Self-Driving Industry

# Technology & the Market Trends
# 技術と市場動向

ムハマド　ムルサリーン



1

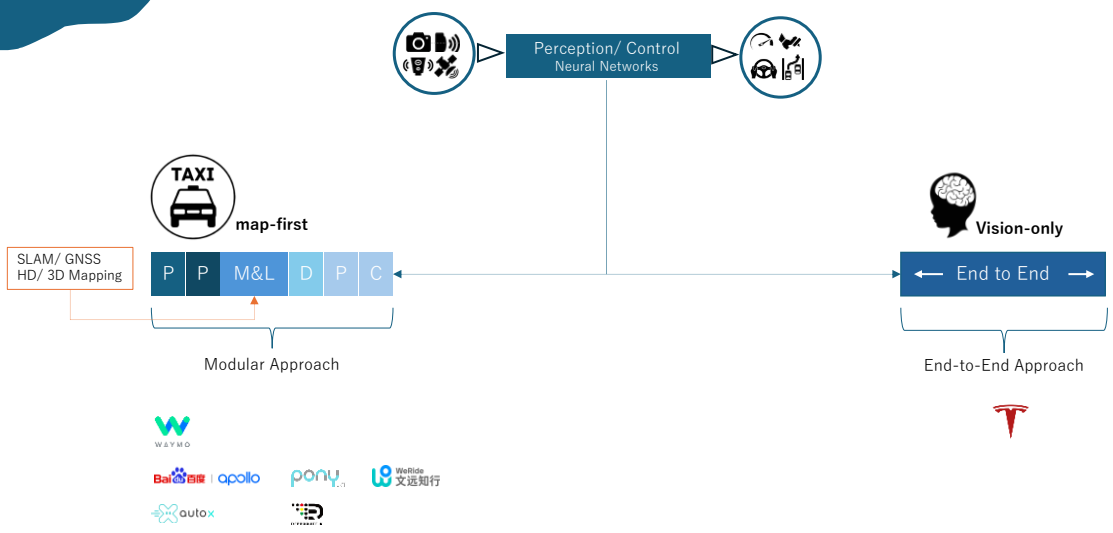
## Agenda

1. Current Landscape
2. Modular Approach (Waymo, ……)
3. End-to-End Approach (Tesla)
4. China Robotaxis
5. Waymo Vs China Robotaxis
6. Future Tech (LLMs, VLMs)
7. Lessons for L4 in Japan



2

## Industry LandScape




3

3

# Waymo

4

4

## Cost Elements
Modular **Vs** E-E

P P M L D P C

Market: Robotaxi
≈ ~5,000 on roads

End to End

Market: End Consumer
~ 5–6 million cars on roads

| | | Sensor Devices | Compute | OEM | Sub Total | 3D Mapping | Example Cost/ Robo |
|---|---|---|---|---|---|---|---|
| | | | | | | | SF ~ 200 sq.km<br>~ 800 Robos |
| Module Base<br>Waymo | 2015 | 122k | 2@250k | 25k | ~ 650k | 2M/sq.km | ~ 500k |
| | 2025 | 13.5k | 2@20k | 75k | ~ 130k | 40k/sq.km | ~ 100k |
| | | | | | | | |
| End-to-End<br>Tesla | 2015 | 7k | 1k | 50 | ~ 60k | ~ | |
| | 2025 | 2.4k | 4k | 50 | ~ 60k | ~ | |

Costs are near approximations ~ +- 20%
Only good for cost comparison

3D_Mapping_Market
3D_Mapping





## Tech Stack
Modular Approach

### Stacked Pipeline

LiDARS
Radars
Cameras
Sonics

GPU/
AI accelerator optimized for CNN/Transformer vision

**Perception**
Object Detection/ Segmentation
**CNN + Transformers**
Real-time (ms)

CPU/GPU combination; sometimes a dedicated AI core

**Prediction**
trajectory prediction
**RNN+GNN+LSTM +Transformers**
Real-time (ms)

FPGA/ fusion processor

**Sensor Fusion**
Combine
**LiDAR, radar, cameras, GPS**
into unified scene

**3D Mapping**
Pre-Built HD 3D-Maps
**Offline LiDAR SLAM**
Most expansive task

Sensor data (LiDAR, GPS) to HD maps

dedicated localization CPU

**Localization**
Localize Vehicle on 3D Env
**GNSS + IMU + Gyro**
LiDAR/ Camera matching/ EKF/ Particle filter/
Real-time

