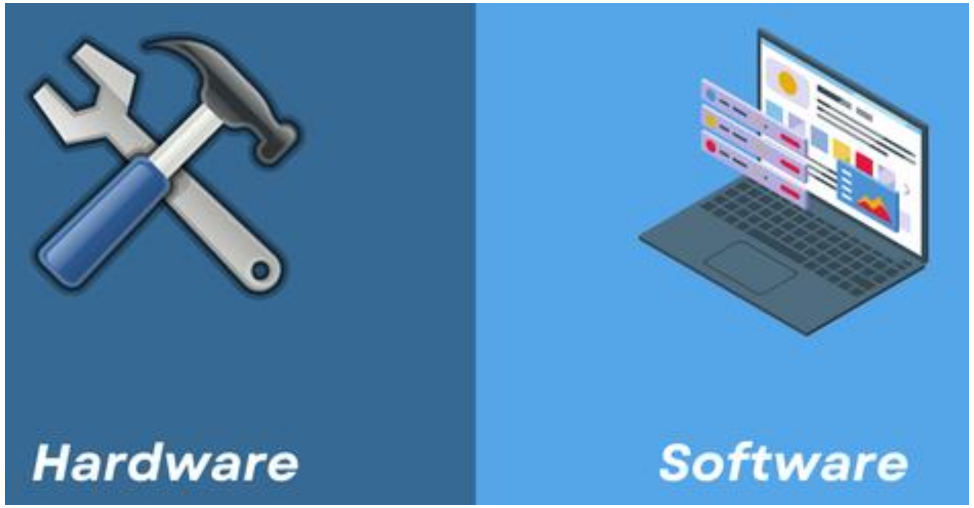
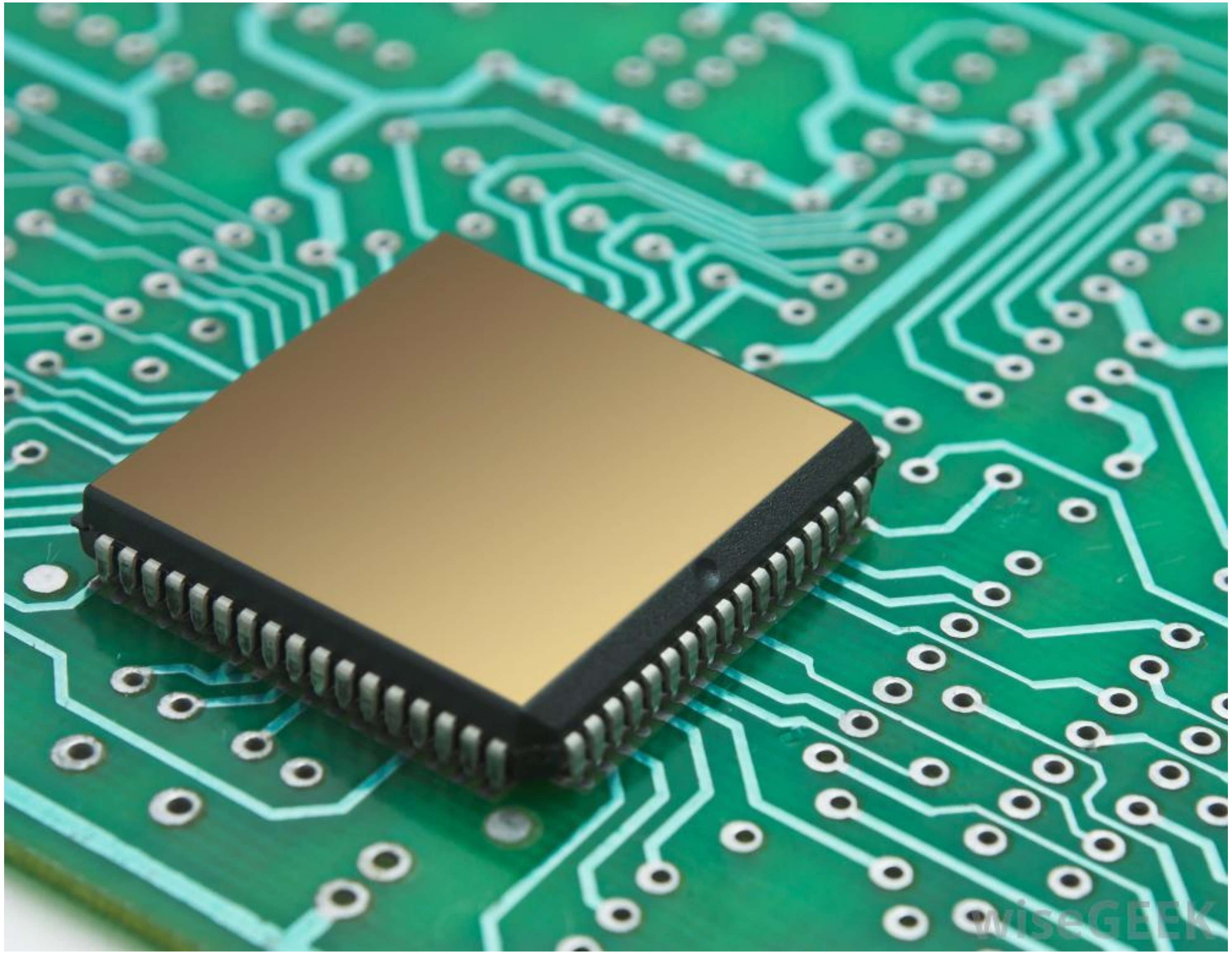
L1.3 HW/SW Platforms, Ethics



