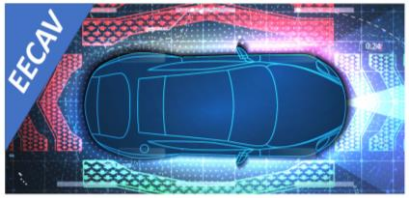
SAND2021-12326O

Semiconductor and Automated Vehicle Development Projections

Matt Marinella, Sandia
Jace Mogill, USCAR

Workshop on Energy Efficient Computing for Automated Vehicles (EECAV)

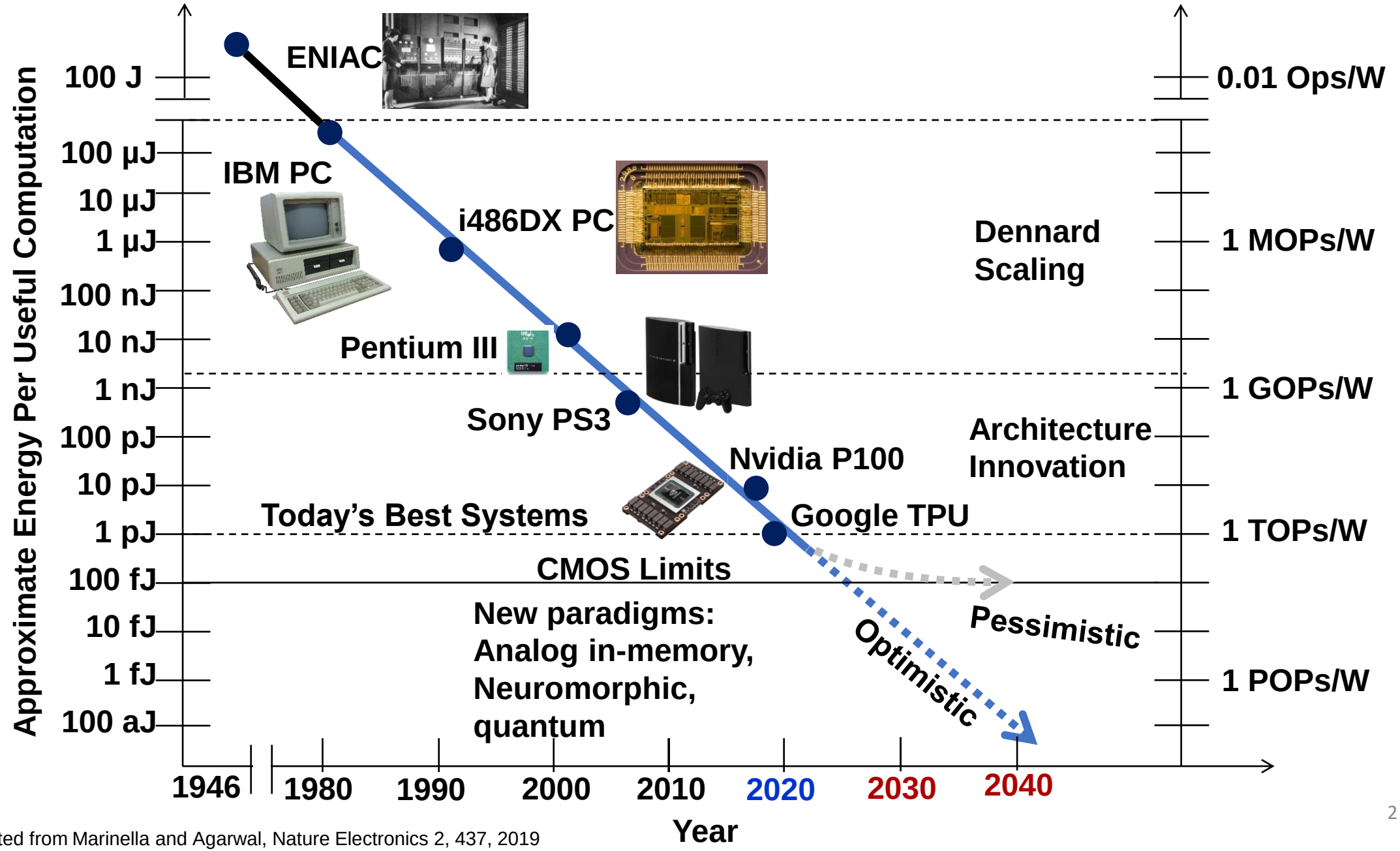
May 11 - 12, 2021



Sandia National Laboratories

UNITED STATES COUNCIL FOR AUTOMOTIVE RESEARCH LLC