Central CPU (multi-threaded real-time planner)

**Decision & Decision**
**Hierarchical FSM, Rule-based RL + Deep RL**
Real-time

Real-time microcontroller or FPGA

**Path Finding**
Generate Safe Trajectory
**A* / D*, RRT***
trajectory optimization
Real-time

### Issues

- Integration Complexity
- High Cost
- Zero scalability
- Poor Generalization
- Limited Adaptability

FPGA: Field-Programmable Gate Array

## 3D Mapping

**Waymo's 3D Mapping Process**
Pre-built centimeter-accurate digital twins
Static HD data stored on disk

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Data Acquisitio | SLAM Geometric Reconstruction | Sementic Annotations | Map Optimization Quality Check | Deployment Contineous Update |
| Mapping cars<br>Collect raw sensor data<br>LiDAR, HR Cameras, GPS, IMU | Point clouds merge with Cameras images to construct 3D world<br>SLAM | AI/ Human Label data<br>Deep Lerning<br>Manual Verification | AI/ Human Label data<br>Smooth merged data<br>Remove dynamic objects | Maps Compressed, stored and distributed<br>Cloud Service<br>OTA map sync |

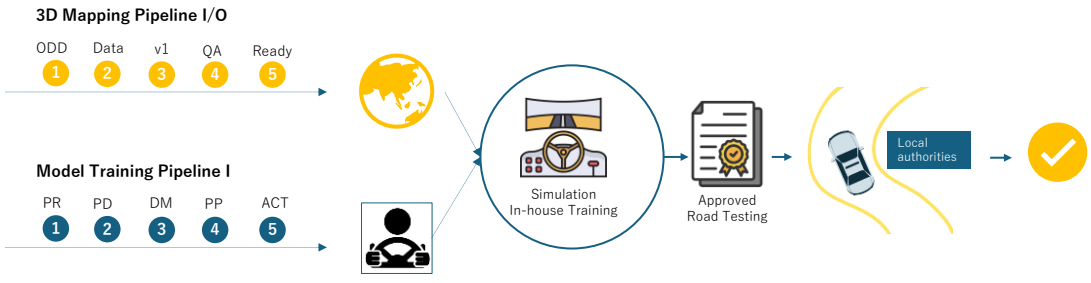
**Google Maps Vs Waymo's 3D Map**

| | Resolution | Purpose | Update | Data Type | Input | Size |
|---|---|---|---|---|---|---|
| Google Maps<br>Waymo HD Maps | Meters<br>Centimeters | Human readable<br>Robotics<br>Machine readable | Periodic<br>Contineous | 2D<br>3D Scan<br>Semntic Layer | Aerial/ Satelite<br>GL-LiDAR/ Camera | MBs/ City<br>TBs/ City |

7

7

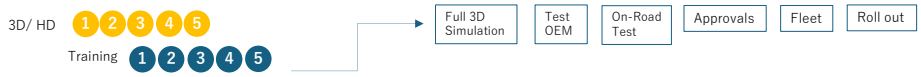
## Complete
Modular Pipeline




8

8

## Robotaxi cost
Tentative Estimate




**Case Study**

**City ABC**
Urban Stretch: 100 sq. km
Road Density: 15 km/ sq. km (1500 km. Road Length)
3D Mapping Cost @ US$5K/ sq.km
Planned fleet size: 100 Robotaxis
Full model-training pipeline cost: US$ 5 billion
Adaptibility training cost: 10% original stack
In-House Simulation Cost on Full 3D Maps: US$ 10 million
Insurance Pledge/ Robo: 5 million

**Ballpark Figure**
deploy-ready robotaxi
(vehicle + mapping + training + licenses + sensors)

= **$250k – $400k**
= (約3,750万円 ~約6,300万円)