Zonghua Gu, Umeå University

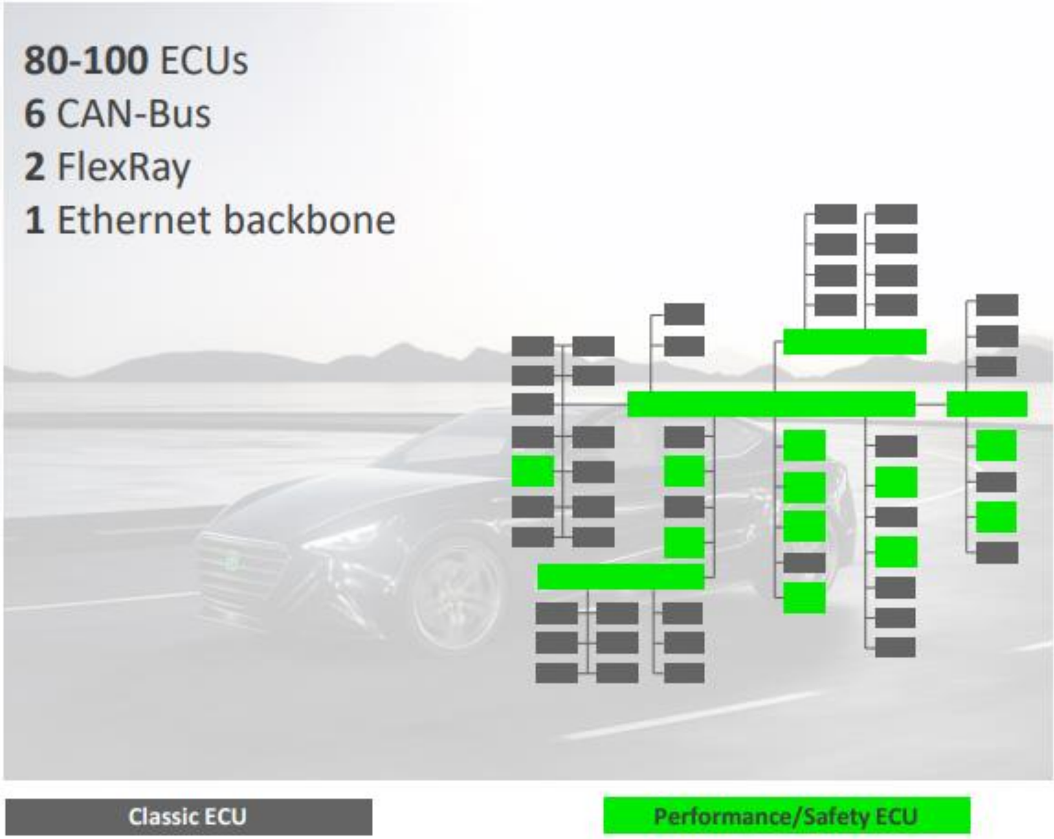
Nov. 2023

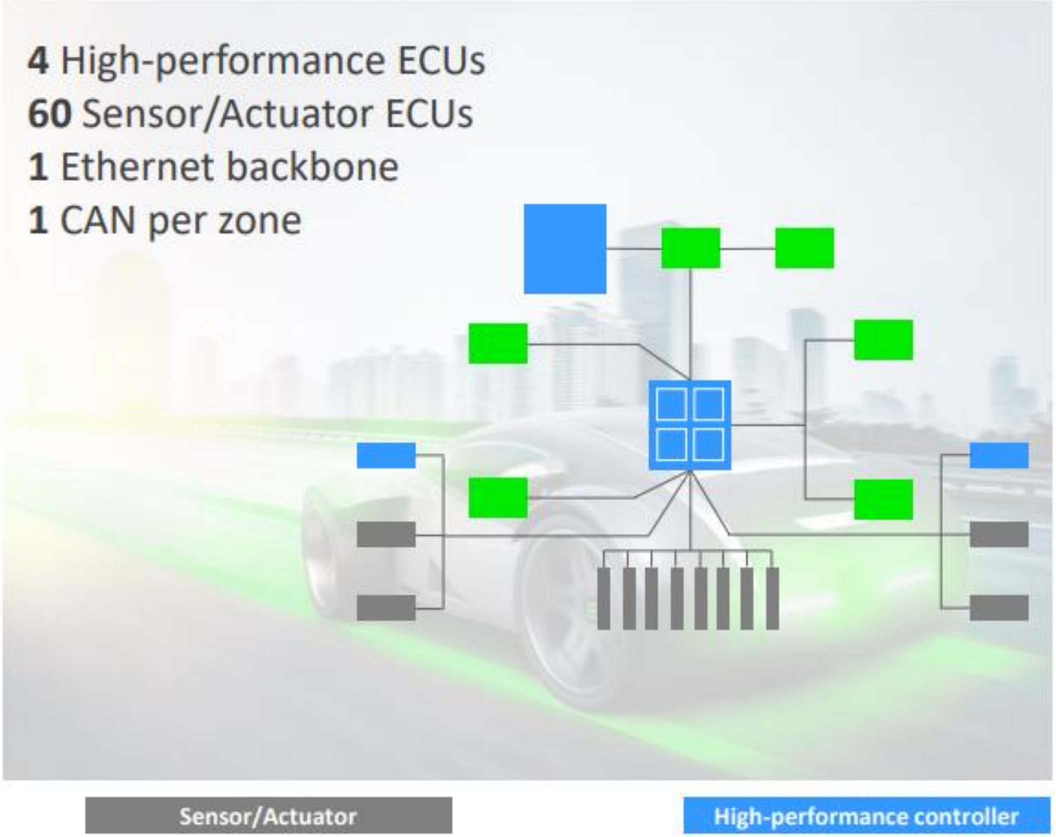
Hardware Platforms



Evolution of Automotive E/E Architecture

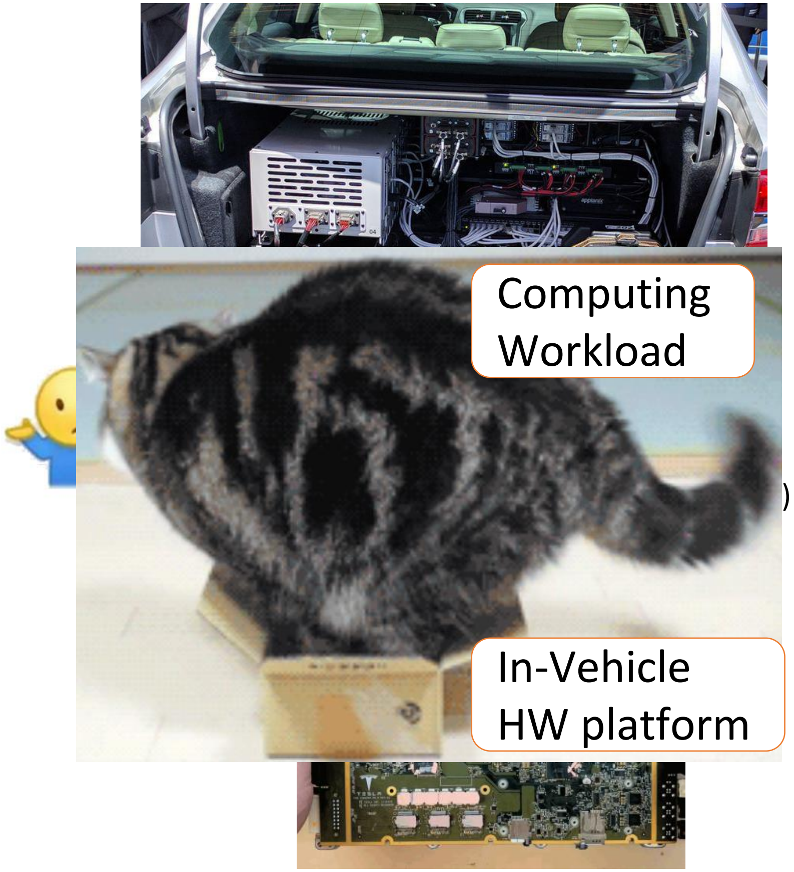
- From many (~80-100) distributed and networked Electronic Control Units (ECUs) to a few (~4) high-performance ECUs with massive computing power, and large number of (~60) small ECUs for interfacing with sensors and actuators





Automotive E/E Architecture Trends

- Automotive E/E architecture trends:
  - HW platform consolidation
    - Centralized architecture with a few ECUs helps simplify system architecture, reduce network load, and improve system reliability
  - HW platform miniaturization
    - A trunkful of electronics hinders consumer acceptance and mass deployment
    - System-on-Chip technology helps to achieve high-performance computing with small form factor and low power consumption
- Significant HW resource constraints in terms of processor speed, memory size, and network bandwidth
  - Level of constraints depends on the application context, e.g., mobile robots and consumer drones have more severe constraints than passenger cars



Tesla’s FSD Processor (2022)

Design Objectives vs. HW Constraints

- **Design Objectives**
- Safety
  - System failures may be deadly
- Hard real-time
  - Deadline misses may compromise safety
- Security
  - Malicious attacks may compromise safety

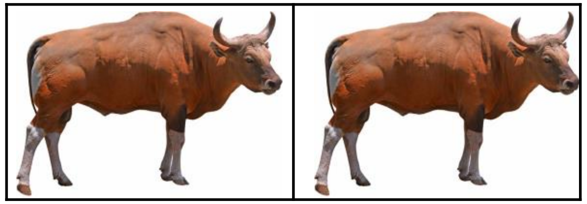
- **HW Constraints**
- Size/Weight
  - Must be compact and lightweight (cannot have a trunkful of electronics)
- Power consumption
  - Power consumption of electronics (sensors and computing hardware) for AD may be 100x that of a vehicle with regular ADAS. This drains battery and implies increased fuel consumption or reduced range for EVs
- Cost
  - Important for mass deployment
  - Cost of electronics in an experimental AV often exceeds cost of the original vehicle.

SoC Hardware for AVs

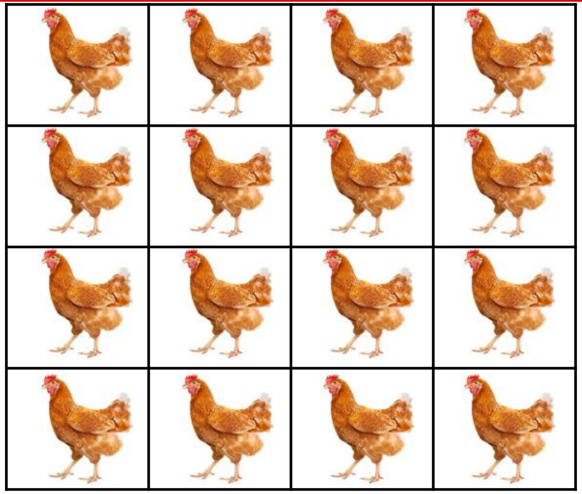
- The most compute-intensive workload is Deep Learning inference
- Many vendors provide SoC (System-on-Chip) products that integrate CPU cores with specialized computational engines for Deep Learning:
  - GPU (Graphics Processing Unit)
    - NVIDIA is the only serious player
    - Other GPU venders, e.g., AMD, ARM, Intel, focus on computer graphics instead of general-purpose computing (GPGPU)
  - FPGA (Field-Programmable Gate Arrays)
    - Xilinx, Intel Altera
  - ASIC (Application-Specific Integrated Circuit)
    - An explosion of specialized ASICs for Deep Learning in recent years, with hundreds of companies and products ranging from high-performance to embedded
  - DSP (Digital Signal Processor)
    - Mainly for image preprocessing, e.g., products from Texas Instruments

CPU vs. GPU/FPGA/ASIC

- CPU is designed for general-purpose workloads
- GPU has much simpler control logic than CPU, hence has more computational elements (Arithmetic Logic Units)
  - Specialized for highly-parallel workloads, e.g., matrix-multiply, which is a core operation in Deep Learning training and inference
  - Similarly for FPGA and ASIC

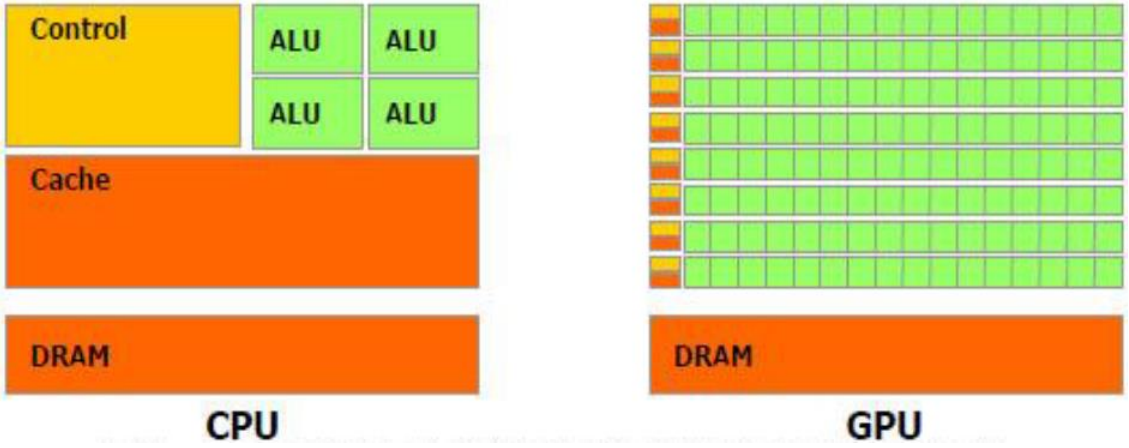


CPU (2 oxen)



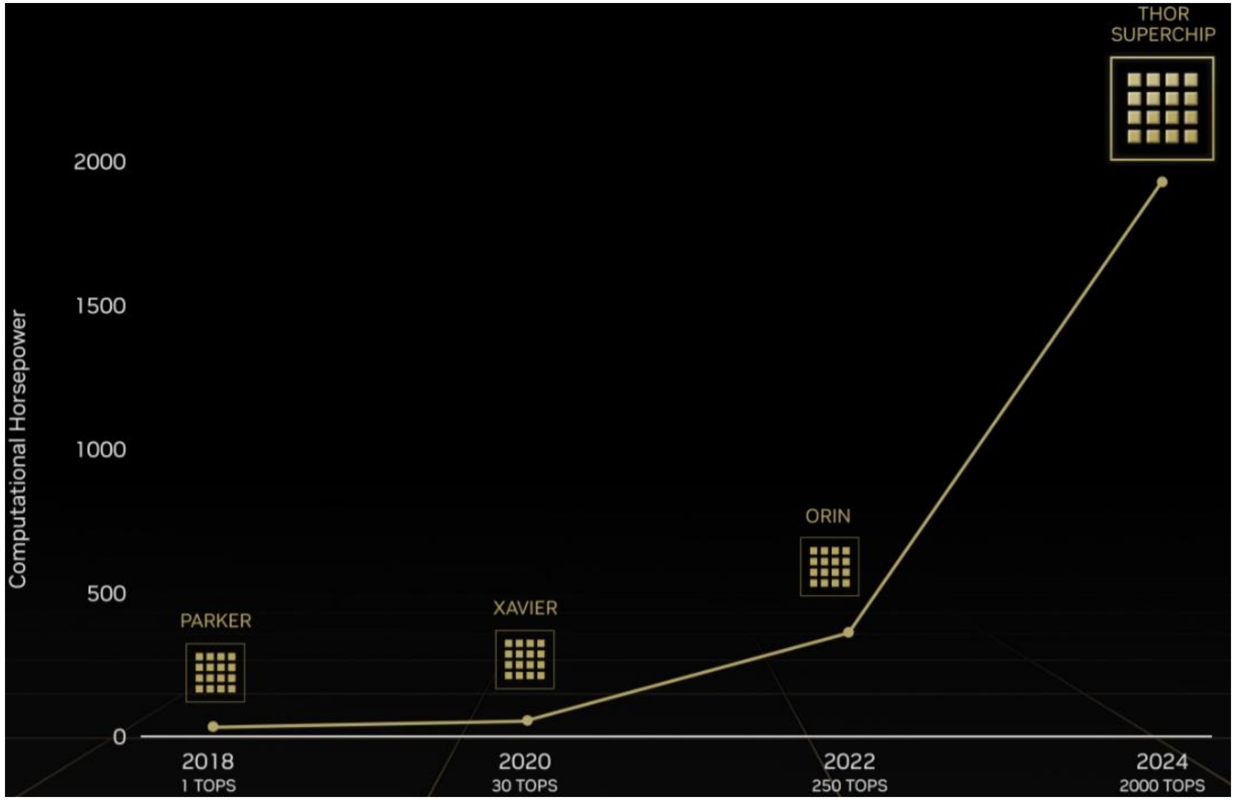
GPU (1024 chickens)

"If you were plowing a field, which would you rather use? 2 oxen, or 1024 chickens?" S. Cray.