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525

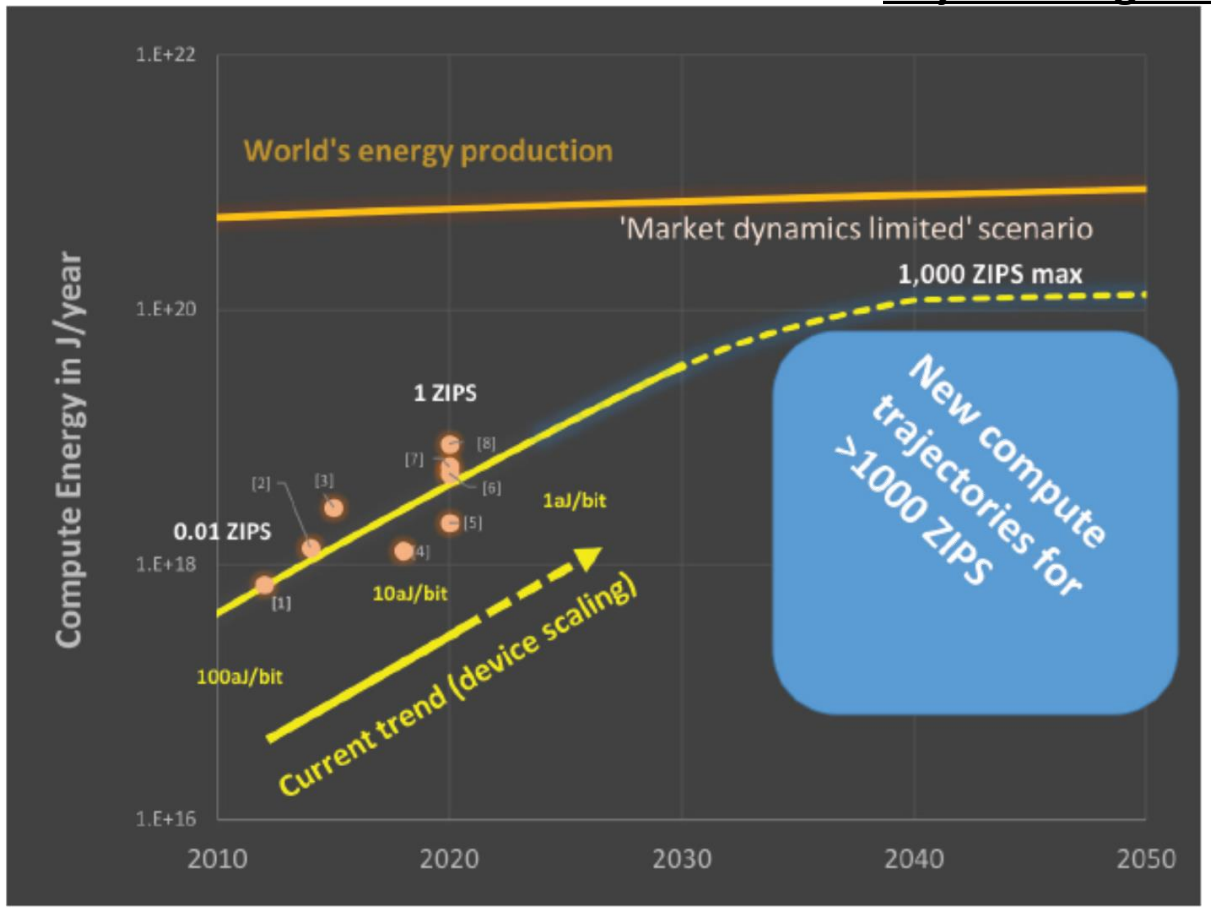


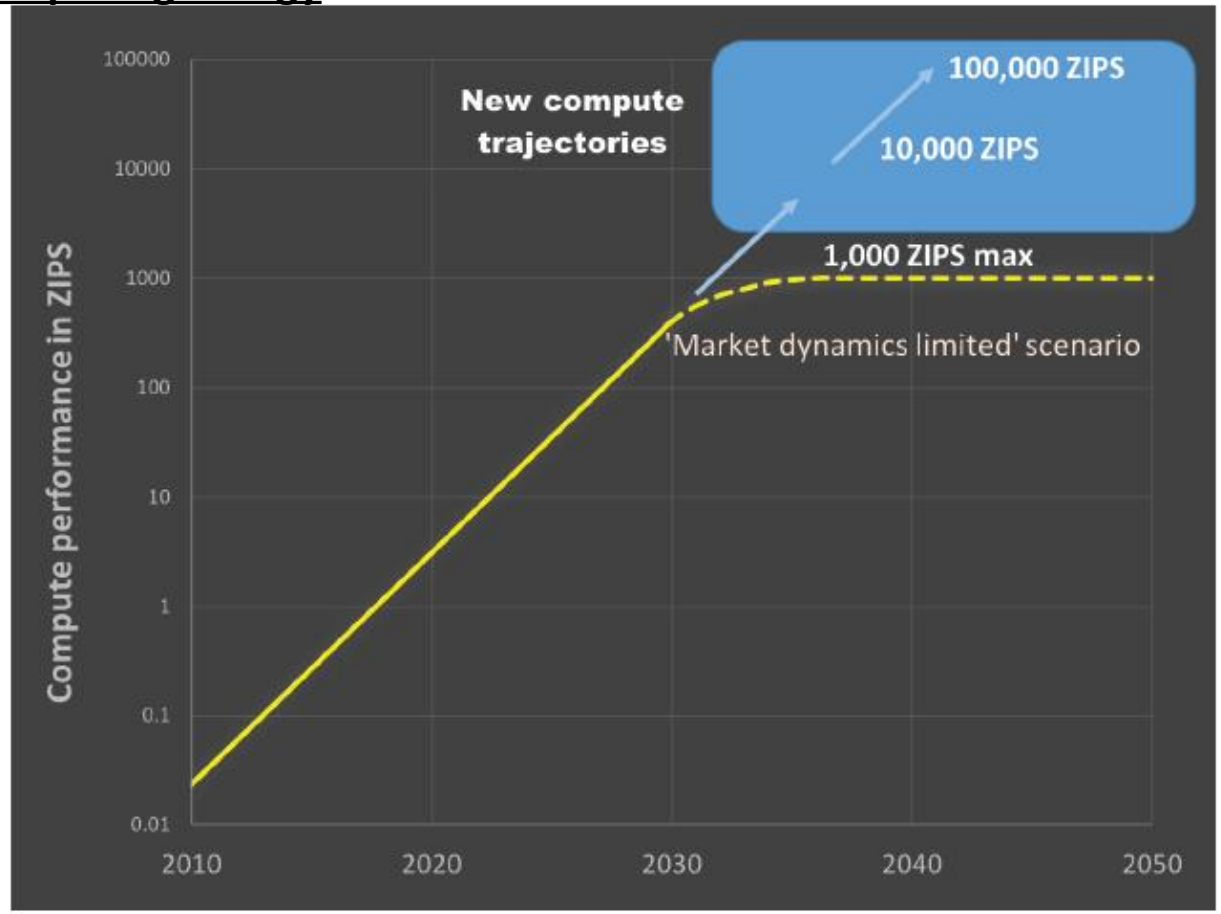
Adapted from Marinella and Agarwal, Nature Electronics 2, 437, 2019

Perspective from the semiconductor industry

Semiconductor Research Corp (SRC) & Semiconductor Industry Assoc (SIA) Decadal Plan