橋本尚久 11/19 19:42
How about the cost so far about each robotaxi

- HD Mapping Costs
- 3D Mapping Costs
- Waymo Driver Cost
- Cruise Driver Cost

9

9

# Tesla

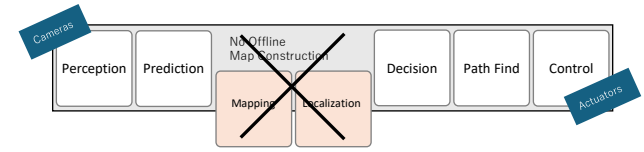
10

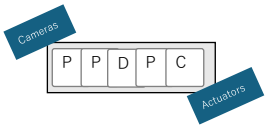
10

## Tech Stack
E2E Approach

### Single Stack architecture

**One unified hardware module**
Tesla FSD Computer (Hardware 3 or 4)

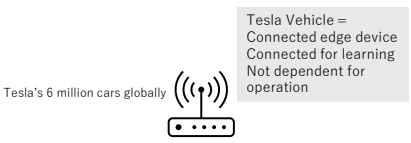


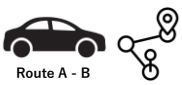

11

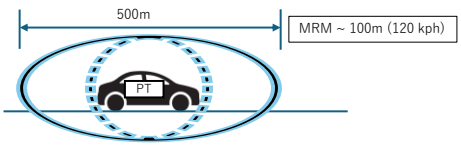
11

## Tech Stack
E2E Approach






**Drive = Cameras + PT**

Cameras handle everything precise:

- lane lines
- road edges
- curbs
- cars
- pedestrians
- traffic lights
- road curvature
- merging lanes
- speed bumps

- detect the correct lane
- center the car inside the lane
- keep distance to edges
- avoid obstacles
- follow curves
- detect intersections
- plan motion

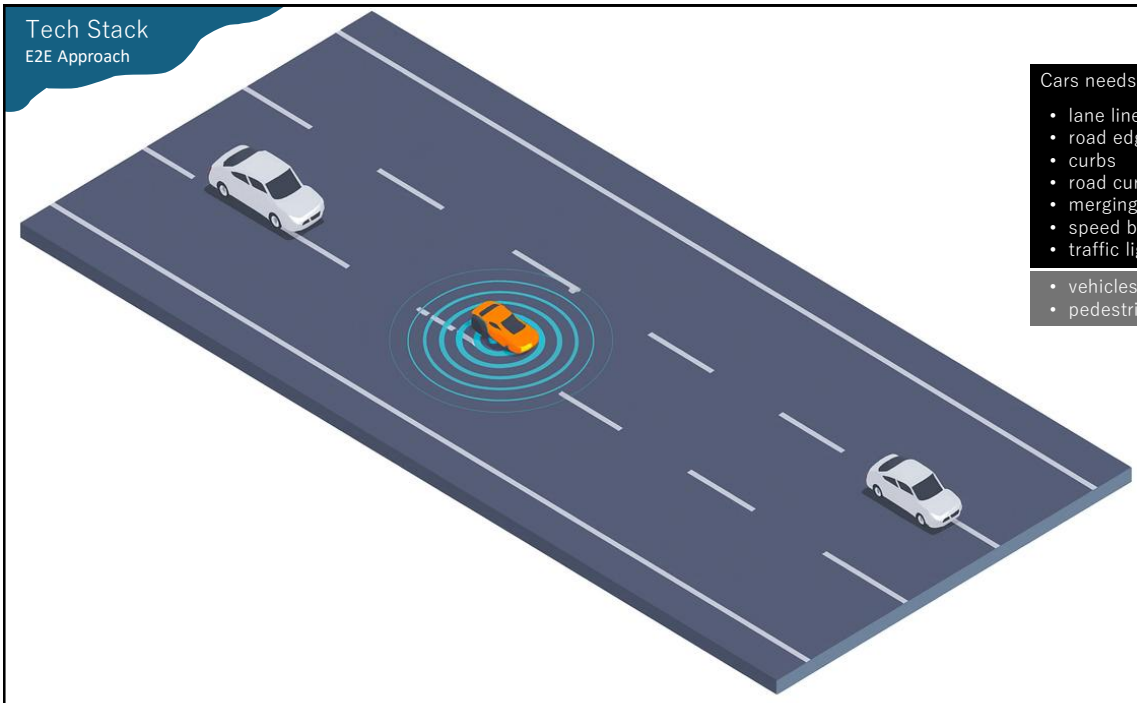
**Edge cases**

- New features
- Driver intervention
- Hidden passenger
- Acute geometery
- Acute weather
- Neural network disagreement