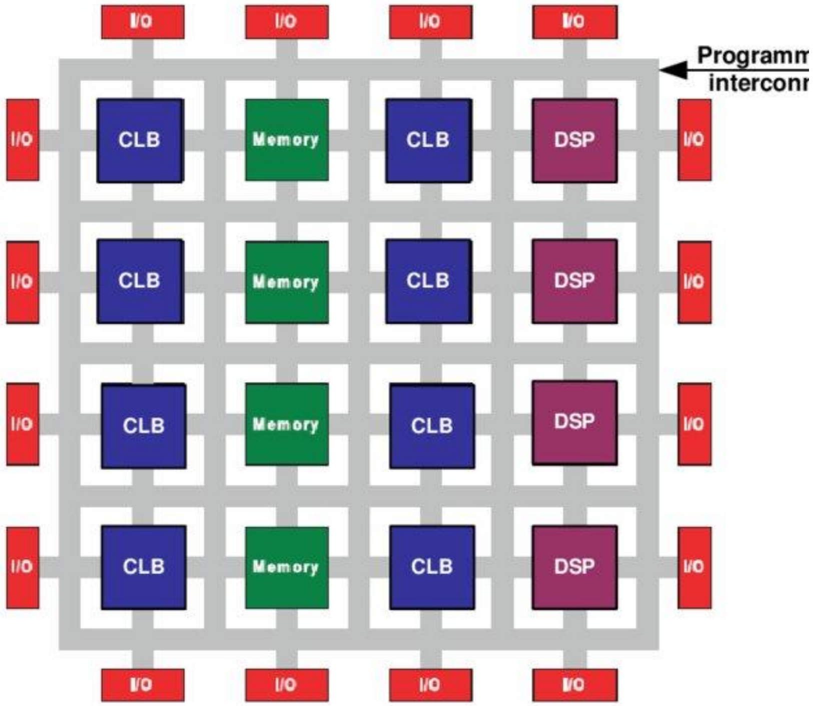
NVIDIA DRIVE GPU-based HW Platform

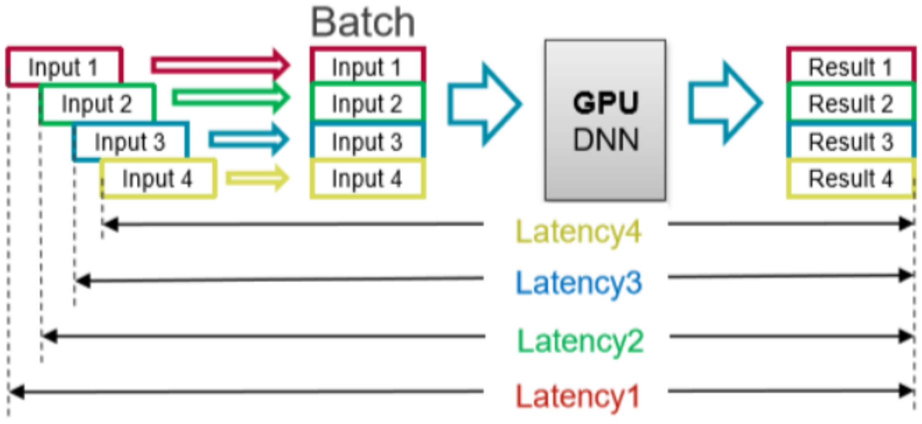
- A family of products, ranging from the low-end Parker to the latest high-end THOR with 2000 TOPS (Tera Operations Per Second)
- Besides CPU and GPU cores, also includes NVDLA (NVIDIA Deep Learning Accelerator), an ASIC for Deep Learning inference

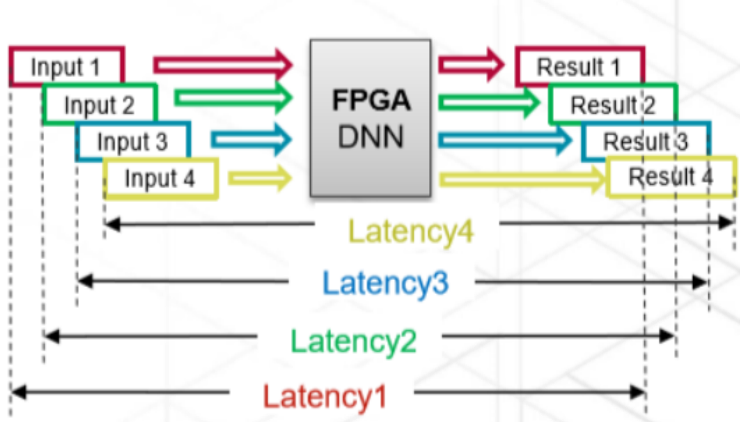


FPGA

- FPGA is reprogrammable hardware, consisting of an array of Configurable Logic Blocks (CLBs) and interconnections which can be configured at design time or runtime. FPGA has advantages over GPU for Deep Learning inference tasks
- GPU performs computation in batches for efficient exploitation of SIMD (Single Instruction, Multiple Data) computation model
  - This is ideally suited for training tasks, with well-known algorithms such as Stochastic Gradient Descent with mini-batches
  - But not ideal for inference tasks
    - Larger batch size leads to high throughput, but also high and nondeterministic latency for each data item
    - Smaller batch size leads to low computation efficiency
- FPGA can perform "batch-less" inference
  - Low and deterministic latency for any batch size

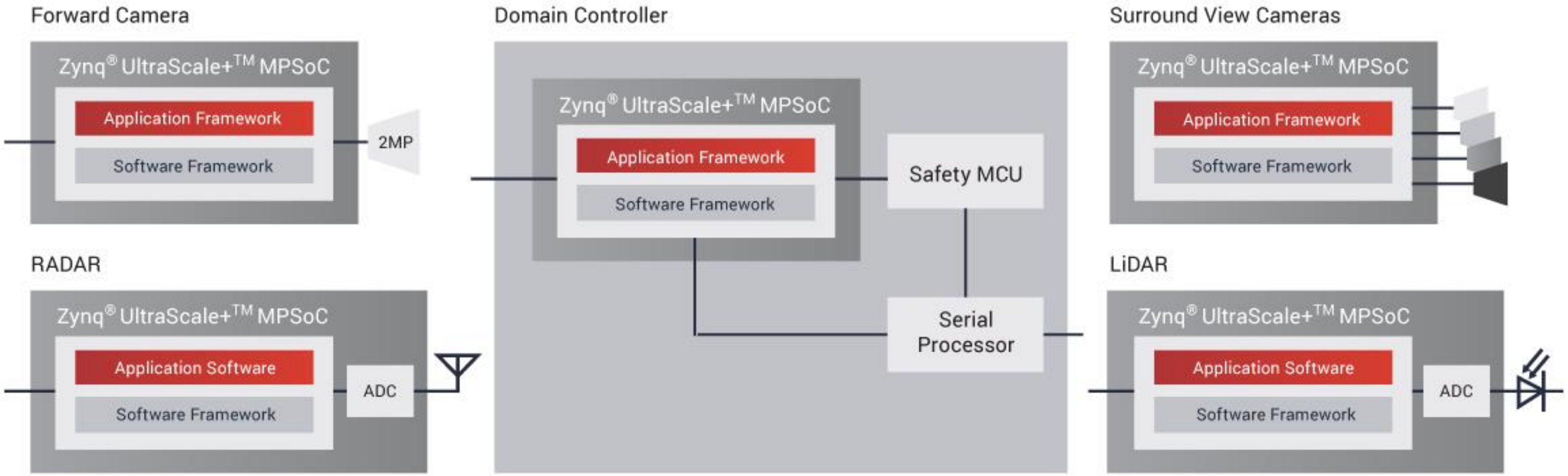






FPGA for Automotive

- FPGAs can be integrated into smart sensors (camera, Lidar, radar), or serve as central compute engine in domain controller or AD computer
  - Xilinx FPGAs have 90% market share in Lidar signal processing
  - Intel went into the automotive market, with acquisition of Altera in 2015



https://www.xilinx.com/applications/automotive/automated-driving.html

ASIC

- ASICs for Deep Learning are often called Neural Processing Units or AI accelerators
- ASICs, thanks to dedicated circuit design, may achieve up to 10x in computation efficiency and power consumption compared to CPU/GPU, and less dramatic, but still significant improvement compared to FPGA. The drawback is loss of programmability and flexibility.
- Almost every chip vendor provides some kind of AI accelerator, e.g. Google’s Tensor Processing Unit (TPU)

MobileEye EyeQ Series

- ASICs for computer vision

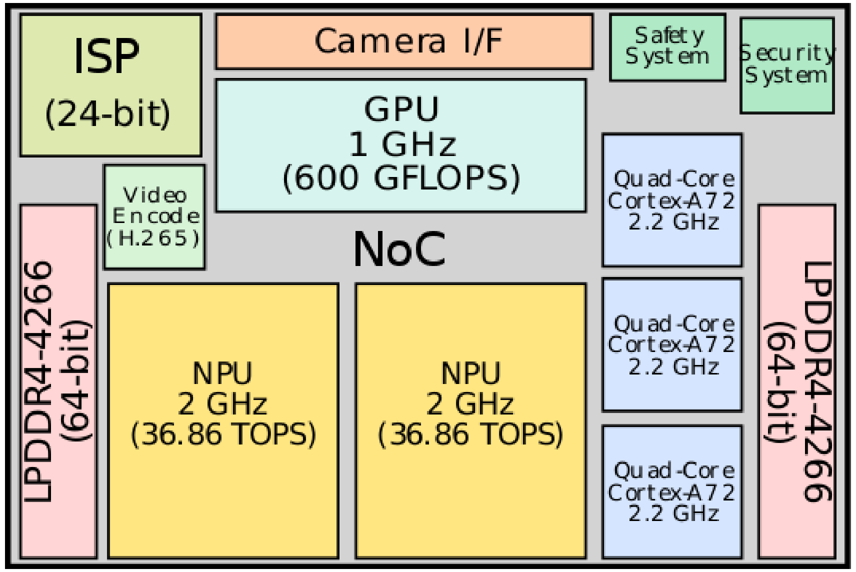
Telsa AutoPilot HW Evolution

- 2014~2016: HW1 based on Mobileye EyeQ3
- 2016~2019: HW2 based on NVIDIA DRIVE PX2
- 2019~2023: HW3 FSD processor
- 2023~now: HW4 FSD processor

	HW Platform	Processor Cores	Sensor Configuration
HW1	Mobileye EyeQ3	NVIDIA Tegra 3	2 cameras (front and back) 1 Radar 12 ultrasonic sensors
HW2	NVIDIA DRIVE PX2	1 NVIDIA Pascal GPU, 2 NVIDIA Parker SoCs, 2 Infineon TriCore CPUs	8 cameras 1 Radar 12 ultrasonic sensors
HW3	FSD	NPU	8 cameras 1 Radar 12 ultrasonic sensors
HW4	FSD	NPU	8 cameras

Tesla FSD Processor

- HW3 incorporates 3 quad-core Cortex-A72 clusters for a total of 12 CPU cores operating at 2.2 GHz, a GPU operating at 1 GHz, 2 NPUs operating at 2 GHz
  - Neural Processing Unit (NPU): ASIC for Deep Learning inference
- HW4 has increased computing power, shown on the right



HW3 floorplan

HW3	HW4
CPU: Samsung Exynos-IP, 12 cores@2.2 GHz	CPU: Samsung Exynos-IP, 20 cores@2.35 GHz
2 NPUs@2.0 GHz, 36 TOPS	3 NPUs@2.2 GHz, 50 TOPS
1 GPU@1.0 GHz	No GPU
Process: 14 nm	Process: 7nm or N4 (4nm class)
Camera resolution: 1.2MP	Camera resolution: 5MP

https://www.autopilotreview.com/tesla-hardware-4-rolling-out-to-new-vehicles/

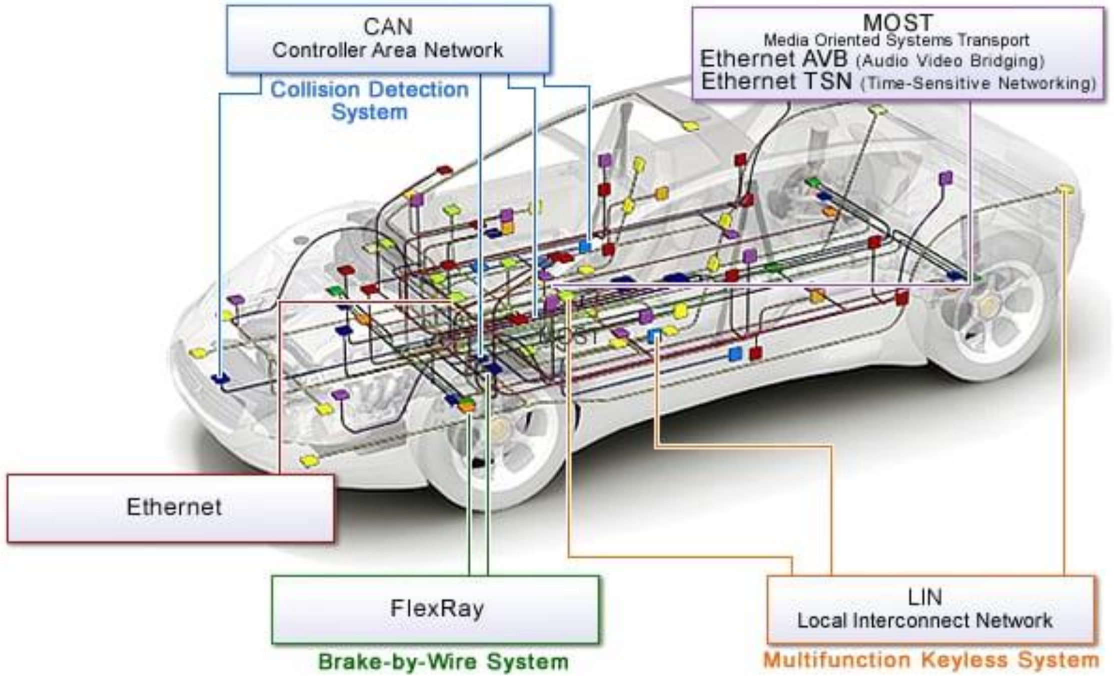
Products from Automotive OEMs and Suppliers