Key Challenge: Computing Energy



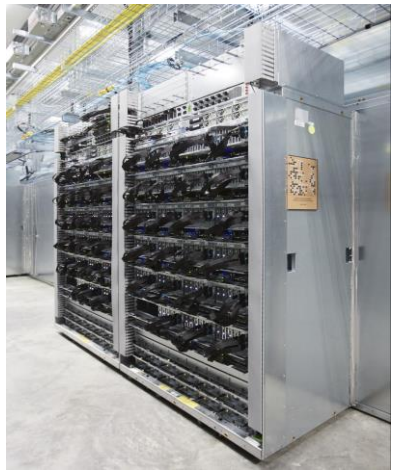


SRC, SIA, "Interim Report for the Decadal Plan for Semiconductors," October 2020

SEMICONDUCTOR INDUSTRY ASSOCIATION

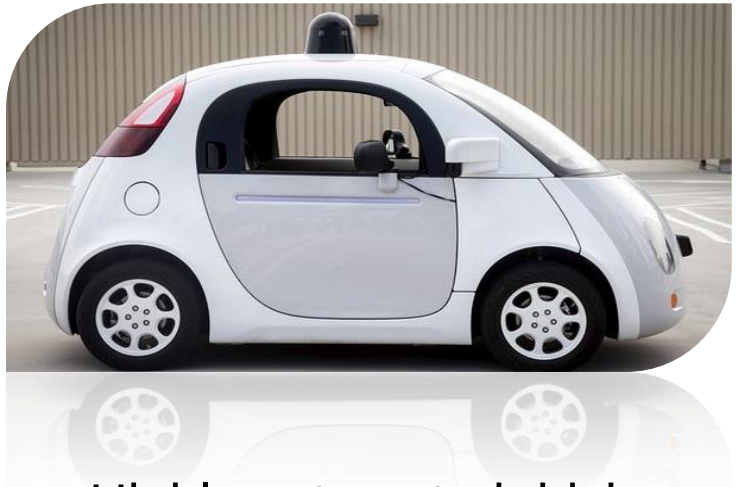
Semiconductor Research Corporation

How does the computing energy affect autonomous vehicles?



computing must meet size, weight, and power constraints

>1 petaflops
~100 W (system)
~100 TOPS/watt (SoC)
TOPS == Trillion (tera) Operations



Highly automated driving
Using neural algorithms

>100x power performance

>10x compute

>10x less power

Early prototype self-driving

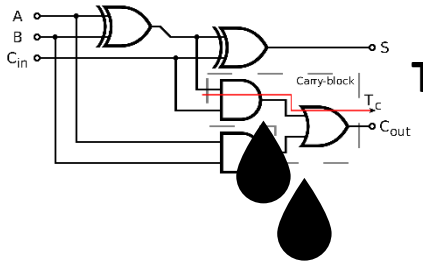
https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip/

~100 teraflops
~1000 W (system)
~1 TOPS/watt (SoC)

Significant innovation will be required in microelectronic materials and devices, sensing and computing architectures, and computer algorithms.

Squeezing every drop out of CMOS: Apple A13

- Apple's iPhone 11 main SoC processor
  - 7nm+ TSMC process
- Includes Lightening AMX 8-core Neural Engine accelerator IP
  - Apple spec: 5 TeraOps/s (TOPS) @ 8 bit precision
  - Power is ~2.5-5W (this is a reasonable estimate)
  - Hence, the best smartphone chip is ~ 1-2 TOPS/W
- **State of the art neural accelerator is about 500fJ to 1pJ per 8 bit operation**
  - Nearing digital limits?

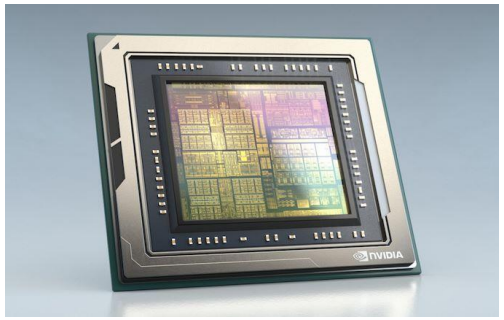


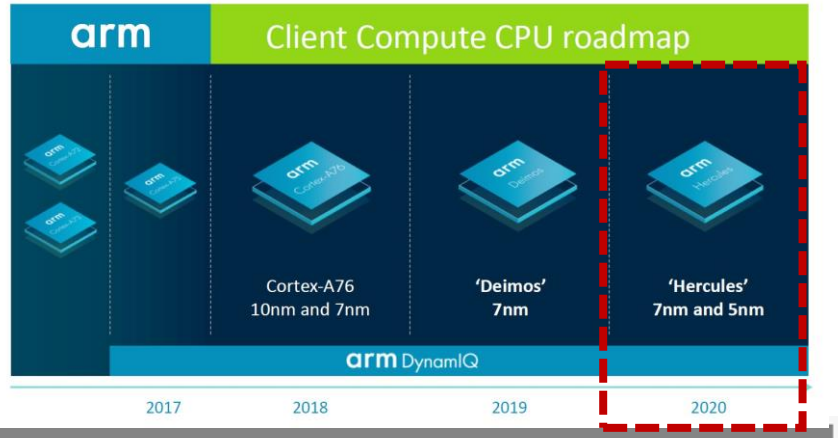
TOPS/W

apple.com, techinsights.com

Nvidia's next Drive platform

- Nvidia DRIVE AGX Orin
  - Available in 2022
  - Nvidia iGPU "AMPERE" + "HERCULES" Most advanced ARM core
  - 7nm and 5nm CMOS processes
  - **DRIVE AXG Orin: 2000 TOPS, 800W**
  - **May approach <u>2-3 TOPS/W</u> @ INT8**
  - Similar to Google TPUv1 (ISCA '17)
  - What will come after this?
  - Converging to ~1-5 TOPS/W for many published accelerator data, modern GPU, and Google TPU, and other benchmarked state of the art digital accelerators*





