENCE

Training a single AI model can emit as much carbon as five cars in their lifetimes

Technology Review 2019

Concern for Emissions from Data Centers and ML Workloads

Data centers could cause serious environmental damage — if we don't regulate them now

NextWeb 2021

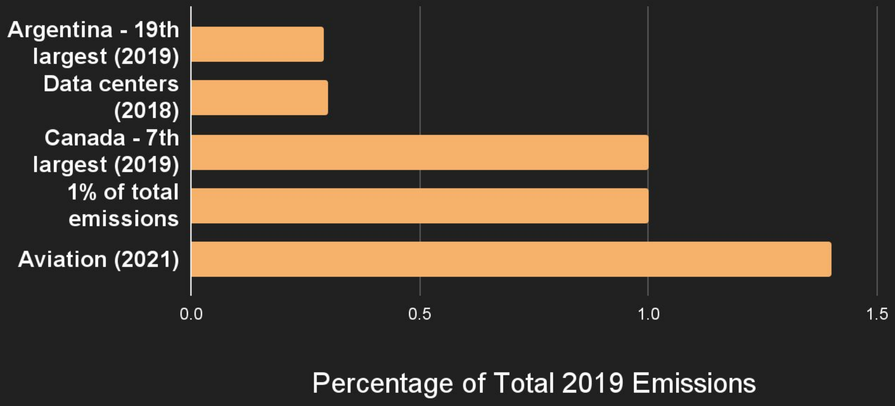
How to stop data centres from gobbling up the world's electricity

Nature 2018

ARTIFICIAL INTELLIGENCE

Training a single AI model can emit as much carbon as five cars in their lifetimes

Technology Review 2019



Concern for Emissions from Data Centers and ML Workloads

Data centers could cause serious environmental damage — if we don't regulate them now

NextWeb 2021

How to stop data centres from gobbling up the world's electricity

Nature 2018

ARTIFICIAL INTELLIGENCE

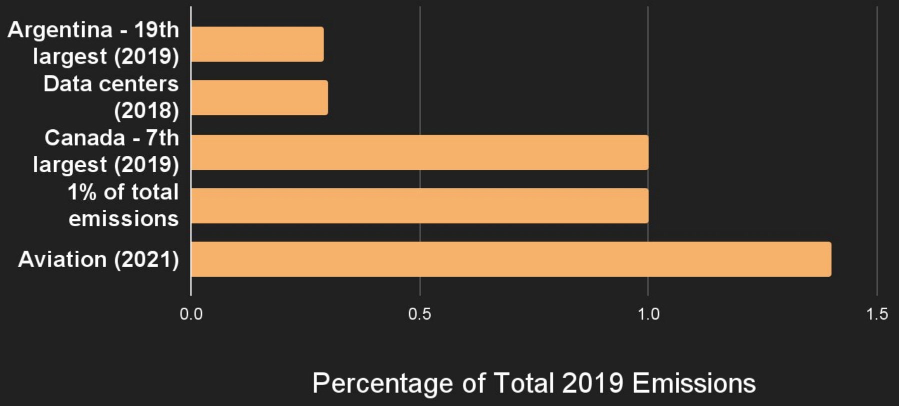
Training a single AI model can emit as much carbon as five cars in their lifetimes

Technology Review 2019

Why Not Use Self-Driving Cars as Supercomputers?

Autonomous vehicles use the equivalent of 200 laptops to get around. Some people want to tap that computing power to decode viruses or mine bitcoin.

Wired 2021



Concern for Emissions from Data Centers and ML Workloads

Data centers could cause serious environmental damage — if we don't regulate them now

NextWeb 2021

How to stop data centres from gobbling up the world's electricity

Nature 2018

ARTIFICIAL INTELLIGENCE

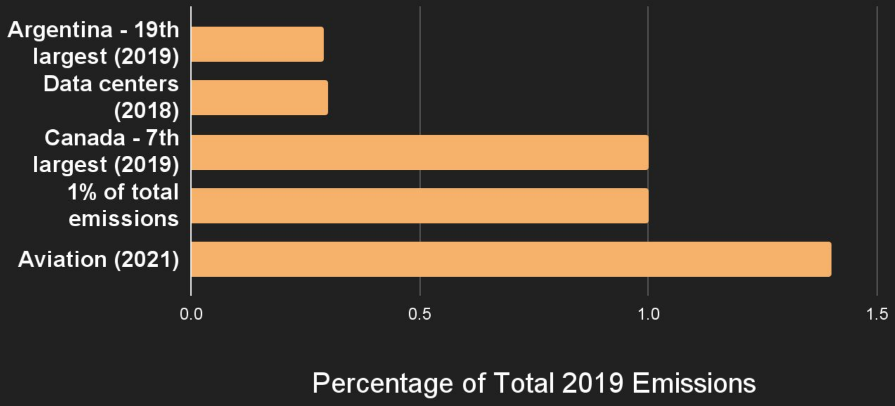
Training a single AI model can emit as much carbon as five cars in their lifetimes

Technology Review 2019

Why Not Use Self-Driving Cars as Supercomputers?

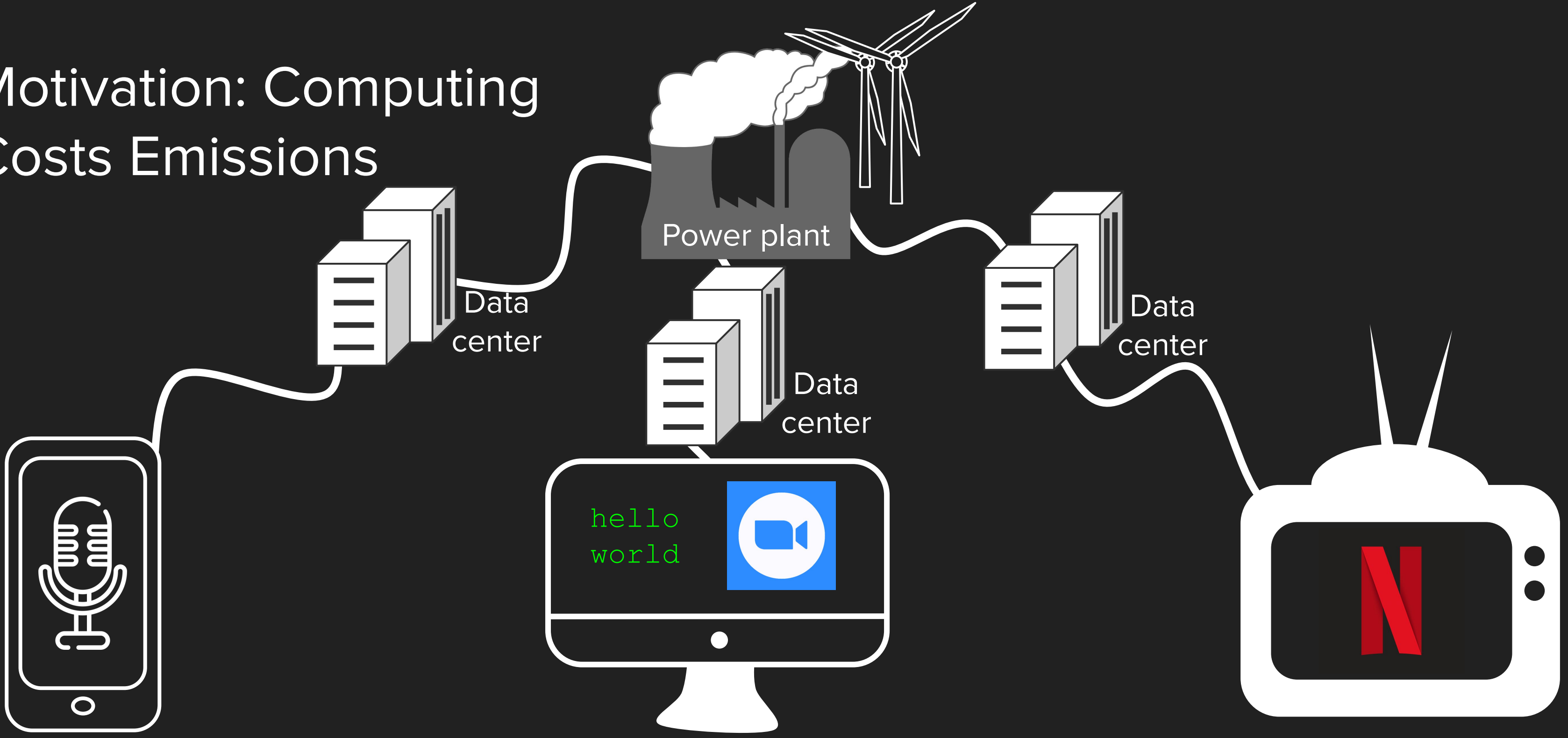
Autonomous vehicles use the equivalent of 200 laptops to get around. Some people want to tap that computing power to decode viruses or mine bitcoin.

Wired 2021

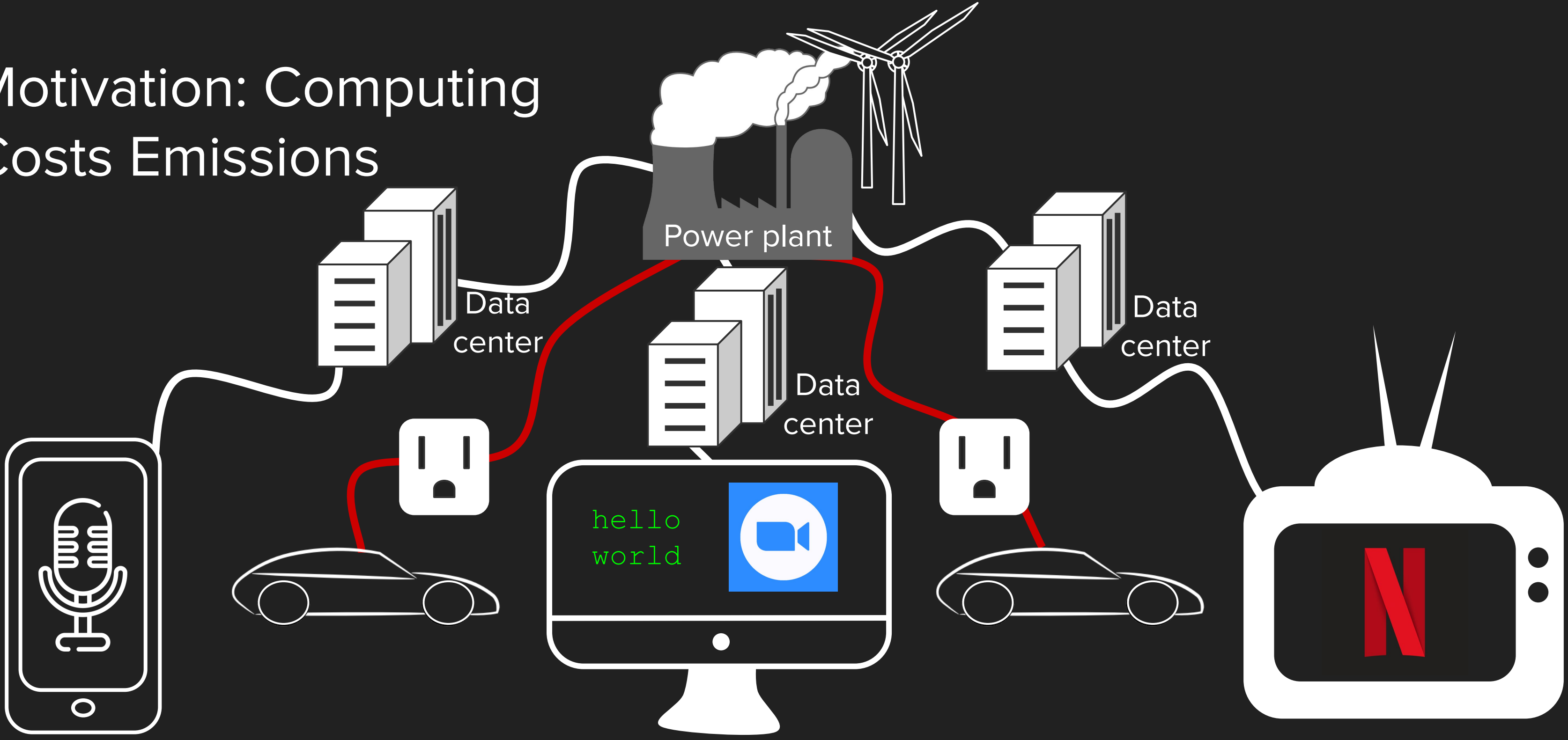


- "[T]rillions of inference per day across Facebook's data centers" (Wu et al. 2021)
- Autonomous vehicles (AVs) w/ 10 deep neural network (DNN) inferences at 60 Hz on 10 cameras:
  - One AV: 21.6 million inferences per hour driven
  - One billion AVs: **21.6 quadrillion** inferences per hour driven!

Motivation: Computing Costs Emissions



Motivation: Computing Costs Emissions



Could the carbon emissions from computing onboard a global fleet of AVs be significant?

Increased Attention to Emissions from Computing

Emissions from...	2018		2020	
Significant computational workloads	Data center emissions [Jones et al. (Nature 2018)]	Training large NLP models [Strubell et al. (2018)]	Carbon emissions and large NN training [Patterson et al. (2021)]	Sustainable machine learning [Wu et al. (ML and Systems 2022)]
Scaling of number of devices			Lifecycle analysis of mobile phone's carbon [Gupta et al. (IEEE Micro 2022)]	Architecture carbon modeling tool [Gupta et al. (ISCA 2022)]
Emerging applications	Bitcoin emissions [Mora et al. (Nature Climate Change 2018)]	Lower Bitcoin emissions [Masanet et al. (Nature Climate Change 2019)]		

Increased Attention to Emissions from Computing

Emissions from... **2018** **2020**

Significant computational workloads

- Data center emissions [Jones et al. (Nature 2018)]
- Training large NLP models [Strubell et al. (2018)]
- Carbon emissions and large NN training [Patterson et al. (2021)]
- Sustainable machine learning [Wu et al. (ML and Systems 2022)]

Scaling of number of devices

- Lifecycle analysis of mobile phone's carbon [Gupta et al. (IEEE Micro 2022)]
- Architecture carbon modeling tool [Gupta et al. (ISCA 2022)]

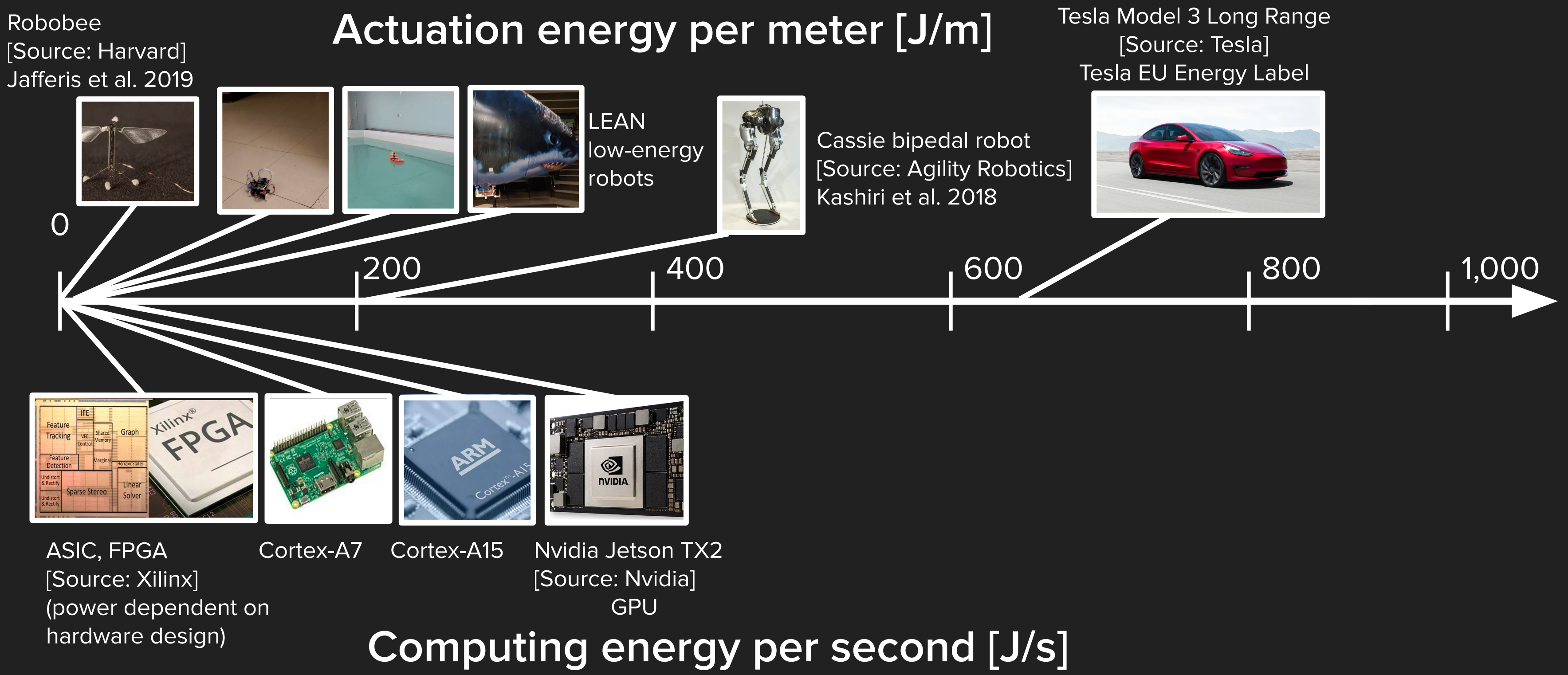
Emerging applications

- Bitcoin emissions [Mora et al. (Nature Climate Change 2018)]
- Lower Bitcoin emissions [Masanet et al. (Nature Climate Change 2019)]