- Some companies offer integration solutions based on chip products from other vendors
  - Delphi/Audi zFAS (Central Driver Assistance Center)
    - based on NVIDIA Tegra K1 and Mobileye EyeQ3
  - ZF ProAI
    - based on NVIDIA DRIVE PX2
  - Bosch AI Car Computer
    - based on NVIDIA DRIVE AGX Xavier
  - Many others
    - Continental ADCU; Visteon DriveCore; NXP BlueBox; Renesas R-Car…

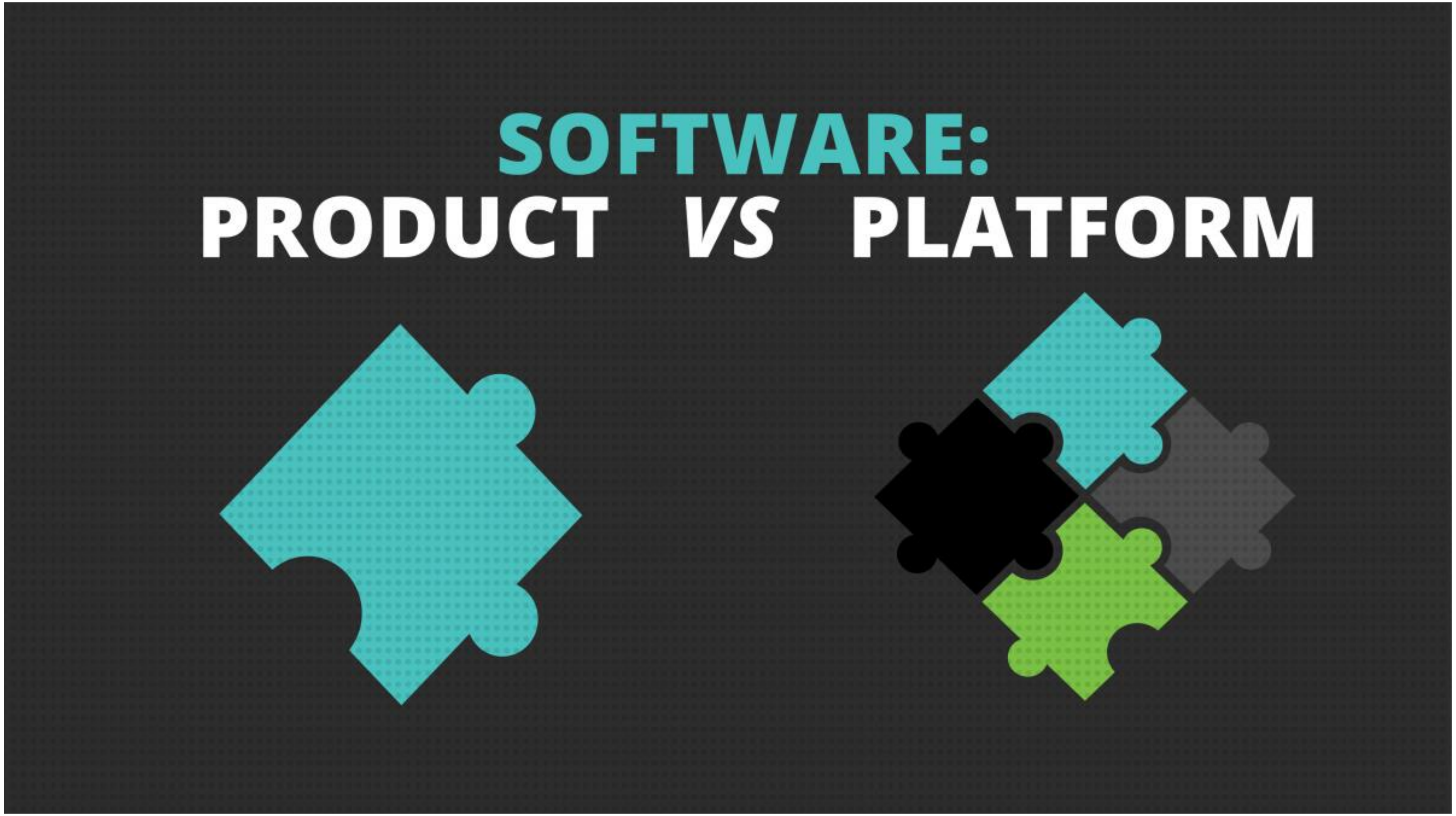
In-Vehicle Networks

- Ethernet as high-bandwidth backbone network
  - Ethernet Time-Sensitive Networking (TSN) posed to be the dominating standard protocol
  - Regular Ethernet is also used for diagnostics
- Media Oriented Systems Transport (MOST) for multimedia (audio/video) transmission
- FlexRay for safety-critical X-by-Wire, where X stands for brake, steer, drive..
- CAN (Controller Area Network) for low-bandwidth network and interfacing with sensors/actuators
- LIN (Local Interconnect Network) for body electronics, e.g., door, light, rearview mirrors…

Protocol	Datarate	Applications
LIN	10 kbps	Sensor/Actuator networking
Low-speed CAN	125 kbps	Body and comfort
High-speed CAN	1 Mbps	Powertrain and chassis
CAN-FD	8 Mbps	Powertrain and chassis
FlexRay	20 Mbps	Powertrain and chassis
MOST150	150 Mbps	Infotainment (audio/video)
Ethernet	100 Mbps to 10 Gbps	Backbone network

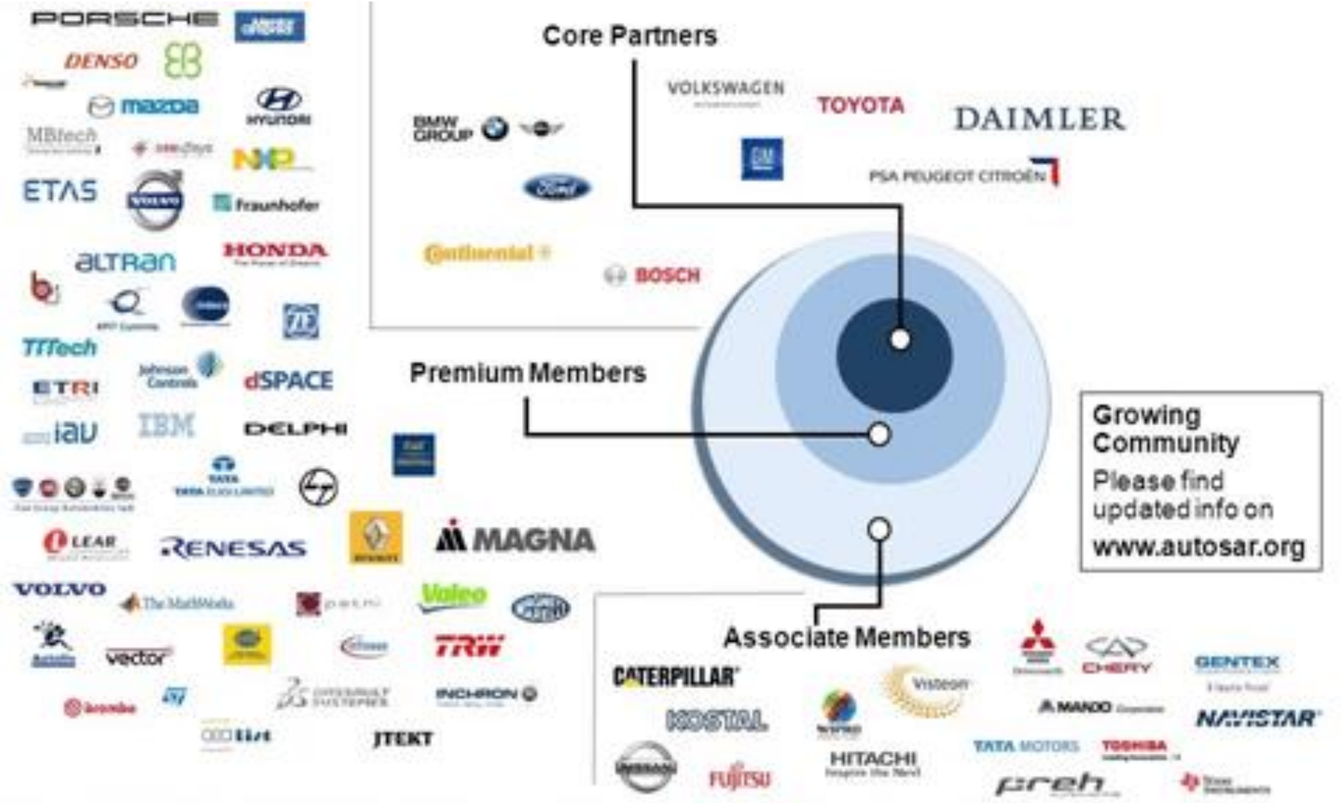


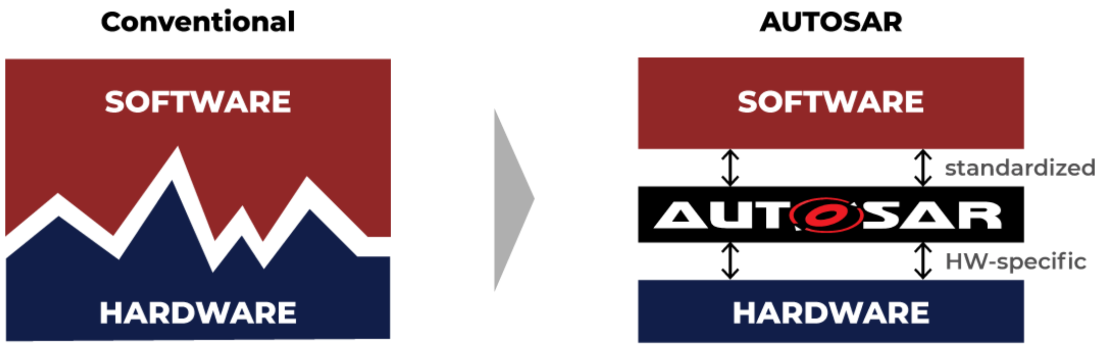
Software Platforms



AUTOSAR Consortium

- AUTomotive Open System ARchitecture (AUTOSAR) is a global development partnership of automotive interested parties founded in 2003. It pursues the objective to create and establish an open and standardized software architecture for automotive Electronic Control Units (ECUs)