12

12

## Tech Stack
E2E Approach



Cars needs to know:

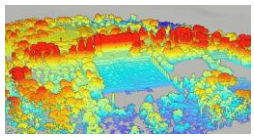
- lane lines
- road edges
- curbs
- road curvature
- merging lanes
- speed bumps
- traffic light

- vehicles
- pedestrians

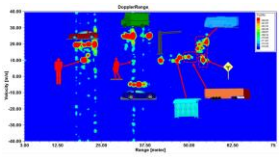
13

## Vision Data Types

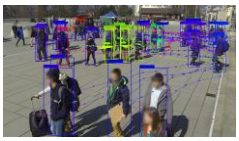
| Sensor | What It Sees / Hears | Data Size (per second) | Processing Load | Typical Range | Strength | Weakness |
|---|---|---|---|---|---|---|
| **LiDAR** | 3D shape & distance (point clouds) | **High** (10–70 MB/s) | **Very High** (3D processing) | 100–300 m | Accurate depth & geometry | Expensive, heavy compute |
| **RADAR** | Distance + speed (low-resolution) | **Very Low** (0.1–1 MB/s) | **Low** | 100–250 m | Works in rain, fog, night | Poor shape details |
| **Camera** | 2D/3D images, textures, color | **Very High** (20–120 MB) | **High** (vision models) | 50–200 m | Rich details, cheap | Sensitive to lighting |
| **Audio** | Sound patterns, screams, alarms | **Very Low** (0.01–0.1 MB/s) | **Low–Medium** | 3–10 m indoors | Excellent for events detection (voice, screams) | No visual info |



LiDAR Point Cloud



RADAR Datasets



Camera Datasets

Sound Waves Datasets

14

14

**Why Cameras?**

- Scaling to Millions: Cameras are the only scalable sensor
- Cameras give *semantic richness* that LiDAR/Radar cannot
- LiDAR and RADAR introduce *sensor contradictions*
  (creates ghost objects, false positives, and frame alignment problem)
- Neural networks love images (Cameras), not point clouds (LiDAR)
- The REAL engine (Dojo): Tesla has the world's largest driving video dataset
  (65 million+ real-world video clips, over 10 billion miles of human driving)

- Vision-only enables global ODD without remapping
- Future-proof: E2E models need consistent visual input
- Camera-only is the only path to global, affordable autonomy

The frontier of deep learning is **vision-language-action models**, not point clouds

**How Issues are addressed?**

- Light Dependance --------- High Dynamic Range (HDR) cameras
- Weather Aspects ---------- Spatiotemporal Vision Transformers (2024–2025)
- Depth ---------- Multi-view Geometric Depth (Stereo-from-motion)
- Depth ---------- Video-based Depth (Monocular depth from motion)
- 3D voxel grid from cameras only -------- 4D Occupancy Network (The real breakthrough)
- Lens Blockage: Dirt, Fog, Water Drops ---------- Tesla Vision never depends on a single frame

15

15

# China
## RoboTaxi

16

16

China
Self-Driving Landscape

## China Robotaxi tech stack Vs US

| Company | Sensors | Philosophy |
|---|---|---|
| US- Mainstream | | |
| Tesla | Cameras only | End-to-end vision; HD map free |
| Waymo | LiDAR + HD maps | Modular/ HD Map |
| China-Maistream | | |
| Xpeng | Cameras + LiDAR + HD | Modular, Moving toward map-free driving |
| Huawei | Cameras + LiDAR + HD | Modular, Citywide map-free autopilot |
| Li Auto | Cameras + LiDAR + HD | Modular, Robust **highway** + city driving |
| Baidu Apollo/Pony.ai/ Wedrive | LiDAR + HD maps | Modular, Robotaxi only |



17

China
Self-Driving Landscape

## Comparative advantages of China Robotaxi Market

| Category | China Advantage | Why It Beats Waymo | China Attributes (Positive) |
|---|---|---|---|
| Infrastructure | uniform roads | U.S. is inconsistent and old | Road Infrastructure is **new and well documented** |
| Regulation | unified, fast approvals | Waymo faces lawsuits & bans | One window Operations, **National Level** |
| Sensors/ Compute | cheap mass-produced LiDAR | Waymo uses expensive custom LiDAR | **Ower Costs**: Economics of scale |
| Compute | domestic supply chain | Waymo depends on costly imports | Local Manufacturing/ Cost efficiency |
| Scale | millions of consumer AD cars | Waymo only has thousands of robotaxis | **Data** Abundancy |
| Maps | map-optional driving | Waymo stuck on HD maps | **National HD** maps available for upgrade |
| Urban training | richer traffic diversity | Waymo trains in limited cities | Public awareness/ **adaptability** (**National Element**) |
| V2X support | built into cities | Waymo has almost none | **Urban V2X** infra, 5G/6G incorporation |
| National Policy | A national product | A private product image | Targetted research and development |