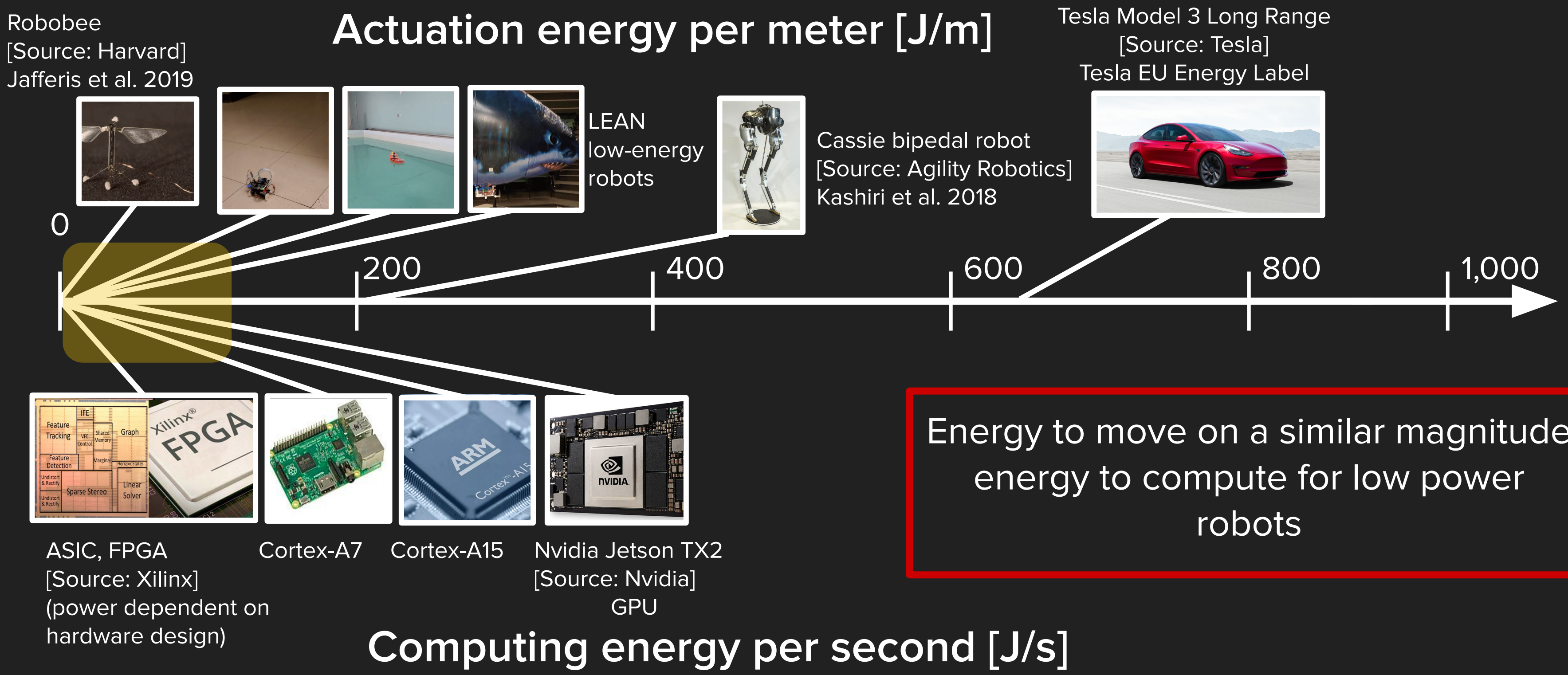
Emissions from AVs (this work)

[Sudhakar et al. (IEEE Micro 2023)]

Computing Energy vs. Actuation Energy in AVs

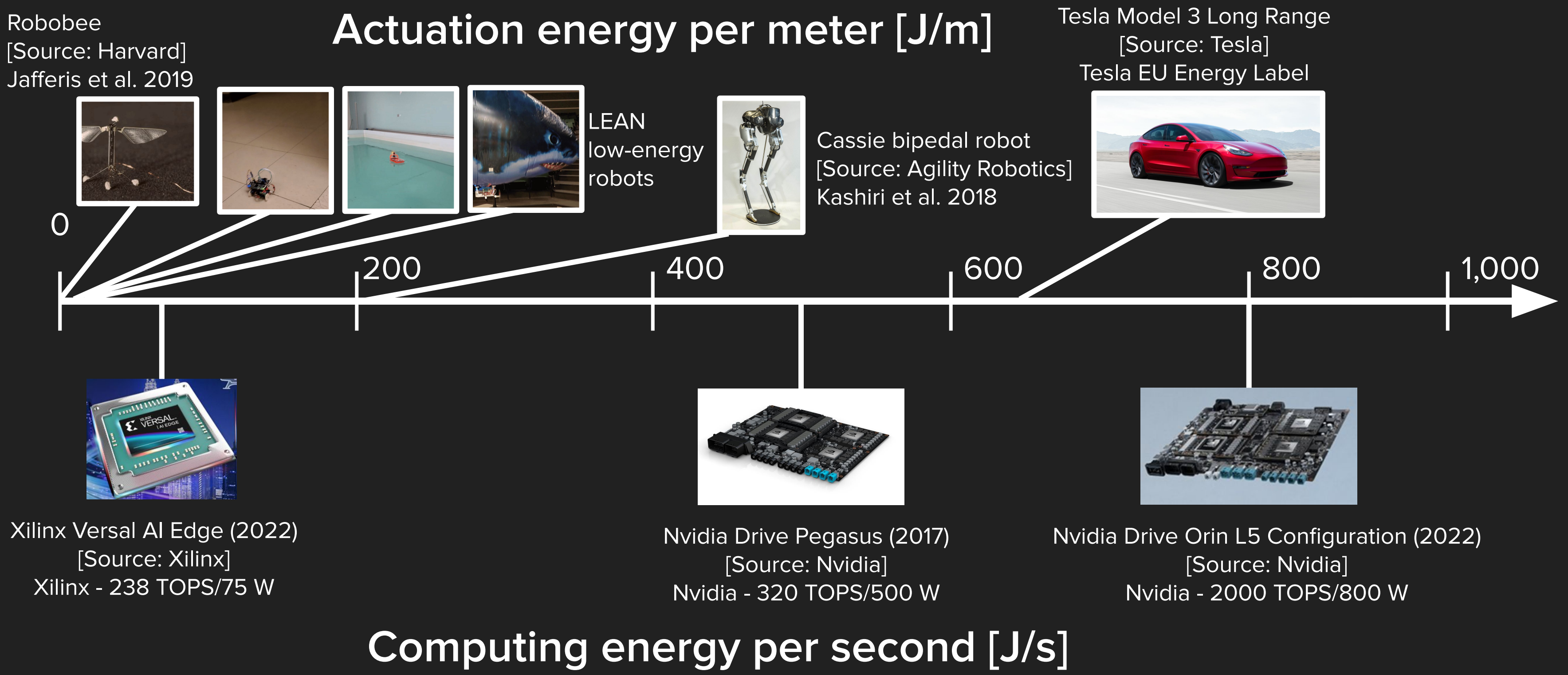


Computing Energy vs. Actuation Energy in AVs



Energy to move on a similar magnitude energy to compute for low power robots

Computing Energy vs. Actuation Energy in AVs



Computing energy may be also comparable to actuation energy for AVs!

Background: Levels of Autonomy

0	1	2	3	4	5
No Automation	**Driver Assistance**	**Partial Automation**	**Conditional Automation**	**High Automation**	**Full Automation**
Zero autonomy; the driver performs all driving tasks.	Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.	Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.	Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.	The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.	The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

Source: SAE/NHTS/NIST

Background: Levels of Autonomy

0	1	2	3	4	5
No Automation	**Driver Assistance**	**Partial Automation**	**Conditional Automation**	**High Automation**	**Full Automation**
Zero autonomy; the driver performs all driving tasks.	Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.	Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.	Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.	The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.	The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

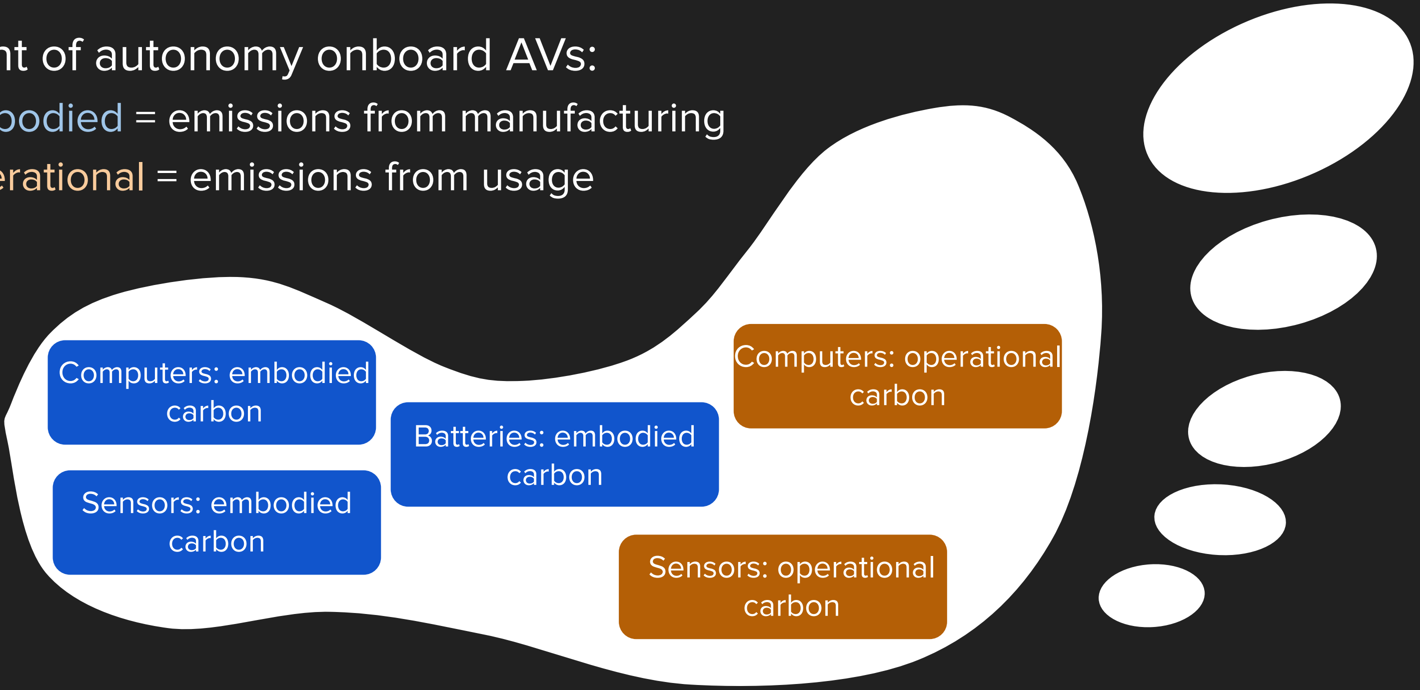
Source: SAE/NHTS/NIST

Scope of Analysis: Operational Carbon

Footprint of autonomy onboard AVs:

embodied = emissions from manufacturing

operational = emissions from usage



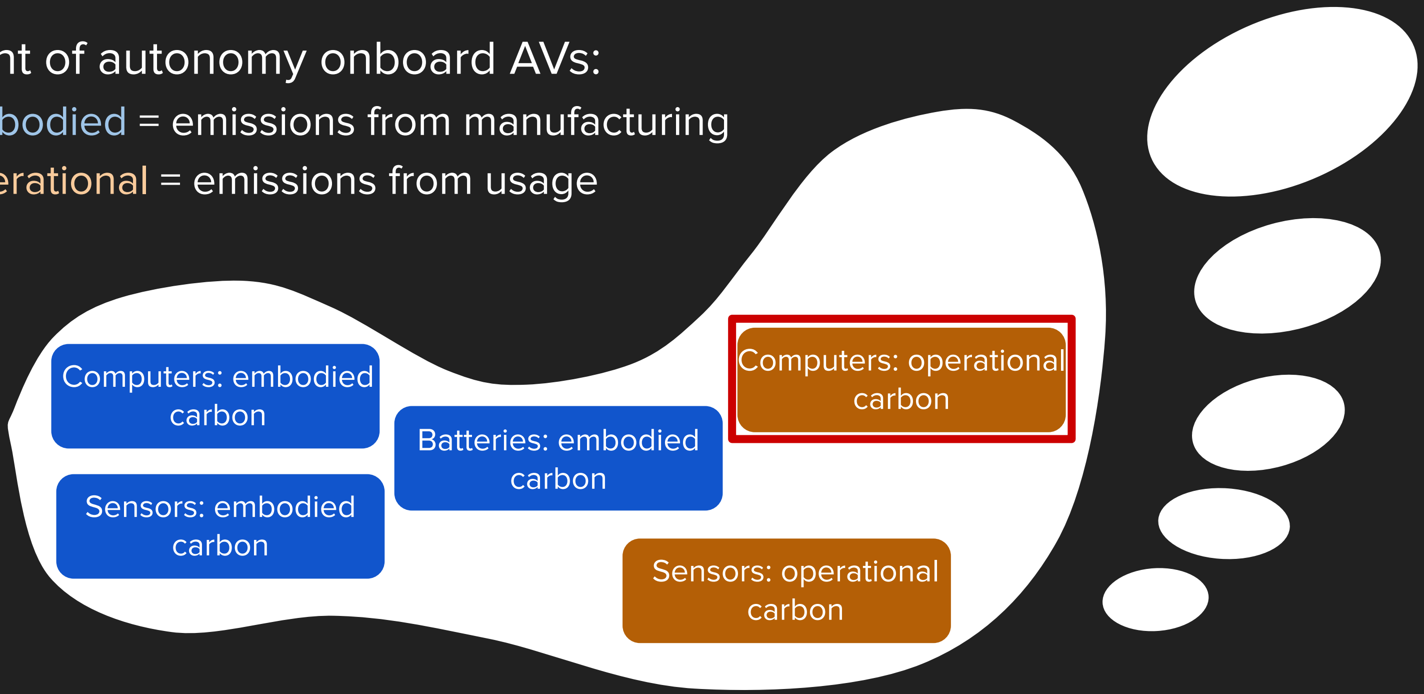
*Not including emissions from prototyping/training AV software stack

Scope of Analysis: Operational Carbon

Footprint of autonomy onboard AVs:

embodied = emissions from manufacturing

operational = emissions from usage



*Not including emissions from prototyping/training AV software stack

Operational Emissions from Computing Onboard AVs

Emissions from computing
onboard AVs (CO_2 eq. tons/yr)

$$G = \alpha NQIP$$

Operational Emissions from Computing Onboard AVs

Emissions from computing onboard AVs (CO_2 eq. tons/yr)

$$G = \alpha NQIP$$

Number of AVs

Operational Emissions from Computing Onboard AVs

Emissions from computing onboard AVs (CO_2 eq. tons/yr)

Average time driven per AV [hrs/day]

$$G = \alpha N Q I P$$

Number of AVs

Operational Emissions from Computing Onboard AVs

$$G = \alpha N Q I P$$

- G: Emissions from computing onboard AVs (CO_2 eq. tons/yr)
- N: Number of AVs
- Q: Average time driven per AV [hrs/day]
- I: Carbon intensity of power source [CO_2 eq. grams/kWh]

Operational Emissions from Computing Onboard AVs

Emissions from computing onboard AVs (CO_2 eq. tons/yr)