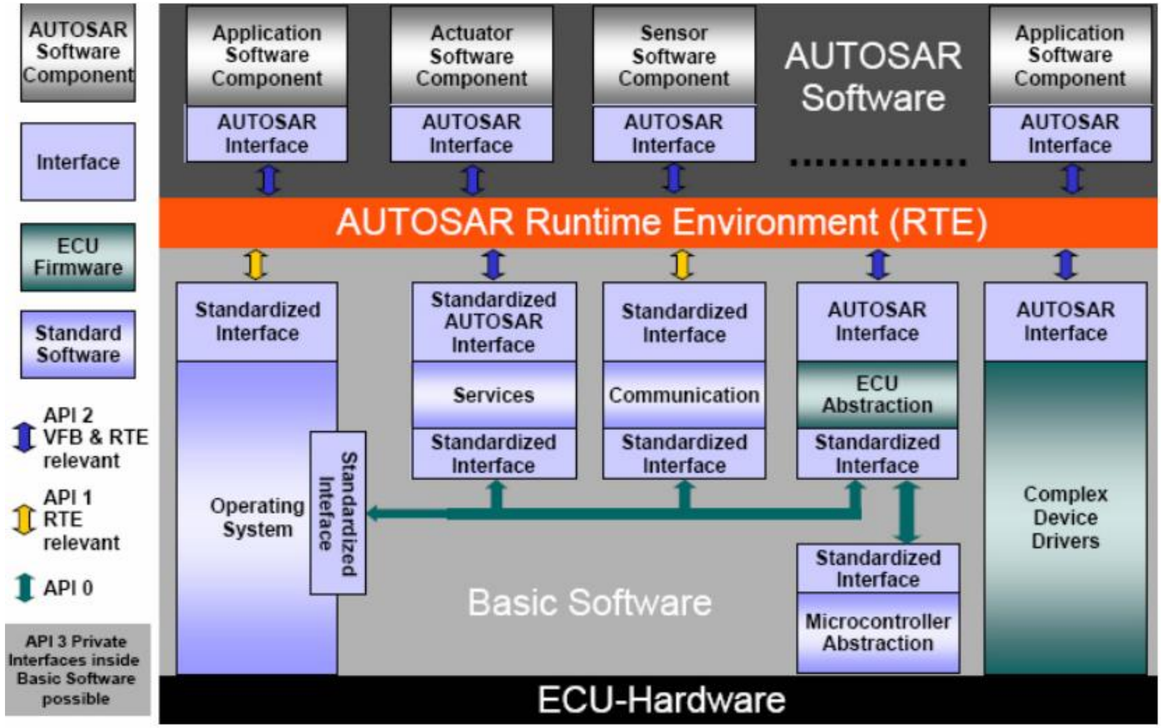


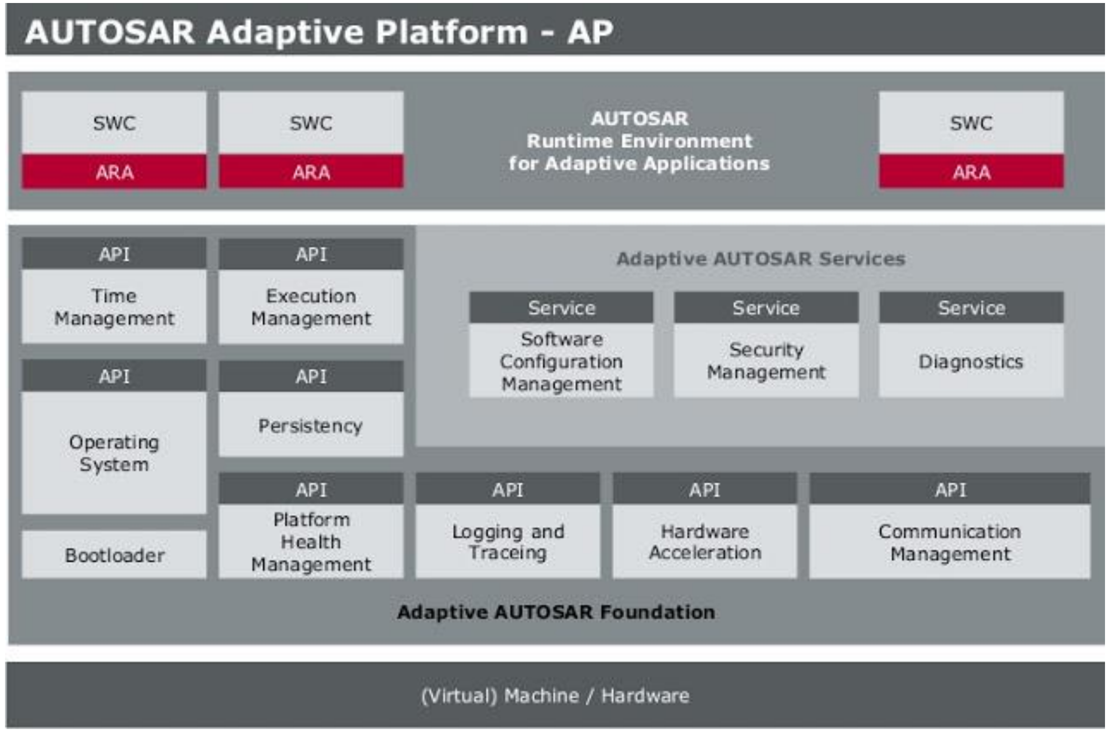


https://promwad.com/news/autosar-layer-architecture

AUTOSAR Classic and Adaptive Platforms

- AUTOSAR-CP (Classic Platform) is an industry standard for resource-constrained safety-critical ECUs
- AUTOSAR-AP (Adaptive Platform) is an industry standard for high-performance multicore automotive ECUs
  - AUTOSAR-AP allows dynamic linking of services and clients during ECU runtime, which facilitates Over-the-Air (OTA) Update

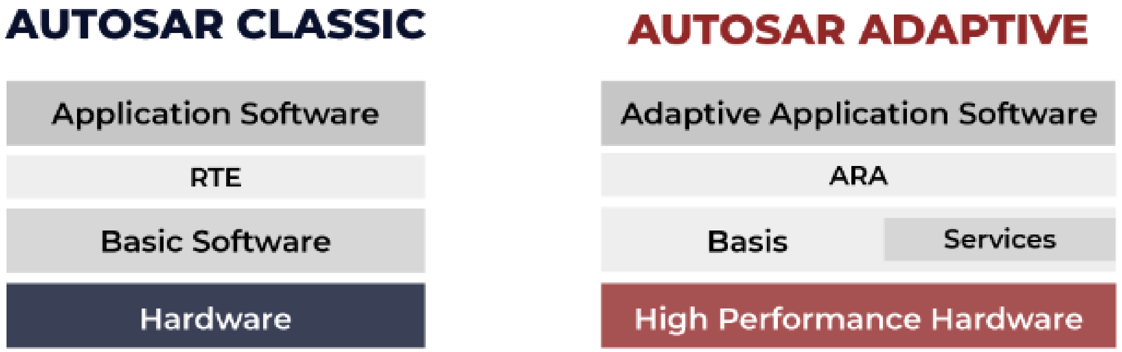




AUTOSAR Classic vs. Adaptive

	AUTOSAR-CP	AUTOSAR-AP
Use Case	Embedded System, e.g., Engine Control, X-by-Wire	High-Performance Computing, e.g., Deep Learning-based Perception/Planning
Operating System	OSEK	POSIX
Communication	Statically Configured Signal-based (CAN, FlexRay)	Dynamically Configured Service-Oriented (SOME/IP)
Code Execution	Code executed directly from ROM	Code loaded into RAM and executed from RAM
Dynamic Updating (for OTA)	N/A	Runtime Configuration Change
API	RTE (AUTOSAR Run Time Environment)	ARA (AUTOSAR Runtime for Adaptive Applications)
Real-Time Constraints	Hard	Soft

Classic vs. Adaptive Autosar https://www.youtube.com/watch?v=xsUCcetEcSk



Integration of Multiple SW Platforms

- AUTOSAR CP (labeled C) is used for safety-critical ECUs for low-level control and interfacing with actuators
- AUTOSAR AP (labeled A) is used for high-performance AD computer.
- Non-AUTOSAR (labeled N) may be Linux or Android, for non-safety-critical IVI (In-Vehicle Infotainment) and COTS (Commercial Off-the-Shelf) applications.