18

Future
2025-2035

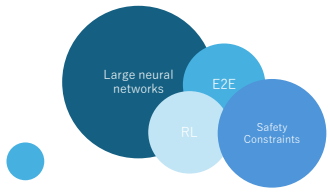
## The future tech stack of autonomy (2025–2035)

Modular E2E LLM/ LVM

Modular

**The modular pipeline has problems:**

- too many hand-engineered rules
- brittle interactions
- difficult to tune at scale
- catastrophic edge-case explosion




**LNN**

- Tesla "FSD v12"
- Wayve in UK
- Huawei ADS 3.0 (announced)
- Xpeng XNGP 3.0 prototype

LLM/ VLA

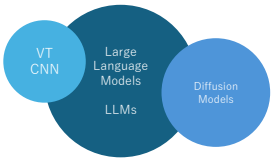
**Vision-Language-Action (VLA) models**

A single model that:
- understands raw video
- reasons in language
- plans actions

Like a "driving GPT" that sees and reasons

**These models combine:**

- LLMs (reasoning)
- Vision transformers (perception)
- Trajectory diffusion models (planning)




- Tesla (end-to-end world model in development)
- Wayve (VLA model with GPT-like backbone)
- DeepMind's Gato → robotics/driving hybrids
- Baidu Apollo's next-gen research
- Huawei "Super World Model" (announced for 2026)

19

19

Future
2025-2035

## The future Sensor stack

**New Sensor Paradigms Beyond LiDAR & Cameras**

**Event cameras**
- microsecond reaction time
- ideal for fast highway scenarios
- used by Huawei research already

**Thermal imaging**
- detects humans at night
- sees through fog/smoke
- becoming standard in China

**Radar imaging (4D radar neural imaging)**
- China is investing heavily
- returns 3D "voxel maps" like LiDAR

**Neural LiDAR Fusion**
- Future LiDAR systems
- output learned features
- not point-clouds
- integrate directly with neural backbones

20

20

## What's Next

**1 map-first/ modular approach**

1. Create a 3D limited universe
2. Train the model for that region offline in Labs
3. Fit-in trained model in car
4. Good to Go.

1. Cost
2. Scalability
3. Safety

**Waymo**

**LLM** (driving reasoning + multimodal fusion)
(LiDAR and HD maps for reliability)

**EMMA**
End-to-End Multimodal Model for Autonomous Driving

Built on top of Gemini (Google's multimodal LLM)
Combines vision + text + reasoning
Produces perception, prediction, and planning outputs

1. Cost
2. Scalability
3. Safety

**2 E/E Approach**

1. Create a basic 2D Map (Navigation)
2. Car is Equipped with AI Single Module
3. The system will learn in field
4. Model updated in "Dojo"
5. Updates are synced OTA

1. Cost
2. Scalability
3. Safety

**Tesla**

**End-to-End** (World Model + Large Vision Transformers)

Tesla does not use LLMs, but instead uses:

FSD v12 / v13
Fully end-to-end neural policy
**Massive video transformers**
No HD maps
Real-time 3D world-modeling
Trained with fleet learning (scale unmatched)

1. Cost
2. Scalability
3. Safety

21

21

## What's Next
## China

**3 map-first modular Vs Waymo US**

1. Create a 3D limited universe
2. Train the model for that region offline in Labs
3. Fit-in trained model in car
4. Good to Go.