Average time driven per AV [hrs/day]

Average computer power [W]

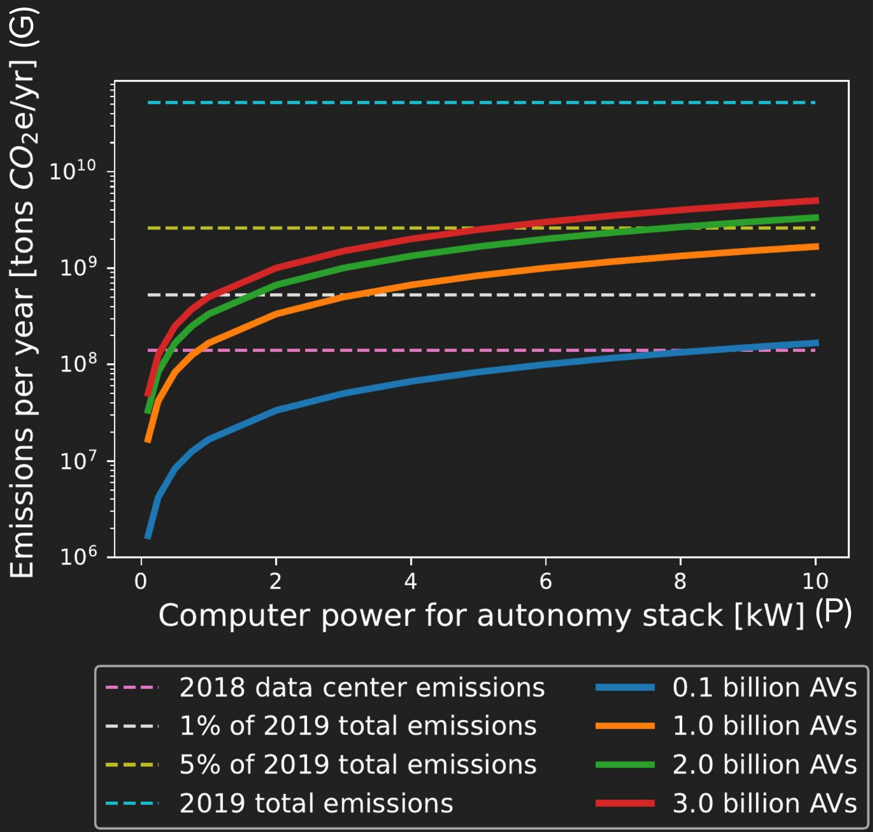
$$G = \alpha N Q I P$$

Constant for unit conversion

Number of AVs

Carbon intensity of power source [CO_2 eq. grams/kWh]

Emissions from Computing Onboard AVs



Assume 2020 global average carbon intensity and each AV driven for 1 hour per day

Emissions from computing onboard AVs (CO_2 eq. tons/yr)

Average time driven per AV [hrs/day]

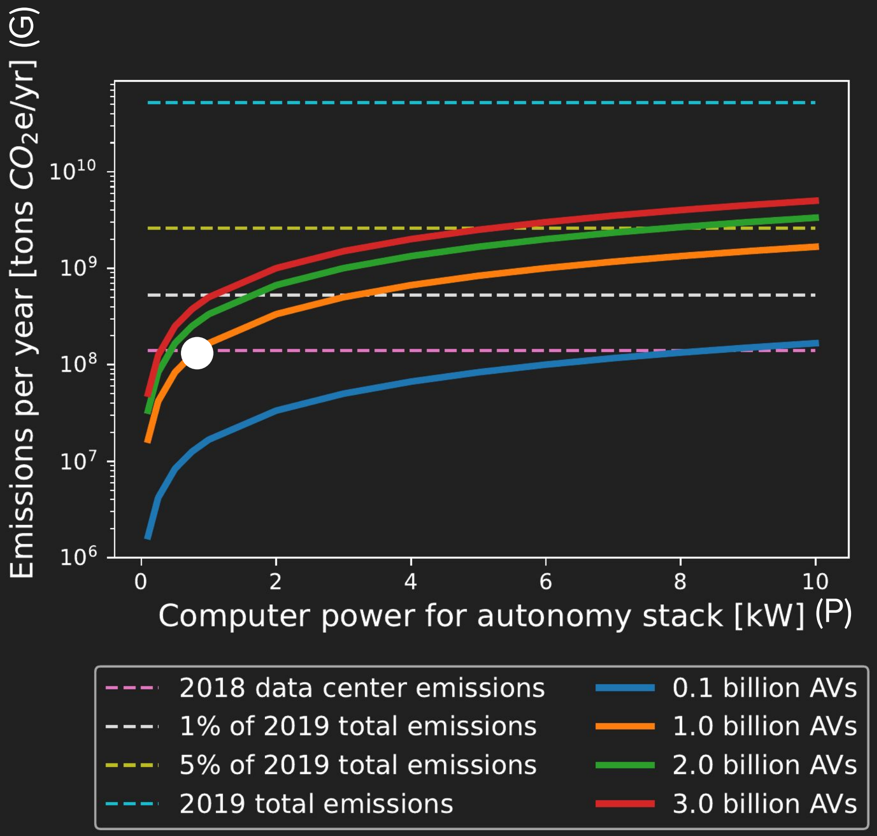
Average computer power [W]

$$G = \alpha NQIP$$

Number of AVs

Carbon intensity of power source [CO_2 eq. grams/kWh]

Emissions from Computing Onboard AVs

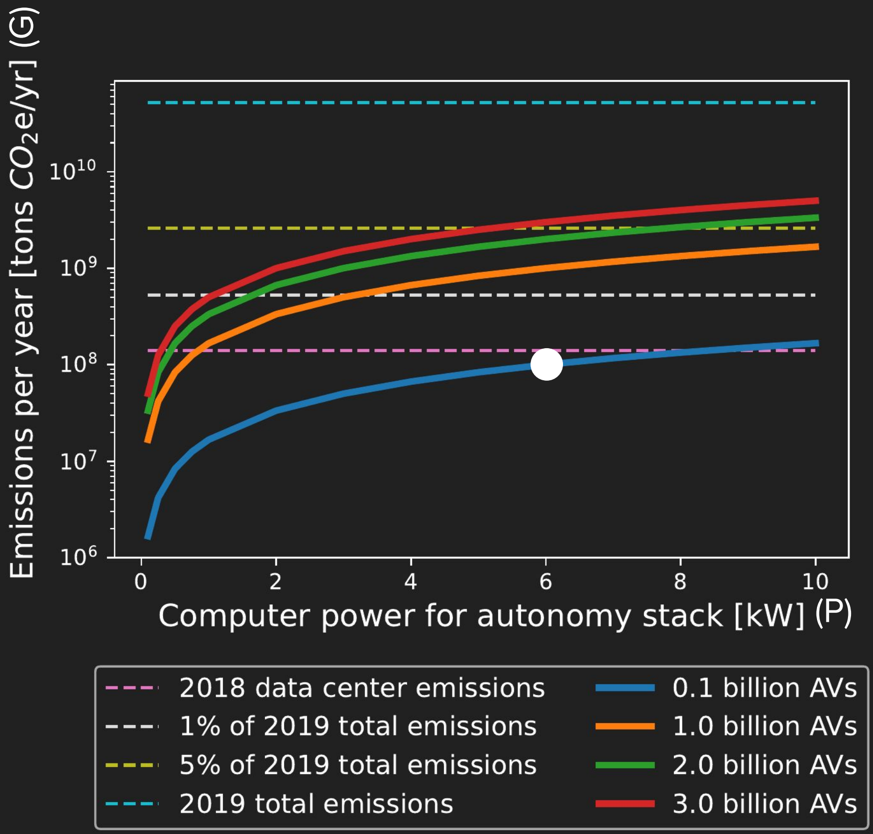


Assume 2020 global average carbon intensity and each AV driven for 1 hour per day

$$G = \alpha NQIP$$

- G: Emissions from computing onboard AVs (CO_2 eq. tons/yr)
- N: Number of AVs
- Q: Average time driven per AV [hrs/day]
- I: Carbon intensity of power source [CO_2 eq. grams/kWh]
- P: Average computer power [W]

Emissions from Computing Onboard AVs



Assume 2020 global average carbon intensity and each AV driven for 1 hour per day

Emissions from computing onboard AVs (CO_2 eq. tons/yr)

Average time driven per AV [hrs/day]

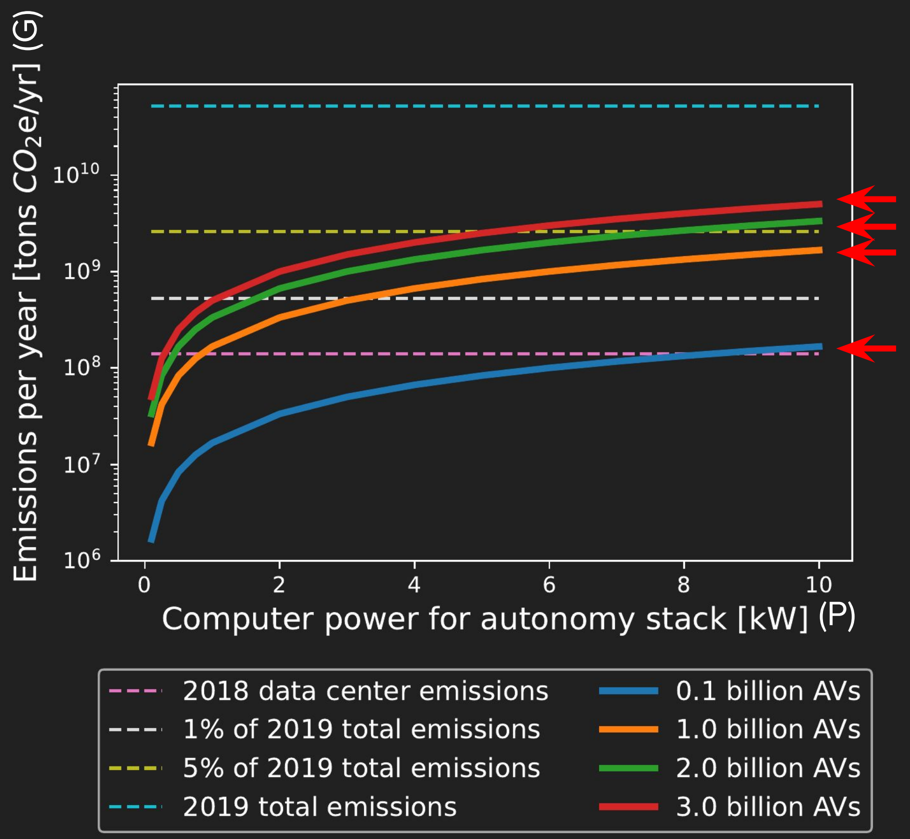
Average computer power [W]

$$G = \alpha N Q I P$$

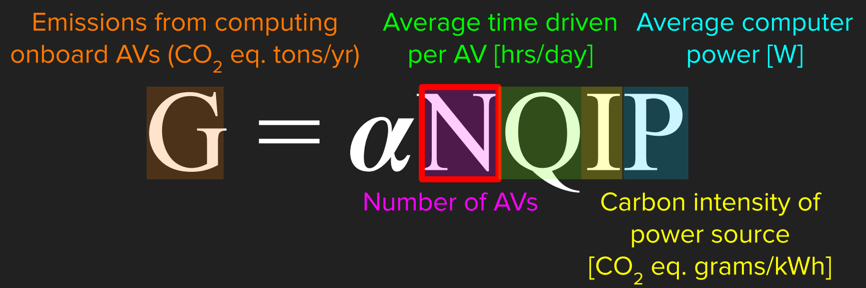
Number of AVs

Carbon intensity of power source [CO_2 eq. grams/kWh]

Emissions from Computing Onboard AVs



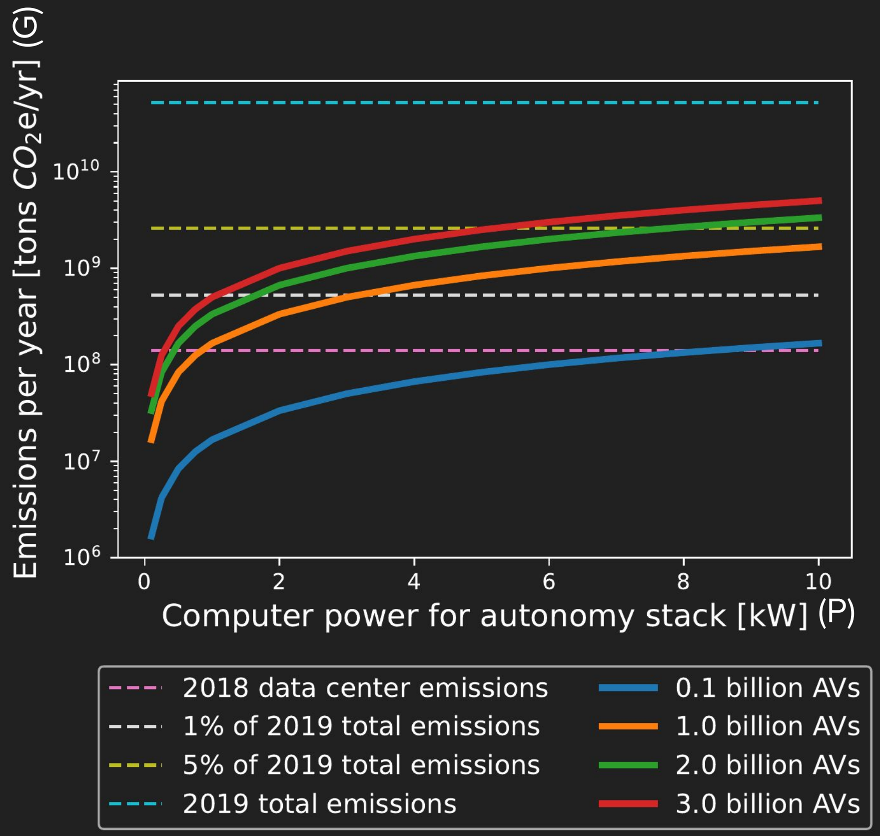
Assume 2020 global average carbon intensity and each AV driven for 1 hour per day



Sources of uncertainty:

- Will Level 5 autonomy be solved?
- How fast will AVs be adopted by the public?

Emissions from Computing Onboard AVs



Assume 2020 global average carbon intensity and each AV driven for 1 hour per day

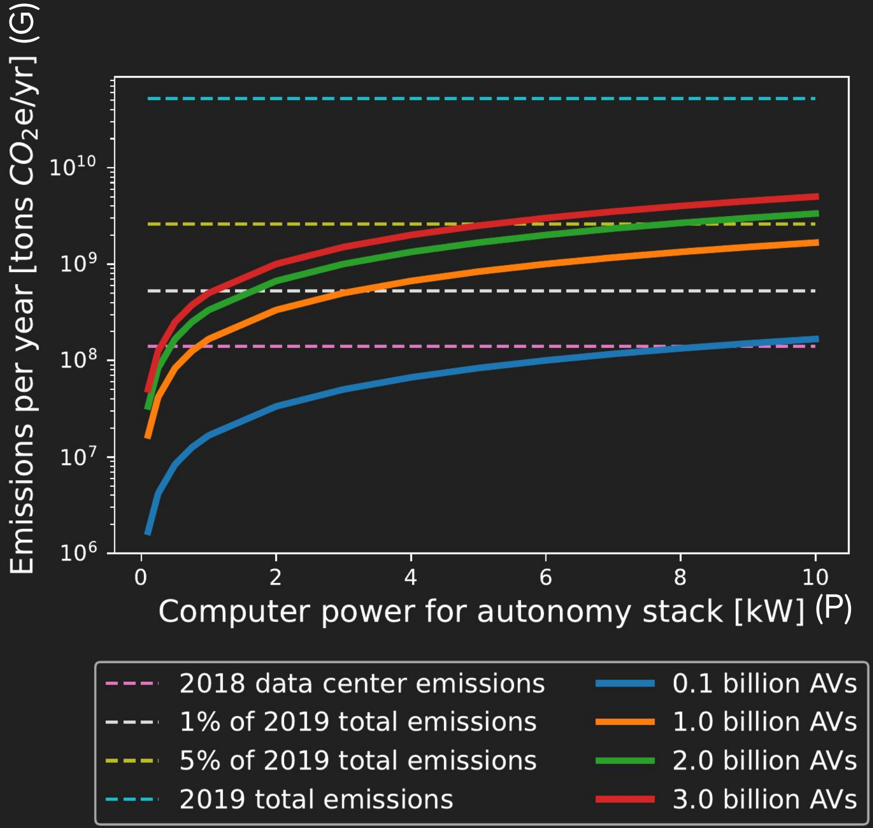
$$G = \alpha N Q I P$$

- G: Emissions from computing onboard AVs (CO_2 eq. tons/yr)
- N: Number of AVs
- Q: Average time driven per AV [hrs/day]
- I: Carbon intensity of power source [CO_2 eq. grams/kWh]
- P: Average computer power [W]

Sources of uncertainty:

- Will driving time increase (*e.g.*, ability to multitask)?
- Will driving time decrease (*e.g.*, optimized routing)?