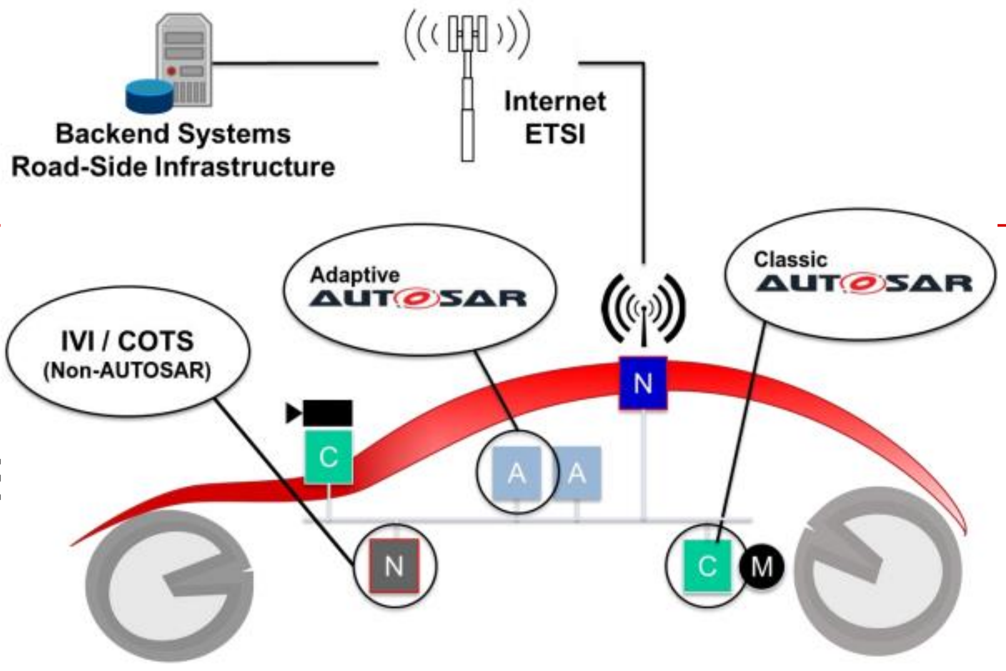


Figure 2-1 Exemplary deployment of different platforms

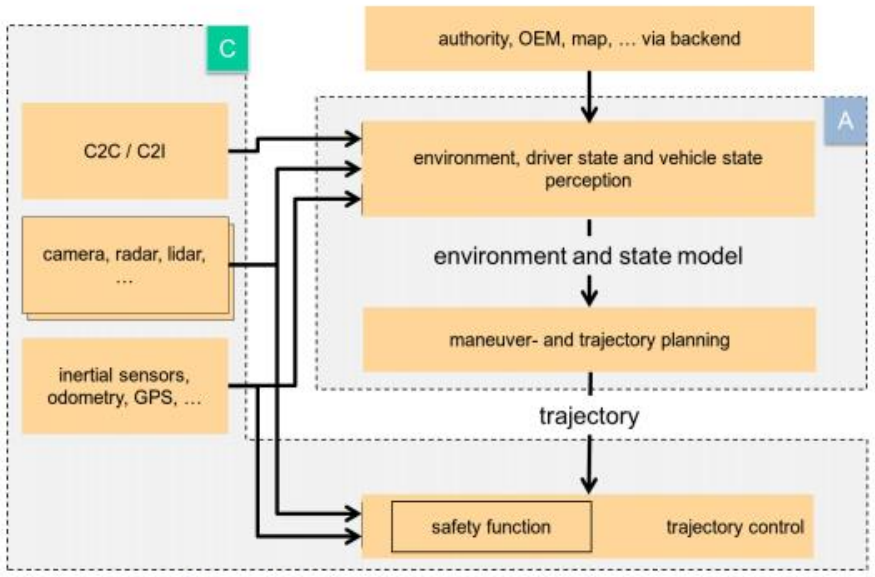
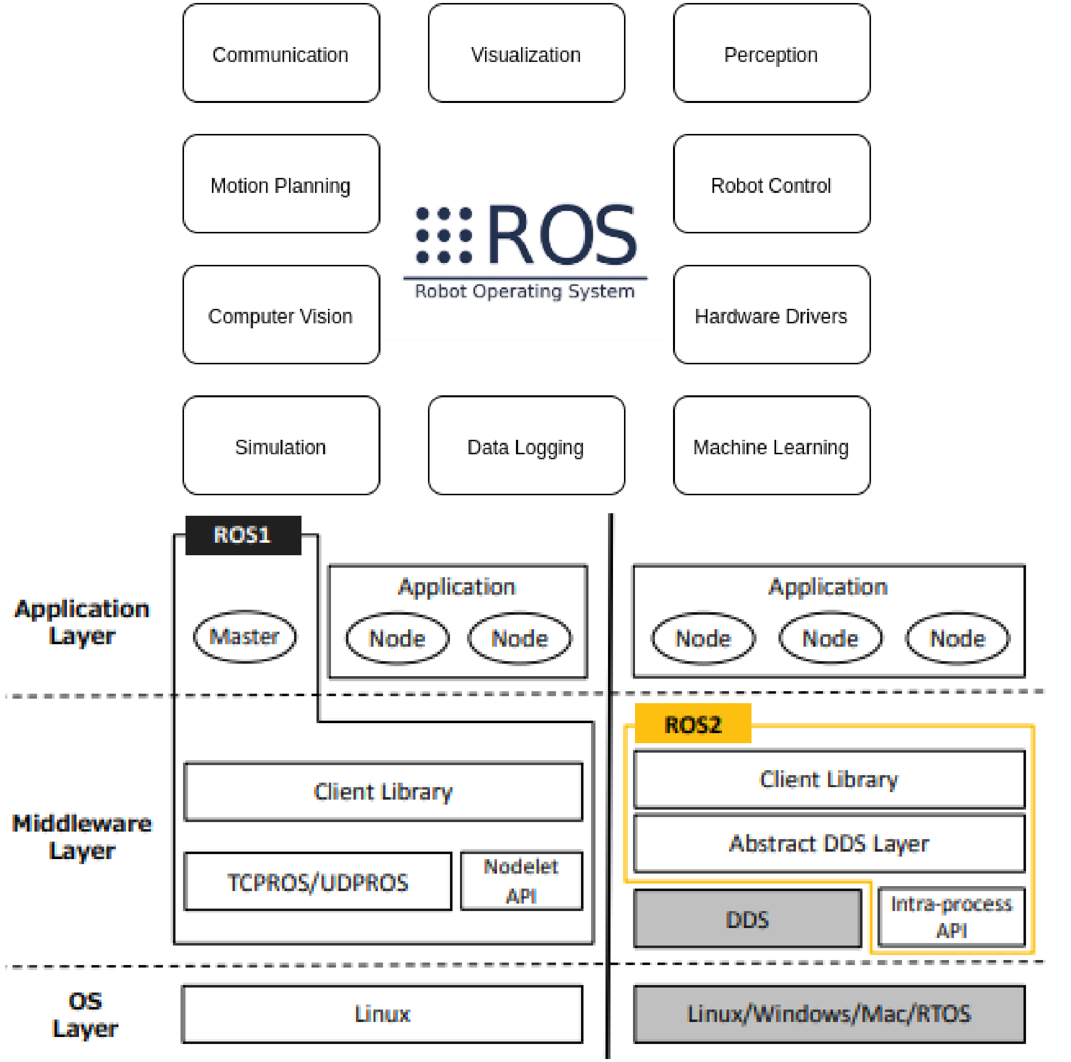


Figure 2-2 Exemplary interactions of AP and CP

Robot Operating System (ROS)

- ROS is a set of software libraries and tools for building robotic applications. Many companies use ROS to develop AVs. It uses the publish-subscribe paradigm for inter-node communication.
  - ROS has a Master node that provides naming and registration services to the rest of the nodes.
  - ROS 2 removed the Master node, and uses publish-subscribe middleware DDS (Data Distribution Service)
- Since ROS uses Linux as the underlying operating system, it is difficult to pass high-level of safety certification



Maruyama Y, Kato S, Azumi T. Exploring the performance of ROS2[C], Proceedings of the 13th International Conference on Embedded Software. 2016: 1-10.

Apex.ai

- "Safe and certified software framework for autonomous mobility systems."
- Certified as a Safety Element out of Context (SEooC) up to Automotive Safety Integrity Level (ASIL)-D
    - Choice among multiple Real-Time Operating Systems (RTOS)

Applications	All mobility applications, including ADAS, automated driving, powertrain, telematics.
Software framework	Apex.Grace
Data transport	**Apex.Ida**
Real-time operating system	QNX, GHS INTEGRITY, eSOL eMCOS, SYSGO PikeOS, Linux RT
ECU	Freedom of choice of SoC, supports ARM, x86, GPU, …

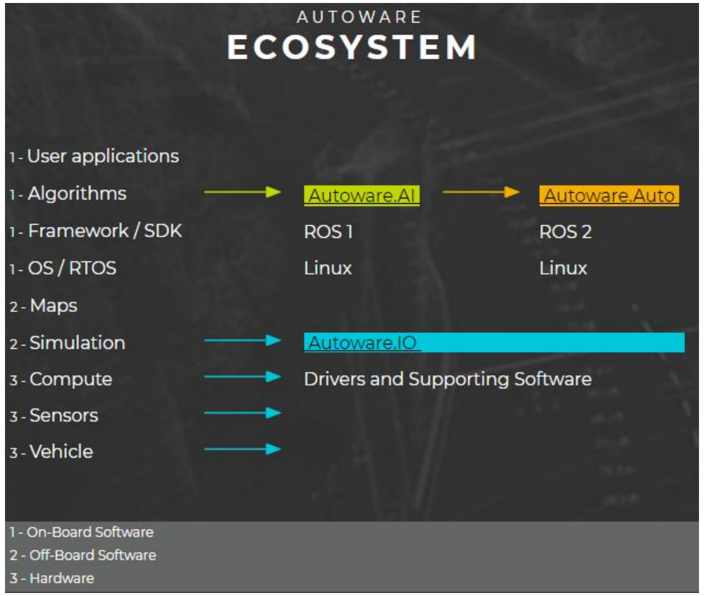
https://www.apex.ai/apexida

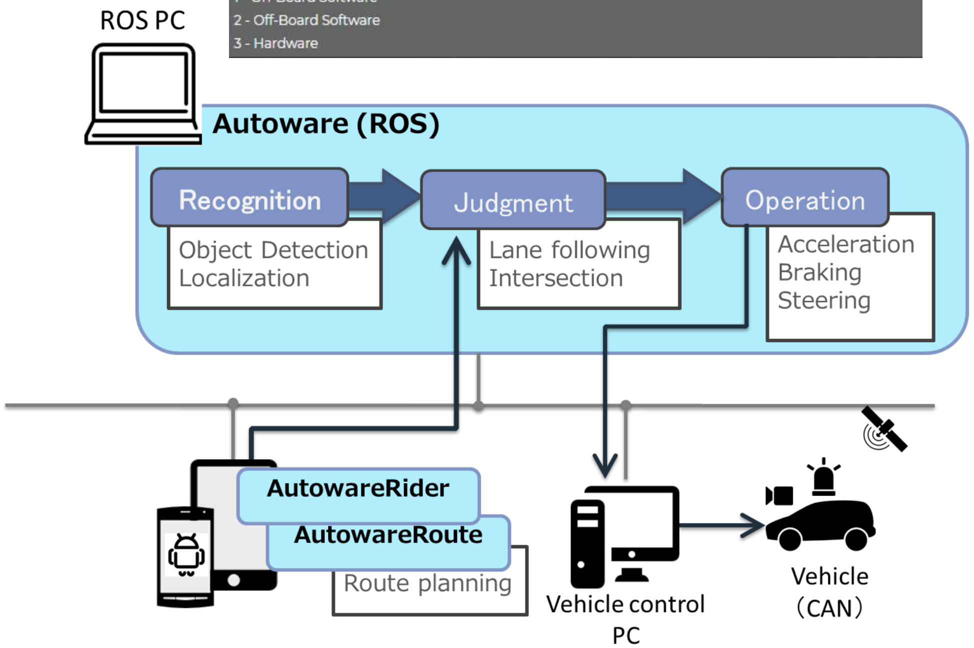
NVIDIA DRIVE Software Framework

- An open-source framework for AD (only for NVIDIA hardware).
  - **DRIVE OS** is a foundational software stack consisting of an embedded Real Time OS (RTOS), hypervisor, CUDA libraries, Tensor RT, and other modules that give you access to the hardware engines.
  - **DriveWorks SDK** enables developers to implement AV solutions by providing a comprehensive library of modules, developer tools, and reference applications.
  - **DRIVE AV** provides perception, mapping, and planning modules that utilize the DriveWorks SDK.
  - **DRIVE IX** provides full cabin interior sensing capabilities needed to enable AI cockpit solution.

Autoware

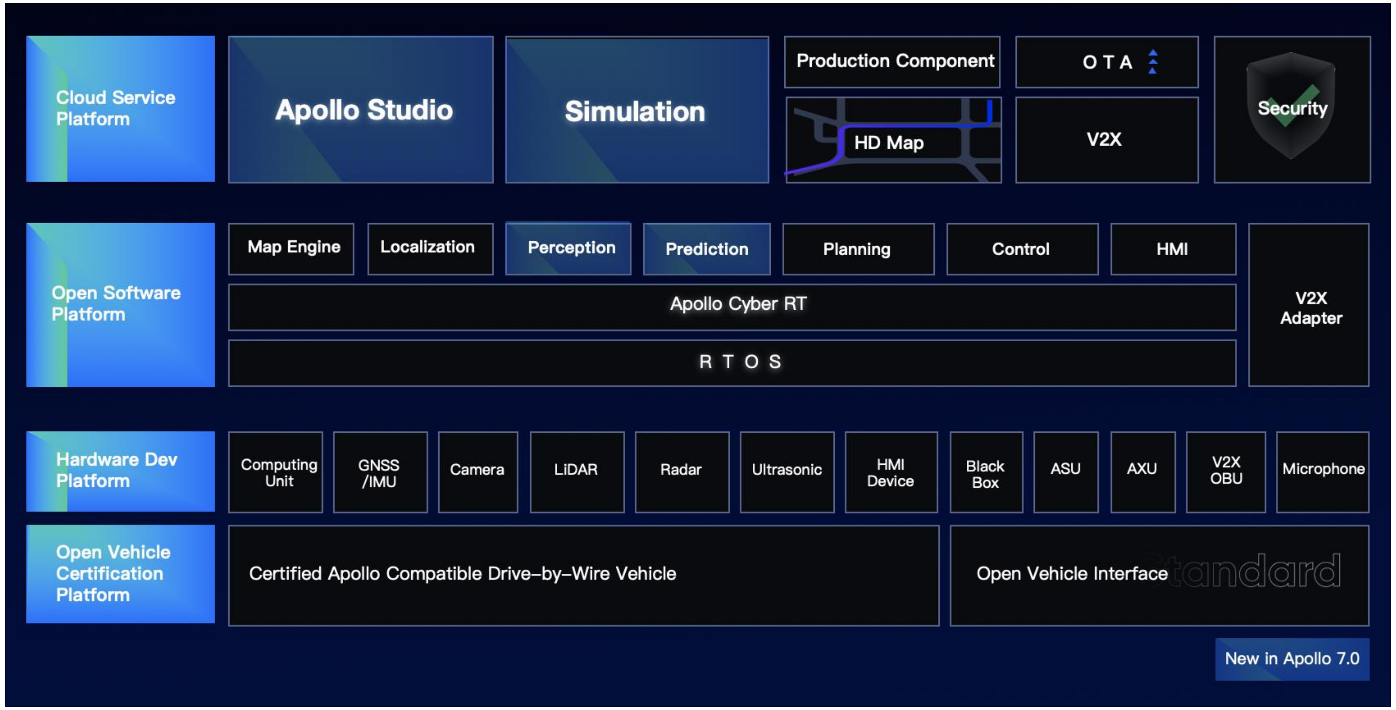
- Open-source AD platform from Japan.
  - Autoware.AI (https://www.autoware.ai) is based on ROS-1
  - Autoware.auto (https://www.autoware.auto) is the new version based on ROS2