1. Cost
2. Scalability
3. Safety

**China**

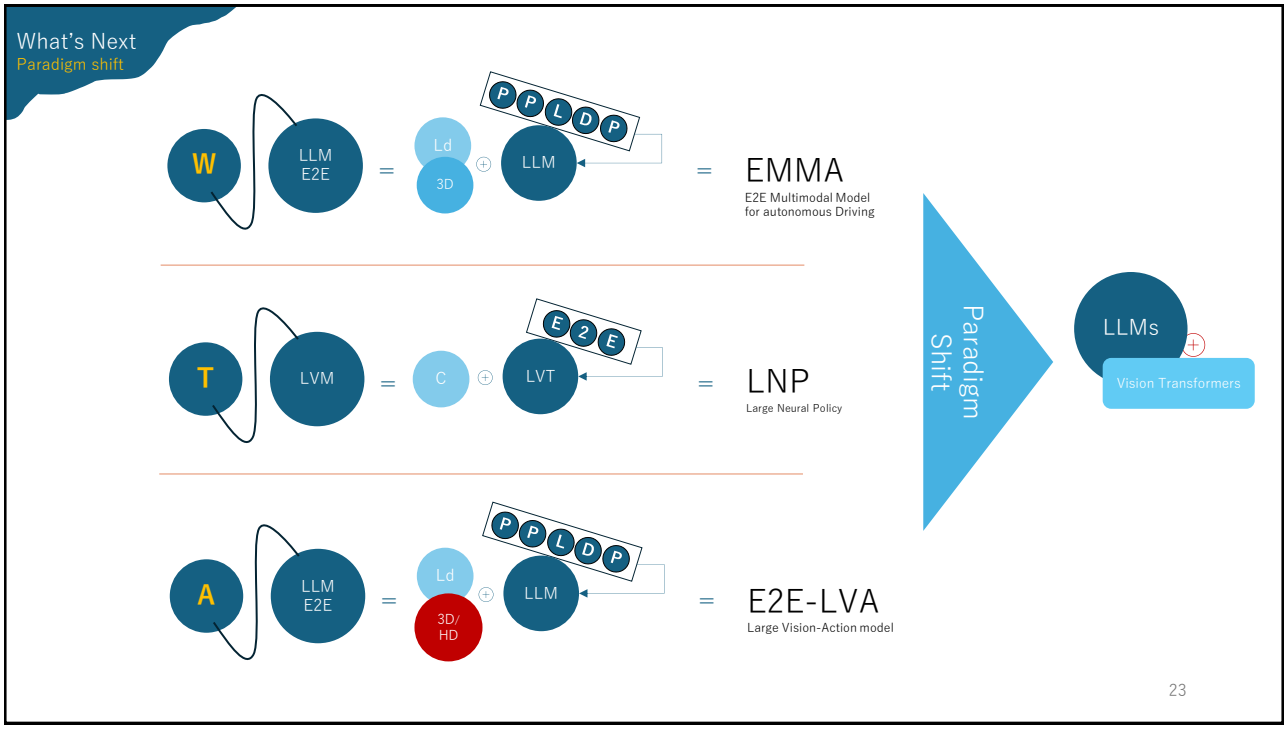
**Hybrid Approach**
HD maps + end-to-end vision + LLMs
(LiDAR and HD maps for reliability)

Combine the strengths of both Tesla and Waymo
Rapid scaling

1. Cost
2. Scalability
3. Safety

22

22


What's Next
Paradigm shift
W
LLM
E2E
=
Ld
3D
P P L D P
LLM
=
EMMA
E2E Multimodal Model
for autonomous Driving
T
LVM
=
C
E 2 E
LVT
=
LNP
Large Neural Policy
A
LLM
E2E
=
Ld
3D/
HD
P P L D P
LLM
=
E2E-LVA
Large Vision-Action model
Paradigm
Shift
LLMs
Vision Transformers
23

23

24

| Phase | Activities | Costs (Millions US$) | Time (Months) |
|---|---|---|---|
| **Planning** | Permissions, ODD,<br>Mapping Settup, Simulations<br>Insurance Pledge 5.0 | 0.3 ~ 0.50 | 1 - 3 |
| **Fleet Caliberaion** | Mapping Vehicles: 10-20<br>LiDAR 360, Camera, IMU, RTK-GPS,<br>Calibration runs | 10 ~ 20 | 3 - 6 |
| **SLAM + HD map v1** | Large-Scale HD Data Collection<br>Multi-pass/ Multi-weather, day/night, Feature positioning,<br>GPS + SLAM data, Raw LiDAR point clouds, Multi-camera video | 2 ~ 4 | 6 - 18 |
| **HD Map v1 Construction** | SLAM processing: stitching 3D environment<br>Multi-pass alignment: removing moving objects<br>Auto-segmentation: curbs, traffic lights, lanes, Semantic labeling<br>QA pipeline | 2 ~ 5 | 6 - 18 |
| **Safety Model Refinement**<br>**DMV Driverless Permit** | Prepare DMV "Driverless Testing" application<br>Build Teleoperation center, Train remote assistance team | 1.5–3 | 0.5–1 |
| Driverless testing | | 0.5–1 | 0.5–1.5 |
| DMV + CPUC commercial | | 0.8–2.2 | — |
| Fleet rollout | | — | 20–70 |