Emissions from Computing Onboard AVs



Assume 2020 global average carbon intensity and each AV driven for 1 hour per day

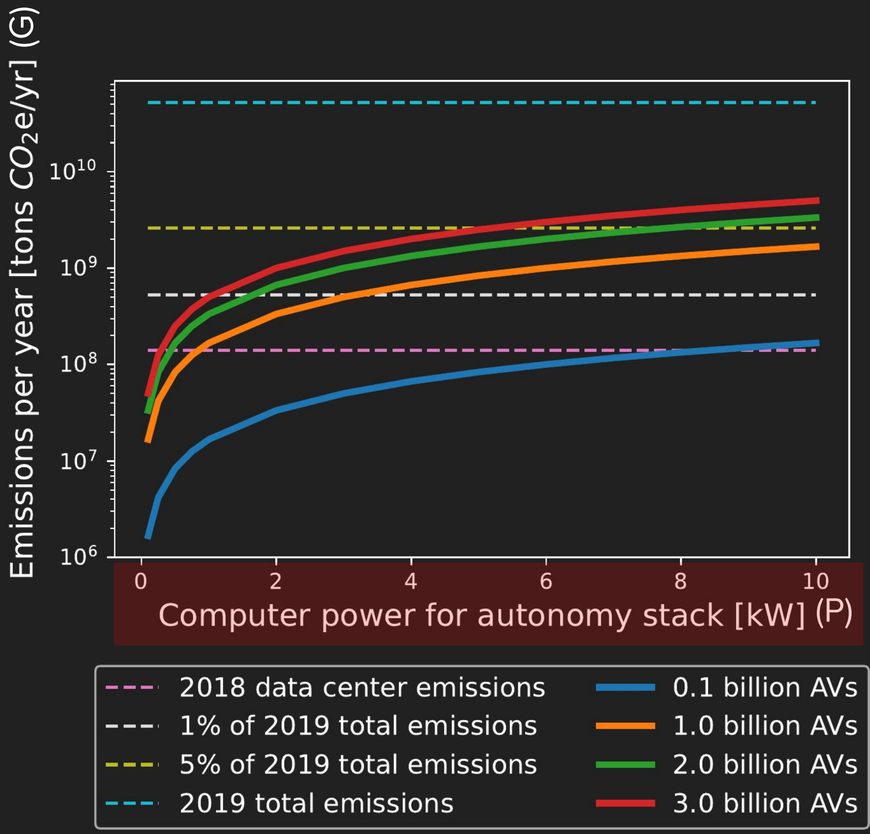
$$G = \alpha NQIP$$

- G: Emissions from computing onboard AVs (CO_2 eq. tons/yr)
- Q: Average time driven per AV [hrs/day]
- P: Average computer power [W]
- N: Number of AVs
- I: Carbon intensity of power source [CO_2 eq. grams/kWh]

Sources of uncertainty:

- How quickly will the world decarbonize over the next decades?

Emissions from Computing Onboard AVs



Assume 2020 global average carbon intensity and each AV driven for 1 hour per day

Emissions from computing onboard AVs (CO_2 eq. tons/yr)

Average time driven per AV [hrs/day]

Average computer power [W]

$$G = \alpha NQIP$$

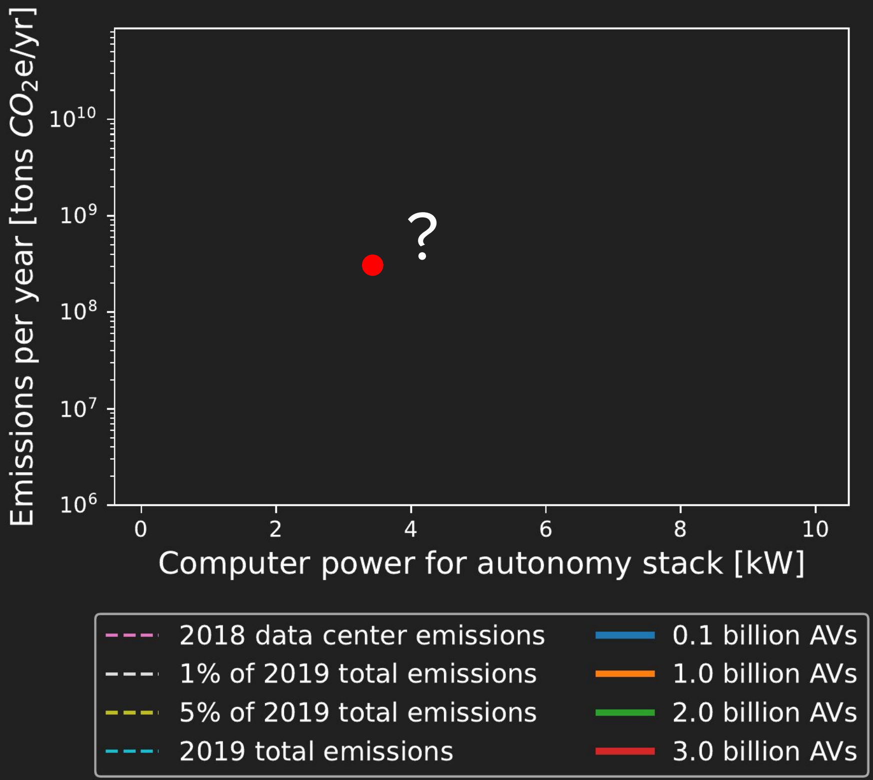
Number of AVs

Carbon intensity of power source [CO_2 eq. grams/kWh]

Sources of uncertainty:

- Proprietary information about industry AV stacks
- AV software stack for Level 5 autonomy:
  - What algorithms will be used? Types and number of sensors? Throughput of the system?
- Hardware energy efficiency?

Technical Gap: Emissions from Computing Onboard AVs



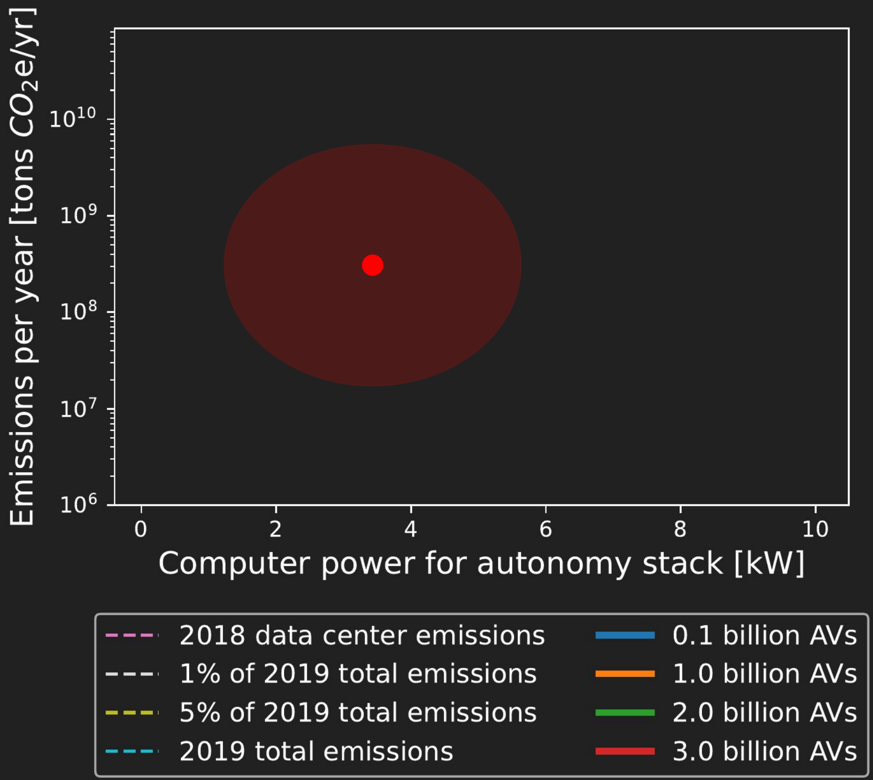
Assume 2020 global average carbon intensity and each AV driven for 1 hour per day

$$G = \alpha NQIP$$

- G: Emissions from computing onboard AVs (CO_2 eq. tons/yr)
- N: Number of AVs
- Q: Average time driven per AV [hrs/day]
- I: Carbon intensity of power source [CO_2 eq. grams/kWh]
- P: Average computer power [W]

Uncertainty on emerging application of Level 5 AVs and future trends of variables

Method: Emissions from Computing Onboard AVs



Assume 2020 global average carbon intensity and each AV driven for 1 hour per day

Emissions from computing onboard AVs (CO_2 eq. tons/yr)

Average time driven per AV [hrs/day]

Average computer power [W]

$$G = \alpha NQIP$$

Number of AVs

Carbon intensity of power source [CO_2 eq. grams/kWh]

Proposed framework **(this work)** explicitly takes uncertainties in each variable into account and produces a distribution of likely emissions scenarios

Method: Probabilistic Open-Source Framework to Model Emissions

- Probabilistically model each variable as a distribution to directly incorporate uncertainty and produce distributions of emissions

$$G = \alpha NQIP$$

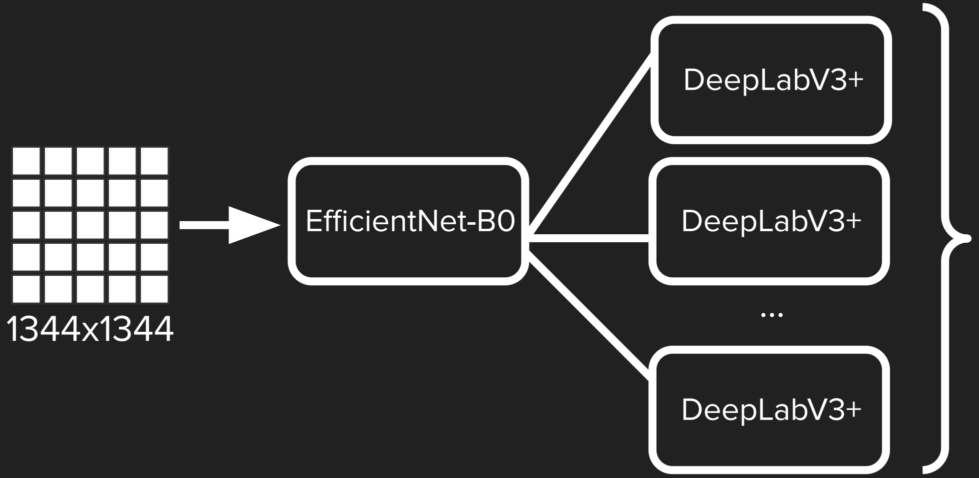
Method: Probabilistic Open-Source Framework to Model Emissions

- Probabilistically model each variable as a distribution to directly incorporate uncertainty and produce distributions of emissions
- Project future scenarios from 2025-2050 based on different adoption rates and annual changes in workload size, hardware efficiency, and carbon intensity

$$G = \alpha NQIP$$

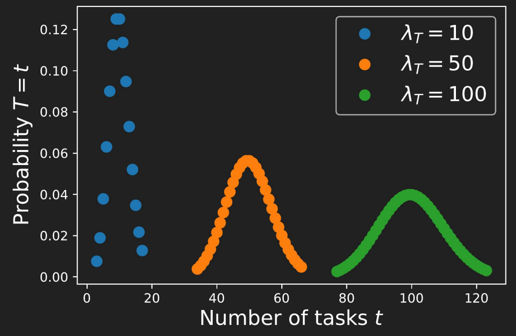
Open-source framework that can be modified as new information is learned or used by industry using internal proprietary numbers

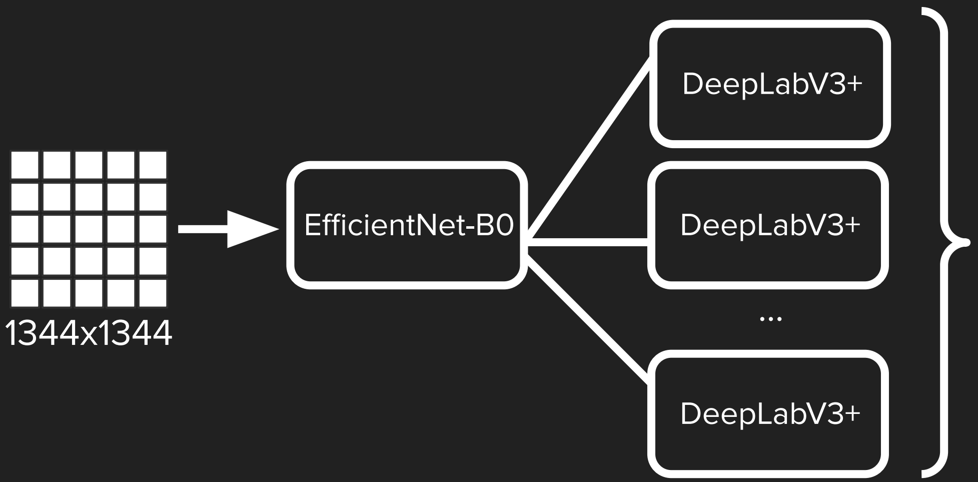
Modeling Computer Power (P)



Number of tasks *T*
(*e.g.*, pedestrian detector)

Modeling Computer Power (P)

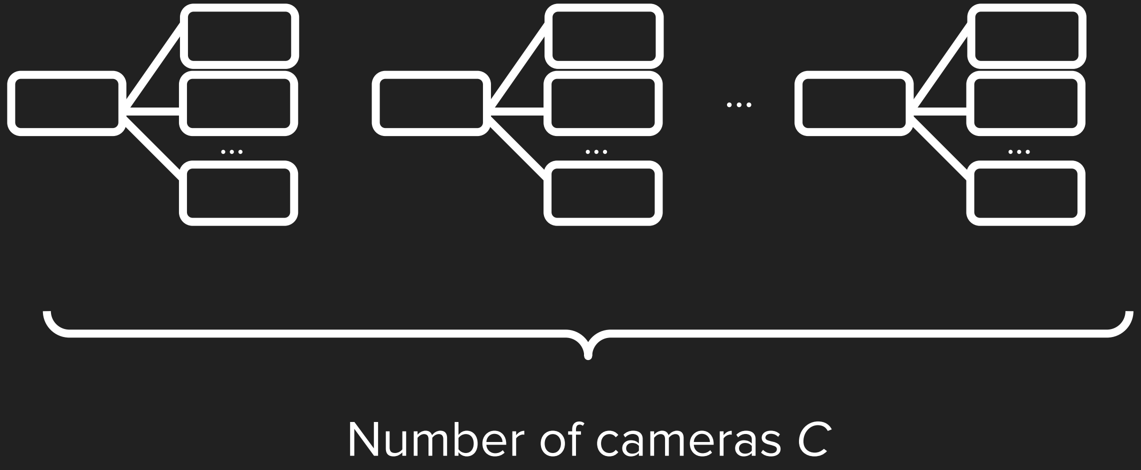


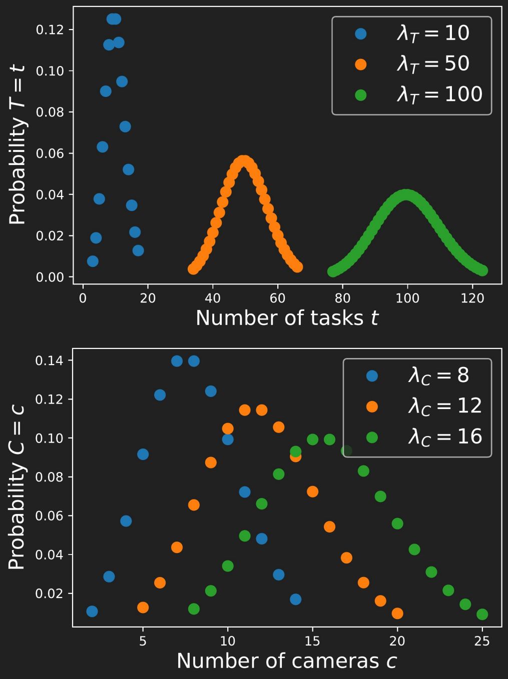


Number of tasks T
(*e.g.*, pedestrian detector)