NVIDIA ARM SoC Specification Comparison

	Orin	Xavier	Parker
CPU Cores	12x Arm "Hercules"	8x NVIDIA Custom ARM "Carmel"	2x NVIDIA Denver + 4x Arm Cortex-A57
GPU Cores	Nvidia AMPERE iGPU	Xavier Volta iGPU (512 CUDA Cores)	Parker Pascal iGPU (256 CUDA Cores)
INT8 DL TOPS	200 TOPS	30 TOPS	N/A
FP32 TFLOPS	?	1.3 TFLOPs	0.7 TFLOPs
Manufacturing Process	5-7nm	TSMC 12nm FFN	TSMC 16nm FinFET
TDP	~65-70W?	30W	15W

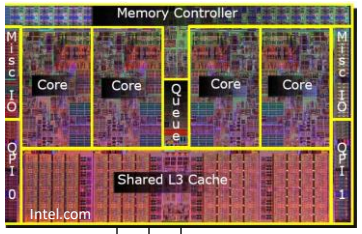
*TP Xiao, CH. Bennett, B Feinberg, S Agarwal, MJ Marinella, Appl. Phys. Rev 7, 2020.

Future Directions

Von Neumann Digital
Separate logic and memory structures

Memory

Logic Unit



Board Level Data Bus (i.e. DDR) : Slow and Power Hungry

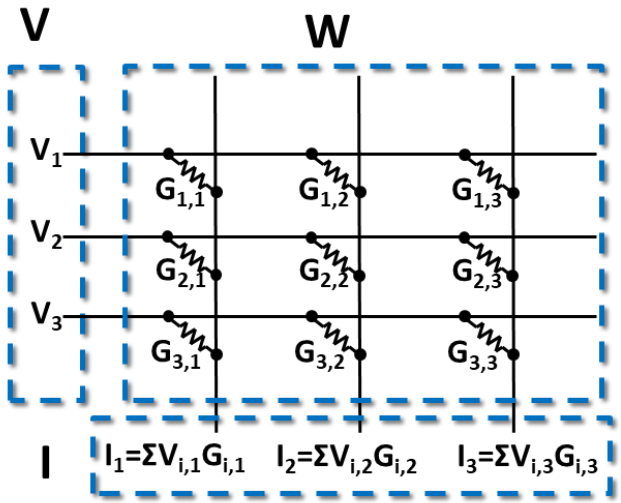
In-memory Parallel Analog
Use non-volatile memory

Mathematical

$$V^T W = I$$

$$\begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix} = \begin{bmatrix} I_1=\Sigma V_{i,1}W_{i,1} & I_2=\Sigma V_{i,2}W_{i,2} & I_3=\Sigma V_{i,3}W_{i,3} \end{bmatrix}$$