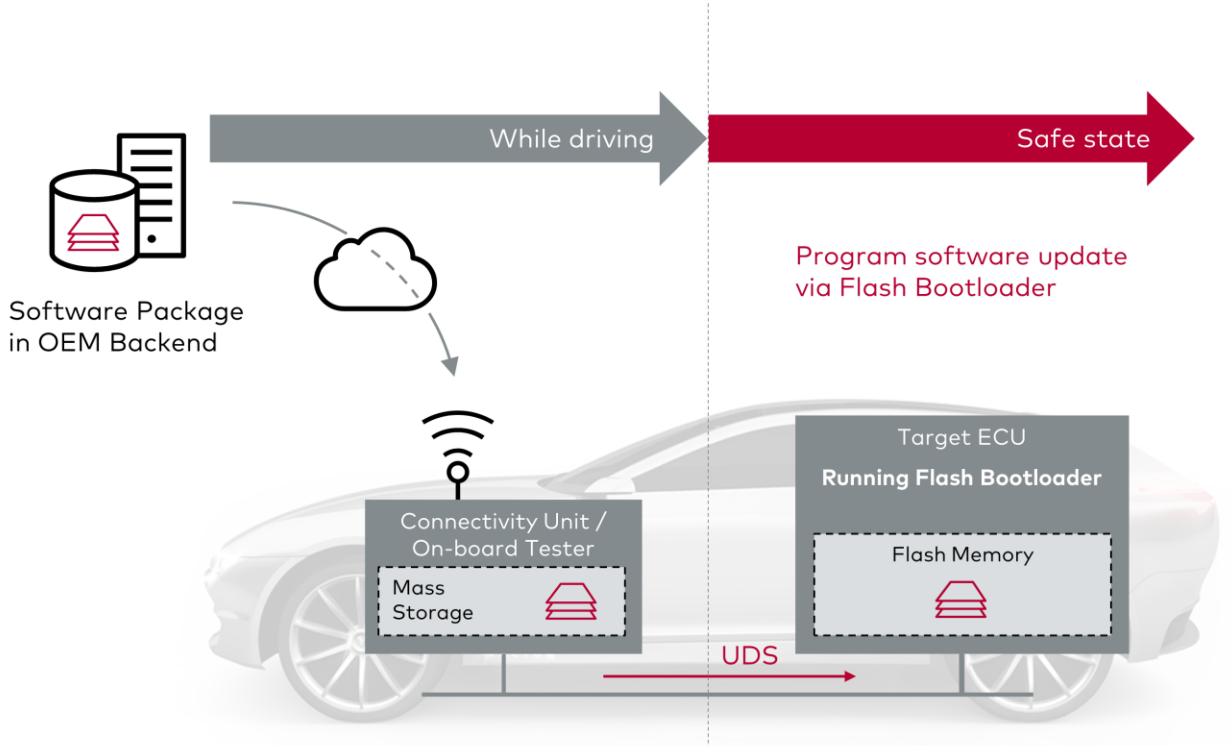
Baidu Apollo

- An open-source, hardware-neutral AD platform from China
- Initially based on ROS, but later replaced ROS with their own components
  - Real-Time OS: Linux kernel with real-time patch
  - Cyber RT: lightweight, high-performance communication middleware



Over-The-Air (OTA) Update w. Connectivity ECU

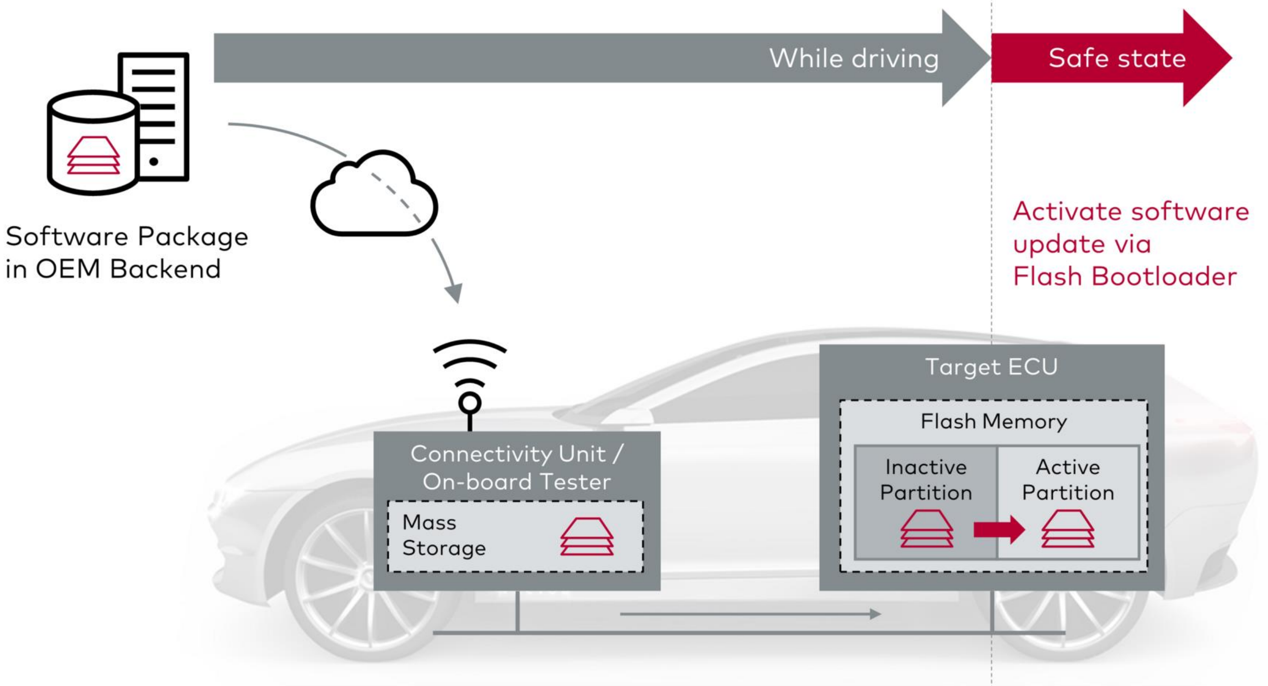
- Flash Bootloaders are small programs used to erase and rewrite the flash memory, for programming an ECU or updating it later in its life cycle
- The new version of software is transferred wirelessly to the vehicle and temporarily stored on a "Connectivity ECU", with sufficiently large memory. When in a safe state, the connectivity ECU starts the update process and loads the software update to the target ECU via a diagnostic sequence - just as the service shop diagnostic tester would do
  - During the update process, the vehicle remains in a safe state and cannot be used
  - The ECUs involved in the update process must be supplied with power. The remaining capacity of the battery limits duration of the update



https://www.vector.com/se/en/products/application-areas/embedded-software/embedded-trends/ota-update-approaches/

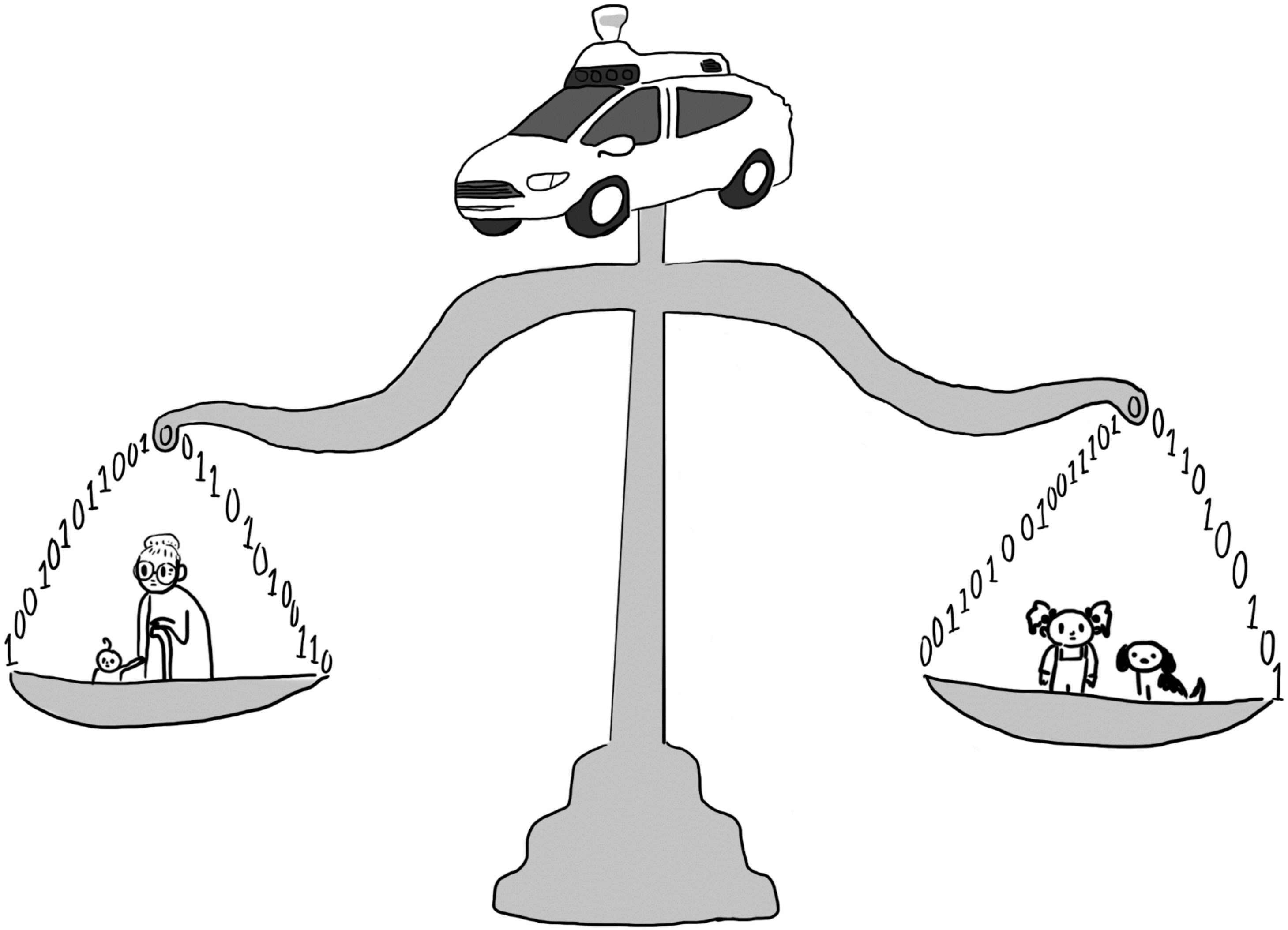
OTA Update w/o Connectivity ECU

- New version of software is directly transferred to the target during normal operation, i.e. while the vehicle is in motion, with storage in a memory area separate from the driving application.
- Advantages:
  - No need for transfer of the update from Connectivity ECU to the target ECU in the safe vehicle state; the vehicle remains ready for operation at all times despite software updates
  - Restoring the previous software is possible without further data transmission