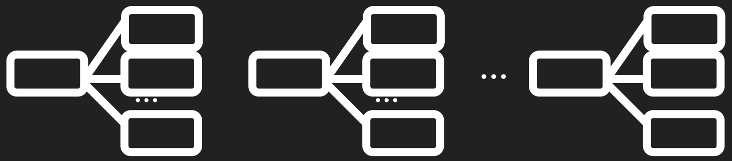
Measure power and latency on hardware platform (*e.g.*, Nvidia 2080 RTX Ti) for different number of tasks

Modeling Computer Power (P)

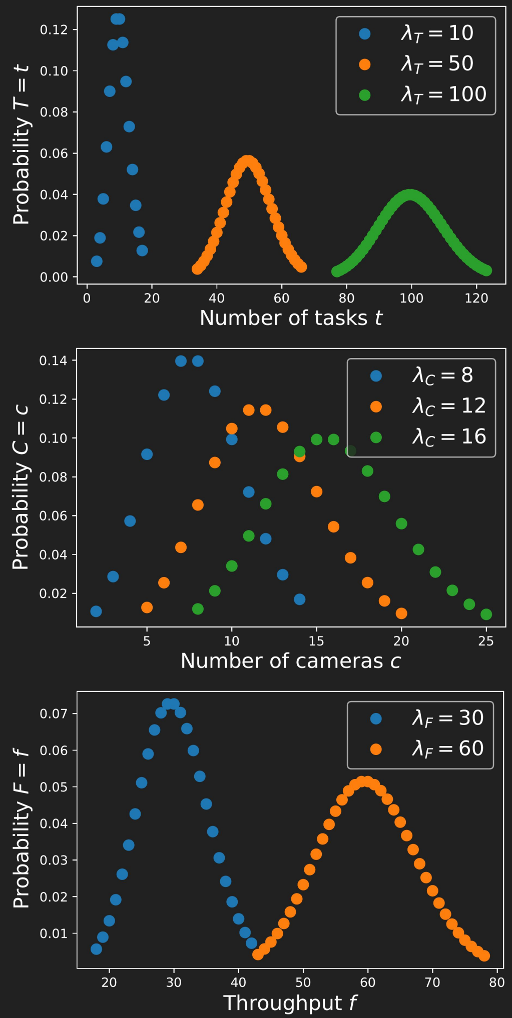




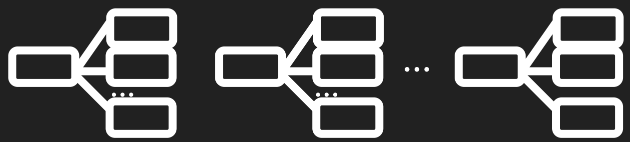
Modeling Computer Power (P)



@ target throughput ***F*** [Hz]



Modeling Computer Power (P)



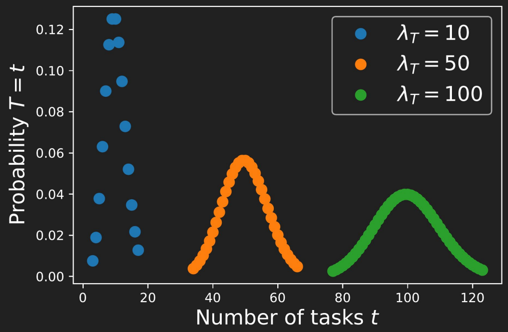
@ ***F*** [Hz] *@ target hardware efficiency*

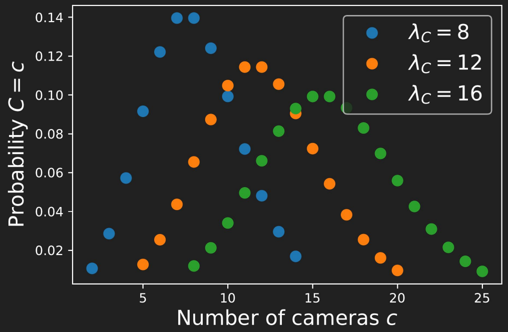
Scale by ratio of TOPS (INT8)/Watt η

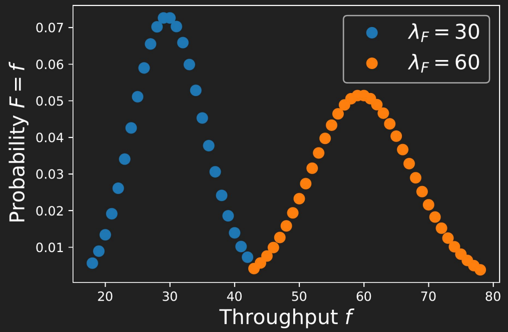
e.g., Nvidia 2080 RTX Ti: 215 TOPS/250 W

Nvidia Drive Orin: 2000 TOPS/800 W

η = 0.344

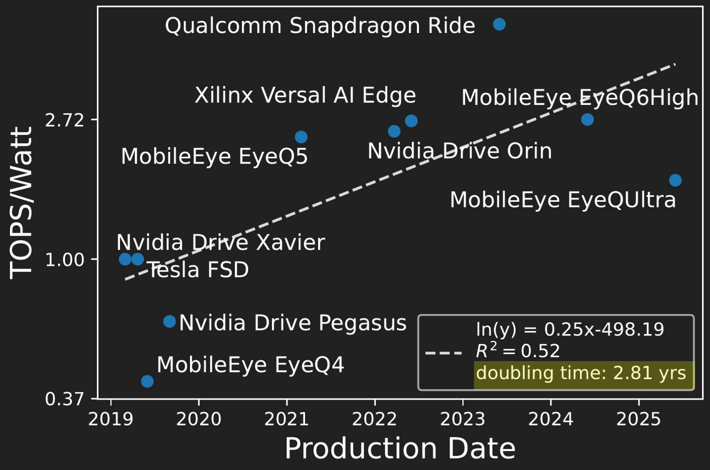






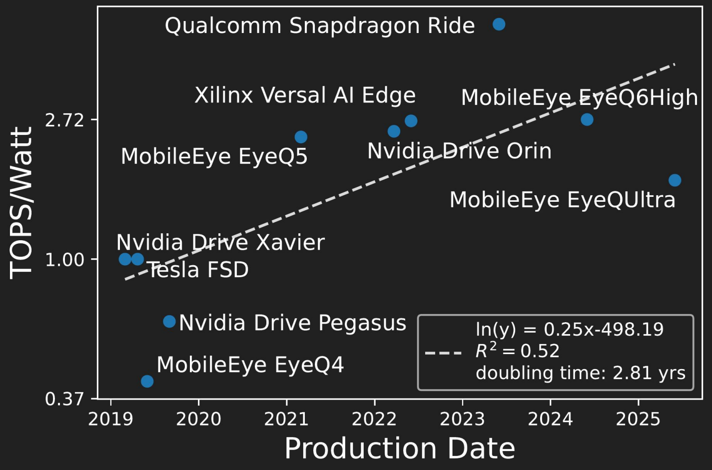
Modeling Future Trends in Computer Power (P)

Hardware energy efficiency:

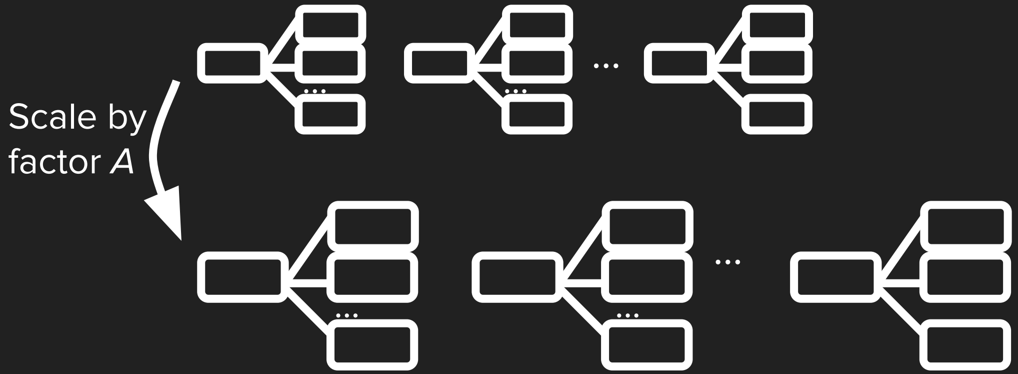


Modeling Future Trends in Computer Power (P)

Hardware energy efficiency:

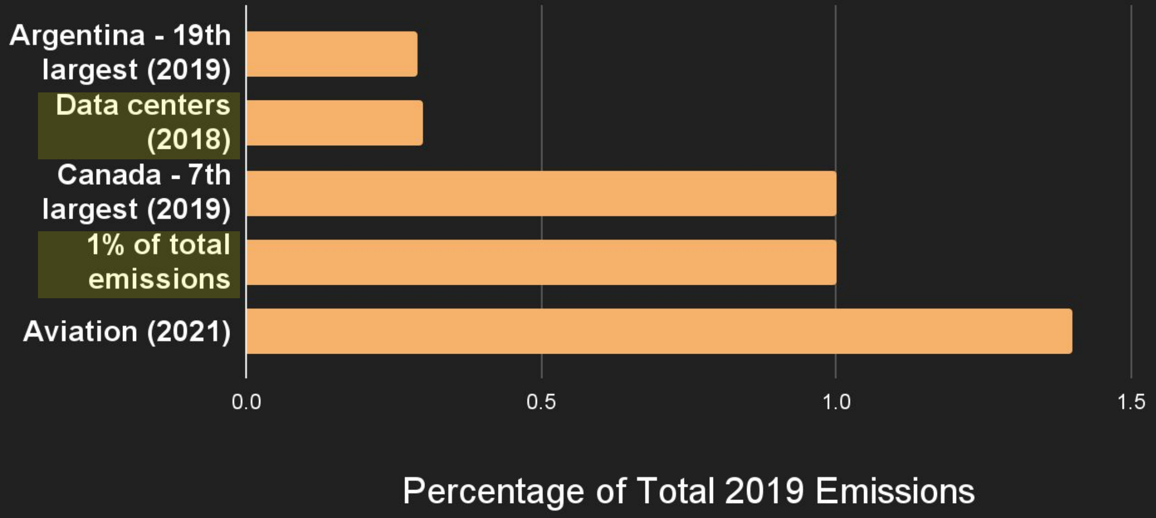


Size of workload:



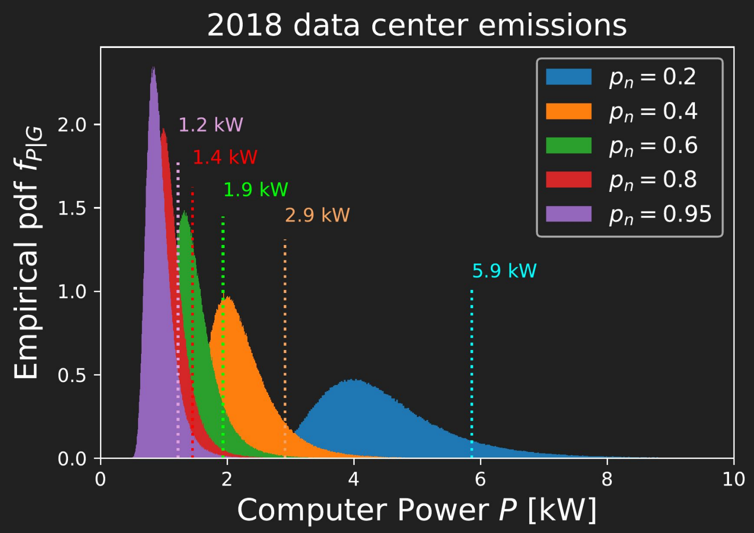
Framework capable of modeling increase in hardware energy efficiency and workload size

Results: Baselines



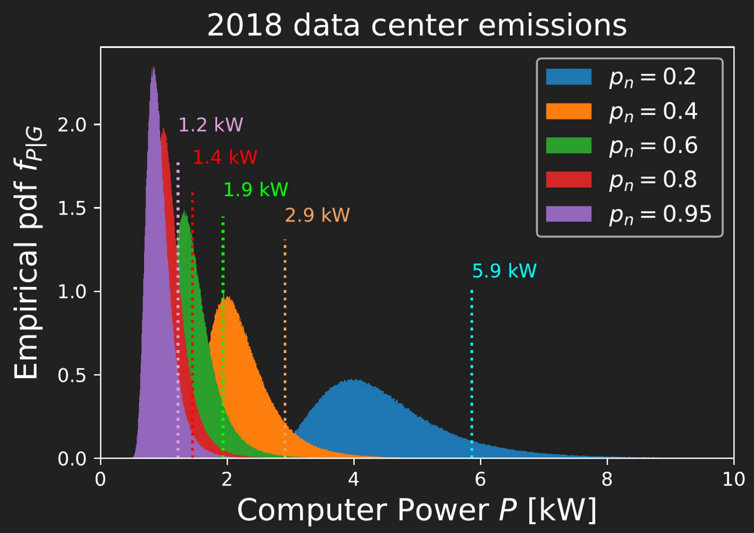
Results: Computer Power to Stay Under Emissions Targets

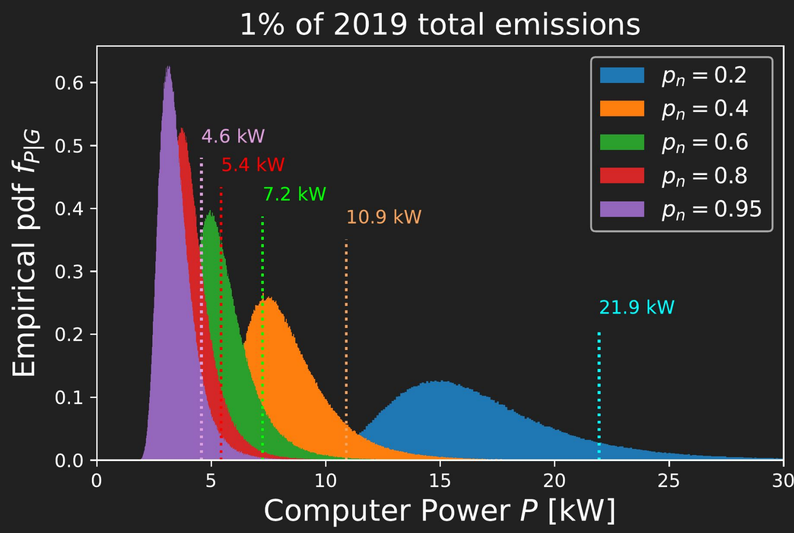
Emissions from computing onboard AVs: $\mathbf{G = \alpha NQIP}$



Results: Computer Power to Stay Under Emissions Targets

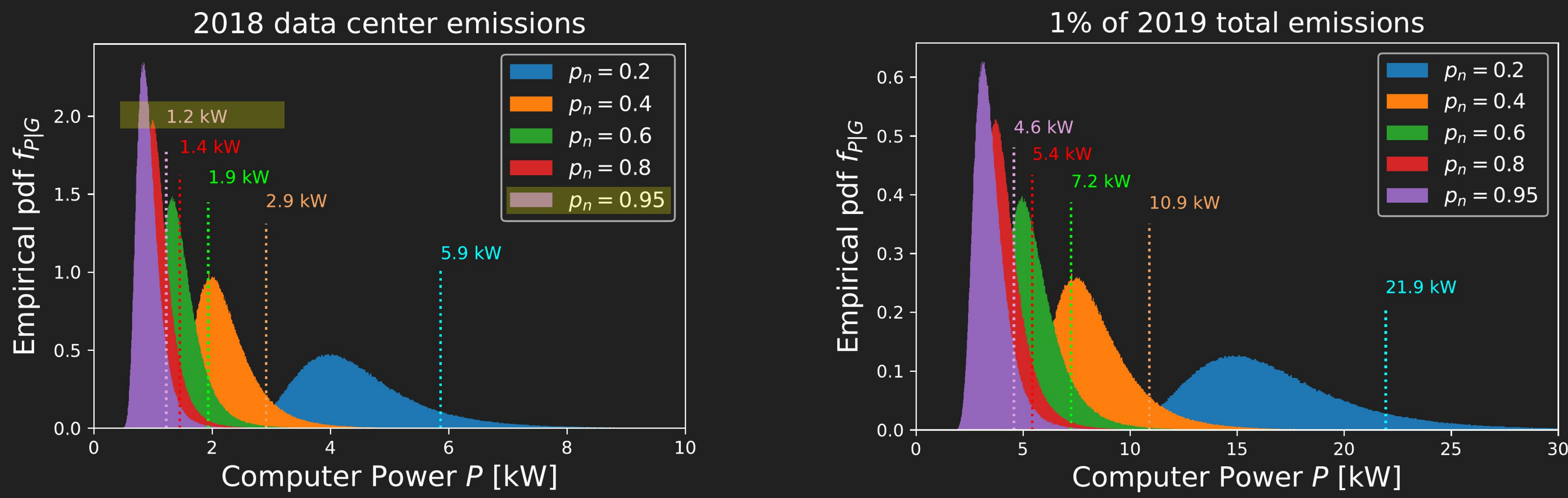
Emissions from computing onboard AVs: $\mathbf{G = \alpha NQIP}$