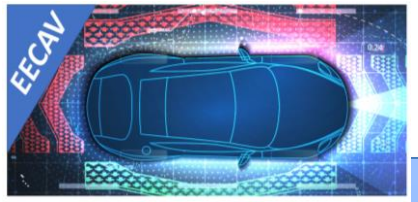
Electrical



Analog Compute in Memory: Eliminates Massive Data Movement

Path to Orders of Magnitude Improvement in TOPS/W

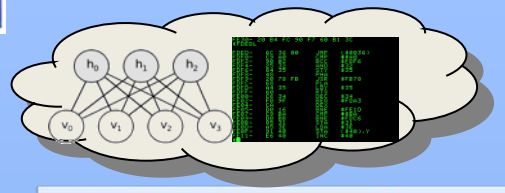
CoDesign

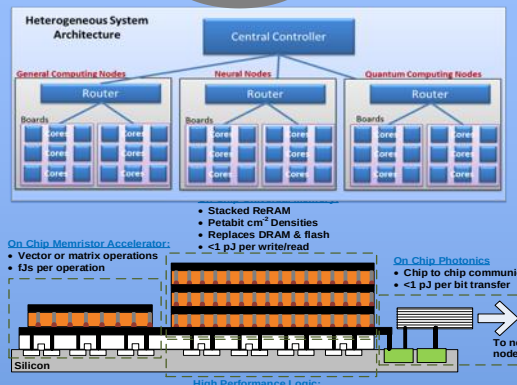


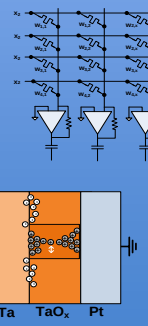
Top Down

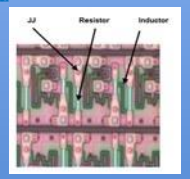
Bottom Up

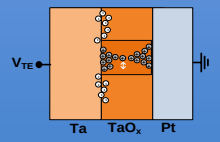
Multiscale CoDesign Framework

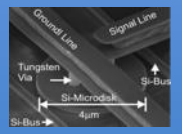


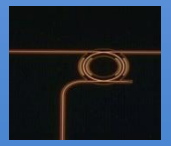


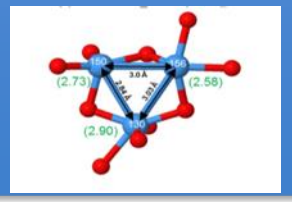












Programming paradigms

System architecture

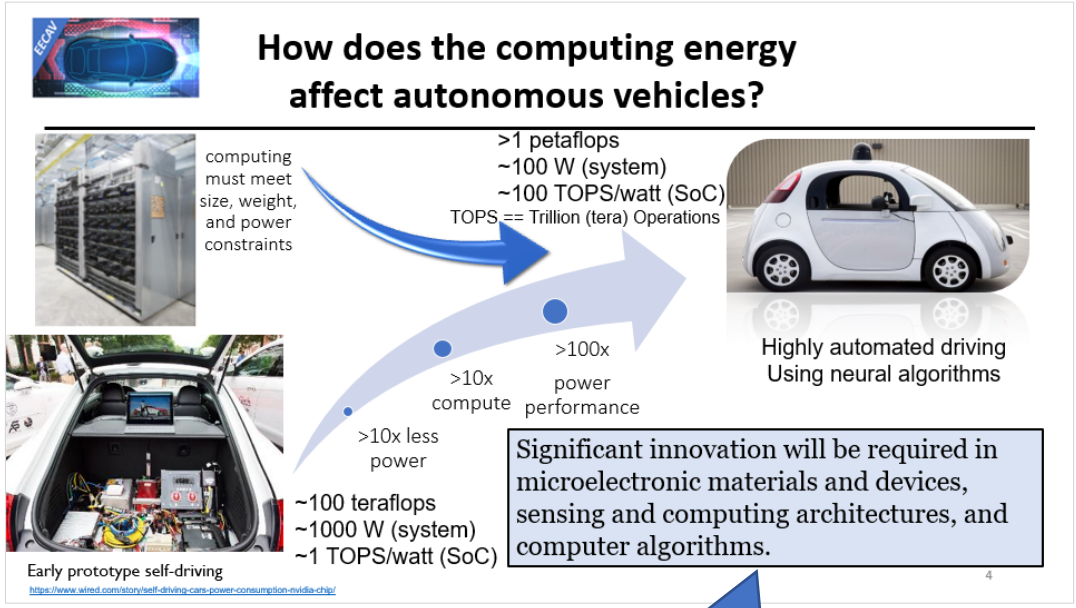
Component integration

Device demonstration

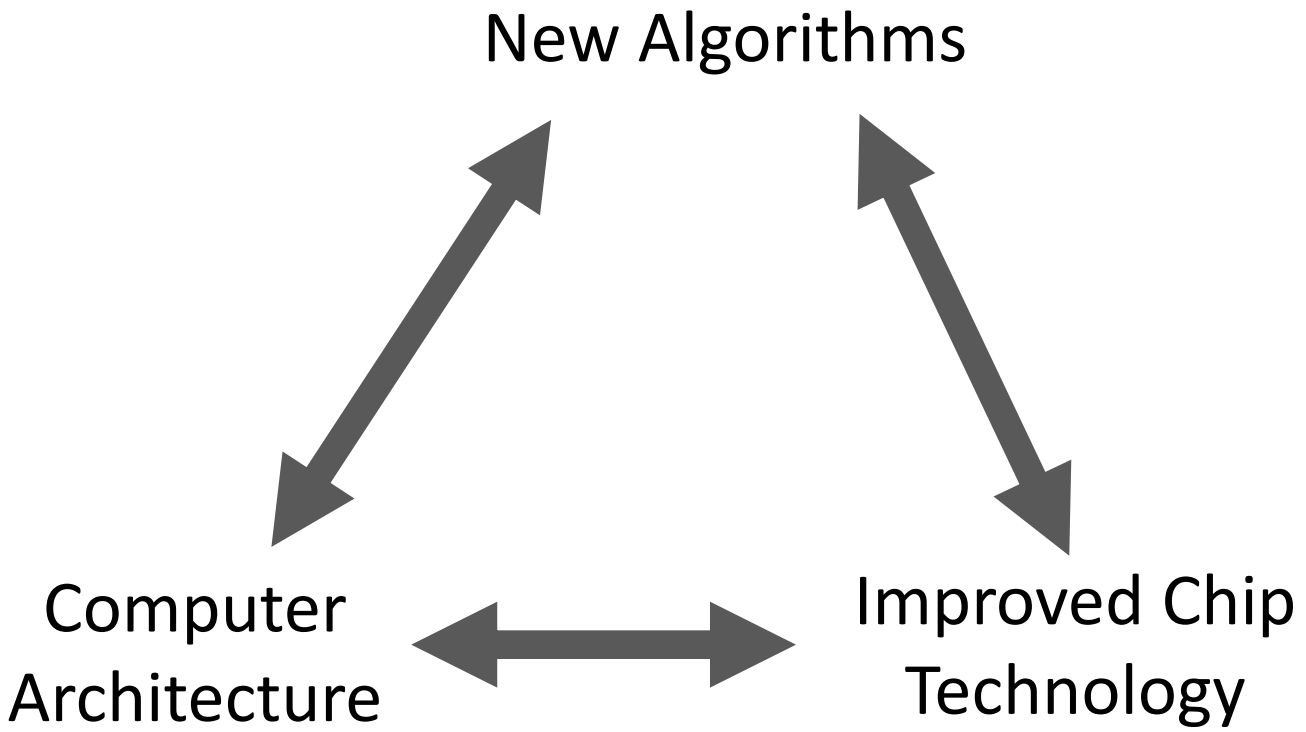
Device physics

Fundamental science

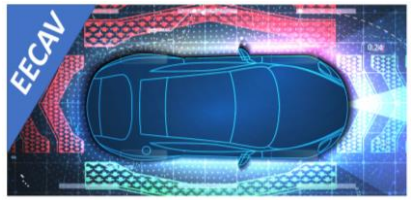
Energy Efficient AVs Require a HW/SW Co-Design Approach



This is not just about silicon



Together, we believe that a Hardware/Software Co-Design approach is essential for developing commercially viable, energy efficient retail AVs.



Safety & Sensing Processors Compared to Existing Market Segments

Embedded

IoT, Cameras,
Smart doorbell

Low Power,
Fast Wake,
Real Time

Mobile

Smart phones
Tablets

5G Networking,
Strong Security

Client

Laptops
Desktop, deskside

Multimedia,
Diverse Applications

Server

Data Center
Departmental Server

High Bandwidth,
Algorithm Accelerators

New Market Segment:

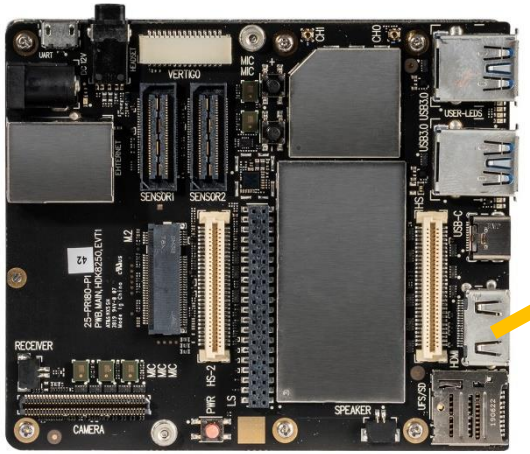
Safety & Sensing Processors

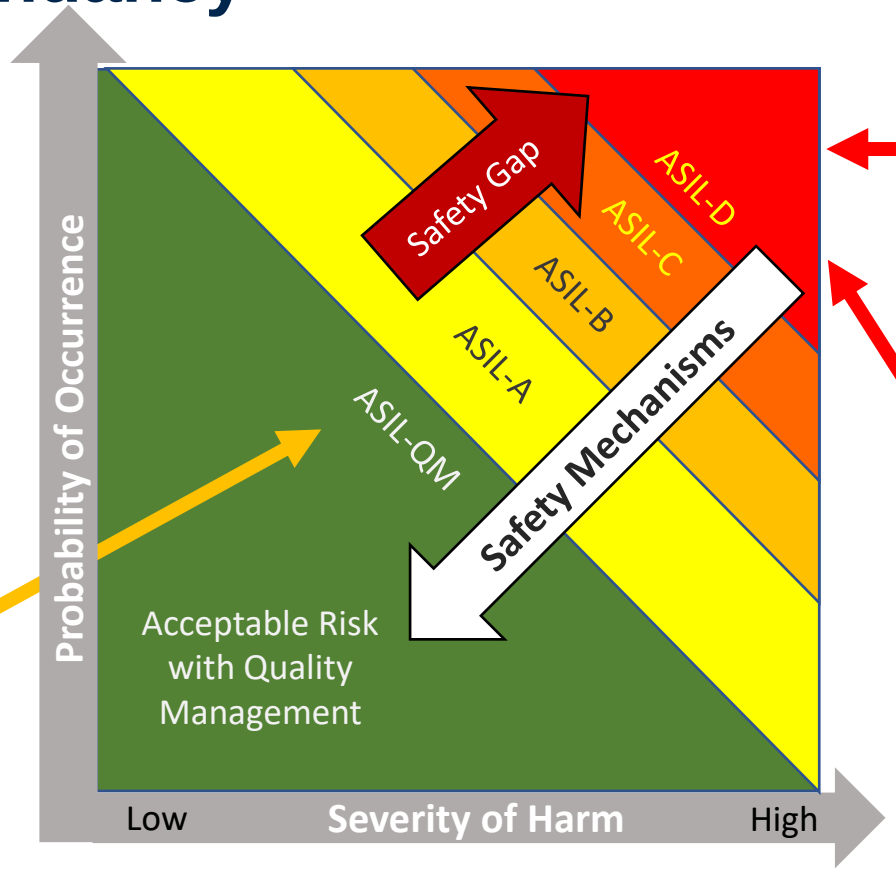
Energy Efficient
Modular and Composable
I/O Designed for High Resolution Sensors
Harsh Environment Ruggedization
Fail Operational Safety

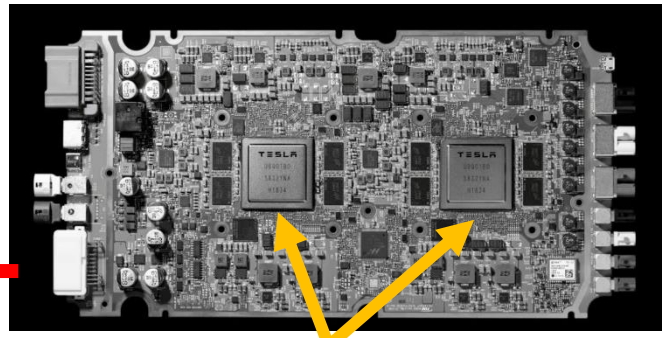
AV Safety Requirements add to Energy Costs, Creating Efficiency Opportunities

Current generation of AV solutions retrofit safety features onto data center devices, often via full module redundancy

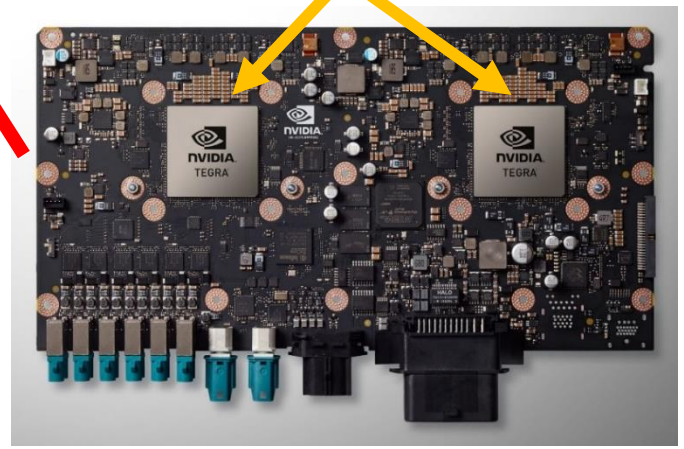
Data Center Hardware:
No built-in safety mechanisms, requires add-on safety solutions







Redundant devices needed for safety doubles power consumption



SAE J3016 Driving Automation Levels Do *NOT* Tell You "What the Car is Doing"

SAE Automation Levels define liability.

L4 is not "doing more" than L3, but the steering wheel goes away.