## Tech Stack
E2E Approach

### Single Stack architecture

Tesla Vehicle = Connected edge device
Connected for learning
Not dependent for operation

| Feature | Modular Systems (Waymo, Cruise, etc.) | Tesla End-to-End |
|---|---|---|
| Compute Units | Multiple (GPU, CPU, FPGA, IMU processor, MCU) | One main FSD computer (dual redundant chips) |
| Architecture | Distributed, message-passing | Centralized neural inference engine |
| Data Flow | Sequential (sensor → perception → planner) | Parallel (multi-camera → unified network) |
| Redundancy | Separate modules for each subsystem | Dual-chip redundancy inside one board |
| Sensors | LiDAR + radar + cameras | Cameras only |

FPGA: Field Programmable gate array
MCU: Microprocessor Unit

## Tech Stack
E2E Pipeline

- 8/12 cameras (forward, rear, sides)
- Radar (older models only)
- Ultrasonic sensors (legacy models)
- GPS + IMU
- Wheel, steering, motor, and brake sensors


**FSD Computer (Hardware 3 & 4)**
A custom, automotive-grade AI supercomputer in the dashboard

**Main components**
- 2 custom Tesla neural network accelerators (NNA chips)
- Redundant CPU clusters
- GPU-like logic for image processing
- Power management + safety microcontrollers

- Parallel processing
- Real-time neural inference
- Redundancy for safety (ASIL-D philosophies)

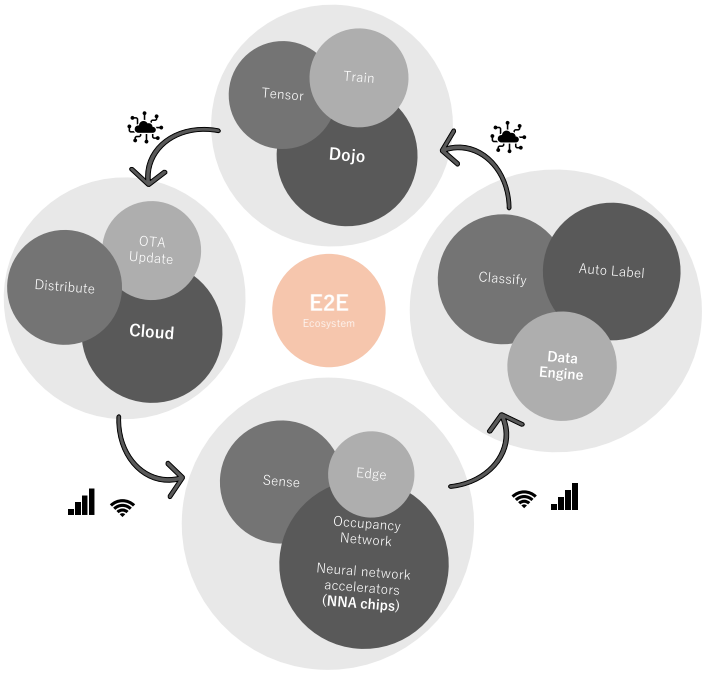
**Logic/ Functions**
- Computer vision
- Occupancy flow (neural fields)
- Transformational neural networks (multi-camera fusion)
- Behavior cloning
- End-to-end planning
- Reinforcement learning (small part)

**Local Data Buffer + Uploader**
The car stores:
- short video clips (last 10–30 sec)
- metadata (speed, NN confidence, planner decisions)
- event triggers (interventions, jerkiness, planner disagreement)
- These clips are uploaded later to Tesla servers (not to Dojo directly).

27

27

## Tesla
E2E Ecosystem




28

28