Results: Computer Power to Stay Under Emissions Targets

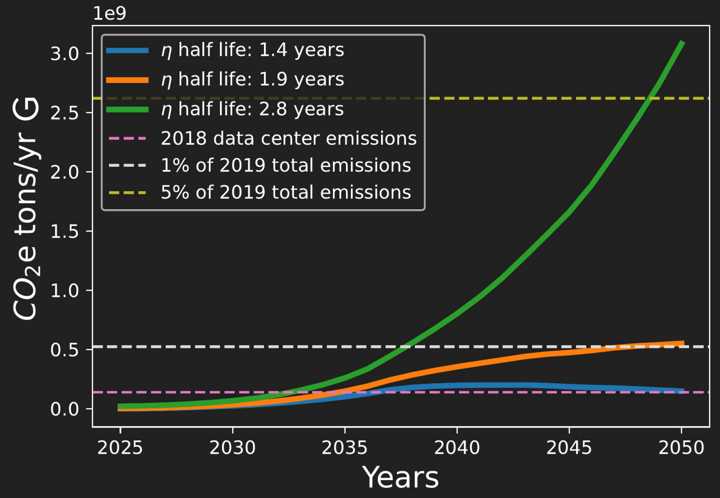
Emissions from computing onboard AVs: $\mathbf{G = \alpha NQIP}$



In 90% of scenarios where 95% of cars are AVs, computer power must stay under 1.2 kW for emissions to stay under 2018 data centers

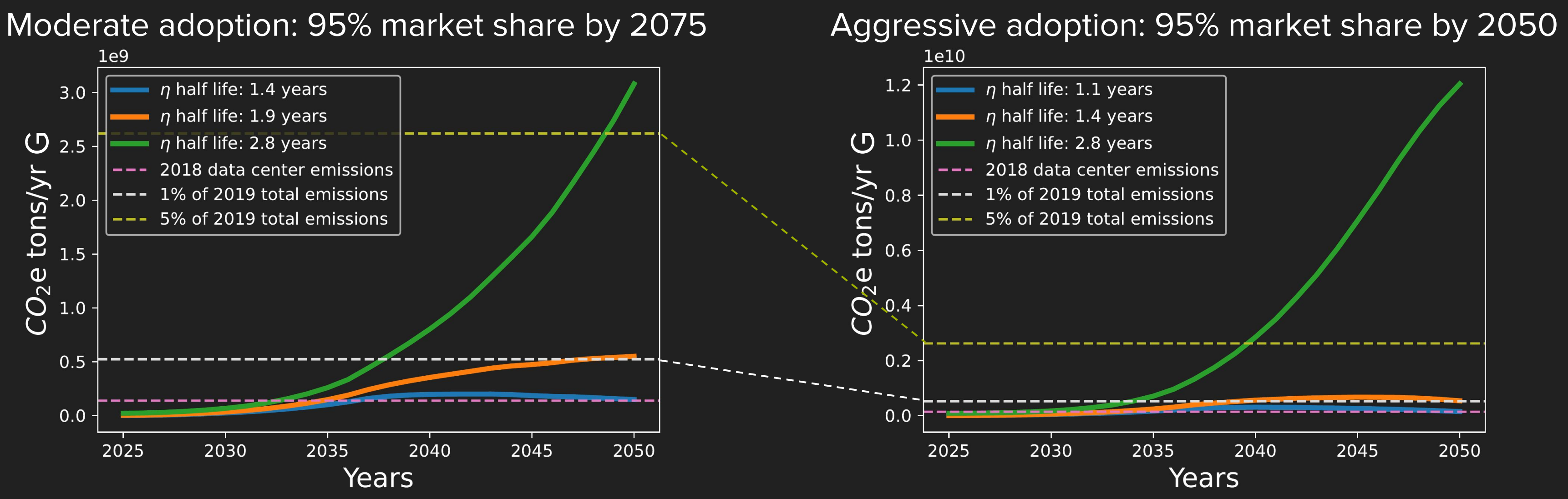
Results: Scenarios with Different Rates of Hardware Efficiency Increase

Moderate adoption: 95% market share by 2075



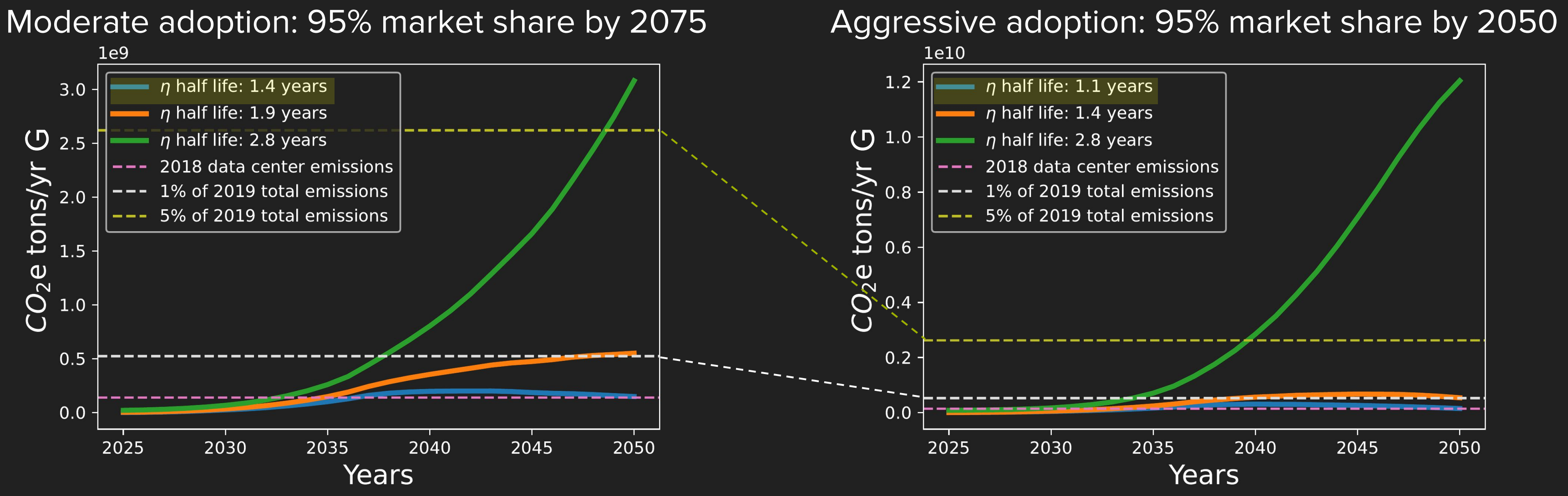
Assuming workload doubles every 3 years, business-as-usual decarbonization

Results: Scenarios with Different Rates of Hardware Efficiency Increase



Assuming workload doubles every 3 years, business-as-usual decarbonization

Results: Scenarios with Different Rates of Hardware Efficiency Increase

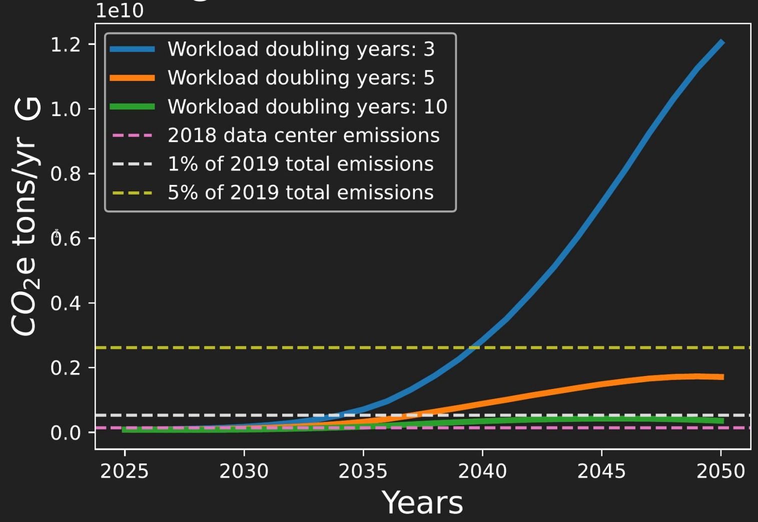


Assuming workload doubles every 3 years, business-as-usual decarbonization

Hardware energy efficiency would need to double every 1.4 or 1.1 years to contain emissions to 2018 data center emissions in moderate or aggressive adoption scenarios

Smaller Workload Growth Rate and Faster Decarbonization Helps

Effect of growth rate of workload:

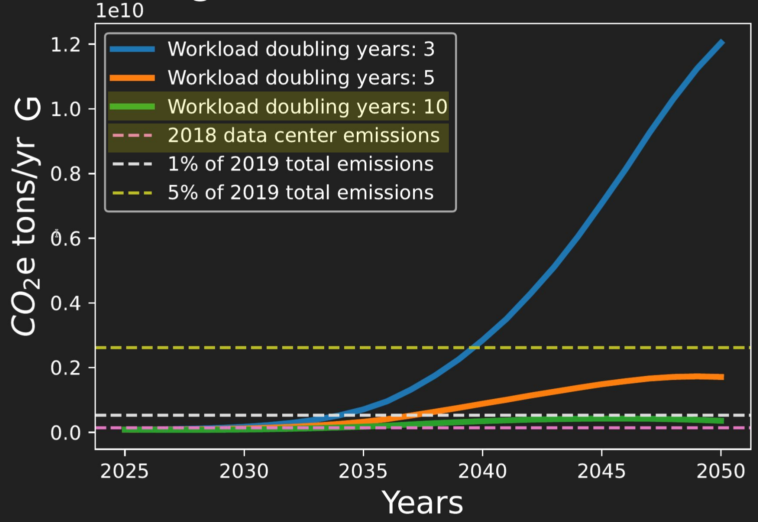


Business-as-usual decarbonization

Hardware energy efficiency doubles every 2.8 years, aggressive adoption scenario

Smaller Workload Growth Rate and Faster Decarbonization Helps

Effect of growth rate of workload:

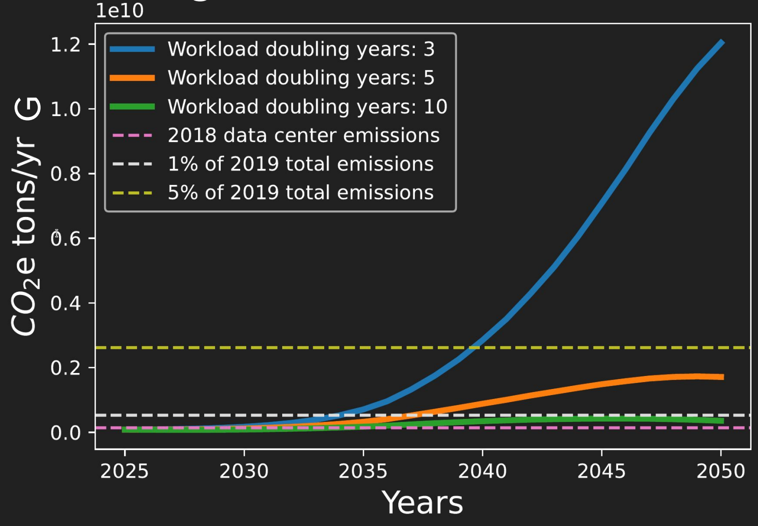


Business-as-usual decarbonization

Hardware energy efficiency doubles every 2.8 years, aggressive adoption scenario

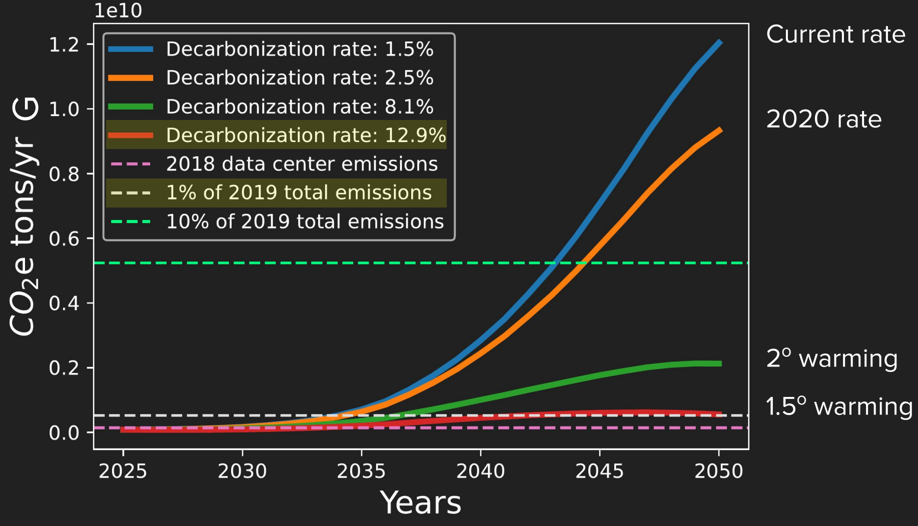
Smaller Workload Growth Rate and Faster Decarbonization Helps

Effect of growth rate of workload:



Business-as-usual decarbonization

Effect of decarbonization rate:



Workload doubles every three years

Hardware energy efficiency doubles every 2.8 years, aggressive adoption scenario

Smaller workload growth and rapid decarbonization reduces potential emissions from AVs, may still rival emissions from data centers

Future Directions: Challenges Unique to AVs

1. Explore algorithmic efficiency improvements without sacrificing safety
2. Characterize emissions from sensing
3. Characterize embodied (manufacturing) carbon vs. operational carbon emissions
4. Explore trade-off between hardware specialization and generalization
5. Encourage an industry standard to release data points to calculate emissions of autonomy stack

Future Directions: Challenges Unique to AVs

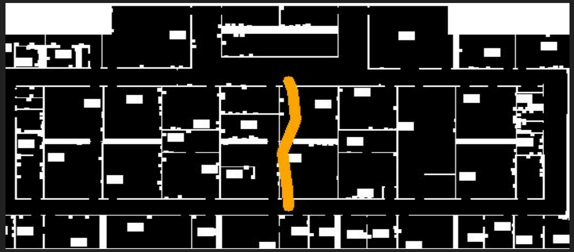
1. ***Explore algorithmic efficiency improvements without sacrificing safety***
2. Characterize emissions from sensing
3. Characterize embodied (manufacturing) carbon vs. operational carbon emissions
4. Explore trade-off between hardware specialization and generalization
5. ***Encourage an industry standard to release data points to calculate emissions of autonomy stack***

Efficient Algorithms That Will Not Compromise Safety

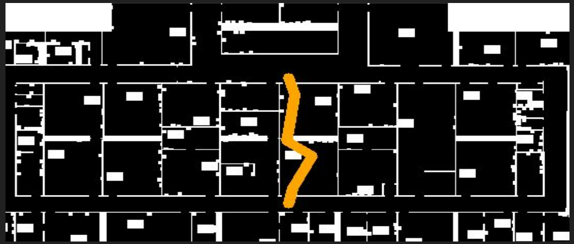
- Cannot tolerate decrease in performance metrics relevant to safety
- Many techniques used to make data centers greener cannot be applied here

Efficient Algorithms That Will Not Compromise Safety

- Cannot tolerate decrease in performance metrics relevant to safety
- Many techniques used to make data centers greener cannot be applied here
- Research needed on which algorithmic efficiency improvements (*e.g.*, compact, pruned, sparse DNNs, choosing to compute less and accept longer paths) can be safely applied