Not Incremental and Stepwise:

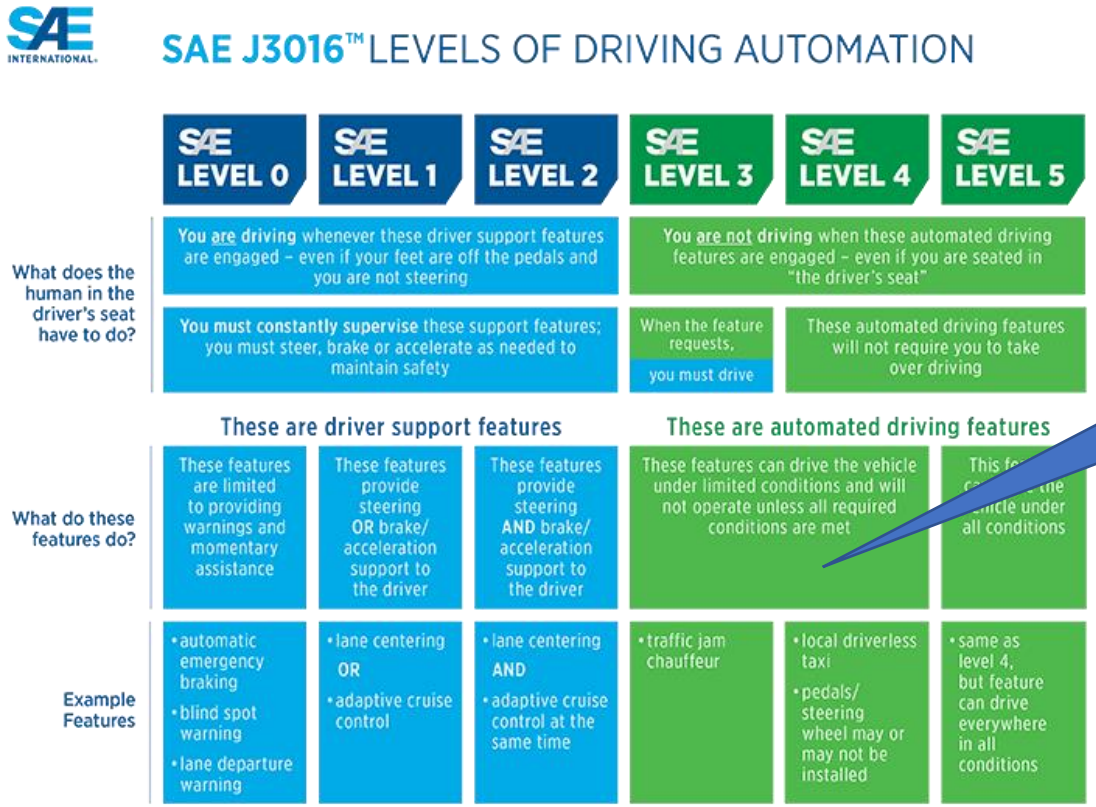
- L4 is L3 w/out Driver Monitoring
- Difference between "limited" and "unlimited" can be large

L4 does not define "limited conditions"

- Parking lot shuttle
- Warehouse forklift
- Geofenced Robo-taxi

L5 operates "everywhere in all conditions"

- Super-human abilities?



"These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met."

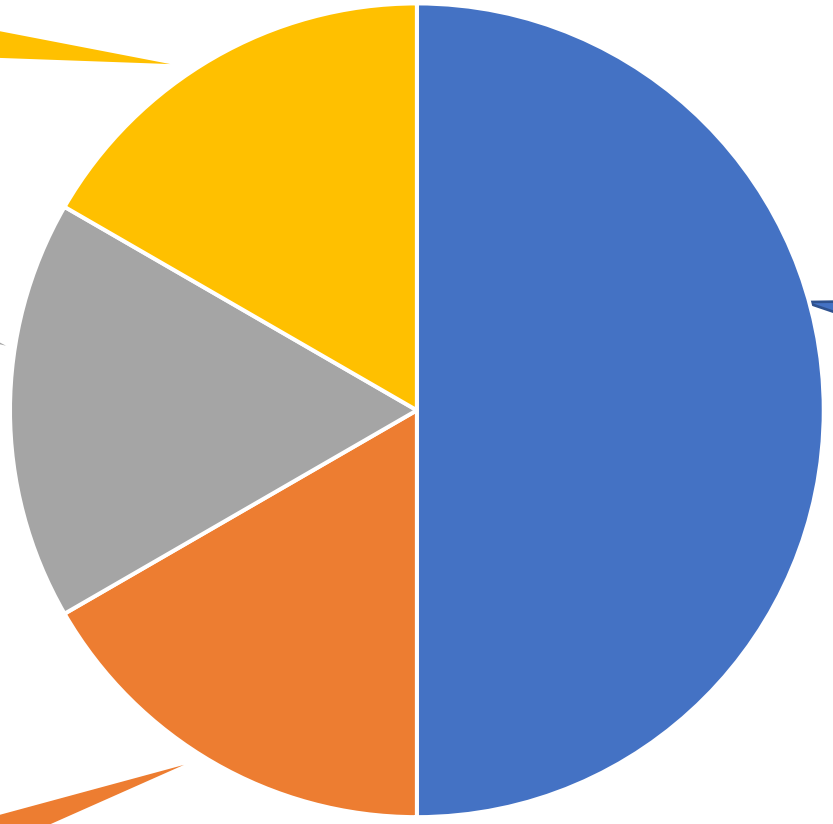
Typical EV-AV Energy Use: Most Efficiency Opportunities are in Computing

Data fusion, planning, actuation

Proportional to sensor data rate and model complexity

ISP & ML are sometimes merged, is this an improvement or just different?

Image processing directly proportional to sensor data rate.