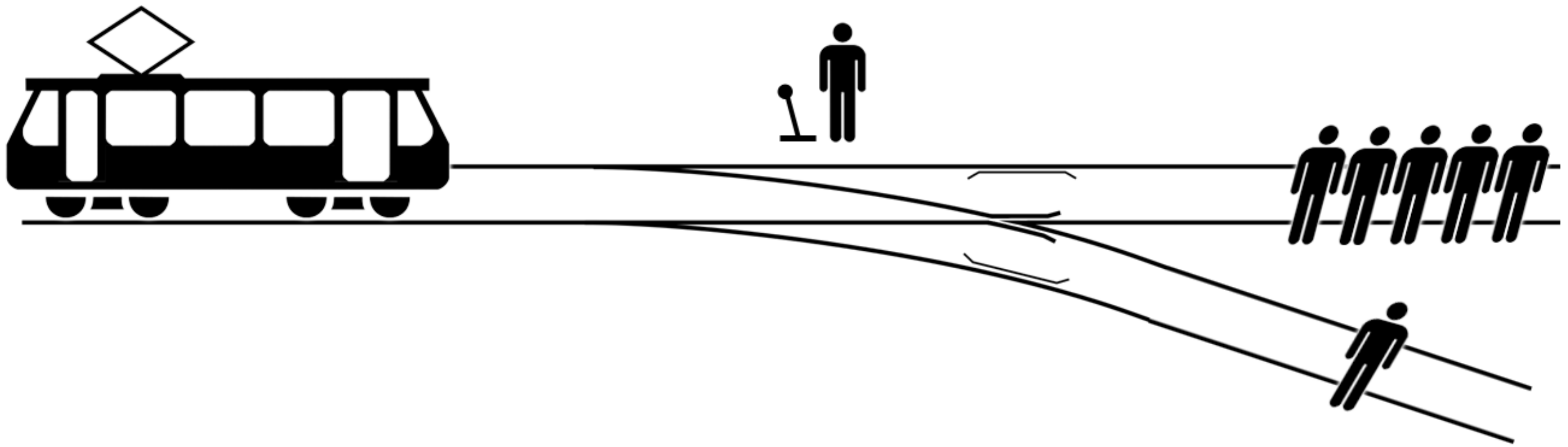
https://www.vector.com/se/en/products/application-areas/embedded-software/embedded-trends/ota-update-approaches/

Ethical Issues



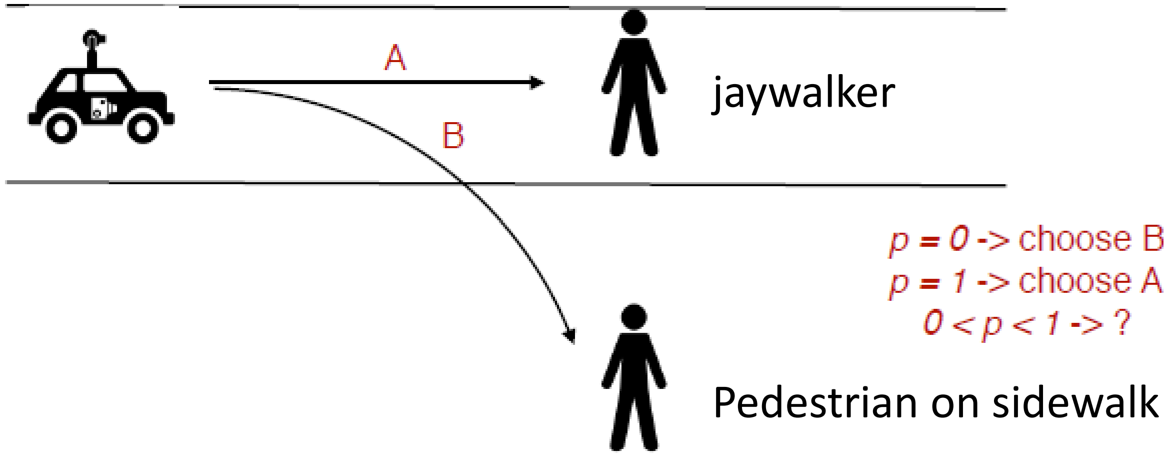
The Trolley Problem

- There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person on the side track. You have two options:
  - Do nothing and allow the trolley to kill the five people on the main track.
  - Pull the lever, diverting the trolley onto the side track where it will kill one person.
- What is the right thing to do?



Variant of Trolley Problem with a Probability Threshold

- You are in a situation where:
  - A. you kill a pedestrian with probability 1, but it's not your fault
  - B. you kill a different pedestrian with probability p, and it is your fault
- What is your threshold value p_{th} for making the choice?
  - if$(p \geq p_{th})$ choose A; otherwise choose B

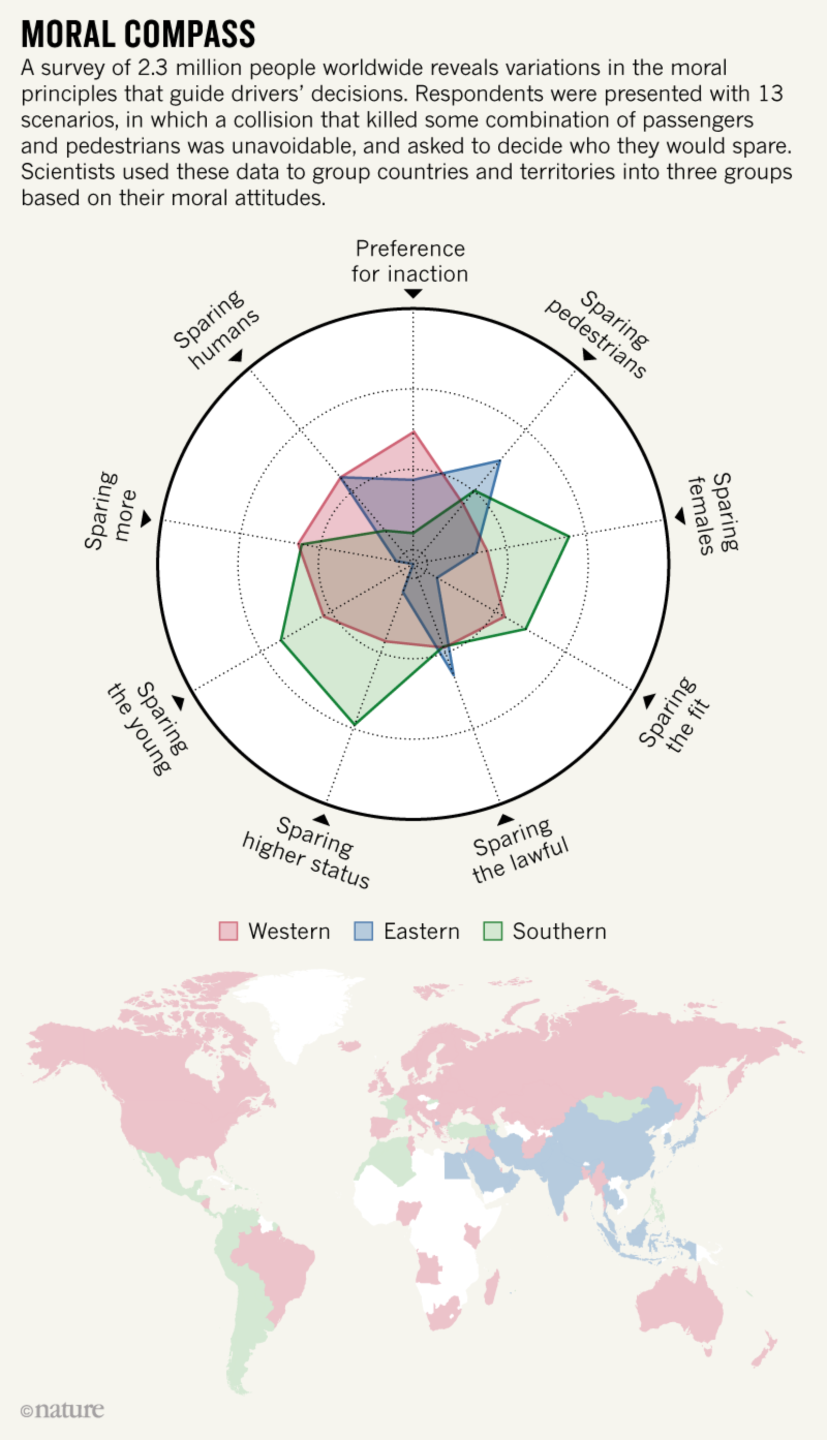


MIT Moral Machine Experiment

- A 2016 survey indicates that people wanted an autonomous vehicle to protect pedestrians even if it meant sacrificing its passengers — but also that they wouldn't buy self-driving vehicles programmed to act this way. This prompted the MIT Moral Machine Experiment, a platform for gathering a human perspective on moral decisions made by machine intelligence (http://moralmachine.mit.edu/)
  - One scenario: an AV must choose between killing two passengers or five pedestrians. An AV experiences a sudden brake failure. Staying on course would result in the death of two elderly men and an elderly woman who are crossing on a 'do not cross' signal (left). Swerving would result in the death of three passengers: an adult man, an adult woman, and a boy (right)
  - You can also design other scenarios. Accident scenarios are generated with nine factors: sparing humans (versus pets), staying on course (versus swerving), sparing passengers (versus pedestrians), sparing more lives (versus fewer lives), sparing men (versus women), sparing the young (versus the elderly), sparing pedestrians who cross legally (versus jaywalking), sparing the fit (versus the less fit), and sparing those with higher social status (versus lower social status)
- This platform gathered 40 million decisions in ten languages from millions of people in 233 countries
- Moral Machines: How culture changes values
  - https://www.youtube.com/watch?v=jPo6bby-Fcg

Cultural Clusters

- Three large clusters
  - Western: Protestant, Catholic, and Orthodox countries in Europe and North America
  - Eastern: Islamic and Confucian (Asian) cultures
  - Southern: Central and South America, as well as France and former French colonies
- The preference to spare younger characters rather than older characters is much less pronounced for countries in the Eastern cluster, and much higher for countries in the Southern cluster.
- The same is true for the preference for sparing higher status characters.
- Countries in the Southern cluster exhibit a much weaker preference for sparing humans over pets, compared to the other two clusters.
- Only the (weak) preference for sparing pedestrians over passengers and the (moderate) preference for sparing the lawful over the unlawful appear to be



AV Ethical Issues: is it Worth the Time?

- Many argue that ethical issues are just a distraction from the real problem of AV safety and security, esp. in the presence of ML/DL algorithms.
  - None of the AV accidents in recent years involved any ethical decisions similar to the Trolley Problem. They are due to failures in sensors or perception algorithms.
- Sebastian Thrun (former head of Google's SDC project, former professor at Stanford who led the development of Stanley, winner of DARPA Grant Challenge in 2005):
  - "I think it's a great thing for philosophers to discuss these kind of problems. They can get tenure at their universities, but it's not of practical relevance. If we manage with certain car technology to halve the traffic deaths in the world, which means if we are able to have 500,000 fewer deaths in total, then for this extremely rare, purely hypothetical trolley problem that might occur once in a hundred years. I think whatever the outcome is, the mental energy that philosophers have spent on discussing it is completely out of proportion to the benefit of others on one problem. I will leave it at that."

AV Testing Legislation (USA)

- It is necessary to test AVs on public roads for technology development, but is it ethical?
- Legislation regulating AV testing differs widely across states. Several states have no proposed legislation, meanwhile states like Nevada, California, Texas, and Arizona are hotbeds for testing AVs (partly due to sunny weather)