Example path returned by baseline

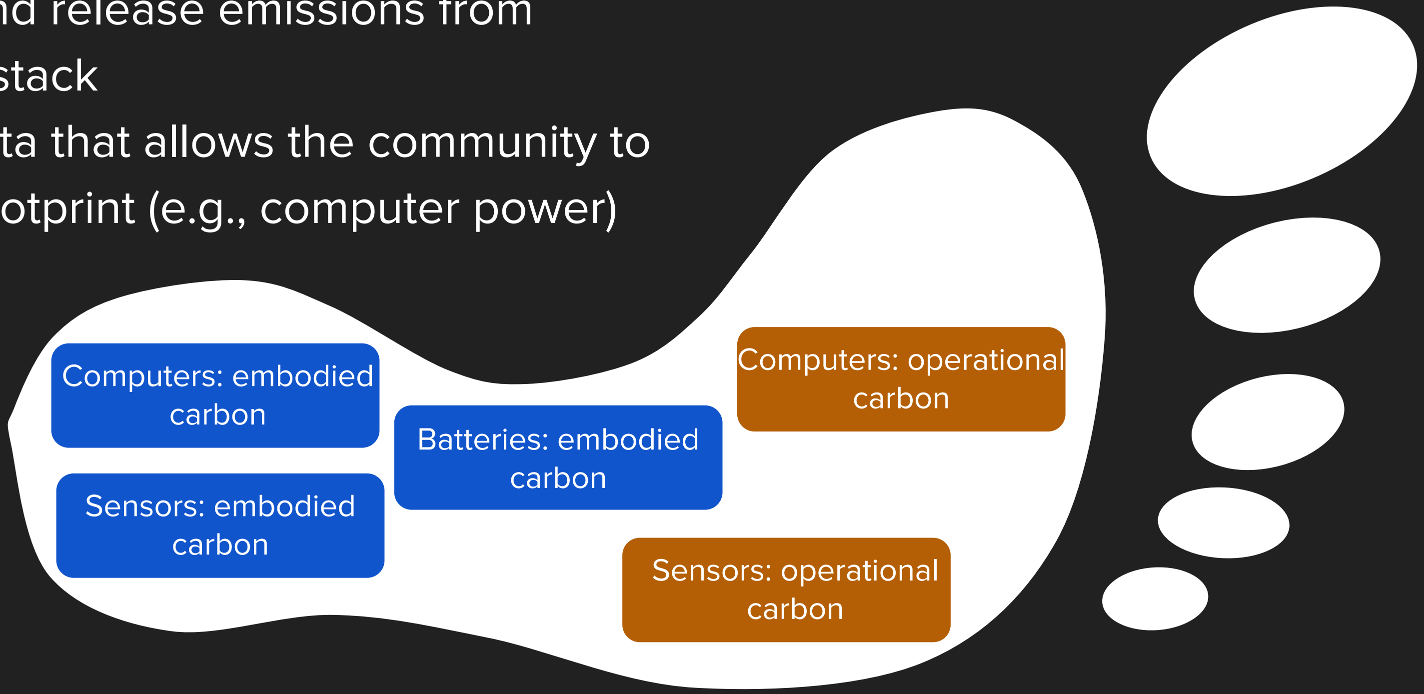


Example path returned by CEIMP (prior work)

Look to resource-constrained robot literature since computing is similarly not negligible

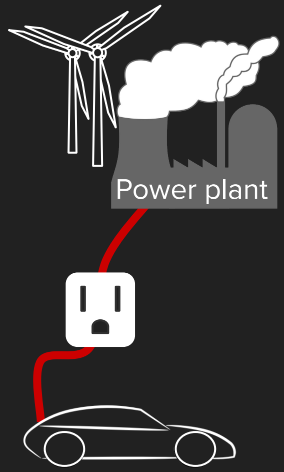
Encouraging an Industry Standard to Evaluate Footprint of Autonomy

- Use framework with internal numbers to evaluate and release emissions from autonomy stack
- Release data that allows the community to evaluate footprint (e.g., computer power)



Key Takeaways:

1. **Emissions from computing on AVs could be significant:** emissions could rival that of data centers due to significant workload size and the size of a global fleet of AVs
2. **Business-as-usual trends are not enough to constrain emissions:** current rates of decarbonization, hardware energy efficiency increase, and workload size increase will likely not constrain emissions to that of data centers today.
3. **Probabilistic framework enables emissions estimates that incorporate uncertainties**
4. **Encourage industry to account and release autonomy carbon footprint**



Sudhakar, Soumya, Vivienne Sze, and Sertac Karaman. "Data Centers on Wheels: Emissions from Computing Onboard Autonomous Vehicles." To be published at IEEE MICRO Special Issue on Environmentally Sustainable Computing. 2023.

Link: https://www.rle.mit.edu/eems/wp-content/uploads/2022/10/2022_micro_carbonAV.pdf

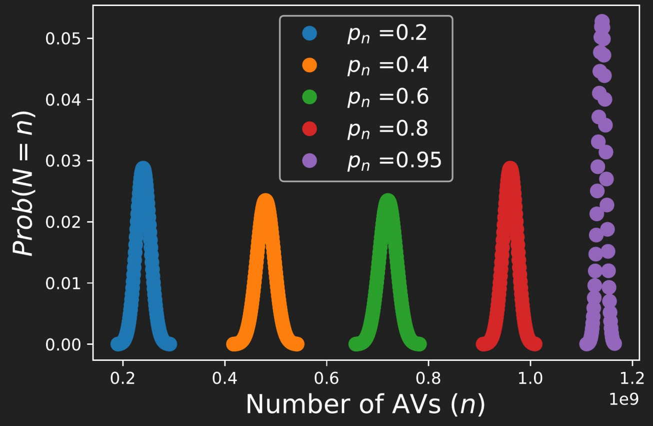
This work was partially supported by National Science Foundation (NSF) Cyber-Physical Systems program grant no. 1837212, NSF Real-Time Machine Learning program grant no. 1937501, and the MIT-Accenture Fellowship.

Back-Up

Modeling Number of Vehicles (N)

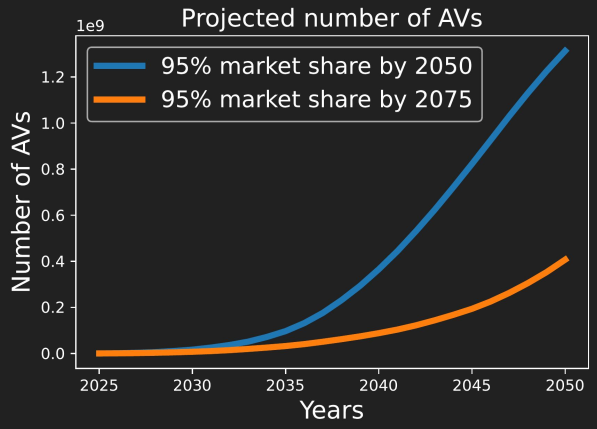
Current trends:

- ~1.2 billion cars on the road
- 1,400 AVs approved for testing in 2019 in US

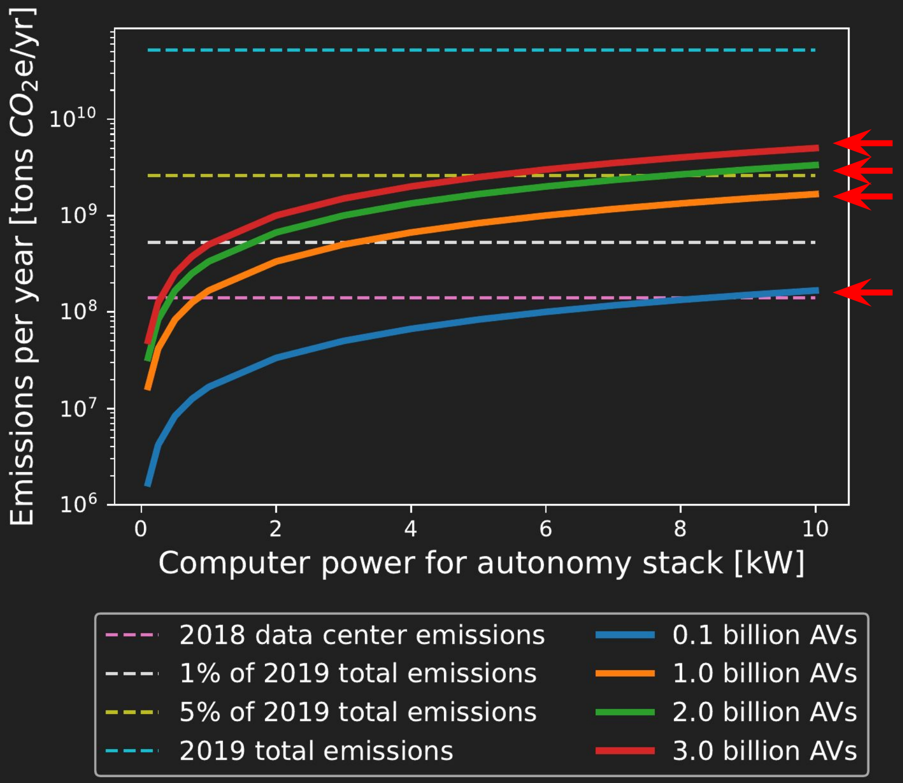


Future trends:

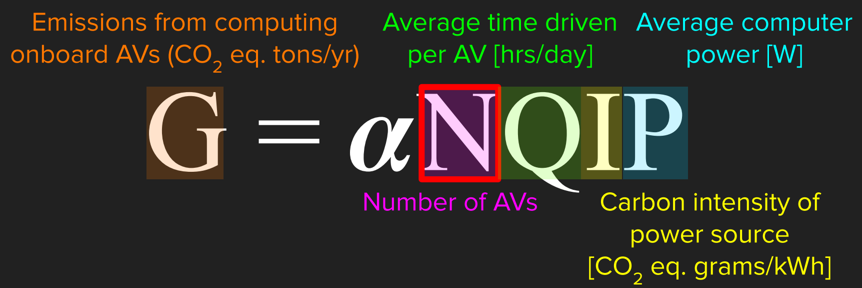
- Aggressive adoption scenario: 95% of market share by 2050
- Moderate adoption scenario: 95% of market share by 2075



Emissions from Computing Onboard AVs



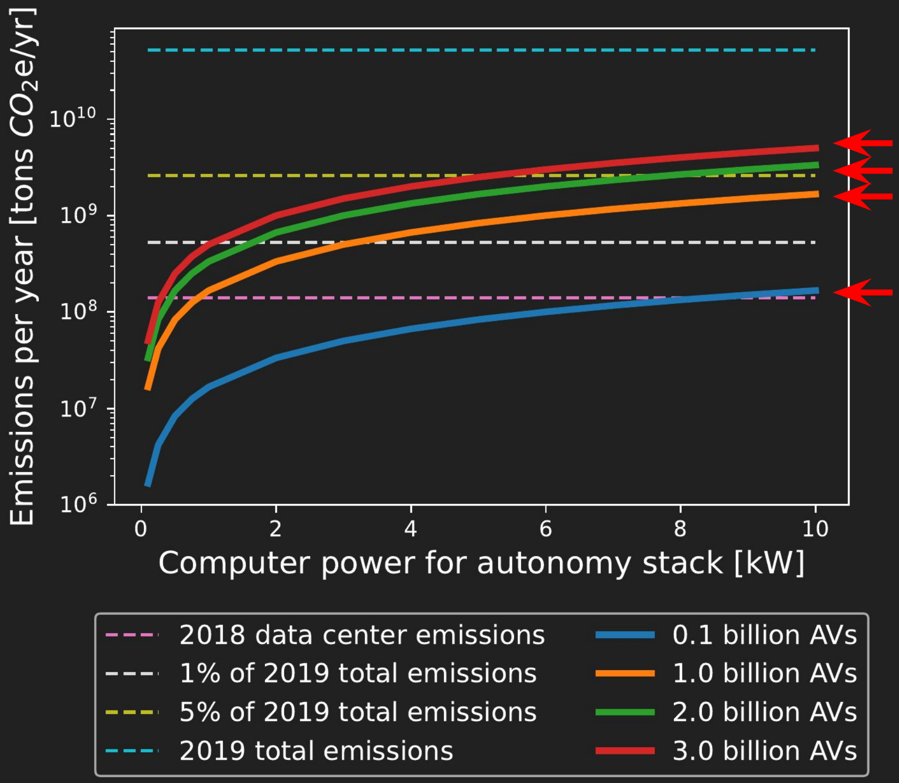
Assume 2020 global average carbon intensity and each AV driven for 1 hour per day



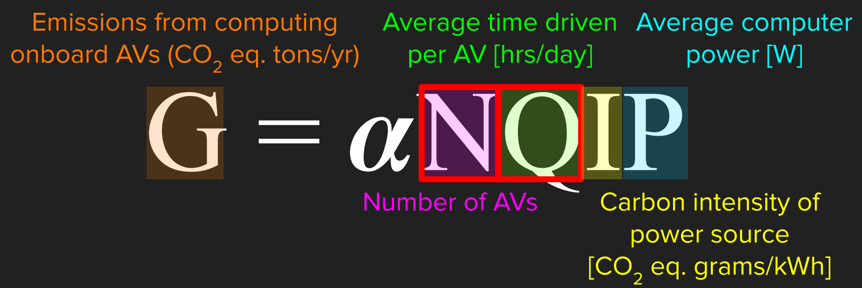
Sources of uncertainty:

- Will Level 5 autonomy be solved?
- How fast will AVs be adopted by the public?
- Growth in population and car ownership?
- Trends toward car sharing vs. private ownership?

Emissions from Computing Onboard AVs



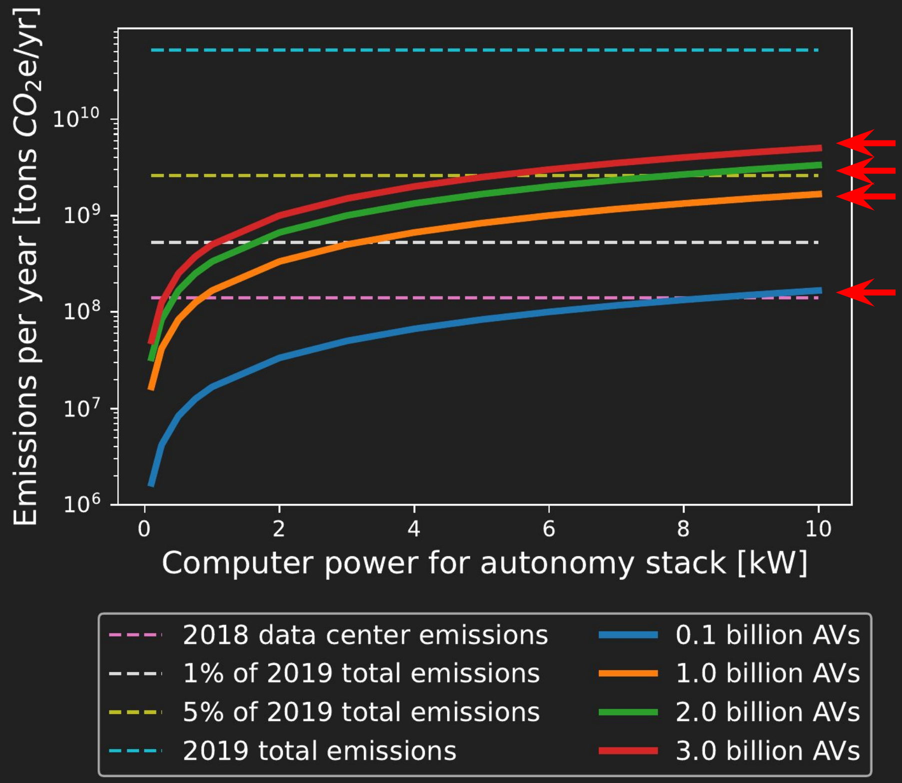
Assume 2020 global average carbon intensity and each AV driven for 1 hour per day



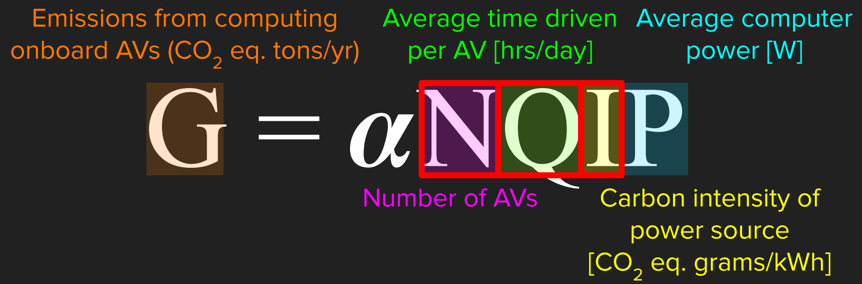
Sources of uncertainty:

- Will driving time increase due to ability to multitask, access to underserved populations who cannot drive?
- Will driving time decrease due to optimized routing, eco-driving strategies?

Emissions from Computing Onboard AVs



Assume 2020 global average carbon intensity and each AV driven for 1 hour per day



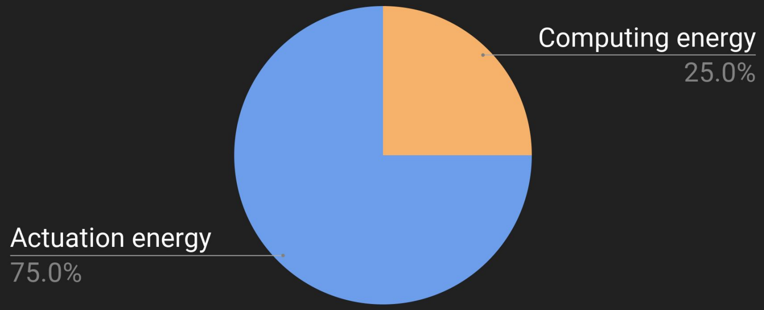
Sources of uncertainty:

- Is the global average the right metric to use for geographically varying carbon intensity?
- How quickly will the world decarbonize over the next decades?