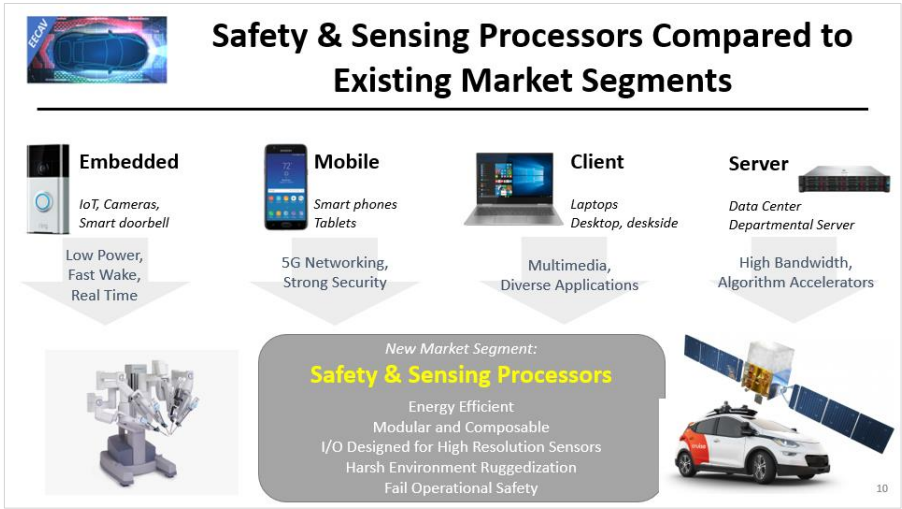


Propulsion *Image Processing*
Machine Learning *Everything Else*

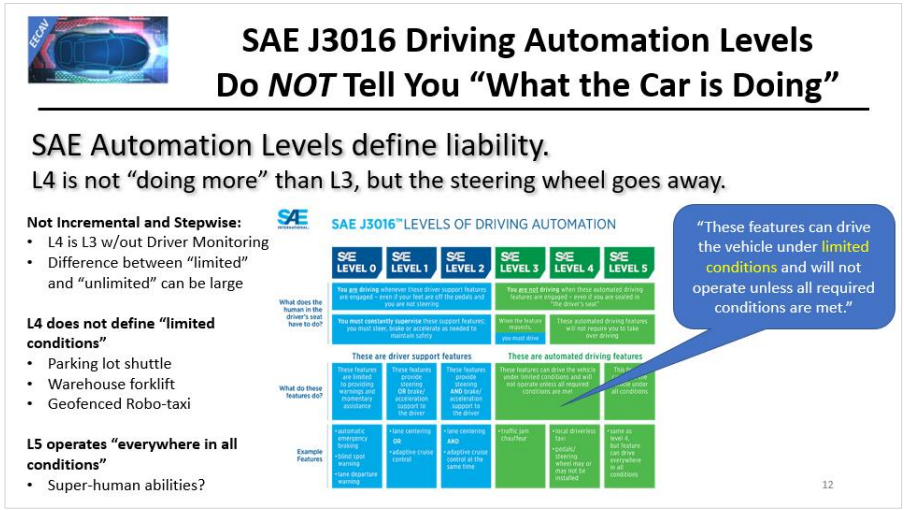
Electric motors already highly efficient, limited room for improvement.

- Only half the energy is used for propulsion.
- Equal energy for image processing, machine learning, and all other processors.

Discussion Topics

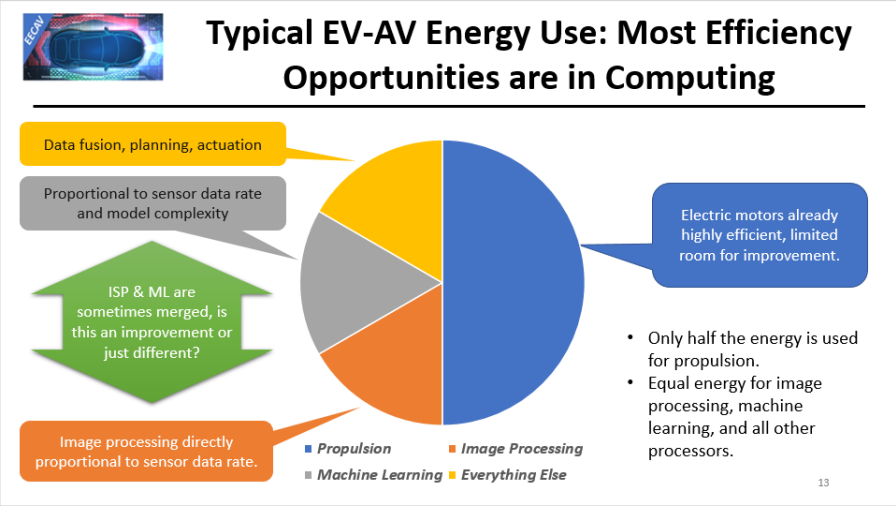


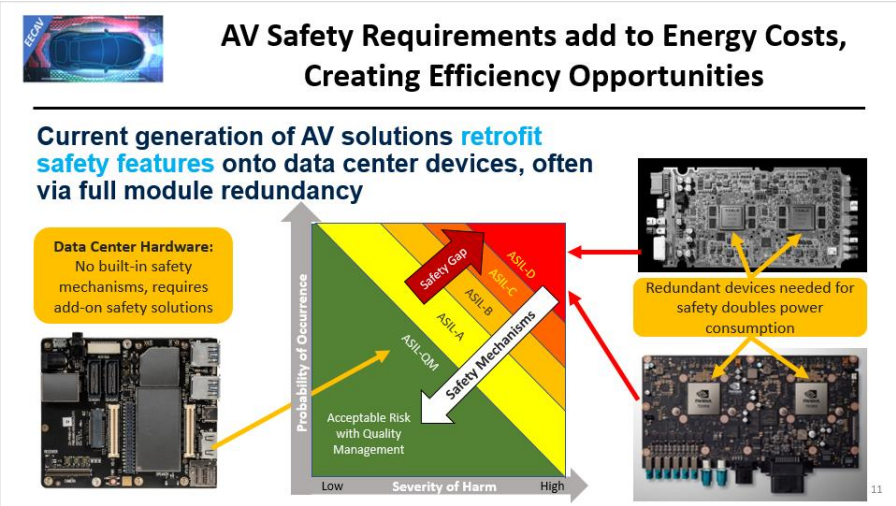
- Do existing product segments cover AV system needs?



- How do you describe "what the vehicle is doing"?
- What is the relationship between what the car is doing and power utilization?

Discussion Topics





- Can merging ISP and ML help?
  - Some ISP still needed for human vision
  - Implications of replacing a sensor?
  - Smart sensors vs. Central Computing
  - Is there a conservation of complexity that makes moving this work pointless?

- Is Fail Operational safety economically feasible?
- Can fail operational redundancy compute capacity also be used to implement capabilities at lower levels of automation?