Computing Energy vs. Actuation Energy for AVs

For an AV that consumes 0.25 kWh/mile that travels 30 miles over one hour,

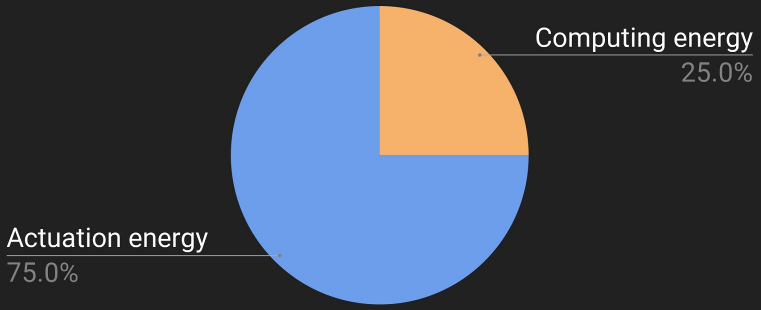
With a computer that consumes 2500 W:



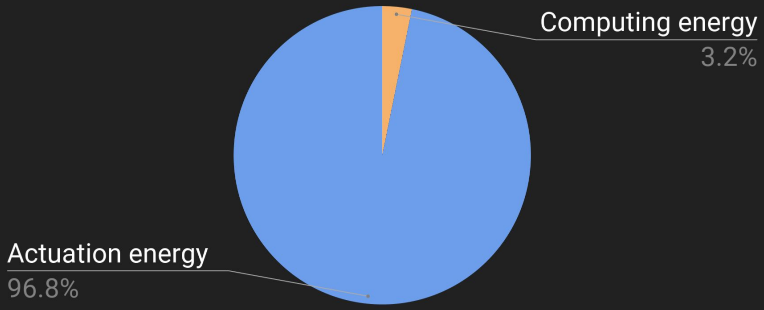
Computing Energy vs. Actuation Energy for AVs

For an AV that consumes 0.25 kWh/mile that travels 30 miles over one hour,

With a computer that consumes 2500 W:



With a computer that consumes 250 W:



Level 5 Autonomy

Tesla Sells 'Full Self-Driving,' but What Is It Really?

As the company deals with government scrutiny of its driver-assistance technology, an add-on kit sold for up to $10,000 is also getting more attention.

NYTimes 2021

FUTURE OF TRANSPORTATION

As Driverless Cars Falter, Are 'Driver Assistance' Systems in Closer Reach?

With investigations and lawsuits over accidents adding skepticism toward fully driverless technology, car companies are betting on systems that take some, but not all, control.

NYTimes 2022

Despite High Hopes, Self-Driving Cars Are 'Way in the Future'

Ford and other companies say the industry overestimated the arrival of autonomous vehicles, which still struggle to anticipate what other drivers and pedestrians will do.

NYTimes 2019

The Costly Pursuit of Self-Driving Cars Continues On. And On. And On.

Many in Silicon Valley promised that self-driving cars would be a common sight by 2021. Now the industry is resetting expectations and settling in for years of more work.

NYTimes 2021

Future Directions: Challenges Unique to AVs

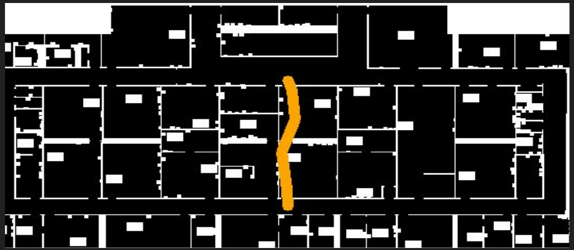
2. Characterize emissions from sensing
   a. Industry proposed AVs have different sensor suites and configurations
   b. LiDAR not negligible power consumption

3. Characterize embodied (manufacturing) carbon vs. operational carbon emissions
   a. Capture the total carbon footprint of computing onboard AVs
   b. AVs will have longer lifespans than that of data center servers or mobile devices – does operational carbon dominate over embodied carbon over the AV lifespan?
   c. Identify impact of strategies like car-sharing that will lower embodied carbon, but not operational carbon

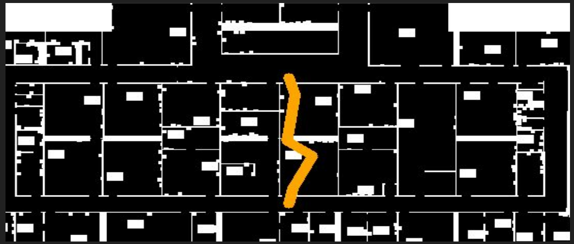
4. Explore trade-off between hardware specialization and generalization
   a. Unlike data centers, AVs handle constant workloads that are known ahead of time, making the case for hardware specialization
   b. Still need to generalize to new workloads over the lifespan of the car

Future Directions: Efficiency with Safety

1. Explore algorithmic efficiency improvements without sacrificing safety
   a. Trading off computing energy vs. actuation energy in tasks such as motion planning
   b. Choosing less perceptually difficult (more compute intensive) paths to spend less energy and lower emissions from computing



Example path returned by baseline



Example path returned by CEIMP (prior work)

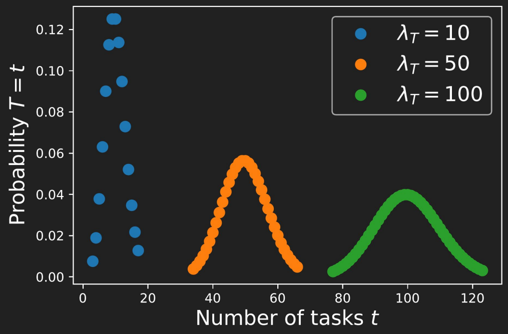
Modeling Computer Power (P)

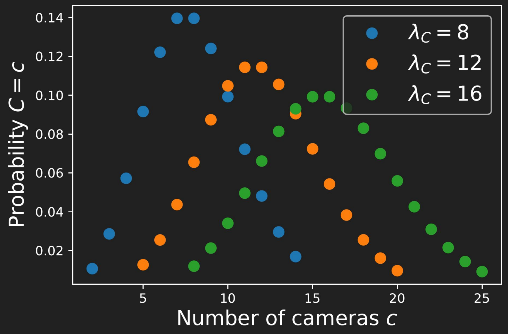
Current trends:

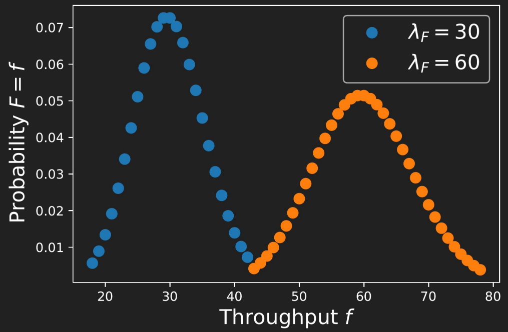
- DNN workload (likely major component of perception)
- Assume a multitask DNN (shared EfficientNet-B0 encoder between tasks, new DeepLabV3+ decoder per task, full autoencoder run per 1344x1344 camera image) at desired throughput
- Extrapolate to power on target hardware (e.g., Nvidia Drive Orin) using ratio of TOPS (INT8)/Watt η

$$P_{target} = P_{meas}(T)L_{meas}(T)\eta FC$$

$$G = \alpha NQIP_{meas}(T)L_{meas}(T)\eta FC$$

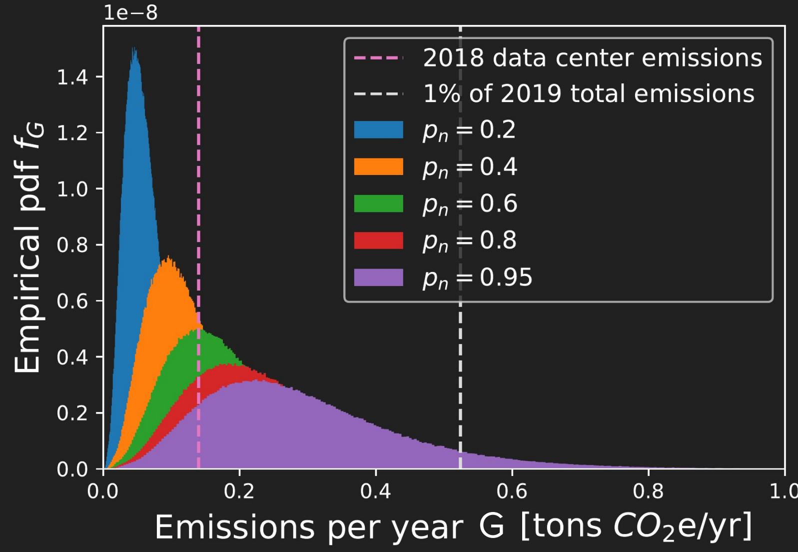






Results: Potential for Significant Emissions

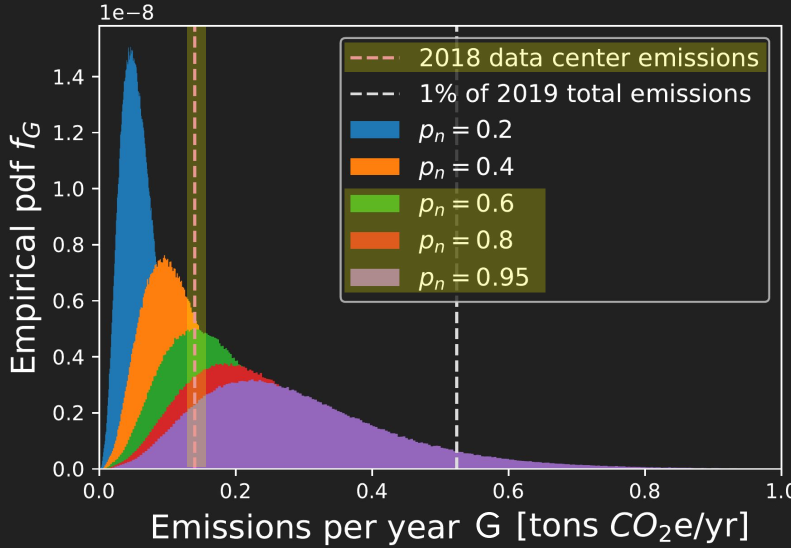
Emissions from computing onboard AVs: **G = αNQIP**



Under current trends, potential for emissions from computing onboard AVs to rival 2018 data centers

Results: Potential for Significant Emissions

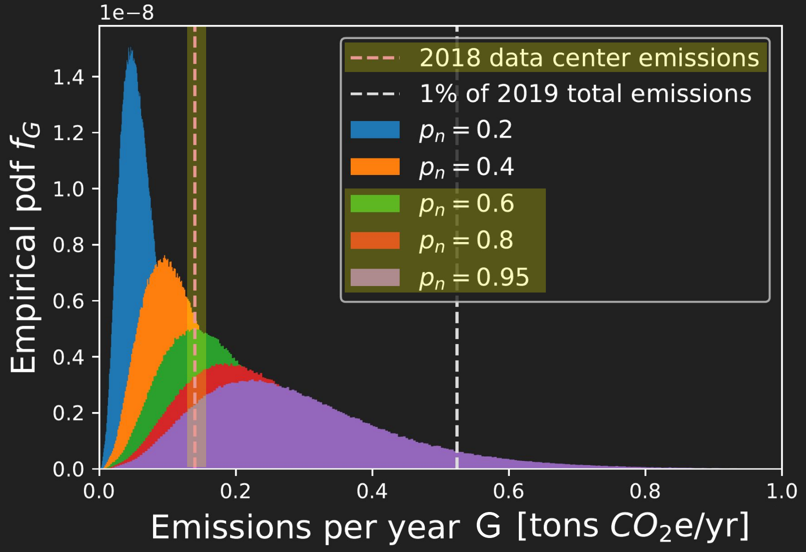
Emissions from computing onboard AVs: **G = αNQIP**

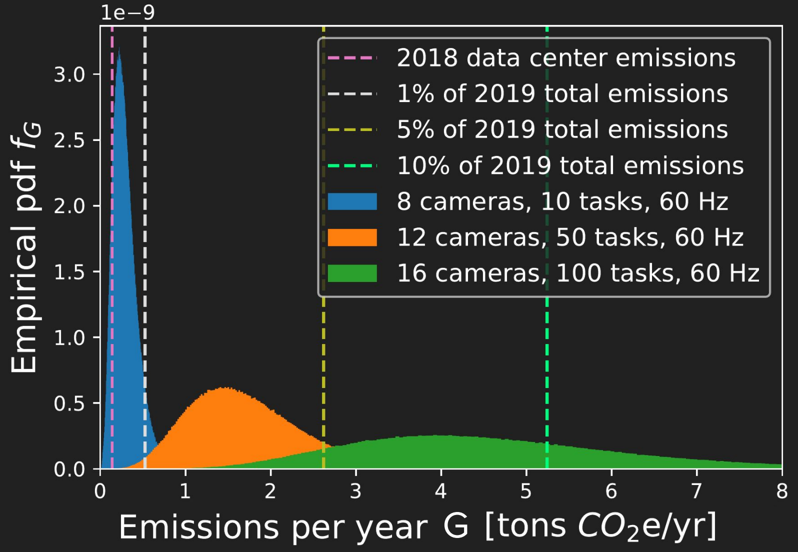


Under current trends, potential for emissions from computing onboard AVs to rival 2018 data centers

Results: Potential for Significant Emissions

Emissions from computing onboard AVs: $\mathbf{G = \alpha NQIP}$

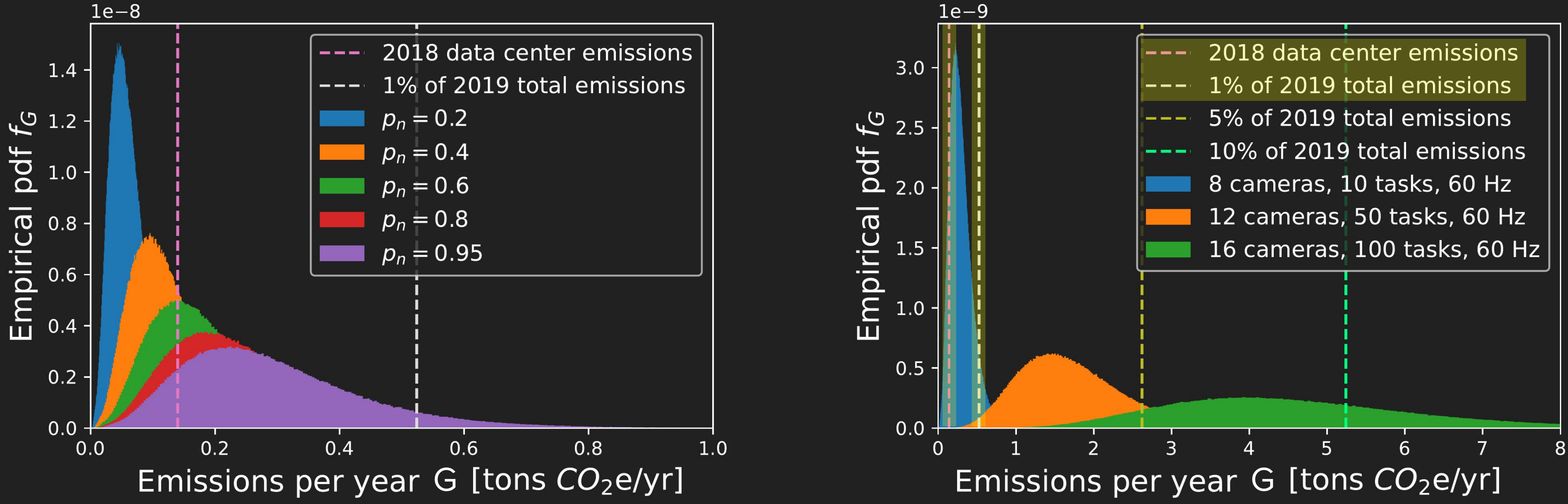




Under current trends, potential for emissions from computing onboard AVs to rival 2018 data centers

Results: Potential for Significant Emissions

Emissions from computing onboard AVs: **G = αNQIP**



Under current trends, potential for emissions from computing onboard AVs to rival 2018 data centers